Performance Evaluation of SRAM Memory Compiler

Major Project Report

Submitted in partial fulfillment of the requirements

for the degree of

Master Of Technology

in

Electronics & Communication Engineering

(VLSI Design)

By
Anuj Patel
(13MECV16)

Electronics and Communication Engineering Branch
Electrical Engineering Department
Institute Of Technology
Nirma University
Ahmedabad-382481
May 2015

Performance Evaluation of SRAM Memory Compiler

Major Project Report

Submitted in partial fulfillment of the requirements

for the degree of

Master Of Technology

in

Electronics & Communication Engineering

(VLSI Design)

By

Anuj Patel

(13MECV16)

Under the guidance of

External Guide	Internal Guide
Mr. Ashish Sharma	**Dr. Usha Mehta**
R&D Manager,	Professor EC,
Synopsys India Pvt. Ltd.	Nirma University.

Electronics and Communication Engineering Branch
Electrical Engineering Department
Institute Of Technology
Nirma University
Ahmedabad-382481

May 2015

Certificate

This is to certify that the Major Project entitled **" Performance Evaluation of SRAM Memory Compiler"** submitted by **Anujkumar Kantilal Patel (13MECV16)** , towards the partial fulfillment of the requirements for the degree of Master of Technology in VLSI Design , Nirma University, Ahmedabad is the record of work carried out by him under our supervision and guidance. In our 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 our knowledge,haven't been submitted to any other university or institution for award of any degree or diploma.

Dr. Usha Mehta
Internal Project Guide

Dr. N. M. Devashrayee
PG Coordinator(VLSI Design)

Dr. P. N. Tekwani
Head of EE Dept.

Dr. Ketan Kotecha
Director, IT-NU

Date:

Place : Ahmedabad

Synopsys India Pvt Ltd.

A-36, Basement and Ground floor,
Sector 64,
Noida -201301
Uttar Pradesh, India

This to certify that **Mr. Anujkumar Knatilal Patel (13MECV16)**, student of M.Tech EC (VLSI Design), Institute of Technology, Nirma University has undergone training in our Solution Group (SG) from **9th June 2014** to **8th May 2015**. He has successfully completed his training on **"Performance Evaluation of SRAM Memory Compiler"**.

During the training period, we found him very sincere and professional in his conduct.

We wish him all the best in his future endeavors.

Signature:

Date & Place:

Declaration

This is to certify that

a. The thesis comprises my original work towards the degree of Master of Technology in VLSI Design at Nirma University and has not been submitted elsewhere for a degree.

b. Due acknowledgment has been made in the text to all other material used.

- Anuj Patel
13MECV16

Acknowledgements

First and foremost, sincere thanks to Dr. N. M. Devashrayee, P.G. Coordinator of VLSI design, Institute of technology, Nirma University, Ahmedabad and my internal guide Dr. Usha Mehta. I enjoyed their vast knowledge and thank them a lot for giving valuable support for project work.

I would like to thank my Project Manager, Mrs. Monila Juneja (Manager R & D, Synopsys Private Limited, Noida) for her valuable guidance. I would also like to thank my project guide, Mr. Ashish Sharma (Manager R & D) for his support throughout my project.

I would like to thank Mr. Nitesh Gautam (Manager R & D) for his helping part in my project. Throughout the training, he has given me much valuable advice on project work which I am very lucky to benefit from. Without them, this project work would never have been completed.

I would also like to thank Dr. Ketan Kotecha, Director, Institute of Technology, Nirma University, Ahmedabad for providing me an opportunity to get an internship at Synopsys India Private Limited, Noida.

I would like to thank my all faculty members for providing encouragement, exchanging knowledge during my post graduate program.

I also owe my colleagues in the Synopsys, special thanks for helping me on this path and for making project at Synopsys more enjoyable.

- Anuj Patel
13MECV16

Abstract

Memory currently occupies a large part of system on chip, approximately sixty percent, therefore the reduction of power and delay in memory become important issues. In such a condition it is require to find the cause of power consumption and delay in periphery and array part, if some how one can remove the sources of power dissipation and delay, it will improve the performance of the system. So aim of this project is to design and analyze of Static Random Access Memory (SRAM), focusing on optimizing delay. Hence, increase the speed of the memory. In typical memory architecture, access time is the sum of clock to word line path delay and word line to output path delay. In this Project, one memory instance is implemented with three different architectures like simple architecture, bank partition and split core with bank partition using a memory compiler. Then analyze and compare the speed of all this, three different architecture. Secondly, it is observed that the address decoder introduce significant delay from clock to word line path delay. Mostly in industry, NAND gate is preferred to implement decoder and hence, simulated the NAND gate with different fan-in. From that it is observed that, for more than three input delay from input to output increase significantly. In new address decoding scheme, whole decoder is divided into two stages, predecode and postdecode. The predecode stage generates intermediate signals that are used by multiple gates in the final decode stage. As a result fan-in for the NAND gate reduces. Here in this project, 4x16 decoder is simulated with two different decoding scheme. From simulation it is observed that delay from clock to word line path delay, for pre-post address decoding scheme is reduces significantly. Above all work is to optimize the delay. Now to improve the performance, like write-ability at lower supply voltages. The operation of the SRAM at lower supply voltages becomes even more challenging. The dominant yield loss from increased device variability occurs at minimum operating voltage. The failures at Vmin can be due to write failure, read failure. In this project, write-assist techniques are described.

Abbreviations

NW	Number of Words
NB	Number of Bit
BK	Bank partition
CM	Column Mux
CD	Split core
RPB	Row per Bank
CPS	Column per side
PR	Physical Row
PC	Physical column
WA	Write assist
SNM	Static noise margin
DRNM	Dynamic read noise margin

Contents

Certificate **i**

Declaration **iii**

Acknowledgements **iv**

Abstract **v**

Abbreviations **vi**

List of Figures **xi**

1 Introduction to Memory Architecture **1**

1.1 Memory architecture 1

1.1.1 Main Array 2

1.1.2 Control Block 2

1.1.3 Row and Column decoder 3

1.1.4 Multiplexer 3

1.1.5 Sense Amplifier 3

1.2 Role of Memory Compiler 3

1.3 Features of Memory compiler 4

1.4 Thesis Organization 5

2 Analysis of 6T SRAM using Different Architecture **6**

2.1 Introduction 6

2.2 Architecture of 1 Kb 6T SRAM 7

2.2.1 Simulation . . . 9
2.2.2 Result . . . 12
2.2.3 Plot . . . 12
2.3 Summary . . . 13

3 Address Decoding Scheme In SRAM Memory Compiler 14
3.1 Introduction . . . 14
3.2 CMOS Decoder . . . 15
3.2.1 Simulation for NAND Gate to compare Delay from Input to Output . . . 16
3.2.2 Result . . . 19
3.2.3 Plot . . . 19
3.2.4 Simulation Waveform . . . 20
3.3 Static address decoder . . . 20
3.4 Pre-post Address Decoding Scheme . . . 22
3.4.1 Important Definition . . . 23
3.4.2 Result . . . 24
3.4.3 Comparision . . . 24
3.5 Summary . . . 25

4 Write Assist technique in SRAM 26
4.1 Introduction . . . 26
4.2 Proposed write assist techniques . . . 27
4.2.1 Reducing the Vddc voltage . . . 28
4.2.2 Increase the Vss core voltage . . . 29
4.2.3 Word-line boosting . . . 30
4.2.4 Negative bit-line capacitive coupling . . . 31
4.3 Impact on half selected cell due to write assist techniques . . . 32
4.4 Summary . . . 33

5 Conclusion And Future scope 34

References 35

List of Figures

1.1 Memory Architecture 2
1.2 Role of Memory compiler 4

2.1 1 Kb SRAM Core 7
2.2 1 Kb SRAM Core with Bank Partitioning 8
2.3 Transistor Count 8
2.4 Memory Architecture with BK=1 and CD=0 9
2.5 Memory Architecture with BK=1 and CD=1 10
2.6 Memory Architecture with BK=2 and CD=1 11
2.7 Delay for three different case 12
2.8 Plot for different architecture Access time 12

3.1 AND based decoder 15
3.2 NOR based Decoder 16
3.3 NAND2 gate with transistor and its RC equivalent 17
3.4 NAND4 gate with transistor and its RC equivalent 18
3.5 Delay for different Fan-in of NAND gate 19
3.6 NAND Gate Delay Vs Fan-in 19
3.7 Delay for different Fan-in 20
3.8 NAND4 based decoder 21
3.9 NAND4 based decoder delay at different temperature 21
3.10 Block Diagram of Pre-post address decoding Scheme 22
3.11 Pre-post address decoder delay at different temperature 23
3.12 Pre-post scheme based address decoder delay at different temperature 24

3.13 waveform which shows the delay from CLK to Wordline at 25 degree temperature 24

4.1 Variation in SNM with supply voltage reduction 26
4.2 6T SRAM bit-cell 27
4.3 6-T SRAM Bitcell 28
4.4 Timing relationships using the Vddc lowering WA scheme 28
4.5 Simulation result for Vddc lowering WA scheme 29
4.6 Timing relationship of Vsscore rising WA scheme 29
4.7 Simulation result for Vsscore rising WA scheme 30
4.8 Timing relationship of Word-line boosting WA scheme 31
4.9 Simulation result for word-line boosting WA scheme 31
4.10 Timing relationship of Negative bit-line WA scheme 32
4.11 Simulation result for negative bit-line WA scheme 32

Chapter 1

Introduction to Memory Architecture

In general, the memory architecture is divided in two part namely, Main array and Periphery. In array part, the Memory Bit-cells are located and to read data from bit-cells or to write data in to bit-cells, the circuits called periphery of memory architecture are designed.Here role of memory compiler is also described.

1.1 Memory architecture

The classical semiconductor memory architecture is shown in Fig.[1.1]. The whole memory architecture is divided into following part.

a. Main Array

b. Control Block

c. Row and column Decoder

d. Sense Amplifier

e. Column Multiplexer

f. Input-Output Block

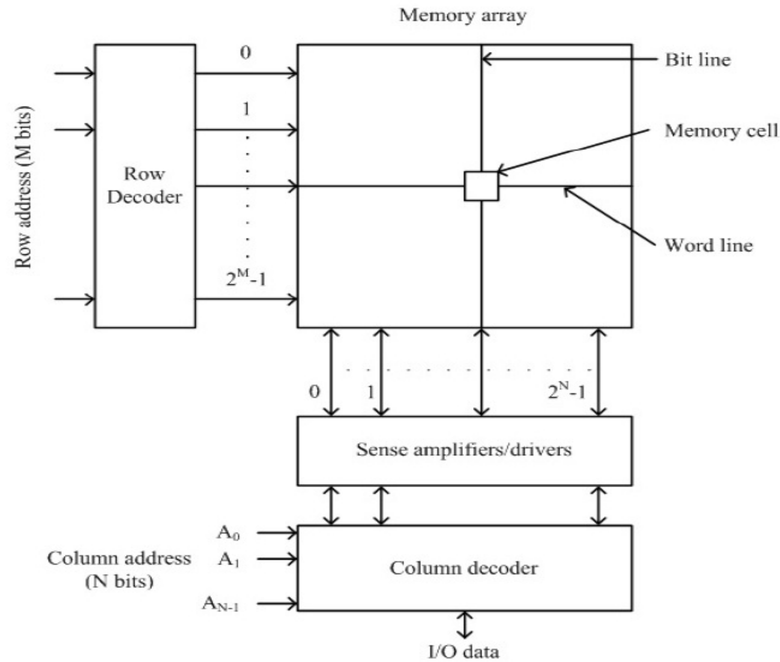


Figure 1.1: Memory Architecture

1.1.1 Main Array

The Main array consists of individual memory cells arranged in an array of horizontal rows and vertical columns as shown in Fig.[1.1]. Each cell is capable of storing one bit of binary information. Each cell is having one common connection with the other cells in the same row (Word line), and another common connection with the other cells in the same column means common bit line).

1.1.2 Control Block

Control block is responsible for generating all control signals like read address signal, column address signal, internal clock generation, write enable, read enable and memory enable. It is having responsibility for latching of all the control signals to provide synchronization throughout memory operation. The generation of the entire signal with valid duty cycle to prevent the set-up and hold time is also taken care by control block. The control block is also generate to power gating signal to reduce the leakage and power when memory is in standby mode.

1.1.3 Row and Column decoder

If there is 2^Nx2^M size of memory, then to there will be 2^N word lines and 2^M bit lines. Now to access a particular bit cell, corresponding word lind and bit line has to be selected through memory address. To do this N$x2^N$ row decoder and M$x2^M$ column decoder are used which will have N bits of row address and M bits of column address as inputs respectively.

1.1.4 Multiplexer

Multiplexer is used to share the Sense Amplifier among some bit lines to reduce the area off the over-all chip design. It is also used to change the aspect ratio of the memory chip. For 2^Nx2^M size of memory, without column multiplexer, there will be 2^N rows and 2^M columns. If column multiplexer (CM) is used, let say CM=2 (2x1 mux), than the physical row will be half of the previous one and physical column will be double of the previous one. Thus the aspect ratio is changed than the previous one. By changing the value of CM, different size of memory (tall, wide, small, big, square) can be generated accordingly.

1.1.5 Sense Amplifier

Sense Amplifier is the most important circuits in the periphery of CMOS memories. Its function is to sense or detect stored data from read selected memory. The performance of sense amplifiers strongly affects both memory access time and overall memory power dissipation. The increased memory capacity are increased bit line capacitance which in turn makes memory slower a,hence by using sense amplifier one can detect data quickly as soon as some differential signal provided to it.

1.2 Role of Memory Compiler

In System on Chip (SoC) design, i.e. customer required memory with different aspect ratio with different size. Memory compiler provide the features to generate the memory instances with different sizes with different features. In this chapter, how user can interface with memory compiler and different features are described.

The working of memory compiler is shown in Fig.[1.2] Memory designer design and architects the memory compiler. For different library, there is a particular compiler is design at particular technology node.The IC designer gives inputs to the memory compiler to generate memory instance.

1.3 Features of Memory compiler

Here is the list of basic features of Memory compiler.

a. Periphery Vt option

b. Dynamic Voltage and Frequency Scaling support

c. Built in self Test

d. Selective Bit-Write

e. Power saving mode

f. Read Margin Control

g. Redundancy Enable

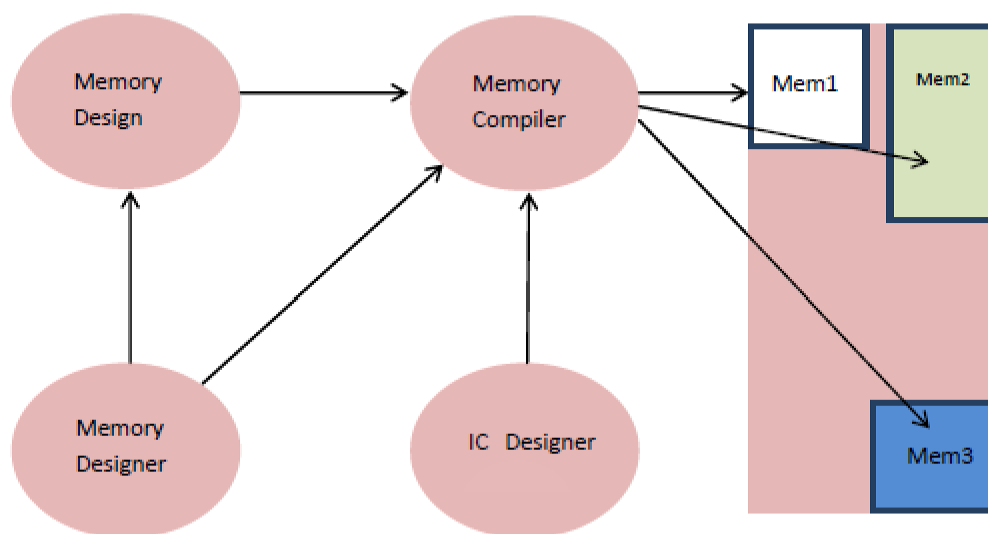


Figure 1.2: Role of Memory compiler

1.4 Thesis Organization

The thesis is organized in 5 chapters, the details of each chapter is as follows

Chapter 1: This chapter describe the Role of memory compiler and memory architecture.

Chapter 2:: Design and Analysis of 6T SRAM using Different Architecture- This chapter include different architecture of Memory along with simulation to analyze the speed of the Memory.

Chapter 3: : Address Decoding Scheme In SRAM Memory Compiler- This chapter describe how Pre-post address decoding scheme for the memory increase the speed of Memory along with simulation.

Chapter 4: This chapter describe the implementation of write assist scheme.

Chapter 5: Conclusion and Future work

Chapter 2

Analysis of 6T SRAM using Different Architecture

Here in this chapter,one memory instance is implemented with different architecture. This chapter include different architecture of Memory along with simulation to analyze the speed of the Memory.

2.1 Introduction

An SRAM is array of memory cells.To access particular Memory address or memory cell, address decoder are used for read and write operation.As a whole memory is divided into two parts array part and periphery part. Address decoder is a part of periphery.The basic block diagram of the SRAM contains arrays of memory cells,control block for address decoder for basic[13]. We can say that performance of system is largely depends on the memory, therefore incrase in speed and reduction in power leakage is important concern as long as performance of system is concern. In cases like this it is very important to reduce the leakage. This issues has been seen in many system on chip memories. As a result it is necessary for designer to determine the cause of delays and leakage in memory blocks. Hence, it can be resolved or reduced, and allowing to find a new better technique. which will improve performance of the system[13].

It is possible to reduce the power leakage by using different techniques. Such as bank structure,such as circuit partitioning, increase the gate oxide thickness so as a result gate oxide reduces, increase the threshold voltage[13]. The bank structure technique increase the speed of memory[13]. The control block, the address decoder and I/o ports are design by using low threshold transistor, while in the design of the bit-cell and sense amplifier high threshold voltage transistor are used[13].

Here the aim is to design and analyze of SRAM, to optimizing the delay[3]. The

whole memory can be partitioned into blocks. The number of words are distribute among with block equally[3]. If somehow we can reduce the word-line capacitance[3] then we can also optimize the power dissipation. The sense amplifier is used to sense the data.To reduce the wordline capacitance we can use centre decoding which we called split-core architecture. And to reduce the bit-line capacitance memory bank architecture is used.

2.2 Architecture of 1 Kb 6T SRAM

The 1 Kb static random access memory structure is shown in fig.[2.1]. It is called random access because any memory location can be accessed in arbitrary order for read and write operation. Here two different types of memory architecture is shown[2]. One is traditional single core architecture and other one is bank structure.The simple architecture contain 32 phsical rows and 32 physical columns. To access each bitcell 5x 32 row decoder is and column decoder (Fig[2.1]) are used.Other fig.[2.2] shows the0 bank structure based architecture[2]. In bank structure entire memory is divided into number of blocks.Here in this figure memory block is divided into the 4 block[2] as shown in Fig [2.2]. Here each block has capacity to store 256 bits.Here actually we distribute the number of words in each block. Traditional architecture contain 32 words of 32 bits. Now in bank structure each block contain 8 words of 32 bits.Here we can see that row decoder of 2x4 is used to select particular one block at a time. To select a particular block block-circuit is used. While one block is accessed,rest of the blocks are in stand by mode.

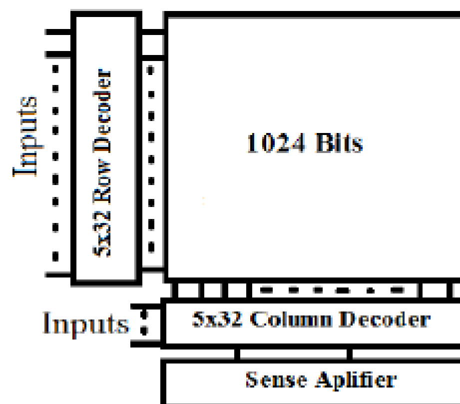


Figure 2.1: 1 Kb SRAM Core

A traditional memory has a only one block with i/o block, control block and row an column decoder [14]. This single block architecture is suited for small amount of memory. But if memory size increases the bit-line and word-line capacitance increases.As a result, the delay and the power dissipation increases.Hence, the speed of the memory reduces. If somehow we can reduce the bit-line and word-line load(resistive and capacitive), we can increase the speed[14]. Till date,there are different kind of techniques are proposed to reduce the bit-line and word-line capacitive and resistive load[14]. One mos used technique is memory bank partition method. In this technique both power and delay reduces, because the bit-line and word-line capacitance load reduces.But it is increase the area of memory.Below fig. [2.2] with Bank Partitioning.

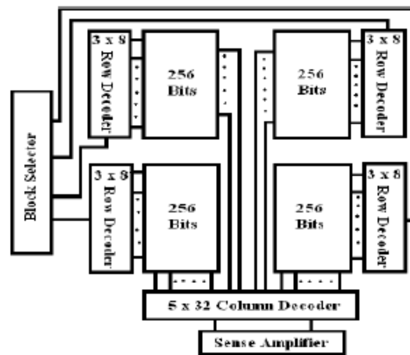


Figure 2.2: 1 Kb SRAM Core with Bank Partitioning

Fig [2.3] shows the comparison of number of transistor used in 1 Kb SRAM core Architecture of Fig [2.1] and Fig[2.2]. In comparison the number of transistor count in Architecture with memory Bank increases by 9.31%.

Sr. No	Component	Transistor Count	
		Without Memory Bank	**With Memory Bank**
1	Precharge circuit	96	384
2	Sense Amp.	160	640
3	Decoder	640	528
4	1 kb SRAM	6144	6144
	Total	**7040**	**7696**

Figure 2.3: Transistor Count

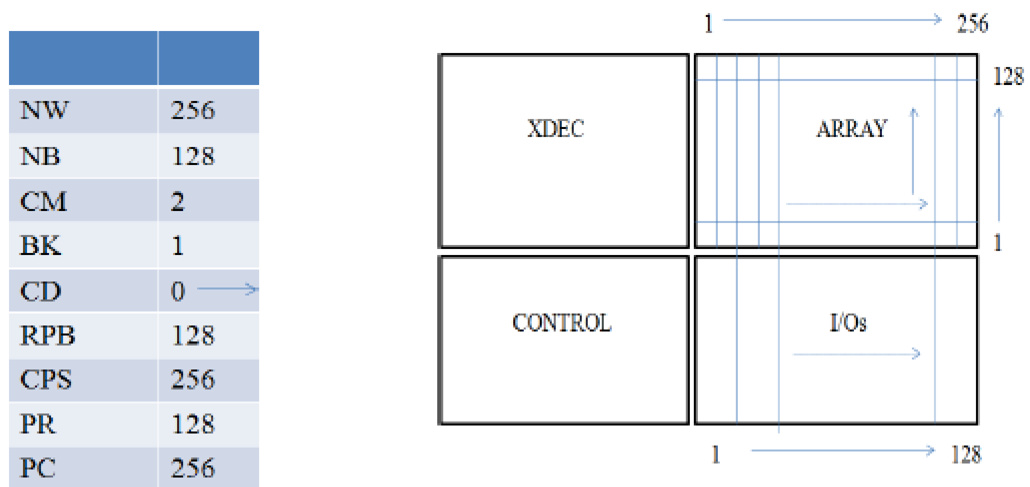
2.2.1 Simulation

I did the simulation with different Architecture i.e. split core and Bank partitioning. In simulation observed the access time of Memory. In memory architecture, access time is depend on how quickly we can access particular bit-cell. As we have seen the memory structure,from that we can say that the clock is generated in control block and it is routed to decoder area to generate the row and column address.based on row address particular bit-cell is selected.Suppose if we are performing read operation the bit-line will discharge and difference of bit-line voltage is sensed by sense amplifier and it resolved the stored data and gives the out-put. So here we can say that the access time is from clock to xdecoder to generate word-line and from selected bit-cell to output delay through sense-amplifier.

Consider the Memory instance of 256 words with each word of 128 bits.

Case:1

First consider Fig[2.4] that this memory instance is implemented with simple block memory Architecture. Fig[2.4] shows the all the specification along with BK(Bank partitioning) and CD(split core).Here BK=1 and CD=0.



NW	256
NB	128
CM	2
BK	1
CD	0
RPB	128
CPS	256
PR	128
PC	256

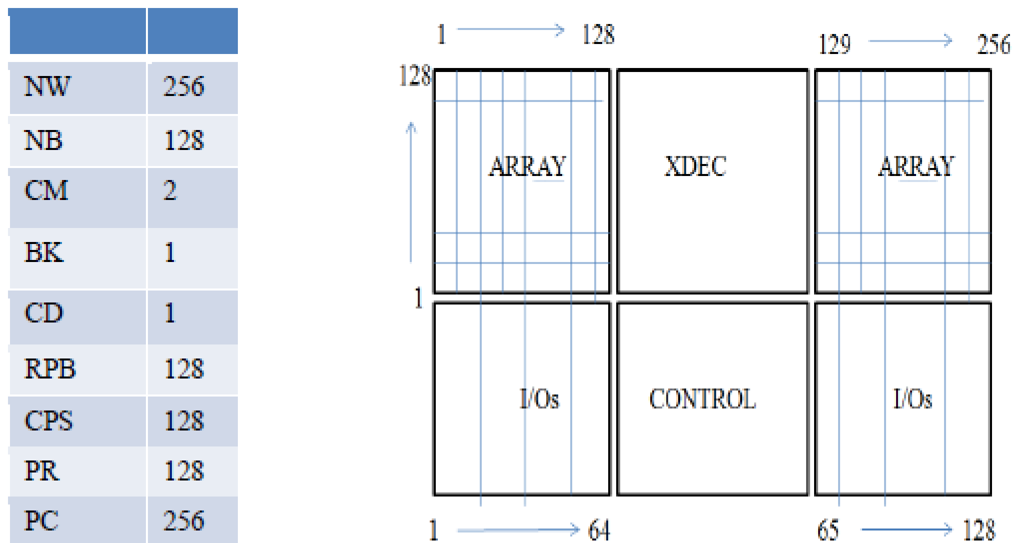
Figure 2.4: Memory Architecture with BK=1 and CD=0

Here in this architecture the Row per Bank is 128 and Column per side is 256. Other way we an say that word line is connected to the pass transistors of the 256 Bitcell and Bitline is connected to the Source and Drain of 128 transistor. So the

capacitive load of this architecture is maximum. As a result time required to access particular memory cell is also maximum.

Case:2

In this case consider that, memory instance is implemented with Split core Architecture. Fig [2.5] shows the all the specification along with BK=1 and CD=1 Architecture.



NW	256
NB	128
CM	2
BK	1
CD	1
RPB	128
CPS	128
PR	128
PC	256

Figure 2.5: Memory Architecture with BK=1 and CD=1

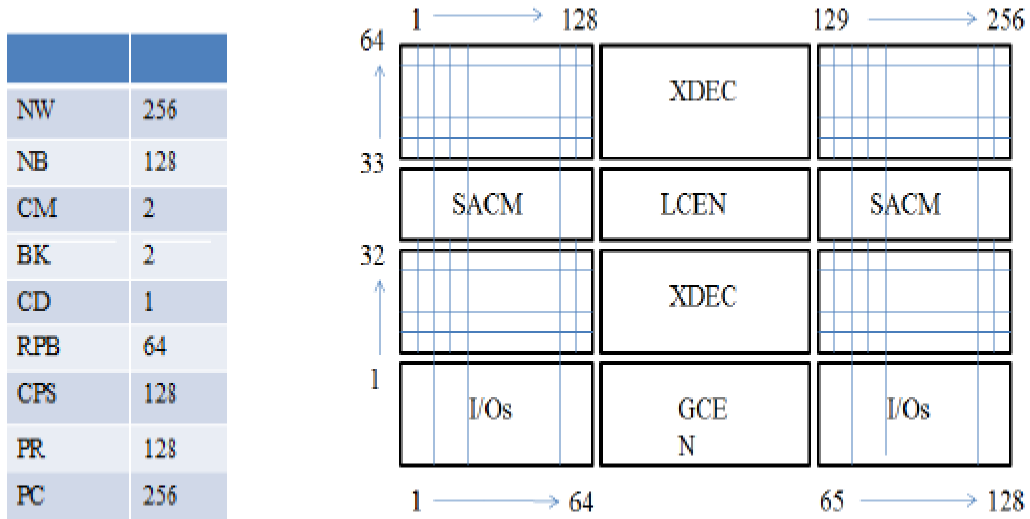
Here in this architecture the Row per Bank is 128 and Column per side is 128. Other way we an say that wordline is connected to the pass transistors of the 128 bitcell at each side and bitline is connected to the Source and Drain of 128 transistor. So the capacitive load of this architecture is reduced compared to case_1 architecture. As a result time required to access particular memory cell is also reduced compare to case_1.

Case:3

In this case consider that, memory instance is implemented with Split core Architecture. Fig [2.5] shows the all the specification along with BK=2 and CD=1 Architecture.

Here in this architecture the Row per Bank is 64 and Column per side is 128. Other way we an say that wordline is connected to the pass transistors of the 64

bitcell at each side per Bank and bitline is connected to the Source and Drain of 128 transistor. So the capacitive load of this architecture is reduced compared to case_2 architecture. As a result time required to access particular memory cell is also reduced compare to case_2.



NW	256
NB	128
CM	2
BK	2
CD	1
RPB	64
CPS	128
PR	128
PC	256

Figure 2.6: Memory Architecture with BK=2 and CD=1

Here in this architecture the Row per Bank is 64 and Column per side is 128. Other way we an say that wordline is connected to the pass transistors of the 64 bitcell at each side per Bank and bitline is connected to the Source and Drain of 128 transistor. So the capacitive load of this architecture is reduced compared to case_2 architecture. As a result time required to access particular memory cell is also reduced compare to case_2.

2.2.2 Result

Fig[2.7] shows the access time of three cases. Access time is Define as Tdelay.

CASE	NW	NB	CM	BK	CD	RPB	CPS	PR	PC	Tdelay
1	256	128	2	1	0	128	256	128	256	6.54E-10
2	256	128	2	1	1	128	128	128	256	5.33E-10
3	256	128	2	2	1	64	128	128	256	4.19E-10

Figure 2.7: Delay for three different case

2.2.3 Plot

Fig.[2.8] shows the Access time for different architecture

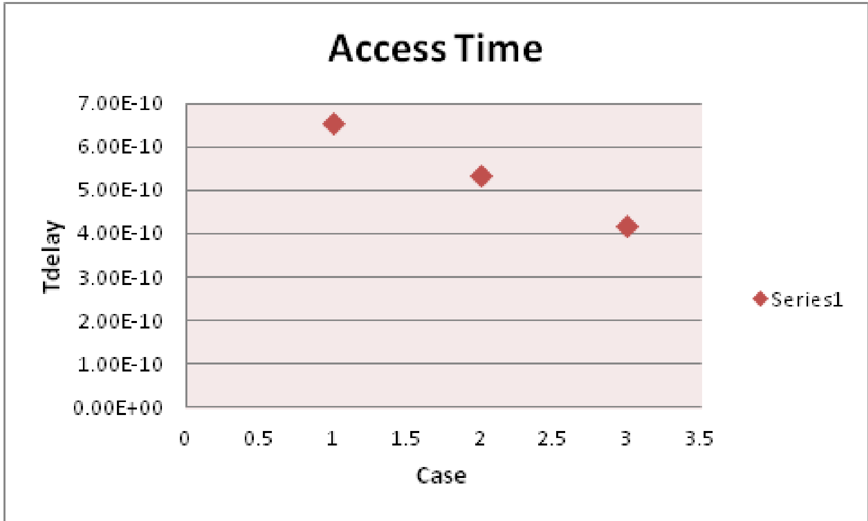


Figure 2.8: Plot for different architecture Access time

2.3 Summary

From above simulation result, we can say that by using Memory Architecture with Bank partition and split core, one can reduce the access time. Hence one can increase the speed of the Memory. As of now if increase the Number of Bank, we can further increase the speed of the memory as expense of Area. In memory bank architecture power consumption is less. The reason behind that, at a time only one memory bank is enabled and other banks remain in standby mode and also the capacitive load on the wordline and bitline decreases.

Chapter 3

Address Decoding Scheme In SRAM Memory Compiler

In this chapter, address decoder is implemented with two different decoding scheme like static decoder and pre-post address decoder. This chapter includes comparison of simulation results of both decoder.

3.1 Introduction

. As now a days processor speed is increases to GHz. So now it is become mandatory for system on chip designer to innovate a new scheme to improve the performance of memory, as a most of the part of soc is occupied by the memory. we can say that the performance of processor is decided by the performance of the memory. Processor speed is always limited by the access time which is time from clock to word-line and from selected bit-cell to output delay through sense amplifier.On each processor memory is their to store the information like instruction, data. This memory has to run at the speed of the microprocessor.Hence, it is necessary to improve the speed of the memory.

In memory architecture, access time is depend on how quickly we can access particular bit-cell. As we have seen the memory structure,from that we can say that the clock is generated in control block and it is routed to decoder area to generate the row and column address.based on row address particular bit-cell is selected.Suppose if we are performing read operation the bit-line will discharge and difference of bit-line voltage is sensed by sense amplifier and it resolved the stored data and gives the output. So here we can say that the access time is from clock to xdecoder to generate word-line and from selected bit-cell to output delay through sense-amplifier. This delay is depend on the Memory architecture implementation. Number of physical rows and physical columns.In this chapter our aim is to reduce the delay from clock

to word-line.

3.2 CMOS Decoder

The row and column decoders in Fig. [3.1] and Fig.[3.2] are necessary part in the RAM(random access memory). Speed of the memory and how much power will consume is largely depends on the architecture of decoder. Memory addresses are applied as an input to the decoder. Decoder generates 2^n outputs from n addresses. Out of all output, based on input memory addresses only one particular output goes high and select one word-line. Fig.[3.1] & [3.2] shows that decoder can be implemented using NAND gate or NOR gate. In industry NAND gate is preferred over the NOR gate because in transistor level NOR-gate PMOS comes in the series. And PMOS is already slow in compare to NMOS, as a result to reduce the large resistance of PMOS series, PMOS sizes must be increased,Hence overall area for NOR gate increases. Here in figure shows for 2- input NAND gate and NOR gate. Here, total four combination are possible. Based on input only one output goes high.

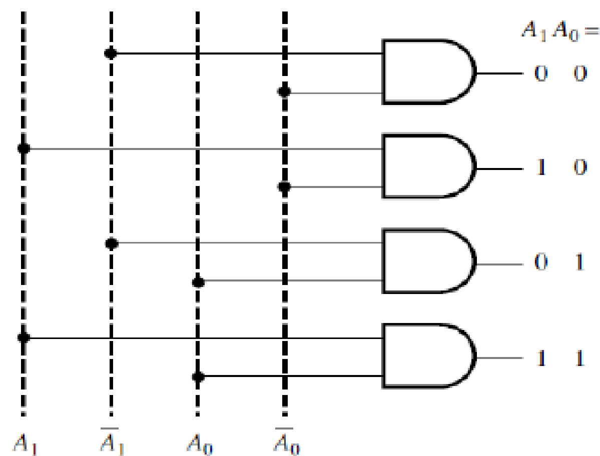


Figure 3.1: AND based decoder

In typical memory architecture, to access a particular bit-cell word-line of that bit-cell and bit-line need to be select. To select a word-line row decoder is used. This decoder can be implemented with different approach. Memory address is the input of the row decoder and it gives 2^n output line (word-line), one of word-line is activated based on input addresses.Column decoder size is depend on column multiplexing value.Like if CM=16 then it require 4x16 column decoder to select a bit-line.

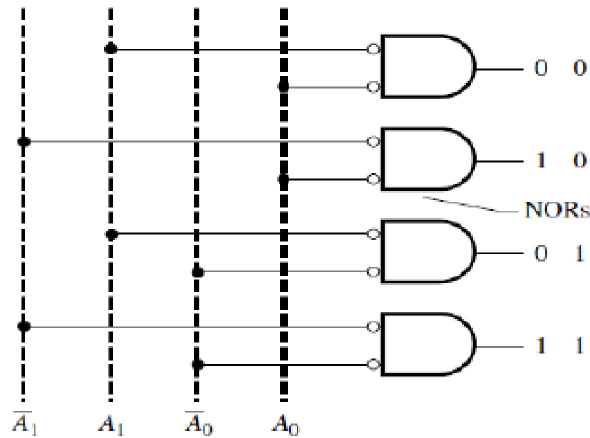


Figure 3.2: NOR based Decoder

In industry NAND gate is preferred over the NOR gate, because two implement NOR gate PMOS comes into the series. As a result two maintain equal rise and fall time, area occupied by the PMOS increases. As a result overall area increases.

To implement n-bit decoder, 2^n NAND or NOR gate requires.Each gate with n-input and 2^n inverter. Consider 1Kb SRAM of 32 words with each word of 32 bit.If CM=1 then it has 32 word-line.Hence, it requires 5x32 decoder,which contain 32 NAND gate with each gate with 5-inputs. In the next section we will see that what is the effect on delays of increasing fan-in for the NAND gate.

3.2.1 Simulation for NAND Gate to compare Delay from Input to Output

In this section to calculate delay of NAND gate for different fan-in, transistors are modeled as an ideal switch is in series with resistance as shown in Fig.[3.3]. The value of series resistance is depend on the size of the transistor, voltage supply.Based on operating region of transistor, resistance will differ.In the linear operating region, it acts as idea resistor while in saturation region ,it will act as large(infinite) resistance.Here two input NAND gate is shown as in Fig.[3.3] with it's all transistor is modeled as RC-switch.It also shows the output load capacitance C_L with internal node (C_int) capacitance.C_int (internal load capacitance) is significance of overlap capacitance, gate to drain and gate to source capacitance.All this capacitance are voltage dependent.

Consider here that output transition take place from low-to-high. Such a condition take place in three different combination.Either one input goes low or both, output

load capacitance will charge to supply voltage.We can say that the time to take output from low to high is depends on state of PMOS. If both PMOS is ON condition, then resistance comes into parallel.Hence, effective resistance reduces,as a result Low-to High delay reduces.The propagation delay Tplh is given by 0.69x($R_p/2$)xC_L.But in any design we have to consider worst case. And worst case becomes only when one PMOS is ON means only one input is Low. In this case propagation delay Tplh is given by 0.69x(R_p)xC_L. Same case if we consider the output transition from High to low, then output load capacitance is discharged through ground only when both NMOS is ON. In this case propagation delay Tphl is given by 0.69x($2R_n$)xC_L. For the Tphl path we can say that, if we goes on adding NMOS(means increase the number of input) then large resistance of NMOS comes into series and delay increases significantly for higher fan-in.

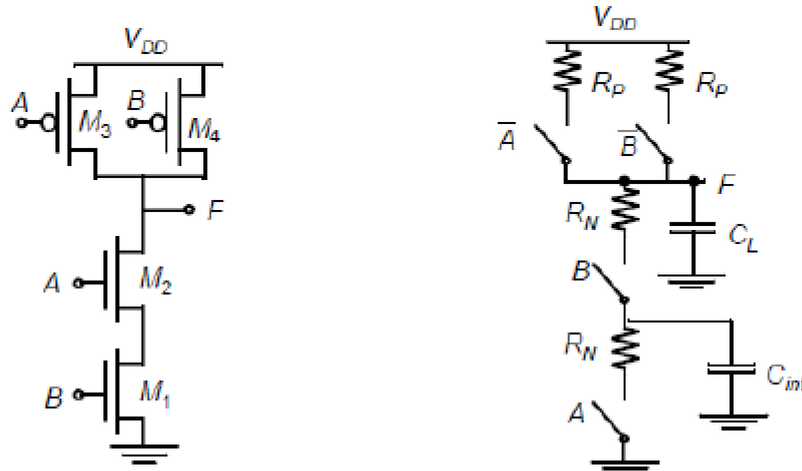


Figure 3.3: NAND2 gate with transistor and its RC equivalent

In the above analysis of High-to-Low and Low-to-high propagation delay, we have considered the first order approximation only.But if we consider the higher order approximation, then internal node capacitance comes into the effect.As Fan-in increases, internal node capacitance effect on to the delay become significant.The proof of this is analysis is given in next section.

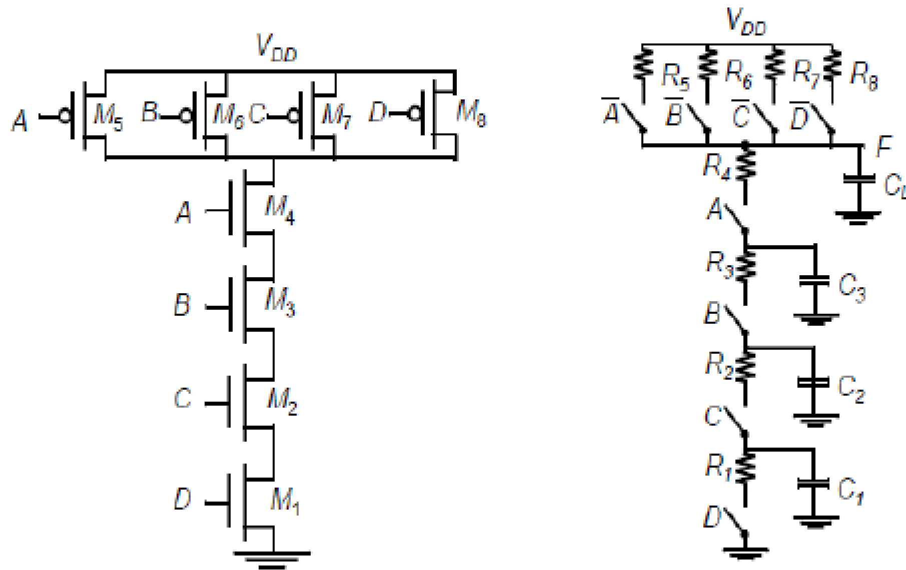


Figure 3.4: NAND4 gate with transistor and its RC equivalent

Now here increase the number of input from 2 to 4.Fig.[3.4] shows a 4-input NAND gate with its equivalent RC switch model. It also shows the internal node capacitance.internal load capacitance is significance of overlap capacitance(Cov), gate to drain(Cgd) , gate to source(Cgs) and junction capacitance.All this capacitance are voltage dependent.The value of capacitance is depend on the operating condition of transistor.Now if suppose i want to find the propagation delay of this complex circuit it difficult compared to first order analysis.It's analysis can be done if we considered as a distributed RC network.Here consider the High-to -low propagation delay only.When all the input to NAND4 goes high output node discharges through ground. Here the initial conditions of the the internal node volatge is very important.For an NMOS to ON internal node must be precharged to VDD-VTN) before input goes high.

Here based on Elmore propagation delay model high-to-low propagation is approximated as below.

$$\text{Tphl}= 0.69(\text{R}_1 C_1+(R_1+R_2)C_2 +(R_1+R_2+R_3)C_3+(R_1+R_2+R_3+R_4)C_L$$

Assuming that all NMOS device size is equal then,

$$\text{Tphl}= 0.69\text{R}_N(C_1+2C_2+3\text{C}_3 +4C_L)$$

3.2.2 Result

Fig.[3.5] shows the Delay from input to output for different fan-in.

	FAN-IN	DELAY FROM INPUT TO OUTPUT
NAND2	2	2.35E-11
NAND3	3	2.85E-11
NAND4	4	4.08E-11
NAND5	5	6.43E-11
NAND6	6	7.45E-11

Figure 3.5: Delay for different Fan-in of NAND gate

3.2.3 Plot

Fig.[3.6] shows the Plot of Delay at different Fan-in.

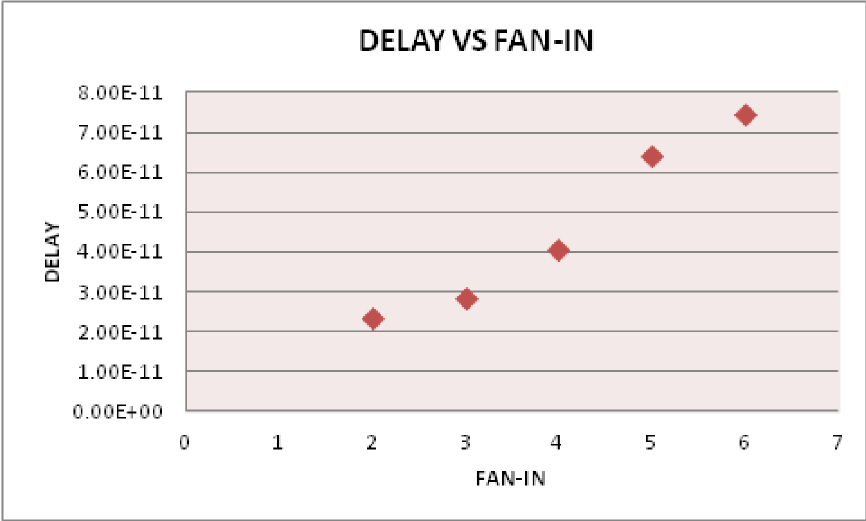


Figure 3.6: NAND Gate Delay Vs Fan-in

3.2.4 Simulation Waveform

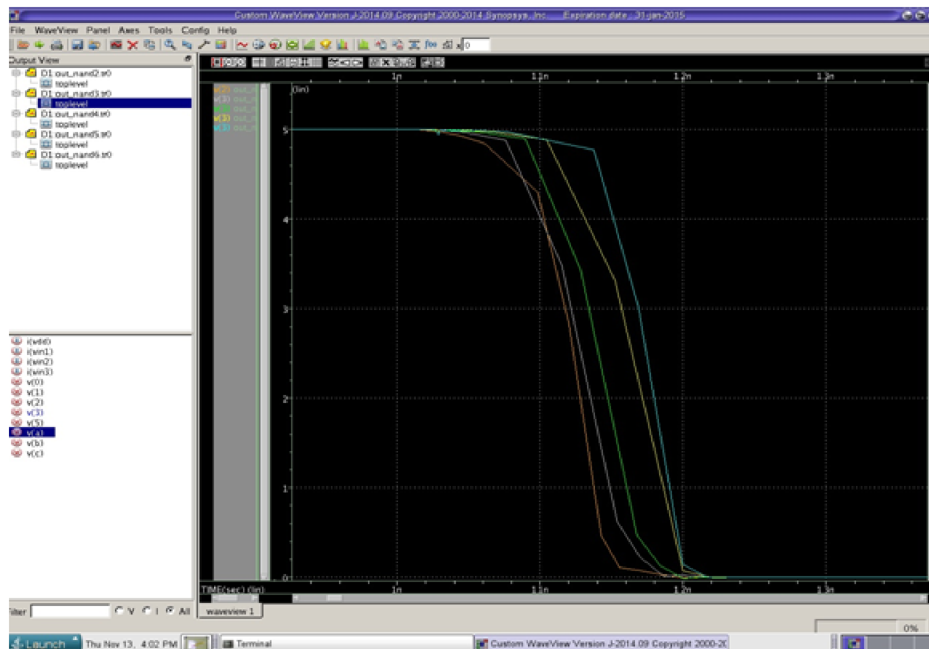


Figure 3.7: Delay for different Fan-in

3.3 Static address decoder

As we know the address decoder is the main periphery part of the memory.Address decoder decide the speed of the memory.Hence, if some how we can reduce the delay of decoder, we can increase the speed of the memory.

Consider that we want to implement 4x16 input decoder. If we implement this 4 to 16 decoder by traditional NAND4 based decoder than, we required 4 input 16 NAND gates and 16 inverters. Block diafram is shown in fig.[3.8]

Simulation waveform for the NAND4 based decoder delay at different Temperature is shown in Fig.[3.9]

Here i have implemented using 4 input NAND gate, as a result the delay for NAND4 is significant compared to NAND3 and NAND2 gates. In the next section i have implemented this decoder with pre-post address decoding scheme and compare the simulation results.

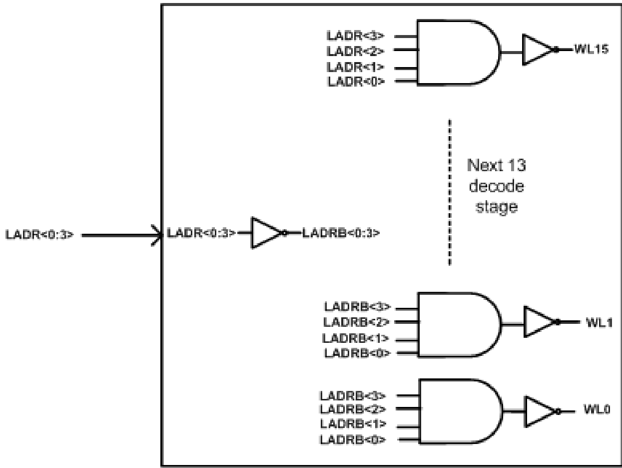


Figure 3.8: NAND4 based decoder

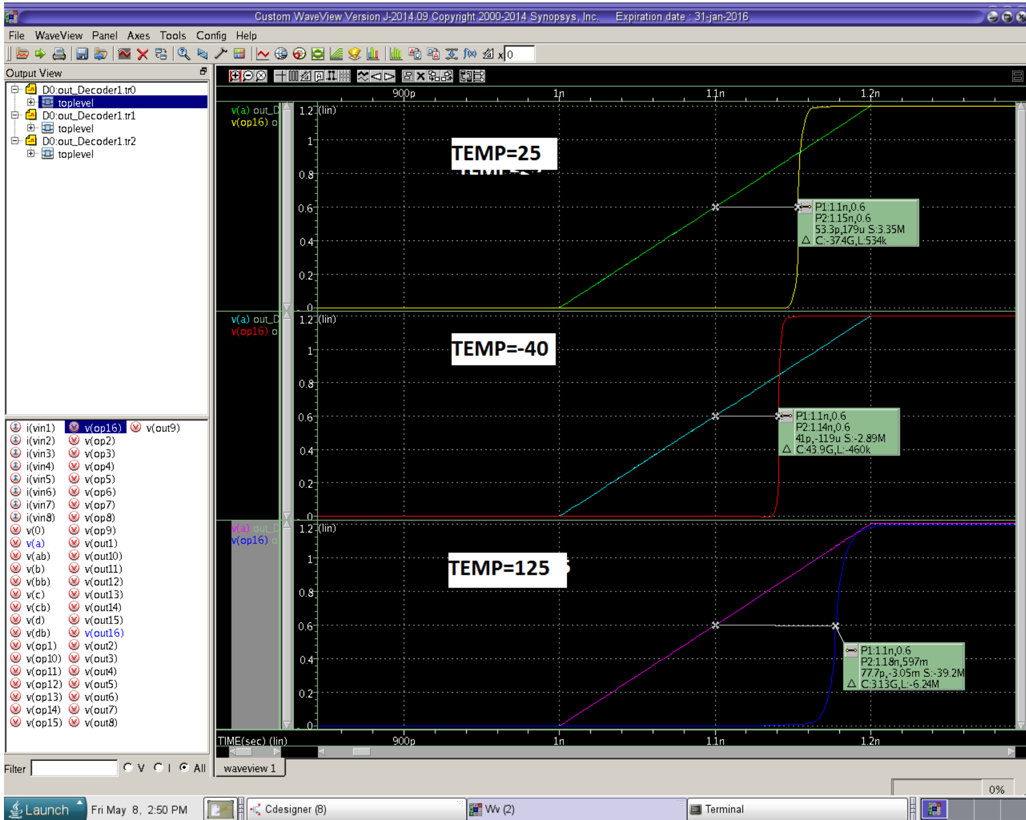


Figure 3.9: NAND4 based decoder delay at different temperature

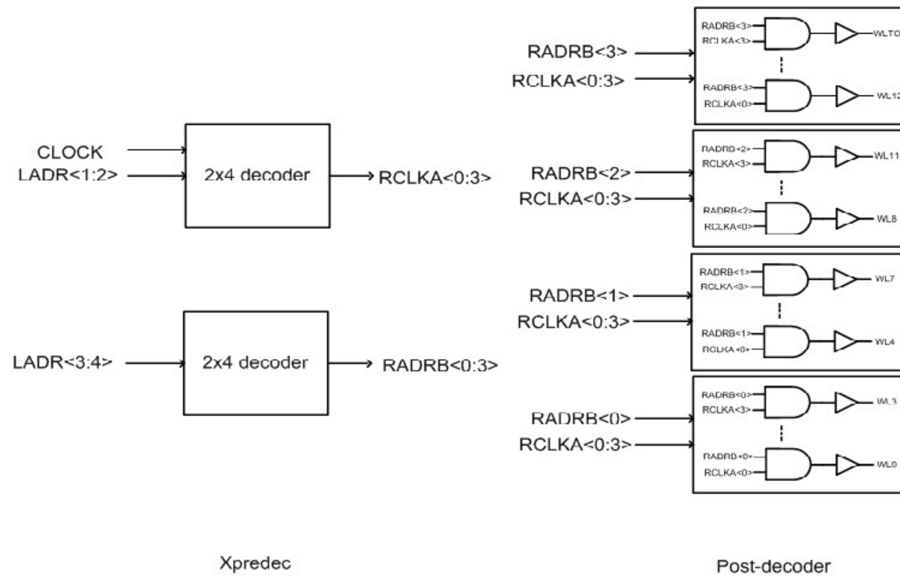


Figure 3.10: Block Diagram of Pre-post address decoding Scheme

3.4 Pre-post Address Decoding Scheme

As we have seen from the simulation result that the delay for NAND gate for more than three inputs delays increases significantly. Hence, instead of using n-input gates,we can use cascading of gates. As shown in fig.[3.10] first stage generates intermediate signals. Second stage reuse this all possible intermediate signal and generate final outputs.

Now implement this 4 to 16 decoder by using concept of predecode and postdecode stages. Here see the below block diagram for the implementation of 4 to 16 decoder.

Block diagram shows the alternate way to implement 4 input NAND gate. Four input are divided in a group of two input and each group is implemented with 2 to 4 decoder in the first Predecode Stage. In predecoded stage 2 input NAND gate is used to implement 2 to 4 decoder. Here output of predecoded stages called intermediate output. Now, this intermediate signal are reused to generate final output in postdecode stage as shown in Block diagram. In postdecode also to generate final output 2 input NAND gate is used.

We can say that in new scheme, we used 2 input NAND gates only compared to 4 input NAND gates as Traditional Decoder. From previous simulation result we observed that as fan-in increase from 2 to 4, the delay increase from 53.3ps to 39.6ps. Hence, we can say that by using this new address decoding scheme we can improve the delay from clock to wordline. Hence, the speed of the memory increases.

3.4.1 Important Definition

Stack Size:

In NAND gate transistor level circuit, each NMOS is in series, and in NOR gate each PMOS is in series.Hence,in the worst case delay is decided by this series transistor's resistance. Stack size is the number of NMOS in series in NAND gate or the number of PMOS in case of NOR gate. If stack size is high delay becomes high. hence,it is necessary to reduce the stack size.

Minimum Prerecorded Line:

First stage of decoder is implemented in control block.Its output is intermediate signals.This all signal are routed through decoder area. As a result, to avoid routing issues intermediate signals should be minimum.

Simulation waveform for New address decoding scheme.Fig.[3.11] shows delay for different temperature.

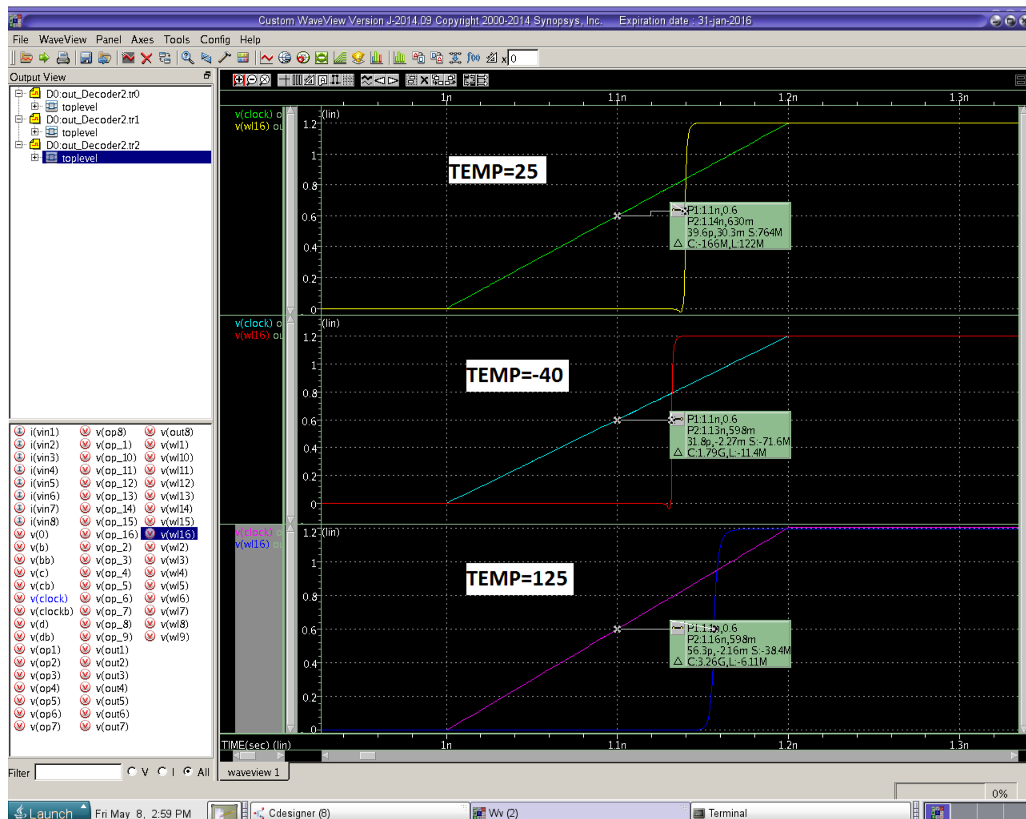


Figure 3.11: Pre-post address decoder delay at different temperature

3.4.2 Result

Fig.[3.12]compares the Delay for both decoding scheme at different temperature.

	Clock to word-line delay	
Temperature	Static decoder	Pre-post decoder
-40	4.11E-11	3.18E-11
25	5.33E-11	3.96E-11
125	7.78E-11	5.63E-11

Figure 3.12: Pre-post scheme based address decoder delay at different temperature

3.4.3 Comparision

Fig.[3.13] shows waveform of which compare both decoding scheme.

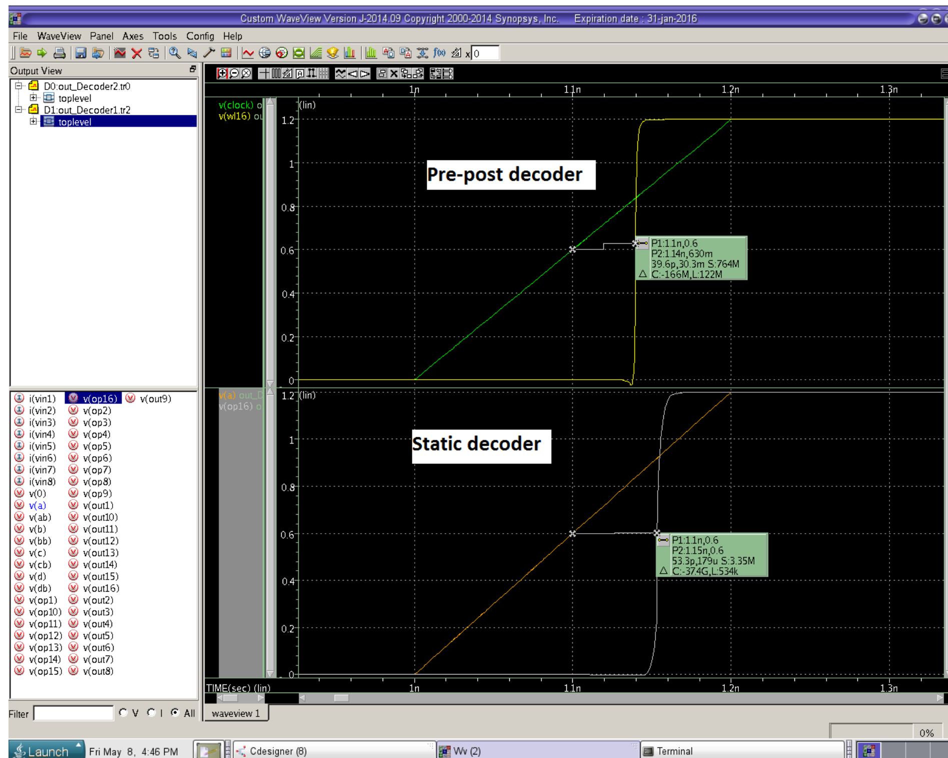


Figure 3.13: waveform which shows the delay from CLK to Wordline at 25 degree temperature

From simulation result we can say that in new address decoding scheme delay from clock to Word-line decreases from 53.3ps to 39.6ps. The advantage in pre-post address decoding scheme is that, first satge generates intermediate signal, this intermediate signals are reused in second stage.Hence, the number of input for NAND gate is reduces.

3.5 Summary

From all above simulation results, it is clear that by using pre-post address decoding scheme one can reduce the delay more than 25% from clock to wordline. Area of decoder implementation can be found by number of intermediate signals.More the number of intermediate signal,higher the area overhead in decoder area.Higher stack size in the second stage decoder results in increase in area. Also if it increases beyond three, the delay increases comparatively. Here we can reduce the stack size of second stage decoder(post-decoder) by make use of first decoder(pre-decoder). If temperature increases the delay will increase because of mobility degradation.

Chapter 4

Write Assist technique in SRAM

In the current tend of scaling, supply voltage is lowered. Hence, it is very difficult for SRAMs to function at low voltage. As supply voltage is reduces the current driving capability of transistor decrease.Hence to ensure write-ability, in this chapter some write assist techniques are described.

4.1 Introduction

As supply voltage is reduces the current driving capability of transistor decrease. Hence, it will take more time to either charging or discharging process. As a result delay will increase. If threshold voltage of bit-cell transistor is lower, then effect of noise is very high.At lower supply voltage static noise margin reduces.So, there might be chances of flipping of bit-cell due to small noise.From all this we can say that due to voltage scaling SNM degrades and time to access particular bit-cell for read and write operation increase. Many techniques have been proposed to overcome this difficulty. Here i have described some of the write assist technique in SRAM. Fig.[4.1] shows the simulation result of static noise margin at Typical process corner and different voltage supply.From that we can say that SNM degrades as we lowered the supply voltage.

Process	Voltage	Temperature	Mean	Sigma	Mean-5.2*Sigma
TT	1.08	25	0.1622	0.019	0.065428
TT	0.99	25	0.1594	0.018	0.063564
TT	0.9	25	0.1514	0.018	0.056968
TT	0.81	25	0.139	0.018	0.044464
TT	0.72	25	0.1214	0.018	0.028944

Figure 4.1: Variation in SNM with supply voltage reduction

In the 6-T SRAM as shown in fig.[4.2] the speed of the read and write operation is mainly depends on pull-up,pull-down and pass transistor. pass transistors are connected to the bit-line. When word-line is activated this transistor becomes ON. In 6-T SRAM as a pass transistor we are using NMOS. As we know NMOS is a weak '1' pass transistor.As a result when bit-line(BL) goes high and word-line(WL) is activated then internal node(XT) charges maximum (VDD-Vth).Which result in a decrease in a driving current.

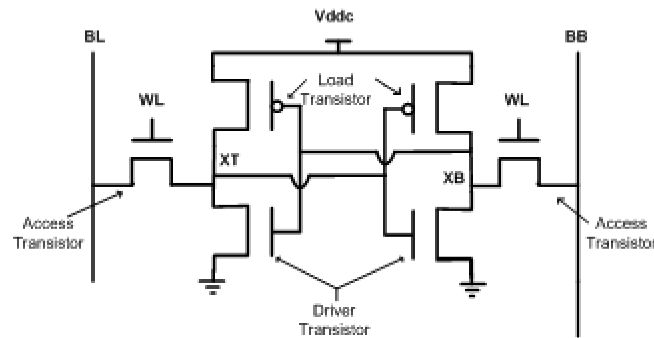


Figure 4.2: 6T SRAM bit-cell

In SRAM bit-cell contain, cross coupled inverter which is connected to bitlines through access transistor.During the read and write operation access transistor play a major role, because current from bit-line to internal node driving through access transistor.If some how we can increase the current through access transistor we can increase the speed of the read and write operation.In the read operation bitlines are precharged to high voltage. then based on data stored, '0' data stored node is discharge the bit-line slowly. The discharge rate is depends on the current drive of the access transistor. So here to increase the driving capability of access transistor gate to source voltage of access transistor should be increased. Next section describe different proposed techniques.

4.2 Proposed write assist techniques

Here is the list of write assist techniques

a. Reducing the Vddc voltage

b. Increase the Vss volatge

c. Boosting of Word-line volatge

d. Negative bit-line capacitive coupling

4.2.1 Reducing the Vddc voltage

Write time in the memory cell is depend on the strength of the access transistor and pull-up transistor. For ensure write operation the size of access transistor is large compare to pull-up transistor to make pull-up transistor weak.Now if by any other means we can more weaken the pull-up compared to access transistor, it results in easy write operation. Hence here in this technique, by lowering the Core voltage of array weaken the pull-up transistor.Fig.[4.4] shows the timing waveform of different signal. This scheme required a second voltage. Using multiplexer we can apply both voltage when required to particular interested column.

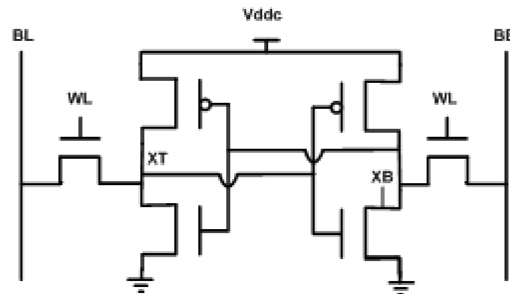


Figure 4.3: 6-T SRAM Bitcell

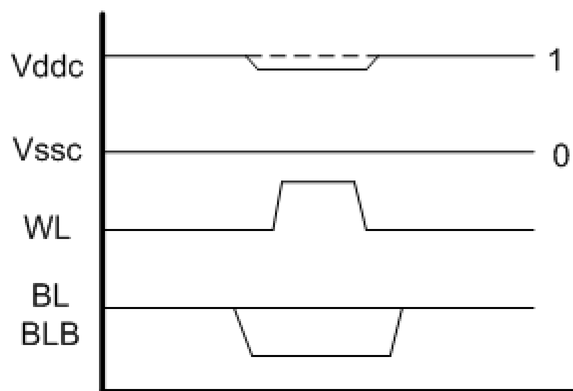


Figure 4.4: Timing relationships using the Vddc lowering WA scheme

There are also other techniques to reduce the core supply voltage. One of that is charge sharing concept at particular selected column.

I took a bit-cell and implemented this technique. Fig.[4.5] shows the simulation result for the same.

Process	Voltage	Temp	Mean	Sigma	Mean + 5*Sigma
TT	1.08	25	5.93E-11	2.59E-12	7.23E-11
TT	0.99	25	6.49E-11	3.43E-12	8.20E-11
TT	0.9	25	7.30E-11	4.53E-12	9.57E-11
TT	0.81	25	8.41E-11	5.20E-12	1.10E-10
TT	0.72	25	9.84E-11	5.80E-12	1.27E-10

Figure 4.5: Simulation result for Vddc lowering WA scheme

The main difficulty in this technique is to check that if any other unselected bit-cell should not flip, otherwise stored data will lost.

4.2.2 Increase the Vss core voltage

In this technique also we are weaken the pull-up device by applying higher voltage above the Vss at the gate of pull-up device.Hence, effective gate to source voltage reduces, and it becomes more weak to drive current compared to access transistor. Fig.[4.6] shows the waveform of relationships using the increase the Vss core WA scheme.

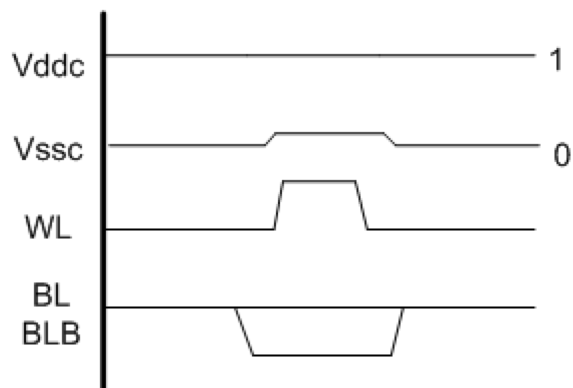


Figure 4.6: Timing relationship of Vsscore rising WA scheme

This extra ground can be routed separately or can be generated internally. Here as compare to normal scheme, the gain in this scheme is very small. In this Write Assist scheme we are weakening the pull-up which is PMOS,which is already weak. And now if we make more weak,then overall it will not significantly weaken it.Fig.[4.7] shows the simulation result for this technique.

Process	Voltage	Temp	Mean	Sigma	Mean + 5*Sigma
	1.08	25	7.90E-11	2.04E-14	7.911E-11
TT	0.99	25	8.08E-11	4.99E-14	8.109E-11
TT	0.9	25	8.32E-11	2.46E-13	8.447E-11
TT	0.81	25	8.72E-11	2.22E-12	9.873E-11
TT	0.72	25	1.01E-10	6.45E-12	1.344E-10

Figure 4.7: Simulation result for Vsscore rising WA scheme

4.2.3 Word-line boosting

Another technique which helps to write into the bit-cell is increased voltage of word-line above the supply.here in this concept we want to increase the effective gate to source voltage of access transistor.The boost in word-line voltage increases the gate to source voltage of the access transistor[11]. Hence, increase the current driving capability. As a result, bit-cell flip-time reduces significantly[11]. The supply boost voltage is needed in this scheme externally,Which is routed in array part or also it can can be generated internally by a capacitive coupling[11]. Fig.[4.8] shows the waveform of timing using the word-line boosting WA scheme.

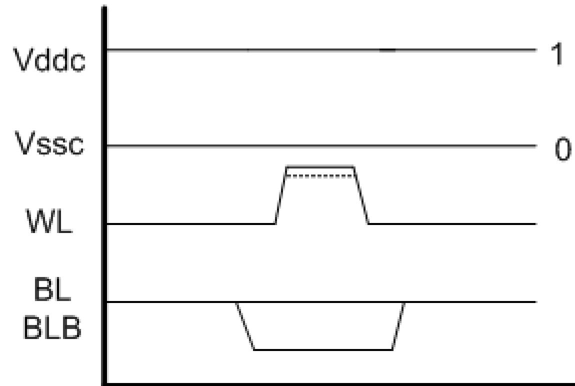


Figure 4.8: Timing relationship of Word-line boosting WA scheme

Above two techniques work on a column, while this technique is worked on row based. Fig.[4.9] shows simulation result for this technique.

Process	Voltage	Temp	Mean	Sigma	Mean+5*Sigma
TT	1.08	25	5.97E-11	4.52E-13	6.20E-11
TT	0.99	25	6.94E-11	8.15E-13	7.35E-11
TT	0.9	25	6.10E-11	4.77E-12	8.48E-11
TT	0.81	25	7.10E-11	7.14E-12	1.07E-10
TT	0.72	25	9.11E-11	6.89E-12	1.26E-10

Figure 4.9: Simulation result for word-line boosting WA scheme

4.2.4 Negative bit-line capacitive coupling

Here in this technique our aim is to increase the gate to source voltage for the access transistor.In above technique we increase the gate to source voltage by increasing the gate voltage of access transistor,while in this technique we decreasing the bit-line voltage up-to negative.Hence,access transistor becomes stronger compared to pull-up transistor. As a result, it can flip the cell easily[11]. The negative bit-line voltage can be generated using a capacitive coupling technique[11]. Fig.[4.10] shows the waveform of timing using Negative bit-line WA scheme[11].

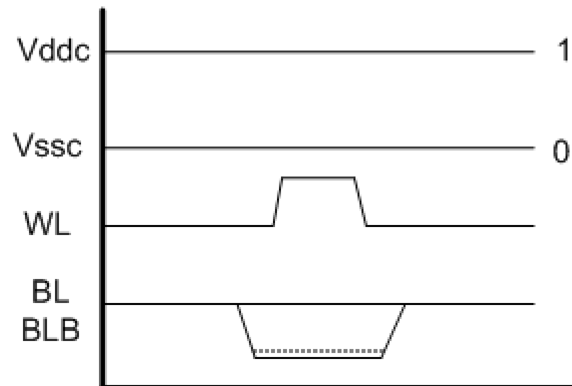


Figure 4.10: Timing relationship of Negative bit-line WA scheme

In this write assist technique, negative voltage is pass through column. Here we have to make sure that,other unselected cells should not alter the stored data. Here value of negative bit-line is limited by the threshold voltage of access transistor,hence for those unselected cells Dynamic read noise margin should no affected.Fig.[4.11] shows the simulation result for this technique.

Process	Voltage	Temp	Mean	Sigma	Mean+5*Sigma
TT	1.08	25	5.36E-11	1.54E-12	6.13E-11
TT	0.99	25	5.62E-11	3.10E-12	7.17E-11
TT	0.9	25	6.72E-11	4.16E-12	8.80E-11
TT	0.81	25	7.91E-11	4.65E-12	1.02E-10
TT	0.72	25	9.43E-11	5.33E-12	1.21E-10

Figure 4.11: Simulation result for negative bit-line WA scheme

4.3 Impact on half selected cell due to write assist techniques

Now a days to reduce the bit-line capacitance column multiplexing is used memory architecture. So here Number of bit-lines are shared between one output column. In this architecture,some time such a condition take place, like word-line is inserted and bit-line goes low to perform write operation in particular bit-cell. But for some bit-cell word-line is inserted but bit-lines are not pulled to low. Here for all these bit-cells read operation is performed.All these cells called half-selected. Here designer

have to make sure that, during read operation the data should be retain.For all the write assist technique dynamic read noise margin is evaluated first. Dynamic read noise margin is defined as minimum voltage difference during read operation between the internal nodes

Here in all above techniques, word-line boosting and negative bit-line capacitive coupling are rows base, means word-line is inserted, which impacts the dynamic read noise margin of those half selected cells. Same condition occurs in case of negative bit-line also. First two techniques decrease the Vdd core of bit-cell and increase the Vss core of bit-cell.If during Write operation whole array voltage is scaled down then it will impact the half -selected cells.However, if the supply voltage is decreased for only selected cell then it would not impact the DRNM.

4.4 Summary

In this chapter, in order to improve the write ability of SRAM at lower supply voltage while maintaining the SNM.I have described write assist techniques which increase the strength of pass transistor. Hence,gate to source voltage of access transistor increases, which increases the driving current from access transistor.Simulation results have shown the proposed approach which reduce the flip-time of bit-cell.

Chapter 5

Conclusion And Future scope

Here, We can conclude that by using Memory Architecture with Bank partition and split core, one can reduce the access time. Hence one can increase the speed of the Memory. As of now if increase the Number of Bank, we can further increase the speed of the memory as expense of Area. In memory bank architecture power consumption is less. The reason behind that, at a time only one memory bank is enabled and other banks remain in standby mode and also the capacitive load on the word-line and bit-line decreases. And also by using proper address decoding scheme one can reduce the delay more than 25% from clock to word-line. Area can be determined by number of predecoder lines in Xdecoder area. More the number of predecoder lines, larger is area overhead in local decoder. Higher stack size in the post decoder results in more area and also if it increases beyond three, the delay increases comparatively.

In order to improve the write ability of SRAM at lower supply voltage while maintaining the SNM, by using write assist techniques Which increases the driving current from access transistor.Hence, the flip-time of bit-cell reduces and ensure the write operation at lower supply voltage.

Bibliography

[1] Sung-Mo-Kang and Yusuf Leblebici. "CMOS Digital Integrated Circuits Analysis and Design

[2] Verma K; Jaiswal S.K. ; Jain D.; Maurya V."Design and analysis of 1-kb 6T SRAM Using differential Architecture" , Computational Intelligence and communication Networks (CCIN),2012 Fourth international conference.

[3] Kumkum Verma."Design of a high performance and low power 1Kb 6T SRAM using bank partitioning method" 2011 international conference on Multimedia signal processing and communication technologies,12/2011.

[4] Sanjeev Kumar Jain"A novel Circuit to Optimize Access Time and Decoding Schemes in Memories",2010 23rd International conference on VLSI Design, 01/2010.

[5] Shobha Singh, Shamsi Azmi , Nutan Agrawal, Penaka Phani and Ansuman Rout "Architecture and Design of a High Performance SRAM for SOC Design Central R&D, STMicroelectronics, Noida 201301 INDIA.

[6] http://scholar.lib.vt.edu/theses/available/etd-72198162528/unrestricted/body.PDF

[7] Vikas Chandra, Cezary Pietrzyk,Robert Aitken, "On the Efficacy of Write-Assist Techniques in Low Voltage Nanoscale SRAMs",ARM R&D,San Jose, CA.

[8] Masaaki Iijima, Kayoko Seto, Masahiro Num and Akira Tada, Takashi Ipposhi,"Low Power SRAM with Boost Driver Generating Pulsed Word Line Voltage for Sub-1V Operation" JOURNAL OF COMPUTERS, VOL. 3, NO. 5, MAY 2008

[9] W. Dehaene et al,"Embedded SRAM design in deep deep submicron technologies",European Solid-State Circuits Conference (ESSCIRC),pp. 384-391, 2007.