

# DESIGN VALIDATION OF SINGLE PORT SRAM COMPILER

By

**ALPESH PATEL**

07MEC011



DEPARTMENT OF ELECTRONICS & COMMUNICATION  
ENGINEERING  
AHMEDABAD-382481

May 2009

# Design Validation of Single Port SRAM Compiler

Major Project Report

Submitted In Partial Fulfillment of the Requirement

For

MASTER OF TECHNOLOGY

in

ELECTRONICS & COMMUNICATION ENGG.

(VLSI DESIGN)

By

Alpesh Patel(07MEC011)

Guided By:

Mr. Nitesh Gautam

STMicroelectronics Pvt. Ltd.



Department of Electronics & Communication Engineering  
INSTITUTE OF TECHNOLOGY  
NIRMA UNIVERSITY OF SCIENCE and TECHNOLOGY  
AHMEDABAD-382481

May 2009

## Certificate

This is to certify that the Major Project entitled "Design Validation of Single Port SRAM Compiler" submitted by Alpesh M Patel (07MEC011), towards the partial fulfillment of the requirements for the degree of Master of Technology in Electronic & Communication of Nirma University of Science and Technology, Ahmedabad at STMicroelectronics Pvt. Ltd., Greater Noida is the record of work carried out by him under my supervision and guidance. In my opinion, the submitted work has reached a level required for being accepted for examination. The results embodied in this major project, to the best of my knowledge, haven't been submitted to any other university or institution for award of any degree or diploma.

Mr. Pinakin Thaker  
Project Guide,  
Scientist / Engg., SATD,  
SAC, ISRO,  
Ahmedabad

Prof. N.P. Gajjar  
Internal Project Guide,  
Department of EC Engineering,  
Institute of Technology,  
Nirma University, Ahmedabad

Prof. A. S. Ranade,  
HOD (E.C. Dept.),  
Department of EC Engineering,  
Institute of Technology,  
Nirma University, Ahmedabad

Dr K Kotecha  
Director,,  
Department of EC Engineering,  
Institute of Technology,  
Nirma University, Ahmedabad

## Abstract

This Project Report explores the Analysis and Verification of a Single Port Static Random Access Memory (SRAM) with 90nm three metal level technology, which includes the basic theory of memory and working principals of SRAM. The Report also describe the post layout extraction, different analysis, measurements and characterization of different parameters like timing, capacitances, race conditions and the results of different analysis. The Memory cell analysis includes static noise margin, write margin, discharge rate and leakage current through memory cell during ON and OFF condition. The Dynamic circuit and latch analysis includes the leakage analysis, charge sharing validation, strength validations and bump validation in Row-decoder and I/O section of memory. The Marginality analysis includes analysis of time race condition using memory characterization flow setup. At the end the reports explains the sense amplifier offset and pulsewidth analysis and the write self time analysis required for the nanometer range technology.

## Acknowledgements

I first would like to acknowledge help and valuable guidance of **Dr. N. M. Devashrayee** ( PG Coordinator,(VLSI design, Nirma University, Ahmedabad ) ,because of whom I got an opportunity to work at STMicroelectronics, Greater Noida. I would like to express my hearty thanks and indebtedness to my guide **Mr. Nitesh Gautam** for his enormous help and encouragement.

I express my soulful gratitude to Mr. Anand Hardi for their invaluable suggestion for my training. I specially want to acknowledge Mrs. Nutan Aggarwal who gives me opportunity to do my thesis work and provide all resources required for my project work. I also want to acknowledge Vivek Asthana, Sachin Gulyani, Jitendra Dasani, Ashish Kumar, Atul Bhargva, Parminder Singh, Deepti Saini, Seema Jaiswal, Swarnima Singh , Manish Patel, Manish Trivedi, Chirag Gulati, Mohit Verma, Varun, Satisbabu, Shivendera, Saurabh Agrawal, Jasvinder Sing, Naveen Batra and Jyoti without whom this project work could not have seen the daylight and help me with their constant involvement during my project tenure here.

I would like to express my sincere gratitude to **Prof. Amisha Naik** (ECE Department, Nirma University, Ahmedabad) for her continuous guidance, support and enthusiasm.

I acknowledge gratefully the help and suggestion of NVSM people and friends, in spite of their busy schedule and huge workload, they were always eager to help me with their warm attitude and technical knowledge.I would also like to thank my colleagues and friend for the things that they have taught me. My greatest thanks are to all who wished me success especially my parents.

**Alpesh M. Patel**

**07MEC011**

## Company Profile

**STMicroelectronics** is the world's fifth largest semiconductor company with net revenues of US \$ 9.84 billion in 2008.

The Company's sales are well balanced among the semiconductor industry's five major high-growth sectors (approximate percentage of ST's sales in 2008 : Communications (36%), Consumer (17%), Computer (16%), Automotive (15%) and Industrial (17%).

According to the latest industry data from iSuppli, ST holds market leadership in many fields. For example, the Company is the leading producer of application-specific analog chips and power conversion devices. It is the No 1 supplier of semiconductors for the Industrial market, set-top box applications, and MEMS (micro-electromechanical systems) chips for portable and consumer devices, including game controllers and smart phones. ST also occupies leading positions in fields as varied as automotive integrated circuits (No 3), chips for computer peripherals (No 3), and the rapidly expanding market for MEMS overall (No 5).

**Product Portfolio** ST aims to be the leader in multimedia convergence and power applications, offering one of the world's broadest product portfolios, including application-specific products containing a large proprietary IP (Intellectual Property) content and multi-segment products that range from discrete devices to high-performance micro-controllers, secure smart card chips and MEMS devices.

The Company provides solutions for a wide array of Digital Consumer applications, with a particular focus on set-top boxes, digital TVs and digital audio, including radio. In the Computer Peripherals arena, ST provides leading solutions in data storage, printing, visual display units, power management for PC motherboards, and power supplies. A wide range of ST's ASSPs (Application Specific Standard Products) power sophisticated Automotive systems such as engine control, vehicle safety

equipment, door modules, and in-car infotainment. The Company also supplies industrial integrated circuits (IC) for factory automation systems, chips for lighting, battery chargers and power supplies, as well as chips for advanced Secure Access applications.

ST pioneered and continues to refine the use of platform-based design methodologies for complex ICs in demanding applications such as mobile multimedia, set-top boxes and computer peripherals. The balanced portfolio approach allows ST to address the needs of all microelectronics users, from global strategic customers for whom ST is the partner of choice, for major System-on-Chip (SoC) projects to local enterprises that need fully-supported general-purpose devices and solutions.

To maximize the benefit of scale that is becoming increasingly important in some semiconductor markets, ST completed the creation of two joint ventures in 2008. In the memory field, ST, Intel and Francisco Partners formed a new company, Numonyx, dedicated to providing non-volatile memory solutions, including NAND and NOR Flash memories as well as MCP (multi-chip package) memory solutions, for a wide variety of consumer and industrial applications. ST holds a 48% share in Numonyx.

ST has also been very active in the wireless arena. In mid 2008, ST and NXP combined their key wireless semiconductor operations in a joint venture, in which ST held 80%. In February 2009, ST acquired the minority stake from NXP and merged its wireless operations with Ericsson Mobile Platforms to create ST-Ericsson, a 50/50 joint venture focusing on semiconductors and platforms for mobile applications.

# Contents

<b>Certificate</b>	<b>iii</b>
<b>Abstract</b>	<b>iv</b>
<b>Acknowledgements</b>	<b>v</b>
<b>Company Profile</b>	<b>vi</b>
<b>List of Figures</b>	<b>x</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Objective and Scope of the Project . . . . .	1
1.2 Applications . . . . .	2
1.3 Definition of Problem . . . . .	2
1.4 Scope of Work . . . . .	3
1.5 Implementation Environment . . . . .	3
1.6 Organization of Report . . . . .	4
<b>2 Exhaustive Survey Of Memory</b>	<b>6</b>
2.1 MOS Memories . . . . .	6
2.1.1 Read-Write Memories . . . . .	7
2.1.2 Read-Only Memories . . . . .	8
2.2 Memory Organization . . . . .	8
2.2.1 Functional SRAM Chip Model . . . . .	9
2.3 SRAM Cell Description . . . . .	11
2.3.1 Static RAM Cell . . . . .	11
2.3.2 CMOS SRAM Cell (6T Memory Cell) . . . . .	13
2.4 Read Operation . . . . .	14
2.5 Write Circuitry of SRAM . . . . .	15
2.6 Timings Waveforms . . . . .	16
2.6.1 Read Cycle Timings . . . . .	16
2.6.2 Write Cycle Timings . . . . .	17
2.7 IO Block Circuitry . . . . .	19
2.7.1 Column Pull Up . . . . .	19



2.7.2	Column Multiplexing . . . . .	21
2.7.3	Latch Based Sense Amplifier . . . . .	22
2.8	SRAM Architecture . . . . .	23
2.8.1	Basic Architecture . . . . .	23
2.8.2	Split Core Architecture . . . . .	25
2.9	Access Time . . . . .	25
2.10	Monte Carlo Analysis . . . . .	26
<b>3</b>	<b>Development Flow Of SRAMS</b>	<b>27</b>
3.1	Features of Memory Compiler . . . . .	27
3.1.1	Default Features . . . . .	27
3.1.2	Optional Features . . . . .	28
3.1.3	Feature Description . . . . .	28
3.2	Symbol of Memory Instance . . . . .	29
3.3	Block Diagram of Memory Instance . . . . .	29
3.4	Functional Description . . . . .	30
3.5	Detail Life Cycle of Project . . . . .	31
<b>4</b>	<b>Implementation</b>	<b>33</b>
4.1	Critical Path Modeling . . . . .	33
4.2	Replica Technique for Wordline and Sense Control in Low Power SRAM	36
4.2.1	Inverter Chain Technique . . . . .	37
4.2.2	Capacitance Ratioing . . . . .	39
4.2.3	Cell Current Ratioing . . . . .	42
<b>5</b>	<b>Extraction and MCF</b>	<b>45</b>
5.1	Post-Layout Extraction . . . . .	45
5.1.1	Need Of Extraction . . . . .	45
5.1.2	Extraction Methodologies . . . . .	46
5.2	Extraction of Single Memory cell . . . . .	47
5.3	Memory Characterization Flow . . . . .	48
<b>6</b>	<b>Analysis Work</b>	<b>51</b>
6.1	Memory Cell Analysis . . . . .	51
6.1.1	Static Noise Margin (SNM) . . . . .	52
6.1.2	Write Margin . . . . .	53
6.1.3	Discharge Rate . . . . .	54
6.1.4	IOFF and ION Currents . . . . .	55
6.2	Latch and Dynamic Circuit Analysis . . . . .	56
6.3	Row-Decoder Validation . . . . .	57
6.3.1	Leakage Validation . . . . .	57
6.3.2	Strength Validation . . . . .	57
6.3.3	Charge Sharing Validation . . . . .	59
6.3.4	BUMP Validation . . . . .	59

6.4	I/O Latch Validation . . . . .	59
6.4.1	Leakage Validation . . . . .	60
6.4.2	Strength Validation . . . . .	61
6.4.3	Static Noise Margin (SNM) . . . . .	62
6.4.4	Coupling . . . . .	63
6.5	Marginality Analysis . . . . .	63
6.5.1	Criteria for Marginality Analysis . . . . .	64
6.6	Power On Analysis . . . . .	65
6.7	Access Time Analysis . . . . .	65
6.8	Sense Amplifier Offset Analysis . . . . .	67
6.9	Sense Amplifier Pulse Width Analysis . . . . .	70
6.10	Write Self Time Analysis . . . . .	70
6.11	Read Self Time Analysis . . . . .	72
<b>7</b>	<b>Results &amp; Waveforms</b>	<b>73</b>
<b>8</b>	<b>Conclusion &amp; Future Scope</b>	<b>85</b>
8.1	Conclusion . . . . .	85
8.2	Future Scope . . . . .	85
	<b>Index</b>	<b>87</b>

# List of Figures

2.1	MOS Memories . . . . .	7
2.2	Memory System Organization . . . . .	9
2.3	Architecture of Memory . . . . .	10
2.4	Static RAM Cell . . . . .	11
2.5	SRAM cell with Polysilicon Load Devices . . . . .	12
2.6	6T Memory Cell . . . . .	13
2.7	Read Operation . . . . .	14
2.8	Write Operation . . . . .	15
2.9	Typical read cycle diagram for SRAM . . . . .	16
2.10	Typical write cycle diagram for SRAM . . . . .	18
2.11	Column Pull Up Configuration . . . . .	20
2.12	Column Multiplexing . . . . .	22
2.13	Latch Based Sense Amplifier . . . . .	23
2.14	Basic 256 X 256 SRAM Architecture . . . . .	24
2.15	Split Bank Architecture for 64K SRAM . . . . .	24
2.16	Block Diagram of Access Time . . . . .	25
3.1	Symbol . . . . .	30
3.2	Block Diagram . . . . .	31
3.3	Detail Life Cycle of Project . . . . .	32
4.1	Local Decoder Modeling . . . . .	35
4.2	Local Core Modeling . . . . .	36
4.3	Common Sense Generation . . . . .	37
4.4	Matching of the bitline delay with the inverter chain delay and the replica cell bitline delay across process fluctuations over varying threshold offsets for the accessed memory cell. . . . .	38
4.5	Latch Type Sense Amplifier . . . . .	40
4.6	Replica Bit Line Column . . . . .	40
4.7	Control circuits for sense clock activation and wordline pulse control. . . . .	41
4.8	Delay Matching of Two Buffer Chain . . . . .	42
4.9	Current Ratio Based Replica Structure . . . . .	43
5.1	Block Diagram for Post Layout Extraction . . . . .	46

5.2	Different Extraction Methodology . . . . .	47
5.3	Single Memcell Extraction . . . . .	48
5.4	Block Diagram of MCF . . . . .	49
6.1	Memcell for SNM . . . . .	52
6.2	Memcell for WM . . . . .	53
6.3	Stimuli for Discharge Rate . . . . .	54
6.4	Stimuli for ION . . . . .	55
6.5	Dynamic Nand Gate . . . . .	58
6.6	Stimuli for Node A . . . . .	58
6.7	Schematic of IO Latch . . . . .	60
6.8	Stimuli for Net BL . . . . .	61
6.9	Stimuli for Net BH . . . . .	62
6.10	IO Latch with Noise Voltage Sources . . . . .	62
6.11	Power On Analysis . . . . .	66
6.12	Cross coupled ( latch type ) Sense Amplifier . . . . .	67
7.1	Results of SNM of Memcell . . . . .	74
7.2	Results of Write Margin of Memcell . . . . .	74
7.3	Results of Discharge Rate of Memcell . . . . .	75
7.4	Waveform of DR with MC Simulation . . . . .	75
7.5	Variation of DR with 400 MC Simulation . . . . .	76
7.6	Results of ION Currents of Memcell . . . . .	76
7.7	Results of IOFF Currents of Memcell . . . . .	77
7.8	Waveform of Leakage Validation of Row Decoder . . . . .	77
7.9	Waveform of Strength Validation of Row Decoder . . . . .	78
7.10	Waveform of Charge Sharing Validation of Row Decoder . . . . .	78
7.11	Waveform of Bump Validation of Row Decoder . . . . .	79
7.12	Waveform of Leak_1 of IO Latch Analysis . . . . .	79
7.13	waveform of Leak_0 of IO Latch Analysis . . . . .	80
7.14	waveform of Write_0 of IO Latch Analysis . . . . .	80
7.15	waveform of Write_1 of IO Latch Analysis . . . . .	81
7.16	waveform of SNM_0 of IO Latch Analysis . . . . .	81
7.17	waveform of SNM_1 of IO Latch Analysis . . . . .	82
7.18	Waveform of Access Time . . . . .	82
7.19	waveform of Sense Amplifier Offset . . . . .	83
7.20	waveform of Sense Amplifier Offset(zoom) . . . . .	83
7.21	Waveform of SA Pulse Width Analysis . . . . .	84
7.22	Waveform of Write Self Time Analysis . . . . .	84

# Chapter 1

## Introduction

### 1.1 Objective and Scope of the Project

Digital systems require the capability of storing and retrieving large amounts of information at high speeds. Memories are circuits or systems that store digital information in large quantity, hence are vital components in modern integrated circuits. Manufacturers of such products demand low-priced memories with low-power consumption, high-speed operation, high density, and small package size.

The semiconductor markets have embraced the fact that the architecture of the memory structure has a considerable impact on the performance of the system. Over the years, technology advances have been driven by memory designs of higher and higher density.

Memory developers have to design memories to address the issues in Bandwidth, Latency, Density, Speed, Power and Cost. Unfortunately, it is not possible for a single memory technology to address all these issues with distinct advantages - this translates to an arsenal of components available for designers to design their system.

The criticality of my project work is to characterize the memory compiler to achieve the target to adjust the things according to international technology road-map for semiconductor in the field of embedded memory. This is somewhat specific

to user requirements.

Memory design takes 40% of total time and its verification takes 60% of total time. To ensure that the designed memory when fabricated work correctly it must be verified for critical issues with worst case possibilities.

## 1.2 Applications

The SRAM Compiler generates different memory cuts according to the customer specifications. These SRAMs are used in various Automotive Products. These memories are used in various automatic parts of car like rain sensor, adaptive cruise control, water pump, LIN battery sensor, mirror control, active suspension, climate control, parking system, central body module, power window, car multimedia, navigation and telematics system, engine control, charging, GPS module, satellite radio, RF tyre pressure system, dash board control, airbag control, engine control, vehicle stability control, power steering, seat control, light sensor, door and roof zone systems, digital mobile video. These memories are also used in mobile processing systems. These memories are also used in various computer periphery applications.

## 1.3 Definition of Problem

This project work dealt with the Design Validation of Single Port SRAM Compiler. It includes various analysis and measurements for a given memory compiler to ensure correct functionality for various process corners at given temperature and supply voltage ranges. Here different types of characterization of the parameters like timings, noise margin and capacitance for various memory cuts are performed, so feedback is provided to the designer if any discrepancy in timing, capacitances or dimensional trend is found to improve the design performance.

## 1.4 Scope of Work

- Study of basic concepts of Memory.
- Study of Critical Path and design concept of the SRAM.
- Study of Replica Technique for Low Power SRAM
- Post-Layout Extraction.
- Memcell Validations
- Row Decoder Validations.
- IO Latch Analysis
- Marginality Analysis
- Power on Analysis
- Access Time Analysis
- Sense Amplifier Offset Analysis
- Sense Amplifier Pulse Width Analysis
- Write Self Time Analysis
- Read Self Time Analysis

## 1.5 Implementation Environment

- **Platform**
  - Linux
  - Solaris
- **Simulators**

- ELDO (Mentor Graphics Corporation)
- NANOSIM (Synopsys)

- **EDA Tools**

- Cadence OPUS
- Cadence Virtuoso
- PLSKit
- MCF Tool (ST Version)

- **Frame Work**

- SPY Wave (ST Version)
- Xelga
- EZwave
- Cosmos Scope (Synopsys)

- **Technology**

- a. M10 ( 90 nm )

## 1.6 Organization of Report

This project report is organized into eight chapters.

**Chapter 2**, *Exhaustive Survey Of Memory* , mainly describes the basic theory of memory, working principles, read and write operation and different architectures of the memory.

**Chapter 3**, *Development Flow Of SRAMS*, presents various features of the given compiler. This chapter includes the symbol and block diagram of the memory instance. It also consists functional discription and detail life cycle of the project.



**Chapter 4**, *Implementation*, presents the critical path modeling of the memory and replica techniques for Wordline and Sense Control in Low Power SRAM.

**Chapter 5**, *Extraction And MCF*, presents the details of the post layout extraction and memory characterization flow. It also explores the various extraction options available.

**Chapter 6**, *Analysis Work*, presents discription of various analysis performe for validation of the compiler. It shows how the different analusis are carried out.

**Chapter 7**, *Results & Waveforms*, presents the results and waveforms of the different analysis.

Finally, in **Chapter 8** concluding remarks and scope for future work is presented.

# Chapter 2

## Exhaustive Survey Of Memory

Modern digital systems require the capability of storing and retrieving large amounts of information at high speeds. Memories are circuits or systems that store digital information in large quantity.

Today, memory circuits come in different forms including SRAM, DRAM, ROM, EPROM, E2PROM, Flash, and FRAM. While each form has a different cell design, the basic structure, organization, and access mechanisms are largely the same. Dual-port memories have the ability to read/write two words in one cycle.

### 2.1 MOS Memories

The ideal memory would be low cost, high performance, high density, with low power dissipation, [10] random access, non-volatile, easy to test, highly reliable, and standardized throughout the industry. The MOS memories fall into two broad categories:

**Read-Write memories:** Dynamic RAMs and Static RAMs, allow the user both to read information from the memory and to write new information into memory while it is still in the system.

**Read Only Memories:** ROMs, EPROM's, EEPROMs, are used primarily to store data; however, the EEPROMs can also be written into a limited number of

times while in the system. Read-Only memories are non-volatile, that is, they retain their information stored in it even if the is turned off.

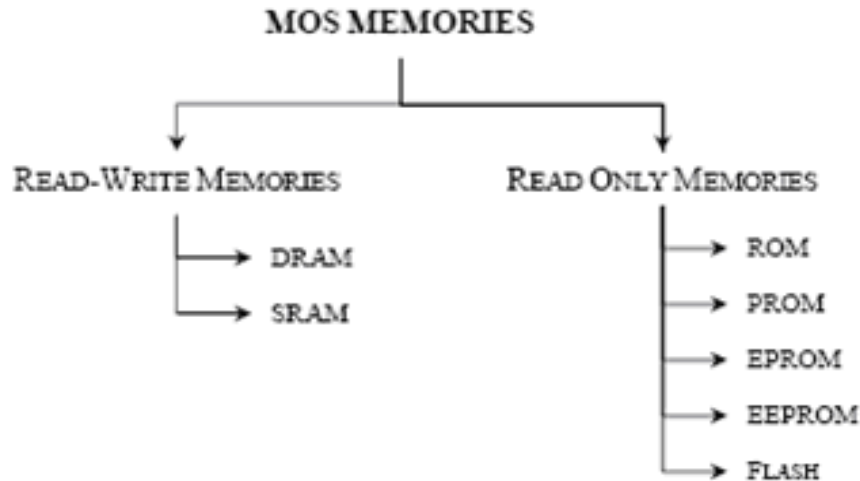


Figure 2.1: MOS Memories

### 2.1.1 Read-Write Memories

Read-write random-access memories (RAM) may store information in flip-flop style circuits or simply as charge on capacitors. Because read-write memories store data in active circuits, they are volatile; that is, stored information is lost if the power supply is interrupted. The natural abbreviation for read-write memory would be RWM. However, pronunciation of this acronym is difficult. Instead, the term RAM is commonly used to refer to read-write random-access memories.

The two most common types of RAMs are the static RAM (SRAM) and the dynamic RAM (DRAM). Static RAMs hold the stored value in flip-flop circuits as long as the power is on. SRAM tends to be high-speed memories with clock cycles in the range of 5 to 50 ns. Dynamic RAMs store values on capacitors. They are prone to noise and leakage problems, and are slower than SRAM, clocking at 50 ns to 200

ns. However, DRAMs are such denser than SRAMs, up to four times denser in a given generation of technology.

### 2.1.2 Read-Only Memories

Read-only memories (ROMs) [10] store information according to the presence or absence of transistors joining rows to columns. ROMs have read speeds comparable to those for read-write memories.

All ROMs are nonvolatile, but they vary in the method used to enter (write) stored data. The simplest form of ROM is programmed when it is manufactured by formation of physical patterns on the chip; subsequent changes of stored data are impossible. These are termed mask-programmed ROMs. In contrast, programmable read-only memories (PROMs) have a data path present between every row and column when manufactured, corresponding to a stored 1 in every data position. Storage cells are selectively switched to the 0 state once after manufacture by applying appropriate electrical pulses to selectively open (blow out) row-column data paths. Once programmed, or blown, a 0 cannot be changed back to a 1.

## 2.2 Memory Organization

The preferred organization for most large memories is shown in Figure 2.2. This organization is a random-access architecture. The name is derived from the fact that memory locations (addresses) can be accessed in random order at a fixed rate, independent of physical location, for reading or writing. The storage array, or core, is

made up of simple cell circuits arranged to share connections in horizontal rows and vertical columns [8]. The horizontal lines, which are driven only from outside the storage array, are called wordlines, while the vertical lines, along which data flow into and out of cells, are called bitlines. A cell is accessed for reading or writing by selecting its row and column. Each cell can store 0 or 1. Memories may simultaneously

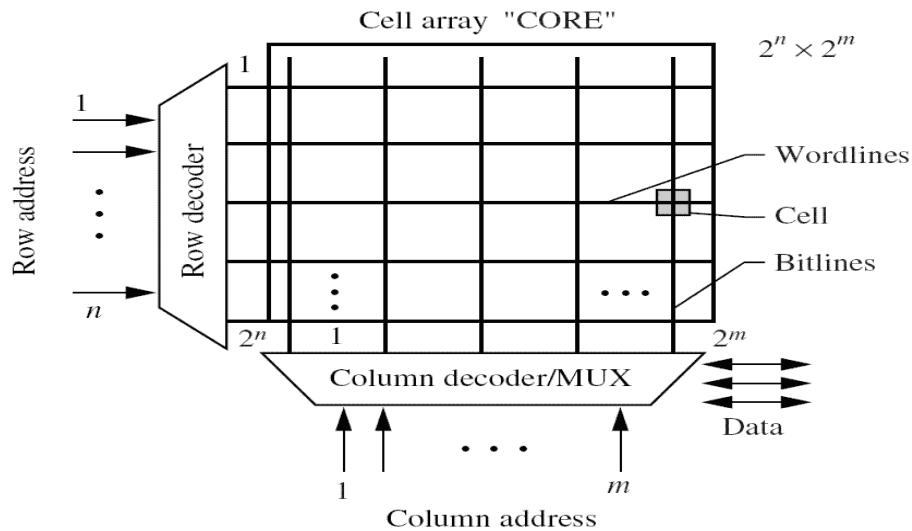


Figure 2.2: Memory System Organization

select 4, 8, 16, 32, or 64 columns in one row depending on the application. The row and column (or group of columns) to be selected are determined by decoding binary address information. For example, an  $n$ -bit decoder for row selection, as shown in Figure 2.3, has  $2n$  output lines, a different one of which is enabled for each different  $n$ -bit input code. The column decoder takes  $m$  inputs and produces  $2m$  bitline access signals, of which 1, 4, 8, 16, 32, or 64 may be enabled at one time. The [8] bit selection is done using a multiplexer circuit to direct the corresponding cell outputs to data registers. In total,  $2n \times 2m$  cells are stored in the core array. An overall architecture of a 64 Kb random-access memory is shown in Figure 2.3. For this example,  $n = m = 8$ . Therefore, the core array has a total of 65,536 cells. The memory uses a 16-bit address to produce a single bit output. Memory cell circuits can be implemented in a wide variety of ways.

### 2.2.1 Functional SRAM Chip Model

Basic Description of Above Block :

- a. Block 'A', address latch, contains the address [10].

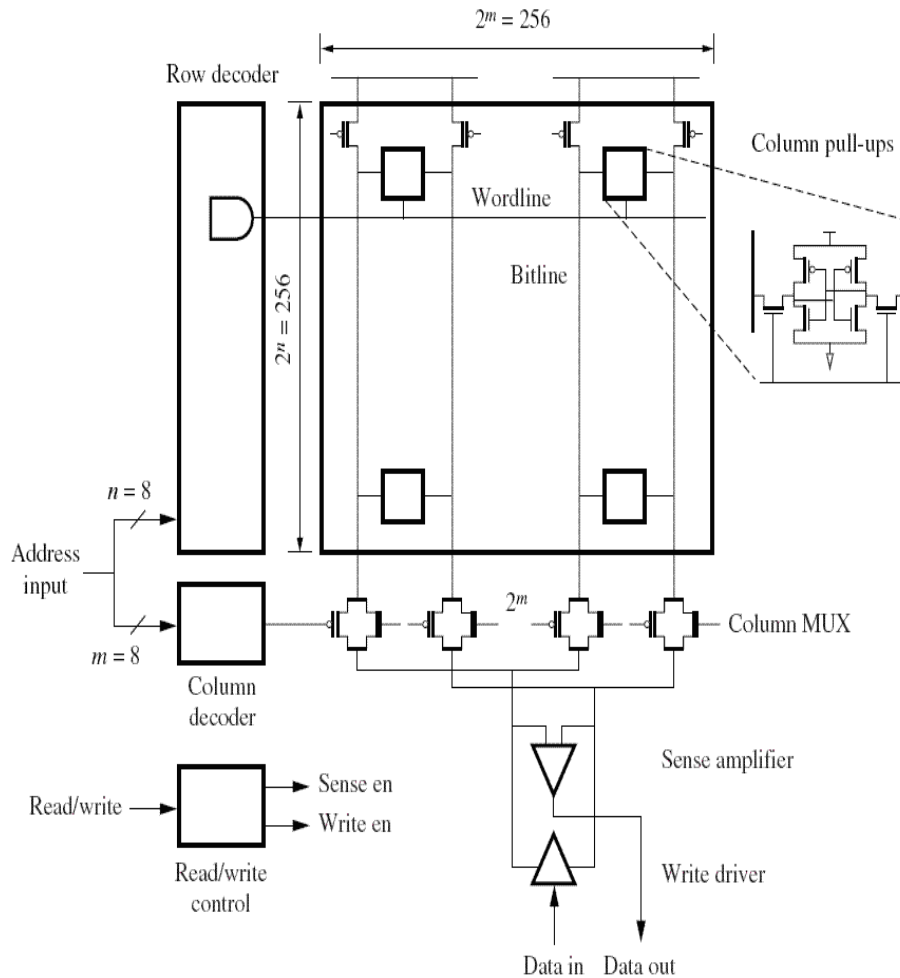


Figure 2.3: Architecture of Memory

- b. The higher order bits of the address are connected to the row decoder 'B', which selects a row in the memory cell array 'D'.
- c. The lower order address bits go to the column decoder 'C', which selects the required columns. The number of column selected depends on the data width of the chip, that is the number of data lines of chip, which determines how many bits can be accessed during a read or write operation
- d. When the read/write line indicates read operation, the contents of the selected cells in the memory cell array are amplified by the sense amplifiers 'F', loaded

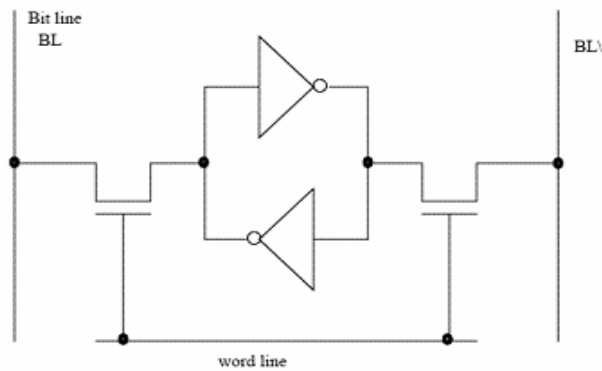


Figure 2.4: Static RAM Cell

into the data register 'G' & presented on the data-out line(s).

- e. During a write operation the data on the data-in line(s) are loaded into the data register & written in to the memory cell array through the write driver 'E'. Usually the data-in & data-out lines are combined to form bidirectional data lines, thus reducing the number of pins on the chip.
- f. The chip-enable line enables the data register & together with read/write line, the write driver.

## 2.3 SRAM Cell Description

### 2.3.1 Static RAM Cell

A fully static RAM cell is a bistable circuit, capable of being driven into one of two states. After removing the driving stimulus, the circuit retains its state. Thus we say the cell is "static" since it does not need to have its data refreshed as long as the dc power is applied. SRAM memory cells are based on the latch structure with two back-to-back connected inverters & two pass transistors. There are two kinds of SRAM memory cells- the 6T cell & the 4T cell. Basic SRAM cell looks like:

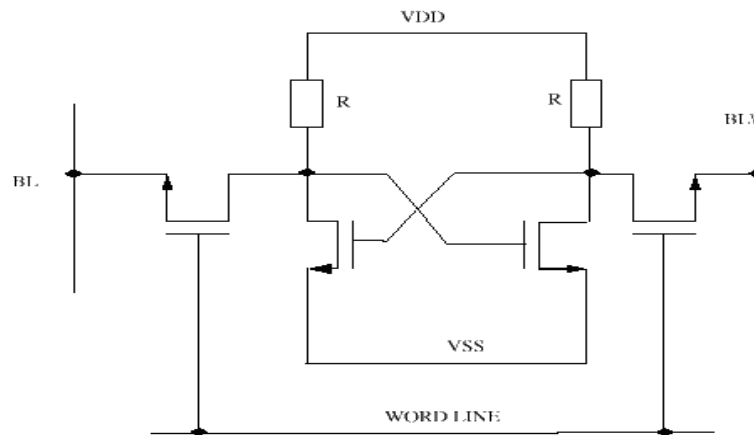


Figure 2.5: SRAM cell with Polysilicon Load Devices

SRAM cell using polysilicon load devices "R": A **4T memory cell** The resistors are made from polysilicon with a high resistivity ( they may have a value of 100 GOHM), causing asymetry in the logic 1 and 0 drive power of latch. A logic 0 level is caused by a conducting transistor, which allows for relatively low ohmic path. A logic 1 level is caused by a non-conducting transistor and has to be maintained by the high ohmic polysilicon resistor [10], which is dimentioned such that it will be able to sup-ply the relatively small leakage current of the non-conducting resistor.

**Advantages:**

- a. Polysilicon resistors in the 4T memory cell are located above the four NMOS devices to save layput area.
- b. Two polysilicon resistors of 1GOHM generated by "ion implantation" are used in 4Tmemory cell to reduce the standby current.
- c. Using the polysilicon resistors & gate structure, double-poly CMOS technology is required to implant SRAMs.

**Problems:**

- a. Smaller noise margin.



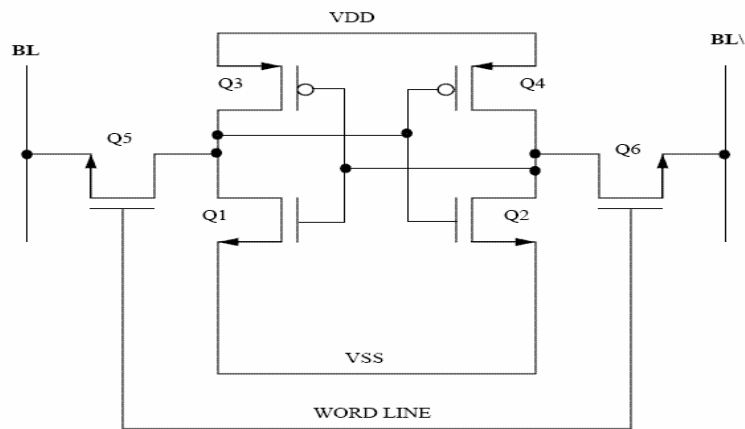


Figure 2.6: 6T Memory Cell

- b. Smaller soft error immunity.
- c. Pull-up of the internal node is via the high resistive load. In order to raise the speed performance, the resistive load should be as small as possible. With the reduced resistive load, the consumed current may rise, thus increasing the power consumption.
- d. Lower stability than 6T cell & decreases as VDD is scaled down.

### 2.3.2 CMOS SRAM Cell (6T Memory Cell)

The polysilicon load devices are replaced by PMOS enhancement mode transistors Q3 & Q4 [10]. This further reduces the power requirements of the cell, except for small leakage current, no power will be dissipated during the time the cell retains the stored logic value. Good static noise margin & stability. Stability is held when VDD is scaled down. The disadvantage of CMOS technology is that it requires more processing steps because of the presence of NMOS & PMOS transistors & requires large area.

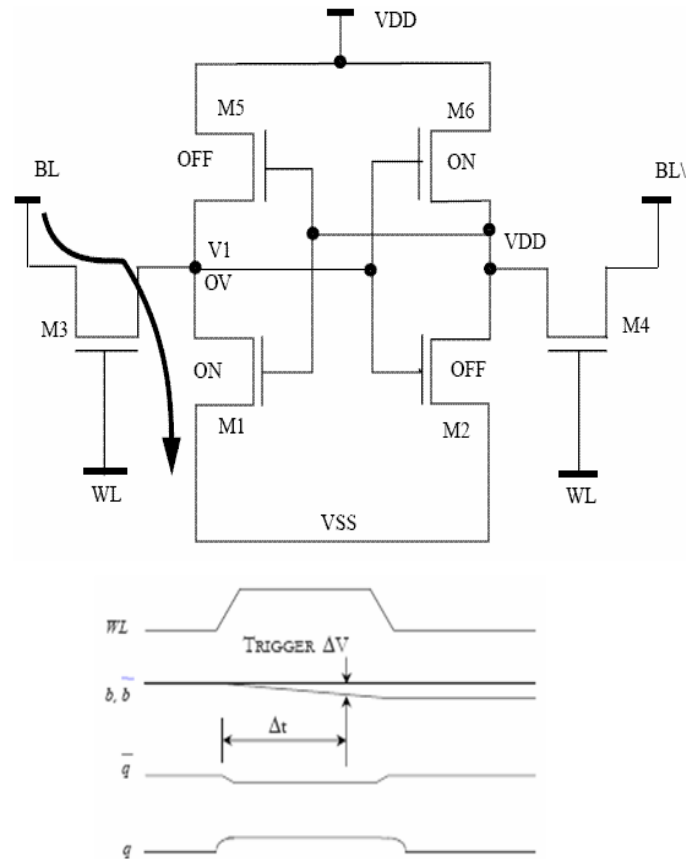


Figure 2.7: Read Operation

## 2.4 Read Operation

The Data-read operation must not destroy the stored information in the cell. The above fig. shows the SRAM at the beginning of the read operation. To prevent spurious read the following conditions must be verified [10]:

- $V1 < V_{tn2}$  (N-channel threshold voltage of M2)
- M3 operates in saturation region, M1 in linear region.

Following activities take place during read cycle:

- First the bit lines are precharged to VDD.

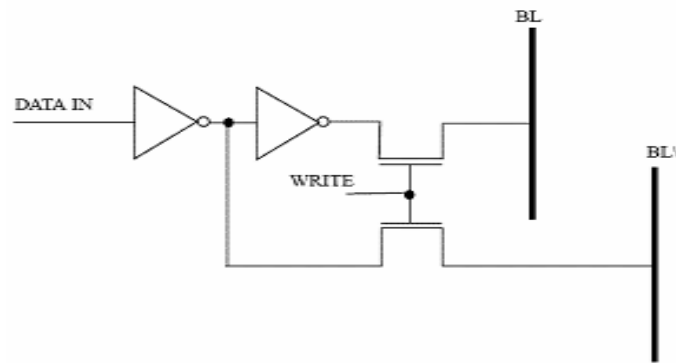


Figure 2.8: Write Operation

- Then the word line is made high. ( Note the precharge is switched OFF prior to WL goes high).
- Enabling word line causes one of the bit line to discharge while the other remains high. BL &  $\overline{BL}$  will discharge, depending on the prior voltage. Like in above case, M3 will switch ON & the current discharges through the path as indicated above. Note the Bit lines has a much higher capacitance than than the capacitor of a single cell, as Bit line is rather long & is connected to many cells. Also note very little current flows through M4.
- The level of BL as a result become lower than  $\overline{BL}$  & this differential signal is detected the differential amplifier(sense amplifier) connected to the BL &  $\overline{BL}$  lines.
- The signal is amplified & fed to the output buffer.

## 2.5 Write Circuitry of SRAM

- Typical write circuitry consists of pair of invertors & a pass gate with write control input signal to the BL &  $\overline{BL}$  lines [10].

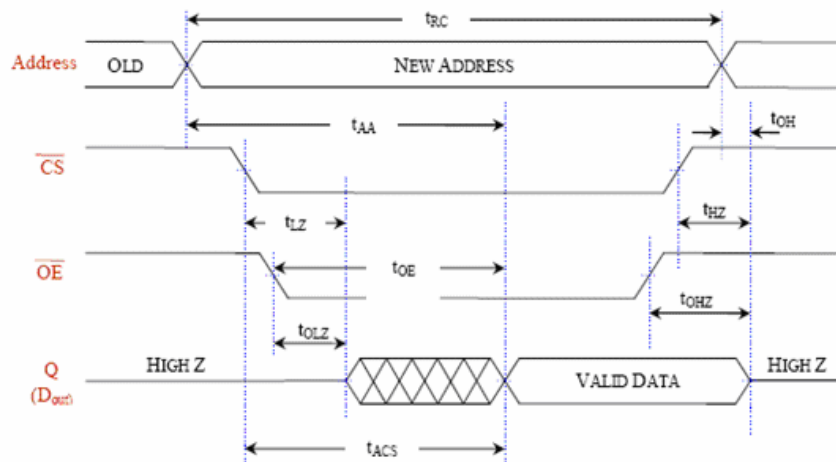


Figure 2.9: Typical read cycle diagram for SRAM

- Data can be written by driving WL(write) high & driving the line BL &  $\overline{BL}$  with data with complementary values.
- Because the bit lines are driven with more than the force with the cell retains its information(the transistors driving the bit lines are more powerful), hence the cell will be forced to the state presented on the lines BL &  $\overline{BL}$ .

## 2.6 Timings Waveforms

### 2.6.1 Read Cycle Timings

A timing diagram for a basic read cycle during which the system reads out information That is stored in a static RAM for a wide bus, common I/O SRAM with chip Enable, and output enable functions. It consists of

- System selects the RAM by turning the chip select on ( $\sim$ CS low)
- System sets the correct addresses (A set).

- System turns the output enable on ( $\sim$ OE low)
- System must make sure that the time that old data from other sources is still on the common I/O bus is less than the minimum of output enable low to output active  $t_{OLZ}$  or chip enable low to output active  $t_{LZ}$ .
- The system must wait a minimum time of address access time  $t_{AA}$  in order to be sure of correct data

The expansions of the acronyms for read timing parameters are:

$t_{RC}$  : Read Cycle Time

$t_{AA}$  : Address Access Time

$t_{ACS}$ : Chip Enable Access Time

$t_{OE}$  : Output Enable Access Time

$t_{OH}$  : Output Hold from Address Change

$t_{LZ}$  : Chip Enable Low to Output Enable

$t_{OLZ}$ : Output Enable Low to Output Enable

$t_{HZ}$  : Chip Enable High to Output Enable - Z

$t_{OHZ}$ : Output Enable High to Output Enable - Z

## 2.6.2 Write Cycle Timings

A simple write cycle for changing data in an SRAM (writing into it) is shown below.

It consists of

- System sets the correct addresses (A set).
- System selects the RAM by turning the chip select on ( $\sim$ CS low).
- System waits a minimum required amount of time after changing the addresses for the RAM to do internal 'set up' of the addresses  $t_{AS}$ , and then turns the write enable on ( $\sim$ WE low).

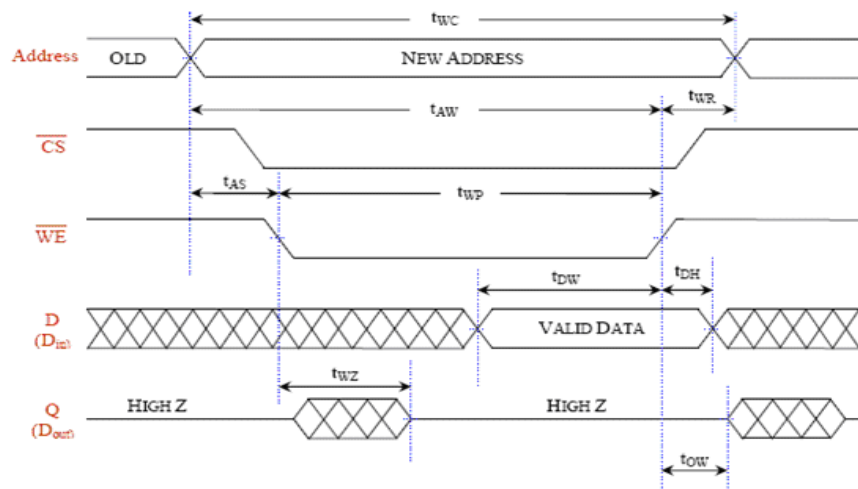


Figure 2.10: Typical write cycle diagram for SRAM

- System waits a minimum required amount of time after turning on the write enable  $t_{WZ}$  for the memory to disable the data output driver 'Q' in preparation for using these lines for data input.
- System inputs the new data and waits a minimum required amount of time for the memory to write the data before turning off the write enable  $t_{DW}$ .
- System waits a minimum required amount of time after turning the write enable on before turning it off  $t_P$ . This is to be sure the write enable pulse width is wide enough for correctly writing the data into the RAM.
- System waits a minimum required amount of time after turning write enable off
- Before changing the data  $t_{DH}$ . This is called the 'data hold time', and it ensures that the data is stable during the entire write cycle.
- Before changing addresses to start the next cycle  $t_{WR}$ . This is called the 'write recovery time' and ensures that the addresses are stable during the entire write cycle.

- Makes sure the data have disappeared before the RAM turns the data output drivers back on ( $t_{ow}$ ) [10].

The expansions of the acronyms for read timing parameters are:

$t_{WC}$  : Write Cycle Time

$t_{AS}$  : Address Setup Time

$t_{AW}$  : Address Valid to End of Write

$t_{WP}$  : Write Pulse Width

$t_{DW}$  : Data Valid to End of Write

$t_{DH}$  : Data Hold Time

$t_{WZ}$  : Write Enable Low to Output High - Z

$t_{OW}$  : Write Enable High to Output Active

$t_{WR}$  : Write Recovery Time

## 2.7 IO Block Circuitry

This section includes the column input/output (I/O) circuitry, which includes bitline precharge circuits, column multiplexers [8].

### 2.7.1 Column Pull Up

In both read and write operations, the bitlines are initially pulled up to a high voltage near VDD. The circuits used to precharge the bitlines depend on the type of sensing that is used in the read operation [8]. Figure 2.11 illustrates three possible precharge configurations. In Figure 2.11a, precharge signal, PC, is applied to the two pull-ups and to a third transistor, called the balance transistor, connected between the two bitlines to equalize their voltage levels. When the wordline (wl) signal goes high, one bitline remains high and the other falls at a linear rate until wl goes low [8]. The difference between the bitlines is fed into a voltage-sensing latch-based amplifier that

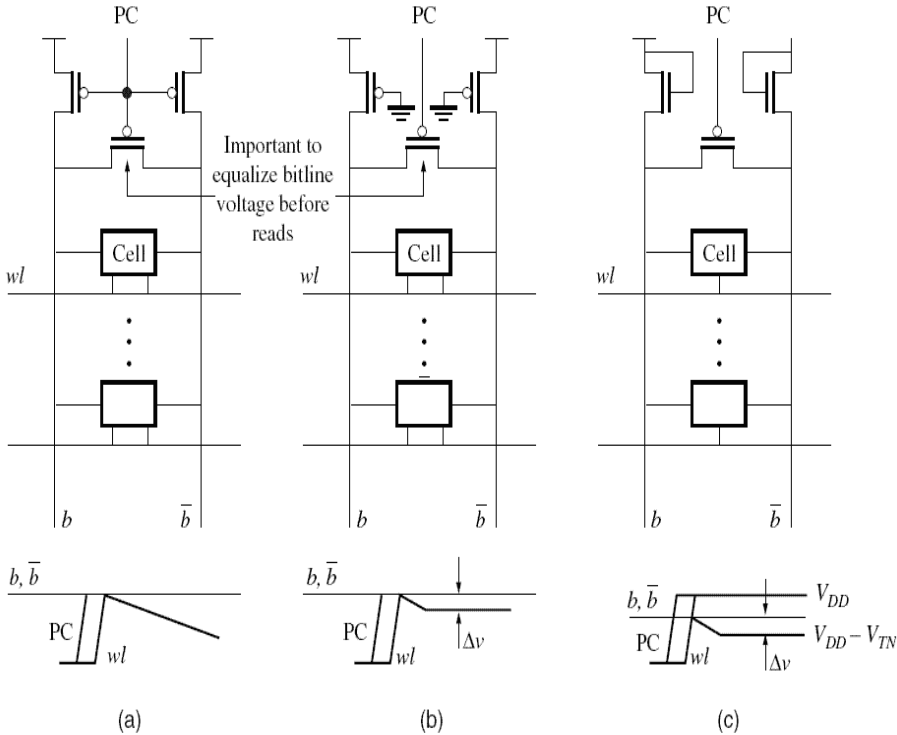


Figure 2.11: Column Pull Up Configuration



is triggered when the differential voltage exceeds a certain threshold. The precharge circuit of Figure 2.11b is reminiscent of the pseudo-NMOS circuits. Two static loads and a balance transistor form the precharge circuit. When PC is applied to the balance transistor, it simply equalizes the two voltage levels. Once the bitlines are precharged, the PC signal is turned off (raised to VDD) and, at this point, the wordline can be activated. Of course, the pull-ups are still on so current will flow through one of them and into the cell side with the stored "0." Eventually, a steady-state output level will be reached by the bitline, as shown in the figure. This type of pull-up is suitable for current-sensing amplifiers since there is continuous current flow, or latch-based voltage-sensing amplifiers since the bitlines will establish a differential voltage,  $dv$  [8].

Figure 2.11c is based on the NMOS saturated enhancement load. Therefore, the maximum possible voltage on the bitline is  $VDD - VT$ . When PC is applied to the balance transistor, it equalizes the two voltage levels. Once the lines are precharged high, the PC signal is turned off (raised to VDD) and then  $wl$  goes high. At this point, the pull-ups are still active so current will flow through one of them into the cell side with the stored "0." Again, a steady-state output level will be reached by the corresponding bitline, as shown in the figure, although this value will be lower than the pseudo-NMOS case. This type of pull-up is suitable for differential voltage sensing amplifiers since the bitline voltages initially start at  $VDD - VTN$ . This lower voltage is needed for a proper biasing and output swing of the differential amplifier [8].

### 2.7.2 Column Multiplexing

Once all the columns have been pulled up to a high voltage, the next step is to select the column(s) that will be involved in the read or write operation. This column selection is performed using a decoder/multiplexer combination. The  $m$ -bit column address is used to select one or more of the  $2m$  columns [8]. This can be performed

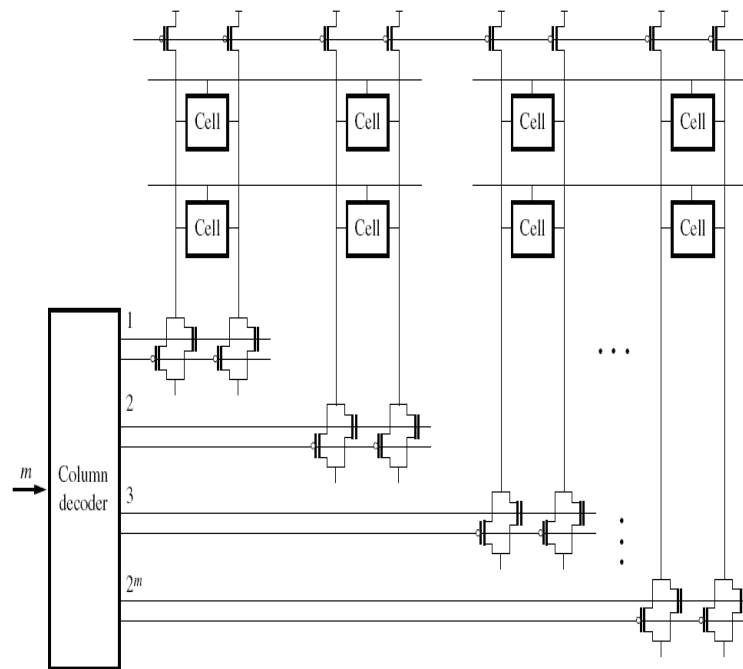


Figure 2.12: Column Multiplexing

using a decoder, similar to the row decoder, driving a number of CMOS pass transistors as shown in Figure 2.12. The pass transistors require complementary signals. Of course, if an 8-bit output is desired, then the decoder outputs would each drive eight CMOS transmission gates, and fewer column address bits would be needed [8].

### 2.7.3 Latch Based Sense Amplifier

The circuit is effectively a cross-coupled pair of inverters with an enabling transistor, M1. This circuit relies on the (slower) regenerative effect of inverters to generate a valid high or low voltage. It is a lower power option since the circuit is not activated until the required potential difference has developed across the bitlines. However, it is slower since it requires a large input voltage difference and is not as reliable in the presence of noise [8]. The bitlines are precharged to VDD. Then the wordline is enabled and one of the bitlines drops in voltage. As the bitline differential voltage

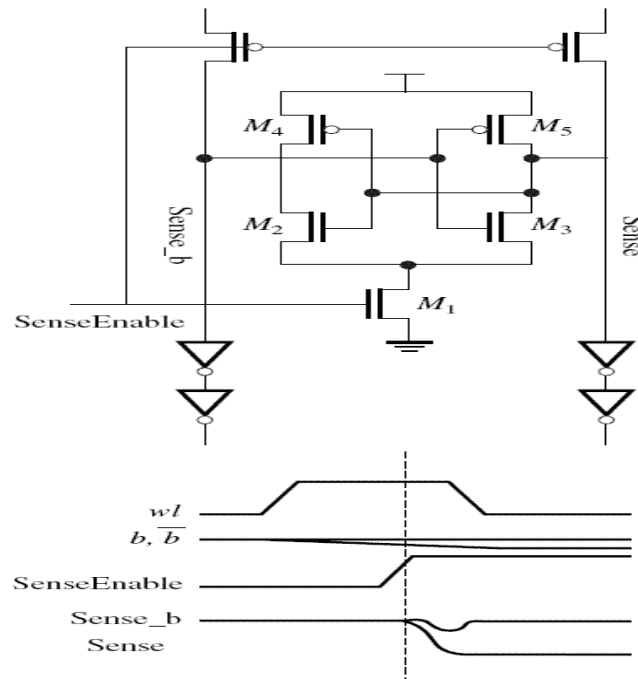


Figure 2.13: Latch Based Sense Amplifier

reaches the prescribed amount, the sense enable is activated. The timing of the sense enable is critical. For now, assume that it arrives at the proper time. At this point, the bitline difference is fed into the cross-coupled inverters. One side drops in voltage faster than the other side, since one side will always have more gate overdrive than the other. When the voltage drops below  $V_{TN}$  on one side, it turns off the pull-down transistor of the opposite side, and the pull-up transistor acts to raise the voltage to VDD. This regenerative process is shown in the timing diagram of Figure 2.13 [8].

## 2.8 SRAM Architecture

### 2.8.1 Basic Architecture

In the conventional architecture when the selected word line is high, all the cells connected to the wordline in the row are active. When the word line is high, all the

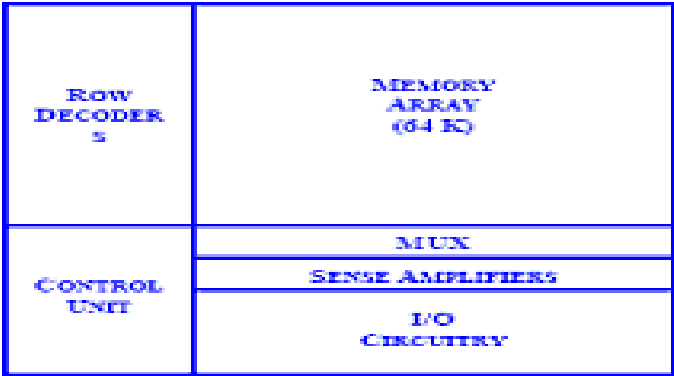


Figure 2.14: Basic 256 X 256 SRAM Architecture

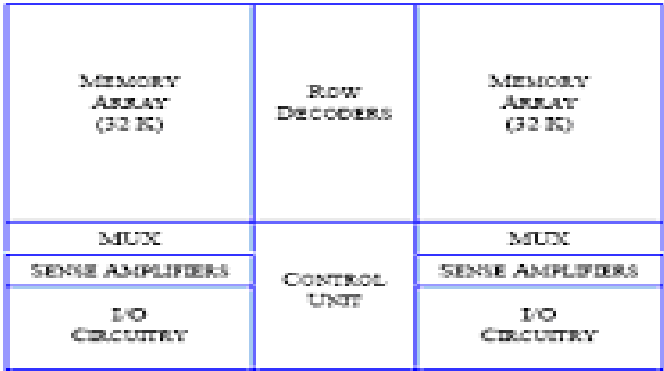


Figure 2.15: Split Bank Architecture for 64K SRAM

cells connected to the wordline become active – thus dissipation increases. In the basic architecture, the two major factors contribute to the read access are the bit access time & the word line access time. When the size of the SRAM increases, the number of cells connected to the word line increases the load is reduced. Therefore, the wordline delay increases because of the increase in the wordline capacitance. These two factors can be improved by reducing the bit line capacitance & the word line capacitance, but this is achieved only after using a different architecture [10].

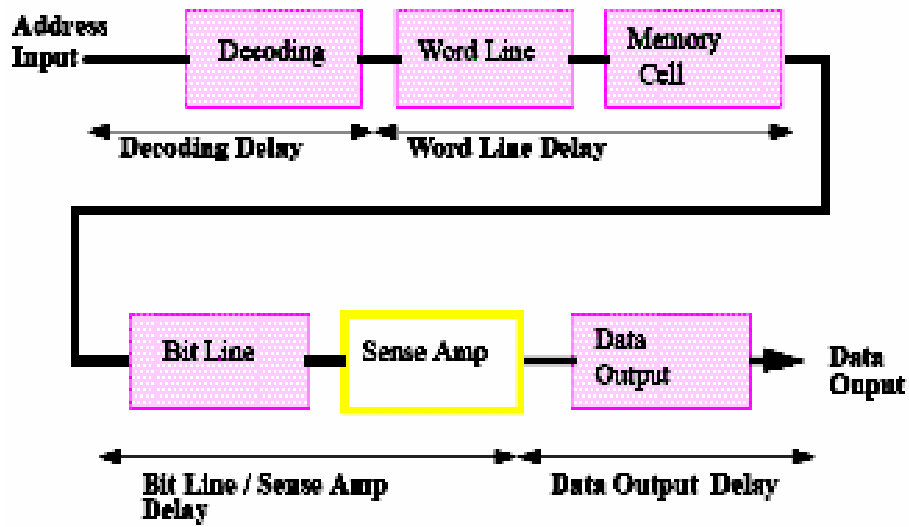


Figure 2.16: Block Diagram of Access Time

## 2.8.2 Split Core Architecture

In this type of architecture, reduction is performed by splitting the matrix in smaller blocks. The resulting architecture is called Split-Core architecture. The reduction in the RC delay is observed because of the split bank, but here too the activation of a wordline activates the entire cell in both of the core areas. So certainly, there is need of a different architecture, which could also provide some advantage in terms of power dissipation.

## 2.9 Access Time

The Access Time is determined by the critical path from the Address input to the Data output as shown in Figure 2.16. It is the sum of following delays [10] : Decoder Delay, Word-Line Delay, Sensing Delay And Data Output Delay.

## 2.10 Monte Carlo Analysis

Monte-Carlo (MC) analysis is a series of DC, AC, Transient analysis, where one or more circuit or model parameters follows a probability distribution. The distribution is uniform, Gaussian, or User-defined. This kind of analysis can be useful for yield analysis. Each MC parameter is assigned a nominal value, a standard deviation (expressed in an absolute value or a relative percentage of the nominal), and LOT/DEV correlation. These are the available probability functions:

**LOT/DEV Correlation** MC parameters can be correlated or not. For example, capacitors in an IC Design may have correlation, e.g. their capacitance values tend to rise or fall together. This is a LOT correlation. However, components on the printed circuit board tend not to be correlated, which is DEV correlation. Parameter can have both LOT/DEV variations. Using both means that affected devices have a degree of independence, even though they are correlated. If a parameter is not a primitive, but depends on parameters with no LOT/DEV specifications, then a LOT/DEV specification can be set on that parameter. A LOTGROUP can be defined to share the same distribution between dissimilar elements. Once a LOTGROUP is defined, it is used in the same way as LOT or DEV. When defining a parameter using MC distribution variation on the parameter differs depending on where the parameter is specified. LOT variation is used when a defined parameter affects model parameter, this means the same random values will be used each time it affects a separate model parameter. DEV variation is used when a parameter affects instance parameters, an independent random value is calculated each time the parameter is specified. MC analysis, when use with .EXTRACT statements, will produce Histograms as a result.

# Chapter 3

## Development Flow Of SRAMS

### 3.1 Features of Memory Compiler

The compiler provides many features in the default configuration and also provides flexibility to the System designer to choose from a set of optional features [10].

#### 3.1.1 Default Features

- Low leakage 3 metal level memory.
- Fully synchronous operation
- Functional voltage range is from 1.0V to 1.45V.
- Functional temperature range is from -40c to 150c
- Worst case Access time of 4.4ns
- Area is 0.12085mm<sup>2</sup>
- Generator range is from 4Kb to 312Kb.
- Word range is from 512 to 8192.
- Word length is 8 bits to 78 bits [10].

### 3.1.2 Optional Features

- Two mux options - Mux8 and Mux16.
- Option for Power-down mode.
- Option for having two VDD domains.
- Two types of transistor options: HVT & MVT.
- Two types of abstract options - Small Pin abstract and Ring Type abstract [10].

### 3.1.3 Feature Description

**Words Parameter :** The Words parameter defines how many address are physically accessible. The aspect ratio selector MUX gives the increment of words. The architecture also imposes the number of rows to be a multiple of four.

**Bits Parameter :** The Bits parameter defines how many bits are contained in each address location. Up to 78 bits can be generated. The number of columns is given by the product Bits \* Mux [10].

**Mux Parameter :** The Mux factor defines how many words are allocated on a single row, enabling the user to specify the aspect ratio of the RAM and hence, tailor it optimally to the floor plan. The higher the MUX is, assuming all the other parameters frozen, the larger the RAM is in the column direction and smaller in row direction.

**Bit Mask Parameter:** The Bit Mask parameter is an optional parameter. This parameter allows the user to specify if masked write operation is required. By default, this parameter is set to No , implying that mask bits are not provided with every data bit in the memory. For each data bit D, there is a mask bit, M. If the mask bit, M = 0, then the data on the corresponding D bit is written into the memory bit location. On the contrary, if the mask bit M = 1, then, no write operation is performed on that bit. The corresponding output pin remains stable. The mask bit, M has similar



setup and hold times as data bit with respect to the clock. Mask bit does not impact read operation anyway [10].

**Section1Words Parameter:** This option allows the user to have the memory array divided into two VDD domains. Power supply of SECTION2 (called VDD2) can be switched OFF for low standby consumption reason. The rest of the memory array (SECTION1) and the periphery is kept ON through the main power supply (called VDD). Access to SECTION1 (read or write) is always possible (even when SECTION2 is shut off). SECTION1 occupies the memory space from word zero to word (Section1Words-1). The memory partition between the two sections should be specified at cut generation level. Section1Words parameter allows the user to specify the number of words in SECTION1 [10].

**Abstract:** In this compiler, two types of abstracts are supported: small pin type and ring type. In Ring Type abstract, power supply is provided through two rings: gnd and vdd and signal pins are of small pin type. Small Pin Type abstract, signal and power pins both are of small pin type.

**Transistors Parameter:** Memory instances can be generated with HVT or MVT options: In HVT option, the core array and periphery both are in HVT devices. In MVT option, core array is in HVT devices and periphery is in SVT except few devices in HVT [10].

## 3.2 Symbol of Memory Instance

The symbol view of a memory instance with all pins is shown in the following Figure 3.1.

## 3.3 Block Diagram of Memory Instance

The Block Diagram of the memory instance is shown in Figure 3.2.

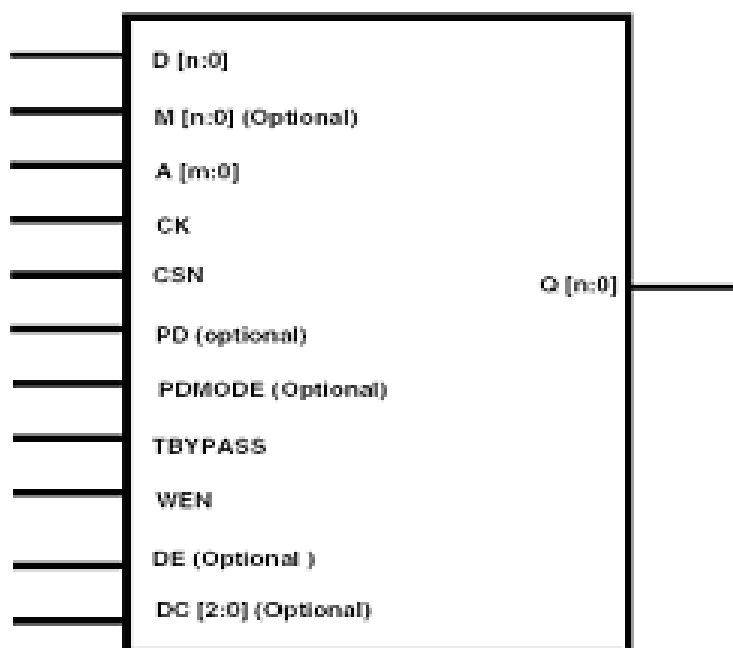


Figure 3.1: Symbol

### 3.4 Functional Description

The SP10LLC RAM operates in Synchronous mode with the rising edge of the clock, CK. Memory access is initiated by the rising edge of the clock, CK with the chip select input, CSN set low. On this clock edge, assuming that minimum setup and hold times are met, address, data and WEN signals are latched. If the latched value of write enable signal, WEN, is high, RAM goes in Read mode, else it enters Write mode[10].

When CSN is high, the memory is in standby state and the external CK cannot propagate inside the control logic circuitry, so that no memory operation is performed. In this condition, the dynamic power consumption is minimum and the output data remains unchanged. The read cycle reads a word from the memory core and places the data on the output Q at the access time,  $t_{aa}$ , following the rising edge of the clock. Output data is internally latched and remains valid till  $t_h$ , after the next

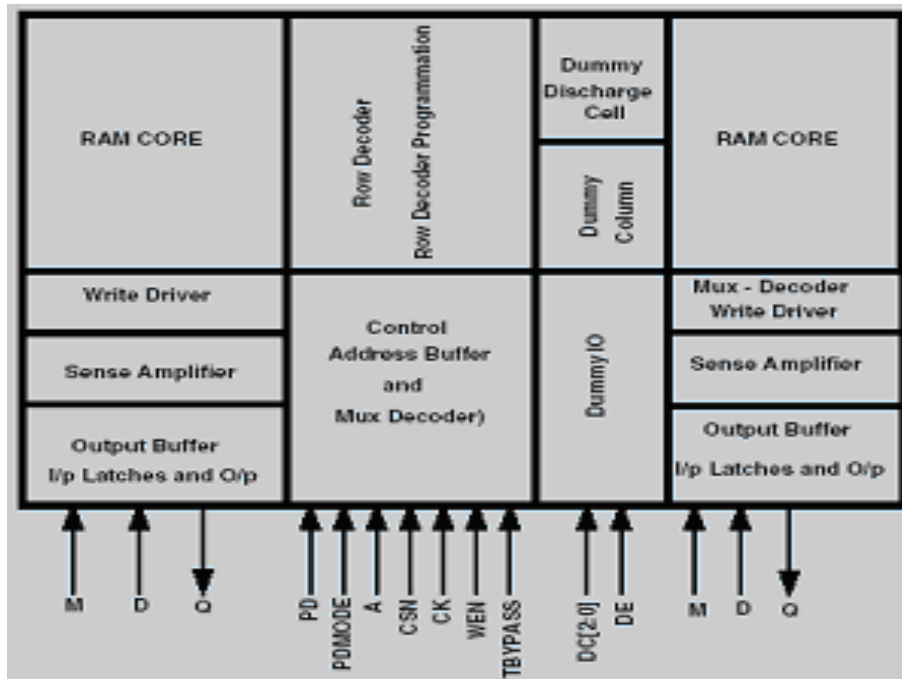


Figure 3.2: Block Diagram

positive edge of the CK. Output drivers are always enabled [10].

The generator has bit write option, implemented using mask bits, that is, for each data bit D, there is a mask bit, M. If the mask bit,  $M = 0$ , then the data on the corresponding D bit is written into the memory bit location. If the mask bit  $M = 1$ , then write operation does not occur on that bit, instead the previous data is held at the output. The mask bit M has similar setup and hold times as data bit with respect to clock. Mask bits do not impact read operation anyway. In write cycle, the output does not change its state, that is, no write-through is implemented. Access to the non-existing address space of the RAM is allowed [10].

### 3.5 Detail Life Cycle of Project

The detail life cycle of the project is shown in Figure 3.3.

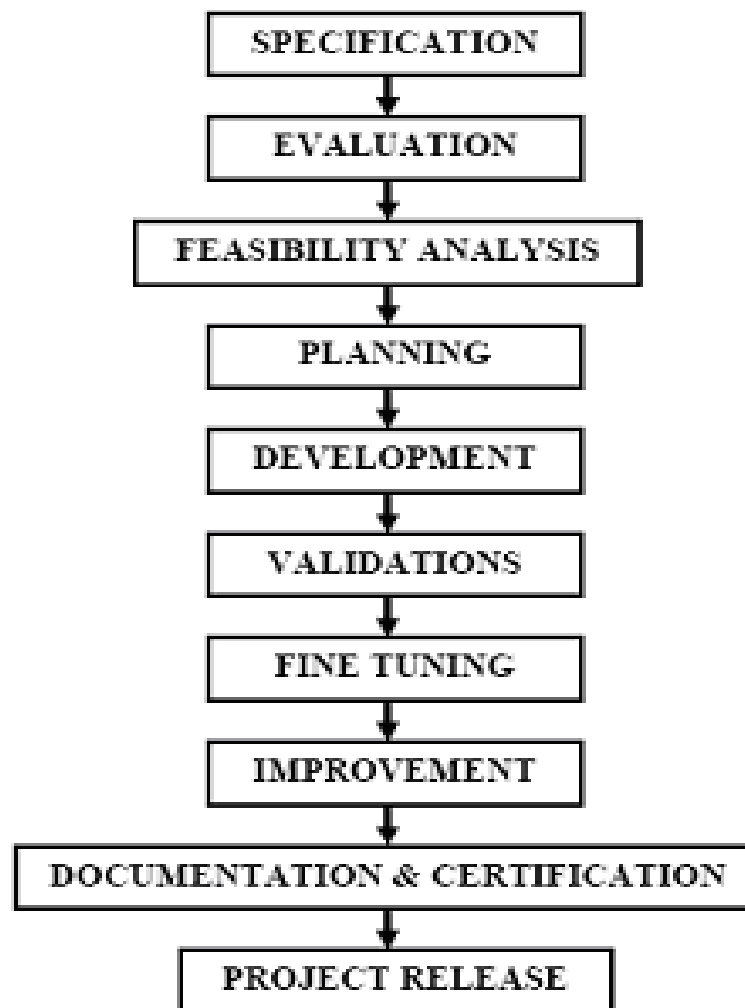


Figure 3.3: Detail Life Cycle of Project

# Chapter 4

## Implementation

### 4.1 Critical Path Modeling

For designing, characterization & validation of Memories Schematic Simulations are needed. Schematic simulation was carried out to verify the functionality of the SRAM. The core matrix contains 256 x 256 memory cells. Each cell contains 6 transistors i.e. the SRAM core has  $256 \times 256 \times 6 = 393216$  transistors [10].

There are two reasons of using the concept of Critical Path Modeling.

- To save the simulation times but still retain a simple but exact view of the SRAM performance (Timing and Power information).
- In a Compiler approach if we wish to provide memory compiler (memory generator) to customer so as customer can pick any size of memory CUT with in the range of compiler, then critical path modeling becomes necessity. For Example: If a single port memory compiler has row range of 128 to 1024 & column range of 8 to 128 then there exists hundreds of possible CUT combinations and it is practically not possible to design these memory CUTs individually. Only critical path model can solve the purpose [10].

Note that for verification purpose it is sufficient to perform two writes cycles and two read cycle's viz. Write Low, Write High, Read Low and Read High. This is carried

out as follows. Thus for a 256 x 256 core, in four cycles we alternately select 1st and the last row, the rest 254 rows are represented by a single load row in our model. The load row has a defined variable called multiplication factor,  $m$ . for the present example  $m = 254$ . We perform a hierarchical simulation of the core. The core is modeled as explained below [10].

- a. **Level 0:** At this level the core is represented as a single block. It is connected to row decoder through global word lines and to the I/O through bit / bit bar lines and the block select lines.
- b. **Level 1:** At this level the core is represented as 16 blocks, each block having its own local core and local decoder which is internal to that block. For verification purposes we simulate the 1st and the 16th blocks. The remaining 14 blocks are represented as a single load block with  $m = 14$ . The net load appearing on the global word lines remain same as it would be in actual core.
- c. **Level 2:** Inside a block we have local decoder and local core.

**Local Decoder:** Local decoder is represented as three blocks. The first and second are active blocks while the third is the load block. Each active block receives inputs from global word lines and block select line and feeds to two local word lines. So we need 128 local decoder blocks to feed 256 local word lines. The multiplication factor,  $m$  for the local decoder is  $m = ((\text{row}/2) - 2) = ((256/2) - 2) = 126$  and load on block select is  $126 + 1 + 1 = 128$  local decoder blocks [10].

**Local Core:** Each local core is matrix of 256 x 32 cells. We read/write 16 bits at a time. For simulation purpose checking for MSB and LSB is sufficient while rest 14 bits can be represented by a single load column. Where the two rows are enough to simulate the core activity to check the timing and power information. These two rows LWL\_0 (WL\_BOT) and LWL\_256 (WL\_TOP) are coming from the active local decoder blocks [10]. **IO Modeling:** Each I/O block corresponds to single bit.

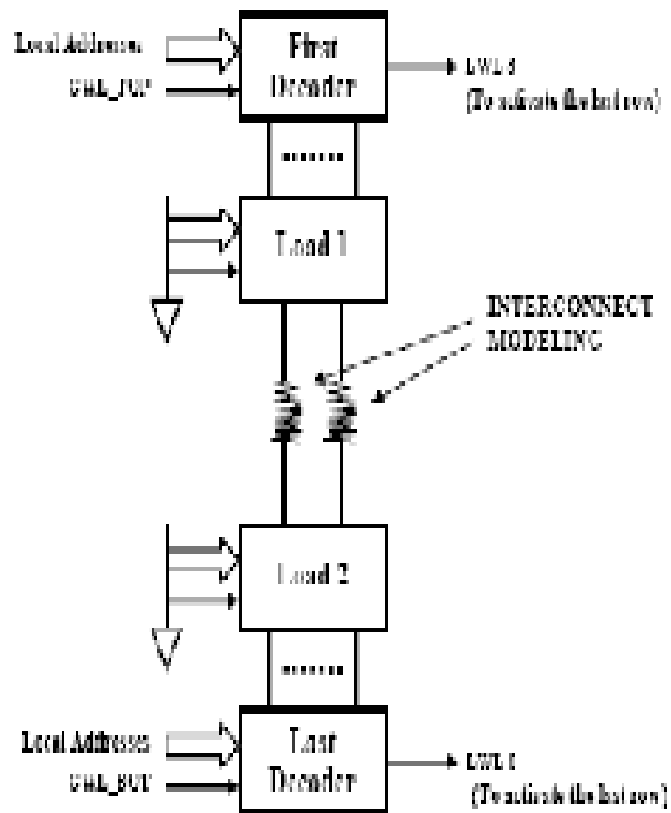


Figure 4.1: Local Decoder Modeling

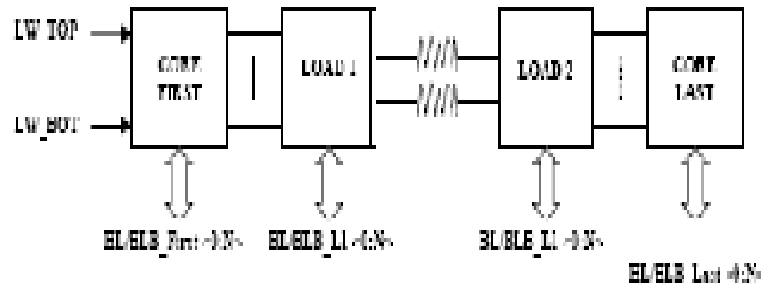


Figure 4.2: Local Core Modeling

Critical path modeling providing the provision to access the first and the last bit so we have to model the I/O block also to match it with entire modeling [10]. Note: Same modeling technique will be used in IO Modeling and DQ

**Logic Level 3:** This is the bottom most level therefore at this level schematics are present.

## 4.2 Replica Technique for Wordline and Sense Control in Low Power SRAM

With the migration toward low supply voltages in low-power SRAM designs, threshold and supply voltage fluctuations will begin to have larger impacts on the speed and power specifications of SRAM's [6].

Replica circuits minimize the effect of operating conditions' variability on the speed and power. Replica memory cells and bitlines are used to create a reference signal whose delay tracks that of the bitlines. This signal is used to generate the sense clock with minimal slack time and control wordline pulsewidths to limit bitline swings.

Low Power circuits have been continually pushing down supply voltages to minimize the energy consumption of chips for portable applications. The same trend has also applied to low-power SRAM's in the past few years. While the supply voltages



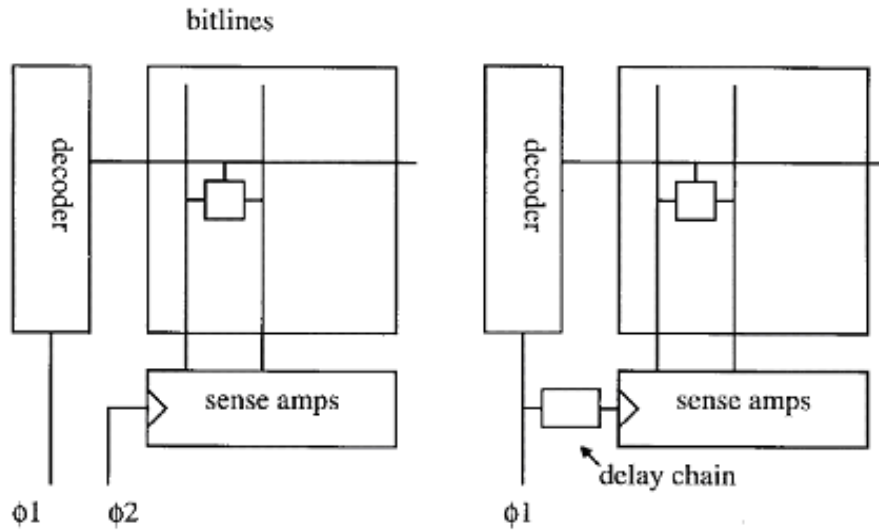


Figure 4.3: Common Sense Generation

are scaling down at a rapid rate, to control subthreshold leakage, the threshold voltages have not scaled down as fast, which has resulted in a corresponding reduction of the gate overdrive for the transistors.

### 4.2.1 Inverter Chain Technique

The prevalent technique to generate the timing signals within the array core essentially uses an inverter chain. This can take one of two forms the first kind relies on a clock phase to do the timing Figure 4.3a, and the second kind uses a delay chain within the accessed block, and is triggered by the block select signal Figure 4.3b or a local wordline. The main problem in these approaches is that the inverter delay does not track the delay of the memory cell over all process and environment conditions. The tracking issue becomes more severe for low-power SRAM's operating at low voltages due to enhanced impact of threshold and supply voltage fluctuations on delays. Delay variations are inversely proportional to the gate overdrive [6]. The memory cell delay is mainly affected by the nMOS thresholds, the inverter chain delay is affected by both nMOS and pMOS thresholds. The worst case matching for the inverter delay

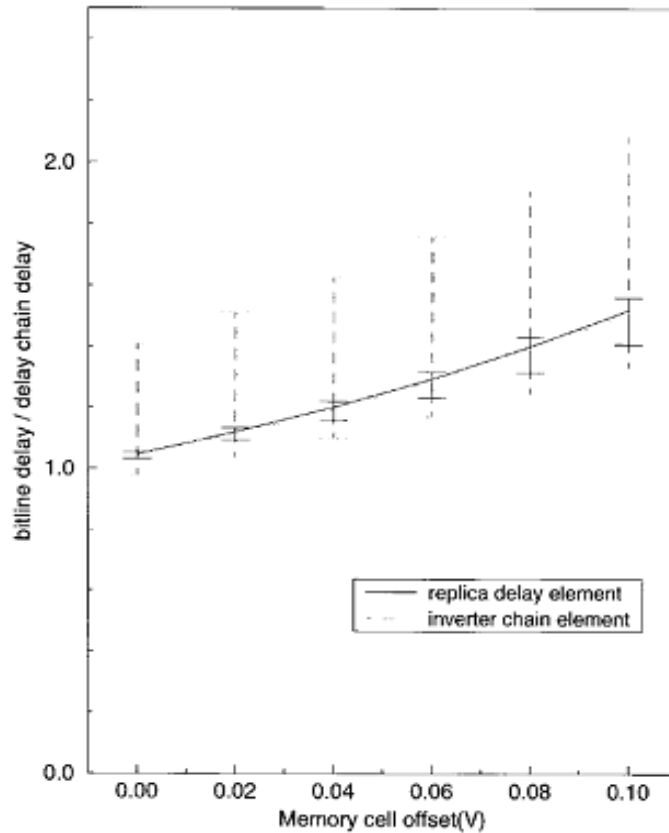


Figure 4.4: Matching of the bitline delay with the inverter chain delay and the replica cell bitline delay across process fluctuations over varying threshold offsets for the accessed memory cell.

chain occurs for process corners where the nMOS and pMOS thresholds move in the opposite direction [6].

The delay element is designed to match the delay of a nominal memory cell in a block. But in an actual block of cells, there will be variations in the cell currents across the cells in the block. Figure 4.4 displays the ratio of delays for the bitline and the delay elements for varying amounts of threshold mismatch in the access device of the memory cell compared to the nominal cell. The graph is shown only for the case of the accessed cell being weaker than the nominal cell as this would result in a lower bitline swing. The curves for the inverter chain delay element (hatched) and the replica delay element (solid) are shown with error bars for the worst case fluctuations

across process corners. The variation of the delay ratio across process corners in the case of the inverter chain delay element is large even with zero offset in the accessed cell, and grows further as the offsets increase. In the case of the replica delay element, the variation across the process corners is negligible at zero offsets, and starts growing with increasing offsets in the accessed cell. This is mainly due to the adverse impact of the higher nMOS thresholds in the accessed cell under slow nMOS conditions. It can be noted that the tracking of the replica delay element is better than that of the inverter chain delay element across process corners, even with offsets in the accessed memory cell [6].

All of the sources of variations have to be taken into account in determining the speed and power specifications for the part. To guarantee functionality, the delay chain has to be designed for worst case conditions, which means that the clock circuit must be padded in the nominal case, degrading performance. Replica-based delay elements, by virtue of their good tracking, offer the possibility of designing SRAM's with tight specifications across all process corners.

### 4.2.2 Capacitance Ratioing

The replica delay stage is made up of a memory cell connected to a dummy bitline whose capacitance is set to be a fraction of the main bitline capacitance. The value of the fraction is determined by the required bitline swing for proper sensing. For the clocked voltage sense amplifiers we use Figure 4.5 the minimum bitline swing for correct sensing is around a tenth of the supply. An extra column in each memory block is converted into the dummy column by cutting its bitline pair to obtain a segment whose capacitance is the desired fraction of the main bitline Figure 4.6 The replica bitline has a similar structure to the main bitlines in terms of the wire and diode parasitic capacitances [6].

Hence, its capacitance ratio to the main bitlines is set purely by the ratio of the geometric lengths  $r/h$ . The replica memory cell is programmed to always store a

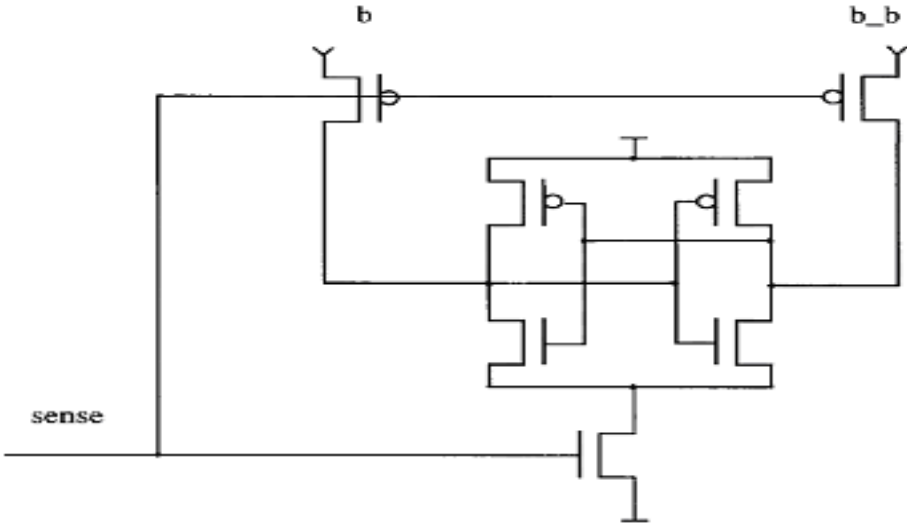


Figure 4.5: Latch Type Sense Amplifier

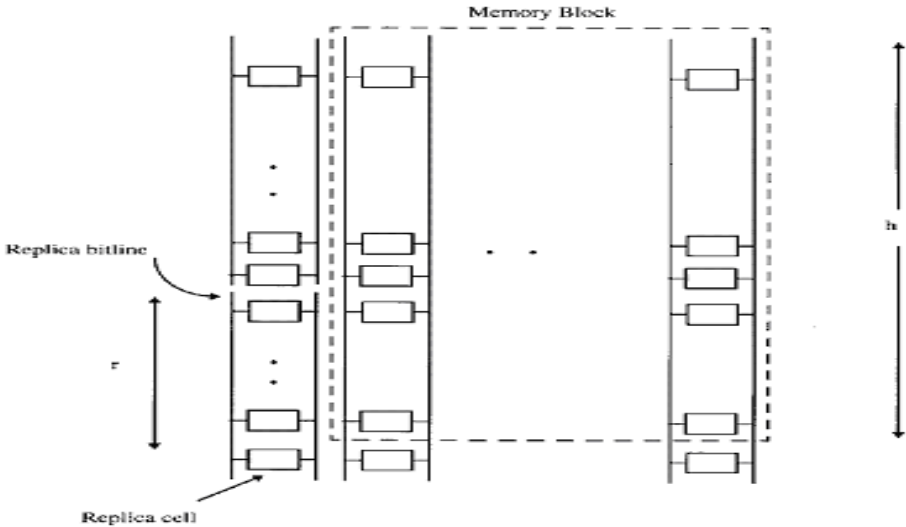


Figure 4.6: Replica Bit Line Column

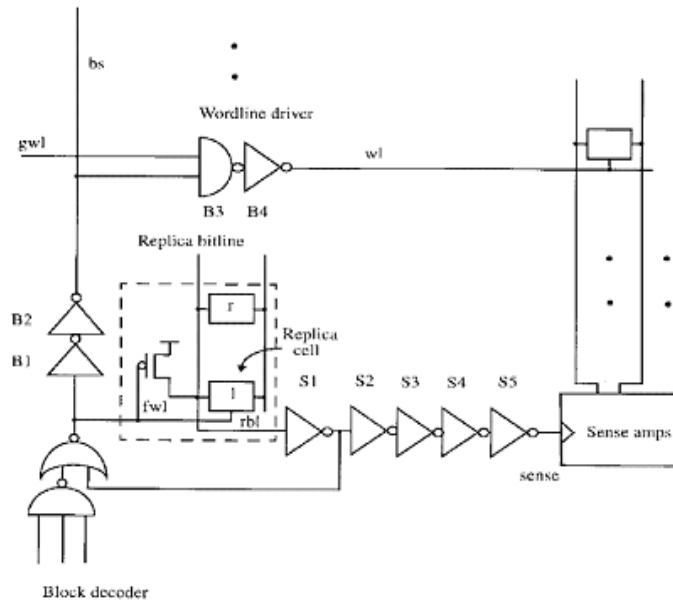


Figure 4.7: Control circuits for sense clock activation and wordline pulse control.

zero so that, when activated, it discharges the replica bitline. The delay from the activation of the replica cell to the 50% discharge of the replica bitline that track of the main bitline very well. The delays can be made equal by fine tuning of the replica bitline height using simulations. The replica structure takes up only one additional column per block, and hence has very little area overhead [6].

The circuits to control the sense clock and wordline pulsewidths are shown in Figure 4.7. The block decoder activates the replica delay cell (node *fwl*). The output of the replica delay cell is fed to a buffer chain to start the local sensing, and is also fed back to the block decoder to reset the block select signal. Since the block select pulse is ANDed with the global wordline signal to generate the local wordline pulse, the latter's pulsewidth is set by the width of block select signal. It is assumed that the block select signal does not arrive earlier than the global wordline. The delay of the buffer chain to drive the sense clock is compensated by activating the replica delay cell with the unbuffered block select signal [?].

The delay of the five inverters in the buffer chain, S1 to S5, is set to match the

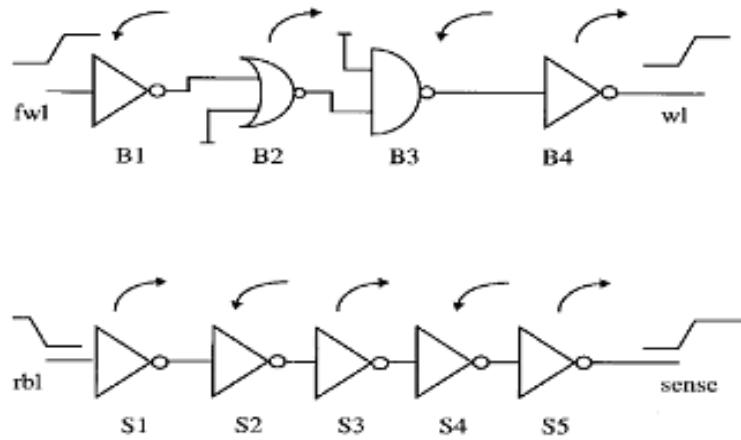


Figure 4.8: Delay Matching of Two Buffer Chain

delay of the four stages, B1 to B4, of the block select to local wordline path (the sense clock needs to be a rising edge). The problem of delay matching has now been pushed from having to match bitline and inverter chain delay to having to match the delay of one inverter chain to a chain of inverters and an AND gate [6].

### 4.2.3 Cell Current Ratioing

While the above technique works well, it can be modified further to improve the access time. If the reset timing signal for the wordline can be generated locally, then the wordline driver can be skewed to speed up the propagation of the rising block select transition, reducing the access time, with the falling wordline transition being triggered off the local reset signal. An extra row and column containing replica memory cells can be used to provide local resetting timing information for the wordline drivers [10].

The extra row contains memory cells whose pMOS devices are eliminated to act as current sources, with currents equal to that of an accessed memory cell Figure 4.9. All of their outputs are tied together, and they simultaneously discharge the replica bitline. This enables a multiple of memory cell current to discharge the replica bitline.

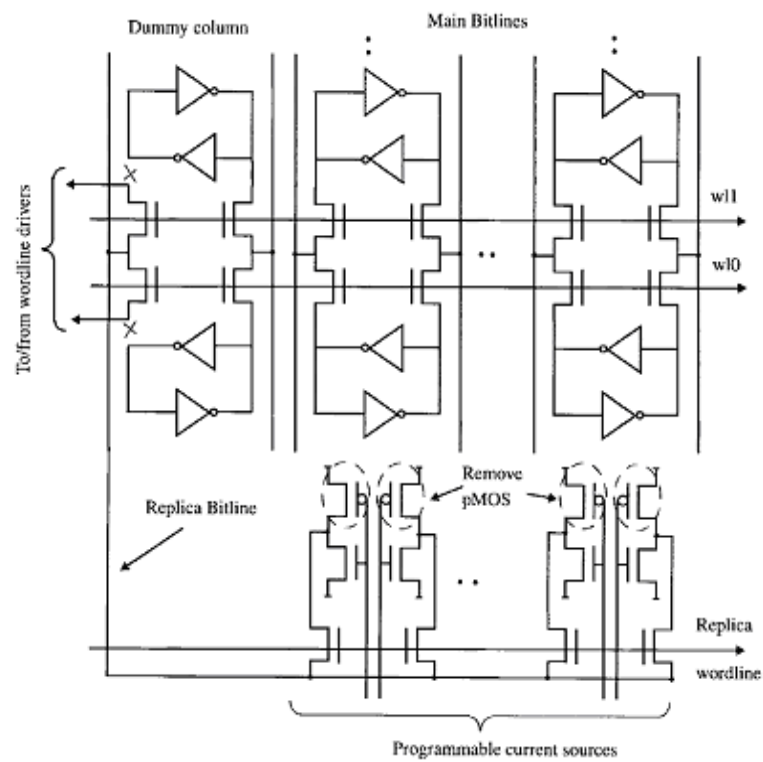


Figure 4.9: Current Ratio Based Replica Structure

The current sources are activated by the replica wordline, which is turned on during each access of the block. The replica bitline is identical in structure to the main bitlines, with dummy memory cells providing the same amount of drain parasitic loading as the regular cells. By connecting 'N' current sources to the replica bitline, the replica bitline slew rate can be made to be 'N' times that of the main bitline slew rate, achieving the same effect as bitline capacitance ratioing described earlier[10].



# Chapter 5

## Extraction and MCF

### 5.1 Post-Layout Extraction

Extraction is the process for finding out parasitic resistances and capacitances. In extraction, it is often helpful to make an (informal) distinction between designed devices, which are devices that are deliberately created by the designer, and parasitic devices, which were not explicitly intended by the designer but are inherent in the layout of the circuit [10].

#### 5.1.1 Need Of Extraction

Extraction is needed to support simulation of the effect due to "parasitic" elements. "Parasitic" effects are due to interconnect and are not intended part of the circuit function. "Parasitic" effects typically only affect the timing behavior of the circuit. The circuit should Perform correctly if the frequency is slow enough [10]. The Post Layout Extraction flow is aimed to provide spice transistor flat netlist back annotated with Parasitic Cs and Rs Extracted on Interconnections. The Post-Layout Simulation (PLS) tool is automating the interconnect RC parasitic extraction from a GDS (Graphical Design Structure) database and a CDL (Circuit Description Language) netlist to a Spice at device (transistor) level [10].

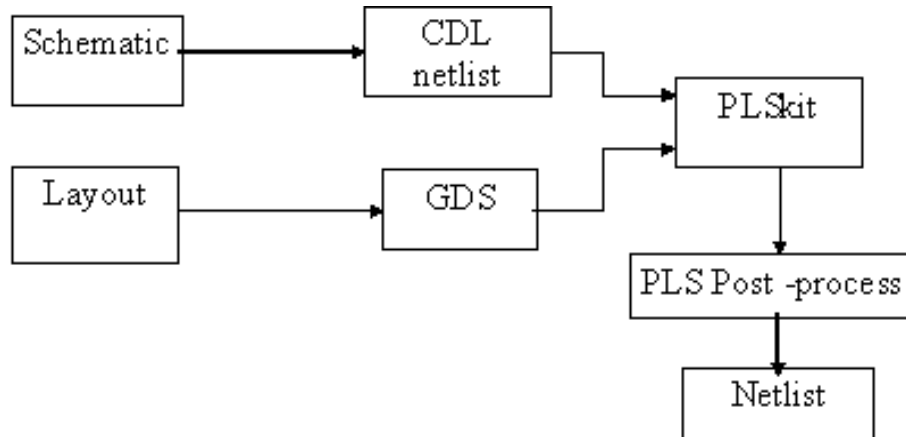


Figure 5.1: Block Diagram for Post Layout Extraction

### 5.1.2 Extraction Methodologies

In the extraction process two input files are required, CDL file and GDS file for the particular block. The CDL file ( netlist file ) is generated from the schematic and the GDS file is generated from layout. This two files are applied to PLS tool which gives the netlist file with all parasitic capacitance and resistance values. Then final step is the post processing which make the extracted netlist file to simulatable file.

The PLS tool is proposing various extraction methodologies (or extraction modes) to extract the parasitic elements on the layout [10].

#### Mode C: net-to-ground coupling

If net A is declared in the Net file then the lumped capacitance associated to this net will be Extracted. The reference node can be chosen in the PLS interface. A threshold "C threshold" can be specified in order to filter small capacitances:

If  $[C(A) \geq C \text{ threshold}]$  then  $C(A)$  is ignored.

#### Mode $C_{tc}$ : net-to-net coupling

If net A is specified in the Net file, then all coupling capacitances between net A and all nets  $N_i$  will be extracted. In that example, nodes A, B and C are declared in the

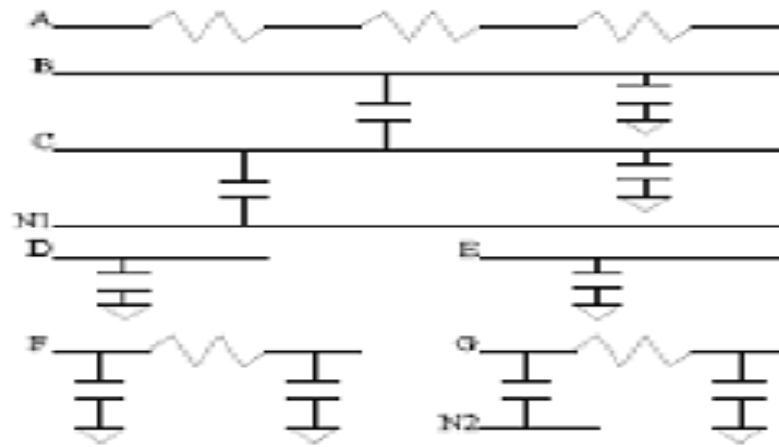


Figure 5.2: Different Extraction Methodology

Net file. There is a coupling capacitance Between A and B. C has no aggressor nets except the substrate.

#### **Mode R: Net resistances**

If net A is declared in the Net file, then parasitic resistances will be extracted.

#### **Mode RC: Sub node-to-ground coupling**

If net A is declared in the netfile, parasitic resistances will be extracted and Lumped capacitances will be netlisted on each sub node of net A.

#### **Mode RCc: subnode-to-subnode coupling**

Parasitic resistances will be extracted on nets declared in the Net file. All coupling capacitances associated to each node declared in the Net file will be extracted, even if the aggressor nets are not declared in the Net file.

## **5.2 Extraction of Single Memory cell**

The extraction of single memory cell is done from the 44 memory cell. The extraction of the single memory cell is very critical point in the design validation. The extracted netlist file of single memory cell is modeled for the all memory cells in the critical path modeling. In the memory core block , memory cells are arranged in the rectangular

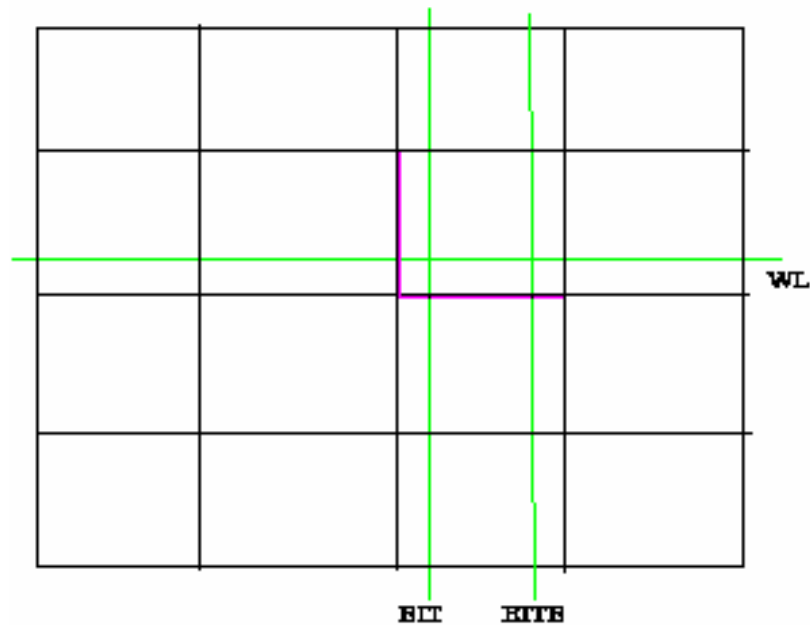


Figure 5.3: Single Memcell Extraction

format as shown in the bellow figure. Considering the bellow figure , each block represent the single memory cell. If the layout file of single memory cell is used for the extraction then it does not contain any environmental capacitance from the neighboring memory cells.

So, the layout file of single memory cell is made from layout file of 44 memory cell by changing the labels of the bitlines and worldlines and deleting unwanted capacitance.

### 5.3 Memory Characterization Flow

Memory Characterization Flow (MCF) is a tool for the characterization of memory cuts. Characterization refers to finding of certain attributes associated with memory cuts like timing values, power consumption, and leakage and pin capacitance. The

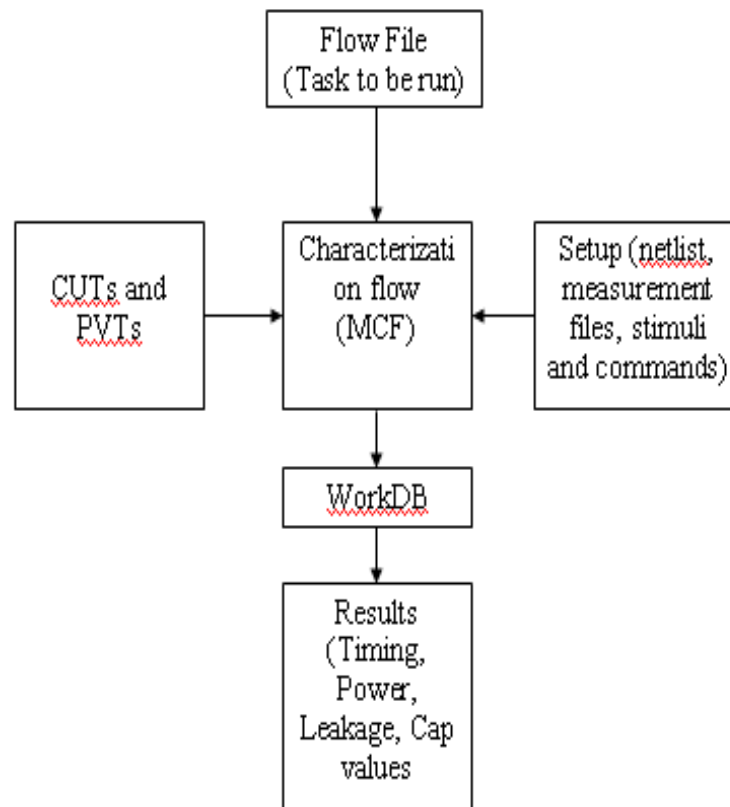


Figure 5.4: Block Diagram of MCF

tool can be put to certain other uses also apart from characterization, like marginality measurements.

The MCF tool is used to measure memory's timing parameters, power, leakage power & pincap. MCF has a flow file .mcf which calls several TASKS & MACROS to perform specific jobs for timing parameters, power, leakage power & pin capacitance. These TASKS & MACROS are being defined in the methodology file which is there in mcf tool. To launch specific jobs for timing, power, leakage & pin capacitance; mcf tool requires files [10].

The major advantage of introduction of MCF, is that, it is OTF (On the Fly) based tool MGC. OTF refers to the fact that when a user requires a memory cut from webgen, actual simulation is performed and then the characterized values are reported to the user. As MGC was not OTF based tool, it employed curve fitting equations to report values MCF structure can be explained as follow: For launching MCF 3 important things are needed as input:

- Flow file
- Source file
- Setup file

**FLOW FILE:** MCF flow is the main file of the setup as it specifies the flow for a task to be performed and the inputs required for the corresponding task. **SOURCE FILE:** This file mentions the product required for running for the setup. **SETUP FILE:** This file contains all the setup required for launching MCF [10].

# Chapter 6

## Analysis Work

### 6.1 Memory Cell Analysis

Memory cell stability analysis means memory cell has to be tested for its reliability or ruggedness in storing bits. For example any stored bit should not change due to change in any parameters. Major causes which can change the stored bit are:

- Noise
- Temperature
- Design of memcell, here due to different sizes of pull down transistors differential noise comes up. In this case one transistor is more affected than other by noise.
- Process variation also affects the stability of memcell. When process varies model parameters of the transistors also varies accordingly.

Before studying the memcell stability, first of all we find out worst case conditions for stability analysis. It is essential to consider the stable SRAM cells when CMOS technology is scaled down to deep-submicron dimensions. Six key parameters which can lead to the change in the stored information that need to be checked in the designing of a memory cell are:

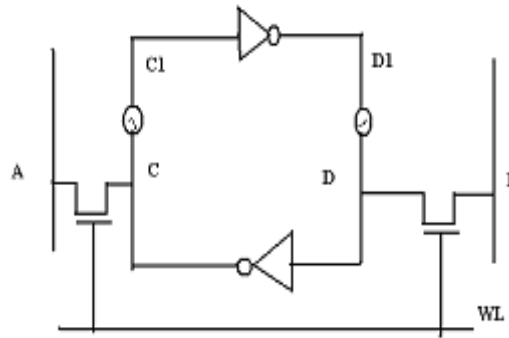


Figure 6.1: Memcell for SNM

- Discharge Rate
- SNM
- Write Margin
- IOFF
- ION

### 6.1.1 Static Noise Margin (SNM)

One most important aspect for SRAM cell design is stability of the cell. The cell stability determines the sensitivity of the memory to process tolerances and operating conditions. Stability of the cell is expressed by Static Noise Margin. Noise margin describes the amount of variation in the signal levels that can be allowed while signal is transmitting. In short, static noise margin describes the noise tolerance range of the memory cell while performing the read operation [2].

Noise margin of the memory cell can be analyzed by adding noise sources (VCVS) of opposite polarity, as shown in figure. Both the noise source will provide the reason for memory cell to be flipped. Static noise margin of the memory cell will be the



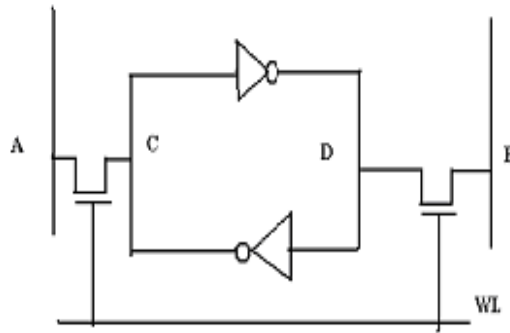


Figure 6.2: Memcell for WM

maximum voltage value of the noise source that can be tolerated by the cell before changing the state.

Node C & C1 is initialized to '0'V and node D & D1 is initialized to VDD. A DC voltage source of voltage VDD is connected to net WL , A & B. A voltage controlled voltage source with reverse polarity is connected between node C-C1 and node D-D1 as shown in figure. Value fo noise source is increased in ramp fashion. The value of the noise source at which memory cell get flipped is the SNM of the memory cell.

### 6.1.2 Write Margin

During a write operation, one of the bitlines is pulled low. The requirement for a successful write operation is to swing the internal voltage of the cell to pass the switching threshold of the corresponding inverter. During write '0' operation one of the bitline gets discharge on the side at which we want to write zero. The value of discharging bitline voltage at which memcell get flipped is called the write margin for the memcell.

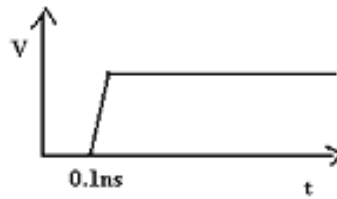


Figure 6.3: Stimuli for Discharge Rate

### 6.1.3 Discharge Rate

The rate at which the Bitline discharges is called the Discharge Rate. There are two points to be taken care while calculating Discharge Rate:

- Additional load on Bitline from
  - I/O Cell
  - Strap Cell
  - Dummy Row Cell
- Domination of fringe capacitance for small height memory. As the technology changes the height of the memory cell changes, thus the capacitance per memory cell also changes. Lesser the height, lower is the capacitance. But effectively the fringe capacitance remains the same and thus it dominates.

Consider the Figure 6.2 shown in write margin . Initialize net A, B to VDD . Node C is initialized to '0'V and node D is initialized to VDD . Following stimuli is applied to WL. Now , when WL turns on bitline A starts discharging towards GND. Time require for bitline to discharge from  $(0.9 * VDD)$  to  $(0.8 * VDD)$  is called the discharge rate .

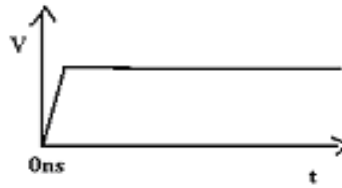


Figure 6.4: Stimuli for ION

#### 6.1.4 IOFF and ION Currents

Ion is the on current of the memory cell, which means the current flowing when memcell is on. Ioff is the leakage current which typically contributes a major portion of the standby current of the chip. This leakage current consumes a fairly large amount of standby power. We generally assume that in "OFF" condition MOS transistors there is no flow of current. But, in practical circuits there exists very high impedance between the source and drain in the "OFF" state. Thus, leakage in this context is calculated as the current flowing through the access transistor when the wordline is "OFF" and the bitlines are precharged.

**ION:** Consider the figure shown in write margin . Initialize net A, B to VDD . Node C is initialized to '0'V and node D is initialized to VDD . Following stimuli is applied to WL Now, the current flowing from bitline A and B is measured, which is called the ON current of the memcell.

**IOFF:** This current is measured in presence of other memcell . Consider the figure shown in write margin . Initialize net A, B to VDD . Node C is initialized to '0'V and node D is initialized to VDD . Here. No signal is applied to WL , means pass transistors remain off and the current flowing through net A and B is measured which is called OFF current of the memcell.

## 6.2 Latch and Dynamic Circuit Analysis

When we use dynamic circuits (to gain on area and speed) we may have floating nodes in our designs (a node is not driven to gnd/vdd by NMOS/PMOS and value is stored only as charge in capacitors). If these nodes are kept floating for long time then due to leakage it may happen that their state gets corrupted. This will corrupt the subsequent nodes and may lead to memory failure [10].

To prevent this from happening, latches are put on such nodes. These latches are typically uncontrolled (since they are not controlled by any clock or external signal). It is very important that they don't mal-function or the memory may enter into a stuck state.

There are two types of latches that we may decide to put, full latch and half latch. Half latches only hold one state means they can either hold gnd or vdd so they will constitute NMOS or PMOS but not both. Full latches on the other hand can hold both gnd and vdd and thus will both involve PMOS and NMOS. Half-latches can be put where the pre-charge state is controlled by CK (or activating signal). Full latches have to be put where both Pull-up network and Pull-down network exist and the output is not pre-charged to any particular state [10].

Validation of these latches requires that they should be able to compensate the leakage, fight-back charge-sharing but should allow the driving circuit to write into them even at cross-corners like SFA and FSA [10].

### Guidelines for Latch Validation

It should be able to compensate the leakage at node even when leakage gets increased by 10 times. This must be checked at cross corners with latch devices at slow corner and leaking device at fast corner [10].

Latch should allow the write at node which it is holding. Again this should be performed at cross corner but this time we want latch devices to be at fast corner and driving device at slow corner [10].

To ensure immunity to charge-sharing (at input signals), it should be verified that the latch devices at slow corner are able to take care of charge-sharing. For this purpose, the parasitic capacitance of internal nodes should be doubled . The effect of all dynamic couplings on the node holding the charge should be analyzed [10].

## 6.3 Row-Decoder Validation

The following validation are done on Dynamic Circuits of Row Decoder block.

- Leakage Validation
- Strength Validation
- Charge Sharing Validation
- Bump Validation

Consider the Figure 6.5 for the above validations.

### 6.3.1 Leakage Validation

First step is to keep node X,Y,Z at gnd and node OUT is initialized to VDD. Then node B is tide to gnd. Next step is to make mfacroe of  $M2 = 30$  . Then check if M1 is able to compensate the leakage of M2. This analysis must be done with FSA corner.

### 6.3.2 Strength Validation

First step is to initialize node OUT to VDD. Then Connect B, C, D to VDD and apply stimuli at A as shown in Figure 6.6. Measure the delay between rising edge of A and falling edge of OUT (WL). Check whether OUT is been able to change its state to gnd. This analysis should be done at SFA corner

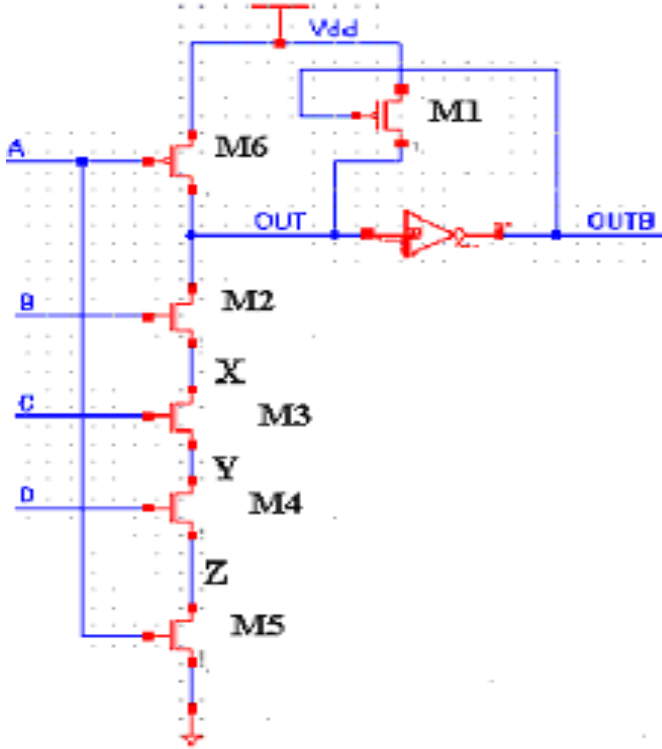


Figure 6.5: Dynamic Nand Gate

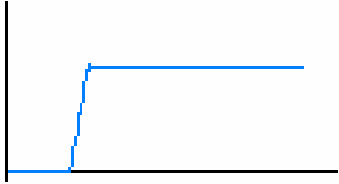


Figure 6.6: Stimuli for Node A

### 6.3.3 Charge Sharing Validation

First step is to initialize node X,Y,Z at gnd. Then connect A, B at logic 0 and connect C and D at logic 1 [1]. Then B is put at logic 1. At this time, node OUT should not observe a dip of more than 10

### 6.3.4 BUMP Validation

First step is to connect extra PMOS between VDD and OUT with mfactor =30. Then make mfactor of M0=30 and initialize node OUT to 0v. Connect A, B, C ,D to VDD. Here, node OUT should not observe a bump of more than 10

## 6.4 I/O Latch Validation

Sometimes, we may have floating nodes in our designs (a node is not driven to GND/VDD by NMOS/PMOS and value is stored only as charge in capacitors). If these nodes are kept floating for long time then due to leakage it may happen that their state gets corrupted which may lead to memory failure [10].

To prevent this from happening, latches are put on such nodes. These latches are typically uncontrolled (since they are not controlled by any clock or external signal). It is very important that they don't mal-function or the memory may enter into a stuck state.

There are two types of latches, full latch and half latch. Half latches only hold one state means they can either hold GND or VDD so they will constitute NMOS or PMOS but not both. Full latches on the other hand can hold both GND and VDD and thus will both involve PMOS and NMOS [10].

Validation of these latches requires that they should be able to compensate the leakage, but should allow the driving circuit to write into them even at cross-corners like SFA and FSA. The following validation are done on Latch of the IO block.

- Leakage Validation

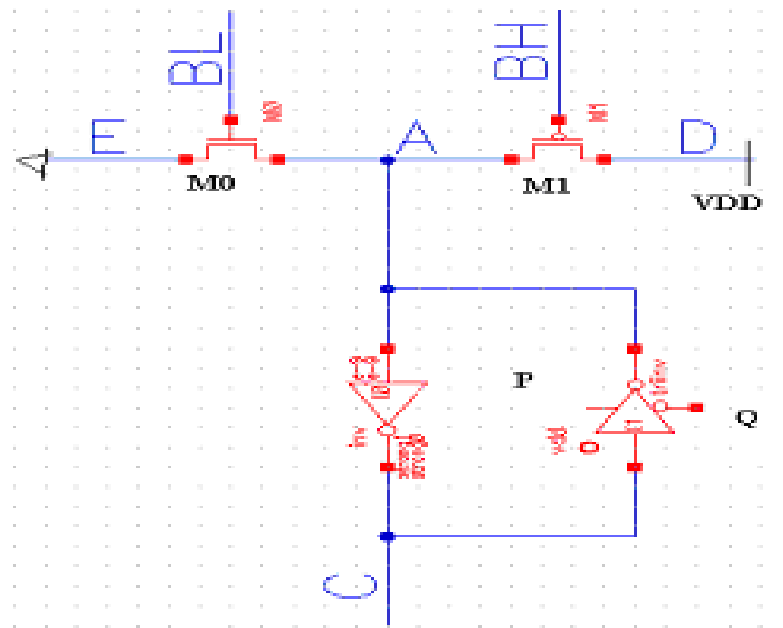


Figure 6.7: Schematic of IO Latch

- Strength Validation
- Charge Sharing Validation
- SNM
- Coupling

### 6.4.1 Leakage Validation

Some voltage dips & glitches occur in latches due to the leakage-effect. In latch-leakage analysis voltage dips and glitches are measured for all the latches to get ensured that these voltage dip and voltage rise is not sufficient to change the data stored in the latch. If there is no switching in the data stored in latch then only latch-leakage analysis qualifies the passing criteria.



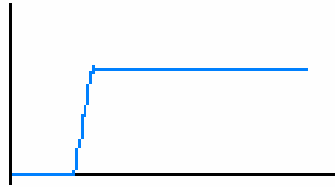


Figure 6.8: Stimuli for Net BL

**Leak\_1** In the Figure 6.7, make the tristate buffer on. For that , make  $P=1$  and  $Q=0$  . Then initialized the node A to vdd and node C to gnd and make  $BH = vdd$  and  $BL = gnd$ . Make the mfactor of MO and nmos of the inverter to 30. Now measure the Dip due to leakage of this nmos transistors

**Leak\_0** In the above Figure 6.7, make the tristate buffer on. For that , make  $P=1$  and  $Q=0$  . Then initialized the node A to gnd and node C to vdd and make  $BH = vdd$  and  $BL = gnd$ . Make the mfactor of M1 and pmos of the inverter to 30. Now measure the Bump due to leakage of this pmos transistors.

## 6.4.2 Strength Validation

Latch-writability analysis is performed to be ensured that the latch is writable or not. In this case driver transistors are turned on one by one to check the writability of '0' or '1' in the latch. If the data stored in latch get flipped (if there is switching) up to an acceptable level then only the analysis qualifies the passing criteria.

**Write 0** Is it possible to write '0' at node A ? To check this initialize node A to vdd and node C to gnd. Make the tristate buffer on. For that , make  $P=1$  and  $Q=0$  . Apply stimuli as shown in Figure 6.8to net BL and keep BH at vdd. Then measure the delay 'd1' and 'd2' as shown in the waveform.

**Write 1** Is it possible to write '1' at node A ? To check this initialize node A to gnd and node C to vdd. Make the tristate buffer on. For that , make  $P=1$  and  $Q=0$

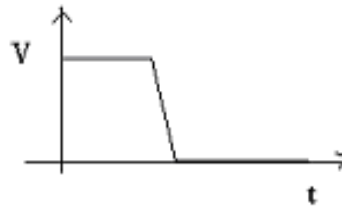


Figure 6.9: Stimuli for Net BH

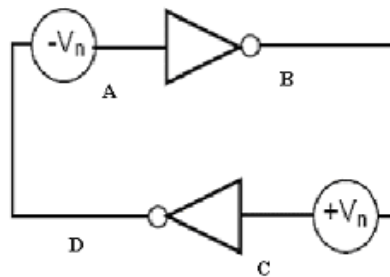


Figure 6.10: IO Latch with Noise Voltage Sources

- . Apply stimuli as shown in Figure 6.9 to net BH and keep BL at gnd. Then measure the delay 'd1' and 'd2' as shown in the waveform.

### 6.4.3 Static Noise Margin (SNM)

The term "Static Noise" in this context indicated disturbances caused by the interference, spurious and transistor mismatch and offsets [2]. Noise immunity of the IO Latch was tested by adding two noise sources as shown in Figure 6.10. Static noise is dc disturbance such as offsets and mismatches due to processing and variations in operating conditions. The SNM of the flip-flop is defined as the maximum value of  $V_n$  that can be tolerated by the flip-flop before changing states. A SRAM should be designed such that under all conditions some SNM is reserved to cope with dynamic disturbances caused by alpha-particles, crosstalk, voltage supply ripple and thermal noise. I did DC analysis to find the Static Noise Margin with a linearly increasing

noise source. The value of noise source at which IO Latch flips is the static noise margin of IO Latch.

**SNM 0** First initialize node A,D to 0v and node B,C to vdd. Then apply a dc ramp voltage up to 1.5v as  $V_n$ . Now measure the value of  $V_n$  when  $V(B) = 0.2$  vdd. This is the voltage at which the given latch may flip its state.

**SNM 1** First initialize node A,D to vdd and node B,C to 0v. Then apply a dc ramp voltage up to 1.5v as  $V_n$ . Now measure the value of  $V_n$  when  $V(D) = 0.2$  vdd. This is the voltage at which the given latch may flip its state.

#### 6.4.4 Coupling

In a layout signals are routed adjacent to each other and they have coupling capacitance. At the time of making a transition, these signals (aggressor) can affect the level of adjacent signals (victim) due to coupling. If the victim signal is floating or held weakly i.e. by a high resistance device, then we may see a change in state of this signal and thus lead to failure [10].

At times, cross-talk may lead to delaying of some operation like in row-decoder where the address lines are placed quite close to each other deselection of one address line delays selection of other address line. This leads to later selection of WL and hence VDIFF loss may happen at Sense Amplifier. So, cross talk analysis and validation is necessary for ensuring proper working of memory compilers.

### 6.5 Marginality Analysis

In memory there are certain circuits/modules (like pulse latched circuits) which work correctly only when proper sequencing in signals is ensured. These circuits/modules must be designed with proper care. For analysis and validation of these types of circuits/modules marginality analysis is done [10].

Marginality analysis will include analysis of **Race Condition** which lead to memory failure and measurement of some performance critical parameters like voltage difference in read cycle etc [10].

This output of this analysis helps designer to tune its design to ensure functionality.

### Example of some Marginality:

**tmarg\_IN\_seno** Data must be latched (seno\_latch) before the read operation end (SAEN goes low)

**tmarg\_wl** Precharge must turn off before wordline goes high (A slight negative value within a few 10s of ps may be OK)

**tmarg\_wr\_drv\_off** Precharge must turn on after wr\_drv turns off. If it is only slightly negative, is tolerable.

**tmarg\_PCH\_SA\_SAEN** Sense amp. latch precharge must be active after the read operation end SAEN should go 'LOW' before PCH\_SA starts .

**tpowmarg\_wr\_drv** Precharge must turn off before the wr\_drv turns on

**tpowmarg\_wl** Precharge must turn on after wordline goes low

### 6.5.1 Criteria for Marginality Analysis

- All marginalities must be positive.
- For any negative marginality: Either Tune the Design Or Logically explain as non-destructive in the Marginality report (e.g in cases where penalty in taa is acceptable and vdiff is sufficient); may decide to qualify with relaxed thresholds in such cases.
- Check the marginality in Tight stimuli simulation if it causes any failures.

- All marginalities which affect the functionality of memory i.e. functional marginalities should have 10%-90% as their threshold value
- All power marginalities must have 20%-80% as threshold value.
- Marginality validation should be run with analog process corners models(SSA, FFA, SFA, FSA).

## 6.6 Power On Analysis

When power supply is given to the memory devices may be in different states of activation. Based on device mismatch, different internal latches (including state-machine latches) can get initialized to "0" or "1". As shown in Figure 6.11 , suppose at the time of power on , BL\_L is at VDD and BL\_H is at '0' V . Now assume that , in three NMOS of write section data value is such that transistor W1 & W2 is 'ON'. Here, BL\_H is at '0'V so, transistor P1 is 'ON' and BL\_L is at VDD so, transistor P2 is 'OFF'. Now, current will flow from VDD to BL\_L to W1 to W3 to GND. So, voltage at BL\_L is start reducing & it makes P2 to turn 'ON'. As, P2 turns 'ON' BL\_H starts charging and it makes P1 to be 'OFF'. As W2 is 'OFF' no current will flow from BL\_H to GND. Now, the current flowing during this time is measured in three forms (1) Integral Current , (2) Peak Current , ( 3) Stable Current . Power dissipation due to this current power dissipated during 'Read' or 'Write' operation of memory [7].

## 6.7 Access Time Analysis

In access time analysis access time of the memory is measured. Access time is defined as time difference between rising edge of the clock and output data at 'Q'. This analysis is done with memory characterization set up with monte carlo simulations.



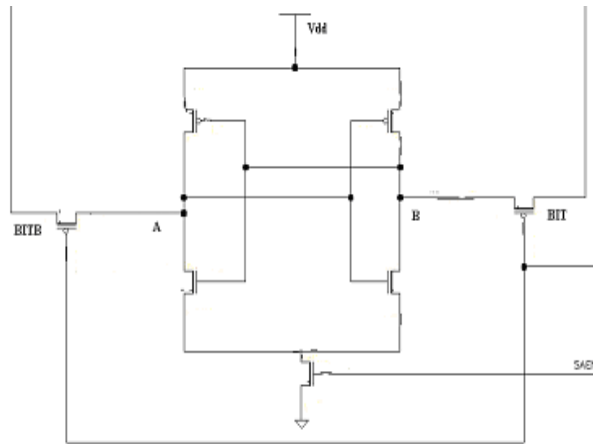


Figure 6.12: Cross coupled ( latch type ) Sense Amplifier

## 6.8 Sense Amplifier Offset Analysis

An ideal Sense-amplifier will have infinitely small offset ( $> 0$  Volt). (Offset is the minimum voltage differential needed at input nodes to resolve the latch in correct state). But in reality, a Sense-amplifier will have finite offset. If this finite offset is not accounted in design, then it may lead to memory failure. It is therefore critical to characterize this offset requirement accurately. This offset may be due to the following reasons [10]:-

**Device Variation:** Sense-amplifier transistors which are fabricated in very identical condition may have variation in their device parameters. This variation is called Device variation. This variation may exist because of mask-misalignments and stress effects or statistical doping concentration differences [10].

**Capacitance Variation:** The total metal capacitance (load) on internal nodes of sense-amplifier (SA) may not be same. This capacitance mismatch can occur due to layout placement or due to fabrication process.

**X-talk:** The unequal coupling capacitance of internal nodes of sense-amplifier with other nodes in layout will contribute to unequal voltage injection due to cross-talk.

As shown in Figure 6.12 A / B are the internal nodes of sense amplifier. SAEN is the sense enable signal which cuts off pass gates and turns Sense Amplifier ON. BIT/BITB are the nodes, one of which will get discharged and transfer its voltage to corresponding sense-amplifier internal nodes A / B through passgates. Sense Amplifier Offset is characterized as

$$\text{Offset of SA} = \text{device\_offset (due to mismatch between devices)} + \text{cap\_variation\_offset (due to total cap mismatch)} + \text{x-talk (due to different coupling)}$$

This analysis has to be done for XTALK, RCMAX, RCMIN and RCTYP extractions separately for the first time. Once the worst extraction mode is finalized for the given architecture, only that may be used for future iterations.

**Device Offset:** The spice models have equations that represent the spread in mobility ( $\mu$ ) and threshold voltage ( $V_t$ ) of transistor within one process corner (lot point). Monte Carlo simulation uses these equations and applies random variation. So, depending on the mismatch the sense-amp will require voltage difference ( $v_{diff}$ ) between its internal nodes above a certain threshold for which circuit will operate correctly [10].

For each run of Monte Carlo, first the differential voltage (VDIFF) between SA internal nodes is kept at zero and the direction of sense resolution is checked. In the same simulation (same set of mismatch), by using the .alter command,  $v_{diff}$  is given in opposite direction and is gradually increased. The  $v_{diff}$  at which the sense resolves in the opposite direction is found. This value of  $v_{diff}$  is the offset value for this particular set of mismatch. Similarly the offset is found for each Monte Carlo run. The number of monte-carlo runs to be launched can be found by a simple experiment. Run simulation for 1000, 300, 200 monte-carlo runs and find the difference in value of standard deviation of distribution with these simulations. The minimum number of simulations at which the sigma value is acceptable can be taken. The pulse width given to sense amp is very relaxed. The absolute VDIFF created between the internal



nodes of the sense amplifier should be measured as a figure of merit [10].

**Capacitance Variation Offset:** Nominal Offset requirement of SA is described as VDIF in between its internal nodes required for SA to work correctly when SA is not under any mismatch. For capacitance mismatch we will be applying 5% metal capacitance mismatch between matched internal nodes of sense-amplifier internal nodes (+/- 2.5% on each internal node. Then perform corresponding read (0 or 1) operation which makes SA to work in worse case environment [10].

For example, Let us suppose that read0 is detected when A discharges and read1 is detected when B discharges. We increase the capacitance of A for modeling capacitance mismatch. This would decrease discharge rate ( $dV/dT$ ) at A. So, this SA would favor read1 than read0. So for calculating impact due to capacitance mismatch we would perform read0.

We would find the VDIF that is needed to make SA (with capacitance mismatch) pass. We say this voltage difference as capacitance mismatch induced offset. Then,

$$\text{Cap\_variation\_offset} = \text{Capacitance Mismatch Induced Offset} - \text{Nominal Offset}$$
 [without any device and capacitance mismatch]

**X-talk:** The SA internal nodes might see unequal or unmatched capacitance from one common node.

For example, In Figure 6.12 A and B may have different coupling capacitance with SAEN. This will cause unequal voltage injection at the SA internal nodes in case this common node makes transition. All such differential injection (during active cycle) should be added in the offset requirement for the SA. While such differential injection of signals that are toggling in the sense amplifier setup will automatically be taken care of, injection due to static signals (data/ mask buses etc.) should be carefully added.

## 6.9 Sense Amplifier Pulse Width Analysis

Reaction time of a Sense Amplifier at a given differential voltage is the time taken by the sense amplifier to resolve the correct output when the said differential is applied at the matched nodes. Reaction time decreases when the differential voltage applied is increased [10].

Sense Amplifier is usually activated by a pulsed signal. The duration of this pulse has to be more than the reaction time of the sense amplifier at the applied  $v_{diff}$  value

The pulse width is usually implemented by a chain of inverters. If the pulse generator is in Dummy or Replica circuits then the designer has to qualify the applied pulse width (measured in the self-timing setup) against the required reaction time (measured in the sense-amp setup) [10]:

$$(\text{Nominal pulse width}) - 3 * (\text{Standard deviation of pulse width}) \geq (\text{Nominal reaction time}) + 3 * (\text{Standard deviation of reaction time})$$

The required reaction time is measured by applying the minimum  $v_{diff}$  (Nom-3\*sigma) obtained in self time simulations to the sense amp setup and running Monte Carlo simulations.

## 6.10 Write Self Time Analysis

Writability of memory cell is a critical issue in nanometer range technologies and process variation is high. The process people are unable to manage huge statistical variation and hence write operation has to be tracked [10].

Write operation tracking is mandatory for all robust memory cells. To qualify write self-time, we need to define how much write window must be provided so that write operation is guaranteed for all memory cells. For write operation, the memcell wordline goes high and depending on the data corresponding bitline goes low.

We can define **write window** as the maximum window available for forced writing into memory cell. It can be represented as:

Minimum of Wordline de-selection time (wordline falling to 95%), Bitline precharge on time (bitline getting precharge to 5% - Maximum of Wordline selection time (wordline rising to 95%), Bitline discharge (bitline discharge to 5%) [10]

We can define **margin-after-write** as :

Minimum of Wordline de-selection time (wordline falling to 95%), Bitline precharge on time (bitline getting precharge to 5% - Maximum of Memcell internal node rising 95%, Memcell internal node falling 5%

Also, we can define **write time** as :

Maximum of Memcell internal node rising 95%, Memcell internal node falling 5% - Minimum of Wordline selection time (wordline rising to 95%), Bitline discharge (bitline discharge to 5%)

For write selftime simulations we use reduced netlist such that it only contains the devices which affect the write time or write window in memory. We then run monte-carlo simulation on this reduced netlist and obtain the distribution of **write window** and **margin-after-write**.

Criteria for passing the write selftime analysis is :

*Nominal value of margin-after-write - 3 \* (standard deviation of margin-after-write) > 0*

## 6.11 Read Self Time Analysis

When Read/Write operations start synchronously with one clock signal and are ended by memory itself, self timing can be implemented by using a model path (Dummy path or Replica path) to track the operation inside the memory. Replica path, as the name suggests, should replicate the real memory cell access path. Basic methodology behind dummy cell is to produce a reset signal for word line. Two advantages of using self-time memory are to gain high speed and reduce dynamic power [10].

The SA amplifies a small differential on Bitlines (BLs) into a digital output correctly when the BLs has a minimum differential voltage (VDIFF) between them.

The SA is switched ON by a replica path (also called self-time (ST) circuit/ loop). The ST circuit needs to be tuned to ensure sufficient VDIFF at internal nodes of SA.

The VDIFF that may appear at SA nodes is not constant and depends upon the mismatch encountered during fabrication i.e. device variation, parasitic mismatch, process mismatch between memcell and logic devices. We need to design ST circuit such that even in the worst case, we get sufficient VDIFF at the SA nodes (more than its requirement). In Self-Time Read analysis VDIFF is measured between internal nodes of sense-amp when sense-enable is rising during read operation, which should be more than Offset of sense-amplifier.

# Chapter 7

## Results & Waveforms

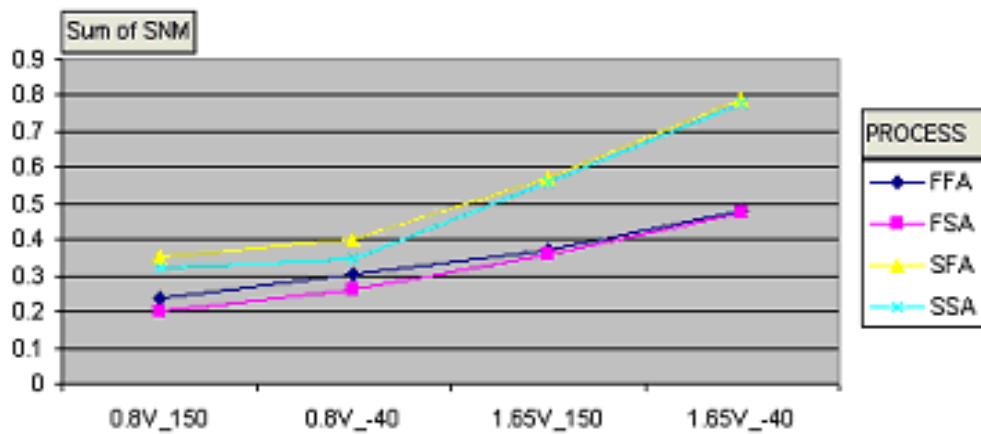


Figure 7.1: Results of SNM of Memcell

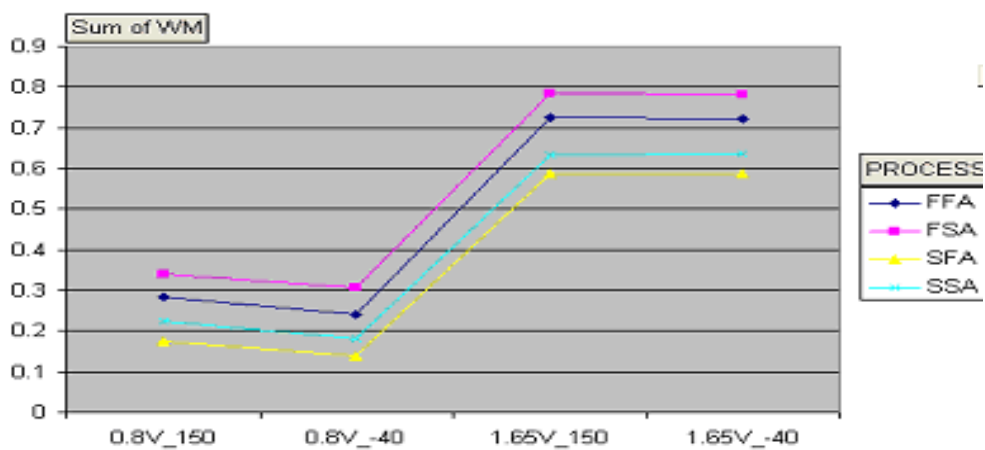


Figure 7.2: Results of Write Margin of Memcell

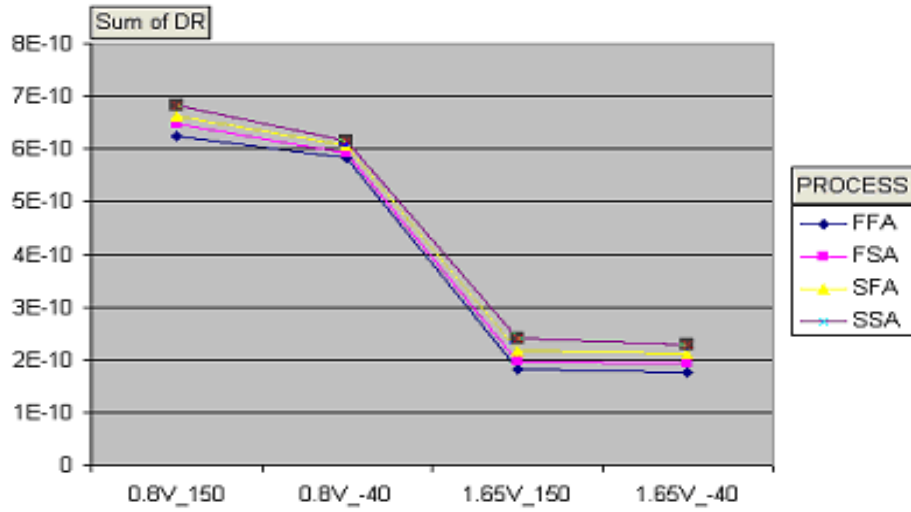


Figure 7.3: Results of Discharge Rate of Memcell

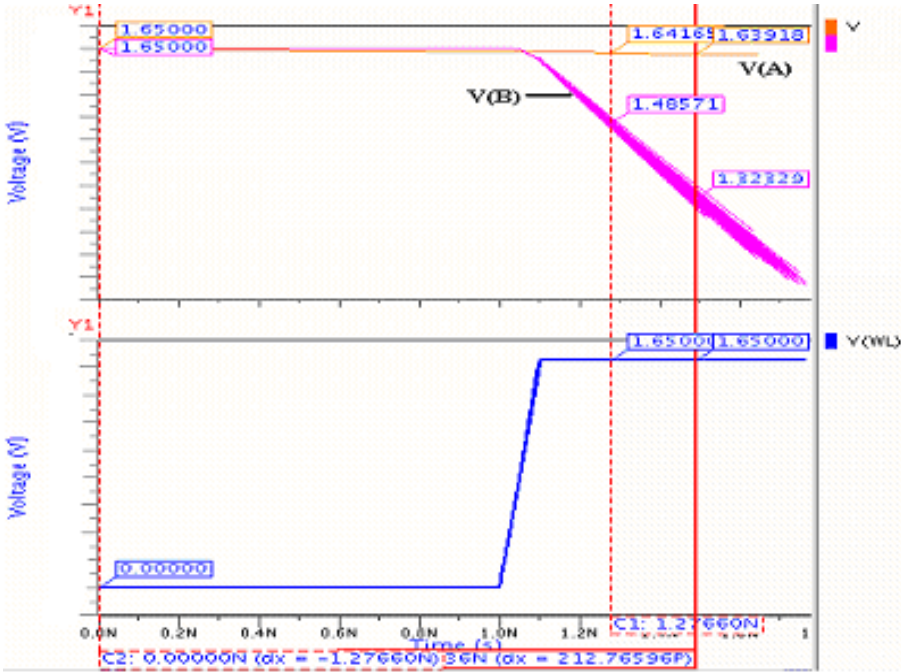


Figure 7.4: Waveform of DR with MC Simulation

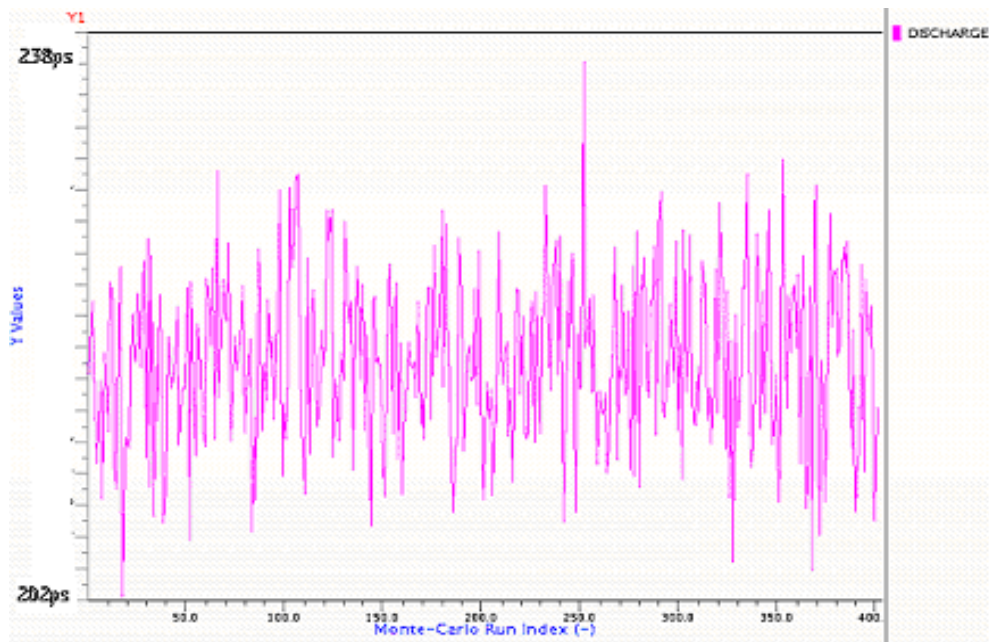


Figure 7.5: Variation of DR with 400 MC Simulation

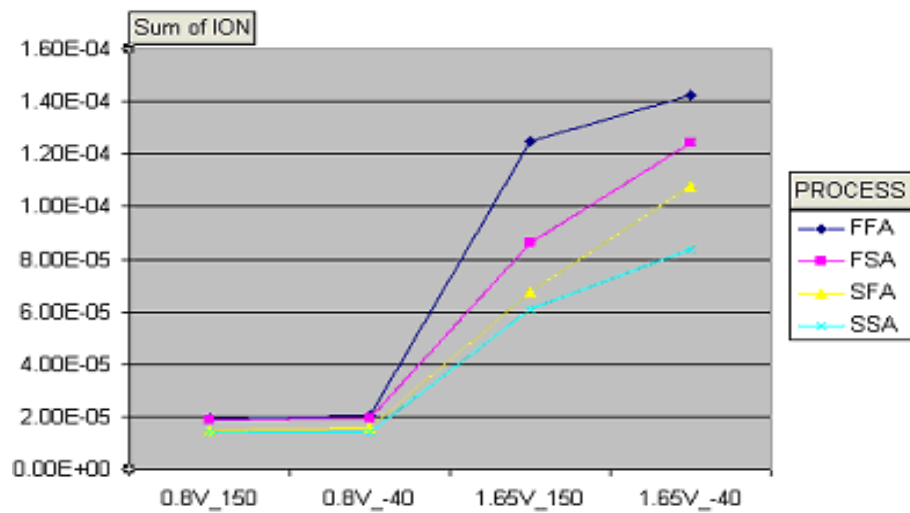


Figure 7.6: Results of ION Currents of Memcell



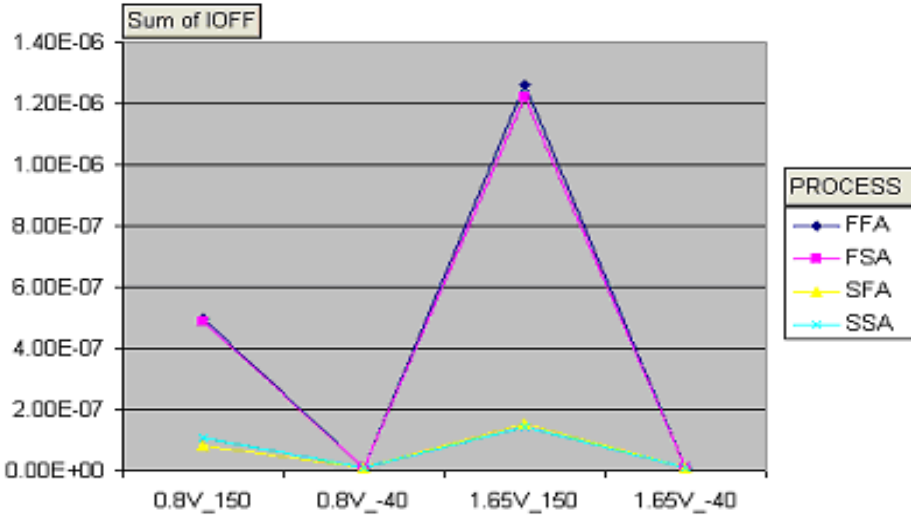


Figure 7.7: Results of IOFF Currents of Memcell

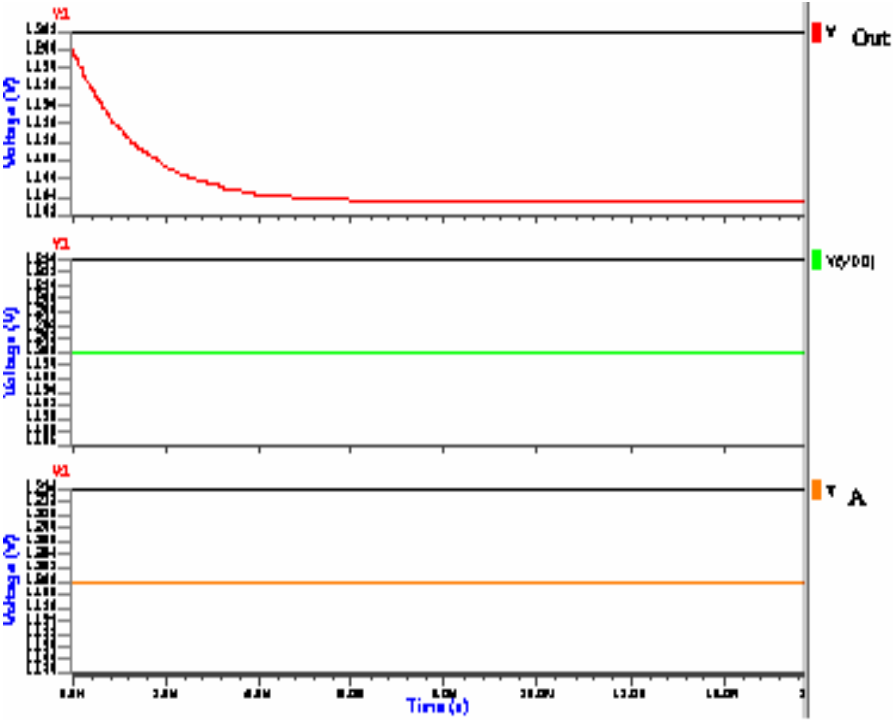


Figure 7.8: Waveform of Leakage Validation of Row Decoder

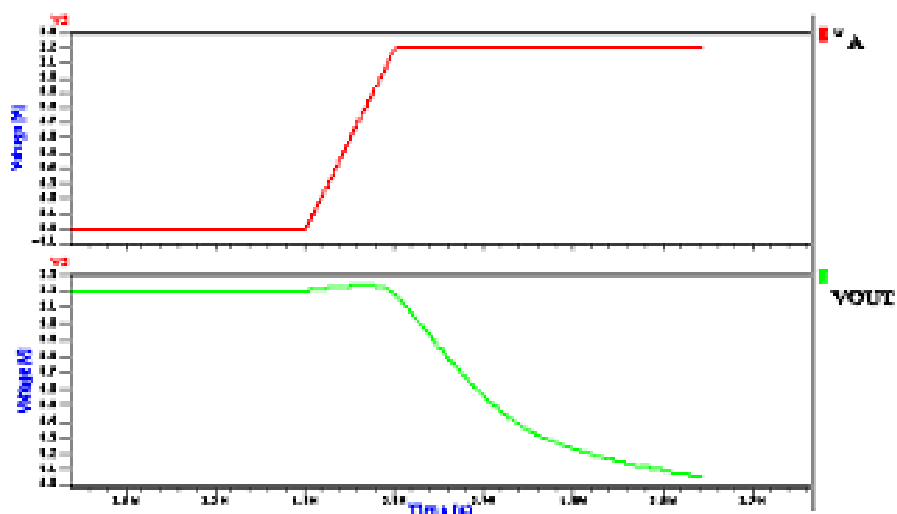


Figure 7.9: Waveform of Strength Validation of Row Decoder

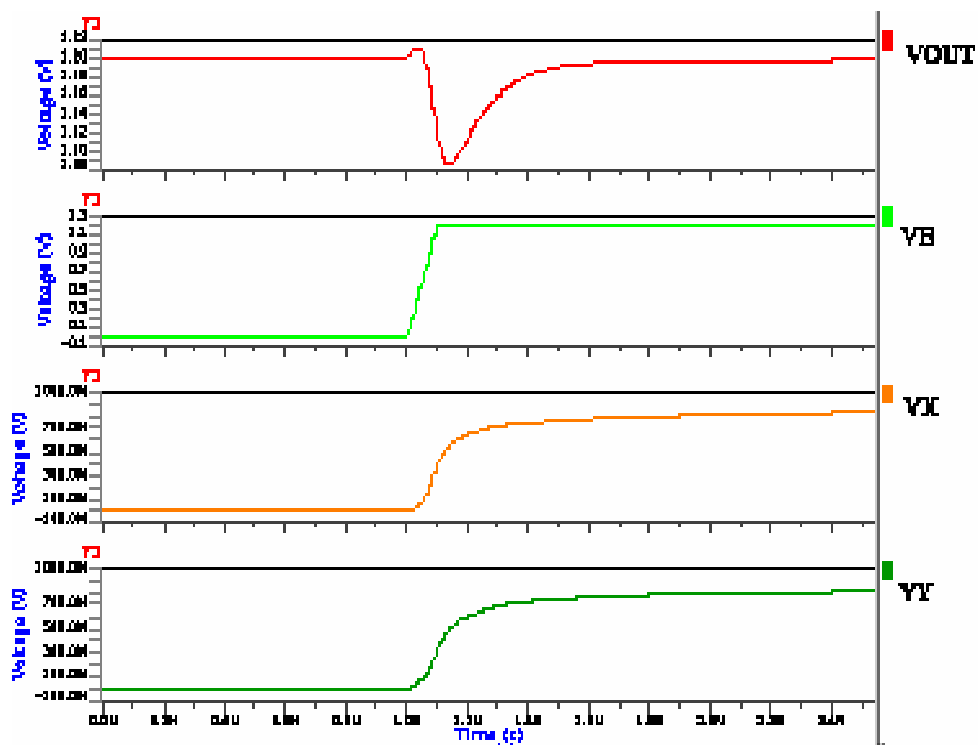


Figure 7.10: Waveform of Charge Sharing Validation of Row Decoder

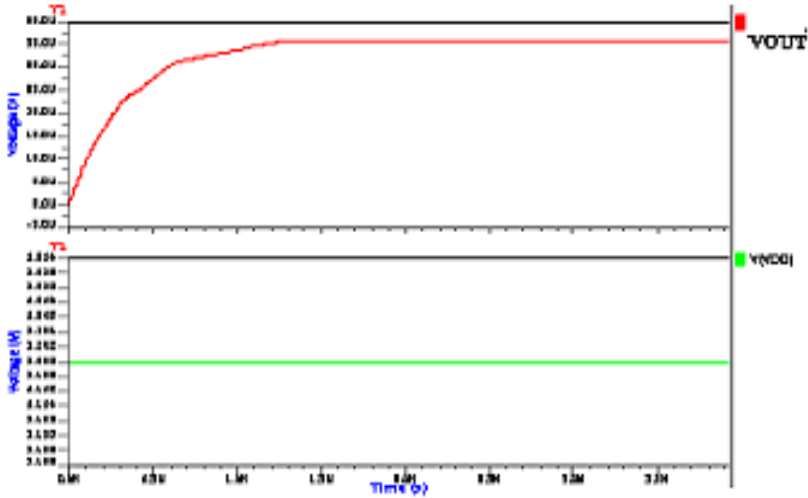


Figure 7.11: Waveform of Bump Validation of Row Decoder

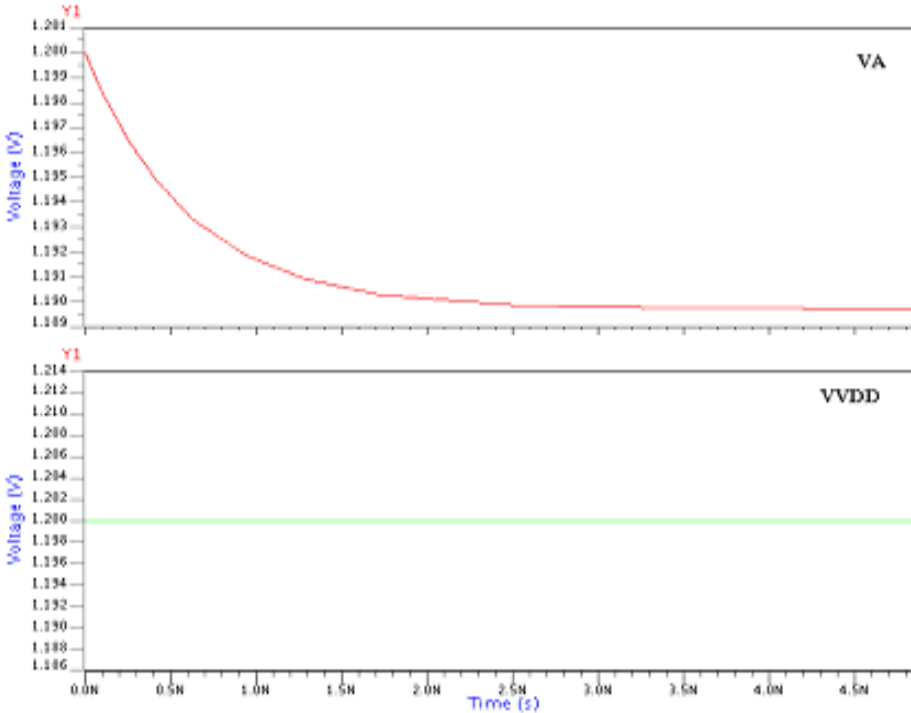


Figure 7.12: Waveform of Leak\_1 of IO Latch Analysis

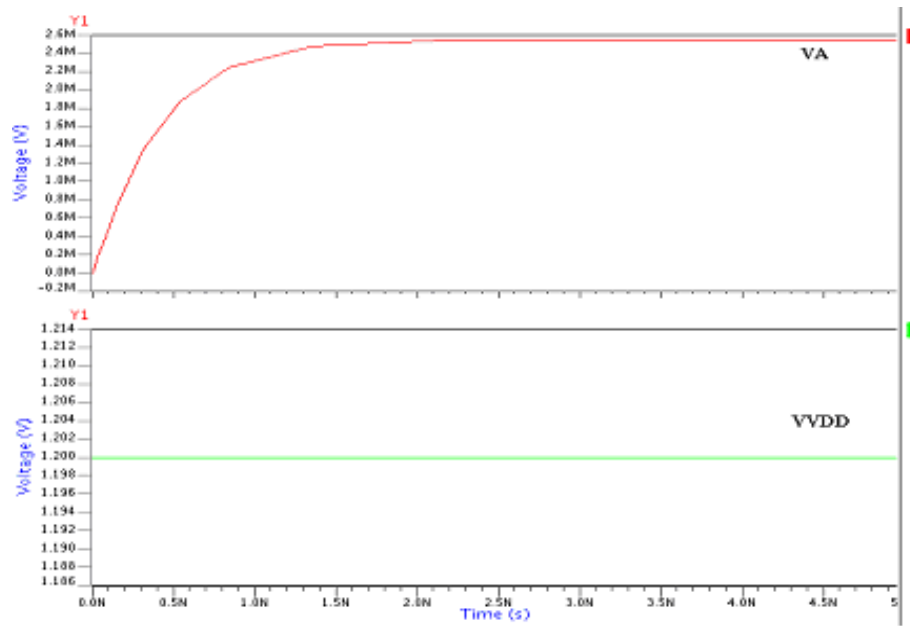


Figure 7.13: waveform of Leak\_0 of IO Latch Analysis

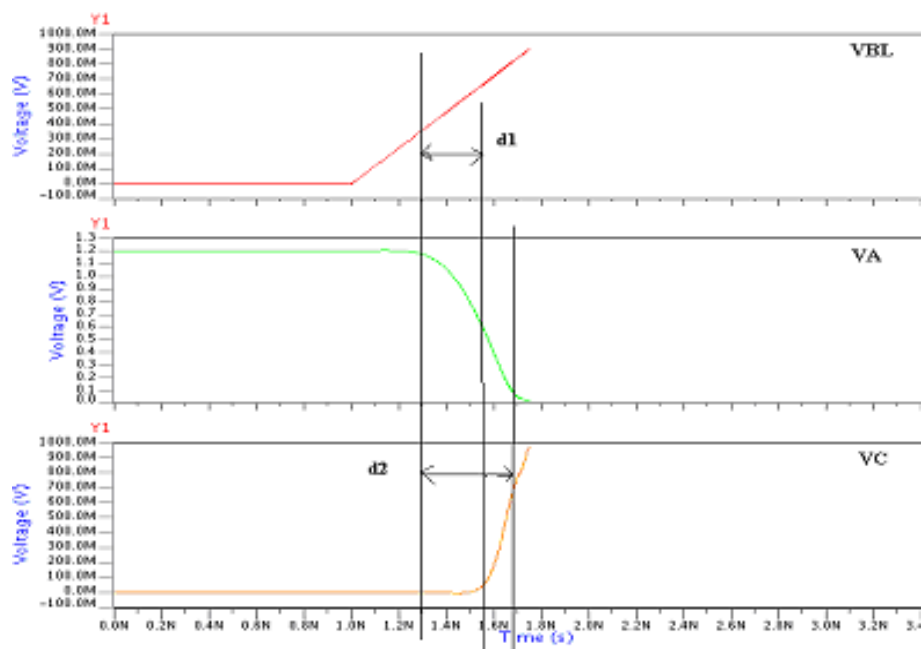


Figure 7.14: waveform of Write\_0 of IO Latch Analysis

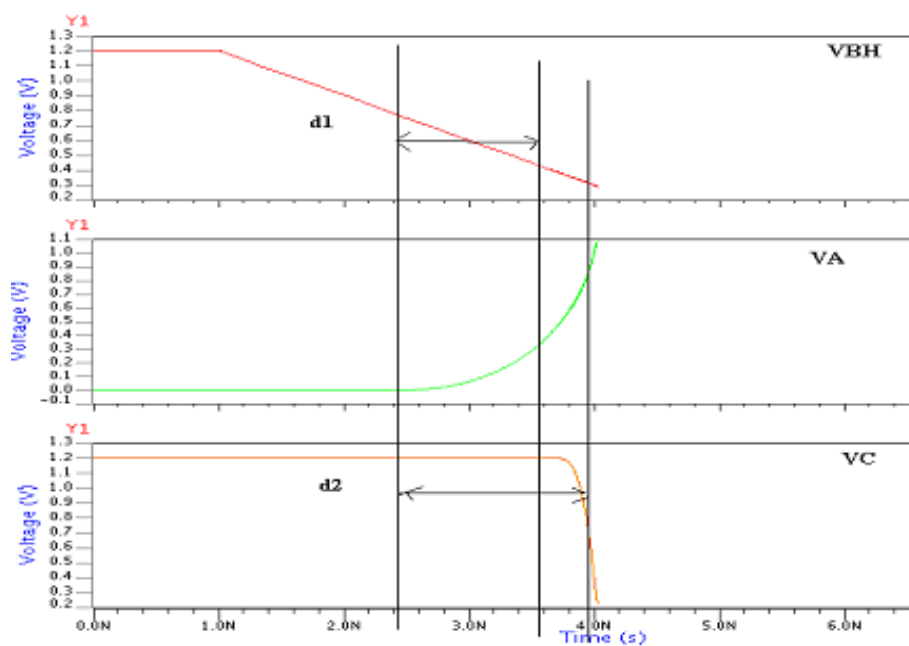


Figure 7.15: waveform of Write\_1 of IO Latch Analysis

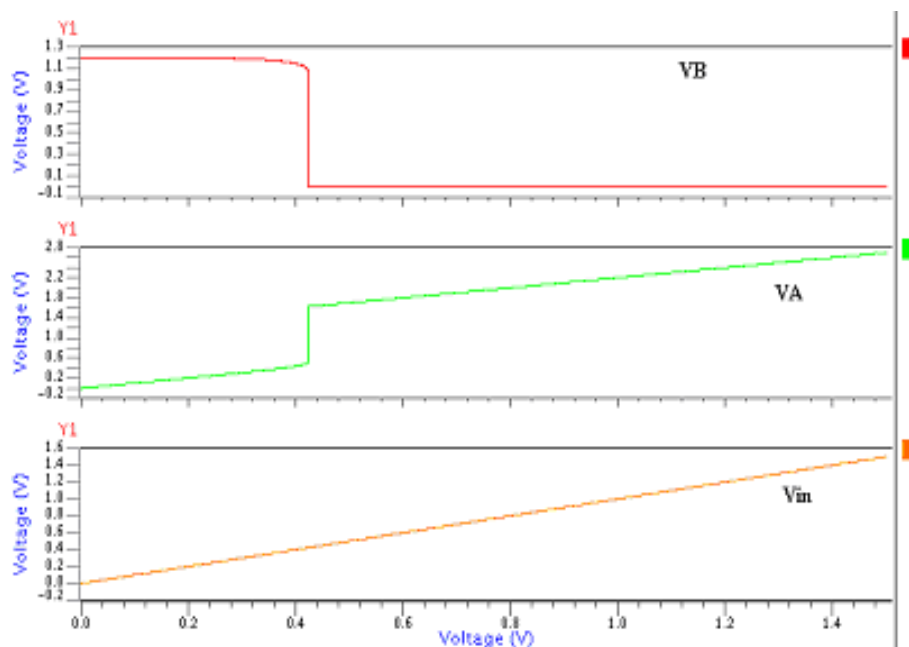


Figure 7.16: waveform of SNM.0 of IO Latch Analysis

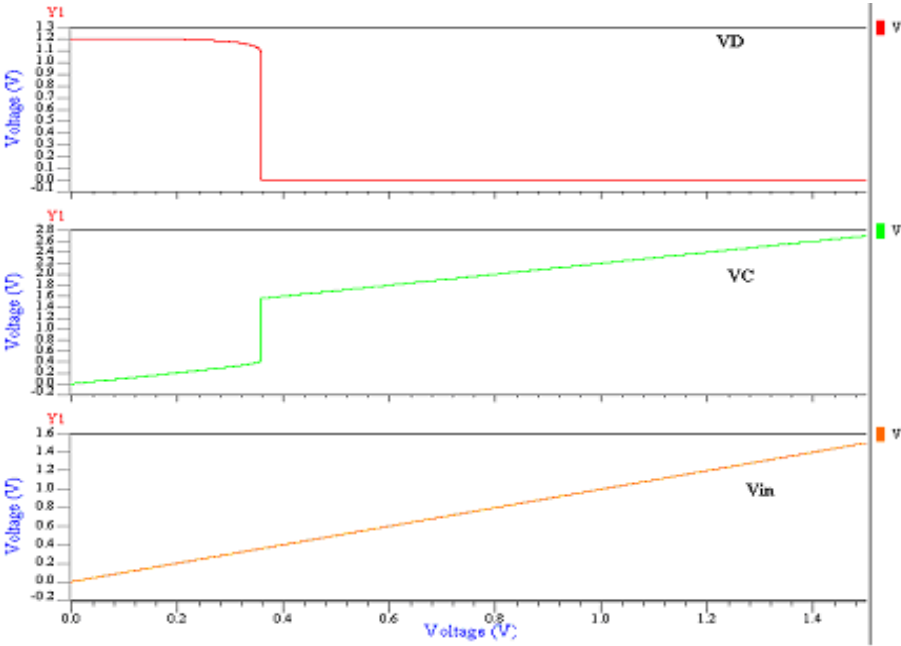


Figure 7.17: waveform of SNM.1 of IO Latch Analysis

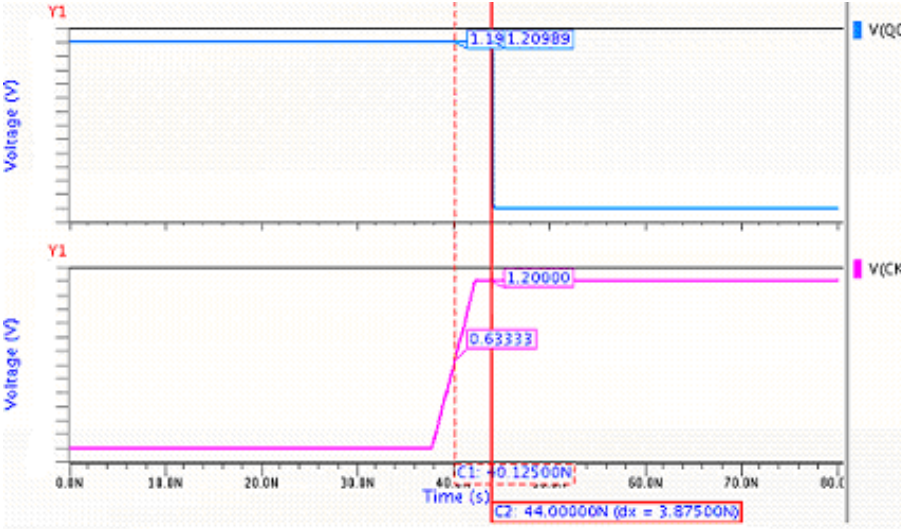


Figure 7.18: Waveform of Access Time

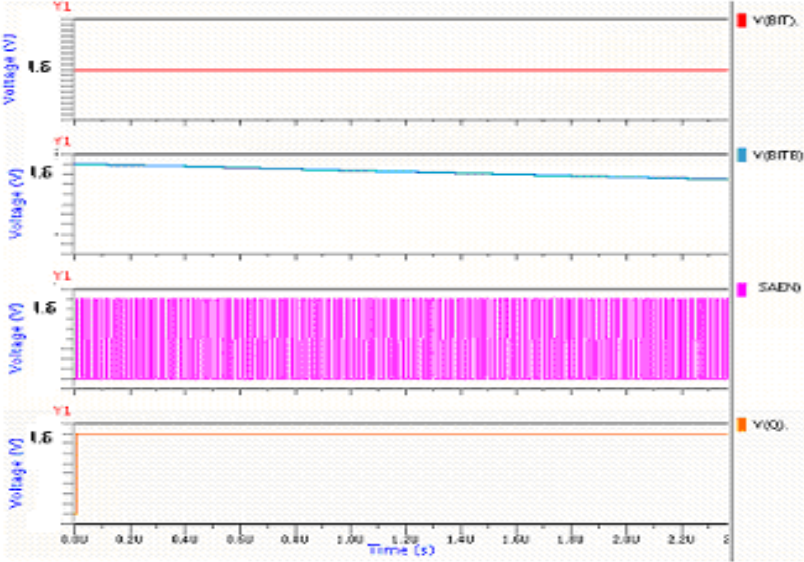


Figure 7.19: waveform of Sense Amplifier Offset

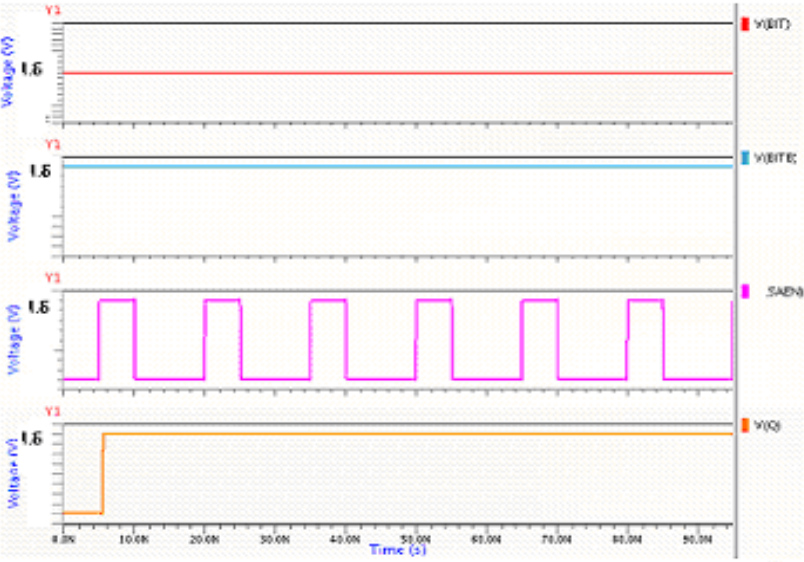


Figure 7.20: waveform of Sense Amplifier Offset(zoom)

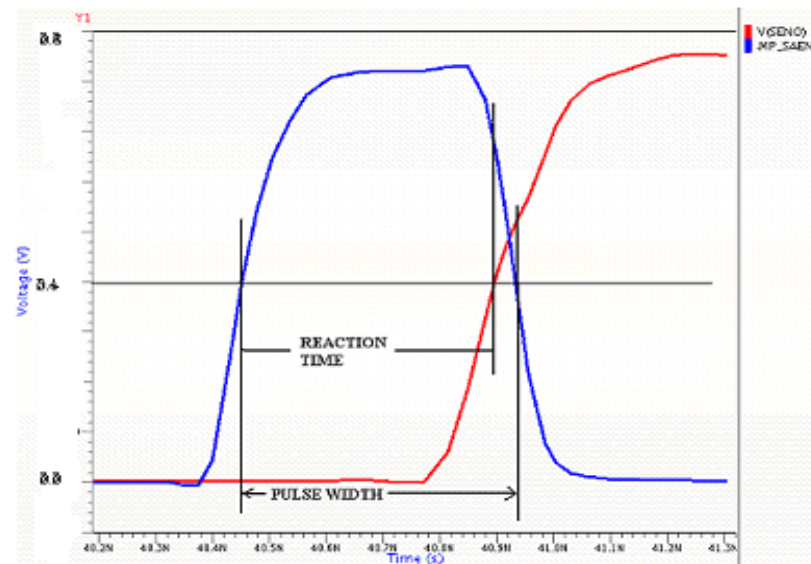


Figure 7.21: Waveform of SA Pulse Width Analysis

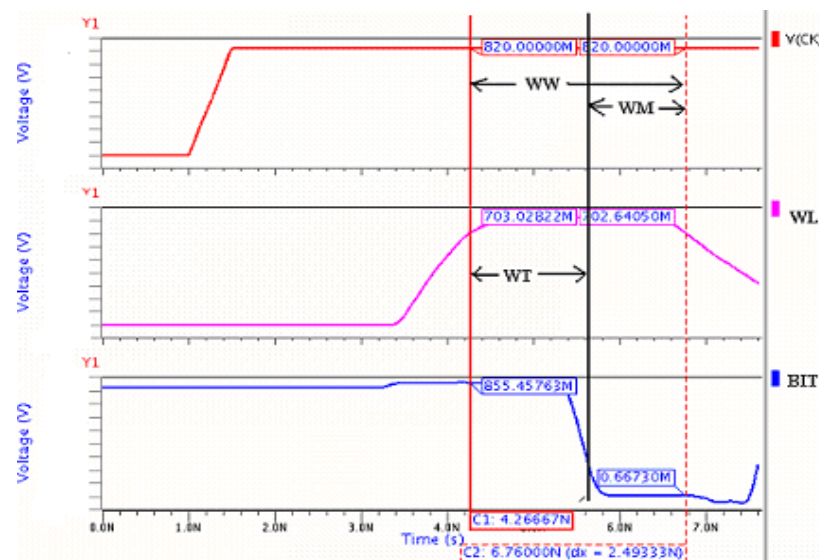


Figure 7.22: Waveform of Write Self Time Analysis



# Chapter 8

## Conclusion & Future Scope

### 8.1 Conclusion

In the company, the work I have done has greatly increased my knowledge in the field of design flow. I have learnt a lot about the working and the architecture of the memory compiler and how the various operations are performed in a memory. The knowledge of UNIX and its commands also helped me a lot in understanding the various scripts. The memory compiler given to me is validated completely through all necessary analysis and validations. I also helped my team member, working with a similar project, in debugging some of the problems giving me a sense of team work. I was made to work on the various live projects that have increased my confidence and sense of responsibility.

### 8.2 Future Scope

The SRAM Compiler generates different memory cuts according to the customer specifications. These SRAMs are used in various Automotive Products. These memories are used in various automatic parts of car like rain sensor, adaptive cruise control, water pump, LIN battery sensor, mirror control, active suspension, climate control, parking system, central body module, power window, car multimedia, navigation

and telematics system, engine control, charging, GPS module, satellite radio, RF tyre pressure system, dash board control, airbag control, engine control, vehicle stability control, power steering, seat control, light sensor, door and roof zone systems, digital mobile video. These memories are also used in mobile processing systems. These memories are also used in various computer periphery applications.

# Index

Bits, 28  
BUMP, 59  
Capacitance Ratioing, 39  
CDL, 45  
Cell Current Ratioing, 42  
CK, 30  
CPM, 33  
CSN, 30  
DR, 54  
GDS, 45  
HVT, 29  
I/O, 19  
Inverter Chain Technique, 37  
IOFF, 55  
ION, 55  
MC, 26  
MCF, 48  
MCP, vii  
MEMCELL, 11  
MOS, 6  
MUX, 21, 28  
MVT, 29  
NXP, vii  
OTF, 50  
PLS, 45  
Replica Technique, 36  
ROM, 8  
ROW-DEC, 57  
RST, 72  
SA, 22  
SNM, 52  
SRAM, 6  
VDIFF, 72  
WEN, 30  
WM, 53  
Words, 28  
WST, 70