Enabling Ultra-Low Power High Density 6T SRAM using Assisted Design Techniques

M.Tech. Thesis

By

ABHISHEK DALAL

DISCIPLINE OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

JUNE 2019

(This page is left blank intentionally)

Enabling Ultra-Low Power High Density 6T SRAM using Assisted Design Techniques

A THESIS

Submitted in partial fulfillment of the
requirements for the award of the degree
of
Master of Technology
in
Electrical Engineering
with specialization in
VLSI Design and Nanoelectronics

by

ABHISHEK DALAL

DISCIPLINE OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY INDORE
JUNE 2019

(This page is left blank intentionally)

INDIAN INSTITUTE OF TECHNOLOGY INDORE

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis entitled **"Emerging Ultra-Low Power High Density 6T SRAM Using Assisted Design Techniques"** in the partial fulfilment of the requirements for the award of the degree of **MASTER OF TECHNOLOGY** and submitted in the **DISCIPLINE OF ELECTRICAL ENGINEERING, Indian Institute of Technology Indore**, is an authentic record of my own work carried out during the time period from July 2017 to June 2019 under the supervision of Dr. Santosh Kumar Vishvakarma , Associate Professor, IIT Indore and Dr. Ashish Kumar, Group Manager, STMicroelectronics.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other institute.

Signature of the student with date
(ABHISHEK DALAL)

This is to certify that the above statement made by the candidate is correct to the best of my/our knowledge.

Signature of the Supervisor of
M.Tech. thesis #1 (with date)
(Dr. Santosh Kumar Vishvakarma)

Signature of the Supervisor of
M.Tech. thesis #2 (with date)
(Dr. Ashish Kumar)

ABHISHEK DALAL has successfully given his M.Tech. Oral Examination held on

Signature(s) of Supervisor(s) of M.Tech. thesis
Date:

Convener, DPGC
Date:

Signature of PSPC Member #1
Date:

Signature of PSPC Member #1
Date:

(This page is left blank intentionally)

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my extreme gratitude to my research supervisor **Dr. Santosh Kumar Vishvakarma,** Associate Professor IIT Indore and **Dr. Ashish Kumar** Sr. Manager at STMicroelectronics. At many stages in the course of this research project I benefited from their advice, particularly so when exploring new ideas. Their positive outlook and confidence in my research inspired me and gave me confidence.

A project of this nature, based on both experiment and theoretical work, is only possible with the help of many people. In particular, I would like to thank **Mohammad Aftab Alam** my mentor in the world of Memory for helping me with the experiments.

The many hours that I spent at the college have been very stimulating and enriching, and thanks to the wonderful that I have been privileged to interact with. Particularly, I would like to thank **Dr. Vishal Sharma** Sr. R&D Engineer at Synopsys, **Pranshu Bisht**, and **Pallab Nath** for the great discussions that we had in the lab and for useful comments that improved my work significantly.

I would like to thank the great people of the **STMicroelectronics** community. The divine inspiration from the Tech Events and study session gave me the confidence and power to stand up against the difficulties that I faced in the course of my Project.

At last, I would like to thank **IIT Indore** for providing the essential research facilities for my research work. I gratefully acknowledge the contribution of all the faculty members and staff of Electrical Engineering for their help and support. I would also like to express my sincere gratitude to Ministry of Human Resource and Development (MHRD), Government of India for providing the Teaching Assistanship during the course of M.Tech.

(This page is left blank Intentionally)

DEDICATION

To my Parents and Teachers

(This page is left blank Intentionally)

ABSTRACT

The explosive growth of battery operated devices has made low-power design a priority in recent years. Moreover, embedded SRAM units have become an important block in modern SoCs. The increasing number of transistor count in the SRAM units and the surging leakage current of the MOS transistors in the scaled technologies have made the SRAM unit a power hungry block from both dynamic and static perspectives. The static power consumption is mainly due to the leakage current associated with the SRAM cells distributed in the array. Moreover, as supply voltage decreases to tackle the power consumption, the data stability of the SRAM cells have become a major concern in recent years.

In embedded memories, Static Random Access Memory (SRAM) is considered as a critical technology enabler for a wide range of applications. However, with the scaling of complementary metal oxide semiconductor (CMOS) technology, the conventional SRAMs have failed to achieve the trade-off between system power and performance for the applications in electronic devices such as portable smartphones, laptops, field programmable gate arrays (FPGAs), IoT edged devices. Another challenge in ULV SRAM is the statistical process variations in transistor parameters such as threshold voltage (Vth), channel length (L), and mobility. Therefore, the statistical device variability in modern SRAM design has become a major concern, as it degrades the performance, reliability, and yield of the system. Moreover, the noise generated from threshold variation, process variation, half-select issue and multiple bit errors reduces the stability of SRAM cell. Therefore, the SRAM in-stability across process-voltage-temperature (PVT) values has also become a challenging issue.

Source biasing is commonly used method for leakage reduction in deep sub-micron SRAM. However, application of such methods result into reduced stability of the SRAM bit cell. Moreover, reducing supply voltage and increasing process parameter variation put a limitation on such usage in deep sub-micron process.

Present scheme describes a method to enhance stability while applying such data retention power gating to SRAM memory core. Method improves stability cross-corner/high-leakage conditions. This scheme is realized in 40 nm CMOS technology.

In this scheme, at TT/0.85/25^{o}C we could achieve the target leakage of 0.53pa/cell and ensuring six-sigma robustness for stability

(This page is left blank Intentionally)

(This page is left blank Intentionally)

LIST OF PUBLICATIONS

Peer Reviewed International Journals

- Vishal Sharma, Pranshu Bisht, Abhishek Dalal, Maisagalla Gopal, Santosh Kumar Vishvakarma and Shailesh Singh Chouchan "Half-select free bit-line sharing 12T SRAM with double-adjacent bits soft error correction and reconfigurable FPGA for low-power applications",International Journal of Electronics and Communication, Feb. 2019
- Mahesh kumawat, Abhishek Dalal, Santosh Kumar Vishvakarma"Wave Combining Driver based serial data link transceiver design for multi standard applications", Journal of Optoelectronics.

Conferences

- Vishal Sharma, Pranshu Bisht, Abhishek Dalal, Shailesh Singh Chouhan, H. S. Jatana and S. K. Vishvakarma, "A Write-Improved Half-Select-Free Low-Power 11T Subthreshold SRAM with Double Adjacent Error Correction for FPGA- LUT Design," 22nd International Symposium on VLSI Design and Test (VDAT), 28th-30th June, 2018, Tamilnadu, India.

(This page is left blank Intentionally)

TABLE OF CONTENTS

3. Assist Circuit to Recover Stability

4. Leakage Reduction Techniques

(This page is left blank Intentionally)

LIST OF FIGURES

(This page is left blank Intentionally)

ACRONYMS

MOSFET: Metal Oxide Semiconductor Field Effect Transistor

VLSI: Very Large Scale Integration

PPA: Power Performance Area

IC: Integrated Circuit

SoC: System-on-chip

SRAM: Static Random Access Memory

FIFO: First In First Out

DRAM: Dynamic Random Access Memory

TTL: Transistor-Transistor Logic

ECL: Emitter Coupled logic

WL: Word line

CS: Chip select

BL: Bit-line

BLB: Bit-line Bar

FPGA: Field Programmable Gate Array

LUT: Look-up-table

CLB: Configurable Logic Blocks

ASIC: Application Specific Integrated Circuit

PVT: Process Voltage Temperature

WLUD: Word Line under Drive

SNM: Static Noise Margin

WM: Write Margin

RNM: Retention Noise Margin

DIBL: Drain Induced Barrier Lowering

SCL: Short Channel Effects

GIDL: Gate Induced Drain leakage

Chapter 1

Introduction

With the aggressive growth of semiconductor market, low power Static Random Access Memory has been an important research area. Performance, Power management and Area (PPA) are the three major entities of an integrated circuit (IC) in very large scale integration (VLSI) research community. All digital electronic devices are battery enabled and manufactured using semiconductor devices. Portable applications such as implantable medical devices and wireless sensor networks require ultra-low power dissipation as it requires long power back-up to perform multiple tasks associated with smart technology equipment. However, over the past few decades, the technology has scaled downed from deep-submicron to nanometer technology, which leads to higher leakage current and hence increases power consumption. Based on Moore's law, the number of transistors on a die is increasing with a higher rate with scaling the device dimensions [1]. The density of the number of transistors per unit area is also increasing, which reflects higher power dissipation. In addition, the speed (clock frequency) of IC is directly proportional to its power dissipation [2]. For circuit designers, these are two contradicting requirements. Portability requires keeping power at minimum and on the other hand, High functionality requires complex designs, such as Multi-Processor System-on-Chip (MPSoC), that demands higher power consumption

Since the performance of IC is improved with the transistor density, room for more number of applications running on a single IC also increases. However, this would increase the heat dissipation through the IC. Consequently, heat generated may damage the device or can cause undesired functioning or may lead to slow the system performance. Therefore, the cooling systems are employed to improve the heating of IC [3]. The packing of IC is also required to consume the amount of heat dissipated from the IC. As a result, due to heating, a large number of cooling fans and device packaging are required, which further increases the area and cost of the system.

Moreover, Fig. 1.1 shows how the scaling of CMOS devices change over the years [4]. Since portable devices are vastly used these days, as a result, better performance batteries are highly required. Subsequently, the devices are becoming more power

hungry due to aggressive scaling [5]. To fulfill the demand of power in portable devices, the battery size needs to be large, which is limited by the size of the device. Therefore, to provide limitless functionalities to the portable device, a better power source is needed. Unfortunately, over the years, the battery technology has not improved with the same speed as that of CMOS technology. Therefore, present batteries are not capable of fulfilling the energy requirements of the portable devices. Consequently, power reduction techniques in CMOS circuits are required to achieve less power-hungry portable devices.

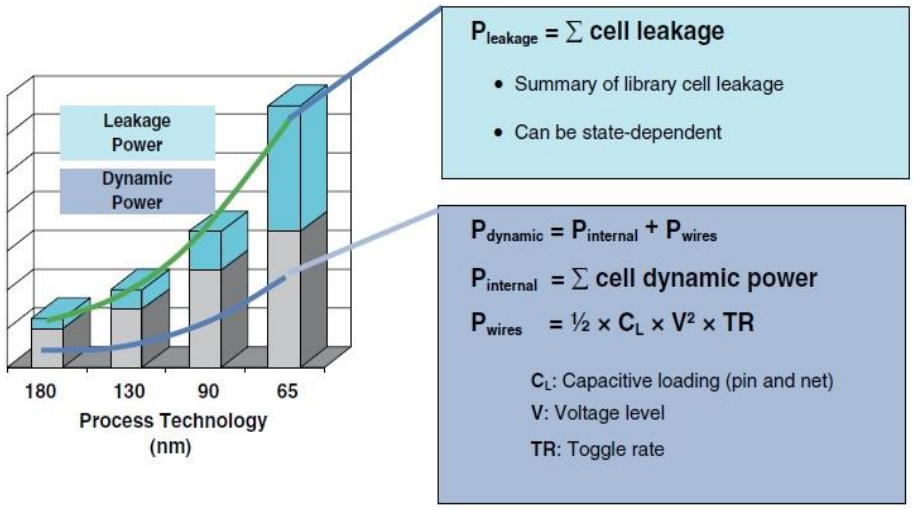


Figure 1.1 Power dissipation Vs CMOS Scaling throughout the years [4]

1.1 SRAM Design Constraints

SRAM cell is considered as the key component to store and read binary information. A typical SRAM cell is formed by a latch (using two cross-coupled inverters) and two access transistors. These access transistors provide entry to the cell during read and write operations and isolation during the hold state. A SRAM cell is designed to provide non-destructive read access, write capability and data retention until the cell is powered. Technology trends has resulted in static and dynamic power dissipation and emerged

as a primary design consideration in microprocessors. However, to keep dynamic power dissipation to a lower level, successive technology generations are depending on reducing the supply voltage. Additionally, in order to maintain performance, consistent reduction in the transistor threshold voltage is required [6], [7]. Since the leakage current increases exponentially with the reduction in threshold voltage, the static power dissipation increases to a significant fraction of overall chip power dissipation in modern nanometer processes.

Nowadays, microprocessor-controlled hand-held devices contain embedded memory which represents a large portion of the system-on-chip (SoC). These portable systems require ultra-low power circuits to utilize the battery for a longer duration. Applications of ultra-low power SRAM are extremely broad including a neural signal processor, sub-threshold processor, biomedical implants, wireless sensing, low voltage cache operation, etc. [7-10]. These applications demand careful design by maintaining the associated trade-off between power and speed. In order to adhere to intense scaling trends, SRAM design is also highly constrained. Moreover, parameter variations in MOSFET and system-level power consumption increase the design challenges. Since the impact of SRAM on the whole processing unit is very significant, modern ultra-low power SRAMs must be developed with their own trade- offs.

1.2 SRAM Architecture

A typical SRAM memory architecture along with a simple schematic illustration of column circuitry is shown in Fig. 1.2. The data storage element, or core, consists of 6 transistors (6T) memory cells arranged in an array of rows and columns. Here, each cell is capable of storing one bit of binary information. Also, each 6T cell shares a common connection with the other cells in the same row called word line (*WL*) and another common signal with the other cells in the same column called column select (*CS*). In this structure, there are 2n rows and 2m columns. Thus, the total number of memory cells in this array is 2m × 2n.

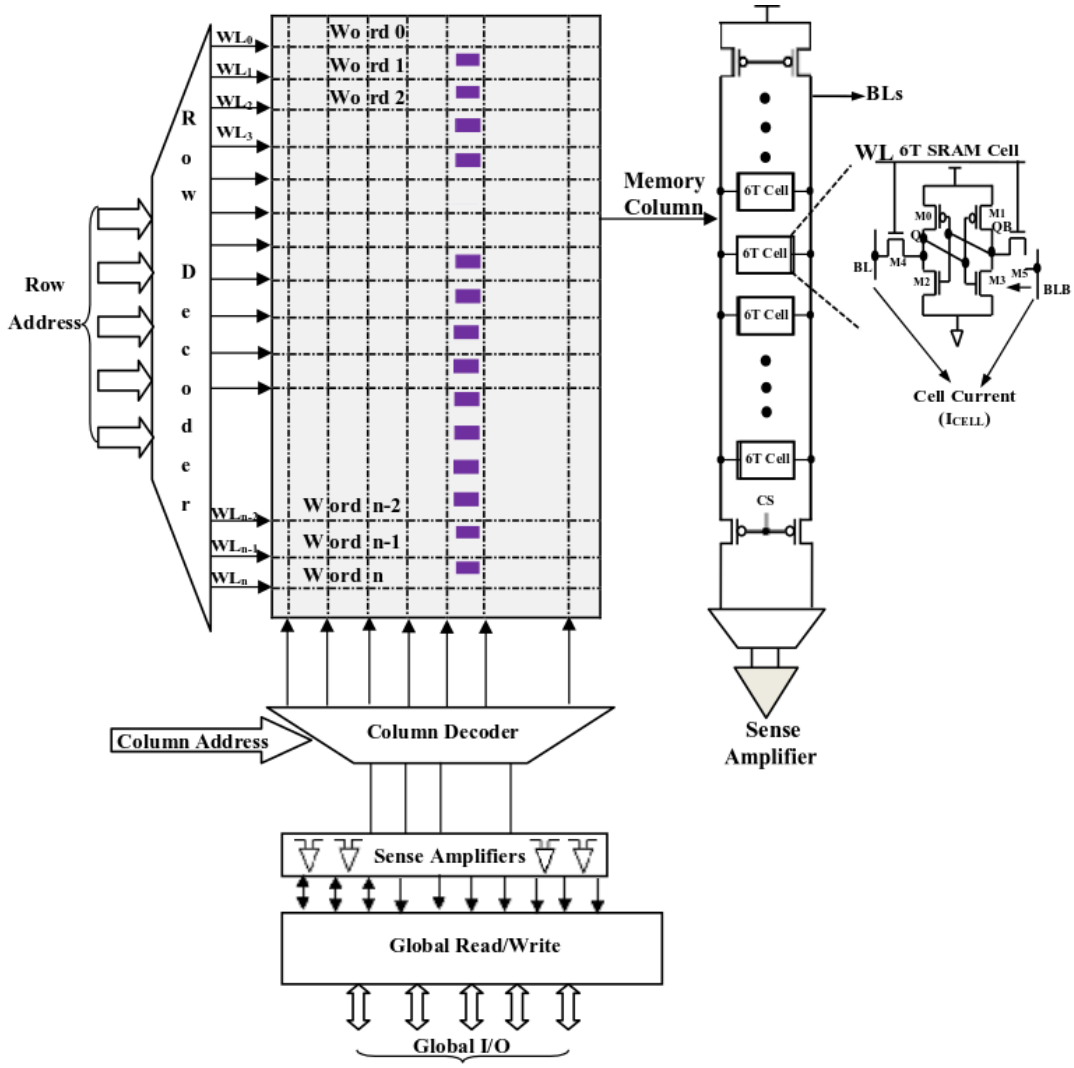


Figure 1.2. Schematic of SRAM with row-column decoders and sense amplifiers.

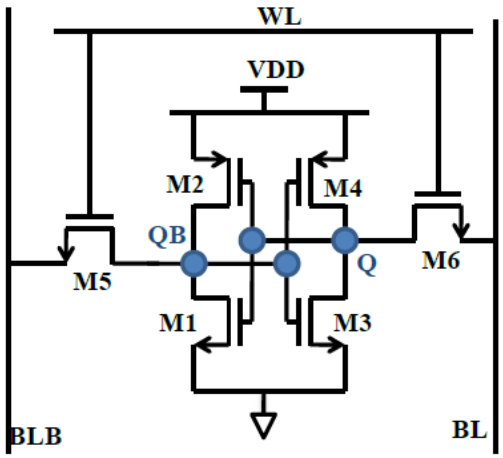


Figure 1.3 Conventional 6T SRAM

The 6T SRAM cell (as shown in Fig. 1.3) contains a pair of weak cross-coupled inverters (M1-M4), which holds the state and a pair of access transistors (M5-M6) to initiate the read or write operation. To access a particular memory cell, the corresponding column select and word line must be activated. The row and column decoders are employed to accomplish row and column selection operations, respectively. The cell is written by driving the desired value and its complement onto the bit lines and then activating the wordline (*WL)*. In addition, the column circuitry also consists of the global read/write circuitry, the bit-line sensing circuitry, and the column multiplexers as shown in Fig 1.2.

1.2.1 Types of SRAM

Static random-access memory (static RAM or SRAM) is a type of semiconductor memory that uses bi-stable latching circuitry (flip-flop) to store each bit. SRAM exhibits data remanence, but it is still volatile in the conventional sense that data is eventually lost when the memory is not powered on.

The term static differentiates SRAM from DRAM (dynamic random-access memory) which must be periodically refreshed. SRAM is faster and more expensive than DRAM;

it is typically used for CPU cache while DRAM is used for a computer's main memory. The various types of SRAM as follows :

Non-volatile SRAM (NV-SRAM)

Non-volatile SRAMs, or NV-SRAMs, have standard SRAM functionality, but they save the data when the power supply is lost, ensuring preservation of critical information. NV-SRAMs are used in a wide range of situations – networking, aerospace, and medical, among many others where the preservation of data is critical and where batteries are impractical

Pseudo SRAM (PSRAM)

PSRAMs have a DRAM storage core, combined with a self-refresh circuit. They appear externally as a slower SRAM. They have a density/cost advantage over true SRAM, without the access complexity of DRAM.

By transistor type

Bipolar junction transistor (used in TTL and ECL) – very fast but consumes a lot of power MOSFET (used in CMOS) – low power and very common today

By function

Asynchronous – independent of clock frequency; data in and data out are controlled by address transition Synchronous – all timings are initiated by the clock edge(s). Address, data in and other control signals are associated with the clock signals

In 1990s, asynchronous SRAM used to be employed for fast access time. Asynchronous SRAM was used as main memory for small cache-less embedded processors used in everything from industrial electronics and measurement systems to hard disks and networking equipment, among many other applications. Nowadays, synchronous SRAM (e.g. DDR SRAM) is rather employed similarly like Synchronous DRAM –

DDR SDRAM memory is rather used than asynchronous DRAM (dynamic random-access memory). Synchronous memory interface is much faster as access time can be significantly reduced by employing pipeline architecture. Furthermore, as DRAM is much cheaper than SRAM, SRAM is often replaced by DRAM, especially in the case when large volume of data is required. SRAM memory is however much faster for random (not block / burst) access. Therefore, SRAM memory is mainly used for CPU cache, small on-chip memory, FIFOs or other small buffers.

1.2.2 SRAM Operation

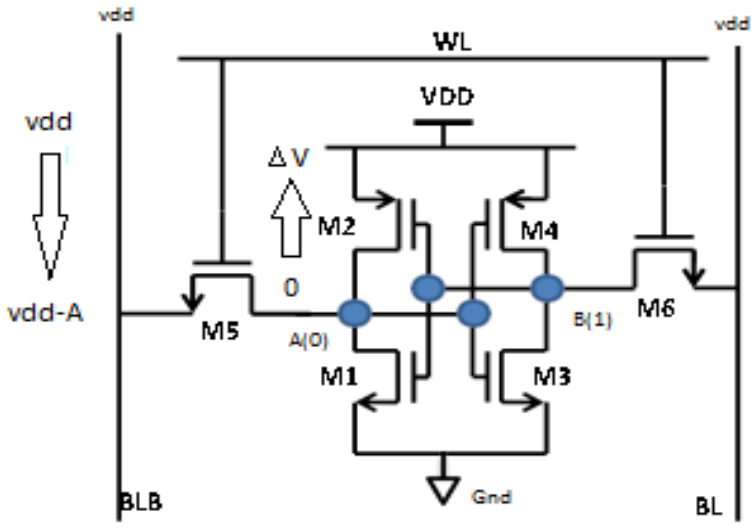


Figure 1.4 An SRAM cell status during read operation

An SRAM cell has three different states: standby (the circuit is idle), reading (the data has been requested) or writing (updating the contents). SRAM operating in read mode and write modes should have "readability" and "write stability", respectively. The three different states of SRAM as follows:

Read Operation

Fig. 1.4 illustrates the operation of the cell during a read access. In this figure, node ‘A’ carries a logic ‘zero’ and node ‘B’ carries a logic ‘one’ before the cell is accessed. Thus,

the transistors, M2 and M3, are ‘off’ while M1 and M4 are ’on’ and compensate for the leakage current of M2 and M3. In conventional design, the bit-lines (BL and BLB) are pre-charged to V_{DD} before the read operation begins.

Activation of the wordlines (WL), i.e., the gate of the access transistors, initiates the read operation. As the wordlines go high, M5 goes to saturation region while M1 operates in triode region. Owing to the short-channel effect, the current associated with M5 has a linear relationship with the voltage of the node ‘A’. Hence, these transistors behave like a resistor in this operation. Therefore, M5 and M1 form a voltage divider and raise node ‘A’ voltage by ΔV. This voltage drives the output of inverter M4-M3. To ensure a non-destructive read operation ΔV is chosen such that it does not trigger the M4-M3 inverter and node B remains at VDD over the entire cell access time. Having a constant voltage of VDD at the gate of M1 warrants the constant resistivity assumption for M4 over the access time.

DC analysis of the operation of the cell transistors is conventionally adopted to ensure the stability of the cell during the read operation. As it was mentioned before, a low enough ΔV ensures that the output of inverter M4-M3 remains constant at node ‘B’. To ensure a non-destructive read operation, the voltage level ΔV is controlled by the resistive ratio of M5 and M1. To assess the stability of the stored data during a read operation, cell ratio (CR) is defined as:

$$\textit{Cell Ratio (CR)} = (W/L)_{M5} / (W/L)_{M1}$$

where W and L are the width and length of the corresponding MOS transistors, respectively. A higher cell ratio leads to a lower ΔV and results in a more stable read operation. The concept of data stability will be treated in the subsequent chapter in detail.

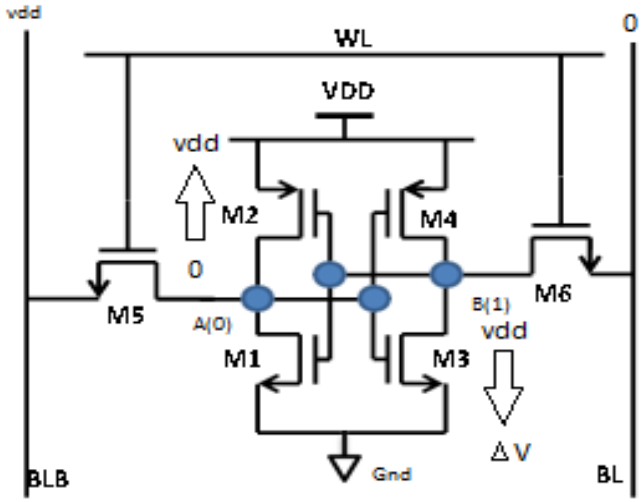


Figure 1.5 An SRAM cell status during write operation

Write Operation

In the Fig.1.5 the initial conditions of nodes ‘A’ and ‘B’ are 0 and VDD, respectively. Re-writing the old data to the cell is trivial so we concentrate on changing the data of the cell. In other words, the write operation is complete only if the voltage level on node ‘A’ and ‘B’ become VDD and 0, respectively.

As it was mentioned in the previous subsection, for an appropriate CR, the activation of the wordline cannot cause a sufficient voltage increase on node A to trigger the inverter M4-M3 if both bit-lines are pre-charged to VDD. Therefore, the write operation is conducted by reducing the bit-line associated with node B, BL, to a sufficiently low voltage (e.g., 0.) This operation forms a voltage divider comprising of M4 and M6 at the beginning of the operation. Pull-up ratio (PR) is defined as:

$$\text{Pull-up ratio (PR)} = (W/L)\ M4 \,/\, (W/L)\ M6$$

To access the voltage that appears at node ‘B’ upon activation of the wordlines in write operation, ΔV. A sufficiently low ΔV triggers the inverter M2-M1 which results in charging up node A to V_{DD}. Since node ‘A’ drives the inverter M4-M3, node ‘B’ is pulled down to zero through M3 and M4 turns off. Hence, the logic state of the cell I changed. The wordline (WL) becomes inactive after the completion of the operation.

A successful write operation can be guaranteed by choosing a proper *PR*. A lower *PR* results in a lower ΔV, and a lower ΔV is associated with higher drive at the input of inverter M2-M1. In order to achieve a low *PR* a wider access transistor is desirable, however, increasing the width of the access transistor threatens the stability of the cell during the read operation by affecting *CR*. This calls for a trade-off between data-stability in the read operation and successfulness of the write operation.

1.3 SRAM Design Trade-Offs

1.3.1 Power Consumptions in SRAM

SRAM makes up a large portion of a system-on-chip (SoC) area, and most of the time, it also dominates the overall performance of a system. In addition to this, the tremendous growth in the popularity of mobile devices and other emerging applications, necessitates the requirement of low-power SRAMs [9-10] Therefore, a robust low-power SRAM has drawn great research attention. However, a design of robust low-power SRAM faces many process and performance related challenges due to increased device variations and reduced noise margins in deep sub-micrometer technology.

One of the most effective methods to reduce the power consumption and thus extend the battery life is lowering the supply voltage. However, operations at reduced voltages degrade robustness due to low noise margins and higher vulnerability to process variations and device mismatch. This is because conventional SRAM bit-cells are generally designed to operate robustly at the nominal supply voltage. Due to the ratioed logic of the conventional 6T SRAM bit cells, if the operating conditions move away from the nominal point, the operating margins that are required for functionality start

to erode quickly. Moreover, local and global transistor variations coupled with aging effects restrict the design space for SRAM functionality, and consequently, it is becoming increasingly difficult to make SRAMs operational at lower supply voltages.

In addition, the key factors of power consumption in sub-nanometer SRAM cell are:

Dynamic power dissipation: There are three ways by which dynamic power dissipation can occur namely, switching power, short-circuit power and glitching power.

Switching power: It occurs due to the charging and discharging of parasitic capacitances across the bit-lines of SRAM. The average energy from VDD to the bit-line capacitance (*CBL*) to charge the capacitor across the storage node is equal to $CBL \times VDD^2$. Hence, the total switching power dissipated in one cycle of input pulse can be computed as:

$$Pswitching = CBL \times VDD^2$$

Short-circuit power: Short-circuit current occurs during transitions of signal when both the NMOS and PMOS of SRAM inverters are ON and there is a direct path between VDD and GND. The power dissipation arises due to this current is known as short-circuit power dissipation.

$$P_{SC} = Imean \times VDD$$

Where *Imean* is the average current passes through VDD to GND in a

Short-circuit condition.

Glitching power: Glitches are undesired signal transitions, which do not contain any useful information. Glitches can be divided into two categories namely, generated and propagated. If the input signal to a gate is skewed in time, there is a clear chance of generated glitch at the output. If a glitch arrives at the input of the gate when it is active, a propagated glitch will occur. The number of glitches in a SRAM depends upon the logic function and the number of inputs to the gates of the transistors associated with SRAM. Generally, a 6T SRAM has only two inputs, the wordline (*WL*) and the bit-line (*BL*). Hence, the glitching power occurs only due to the timing mismatch between *WL* and *BL* inputs.

Static power dissipation: The static power components become active when the SRAM is in a retention state, i.e. when SRAM is in no operation mode or it is biased to a specific state. The static power dissipation includes sub-threshold and reverses biased diode leakage currents. In addition, with each technology node, the share of leakage power in the total power dissipated by circuit is increasing. Since, most of the time, SRAM cell stay in the standby mode, thus leakage power is very important.

1.3.2 Sub-threshold SRAMs

To reduce power consumption, researchers have proposed several methods and one of the most effective techniques for power reduction in active as well as standby mode is supply voltage scaling which considerably reduces the power consumption. However, there are several issues like stability, low static noise margin (SNM) and process parameter variations [12-15]. The sub-threshold operation can be achieved by fixing the supply voltage below the threshold voltage (Vth) of the MOSFET which shows the exponential dependency of the drain current on the gate voltage. A device in the sub-threshold region exhibits higher trans-conductance as current increase exponentially with gate voltage.

The basic equation of sub-threshold current and total off current of the MOSFET is as follows:

$$I_{SUB} = I_O e^{\frac{V_{GS} - V_{THO} - \eta V_{DS} + \gamma V_{BS}}{n V_T}} \left(1 - e^{\frac{-V_{DS}}{V_T}}\right)$$

Where V_{THO} is the transistor threshold voltage for zero substrate bias and η is the sub-threshold Slope factor. Io is the off current at $V_{GS} = V_{th}$, VDS is the drain to source voltage, V_{GS} is the gate to source voltage and V_{BS} is the body to source voltage.

It is noted from the above equations that *Ids* is measured as the basic circuit design parameter. As *Ids* α *W/L*, the transistor sizing aspect ratio *W/L* is not so effective in

changing *Ids*. On the other hand, threshold voltage variation can be very effective in changing *Ids* while designing a sub-threshold SRAM circuit. Gate current due to carrier tunneling through the oxide is negligible compared to *Ids*. Also, the junction leakage current is negligible in the sub0threshold regime related to *Ids*.

The conventional 6T SRAM architectures have benefited memory design industries which targeted designing high-speed, a high-performance cache memory for high-end processors for computer applications. However, with development in new-generation bulk CMOS technologies and the requirement of ultra-low power SRAM for applications in hand-held electronic devices i.e. mobiles, laptops, FPGAs, IoT edge devices, these conventional SRAM fails to achieve the trade-off between system power and performance. Thus, the development of application-oriented SRAMs architectures working at the sub-threshold voltages has emerged significantly.

1.4. Applications of SRAM

1.4.1 Application of SRAM in FPGA

SRAM cells are the basic cells used for SRAM-based FPGA. These cells are scattered throughout the design in form of an array and mainly used to program: (1) the routing interconnects of FPGAs and (2) configurable logic blocks (CLBs) that are used to implement logic functions. SRAM-based programming technology has become the dominant approach for FPGAs because of its re-programmability and the use of standard CMOS process technology, which results in larger package density and higher speed. Due to the volatile nature of SRAM technology, SRAM-based FPGAs lose their configured data whenever power supply is switched off and need to be reprogrammed every time when the power supply is turned on. Digital devices use FPGA as a platform to implement digital systems [21]. Since customer demands longer power backup for their portable devices, research on power reduction in SRAM and FPGA has emerged significantly. FPGAs offer a short time- to-market and low design cost, which makes them gradually more prominent in the present market. However, due to their design flexibility, FPGAs do not achieve comparable area, power consumption and delay as compared to application specific integrated circuits (ASICs) [22]. This is primarily due

to the overheads introduced for re-configurability. The configurable logic block (CLB) and look up table (LUT) are the basic building blocks of FPGA as shown in Fig. 1.6(a).

It is observed from Fig 1.6 (a) that the LUTs occupy a huge amount of area in a FPGA. Therefore, to reduce the overall power of FPGA, the reduction in LUT power is essential. SRAM-based FPGAs those manufactured by Xilinx and Altera comprise the largest fraction of the overall market. These FPGAs utilize SRAM for expressing routing and core programmable computational functions, typically through the use of lookup tables (LUTs) and multiplexers as shown in Fig. 1.6(b). The SRAM-based FPGAs provide ideal prototyping medium and are widely used to integrate FPGAs in an embedded system due to the use of standard CMOS technologies, higher performance, and re-programmability. A typical logical element consists of a 6-input lookup table (LUT), multiplexers, and flip-flops [18]. Fig. 1.6(b) shows examples of a generic logic element with a 6-input LUT that can be used to implement any 6-input function. For example, the 6-input LUT requires 64 look-up values, and the multiplexer requires whether to select the '1' input or the '0' input. The control logic '*M*' characterizes, whether the digital system is sequential or combinational (sequential for $M = 1$ and combinational for $M = 0$). This configuration data must be supplied to the associated PEs before the logical elements can be used. However, the SRAM based FPGA devices are designed at different technology nodes by different FPGA manufactures [20-25].

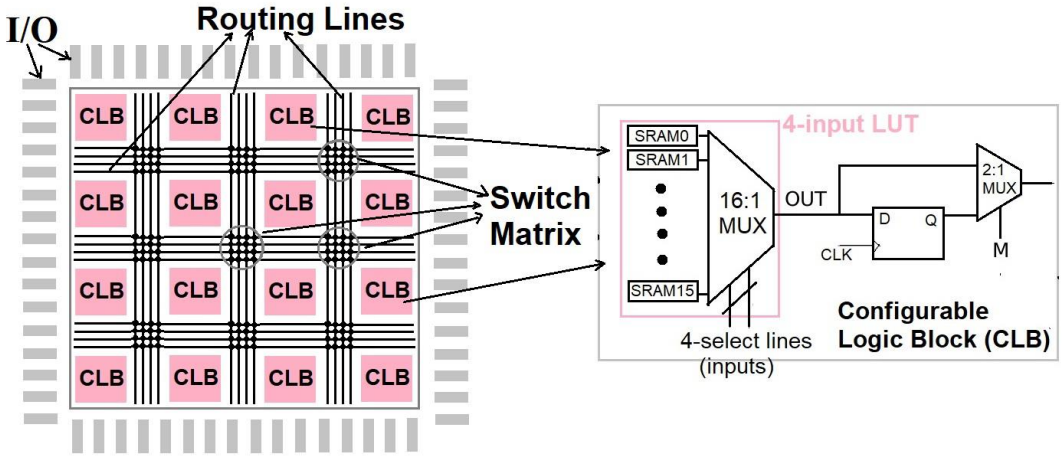


Figure 1.6. (a) FPGA logic architecture, and (b) Configurable logic block (CLB) consist of 4-input LUT, a D-FF and a 2×1 Multiplexer [20-25].

SRAM memory contributes as the major source of power consumption in an electronic device due to the introduction of leakage in the sub-threshold region. Thus, power consumption in SRAM based FPGAs and cache memories need to be revived in ULV operations.

1.4.2 Application of SRAM in IoT Devices

For the last few decades, the SRAM space has been divided between two distinct product families – fast and low-power, each with its own set of features, applications, and price. The devices where SRAMs are used need it for either it's high-speed or low power consumption, but not both. However, there is an increasing demand for high-performance devices with low power consumption to perform complex operations while running on portable power. This demand is driven by a new generation of medical devices, handheld devices, consumer electronics products, communication systems, and industrial controllers, all driven by the Internet of Things (IoT). The constraint of high standby power present in the internet of things (IoT) devices has directed towards the development of ultra-low power (ULP) systems-on-chip (SoCs) that are capable of operating at sub-threshold voltages [26]. The growth of IoT is headed in two distinct directions – smart wearables and automation. Wearables, will be serviced best by SRAMs that have a small footprint and low power consumption. At the same time, the impact of the Internet of Things will be felt in industrial, commercial, and large-scale operations, and for automating individual houses to vast factories and entire cities. SRAMs that can retain high-speed performance while reducing power consumption in a small package will offer significant value in IoT applications. It is evident that exciting times are ahead for standalone SRAM manufacturers, provided they innovate to align their products with the new-age application requirements. The key areas of innovation for SRAMs include: smaller sized chips, lower pin count, high performance chips that consume less power on-chip soft-error correction

Moreover, the IoT portable devices communicate with each other and thus require a huge amount of memory to store and process data. The memory requirement in IoT depends upon the applications related to the market. For instance, in case of huge

amount of data storage and handling information for a long period of time, a low power high-density memory is required. On the other hand, for high data transfer rate systems, a fast SRAM memory is required, where a high-speed is essential to communicate between IoT devices. Therefore, SRAM is always preferred as a cache memory due to its faster response. Moreover, the robustness of such memory systems irrespective of the variations in process-voltage-temperature (PVT) values of metal oxide semiconductor devices (MOS) and power efficiency are two of the most important design constraints [27].

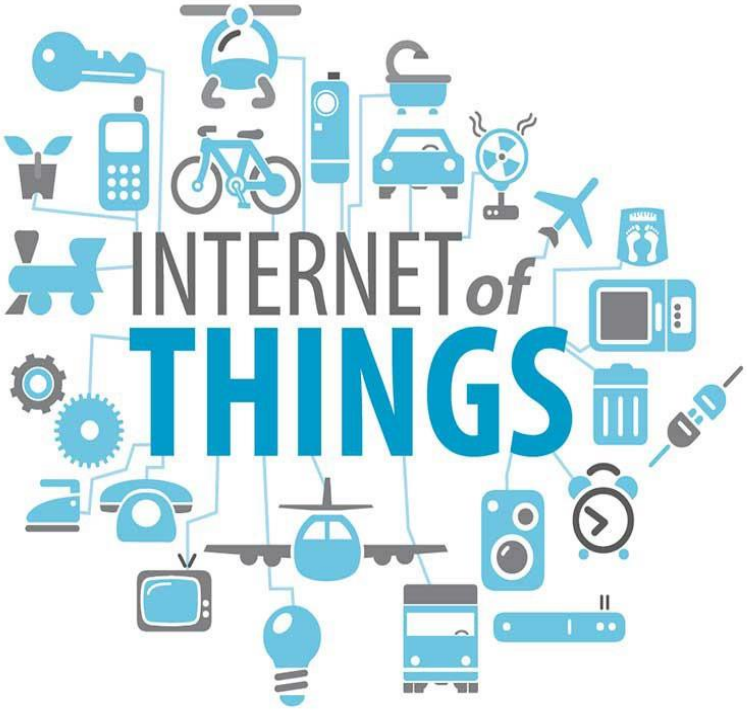


Fig. 1.7. The IoT world where all portable devices connected [26].

1.5 Challenges in SRAM Design

As the technology has scaled to nanometer regime, device density of System on Chip (SoC) has also increased. Increasing number of transistors on a SoC results in higher power density. This motivates for the reduction in supply voltage to reduce power

consumption. Along with dynamic power consumption, standby power consumption is also a key issue. Leakage power constitutes a major portion of overall power consumption of SoC due to increased device count. Reduction of leakage both during active and standby mode of operation is important. However, for battery operated and devices kept in standby mode for larger duration of their operational life, leakage power becomes a key consideration. Hence, this is desired to keep SoC at minimum supply voltage (Vmin) during standby mode of operation to minimize leakage power. This Vmin is limited mostly by the retention Vmin of SRAM. Retention Vmin of SRAM is the minimum supply voltage that ensures data in SRAM to be retained.

Low supply voltage results in near sub-threshold device operation and hence causes large statistical variation in device characteristics. The 6T-SRAM cell using smallest geometry and packed with the highest density, is affected most by the statistical variation. Ensuring data retention of SRAM at low voltage becomes a challenge, putting a limitation on minimum retention voltage (Retention Vmin) for SoC. This is due to reduced retention noise margin (RNM) of the SRAM cell at low supply voltage.

1.6 Organization of the Thesis

The rest of thesis is organized as follows:

Chapter 2 presents a brief background of our work, the challenges faced by traditional state of art memory architectures, stability issues during read and write operations of SRAM at lower voltages. Results in terms of sigma Qualification for SNM and WM are discussed.

In Chapter 3, the proposed assist schemes are discussed to recover the loss in stability specifically for SNM and WM for proper read and write operation. Simulation results are discussed in terms of SNM variation with WLUD circuit and sigma Qualification across PVTs.

Chapter 4, presents the different leakage current mechanism and proposed leakage reduction architectures for low power SRAMs. Results have been discussed for

different proposed schemes with the gain in Leakage current and improved RNM sigma Qualification across different process, voltage and temperatures.

Finally, Chapter 5 delivers the conclusion on the thesis work and suggests directions for future research work.

(This page is left blank Intentionally)

(This page is left blank Intentionally)

Chapter 2

Need and Challenges for low Voltage/Leakage Operation of SRAM

2.1 Introduction

SRAM is the most common embedded-memory option for CMOS ICs. As the supply voltage of low power ICs decreases, it must remain compatible with the operating conditions. At the same time, increasingly parallel architectures of such low-power systems demand more on-chip cache to effectively share information across parallel processing units [28]. Supply voltage scaling improves the energy consumed by SRAM and dramatically reduces its leakage power. Achieving low-voltage operation in SRAM faces a confluence of challenges, originating from process variation, and related to bit cell stability, sensing, architecture, and efficient CAD methodologies. The trend toward increased quantity of embedded SRAM in scaled technology compounds the specific need of SRAM in low-power systems. Integrating more memory on chip provides an effective means to use silicon because of memory's lower power density, layout regularity, and performance and power benefits from reduced off-chip bandwidth. As a result, the ever increasing integration of embedded SRAM continues. Low-power systems benefit from SRAMs that function at very low voltage in the state of the art, but such design solutions of low voltage SRAM significantly impact area and performance. Reducing this area overhead and further improving the metrics of energy per accessed bit and leakage power will enable new opportunities for low-power electronics in mobile platforms. Wearable electronics, portable medical monitors, and implantable medical devices are some of the applications requiring the storage of significant quantities of information (e.g., patient data)[29][30], low-access energy caches, and a long operating lifetime from a battery Consequently, SRAM power dissipation is becoming widely recognized as a top- priority issue for VLSI circuit design. The SRAM instability at various process- voltage-temperature (PVT) corners at reduced supply voltages has also came as a challenging issue to take care of. It is observed that the standby power and cell stability are the key concern in sub-threshold

SRAM architectures to improve reliability, yield and susceptibility of portable electronic devices for application in medical equipment's, field programmable gate arrays (FPGAs), and IoT edge devices. Moreover, in these portable devices, at most of the. Moreover, in these portable devices, at most of the time SRAM appears to be in hold state, which contributes to more and more leakage power. The cell stability is also a foremost concern in sub-threshold region. The noise generated from threshold variation, process variation, half select issue and multiple bit errors, reduces the stability of SRAM cell. Other challenges for SRAM include V_{DD} min, leakage and dynamic power reduction. However, stability has long been a major concern for SRAM architectures [31-33]. Low voltage operation and increased process variation caused by random dopant fluctuation (RDF) & line edge roughness (LER) have been shown to degrade the stability and performance of SRAM, and may lead to functional failure. Aggressive power reduction can be achieved by sub-threshold operation. However, operation at these reduced voltages degrades robustness, due to depleted noise margins and higher susceptibility to process variations and device mismatch.

2.2 Power/Energy Consumption in SoC

In past years, the most serious concerns for the VLSI designer were performance, cost, and reliability. Recently, however, this paradigm has shifted. More specifically, reducing power and/or energy consumption has become one of the most important themes in SoC design. The driving factors includes the popularization of portable electronic devices, raising demand for reliable and stable computer systems, Worldwide environmental destruction. One of the biggest factors which motivates the need for low power SoC is the popularization of portable electronics. The typical power consumption for a portable multimedia terminal is around the range of 10-50 [W] when employed chips are not optimized for low-power.

Therefore, it is clear that the power consumption has a strong impact on a value of the portable electronic products. The second need for low power comes from a strong pressure for designers of high-end products to reduce their temperature. The Mean Time to Failure (MTTF) [37] of aluminum interconnects exponentially decreases as the

temperature of a chip increases. Therefore, cooling down the chip temperature is essential for a reliable and stable operation of computer systems. Contemporary performance-optimized microprocessors dissipate as much as 15-50W at 100-200MHz clock rates. The leakage power issue makes this situation worse, because the leakage power increases exponentially as the temperature of the chip increases. Cost for cooling such chips is huge. Consequently, there is a clear advantage to reducing the power consumed in computer systems. Especially for consumer products whose sales are strongly affected by its price, lowering the power is indispensable.

So, we need to come up with innovative solutions which drastically save the energy of the electronic devices with accelerating the growth of IT population. Recently, many energy reduction techniques at various levels of abstraction, such as at device, circuit, layout, architectural, and software levels are proposed. Regarding the physical design, energy optimization techniques are well studied. However, there is much scope left to study in the system level such as architectural, algorithm, or software level. In this work, we present Circuit level energy reduction techniques which might be an essential in SoC design. Static Random Access Memory is the major contributor to the power dissipation of the digital design in System-on-Chip (SoCs). Major share of the static Power consumption is attributed to the SRAM as they occupy more than 50% of the chip area. Furthermore, in the present scenario, the state of the art SoCs require larger embedded memories (mainly SRAM) to support wide range of capabilities puts a major setback on the power, performance, and energy of the design [38-40]. The tremendous increase in the demand of low power devices like wireless sensor networks, implantable biomedical devices and other battery operated portable devices has put a challenge on the design of ultra-low power memories.

2.3 Technology scaling and Thermal Issues in SoCs

The Junction temperature depends on power consumption and thermal resistance. With manufacturing process scaled to finer geometries, power consumption in a single transistor decreases. Thermal resistance for a single transistor, on the other hand, increases due to the reduction in the transistor's geometrical size. Transistor's

temperature in a new process is thus dependent on the relative rate of changes of the two parameters. To estimate full chip temperature increase in a new process, one also has to take into account the increase in transistor density. In [41], the authors using industrial technology data and ITRS prediction for future technologies showed that the normalized temperature increase of a chip is significantly elevated when CMOS technology scaled from 350 nm to 90 nm. The estimated junction temperature of a 90 nm process CMOS chip is about 4.5 times higher than that of a 350 nm process CMOS chip [42-43]. The rapid increasing junction temperature can affect several aspects of circuit design as many CMOS circuit parameters are temperature dependent. The mobility of a transistor decreases with increasing temperature which lowers the drive current and leads to increased delays. On the contrary, transistor's threshold voltage decreases with temperature which improves transistor switching time. The performance of a transistor is therefore dependent on which of the two factors dominate. The unevenly distributed heat caused by large spatial variations in power consumption at different locations can make performance analysis difficult. Thermally induced device mismatch is a major concern in high speed and high precision IC design such as clock distribution networks, Arithmetic Logic Units (ALU), data converters, amplifiers, etc. Containing temperature and thermal gradient is also critical to the design of mixed signal and analog ICs as they are more sensitive to temperature. Sub-threshold leakage, has an exponential dependence on temperature. It has been shown that for every 30^{o}C increase in temperature, the amount of leakage more than doubles. The induced leakage in turn increases total power consumption and causes further temperature rise. If the cooling system is inadequate to remove the generated heat fast enough, the positive feedback loop between temperature and leakage will eventually cause thermal runaway and burn down the chip. Temperature is a vital factor in microelectronics system's reliability. Higher junction temperature reduces mean time to failure (MTTF) for the devices, which has a direct impact on the overall system reliability. It is reported that a small increase in operating temperature (10 – 15^{o}C) can decrease the lifespan of devices by 2 times. Many physical effects that cause reliability degradation are thermally activated processes [44, 45, 46].

Negative Biased Temperature Instability (NBTI) and Hot Carrier Injection (HCI) effects, which are strongly dependent on temperature, degrade the performance of transistors in an irreversible manner over time. These effects reduce a circuit's lifetime and cause timing violations eventually. Other failure mechanisms such as electromigration, stress migration and dielectric breakdown are accelerated by high temperature and temperature gradients and cause permanent device failures [47]. According to ITRS [6], the junction temperature of a semiconductor device must be kept at 85^{o}C or lower to ensure long term reliability.

All these factors dictate that the excessive heat generated from the circuit must be dissipated to the ambient at a reasonable speed and the circuit should be operated within a specified temperature range**.**

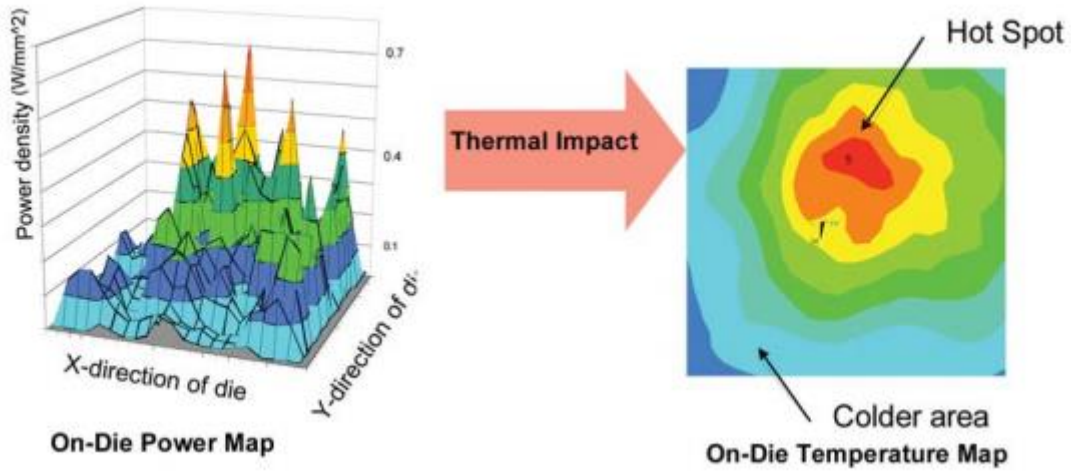


Figure 2.2: A typical power map and the corresponding thermal map [47]

In Fig. 2.2 illustrates the correlation between the power profile and the corresponding thermal map, which schematically shows the temperature rise at different locations in the chip. The non-uniformity of power consumption can cause a much higher local power density (typically referred to as hotspots). In microprocessors regions on the die with a temperature higher than 85^{o}C are usually called hotspots. Temperatures in the hotspots rise much faster than the full chip heating and can in the worst case cause severe damage to the chip.

2.4 Low Stability of SRAM at Low Voltages

A SRAM cell consist of a latch, therefore the cell data is kept as long as power is turned on and refresh operation is not required for the SRAM cell. SRAM is mainly used for the cache memory in microprocessors, mainframe computers, engineering workstations and memory in hand held devices due to high speed and low power consumption. Each bit in an SRAM is stored on four transistors that form two cross coupled inverters. With increased device variability in nanometer scale technologies, SRAM becomes increasingly vulnerable to noise sources. The wider spread of local mismatch leads to reduced SRAM reliability. However, SRAM reliability is even more suspect at lower voltages. Vmin is the minimum supply voltage for an SRAM array to read and write safely under the required frequency constraint. Therefore, the analysis of SRAM read/write margin is essential for low-power SRAMs. In recent years, research on sub-threshold SRAMs has shown the promise of SRAM design for energy-efficient and ultra-low-power applications. The most challenging issue for sub threshold SRAM is increasing reliability during read/write. A good metric for read/write margin is critically important to all kinds of SRAM designs. Moreover, the stability of the SRAM cell is seriously affected by the increase in variability and by the decrease in supply voltage.

SRAM cell read stability and write-ability is major concerns in nanometer CMOS technologies, due to the progressive increase in intra-die variability and VDD scaling. Data retention of the SRAM cell, both in standby mode and during a read access, is an important functional constraint in advanced technology nodes. The cell becomes less stable with lower supply voltage increasing leakage currents and increasing variability, all resulting from technology scaling. The stability is usually defined by the SNM [6] as the maximum value of DC noise voltage that can be tolerated by the SRAM cell without changing the stored bit.

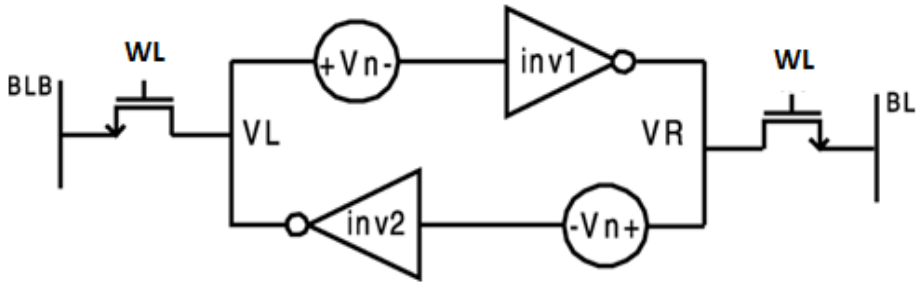


Figure 2.3: The standard setup for the SNM definition

In Fig. 2.3, the equivalent circuit for the SNM definition is shown. The two DC noise voltage sources are placed in series with the cross-coupled inverters. The minimum value of noise voltage (Vn) which is necessary to flip the state of the cell is recorded as SNM.

The noise tolerating capability of an SRAM cell largely determines the stability of the cell to retain its data. It is measured by the SNM, which in turn depends on device parameters like supply voltage, cell ratio (CR), pull-up ratio (PR), threshold voltage of the device and temperature.

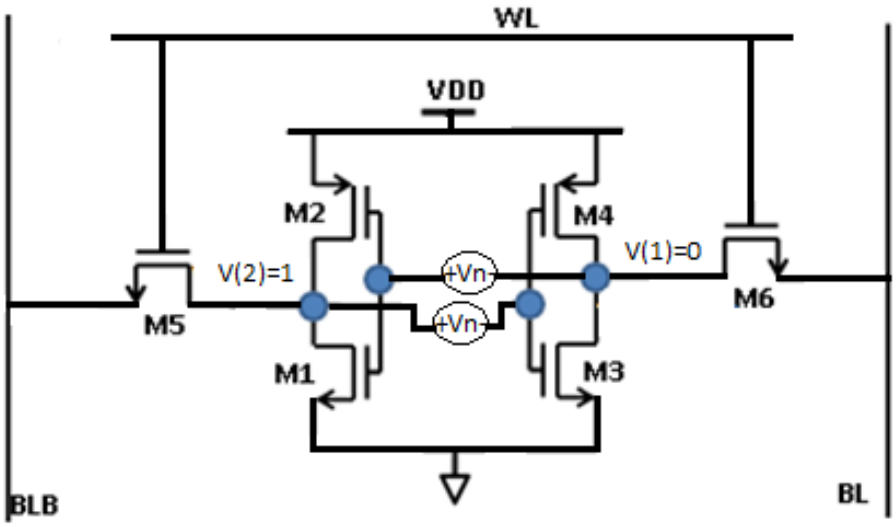


Figure 2.31 standard SNM setup in 6T SRAM cell

2.4.1 The Statistical Simulation

Monte Carlo methods are a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results [49-50]. Simplifying the definition, Monte Carlo algorithms are used for introducing random variations within the given limits to explore the corner cases of any problem.

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°C to as high as 125°C,
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 V_t), F for Fast (Low V_t) and S for Slow (High V_t).

Running the design over different PVT corners cover the environmental variations (voltage and temperature) as well as manufacturing variations (process).

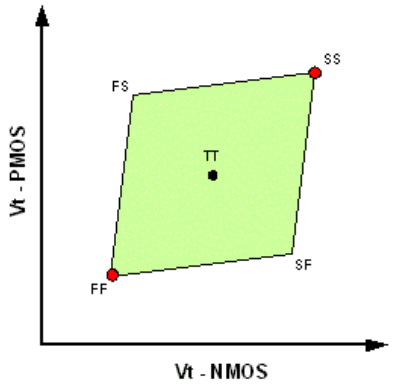


Figure 2.4 Illustrate the process corner of NMOS and PMOS

Suppose, In a design which has 1000 NMOS and 1000 PMOS and we are running this design at FS corner, considering all the 1000 NMOS are identically FAST and all the 1000 PMOS are identically SLOW. This is not true in real silicon where no two transistors are identical due to Systematic and Random variations. So even after running the design across process corners we are leaving behind the corner case where there is variations across different transistors in the same process corner. Here we need Monte Carlo which aids in introducing the randomness into the transistors by changing its V_t in different directions such that all the 1000 NMOS/PMOS are different at a time, depicting the real silicon behaviour.

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.

Global Monte: This Monte run defined in a way that the variations in this case can span over different process corners.

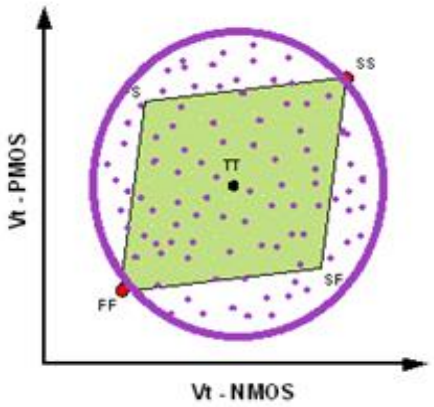


Figure 2.5. Global Monte in which the variation spreads across the process corner.

In the Fig.2.5, each dot represents one Monte Carlo run and as we can see it will spread the variation by introducing a V_t change in its every single run

Local MC: This Monte run is constrained to a 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

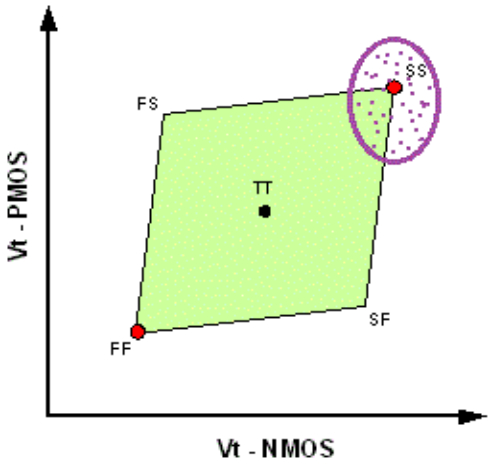


Figure 2.6. This is Local Monte as the scope of variations is limited to a particular corner

Both the methods have their own set of applications and used across industry to emulate the silicon behavior during simulation.

2.4.2 Stability Issues in 6T SRAM Cell

This most commonly used SRAM cell implementation has the advantage of very less area. However, the potential stability problem of this design arises during read and writes operation, where the cell is most vulnerable towards noise and thus the stability of the cell is affected [51-52]. If the cell structure is not designed properly, it may change its state during read and write operation. There are two types of noise margin which affects the Cell stability are as follows:

Read Static-Noise-Margin

During read accesses, the Read-SNM decreases. This is due to the reason that Read-SNM is calculated when the word line is set high and both bit line are still pre-charged high. At the onset of a read access, the access transistor (WL) is set to "1" and the bit-lines (BL and BLB) are already pre-charged to "1". The internal node of the bit-cell representing a zero gets pulled upward through the access transistor due to the voltage dividing effect across the access transistor and drive transistor. This increase in voltage severely degrades the SNM during the read operation

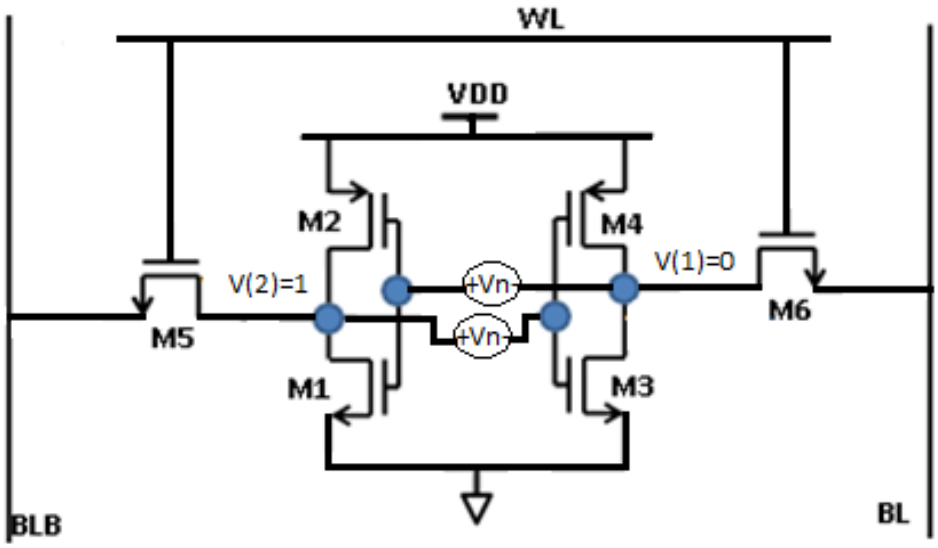


Figure 2.7: Voltage Stability Problem of 6T SRAM cell in read mode

as shown in the Figure 2.7. During the read operation, a stored "0" can be overwritten by a "1" when the voltage at node V1 reaches the Vth of M1 to pull node V2 down to "0" and in turn pull node V1 up even further to "1" due to the mechanism of positive feedback. This results in wrong data being read or a destructive read when the cell changes state.

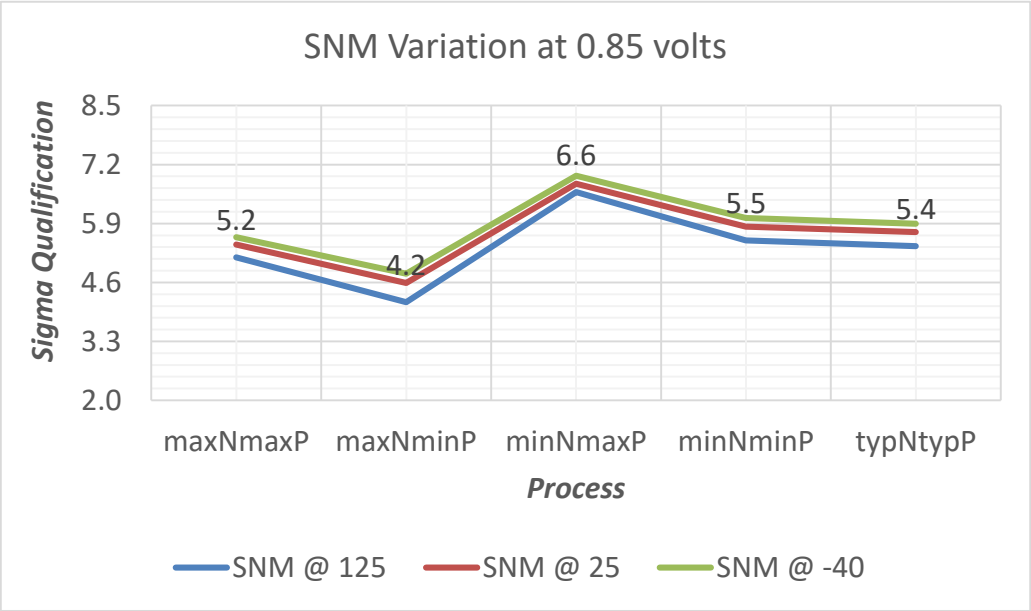
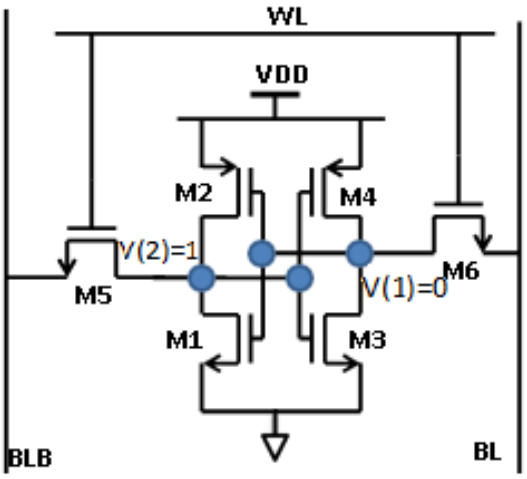


Figure 2.8 SNM variation across the PVTs

To ensure robust operation, we target six-sigma qualification and from the figure 2.8 we can observed that the SNM fails to qualify six-sigma at worst process corner FS/0.85/125℃.

Write Static-Noise-Margin

The write noise margin is defined as the minimum bit-line voltage needed to flip the state of cell. During a write operation, the input data are sent to the bit-lines (BL and BLB), and then the word lines are activated to access the cell. The bit-line that is charged to '0' pulls the node of the cell storing '1' to '0' causing the cell to flip state. Since the cross-coupled inverters have complementary data. WL is held at vdd, BL (the side storing gnd) is tied to vdd. The other bit line (BLB) is swept from vdd to gnd. The Value for BLB at which memcell contents are flipped is defined as WM.



.

Figure 2.71: Voltage Stability Problem of 6T SRAM cell during write mode.

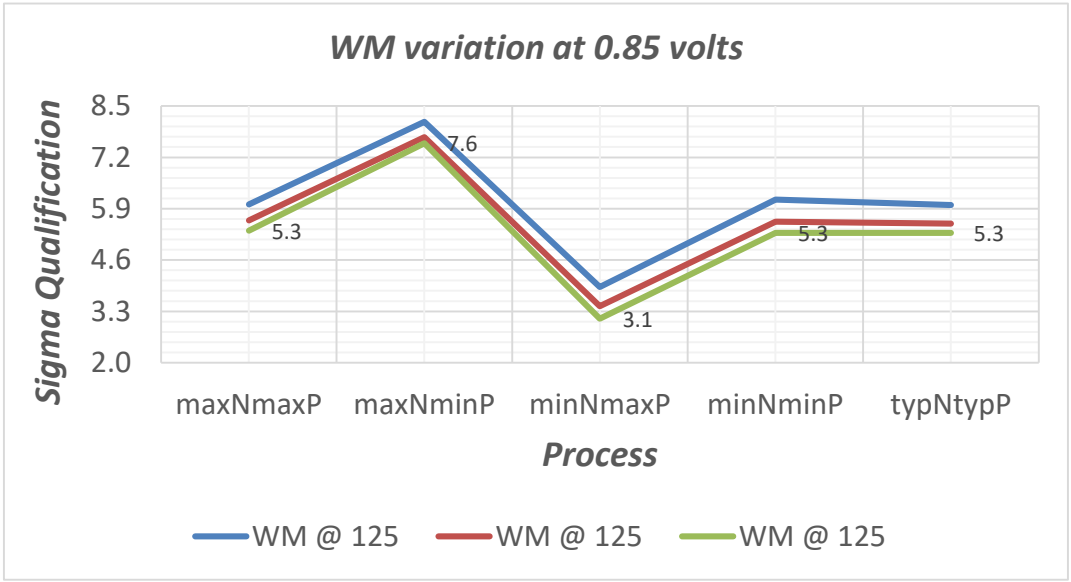


Figure 2.9 WM variation across the PVT

The 6T-SRAM cell using smallest geometry and packed with the highest density, is affected most by the statistical variation. In order to enable the SoC to work at lower voltages, all components must be ensured to function correctly at reduced supply

voltage. Ensuring correct functionality of SRAM at low voltage becomes a challenge, putting a limitation on minimum operational voltage (Vmin) for SoC. This is due to reduced static noise margin (SNM) and write margin (WM) of the SRAM cell at low supply voltage. This reduction in SNM and WM is illustrated in Fig.1 and Fig.2. To ensure a robust operation, we target a six-sigma qualification for SNM and WM.

Stand-by-mode/Retention

One of the negative side effects of technology scaling is that leakage power of on-chip memory increases dramatically and forms one of the main challenges in future system on-a-chip (SoC) design. In battery-supported applications, leakage power can dominate system power consumption and determine battery life. Therefore, an efficient memory leakage suppression scheme is critical for the success of ultra-low power design. From the Fig.2.11 of RNM we can see the variation of leakage current / stand-by-power across the process corner which is very essential for low power design circuits.

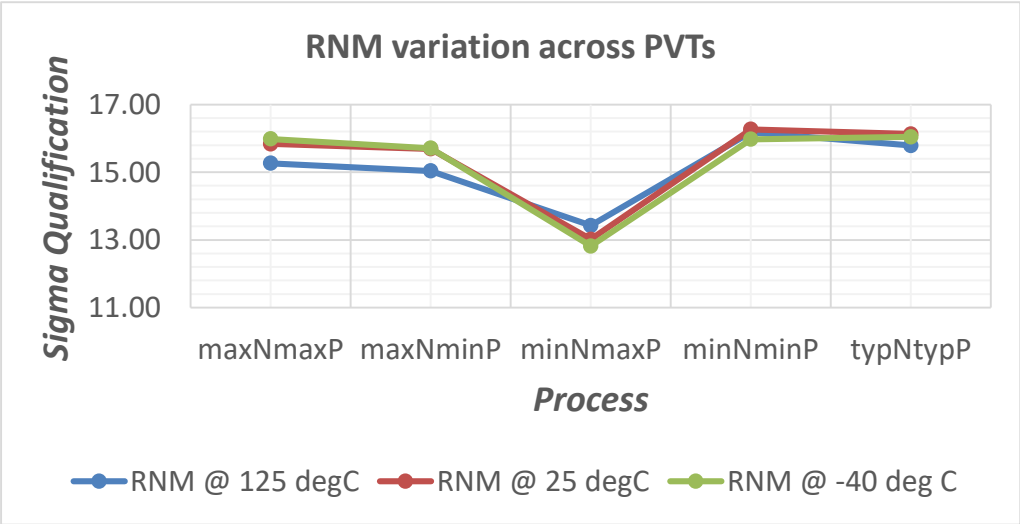


Figure 2.10. RNM variation across PVTs

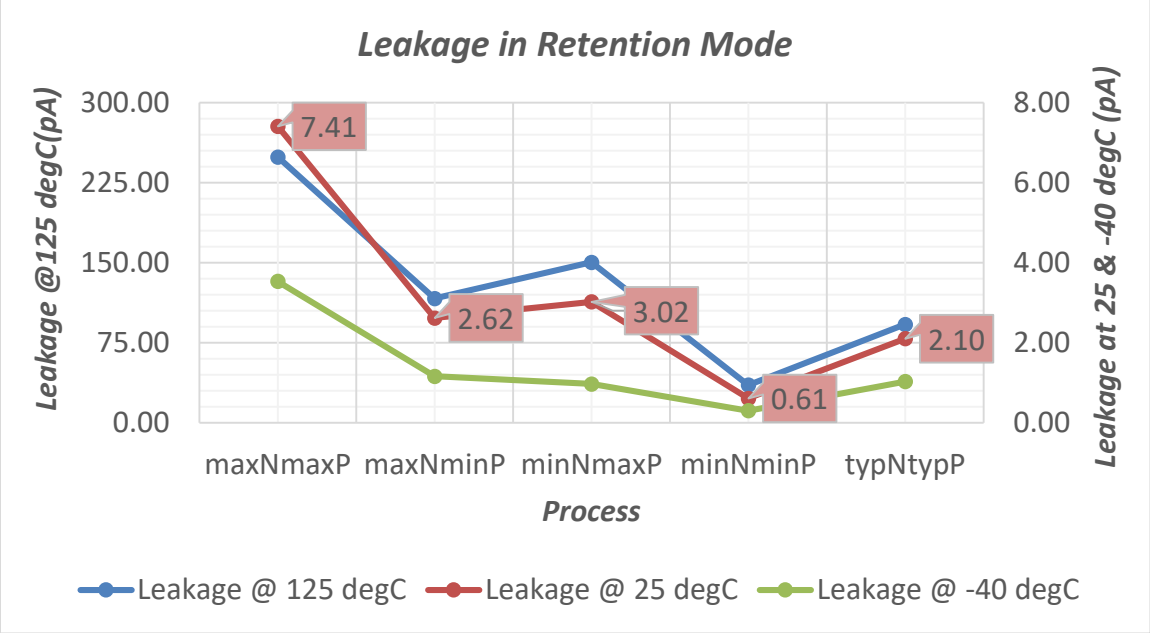


Figure 2.11. Leakage variation across PVTs

Application of SB reduces the rail-to-rail voltage of SRAM array and thus reduces the RNM. Source bias is applied only in the standby mode of operation and helps to reduce the leakage by a factor of three to five depending on the Process, Voltage and Temperature (PVT). Applying SB becomes essential for the SRAM array as leakage of SRAM is a major portion of the total SoC leakage in standby mode. From Fig.2.11 we can see that the variation of leakage across the PVTs and at TT/25°C it is 2.10 pA. This work presents a method to reduce leakage while operating under SB condition for the SRAM array by selectively reducing or removing the source bias for low RNM and low leakage PVT conditions.

2.5 Summary

In this chapter, we have discussed, the need for low voltage SRAM and challenges in low power SRAM design. Leakage and the stability are the main concerns in low power SRAMs. Energy consumption has become serious concerns in SoC design. The trend toward increased quantity of embedded SRAM in scaled technology compounds the specific need of SRAM in low-power systems. From table of RNM, SNM and WM we can conclude that the standby power and the stability are the key concerns at reduced supply voltage.

(This page is left blank intentionally)

Chapter 3

Assist Circuits to Recover Stability

3.1 Introduction

As we move into deep nanometer CMOS technology, associated challenges are also increasing. Increased integration density along with large size of SoC (system on chip) is resulting in very high power density. This requires reduction of supply voltage to make the integration feasible. Low voltage operation, however, results in near sub-threshold device operation, resulting in large statistical variation in current. Very small geometry adds to the parametric variation of the device. The 6T-SRAM cell using smallest geometry and packed with the highest density, is affected most by the statistical variation. In order to enable the SoC to work at lower voltage, all components must be ensured to function correctly at reduced supply. Ensuring correct functionality of SRAM at low voltage becomes a challenge, putting a limitation on minimum operational voltage (Vmin) for SoC. This is due to reduced static noise margin (SNM) and write margin (WM) of the SRAM cell at low supply voltage, as discussed in previous chapter. Six-sigma robustness of the design is not achieved at low supply voltage. To recover the loss in stability, specifically SNM and WM, there is need for assist schemes that can help for the correct read and write operations. Several assist schemes has been explored and implemented to provide the solution [55- 60].

The assist schemes implemented so far have used methods such as wordline (WL) lowering to improve SNM at low voltage, negative bit-line, column supply lowering and WL boosting as write assist [61]. WL lowering results in performance loss and also reduces the efficiency of write assist (WA). Schemes, such as partial suppression of WL are proposed to reduce this loss. Write assist using negative bit-line results in excessive undershoot at higher supply levels and its application is limited by reliability concerns.

3.2 Read Assist Circuit Techniques

To make a successful non-destructive read operation, one option is to reduce the strength of the pass transistor or/and strengthen the pull-up transistor during the read operation. We can improve the read margin by using one of the below mentioned read assist circuit techniques [63].

Lowered wordline voltage (LWL)

Word line voltage is reduced using an MOS device [64], which is applied to the gate to source voltage of pass transistor and this voltage level is lower than the supply voltage. Therefore, the pass transistor will be weakly driven. Reducing WL voltage helps SNM, but it degrades WM for the selected bit cells.

Cell V_{DD} boost

Supply voltage across the cell is increased above the VDD [64]. Thus, it improves the Vgs there by the strength of the pull up transistor and thus sufficiently improves read stability in read mode. However, VDD boost degrades to write margin during write mode. So, VDD boost should not be used for the selected columns during write.

Negative VSS

Reducing GND below the ground level improves read stability [64], [66]. Negative GND is the most effective of all readability assist techniques as it increases the Vgs on both the pull down and pass gate transistor by pulling the internal node holding '0' below ground. Unfortunately, this technique has a very high-energy cost for memory arrays, because GND lines have large capacitance.

3.3 Write assist Circuit techniques

We can improve the write margin by using one of the below mentioned write assist circuit techniques

Word-Line boosting

Boosting the word-line higher than the supply voltage, increases the Vgs of the access transistor and hence increases its drive strength. The increased drive strength of the

access transistor aids significantly in flipping the bit cell. The word-line boosting technique works on a row. Hence all the half-selected cells in a row are more prone to an upset due to reduction in their dynamic read noise margins.

Negative bit-line

The approach of negative bit-line based WA swings the bit-line voltage below 0. During the write operation. The increase in Vgs causes the access transistor to become stronger and hence can flip the bit. WA technique works on a column, hence all the un-accessed bit cells in the column see an increase in Vgs on the pass transistor [65]. Since the word-lines are not asserted for those cells, the increase in the pass gate Vgs is well lower than the Vt of the access transistor, and hence their DRNMs on are not affected.

Vss raising

A raised ground scheme is another way to aid the write operation. The idea is still to weaken the pull-up PMOS but in this scheme it is done by weakening the PMOS gate voltage instead of the source voltage [66]. The core ground is raised during the write operation. This WA technique also impacts the DRNMs of half-selected bit cells, if implemented globally for the whole array.

Vdd lowering

This scheme is based on weakening the pull-up device with respect to the pass-gate device. Once the pull-up device is weakened, it is easier to write a new data to the bit cell. This WA scheme is implemented using a second external lower supply which is connected via a multiplexer to the write-selected columns [67].

3.4 Read assist circuit technique

3.4.1 Lowered word line voltage read assist

WLUD is commonly used to recover SNM for the 6T SRAM cell. This results into a weaker pass-transistor (PG) in 6T-SRAM cell, resulting in a smaller cell current. This

in turn results into smaller rise of the node '0' and thus improving the noise margin of the SRAM cell. We observe that the cell is hardly four-sigma qualified without any WLUD in previous chapters. We need a WLUD read assist to ensure six-sigma robustness of the cell. This gain in SNM is associated with tradeoff between various parameters. At very first sight, cell current decreases, and thus reducing the operational speed of the memory. Also, during write operation, to ensure the stability of half-selected cells (cells sharing the same WL but not belonging to the selected column for write in a multiplexed architecture), WLUD must be applied. WLUD reduces the drive of PG and thus reducing the WM which affects the write-ability. SNM is worst at FS process corner where NMOS is fast and PMOS is slow. FS process corner associated with high temperature results in least SNM. Thus FS/125 ^{0}C is the worst PVT for SNM [68, 69, 70]. WM is worst for the SF process corner where NMOS is slow and PMOS is fast. SF process corner associated with cold (-40 ^{0}C in our case) results in worst WM. Thus SF/-40^{0}C is the worst PVT for WM in our case. Hence we need to find a mechanism whereby we achieve WLUD only around FS/125^{0}C to recover the SNM. WM and cell current should not see WLUD at their respective critical PVTs.

To ensure that WLUD is implemented only around FS/125°C, a process compensated WLUD circuit is designed as shown in Fig.3.1. Total current through N2 is a sum of currents through N1 and P1. Considering the case when process is FS, N1 and N2 are fast whereas P1 is slow. Under this condition I1 is large and providing the required WLUD. The circuit is designed to ensure the required WLUD at SNM critical PVT. Compensation through P1 is designed to ensure least WLUD at SF and SS conditions.

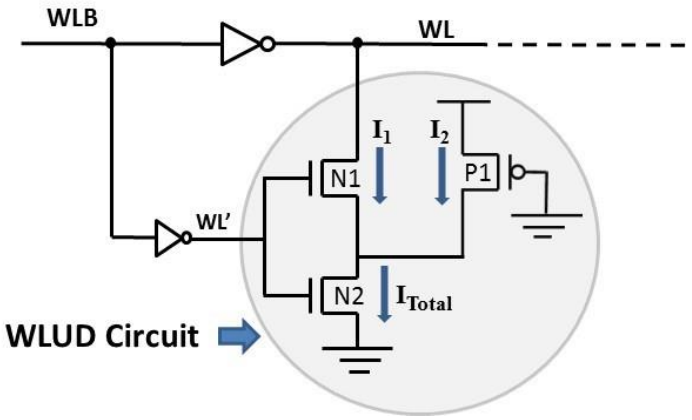


Figure3.1 Process compensated WLUD circuit

currents through N1 and P1. Considering the case when process is FS, N1 and N2 are fast whereas P1 is slow. Under this condition I1 is large and providing the required WLUD. The circuit is designed to ensure the required WLUD at SNM critical PVT. Compensation through P1 is designed to ensure least WLUD at SF and SS conditions.

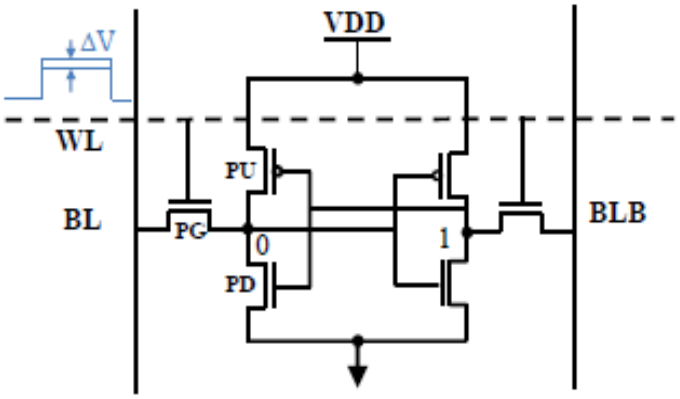


Figure 3.2. 6T SRAM Cell (ΔV Under-drive for SNM Recovery)

The variation of WLUD across PVT is shown in Fig.3.3. We observe that WLUD is suppressed at lower temperatures which is critical for write margin.

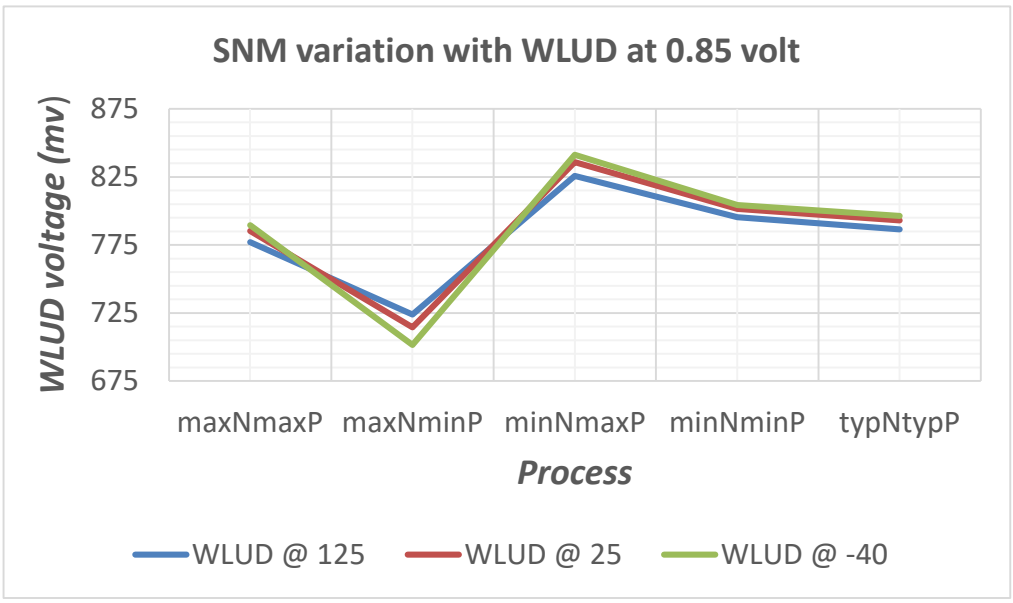


Figure 3.3. WLUD variation across PVT

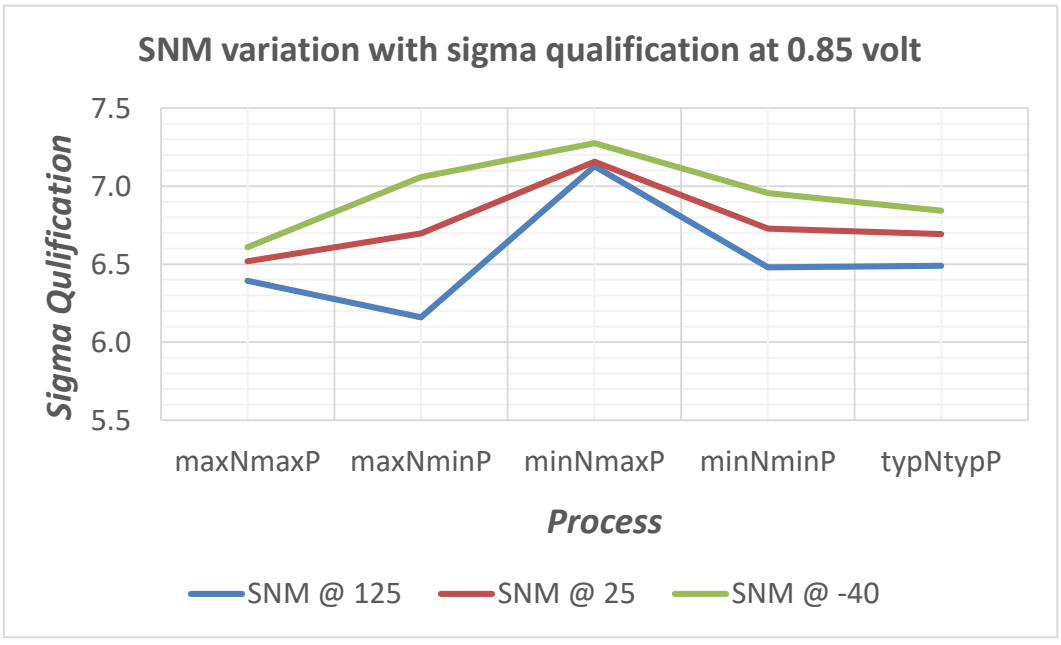


Figure3.4. SNM sigma variation across PVT

We observe from Fig.3.4 that the WLUD with respect to 0.85 volt is least in case of write critical PVT. Stability of cell increases from 4.2-sigma to 6.2-sigma, ensuring 99% yield capacity. The circuit is designed to ensure the required WLUD at SNM critical PVT. Compensation through P1 is designed to ensure least WLUD at SF and SS conditions

3.4.2 Write Assist Using Bit-line Under-drive

To design a write assist circuit, we need to understand the correlation of WLUD with write operation in SRAM. To have an area efficient architecture for moderate to high capacity memory requirement, memory-cell columns are multiplexed and the sense amplifier reads the selected column. This necessitates the application of WLUD even during write operation to ensure the stability of half-selected cells. WLUD deteriorates the WM of SRAM cell [71-72]. Hence, the required magnitude of undershoot to ensure a six-sigma robustness for write also increases. This is a major problem for the application of conventional WLUD scheme. In the scheme WLUD reduces to zero for the write critical PVT and does not put any extra penalty on the required undershoot**.**

3.5 Chapter Summary

In this chapter, we addressed the low-voltage operation of SRAM due to large statistical variations in the transistor parameters. Ensuring correct operation of SRAM at low voltages, the process compensated WLUD (word line under drive) scheme are explained. The assist scheme is designed in 40-nm standard CMOS technology. It is observed from the experimental results that the proposed cell shows a superior performance in terms of cell stability ensuring six-sigma robustness across the PVTs.

(This page is left blank intentionally)

Chapter 4

Leakage Reduction Techniques

4.1 Introduction

In this chapter, to achieve higher density and performance and lower power consumption, CMOS devices have been scaled for more than 30 years. Transistor delay times decrease by more than 30% per technology generation, resulting in doubling of microprocessor performance every two years. Supply voltage has been scaled down in order to keep the power consumption under control. Hence, the transistor threshold voltage has to be commensurately scaled to maintain a high drive current and achieve performance improvement. However, the threshold voltage scaling results in the substantial increase in the leakage current [70].

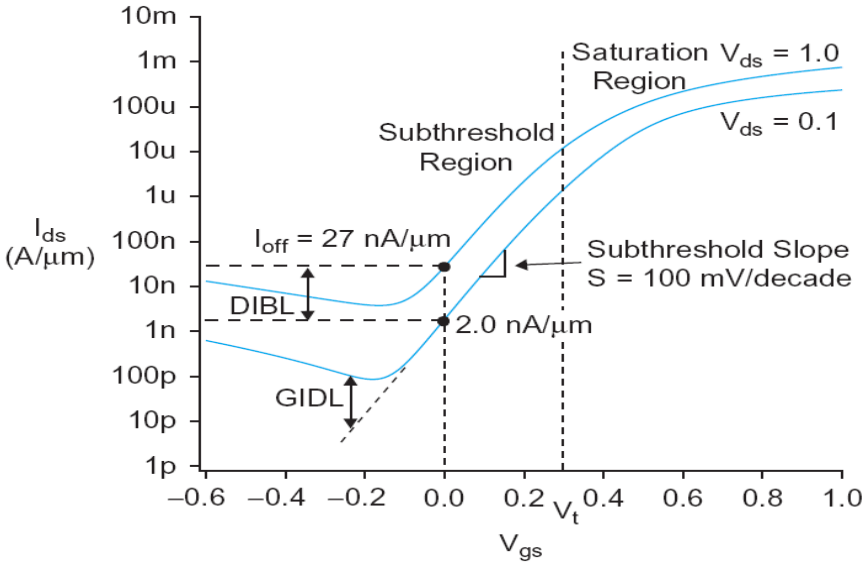


Figure 4.1 shows Log Id vs Vgs

In Fig.4.1 shows a typical curve of drain current (ID) versus gate voltage (VG) in logarithmic scale [2]. It allows measurement of various device parameters such as IOFF (off-state current), Vth (Threshold Voltage), and sub-threshold slope (St), that is, the

slope of VG versus ID in the weak inversion state. Transistor off-state current (IOFF) is the current when the applied gate voltage VG is zero. IOFF is influenced by the threshold voltage, channel physical dimensions, doping profile, junction depth, gate oxide thickness, and V_{DD}. In long-channel devices I_{OFF} is dominated by leakage from reverse-bias pn junction diodes [71]. Short-channel transistors require lower supply voltage levels to reduce their internal electric fields and consequently power consumption. This forces a reduction in the threshold voltage levels thus causes a substantially large increase in off-current I_{OFF} [71]. Other leakage mechanisms are peculiar to the small geometries themselves. As the drain voltage increases, the drain to channel depletion region widens, resulting in a significant increase in the drain current. This increase in I_{OFF} is typically caused by drain-induced barrier lowering (DIBL) or due to channel punch through currents [71]–[72]. Moreover, as the channel width reduces, the threshold voltage and the off state current both get modulated, giving rise to narrow-width effect. All these adverse effects cause reduction in threshold voltage and increase in the leakage current in deep sub-micrometer technology known as short-channel effects (SCE). To maintain a reasonable SCE immunity, oxide thickness has to be reduced nearly in proportion to the channel length but decrease in oxide thickness results in increase in the electric field across the gate oxide results in considerable current flowing through the gate of a transistor. Major contributors to the gate leakage current are gate oxide tunneling and injection of hot carrier from substrate to the gate oxide. Gate-induced drain leakage (GIDL) is another significant leakage mechanism, is proportional to gate-drain overlap area and hence to transistor width.

4.2 Transistor Leakage Mechanism

We describe three short-channel leakage mechanisms as illustrated in Fig. 4.2. $I_{junction}$ is the reverse-bias pn junction leakage; I_{sub} is the sub-threshold leakage; I_{gate} is the oxide tunneling current and the gate current due to hot-carrier injection. These are off-state leakage mechanisms, but more typically occurs during the transistor bias states in transition.

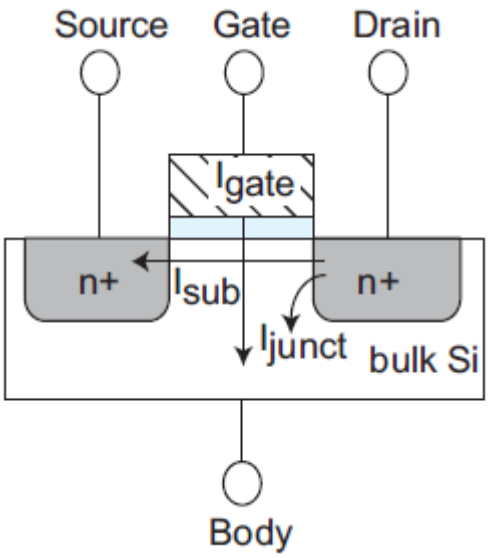


Figure4.2. Leakage current Components in MOS Transistor

Sub-threshold Leakage

The long-channel transistor I-V model assumes current only flows from source to drain when Vgs > V_{th}. In real transistors, current does not abruptly cut off below threshold, but rather drops off exponentially, as seen in Figure 4.1. When the gate voltage is high, the transistor is strongly ON. When the gate falls below Vth, the exponential decline in current appears as a straight line on the logarithmic scale. This regime of Vgs < Vt is called weak inversion. The sub-threshold leakage current increases significantly with Vds because of drain-induced barrier lowering

Sub-threshold conduction is used to advantage in very low-power circuits. It affects dynamic circuits and DRAMs, which depend on the storage of charge on a capacitor. Conduction through an OFF transistor discharges the capacitor unless it is periodically refreshed. Leakage also contributes to power dissipation in idle circuits. Sub-threshold leakage increases exponentially as Vt decreases or as temperature rises, so it is a major problem in low power applications as the chips using low supply and threshold voltages and for chips operating at high temperature.

Gate Leakage

For gate oxides thinner than 15–20 Å, there is a nonzero probability that an electron in the gate will find itself on the wrong side of the oxide, where it will get whisked away through the channel. This effect of carriers crossing a thin barrier is called tunneling, and results in leakage current through the gate [72].

Two physical mechanisms for gate tunneling are called Fowler-Nordheim (FN) tunneling and direct tunneling. FN tunneling is most important at high voltage and moderate oxide thickness and is used to program EEPROM memories. Direct tunneling is most important at lower voltage with thin oxides and is the dominant leakage component [73].

Figure 4.3 plots gate leakage current density (current/area) J_G against voltage for various oxide thicknesses. Gate leakage increases by a factor of 2.7 or more by reducing in thickness. Large tunneling currents impact not only dynamic nodes but also quiescent power consumption and thus limits equivalent oxide thicknesses tox to at least 10.5 Å to keep gate leakage below 100 A/cm^2.

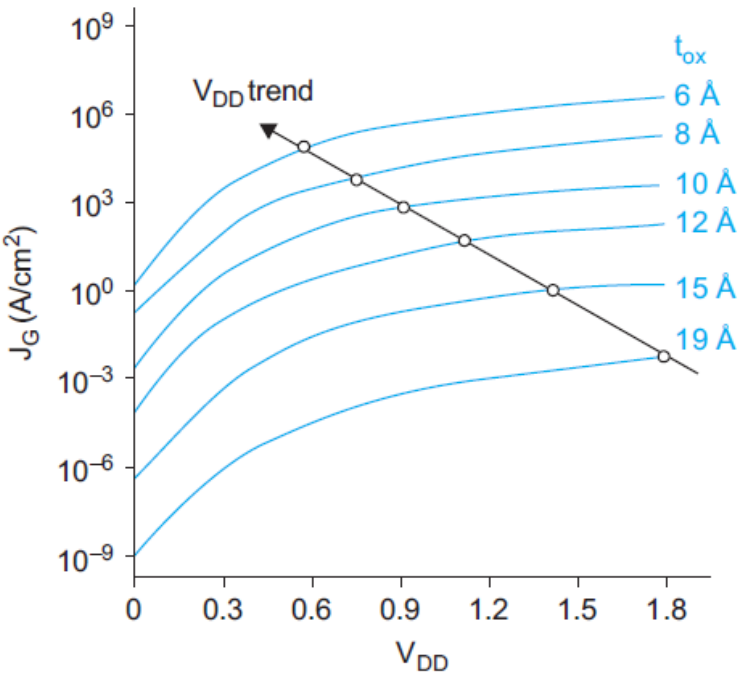


Figure 4.3 Gate Leakage current

Junction Leakage

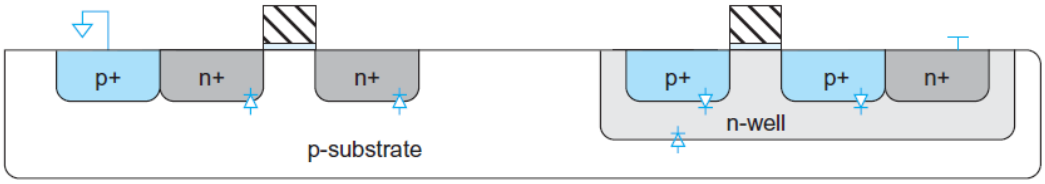


Figure 4.4 Substrate to diffusion diodes in CMOS circuits

The p–n junctions between diffusion and the substrate or well form diodes, as shown in Figure 4.4. The well-to-substrate junction is another diode. The substrate and well are tied to GND or *VDD* to ensure these diodes do not become forward biased in normal operation. However, reverse-biased diodes still conduct a small amount of current.

$$I_D = I_S \{ \text{Exp}^{VD/VT} - 1 \}$$

where I_S depends on doping levels and on the area and perimeter of the diffusion region and *VD* is the diode voltage (e.g., *–Vsb* or *–Vdb*). When a junction is reverse biased by significantly more than the thermal voltage, the leakage is just $-I_S$, generally in the 0.1–0.01 fA/um^2 range, which is negligible compared to other leakage mechanisms. More significantly, heavily doped drains are subject to *band-to-band tunnelling* (BTBT) and *gate-induced drain leakage* (GIDL).

BTBT occurs across the junction between the source or drain and the body when the junction is reverse-biased. It is a function of the reverse bias and the doping levels. High halo doping used to increase *Vt* to alleviate sub-threshold leakage instead causes BTBT to grow [73].

GIDL occurs where the gate partially overlaps the drain. This effect is most pronounced when the drain is at a high voltage and the gate is at a low voltage.

4.3 LEAKAGE REDUCTION TECHNIQUES

For a CMOS circuit, the total power dissipation includes dynamic and static components during the active mode of operation. In the standby mode, the power dissipation is due to the standby leakage current. Dynamic power dissipation consists of two components. One is the switching power due to charging and discharging of load capacitance. The other is short circuit power due to the nonzero rise and fall time of input waveforms. The static power of a CMOS circuit is determined by the leakage current through each transistor. The dynamic (switching) power (P_D) and leakage power (P_{LEAK}) are expressed as

$$P_D = \alpha f C_L V_{DD}^2$$

$$P_{LEAK} = I_{LEAK}.V_{DD}$$

where α is the switching activity; f is the operation frequency; C_L is the load capacitance; V_{DD} is the supply voltage; I_{LEAK} is the cumulative leakage current due to all the components of the leakage current. Particularly, with reduction of threshold voltage (to achieve high performance), leakage power becomes a significant component of the total power consumption in both active and standby modes of operation (see Fig. 4.5 [74]). Hence, to suppress the power consumption in low-voltage circuits, it is necessary to reduce the leakage power in both the active and standby modes of operation.

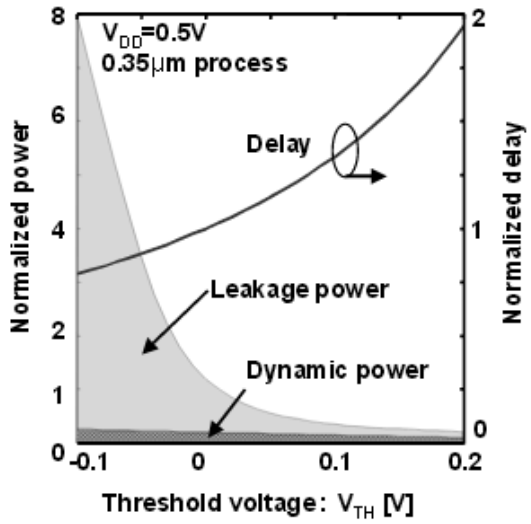


Figure 4.5 Power and delay dependence on threshold voltage (V_{th}) [74]

4.3.1 Supply Voltage Scaling

Supply voltage scaling was originally developed for switching power reduction. It is an effective method for switching power reduction because of the quadratic dependence of the switching power on the supply voltage. Supply voltage scaling also helps reduce leakage power, since the sub-threshold leakage due to DIBL decreases as the supply voltage is scaled down.

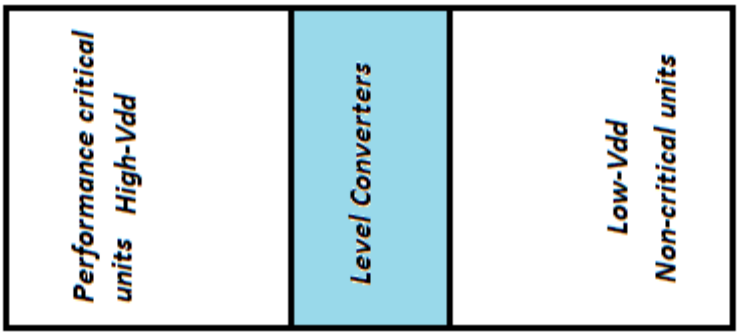


Figure 4.6 Two level multiple supply voltage scheme [75].

To achieve low-power benefits without compromising performance, two ways of lowering supply voltage can be employed: static supply scaling and dynamic supply scaling. In static supply scaling, multiple supply voltages are used as shown in Fig. 4.6 Critical and noncritical paths or units of the design are clustered and powered by higher and lower supply voltages, respectively [75]. Since the speed requirements of the noncritical units are lower than the critical ones, supply voltage of noncritical units can be lowered without degrading system performance. Whenever an output from a low V_{DD} unit has to drive an input of a high V_{DD} unit, a level conversion is needed at the interface [76]. The secondary voltages may be generated off-chip [77] or regulated on-die from the core supply [78] Dynamic supply scaling overrides the cost of using two supply voltages by adapting the single supply voltage to performance demand. The highest supply voltage delivers the highest performance at the fastest designed frequency of operation. When performance demand is low, supply voltage and clock frequency is lowered, delivering reduced performance but with substantial power reduction [79-80]. There are three key components for implementing dynamic voltage scaling (DVS) in a general-purpose microprocessor: an operating system that can intelligently determine the processor speed, a regulation loop that can generate the minimum voltage required for the desired speed, and a microprocessor that can operate over a wide voltage range. Fig. 4.7 shows a DVS system architecture [81]. Control of the processor speed must be under software control, as the hardware alone may not distinguish whether the currently executing instruction is part of a compute-intensive task or a non- critical task. Supply voltage is controlled by hard-wired frequency-voltage feedback loop, using a ring oscillator as a replica of the critical path. All chips operate at the same clock frequency and same supply voltage, which are generated from the ring oscillator and the regulator.

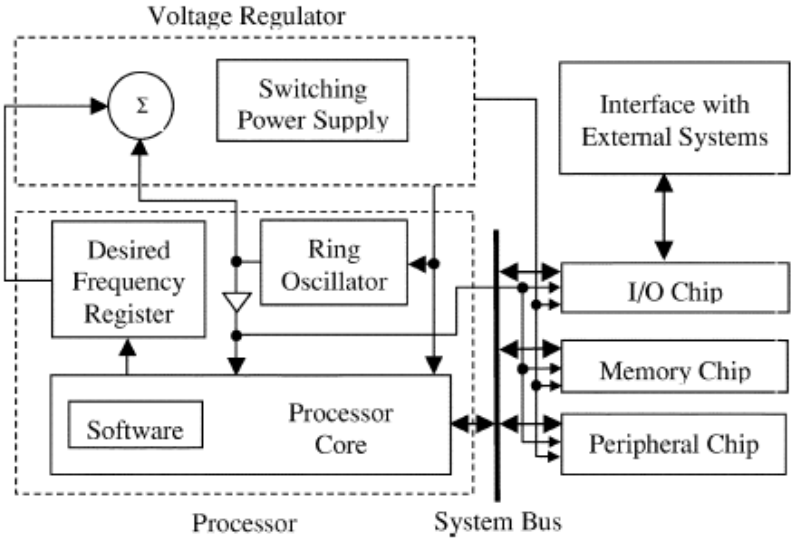


Figure 4.7 DVS architecture [88].

4.3.2 Leakage Reduction in Cache Memory using Voltage Scaling

Drowsy cache

Significant leakage reduction can also be achieved by putting the cache into a low-power drowsy mode [82-84]. In the drowsy mode, the information in the cache line is preserved. However, the line has to be reinstated to a high-power mode before its contents can be accessed. One technique for implementing a drowsy cache is to switch between two different supply voltages in each cache line [85]. Due to SCE in deep sub-micrometer devices, sub-threshold leakage current reduces significantly with voltage scaling [85]. The combined effect of reduced leakage and supply voltage gives large reduction in the leakage power.

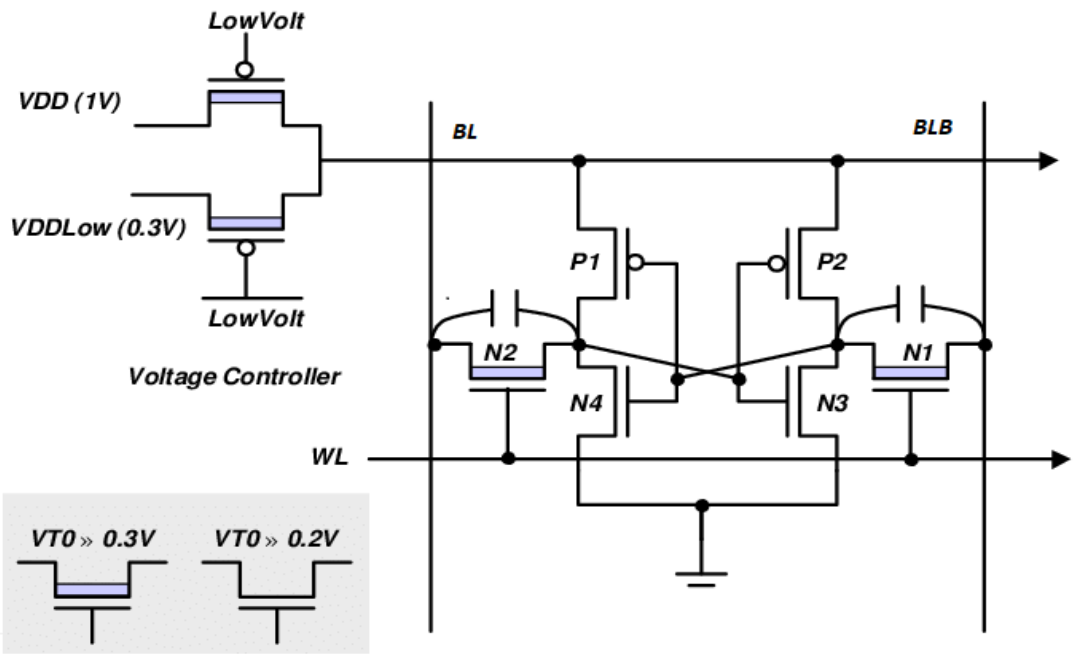


Figure 4.8 Schematic of drowsy memory circuit [85].

Fig. 4.8 illustrates the circuit schematic of a SRAM cell connected to the voltage controller. One PMOS pass gate switch supplies the normal supply voltage (VDD) (in the active mode), and the other supplies the low supply voltage (VDD Low) (in the standby mode) for the drowsy cache line. Each pass gate is a high V_{th} device to prevent leakage current from the normal supply to the low supply through the two PMOS pass gate transistors. A separate voltage controller is needed for each cache line. By scaling the voltage of the cells to approximately 1.5 times of V_{th} the state of the memory cell can be maintained.

4.3.3 Standby Leakage Control Using Transistor Stacking

Leakage Power Dissipation

The leakage current in cache memory is of major concern in deep sub-micrometer technology. The power dissipated in an SRAM cell is mainly contributed by the leakage current. This is because most of the time, a major part of the cache memory remains idle except the row, which is accessed. A six-transistor SRAM (6T-SRAM) cell consists of pairs of PU, PD and PG devices as shown in Fig.4.9. Major components of Leakage at moderate and high temperature for a 6T-SRAM cell are illustrated in Fig.4.9. I1, I2 and I3 representing the sub-threshold leakage while I3 and I4 represents the major gate leakage components.

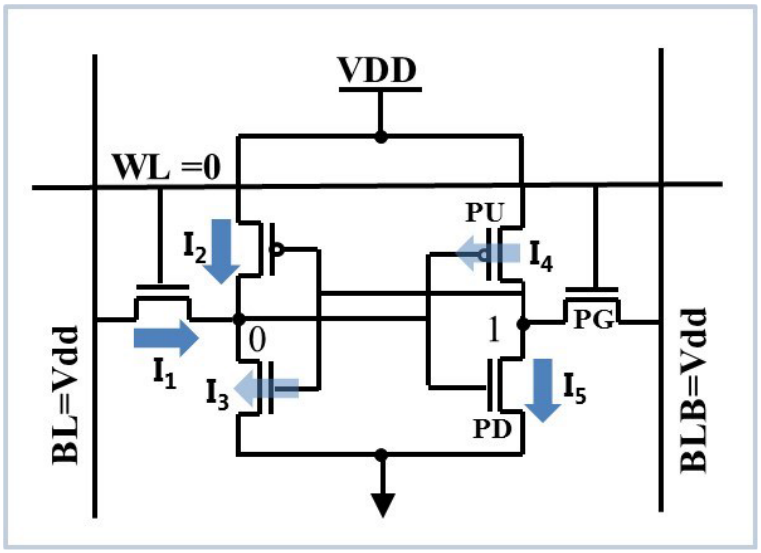


Fig.4.9. Leakage Components of a 6T-SRAM cell

Apart from the displayed components, there are other components of leakage also present in SRAM cell but their contribution is negligible at moderate and high temperature conditions [86-87]. The leakage current in 6T-SRAM cell as shown in Fig.4.10.

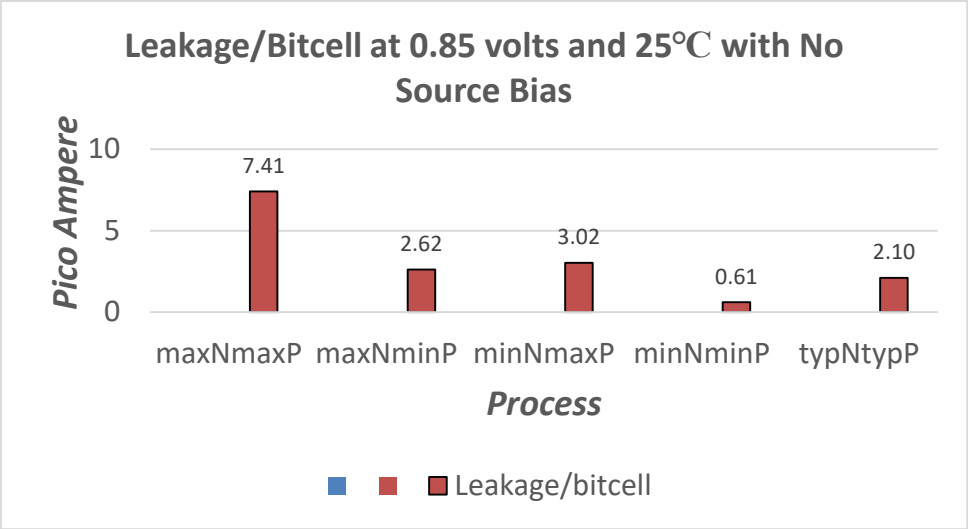


Figure4.10. Leakage current in Retention mode.

In order to reduce the leakage in standby condition, source bias (SB) is applied. Fig.4.11 illustrates the method to put memory array under SB. Memory array ground is gated through NMOS transistor and is controlled by an enable signal En (Enable). En is kept at logic '0' level in standby mode of operation. Leakage current of the array passes through the diodes M1 and M2. Due to leakage current, virtual ground level raises to the threshold voltage of diode structure and memory array goes into source bias condition. Stability of the cell is proportional to the rail-to-rail voltage, where Vrail-to-rail is defined as,

$$\text{Vrail-to-rail} = V_{DD} - \text{Virtual Ground}$$

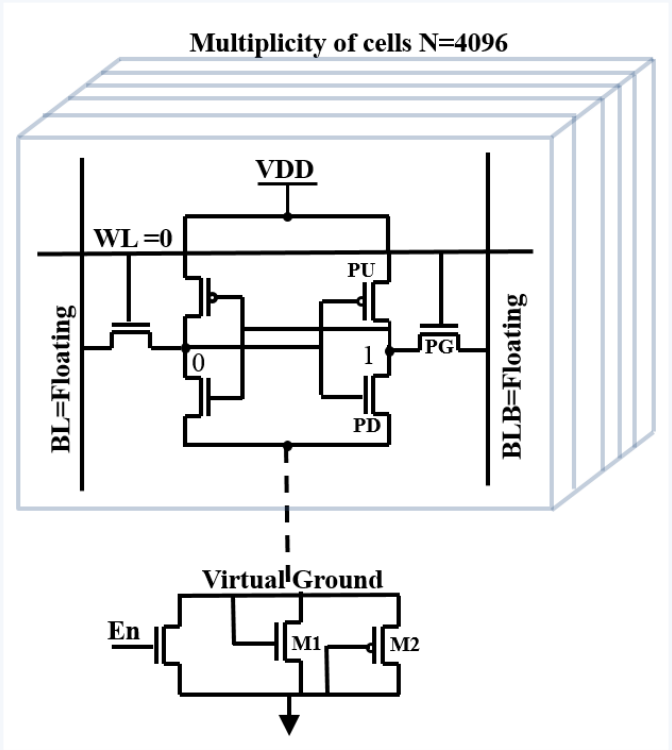


Fig.4.11. Source Bias (SB) scheme for SRAM array using Pmos-Nmos diodes

This positive Virtual Ground potential have the following effects :

1) Due to positive Virtual Ground potential, the gate-to-source voltages of PD transistors become negative, which reduces sub-threshold leakage (I_{SUB}) by certain amount.

2) Due to positive Virtual Ground potential, body-to-source voltages of PD transistor becomes negative. Therefore, threshold voltage of transistors increases, due to reverse body bias which reduce sub-threshold leakage (I_{SUB}) by certain amount.

3) Due to positive Virtual Ground, drain-to-source voltage of PD transistor decreases, which results in an increase in the threshold voltage of PD transistor, thereby reducing I_{SUB} by certain amount.

Power gating reduces the supply voltage, and increases the source voltage up to certain extent. This reduces the bit-line leakage, cell-leakage-n and cell-leakage-p, the main components of leakage at high leakage conditions, along with gate-leakage and GIDL.

Reduction of rail-to-rail voltage results in the reduced noise margin of the cell in standby condition, measured as retention noise margin (RNM). Sufficient RNM is needed for the cells in sleep/standby mode so as to ensure data integrity once the memory is reactivated from sleep mode [88]. At low voltages and slow process corners the high Vt of p-diode and n-diode devices results in low rail-to-rail voltage hence reducing RNM to unacceptable values. This scheme proposes a method to reduce leakage in retention mode while operating under SB condition for the SRAM array by selectively reducing or removing the source bias for low RNM and low leakage PVT conditions. This results in Vt lowering of diode devices making rail-to-rail voltage higher without too much loss in leakage current. To take care of cross-corner instabilities, pn combination of diode device structure is used.

4.4 Proposed SRAM Architectures for Low Leakage in Stand-by-mode

4.4.1 Normal Sleep Circuit using Footer n-diode.

Using data-retention gated power is a widely used feature for deep sub-micron SRAM [89], [90]. This Scheme uses footer devices for providing lower supply across memory cells in sleep mode (Fig.4.12). NMOS is used to provide normal mode active sink for the memory columns, controlled by EN. A low EN triggers sleep mode for the memory

core. Normal mode current is no more available. Due to leakage current of the memory core, virtual ground plane starts rising.

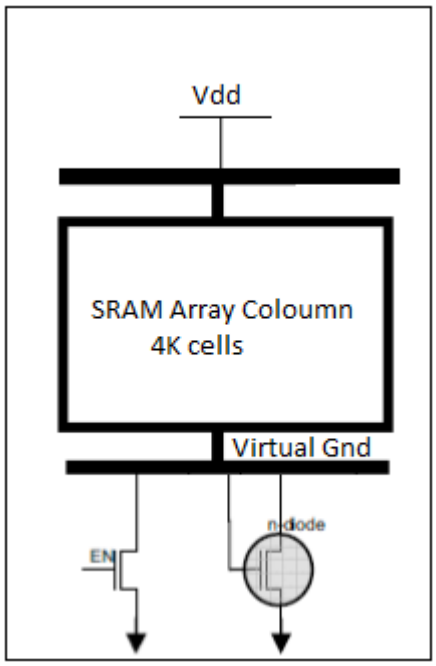


Figure4.12 Sleep circuit using Footer ndiode

The rail-to-rail voltage is given by,

V rail-to-rail = VDD – Virtual ground

For a stable and usable cell, we define the criteria,

RNM (Mean) - 6*Sigma (RNM) > 0

"Sigma Qualification > 6"

"Across the PVT's for Stabilty"

Sigma (RNM) is the standard deviation of RNM value under statistical

Simulations. From Fig. 4.13 the leakage per bit-cell at TT/25℃ is 0.83 pA.

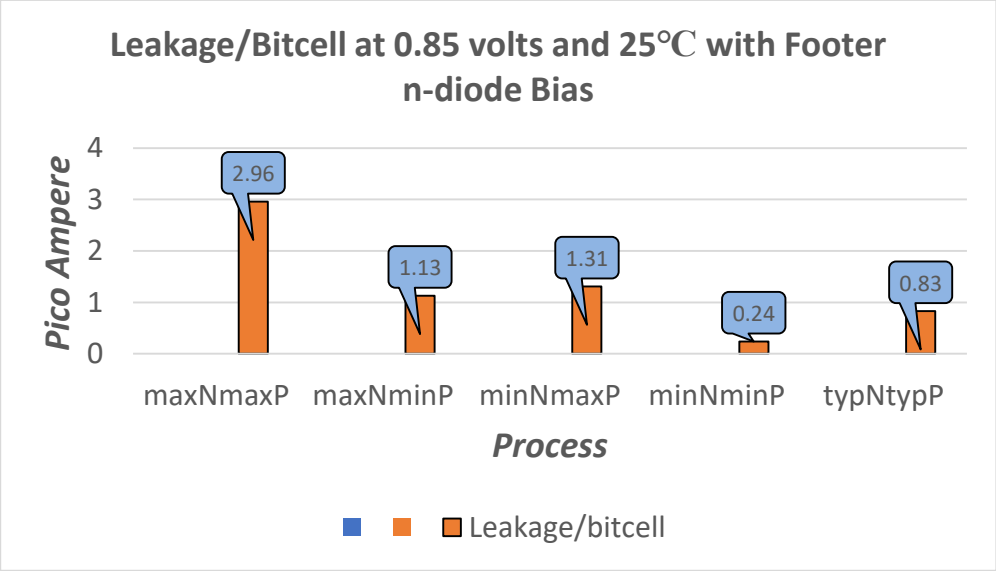


Figure4.13 Leakage variation across PVT for Footer ndiode

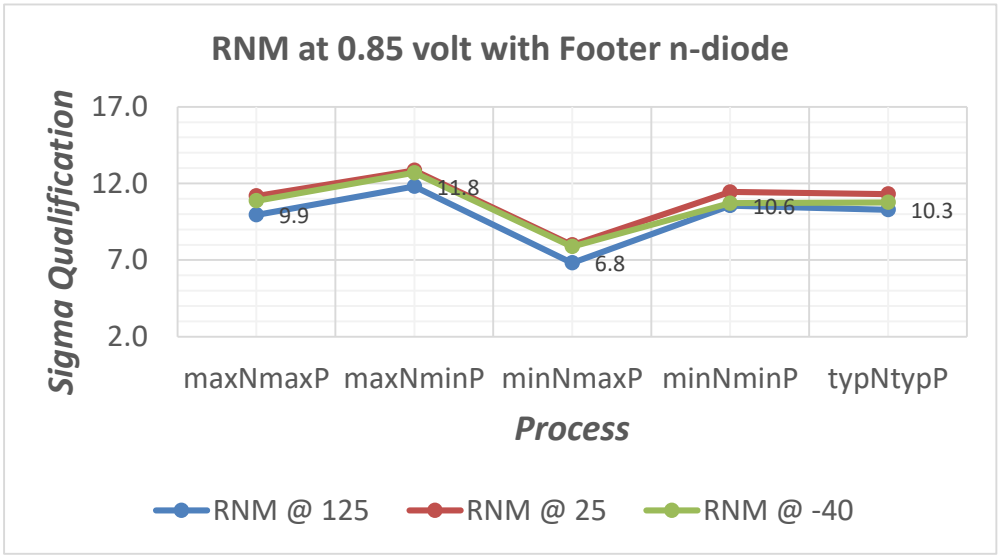


Figure 4.14 Sigma Qualification for RNM Stability for Footer ndiode

4.4.2 Sleep Compensation Circuit using Footer n-diode and Header p-diode

With the use of normal sleep circuit explained above, sigma (RNM) is above 6 as shown in fig.4.14 and the leakage at TT is 0.83 pA. At SF corner the hight Vt of nmos causes

reduction in rail-to-rail voltage. In this scheme we propose a process compensation circuit, to achieve gain in leakage without affecting RNM.

The problem arises at cross corner situations. When NMOS becomes slow and PMOS becomes fast (SF corner condition), virtual ground value rises to higher value, despite of low leakage. Thus rail-to-rail voltage reduces, decreasing the RNM and hence putting a limitation on Data Retention Voltage. Reduction in rail-to-rail voltage can be due to excessive leakage and hence increase in virtual-ground with reduction in virtual-VDD level. In deep submicron processes, due to large process variations, slow device achieves very high threshold value reducing the rail to rail voltage though leakage is very less in this condition. This phenomena creates a situation, when at cross corner conditions (SF/FS) or at slow (SS) corner, rail-to-rail voltage and RNM becomes minimum.

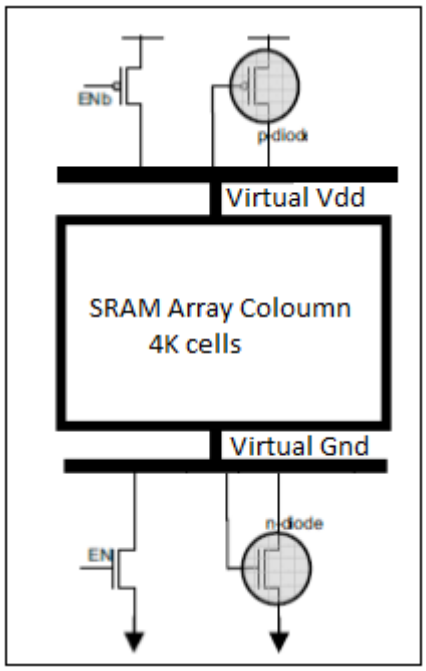


Figure4.15 Sleep circuit using Footer ndiode and Header-pdiode

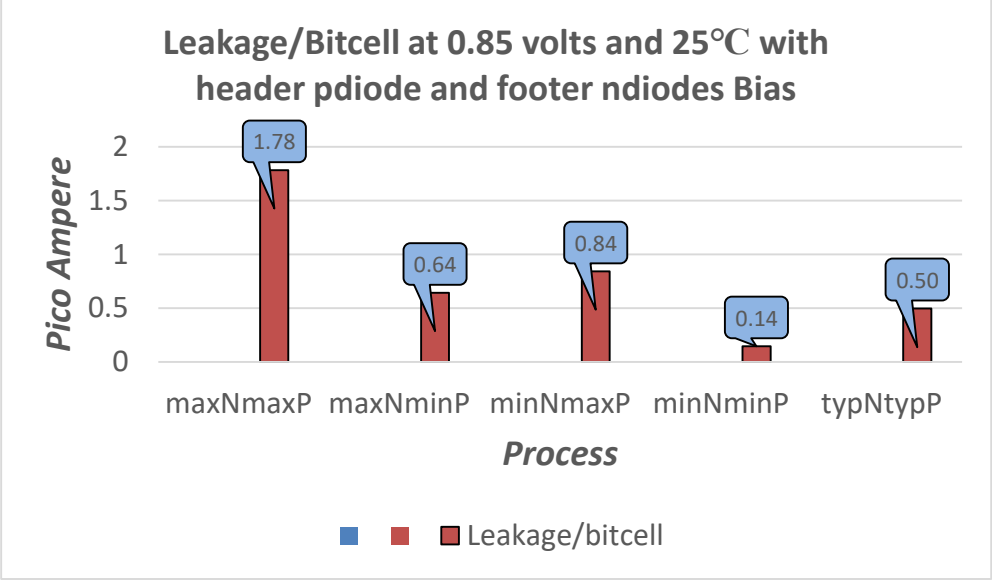


Figure4.16 Leakage variation across PVT for Footer ndiode and Header pdiode

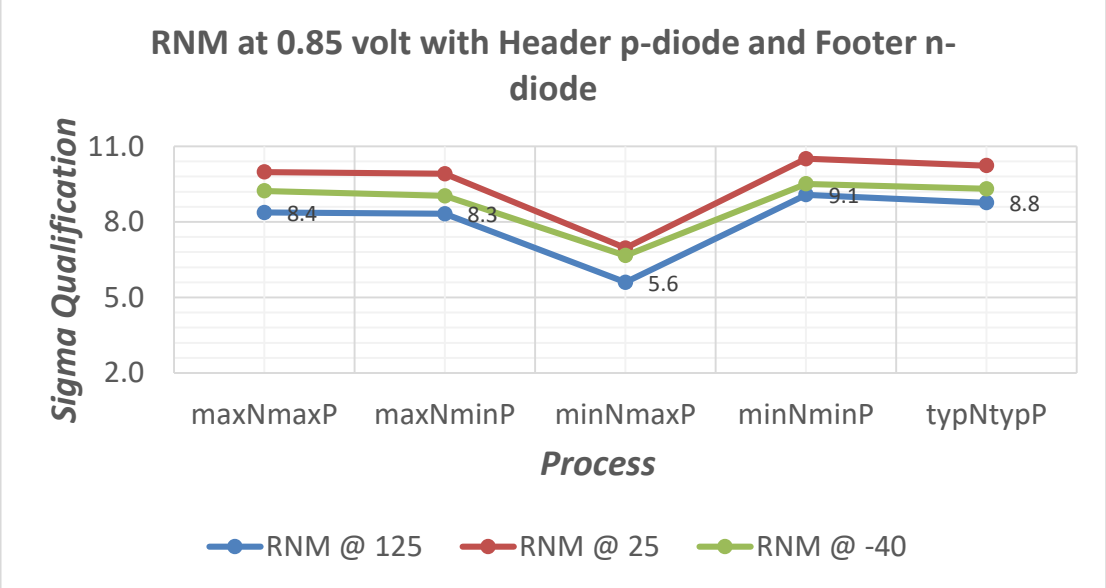


Figure 4.17 Sigma Qualification for RNM Stability using Footer n-diode and Header p-diode

4.4.3 Sleep Compensation Circuit using Footer pn-diode

In this scheme PN compensated structure takes care of cross-corner conditions. Fig. 4.18 explains the implementation of the scheme. At SF condition, n-diode in footer

becomes less effective, and most of the leakage current is siphoned by the p-diode put in parallel.

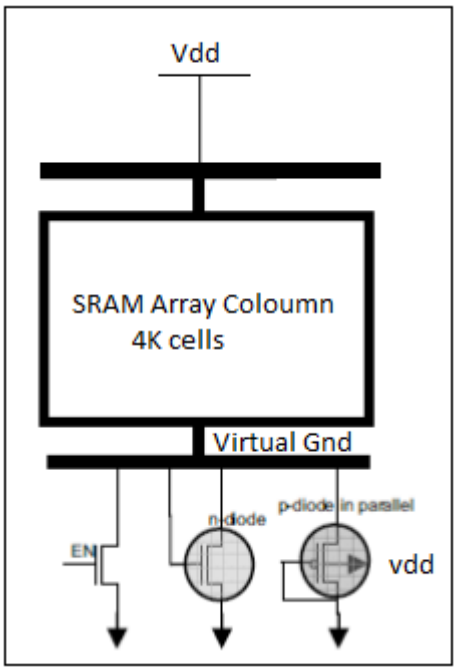


Figure4.18 Seep circuit using Footer pn-diode. Fig 4.19 and Fig 4.20 explains the SRAM stability with improved sigma Qualification and the leakage at the critical PVTs.

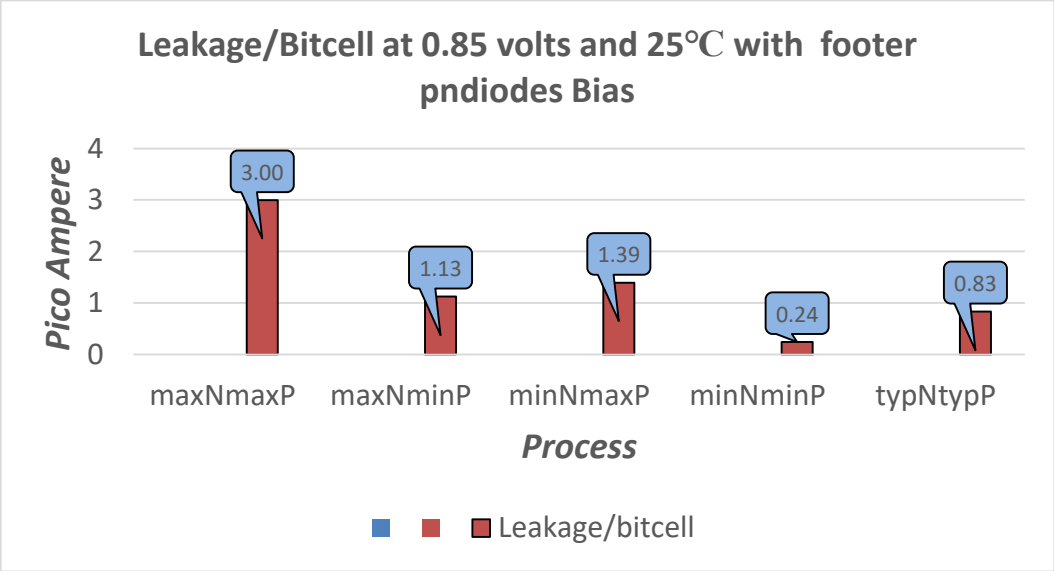


Figure 4.19 Leakage variation across PVT for Footer pn diode

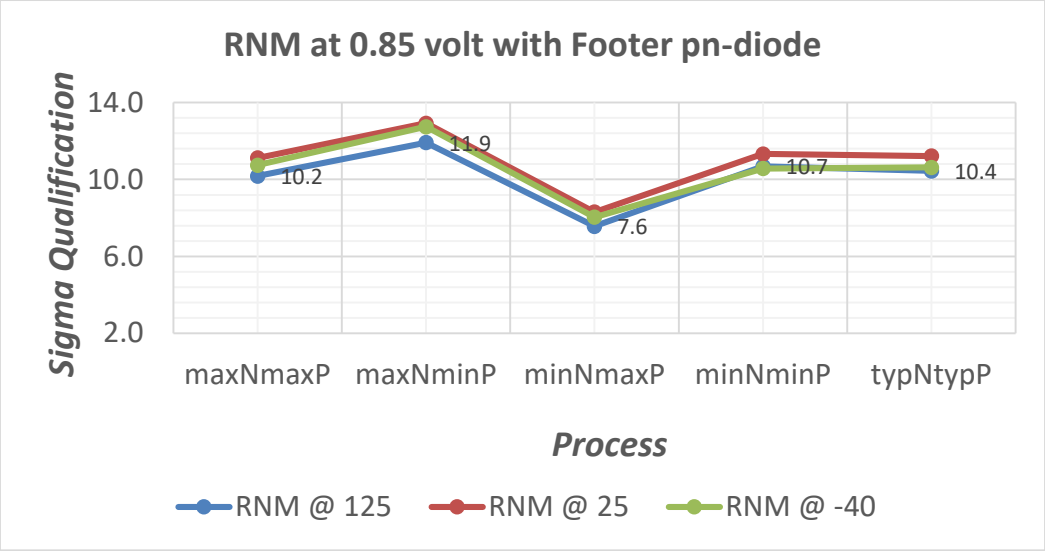


Figure 4.20 Sigma Qualification for RNM Stability using Footer pn-diode

4.4.4 Sleep Compensation Circuit using Footer pn-diode and Header pn-diode

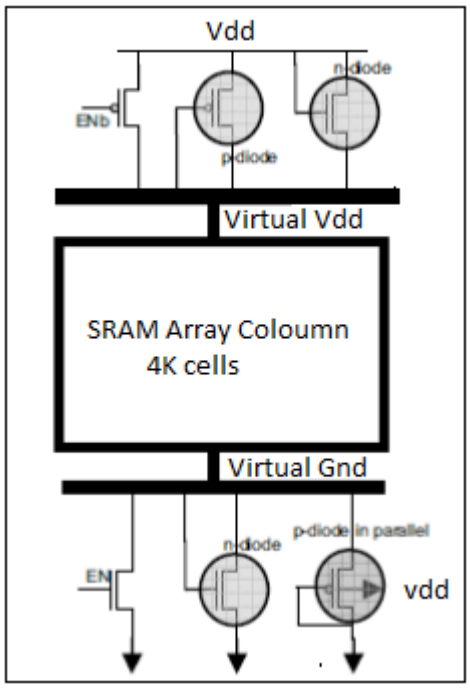


Figure4.21 Sleep circuit using Header Footer pn diodes

This scheme will solve the above two contradicting cases 2&3 where improvement in one parameter affects the other parameter. Addition of n-diode in parallel with the p-diode of header from virtual VDD takes care of excessive drop in its value. This feedback reduces threshold voltage of p-diode (header) and hence prevents virtual-VDD from falling to a certain extent. The similar arrangement can be done at footer side with parallel pn –diode combination.

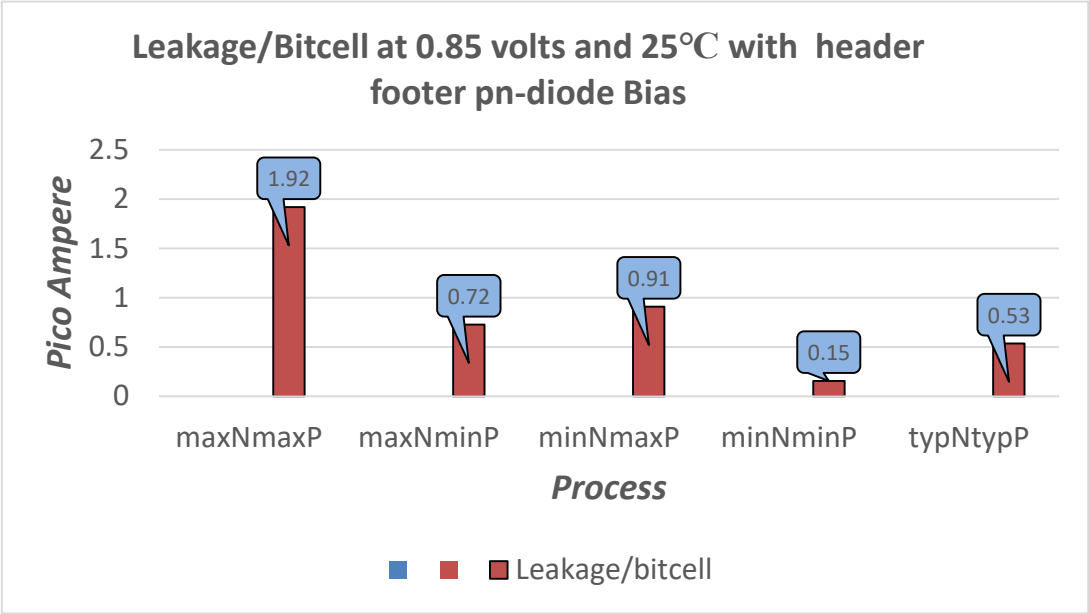


Figure4.22 Leakage variation across PVT using Header Footer pn diodes

In Fig.4.22 Leakage variation across PVT using Header Footer pn diodes.From the given figure we can say that the 0.53 pA of leakage is achieved with improved RNM stability ensuring six-sigma robustness.

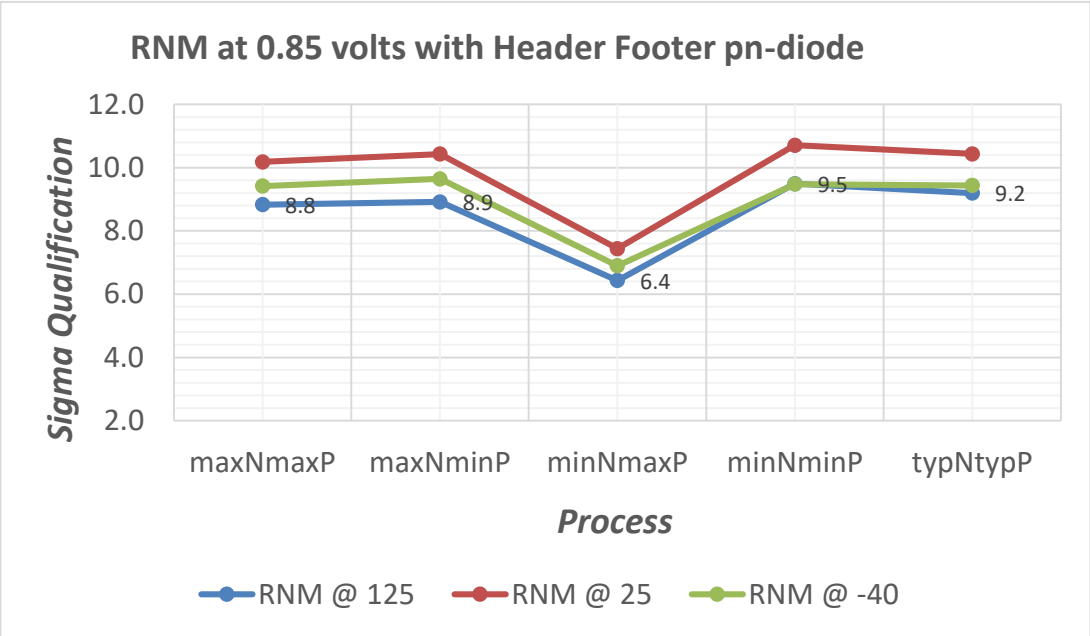


Figure 4.23 Sigma Qualification for RNM Stability using Header Footer pn diodes.

4.4.5 Result and Discussion

We have discussed four different sleep circuit schemes. Stability of memory cells under source-biased condition is analyzed through statistical simulations using Eldo-simulator (Mentor Graphics Tool). From the Fig.4.24 we can analyzed that at typical corner conditions, TT/0.85V/25°C, leakage of 0.53pA/cell is achieved and ensuring six sigma robustness for stability. Reduction in retention Vmin is extremely important for overall reduction in SoC leakage during standby mode of operation. The proposed scheme is demonstrated in a 40nm CMOS technology.

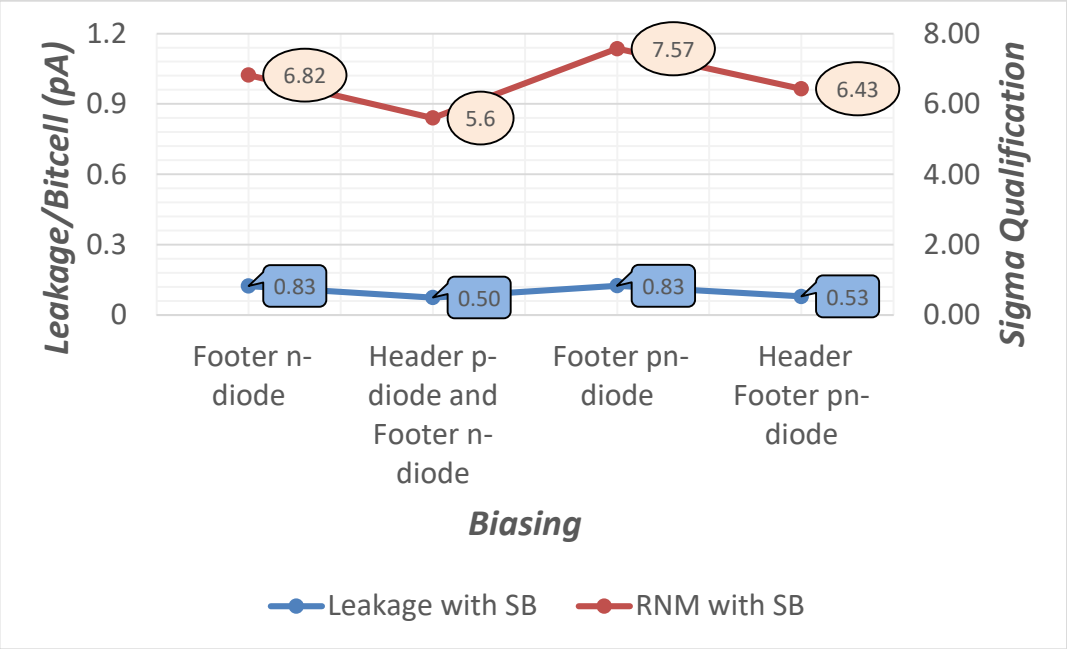


Figure4.24 Leakage reduction and Improved Stability

4.5 Chapter Summary

We have analysed sleep circuits and stability of SRAM. In this chapter a novel low power SRAM architectures based on source biasing scheme are proposed for IoT applications. This scheme proposes four different cases for sleep circuits. In the proposed schemes we target six sigma robustness for SRAM stability across PVTs without loss in leakage current.

(This page is left blank intentionally)

Chapter 5
Conclusion

With growing demand for low-power SRAMs different sources of power consumption are the targets for the low-power design techniques. Increasing leakage current in the deep submicron technology demands for low-power techniques that reduces the leakage current of the SRAM cell. The objective of the thesis is to design new SRAM architectures that can achieve low power consumption with high cell stability at reduced supply voltage

As we know that the leakage power and stability are the major concerns in portable electronic devices such as laptops, wireless sensors, high-end servers, FPGAs, IoT edge devices. These portable devices use embedded SRAM cache to improve the performance and speed of operation. All of these devices have distinctive requirements and applications. Therefore, the embedded SRAMs used in these electronic devices are designed as per the requirements and applications. In this thesis, the main focus is on leakage reduction schemes keeping sufficient margin for stability.

Power gating is a commonly used method for leakage reduction in deep submicron SRAM. However, application of such methods reduced stability of the SRAM bit-cell. Reducing supply voltage and increasing process variation put a limitation on such usage in deep submicron processes. This work proposed a method to improve leakage keeping sufficient margin for stability using Header and Footer pn-diodes to SRAM memory array. This method improves stability under cross-corner/high-leakage conditions using a feedback mechanism ensuring six-sigma robustness for RNM.

(This page is left blank intentionally)

References

1. N. S. Kim et al. (2003), Leakage current: Moore’s law meets static power, IEEE Computer Society, pp. 68-75.
2. R. K. Cavin et al. (2012), Science and engineering beyond Moore's law, Proc. of the IEEE, vol. 100, Special Centennial Issue, pp. 1720-1749.
3. K. Sung-Mo, and Y. Leblebici (2003), CMOS digital integrated circuits, Tata Mc-Graw-Hill Education.
4. M. Horowitz, et al. (2005), Scaling, power, and the future of CMOS, IEEE Electron Devices Meeting, IEDM Technical Digest.
5. G. E. Moore (1965), Cramming more components onto integrated circuits, Electronics, vol. 38, no. 8. pp. 1-4
6. Solid State Technology, Insight for Semiconductor Technology, Webcast.
7. Hook et al. (2010), Channel length and threshold voltage dependence of transistor mismatch in a 32-nm HKMG technology, IEEE Trans. on Electronics Devices, vol. 7, no. 10, pp. 2440-2447
8. Zhang et al. (2008), Embedded memory design for nano-scale VLSI systems, Integrated Circuits and Systems, Springer
9. Takeuchi et al. (2007), Understanding random threshold voltage fluctuation by comparing multiple fabs and technologies, Proc. of IEEE Inter. Device Devices Meeting (IEDM) Tech. Dig., pp. 467-470.
10. Y. Nakagome et al. (2003), Review and future prospects of low-voltage RAM circuits, IBM J. of Research and Development, vol. 47, no. 5.6, pp. 525-552.
11. P. Kulkarni and K. Roy (2012), Ultra-low-voltage process-variation-tolerant Schmitt-trigger-based SRAM Design, IEEE Trans. on Very Large Scale Integration System, vol. 20, no. 2, pp. 319-332
12. B.Islam and M. Hasan (2012), Leakage characterization of 10T SRAM
13. cell, IEEE Trans on Electron Devices, vol. 59, no. 3, pp. 631-63

14. Mukhopadhyay et al. (2005), Modeling of failure probability and statistical design of SRAM array for yield enhancement in Nano-scaled CMOS, IEEE Trans. CAD of Integrated Circuits & Systems, vol. 24, no. 12, pp. 1859–1880
15. Paul et al. (2014), A Variation-aware preferential design approach for memory-based reconfigurable computing, IEEE Trans. Very Large Scale Integration (VLSI) Systems, vol. 22, no. 12, pp. 2449-2461.
16. Zhang et al. (2008), Embedded memory design for nano-scale VLSI systems, Integrated Circuits and Systems, Springer.
17. Choi et al. (2008), New non-volatile memory structures for FPGA
18. architectures, IEEE Trans. on VLSI Systems, vol. 16, no. 7, pp.881
19. Altera (2006), FPGA architecture, White Paper, version 1.0, Altera.
20. X. Wu and P. Gopalan (2013), Xilinx next generation 28 nm FPGA technology overview, White Paper, Xilinx, WP312, v1.1.1.
21. Altera (2010), Memory System Design, Embedded Design Handbook.
22. N. Mehta (2012), Xilinx 7 series FPGAs: The logical advantage, White Paper: 7 Series FPGAs, Xilinx, WP405, v1.0.
23. N.Mehta (2012), Xilinx redefines power, performance, and design productivity with three innovative 28nm FPGA families: Virtex-7, Kintex-7, and Artix-7 Devices, White Paper: 7 Series FPGAs, Xilinx, WP373, v1.4.
24. M.Fernandez and P. Abusaidi (2010), Virtex-6 FPGA routing optimization design techniques, White Paper: Virtex-6 FPGAs, Xilinx, WP381, v1.0.
25. P.Abusaidi et al. (2008), Virtex-5 FPGA system power design considerations, White Paper: Virtex-5 FPGAs, Xilinx, WP285, v1.0.
26. Altera (2003), An analytical review of FPGA logic efficiency in Stratix, Virtex-II & Virtex-II Pro Devices, White Paper, ver.1.1.
27. Ottawa Product Design Company, New guidelines for developing.
28. Internet of things devices (2016).
29. Synopsis, Design Ware IP for IoT (2017), https://www.synopsys.com.
30. J. R. Black, “Electromigration Failure Modes in Aluminum Metallization for Semiconductor Devices, ” in Proc. of IEEE, vol. 57, no. 9, pp.1587-1594, Sep. 1969.

31. N. Weste and K. Eshraghian, “Principles of CMOS VLSI design”, Addison-Wesley, 1993.
32. Chatterjee, M. Nandakumar, and I. Chen, “An Investigation of the Impact of Technology Scaling on Power Wasted as Short-Circuit.
33. Current in Low Voltage Static CMOS Circuits,” in Proc. ISLPED, pp.145-150, Aug. 1996.
34. K. Usami, and M. Horowitz, “Clustered Voltage Scaling Techniue for Low-Power Design”, in Proc. of Int’l Symposium on Low Power Design, pp.3-8, April, 1995.
35. M. C. Johnson and K. Roy, “Datapath Scheduling with Multiple Supply Voltages and Level Converters,” ACM TODAES, vol.2, no.3, pp.227-248, July, 1997.
36. P. Chandrakasan, M. Potkonjak, R. Mehra, J. Rabaey, and R. W.Brodersen, “Optimizing Power Using Transformations,” IEEE Trans. on CAD, vol.14, no.1, pp.12-31, Jan., 1995.
37. Raghunathan and H. K. Jha, “Behavioral Synthesis for Low Power”, in Proc. of Int’l Conference on Computer Design, pp.318-322, Oct., 1994.
38. Raghunathan and H. K. Jha, “An Iterative Improvement Algorithm for Low Power Data Path Synthesis”, in Proc. of Int’l Conference on Computer Aided Design, pp.597-602, Nov., 1995.
39. L. Goodby, A. Orailoglu, and P. M. Chau, “Microarchitectural Synthesis of Performance-Constrained Low-Power VLSI Designs”, In Proc. of Int’l Conference on Computer Design, pp.323-326, Oct., 1994.
40. T. Okuma, T. Ishihara, and H. Yasuura, “Real-Time Task Scheduling for a Variable Voltage Processor”, in Proc. of Int’l Symposium on System Synthesis, pp.24-29, Nov., 1999.
41. T. Austin, D. Blaauw, T. Mudge and K. Flautner, “Making Typical Silicon Matter with Razor”, IEEE Computer Magazien, pp.57-65, March 2004.
42. Bertozzi, L. Benini and G. De Micheli, “Low-Power Error-Resilient Encoding for On-Chip Data Busses”, in Proc of Dasign Automation and Test in Europe Conference, pp.102-109, March, 2002.

43. F. Worm, P. Lenne, P. Thiran and G. De Micheli, “An adaptive low-power transmission scheme for on-chip networks”, Proc. of Int’l symposium on system synthesis, pp.92-100, Oct. 2002
44. X. Tang, V. De, J. Meindl, “Intrinsic MOSFET parameter placement due to random placement”, IEEE Trans on VLSI, 1997, pp. 369-376.
45. Seevinck, F. List, J. Lohstroh, “Static-noise margin analysis of MOS transistors”, JSSC, 1987, pp. 748-754.
46. Current reduction in VLSI systems. Journal of Circuits, Systems, and Computers, 5. 11(6):621–636, 2002.
47. B Calhoun and A Chandrakasan. Static noise margin variation for sub-threshold SRAM in 65-nm CMOS. IEEE Journal of Solid State Circuits, 41(7):1673, 2006.
48. Li Ding and P. Mazumder. Dynamic noise margin: definitions and model. In VLSI 9. Design, 2004. Proceedings. 17th International Conference on, pages 1001 – 1006, 2004.
49. S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi and V. De, “Parameter Variations and Impact on Circuits and Microarchitecture”, Proc. DAC, 2003, pp. 338-342.
50. M. Khellah, Y. Ye, N. S. Kim, D. Somasekhar, G. Pandya, A. Farhang, K. Zhang, C. Webb and V. De, “Wordline & bitline pulsing schemes for improving SRAM cell stability in low-Vcc 65nm CMOS designs,” Symp. on VLSI Circuits, pp. 9-10, June 2006.
51. [1] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi and V. De, “Parameter Variations and Impact on Circuits and Microarchitecture”, Proc. DAC, 2003, pp. 338-342. [1] J.M. Rabies, Digital integrated circuits, Prentice Hall, (1996)
52. K .Itch, VLSI Memory Chip Design,Springer-Verlag,NY,2001
53. K. Roy and S.C. Prasad. Low-Power CMOS VLSI Circuit Design. John Wiley and Sons, 2000.
54. Chandrakasan and R. Brodersen. CMOS Low Power Digital Design. Kluwer Academic Pubs., 1996

55. Makoto Yabuuchi et al, “20nm High-Density Single-Port and DualPort SRAMs with Wordline-Voltage-Adjustment System for Read/Write Assists”, ISSCC 2014, p.p. 234-235.
56. Jonathan Chang et al, “A 20nm 112Mb SRAM in High-κ Metal-Gate with Assist Circuitry for Low-Leakage and Low-VMIN Applications”, ISSCC 2013, p.p. 316-317.
57. Robert Aitken et al, “On the Efficacy of Write-Assist Techniques in Low Voltage Nanoscale SRAMs”,DATE 2010.
58. Mudit Bhargawa et al, “Low VMIN 20nm Embedded SRAM with Multi-voltage Wordline Control based Read and Write Assist Techniques”, Symposium on VLSI Circuits Digest of Technical Papers,2014.
59. R.Ranica et al, “FDSOI Process/Design full solutions for Ultra Low Leakage, High Speed and Low Voltage SRAMs”, Symposium on VLSI Technology Digest of Technical Papers, 2013.
60. Vivek De et al, “Capacitice Coupling Wordline Boosting with SelfInduced Vcc Collapse for Write Vmin Reduction in 22-nm 8T SRAM” ISSCC 2012, p.p. 234-235
61. Pramod Kolar et al, “A 32 nm High-k Metal Gate SRAM With Adaptive Dynamic Stability Enhancement for Low-Voltage Operation”, IEEE Journal of Solid-State Circuits, Vol. 46, No. 1, Jan 2011, p.p. 76-84.
62. Yen-Huei Chen et al, “A 16nm 128Mb SRAM in High- κ Metal-Gate FinFET Technology with Write-Assist Circuitry for Low-VMIN Applications”, ISSCC 2014, p.p. 238-239.
63. Eric Karl et al., “A 4.6GHz 162Mb SRAM Design in 22nm Tri-Gate CMOS Technology with Integrated Active Vmin Enhanced Assist Circuitry,” ISSCC 2012, pp 230-231
64. H. Pilo, et al., “A 64Mb SRAM in 32nm High-k metal-gate SOI technology with 0.7V operation enabled by stability, write-ability and read-ability enhancements,” ISSCC 2011, pp. 254-256.
65. Hu et al, “Analysis of GeOI FinFET 6T SRAM Cells”,IEEE Transactions on Electron Devices, 2015.

66. Yi-Wei Lin et al, “A 55nm 0.55V 6T SRAM with Variation-Tolerant Dual-Tracking Word-Line Under-Drive and Data-Aware WriteAssist”,ISLPED 2012.
67. Kumar et al, “A 6T-SRAM in 28nm FDSOI technology with Vmin of 0.52V using assisted read and write operation”, ICICDT 2015
68. J. Bhavnagarwala et al, “The impact of intrinsic device fluctuations on CMOS SRAM cell stability,” IEEE Journal of Solid-State Circuits, Vol. 36, pp. 658-665, Apr. 2001.
69. L. Chang et al, “An 8T-SRAM for Variability Tolerance and LowVoltage Operation in High-Performance Caches,” IEEEJournal of SolidState Circuits, Vol. 43, pp. 956-963, Apr. 2008.
70. W. Dehaene et al, “Embedded SRAM design in deep deep submicron technologies,” European Solid-State Circuits Conference (ESSCIRC), pp. 384-391, 2007.
71. W. Dong et al, “SRAM Dynamic Stability: Theory, Variability and Analysis,” Intl. Conf. on Computer Aided Design (ICCAD), pp. 378385, 2008.
72. M. Iijima et al, “Low Power SRAM with Boost Driver Generating Pulsed Word Line Voltage for Sub-1V Operation,” Journals of Computers, Vol. 3, No. 5, May 2008
73. M. Khellah et al, “Read and Write Circuit Assist Techniques for Improving Vccmin of Dense 6T SRAM Cell,” Intl. Conf. on IC Design and Technology, June 2008.
74. K. Nii et al, “A 45-nm Bulk CMOS Embedded SRAM with Improved Immunity Against Process and Temperature Variations,” IEEE Journal of Solid-State Circuits, Vol. 43, Jan 2008.
75. H.Pilo et al, “An SRAM design in 65nmtechnology node featuring read and write-assist circuits to expand operating voltage,” IEEE Journal of Solid-State Circuits, Vol. 42, Apr 2007.
76. N. Shibata et al “A 0.5V 25 MHz 1mW 256kb MTCMOS/SOI SRAM for Solar-Power-Operated Portable Personal Digital Equipment - Sure Write Operation by Using Step-Down Negatively Overdriven Bitline Scheme,” IEEE Journal of Solid-State Circuits, Vol. 41, Mar 2006.

77. C.Wang et al, “A Boosted Wordline Voltage Generator for Low Voltage Memories,” ICECS, 2003. [11] J. Wang et al, “Analyzing Static and Dynamic Write Margin for Nanometer SRAMs,” ISLPED, 2008.
78. M. Yamaoka et al, “90-nm Process-Variation Adaptive Embedded SRAM Modules with Power-Line-Floating Write Technique,” IEEE Journal of Solid-State Circuits, Vol. 41, Apr 2006.
79. H. S. Yang et al, “Scaling of 32nm Low Power SRAM with High-K Metal Gate,” IEEE Intl. Electron Devices Meeting, 2008.
80. K. Zhang et al, “Low Power SRAMs in nanoscale CMOS technologies,” IEEE Trans. Electron Devices, Vol. 55, No. 1, pp. 145-151, Jan. 2008.
81. K. Zhang et al, “A 3 GHz 70 Mb SRAM in 65nm CMOS technology with integrated column-based dynamic power supply,” IEEE Journal of Solid-State Circuits, Vol. 41, No. 1, Jan 2006.
82. V. De and S. Borkar, “Technology and design challenges for low power and high performance,” in Proc. Int. Symp. Low Power Electronics and Design, 1999, pp. 163–168.
83. K. Roy and S. C. Prasad, Low-Power CMOS VLSI Circuit Design. New York: Wiley, 2000, ch. 5, pp. 214–219.
84. Mead, “Scaling of MOS technology to submicrometer feature sizes,” Analog Integrated Circuits Signal Process., vol. 6, pp. 9–25, 1994.
85. R. Dennard et al., “Design of ion-implanted MOSFET’s with very small physical dimensions,” IEEE J. Solid-State Circuits, vol. SC-9, p. 256, Oct. 1974
86. J. Brews, High Speed Semiconductor Devices, S. M. Sze, Ed. New York: Wiley, 1990, ch. 3.
87. (2001) International Technology Roadmap for Semiconductors. International SEMATECH, Austin, TX. [Online]. Available: http://public.itrs.net/
88. S. Thompson, P. Packan, and M. Bohr, “Linear versus saturated drive current: Tradeoffs in super steep retrograde well engineering,” in Dig. Tech. Papers Symp. VLSI Technology, 1996, pp. 154–155.
89. S. Venkatesan, J. W. Lutze, C. Lage, and W. J. Taylor, “Device drive current degradation observed with retrograde channel profiles,” in Proc. Int. Electron Devices Meeting, 1995, pp. 419–422.

90. J. Jacobs and D. Antoniadis, “Channel profile engineering for MOSFET’s with 100 nm channel lengths,” IEEE Trans. Electron Devices, vol. 42, pp. 870–875, May 1995.
91. W. Yeh and J. Chou, “Optimum halo structure for sub-0.1 m CMOSFET’s,” IEEE Trans. Electron Devices, vol. 48, pp. 2357–2362, Oct. 2001.
92. Keshavarzi, K. Roy, and C. F. Hawkins, “Intrinsic leakage in low power deep submicron CMOS ics,” in Proc. Int. Test Conf., 1997, pp. 146–155.
93. R. Pierret, Semiconductor Device Fundamentals. Reading, MA: Addison-Wesley, 1996, ch. 6, pp. 235–300.
94. S. Grove, Physics and Technology of Semiconductor Devices. New York: Wiley, 1967.
95. Y. Taur and T. H. Ning, Fundamentals of Modern VLSI Devices. New York: Cambridge Univ. Press, 1998, ch. 2, pp. 94–95.