Power and Leakage Saving Technique in Memory

Major Project Report

Submitted in partial fulfillment of the requirements
for the degree of

Master of Technology
In
Electronics & Communication Engineering
(VLSI Design)
By
Prachi Vyas
(17MECV17)

Electronics & Communication Engineering Department
Institute of Technology
Nirma University
Ahmedabad - 382 481
May,2019

Power and Leakage Saving Technique in Memory

Major Project Report

Submitted in partial fulfillment of the requirements
for the degree of

Master of Technology
In
Electronics & Communication Engineering
(VLSI Design)
By

Prachi Vyas
(17MECV17)

Under the Guidance of

Internal Guide	**External Guide**
Dr. N M Devashrayee	Mr. Naveen Batra
Professor (VLSI Design)	Engineering Manager
Nirma University	Synopsys India Pvt Ltd.

NIRMA
UNIVERSITY
INSTITUTE OF TECHNOLOGY
NAAC ACCREDITED 'A' GRADE

Electronics & Communication Engineering Department
Institute of Technology
Nirma University
Ahmedabad - 382 481
May, 2019

Declaration

This is to certify that

1. 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.
2. Due acknowledgment has been made in the text to all other material used.

Prachi Vyas
(17MECV17)

तमसो मा ज्योतिर्गमय

NIRMA
UNIVERSITY

INSTITUTE OF TECHNOLOGY

NAAC ACCREDITED 'A' GRADE

Certificate

This is to certify that the Major Project entitled **"Power and Leakage Saving Technique in Memory"** submitted by **Prachi Vyas (17MECV17)**, 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 her 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 N. M. Devashrayee
Internal Guide

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

Dr D. K. Kothari
Head, EC Dept.

Dr Alka Mahajan
Director, IT - NU

Date :

Place : Ahmedabad

Acknowledgment

Let me take the opportunity to express my deep regards to Mr. Naveen Batra (Project Manager) for assigning me such project and providing his guidance and constant encouragement during the project. I would also like to thank Mr. Rohit Sharma (Mentor) for his guidance, help and inspiring me to put my best efforts.

I would like to express my gratitude & sincere thanks for generous assistance to my guide Dr. N M Devashryee, Professor, VLSI Design, Institute of Technology, Nirma University, Ahmedabad for his guidance and constant encouragement during my course of project. Special thanks to all faculties who has always been an inspiration and guided us with their experience.

I would like to express my gratitude towards my parents for constant support and encouragement in life. I also wish to express my heartfelt appreciation to my friends and colleagues at Intel who have rendered their support throughout my project, both explicitly and implicitly.

- Prachi Vyas
(17MECV17)

Abstract

The performance and cost of an ASIC depends heavily on the quantity and quality of embedded SRAM included in the design. SRAM memory, there are still several options available to configure the memory instances to dramatically reduce cost or improve performance. If these configuration options are either unavailable in the memory compiler or are not chosen properly, the overall design will suffer. Power and Leakage plays a major role to check the product quality before it goes to market. While advanced Pre-Si Quality Assurance checks can catch design errors, it can miss physical bugs which can only be identified by Post-Si validation. Today, low power memory is given most priority in VLSI design. Low power feature for on-chip SRAMs is becoming increasingly important, especially for battery-operated portable application. So, the power reduction for one cell is the vital role in memory design techniques. As the technology growing portable device (e.g. Cell phone, PDA) increases, the Static Power Consumption (Leakage Power) and dynamic power became a significant issue. Leakage current in standby mode is the major part of power loss. SRAM continues to be an important macro block of SoCs. There is some technique through which we can reduce the power dissipation like by using dualrail voltage, address decoding schemes, assist technique, by applying different rmi settings, power gating, etc. Analyzing the memory operation by applying different power supply to periphery and array by inserting level shifter in design and read and write margin is very important to achieve also done some analysis by changing wordline and bitline voltage and try to reduce leakage in memory and also analyze the 6T SRAM parameters and 6T SRAM with different schemes to achieve less leakage and power.

This thesis outlines the architecture of the Memory Chips and its Design Flow and how the efficient methodologies have been proposed for reducing leakage current and Power in Memory design. Its built to check behavior of chip with the different technology on die. The thesis will focus more on the various approaches used for Pre-Silicon Power and Leakage Saving strategy for the Memory based Designs and extend the lifetime of digital circuits.

Table of Contents

List of Tables

List of Figures

List of Abbreviation

PMU	- Performance Monitoring Unit
IoT	- Internet of Things
HSSP	- High Speed Single Port
MC	- Monte Carlo
BIST	- Built in Self-Test
WM	- Write Margin
RM	- Read Margin
PVT	- Process, Voltage, and Temperature
SRAM	- Static Random-Access Memory
HDSP	- High Density Single Port
SoC	- System on Chip
SNM	- Static Noise Margin
VLSI	- Very Large-Scale Integration
DS	- Deep Sleep Mode
LS	- Light Sleep Mode
SAE	- Sense Amplifier Enable
DRAM	- Dynamic Random-Access Memory
RPB	- Row Per Column
CPS	- Column per Segment
DSM	- Deep sub-micron
SA	- Sense Amplifier
CM	- Column Muxing
DIBL	- Drain Induced Barrier Lowering
GIDL	- Gate Induced Drain Lowering
MB	- Multi Bits
PG	- Power Gating
SA	- Sense Amplifier
WE	- Write Enable
SD	- Shut Down Mode

Chapter 1

Introduction

Batteries power mobile devices and many electronic systems. In such systems, optimization of power is a key design constraint. System-on-chip (SoC) designers spend significant amounts of time trying to save battery life in such systems. Along with technology scaling down and higher operating speeds of CMOS VLSI circuits, the leakage power is getting enhanced. As process geometries are becoming smaller, device density increases and threshold voltage as well as oxide thickness decrease to keep pace with performance. Thus Static power consumption is a major concern in nanometer technologies. As memories occupy a larger and larger percentage of SoCs, it is imperative to select memories and memory compilers that provide flexible power management techniques. Power management within memories is key to continuing to extend the battery life, especially in all the mobile devices we use every day.

The embedded Static Random-Access Memory (SRAM) has become vital part of complex processors now days. The dominant area on chip is consumed by the memories. According to SIA road map, this area consumption will increase to 94% by 2014. One of the biggest advantages of SRAM over DRAM and other primary memories is its speed due to static storage. Similarly, the dissipation is the prime threat.

So, efforts should be made for designing and testing of new designs used on different technology on die and different methodology gives the reduction in power, leakage.

1.1 4T SRAM cell with Polysilicon Resistor Load

The main advantage of static 4t cells with polysilicon resistor load is the approximately 30% smaller area as compared to 6t SRAM cells. Due to higher electron mobility, all transistors in a PRL cell are normally NMOS. The load resistor serves to compensate for the off-state leakage of the pull-down devices. On one hand, the value of RL must be as high as possible to retain a reasonable noise margin NML, i.e., to limit the 0 level rise and reduce the static power consumptions. On the other hand, a high RL severely increase the low-to-high propagation delay if Vdd/2 precharge is used and it also increase the cell size. Furthermore, precharging the bit lines to Vdd/2 can compromise the cell stability with scaling of the Vdd/Vth ratio. Precharge of bit lines to full Vdd can alleviate the requirement for low-to-high cell transition current at the cost of the additional precharge time required for a full-Vdd precharge and the associated power consumption. The upper resistance limit on RL is put by the requirement to provide a pull-up current of at least two orders of magnitude larger than the leakage current. The lower limit on RL is put by the required noise immunity and power consumption requirements. The technological variations of RL caused by the limitations of doping and annealing techniques pose another constraint on the increase of RL.

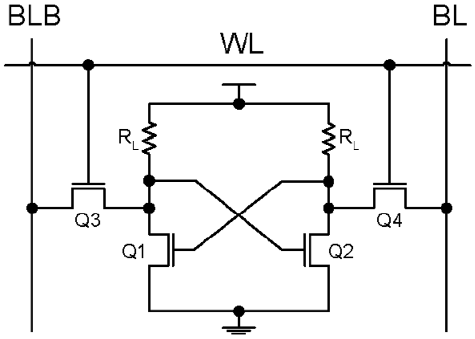


Figure 1.1.1: 4T SRAM cell with Polysilicon Resistor Load

Ratioed inverters comprising the cell have lower gain in the transition region and produce inherently less steep VTCs, which reflects on the SNM values and the recovery time from a metastable state of such cells. The reduction of Vdd from the standard 5v to 3.3v,2.5v and so on, i.e., the switch from constant-voltage scaling to constant-filed scaling to combat the short-channel effects, revealed non-satisfactory low-voltage stability of PRL cells. Moreover, the extra technological steps of forming high-resistivity polysilicon are not a part of the standard logic technological steps of forming high-resistivity polysilicon are not a part of the standard logic technological process. Insufficient tolerance to soft errors, which is directly linked to the SNM, adds to the list of disadvantages of a PRL cell. These factors prohibit using the PRL SRAM cell in SoCs traditionally implemented using standard full CMOS process.

1.2 Conventional SRAM

SRAM cell is the most important block of memory. SRAM called static random-access memory. It has a property to store the data for the long duration of time till power supply is provided. A 1-bit SRAM cell is made up of two inverters connected back-to-back (i.e. CMOS latch) with two complementary access transistors, with two stable states, whose output depend on the data stored in two inverters. The written data in the memory cell can be read as logic1or logic 0. The data can be accessed through bit lines (read and write operation) in SRAM cell as shown in Figure. It is most popular memory cell due to its high stability and lowest static power dissipation. When WL is enabled (i.e. WL=1) access transistors are turned ON for read and write operation which is connected to the bit lines of the cell.

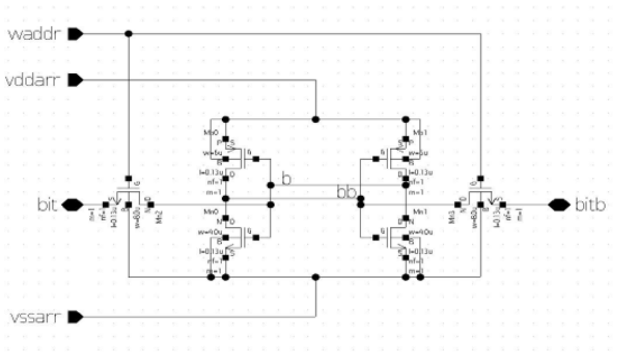


Figure 1.2.1: SRAM Cell (6T)

Write Operation: In a write operation bl is charged to high (i.e. VDD) and blb is a charge to low, now word line is enabled (i.e. WL=1) and access transistors are turned ON and the information is stored into the SRAM Cell. After the fulfilment of write operation, the access transistors turn off, and data is stored for read operation.

Read Operation: In a read operation, both the bitlines is precharged to high (i.e. VDD) and the word line (i.e. WL=1) is enabled. It is well known that driver transistor has stronger driving strength than the access transistor because of one bitline is pulled down via memory cell. Thus, the information at bit lines will be sensed by SA.

SRAM v/s DRAM: RAMs can be categorized mainly into two types, each having its own advantages and disadvantages: SRAM and DRAM. Each of them holds the data in different way. Periodic refreshing of data is required in DRAM to retain the information while SRAM does not requires refreshing because the transistors indeed will continue to hold the data till the power supply is ON. This leads SRAM to have some advantages, one of which is faster speed for the read and write operation. In DRAM for refresh cycle, additional circuit and timing are needed which creates some complications and makes it slower as well as less desirable than the SRAM. One major disadvantage of DRAM is the much higher power consumption and dissipation because of the charging and discharging of capacitors. SRAM structure is rather simple as compared to DRAM and hence it is easier to create an interface to access the memory. A large number of transistors are required by the SRAM structure in order to store a certain amount of memory. On the other hand, DRAM structure needs a capacitor and transistor to store one bit of data whereas SRAM requires 6 transistors. The number of transistors used determines the storage capacity in a memory module.

So, a DRAM module will have 6 times more the capacity when compared to a SRAM module. This ultimately reduces the price which is the major concern for most of the buyers. As DRAM is cheaper, it has been the mainstream in computer for main memory despite of being slower and consuming more power compared to SRAM. On the other hand, SRAM is still used in devices where speed is a major concern rather than the capacity. SRAMs are mostly used in the cache memory of the processors where speed is more important. Optical drives, Hard drives and other devices use SRAM for cache memory and buffers.

1.3 SRAM Design parameters

The proper designing of 6T SRAM bitcell is very important because it directly affects various parameters. For e.g. if we increase the pull-down transistor size significantly to improve the read operation, then there will be difficulty in write operation and the cell wont flip. Similarly, if we increase the pass transistor size to increase the read current then leak current and power dissipation also increases. So, one needs to size the transistors such that every parameter is optimized. The various design parameters include Read current, leak current, Read SNM (Static Noise Margin), Flip time, Read Margin etc. Here we will discuss them in detail.

1.3.1 Read current

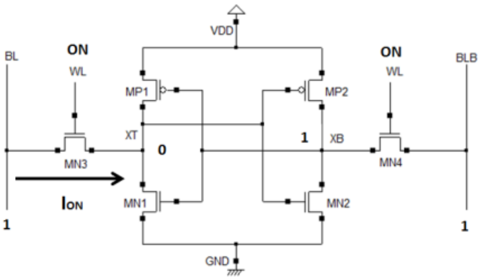


Figure 1.3.1: Iread in 6T SRAM

Figure shows how read current is measured. Read current is the current flowing through the pass transistor when word line turns ON for the read operation. When word line is turned ON, the pass transistor connected to node storing 0 is turned ON and current flows from bitline to ground via pull down transistor. As shown in the figure, XT is at 0 and XB is at 1. BL and BLB are precharged to VDD. Read current is the current flowing through MN3 when WL is turned ON. Read current should be as high as possible because it determines the worst read/leak ratio through which we can estimate the maximum number of physical rows possible.

1.3.2 Leak current

Leak current is the current flowing through the pass transistor when word line is OFF, and the cell is in idle state. Even when the word line is OFF, leakage current flows from bitline to ground via pass gate and pull-down transistor. As shown here in the figure, the cell is in idle state with XT and XB storage nodes at 1 and 0 respectively.

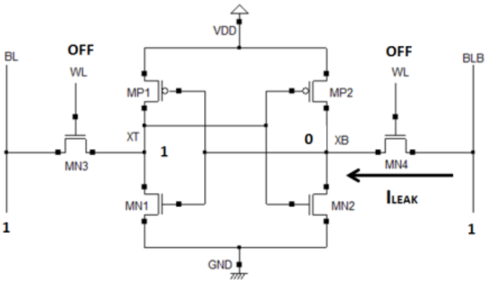


Figure 1.3.2: Ileake in 6T SRAM

Leak current is the current flowing through MN4 when wordline is OFF. Leak current should be as low as possible because it determines the worst read/leak ratio through which we can estimate the maximum number of physical rows possible. Leak current also determines the static power dissipation in the memory array. So large leakage current is not affordable since it can also flip the cell.

1.3.3 Flip time

Flip time is the time required by the cell to flip its value. When we are performing write operation, one bitline is at VDD while other is at ground. When WL turns ON, bitline which is at VDD discharges and after some time cell flipping takes place. This time required to flip the cell is flip time. There are two ways to determine flip time. 1) WL driven: - For WL driven, it is the time required for cell to flip from WL rise 50% to the storage nodes getting flipped. 2) BL driven: - For BL driven, it is the time required for cell to flip from 50% fall of BL to the storage nodes getting flipped.

1.3.4 Static Noise Margin (SNM)

Static Noise Margin in 6T SRAM is the maximum noise voltage that can be tolerated at the storage nodes which does not flip the cell value. It is a very important design parameter as it determines the stability of the cell. The NM defined using the VTC. In general, the NM is the maximum spurious signal that can be accepted by the device when used in a system while still maintaining the correct operation. It is assumed that noise is present long enough for the circuit to react, i.e. the noise is static or dc.

An ideal inverter would have tolerated a change in the input voltage (Vin) without any change in the output voltage (Vout) until the input voltage reaches the switching point. The switching point is presented in fig. as dVout/dVin=1 the switching point of an ideal inverters is equidistant from the logic levels. However in real inverters the switching point is not equidistant from the logical levels and the transition region is characterized by the finite slope as shown in fig. SNM Definition: Noise margin high and noise margin low are defined as below, Where VIL is the maximum input voltage level recognized as logical 0, VIH is the minimum input voltage level recognized as a logical 1, VOL is the maximum logical 0 output voltage, VOH is the minimum logical 1 output voltage as shown in fig. NMH = VOH -VIH and NML = VIL - VOL

SNM can be determined by the butterfly curve as shown in figure above. This curve is nothing but VTC (Voltage Transfer Characteristic) of two back to back connected inverters. SNM is the largest possible square which can be fit in these curves. The next figure shows how we can actually measure SNM in 6T SRAM circuit. Noise can occur at any of the storage nodes. Here we take worst case possible where noise voltage is applied at both the storage nodes. Vx is voltage source and E1 is VCVS. When Vx is increased, E1 also increases. To measure SNM for this circuit WL is turned ON for read operation and simultaneously Vx is also increased.

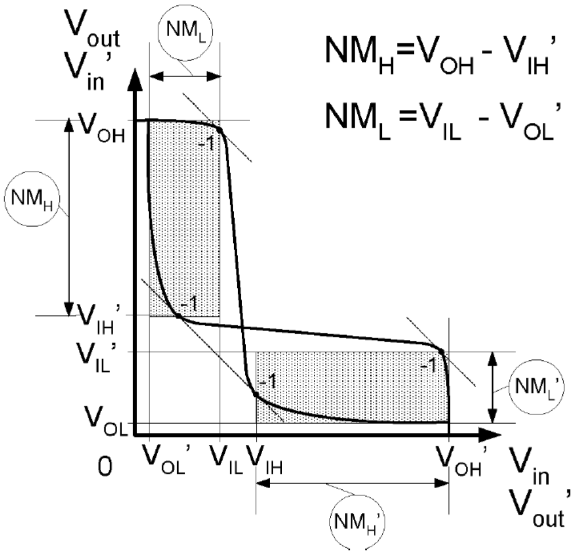


Figure 1.3.3: SNM with VOH and VOL

The voltage Vx at which the cell flips are the SNM of 6T.

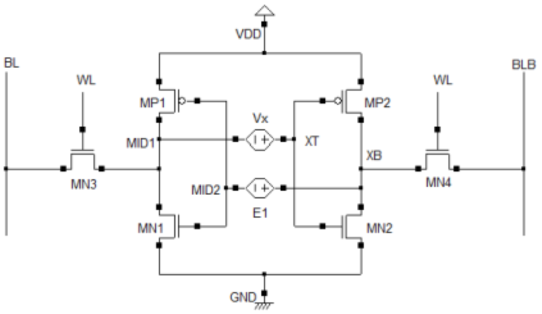


Figure 1.3.4: 6T SNM

1.3.5 SRAM SNM and Operating Voltage Variation

High temperature; the best case SNM occurs for the slow process corner and low temperature. Variation in the operating voltages, such as the supply (Vdd), bit line (Vbl) or word line (Vwl) voltages, strongly impact and SRAM cells SNM. The worst case SNM is typically observed for the fast process corner and high temperature the best case snm occurs for the slow process corner and low temperature. We swept Vdd,Vbl and Vwl one at a time from 0 to 1.5v and measured the corresponding snm. Fig depicts the SNM depends on Vbl while the Vdd, Vwl and Vblb are at the typical 1.2v the situation when one of the bit lines is driven from Vdd to the ground corresponds to a wrte operation. Overwriting the data stored in a sram cell becomes zero at Vbl ¡0.3v for the typical process corner. Note that the SNM does not decreases immediately once Vbl start decreasing. The reduction of Vbl begins to reduce the SNM once Vbl-Vwl will be less then Vthaccess and the access transistors enters the linear mode.

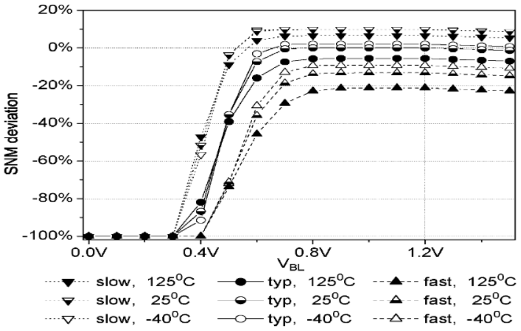


Figure 1.3.5: SRAM cell SNM deviation vs. bit line voltage

Fig a show the read and write regions of a SRAM cell as a function of the bit line voltage. A write operation is possible in the region where the bit line voltage is at or below the voltage point where the SNM is zero. This voltage region is called the write margin. The WM is an important design parameter as its also define the cell stability to various disturbances. A balance between cell stability (SNM), cell area and access speed (read current) must be found, which may not maximize the cell stability.

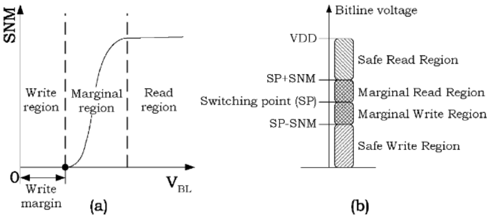


Figure 1.3.6: Read and Write safe and marginal regions of an SRAM cell

1.4 Disadvantages of 6T SRAM

6T SRAM is the most widely used for memory but with the advancement in CMOS technology and decreasing channel length, CMOS is facing lot of problems. At low channel length, stability, leakage current and power dissipation are of major concern. So, a more robust SRAM is required. The disadvantages of 6T are:

- Low cell stability
- Low read current
- High read time
- High leakage
- More power dissipation
- Cannot be operated at ultralow voltage

1.5 Motivation

Today, low power memory is given most priority in VLSI design. The power is most important aspect for todays technology. With the rapid progress in semiconductor technology, chip density and operation frequency have increased, making the power consumption in battery-operated portable devices a major concern. High power consumption reduces the battery service life. Speed, Size, and Price are the three major concern. By emerging battery-operated application on one hand and shrinking technologies of deep sub-micron (DSM) regime on the other hand, leakage power dissipation is rapidly playing a significant role in the total power dissipation as threshold voltage become small. Transistor leakage is a growing problem in re-configurable devices and will soon become the dominant source of power dissipation.

1.6 Objectives

The objective is to reduce the power requirement and leakage in memory by understanding the different parts of memory architecture and sub circuits used in memory chips.

- Optimize Memory architecture and bit-cell design.

- Low power memory techniques
- Test the design under the different checks by varying PVT.
- Understanding of the methodologies and tools used for testing.

1.7 Overview of the Thesis

In this thesis,
Chapter 2 is the Literature Survey which will discuss about the common challenges faced in memory chip design, Process variation testing, memory architecture used for testing and importance of each component in memory architecture.
Chapter 3 will discuss about the read and write operation of memory.
Chapter 4 will discuss about the types of leakage current in device and power.
Chapter 5 will discuss about the implementation of few of the test scenarios and the results obtained.
Chapter 6 will cover result discussion and work conclusion.

Chapter 2

Literature Survey

2.1 Introduction to Compiler

Compiler is software system that takes a users specification and automatically generates and IC. Compiler can be parameterized by number of words, number of bits per word, desired aspect ratio, number of sub banks, degree of column muxing, etc. Area, delay, and energy consumption complex function of design parameters of compiler. Memory Compilers and Logic Libraries supporting a wide range of foundries and process technologies from 250-nm to 7-nm FinFET. Optimized for low power, high performance and high density, Design Ware Memory Compilers offer advanced power management features such as light sleep, deep sleep, shut down and dual power rails, allowing designers to meet the stringent low-power requirements of today's SoCs.

- Pomwerful performance: A variety of memory compilers with multiple periphery VT options, power management modes and rich feature sets optimized for a wide range of applications.

- Proven Architecture, Flexible Design Options: Silicon-proven architecture provides optimized PPA for all types of SoC designs, ranging from performance-critical too cost-sensitive and low-power applications.

- Improve Yield: Memories support BIST and redundancy to improve yield and post-silicon performance tuning.

- Energy Efficient Design: Multiple power management modes support low leakage requirements for IoT SOC design.

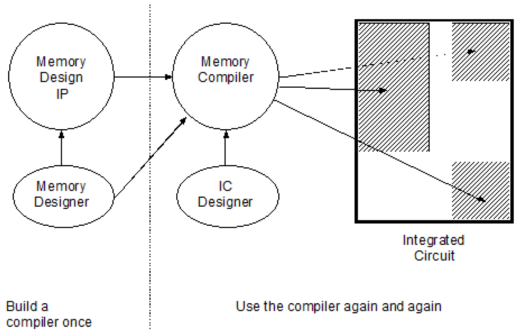


Figure 2.1.1: Compiler Architecture

2.2 High Speed Memory Compiler

7nm product family of memory compilers provides a powerful dashboard of options that enables SoC designers to explore trade-offs between performance, area, power, and statistical yield to generate optimal memory configurations.

HSSP features: -

- Memory instances optimized for speed, without compromising quality.
- Memory compilers that leverage the standard foundry delivered bit cells to ensure high yield and reliability.
- Memory compilers that provide an option for including redundancy capabilities for repair purposes.
- Memory compilers with options for advanced power management modes, such as Light Sleep, Deep Sleep and Shut Down.
- Proprietary circuit design techniques, including high-speed sense amplifiers, fast clocking, and fast bit line recovery, to achieve the high-speed required by todays high-performance applications.

High-Speed memory compilers offer six configuration modes, some of which provide features to support Test and Repair. Each of these features may be combined with advanced power management. They are classified based on built in muxes or bist enable, redundancy enable scan enable. They are defined as below:

- Lite mode without Built-in test muxes or repair.
- Lite Redundancy mode that adds redundancy to Lite mode.
- Integrated Test Lite mode that incorporates fully scannable input and output signals.
- STAR Lite mode that adds redundancy to IT-Lite.
- Integrated Test mode that incorporates built-in test muxes and other test circuitry to enable at-speed testing along with fully scannable input and output signals.

Redundancy: Redundancy is a system design in which a component is duplicated so if it fails there will be a backup. This feature enables the memory to repair bit-cell which are faulty. So the bit-cell which are faulty would be replaced by additional bit-cell.

Read Margin Control: All the arrays of bit-cell would have pre-charged before any operation will going to happen. During the read operation of memory disconnect the pre-charge signal for the selected cell and when the pre-charge is disconnected, one bit-line will start discharging towards the node where '0' is stored. the difference signal between two bit-lines will fed to sense amplifier and it will resolve the output, but it will resolve the output correctly when sufficient difference voltage will be created between two bit-line and that will ensure by RM-control.

Dual Rail Functionality: Separate voltage rails for the array and the periphery may be enabled at the instance level, with level shifters in the periphery. This will reduce dynamic power as well as static power but somehow these features will degrade the performance.

BIST Interface: This is an important feature of the memory compiler. It is the methodology at which circuit can verify its operation itself. After enabling this feature all bist design is activated and this would be test mode, so all the data or address comes from the external pin will isolated.

Bit-write Feature: In this feature can write a specific bit of word in memory instead of whole word. This feature would enable by external pin and it will all other bit of a word and the specific bit that will write in a memory will AND with the control signal and the specific bit will be written into the memory.

Performance Boosters: There are several performance boosting features available within the 7nm compilers. The performance enhancement obtained with each option. The Centre Decode option allows for a trade-off between area and performance. When center decode is set too TRUE, it improves instance performance at the cost of area. Banks (BK) Setting the number of banks provides a compile time option to split the memory into more than one bank. Memory banking is efficient for large instances. It improves performance and active power at the cost of area. Column Mux (CM) Allows you to change the aspect ratio of the instance for chip floorplan for a trade-off between area and performance.

Optional Periphery Transistor Threshold Voltage Selection: A compile-time option, periphery Vt, is offered to select the periphery transistor threshold voltage implant (Vt). The array transistor threshold implants remain unchanged by this compile-time option.Complier Provides power saving design techniques to implement these features:

- Source Biasing
- Fine-grained Power Gating
- Use of Long channel devices
- Integrated Power Gating
- Integrated Level Shifters

2.3 Monte Carlo Simulation

Leakage can easily be considered by this methodology by using Monte Carlo to characterize the leakage current of N worst-case cells as a lognormal. As technology is scaled down to leading edge nodes, chip design engineers face numerous challenges in maintaining the historical figures of performance improvement and the density increase. Performance of the chip, its lifetime and the product yield cannot be determined accurately at the design stage since the chip parameters- such as impurity doping and oxide thickness- cannot be precisely resolute. Process variation is one of the perilous aspects of semiconductor fabrication which impacts the chip performance, yield and the reliability. Process variation is basically defined as the variance in the intended design parameters and the actual (fabricated) design parameters of a circuit and hence reduce the reliability's and yields of chip designs. The relationship between parameter variation and the yield can be viewed in this figure.

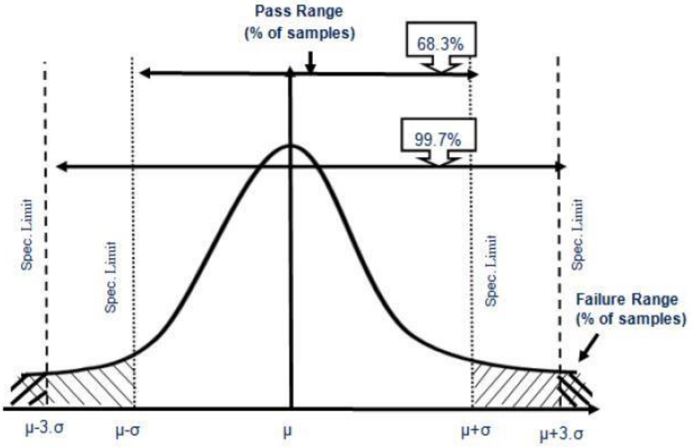


Figure 2.3.1: Relationship between parameter verification and yield

It is important to use such design methodologies which consider the impact of process variation, environmental variation and the uncertainty in temperatures on the chip performance. To ensure the functionality of the design on silicon one must be able to replicate the variability on CAD tools in order to predict the yield. In other words, the design tools and the methodologies from system level down to physical level have to embrace variability impact on VLSI chips which requires ”Statistical-analysis”. For advanced processes, number of corners have expanded which leads to increased design efforts and prolonged time-to-market. Pre-defined corner simulations cannot predict the product yield as they characterize worst case corner only and are pessimistic in nature.

In VLSI circuit design during simulation, we run the design through various PVT (Process, voltage, and Temperature) corners with an aim that the circuit should be able to reliably operate at all the extreme conditions. These PVT variations can be generalized as,1) Temperature from as low as -40 to as high as 125C; 2)Voltage 10% variation from its nominal value; 3) Process This is generally two letter convention where first letter is the behaviour of NMOS and second letter is of PMOS. TT, SS, FF, SF and FS are the corners generally used. Letter T stands for Typical (Nominal VT), F for Fast (Low VT) and S for Slow (High VT).

2.3.1 Global and Local MC

The Monte Carlo simulations can be done in two ways for any given design, Global Monte, and Local Monte. Again, the corner files for these two will be different. These are:

Global Monte: We can think of this Monte run as unconstrained in a way that the variations in this case can span over different process corners. In the figure, each dot represents one Monte Carlo run and as we can see it will spread the variation by introducing a VT change in its every single run. The span of the variations in this global Monte run is spread across the process corners as its name also suggests, Global MC.

Local MC: This Monte run is constrained to a particular process corner. In general, first step is to run the design at various PVT coroners to find the worst one. Then second step is to run the Monte on this particular corner to see the functionality on worst of worst corner. Let us say the worst corner in the first step was found to be SS then the Monte variations will look something like this, this is Local Monte as the scope of variations is limited to a particular corner. Both the methods have their own set of applications and used across industry to emulate the silicon behaviour during simulation and have a working silicon in one go.

2.4 Memory Architecture Blocks

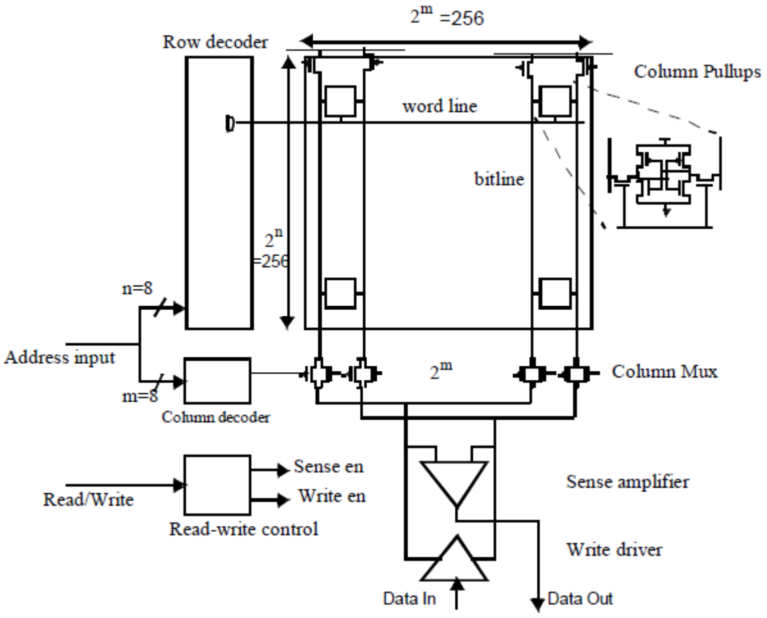


Figure 2.4.1: Basic SRAM block Structure

A row decoder gated by the timing block decodes X row address bits and selects one of the word lines WL0 to WLn-1. If an SRAM array of N rows and M buts is arranged in a page manner, an additional Z-decoder activates the accessed page. Above figure shows an example with four pages of NxM arrays with the corresponding I/O circuitry. Memories can be bit-oriented or word-oriented. In a bit-oriented memory, each address accesses a single bit. Whereas in a word-oriented memory, a word consisting of n (8,16,32 or 64) bits is accessed with each address. Column decoders or column MUXs (YMUX) addressed by Y address bits are often used to allow sharing of a single sense amplifier among 2,4 or more columns. Most of the modern SRAMs are self-timed. All the internal timing is generated by the timing block within an SRAM instance. The main SRAM building blocks will be described in more detail in the following sections.

2.4.1 SRAM Design

The Memory chip architecture and its design flow should be understood properly for the integration in the Memory chip and its tastings. We will consider some of the more recent SRAM cells: a resistive load four-transistor(4T) SRAM cell, a six-transistor (6T) CMOS SRAM cell are already described in section 1 introduction.

CR (cell ratio): The deltaV depended on the CR. Large CR provides higher read current Iread (and hence -the speed) and SNM at the expense of larger area taken by the driver transistors. Whereas smaller CRs make for a more compact cell with moderate speed and noise margins. Both for ensuring cell stability and reducing the leakage current of the access transistors.

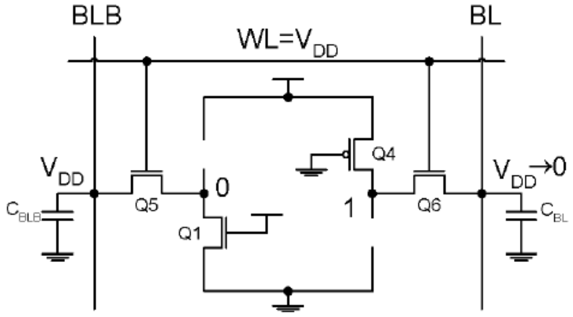


Figure 2.4.2: CR in SRAM

PR (pull-up ratio): To pull the write 1 node below Vthn, the W/L of the pull-up transistors has to be less than 3-4 W/L of the access transistors. The exact maximum allowed PR is defined by the Vthn process option and by the switching threshold of inverters.

2.4.2 Precharge Circuit

To perform the read and write operation in SRAM the bit lines are charged through pre- charge circuit (PCH). PCH Circuit is made up of two PMOS transistors, whose output depends on PCH clock. During the charging of bit lines, a third transistor is placed between them for equalizing of any voltage difference on bit lines. It is mandatory to equalize the voltage on both bit lines

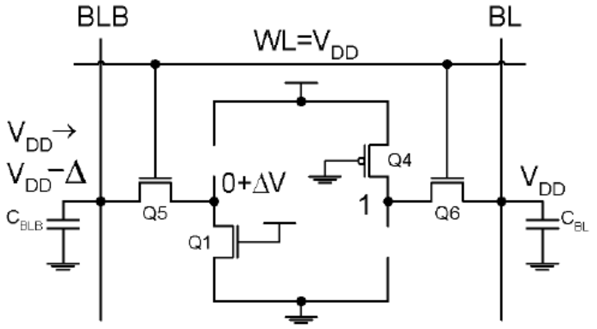


Figure 2.4.3: PR in SRAM

for proper read operation. As the read and write operations are completed a small variation in voltage is developed across both bit lines which require the PCH circuit to again pre-charge the bit lines equal to the supply voltage.

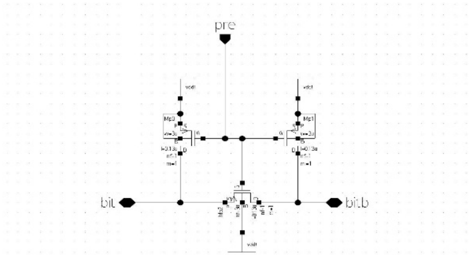


Figure 2.4.4: Precharge Circuit

2.4.3 Sense Amplifier

Sense Amplifier analysis to determine the offset, sense delay and its characterization methodology. Amplifies the differential voltage between bit-lines. Typically consists of cross-coupled inverters with stacked NMOS pull-down transistor. Major part of power consumption in memory system. Play dominant role during read operation. Different sensing schemes are there. The working principal of the SA is as follows: during read operations it senses small voltage difference between beltlines. Offset Voltage: It is a minimum differential voltage between beltlines.

After accessing the memory (Read cycle), a differential signal develops between the bitlines. This differential signal is fed to the input of the bitline sense amplifier to determine what data is stored in the bitcell. Since it takes time for the differential signal on the bitlines to develop, the greater the time delay prior to strobing the sense amplifier, the greater will be the differential signal at the input to the sense amplifier. A low sense amplifier differential signal is susceptible to noise and sense amplifier input voltage offset. Higher input differential voltage results in greater reliability of the sensed data. However, delaying the time when the sense amplifier is strobed results in a longer cycle time, reducing maximum operating speed and increasing access time. Hence the trade-off memory speed verses yield/reliability. The longer you wait, the easier it is for the sense amplifier to determine what was stored in the memory cell. Thus, the term Robustness. The longer you wait, the longer it takes to access the cell (i.e., access time). Thus, the term Speed Trade-off.

For read operation, we precharge RLB and RBLB both to VDD. Now suppose data stored at XT is 0 and that at XB is 1. We want to read XT i.e. 0. When wordline is turned ON, current starts to flow from RBL to pass gate to pull down transistor to ground. On the opposite side, no current flows because XB and RBLB both are at VDD. As current flows through XT, RBL discharges to ground. This discharge rate is very low since the sizing of transistors is very small. Hence, we need sense amplifier to sense this differential voltage and increase the read speed by pulling RBL to ground. When a sufficient differential voltage of about 50-100 mv is established between RBL and RBLB, SAE signal goes high and sense amp comes in the picture. The voltage at node QB is VDD and voltage at node Q is VDD-dV. Hence the Vgs of N2 will be lower than the Vgs of N1. So according to the drain current equation which depends directly on Vgs, the current flowing through N1 will be larger than the current flowing through N2. As these are cross coupled inverters, it forms a positive feedback and more current starts flowing through N1 as compared to N2. In this way node Q is pulled down to logic 0 before node QB because of the bistable nature of this structure. The node Q is connected to output data buffer. In this way sense amplifier increases the read speed because its size is much larger than the bitcell. For each column there is a sense amplifier, so we can afford larger sizing of its transistors because at a time only 1 cell will be operating in a column. Normally the size of sense amplifier transistors is about 10 times that of bitcell, but it may vary depending on the bitline load and speed requirement.

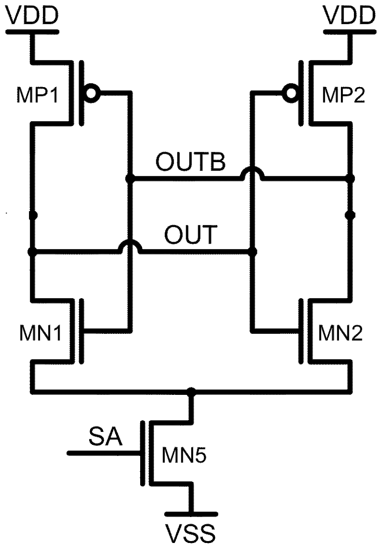


Figure 2.4.5: Sense Amplifier

2.4.4 Write Driver

The function of the SRAM write drivers is to quickly discharge one of the bit lines from the precharge level to below the write margin of the SRAM cell. Normally write driver is enabled by the write enable (WE) signals and drivers the bit line using full swing discharge from the precharge level to ground. A greater discharge of the highly capacitive bit lines is required for a write operation, a write operation can be carried out faster than a read operation. Only one write driver is needed for each SRAM column. Thus, the area impact of a larger write drivers is not multiplied by the number of cells in the column and hence the wryer drivers can be sized up if necessary. Some of the typical write driver circuits are presented in below figure.

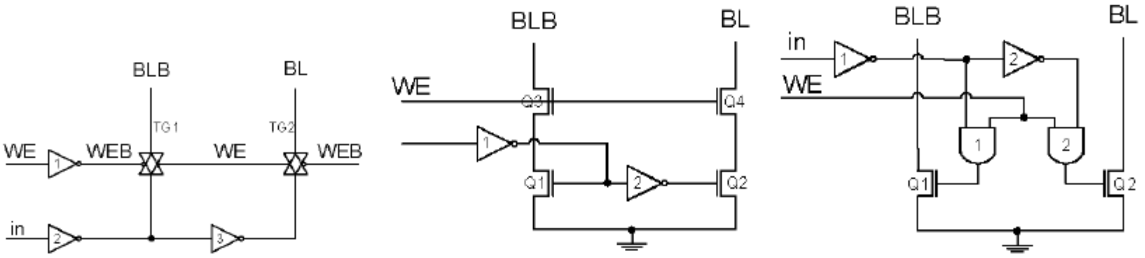


Figure 2.4.6: Write Drivers

2.4.5 Address Decoding Schemes in Memory

A decoding scheme is also discussed in this paper which describes how to choose the best pre-decoding and post-decoding schemes based on minimum pre-decoded lines, minimum stack size in post decoder and maximum granularity of xdecoders. In a typical random-access memory architecture, the row and column to be selected are determined by decoding binary address information. For example, an n-bit decoder for row selection has 2n output lines, one of which is activated. The column decoder takes m inputs and produce 2m bitline access signal. The bit selection is done using multiplexer circuit to direct the corresponding memory cell output to data registers. In total 2n * 2m cells are stored in memory array. An n-bit decoder requires 2n logic levels, each with n inputs.

- Side Decode Addressing: Word lines other signals being driven from side. (Unbalanced, especially for wider memories) Power Compromised performance on wider memories.
- Center Decode Addressing: Word lines other signals being driven from center. (Balanced) Reduces RC delay Power improves performance.
- Memory Addressing: Addressing is considered as 2n, where n is the number of address bits.

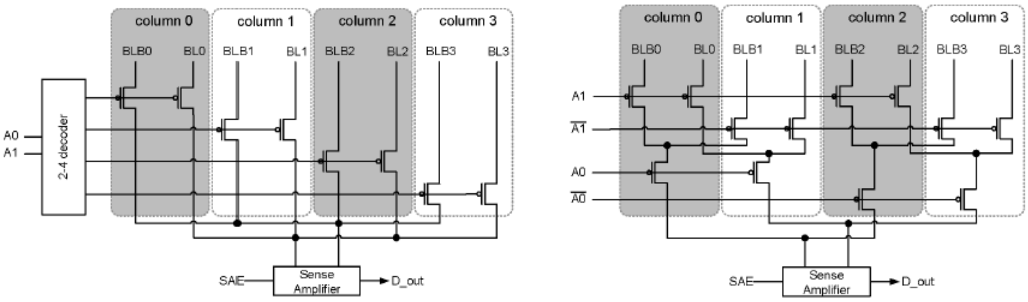
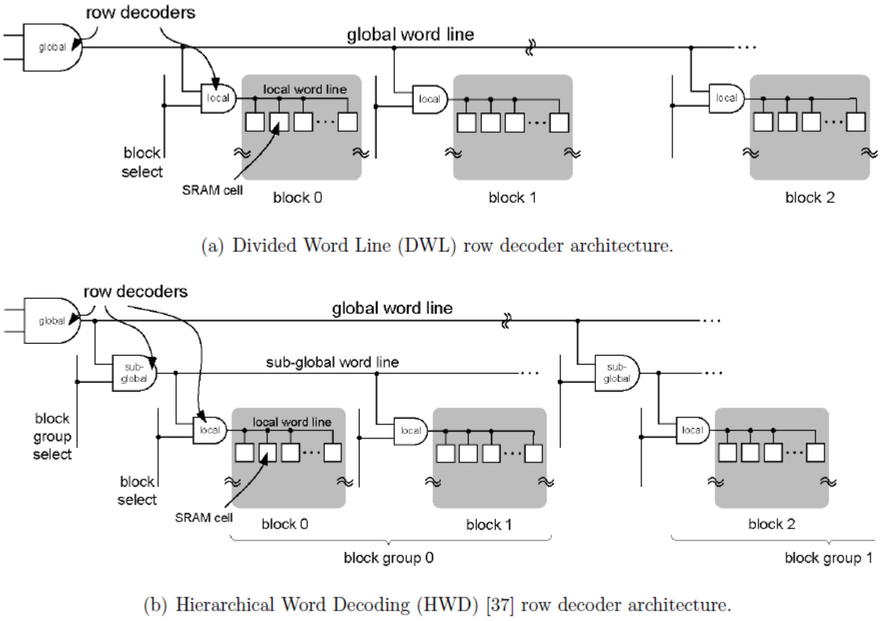


Figure 2.4.7: 4-1 pass transistor Column Decoder with a read predecoder

For a 11 bit of address you can help manage addressing scheme in below fashion. LSB used for column decoding, MSB for Row decoding. If CM=4, A0 and A1 will be used for column decoding. A2-A10 will be used for 512 rows. If bank type architecture supported and bank=2, A2 will be used for bank and A3-A10 will be used for 256 rows.



(a) Divided Word Line (DWL) row decoder architecture.

(b) Hierarchical Word Decoding (HWD) [37] row decoder architecture.

Figure 2.4.8: DWL HWD Multi-Stage Row Decoder Architectures

2.4.6 Timing control schemes

The timing control block controls precharge, word line, sense amplifier clocking and write drivers activation to ensure the correct write and read operation. If timing relation not ensure then timing hazards may arise like if the address changes before the read operation is complete. Suppose precharge is deactivated in this case more than one SRAM cell will be discharging the bit lines which may lead to reading erroneous data. Basic timing control methods employed in SRAM include Direct clocking, Self-timed replica loop and Delay line using a multitude of inverter to define the timing intervals.

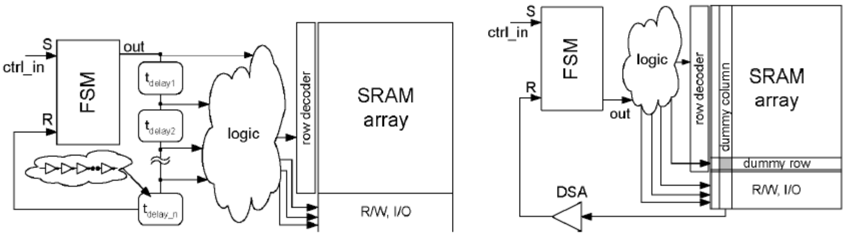


Figure 2.4.9: Delay line timing Loop and Replica timing Loop

2.4.7 Muxing and Banking

The two main options available to a memory compiler are the horizontal muxing (or banking depth) and the vertical partitioning. Consider an example 64-kbit memory logically organized as 2048 words of 32 bits each it would be impractical both from the memory design viewpoint and the system design viewpoint to implement this memory directly as an array of memory cells 2048 tall by 32 wides. Each word-width slice of memory can be referred to as a bank, and the total number of words on a row is the banking depth. By using one of these two techniques, the aspect ratio of the memory may be adjusted to a preferable physical aspect ratio.

This results in wasted dynamic power dissipation, which is the largest drawback of deep muxing. The main advantage of the banked approach is that extra circuitry may be added in between bank slices to disable unused memory banks when they are not accessed thereby saving dynamic power when compared to the interleaved muxing approach. Unfortunately, much of the power savings will be offset by the need to drive the bus to the mux elements at the bottom of the memory. The extra circuitry also requires more area to implement. For these reasons, interleaved muxing is the preferred solution for all but the very largest of memory instances (in the multi-megabit range).

2.4.8 Level Shifter

Level Shifter circuitry is used to switch the voltage domain of the word line passing into the array of bit cells to a desired voltage. With this split, memory is able to execute in a dual rail mode which provides the flexibility to execute the core of the memory with a different voltage then the peripheral section of the memory. Generally Buffer type or Latch Type Level Shifters are available. A Level shifter is placed in the rail voltage domain of the cell. Otherwise a new voltage area is required for such Level Shifter placement. Level shifters are added to ensure that blocks operating at different voltages will operate correctly when integrated together in the SoC. Level shifters must ensure the proper drive strength and accurate timing as signals transition from one voltage level to another. Level shifters can be inserted during the synthesis or implementation stage. In memory for each block separate level shifter is being used mostly we analyses the 3 blocks level shifter one wordline selection second control block and 3rd is lcen control block.

Table 2.4.1: Power Modes

Mode	BC2	BC1	LS	DS	SD	BC0	Mode
LS	0/1	0/1	1	0	0	1	Leakage saving only because of periphery
DS	0/1	0/1	x	1	0	1	Leakage saving because of complete periphery shutdown
SD	X	X	X	X	1	X	Leakage saving because of array and perphery shutdown

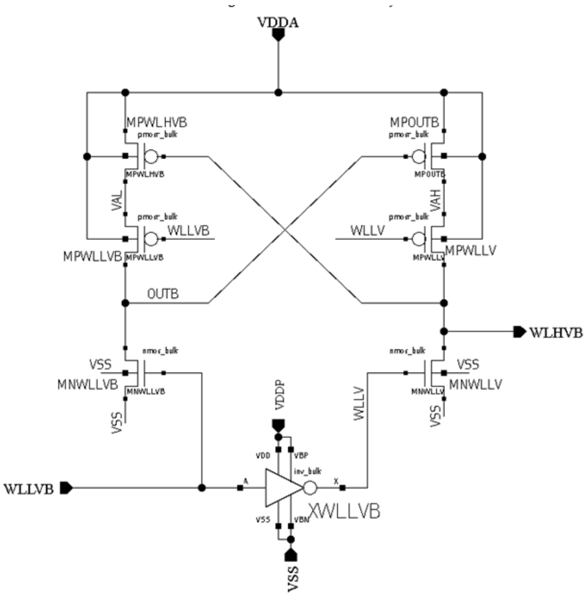
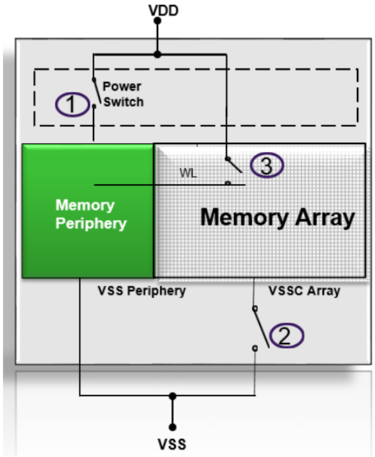


Figure 2.4.10: Level Shifter

2.4.9 Corebias

Objective of corebias design is to design diodes which will provide Array leakage reduction while maintaining the retention voltage. For e.g. if retention voltage for a given technology is 0.7V and Vmin for operation is 0.8V then Diode must be design for 100mV rise on the VSSC node. Different power modes controlling is done by corebias circuit.

Biasing and retention voltage: The information is stored only in the core, Periphery consists of pure combinational logic. There is no disadvantage in switching off the power supply to the periphery (no supply = no leakage). We can employ two different power supplies one to the core (array of bitcells) and another to periphery. The core which is an array of volatile bitcell, needs power supply in order to retain data stored in the bi-stable latch. The voltage needed to retain the data (retention voltage), need not be the same as the operating voltage of the chip, it is often less. Leakage through the array can be reduced by lowering the potential difference across it. We will also have to ensure that this reduced potential difference is not below the retention voltage of the bitcell.



Light Sleep (LS) – (3) OFF
Vssc > Vss (Retention ensured)

Deep Sleep (DS) – (3) & (1) OFF
Vssc > Vss (Retention ensured)

Shut Down (SD) – (3) & (1) & (2) OFF
Vssc floating (No retention)

Figure 2.4.11: Corebias

2.4.10 Level Detector

LD analysis technique to ensure sizing of weak wake-up PMOS is sufficient to pull-up the VDDPI (power gated supply) above the threshold of LD circuitry. Determination of VDDPI level at which the output of the level detector circuitry trips to the intended low logic. This is a voltage trip point analysis to determine the trip point of the level detector circuitry while VDDPI is being ramped up. Once the output of the level detector circuitry is tripped to intended low logic, that level of the VDDPI is captured. The VDDPI level achieved at this stage is due to the ramp-up using the smaller sizing of the power-up pmos known as weaker pmos. The intention of this analysis is to capture this level of VDDPI on various PVTs using foundry models. Rest of the level of the VDDPI is ramped up using relatively larger size power-up pmos known as stronger pmos. The concept of using different pmos is a trade-off to achieve the required level of VDDPI (while transitioning from power down mode to active state) and the peak current.

2.4.11 Address Transition Detector

In an asynchronous SRAM a read or a write operation is initiated by an address change or chip enable signal, where in synchronous SRAMs a read or write operation is initiated by the systems master clock. An asynchronous SRAM features an ATD that produces a pulse td of a controlled duration on every address transition. Every transition of address bus A0-An-1 produces a pulse on the output which initiates a read or a write operation.

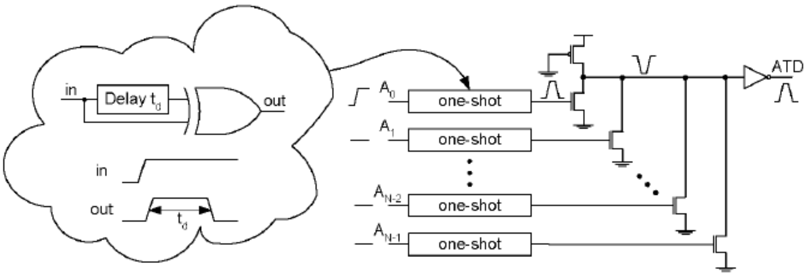


Figure 2.4.12: Address Transition Detector

Chapter 3

Memory Read and Write Margin

In compiler designing this read and write margin decides the memory performance. So, tuning of read and write need to be imp. By read tuning can control the memory access time and cycle time. While write will decide the memory cycle time and write operation. In read margin there are 2 settings internal and external rm settings.

3.1 Memory Read Operation

The memory array bitlines are pre-charged prior to a memory cell access. After accessing the memory (Read cycle), a differential signal develops between the bitlines (bitline and bitline bar). This differential signal is fed to the input of the bitline sense amplifier to determine what data is stored in the bitcell. Since it takes time for the differential signal on the bitlines to develop, the greater the time delay prior to strobing the sense amplifier, the greater will be the differential signal at the input to the sense amplifier. However, delaying the time when the sense amplifier is strobed results in a longer cycle time, reducing maximum operating speed and increasing access time. Hence the tradeoff of memory speed verses yield/reliability. The longer you wait, the easier it is for the sense amplifier to determine what was stored in the memory cell. Thus, the term Robustness. The longer you wait, the longer it takes to access the cell (i.e., access time). Thus, the term Speed Tradeoff. So, Read Margin is one of the most important criteria while designing because we need to turn ON the sense amplifier at the right time otherwise read data will be incorrect. It also affects the speed. In memory compiler there are different modes of operation i.e. Default, Fast, Slow. The SAE signal is determined according to the mode selected. There are 4 pins available to user for RM.

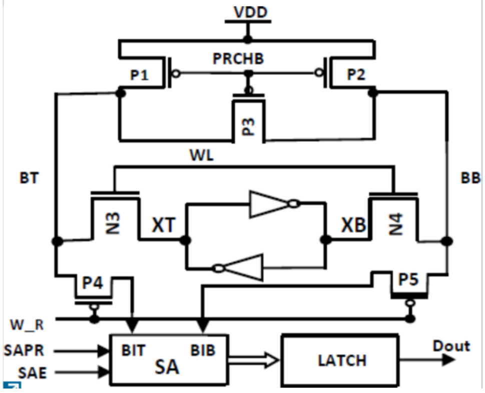


Figure 3.1.1: Memory Read Operation

Depending on the configuration selected, the time required till the SAE signal goes high changes. This concept of Read Margin is shown in below figures. As seen from the waveforms below, initially BT and BB both are precharged to VDD. Now when WL is turned ON for read operation, BT discharges to ground while BB remains at VDD and hence a differential voltage is developed between two bitlines. Now in first waveform, SAE is enabled fast and so we get output very fast. However, in second waveform, SAE is enabled late due to which enough differential voltage is developed and we can ensure a correct read operation. But at the same time the time also increases. So, depending on when SAE is enabled, we have three modes i.e. DEFAULT, FAST, SLOW.

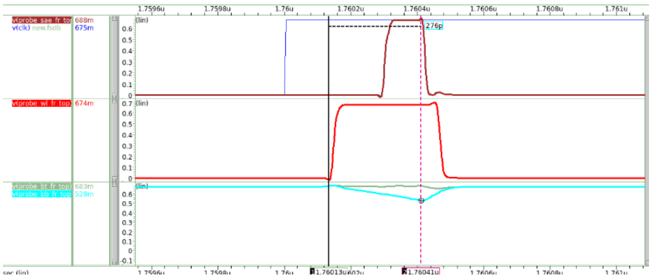


Figure 3.1.2: Memory Read Operation Measurement

For fast mode we want SAE signal to be fastest which turns ON Sense amplifier quickly when low differential voltage is developed. Similarly, for SLOW mode settings are kept so that

SAE reaches very late to ensure enough differential voltage is developed. For DEFAULT mode in between settings are kept. Here all those settings work only when clk is active and bitlines are precharged.

3.2 Memory Write Operation

In 6T SRAM cell write and read operations are initiated by activating the WL and deactivating the precharge of bit lines. At a time either one of the two operations can be performed which is selected by a write read select signal. Read write multiplexer controlled by write and read selected signal connects bit lines of the 6T SRAM cell either to the write driver or to the SA. If write and read selected signal equal to 1 then write operation will be performed and when write and read selected signal 0 then read operation will take place. The schematic setup for write operation for 6T SRAM cell. Here circuit diagrams of precharge block and write driver are connected to the bit lines of 6T SRAM cell to perform the write operation. Thus write driver pulls down either one of the two-bit lines to logic 0 and keeping other one at logic l depending on the value of data (Di) supposed to be written in the cell. In this operation the storage node of the 6T SRAM cell holding logic 1 is pulled down to logic 0 and storage node holding logic 0 is pulled up to logic l then regenerative action of cross coupled inverters will force to flip the cell. Thus the cell is written properly.

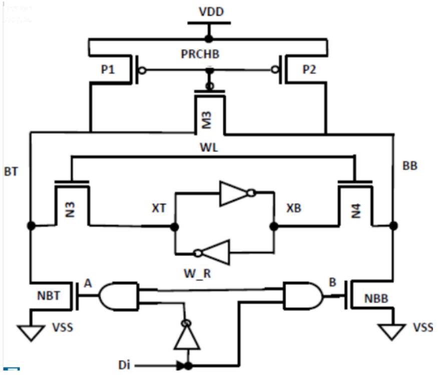


Figure 3.2.1: Memory Write Operation

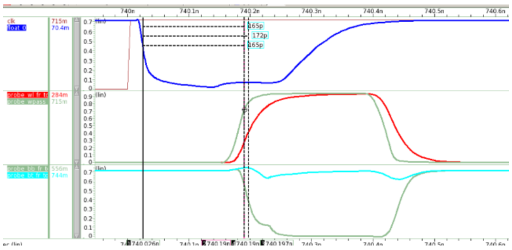


Figure 3.2.2: Memory Write Operation Measurement

Chapter 4

Basic of Power Leakage

4.1 Power

One major limitation on MOSIC is internal power dissipation. Power dissipation in ICs converted into heat. Power dissipation in digital CMOS can be classified into three categories.

$$Ptotal = \text{Pdynamic } + \text{Pstatic} \tag{4.1}$$

Dynamic power: Dynamic power dissipation (Pdynamic) is due to charging and discharging of output node capacitance during active mode and During the switching of transistors. Depends on the clock frequency and switching activity. Consists of switching power and internal power.

$$Pdynamic = \text{Pswitching } + \text{Pshortcircuit} \tag{4.2}$$

Short-circuit Power: short circuit power dissipation (Pshort-circuit) is due to the existence of any conducting path between power supply and ground. Both PMOS and NMOS are conducting for a short duration of time. When Vin bteween Vtn and Vdd-Vtp more rise/fall time and Lower the threshold voltage case the more short circuit.

Static power: Static power is consumed even when chip is quiescent. Leakage draws power from nominally OFF devices. Ratioed circuits burn power in fight between ON transistors. Transistor leakage current that flows whenever power is applied to the device. Independent of the clock frequency or switching activity.

$$Pstatic = (\text{Isub} + \text{Igate} + \ + \text{Ijunc} * \text{Icontention})\text{VDD} \tag{4.3}$$

Internal power not equal to Short Circuit Power: Internal power is any power dissipated within the boundary of a cell. During switching, a circuit dissipates internal power by charging or discharging any existing capacitances internal to the cell. Internal power includes power dissipated by a momentary short circuit between the P and N transistors of a gate, called short-circuit power. When signal transitions from low to high, the N-type transistor turns on and the P-type transistor turns off. However, for a short time during signal transition, both the P- and N-type transistors can be on simultaneously. During this time, current flows from Vdd to GND, causing the dissipation of short-circuit power. For circuits with fast transition times, short-circuit power can be low. However, for circuits with slow transition times, short-circuit power can account for more than 50 percent of the total power dissipated by the gate. Short-circuit power is affected by the dimensions of the transistors and the load capacitance at the gates output. In most simple library cells, the input transition time and the load internal power are due mostly to short-circuit power. For more complex cells, the charging and discharging of internal capacitance can be the dominant source of internal power.

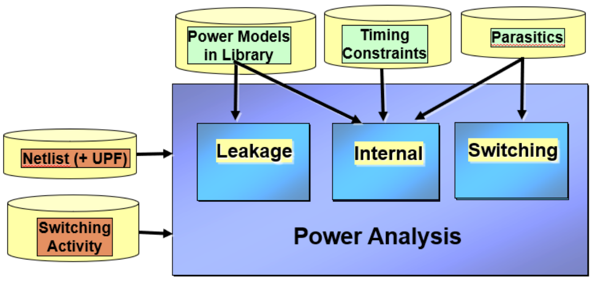


Figure 4.1.1: Power Analysis Requirement

4.2 Leakage Current in Sub-Micrometre Gate

Leakage is a static current flowing from power supply, passing through MOSs and going into ground. Leakage is always existing in circuits when there is a level drop between power supply and ground. Leakage happens at whatever status circuits is, circuit net can be either in toggling/developing status or in static/settled status. Leakage is always determined by the off-status MOSs. Leakage power becomes more along with less channel length but decrease on the GATE thickness increment. And Leakage increase with MOSs Channel Width. Leakage power is a very important performance data. Voltage much impacts leakage drops, much sensitive with voltage changes and temperature. Actual leakage levels vary depending on biasing and physical parameters at the technology node (doping, tox, VT, W, L, etc.)

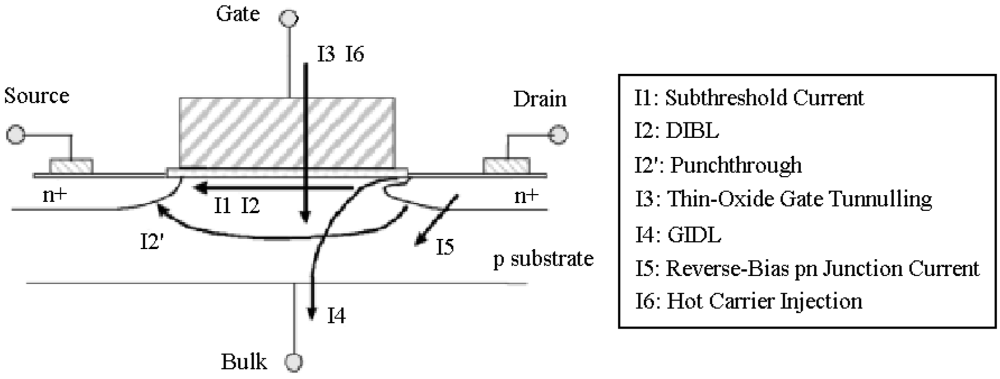


Figure 4.2.1: Leakage in MOS

Source and drain junctions: Are normally reverse-biased so they will leak current typically very small but may increase with scaling since doping levels are very high in future technologies.

Gate Oxide Tunnelling: Due to high electric field across a thin gate oxide. tox has been scaling with each technology generation so We reached the point where tox is so small the direct tunnelling occurs.

Subthreshold Leakage: Is the most important contributor to static power in CMOS. Note that it is primarily a function of VT; Higher VT, exponentially less current. Always flow from source to drain. Carrier diffusion when Vgs less than Vth causes sub threshold leakage. Higher Vth results in lower leakage, longer delay.

DIBL: For long-channel device, the depletion layer width is small around junctions, so VT does not change noticeably. For short-channel devices, as we increase VDS, the depletion layer will continue to increase and help to reduce the VT and VT will continue to decrease as depletion layer thickness grows. If source and drain depletion regions merge even without bias Punch-through occurs. As the drain/source depletion region continues to increase with the bias, it can interact with the source to channel junction and hence lowers the potential barrier.

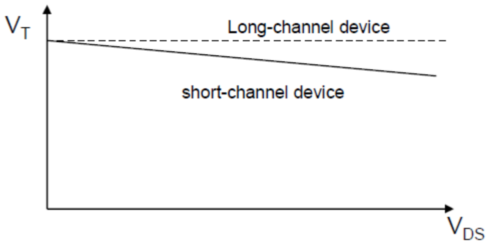


Figure 4.2.2: DIBL

Punchthrough: The punch through mechanism is described as reverse bias applied to drain, which results into extended depletion region. The two depletion regions of drain and source therefore are intersectional with each other, and this results into ”one” depletion region, and flow of leakage current and consequently breakdown of MOSFET.

Hot carrier injection: Electric fields tend to be increased at smaller geometries, since device voltages are difficult to scale to arbitrarily small values. hot carrier effects appear in short channel devices. The field in the reversed biased drain junction can lead to impact ionization and carrier multiplication. The resulting holes contribute to substrate current and some may move to the source, where they lower source barrier and result in electron injected from source into p-region.

Reverse Biased Diode Current (Junction Leakage): Parasitic diodes formed between the diffusion region of the transistor and substrate. Reverse biased diode current can be reduced by decreasing the junction area it depends on material used.

Gate Induced drain leakage (GIDL): Caused by high field effect in the drain junction of MOS transistors. When Vgs = 0V; Vd = Vdd. Avalanche multiplication and band-to-band tunneling. Minority carriers underneath the gate are swept to the substrate. GIDL increases with: Higher supply voltage, thinner oxide and increase in Vdb and Vdg.

Latch-up: Creation of low impedance path between power rails nets as consequence of triggering the parasitic thyristor structure SRC (Silicon Control Rectifier) the PNPN device that is created with two bipolar transistors. This leads to high current flow and permanently damage of device if not limited. Once it occurs the current is not sensitive for any control signals. Triggering may occur due to: input or output voltage that exceed the power supply and ESD event (high current/voltage injection). The pulse supposed to be wide otherwise thyristor may not be fast enough to respond and trigger. If it comes to ESD event which is in fact very short the high current/voltage present at the input/output cause the charge to flow away slowly the pulse is wider, latchup may occur.

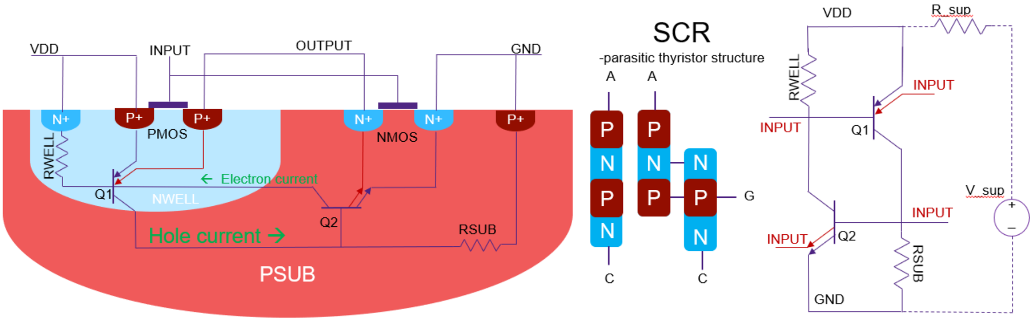


Figure 4.2.3: Latchup in CMOS

Q2 NPN transistor may be turned on by injecting sufficient current into its base (base emitter junction will be forward biased) this leads to obtain IR drop on RWELL resistor and finally trigger Q1. As a consequence positive feedback structure is created leading to excessive current flow from VDD to GD. The only way to stop this is to shut off and then turn on VDD supply again (increasing the resistance in current path or decreasing the power supply to value for which the Q1 emitter-base junction will not be forward biased).

Chapter 5

Techniques to reduce Power and Leakage in Memory

5.1 Increasing Challenges of Power

- Increases the device densities and clock frequencies
- Lowering the supply voltage and transistor threshold voltage

5.2 Supply Voltage Reduction (Voltage Scaling)

- Scale both Vdd and Vth to maintain performance
- Quadratic reduction in supply voltage; cubic reduction of power
- This equation deviates when Vdd reaches sub threshold voltage level (vdd=vth)
- Dynamic power reduction decreases... sub threshold leakage increases. puts limit on scaling.

5.3 Power Saving Features in SRAM compilers

- Memory Architecture

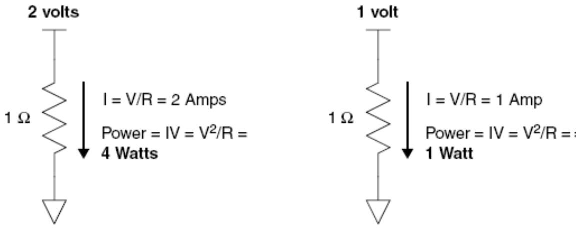


Figure 5.2.1: Voltage Scaling

- Periphery Device Selection
- CM/Bank/CD Selection
- Standby Power Management Advance Modes
- Dual Rail
- Dynamic Voltage Frequency Scaling (DVFS)
- Write Assist
- VDDP OFF

5.4 Low Power design Techniques

Table 5.4.1: Low Power Design Technique

Dynamic Power	Leakage Power	Design	Architectural	Process Technology
Clock gating Variable frequency Variable power supply Multi Vdd Voltage islands DVFS	Multi Vt Power gating Back (substrate) bias Use new devices-FinFet, SOI	Multi Vt Clock gating Power gating Multi Vdd DVFS	Pipelining Asynchronous	Multi Vt PD SOI FD SOI FinFet Body Bias Multi oxide devices

5.5 Power Saving Features Periphery VT Selection

Threshold voltage options in the periphery devices provide power and performance tradeoffs. The threshold voltage (Vt) is the voltage at which the device effectively turns on. Higher the

Vt, the higher the turn-on voltage the lower the leakage, and the slower the device. Memory compilers support two to three different Vt devices for the memory periphery. Typical devices supported for the periphery are:

Standard Vt + Low Vt: Highest performance, higher leakage

Standard Vt: Mid-range performance, mid-range leakage

High Vt: Lowest leakage, lowest performance

5.5.1 Multiple-Vt Library Cells

Some CMOS technologies support the fabrication of transistors with different threshold voltages (Vt values). In that case, the cell library can offer two or more different cells to implement each logic function, each using a different transistor threshold voltage. For example, the library can offer two inverter cells: one using low-Vt transistors and another using high-Vt transistors. In this technique, a low-threshold-voltage library is used for a first pass through synthesis to get maximum performance and meet timing goals. Thereafter, the critical paths in the design, that is, the path or paths in the design that require the highest performance, are determined. Later, areas that do not require low-threshold-voltage cells are located and low-voltage cells are swapped for high-voltage cells to reduce overall power and leakage of the design.

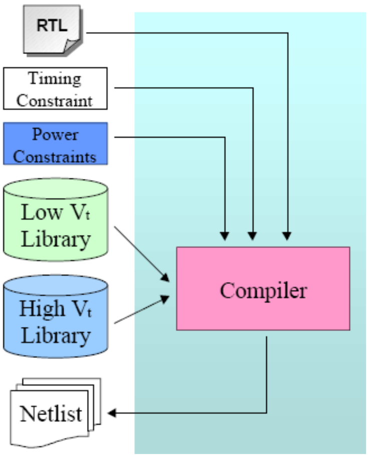


Figure 5.5.1: Multi-Vt Synthesis Flow

5.6 Power modes in Memory

1) Single Rail Option: The following diagram provides a general representation of how the three power management modes are implemented for the Single Rail option.

2) Dual Rail Option: When the Dual Rail option is enabled, the Periphery Power Rail (VDDP) and the Array Power Rail (VDDA) are physically separated. Level Shifters are added to interface between Periphery (VDDP) and Array (VDDA) voltage domains. The table below describes the valid combinations of VDDA and VDDP.

Table 5.6.1: LS,DS and SD Modes

VDDA	VDDP	Comments
Off	Off	Valid
Off	On	Permitted as per POFF table
On	Off	Not permitted
On	On	Valid

- **LS:** This mode is always available, regardless of the state of pg-enable. Leakage reduction with fine grained power gating and source biasing. Data Retention is valid for voltages greater than Vnom-10%.

- **DS:** When the DS pin is asserted, integrated periphery power gating with data retention available and the memory outputs are held low. Data Retention is valid for voltages greater than Vnom-10%. After DS is deactivated, the output Q will remain low but the output QP will be unknown until the next valid functional read/write operation.

- **SD:** When the SD pin is asserted, there is a complete shutdown (both the periphery and array are power gated), with no data retention, and the memory outputs are held low. After SD is deactivated, the output Q will remain low but the output QP will be unknown until the next functional read/write operation.

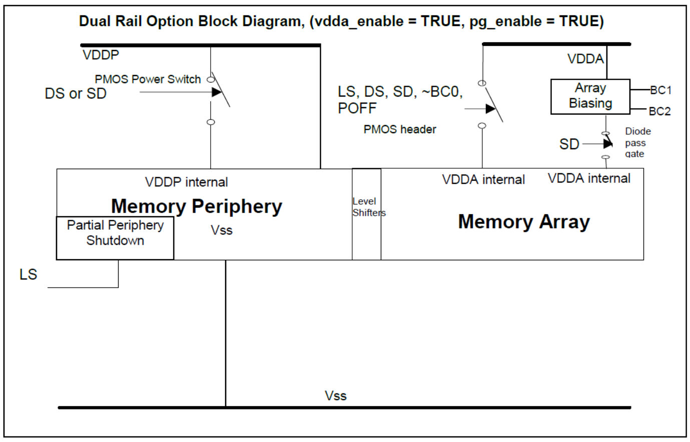


Figure 5.6.1: Dual Rail Block Diagram

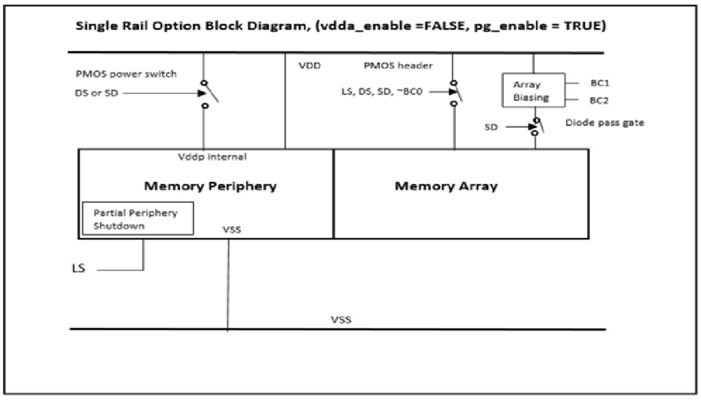


Figure 5.6.2: Single Rail Block Diagram

5.7 Leakage during Power Modes

Three major power management modes are supported in MC. LS mode is available always. DS and SD modes are only available when the power management feature is enabled by setting the compiler option flag pg-enable is TRUE, prior to memory instance generation. The LS, DS and SD pins are level sensitive to facilitate sleep mode exit without requiring a clock. In this table values shows the how much amount of power saved in LS, DS and SD mode compare to normal mode.

Table 5.7.1: Power and Leakage in LS, DS SD mode

Mode	Total Power	Array Power	Periphery Power
Normal	-	-	-
LS	3.20E+00	2.83E+00	3.00E-01
DS	1.07E+01	3.12E+00	7.60E+00
SD	1.14E+01	3.43E+00	7.90E+00

Table 5.7.2: Power and Leakage in LS, DS SD mode

Mode	Leakage
DS	80%
SD	80%
LS	25%

5.8 Power Saving Features- CD/Bank/CM Selection

These are options at compiler run time to select the instance for area/power/speed trade off. If speed is not critical, Instance can be generated for lower bank (max rpb), CD0 and Max CM option. Banking which can save significant power, Center decode for speed and Column mux for adjusting aspect ratio and optimizing performance.

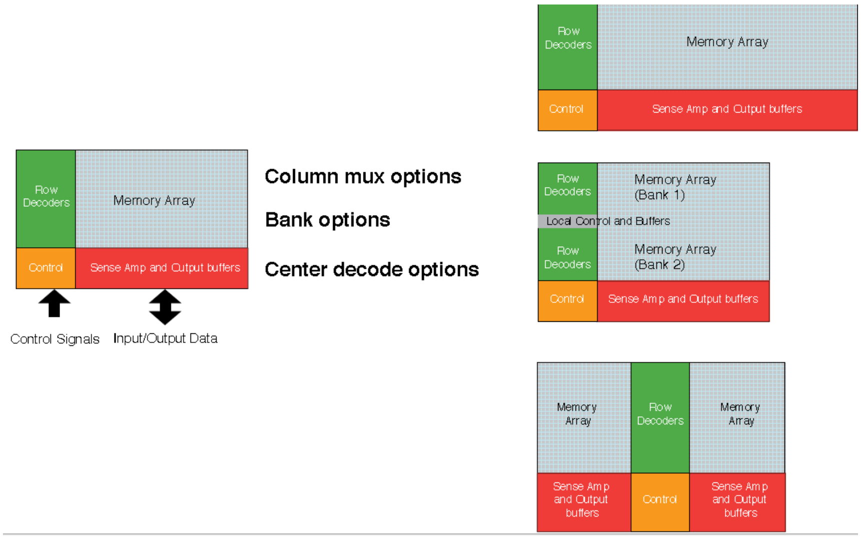


Figure 5.8.1: Different Memory Design Wide and Tallest

As we increase the number of words and number of bits power and leakage will increases. So, to reduce it we do the Column muxing and banking by column muxing we are reducing the load on column so ultimately leakage will be reduced.

5.9 Assist Techniques

SRAM cell stability is the primary concern for future technologies due to process variations like threshold voltage and supply voltage scaling etc. The increased effect of process variation and increase in parasitic resistance and capacitance in Nano scale technologies, the lower supply voltages, continuous increase in the size of SRAMs requires additional techniques such as write assist and read assist to improve the write-ability, readability and stability of SRAM memories. In modern chips memory circuits occupies about 70% of the chip area, thus reducing the supply voltage, results in significant improvement in leakage and switching power of the SoC. However a reduced supply voltage comes at the cost of degraded read static noise margin (SNM), which is to be optimized for the successful read operation of memory. If the circuit is designed to have good read SNM then due to the conflict requirements for read and write operations in a typical SRAM cell, write margin (WM) is degraded. This conflict requirement for read and write operations along with process variations in a memory chip sets the foundation for the need of write assist circuits. The below discussed techniques can be used for low power devices as these techniques can be used to gain the same or higher write margin at the lower power supply as compared to the conventional method at the higher power supply.

5.9.1 Read Assist Technique

When our SNM fails (Vdd scaling). Aadvantage is to improve reliability, affect the DC write margin. Reliability of read assist depends on SNM and Iread current. Technique: under-driving the word-line (WL) voltage to a value less than Vdd. (Lower the word line voltage). At which time need to turn on that we can find out by ewt wl driven. Read Assist Pins to control WL under-drive. WL dip voltage for wc read current: In this we are running simple read current analysis on bitcell and measures the WL voltage here we are taking wldip=Vdd- Vwl. Sa not scaled Sa input must have minimum of this threshold. (means SA input signal is close to SA threshold signal so dont need to scale the SA inputs). During read margin to increase the speed and input at sa we are dividing memory into clusters. Clusters are nothing but the small rpb and cps. Across all clusters we meet the minimum signal requirement decided by bitcell and sa. The difference signal between two bit-lines will fed to sense amplifier and it will resolve the output, but it will resolve the output correctly when sufficient difference voltage will be created between two bit-line and that will ensure by RM-control.

Table 5.9.1: Read Assist wl Driver

RA[1:0] Settigns	WL under driver %
00	0%
01	5%
10	7.5%
11	10%

Table 5.9.2: Write Assist Modes

WA[2]	WA[1]	WA[0]	Comments
0	x	x	Disabled
1	0	0	15-20% less coupling
1	0	1	capacitance design for Vddnom-10%
1	1	0	capacitance design for Vddnom-10%
1	1	1	15-20% more coupling

5.9.2 Write Assist Technique

We are using write assist technique to ensure write operation in bitcells, to enable write operation below the bitcell VDDmin, writability improvement is required in case of WL lowering technique is used for SNM issues. Also, when our Vdd nearly equals to Vt (write fail) at that point of time we are using Write assist Technique. By this we will improve reliability and it will affect the DC write margin. Reliability of wr-assist depends on DC write margin. There are different technique VDD Lowering, VSS raising, WL boosting and Negative Bit Line. At which time need to turn on the write assist control signal find out by ewt Bl driven. Key point is increase Vgs. But not go beyond the Vt of pass gate. pmos is less then nmos in capacitance.

Negative voltage determined: By snm (Apply less then vss to bitline (ewt bl driven)). WA capacitance design, as RPB increase capacitance size increases.

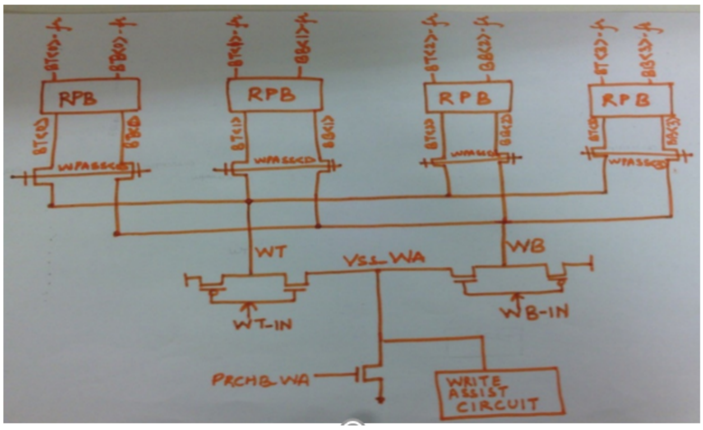


Figure 5.9.1: Write Assist Schematic

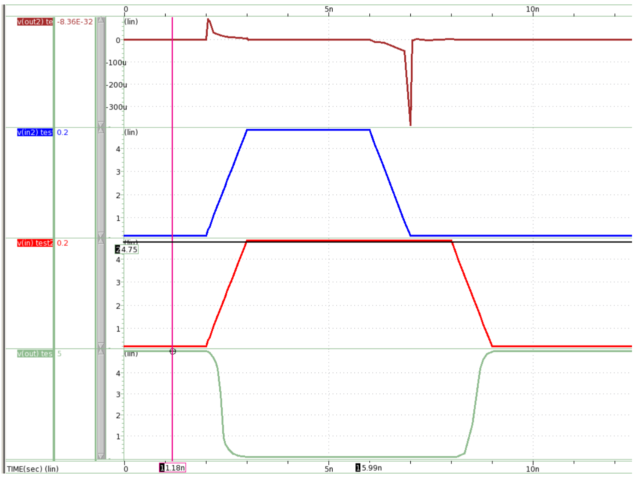


Figure 5.9.2: Write Assist Negative Bump

5.10 Level shifter (Dual Rail)

Reducing the supply voltage is not allowed below minimum array voltage (recommended by foundry). Dual Rail provides separated supply for Array and periphery. Level shifters are used at boundary of array allowing to run memory array and periphery at different voltages. If the performance is not critical then periphery voltage can be reduced till minimum logic functional voltage. And keeping array at minimum array voltage. This way both Dynamic and static power are reduced in periphery.

5.10.1 Tradeoff between Slope, Delay, and Leakage

Working of Level Shifter already explain in previous Chapters. In this we are trying to reduce dynamic leakage. The level shifter cells are treated as always-on cells, and no implicit connections are made to the PG pins. As we can see in figure that as we increase the slope, time to rise or fall will increase means delay will increasing so that dynamic leakage is also increasing.

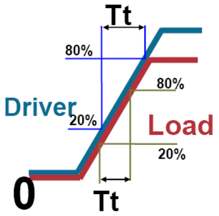


Figure 5.10.1: Level Shifter Driver and Load

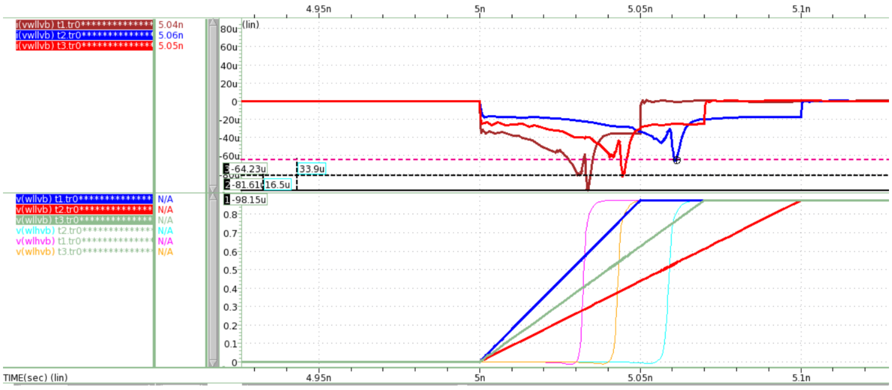


Figure 5.10.2: Leakage in Dual Rail (LS)

5.10.2 Power Reduction by reducing Array Voltage

In this method, reducing voltage of array and analyse that at what minimum amount voltage, latch would be a stable. So, it would analyse on different processes and at which process latch stability would low that is worst case analysis. The worst-case analysis observed by running 10000 MC simulation which varies the device extraction parameter. the data of this analysis would show below: -

Table 5.10.1: Pass/Fail criteria by decrease array voltage

0.6	0.75	1.40	Pass
0.6	0.65	1.37	Pass
0.6	0.6	1.27	Pass
0.6	0.57	1.18	Pass
0.6	0.55	1.10	Fail
Periphery Voltage	Array voltage	Mean	Criteria

Table 5.10.2: Pass/Fail by reducing periphery voltage

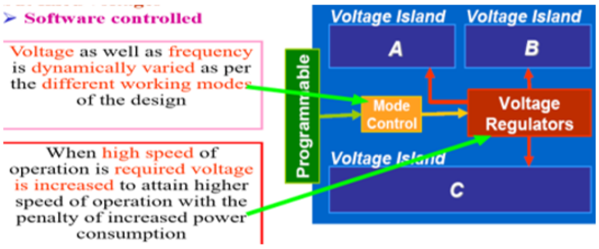
Periphery voltage	Array voltage	Level shifter
0.55	0.55	Pass
0.5	0.55	Pass
0.45	0.55	pass
0.4	0.55	Fail
0.35	0.55	Fail
0.3	0.55	Fail
0.25	0.55	Fail

5.10.3 Power reduction by reducing periphery voltage

There would be a limitation of reducing array voltage. So, in the array voltage we cannot go beyond 0.55, but it is possible to reduce periphery voltage. In the digital logic Vmin of an inverter can be Vth of Nmos or Pmos which having large threshold voltage. when the array and periphery supply voltages are different then Interface between VDD array and VDD periphery comes into picture. Level shifter have some limitation that it can shift the supply voltage up too certain level. So, analysis of level shifter shown below.

5.11 Dynamic Voltage and Frequency Scaling

DVFS is a power-management scheme which jointly optimizes performance and power consumption for energy-constrained applications. The main idea is to reduce the supply voltage (and operating frequency) when the design is not doing critical tasks. This leads to significant savings in power consumption, both dynamic and leakage. Many memory compilers support under-drive and over-drive voltage domains. This allows the memories to operate across large range of voltages, which then can be utilized to implement DVFS. One of the key reasons to push the SRAM Vmin lower is to enable DVFS to save power and energy. Typically, however, the lowest limit of Vmin is set by SRAM arrays, and hence the supply voltage of the whole system cannot go lower than the SRAM Vmin. It is imperative to push the SRAM Vmin lower so that the DVFS scheme can be more efficient. With technology scaling, it is becoming difficult to write to SRAMs even at nominal supply voltage, and the challenges energy-saving techniques become more apparent at lower supply voltages. Write assist techniques can compensate for some of the low supply voltage challenges; these techniques can significantly increase the range of applications that can utilize low-voltage operation.



➢ Software controlled

Voltage as well as frequency is dynamically varied as per the different working modes of the design

When high speed of operation is required voltage is increased to attain higher speed of operation with the penalty of increased power consumption

Figure 5.11.1: DVFS

5.12 Clock-Gating

Clock gating is one of the major dynamic power saving technique in todays low power digital circuit design. Can be specified as clock-, fanout-, and stage-based. Ensures that the enable path timing is not too optimistic. Clock tree consume more than 50 % of dynamic power. Turn off the clock when it is not needed. Gate the clocks of flops which have common enable signal. There are two types of clock gating styles available: 1) Latch-based clock gating, 2) Latch-free clock gating. Latch free clock gating: Uses a simple AND or OR gate. Glitches are inevitable and less used. Latch based clock gating: Adds a level-sensitive latch can holds the enable signal from the active edge of the clock until the inactive edge of the clock. Less glitch.

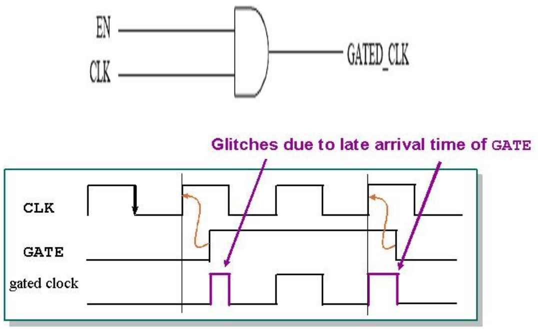


Figure 5.12.1: Latch free Clock Gating

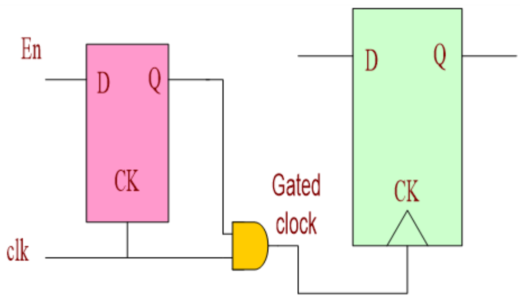


Figure 5.12.2: Latch based Clock Gating

5.13 Power-Gating Design

Power-gating reduce leakage by reducing transistor gate (G) to source (S) voltage. Circuit blocks that are not in use are temporarily turned off. Affects design architecture more compared to the clock gating. It increases time delays as power gated modes have to be safely entered and exited. How to shut down?

- Either by software or hardware
- Driver software can schedule the power down operations
- Hardware timers can be utilized
- A dedicated power management controller is the other option.
- Switch off the block by using external power supply for long term
- Use CMOS switches for smaller duration switch off

- Header switch (PMOS) or footer switch (NMOS)

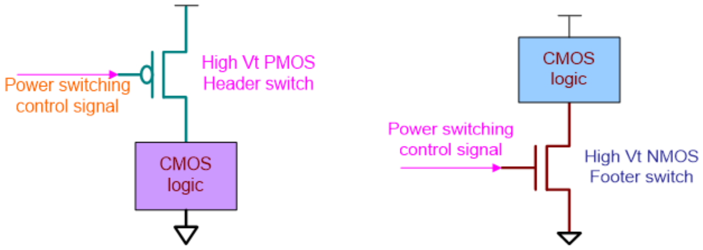

Power switching control signal
High Vt PMOS Header switch
CMOS logic
CMOS logic
Power switching control signal
High Vt NMOS Footer switch

Figure 5.13.1: Power Gating

Power-gating parameters

- **Power gate size:** Should handle the switching (rush) current its big enough not to have IR drop. Footer gates are smaller for the same amount of current (NMOS has twice mobility of PMOS)

- **Gate control slew rate:** Larger the slew larger the time taken to switch off or switch on

- **Simultaneous switching capacitance:** Refers to the amount of circuit that can be switched simultaneously without affecting the power network integrity. Rush current may damage the circuitry. Switch the block step by step.

- **Power gate leakage:** Should have less leakage and use High Vt transistors (slower switching)

Power-Gating Topologies: Fine-Grain, Coarse-Grain (column or ring based), Isolation Cells and State Retention. **Fine-Grain Power-Gating:** Fine-grain approach is that the switch is placed inside each cell. Larger area overhead and Large loading of the control signal. Add a sleep transistor to every cell. Switching transistor as a part of the standard cell logic. There is 10X leakage reduction but Creates timing issues. **Coarse-Grain Power-Gating:** Coarse-grain approach implements the switches through locally shared virtual power networks. Grid style sleep transistors. Power-gating transistor is a part of the power distribution network. Less sensitive to PVT variation. Introduces less IR-drop variation. Imposes a smaller area overhead. Switching capacitance is a major issue switch on blocks one by one, use counters logic. The global power is the higher layers of metal (Metal 5 and 6 in a 6-metal layer process), while the switched power is in the lower layers (Metal 1 and 2). Smaller area overhead. Design style: Ring-based, Array-based.

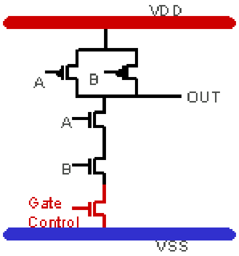


Figure 5.13.2: Fine Grain Topologies

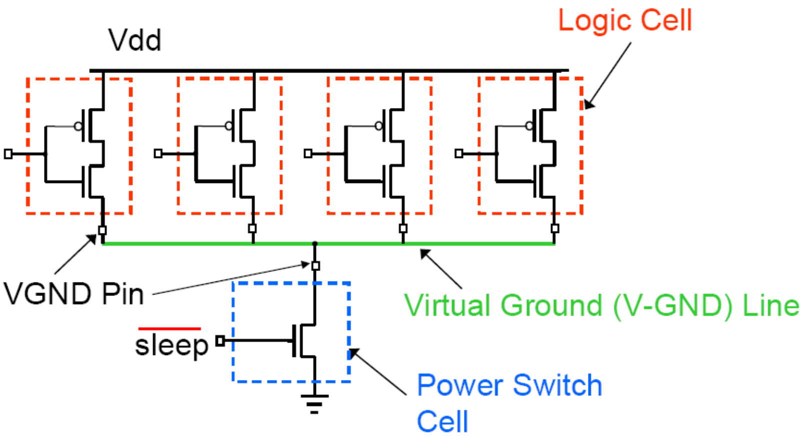


Figure 5.13.3: Cross Grain Topologies

Arrangement of Power Switches

- **Ring-based:** A ring of switches connects VDD to a switched or virtual VDD power mesh that covers the power gated block. (Power gates are placed around the perimeter of the module).**Benefits:** Simplified power grid creation for the block. Easy to implement on Hard IP without modification inside the block. **Limitations:** Large IR drop impact because no gating cell inside the voltage area.

- **Array-based:** The switches are placed in a uniform grid array inside the voltage area (Column-based methodology and Gates are inserted within the module. **Benefits:** Greater control over IR-drop distribution and the leakage through the switches. Reuse of the power routing for always-on cells. Greater scalability of a power gated blocks size. **Limitations:** Reduction of the original available cell placement area and increase of routing area, congestion and power grid complexity.

- **Isolation Cells:** When the power is shut off, each power domain must be isolated from the rest of the design, so that it does not corrupt the downstream logic. Power shutdown results in slow output from the power gated blocks. These outputs spend a significant time at the threshold voltage, causing large crowbar currents in the always on blocks. Isolation cells are used to prevent these crowbar currents. The isolation cells are placed between the outputs of the power gated blocks and inputs of the always on blocks. In a design with power gating, an isolation cell is required where each logic signal crosses from a power domain that can be powered down to a domain that is not powered down. The cell operates as a buffer when input and output sides of the cell are both powered up but provides a constant output signal during times that the input side is powered down. An enable input controls the operating mode of the cell. An isolation cell is illustrated in the following figure:

5.14 Multi-Voltage Design (Multi-VDD)

In this technique, critical paths and blocks in the design are given access to maximum voltage for the process and the specification, but voltage is reduced for those blocks that need less power. Different parts of a chip might have different speed requirements. For example, the CPU and RAM blocks might need to be faster than a peripheral block. As mentioned earlier, a lower supply voltage reduces power consumption, but also reduces speed. To get maximum speed and lower power at the same time, CPU and RAM can operate with a higher supply voltage while the peripheral block operates with a lower voltage, as shown in the following figure.

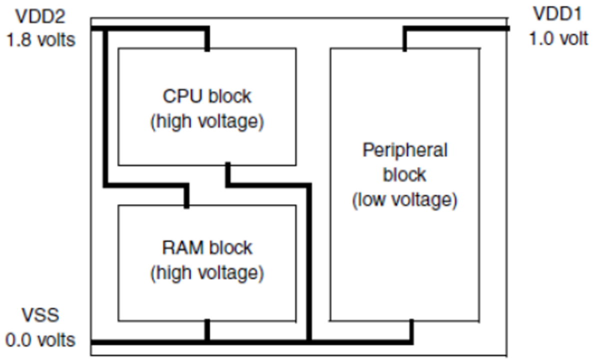


Figure 5.14.1: Multi Vdd

5.15 Adaptive Body-Biasing (ABB)

This technique is used to control the leakage current during stand-by and active mode. Reverse body bias is used for stand-by mode and zero body bias is used for active mode. ABB method reduces the leakage current exponentially. When body-to-source junction is reversed biased, the voltage threshold increases, thus reducing the leakage.

5.16 Power-Aware Placement

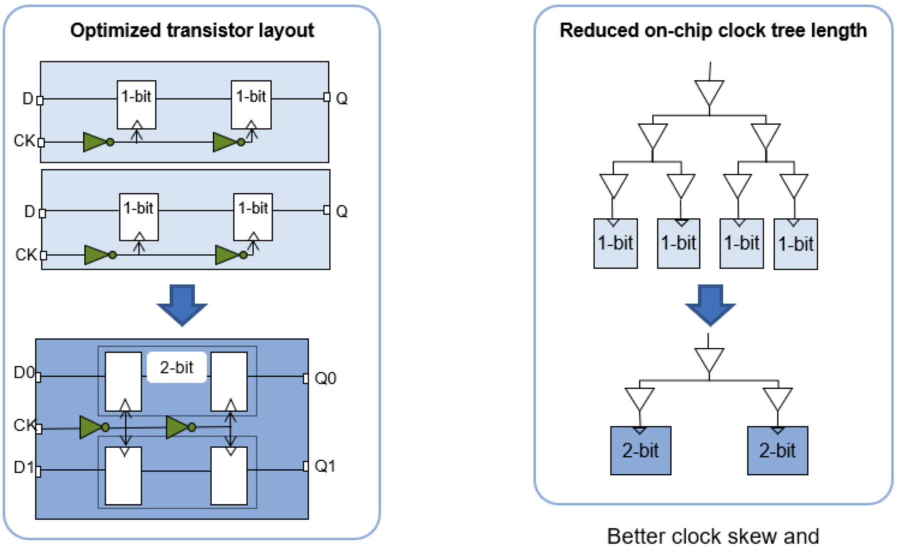
Cells with high switching activity are placed closer together. Shorter nets achieve better dynamic power savings with comparable timing. Low power placement places register closer to shorten the nets that have high switching activity to reduce leaf-level capacitance. It helps better dynamic power saving with comparable timing QoR. Low power placement can be run with or without power option in placeopt. With power, leakage power optimization is enabled in addition to low power placement.

5.17 Multibit Register Support

Dynamic power is an increasing concern particularly at advanced nodes. Designers aim to replace single-bit with multi-bit registers for better area and power savings. MB registers offer small area due to shared transistors and optimized layout simple example of 2 1-bit registers. would look when converted to a single 2-bit register in the vendor library. Having a richer variety of MB registers in the library (4-bit, 8-bit, 16-bit for instance). Will provide significant savings in power when used effectively in layout. Picture on the right illustrates the benefit you would see in layout. IC Compiler provides MB support through banking and debanking capabilities. Benefits of Merging (Banking) single bit registers to Multibit register: Smaller area, due to shared transistors and optimized transistor-level layout. Reduction in total clock tree net length and power consumption.

- **Banking:** Merging a group of single bit registers into one big register (Multi-Bit register), to reduce power especially clock power. Physically-aware multibit register identification with user guidance. Automatic connection of clock, data, and scan connections for multi-bit registers.

- **Debanking:** Split multi-bit registers into single bit registers or into smaller banks when needed. User identification of multibit registers to be split. Automatic splitting and re-connection of debanked registers.



Provides smaller area and power

Better clock skew and lower dynamic power

Figure 5.17.1: Multibit Registers

5.18 Variable Threshold CMOS (VTCMOS)

Pros: Considerable power reduction and Negligible area overhead **Cons:** Requires either twin well or triple well technology to achieve different substrate bias voltage levels at different parts of the IC.

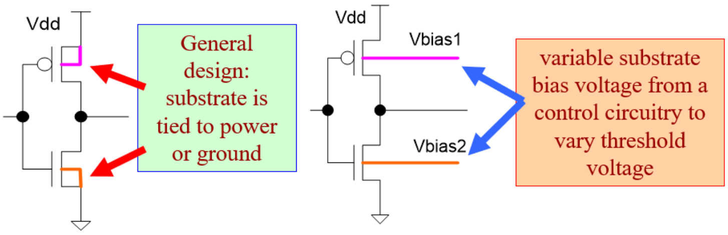

Vdd
General design: substrate is tied to power or ground
Vdd
Vbias1
Vbias2
variable substrate bias voltage from a control circuitry to vary threshold voltage

Figure 5.18.1: VTCMOS

5.19 Multiple Threshold CMOS (MTCMOS) Circuits

Use Hvt and Lvt cells. Mainly used ion CMOS full custom design and Extensively used in Power gating Called as sleep transistor.

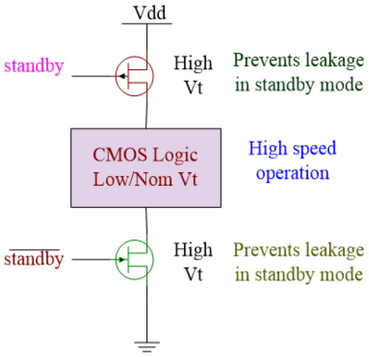


Figure 5.19.1: MTCMOS

5.20 RM Settings

Cycle and Access time are the Key timing Parameter in SRAM/RF/ROM Memories. Read and Write Operation are Primary feature of Memory. Read Margin Settings are internal setting

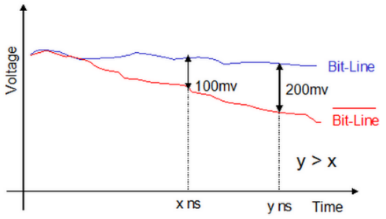


Figure 5.20.1: Voltage Vs Time Behaviour of Bit Lines Signals

which decide the Memory cycle time and access time. Write Margin settings decide the cycle time for Write Operation. These setting Decides the Performance Hence selection/tunning of these settings are very important and critical to the development of Memory. ROM has only Read Margin Setting, No Write Margin Setting required as No write operation associated. There is tradeoff between speed and robustness of Memory. These are 2 key parameters of memory. Speed = Cycle time or Access time; Robustness: Yield; These 2 parameters are contradicting Each other. Means if Memory has high speed then there be chance of low Yield (SA may not detect content of bitcell) and if we have high reliability (more Signal @ SA) then speed is sacrificed. Bitlines are precharged at VDD before WL is fired. With WL high, access bitcell develops signal on Bitline, One Bitline discharge and other complementary stay at VDD. Differential develop on Bitline This differential is sensed at SA Input. SA detects the Differential and send the Data read at output according the content of Bitcell. Sensing timing for Differential is very critical. If it is sense later then More Differential is generated and Sense Amplifier yield will improve Loss in Cycle time and Access time. If it is sense early then Less Difference Low Robustness and High Speed. This is we call Speed Vs Robustness Trade off. So Read Margin Setting plays critical Role to decide on these 2 Parameters. We have 4 RM mode to control the voltage and frequency scaling. Customer can control the RM Mode using RM pins and RME pin. See the below table for RM pin Use.

Table 5.20.1: RM Settings Description

RM[1:0]	Timing Mode
11	FAST
10	DEFAULT
01	SLOW
00	VDDMIN

Table 5.20.2: RM Settings Description

RM[3:2]	Timing Mode
11	Not Valid
10	Disables and bypass internal bias circuit
01	write timer disabled, read timer is used for write self time
00	Normal Operation

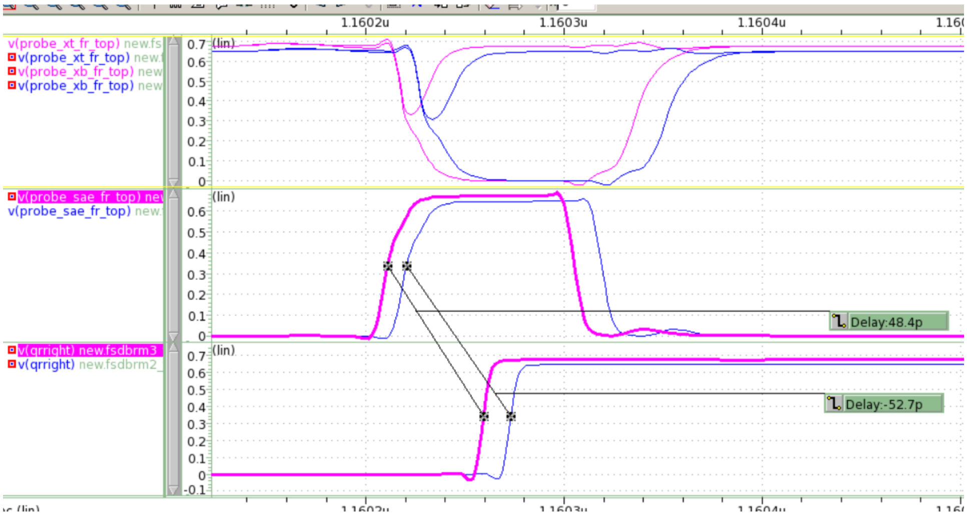


Figure 5.20.2: Read Margin Settings for Read Margin

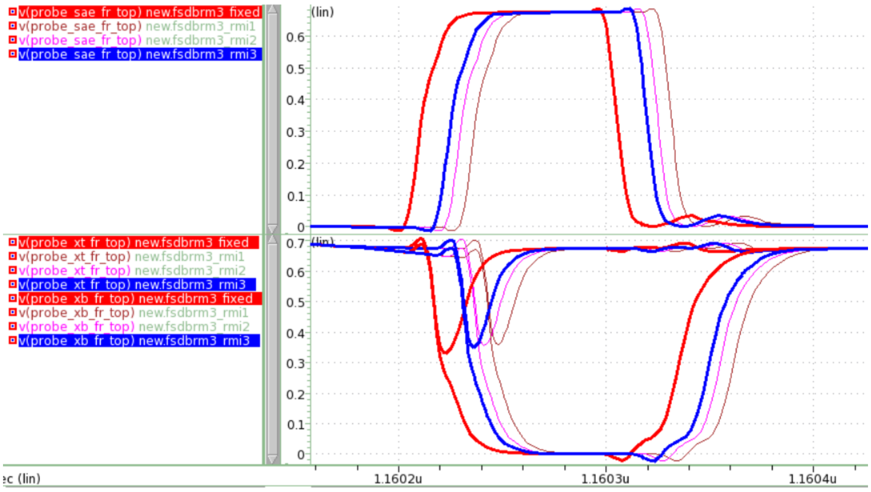


Figure 5.20.3: Read Margin Internal Settings

Chapter 6

Results

6.1 Conclusion

Table 6.1.1: Result

Power Reduction Technique	Power Benefit	Timing Penalty	Area Penalty	Architecture
Multi-Vt Optimization	Medium	Little	Little	Low
Clock Gating	Medium	Little	Little	Low
Multi-Supply Voltage	Large	Some	Little	High
Power Shut-Off	Huge	Some	Some	High
DVFS	Large	Some	Some	High
Substrate Biasing	Large	Some	Some	Medium

- Different modes of memory give Leakage reduction with fine grained power gating and source biasing.
- After separating the supply voltage for array and periphery it is possible to reduce static power at voltages. In section 4 the data would be observed when array voltage is 0.75, if array voltage reduces then power saving would be more in static mode.
- Using this decoding scheme, the best pre-decoding and post decoding scheme is decided based on the memory architecture specification in terms of number of predecoder lines, number of stack size to be used in post decoder and the granularity. which gives huge saving in clock to wordline path delay as well as better write margin (2.26 times saving)

especially at lower voltages increasing robustness of memory architecture. If access time is to be kept same, wordline to sense amp enable time can be increased to develop more differential signal, which in turn provides better yield.

6.2 Future Work

For Asynchronous Design Solution is Dynamic Power. Clock is a third to half the total dynamic power. Lets get rid of the clock. Micro pipeline: A Simple Asynchronous Design Methodology. Use Hi-k material it is sufficient for 22nm and 16nm. Whether this type of transistor structure (hi-k, metal gate) will continue to scale to the next two generation. A simple, coherent power strategy that unifies the best of DVFS, power gating, for asynchronous. Verify very complex power intent such as asynchronous.

References

[1] B.K. Kaushik, D. S. Chauhan, Sanjay Kr Singh, ”Characterization of 6T SRAM cell DRV for ULP Application ” International Journal of Computer Science (0975-8887) Volume 72, June 2013.

[2] Design and test of embedded SRAMs by Andrei S. Pavlov Dual rail Methodology and its Analysis, Synopsys Dual rail analysis for different VDDA and VDDP synopsys Memory Compiler Architecture and Characterization Flow, Synopsys Portal.

[3] Optimizing SRAM for Speed, Power, and Density Gil Winograd, Chief Operating Officer, Novelics March 2008 Sung-Mo-Kang and Yusuf Leblebici: CMOS Digital Integrated Circuits Analysis and Design.

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

[5] S. K. Singh, S. V. Singh, K.Chauhan, Tripathi,” Characterization Improvement of SNM in Deep Submicron SRAM Design” 2014 International Conference on Signal Processing and Integrated Networks (SPIN).

[6] Design and test of embedded SRAMs by Andrei S. Pavlov.

[7] Dual rail Methodology and its Analysis, Synopsys

[8] Dual rail analysis for different VDDA and VDDP synopsys.

[9] Memory Compiler Architecture and Characterization Flow, Synopsys Portal.