IMPLEMENTATION OF WRITE ASSIST TECHNIQUE ON LOW VOLTAGE DISTRIBUTED MEMORY

By

CHAN GAIK MING

A Dissertation submitted for partial fulfilment of the requirement for the degree of Master of Science (Microelectronic Engineering)

August 2016

ACKNOWLEDGEMENTS

First of all, I would like to express the deepest appreciation to my research supervisor, Professor Dr. Nor Ashidi Mat Isa for his guidance and advices had given to me throughout this research study. I am thankful to be under his supervision as a lot of proper research skills have been gained through his constructive comments given to me. Besides, his continuous guidance on research writing has also helped me in gaining proper technical writing skill and completing this research writing within the time frame. Without his supervision, this research would not have been possible.

Secondly, I would like to thank my family members for their spiritual support and warm encouragement given to me through the difficult time of this research study. Their support and encouragement have become my willpower to complete this research study.

Lastly, I would like to thank my friends who have helped me in completing this research study and writing. Their knowledge and experience sharing have helped me in speeding up my learning curve during this research study.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS ii

TABLE OF CONTENTS iii

LIST OF TABLES vi

LIST OF FIGURES vii

LIST OF ABBREVIATIONS xi

ABSTRAK xiii

ABSTRACT xv

CHAPTER 1 INTRODUCTION 1

1.1 Background 1

1.2 Current Trend in SRAM Write Technology 2

1.3 Problem Statements 3

1.4 Research Objectives 4

1.5 Research Scope 4

1.6 Thesis Outline 5

CHAPTER 2 LITERATURE REVIEW 6

2.1 Introduction 6

2.2 Architecture of Memory design 6

2.3 Memory Array Organization 7

2.4 SRAM Cell Structure and Operations 10

2.5 SRAM Write Failure and Write Margin 17

2.6 SRAM Write Assist Techniques 20

2.6.1 Boosted Word-line Voltage Write Assist Technique 21

2.6.2 Negative Bit-line Voltage Write Assist Technique 22

2.6.3 Reduced Cell VDD Write Assist Technique 23

2.6.4 Raised Cell VSS Write Assist Technique 24

2.7 SRAM Power Consumption 25

2.7.1 Static Power 25

2.7.2 Dynamic Power 28

2.8 Conclusion 29

CHAPTER 3 METHODOLOGY 30

3.1 Introduction 30

3.2 Conventional Boosted Word-Line Write Assist Technique 30

3.3 Proposed Boosted Word-Line Write Assist Technique 31

3.4 Design Architecture 35

3.5 Write Performance Analysis 37

3.6 Power Consumption Analysis 43

3.6.1 Static Power Consumption Analysis 44

3.6.2 Dynamic Power Consumption Analysis 47

3.6.3 Total Power Consumption Analysis 49

3.7 Design Simulation Condition 54

3.8 Conclusion 56

CHAPTER 4 RESULTS AND DISCUSSION **57**

4.1 Introduction 57

4.2 Schematic Design 57

4.3 Simulation Results of Write Performance 58

4.3.1 Identification of the Worst Case Corner of Write Access Time 60

4.3.2 Determination of Boosted Gate Voltage 61

4.3.3 Verification of Write Performance 63

4.3.4 Write Performance Scaling with Supply Voltage 68

4.3.5 Write Performance in Monte-Carlo Simulation 69

4.4 Simulation Results of Static Power Consumption 71

4.5 Simulation Results of Dynamic Power Consumption 73

4.6 Simulation Results of Total Power Consumption 75

4.7 Overall Result Comparison between the Conventional and Proposed Techniques 81

4.8 Conclusion 82

CHAPTER 5 CONCLUSION **84**

5.1 Conclusion 84

5.2 Future works 86

REFERENCES **88**

APPENDIX A **91**

LIST OF TABLES

Table 3.1 Relationship between Vth, Vgs and overdrive voltage 33

Table 3.2 Boosted word-line voltage evaluating condition..... 35

Table 3.3 Truth table of write address decoder 37

Table 3.4 Design simulation condition for write performance analysis..... 54

Table 3.5 Design simulation condition for power consumption analysis 55

Table 4.1 Selected vccboost voltage for conventional and the proposed techniques..... 63

Table 4.2 Overall result comparison between conventional and the proposed techniques 82

LIST OF FIGURES

Figure 1.1 Generic FPGA with its embedded elements 1

Figure 2.1 General architecture of memory design 7

Figure 2.2 Typical Random Access Memory array organization 8

Figure 2.3 An example of 256 x 256 SRAM architecture with N=M=8 9

Figure 2.4 Storage element in an SRAM cell 10

Figure 2.5 Basic 6T-SRAM cell 11

Figure 2.6 Voltage Transfer Characteristic of SRAM cell 11

Figure 2.7 Schematic of 6T-SRAM cell 12

Figure 2.8 6T-SRAM cell in read operation 13

Figure 2.9 6T-SRAM cell in write operation 14

Figure 2.10 8T-SRAM cell 17

Figure 2.11 Memory cell storage nodes during write operation (a) Unsuccessful write operation (b) Successful write operation 19

Figure 2.12 Write fail bit count dependence on VDD and word-line pulse width 20

Figure 2.13 Change in WLcrit with voltage scaling at 32nm technology node 20

Figure 2.14 Write assist based on boosted word-line voltage (a) Schematic and waveform (b) Impact of write assist on the WLcrit 22

Figure 2.15 Write assist based on negative bit-line voltage (a) Schematic and waveform (b) Impact of write assist on the WLcrit 23

Figure 2.16 Write assist based on reduced cell VDD (a) Schematic and waveform (b) Impact of write assist on the WLcrit 24

Figure 2.17 Write assist based on raised cell VSS (a) Schematic and waveform (b) Impact of write assist on the WLcrit 25

Figure 2.18 Subthreshold leakage and gate tunneling leakage in a 6T-SRAM cell storing bit "0"..... 27

Figure 2.19 Dual-VT SRAM cell configurations 28

Figure 3.1 8T-SRAM cell for the conventional and proposed techniques (a) Single-VT configuration (b) Dual-VT configuration 32

Figure 3.2 Conventional and proposed write word-line drivers 34

Figure 3.3 SRAM design architecture for evaluation study 36

Figure 3.4 SRAM write access cases for write performance evaluation (a) Write-0 case (b) Write-1 case..... 39

Figure 3.5 Five phases of SRAM write performance analysis 40

Figure 3.6 SRAM cell in inactive condition for static power evaluation (a) Word-line off case 1 (b) Word-line off case 2 (c) Word-line off case 3 (d) Word-line off case 4 46

Figure 3.7 Write address decoder in inactive condition for static power evaluation..... 47

Figure 3.8 SRAM cell in active condition for dynamic power evaluation (a) Write case 1 (b) Write case 2..... 48

Figure 3.9 Write address decoder in active condition for dynamic power evaluation 49

Figure 3.10 Memory design setup for total power evaluation (a) Write case 1 (b) Write case 2 (c) Write case 3 (d) Write case 4 52

Figure 4.1 Schematic hierarchies of the memory design under evaluation 58

Figure 4.2 Memory write access time measurements in simulation waveform view (a) write-0 access time (b) write-1 access time 59

Figure 4.3 SRAM write access time across transistor corners and temperature at vcc = 0.80V and vccboost = 1.00V (conventional technique) 61

Figure 4.4 Boosted voltage effect on SRAM write access time at SS corner, vcc = 0.80V and temperature = -40C (Low-VT write access transistor) 62

Figure 4.5 Boosted voltage effect on SRAM write access time at SS corner, vcc = 0.80V and temperature = -40C (Ultra-Low-VT write access transistor)..... 62

Figure 4.6 SRAM write-0 access time across transistor corners and temperature at vcc = 0.80V (a) Temperature = -40C (b) Temperature = 85C (c) Temperature = 125C 64

Figure 4.7 SRAM write-1 access time across transistor corners and temperature at vcc = 0.80V (a) Temperature = -40C (b) Temperature = 85C (c) Temperature = 125C 65

Figure 4.8 SRAM write access time across transistor corners and temperature at vcc = 0.80V (a) Low-VT write access transistor, vccboost = 0.94V (b) Ultra-Low-VT write access transistor, vccboost = 0.90V..... 67

Figure 4.9 Change in write access time with supply voltage scaling at SS corner and temperature = -40C (a) Write-0 access time (b) Write-1 access time 69

Figure 4.10 Monte-Carlo simulation on write access time for the conventional and proposed techniques at SS corner, vcc = 0.80V, temperature = -40C and 6 sigma variation (a) Standard deviation of write access time (b) Upper limit of write access time 70

Figure 4.11 Static power consumption of SRAM cell across temperature at TT corner and vcc = 0.80V (a) Temperature = -40C (b) Temperature = 85C (c) Temperature = 125C 72

Figure 4.12 Static power consumption of write address decoder across temperature at TT corner and vcc = 0.80V 73

Figure 4.13 Dynamic power consumption of SRAM cell across temperature at TT corner and vcc = 0.80V (a) Temperature = -40C (b) Temperature = 85C (c) Temperature = 125C 74

Figure 4.14 Dynamic power consumption of write address decoder across temperature at TT corner and vcc = 0.80V 75

Figure 4.15 Total static power consumption of memory design under evaluation across temperature at TT corner and vcc = 0.80V (a) Temperature = -40C (b) Temperature = 85C (c) Temperature = 125C 77

Figure 4.16 Total dynamic power consumption of memory design under evaluation across temperature at TT corner and vcc = 0.80V (a) Temperature = -40C (b) Temperature = 85C (c) Temperature = 125C 78

Figure 4.17 Final total power consumption of memory design under evaluation across temperature at TT corner and vcc = 0.80V (a) Temperature = -40C (b) Temperature = 85C (c) Temperature = 125C 80

LIST OF ABBREVIATIONS

6T-SRAM	Six Transistors-Static Random Access Memory
8T-SRAM	Eight Transistors-Static Random Access Memory
CMOS	Complementary Metal-Oxide Semiconductor
Dual-VT SRAM	Dual-Voltage Threshold Static Random Access Memory
DUT	Design Under Test
DRAM	Dynamic Random Access Memory
DSP	Digital Signal Processing
FIFO	First-In, First-Out
FPGA	Field Programmable Gate Array
Low-VT	Low Threshold Voltage
NMOS	Negative-Channel Metal Oxide Semiconductor
PMOS	Positive-Channel Metal Oxide Semiconductor
RAM	Random Access Memory
ROM	Read Only Memory
Single-VT	Single-Voltage Threshold
SPICE	Simulation Program with Integrated Circuit Emphasis
SRAM	Static Random Access Memory
Std-VT	Standard Threshold Voltage
Ultra-Low-VT	Ultra-Low Threshold Voltage
vccboost	Boosted Voltage
VDD	Supply Voltage
Vgs	Gate-to-Source Voltage
Vmin	Minimum Operating Voltage
VSS	Ground Voltage
VTC	Voltage Transfer Characteristic

WL	Word-Line
WLcrit	Critical(Minimum) Width of Word-Line Pulse

IMPLEMENTASI TEKNIK BANTUAN PENULISAN UNTUK MEMORI TERAGIH VOLTAN RENDAH

ABSTRAK

Dalam teknologi nod dan voltan bekalan yang diskala, keupayaan penulisan untuk SRAM direndahkan dan menjadi satu kebimbangan reka bentuk yang kritikal. Pelbagai teknik bantuan penulisan dibangunkan untuk meningkatkan keupayaan penulisan untuk SRAM. Dalam penyelidikan ini, satu teknik bantuan penulisan yang lebih baik mengimplementasikan idea voltan baris-perkataan dipertingkatkan konvensional ke atas sel Dual-VT SRAM telah dicadangkan. Dengan menggunakan transistor akses tulis yang mempunyai voltan ambang yang rendah dalam teknik yang dicadangkan, masa akses tulis yang setanding dengan teknik konvensional boleh dicapai dengan menggunakan voltan dipertingkatkan yang lebih rendah. Voltan dipertingkatkan yang lebih rendah pada pemandu baris-perkataan boleh membantu dalam pengurangan kuasa statik dan dinamik.Walau bagaimanapun, teknik yang dicadangkan mempunyai kelemahan iaitu kuasa statik yang lebih tinggi pada sel SRAM. Ini disebabkan oleh penggunaan transistor yang mempunyai voltan ambang yang rendah. Simulasi SPICE telah dijalankan untuk menilaikan dan membandingkan prestasi penulisan dan penggunaan kuasa teknik konvensional dan teknik dicadangkan. Keputusan simulasi menunjukkan bahawa voltan dipertingkatkan yang lebih rendah boleh digunakan pada transistor akses tulis yang mempunyai voltan ambang yang lebih rendah dengan peningkatan sebanyak 1% hingga 2% berbanding teknik konvensional. Voltan dipertingkatkan untuk transistor akses tulis yang menggunakan Std-VT, Low-VT dan Ultra-Low-VT masing-masing adalah 1.00V, 0.94V dan 0.90V. Apabila operasi tulis SRAM diaktifkan, terdapat purata sebanyak 6% jumlah pengurangan kuasa diperhatikan dengan pelaksanaan transistor akses tulis yang menggunakan Low-VT. Pelaksanaan transistor akses tulis yang menggunakan Ultra-Low-VT menghasilkan 10% jumlah pengurangan kuasa. Apabila operasi tulis SRAM adalah tidak diaktifkan, pelaksanaan transistor akses tulis yang menggunakan Low-VT boleh menjimatkan 7% jumlah penggunaan kuasa.

Walau bagaimanapun, pelaksanaan transistor akses tulis yang menggunakan Ultra-Low-VT menyebabkan sehingga 4% kenaikan dalam jumlah kuasa. Hasilnya, teknik bantuan menulis yang dicadangkan sesuai untuk memori teragih voltan rendah dalam aplikasi kelajuan yang tinggi dan aktiviti yang tinggi dalam operasi tulis memori.

IMPLEMENTATION OF WRITE ASSIST TECHNIQUE ON LOW VOLTAGE DISTRIBUTED MEMORY

ABSTRACT

In scaled technology nodes and scaled supply voltages, the SRAM write ability is being degraded and becomes a critical design concern. Various write assist techniques are developed to improve SRAM write ability. In this study, an improved write assist technique that implements the conventional boosted word-line voltage idea on the Dual-VT SRAM cell is proposed. By adopting the low threshold voltage write access transistors in the proposed technique, the comparable write access time as the conventional technique could be achieved with the lower boosted voltage. The lower boosted voltage on word-line drivers could help in static and dynamic power reductions. However, the proposed technique has drawback of higher static power on SRAM cell due to the adoption of low threshold voltage transistor. The SPICE simulation has been performed to evaluate the write performance and power consumption of the conventional and proposed techniques for comparative study. The simulation results have shown that a lower boosted voltage could be applied on lower threshold voltage write access transistor with 1% to 2% improvement as compared to the conventional technique. The boosted voltages for Std-VT, Low-VT and Ultra-Low-VT write access transistors are 1.00V, 0.94V and 0.90V respectively. When SRAM write operation is activated, there is an average of 6% total power reduction observed with the implementation of Low-VT write access transistor. The implementation of Ultra-Low-VT write access transistor produces 10% of total power reduction. When SRAM write operation is inactivated, the Low-VT write access transistor implementation could save 7% total power consumption. However, the implementation of Ultra-Low-VT write access transistor causes up to 4% total power increment. As a result, the proposed write assist technique is suitable for the low voltage distributed memory in the applications of high speed and high activity of memory write operation.

CHAPTER 1

INTRODUCTION

1.1 Background

Two primary memory types are available in the Field Programmable Gate Array (FPGA). They are distributed memory (Distributed SRAM) and dedicated block memory (Block SRAM) (Mehta, 2010, Altera, 2016). Figure 1.1 shows the generic FPGA with its embedded elements. It consists of configurable logic blocks, configurable routing, embedded memories and embedded digital signal processing (DSP) blocks. The embedded memories shown are the block memories. The distributed memory is formed with the look-up table inside certain configurable logic blocks that normally used for logic function (Xilinx, 2005, Altera, 2016).

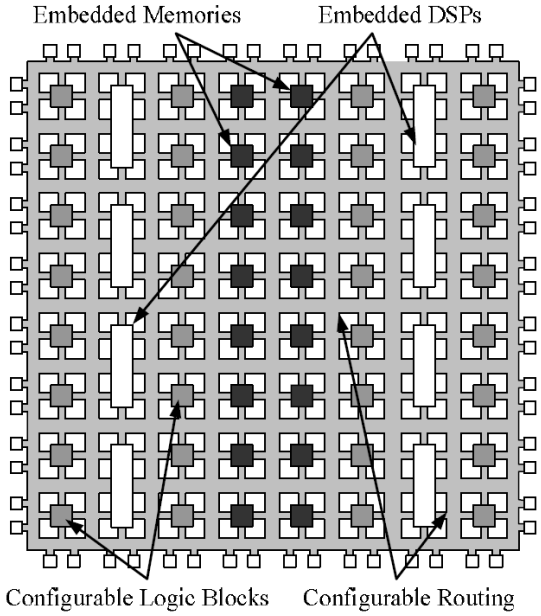

Figure 1.1 Generic FPGA with its embedded elements (Lamoureux and Luk, 2008)

The distributed memory can be programmed to function as Random Access Memory (RAM) or Read Only Memory (ROM). This type of memory is distributed throughout the FPGA, but block memory is located at certain area only. The distributed memory has much

smaller capacity compared to block memory that fixed to 10K Bits, 20K Bits or other amount of K Bits. Thus, it has the advantage over block memory when the full capacity of block memory is not needed. Moreover, distributed memory is ideal for small memory design. For instance, distributed memory is crucial to many high performance applications that require relatively small memory blocks, such as small register files or FIFOs (First-In, First-Out) (Xilinx, 2005, Altera, 2016).

As the need for low power systems increases, lowering the supply voltage (VDD) becomes popular and effective approach to reduce both static and dynamic power consumptions. However, the supply voltage scaling has been raising design challenges to SRAM as it degrades SRAM read stability and write ability. Read and write are the two critical operations of SRAM cell. The read stability issue can be eliminated by utilizing 8T-SRAM cell which has the structure of a dedicated read port where the memory cell storage nodes are isolated from read current path (Ching-Te et al., 2007, Chandra et al., 2010, Keshavarapu et al., 2012, Zimmer et al., 2012). On the other hand, solving the write failure issue is more challenging. With technology node scaling, it is difficult to perform successful write to SRAM cell even at nominal voltage. The design challenges become more apparent when the supply voltage is scaled continuously with technology node (Chandra et al., 2010). Therefore, there is a need to have design assists to enable robust write operation in low voltage SRAM design.

1.2 Current Trend in SRAM Write Technology

As a consequence of technology node and supply voltage scaling, the write performance of SRAM cell is being challenged. In order to address the challenge of writing data into the SRAM cell, several design techniques have been proposed from cell level to architecture level. The novel SRAM cell topologies such as 7T, 8T, 9T, 10T and 11T have been proposed at the cell level for write performance improvement (Moradi and Madsen, 2014, Farkhani et al., 2015).

However, they have the drawback of higher SRAM cell area due to more transistors are utilized to form a memory cell (Farkhani et al., 2015). The proposals at the architecture level are using write assist technique to strengthen the write access transistors or weaken the latch strength (Sinangil et al., 2016, Mann et al., 2010, Moradi and Madsen, 2014, Farkhani et al., 2015, Chandra et al., 2010, Goel et al., 2012, Sharma and Kumar, 2013, Kim et al., 2016). The architecture level of techniques has advantages of less area consumption compared to the cell techniques and suitable to apply in any SRAM cell type (Farkhani et al., 2015).

1.3 Problem Statements

As the technology scales in deep nanometer era, the challenges in designing a robust write and read SRAM cell increase substantially. Both the device variations and leakage are increasing with each shrinking technology node. Furthermore, the supply voltage is scaled down to meet the low power requirements. The robust operation of the SRAM cell at lower supply voltages becomes even more challenging. The read stability issue can be eliminated with SRAM cell topology such as 8T-SRAM cell. The more challenging part is to solve the write failure issue with both the technology and voltage scaling. High minimum operating voltage of SRAM cell will limit its applicability in low power designs. Therefore, it is a need for SRAM cell to implement write assist technique to enable robust write operation at lower supply voltages. The write assist techniques are implemented to aid the memory cell in changing the state during write operation. The existing write assist techniques for SRAM cell include boosted word-line voltage, negative bit-line voltage, reduced cell VDD and raised cell VSS.

1.4 Research Objectives

The objectives of this study are:

1. To propose an improved write assist technique for low voltage distributed memory.
2. To compare the write performance and power consumption of the proposed write assist technique with the conventional write assist technique.

1.5 Research Scope

The scopes of this study are described in this section. A schematic design of memory architecture that consists of small memory array, write address logics and write bit-line logics is constructed. The memory design will be the design under test (DUT) in this study. The same memory architecture is used for both the conventional and the proposed techniques for equitable comparison. The schematic is used to generate the pre-layout netlist for write performance and write power consumption evaluations. Hence, the SPICE simulations that executed for SRAM write access time and write power consumption are pre-layout type of simulations. The SPICE simulation of SRAM write access time is executed to evaluate the write performance of the conventional and the proposed write assist techniques. Besides, the SPICE simulation of SRAM write power consumption is executed to evaluate the power consumed by the DUT that implemented with the conventional and the proposed write assist techniques. The static power, dynamic power and total power consumptions are evaluated. The comparative study of SRAM write access time and write power consumption of the proposed technique with the conventional technique is performed based on the SPICE simulation results.

1.6 Thesis Outline

This thesis is presented with five chapters.

In Chapter 1, the research background, current trend in SRAM write technology and problem statement are presented. The research objectives and scopes are also outlined here.

In Chapter 2, the memory architecture, memory organization, SRAM cell structure and supporting operations are discussed. The discussions on write failure and write margin are presented here. Review on the existing write assist techniques such as boosted word-line voltage, negative bit-line voltage, reduced cell VDD and raised cell VSS is also stated here. Power consumption includes static power and dynamic power consumptions are discussed here as well.

In Chapter 3, the benefit and drawback of the conventional boosted word-line voltage write assist technique are highlighted. The idea of the proposed write assist technique is presented here with theory. The methodologies that used to evaluate the SRAM write performance and power consumption of the proposed write assist technique are presented. The design simulation condition is also stated here.

In Chapter 4, the SPICE simulation results for SRAM write performance and power consumption of the conventional and the proposed write assist technique are presented. The simulation results for both techniques are compared and discussed in details.

Lastly, the Chapter 5 concludes the overall findings from the study. The recommendations for future works on this area are stated as well.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter is mainly to discuss SRAM cell design and its challenging in support robust write operation at lower supply voltage. Firstly, the memory architecture and organization are discussed. It is then followed by discussions on SRAM cell structure, operations, write failure and write margin. Next, various write assist techniques to improve the write performance of SRAM cell at lower supply voltage are discussed. The power consumption of SRAM cell is discussed in the end of this chapter.

2.2 Architecture of Memory design

A memory is a storage unit. It mainly supports write and read operations. Figure 2.1 shows a general architecture of memory design for write and read operation supports. It typically consists of functional blocks such as address decode logic, memory core, write column logic, read column logic, read control logic and write control logic. The function of row decoder is to select the desired word-line which corresponding to the input address and thereby activates the row in the memory array. Prior to read operation, the bit-lines are charged to a supply voltage. In read cycle, the pre-charged bit-lines are either staying charge or discharge is determined by the data stored in the memory cells selected by the word-line. The voltage changes in the bit-lines are detected by sense-amplifier in the read logic and the appropriate data is multiplexed to the data output. The signals to the sense-amplifier and bit-line pre-charged logic are controlled by the read control logic. In write cycle, the bit-line drivers drive the bit-lines with the data to be written into memory location where corresponding to the write address.

At this time, the sense-amplifiers are isolated. After completion of a write or read operation, the bit-lines are pre-charged to supply voltage and waiting for another write or read operation in the next cycle. In case of there is no write or read operation being performed in a clock cycle, all the word-lines are de-activated and the bit-lines are stay pre-charged. This generic design architecture is usually common for every memory design. Typically, their differences are in array size and organization of the memory core that in terms of number of rows and columns (Mamidipaka et al., 2003).

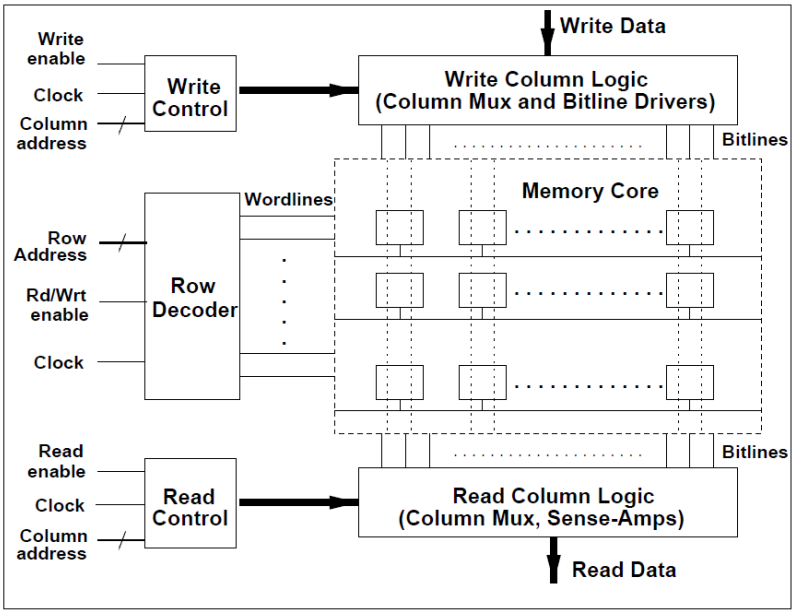

Figure 2.1 General architecture of memory design (Mamidipaka et al., 2003)

2.3 Memory Array Organization

The memory core consists of individual memory cells that are capable of storing one bit binary data information, either logic “0” or logic “1”. Figure 2.2 shows a typical memory array organization. Physically, the memory cells are arranged in two-dimensional array of horizontal rows and vertical columns. Therefore, a row select signal and a column select signal need to be selected in order to access a particular memory cell. All the memory cells in a row are sharing the same row select signal, or also called as word-line. On the other hand, all the memory cells

in a column are sharing the same column select signal, or also called as bit-line (Kang and Leblebici, 2002, Hodges et al., 2004, Shih-Lien, 2006, John, 2007). In the memory array shown in Figure 2.2, there are 2^N rows (word-lines) and 2^M columns (bit-lines). Thus, the total number of memory cells in this array is 2^M x 2^N(Kang and Leblebici, 2002). Since the memory cells can be accessed for data writing or data reading in random order at a fixed rate and independent of their physical locations in the memory array, this kind of array organization is called as a Random Access Memory (RAM) architecture (Kang and Leblebici, 2002, Hodges et al., 2004).

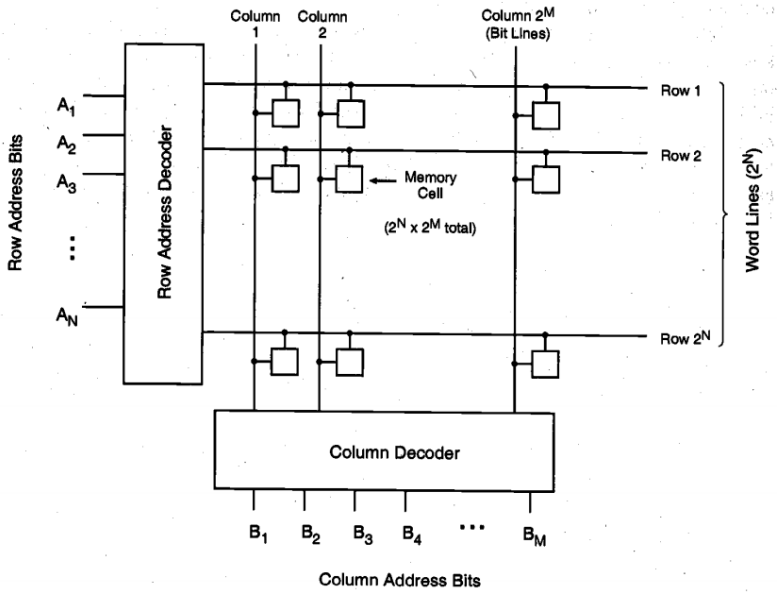

Figure 2.2 Typical Random Access Memory array organization (Kang and Leblebici, 2002)

Static Random Access Memory (SRAM) is a static memory circuit that the memory content is always retained as long as the power supply is being supplied. SRAM does not require periodic refreshing to retain memory content, which is in contrast to Dynamic Random Access Memory (DRAM). In addition, SRAM is categorized as volatile memory which means that its content will be lost if the power supply is interrupted (Shih-Lien, 2006, John, 2007, Patel et al., 2013). Figure 2.3 shows an example of 256 x 256 SRAM architecture with N=M=8. The memory core array contains a total of 65,536 memory cells. A 16-bit address is used by the memory to produce a single bit data output (Hodges et al., 2004).

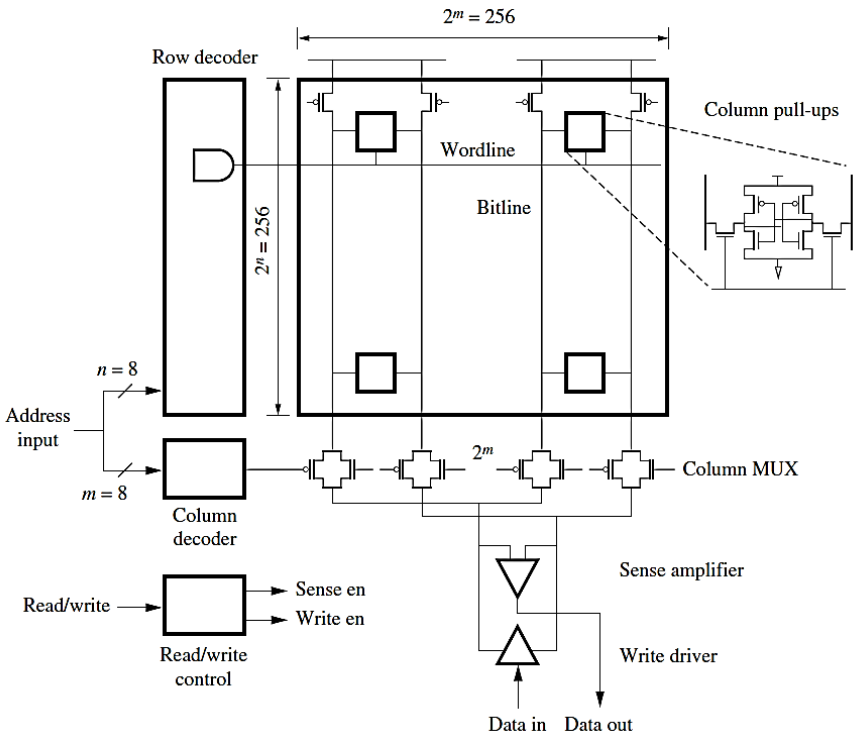

Figure 2.3 An example of 256 x 256 SRAM architecture with N=M=8 (Hodges et al., 2004)

The memory array organization will impact overall SRAM effectiveness. The study in Zimmer et al. (2012) proves that bit-line capacitance which is determined by the number of memory cells on a bit-line has a significant effect on SRAM read ability. When the number of memory cells on a bit-line is large, this causes a large bit-line capacitance that a pull-down transistor must discharge and thus degrade the SRAM read ability. A SRAM content-0 read requires a certain amount of charge to be removed from the pre-charged bit-line. The bit-line with large capacitance requires more charge removed to produce the same differential voltage for sense-amplifier detection when compared to bit-line with smaller capacitance. It has been noted in Yadav et al. (2013) that the bit-line capacitance is increased with the memory capacity. Zimmer et al. (2012) suggests that memory array organized with small number of rows (short column) to keep prevent excessive bit-line capacitance. This is critical to improve low voltage SRAM operation. But, the drawback is area overhead increase due to the column and control circuitry can be amortized over less cells. Besides, SRAM read performance depends on effective read current. The faster SRAM data readout operation can be achieved with higher

effective read current. However, excessive bit-line leakage decreases the effective read current. The impact of bit-line leakage on effective read current can be eliminated by lowering the number of rows per bit-line (Ye et al., 2003).

2.4 SRAM Cell Structure and Operations

There are many SRAM cell structures being designed, but almost every SRAM cell has a pair of cross-coupled inverters (I1 and I2) that act as a storage element. Both the true and complementary value of data is stored on the two different storage nodes (Q and Q'), as shown in Figure 2.4. With the existence of cross-coupled inverters and power supply is supplied, any voltage change on a storage node is counteracted by the feedback path and tends to go back to the original value. For example, the storage node Q is at 0V and storage node Q' is at VDD. Q is discharged through NMOS in I1 if there is any positive charge deposited on storage node Q. Thus, the storage bit information is retained.

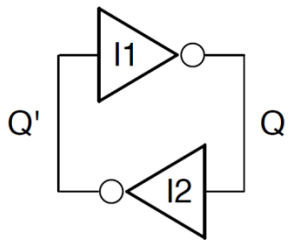

Figure 2.4 Storage element in an SRAM cell (Zimmer, 2012)

The SRAM cell structures differ by the number of transistors used to provide access to these storage nodes. Typically, the SRAM cell is named by the total number of transistors in the memory cell (Yadav et al., 2011, Zimmer, 2012). Figure 2.5 shows the basic 6T-SRAM cell. The 6T-SRAM cell has total of six transistors in the memory cell. It composes of a pair of cross-coupled inverters and two access transistors. The access transistors are connected to the bit-lines at their source or drain terminals. Bit-lines are used to transfer data for both read or write

operations. The word-line is connected to the gate terminal of access transistors and used to select the write or read intended memory cell. The stored value (q) and its complement ($\overline{q}$) are held internally in the memory cell.

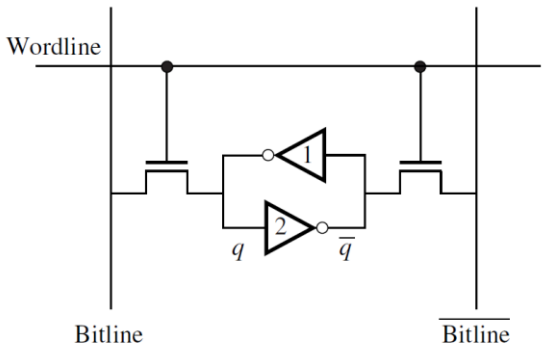

Figure 2.5 Basic 6T-SRAM cell (Hodges et al., 2004)

Figure 2.6 shows the voltage transfer characteristic (VTC) of cross-coupled inverters. The VTC delivers the key design considerations for write and read operations of SRAM cell. In the configuration of cross-coupled inverters, the stored values are represented by the two stable states in the VTC, namely Stored 0 and Stored 1. The memory cell retains its current state until the switching threshold, VS is crossed by one of the storage nodes. When this takes place, the internal state of memory cell is flipped. Therefore, current state cannot be disturbed during a read operation in order to show good read stability. During a write operation, the internal voltage is forced to swing over VS to change the state for showing good write ability.

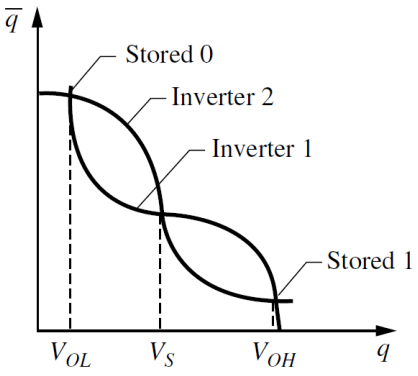

Figure 2.6 Voltage Transfer Characteristic of SRAM cell (Hodges et al., 2004)

The schematic for 6T-SRAM cell in CMOS technology is shown in Figure 2.7. The 6T-SRAM cell is composes of two PMOS and four NMOS. The cross-coupled inverters, M1/ M5 and M2/M6, function as the storage element. The NMOS pass-gates, M3 and M4, act as the access transistors to the storage element. SRAM cell performs three different operations, namely read, write and hold operations (Prakash et al., 2012, Singh et al., 2012). It is in reading mode when the memory content has been requested. And, it is in writing operation when the memory content needs to be updated. SRAM is said to be in hold mode when the circuit is idle, no read or write operation is performed.

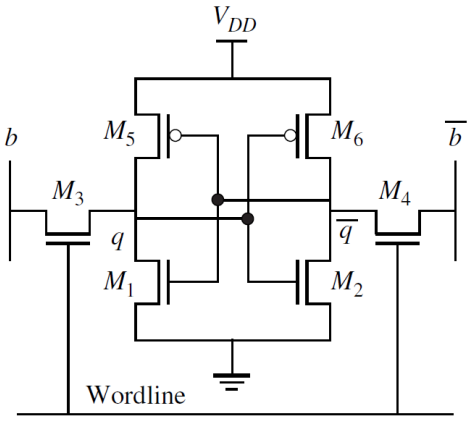

Figure 2.7 Schematic of 6T-SRAM cell (Hodges et al., 2004)

For 6T-SRAM cell in read operation as shown in Figure 2.8, assuming that a data (bit "0") is stored on the left side of the cell (q), and its compliment (bit "1") is stored on the right side ($\bar{q}$). At this time, M1 is turned on and M2 is turned off. Prior to read operation, bit-lines (b and $\bar{b}$) are pre-charged to a high voltage around VDD. The word-line (wl) that held low in the standby state is raised to VDD which turns on access transistors M3 and M4. The current ($Icell$) begins to flow through M3 and M1 to ground. The capacitance on bit-line b (C_{bit}) is slowly discharged due to the resulting cell current. Meanwhile, the voltage on bit-line $\bar{b}$ remains high on the other side of the cell since there is no conduction path to ground through M2. The voltage difference between b and $\bar{b}$ (ΔV) is fed to a sense amplifier to convert to a logic low

level output. Upon completion of the read cycle, the word-line is returned to zero and the bit-lines are pre-charged back to a high value (Hodges et al., 2004, Qazi et al., 2011).

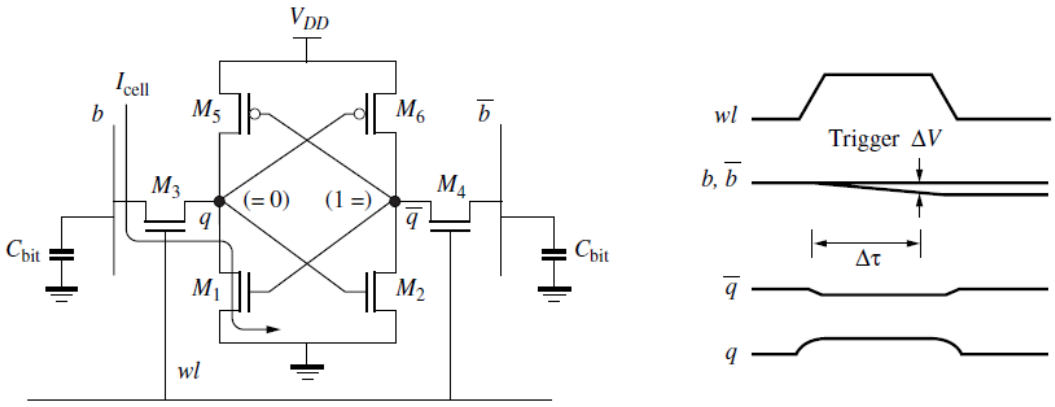

Figure 2.8 6T-SRAM cell in read operation (Hodges et al., 2004)

In order to achieve read stability during read operation, the memory cell must be designed to ensure that the stored values are not disturbed and corrupted during the read cycle. The problem is that the voltage at storage node q is raised when current flows through M3 and M1. This raised voltage could turn on M2 and bring down the voltage at storage node $\bar{q}$, as shown in Figure 2.8. The memory content is being altered if M2 is turned on by the raised voltage at node q. The state altering in memory cell during read cycle is called as read disturb (Ching-Te et al., 2007). In order to avoid read disturb, the memory cell must be designed such that the conductance of M1 is several times larger than M3. This is to avoid the drain voltage of M1 does not rise above VS as mentioned in Figure 2.6. In effect, the read stability requirement establishes the cell ratio which is the ratio of pull down NMOS transistor to the access NMOS transistor in SRAM cell (Hodges et al., 2004, Keshavarapu et al., 2012, Moradi and Madsen, 2014). The other design consideration for SRAM read is the cell current (I_{cell}) for bit-line capacitance discharge sufficiently within a specified of time. The rate of change of the bit-line voltage can be approximated as follows (Hodges et al., 2004):

$$I_{cell} = C_{bit} \frac{dV}{dt} \tag{2.1}$$

$$\frac{dV}{dt} = \frac{I_{cell}}{C_{bit}} \tag{2.2}$$

where I_{cell} is cell current, C_{bit} is bit-line capacitance and $\frac{dV}{dt}$ is rate of change of bit-line voltage. It is clearly shown in Equation (2.2) that cell current and bit-line capacitance control the rate of bit-line discharge. Larger cell current and smaller bit-line capacitance speed up the bit-line discharge process.

For 6T-SRAM cell in write operation, the operation of writing data bit "0" or bit "1" is accomplished by forcing one bit-line (b or $\bar{b}$) low while the other bit-line remains at about VDD. This also means that $\bar{b}$ is forced low for write-1, and $b$is forced low for write-0. Figure 2.9 shows the conditions when SRAM perform write-1. The word-line (wl) that held low in the standby state is raised to VDD which turns on access transistors M3 and M4. The current starts to flow through M4 and M6 causing the voltage at storage node $\bar{q}$ drops from its initial voltage of VDD. Once the storage node $\bar{q}$ dropped below VS as mentioned in Figure 2.6, the regenerative effect between the two inverters is initiated and the cell is forced to switch. The regenerative operation is M1 turned off and its drain voltage rises to VDD due to the pull-up operation of M3 and M5. At the same time, M2 is turned on and helping M4 to pull storage node $\bar{q}$ to its intended low value. When the memory cell finally flips to the new state, the world-line can be de-asserted by returning it to the low standby level.

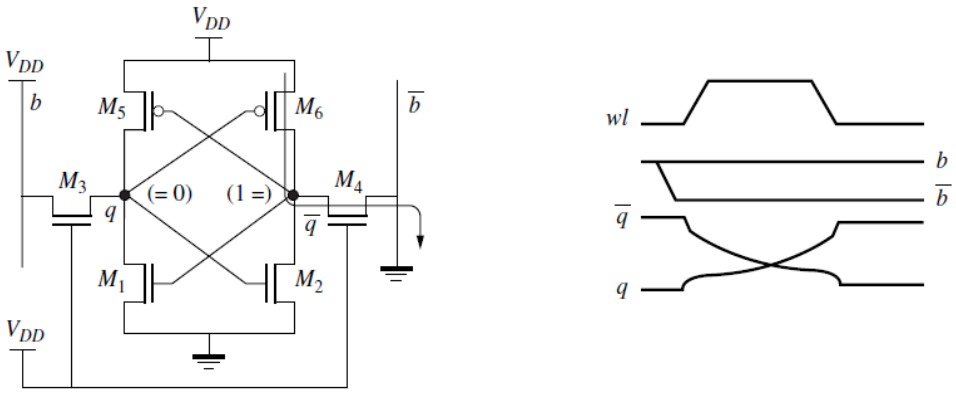

Figure 2.9 6T-SRAM cell in write operation (Hodges et al., 2004)

SRAM write ability needs to be taken care during design phase. The regenerative action required for cell content switch is initiated when the drain of M2 is pulled below VS. Therefore, the memory cell must be designed such that the conductance of M4 is several times larger than M6 in order to ensure proper write operation. In effect, the write ability requirement establishes the pull up ratio which is the ratio of pull up PMOS transistor to the access NMOS transistor in SRAM cell (Hodges et al., 2004, Keshavarapu et al., 2012).

During hold operation, the cell storage nodes are disconnected from the bit-lines by the access transistors M3 and M4 due to the de-assertion of the word-line. The two cross-coupled inverters will continue to reinforce each other to retain the memory content whenever there is sufficient power supply connected to them (Patel et al., 2013, Yadav et al., 2013). If the power supply becomes too low, the feedback provided by the cross-coupled inverters becomes too weak to retain data. The memory cell could flip and fail to perform hold characteristic if this is happen. The minimum voltage at which the memory cell is able to retain their voltage is known as the data retention voltage (DRV) (Amelifard et al., 2006, Keshavarapu et al., 2012, Zimmer, 2012, Sinha and Samanta, 2015).

The 6T-SRAM cell is using the same port for write and read operations. One key conflict requirement for read and write operations can be observed through the cell ratio and pull up ratio. It is desirable to have strong storage inverters and weak access transistors to minimize read disturb. However, it is desirable to have weak storage inverters and strong access transistors in order to improve write ability. The conflict between read and write requirement is an inevitable design constraint that needs to be considered during design phase (Ching-Te et al., 2007, Goel et al., 2012, Keshavarapu et al., 2012, Moradi and Madsen, 2014). However, it is difficult to design 6T-SRAM cell which is stable for both read and write especially in the scaled technology nodes. This is due to the ratio requirements can be severely impacted by device variation (Birla et al., 2010).

In order to isolate the conflict requirement between SRAM write and read, 8T-SRAM cell becomes a promising candidate to replace 6T-SRAM cell. The 8T-SRAM cell is the 6T-SRAM that adding with two stacked NMOS for separating the read word-line (RWL) and write word-line (WWL), as shown in Figure 2.10. The 8T-SRAM write operation is the same as 6T-SRAM at which using write word-line to access to the cell storage nodes. The 8T-SRAM cell read operation is initiated by pre-charging the read bit-line to a high voltage. Read word-line is asserted to turn on transistor RPG after pre-charging read bit-line. If BLLI = 0, RPD is turned on and read bit-line is discharged through transistors RPG and RPD to ground. The voltage on read bit-line is decreased and sensed by the sense amplifier. On the other hand, if BLLI = 1, RPD is turned off, so there is no discharge current flow through the read path, the read bit-line should remain at acceptable high voltage level. At this time, only a small amount of leakage current flows which is called as bit-line leakage (Prakash et al., 2012). By separating the read word-line and write word-line and isolating the cell storage nodes from read current path, the read disturb is completely eliminated (Ching-Te et al., 2007, Chandra et al., 2010, Keshavarapu et al., 2012, Zimmer et al., 2012). However, 8T-SRAM cell does not improve write ability. Furthermore, the 8T-SRAM cell consumes more area overhead when compared to the conventional 6T-SRAM cell (Chandra et al., 2010, Keshavarapu et al., 2012). In Ching-Te et al. (2007), a study reports that the area of 8T-SRAM cell becomes smaller compared to 6T-SRAM cell at the 32nm technology node due to the 6T-SRAM need to trade the transistor area to fulfill the ratio requirements.

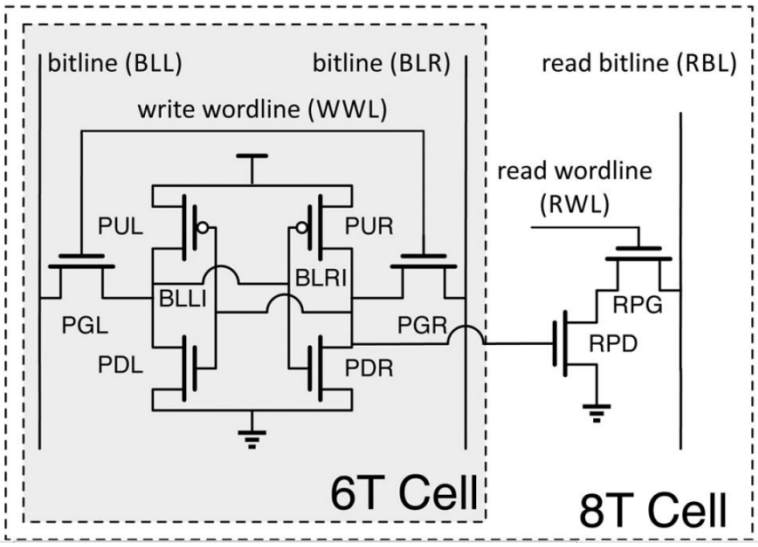

Figure 2.10 8T-SRAM cell (Zimmer et al., 2012)

2.5 SRAM Write Failure and Write Margin

SRAM is an important element of most VLSI systems. As the demand for low power systems grows, lowering the supply voltage (VDD) becomes one of the most effective approaches to reduce both static and dynamic power consumptions. However, the supply voltage scaling has been posing significant challenges to SRAM design as it degrades SRAM read stability and write ability. Read stability failures occur when the memory content flip accidentally during a read operation. Write ability failures occur when the cell storage node voltage does not reach the desired write data value during a write operation (Zimmer et al., 2012). Hence, SRAM design work must be carefully carried out in order to minimize failure and increase yield in production. The read stability concern can be mitigated by utilizing the 8T-SRAM cell structure as mentioned in the previous section. Thus, this indicates that solving the issue of write failure is more challenging and it is the area that needs more efforts to look into.

SRAM write failure is defined as the failure to intentionally alter the content of the memory cell during the write operation (Chandra et al., 2010, Zimmer, 2012). Write margin is a metric used to evaluate the write ability of an SRAM cell in write operation (Goel et al., 2012,

Moradi and Madsen, 2014, Sinha and Samanta, 2015). There are static write margin and dynamic write margin. Static metric is traditionally used to evaluate write ability, but several studies have shown that static metric is poor match to silicon failures. It fails to predict the outliers in the critical write ability, and they could probably underestimate write failure rate (Toh et al., 2011, Zimmer et al., 2012). However, write margin measured using static method is easy and fast for simulation and testing (Wang et al., 2008). The study conducted by Wang et al. (2008) shows that word-line voltage sweep static matrix is the best among the existing static metrics for evaluating the dynamic write behaviour of SRAM cells for lower VDD and future nanometer technologies. In word-line voltage sweep method, the bit-lines are connected to the appropriate voltages to enable cell storage node flipping when word-line is swept from 0V to VDD. The static write margin is calculated as the difference between VDD and word-line voltage when the stored data is flipped $(VDD - VWL_{flip})$ (Mann et al., 2010, Moradi and Madsen, 2014). Static metric does not account for important transient effects such as write access time where increases exponentially in lower voltage SRAM design (Iijima et al., 2008). Thus, dynamic metric which derived from the SRAM under dynamic access provides a better estimate of SRAM write ability (Toh et al., 2011).

The effectiveness of an SRAM write operation is typically quantified by the minimum (or critical) width of word-line pulse (defined as WLcrit) during which the bit cell changes state. The definition is important because it captures the dynamic write margin which is more accurate. The static approaches to measure the write margin assume that the word-line pulse width is infinite which could lead to erroneous conclusions. Figure 2.11 shows how the internal storage nodes (q and $\bar{q}$) of a memory cell change when word-line (WL) of different pulse widths are applied during the write operation. In Figure 2.11(a), the pulse width of WL is smaller than the WLcrit whereas in Figure 2.11(b), the pulse width of WL is equal to or larger than WLcrit. In the case of Figure 2.11(a), the nodes $q$and $\bar{q}$ tend to move towards the other state but they return back to the original state in the end of operation. This indicates SRAM write failure scenario. In Figure 2.11(b), the nodes $q$and $\bar{q}$ are able to move to other stable state, indicates write success

scenario. Moreover, there is potential write failure happen in the marginally write-able memory cell. This is because the voltage value of storage node possible has not been completely written and considered near to the meta-stability point (Zimmer, 2012). In the other word, WL needs to be larger than WLcrit and long enough in order to have a robust and successful write operation.

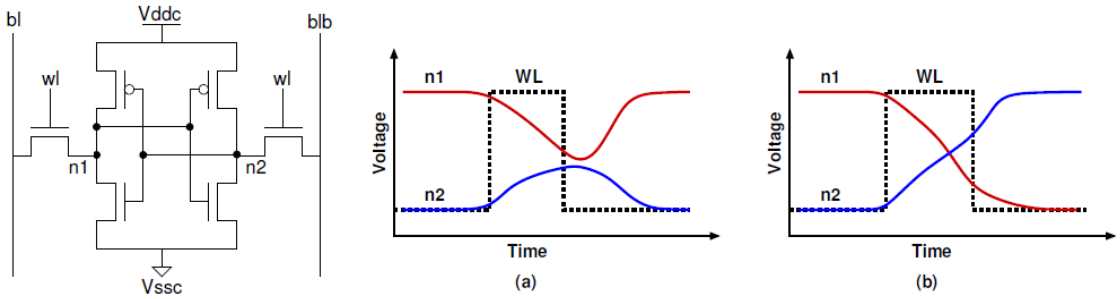

Figure 2.11 Memory cell storage nodes during write operation (Chandra et al., 2010)
(a) Unsuccessful write operation (b) Successful write operation

Toh et al. (2011) compares the fail bit count between static and dynamic conditions and demonstrates the optimistic of static write margins. Figure 2.12 shows the fail bit count dependence on VDD and word-line pulse width. The infinite word-line pulse width case is represented by the "Static" line. The "1ns" line represents the shorter word-line pulse width case while the "20ns" line represents the longer word-line pulse width case. The result shows that the voltage scaling has no impact on write ability (zero write failure) with using static metric that assuming infinite word-line pulse width. On the other hand, dynamic metric shows that the write fail bit increases when the supply voltage scales down and shorter word-line pulse width is used.

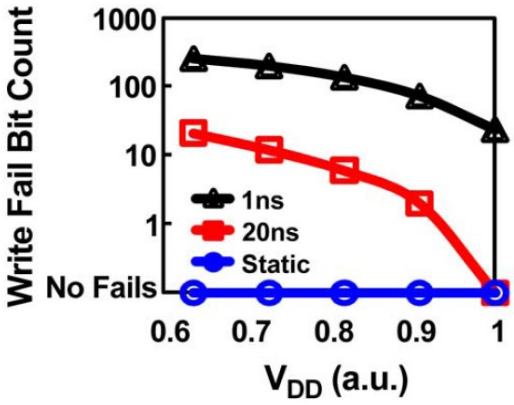

Figure 2.12 Write fail bit count dependence on VDD and word-line pulse width (Toh et al., 2011)

2.6 SRAM Write Assist Techniques

With technology node scaling, it becomes difficult to write to SRAM cell even at nominal supply voltage and the challenges become more apparent at lower supply voltages. Figure 2.13 shows the trend of minimum (or critical) width of word-line pulse (WLcrit) with respect to supply voltage for a 32nm SRAM cell. It shows around 10 times increase in WLcrit as the voltage scales down from 1V to 0.7V. This increment trend is very alarming since supply voltage is frequently dynamically scaled to reduce power consumption.

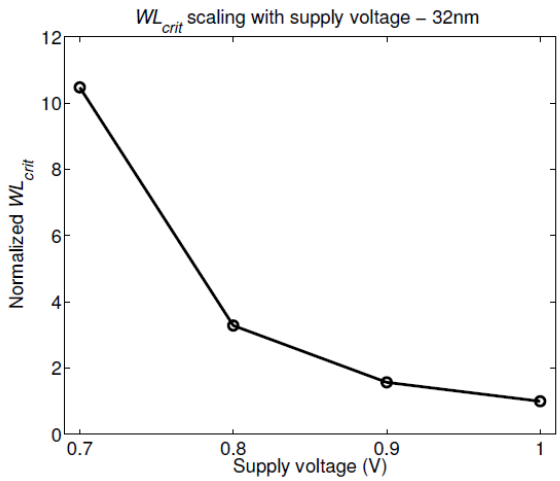

Figure 2.13 Change in WLcrit with voltage scaling at 32nm technology node (Chandra et al., 2010)

The substantial increase in WLcrit increases the minimum operating voltage (Vmin) of the SRAM cell and limits its applicability in low power designs. Hence, the write assist techniques to improve the SRAM cell write performance at lower supply voltages are introduced. The existing write assist techniques include boosted word-line voltage, negative bit-line voltage, reduced cell VDD and raised cell VSS. They are now commonly used to lower the minimum operating voltage of SRAM cell.

2.6.1 Boosted Word-line Voltage Write Assist Technique

This technique assists the memory cell to flip during a write event by boosting the word-line higher than the supply voltage Vddc, as shown in Figure 2.14(a). The word-line boosting increases the gate-to-source voltage (Vgs) of the access transistor. Hence, it increases drive strength of access transistors and reduces its gate delay time. Moreover, the voltage overdriven of access transistors leads them stronger than the pull up transistors reside the memory cell and therefore helps significantly in flipping the memory cell (Mann et al., 2010, Iijima et al., 2008, Chandra et al., 2010, Zimmer, 2012, Zimmer et al., 2012). This is the pull up ratio requirement for write ability as discussed in Section 2.4. Therefore, this boosted word-line voltage assist technique causes the memory cell easier and faster to be written with the write data. Figure 2.14(b) shows the impact of word-line boosting based write assist on the WLcrit. The WLcrit in this case is substantially better than the nominal case with no write assist. The result also shows that the benefits of this write assist scheme increases significantly as the supply voltage is scaled down. It has been found that boosted word-line voltage scheme is one of the most effective write assist techniques (Chandra et al., 2010, Sharma and Kumar, 2013). This write assist technique requires an extra supply voltage for the voltage boosted word-line drivers. The boosted voltage can be generated internally by a charge pump or by capacitive coupling. Also, a separate power supply can be routed for its implementation (Iijima et al., 2008,

Chandra et al., 2010, Zimmer, 2012). Besides area overhead, energy overhead is caused by a larger voltage on the word-line drivers (Zimmer, 2012).

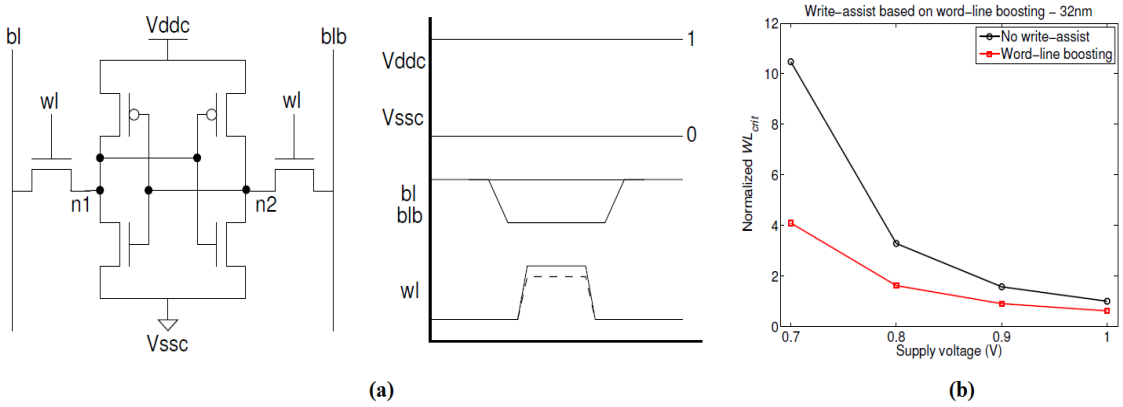

Figure 2.14 Write assist based on boosted word-line voltage(Chandra et al., 2010) (a) Schematic and waveform (b) Impact of write assist on the WLcrit.

2.6.2 Negative Bit-line Voltage Write Assist Technique

There are two ways to create a larger gate-to-source voltage (Vgs) for the NMOS access transistor. It can be achieved with increasing the gate voltage or decreasing the source voltage. The approach of negative bit-line based write assist swings the bit-line voltage below zero volt during the write operation, as shown in Figure 2.15(a). The increase in Vgs causes the access transistor to become stronger and hence can flip the memory cell easily (Mann et al., 2010, Chandra et al., 2010, Goel et al., 2012, Zimmer, 2012, Zimmer et al., 2012). This is the same theory as in the boosted word-line voltage technique. Figure 2.15(b) shows the impact of negative bit-line based write assist on the WLcrit. The benefits of this write assist scheme increases as well when the supply voltage is scaled down. It has been found that negative bit-line voltage scheme is one of the most effective write assist techniques (Chandra et al., 2010, Sharma and Kumar, 2013). But the write stability of memory cells arranged in the same column is decreased because a negative bit-line causes a small overdriven voltage on the un-accessed NMOS pass-gate transistors in the same column which supposed to remain at off state (Gate

terminal at 0V, source terminal at negative voltage). There is probably potential that the memory cell could be accidentally written if the pass-gate transistor has a low threshold voltage (Zimmer, 2012, Zimmer et al., 2012). Therefore, it is important to take into the consideration of the variation in the negative voltage during design phase. Similar to the word-line boosting technique, this write assist technique requires an extra supply voltage for the bit-line. The negative bit-line voltage can be generated on-chip by a charge pump or by capacitive coupling (Farkhani et al., 2015, Chandra et al., 2010, Goel et al., 2012). The energy overhead is minimal for negative bit-line voltage scheme (Zimmer, 2012).

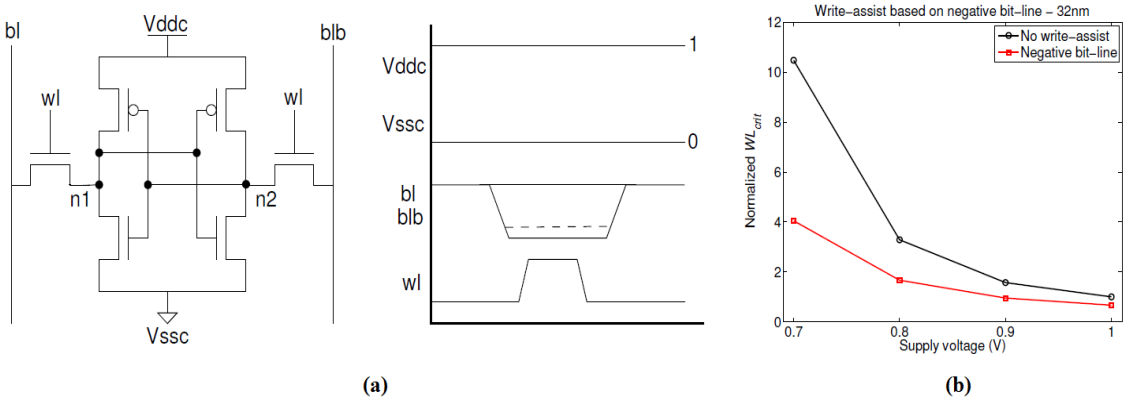

Figure 2.15 Write assist based on negative bit-line voltage(Chandra et al., 2010)
(a) Schematic and waveform (b) Impact of write assist on the WLcrit.

2.6.3 Reduced Cell VDD Write Assist Technique

This write assist technique targets on decreasing the strength of cross-coupled inverters in the SRAM cell. The WLcrit can be decreased by weakening the pull-up transistors with respect to the NMOS access transistors. It is easier to write a new data to the memory cell once the pull-up transistor is weakened (Chandra et al., 2010, Zimmer, 2012, Zimmer et al., 2012). Figure 2.16(a) shows the timing relationships using the VDD lowering write assist scheme. Figure 2.16(b) shows the impact of the VDD lowering based write assist on the WLcrit. The gain is not significant although the write assist based scheme consistently performs better than the nominal case. This is due to the fact that the pull-up PMOS is already very weak in the

SRAM cell and hence making it further weak does not help much. The lower supply voltage can be generated by on-chip regulator or using a second external supply voltage (Chandra et al., 2010). The main challenge with this assist technique is to make sure the reduced cell VDD voltage is still higher than the data retention voltage of the SRAM cell (Chandra et al., 2010, Zimmer, 2012).

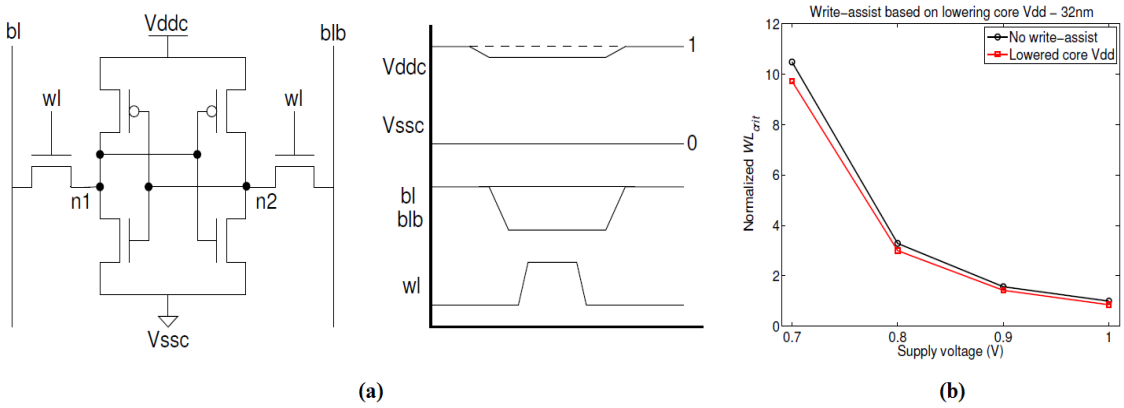

Figure 2.16 Write assist based on reduced cell VDD(Chandra et al., 2010)
(a) Schematic and waveform (b) Impact of write assist on the WLcrit.

2.6.4 Raised Cell VSS Write Assist Technique

A raised ground scheme is another way to help the write operation and decrease the value of WLcrit. The write assist idea is still to weaken the pull up PMOS but in this scheme it is implemented by weakening the PMOS through gate voltage instead of the source voltage (Chandra et al., 2010, Zimmer, 2012, Zimmer et al., 2012). The cell ground, Vssc as shown in Figure 2.17(a) is raised during the write operation. Figure 2.17(b) shows the impact of the VSS raising based write assist on the WLcrit. The WLcrit for the write assist case is better than the non write assist case but the gain is very small. The reason for this marginal increase is similar to that of the reduced cell VDD write assist scheme which is that the PMOS is already weak and hence further weakening the gate drive does not further weaken it. The extra ground voltage can be generated by on-chip regulator or routed as a separate ground. This kind of pull up PMOS