Improvisation & deployment of methodologies to have productivity gain in characterization and design verification flow of static memories

Major Project Report

Submitted in partial fulfillment of the requirements

for the degree of

Master of Technology

in

Electronics & Communication Engineering

(VLSI Design)

By

Jain Anmol Sunilkumar

(15MECV10)

Electronics & Communication Department
Institute of Technology
NIRMA University
Ahmedabad-382 481

May 2017

Improvisation & deployment of methodologies to have productivity gain in characterization and design verification flow of static memories

Major Project Report

Submitted in partial fulfillment of the requirements for the degree of

Master of Technology
in
Electronics & Communication Engineering
(VLSI Design)

By

Jain Anmol Sunilkumar

(15MECV10)

Under the guidance of

External Project Guide:	Internal Project Guide:
Mr. Sachin Gulyani	**Dr Usha Mehta**
Senior Manager	Institute of Technology
ST Microelectronics India Limited	Nirma University
Greater Noida	Ahmedabad

Electronics & Communication Engineering Department
Institute of Technology
NIRMA University
Ahmedabad-382 481

May 2017

Certificate

This is to certify that the Major Project entitled **"Improvisation & deployment of methodologies to have productivity gain in characterization and design verification flow of static memories "** submitted by **Jain Anmol Sunilkumar (15MECV10)**, 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 Guide

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

Dr. Dilip Kothari
Head, EC Dept.

Dr. Alka Mahajan
Director, IT-NU

Date:

Place: Ahmedabad

Certificate

This is to certify that the Project entitled "**Improvisation & deployment of methodologies to have productivity gain in characterization and design verification flow of static memories**" submitted by **Jain Anmol Sunilkumar (15MECV10)**, towards the submission of the Project for 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.

Mr. Sachin Gulyani
Senior Manager,
ST Microelectronics India Limited,
Greater Noida

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.

-Jain Anmol Sunilkumar

Acknowledgement

Completion of any project would bring a sense of satisfaction.However,it would not have been possible without the kind support and help of many individuals and organization. It gives us immense pleasure to extend our sincere thanks to all of them.

I would like to express my gratitude and sincere thanks to my mentor **Mr. Sachin Gulyani** at STMicroelectronics Private Limited, Greater Noida for his valuable guidance throughout this period. He has given me valuable advices and support for my project work which I am very lucky to benefit from.

I would like to thanks to **Kshitij Verma** and **Yagnesh Vaderiya** at STMicroelectronics Private Limited, Greater Noida for supporting me in my project work

I would like to thank my Program Coordinator, **Dr. N. M. Devashrayee**, Professor, EC (VLSI Design), Institute of Technology, Nirma University, Ahmedabad for giving valuable support and motivation throughout the academic period.

I would also thank to my Project Guide, **Dr. Usha Mehta**, Professor, VLSI Design, Institute of Technology, Nirma University, Ahmedabad for being a source of inspiration, giving valuable support and timely guidance for the project work.
I wish to thank my classmates and colleagues for their delightful company which kept me in good humor throughout the journey.

Last, but not the least, no words are enough to acknowledge constant support and sacrifices of my family members because of whom I am able to complete the degree program successfully.

- Anmol Jain
15MECV10

Abstract

Memories are leading semiconductor component which occupy around 70-80% area on a typical SoC store digital information in massive quantity, hence are essential subsystem in modern integrated circuits. The ever-increasing demand for low priced memories with low power consumption, high performance, high density and small package size has compelled the fabrication technology and memory development towards more compact design rules and consequently towards higher data storage densities. Advanced technology nodes enable the designers to integrate more functionality but this integration comes at a cost. Hence nanometer process designers are facing many challenges, which may cause negative impact on product yield and time-to-market constraint. The main motive of my project is the improvement in methodology to mitigate the challenges faced in post layout characterization and design verification flow in advanced technology nodes for a full custom SRAM memory design.

Characterization of memory means to get information about its behaviour in terms of different timing, power, leakage, capacitances and marginalities. This helps in evaluating the performance of memory and improves upon the design. When a set of specified inputs is applied to the memory it includes running simulation on post layout netlist and then doing measurement from the simulated values. We can bucket post layout challenges into two main categories: Extraction related and Simulation related. Characterization to be done with good accuracy and reasonable run-time. Characterization is done with help of simulations at circuit level. Two types of simulators available: True SPICE - Golden (equation based) and Fast SPICE (Optimization based upon algorithms). With True SPICE simulator, high accuracy is obtained - equation based. With Fast SPICE, significant reduction in run-time - accuracy compromise. Also fast SPICE have some optimization options with which trade-off can be done between accuracy and run-time. Enabling relevant algorithm in fast SPICE for design is important in order to have reasonable accuracy and runtime. Also improvement of existing characterization methodology to make it more efficient.

Contents

Certificate iii

Certificate iv

Declaration v

Acknowledgements vi

Abstract vii

List of Figures x

Abbreviations xi

1 Introduction 1
- 1.1 Introduction 1
- 1.2 Motivation 3
- 1.3 Objective of the Project 3

2 Brief Literature Review 5
- 2.1 Introduction to memories 5
  - 2.1.1 MOS Memories: Introduction 5
  - 2.1.2 MOS Memories: Read-Write Memories 6
  - 2.1.3 Read-only Memories 6
  - 2.1.4 Memory Organization 8
  - 2.1.5 Timing Diagrams 9
  - 2.1.6 Critical Timing Path 9
  - 2.1.7 Functional SRAM Chip Model 10
- 2.2 Core of the memory 12
  - 2.2.1 SRAM Cell description 12
  - 2.2.2 Voltage Transfer Characteristics 12
  - 2.2.3 6T SRAM Cell 13
  - 2.2.4 SRAM Array Operation 13
  - 2.2.5 Read Operation 15

2.2.6 Write Operation 16
2.3 Power consumption in SRAM 17
2.4 Analysis of SRAM architecture 19
2.4.1 Basic Architecture 19
2.4.2 Split core architecture 20
2.4.3 Page Type Architecture 20
2.4.4 Bank Architecture 20

3 Benchmarking and Options Tuning 22
3.1 Introduction 22
3.1.1 Run time optimization and accuracy improvement for memory characterization: 23
3.1.2 Simulator option tuning and benchmarking 23
3.1.3 Need of option tuning and benchmarking 24
3.1.4 Benchmarking of memory characterized result and option tuning of simulator 24

4 Full CUT vs Critical Path Verification 26
4.1 Critical Path Modeling 26
4.2 Need of FC vs CP verification 26
4.3 Flow of verification 27

5 AVM Methodology for IR Drop Analysis 28
5.1 Introduction to IR drop analysis 28
5.2 AVM Methodology for analysis 29
5.3 Deployment of AVM Methodology in existing characterization flow . . 29

6 Implementation and Results 31
6.1 Benchmarking & Option Tuning 31
6.2 AVM Methodology Implementation 32

7 Conclusion 33

8 Future Work 34

References 35

List of Figures

2.1 Critical Timing Path for SRAM 10
2.2 Functional SRAM chip model 11
2.3 Basic SRAM Cell 12

6.1 Option Tuning Result 31
6.2 AVM Methodology Implementation results 32

Abbreviations

SOC: System On Chip

IPs: Intellectual Properties

CMOS: Complementary Metal Oxide Semiconductor

RAM: Random Access Memory

ROM: Read Only Memory

SP: Single Port

DP: Dual Port

IO: Input Output

FC: Full CUT

CP: Critical Path

AVM: Apache Voltage Modeling

Chapter 1

Introduction

1.1 Introduction

The electronics industry has achieved a phenomenal growth over the last few decades, mainly due to the rapid advances in integration technologies and large-scale systems design. The use of integrated circuits in high-performance computing, telecommunication and consumer electronics has been growing at a very fast pace. There are various factors involved for the development of VLSI field, most are reliant on increasing device complexity, as the customer demands for more functionality and high reliability at low cost.

CMOS integrated circuits have been widely used to develop random access memory (RAM) chips, microprocessor chips, digital signal processor (DSP) chips, and application-specific integrated circuits (ASIC) chips. In mobile computing era with increased System on Chip complexity consumers come to expect smaller, higher performing devices with long battery life, the chips that fuel these products must also be functionally rich, consume less power, and come in smaller form factors. For design engineers, these changes mean applying advanced power budgeting which enables a design to meet system power requirements by analyzing, reducing, and tracking down power through the entire design cycle dealing with tighter design margins (such as

sub-1V power supplies). On the other hand, the time-to-market requirement continues to be aggressive, which makes the end-of-design cycle's power signoff step as critical as ever for a successful tape-out.

Having fast and accurate models at all stages of a design is essential if SoC designers have to succeed in designing chips with embedded memories. Therefore, embedded memory characterization is of increasing concern to design teams. However, the move to new process geometries is intensifying the challenge - the number of memory instances per chip increases considerably at advanced process nodes. The parasitics are also becoming more significant in advance process geometries and have started impacting the timing performance of the device. To support the full range of process, voltage, and temperature corners (PVTs) and to cater the sensitivity of process variation, designers have to perform more and more memory characterization runs. On top of that, the data processing per characterization grows exponentially.

Computer aided design (CAD) tools are used for design automation and optimization. Computer simulation is, and will continue to be, an essential part of the design process, both for performance verification and for fine-tuning of circuits. However, the emphasis on simulation must be well-balanced with the emphasis on hands-on-design and analytical estimates, so that the significance of the latter is not overwhelmed by the extensive use of computer-aided techniques. In addition to the transistor-level circuit design issues, the accurate prediction and reduction of interconnect parasitic has become a very significant topic in high performance digital integrated circuits, especially for deep sub-micron technologies.

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 in today's SoC era, nearly 70% of the chip area are occupied by the memory itself. The semiconductor markets have embraced the fact that the architecture of the memory structure has a considerable impact on the performance of the system and any yield loss of memory IP cause the failure of entire chip.

1.2 Motivation

The characterization and testing of the CMOS memory compilers is becoming increasingly difficult as the advanced technologies are ramping up and transistor size is shrinking down to the leading-edge nodes (28nm/18nm or below) to meet the customer requirement. The use of new device structures and increasing number of metal layers are introducing millions of new parasitic effects. These parasitic effects are posing new challenges to the designers especially in parasitic extraction and post layout simulation domain and they are caused by the exploding number of process corners and process temperature and hence extraction and simulation needs to be run multiple time for multiple corner and temperature. This will highly impact the performance time and disk usage. For simulation additional care needs to pick the right DSPF file and right PVT.

To ensure the time to market constraint we are motivated to implement new methodology, which will reduce the performance time as well as disk usage. Device mismatch plays a key role in process variation, which needs to be considered while estimating circuit performance. As technology is progressing the design constraints or marginalities are getting tightened up. All these factors are posing new challenges to the designers as these variations affect the design performance to a great extent and hence the yield is affected. Exhaustive simulation runs are required in order to see the impact of variation on the circuit performance.

1.3 Objective of the Project

Objective of my project is to evaluate the existing design methodology for SRAM memory compilers, identification of pain areas and risk analysis with existing methodologies in upcoming technology nodes. The scope was to improvise the methodology used for parasitic extraction and post-layout simulation & characterization so that IC designers can ensure high-yielding successful silicon design and meet time to market

constraints with golden accuracy. Also to make the existing flow of characterization more efficient by exploring and deploying new methodologies to have gain in productivity of the charcaterization flow.

Chapter 2

Brief Literature Review

2.1 Introduction to memories

Prior to 1970's, magnetic-core technology was used to store digital data, where data bits were stored in magnetic wires wound coil. This type of technology had many drawbacks in terms of performance, size, cost, area, speed, reliability etc. The 1970s saw the dawn of electronic industry mainly because of the introduction of the semiconductor memories. The semiconductor memories are showing continuous improvement in the performance and good reliability with the advancement of technology.

2.1.1 MOS Memories: Introduction

The ideal memory would be low cost, high performance, high density, with low power dissipation, 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, EPROMs, 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 power is turned off.

2.1.2 MOS Memories: 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 much denser than SRAMs, up to four times denser in a given generation of technology. There are many methods for modelling is available. One form is one port view or we can say negative resistance model and second model is two port view or say feedback model which consisting of an amplifier with gain A and a frequency selective filter network with linear transfer function via positive feedback path.

2.1.3 Read-only Memories

Read-only memories (ROMs) 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 to 1.

Erasable programmable read-only memories (EPROMs) also have all bits initially in one binary state. They are programmed electrically (similar to the PROM), but all bits may be erased (returned to the initial state) by exposure to ultraviolet (UV) light. The packages for these components have transparent windows over the chip to permit the UV irradiation.

Electrically erasable programmable read-only memories (EEPROMs, E2PROM, or Esquared PROMs) may be written and erased by electrical means. These are the most advanced and most expensive form of PROM. Unlike EPROMs, which must be totally erased and rewritten to change even a single bit, E2PROMâĂ§s may be selectively erased. Writing and erasing operations for all PROMâĂ§s require times ranging from microseconds to milliseconds. However, all PROMâĂ§s retain stored data when power is turned off; thus they are termed nonvolatile.

A recent form of EPROM and E2PROM is termed Flash Memory, a name derived from the fact that blocks of memory may be erased simultaneously. Their large storage capacity has made this an emerging mass storage medium. In addition, these types of memories are beginning to replace the role of ROMs on many chips, although additional processing is required to manufacture Flash memories in a standard CMOS technology.

2.1.4 Memory Organization

The preferred organization for most large memories is the 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. 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 select 4, 8, 16, 32, or 64 columns in one row depending on the application. The row and column (or columns) to be selected are determined by decoding binary address information.

Memory exists as stand-alone component, but also as embedded blocks in system-onchip. Memory cell circuits can be implemented in a wide variety of ways. In principle, the cells can be based on the flip-flop designs since their intended function is to store bits of data. However, these flip-flops require a substantial amount of area and are not appropriate when millions of cells are needed. In fact, most memory cell circuits are greatly simplified compared to register and flip-flop circuits. While the data storage function is preserved, other properties including quantization of amplitudes, regeneration of logic levels, input-output isolation, and fanout drive capability may be sacrificed for cell simplicity. In this way, the number of devices in a single cell can be reduced to one to six transistors.

At the level of a memory, the desired logic properties are recovered through use of properly designed peripheral circuits. Circuits in this category are the decoders, sense amplifiers, column precharge, data buffers, etc. These circuits are designed so that they may be shared among many memory cells. Read-write (R/W) circuits determine whether data are being retrieved or stored, and they perform any necessary

amplification, buffering, and translation of voltage levels.

2.1.5 Timing Diagrams

Timing diagrams specify the minimum required and maximum expected timing requirements for system actions. The two sets of timing symbols are self-explanatory, one being the standard for timing symbols and the other older one in widespread usage. The operation of the SRAM starts with the detection of an address change in the address register. An address change activates the SRAM circuits, the internal timing circuit generates the control clocks, and the decoders select a single memory cell.

At write, the memory cell receives a new datum from the data input buffers; at read, the sense amplifier detects and amplifies the cell signal and transfers the datum to the output buffer. Data input/output and write/read are controlled by output enable OE and write enable WE signals. A chip enable signal CE allows for convenient applications in clocked systems.

In some systems, power consumption may be saved by the use of the power down signal PD. The power down circuit controls the transition between the active and standby modes. In active mode, the entire SRAM is powered by the full supply voltage; in standby mode, only the memory cells get a reduced supply voltage. In some designs, the memory-internal timing circuit remains powered and operational also during power down.

2.1.6 Critical Timing Path

The critical path determining cycle times comprises the delays through the

i Row address buffer

ii Row address decoder

iii Wordline

iv Bitline

v Sense amplifier

vi Output buffer circuits

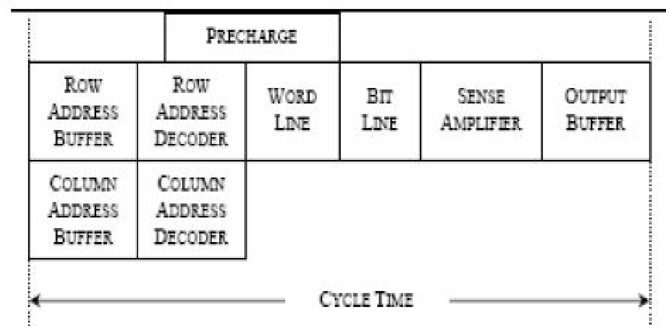

Figure 2.1: Critical Timing Path for SRAM

Precharge and initiation times for sensing as well as column address buffer and decoder delays can be hidden in the critical timing of an SRAM.

The memory clock cycle time, is the minimum time needed to complete successive read or write operations. Maximum read access time should not exceed the memory cycle time since there are write setup operations needed before each memory operation. The cycle time is essentially the reciprocal of the time rate at which address information is changed while reading or writing at random locations.

2.1.7 Functional SRAM Chip Model

Memories are said to be static if no periodic clock signals are required to retain stored data indefinitely. Memory cells in these circuits have a direct path to VDD or Gnd or both. Readwrite memory cell arrays based on flip-flop circuits are commonly referred to as Static RAMs or SRAMs

A functional block diagram for the SRAM chip is shown in the figure 2.2

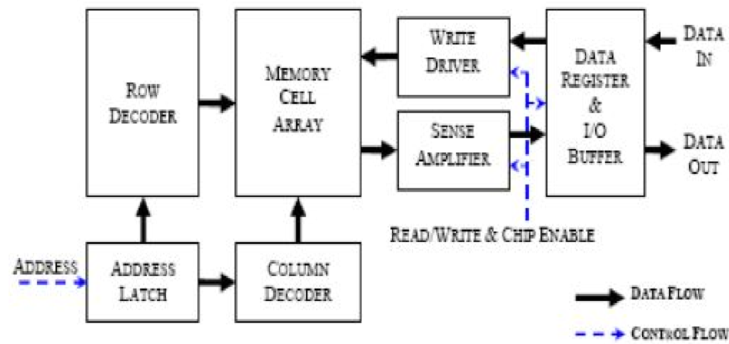

Figure 2.2: Functional SRAM chip model

1. The address latch block, receives the address.

2. The higher order bits of the address are connected to the row decoder, which selects a row in the memory cell array.

3. The lower order address bits go to the column decoder, 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

4. 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, loaded in the data register & presented on the data-out line(s).

5. 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. Usually the data-in & data-out lines are combined to form bidirectional data lines, thus reducing the number of pins on the chip.

6. The chip-select line enables the data register, together with read/write line, the write driver.

2.2 Core of the memory

2.2.1 SRAM Cell description

The basic static RAM cell is consists of two cross-coupled inverters and two access transistors. The access transistors are connected to the wordline at their respective gate terminals, and the bitlines at their source/drain terminals.

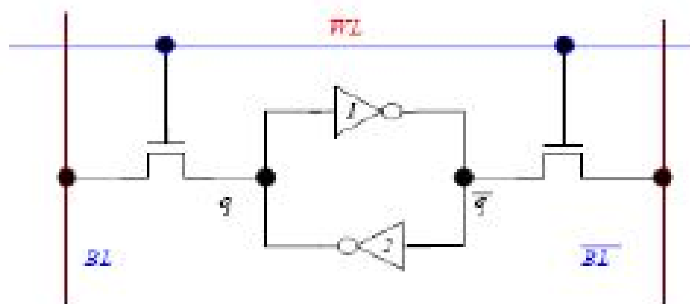

Figure 2.3: Basic SRAM Cell

The wordline is used to select the cell while the bitlines are used to perform read or write operations on the cell. Internally, the cell holds the stored value on one side and its complement on the other side. For reference purposes, assume that node q holds the stored value while node ∼q holds its complement. The two complementary bitlines are used to improve speed and noise rejection properties.

2.2.2 Voltage Transfer Characteristics

The Voltage Transfer Characteristics (VTC) conveys the key cell design considerations for read and writes operation. In the cross-coupled configuration, the stored values are represented by the two stable states in the VTC.

The cell will retain its current state until one of the internal nodes crosses the switching threshold, VS. When this occurs, the cell will flip its internal state. Therefore,

during a read operation, its current state must not be disturbed, while during the write operation the internal voltage is forced to swing past VS to change the state.

2.2.3 6T SRAM Cell

The six transistor (6T) static memory cell in CMOS technology is used in majority of the designs, today. The cross-coupled inverters, M1, M5 and M2, M6, act as the storage element. Major design effort is directed at minimizing the cell area and power consumption so that millions of cells can be placed on a chip. The steady-state power consumption of the cell is controlled by sub-threshold leakage currents, so a larger threshold voltage is often used in memory circuits. To reduce area, the cell layout is highly optimized to eliminate all wasted area.

2.2.4 SRAM Array Operation

In an array the row select lines, or wordlines, run horizontally. All cells connected to a given wordline are accessed for reading or writing. The cells are connected vertically to the bitlines using the pair of access devices to provide a switch able path for data into and out of the cell. Two column lines, b and ∼b, provide a differential data path.

In principal, it should be possible to achieve all memory functions using only one column line and one access device, but due to normal variations in device parameters and operating conditions, it is difficult to obtain reliable operation at full speed using a single access line. Therefore, the symmetrical data paths b and ∼b are usually used.

Row selection in CMOS memory is accomplished using the decoders. For synchronous memories, a clock signal is used in conjunction with the decoder to activate a row only when read-write operations are being performed. At other times, all wordlines are kept low. When one wordline goes high, all the cells in that row are selected. The access transistors are all turned on and a read or write operation is performed. Cells

in other rows are effectively disconnected from their respective wordlines.

The wordline has a large capacitance, C_{word} that must be driven by the decoder, comprises of two gate capacitances per cell and the wire capacitance per cell:

$$C_{word} = (2\,X\,C_{gate} + C_{wire})\,X\,no.\,of\,cells\,in\,a\,row$$

Once the cells along the wordline are enabled, read or write operations are carried out. For a read operation, only one side of the cell draws current. As a result, a small differential voltage develops between b and ∼b on all column lines. The column addresses decoder and multiplexer select the column lines to be accessed. The bitlines will experience a voltage difference as the selected cells discharge one of the two bit- lines. This difference is amplified and sent to output buffers.

It is noted that the bitlines also have a very large capacitance due to the large number of cells connected to them. This is primarily due to source/drain capacitance, but also has components due to wire capacitance and drain/source contacts. Typically, a contact is shared between two cells.

The total bitline capacitance, C_{bit}, can be computed as follows:

$$C_{bit} = (C_{s/d} + C_{wire} + C_{contact})\,X\,no.\,of\,cells\,in\,a\,column$$

During a write operation, one of the bitlines is pulled low if 0 is to be stored, while the other one is pulled low if 1 is to be stored. The requirement for a successful write operation is to swing the internal voltage of the cell past the switching threshold of the corresponding inverter. Once the cell has flipped to the other state, the wordline can be reset back to its low value.

The design of the cell involves the selection of transistor sizes for all six transistors (rather being symmetric only three transistors M1, M3, and M5 or M2, M4, and M6) to guarantee proper read and write operations. The goal is to select the sizes

that minimize the area, deliver the required performance, obtain good read and write stability, provide good cell read current, and have good soft error immunity.

2.2.5 Read Operation

For a "0" stored on the left side of the cell, and a "1" on the right side in the 6T RAM cell, M1 is on and M2 is off. Initially, b and ∼b are precharged to a high voltage around VDD by a pair of column pull-up transistors. The row selection line, held low in the standby state, is raised to VDD which turns on access transistors M3 and M4. Current begins to flow through M3 and M1 to ground. The resulting cell current slowly discharges the capacitance Cbit. Meanwhile, on the other side of the cell, the voltage on ∼b remains high since there is no path to ground through M2. The difference between b and ∼b is fed to a sense amplifier to generate a valid low output, which is then stored in a data buffer.

Upon completion of the read cycle, the wordline is returned to zero and the column lines can be precharged back to a high value. When designing the transistor sizes for read stability, it is ensured that the stored values are not disturbed during the read cycle. The problem is that, as current flows through M3 and M1, it raises the output voltage at node q which could turn on M2 and bring down the voltage at node ∼q. The voltage at node ∼q may drop a little but it should not fall below VS. To avoid altering the state of the cell when reading, the voltage at node q is controlled by sizing M1 and M3 appropriately. This is accomplished by making the conductance of M1 about 3 to 4 times that of M3 so that the drain voltage of M1 does not rise above VTN. In theory, the voltage should not exceed VS, but this design must be carried out with due consideration of process variations and noise. In effect, the read stability requirement establishes the ratio between the two devices.

The other consideration in the read cycle design is to provide enough cell current to discharge the bitline sufficiently within 20 to 30% of the cycle time. Since the cell current, Icell, is very small and the bitline capacitance is large, the voltage will drop

very slowly at b. The rate of change of the bitline can be approximated as follows:

$$\frac{\mathrm{d}V}{\mathrm{d}t} = \frac{I_{cell}}{C_{bit}}$$

Clearly, Icell controls the rate at which the bitline discharges. If a rapid full-swing discharge is desired, Icell is made large. However, the transistors M1 and M3 would have to be larger. Since there are millions of such cells, the area and power of the memory would be correspondingly larger. Instead, a different approach is taken, attaching a sense amplifier to the bitlines to detect the small difference, ΔV between b and ∼b and produce full-swing logic high or low value at the output. The trigger point relative to the rising edge of the wordline, Δt, for the enabling of the sense amplifier is chosen by based on the response characteristics of the amplifier.

2.2.6 Write Operation

The operation of writing 0 or 1 is accomplished by forcing one bitline, either b or ∼b, low while the other bitline remains at about VDD. For SRAM cell taken above, to write 1, b is forced low, and to write 0, ∼b is forced low.

The cell must be designed such that the conductance of M4 is several times larger than M6 so that the drain of M2 is pulled below VS. This initiates a regenerative effect between the two inverters. Eventually, M1 turns off and its drain voltage rises to VDD due to the pull-up action of M5 and M3. At the same time, M2 turns on and assists M4 in pulling output ∼q to its intended low value. When the cell finally flips to the new state, the row line can be returned to its low standby level.

The design of the SRAM cell for a proper write operation involves the transistor pair M6-M4. When the cell is first turned on for the write operation, they form a pseudo-NMOS inverter. Current flows through the two devices and lowers the voltage at node ∼q from its starting value of VDD. The design of device sizes is based on pulling node ∼q below VS to force the cell to switch via the regenerative action.

In the switching process is note that the bitline ∼b is pulled low before the wordline goes up. This is to reduce the overall delay since the bitline will take some time to discharge due to its high capacitance.

The pull-up to pull-down ratio for the pseudo-NMOS inverter can be determined by writing the current equation for the two devices and setting the output to VS. To be conservative, a value much lower than VS should be used to ensure proper operation in the presence of noise and process variations. Based on this analysis, a rule of thumb is established for M6-M4 sizing: W4 = 1.5 X W6

The two ratios M1:M3 and M2:M4 are only estimates. The actual values will depend on a number of factors such as area, speed, and power considerations.

2.3 Power consumption in SRAM

With continuous advancements in technologies, the SRAM memories have undergone changes with respect to following parameters:

- Decrease in geometric cell size
- Increased transistor density
- Higher complexities of the peripheral & control circuitry
- High frequency

Such circuits consume an excessive amount of power and generate increased amount of heat. In case of reduced power processors, memories contribute significantly to the system level power consumption by taking a share of 43%-50%.

The circuits with more power dissipation are more susceptible to run-time failure and reliability problems. In addition, at increased temperatures, high power processors tend to create several silicon failures. As per studies, component failure rate double every 10°C increase in temperature.

The solution for the problem is either to pursue expensive packaging or apply cooling strategies. However, another better option is to restrict the extensive heat generation. For it, the main power consuming areas are studied and efforts are being focused to minimize the same at the extreme nodes.

The power consumption in SRAM can be divided in two modes of its operation i.e. Active and Standby. The power consumption in active mode is in the following sections:

- The core area, it is the main location of power consumption in the SRAM memory.
- The I/O section which uses power during precharge, multiplexer toggling, sensing and output driving also have a significant percentage of power consumption.
- The others sections which include predecoded line toggling and remaining periphery, the control and row decoder section have a small usage of power.

During the standby mode, the power consumption is very low which is used for the purpose of data retention. The main source in this mode is the leakage current in the Memcell. Static currents from other sources are negligible, sense amplifier also being disabled.

The following techniques can be deployed for low power operation of SRAM memories:

1. Capacitance reduction of wordline and bitlines. This helps in reduction of main power consumption in the active mode of operation of memory.

2. Leakage current reduction by utilizing higher threshold voltage devices in the core. This factor helps in reduction of power usage in both of the active and standby modes.

3.Operating voltage reduction. This also needs to improve the periphery circuits accordingly.

4. AC current reduction by using new decoding schemes.

5. DC current reduction by improving pulse operation techniques for wordlines and periphery.

The analysis of SRAM based on various architectures focuses the first point above, i.e. the nerve point of maximum power consumption.

Reduction in power dissipation provides following advantages:

- Better system efficiency is achieved.
- Performance of the system is improved.
- Reliability in enhanced.
- Overall cost is reduced.

2.4 Analysis of SRAM architecture

Fast low power SRAMâĂ§s have become a critical component of many VLSI chips. This is especially true for microprocessors, where the on-chip cache sizes are growing with each generation to bridge the increasing divergence in the speeds of the processor and the main memory. Simultaneously, power dissipation has become an important consideration due to both the increased integration and operating speeds, as well as due to the explosive growth of battery operated appliances.

2.4.1 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 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.

2.4.2 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.

2.4.3 Page Type Architecture

In split-core, architecture although the bank is split is two parts but the word line activates the cell in both and no gain in power is observed. Thus to reduce the run length of the word lines, a new architecture is analyzed.

Here the control unit and row decoder sections are divided in global and local sections. This benefits by activation of cells only one page and thus the wordline capacitance is reduced.

2.4.4 Bank Architecture

This technique to reduce the run length of the bitlines and divided core structure helps in gain in both of speed and power.

Here the control and the Input/Output sections are divided. So for a selected word-line, the cells of only one bank are activated. Also in case of bitline, the numbers of cells activated are reduced. Thus, a significant improvement is observed in case of wordline cap and the bit line cap. There is also reduction in the power consumption to a very significant value. But in this type of architecture, the area used is more and hence a less dense memory is obtained.

Chapter 3

Benchmarking and Options Tuning

3.1 Introduction

Semiconductor memories are capable of storing large amount of Digital information. The data storage capacity of a single chip doubles almost every two years. The number of data bits stored per unit area is one of the key criteria that determine the overall storage capacity, hence, the memory cost per bit. Another important aspect is the memory access time, i.e., the time taken to store or retrieve data in the memory array. The access time determines the memory speed. Static and dynamic power consumption of the memory array is also a significant factor to be considered in the design because of the increasing demand of low power applications.

Memory characterization means to obtain the information about the behaviour of memory in terms of different timings, power and pin cap.

"Performance Characteristics" of SRAM is categorized as:

Timing Characterization: Setup, Hold, Access, Cycle time performance of a given memory at any given PVT
Power Characterization: Dynamic power, Stand-by Power Leakage (Static) power of given memory at any given PVT
Pin Cap Characterization: Capacitance offered by input pins of a memory to the

stage driving it.

For the purpose of providing various timings like cycle-time, address setup and address hold time, power, leakage etc. to the customer, the memory needs to be characterized. To characterize the memory, simulations are done, where a stimuli (types of signals which are applied to different inputs of memory) is given to a netlist. Netlist is basically whole memory circuit written in text form which describes which type of transistors, resistances and capacitors are connected at different places. Simulator is a tool which passes the stimuli (input) to the netlist (memory in textform) and output waveforms are obtained and using another tool and scripts measurement of different timings, power and leakage is done.

3.1.1 Run time optimization and accuracy improvement for memory characterization:

Higher levels of integration, and shrinking IC manufacturing process technology, demands increased transistor-level accuracy as well as optimized run time to achieve time-to-market constraint. To full fill the outlined goal for memory characterization, Option tuning and benchmarking of Fast SPICE and True SPICE simulator for different memory compilers is a very important task.

3.1.2 Simulator option tuning and benchmarking

We can characterize SPICE simulators as True spice (Traditional SPICE) simulators and Fast spice (accelerated transistor level) simulators. From the name itself it is clear that true spice simulators have high accuracy and large run time and vice-versa for fast spice simulator. Since achieving highest possible accuracy of the results is priority (aligned with expectations of SPICE users), true spice simulators include no approximations to device models, uniform time discretization across the entire simulation at every time point. This in turn means constructing and solving a single

system of equations representing the entire circuit at every time-point which results in high run time and huge disk usage. Whereas to reduce run time fast spice simulator uses approximate MOSFET models, does reduction of parasitic networks and are smart enough to partition the circuit dynamically and do hierarchical simulation. Once the simulation is done, it stitches the outputs from individual part as per circuit design.

3.1.3 Need of option tuning and benchmarking

As per the required specification of customer, memory compilers has to be delivered on time with golden results. But Golden simulators take much time in simulation. Our need is to get an accurate result in a less time so we do benchmarking of results of these fast spice simulators and true spice simulators with all worst, best and typical cases for different memory complier. If the results of fast spice simulator differ from golden result (true spice simulator result) then option tuning is to be done with option provided by vendor. Hence to get the accuracy at level of golden simulator's, fast spice simulators should be tuned with proper options. Once it is done we can do memory characterization for any memory cuts and PVT with fast spice simulator and get golden result with less time duration and can full-fill customers need on time.

3.1.4 Benchmarking of memory characterized result and option tuning of simulator

Spice simulators offer various features for optimizing runtime and accuracy based on designer's need. For a circuit simulation various steps are followed by simulator, like netlist parsing, identification of the circuit, DC initialization, transient analysis etc. To start transient analysis all device nodes should be initialized to a particular DC value. If designer has already initialized the nodes as per the requirement of the circuit then simulator has nothing to think upon, but if designer leaves the node uninitialized then SPICE simulator initializes it based on the circuit functionality but fast spice

simulator some time faces trouble in generating accurate result. Similarly floating gates, unbiased bulk create problem for fast spice simulator to generate accurate results. Analysis of the circuit behavior is captured in a waveform format, where simulator offers a feature to probe the nodes so that you can track the signals at a particular node at each instance of time. The current waveform at a circuit node can also be probed. While doing benchmarking designer have to take care that difference between the two results are whether coming under comparison criteria or not and If not then the designer can see the problematic signal and identify the problem. If it is simulator bug then he can fix with several options provided by the vendors to achieve a desired level of accuracy and if its design fault (like uninitialized node, unbiased bulk, floating gate) then designer can correct it. Whenever a new version or new feature of simulator is available, before deploying the simulator to the team benchmarking with respect to golden accuracy (True Spice) is done.

Chapter 4

Full CUT vs Critical Path Verification

4.1 Critical Path Modeling

For characterization of a memory IP, usually a critical path is modelled for a particular instance. Critical path (CP) is optimized RC modeled view of actual design.Full-Cut (FC) is the actual design. Repetitive leaf-cells in FC are replaced by load blocks in CP modeling.

Reasons for critical path modeling are :

- FC View not available at initial stage of design.
- Complex design with over millions of devices.
- Large simulation time for multiple verifications at initial stages.

CP is used for characterization and multiple verifications.

4.2 Need of FC vs CP verification

Full-cut (FC) vs Critical path (CP) verification is performed to verify that CP model used during characterization/verifications and actual layout are fully aligned. Any

mismatch impacts characterization values and/or silicon failure on internal marginalities. While verifying the CP modeling, the figures of merit should include all global signal delays, pulse widths or slews, access time and other major timing labels. For different variants of CP modeling, different comparison criteria should be taken into consideration. Common failure mode includes improper extraction environment, incorrect load factor and other modeling mismatches.

4.3 Flow of verification

Initially extraction for whole full-cut netlist and for individual leaf cells of CP needs to be done. Type of extraction i.e. RCc or Cc depends on the design technology. Extraction corner also needs to be wisely chosen so as to cater worst and best impact of mismatch. After that post-layout simulation is done on that extracted FC etracted netlist. Also simulation of CP modelled netlist along with extracted leaf cells plugged is done. After that measurements are done for the figures of merit (FOM) for which comparison needs to be done. FOM includes All global signals (clocked/non-clocked, vertical /horizontal) delays/pulse widths/slews, Vdiff(offset) and taa(access time)

Chapter 5

AVM Methodology for IR Drop Analysis

5.1 Introduction to IR drop analysis

Advances in process technology and in design styles are increasing the impact of IR-drop effects on the performance and reliability of analog, mixed-signal, RF designs, memory and custom digital IP blocks.IR drop is a signal Integrity effect caused by wire resistance and the current drawn off from power (VDD) and ground (GND) grids. IR drop analysis is of major concern to SoC designer, because Increase in current due to more devices in a design and higher current through each device, Increase in wire and contact/via resistance due to narrower wires and fewer contacts/via, and Voltage drop on the design has more impact on functionality in lower voltage range. For eg. A 100mV drop is acceptable at 3V supply as device will operate in sub threshold region but it is considerable at 0.6V supply, devices may operate in near threshold region (assuming Vt = 400mV). IR drop at SOC level is of major concern to the designers working in nano range. With advanced technology nodes, the metal mesh supplying power to the design also becomes complex. As the resistance of the metal line increases, the current drawn off from power supply also increases and in turn a high

amount of voltage drops across the power grid. This voltage drop may cause timing failures and may alter the gate sub threshold voltages. To estimate this voltage drop for a particular design, i.e., memories in our case, IR drop analysis becomes essential. For a SOC designer, information about the physical layout of memory, its electrical behavior, i.e., current drawn from power supply and its associated decoupling capacitance and the functional behavior in terms of timing and peak power requirement in different operating modes needs to be provided. Excessive IR drop may result in functional failures and/or timing violations.In order to calculate IR drop at SoC level we need to characterize Supply Current and Decoupling Capacitance (De-Cap) values on each supply.

5.2 AVM Methodology for analysis

As there are multiple memory IPs on a SOC, simultaneous operation of these memory IPs requires a high peak current at particular time instant. Therefore, the information about only the peak current value and its occurrence time may not suffice. So, we need to provide the current requirement at each time instant and hence overall current profile during a particular operating mode is provided in form of a current waveform and the associated decoupling capacitance. Simulations are highly time consuming and EDA environment dependent. To avoid these dependencies, we an interpolation based methodology AVM is used which generates the IR information in form of a current waveform and its associated de-cap value.

5.3 Deployment of AVM Methodology in existing characterization flow

It becomes cumbersome to give whole current profile as a waveform as interpolation can not be done for different instances using that. AVM methodology's two triangle approximate approach helps in interpolation too. For dynamic IR drop analysis,

memories active operation modes and standby operation modes need to be considered. These include active read/write, inactive read/write, bypass and shift mode. Also all possible supply voltage needs to be considered for current profile. Two triangular AVM approach uses signal toggling based event to model base of the triangle and area is calculated by integrating current profile in that particular base duration. From this peak can also be calculated. Implementation of this methodology requires a thorough knowledge of design and operation being done. From this event of particular signal toggling and base can be extracted. Also, for different design, the signal may differ as per the architecture being used.

Chapter 6

Implementation and Results

6.1 Benchmarking & Option Tuning

Benchmarking activity was performed on 28FDSOI (28 nm) , B55 (55nm) and C40LP (40 nm) technology compilers. Compilers were chosen in such a way for Benchmarking so as to cover different designs features SP(Single Port), DP(Dual Port) and Low voltage designs. All corner cuts(memory instances) supported by compilers and corner PVTs (SS FF) were chosen for each compiler.

CASE	Fast SPICE Option set	FAST SPICE Measurement (s)	True SPICE (s)	Abs Diff with True SPICE (ps)	Diff with True SPICE(%)	FAST SPICE Runtime
CASE 1	original option	1.06E-09	1.07E-09	16.80	1.56	5min
	Proposed Options	1.07E-09		3.50	0.33	8min
CASE 2	original option	1.69E-10	1.81E-10	12.32	6.79	4min
	Proposed Options	1.81E-10		0.11	0.06	7min
CASE 3	original option	2.01E-11	1.76E-10	155.80	88.57	3min
	Proposed Options	1.76E-10		0.30	0.17	12min
CASE 4	original option	1.55E-09	1.58E-09	34.61	2.18	7min
	Proposed Options	1.59E-09		1.54	0.10	12min

Figure 6.1: Option Tuning Result

The solution found is too accurate and very close to true spice. But Runtime impact is found too high i.e around 3X.There is need of an intermediate solution in terms of accuracy and performance. 5-10 % increase in runtime is considered viable to have.

6.2 AVM Methodology Implementation

AVM Methodology for providing current profile or CAD views enabling IR drop analysis at SOC level was implemented for two different variants of design i.e. for Single Port and Dual Port comprising of different architecture. The methodology implemented was found accurate when compared with original waveform. Results for current profile of a supply voltage of a design in active read operation ca be seen in the figure below.

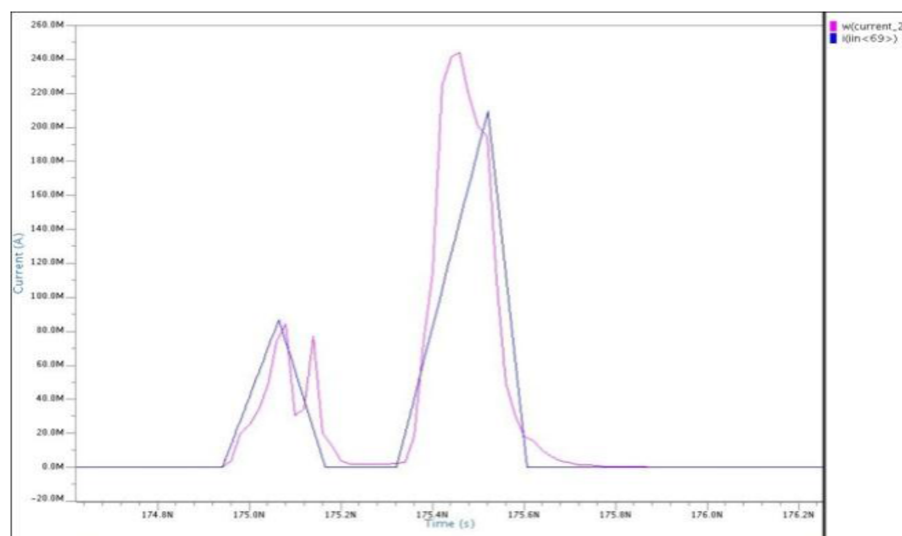

Figure 6.2: AVM Methodology Implementation results

Chapter 7

Conclusion

Various excurses are done for the simulator option tuning for achieving best combination of run time and accuracy. The present characterization setup gives simulated results more accurate as compared to the previous setup. This allows customer to get any memory cut's datasheet having cut's different timings, power, and leakage and dimensional (height, width, & area), value with all pins diagrams and timing diagrams accurately and in very short duration. Memory is used as an embedded block in system on chip (SOC). While designing, customer sets some specification for the memory chip like maximum power consumption, access time, setup time, area etc. Therefore, Characterization of memory is an important step in design flow.
A deeper exploration of the effects of the parameter on the basic memory cell is needed as the technology is shrinking. In particular, a better understanding of the memory cell parameters which are used to characterize the performance, speed and density. Thus, better methods for characterizing the basic memory cell are a rewarding topic for further research.

Chapter 8

Future Work

Future work may include benchmarking activity and options tuning for newer design and simulator release. Also inclusion of AVM Methodology in characterization flow for providing CAD views for different modes of operation of Memory IP for other designs. Also new methodologies improving the characterization and verification flow of embedded memory IP may be explored and deployed for improving the efficiency of characterization flow.

References

[1] ST Internal Documents

[2] J. W. Sleight and R. Rios, *"A continuous compact MOSFET model for fully- and partially- depleted SOI devices"*, in IEEE Transactions on Electron Devices, vol. 45, no. 4,pp. 821-825, Apr 1988.

[3] S. Mukhopadhyay, A. Agarwal, Q. Chen and K. Roy, *" SRAMs in Scaled Technologies Under Process Variations : Failure Mechanisms, Test Variation Tolerant Design"*, IEEE Custom Integrated Circuits Conference 2006, San Jose, CA, 2006, pp. 547-554.

[4] K. Verma, S. K. Jaiswal, D. Jain and V. Maurya, *"Design and Analysis of 1-Kb 6T SRAM Using Different Architectre"*, Computational Intelligence and Communication Networks (CICN), 2012 Fourth International Conference on, Mathura, 2012, pp. 450-454.

[5] Shen Lin and Norman Chang, *"Challenges in power-ground integrity"*, International Conference on Computer Aided Design, Pages: 651 - 654 Year of Publication: 2001