DESIGN OF LOW-VOLTAGE LOW-POWER NANO-SCALE SRAMS

DO ANH TUAN

School of Electrical and Electronic Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirement for the degree of
Doctor of Philosophy

2010

Design of Low-Voltage Low-Power Nano-scale SRAMs

NANYANG TECHNOLOGICAL UNIVERSITY

Submitted by **DO ANH TUAN**

Supervisor **Professor YEO KIAT SENG**

A thesis Submitted to School of Electrical and Electronic Engineering of Nanyang Technological University in partial fulfillment of the requirement for the degree of Doctor of Philosophy in Engineering

September, 2010

ABSTRACT

Static Random Access Memory (SRAM)-based cache is one of the most important components of state-of-the-art VLSI systems. It is responsible for increasing the speed of data flows, and hence the speed of the whole electronic system. SRAM is prevalently utilized in the design of modern microprocessors for bridging the widening divergence between the performances of the Central Processing Unit (CPU) and the Dynamic RAM (DRAM)-based main memory. This trend is accentuated by the never-ending market demand for sophisticated communication and multimedia applications, which require high-tech portable electronic gadgets with high-performance as the requisite feature. As the on-chip memory occupies a large portion of the chip area, the power dissipated within the memory, both active and standby, will become a dominant part of the chip's total power consumption. In view of the above, there is invariably an apparent urgency to address these two often-conflicting power and performance requirements.

Our research focuses on SRAM cell design for ultra low-voltage ultra low-power applications. Two SRAM cell topologies have been proposed with an improved noise margin and hence can operate at very low supply voltages to save power. Our first proposal is a 10T SRAM cell design that has 4x read and 10x leakage power reduction when compared to the conventional 6T design. The proposed cell has separate read/write ports and hence its read noise margin is 31% higher than that of the conventional 6T design. We also proposed a special read scheme that accesses only one cell during read out and hence more than 98% of cell active current is saved (assuming that each row has 128 cells and operating frequency is 250 MHz). This hefty amount of power reduction is obtained at the cost of 33% cell area overhead and 23% of write-ability. Extensive simulations on two SRAM macros using a standard 65nm / 1V CMOS process showed that the total read power consumption of the proposed design has reduced to 25% of that of the conventional 6T design. It also has smaller write power consumption. More

importantly, its leakage current per cell is only one tenth of that of the conventional 6T design. This property is a major advantage as leakage current will soon dominate the active current in the future technologies.

Our second SRAM cell proposal is a differential 8T cell with a column-based dynamic cell supply. The 8T cell consists of a conventional 6T cell and an inverter. Unlike the conventional 6T design, the word-line (WL) signal in our 8T design is used to control the input of the inverter, the output of which is tapped to the access transistors of the core 6T cell. As a result, the proposed 8T cell maintains the differential read/write property of the 6T cell and eliminates the inherent half-access issue of the previously published designs. Thus, it can be bit-interleaved for an efficient conventional Error Correcting Code (ECC) implementation. In order to improve the cell's static noise margin (SNM), a dynamic cell supply is raised to a higher voltage during the read operation. Furthermore, all transistors used in the 8T cell are of minimum size and hence its leakage is comparable to that of the 6T and its area overhead is only 14%. The cell thus can operate properly in a wide range of supply voltages from 0.2 V to 1V. Its V_{DDmin} is only 0.15V, which is similar to the above-mentioned 10T cell and much lower than that of the 6T design. Detailed analysis has proved that the proposed 8T is the most suitable design for ultra low supply voltage applications.

Other than SRAM cell design, we also investigated several sense amplifier (SA) circuits to further assist the read operation of the SRAM. A new hybrid-mode SA that offers high-speed sensing to improve the read operation of the SRAM was introduced. As the read delay in SRAM is normally longer than its write delay, SRAM speed is limited by its read operation. Our proposed SA improves the reliability of the conventional current-mode SA while maintaining its speed. Simulation results using a standard 0.18 μm/ 1.8 V CMOS process showed that the proposed design offers more than 70% sensing delay and power consumption reduction when compared to the other current-mode SA designs. Another SA that is very much insensitive to process variation was also proposed. The design of this SA is based on the conventional cross-coupled structure with a novel input/output port

topology. Simulation results using a standard 65 nm/ 1V CMOS process showed that our design outperforms the SAs in comparison by at least 27% in terms of speed and 30% in terms of power consumption. Sensitivity analysis has proven that the new design offers the best reliability with the smallest standard deviation and bit-error-rate (BER). Four 32-bit SRAM macros have been used to validate the proposed design, in comparison with three other circuit topologies. The new design can operate at a maximum frequency of 1.25 GHz at 1 V supply voltage and a minimum supply voltage of 0.2 V. Finally, we analyzed the impact of process variations on a conventional latch-type SA. Although this SA topology is significantly faster than the conventional current-conveyor SA, its input-offset voltage is heavily dependent on the random device mismatches. Our analytical work proposed a criterion to quantify the offset of the SA using the pre-defined process variations. As these data are normally available from the foundry, the proposed formula allows us to calculate the worst-case offset voltage of the SA. This greatly reduces the amount of work in the design flow of SA as its offset voltage can be predicted easily. Our formula was evaluated using three different predictive FET models (180 nm CMOS, 32 nm CMOS, 22 nm CMOS) and the analytical results closely correlate with the simulation results.

ACKNOWLEDGMENTS

I would like to express my deepest thanks to my research supervisor, Professor Yeo Kiat Seng, for his knowledge, advice, expertise, professionalism and patience throughout my research. Through numerous fruitful and intellectual discussions, I have been motivated by his innovation and encouragement through many times of good discussions. The positive thinking and can-do attitude he has taught me also helps me carry on my research despite all of the difficulties and problems that I have faced.

I would also want to extend my appreciation to Dr. Kong Zhi Hui for her support. I am grateful to have her guidance since the first day I joined Prof. Yeo Kiat Seng's research group four years ago.

It would also be very difficult to do my job without the generous assistance from all the technicians in Center for Integrated Circuits and Systems and IC Design. I thank them for their technical support.

I would like to thank my girlfriend Melissa for her love and the sleepless nights she spent to proof-read my publications. I am so thankful to have her constructive comments, without which none of my works would be published in such a short time.

Last but not least, I give thanks and glory to my family for giving me wisdom and supports to research in this topic.

TABLE OF CONTENT

NOMENCLATURE

γ	Gamma, the body effect parameter
/BL	Bit-line bar (the complementary bit-line)
/CS	Column Select
/DL	Data-line bar (the complementary bit-line)
/OE	Output Enable
/WE	Write Enable
BER	Bit-Error-Rate
BL	Bit-line
C_{BL}	Bit-line parasitic capacitance
C_{DL}	Data-line parasitic capacitance
CMOS	Complementary Metal-Oxide Semiconductor
DL	Data-line
DNM	Dynamic noise margin
ECC	Error Checking Code
F	Operating frequency
GFS	Global Foundries Singapore (Former known as Chartered Semiconductor)
G_m	Small-signal trans-conductance
Grnd	Ground voltage level
I	Current
L	Channel length
r	Cell ratio
RBL	Read Bit-line
RWL	Read Word-Line
SA	Sense Amplifier
SNM	Static noise margin
SRAM	Static Random Access Memory
STM	STMicroelectronics
t_{AA}	Address access time
t_{RC}	Read cycle time
t_{WC}	Write cycle time
V_{BS}	Body-to-Source voltage
V_{CM}	Common-mode voltage
V_{DD}	Supply voltage
V_{DDmin}	Minimum data retention voltage
V_{DS}	Drain-to-Source voltage
V_{GS}	Gate-to-Source voltage
VLSI	Very Large Scale Integration
V_n	Noise voltage
V_{SS}	Zero voltage reference
V_{th} or V_T	Threshold voltage
W	Channel width
WBL	Write Bit-line
WL	Word-line
WTP	Write trip point
WWL	Write Word-Line
β or K	Trans-conductance of the transistor
λ	Lambda, the short-channel length effect parameter

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 INTRODUCTION

1.1 SRAM overview and applications

SRAM is a type of semiconductor memory that theoretically can store data as long as the supply voltage is available. At the heart of any SRAM is the cross-coupled inverters (**Fig. 1-1**) which holds the true (Q) and the complementary (/Q) values of any stored bit. During the SRAM operation, a memory cell is addressed using its row and column address and hence any physical location can be accessed randomly. SRAM has three features that make it an *indispensable* component in computer memory hierarchy: **1)** It is fully compatible with the bulk CMOS process and hence can be fabricated as on-chip memory. **2)** It is stable due to the positive feedback storage structure, and thus, refresh cycle is not required. **3)** It has a very high speed operation and consequently forms an important bridge between the microprocessor and the main memory.

Due to its more complex internal structure, SRAM is less dense than DRAM (DRAM) and is therefore not used for high-capacity and low cost applications such as main memory.

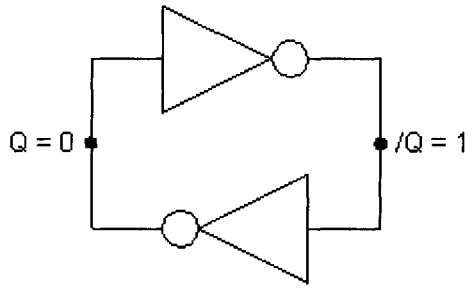


Figure 1-1 The cross-coupled inverters

Since SRAMs are faster than other RAMs, they are normally located at the top of the system's memory hierarchy (**Fig. 1-2**). SRAM can be implemented as a stand-alone component or an embedded memory. Embedded SRAMs are normally smaller and are used as a buffer to speed up the systems. SRAMs can be found in a wide range of systems,

from simple electronic devices such as toy cars and network routers, to more complex systems such as computers, digital cameras and automotive electronics. In computer architecture, SRAMs are used as Level 1 and Level 2 caches (L1 and L2) as shown in **Fig. 1-2**.

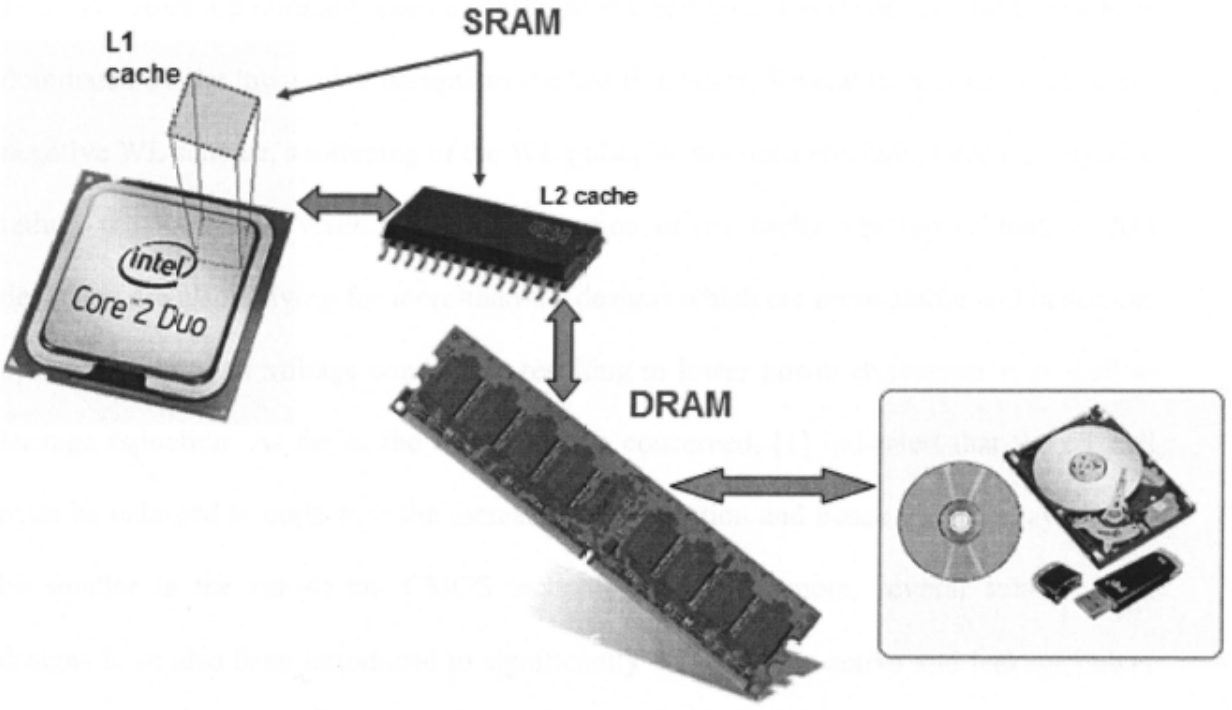


Figure 1-2 Generic computer memory hierarchy. SRAMs are located at the top of the hierarchy to boost up the speed of the whole system

1.2 Trends in SRAM design

Over the last decade, the development of SRAM has shifted from stand-alone component to the embedded cache memory. With the continuous increase of microprocessor speed, SRAM's performance has also undergone rapid enhancement, thanks to the incessant scaling in CMOS fabrication technology. This, however, leads to three unavoidable consequences: **(1)** Capacity of SRAM is escalated after each generation to keep up with the boundless demands from the complex electronic systems. **(2)** Given the fact that both size and speed of SRAM are growing, the power consumption of the SRAM cache will eventually dominate the total power consumption of a chip. **(3)** As technology scales down to sub-100 nm regime, threshold voltage (V_{th}) variations and sub-

threshold leakage current become problematic as the former complicates voltage scaling while the latter dictates the use of lower supply voltage in order to save power.

The above-mentioned issues have shaped the trends in SRAM designs in recent years. In particular, quite a lot of less-than-6T SRAM designs have been proposed to reduce the size of the cell and hence the total silicon area occupied by the cache. Moreover, minimum size transistors are normally used to optimize the cell area. However, this trend has been dominated by the low-power designs in the last five years. Several techniques such as the negative WL scheme, shortening of the WL pulse, Write-assist etc. have been employed to reduce the Read and Write power consumption of the cache. On top of that, SRAM designers are also striving for more-than-6T designs which are more stable and hence can operate at ultra-low voltage conditions, resulting in lower power consumption as well as leakage reduction. As far as the layout area is concerned, [1] indicated that the 6T cell must be enlarged to cope with the increasing V_{th} variation and hence the 8T's layout will be smaller in the sub-45-nm CMOS technologies. Furthermore, several sub-threshold designs have also been introduced to significantly cut down the active and leakage power consumption of the cache. These designs centered around two major concerns in low-power nano-scale SRAM designs: process variations tolerance and leakage minimizations. As a result, the research community has seen 9T [2], 10T[3-6] or even 11T[7] SRAM designs which are able to work at very low voltage supplies in order to suppress the leakage current. In the near future, SRAM will continue to play its major roles in embedded cache memory. Therefore, designing SRAM for lower-power consumption, higher speed performance and better stability are the utmost requirements for researchers and circuit designers in nano-scale CMOS technologies.

1.3 Motivation

Naturally, caches are expected to operate as fast as possible because they communicate directly with the microprocessors. From a circuit's perspective, this can be done by redesigning circuit topologies, using a higher power supply or simply utilizing

larger transistor sizes. These options also improve the stability of the memory cell as well as the read/write operations. However, cache designers face the serious constraint of stringent silicon area allocation as well as power budget limitation from the battery and thus, cache speed has to be compromised.

As mentioned before, one of the most important requirements in SRAM cache design is to *obtain a small cell area*. Thus, near-minimum device sizes are normally used in SRAM cells. This, however, leads to slow read/write operations as well as unstable cells against various type of disturbances. As a consequence, one must use larger-than-minimum-sized transistors in memory cells to enhance their stability [8-10]. Despite this enhancement, cell stability still cannot cope with the excessive process variations and hence, more-than-6-T cell designs must be used to separate the read/write ports of the cells, therefore greatly improving their noise margin during read and write operations [11-12]. This trend has extended to 11-T cell design to enhance cell read stability, write margin as well as read reliability [7] at the cost of more than 60% cell area overhead. Consequently, it is desirable to have more stable yet smaller area memory cell design to cope with rapidly increasing process variations and read/write disturbances.

Another challenge that has recently received widespread attention in every cache design is to *reduce the total power consumption*, which for the sake of simplicity can be broadly categorized into read/write dynamic power and leakage power. As technology and supply voltage scale down, the threshold voltage of the MOS devices must also be reduced. Hence, their leakage currents increase exponentially [7, 13]. Since only a small portion of the millions of cells in the cache are accessed at any one time while the rest are at standby, the standby leakage current contributes a significant portion to the total power consumption. This tendency is moving faster along with the scaling down of V_{th} and the escalating cache capacity. Therefore, to some extent, reducing cache's standby leakage is becoming more crucial than managing its dynamic power. In this project, we are going to

explore the circuit techniques used to reduce the following three power components of the cache, namely the leakage power, read power and write power.

The third (and possibly the most difficult) challenge of cache design in the state-of-the-art technology is *ensuring its high yield*. As we are moving to sub-32 nm gate length technologies, device fabrications and operations are showing more profound statistical behaviors [10, 14-19]. Thus, deterministic design methodology is no longer suitable to predict the yield of the device. One must therefore comprehend the statistical nature of these variations in order to manage the yield of the cache. Statistical approaches [10, 16-18, 20-23] are more reliable but appear to be computationally costly. Important sampling methodology [24] has recently been employed to estimate the yield of SRAM design with more than several orders faster than traditional Monte Carlo simulations. This method however requires a good choice of the evaluation function. Nevertheless, it allows an estimation of failure probability smaller than 10^{-10}, which is almost impossible to be carried out by using the traditional Monte Carlo simulations.

1.4 Contributions

This research work achieved the following contributions:

1. A new 10T SRAM cell has been proposed. It has separate read/write ports hence it is more stable than the conventional 6T during the read operation. The proposed design offers 75% read power reduction and 90% leakage reduction at 65 nm/ 1V CMOS process.
2. A new 8T SRAM cell has also been proposed. This is the first differential 8T cell reported in the literature. In this design, a dynamic cell supply control is also employed in order to improve both read and write noise margin of the cell. As a result, our proposed design is more stable and is able to work at sub-threshold supply voltage conditions. Moreover, the proposed design also solves the half-access issue in the previously published design and hence can be bit-interleaved for an efficient ECC.

Simulation results has reaffirmed that the proposed design offer 2X noise margin and 50% power reduction when compared to the conventional 6T SRAM design.

3. A hybrid-mode sense amplifier (SA) has been proposed to improve the reliability of the conventional current-mode SA. Circuit performance of the new design has been evaluated using a standard 0.18 μm/ 1.8 V CMOS process.
4. A latch-type SA has been proposed with a novel input/output port topology. The proposed design is less sensitive to process variations and hence is more reliable in sub-100 nm CMOS processes.
5. A criterion has been proposed to closely predict the input-offset voltage of the latch-type SA.

1.5 Thesis organization

The rest of the thesis is organized as follows: **Chapter 2** introduces basic SRAM operations and studies two of the most important issues in SRAM design: Power consumption and reliability. It reviews the most popular techniques and published works in SRAM designs. **Chapter 3** presents our 10T SRAM cell which is assisted by a proposed read scheme to minimize the read power consumption. **Chapter 4** proposes another SRAM cell design that offers high noise margin and low power consumption using a column-based dynamic supply voltage scheme. This cell design eliminates the half-accessed issue in the conventional SRAM and hence can be bit-interleaved to reduce the multi-bit soft error rate. In **Chapter 5,** we propose two latch-based SAs to further improve the power and speed performances of the memory read out circuit. **Chapter 6** investigates the input-offset voltage of a latch type SA. **Chapter 7** forms the conclusion of the thesis.

CHAPTER 2 LITERATURE REVIEW

As discussed in **Chapter 1**, SRAM designers must address cell area, total power consumption, cell stability, read/write reliability and speed performance simultaneously in order to meet the requirements of an advanced system. Previously, circuit designers have strived for higher speed performance [25]. However, with the explosion of handheld gadgets market and the continuous miniaturization of device feature size, power consumption management has been given the highest priority followed by the system reliability [11]. Nevertheless, these two factors cannot be optimized without jeopardizing each other (at least, there is not yet an obvious solution). Furthermore, they both face the stringent requirements of limited silicon area and latency, satisfying which normally results in degraded stability.

This chapter begins with the basic operation of the Metal Oxide Semiconductor Field Effect Transistor (MOSFET), focusing on leakage mechanisms and V_{th} variations in deep sub-micron technologies. Based on this knowledge, we then study the main sources of power consumption and circuit techniques to tackle the power-performance constraints. Subsequently, the issue of cell stability and its analytical model will be studied and demonstrated. The last subsection will be dedicated to review the most recently published works on SRAM cell designs.

2.1 MOS characteristics

This section discusses the operations of an n-type MOS transistor, i.e. nMOS, its I-V characteristics and leakage mechanisms. In a similar manner, the discussion is applicable to the pMOS as well. **Fig. 2-1(a)** shows the cross-sectional and symbolic view of an nMOS transistor [26]. It has four terminals, namely the Source (S), Drain (D), Gate (G) and Body (B). Its channel length (L) and channel width (W) are also presented in the same figure.

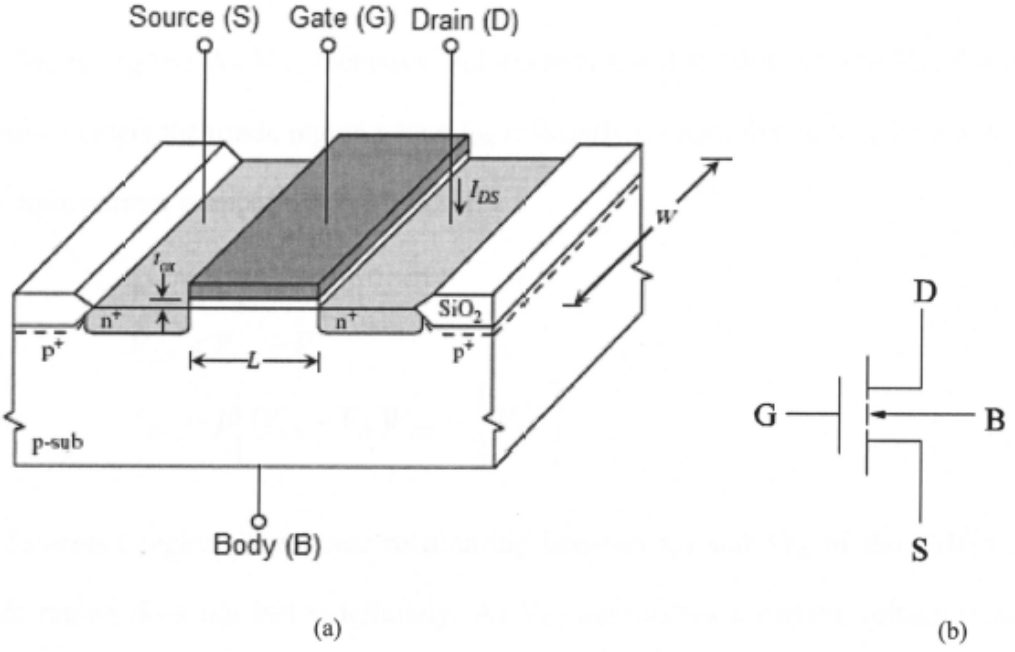


Figure 2-1 nMOS a) cross-section b) symbol view [26]

The nMOS has three regions of operation: Cutoff, Triode and Saturation.

Cutoff region: This region corresponds to $V_{GS} < V_{th}$ and $I_{DS} \approx 0$, where V_{GS} is the gate-to-source voltage, V_{th} is the threshold voltage and I_{DS} is the drain-to-source current. The cutoff region is also referred to as the sub-threshold region where the drain current (I_{DS}) in this region is much smaller than those in the triode and the saturation regions. Thus, the transistor is considered to be turned off. However, I_{DS} is not strictly zero and increases exponentially as the threshold voltage V_{th} decreases. The sub-threshold current is given by [26]:

$$I_{sub} = W\frac{I_o}{W_o}10^{-V_{Th}/S} \tag{2.1}$$

$$S \cong \frac{kT_j}{q}(1+\frac{C_{DP}}{C_{OX}})\ln 10 \tag{2.2}$$

where W is the gate width of the MOS, I_o/W_o is the current density to define V_{th}, S is the sub-threshold slope, C_{OX} is the gate capacitance, C_{DP} is the depletion-layer capacitance and T_j is the junction temperature. Eq. (2.1) shows that I_{sub} is exponentially dependent on the threshold voltage V_{th}.

Triode region: As V_{GS} increases and exceeds the threshold voltage V_{th}, the nMOS transistor enters the triode region where I_{DS} is linearly proportional to V_{DS} for a given V_{GS}. The drain current is approximated by [26]:

$$\begin{aligned} &V_{GS} \geq V_{th} \\ &V_{DS} < V_{GS} - V_{th} \\ &I_{DS} = \beta\left[(V_{GS} - V_{th})V_{DS} - \frac{1}{2}V_{DS}^2\right] \end{aligned} \tag{2.3}$$

Saturated region: The linear relationship between I_{DS} and V_{DS} of the nMOS in the triode region does not last indefinitely. As V_{DS} approaches a certain voltage value, I_{DS} saturates and does not increase with the increasing V_{DS}. Drain current (I_{DS}) of an nMOS in the saturated region is given by [26]:

$$\begin{aligned} &V_{GS} \geq V_t \\ &V_{DS} \geq V_{GS} - V_{th} \\ &I_{DS} = \frac{\beta}{2}(V_{GS} - V_{th})^2 \end{aligned} \tag{2.4}$$

$$\beta = \frac{W}{L}\mu_n\frac{\varepsilon_{OX}}{t_{OX}} = \frac{W}{L}\mu_n C_{OX}$$

$$\begin{aligned} V_{th} &= V_{to} + K\left(\sqrt{|V_{BB}| + 2\Psi} - \sqrt{2\Psi}\right) \\ &= \left(V_{FB} + K\sqrt{2\Psi} + 2\Psi\right) + \sqrt{2\varepsilon_S qN}\left(\sqrt{|V_{BB}| + 2\Psi} - \sqrt{2\Psi}\right) \end{aligned}$$

Where

β: is the conductance of the nMOS.

ϵ_{ox} and t_{ox}: permittivity and thickness of the gate oxide.

V_{FB}: flat-band voltage.

Ψ: Fermi potential,

N: substrate doping concentration.

ϵ_s: permittivity of silicon

q: magnitude of electric charge.

I-V characteristic curves of a typical nMOS operating in the triode and the saturated regions are illustrated in **Fig. 2-2(a)**.

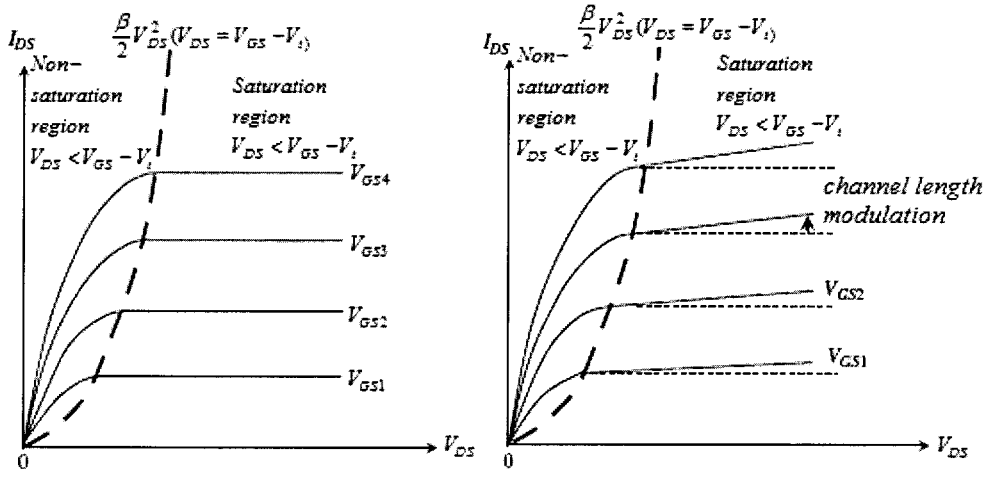


Figure 2-2 I-V characteristic of an nMOS (a) ideal (b) actual [26]

Channel Length modulation: In the actual MOS characteristics shown in **Fig. 2-2(b)**, in fact, I_{DS} increases slightly with V_{DS} even in the saturation region. This phenomenon can

be accounted for by including the empirical channel length modulation parameter (λ) into Eq. (2.5):

$$I_{DS} = \frac{\beta}{2}(V_{GS} - V_{th})^2(1 + \lambda V_{DS}) \tag{2.5}$$

Gate-Tunneling current: Ideally, the gate current of the MOS device is zero. However, in practice, there is a very small current flowing through the gate of the MOS which increases with the reduction of the gate oxide thickness [26]. The gate tunneling current flows from the gate to the source in an nMOS or from the source to the gate in a pMOS, via the channel. Since the gate-oxide thickness has been rapidly decreased to less than 3 nm [26], this current becomes prominent. **Fig. 2-3** illustrates the dependency of the gate tunneling current on V_{GS} and $t_{ox.}$ Although the absolute value of this current is considerably smaller than those of the sub-threshold current and the active current of the MOS devices, the presence of millions of transistors in the cache makes it an important factor in power management strategy.

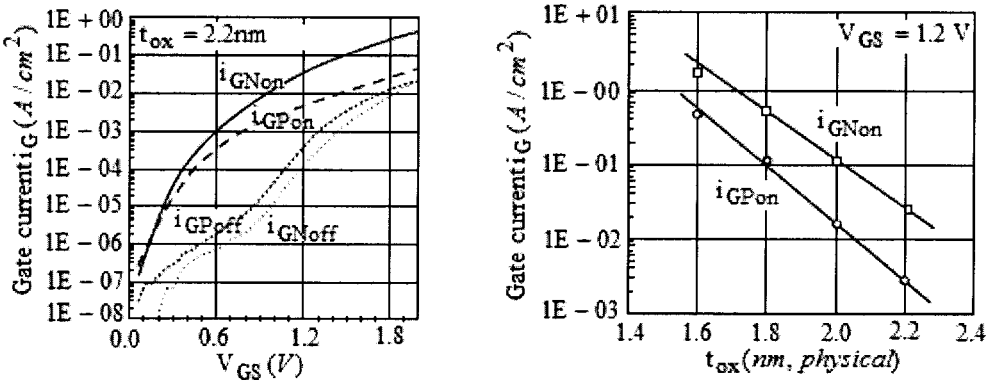


Figure 2-3 Gate tunneling current i_{GN} and i_{GP} for nMOS and pMOS, respectively [26-27]

Physical meaning and explanation of these operating characteristics can be found in some popular textbooks, such as [28-29]. These equations will be used in this report as guidelines to highlight the sources of power consumption in the cache in the next chapters.

2.2 SRAM operation

SRAM occupies a large portion of the silicon area of many contemporary digital systems. Thus, their power consumption constitutes a significant part of that of the chip [30-31]. This situation is getting more severe with the rapid development of semiconductor audio, video, game players and mobile devices that dictates high performance, high-capacity memory and low power consumption with equal priorities [28]. **Fig. 2-4** shows a simplified SRAM organization. Functionalities and constituent circuits of SRAM and its design challenges are briefly discussed in **Subsections 2.2.1** and **2.2.2** below.

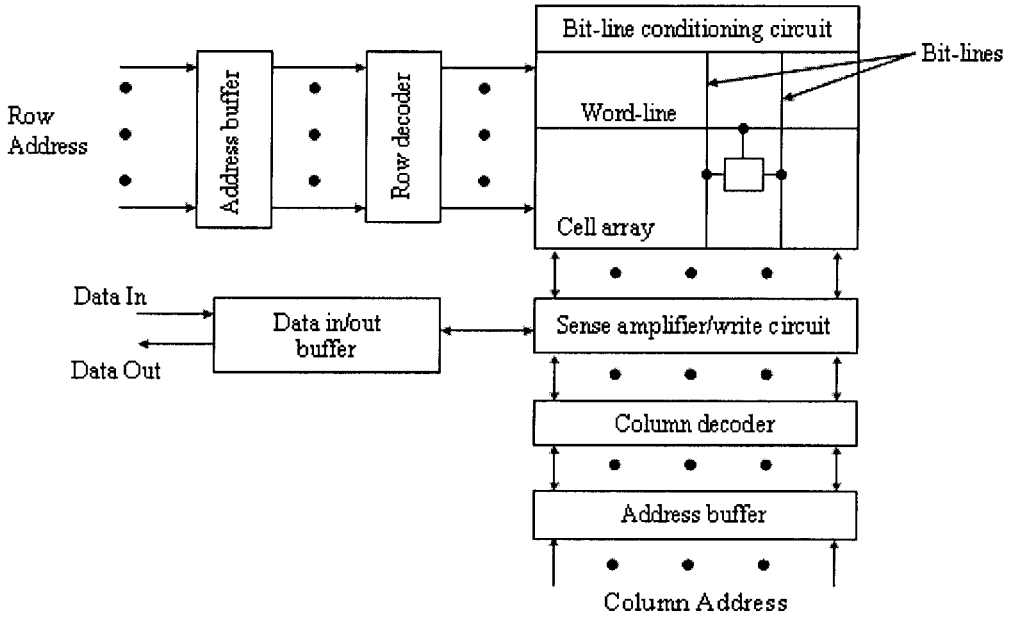


Figure 2-4 A typical SRAM structure with simplified Input/Output interfaces

2.2.1 Functionality and structure of the SRAM

On-chip SRAMs are used to store temporary data of the processor. As they receive direct instructions from the processor, it is required that data are written to or read from the cache as fast as the speed of the processor. In general, the nearer they are to the processor in terms of proximity, the faster they should operate. However, high speed operation normally incurs additional power consumption, which is not desirable for

battery-powered devices. CMOS SRAM features very fast write and read operations and can be designed to have extremely low standby power consumption [30, 32-33]. Therefore, it is used extensively as on-chip cache at different hierarchical levels, each of which has a different speed and size as shown in **Fig. 1-2**. A typical SRAM structure **(Fig. 2-4)** consists of a SRAM cell array, row/column decoders, SAs and read/write driver circuits. Operations of these sub-circuits will be discussed in the next sections. When an instruction is sent from the microprocessor, both row/column decoders are activated by the control circuitry to address a specific memory cell. Data will be written to or read from the chosen cell by the write driver or the SA. Since refresh cycles are not required in SRAM, it is generally faster and consumes less power than DRAM. It is also more stable and be able to operate in radiation-hardened environments [32].

These privileges however come at a cost. A conventional SRAM cell **(Fig. 2-5)** is bulky with 6 transistors per cell. It consists of a cross-coupled structure (N1, N2, P1, and P2) to store data and two pass-gate transistors (N3, N4) for accessing, also shown in **Fig. 2-5**. Thus, SRAM cache requires a substantial area usage.

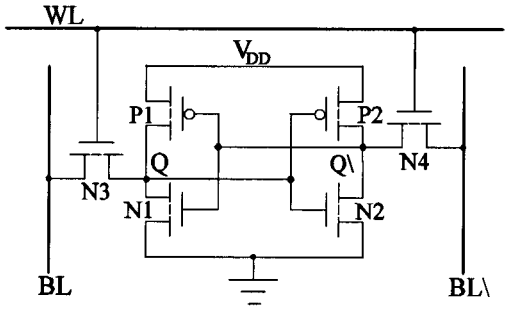


Figure 2-5 Conventional 6T SRAM cell

Logically, circuit designers use minimum-sized devices in the SRAM to save area. However, this cannot be done indefinitely when performance and stability come into consideration. Another design concern is the power consumption. Previously, attentions were drawn mostly to the dynamic power during the active cycles. However, this factor is becoming less dominant as technology advances to the nanometer regime. When the MOS

devices get smaller and thinner, their leakage currents are higher, which constitute a significant portion of power dissipation [30, 33]. Furthermore, in the state-of-the-art electronic systems where capacity of cache memory reaches tens of MB, only a small fraction of memory cells is in active mode while the rest stores data in the standby mode. This results in a considerable amount of standby power consumption. Consequently, SRAM designers must reduce the supply voltage V_{DD} in order to decrease both dynamic and static power dissipations which unavoidably lead to degradations of both stability and speed performance. At a low voltage operation, one of the solutions to enhance SRAM's performance and stability is to employ larger MOSFET devices. Ironically, this solution faces the constraint of the chip's stringent silicon area. Therefore, a smart compromise must be made to accommodate the requirement of the system as a whole, be it from the view point of high-speed, low-power or small-area.

2.2.2 SRAM core

SRAM core consists of a sea of identical cells each of which normally has a 6-T topology as shown in **Fig. 2-5**. More SRAM cell designs are going to be mentioned in **Section 2.6**. Each bit-line (BL) pair of the SRAM is shared by all cells in the same column while each data-line (DL) is shared by all columns in the same macro. To reduce the parasitic capacitances associated with these lines, memory core is partitioned into smaller banks. The purposes are two-fold: Firstly, it shortens the BLs and DLs and hence reduces the parasitic capacitances. This results in faster access time, better reliability and lower power consumption during the read/write operations [34-36]. Secondly, during each read/write access, all memory cells on the same row are semi-active (i.e. turned on by the WL but not accessed by the CS signal) and hence consumes a significant amount of power [28]. Partitioning the memory core into multiple banks reduces the number of cells per world-line (WL) and hence greatly reducing the dynamic power per read/write cycle. However, this approach incurs additional silicon area because each bank requires

individual column and row sub-decoder [28, 37]. Therefore, circuit designers must consider the area-power consumption trade-off to obtain the optimum partitioning.

2.2.3 Read/write operation

In each Read/Write (R/W) operation, a particular cell is selected by triggering its WL and BLs. This task is performed by the row decoder and the column decoder driven by the address buffers. The selected cell is positioned at the point of intersection between the activated WL and BL [29]. A column SA will then detect the contents of the selected cell in the form of small voltage or current variations via the complementary BLs. The detection must occur as speedily as possible to achieve minimal access time.

The commonly used control signals in an SRAM are /WE, /CS and /OE. The /WE signal decides between the read and write modes. For the read operation, the BL signals are transported to the output, whilst for the write operation, the BLs are driven from the input.

Timing diagram of a read operation is shown in **Fig. 2-6 (a)**. During this time, the data stored in a specific SRAM location (defined by the address) is read out. The read cycle time, t_{RC}, and the address access time, t_{AA}, are indicated. **Fig. 2-6 (b)** shows the write cycle, which permits changes to the information within an SRAM. Two timings are indicated, the write cycle time, t_{WC}, and the write recovery time, t_{WR}.

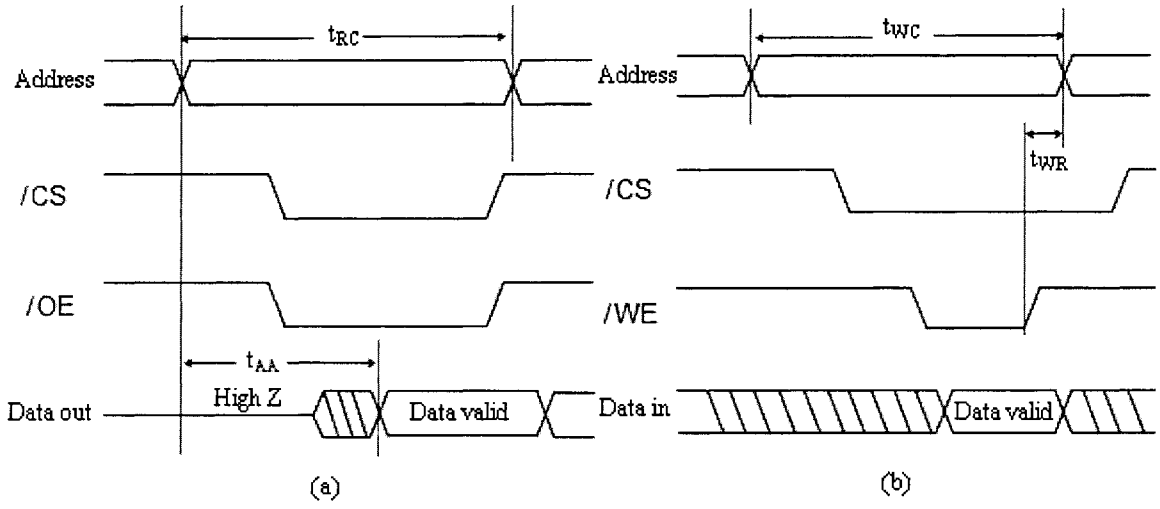


Figure 2-6 Typical timing diagram of an SRAM (a) Read cycle (b) Write cycle [37].

2.2.4 Peripheral circuits

2.2.4.1 Decoder

Because of the random access nature of the cache, address decoders must be present in the constituent circuits of SRAM. These address decoders choose one specific memory cell in order to readout the stored datum or to write a new datum into the cell. When designing these decoders, it is crucial to keep a geometry matching between the dimensions of the decoders and the memory core. Failing to do so would result in a dramatic wiring overhead as well as additional power consumption, timing mismatches and silicon area [28]. These decoders can be implemented statically or dynamically.

2.2.4.2 Write driver

During a write cycle, the input datum (be it a "1" or a "0") must be transferred to one of the DLs while its complementary is transferred to the other. As a result, the write driver behaves like a strong inverter buffer that is able to drive one of the DL to V_{DD} while the other to 0, i.e. full swings, in a short time. **Fig. 2-7** shows a simplified schematic of a SRAM write circuitry.

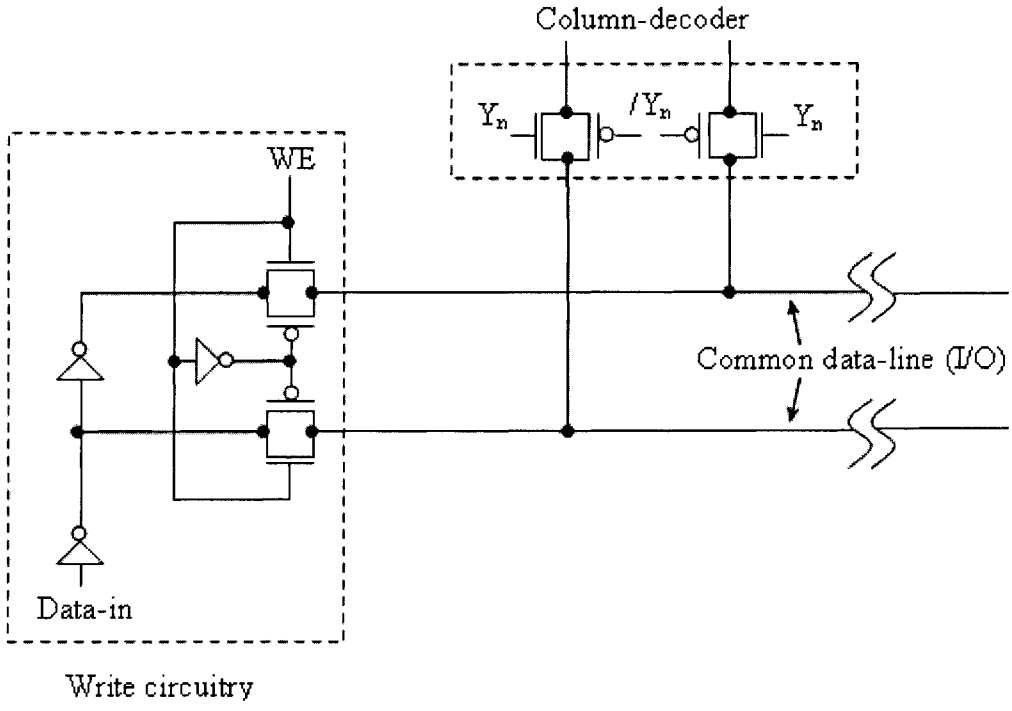


Figure 2-7 A SRAM write circuitry

2.2.4.3 SA

SA, as its name suggests, is responsible for sensing the input signal(s) and amplifying it (them) to the desirable output level(s). It is the key element in defining the performance and environment tolerance of CMOS memory [32] and performs the following functions: 1) Amplification 2) Delay reduction 3) Power reduction [28]. SA can be categorized according to their circuit types (differential or non-differential) or modes of operation (voltage, current or charge) [32]. The minimum signal amplitude that can be distinguished by the differential amplifier is much smaller than that of the non-differential topology [32]. Thus, differential SA can be enabled sooner than the non-differential SA, resulting in a faster sensing delay and lower power consumption. Therefore, differential sensing is more favorable than its non-differential counterpart, especially in the state-of-the-art memory where extremely high level of noise is present.

Conventionally, voltage-mode sensing is used because the very large input resistance of the MOS devices offers high voltage gain. However, its sensing speed is limited by the charging/discharging time of the circuit-inherent parasitic capacitive elements. Thus,

current-mode sensing is more suitable in contemporary, highly packed, large capacity memory systems to obtain a superior power-delay product (PDP). Charge-transfer preamplifier may be used as an alternative to improve the sensing speed of the voltage-mode amplifiers. Nonetheless, in most cases, the performance of the preamplifier-plus-voltage-amplifier is inferior to that of purely current-mode sensing design [32]. In this thesis, two newly proposed SA designs will be presented in **Chapter 5**.

2.3 Sources of power dissipation in SRAM

Power dissipated in an SRAM can be categorized into two components: active (or dynamic) power and standby (or leakage) power [37]. Each of these components comes from the memory array, the decoders and the periphery circuits such as write drivers, SA and the pre-charge circuits, etc. The standby power is dominated by the leakage current from the memory array, thus, static current from other sources can be ignored [37]. The active and standby power of an m × n-bit SRAM can be expressed as follows:

$$\begin{aligned} P_{active} &= I_{DD}V_{DD} \\ &= \left(I_{array} + I_{decoder} + I_{peripheral}\right)V_{DD} \\ &= \left\{\left[mi_{active} + m(n-1)i_{hold}\right] + \left[(m+n)C_{DE}V_{\text{int}}f\right] + \left[C_{PT}V_{\text{int}}f + I_{DCP}\right]\right\}V_{DD} \end{aligned} \tag{2.6}$$

$$P_{s\,\text{tan}\,dby} = m \times n \times i_{hold} \times V_{DD} \tag{2.7}$$

V_{DD}: External power supply to the SRAM.

V_{int}: Internal power supply to the SRAM.

i_{active}: Active current of the selected cells.

i_{hold}: data retention leakage current.

C_{DE}: output node capacitance of the decoders.

F: the operating frequency.

C_{PT}: is the capacitance of the logic and driving circuits in the periphery.

I_{DCP}: total static or short-circuit current of the periphery [19, 28].

In Eq. (2.7), we assumed that leakage currents from other sources are negligible [19].

Eq. (2.6) suggests some solutions to reduce the total power consumption of a cache. For example, cell active current can be reduced by reducing the number of cells per accessed row by re-partitioning the memory array (i.e. the mi_{active} component). Periphery and decoder power consumptions can be trimmed down by using two-stage decoder structure to reduce C_{PT} and C_{DE} components or using smaller V_{int}.

On the other hand, as shown in Eq. (2.6) and Eq. (2.7), the standby leakage component is rather constant and only dependent on the capacity of the memory array, i.e. $m \times n$. In principle, this static dissipation should be negligible. However, its value is escalating in the nano-meter regime [38] where transistor leakage currents and cache size are growing exponentially, making leakage a major source of energy dissipation in the cache [28, 33].

2.4 SRAM power reduction techniques

Power reduction techniques are essential in cache design. In this section, the most popular and effective circuit techniques to manage cache power consumption will be discussed. We will look into both active and standby power as they are equally important in nano-meter CMOS technologies.

2.4.1 Macro partitioning

From the architecture point of view, macro partitioning is the most effective way to reduce active power of the cache. By dividing the memory core into multiple sub-modules, number of cells activated per read/write operation is proportionally reduced and so is the active cell current, i.e. the mi_{active} component in Eq. (2.6). It also trims down the dynamic power consumed by the decoders and peripheral circuits due to the reduced parasitic capacitance associated with the WLs, BLs and DLs [28, 35-36, 39-42]. From the reliability perspectives, macro partitioning enhances the read/write yields since it involves smaller BL and DL capacitances. Unfortunately, this multistage hierarchical architecture incurs additional silicon area for the decoders and peripheral circuits. As a result, sub-module

size cannot be reduced infinitely and must be carefully considered to balance its performance and area.

2.4.2 Power reduction by modulating power supply voltage

Lowering V_{DD} is a remarkably effective way to save power, both in memory and logic circuitries [1, 28, 43-52]. However using a single low-voltage power supply significantly degrades the speed and reliability performances, especially in SRAM since the scaling limit of V_{DD} in SRAM is higher than that of digital circuits [53-54]. Thus, dynamic-V_{DD} and dual-V_{DD} schemes have been proposed to combat this drawback. The former dynamically modulates the power supply to the SRAM circuits so that high-V_{DD} is available in the active mode to enhance the performance while low-V_{DD} is applied in the standby mode to reduce power dissipation [31, 49, 55-59]. The latter, on the other hand, has two different supplies, each for different circuit components to overcome the speed-power trade-off [33, 60-62]. Details operation of each scheme can be found in the respective references.

2.4.3 Power reduction by using dynamic *sleep* transistors

This approach utilizes a gated *sleep* transistor to switch between the active and standby modes [33, 62-68]. Various dynamic *sleep* transistor implementations are presented in **Fig. 2-8**. In the active mode, the Wake signal activates the *sleep* NMOS transistor, pulling the virtual ground (vgnd) to zero potential. Thus, full voltage swing is applied across the SRAM array. In the standby mode, the Wake signal is triggered low but not to zero to limit the leakage current from the SRAM array. Vgnd is therefore raised to a voltage level between V_{DD} and ground. The implementation in **Fig. 2-8 (a)** [69] is simple but the *sleep* transistor needs to be sized properly to meet the wake-up timing requirement and to maintain a low *IR* drop in the current path during an active mode. This causes a severe constraint on the size of the *sleep* transistor to accomplish the optimal vgnd voltage level during the data-retention mode. The vgnd is determined by the leakage path from the SRAM array to the ground through the *sleep* transistor. In order to avoid this,

Bhavnagarwala A. *et. al* [70] used a diode-connected pMOS transistor to force the vgnd to stay at V_{tp} during standby (**Fig. 2-8 (b)**). However, this solution does not give optimal leakage reduction and the level of accuracy of the vgnd control is also degraded by the impact of V_{tp} variations. Programmable bias transistors are effective in overcoming these issues [65] (**Fig. 2-8(c)**). Yet, this design still faces the inter-die, intra-die and temperature-induced variations [33]. Finally an external reference voltage V_{ref} is used to maintain a well-controlled vgnd through the means of an op-amp, as shown in **Fig. 2-8 (d)** [33, 62].

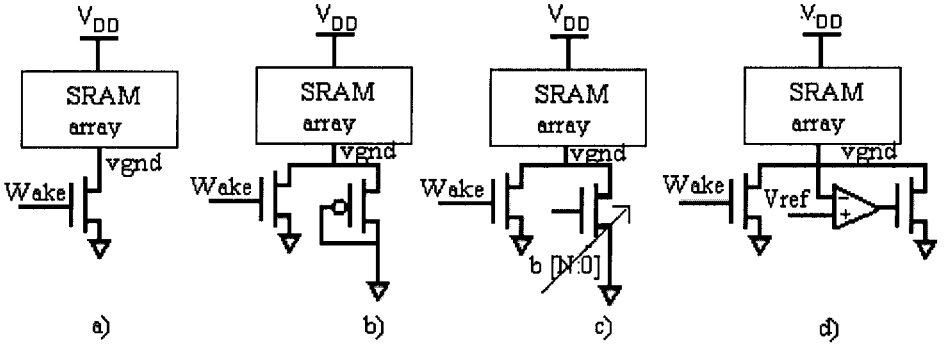


Figure 2-8 Various vgnd control schemes for sleep-transistor design. (a) Sleep-transistor only. (b) Diode-connected PMOS bias transistor. (c) Programmable bias transistors. (d) Active feedback with op-amp based control [33].

2.4.4 Pulse operation

In conventional SRAM, all control signals (i.e. WL, CS, etc) have full swing operations and 50% duty cycles. This draws some unnecessary power consumption. In fact, low-power SRAM designs can utilize smaller-than-50%-duty cycle and half-V_{DD} pulse to reduce the power dissipated from these heavily capacitive lines [37, 71]. Positive and negative pulses can be used concurrently to overcome the reduced gate-overdrive at the receiver ends of the signals [37, 72]. However, this technique is not very widely used in the sub-100 nm technologies since the corresponding power supply is scaled to near 1V, half of which is very close to the threshold border, resulting in an extremely weak current drive-ability.

2.4.5 Dual- and multiple-threshold schemes

Dual- or multiple-threshold processes allow different MOSFET devices on the same die to have different threshold voltages. This enables the use of fast, i.e. low threshold, devices in circuitries where speed is more important while slow devices are used in the circuitry where power saving is critical, thus overcoming the trade-off between speed and power consumption [44, 73]. Several works on SRAM cell designs have been reported utilizing dual threshold process [74-84] to reduce leakage current while maintaining cells' performance. As a result, SNM and the read speed can be enhanced by up to 87% and 17%, respectively, as compared to the conventional 6T SRAM circuits [84]. The leakage and the write power consumption of the dual-threshold SRAM circuit can also be reduced by up to 66% and 35% [84]. The main limitation of this approach is that multiple-threshold processes are not always available and usually cost more than single-threshold processes.

2.4.6 Active write power reduction using write-assisting schemes

Write-assist SRAM designs usually weaken the cross-coupled structure in the 6T bit-cell during a write cycle, making it easier to be written. As a result, less-than-full-swing BLs and DLs are needed to perform the write operation, resulting in a significant amount of power saving [85-90]. This can be done by adding one additional MOS device into the cross-coupled loop [89] or by raising the dynamic ground of the cell to near-supply potential [86]. Sakurai T. *et. al.* reported a 90% write power saving in a write-assisted SA SRAM cell [91-92]. One shortcoming of these approaches is that they also weaken the half-accessed cells in the same row and these cells may be written accidentally.

2.4.7 Active read power reduction using low-power SA

As mentioned in **Section 2.4.3**, SAs are effective in reducing the total SRAM power dissipation. Numerous current-mode SA designs have been proposed to enhance the power performance of the SRAM [93-112]. Since SAs have the ability to amplify small

voltage/current signals on the BLs and DLs and rapidly produce full swing outputs, power dissipation during the read operation of an SRAM can be greatly reduced [37], i.e. the $I_{peripheral}$ component in Eq. (2.6). Furthermore, the I_{array} component in the same equation can also be lowered by using a fast SA coupled with an ATD circuit to reduce the pulse width of the WL signal.

2.4.8 Leakage reduction by redesigning memory cell

Scaling has allowed more transistor counts per die and increases leakage at an exponential rate, making power a primary constraint in all integrated circuit designs. Future designs must address emerging leakage components due to direct band to band tunneling, through MOSFET oxides and at steep junction doping gradients. In SRAM design, leakage currents mainly come from the memory cells since they are in the standby mode most of the time. **Fig. 2-9** illustrates the leakage components in a 6T memory cell storing a '0'. High-threshold and thick gate oxide transistors can be used to maintain low-leakage current from the memory array. Nonetheless, they jeopardize the SNM of the memory cell and thus are not favorable in low supply applications. Alternative solutions including novel cell topologies, negative biasing, reversed-biasing and dynamic V_{th}-control have been proposed to overcome these challenges.

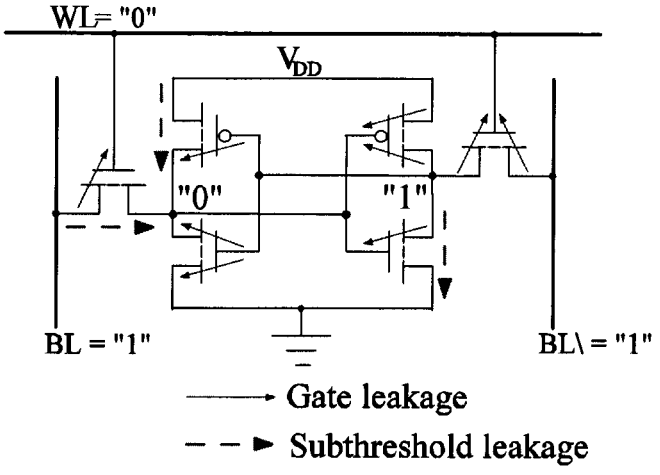


Figure 2-9 Sub-threshold and tunneling gate leakage of an SRAM cell storing "0"

For example, Jain S. K. *et. at.* [113] explored the strong bias towards "0" in the stored data in the cache to propose an asymmetric 8T memory cell that has low leakage when it stores a '0'. Although its leakage when storing a '1" is higher than that of a 6T cell, the net leakage in the memory array is still reduced by more than 20% [113]. Similarly, Azizi N. *et. al.* selectively assigned high-threshold to some transistors and normal threshold to the rest of the 6T SRAM cell to reduce its leakage by 40x when storing a '0' [114-115]. Tawfik S. A. *et. al.* [84] proposed a 7T, dual V_{th} SRAM cell that has separate read and write ports to enhance its stability and reduce its leakage current. As a result, the proposed cell offers more than 40% leakage reduction in a standard 65 nm CMOS process. The authors also proposed a dynamic WL scheme which has low voltage swing during a read and full swing during a write to enhance the stability of the cell. Thus, transistor size can be kept minimum, resulting in a 51% leakage reduction [55]. A portless 5T SRAM cell was reported by Wieckowski M. *et. al.*, reducing the cell leakage by 6x [116]. Kulkarni J. P. *et. at.* [48] demonstrated a 160 mV robust Schmitt-Trigger based sub-threshold SRAM with 18% leakage reduction using a 0.13 μm CMOS process. This design has more transistors stacking in series from V_{DD} to ground and so it has lower leakage current when compared to its 6T counterpart. Amelifard B. *et. al.* dedicated their works to minimize the leakage dissipation of the 6T SRAM cache without incurring any area, delay or design flow overhead. This work deploys different cell topologies at different physical locations in a dual-V_{th} dual-T_{ox} process to reduce cell leakage while maintaining its performance. As a result, total leakage power of the cache is reduced by more than 30% in a 65 nm CMOS process [117]. Zhiyu L. *et. al.* illustrated a novel 9T cell with negative WL scheme to suppress the cell leakage [2]. Zhang K. *et. al.* recently summarized several important low-power techniques to reduce both dynamic and leakage currents in the nano-scale CMOS technologies [33].

2.5 Cell stability and data retention limit in SRAM

Intra-die and inter-die variations result in device parameter mismatches such as V_{th}, L, and W. These mismatches lead to the following failures in memories [118]: 1) Data retention failure e.g. a cell fails to store its value. This happens when V_{DD} is too low, cell mismatches are too severe or a cell is disturbed by a significant amount of noise that exceeds its margin. 2) Read failure e.g. a cell fails to keep its original value and flips during the read operation. 3) Write failure e.g. unsuccessful write operation due to the deviation of the strength of the access devices and the trip point of the cross-coupled inverters. 4) Access failure: e.g. a decrease in the cell current, BL capacitance mismatch, and excessive cell leakage along the BLs or the mismatches of the devices in the SA may cause wrong evaluation in the SA.

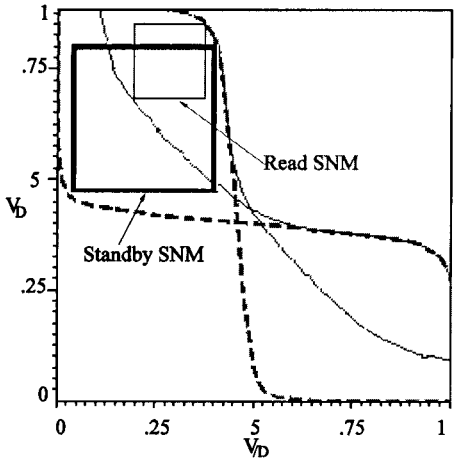


Figure 2-10 Graphical representation of SRAM SNM

Static noise margin (SNM) and Write Trip Point (WTP) are the two most commonly used metrics to measure SRAM reliability. SNM indicates how stable a cell is. It is defined as the worst case noise level present at the gates of the inverters that does not cause the cell to flip. Therefore, SNM is normally associated with the Read operation and it is desirable to have as high SNM as possible. SNM is visually equal to the largest square that can fit into the eye of the butterfly curve formed by the voltage transfer characteristic curves of the inverters [119], as shown in **Fig. 2-10**. From now on, we refer to the cell's

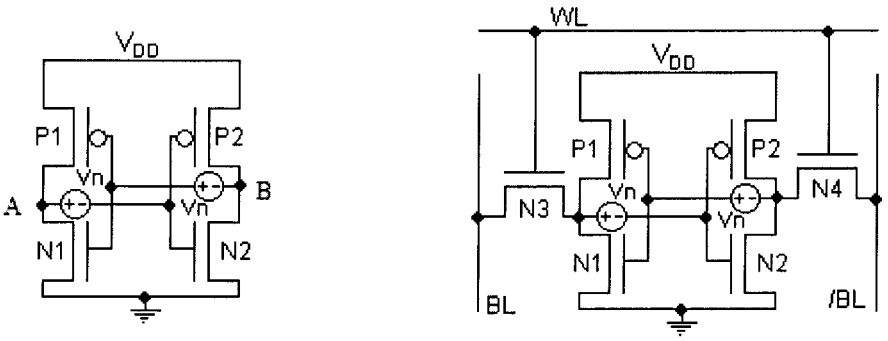
SNM as its worst-case SNM. In the case of 6T design, it is the Read SNM. In the proposed design, as will be shown later, its SNM is stand-by SNM.

WTP on the other hand, measures how easy it is to write into a cell. When a cell is to be written, one of the BLs (e.g. /BL) remains at V_{DD} while the other (e.g. BL) is pulled to ground. WTP is defined as the highest BL potential that causes the cell data to flip successfully. It is also preferable to have as high WTP value as possible. A high WTP indicates that the cell can be written to easily and less voltage swing on the BL is required.

Conflicting nature of SNM and WTP originates from the fact that the access transistors access the cell in the same manner during a read and a write operations but require two opposite outcomes. While accessing in the read mode requires the cell to be stable enough to hold its data, the write operation needs it to be as weak as possible to flip the data. Therefore, improving one factor will unavoidably jeopardize the other. As a result, several novel SRAM cell designs have been proposed to tackle this issue, most of them open a separate port to perform the Read operation while the original 6T design is used for data retention and write operation only.

2.5.1 Static Noise Margin and V_{DDmin}

JEDEC dictionary defines noise margin as the maximum voltage amplitude of extraneous signal that can be algebraically added to the noise-free worst-case input level without causing the output voltage to deviate from the allowable logic voltage level [120]. This concept has been adopted by C. F. Hill [121-122] and J. Lohstroh [122] to define the worst-case static noise as DC disturbance which is adversely present in all logic gates in an infinitely long chain of gates [123]. J. Lohstroh later pointed out that the behavior of the output nodes in the chain, as the number of gates approaches infinity, is equivalent to a steady state of a flip flop [123]. As a result, the author developed a discussion on the worst-case SNM of a flip-flop by adding two DC noise voltage sources into the flip-flop loop in an opposite manner, as shown in **Fig. 2-11 (a).**



(a) (b)

Figure 2-11 Standard circuit set-up for defining the noise margin. (a) flip-flop (b) 6T SRAM cell.

Assume that in **Fig. 2-11 (a)**, node A stores a "1" and node B stores a "0", resulting in voltage potentials at nodes A and B to be V_{DD} and 0 respectively. If noise sources are not present, i.e. $V_n = 0$, then voltages of nodes C and D are V_{DD} and 0, respectively. Hence, the inverter P1- N1 has a strong "0" input and its output, i.e. node A, is kept at a strong "1". Similarly, node B is kept at a strong "0". However, if some noises are present, i.e. $V_n > 0$, voltage at node C is lower than V_{DD} and voltage at node D is higher than 0. Consequently, weak "1" and "0" are applied to the inputs of the inverters (P2, N2) and (P1, N1), respectively. The maximum DC noise source V_n below which the flip-flop can still maintain its data is defined as the SNM of the flip-flop [123]. Several criteria have been proposed to determine the worst-case SNM such as 1) Coincidence of roots of the flip-flop equation [124]; 2) Small-signal closed-loop gain is unity [125-126]; 3) Jacobian of the Kirchhoff Equations is zero [127]; 4) Maximum square between normal and mirrored voltage transfer characteristic curves [121]. These criteria have been proven to be equivalent in [123]. E. Seevinck *et. al.* took a further step to develop an analytical model to determine the SNM of a 6T SRAM cell, with DC voltage noise sources connected as shown in **Fig. 2-11 (b)**. Since the 6T cell is more susceptible during the read access, its SNM is defined as the maximum DC noise bearable by the cell when its WL, BL and /BL are kept at V_{DD} [119]. As a result, its SNM is:

$$SNM = V_{th} - \left(\frac{1}{k+1}\right)\left\{\frac{V_{DD} - \frac{2r+1}{r+1}V_{th}}{1+\frac{r}{k(r+1)}} - \frac{V_{DD} - 2V_{th}}{1+k\frac{r}{q}+\sqrt{\frac{r}{q}\left(1+2k+\frac{r}{q}k^2\right)}}\right\} \quad (2.8)$$

Where

$$r = cell\ \text{ratio} = \frac{\beta_d}{\beta_a}$$

$$q = \frac{\beta_p}{\beta_a}$$

V_{th} = *threshold voltage*

$$k = \left(\frac{r}{r+1}\right)\left\{\sqrt{\frac{r+1}{r+1-V_s^2/V_r^2}} - 1\right\}$$

$$V_s = V_{DD} - V_{th}$$

$$V_r = V_s - \frac{r}{r+1}V_{th}$$

It can be seen that SNM is dependent only on the threshold voltage (V_{th}), V_{DD} and the β ratio r. Furthermore, increasing r will increase the cell stability. Therefore, 'r' must be maximized to optimize the SNM. Nevertheless, it is constrained by the cell area and the reliability of the write operation [119, 128-129].

The expression presented in Eq. (2.8) is based on the assumption that all transistors are matched and both NMOS and PMOS have the same threshold voltage, V_{th}. Therefore, it cannot be used to study the effect of parameter variations of the transistors in the cell. Ichikawa T. *et. al.* [130] developed a new analytical model of SRAM cell stability in low-voltage operation and for the first time, mutual effects of cell-parameter variations have been clarified. This work assumed that drain currents from the pull-up PMOS devices are negligible, resulting in a cross-coupled structure of two enhance-enhanced inverters during the read operation, as shown in **Fig. 2-12** [130].

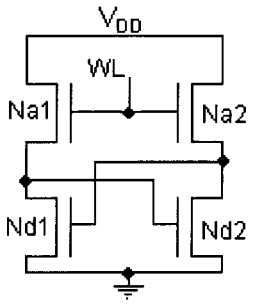


Figure 2-12 Circuit schematic of the ideal SRAM cell with assumption that drain currents from the pull-up pMOS devices are negligible

In order to describe the transfer characteristic of the inverters, the alpha-power law has been used to model the I-V relationship of the inverters in consideration [130-131]. Thus, nMOS drain current is expressed as:

$$I_D = \begin{cases} \frac{1}{2}\beta(V_{GS} - V_t)^{\alpha} & (V_{GS} >= V_{th} : \text{saturation region}) \\ 0 & (V_{GS} < V_{th} : \textit{cutoff region}) \end{cases} \tag{2.9}$$

$$V_{th} = V_{to} - \gamma V_{BS} \tag{2.10}$$

Using these equations in the low-voltage supply region, assuming all transistor pairs are matched, the corresponding SNM of the ideal cell was obtained as:

$$\text{SNM} = \frac{1}{1+\gamma_a + \left(\frac{\beta_d}{\beta_a}\right)^{\frac{1}{\alpha}}} \left\{ \frac{1}{1+\gamma_a}\left(\frac{\beta_d}{\beta_a}\right)^{\frac{1}{\alpha}} - 1 \right\} \left[V_{DD} - \{(1+\gamma_a)V_{tod} + V_{toa}\} \right] \tag{2.11}$$

Where V_{tod} and V_{toa} are the threshold voltages of the nMOS devices denoted as N_d and N_a, respectively in **Fig. 2-12**.

Eq. (2.11) shows that SNM of the SRAM cell is only dependent on the intrinsic parameters of the transistors and the supply voltage V_{DD}. Since SNM must be positive for the cell to be stable, the lower limit of the supply voltage, V_{DDmin} is:

$$V_{DD\,min} = (1+\gamma_a)V_{tod} + V_{toa} \tag{2.12}$$

$$\frac{1}{1+\gamma_a}\left(\frac{\beta_d}{\beta_a}\right)^{\frac{1}{\alpha}} > 1 \tag{2.13}$$

A zero SNM means that the cell is on the verge of instability and hence its supply voltage cannot be reduced any further. Eq. (2.13) on the other hand suggests that the access and driver devices must be sized so that the driver device is strong enough to hold the stored data. With a fixed V_{DD}, the stronger the cell ratio, i.e. $r = \frac{\beta_d}{\beta_a}$, the higher its SNM and thus, the more stable the cell. This agrees with what has been reported before in [119]. Based on this result, similar analytical steps were used to form the SNM expression in a non-ideal cell:

$$SNM = \frac{1}{1+\gamma_{a2}+r_\beta}\left\{\left(\frac{r_\beta}{1+\gamma_{a1}}-1\right)V_{DD} + (1+\gamma_{a2})V_{tod1} - r_\beta V_{tod2} - \frac{r_\beta}{1+\gamma_{a1}}V_{toa1} + V_{toa2}\right\} \tag{2.14}$$

$$r_\beta = \left(\frac{\beta'_{d2}}{\beta_{a2}}\right)^{\frac{1}{\alpha}} + \frac{1}{2}(1+\gamma_{a2})\beta'_{d2}R_{N2} \tag{2.15}$$

$$\beta'_{d2} = \frac{1}{1+\frac{1}{2}(1+\gamma_{a2})\beta_{d2}R_{S2}}\beta_{d2} \tag{2.16}$$

Where r_β is the practical β-ratio of the inverter 2, R_S and R_N are the parasitic resistance at the sources of the driver and the access transistors, respectively. Similar to the ideal case, the condition $\left(\frac{r_\beta}{1+\gamma_{a1}} > 1\right)$ must be satisfied to ensure a positive SNM. The corresponding V_{DDmin} was derived as:

$$V_{DDmin} = \frac{1}{\left(\frac{r_\beta}{1+\gamma_{a1}}-1\right)}\left\{-(1-\gamma_{a2})V_{tod1} + r_\beta V_{tod2} + \frac{r_\beta}{1+\gamma_{a1}}V_{toa1} - V_{toa2}\right\} \tag{2.17}$$

By differentiating the V_{DDmin} obtained in Eq. (2.17) with respect to different coefficients, one can investigate the mutual effects and sensitivity of each cell parameter. Detailed discussions on this matter have been reported in [130]. This model is very useful in explaining the effects of the cell parameters on V_{DDmin} of the cell and hence can be used to predict the V_{DDmin} of the designed circuits.

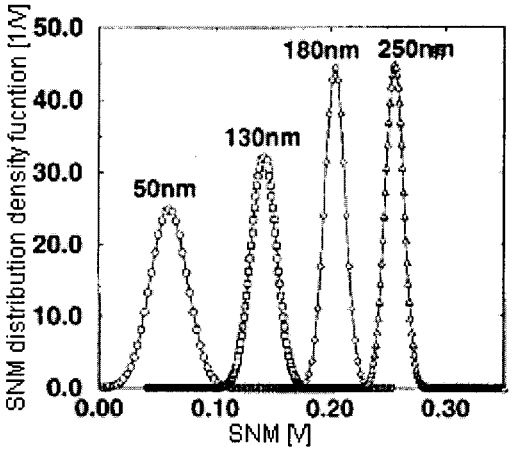


Figure 2-13 Projected PDF of SNM due to intrinsic threshold voltage fluctuations in all cell transistors [128]

As CMOS technologies advance into the sub-100 nm regime, intrinsic parameter fluctuations become excessively high and hence cell SNM is reduced due to threshold variations in uniformly doped minimum-geometry devices. Bhavnagarwala A. J. *et. al.* [128] investigated the impact of intrinsic device fluctuations on CMOS SRAM cell stability. Compact physical and stochastic models of the SNM have been proposed for the first time. Probability distribution function (pdf) of the SNM of the cell has also been derived [128]. As a result, SNM distributions have been projected for sub-100 nm technology generations, as shown in **Fig. 2-13**. This statistical approach is critical in sub-100 nm design since nominal SNM does not make any sense.

2.5.2 Write failure and Write Trip Point

Besides read stability for the SRAM cell, write-ability is equally important to guarantee a correct write operation without spending too much energy on pulling down the BL voltage to zero [129]. WTP is a commonly used metric to measure the cell's ability to write [129, 132]. It is defined as the maximum voltage on the BL that the cell content can be flipped successfully, assuming that the other BL is kept at V_{DD}. The WTP is mainly determined by the strength of the access transistors, or more specifically, the pull-up ratio of the cell. The stronger the access transistor compared to the drive transistor, the higher the WTP, i.e. the easier it is to write into the cell. Ironically, this weakens the cell ratio (**Section 2.3.1**) and hence jeopardizes the cell's SNM. This results in the well-known read-write conflicting design criteria of SRAM cell [129]. **Fig. 2-14** illustrates a WTP of 0.25 V of a typical 6T SRAM cell.

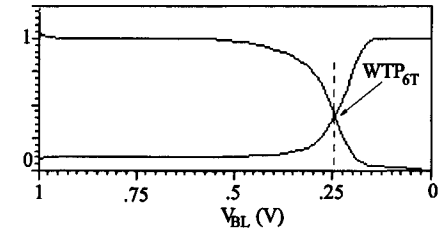


Figure 2-14 WTP simulation waveforms

In a conventional 6T SRAM design, one of the BLs is pulled to zero while the other is kept at V_{DD} during a write operation [37]. Since the trip-point voltage is usually slightly less than $V_{DD}/2$, cell data can be written with little difficulty. However, in a high-speed, high-capacity cache, write failures may occur in the following scenarios: 1) The WL pulse is too short and hence the access transistors are turned off before the cell data flips. 2) The write driver is not strong enough to pull the heavily-loaded BL and DL to zero. 3) Cells are not sized properly and hence its write-trip point is too low. 4) Timing mismatches cause the cell to be activated long before or after the BL is pulled to zero, resulting in unsuccessful write operation.

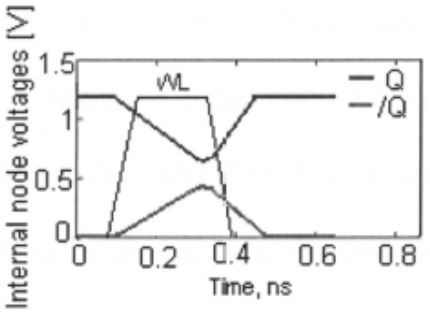


Figure 2-15 Write failure mechanisms of a 6T SRAM cell

Fig. 2-15 illustrates a failed write cycle in which the WL pulse is too short to flip the memory content. These issues must be addressed properly in designing steps to guarantee the fabrication yield of the cache.

2.5.3 Dynamic noise margin

Although SNM is simple and effective in predicting SRAM cell's failures, the model is conservative. Furthermore, this model is based on the static property of the cell which in a real can hardly be archived because read/write operation of the cell is normally very short. Recently, several works have been published reporting that Dynamic Noise Margin (DNM) is a more accurate measurement of the cell stability [133-135]. While SNM simulation assumes that both Read and Write operations of the cell are long enough so that it can settle down at final static state, DNM simulation considers the cell directly at its real read/write operation. This leads to some differences in these two metrics.

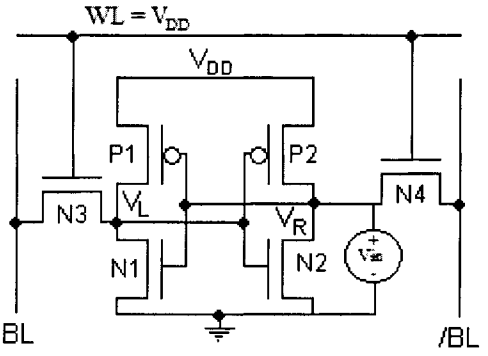
It has been reported that DNM is normally higher than SNM [133]. One can recall that a cell is the most prone to error when the WL is turned on and the data storing nodes are disturbed. If the cell is not stable enough, the node storing a "0" will be pushed to "1" while the node storing a "1" will be pulled to "0". However, because of the parasitic capacitances at these nodes, the cell takes time to flip [133]. As a result, the longer the read duration, the more easily cell flips. This

shows that SNM is somewhat too pessimistic. For example, if the cell is not stable enough but needs 5 ns to flip, it will definitely flip under SNM consideration. However, stored data will be maintained if the WL is only turned on for 1 ns. This explains why under the DNM's perspective, the cell can tolerate a higher noise level as the WL pulse width is shortened[133]. Nonetheless, this model is more time consuming when compare to SNM. Furthermore, both metrics will return the same result if relative noise margin of two cells are to be accessed. As a result, in this thesis, we only consider SNM as a measurement to compare the stability of two different SRAM cells.

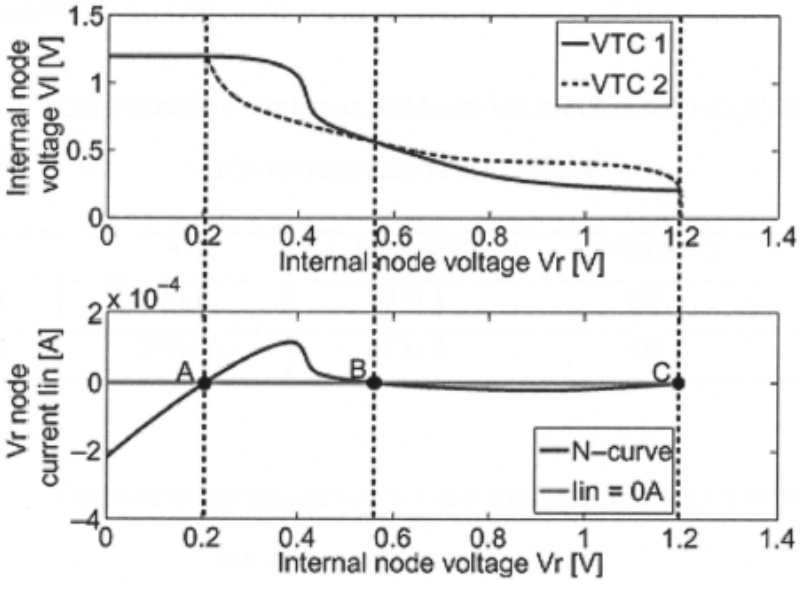
2.5.4 N-Curve as a new metric to measure SNM and WTP

Two drawbacks of the SNM are the inability to be measured with automatic inline testers and the inability to generate statistical information on SRAM [129, 136]. Alternatively, the SRAM "N-curve" provides a way to satisfy both needs [129, 136]. Furthermore, the N-curve metric contains information on both read stability and write ability, thus allowing a complete functional analysis of the SRAM cell [19, 129, 137]. **Fig. 2-16** shows the comparison between the butterfly curves and the N-curve of a SRAM cell. A circuit setup as shown in **Fig. 2-16(a)** is used to extract the N-curve which contains both voltage and current information. A voltage sweep V_{in} from 0V to V_{DD} is applied at node V_r which is at zero potential and the corresponding current I_{in} is measured, forming the N-Curve. At three points A, B and C on the N-curve, the current injected to node V_r is zero. The three points A, C, B are the two stable points and the meta-stable point, respectively **(Fig. 2-16 (b))**. The voltage difference between points A and B indicates the maximum tolerable voltage at V_r of the cell before its content changes, i.e. the static voltage noise margin (SVNM). The peak current between points A and B is the static current noise margin (SINM), which refers to the maximum value of the static current that can be

injected into the SRAM cell before its content flips [129]. SVNM and SINM are now combined to convey a complete definition of the noise margin of the cell, i.e. its stability.



(a)



(b)

Figure 2-16 Circuit setup to extract the N-Curve during the read operation. (b) Corresponding butterfly curves (upper) and N-curve (lower) [129]

The SRAM N-curve can also be used to evaluate the write-ability of the cell. For a write operation, pulling down the BL to ground discharges the '1' node V_l. Therefore, the N-curve is now analyzed from the right to the left. At point C, the internal node V_r is 1. The negative current peak between points C and B, or the Write-Trip-Current (WTI) is the amount of current needed to write the cell when both BLs are kept at V_{DD} [129]. The

voltage difference between points C and B, or the Write-Trip-Voltage (WTV), is the voltage drop needed to flip the internal node '1' of the cell with both BLs clamped at V_{DD}.

Tables I and II confirm the equivalence between the conventional SNM and WTP metrics and the corresponding SVNM and WTV. In both cases, the transistors in the memory cell have the same aspect ratios. However, the transistors in the second case have the channel width and channel length two times larger than those in the first case. As shown in these two tables, SNM and WTP of the two cases are the same and so are their SVNM. However, the transistors in the second case are twice as strong in terms of SINM and WTI. Determining the read stability and write ability therefore requires both voltage and current information [129].

TABLE I. COMPARISON OF THE USUAL SNM AND THE N-CURVE METRICS SVNM AND SINM FOR TWO DIFFERENT CELL DESIGNS

	W/L	SNM (MV)	SVNM (MV)	SINM (MA)
CASE1	W_1/L_1	253.1	429	0.296
CASE2	$2W_1/2L_1$	253.5	428	0.586

TABLE II. COMPARISON OF THE USUAL SNM AND THE N-CURVE METRICS SVNM AND SINM FOR TWO DIFFERENT CELL DESIGNS

	W/L	WTP (mV)	WTV (mV)	WTI (μA)
Case1	W_1/L_1	366.5	630.4	-93.8
Case2	$2W_1/2L_1$	366.6	631	-186

Similar to the previous approach, analytical model for the N-curve is derived by equating the transistor currents at the two storage nodes. Grossar *et. Al.* provided both analytical expression of the N-curve metrics as well as statistical approach to optimize the performance of the cell in [129]. Samon *et. at.* [137] investigated the N-curve metrics for sub-threshold operations in 65 nm CMOS technologies. It has been concluded that N-curve metrics are better even under lower supply voltage conditions since the current

metrics provide more information regarding read stability and write ability issues and enable a complete functional analysis. However magnitudes of SINM and WTI degrade at very low sub-threshold power supply voltages, higher cell ratio of SRAM cell and higher oxide thickness in sub-threshold region. These observations are valuable for designing ultra-low power cache using sub-65 nm technologies and sub-threshold power supply voltages.

2.6 SRAM cell designs

2.6.1 Overview

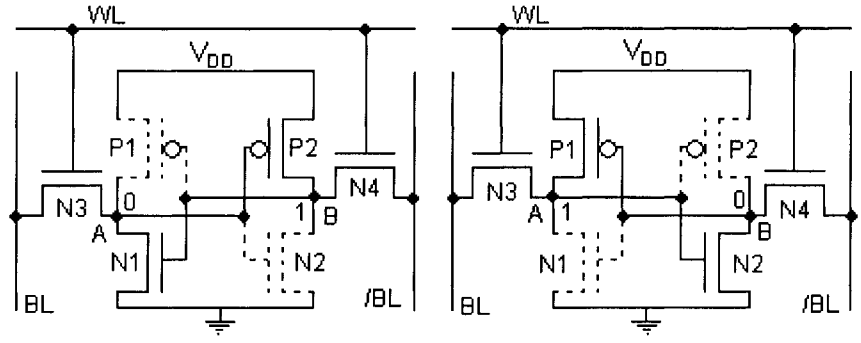
During the last decade, researchers have tried to develop solutions to overcome CMOS technology scaling issues. As the impact of technology scaling and its related problems propagate to circuit level, numerous state-of-the-art SRAM cell designs have been reported, mainly to tackle the issue of sub-threshold leakage current and the stability of the memory cell under the stress of low supply voltage and high process variation [19]. Although most of them share the same storage structure, each has distinctive read/write access that either enhances its stability or reduces its power consumption. In this chapter, we turn our attention to several contemporary SRAM cell designs, their operations and characteristics.

2.6.2 Conventional 6T cell

Two of the most common SRAM storage structures are the Six-Transistor (6T) and Four-Transistor (4T) cells [37]. In the past, the 4T cell dominated the stand-alone market thanks to its small area. However, as technology scales down, the 6T SRAM cell (**Fig. 2-17**) has gained popularity among circuit designers because of its stability, low power consumption and simple fabrication process [37, 138-145]. The 6T cell is made up of a flip-flop formed by two cross-coupled inverters (M1, M3 and M2, M4) and two pass-gate-transistors (M5 and M6). The pass-gate transistors, which are connected to the two

complementary BLs are controlled by the WL. They act as access devices during read and write operations of the SRAM [19, 29, 32, 37].

During standby, the pass-gate transistors are turned off by the WL signal to isolate the storage structure from the BL disturbances. The strong positive feed-back structure of the flip-flop allows the SRAM cell to retain its complementary data as long as the power supply is on [19, 28, 32, 37]. Therefore, no refresh cycles are required, making the SRAM faster and consume less power when compared to DRAM [32].



(a) (b)

Figure 2-17 Conventional 6T design (a) storing a "0". (b) Storing a "1".

During a read cycle, the WL signal is triggered high (**Fig. 2-18**), turning on the two pass-gate devices. As the BLs are pre-charged to the same potential prior to each read cycle, and the cell contains complementary data at its store nodes, this causes a differential signal to be induced on the BLs, also shown in **Fig. 2-18**. This signal is subsequently sensed and amplified before transferring to the output of the SRAM. When a cell is written, one of the BLs is charged to V_{DD} while the other is discharged to 0. Concurrently, the pass-gate transistors are turned on to force one of the store nodes (i.e. A or B) to V_{DD} and the other to 0 [19, 29, 32, 37].

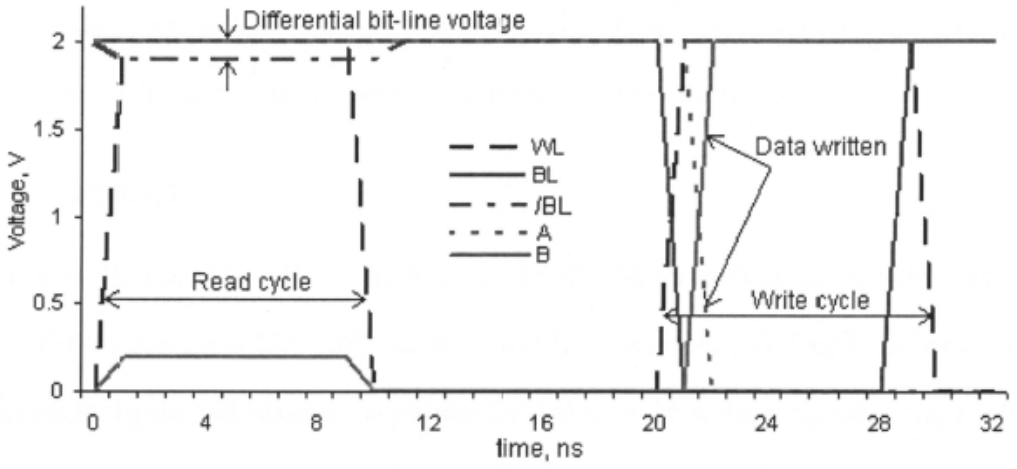


Figure 2-18 Waveforms of several nodes during a read and write cycle of the SRAM.

The 6T cell has the following advantages: 1) It has low-power consumption since both inverter branches are cut-off once the flip-flop reaches the bi-stable stage. 2) Its structure is symmetrical and hence has a very compact layout. 3) It offers differential read/write, which results in very fast and reliable operations. However, it consists of 6 transistors and is considered bulky when compared to the 1 Transistor- 1 Capacitor DRAM or the 4T SRAM. As technology scales down to sub-100 nm feature size, this shortcoming is alleviated but unfortunately, its advantages also diminish. Firstly, the leakage current in the sub-100 nm is substantial and hence accumulate to a significant amount of power consumption in the SRAM. Secondly, low-voltage operation in the state-of-the-art technologies degrades the noise margin of the inverters and hence the stability of the SRAM cell. Thirdly, small transistors in the SRAM cell can only sink a limited current during a read/write cycle. This, couples with the rapidly increasing leakage current along the BLs, leads to slow and unreliable read/write operations. As a result, larger transistors must be used to ensure a correct operation of the memory which evidently requires both additional power consumption and silicon area. Countless effort has been made to tackle these issues [13, 17, 19-20, 25, 29, 32, 37, 41, 53, 146-192], each of which focuses on area, power, speed or reliability improvement. In the next sections, we are going to explain

their working principles, advantages and drawbacks. Detailed operation and design steps will not be discussed but can be found in the respective references.

2.6.3 1T SRAM cell

The single transistor bit cell used in the 1T-SRAM technology [192] attains much higher density since it is 25% - 35% smaller than the conventional 6T SRAM. It comprises an access transistor and basically is a planar DRAM with the storage capacitor realized by using a MOS structure and a capacitor, as depicted in **Fig. 2-19**. It alleviates the issue of process incompatibility and in the meantime offers the SRAM-like interface and high-performance characteristics often associated with traditional SRAMs. However, the main issues of stability and the need for refreshing render it not as popular as the 6T SRAM.

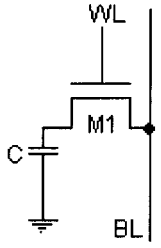


Figure 2-19 1T SRAM cell implementation

2.6.4 Loadless 4T SRAM cell

The concept of 4T SRAM was first introduced in 1987. However, either a specialized WL driver circuit or additional photo masks are required in the fabrication process [37, 193-196]. In addition, it depends on the leakage mechanism of the access transistors to hold data [196-197], which incurs extra power consumption and may not be suitable for the future technologies. The 4T SRAM however has at least 20% area reduction when compared to the bulky 6T counterpart [194, 196], as shown in **Fig. 2-20 (a)**. Furthermore, they have a higher SNM and a lower sensitivity to V_{th} fluctuations [198]. To date, several 4T SRAM structures have been reported in literature [165, 194-196, 198-207], three of which [194-196, 203, 207] are presented in **Fig. 2-20**. Their read/write operations are more

or less the same as that of the 6T SRAM cell with a slight difference in biasing the control signals [194].

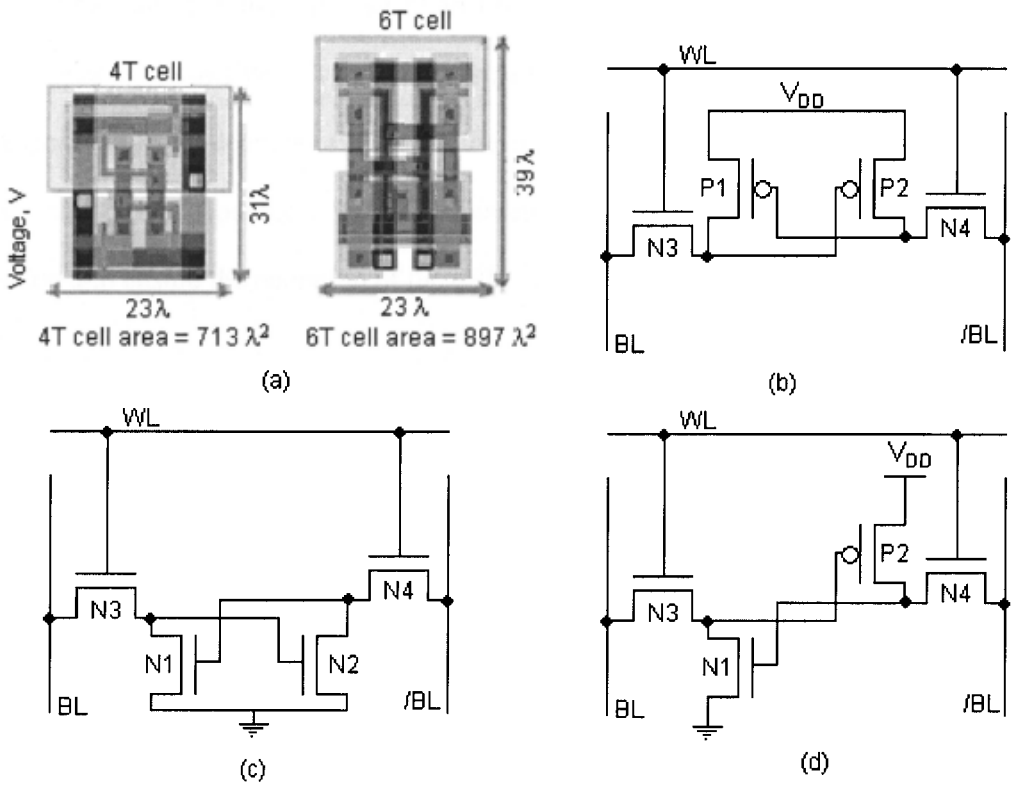


Figure 2-20 Loadless 4T SRAM cells. (a) Layout comparison between 4T and 6T cell. (b) N-type access. (c) P-type access. (d) zero-aware type [196]

2.6.5 5T SRAM cell

The first conventional 5T SRAM cell was introduced in 1988 by Yang *et. al.* [208] where one of the nMOS access transistors in the 6T cell is replaced with a BJT. True single ended 5T cell as shown in **Fig. 2-21 (a)** is then revised in [209] and [210]. This approach eliminates the complementary access transistor and BL, thus, evidently reducing the total silicon area (23% [210]) associated with the BL wiring and the fifth transistor. It also benefits from less dynamic and leakage power [210]. However, its asymmetric structure results in unreliable read/write operations in the sub-65 nm regime where excessive process variations are present. Furthermore, the asymmetrical structure does not allow it to scale very well with the state-of-the-art process.

Wieckowski *et. al.* proposed a symmetrical 5T (**Fig. 2-21 (b)**) cell which is claimed to have both a higher SNM and a lower power consumption [116, 211-212]. Although this approach has only 5 transistors, its cell area is almost the same as the conventional 6T cell [116]. Besides, its active cell current is smaller than that of the 6T cell and thus, the corresponding I_{on}/I_{off} degrades as V_{DD} and device size scale down. This apparently leads to a less reliable read operation. Therefore, careful technology-aware design steps must be considered to ensure the operation of the cell.

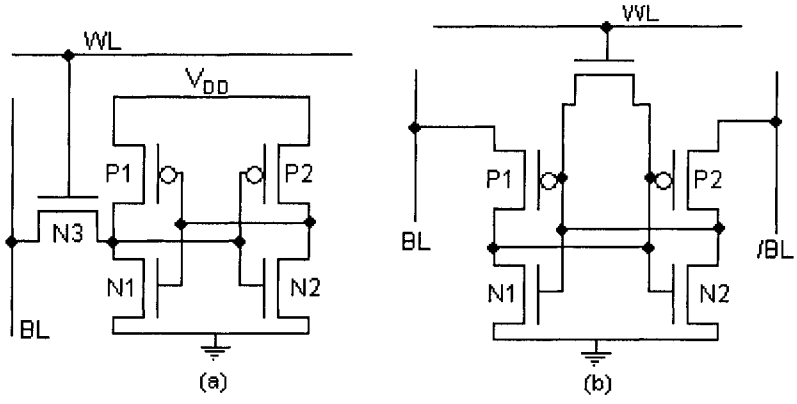


Figure 2-21 5T SRAM cell (a) single-ended cell (b) port-less cell

2.6.6 7T SRAM cell

The above-mentioned SRAM cell designs attempt to eliminate one or more transistors from the conventional 6T cell and thus obtaining smaller cell area. However, as mentioned in the previous chapter, other crucial concerns include cell stability and write/read margin which are being severely degraded along with the rapidly increasing process variations and steep V_{DD} downscaling. This urges for another trend in SRAM design in which cell stability and power consumption (both dynamic and leakage) are placed at the highest priority [13, 19, 25]. From this section onwards, we are going to introduce some more-than-6T SRAM designs which are able to work at sub-threshold supply voltage and hence, significantly reduce total power consumption of the cache.

Several typical 7T SRAM cells are depicted in **Fig. 2-22**. Although these designs appear to be different, they share a similar storage topology as the 6T cell. In [84] and

[85], the authors proposed a SRAM cell which has different read and write ports, as shown in **Fig. 2-22(a)**. As a result, both its read and write operations are single ended and an additional WWL signal is required. During a write operation, cell data are driven into the WBL and written into node A through the pass-gate transistor N3. On the other hand, in a read cycle, N5 is turned on by the RWL signal and the cell current will flow from the RBL to ground through N5 and N4 if node B is high. Otherwise, no cell current is available and the RBL remains at its pre-charged level. Data are then sensed depending on the changes of the RBL. Its area is 16% larger than the 6T cell [84]. In exchange, it benefits from both lower power consumption and higher stability as the read- and write- ports are separated and optimized independently. This design is almost the same as the 8T cell (which will be discussed in the next section) but has slightly smaller area.

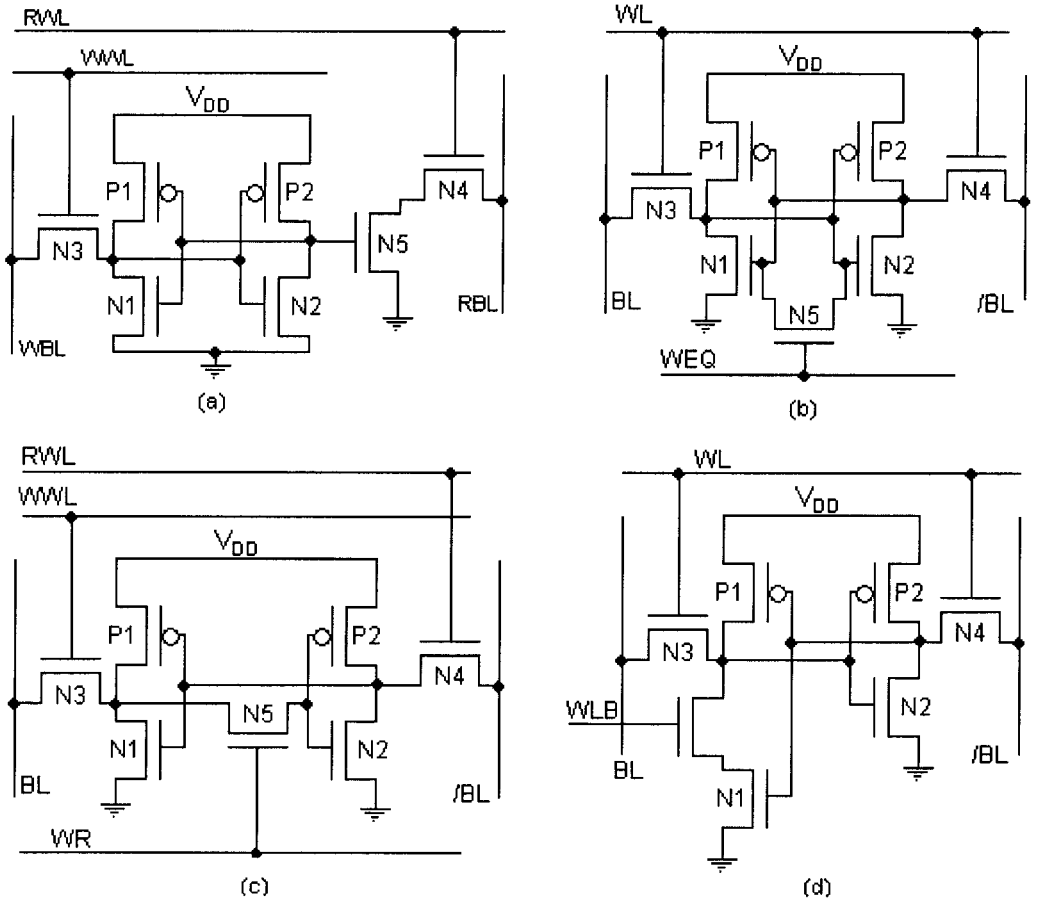


Figure 2-22 7T SRAM cell. (a) dual-port asymmetrical. (b) single-port. (c) decoupled cell. (d) SNM-free.

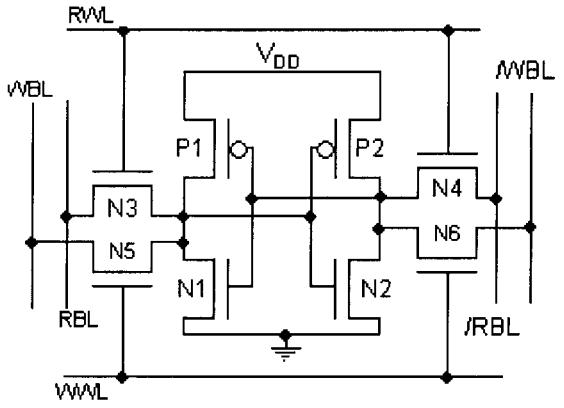
Other approaches such as those depicted in **Fig. 2-22 (b)-(d)** can be found in [213-216]. These designs insert an additional nMOS device into the cross-coupled loop to either assist the write operation [213] or enhance the read-noise-margin [214-216]. Therefore, they have superior performance to the 6T design. However, as discussed in the previous chapter, having the same read- and write- port prevents these designs from optimizing both operations simultaneously. Consequently, it is difficult to scale these designs in the deep sub-micron technologies of which process variations, both inter-die and intra-die, are inevitably becoming a major concern.

2.6.7 8T SRAM cells

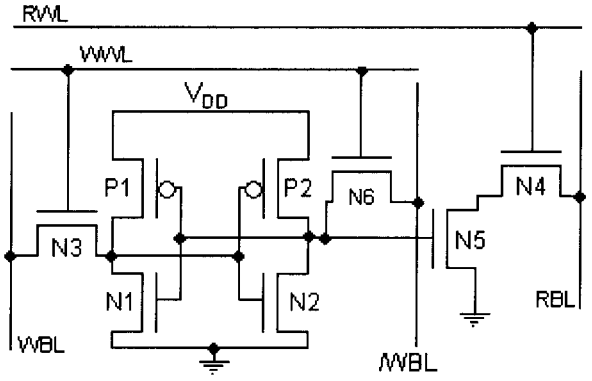
One cannot stress enough the importance of power-saving designs in the era of portable devices [13, 25, 217-220]. The demand for hand-held digital gadgets with long battery-life has motivated circuit designers to take advantage of power supply scaling. However, this trend cannot continue infinitely since the stability and reliability of the memory cell are highly dependent on the supply voltage, which if reduced would lead to a severe degradation in the cell's operation margin. Cache designers therefore must sacrifice some circuit areas to obtain a reasonable margin at a very low supply condition. Fortunately, this is assisted by the rapid miniaturization of the device size.

The 8T cell designs seem to have received the most attention from cache designers, only second to 6T design. It is claimed to have superior performance compared to the 6T cell in sub-65nm CMOS technologies regime [1, 14, 36, 39, 85, 113, 150, 213, 221-236]. Several 8T topologies are available (three of them are presented in **Fig. 2-23** in the next page), which are claimed to be more stable and work better at a very low supply voltage [1, 13-14, 19, 25, 36, 39, 85, 113, 150, 213, 221-236]. Due to the larger number of transistors, 8T cell design's area overhead is about 30% compared to the 6T cell. Nevertheless, they can operate at a sub-threshold supply voltage level [1, 36, 229, 232] and its layout area can be comparable to that of the 6T at 45 nm CMOS process or below [11, 237]. In addition, its write and read operations have a reasonable margin at a very low-supply condition. As a

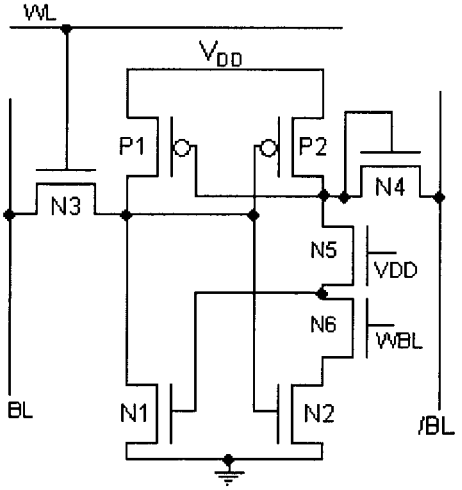
result, it is considered to be best suited for sub-threshold supply applications where performance, power consumption and cell area are all under constraint.



(a)



(b)



(c)

Figure 2-23 8T SRAM designs (a) conventional (b) isolated read gate (c) zero-aware.

Having said so, one must understand that these designs need meticulous considerations and under normal circumstances, several supporting techniques are required to enhance the stability and the bit-error-rate performance. One of the most critical drawbacks of these designs (except that of **Fig. 2-23(a)**) is the single-ended read port, which is highly error-prone due to the mismatches within the SA and the BL leakage.

2.6.8 More-than-8T SRAM cell

As mentioned above, the 8T cell designs suffer from various issues associated with the single-ended read port. To address these issues, one or more transistors have been added to the isolated read gate 8T design [2, 5-6, 47-48, 52, 229, 238-244]. These designs either enhance the stability of the cross-coupled structure or improve the read/write operations through the means of additional MOS devices. As a result, they can achieve a faster operation when compared to the 8T designs [229]. This advantage however comes at a cost of about 20% area overhead when compared to the 8T counterpart [244]. Circuit designers therefore must compromise between area, power consumption and performance so that their designs are best suited for the respective applications.

2.7 Conclusion

This chapter reviews the most frequently used techniques in literature to enhance the performance of SRAM. It reveals that V_{DD} scaling and SRAM macro partitioning are the most effective methods to manage the active power dissipation in SRAM. On the other hand, SRAM leakage is mainly reduced by redesigning the SRAM cell. As a result, memory cell design plays a crucial role in reducing the total power consumption of the chip. Another design concern is to determine the stability of the memory cell under process variations. In order to cope with the continuously increasing device mismatches, different stability metrics has also been analytically proposed to comparatively measure the stability and reliability of the memory cell designs. Finally, this chapter provides a

comprehensive review of the recent SRAM cell developments, which were proposed to solve the power and reliability challenges in nano-scale technologies.

CHAPTER 3 A 10T SRAM WITH IMPROVED SNM AND REDUCED POWER CONSUMPTION

This design aims to reduce the read power consumption as well as to enhance the stability of the SRAM cell during the read operation. A 10-Transisor SRAM cell is proposed with a new read scheme to minimize the power consumption within the memory core. It has separate read and write ports, thus cell read stability is significantly improved. A 16Kb SRAM macro operating at 1V to 0.4 V supply voltage is demonstrated based on a multi-V_{th} 65 nm CMOS process. Its read power consumption is reduced to 24% of the conventional design. The new cell also offers 90% leakage current reduction. Therefore, it is suitable for low-power mobile applications whose power supply is restricted by the battery.

3.1 Introduction

CMOS SRAM memory has been and will continue playing a critical role in modern microprocessors. As previously mentioned, due to its complex 6T structure (**Fig. 3-1**), SRAM cache is one of the most area-consuming components in the state-of-the-art system-on-chip (SoC) [113]. As a result, SRAM cell transistors normally use minimum width-to-length ratios to meet this stringent area constraint. This, coupled with the increasing fluctuations in transistor parameters (e.g. V_{th}) as device dimensions and supply voltage scales down in the nanometer regime, leads to an urgency to increase the cell stability for future technology.

Another major concern is the power consumption of high density SRAMs. Eq. (3.1) and (3.2) show a simplified model of total power consumption of an SRAM which is inclusive of active and passive powers P_{active} and $P_{passive}$, respectively [19]. Assuming that the SRAM macro in consideration has *m* columns and *n* rows, its standby or leakage power will be proportional to the leakage current per cell, as shown in Eq. (3.1). As a result, it is desirable to have a small leakage current as it is becoming a dominant

component of the total power consumption in sub-100 nm CMOS technologies [2, 55, 245]. The second component is P_{active} which is consumed when an SRAM is read or written. During these operations, a row is chosen by triggering one of the WLs high and hence the access transistors (N3 and N4 in **Fig. 3-1**) of all the cells on that row will be turned on. Each cell then draws an active current i_{cell}, hence a current of $(m \times i_{cell})$ is consumed. Therefore, it is necessary to partition the macro into smaller sub-macros in order to reduce this component. During these active cycles, the decoder and other peripheral circuits such as SA and write driver also contribute a significant amount of power consumption, as shown in Eq. (3.2).

$$P_{standby} = m \times n \times i_{hold} \times V_{DD} \tag{3.1}$$

$$\begin{aligned} P_{active} &= I_{DD} V_{DD} \\ &= \left(I_{core} + I_{decoder} + peripheral \right) V_{DD} \\ &= \left\{ \left[m i_{cell} + m(n-1) i_{hold} \right] + I_{decoder} + peripheral \right\} V_{DD} \end{aligned} \tag{3.2}$$

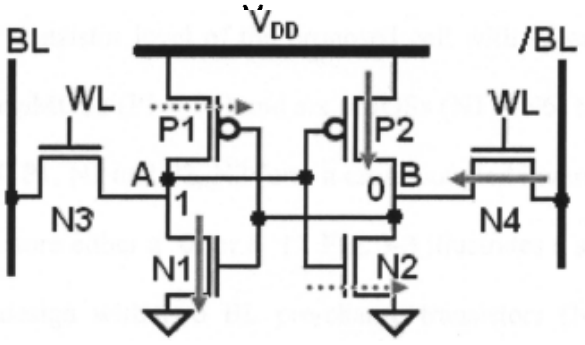


Figure 3-1 A conventional 6T SRAM cell. Cell leakage currents are illustrated by the red arrows. The solid and dotted arrows represent the sub-threshold and gate leakage current, respectively.

In this chapter, we focus on the first component of the active power, i.e. I_{core} ,by redesigning the memory cell and the leakage power, i.e. $P_{standby}$. We propose a new 10T SRAM cell and a new BLs pre-charge scheme that can reduce the $(m \times i_{cell})$ component into $1 \times i_{cell}$ and thus I_{core} is drastically reduced. The new cell also offers a lower leakage and hence can be suitable for applications where the system is in standby mode most of the time.

This work contributed in the following aspects:

1. We propose a 10T SRAM cell to reduce the unnecessary cell currents during the read operation. Transistor optimization and layout is discussed using a multi-V_{th} 65 nm CMOS process from STMicroelectronics (STM) to reduce the leakage current without compromising the read speed .

2. We propose a new BL pre-charge scheme that leads to a more than 90% read power consumption reduction within the memory core and at the same time reduces the cell leakage.

3. Cell stability using noise margin is extensively studied: statically, dynamically and statistically. Read reliability is also investigated via the BL leakage current to ensure that the proposed design can be implemented using smaller technologies where leakage current dominates the active cell current.

3.2 The new 10T SRAM cell

3.2.1 Read operation

Fig. 3-2 shows the transistor level of the proposed cell with separate read and write ports. It consists of four pMOSs (P1 – P4) and six nMOSs (N1 – N6) transistors. Like the conventional 6T SRAM, P1, N1 and P2, N2 form a cross-coupled inverters flip-flop which has two stable states to store either a '0' or a '1'. **Fig. 3-3** illustrates a simplified read data path of the proposed design with two BL pre-charge transistors (N11, N12), pull-up transistors (P10-P12), a BL SA and four DLs driving transistors (P13-P16).

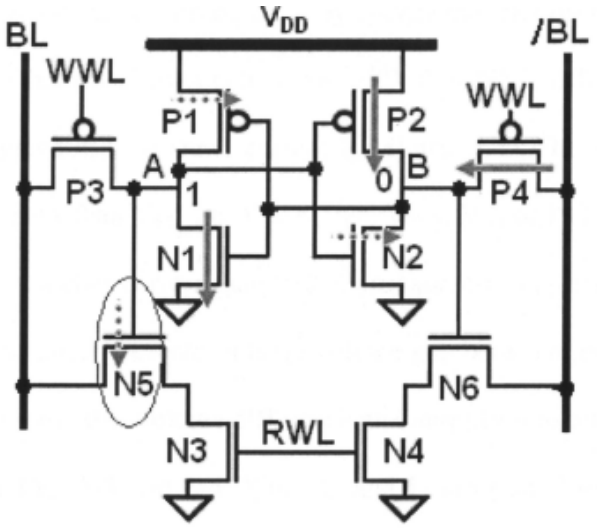


Figure 3-2 Proposed 10T cell with separate write/read ports. Cell leakage currents are illustrated by the red arrows. The solid and dotted red lines represent the sub-threshold and gate leakage currents respectively.

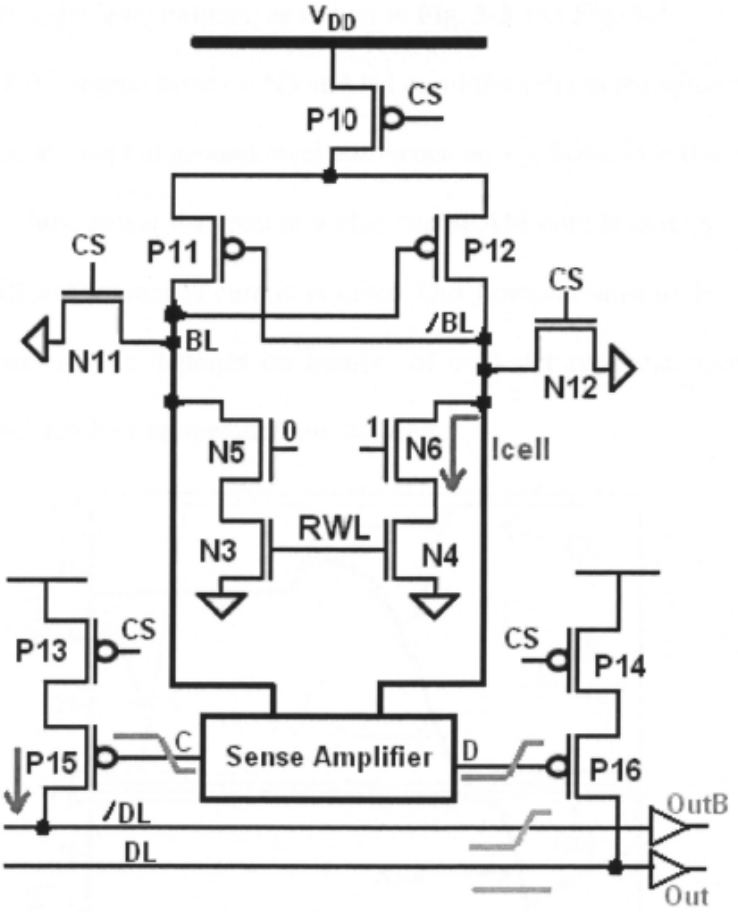


Figure 3-3 Data path in the read cycle of the proposed SRAM.

During standby, both BLs are pre-charged to ground, as shown by the light blue lines in **Fig. 3-4.** When a read operation is activated, a specific memory cell is chosen by its corresponding RWL and CS signals (**Fig. 3-3**). Consequently, N11 and N12 are turned off

to release the BLs. As P10-P12 are turned on, they charge the BLs up from ground level. Henceforth, we assume that the chosen cell stores a '0', thus N5 is off whereas N6 is on. Since the RWL is triggered high, a small current I_{cell} flows from /BL to ground, causing $V_{/BL}$ to rise at a slowe rate r than V_{BL}, i.e. $V_{/BL} < V_{BL}$. Thus, V_{GS} of P11 is larger than that of P12 and P11 sources a higher current than P12. Consequently, V_{BL} continues to rise at a higher rate than $V_{/BL}$ and quickly creates a large voltage gap between these two lines. The SA is then turned on to sense this voltage difference and amplify it to intermediate outputs C and D, as shown in **Fig. 3-3** and **3-4**. Since C and D are pulled to ground and V_{DD} respectively, P15 is turned on and P16 is cut-off. P15 sources a current to /DL and pulls it to a high voltage level while DL remains unchanged. A simple buffer is then used to provide full CMOS logic level outputs, as shown in **Fig. 3-3** and **Fig. 3-4**.

Although the RWL signal turns on N3 and N4 of all the cells in the same row, the BLs of the other column are kept at ground level and hence no I_{cell} flows into the other cells of the accessed row. Thus, power dissipation within the SRAM core is mainly consumed by the SA and a significant amount of current is saved. One however must understand that the exact level of power saving depends on number of cells per row and number of cells accessed during each read/write operation on that row.

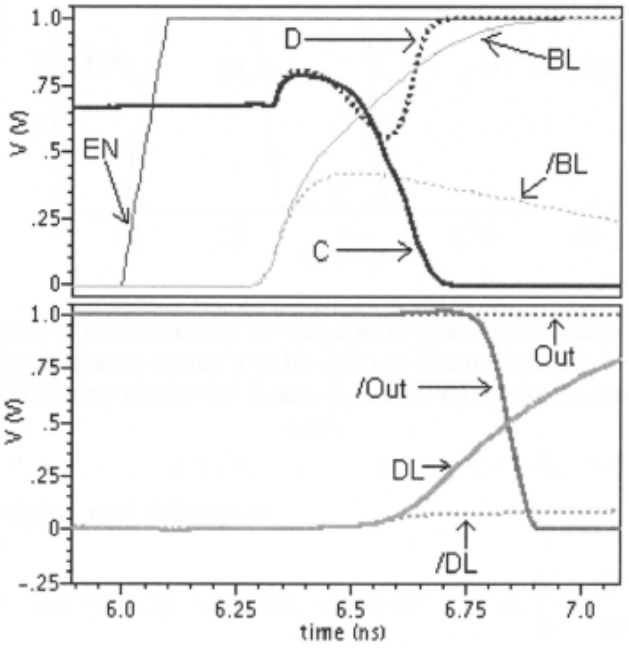


Figure 3-4 Waveforms of several nodes during a read cycle

3.2.2 Write operation

The proposed SRAM design has a similar write operation as the conventional design. When data is transferred to the BLs, the WWL turns on the access-transistors of the cells and data is written. However, since the pre-charge level of BLs is ground, PMOS transistors (P3 and P4 in **Fig. 3-2**) are used to access the memory instead of NMOS transistors (N3 and N4 in **Fig. 3-1**). This results in a 10% smaller cell current during a write operation and hence the proposed design has a 5% longer write delay and smaller write power when compared to the conventional 6T design. Simulated waveforms of several nodes of the proposed and the 6T designs during a write operation are illustrated in **Fig. 3-5**.

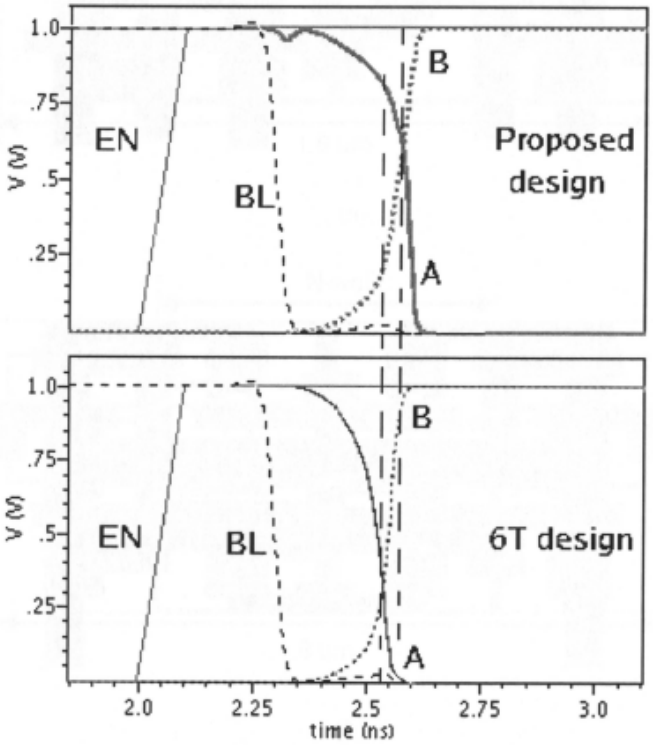


Figure 3-5 Waveforms of several nodes during a write cycle of the proposed (above) and 6T designs (below). The proposed design's write delay is about 5% slower than that of the 6T design due to the PMOS access device. A and B are the data storing nodes of the memory cells.

3.2.3 Transistor sizing and cell layout

All transistors in the 10T cell have the minimum size of $\frac{W}{L} = \frac{120\,nm}{60\,nm}$. On the other

hand, the pull-down transistors of the 6T cell have the size of $\frac{W}{L} = \frac{360\,nm}{60\,nm}$ while the others have the minimum size. As the proposed design has separate read/write ports, its noise margin during the read operation is still higher when compared to the conventional 6T cell, as will be discussed in **Subsection 3.3.2**. However, both read and write delays are 3%-5% longer than those of the 6T cell design. The proposed design has 33% layout area overhead when compared to the conventional 6T layout due to the four additional NMOS transistors and wiring of the RWL, as shown in **Fig. 3-6**.

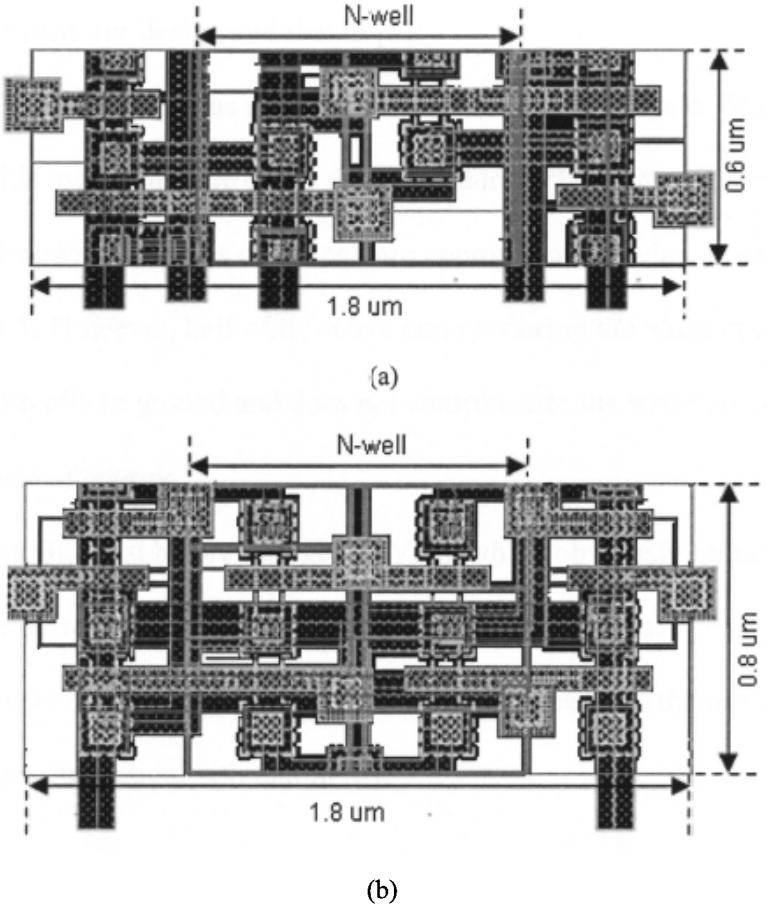


Figure 3-6 Layout of the proposed SRAM cell (a) 6T with the pull-down transistors have a W/L = 360nm/60nm (b) 10T with all transistors have a minimum size of W/L = 120nm/60 nm

It is worth mentioning here that this area comparison is based on our available standard logics rule from the foundry and no pushed layout rules are used. If memory layout rules are used to mimic the industry, memory size of the 6T and 10T are 0.41 um ×

1.26 um and 0.62 um × 1.26 um, respectively, i.e. the proposed design's area over head is 50%. Nonetheless, as technology scales down, excessive fabrication fluctuations require 6T design to use larger transistor size to maintain a reasonable noise margin [12]. Thus, the 10T design will have its advantage as the area overhead reduced thanks to the minimum-sized transistors.

Our proposed design has a similar structure to that in [5] but we use PMOS access devices with separate Read and Write WL. As a result, there are three major differences between our design and that in [5]:

1. The proposed design has separate Write and Read WL while [5] share one WL. This means that the cell is disturbed during the read operation.
2. Both read/write currents of the cell are approximately twice as large as that of the 6T. However, half of its active current during the write operation flows directly to ground and does not contribute to the write process. This is a waste of energy.
3. In [5] its BL must be pre-charged to VDD which obviously induces more leakage current due to the four additional NMOS devices.

Since our work focus on low-power low-leakage property of the design, we will not compare our design with that in [5].

3.3 Leakage and noise margin analysis

3.3.1. Cell and BL Leakages

3.3.2.1 Cell leakage

Leakage current is one of the major concerns in nano-scale SRAM where most of the transistors are in the standby mode [207]. **Figs. 3-7** illustrates the major leakage currents in the 6T and 10T cells where the solid and dotted red lines represent the sub-threshold and gate leakage currents respectively. Since both BLs of the 10T cell are pre-charged at 0

V, it only incurs one additional leakage current through the gate of N5, as highlighted by the blue oval. However, the pull-down transistors N1 and N2 of the new cell are smaller than those in the conventional cell, hence reducing its total leakage. For example, at 40 °C, leakage currents of the 6T and 10T cells are 300 pA and 210 pA respectively.

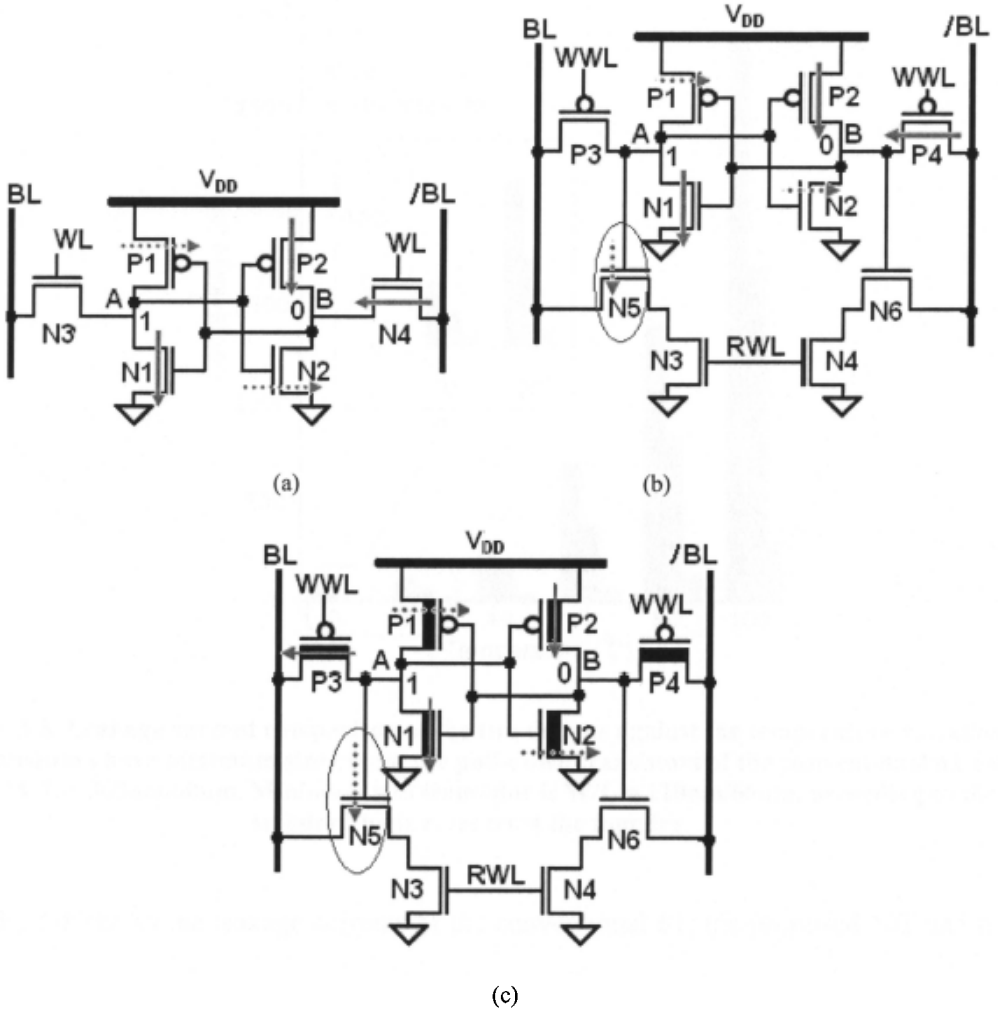


Figure 3-7 (a) Conventional 6T cell. (b) Proposed 10T cell with standard V_{th} transistors (c) Multi-V_{th} 10T SRAM cell. Standard V_{th} transistors have thin channel while high-V_{th} transistors have bold channel. Leakage currents are illustrated by the red arrows. The solid and dotted red lines represent the sub-threshold and gate leakage currents, respectively

In an attempt to further reduce the leakage current of the 10T SRAM cell, we use a multi-threshold CMOS process and apply different V_{th} for different transistors in the cell. Since among the 10 transistors within the memory cell, only four transistors N3-N6 contribute to the read delay, it is necessary for these transistors to have as low-V_{th} as possible. Fortunately, they do not contribute any significant leakage current to the cell. On

the other hand, the other six transistors are the main contributors to the total leakage current, as illustrated in **Fig. 3-7**. Hence, we use high V_{th} for N1-N2 and P1-P4, and standard V_{th} for N3-N6. Nonetheless, this extended write delay is still shorter than the read delay and thus does not affect the operating frequency of the system.

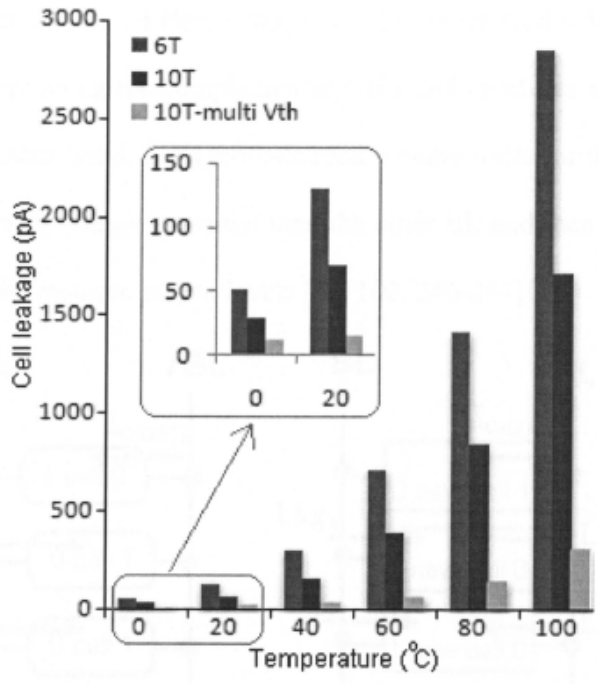


Figure 3-8 Leakage current comparison of the two designs against the temperature variation. All transistors have minimum size except the pull-down transistors of the conventional 6T cell with W/L = 360nm/60nm. Minimum size transistor is W/L = 120nm/60nm, according to the standard logic rules from the foundry.

Fig. 3-8 shows the leakage currents of the conventional 6T, the proposed 10T and the proposed 10T with multiple-V_{th} CMOS process. Despite having more transistor count, the proposed 10T cell has less leakage current when compared to the 6T cell, with about 30% leakage reduction. This is because all transistors used in the 10T designs are minimum size and BLs are pre-charged to ground. When multiple-V_{th} is used, its leakage current is reduced to 10% of that of the 6T cell. It is worth mentioning here that this excellent improvement incurs no additional latency or worse case delay of the system, which is determined by the read operation. All subsequent simulations on the 10T design are performed on the multi-V_{th} cells.

3.3.2.2 BL leakage

One of the fundamental requirements of high density SRAM is a reliable read operation. In any SRAM architecture, the cell read current I_{cell} is used to identify the stored data in the accessed memory cell. There are two basic read schemes in the literature which use the above-mentioned I_{cell} in different ways: current-mode read scheme translates I_{cell} into a current difference along the complementary BLs and amplifies it into CMOS logic output levels. On the other hand, voltage-mode read scheme waits for the I_{cell} to discharge one of the BLs to a lower voltage potential than the other BL and then amplify this small voltage difference to the required output levels [96, 102, 246-247].

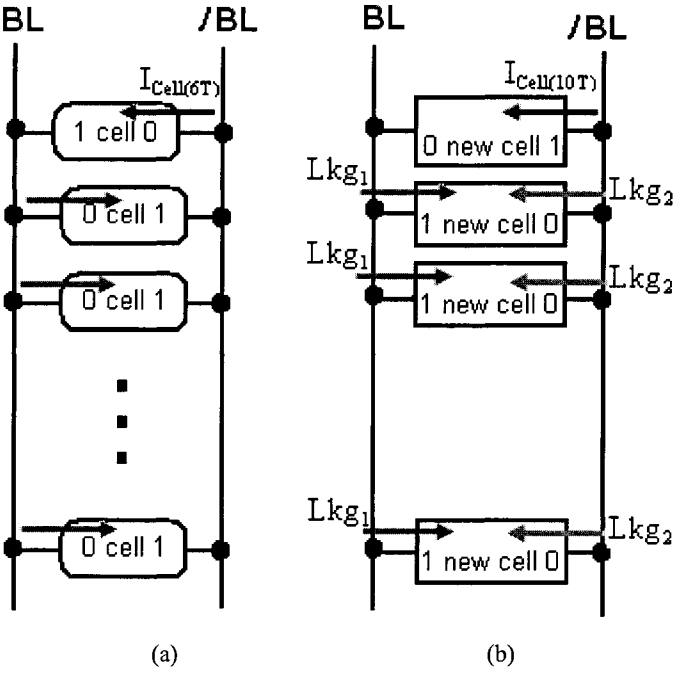


Figure 3-9 BL leakage current (a) 6T cell (b) 10T cell

It is desirable to have as large I_{cell} as possible to ensure a fast and reliable read operation. However, because of the stringent constraint of the cell area, SRAM designers usually use minimum-size transistors and thus I_{cell} is usually very weak. **Fig. 3-9** demonstrates the flow of I_{cell} (blue arrow) in one column when the first cell is accessed. In theory, this is the only available current along the BLs. Nonetheless, one must bear in mind that sub-threshold leakage currents are always present, also illustrated in **Fig. 3-9** as red and purple arrows.

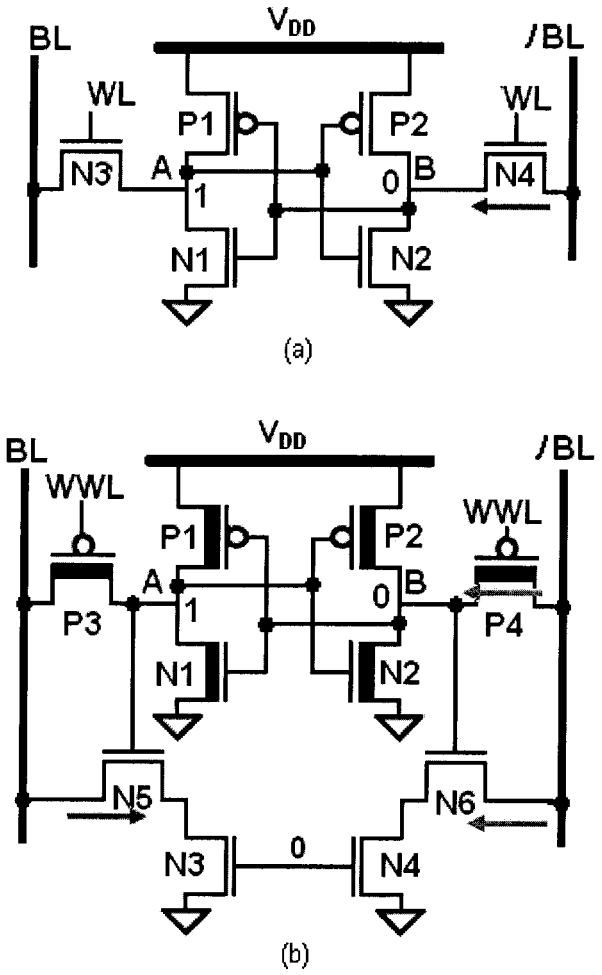


Figure 3-10 BL leakage current within the SRAM cell (a) 6T (b) 10T

Fig 3-10 details the leakage path in the 6T and 10T cells during a read operation. Assuming that the cells in **Fig. 3-10** store a '1' at node A and thus a '0' at node B, the leakage current will flow from the /BL through N4/P4 to node B, as shown in **Fig. 3-10 (a)**. In the 10T cell, there are two other leakage currents flowing from both BLs to ground, through N5-N3 and N6-N4, as shown in **Fig. 3-10 (b)**, and hence its total BL leakage is higher than that of the 6T design. In the 6T SRAM design, the worst case happens when the accessed cell stores a '1' and all the other cells in the same column store a '0', as illustrated in **Fig. 3-9 (a)**. The condition to have a correct output is:

$$Ratio_1 = \frac{I_{cell}}{I_{Leakage}} > n-1 \qquad (3.3)$$

Where n is the number of rows. In case of the 10T cell, since $I_{leakage(N5,3)} > I_{leakage(P4)} + I_{leakage(N6,4)}$, Eq. (3.3) becomes:

$$Ratio_2 = \frac{I_{cell}}{\left[I_{leakage(N5,3)} - I_{leakage(P4)} - I_{leakage(N6,4)}\right]} = \frac{I_{cell}}{\left(I_{leakage1} - I_{leakage2}\right)} > n-1 \quad (3.4)$$

As technology and supply voltage scale down, I_{cell} becomes smaller while $I_{leakage}$ rapidly increases [52] and hence, these ratios will shrink to a level where the number of cells per row is not limited by the BL parasitic capacitance but by the ratio between I_{cell} and $I_{leakage}$ instead. Therefore, it is desirable to have these ratios as high as possible.

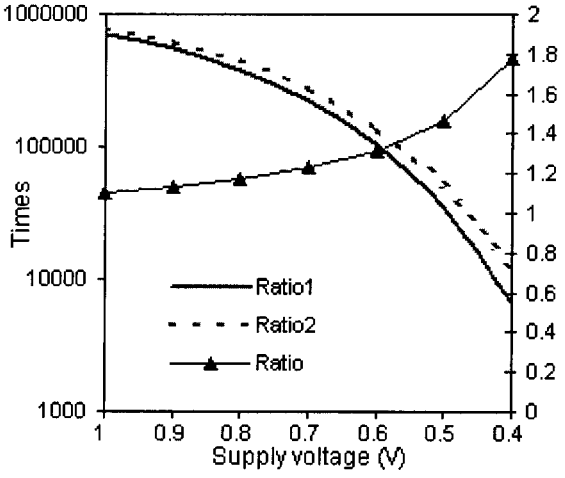
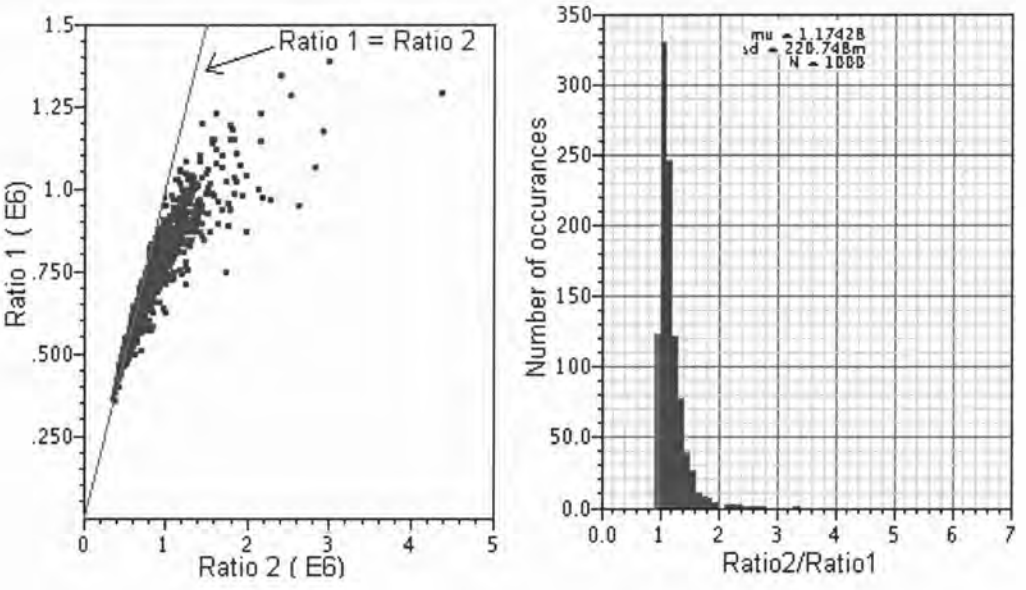


Figure 3-11 $\frac{I_{on}}{I_{off}}$ **Ratios of the memory cells in comparison at different supply voltages.**

$$\textbf{Ratio} = \frac{Ratio_2}{Ratio_1}$$

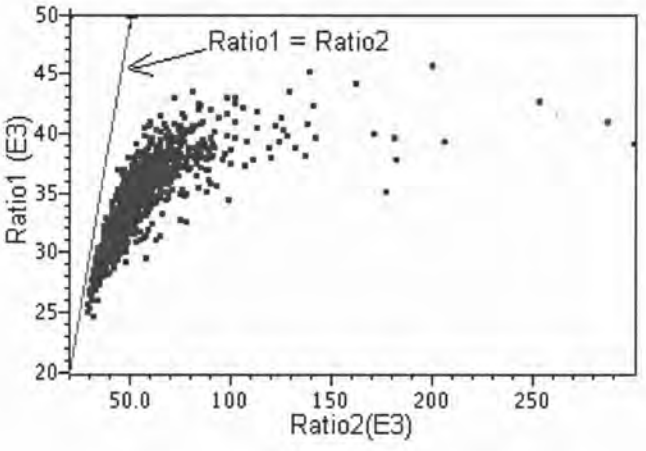
Fig. 3-11 represents the simulation data of $Ratio_1$ (6T SRAM) and $Ratio_2$ (10T SRAM) at different supply voltages. It is apparent that although the proposed cell has more leakage current, as shown in **Fig. 3-9** and **3-10**, the leakage on the right, i.e. $Leakage_1$- denoted as Lkg_1 in **Fig. 3-9**, cancels its left counterpart, i.e. $Leakage_2$ - denoted as Lkg_2 in **Fig. 3-9**, and thus the cell's net leakage current is less than that of the 6T cell. As a result, $Ratio_2$ is higher than $Ratio_1$, as shown in **Fig. 3-11**. It is important to note that these leakages of the new cell are only available in one column during a read cycle when both BLs are raised to high voltage levels. The term Ratio=$\frac{Ratio_2}{Ratio_1}$ shown on the

secondary axis in **Fig. 3-11** is used to emphasize the relative performance of the two designs. It is clear that the lower the supply voltage, the more the new design outperforms the 6T cell. For example, at 1V, $Ratio_2$ equals to 1.1x $Ratio_1$ but at 0.4 V, it increases to 1.8x $Ratio_1$.



(b)

(b)



(c)

Figure 3-12 $Ratio_2$ using Monte-Carlo simulations. (a) V_{DD} = 1V. (b) Histogram plot of $\frac{Ratio_2}{Ratio_1}$, V_{DD} = 1 V. (c) V_{DD} = 0.4 V

In the sub-100 nm CMOS processes, it is necessary to understand the behavior of the designs under excessive process variations. We have performed various Monte-Carlo simulations to evaluate $Ratio_1$ and $Ratio_2$ under process fluctuations using standard data from the foundry. **Fig. 3-12 (a)** shows a scattering plot between $Ratio_1$ and $Ratio_2$ at 1V

supply voltage. It is apparent that most of the points lie below the $Ratio_1$=$Ratio_2$ line and this means that most of the time $Ratio_1$ < $Ratio_2$. An auxiliary histogram plot is shown in **Fig. 3-12 (b)**, indicating that $Ratio_2$ is higher than $Ratio_1$ in only 87% of the time as illustrated by the blue bars in **Fig. 3-12 (b)**. At 0.4V, $Ratio_1$ is always smaller than $Ratio_2$ as all of the points are below the red line. It therefore has conclusively proved the new cell's advantage especially in applications where ultra low supply voltages are preferred.

3.3.2. Noise Margin

3.3.2.1 SNM versus DNM

SNM is the most popular measure to evaluate the stability of the memory cell [119, 248-249] as it indicates how much noise is needed to malfunction the cell content under the worst case scenario. However, recent works on cell stability have pointed out that SNM is only a special case under a broader class of DNM in which NPW extends to infinity [5, 134-135]. In this work, we performed a simple DNM analysis of the proposed design and the conventional 6T cell using different NPW at 40^0 C and 100^0 C, respectively. Simulation results are presented in **Fig. 3-13** and strongly agree with the conclusion in Ref. [134-135].

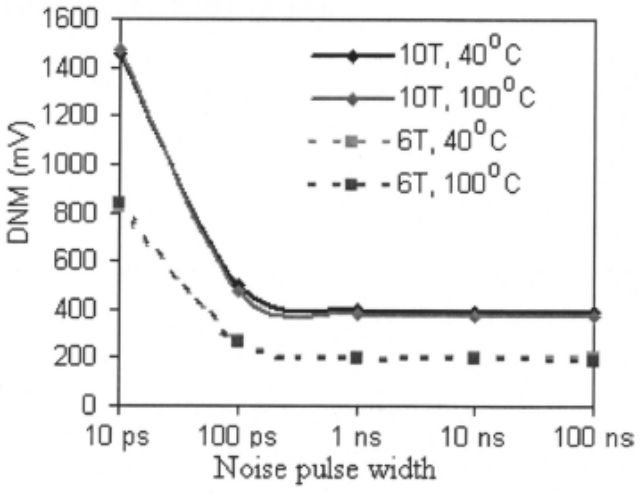
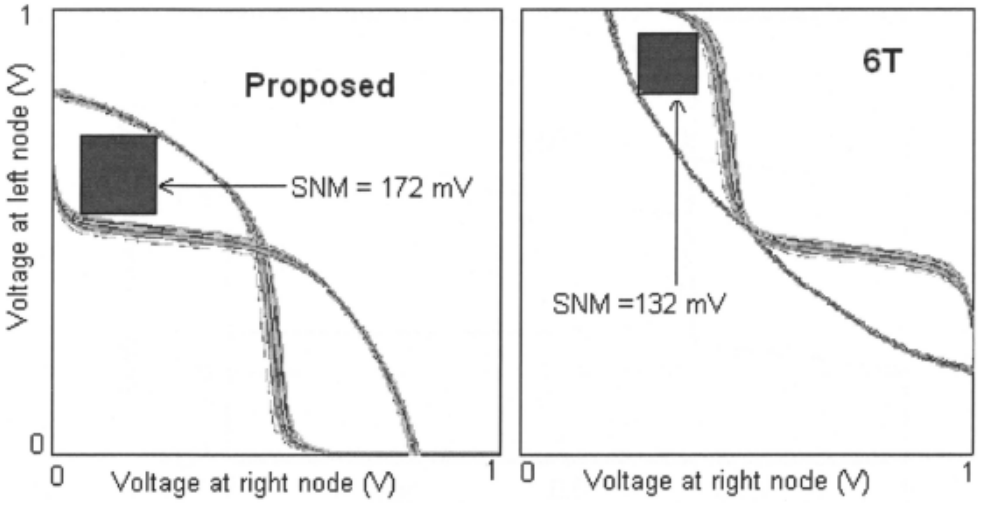


Figure 3-13 Dynamic Noise Margin of the conventional and the new 6T cells versus the cell ratio and the access transistor's width variations.

Both designs have very high noise margins when the NPW is exceptionally short, i.e. about 10 ps but they drop by more than 70% when the pulse width increases to 100 ps. For example, at 40^0 C, 10 ps NPW, the DNMs of the 10T and 6T cells are 1460 mV and 820 mV, respectively. However, for NPW higher than 100 ps, these DNMs saturate and approach their SNM values of 390 mV and 198 mV, respectively. This has reaffirmed that the proposed design has about 2X superior stability when compared to the conventional 6T design, dynamically and statically.

3.3.2.2 Statistical Simulation of SNM

As mentioned in the previous section, deterministic simulations are not sufficient to evaluate the performance of the design, especially in nano-scale SRAM [129]. Although both DNM and SNM of the proposed design are higher than those of the conventional 6T SRAM, statistical simulations are necessary to predict their relative stability under process variations. **Fig. 3-14** shows the butterfly curves of both 10T and 6T SRAM cells at 1 V and 0.4 V supply voltages. The SNM is now defined as the biggest square that can fit into the smaller eye of the butterfly curve [129]. In other words, it is the worst-case-SNM under process variations. As shown in **Fig. 3-14**, the proposed design has a higher SNM of 172 mV when compare to 132 mV of that of the 6T design. This is because the access transistors of the proposed design are PMOS and hence the disturbances of the half-accessed cells are less than that of the 6T design.



(a) (b)

Figure 3-14 Monte-Carlo simulations of the butterfly curves of the two designs. (a) 10T cell. (b) 6T cell.

3.3.3. Write Trip Point

As we use high-threshold transistors N1-N2 and P1-P4 to significantly reduce leakage current, writing into the proposed cell is rather difficult. Furthermore, the PMOSs access transistors are weaker than the conventional NMOSs in the 6T design which further worsens the situation. As a result, the 10T's WTP is inferior to that of the 6T design. Conventionally, WTP is defined as the highest BL voltage that can flip the memory cell successfully (assuming that the /BL is kept at V_{DD}). In our design, one of the BLs is kept at ground while the other is raised from ground to V_{DD} during the write operation. Thus, using the conventional definition may create confusion. Instead, we define WTP as the minimum BL swing that can flip the memory cell successfully. **Fig. 3-15** shows the simulation waveforms of the conventional 6T and the proposed 10T SRAM cells during a write operation. The 10T cell takes more time to write and also requires a higher BL voltage. **Fig. 3-16** summarizes the corresponding WTP of the two designs at different supply voltages. The proposed design requires about 23% higher BL voltage swing.

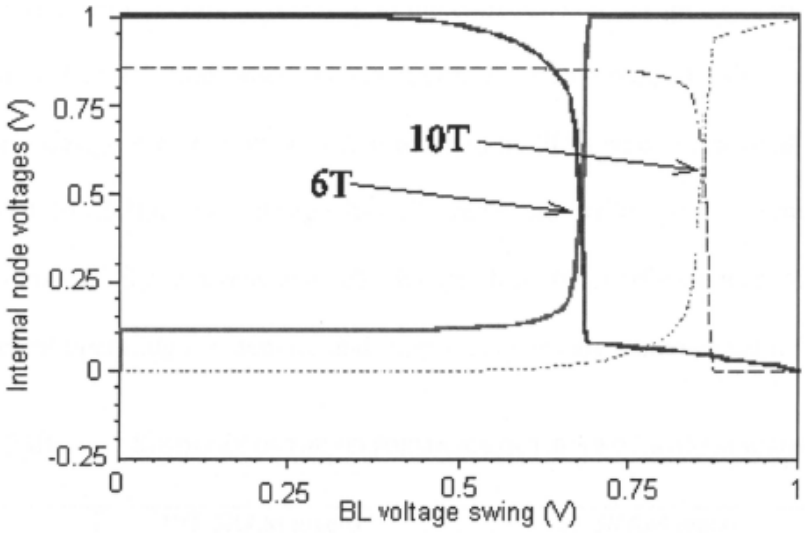


Figure 3-15 WTP of the 6T and the proposed 10T design during a write operation

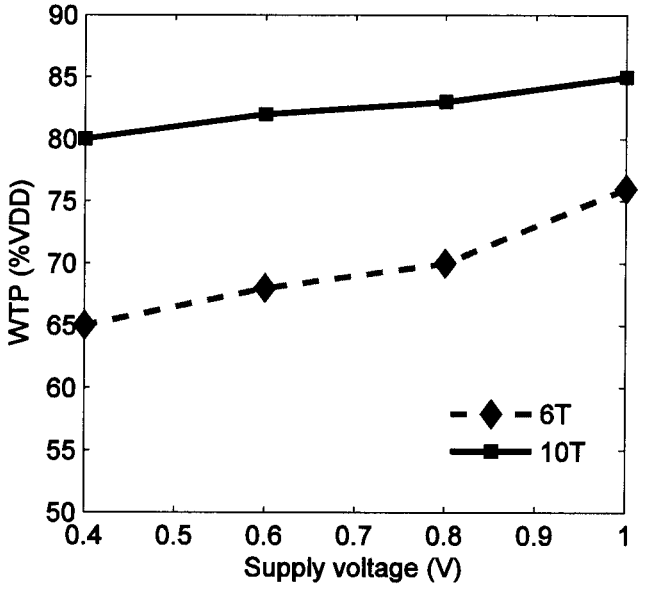


Figure 3-16 WTP of the 6T and 10T cell versus VDD variation

3.4. Performance comparison

Two 128x128 SRAM macros have been designed and simulated in a standard 65 nm CMOS process from STM using the conventional 6T and the multi-V_{th} 10T cells. Both macros have identical address decoders, DL drivers and SA design. Extensive Read/Write operations have been simulated at 40^0 C to evaluate the performance of the newly proposed cell. All results are recorded at (250 MHz, 1 V), (500 MHz, 1V) and (1 MHz, 0.4 V) as shown in Table III. It is apparent that under any operating condition, the proposed design has significantly less read power consumption. This is because only one cell is

turned on instead of all the cells in one row in the conventional design. At 250 MHz, the 6T SRAM macro has 4X read power consumption when compared to that of the new design. The new design's write power is also reduced by 30% when compared to that of the conventional 6T design. Our design has 5% read and write speed overhead when compared to those of the conventional 6T design. Relative performances of the two macros at different operating frequencies and supply voltages are summarized in Table III.

TABLE III. SUMMARY OF THE PERFORMANCE OF THE TWO SRAM MACROS.

	10T SRAM macro	6T SRAM macro
0.4V, 1MHz	Read Power: **11 uA** Read Delay: 96 ns Write Power: **38 uA** Write Delay: 87 ns	Read Power: **50.7 uA** Read Delay: 90 ns Write Power: **50.9 uA** Write Delay : 79 ns
1V, 250 MHz	Read Power: **0.78 mA** Read Delay: 820 ps Write power: 2.0 mA Write delay: 600 ps	Read Power: **3.25 mA** Read Delay: 780 ps Write power: 3.3 mA Write delay: 580 ps
1V, 500 MHz	Read Power: **1.6 mA** Read Delay: 790 ps Write power: 3.2mA Write delay: 600 ps	Read Power: **4.4 mA** Read Delay: 760 ps Write power: 4.5 mA Write delay: 700 ps

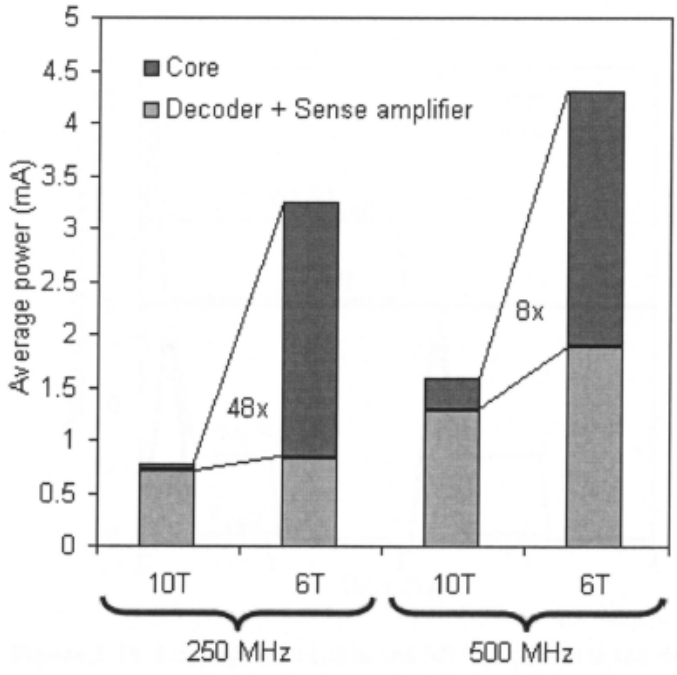


Figure 3-17 Average read power of the two design during a read cycle

Fig. 3-17 breaks down the read power consumption of the two designs in consideration. It consists of two components: Core dissipation and peripheral dissipation. Theoretically, by turning on only one cell during the read operation, power dissipation of the proposed design would be 1/128 of that of the conventional design. Nevertheless, some circuit components also draw currents when the core is activated such as the RWL or the CS, etc. These are short pulse currents and quickly diminish after a few tens of picoseconds. In the conventional 6T design, a significant read current flows into all the half-accessed cells on the same row and lasts until the end of the read cycle, as shown in **Fig. 3-18**. It also shows that the "pulsing period" dominates the total read cycle when the frequency is higher. This explains why at 500 MHz, the power reduction within the core is 8X whereas that at 250 MHz is 48X. Although these numbers are far below the optimum value of 128, our proposed design has made a measurable power reduction within the core. This implies that power consumed due to half-accessed cells during the read operation is no longer a design bottleneck and circuit designers can partition the macro differently with more cells per row, and hence its layout is more efficient and can be used to compensate the area overhead induced by the larger 10T cell layout.

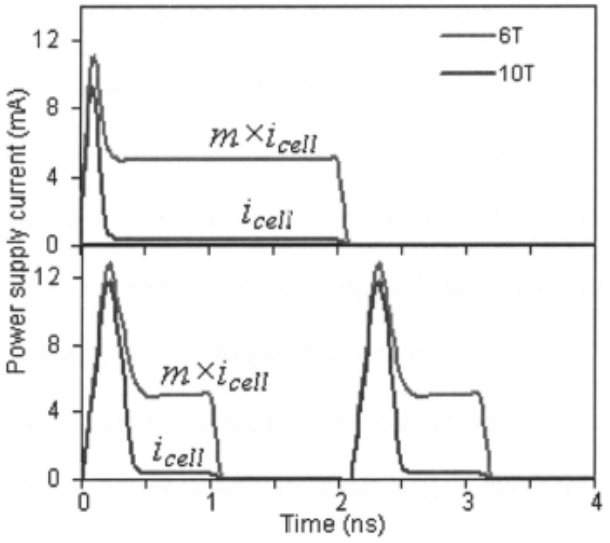


Figure 3-18 Leakage current in the SRAM cell. m is the number of cell per row and i_{cell} is the read current of one accessed cell.

Fig. 3-19 approximates the percentage power saving versus the hit ratio of the cache. Each time data are found within the cache, a hit occurs and a read operation is performed. Otherwise, data must be written to the cache before read out. As a result, average power reduction of the proposed design is a function of the hit ratio. **Fig. 3-19** also shows that in slow operations, the average power reduction is very high, from 65% to 77%. When operating at 500 MHz, these number ranges from 57% to 63%.

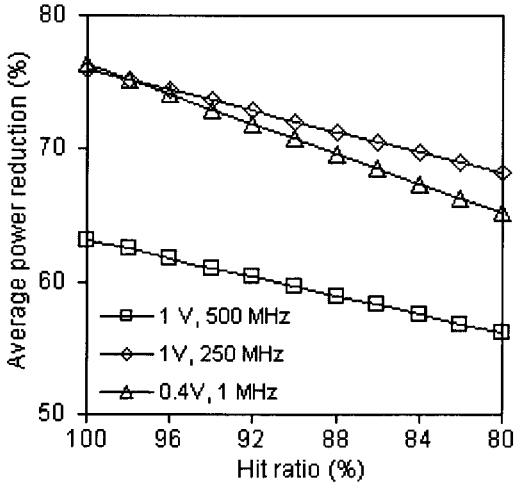


Figure 3-19 Average power reduction versus cache hit ratio

As mentioned in **Subsection 3.3.1**, SRAM speed is limited by its read delay. **Fig. 3-20** represents the read delay of both designs at different process corners and supply voltages. It is noticeable that the 10T design has slightly longer delay when compared to the 6T design. This is because it has smaller pull-down transistors within the memory cell during a read cycle when using the same SA. However, this additional delay is only marginal and remains almost the same at all process corners. Both designs have the longest read delay at Slow-Slow corner, which is up to 70% longer than that at the Fast-Fast corner. As a result, the SA enable (SEN) signal must be delayed further to cope with this variation and ensures a correct sensing. As this is not the emphasis of this project, we only used inverters as a simple way to increase the delay of this signal.

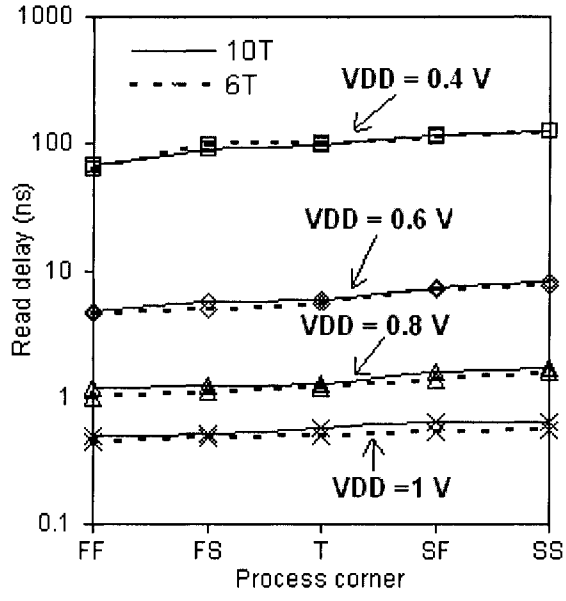


Figure 3-20 Read delay of the two SRAM macros in consideration at different process corners: Fast-Fast (FF), Fast-Slow (FS), Typical (T), Slow-Fast (SF) and Slow-Slow (SS).

3.5 Conclusion

A novel 10T SRAM cell has been proposed and analyzed. It separates the write and read operations of the SRAM and hence has improved the noise margin problem during the read cycle. As a result, its noise margin is 1.3x of that of the conventional 6T design. Concurrently, it reduces 76% of the total read power at 250 MHz, 1 V supply voltage.

Considering the active current within the core, the proposed design offers more than 90% reduction. Its write and read delays are also compatible to that of the conventional 6T. In addition, its leakage is only 10% of that of the conventional 6T design, which is attributable to the use of multiple-V_{th} CMOS process. This achievement is made with 23% sacrifice in WTP. However, the cell still can be written properly if full BL voltage swing is obtained during the write operation. In view of the above-mentioned advantages, it can be concluded that the new design is a more suitable choice for applications that require ultra low-power ultra low-leakage applications where power is the most critical design factor.

CHAPTER 4 AN 8T DIFFERENTIAL SRAM WITH IMPROVED SNM FOR BIT-INTERLEAVING

Lowering power consumption and increasing noise margin have become two central topics in every state of the art SRAM design [36, 250]. Due to parameter fluctuations in scaled technologies, stable operation is critical to obtain high yield low-voltage, low-power SRAM [16, 52, 250]. Recent published works in literature have shown that the conventional 6T SRAM suffers a severe stability degradation due to access disturbances at low-power mode [11, 16]. Thus, several 8T and 10T cell designs have been reported, improving the cell stability [1, 225, 232, 237, 251]. However, they either employ single-ended read port or require too large area. In this paper, we use a fully differential 8T SRAM that allows efficient bit-interleaving to achieve soft-error tolerance with conventional Error Correcting Code (ECC). It also consumes less power when compared to the conventional 6T design. A column-based dynamic supply voltage scheme is utilized to improve both the read noise margin and the write-ability. To verify the technique, a 128x64-bit of the proposed SRAM has been implemented in a standard 65 nm/ 1V CMOS process. Simulation results reaffirmed that the proposed design has 2x higher noise margin and consumes 54% less power when compared to the conventional 6T design.

4.1 Introduction

Low-power and high-stability have been the main themes of SRAM designs in the last decade [251]. The explosion of the portable electronic market constantly urges for less power-hungry architectures. Thus, many techniques have been employed to deliver this requirement such as scaling the supply voltage, using multi-threshold CMOS process to minimize the leakage, dividing the SRAM macro into multiple sub-macros to enhance its

stabilities and to reduce the dynamic power. Among these, supply voltage down scaling offers the highest effectiveness since the dynamic power is a quadratic function of voltage. Furthermore, it also exponentially reduces the leakage current which dominates the active current in the sub-100 nm CMOS processes. Supply voltage scaling has recently extended to sub-threshold circuit operations to significantly reduce the total power consumption [1, 3-4, 12, 251]. As a result, a hefty amount of energy has been saved, at the cost of the speed performance. Several silicon results have been successfully measured at 0.3 V or lower [4, 251]. However, it is arguable to further scale down the supply voltage since the speed is exponentially degraded and hence total energy per read/write is increased. For example, [251] reported a 32 kb 10T SRAM which dissipates 1 μW at 0.3V, 500 kHz and 0.15 μW at 0.16 V, 500 Hz. A simple calculation shows that at 0.3 V, the memory consumes 1.7 pJ per write, whereas at 0.16 V, it consumes 246 pJ per write. Thus, the energy consumption per write increases 144x when the supply voltage is reduced from 0.3 V to 0.16 V. In this paper, we will investigate the optimum supply voltage for the proposed SRAM cell at which the energy consumption per read/write is minimized.

The second challenge in designing a robust SRAM is to ensure a reasonable noise margin, which is normally measured by the SNM and the WTP [23, 129, 249, 252-255]. According to [129], these two design factors degrade when the threshold voltage variation increases. Furthermore, they are linearly dependent on the supply voltage, which if being reduced in order to save power, has a negative impact on the cell stability. As a result, it is extremely difficult to maintain the cell stability as technology enters the sub-100 nm regime. Vigorous efforts have been put forth to improve the SNM and the WTP of the SRAM cell. Unfortunately, these two factors conflict with each other and hence improving one is likely to jeopardize the other.

Several more-than-6T SRAM cell designs have been proposed in literature, emphasizing on solving the above-mentioned conflict in SRAM design [1, 3, 6, 36, 84,

113, 165, 214, 224, 241, 251]. They employ separate Read and Write ports and hence the cell's SNM and WTP can be optimized individually without affecting each other. Nevertheless, most of these designs can only be implemented in WL sharing architecture which is not preferable as the conventional ECC requires bit-interleaving to address multiple bit soft-error [3, 251, 256].

In this paper, we use a disturb-decouple differential 8T SRAM cell [224] to address the above-mentioned issues. It also eliminates the half-access issues associated with previously published designs and thus both SNM and WTP can be improved concurrently. A true single column-based dynamic supply voltage technique is employed to simultaneously improve the cell read stability and write-ability. The proposed design also consumes less power when compared to the other differential cell designs as only one cell is activated during any read or write operation, which is unique when compare to all previously published works. In addition, it can operate in a wide range of supply voltage (0.2 V to 1V) without any significant circuit modification.

The rest of the chapter is organized as follows: **Section 4.2** reviews the basic operations of the conventional 6T SRAM cell and the recent developments in SRAM cell design. **Section 4.3** presents the operating principles of the proposed cell and its circuit implementation. **In section 4.4**, we describe in detail the simulation methodology and performance evaluation of the proposed design, in comparison of the conventional 6T, the original 8T [224], the 10T [251] and the 6T_1[257] SRAM cells. **Section 4.5** concludes the chapter.

4.2 Recent SRAM designs for bit-interleaving

4.2.1 Shared BL versus Bit-interleaving

There are two commonly used ways to arrange the words in SRAM architecture: shared WL (**Fig. 4-1 (a)**) and bit-interleaving (**Fig. 4-1(b)**).

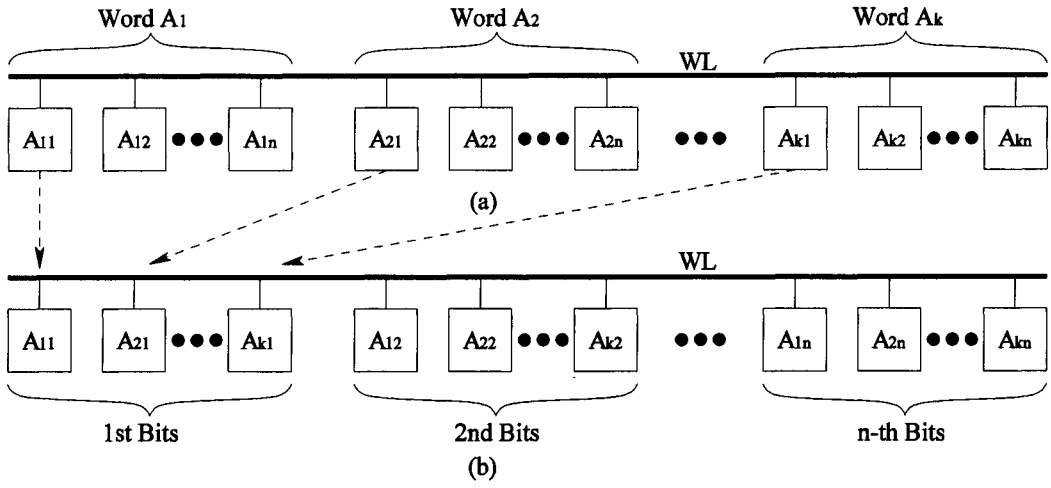


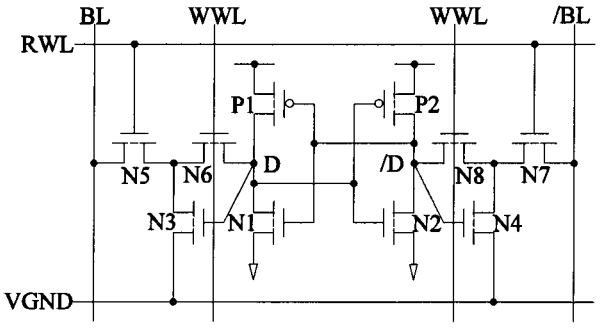
Figure 4-1 SRAM word organization (a) Shared WL (b) Bit-interleaving

In the shared WL architecture, all the bits of the same word are located next to each other. For example, **Fig. 4-1(a)** shows a row of k words, with each having n bit cells. During access, the WL is activated and all the bits on the row are turned on. A NAND gate can also be used to choose only one word out of k words on a row. This architecture is widely used because of its simplicity and compactness. However, as all the bits of a word are adjacent to each other, the probability of multi-bit soft errors is very high, especially in sub-100 nm CMOS technologies. This has a negative impact on the yield of the chip. As a result, bits of adjacent words must be interleaved to avoid the multi-bit errors in the accessed word so that the conventional ECC can be implemented to detect the single error bit. For example, in **Fig. 4-1(a)**, if we read from word A_1 and the ionized radiation is focused on the first bit of the row (i.e. A_{11}), the adjacent cells are also affected. Thus A_{12}, $A_{13}...A_{1p}$ may be erroneous too. This results in the multi-bit error phenomenon that cannot be efficiently solved by the conventional ECC. On the other hand, in **Fig. 4-1(b)**, bits A_{12}, $A_{13}...A_{1p}$ are very far away and hence only one bit of A_1 is affected. Since the rate of soft error increases as technology scales down, bit-interleaving is preferable in state-of-the-art designs. Most of the low-power SRAM designs published in the literature, however, have the half-access issue during the write operation and therefore cannot be bit-interleaved [251]. Recently, there are two 10T cell designs [3, 251] that successfully mitigate the half-

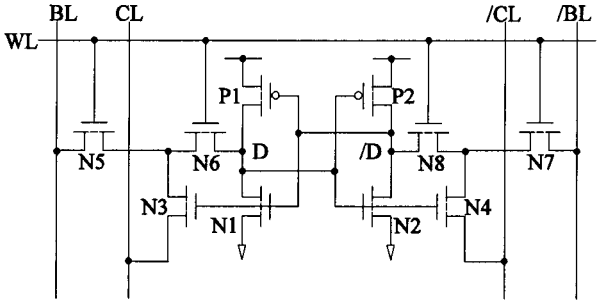
access problem and bit-interleaving becomes feasible at very-low supply voltages. Their operations are going to be discussed in **Section 4.3.2.**

4.2.2 Recent SRAM cell designs

Fig. 4-2 represents the recently published 10T SRAM cells that successfully addressed the half-accessed issue. The 10T_1 design in **Fig. 4-2(a)** has two additional transistors (N6 and N8) in the write port (when compared to the conventional 6T) which are only turned on by the column-based Write WL (WWL). As a result, unselected cells on the accessed row are isolated from the BL disturbances as their WWLs are inactive. It also has a separate read port (N3-N5 and N4-N7) that does not affect the internal nodes of the cells. Consequently, its noise margin is very high and similar to that of the conventional 6T in the standby mode [251]. This design however has two major drawbacks: (1) Its area overhead is very large as it uses 10 transistors and its layout is not very compact; (2) The write port uses two cascaded nMOSs and hence its write strength is weaker than that of the 6T design. Consequently, it has a lower WTP. The 10T_2 design has partially solved the above mentioned drawbacks by rearranging the connections between the transistors and applying a column line assist. Nevertheless, its SNM is compromised. In summary, both designs have better SNMs than the 6T design and the half-access issue has been mitigated but their area and WTP are not as good as those of the 6T cell. Along with the conventional 6T, these two cells will be discussed in the paper as references to evaluate the performance of our design.



(a)



(b)

Figure 4-2 SRAM cells (a) Differential 10T_1 [251] (b) Differential 10T_2 [3]

The 8T disturb-decouple SRAM cell was first introduced in [224]. Similar to the above-mentioned 10T cells, it has successfully eliminated the half-access problem in the conventional 6T design. However, the potential of this 8T cell has not been fully explored to maximize it performance. In [224], the only accessed cell in a row is turned on by its respective WL and CS signals. As a result, this accessed cell is disturbed in a similar manner as the conventional 6T. To improve its read noise margin, a voltage mode sensing scheme must be used to fully discharge the BL to ground. Thus, disturbances to the cell are mitigated and simulations showed that the cell is more stable and less prone to error during operation. However, this scheme has three main drawbacks: *1) The BL must be discharged to ground very quickly. Otherwise, the cell is still flipped. The stability of the cell depends on the timing of the BL discharge and the WL turn on.* 2) *Normal SRAM operation cannot be used as it removes all the benefit of the 8T cell. 3) Noise margin of the cell is only marginal and is limited by the amount* V_{bump} *reduction [224].* Next section will described the operation of the proposed cell with a dynamic column-based supply control

[257] that significantly improves the performance of the cell. The combination of the half-access free cell with a column-based supply control fully explores the cell's potential for implementing in bit-interleaving manner.

4.3 Operating principles of the proposed 8T SRAM

4.3.1 Operating principles

The proposed SRAM cell consists of 8 transistors, N1-N5 and P1-P3, as shown in **Fig. 4-3(a)**, with all having minimum size of $\frac{W}{L} = \frac{120\,nm}{60\,nm}$ to save area. Four transistors N1, N2, P1 and P2 form a cross-couple structure to store data. Four transistors P3 and N3-N5 are used to access to the internal nodes D and /D of the cell. N3 and N4 connect the cell internal nodes to the BLs while P3 and N5 form an inverter to control the voltage of node C1. The source terminal of P3 is connected to a column select (CS) line while the gates of P3 and N3 are connected to the WL. As a result, N4 and N3 are turned on if and only if both the WL and the CS are triggered. Unlike the conventional design, the sources of P1 and P2 are connected to a dynamic cell supply (Cell_supply) line which is raised to the higher voltage during the read operation in order to obtain a higher noise margin.

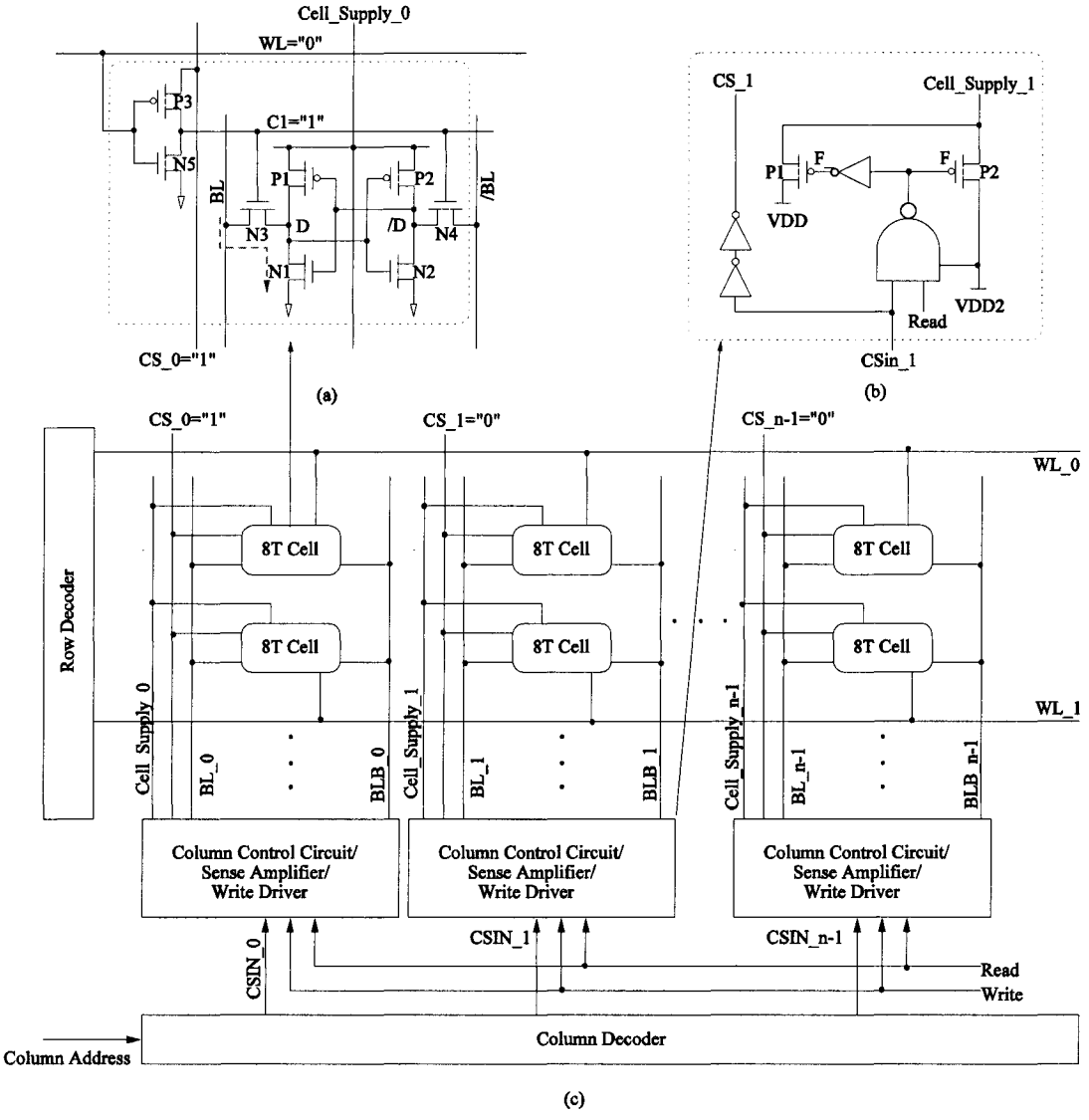


Figure 4-3 Proposed SRAM cell topology and array organization

Similar to the conventional design, it has three modes of operations: *Standby*, *Read* and *Write*; as follows:

4.3.1.1 Standby

During standby, Cell_Supply voltage is kept at V_{DD} to maintain a high noise margin. At the same time, all the WLs are pre-charged high while all the CSs are pre-charged low. As a result, transistor N5 of each cell is turned on to pre-charge node C1 to ground. Thus, both access transistors N3 and N4 are turned off, isolating the storing element from any

BL disturbances. Moreover, the BLs are pre-charged to V_{DD} in preparation for the next read/write operation.

4.3.1.2 Read operation

A read operation starts by asserting the Read signal and the row and column addresses. We assume in this discussion that the first cell of the first column is accessed. As a result, the column control circuit of the first column raises the CS_0 from ground to V_{DD} and the Cell_supply_0 is raised from V_{DD} to V_{DD2}. It is important to note that V_{DD2} must be higher than V_{DD} to improve the noise margin of the cell during the read access. At the same time, WL_0 is pulled low to drive node C1 to V_{DD} and hence turning on N3 and N4. Once N3 and N4 are turned on to read the cell data, subsequent circuit operations are the same as those of the conventional design. **Fig. 4-3(b)** shows the gist of the column control circuit in which the SA and the write driver are excluded for the sake of simplicity.

Fig. 4-4 illustrates the timing diagram of the proposed 8T design during a read operation. It is worth mentioning here that the Cell_Supply_0 must be raised to V_{DD2} before the access transistors N3 and N4 are turned on. In this work, we use an additional external voltage source V_{DD2} that provides a voltage potential of 1.5 V_{DD}. The Column control circuits (**Fig. 4-3(b)**) must ensure the correct timing sequence of the CS_0 and the Cell_supply_0. When both the Read and the CS_{in0} are asserted high, the NAND2 gate quickly pulls node F to ground, switching the Cell_supply_0 from V_{DD} to V_{DD2}. Two inverter buffers are responsible for delaying the CS_0 so that Cell_supply_0 goes high first. This column control block introduces some additional delay into the critical path of the read operation. To compensate this delay, the SA is also powered by Cell_supply_0 to reduce the sensing time. Therefore, the overall speed of the proposed design is higher than that of the conventional 6T, especially at low supply voltages.

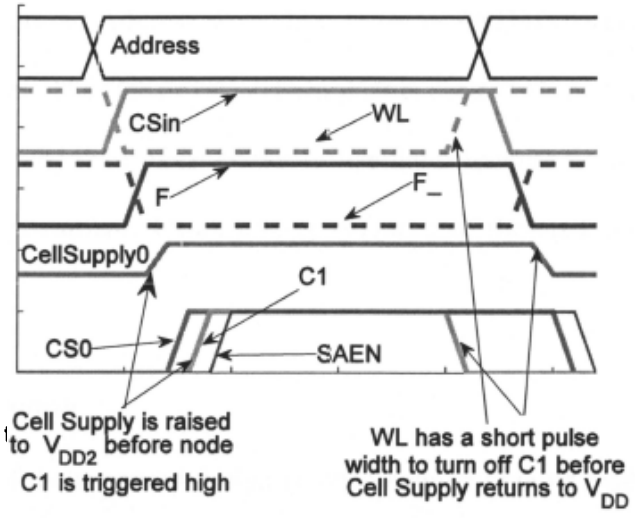


Figure 4-4 Timing diagram of the proposed design

Because only one CS_0 is turned high, only node C1 of the first cell of row 1 is pulled high to turn on the access transistors. The other cells on the same row receive the same WL signal but are not turned on since the respective CS_i signals are at ground level. Thus, only one cell on the accessed row is turned on while the others are not disturbed. This leads to two special features of the new cell: 1) They can be implemented in a bit-interleaving fashion. 2) A hefty amount of current is saved as only one cell is activated instead of the whole row. This is very much different from all previous designs in which all (or 50% in the case of single ended SRAM cells) of the cells in the access rows consume a certain amount of power even though only one of them is addressed. It is obvious that the percentage of power saving depends on the number of cells sharing a WL. In this work, we purposely choose the SRAM architecture of 128 rows × 64 columns to emphasize the difference between the proposed and the conventional design. One can easily change this arrangement with no difficulty.

At the end of the read cycle, node C1 must go low before the Cell_supply_0 signal returns to V_{DD}. This is to ensure that the cell is powered by a voltage of V_{DD} during standby and V_{DD2} during read access. Since CS_0 signal lagged behind the Cell_supply_0 signal, WL must go high to turn off N3 and N4 before the Cell_supply_0 signal goes low.

This is done by using an Address Transition Detector (ATD) to generate a pulse width which is shorter than 50% clock cycle. This signal is used to shape the pulse width of the WL. WL pulse width however must be wider than the read/write delay in order to perform reliable read/write operations. Detail timing sequence of the read operation is shown in **Fig. 4-4**.

4.3.1.3 Write operation

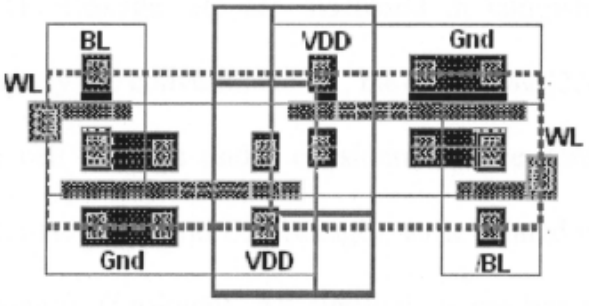
Write operation of the proposed design is much simpler than its read operation. A write starts by asserting the Write and the address signals. Concurrently, input data is available at the input stage. The CS_{in0} signal drives the CS_0 line to V_{DD} while the WL_0 is pulled down. In the meantime, one of the BLs is pulled to ground while the other is kept at V_{DD}. When node C1 is charged up to V_{DD}, both N1 and N2 are turned on and the input data is written into the memory similar to that of the conventional 6T design. It is worth mentioning here that during the write operation, Cell_supply_0 is kept at V_{DD}. In fact, one can reduce the cell supply voltage to assist a write operation. However, this requires additional circuits and is not in the scope of this work. Furthermore, as all transistors in the proposed design have the minimum size, its write-ability is better than that of the conventional 6T design. Detailed comparison of the new design's WTP will be discussed in the next section.

4.3.2 Power consumption discussion

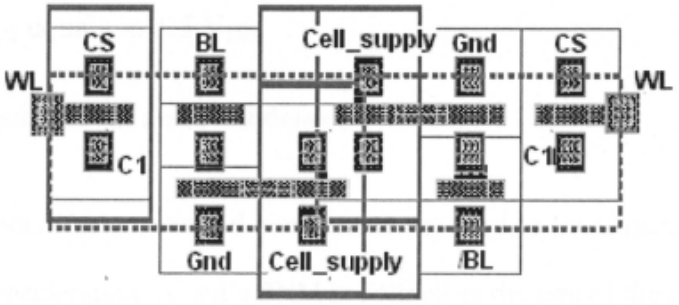
By using a dynamic cell supply scheme coupled with a special access topology, the proposed design requires extra wiring as well as power consumption in the column control circuit. Firstly, the CS signal has a full voltage swing during each active cycle. Although this is comparable with the WWL in [251] and CSL in [3], it consumes more power than the conventional 6T design at the same operating voltage and frequency. Secondly, the Cell_supply signal is driven from V_{DD} to V_{DD2} in each read operation. The additional power dissipation is proportional to the parasitic capacitance of the Cell_supply line and

the voltage swing. However, since only one cell in a row is activated during each read/write cycle, power saving from this special feature can be used to compensate the abovementioned drawbacks. Our calculation and simulation show that if the ratio between the number of cells per row and the number of cells per column is equal or higher than one twelfth, our design in fact consumes less power than the conventional design at a typical working condition of 1V/ 500 MHz in a 65 nm CMOS process.

4.3.3 Transistor sizing and layout



(a)



(b)

Figure 4-5 Layout of the SRAM cells (a) 6T, $\beta = 3$ (b) proposed 8T

As mentioned before, all transistors in the proposed cell have a minimum size of $\frac{120\,nm}{60\,nm}$. The proposed design has the same height as that of the conventional 6T but 14% longer width, using standard logic layout rules, as shown in **Fig. 4-5** (cell pull down ratio of the 6T cell is $\beta = 3$). If the cell pull down ratio of the 6T cell is $\beta = 2$, the corresponding area overhead is 33%, using the same layout rules. This is similar to that of the existing 8T design in [1] and much smaller than those of the 10T designs. Furthermore, WL, CS, BL,

Cell_supply and Gnd nodes of the proposed design can be merged with horizontally and vertically adjacent cells, and hence its layout is as compact as that of the 6T layout. In this study, to ensure a stable 6T design, we used $\beta = 3$ for the 6T and $\beta = 1$ for the proposed and the 10T designs.

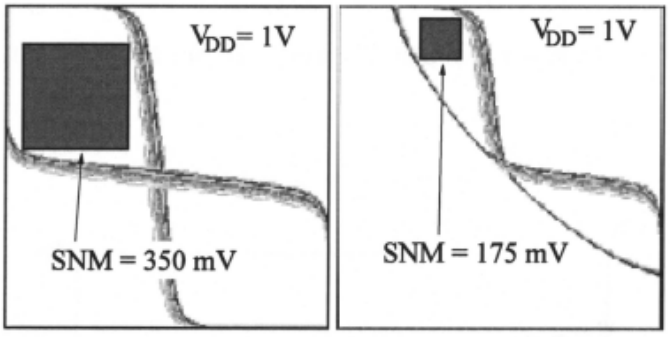
4.4 Cell performance analysis

Performance of the proposed design has been extensively evaluated as an individual as well as in a working SRAM macro. In this section, cell properties such as SNM, WTP, leakage, etc are discussed in comparison with the four existing designs, namely the conventional 6T, the original 8T [224], the 10T [251] and the 6T_1[257]. The cell features under consideration are read-related properties (SNM, read current, read delay and BL leakage), write-related property (WTP) and stand-by related property (Leakage). All simulations are based on a standard 65 nm/ 1V CMOS process using Cadence simulator. In these simulations, we always assume that V_{DD2} equals to 1.5 V_{DD}.

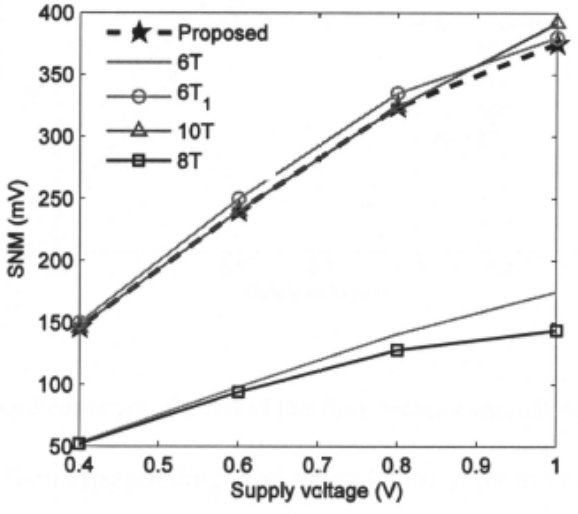
4.4.1 SNM, read current and read delay

8000-cycle Monte Carlo statistical simulations are used to investigate the SNM of the SRAM cells in consideration. A cell's SNM is defined as the size of the largest square that can fit into the eyes of the statistical butterfly curves. For example, **Fig. 4-6 (a-b)** illustrates the SNM of the conventional 6T design during standby and read modes. The 6T'SNM is therefore equal to 175 mV as its SNM during read is smaller than that during standby. **Fig. 4-6 (c)** summarizes the SNM of the five designs in consideration against V_{DD} variations from 1 V to 0.4 V. It is apparent that the proposed, the 6T_1 and the 10T designs are the best performers with similar SNM. This is because the 10T design only accesses the cross-coupled structure during the write operation; otherwise it is isolated from any disturbances. The proposed design raises cell supply to 1.5 V_{DD} and hence its

SNM is significantly improved, in the same manner as the 6T_1 design. At 1V, its SNM is slightly smaller than that of the 10T but much higher than those of the 8T and 6T cells. As the supply voltage scales down (i.e. < 0.9 V), the proposed design's SNM equals to that of the 10T. This is because of the assumption that V_{DD2} equals to 1.5 V_{DD} and thus at low supply voltage, the proposed design's SNM is limited by its standby SNM. In practice, circuit designers can choose a different value of V_{DD2} which may result in different SNM but as long as V_{DD2} is higher than V_{DD}, cell's SNM is significantly improved and the cell works much better than the conventional 6T and the 8T at very low-voltage operations.



(a) (b)



(c)

Figure 4-6 SNM of the 6T design during (a) Standby (b) Read. (c) SNM comparison of the four designs against supply voltage variation

Other than SNM, cell current is also an important property of an SRAM cell design. It is desirable to have a large cell current so that the read operation can be performed as quickly and reliably as possible. **Fig. 4-7** graphically compares the strength of each cell current during the read operation. At 1 V, all five designs have similar cell currents, except the 10T [251] which has two minimum-sized transistors cascaded in the read port [251]. Although the proposed design also uses minimum-sized transistors, its cell supply is raised to a higher voltage and hence its read current is comparable to that of the conventional 6T design. When supply voltage reduces to sub-threshold region, our dynamic cell supply shows its effectiveness clearer. For example, at 0.4 V, the read current of the 6T, the proposed, the 8T, the 10T and the 6T_1 are 173 nA, 430 nA, 173 nA, 173 and 430 nA, respectively.

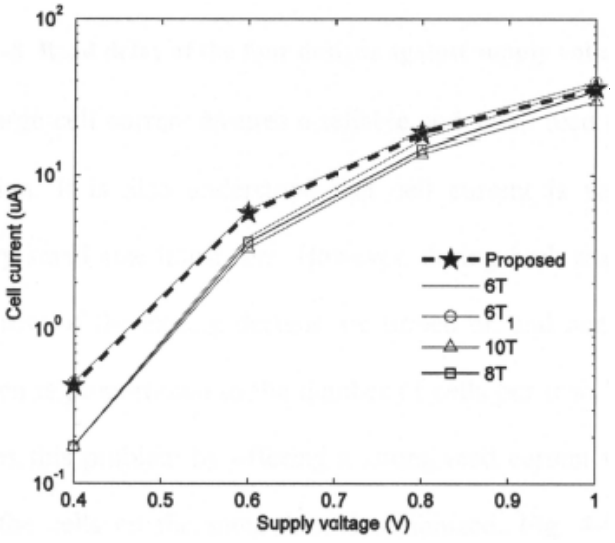


Figure 4-7 Active read currents of a cell of the four designs against supply voltage variation.

Fig. 4-8 shows the corresponding read delay of the cells in consideration. This delay refers to the time required for the cell to discharge the *BL* from V_{DD} to 90% V_{DD}. It agrees with the data reported in **Fig. 4-7** as the discharging time is reciprocally proportional to the cell current. At any supply voltage, the proposed design has a read delay similar to that of the 6T designs, which is better than the 10T and 8T cells Our design is slightly slower than

the conventional 6T since the CS line must be driven to V_{DD} before the accessed transistors of the memory cell are turned on.

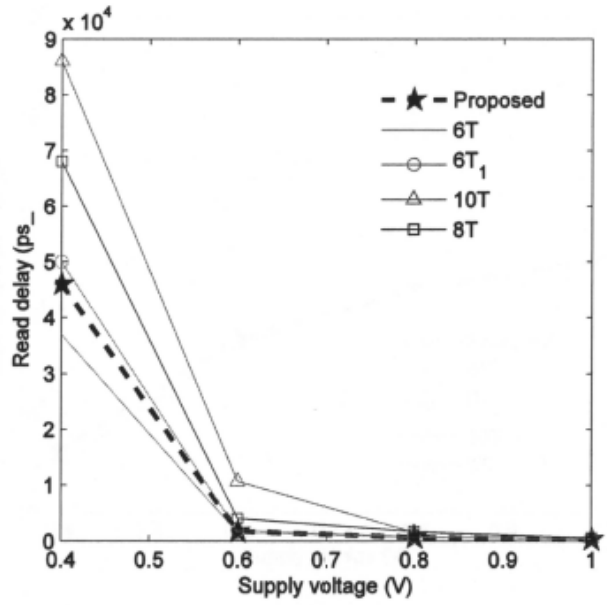


Figure 4-8 Read delay of the four designs against supply voltage variation

Although a large cell current ensures a reliable and small read delay, it implies high power consumption. It is also understood that cell current is very small as the cell constitutes of very small size transistors. However, during each read/write operation, all the cells on one row of the exiting designs are turned on and hence the corresponding power consumption is proportional to the number of cells per row. The proposed 8T has successfully solves this problem by offering a strong read current while the total power consumption by the cells on the same row is minimized. **Fig. 4-9** illustrates the total current drawn by a row of 64 cells during read operations of the five designs. Comparing **Fig. 4-7** and **Fig. 4-9**, it is quite clear that the proposed design not only has the strongest read current but also consumes the least power on the accessed row. For example, at 1 V supply voltage, read current per row of the proposed and the 8T are 35 uA and 33 uA while those of the 6T, 10T, and 6T_1 are 2.3 mA, 1.88 mA and 2.5 mA, respectively. As mentioned previously, this power saving is very valuable and can be used to compensate the additional energy used to drive the *Cell_supply* and *CL* signals. Comparing **Fig. 4-9**

and **Fig. 4-7,** it is shown that our design has the low-power feature of the 8T design but offers the stability of the 6T_1 and the 10T designs.

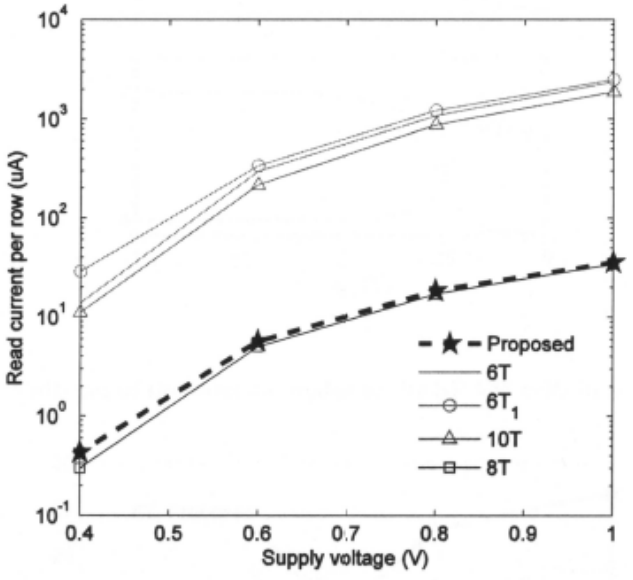


Figure 4-9 Active read current of a row of 64 cells of the four designs against supply voltage variation.

4.4.2 WTP

During the write operation, both *CS* and *Cell_supply* of the proposed design are at V_{DD}. Thus, biasing voltages of the transistors involving in the write cycle of the proposed 8T and 6T are exactly the same. However, the pull-down transistors of the proposed design is smaller than that of the conventional 6T, making it easier to be written to. As a result, the proposed design has a 20% higher WTP when compared to the 6T designs at 1 V, as shown in **Fig. 4-10** and **Fig. 4-11**. These four designs in turn have better WTPs than the 10T design, as the 10T cell has two transistors cascaded in a write cycle[251]. This reaffirms that the proposed design improved both SNM and WTP during the read and the write operations.

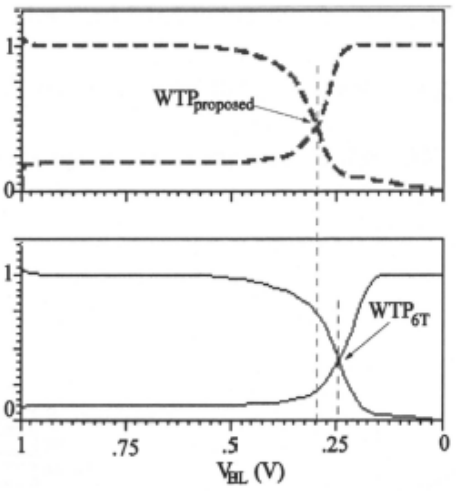


Figure 4-10 Voltages of the internal nodes of the SRAM cells in a write operation

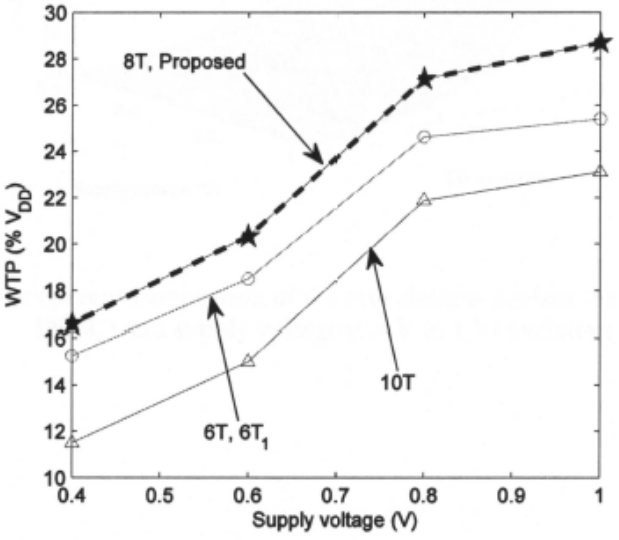


Figure 4-11 WTP comparison of the four designs against supply voltage variation. WTP are normalized to VDD.

4.4.3 Cell Leakage

Leakage current is one of the major concerns in nano-scale SRAM where most of the transistors are in the standby mode [207]. **Fig. 4-12** shows the leakage currents of the five designs in consideration in a wide range of temperate (0^0 C to 100^0 C) and supply voltage (0.4V to 1V). Despite having more transistor count, the 8T cells has less leakage current when compared to the 6T cells with about 10% leakage reduction at 1V and 3% reduction

at 0.4 V, respectively. It can be explained using the fact that the *CS_i* signals are at ground level during standby and all transistors in the proposed design have minimum size. Our design also has a smaller leakage when compared to the 10T design, as shown in **Fig. 4-12**.

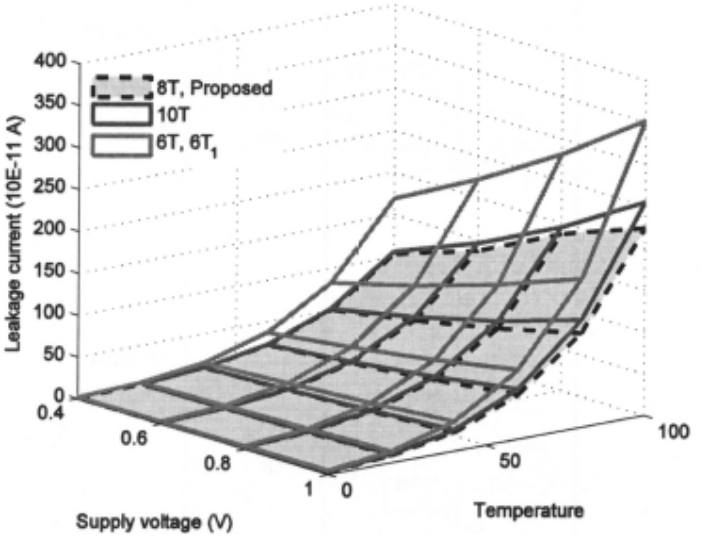


Figure 4-12 Leakage current comparison of the four designs against the temperature (0 ^{0}C to 100 ^{0}C) and supply voltage (0.4 V to 1 V) variation.

4.4.4 BL leakage

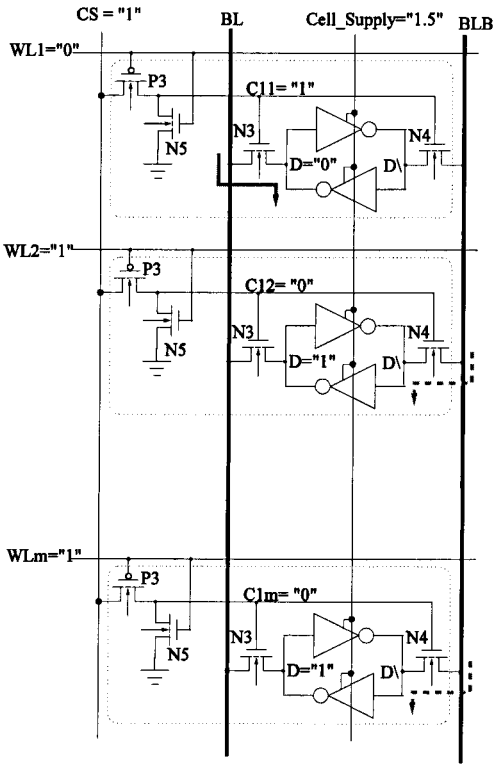


Figure 4-13 BL leakage paths that can malfunction the read operation in the worst case scenario

As mentioned in *section 3.3.2.2*, BL leakage refers to the leakage currents of the cells in the column. **Fig. 4-13** shows the worst-case leakage path of the proposed cell during a read cycle. Similar to that in chapter 3, the ratio $\frac{I_{on}}{I_{off}}$ (also known as $\frac{I_{cell}}{I_{leakage}}$) is used to measure the reliability of the read current during a read operation. Detailed explanation of this term can be found in *section 3.3.2.2*.

Fig. 4-14 represents the simulation results of the five designs inconsideration. It is noticeable that the proposed 8T has the best $\frac{I_{on}}{I_{off}}$ ratio at supply voltages smaller than 0.9V while the 10T design has the worst ratio at every operating voltage. This can be explained as follows: The 10T has two cascaded minimum sized transistors in a read path but has two leakage currents: One flows from the BL to the read port, the other flows into the cell

through its write port. As a result, its current ratio is worse than the other design. The proposed design has the same topology as the conventional 6T design during a read operation. However, its cell supply is raised to a higher voltage and hence its read current is better when the supply voltage scales down. As a result, its cell current ratio is only slightly smaller than that of the 6T at 1 V but becomes better when the supply voltage reduces to less than 0.9 V. It is similar to that of the 6T_1 design, as this design also renders a raised cell supply during the read operation. This suggests that the proposed and the 6T_1 circuits work better than the other designs at very low supply voltage condition and has a similar performance at high supply voltage operation (when $\frac{I_{on}}{I_{off}}$ is less important).

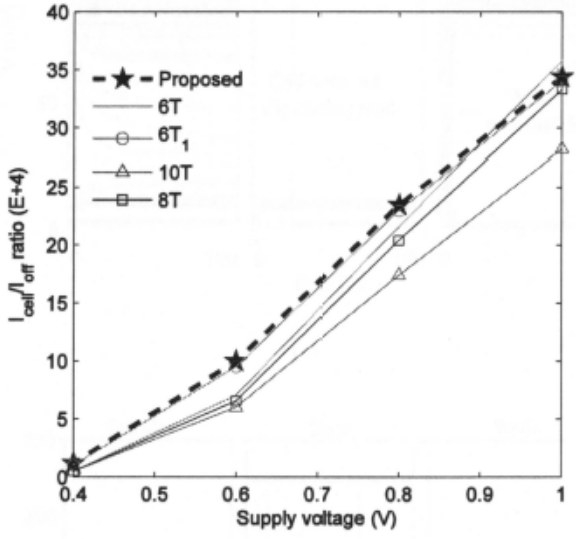
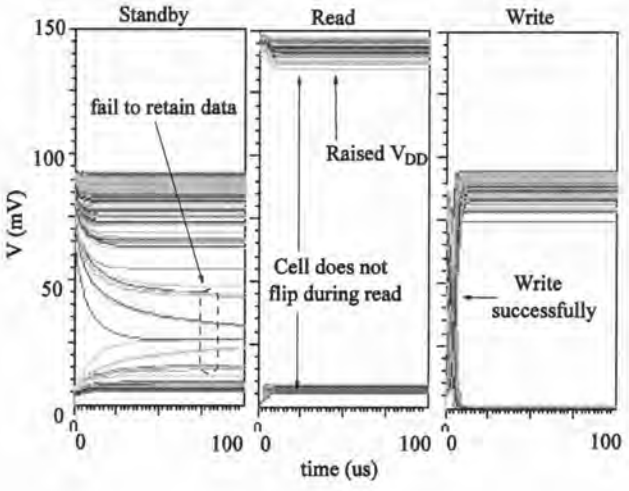


Figure 4-14 $\frac{I_{on}}{I_{off}}$ ratios of the memory cells in comparison at different supply voltages

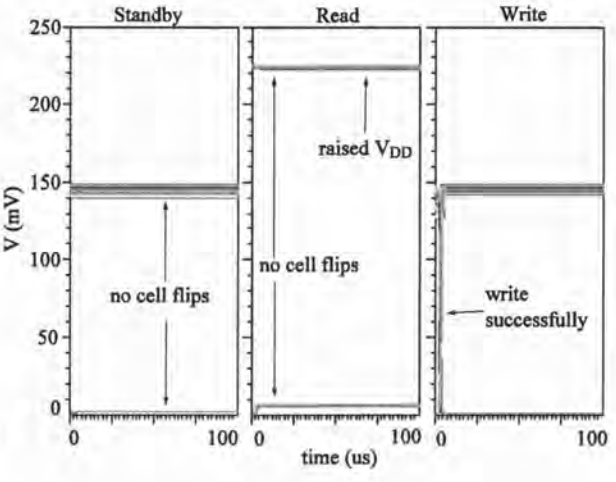
4.4.5 V_{DDmin}

V_{DDmin} refers to the lowest supply voltage at which the cell can retain data, read and write. To find the V_{DDmin} of a cell design, three basis operations (i.e. standby, read and write) are performed at very low supply voltages. We also assume that there is no time limit to perform these tasks. A cell fails to retain data or read if its value flips after a

certain duration of time (a few hundred μs or ms). On the other hand, a cell is considered fail to write if its data does not flip after a certain duration of time. To obtain a reliable conclusion, we performed Monte-Carlo simulations on each cell using the standard process variation data from the foundry. For example, **Fig. 4-15** shows the simulation waveforms of the proposed 8T design during standby, read and write in 8000-cycle Monte-Carlo simulations at 100 mV and 150 mV, respectively. At 100 mV supply voltage, all cycles pass the read and write operation but 9 cycles out of 100 fail to retain the stored data during standby. At 150 mV, all 8000 cycles read, write and retain data successfully.



(a)



(b)

Figure 4-15 8000-cycle Monte-Carlo simulation of the proposed design to determine its V_{DDmin}. (a) V_{DD} = 100 mV→failed. (b) V_{DD} = 150 mV→pass

For each design, the V_{DDmin} is defined as the minimum supply voltage at which it retains data, reads and writes successfully in all 8000 cycles of Monte-Carlo simulations. This provides a much more reliable conclusion than using only the process corner simulations. An interesting finding is that none of the five designs fails to work due to its write operation. It is different from the our initial impression that the two 10T design's V_{DDmin} is limited by their write operations as they use two cascaded NMOS to access to the cell during a write operation. Our simulation results show that the proposed, the 6T_1 and the 10T have the lowest V_{DDmin}. The reason is that the 10T is isolated while the proposed and the 6T_1 have a raised V_{DD} during a read operation. As a result, both are limited by its standby mode when the isolated cross-coupled structure is no longer able to retain the written data. On the other hand, the 6T and 8T designs have disturbed read operations and hence this mode of operation limits its V_{DDmin}. Table IV summarizes the V_{DDmin} values of all five designs.

TABLE IV. SUMMARY OF V_{DDMIN} OF THE FIVE SRAM DESIGNS.

	Proposed	6T	6T_1	10T	8T
VDDmin	150 mV	250 mV	150 mV	145 mV	250 mV
Limited by	Standby	Read	Standby	Standby	Read

4.5 VLSI implementation

4.5.1 Macro architecture

Two 128-row × 64-column SRAM macros have been designed in a standard 65 nm CMOS process from STM using the conventional 6T and the 8T cells. Both macros have identical address decoders, data-line drivers and SA design. Extensive Read/Write operations have been simulated at 40^0 C to evaluate the performance of the newly proposed cell. For a fair comparison, we do not include the 10T SRAM designs in this

comparison. The two SRAM macros are tested using supply voltages from 1V to 0.2 V. At each supply voltage, extensive trials were carried out to find out the maximum frequency at which the macros work correctly. The corresponding results are presented in **Section 4.5.2** below.

4.5.2 Performance summary

Fig. 4-16 shows the maximum operating frequency of the 6T and the 8T designs at different supply voltages. At 1 V supply voltage, the proposed design has a maximum frequency of 714 MHz whereas that of the 6T design is 500 MHz. As supply voltage scales down, the proposed design consistently outperforms the 6T design, as its read operation is backed by the high-speed SA, which is powered by the *Cell_supply* line. Furthermore, the 6T design ceases to work at 0.3 V while the proposed design continues to work properly even at 0.2 V supply. The minimum V_{DD} of the SRAM macros is limited by its readout circuits as the active read current becomes exceptionally weak and hence full output swing is not obtainable. Thus, both macros cease to work at a supply voltage higher than V_{DDmin} of the cells reported in **Section 4.4.5** above.

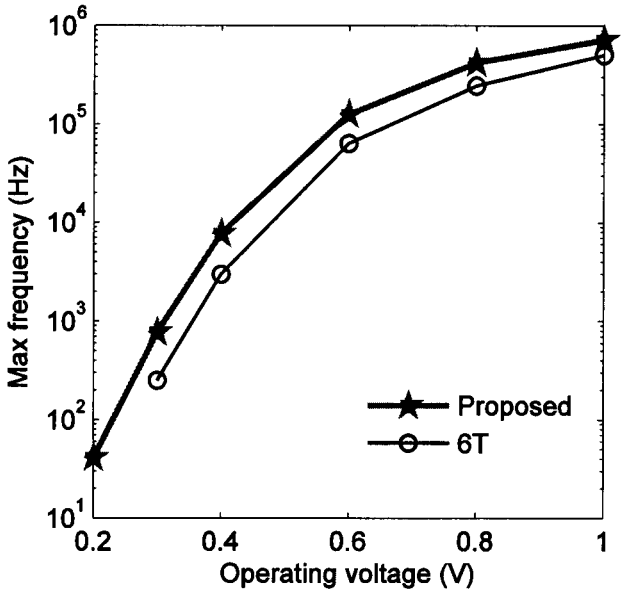


Figure 4-16 Maximum frequency of the 6T and 8T designs at various operating supply voltages.

Other than the maximum frequency, we have also investigated the average power consumed by each macro at different operating voltages, as shown in **Fig. 4-17**. As mentioned in **Section 4.4**, the proposed design consumes much less power than the 6T design as only one cell is activated per row. For example, at 1V, 700 MHz operating frequency, the proposed design consumes 659 μW in a read and 718 μW in a write operation whereas those of the 6T designs are 1.56 mW and 155 mW, respectively, at 1V, 500 MHz. At 0.3 V supply, the proposed design has a maximum frequency of 770 kHz and consumes 117 nW per read while the 6T design has a maximum frequency of 250 kHz but consumes 190 nW per read. This confirmed that the 8T design is not only faster but also consumes less power than the 6T design.

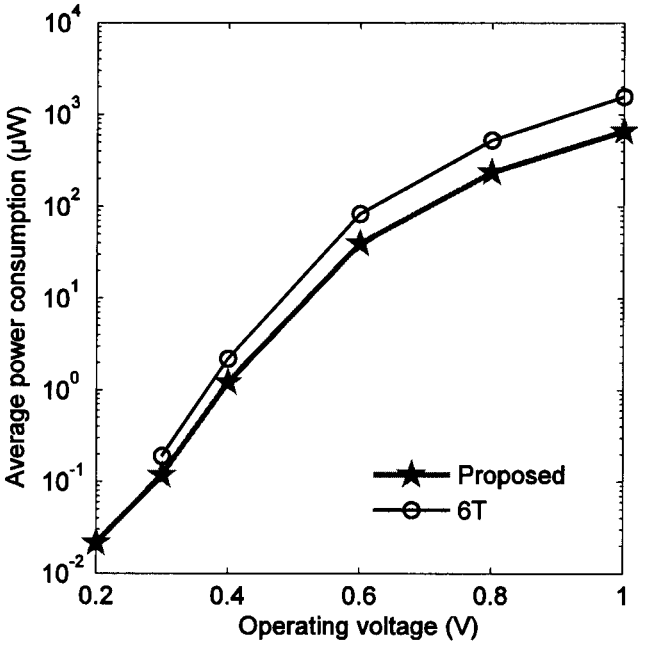


Figure 4-17 Average active power of the 6T and 8T designs.

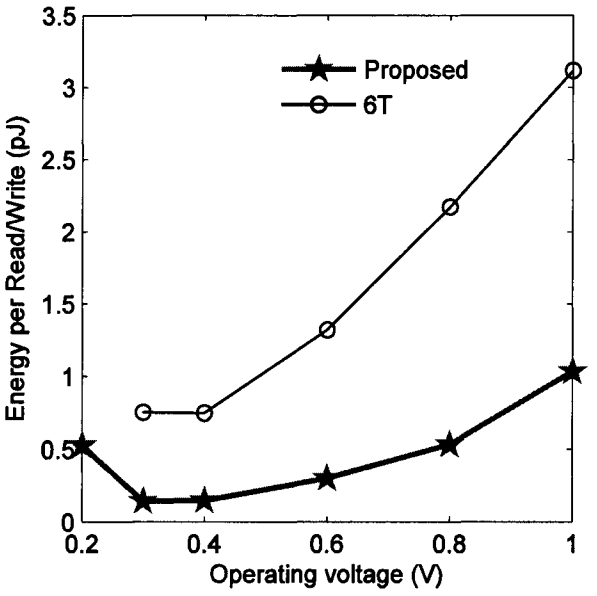
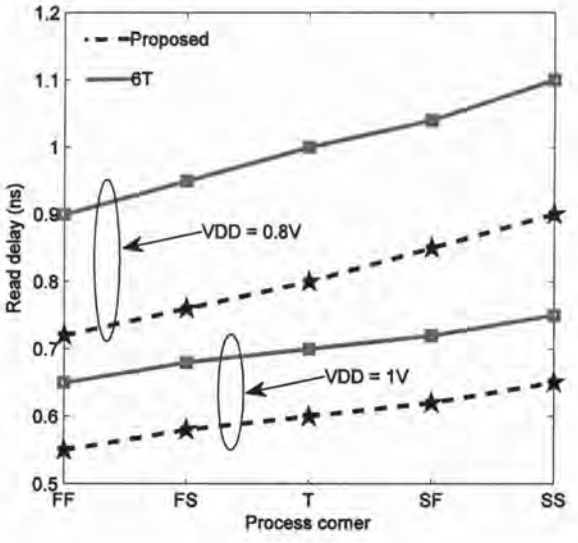


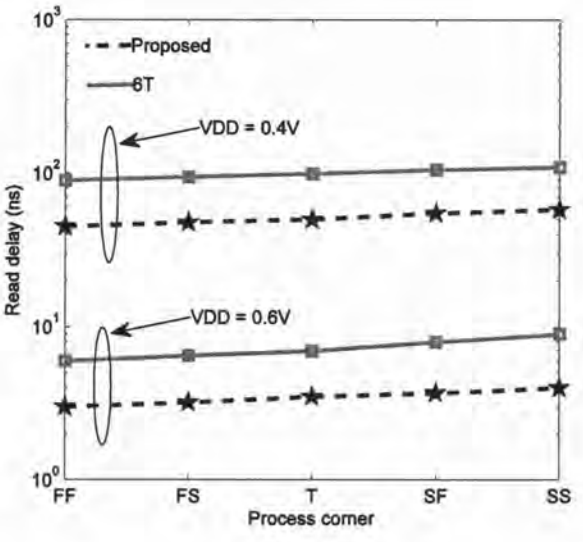
Figure 4-18 Energy consumed per active cycle of the 6T and 8T designs.

Fig. 4-18 emphasizes the impact of the operating voltage on the average energy consumption of the chip. As supply voltage scales down, the chip will certainly require less energy to perform a task. However, there is a limit at which the V_{DD} should not be further reduced. **Fig. 4-18** clearly indicates that at 0.4 V, the 6T design obtains is minimum energy consumption point. Further reducing V_{DD} to 0.3 V does not improve the actual energy consumption. This is because the corresponding delay is lengthened proportionally to the amount of power saving. The proposed design behaves the same way as the 6T design, as illustrated by the bold blue line in **Fig. 4-18**. Furthermore, as the voltage supply is reduced to 0.2 V, the consumed energy drastically increases. This is because at this voltage, the operation of the chip is exceptionally slow, which is more than 10x longer that that at 0.3 V, while the power reduction is only marginal. This finding agrees with the data reported in [251], as previously discussed in **Section 4.1**. It suggests that unless constrained by the power consumption, reducing the V_{DD} to below 0.3 V is not energy-saving. We also carried out the corner simulations on the two designs to make sure that both circuits works correctly at every process corner.

Fig. 4-19 confirms this and on top of that the proposed design always has smaller delay when compared to the 6T SRAM. All of the previously observed relationships between the speed of the two design continue to hold at all five process corners with the proposed design being the faster one and this advantage becomes more profound as the supply voltage scales down. Table V summarizes the performance of the five SRAM designs in consideration.



(a)



(b)

Figure 4-19 Read delay of the two SRAM macros in consideration at different process cornes: Fast-Fast (FF), Fast-Slow (FS), Typical (T), Slow-Fast and Slow-Slow (SS). (a) V_{DD} equals to 1V and 0.8V (b) V_{DD} equals to 0.6 V and 0.4 V

TABLE V. SUMMARY OF THE PERFORMANCE OF THE FOUR SRAM DESIGNS.

	Conventional 6T	**Proposed**	8T	10T	6T_1
CMOS Process	65 nm / 1V	**65 nm / 1V**	90 nm PD/SOI	90 nm *	65 nm
Read/write scheme	Differential	**Differential**	Differential	Differential *	Differential
Capacity	8K	**8K**	32K*	32K *	70M*
VDD_{min}	0.25 V	**0.15 V**	0.25 V (0.55 V*)	0.145 V	N. A
Operating Voltage	0.3 V	**0.2 V**	N. A	0.3V *	N. A
Maximum frequency	3.3MHz	**41KHz**	N. A	500KHz *	N. A
Normalized Cell area	1	**1.14**	**1.14**	2.08 *	1
Leakage per bit at 1V	0.60 nA	**0.53 nA**	**0.53 nA**	0.83 nA	0.6 nA
SNM at 0.4V	53 mV	**190 mv**	50 mV	146 mV	190 mV
WTP	61 mV	**66.3 mV**	61 mV	46 mV	61 mV
* : *numbers are based on the respective published paper.*					

4.6 Conclusion

A fully differential 8T SRAM with a single column-based dynamic supply has been proposed and analyzed. The proposed design improved both SNM (2x) and WTP (10%) with a marginal area overhead (14%) when compared to the conventional 6T design. In addition, only one cell is accessed during a read or write operation, allowing it to be bit-interleaved for an efficient ECC. On top of that, its power consumption is also reduced to 46% when compared to the conventional 6T design. Our design offers the stable features of the 10T and the 6T_1 design while maintaining the power consumption as low as the 8T design in [224]. Furthermore, the column-based dynamic supply control can be modified to accommodate different requirement. For example, instead of having a high and a normal column supply, one may employ a high column supply during a read, low column supply for a write and a normal supply voltage during standby. By using the 8T cell

coupled with the column-based dynamic power supply, this work has successfully separate the read, write, standby operation of the memory cell. Besides, the accessed cell is also separated from its neighbors so that it they can be bit-interleaved. These attractive features make it the best choice in the five cells in Table V for state-of-the-art application where a robust performance at ultra-low supply voltages is of critical requirement.

CHAPTER 5 LATCH-BASED CURRENT-MODE SA DESIGNS

In **Chapter 3** and **4,** two SRAM cell designs have been introduced. Several SRAM macros have also been implemented to evaluate the performance of the proposed SRAM cells. However, for a fair comparison, only conventional write circuits and conventional current-mode SA are used. In this chapter, we propose two SA configurations that improve both sensing speed and power consumption of the read operation of the whole SRAM macro. Our proposed circuits are based on the cross-coupled structure, which is widely used in contemporary SRAM designs. These designs, if used, will certainly enhance the performance of the memory reported in **Chapter 3** and **4**.

5.1 Hybrid-mode SA: a new perspective on transistor sizing

A novel high-speed SA for ultra-low-voltage SRAM applications is presented. It introduces a completely different way of sizing the aspect ratio of the transistors on the data-path, hence realizing a current-voltage hybrid mode SA. Extensive post-layout simulations have proved that the new SA provides both high-speed and low-power properties, with its delay and power reduced to 25.8% and 37.6% of those of the best prior art. It also offers a much better read-effectiveness and robustness against the bit- and DL capacitances as well as V_{DD} variations. Furthermore, the new SA is able to tolerate a large difference between the parasitic capacitances associated with the complementary DLs. It can operate down to a supply voltage of 0.9 V, the lowest reported for a 0.18 μm CMOS process. A modified cross-coupled amplifier is also introduced, allowing the SA to operate down to 0.55 V.

5.1.1 Current-mode SA and its derivatives

The current-mode sensing scheme **(Fig. 5-1)** in SRAM applications was first introduced in [258]. The current-mode SAs, marked by the presence of the conventional current-conveyor [258], is insensitive to the C_{BL} and hence, offers a higher sensing speed and consumes less power compared to the voltage mode counterparts [258]. Over the last two

decades, a number of current-mode SAs have been proposed, aiming at improving the sensing speed and the power consumption during the read operation of the SRAM [44, 259]. Since all of these SA designs utilize the differential output currents of the current conveyor, their improvement is only incremental. Ref. [258] clearly indicated that the above-mentioned differential current is equal to the current flowing into the cell node where a '0' is stored, i.e. I_{cell}. However, our analysis and simulations have proved otherwise. In fact, the differential current is much smaller than the I_{cell}, due to the imperfection of the current conveyor. This issue will be discussed in more details in the next sections.

5.1.2 Imperfections of the current-conveyor based current-mode designs

The current conveyor consists of four identical pMOS transistors P2, P3, P4 and P5 (**Fig. 5-1**). Since this configuration is common to all current-mode SAs, its drawback is also their shared weakness. As mentioned in [258], the current conveyor realizes a virtual short-circuit across the complementary BLs, i.e. $V_{BL} = V_{/BL}$ during the read operation. Therefore, the currents I_0 and I_1, which are sourced by the large-sized BL load transistors P0 and P1 respectively, are equal and a current difference of I_{cell} is realized at the inputs of the current conveyor. This is only true in an ideal case where the process variations and short channel length effect are not present. In deep submicron technologies, these effects become significant and there is a slight difference between the BL voltages. For instance, at V_{DD} = 1.8 V, $W_{P0,1}$ = 15 µm and $W_{P2,3,4,5}$ = 1 µm (**Fig. 5-1**), the corresponding V_{BL} and $V_{/BL}$ are 1.79 V and 1.77 V respectively. This 0.02 V difference makes a significant impact on the effectiveness of the read process. It can be explained in the following manner: Since V_{BL} and $V_{/BL}$ are close to V_{DD}, both P0 and P1 operate in the triode region, where their drain currents are proportional to their drain-to-source voltages. Thus, by using V_{DD} = 1.8 V, V_{BL} = 1.79 V and $V_{/BL}$ = 1.77 V, we have I_0 three times as large as I_1 (Table IV, second column). Consequently, the effective difference between the BL currents after taking the I_{cell} into account is very small (ΔI in Table IV).

We use $\%\,\text{utilization} = \frac{\Delta I}{\text{Total current}} = \frac{I_1 - I'_0}{I_0 + I_1}$ (5.1) as a figure of merit to measure the effectiveness of the read scheme. In our calculations, we assumed that I_{cell}, which is measured from the standard 6T cell, does not change during the sensing process. As indicated in Table VI, at various transistor sizes, *% utilization* of the current-mode SA is only around 10% with ΔI being much smaller than I_{cell}. Therefore, it can be concluded that the differential current in the current-mode needs to be improved in order to obtain a more effective read operation, both in terms of speed and power consumption.

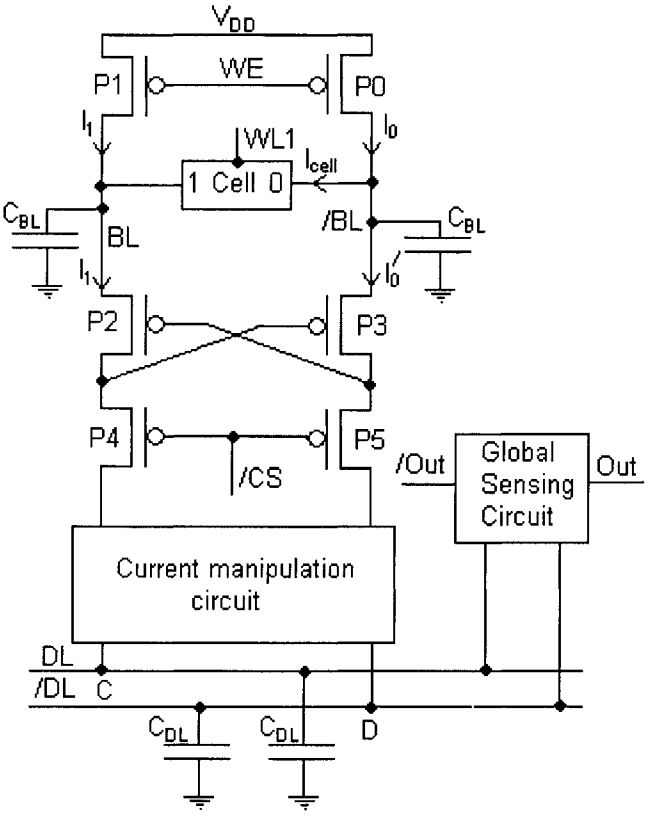


Figure 5-1 Current-mode SA with a current-conveyor incorporated

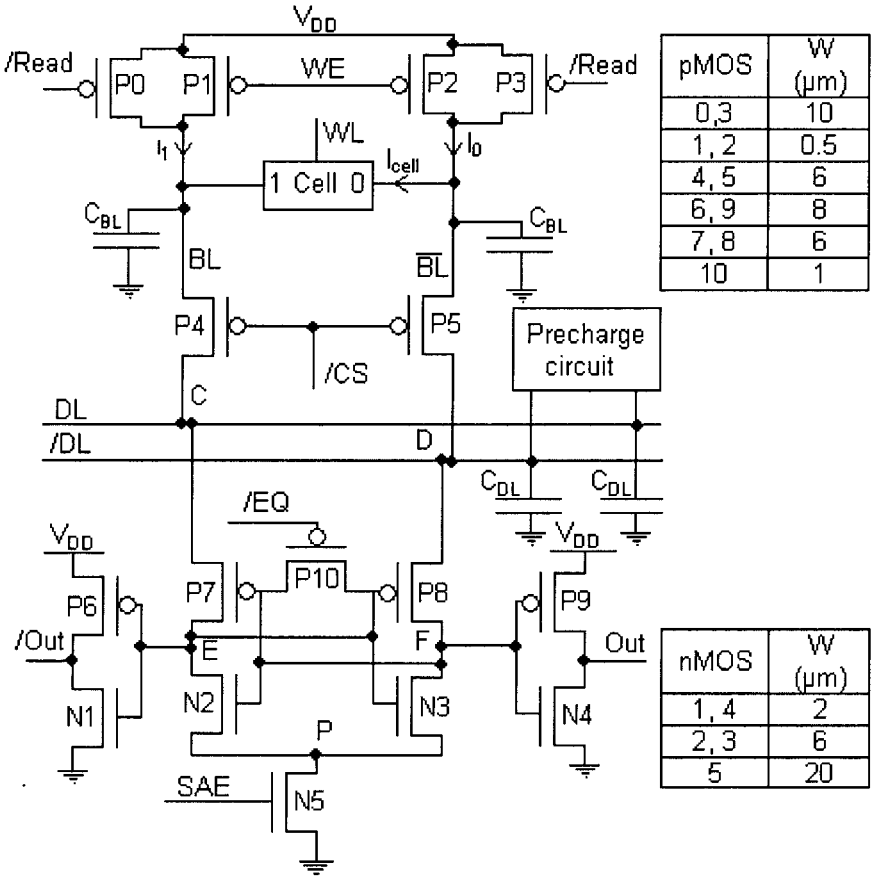
TABLE VI. SUMMARY OF CURRENTS CONSUMED DURING A READ CYCLE IN A CURRENT-MODE SA. $W_{P0,1}$ = 15 μM, I_{CELL} = 92 μA. $I_0' = I_0 - I_{CELL}$. $\Delta I = I_1 - I'_0$.

$W_{P2-,5}$ (μm)	1	2	3	4	5	6	7
I_0 (μA)	119	146	174	198	221	244	267
I_1 (μA)	38	76	113	143	172	199	226
I_0' (μA)	27	54	82	106	129	152	174
ΔI (μA)	11	22	31	37	43	47	52
% utilization	*7.0*	*9.9*	*10.8*	*10.9*	*10.6*	*10.5*	*10.2*

5.1.3 The proposed hybrid current-mode SA

5.1.3.1 Circuit operation

Our proposed SA is presented in **Fig. 5-2**. It consists of 11 pMOS and 5 nMOS transistors (P0 - P10 and N1 - N5). P0 and P3 are responsible for pre-charging the BLs to V_{DD} while P1 and P2 are for holding the BLs at V_{DD} during the read-cycle. P4 and P5 act as a switch to connect the BLs to the DLs. The other 10 transistors (P6 - P10 and N1 – N5) form a cross-coupled inverter which amplifies the small voltage difference on the DLs to the full CMOS logic levels [258].



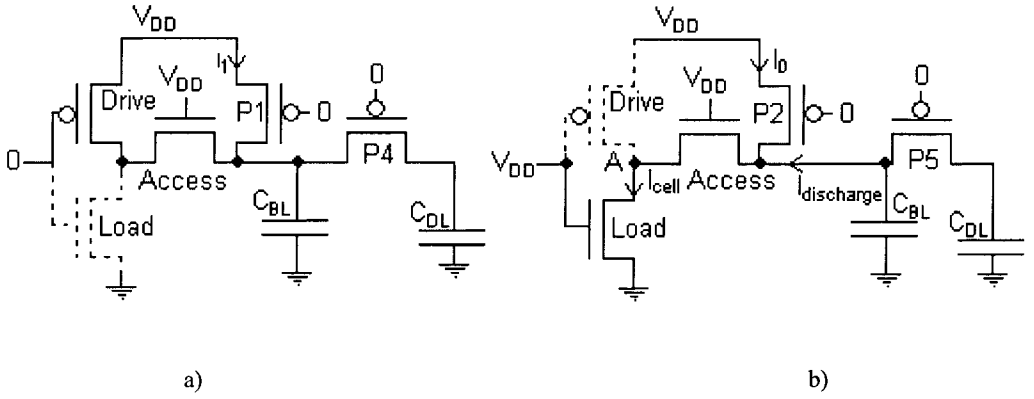
pMOS	W (µm)
0,3	10
1, 2	0.5
4, 5	6
6, 9	8
7, 8	6
10	1

nMOS	W (µm)
1, 4	2
2, 3	6
5	20

Figure 5-2 The proposed SA with a simplified read-cycle-only memory system. Channel lengths of all transistors are 0.18 µm

The new SA is unique in a way that it eliminates the current conveyor and transforms the normally-large BL loads (P0, P1 in **Fig. 5-1**) into small sized transistors (P1, P2 in **Fig. 5-2**). They serve to hold the BLs at V_{DD} and not to source the BL currents unlike in the conventional designs. P1 and P2 are purposely sized small so that it is not strong enough

to keep the BL at V_{DD} if I_{cell} is present. As a result, one of the BLs will drop to a lower level than V_{DD} during a read access.



a) b)

Figure 5-3 Current paths during the read cycle in the proposed SA on the side where a) a '1' is stored. b) a '0' is stored

Before any read cycle, the BLs and DLs are pre-charged to V_{DD} by (P0, P3) and the pre-charge circuit respectively (**Fig. 5-2**). Meanwhile, the SA Enable (SAE) signal turns off N5 to prevent any DC current from flowing to the ground to save power whilst the Equalization (/EQ) signal turns P10 on to hold the two nodes E and F at the same potential. When a cell is accessed by the WL (WL) and Column Select (/CS) signals, the /Read signal is triggered high to deactivate P0 and P3. The pre-charge circuit is then turned off, but P1 and P2 are still on in order to hold the BLs and the DLs at V_{DD}. **Fig. 5-3** presents the current paths during a read cycle. The Access, Drive and Load are transistors of the accessed memory cell. We assume that the first cell of the shown column (**Fig. 5-2**) is accessed. On the left side where a '1' is stored, the BL and DL remain at V_{DD} since they are held by P1 while no discharge current is available (**Fig. 5-3a**). Also, since both DL and BL are kept at V_{DD}, the drain-to-source voltage of P1 is very small and hence, it only sources a negligible current I_1 during a read cycle, just enough to complement the leakage current along the BL. On the side where a '0' is stored, the cell sinks a current I_{cell}, which is larger than the current I_0 sourced by P2. As a result, a discharge current of $I_{discharge}$ ($I_{discharge} = I_{cell} - I_0$) is available to discharge the /BL and /DL. $I_{discharge}$ then discharges both the /BL and /DL to voltage levels lower than V_{DD}. The cross-coupled amplifier (P6 – P10

and N1 – N5) will sense the difference between the DLs and amplify it to a full CMOS logic level. It is worth mentioning here that the /Read is triggered high during the write cycle to turn off P0 and P3.

5.1.3.2 Read effectiveness

Similar to the current-mode SA in **Fig. 5-1**, *% utilization* is also used to measure the new design's read effectiveness. All simulated results are presented in Table V. It is found that the *% utilization* of the new design is higher than 100%. This can be explained as the total supplied current ($I_1 + I_0$) is much smaller than I_{cell}. However, during standby, a current equivalent to $I_{discharge}$ mentioned above is used to pre-charge one of the BLs to V_{DD} (the other BL is already at V_{DD} and no charging-up is needed). Therefore, the *% utilization adjusted* $= \frac{\Delta I}{\text{Total current}} = \frac{\Delta I}{I_1 + I_0 + I_{discharge}}$ is a more appropriate measure of the effectiveness of the read scheme. It is shown in Table VII that the new design offers a much better % utilization with less power consumption and stronger discharge current than the current-mode scheme as presented in Table VII.

TABLE VII. SUMMARY OF CURRENTS CONSUMED DURING A READ CYCLE IN THE PROPOSED SA.

$I_{CELL} = 92\ \mu A$. $I_0' = I_0 - I_{CELL}$. $\Delta I = I_1 - I'_0$

$W_{P1,2}$ (μm)	0.3	0.5	0.7	0.9	1.1	1.3	1.5
I_0 (μA)	4	6	8	10	13	15	17
I_1 (μA)	3	4	4	4	5	5	5
I_0' (μA)	-88	-86	-84	-82	-79	-77	-75
ΔI (μA)	91	90	88	86	84	82	80
% utilization	1308	900	735	614	467	410	363
% utilization (adjusted)	*96*	*94*	*92*	*89*	*86*	*84*	*82*

5.1.3.3 Tolerance to the difference between C_{DL} and $C_{/DL}$

All of the SA designs in the comparison (i.e. Ref. [259-261]) are based on the assumption that $C_{DL} = C_{/DL}$. Therefore, the voltage difference developed on the DLs is governed by Eq. (3.2):

$$V_{DL} = V_o + \frac{I_{BL} * \Delta t}{C_{DL}} \qquad V_{\overline{DL}} = V_o + \frac{I_{/BL} * \Delta t}{C_{/DL}} \tag{4.2a}$$

$$\therefore \Delta V = V_{DL} - V_{/DL} = \Delta t * \frac{(I_{BL} - I_{/BL})}{C_{DL}} = \Delta t * \frac{\Delta I}{C_{DL}} \tag{4.2b}$$

When C_{DL} differs from $C_{\overline{DL}}$, the SA works properly only if $\frac{I_{BL}}{C_{DL}} \geq \frac{I_{/BL}}{C_{/DL}}$. With the BL currents presented in Table IV, it can only tolerate up to 10% difference between the two DL capacitances with reasonable sensing delay. The new design in contrast does not rely on the relationship between C_{DL} and $C_{/DL}$ since only one discharge current is available on one branch while on the other branch, voltage levels (of both the DL and BL) are kept at V_{DD}. **Fig. 5-4** illustrates how the sensing delay changes with the difference between C_{DL} and $C_{/DL}$. We varied $C_{/DL}$ up to ±50% of C_{DL} and the new SA still works with minimum sensing delay variations.

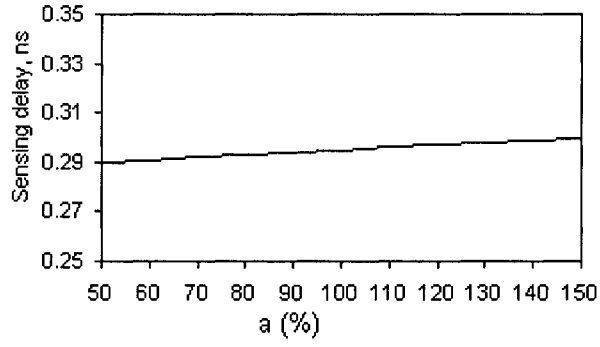


Figure 5-4 Sensing delay versus the difference between C_{DL} and $C_{/DL}$ at C_{DL} = 3 pF, $C_{/DL}$ = a × C_{DL}, C_{BL} = $C_{/BL}$ = 3 pF, V_{DD} = 1.8 V

5.1.4 Performance comparison

The proposed SA and other existing designs [259-261] (the high speed [259], the ultra low-power [260] and the charge-transfer [261]) have been optimized and extensively simulated using Cadence's Affirma Spectre circuit simulator based on a 0.18 µm CMOS process from Global Foundries Singapore (GFS). All four circuits were simulated using a simplified read-cycle-only two columns, two rows memory system. The standard 6T SRAM memory cells were used. The new SA's active layout area is the smallest among

the four designs (Table VIII, **Fig. 5-5**). This is due to its simple structure without the current conveyor and current mirror pairs.

TABLE VIII. COMPARISON SUMMARY OF FOUR CIRCUITS FOR C_L = 0.1 PF, C_{BL}= 1 PF, C_{DL} = 1 PF AT 0.18 µM CMOS TECHNOLOGY AND 50 MHZ FREQUENCY

	Sensing delay, ns	Average power, mW	Layout area, μm^2
Proposed (1.8 V)	0.26	0.244	376
Proposed (0.9 V)	1.6	0.019	376
Charge-transfer (1.8V)	1.01	0.597	568
Ultra low-power (1.8V)	1.28	0.526	579
High-speed (1.8 V)	0.91	0.983	659

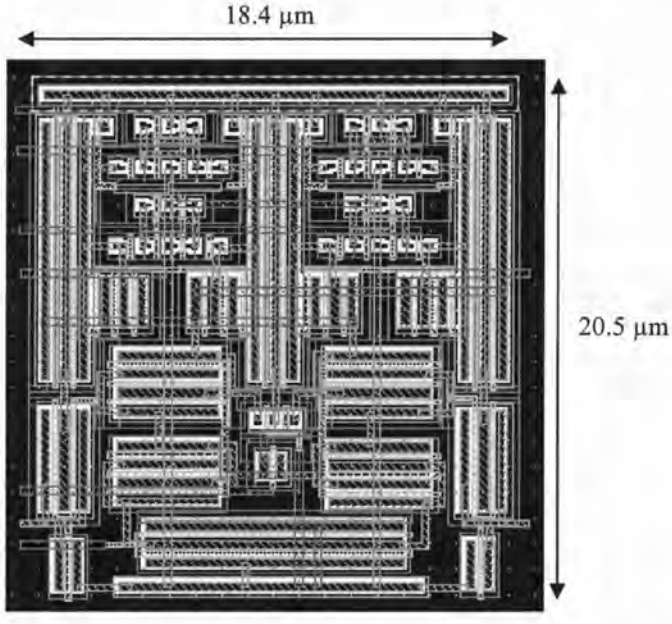


Figure 5-5 Layout of the proposed design. The layout includes four memory cells, pre-charge and equalization circuits as well as the sense amplifier. Its layout has a dimension of 20.5 µm × 18.2 µm.

All four designs were tested against C_{BL}, C_{DL} and V_{DD} variations. In order to gauge the actual behavior of the circuits, a wide range of C_{BL} and C_{DL} (from 1 pF to 5 pF) have been used to model the actual parasitic capacitances of the memory array. **Fig. 5-6** shows that only the proposed design can operate down to a V_{DD} of 0.9 V while [261], [260] and [259]

cease to work at V_{DD} equals to 1.2 V, 1.3 V and 1.3 V respectively. Furthermore, at any supply voltage, the new design outperforms the rest with the smallest sensing delay. **Figs. 5-7** and **Fig. 5-8** demonstrate the superiority of the proposed design over the other circuits at 1.8 V supply voltage against C_{BL} and C_{DL} variations. For example, at C_{BL} = 1 pF, C_{DL} = 5 pF and the load capacitor C_L = 0.1 pF, its sensing delay is reduced to 25.8%, 20.3% and 28.6% and its power consumption is decreased to 37.6%, 46.5% and 24.9% as compared to [261], [260] and [259] respectively. In addition, the new SA offers an enhanced speed robustness against the varying C_{DL}, giving a sensitivity of only 3 ps/pF, which is better than that of [261], [260] and [259] designs, being at 3.5 ps/pF, 32 ps/pF and 45 ps/pF respectively. Table VIII provides a summary of performance metrics comparisons for all of the SAs working at 1.8 V and the proposed design working at 1.8 V and 0.9 V.

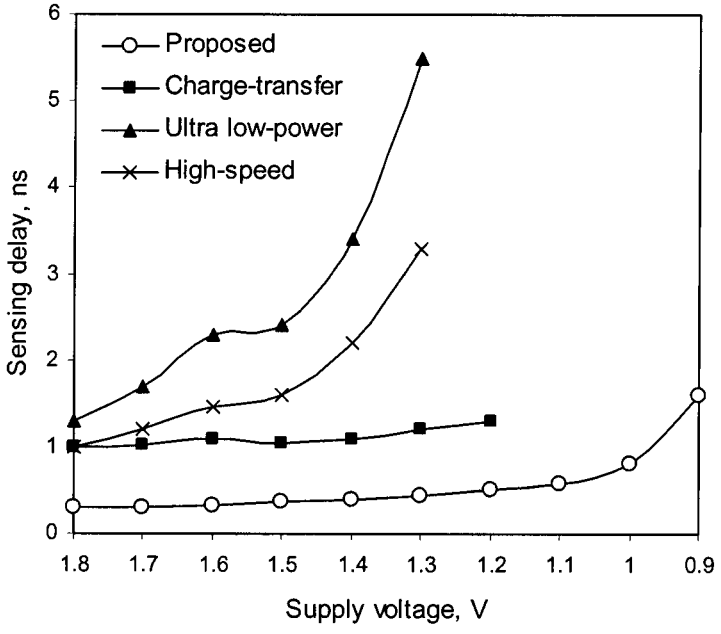


Figure 5-6 Sensing delay versus V_{DD} variation for the circuits in comparison at C_{DL} = 1 pF and C_L = 0.1 pF.

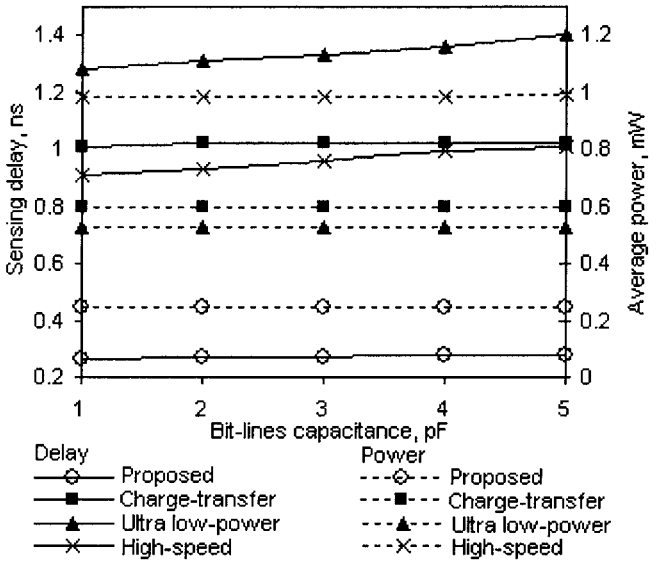


Figure 5-7 Sensing delay and average power at 50 MHz versus C_{BL} variation for the circuits in comparison at C_{DL} = 1 pF and C_L = 0.1 pF.

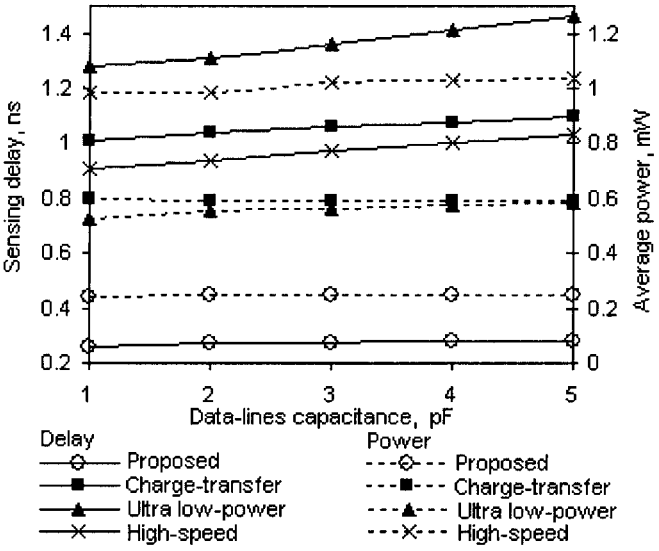
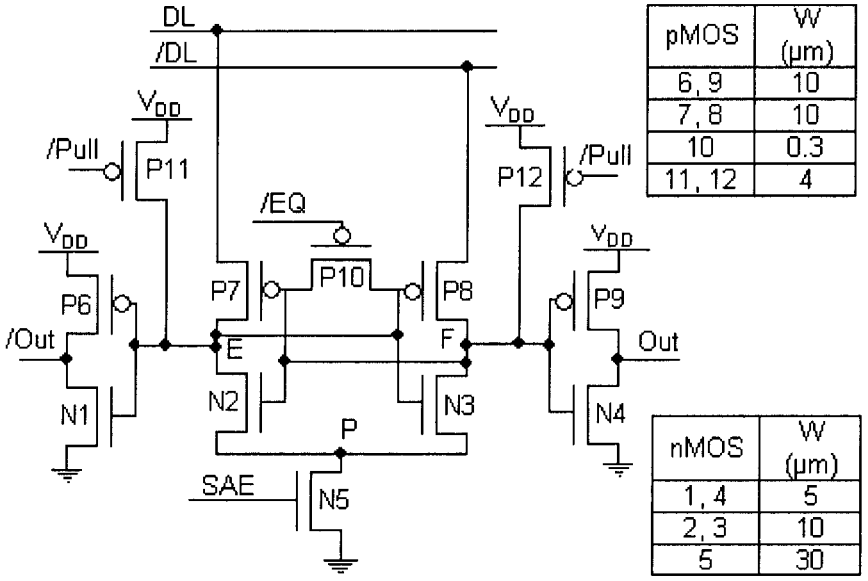


Figure 5-8 Sensing delay and average power at 50 MHz versus C_{DL} variation for the circuits in comparison at C_{BL} = 1 pF and C_L = 0.1 pF



pMOS	W (µm)
6, 9	10
7, 8	10
10	0.3
11, 12	4

nMOS	W (µm)
1, 4	5
2, 3	10
5	30

Figure 5-9 Improved version of the cross-coupled amplifier. Channel lengths of all transistors are 0.18 µm

In an attempt to reduce the supply voltage to lower than 0.9 V, we proposed a modified cross-coupled inverter which can work at a supply voltage of 0.55 V. Two additional pMOS transistors P11 and P12 were added to the conventional design as shown in **Fig. 5-9**. Furthermore, other transistors were resized so that they are strong enough to work at 0.55 V. P11 and P12 serve as a pre-charge circuit to pre-charge the two nodes E and F to the same potential, i.e. V_{DD}. As a result, a small-sized transistor P10 can be used, and thus, significantly reducing the switching time of the cross-couple inverters [262] . At 0.55 V supply voltage, its sensing delay is 4.53 ns and consumes 6.72 µW. We only optimize the SA for it to work at a very low voltage, rather than over a wide range (from 1.8 V to 0.55 V) since large transistors will consume a huge amount of power if it operates at 1.8 V. It is, however, enough to prove the superiority of the proposed design over its state-of-the-art counterparts. Transistor sizing for the improved circuit is also shown in **Fig. 5-9**.

5.1.5 Summary

A robust, high performance SA is presented. It introduces a new read scheme that creatively combines the current and voltage sensing schemes to maximize the utilization of the I_{cell} and hence, offering a much better performance, in terms of both sensing speed and power consumption. Since only one of the BLs and one of the DLs are discharged to voltage levels lower than V_{DD} while their complementary lines are kept at $V_{DD,}$ the new SA is insensitive to the difference between C_{DL} and $C_{/DL}$. This feature helps the new SA to cope with the increasing fluctuation of these parasitic capacitances due to the layout and fabrication processes. The new design can operate in a wide supply voltage range, from 1.8 V to 0.9 V with minimum performance degradation. Furthermore, a modified cross-coupled inverter is introduced, which brings down the operating voltage to 0.55 V. Although this modified version needs larger transistor sizes and only work in a small supply voltage range, both versions of the proposed SAs have conclusively proved the robustness and the suitability of the new read scheme for applications where ultra low-voltage, ultra low-power, and high-speed are of crucial design considerations.

5.2 A variation-tolerant SA using a novel cross-coupled topology

In this section, another current-mode SA is presented. It extensively utilizes the cross-coupled inverters for both local and global sensing stages, hence achieving ultra low-power and ultra high-speed properties simultaneously. Its sensing delay and power consumption are almost independent of the bit- and DL capacitances. Extensive post-layout simulations, based on an industry standard 65 nm/1 V CMOS technology, have verified that the new design outperforms other designs in comparison by at least 27% in terms of speed and 30% in terms of power consumption. Sensitivity analysis has proven that the new design offers the best reliability with the smallest standard deviation and Bit-Error-Rate (BER). Four 32×32-bit SRAM macros have been used to validate the proposed design, in comparison with three other circuit topologies. The new design can operate at a maximum frequency of 1.25 GHz at 1 V supply voltage and a minimum supply voltage of 0.2 V. These attributes of the proposed circuit make it a wise choice for the contemporary high-complexity systems where reliability and power consumption are of major concerns.

5.2.1. Existing designs

This section briefly describes the operations of three existing designs studied in comparison with the proposed work. The schematic of these designs are depicted in **Fig. 5-10.**

5.2.1.1 Current-conveyor- based SA

The first conveyor-based SA was proposed by E. Seevinck *et. at.* in [258]. It consists of four identical pMOS transistors (P1-P4 in **Fig. 5-10(a)**) connected in a feedback structure. It is assumed that the complementary BLs (BL and /BL) are pre-charged to V_{DD} and all four pMOS transistors operate in saturation region during the read cycles. The current conveyor is enabled by triggering the CS signal low. Since all four transistors are in saturation, their source-to-drain currents are only dependent on their gate-to-source

voltages. As a result, voltages at the BL terminals (V_{BL} and $V_{/BL}$) are the same and equal to (v1 + v2). The current conveyor therefore has the ability to convey the differential current from the BLs to the DL without waiting for the discharging of the highly capacitive BLs. Thus, this design achieved both higher sensing speed and lower power consumption when compared to the conventional voltage mode designs in which large voltage differences must be developed between the BLs [258]. Based on this basis structure, several improved versions of this design have been reported, mainly by adding current-mirrors to the feet of the current-conveyor to enhance its current drive-ability [98, 263]. In this paper, we will compare our work with the high-speed design [263] which consists of four additional nMOS transistors, also shown in **Fig. 5-10(a)**. These nMOS devices form two current-mirrors to intensify the output currents I_1 and I_2 to the DLs. This design will be used as the benchmark to evaluate the performance of the proposed design, the alpha latch and the decoupled latch SAs mentioned below. However, because of its current-mode nature, we do not study its input-offset voltage. As a result, input-off set analysis (**Fig. 5-16**) and latching delay analysis (**Fig. 5-14**) are not applicable to this design.

5.2.1.2 Alpha-latch SA

The alpha latch [264] is depicted in **Fig. 5-10(b)**. The nMOS transistor N5 is used to turn the amplifier off during standby, thus saving power. When the SA is activated by the Enable signal (EN), the differential input from the complementary BLs induces a differential trans-conductance in N3 and N4. As a result, voltage and current differences will appear at the drains of N3 and N4, i.e. the sources of N1 and N2. Since the CS signal turns off N6, the flip-flop structure will latch and full swing voltages will be available at nodes A and B, turning one of the transistors N7 and N8 on while the other is off. During standby, /EN is kept high to turn P3 and P4 off. During operation, both P3 and P4 are turned on but one of N7 and N8 is turned off, thus only one current will flow to the DLs (i.e. I_1 or I_2 in **Fig. 5-10(b)**). A global SA is also used to quickly amplify the voltage difference on the DLs to the output of the SRAM.

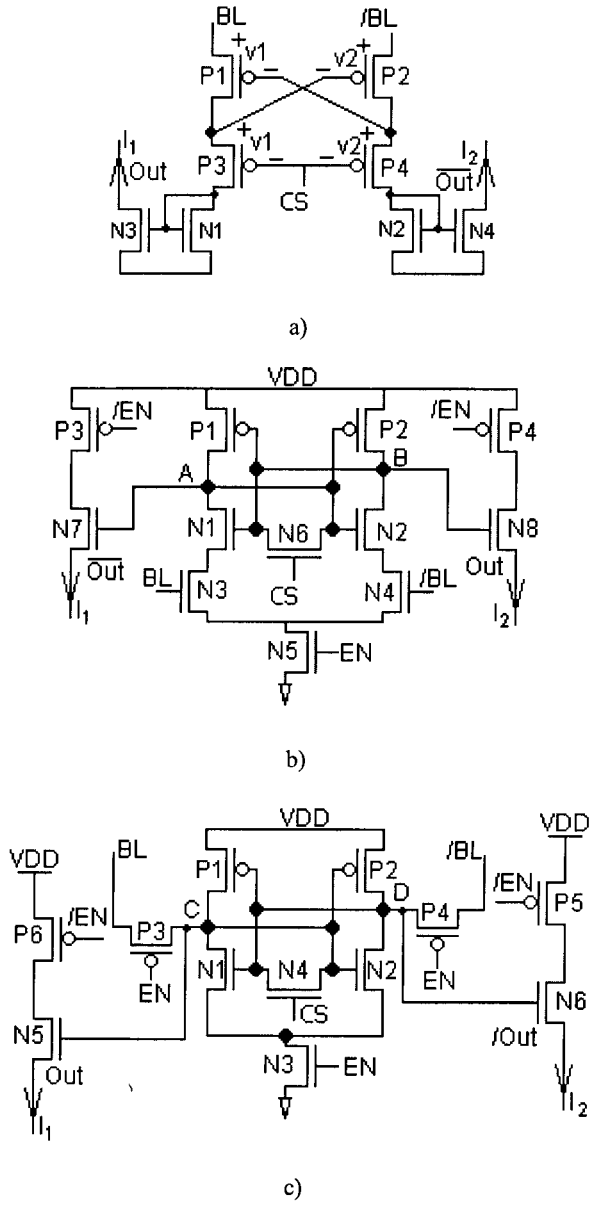


Figure 5-10 Local sensing stage of existing SRAM SA. a) current-conveyor b) alpha latch c) BL decoupled latch.

5.2.1.3 Decoupled latch SA

The decoupled-latch consists of six nMOS and two pMOS transistors, as shown in **Fig. 5-10(c)**. Similar to the alpha-latch, its N3 is used to save power. The reason we use a tailed nMOS device in **Fig. 5-10(b)** and **Fig. 5-10(c)** is because it gives a smaller area compared to a pMOS with the same current strength. Furthermore, the BLs are pre-charged to V_{DD} and hence nMOS tail device is required. To tackle the heavily loaded BLs issue, these BL signals are tapped to the input ports of the amplifier through two

decoupled devices, i.e. P3 and P4. Once the BL differential signal is induced at nodes C and D, the latch is enabled by turning off N4 but turning on N3. Concurrently, P3 and P4 are turned off to decouple the BLs from the high-swing output nodes. The use of P3 and P4 helps in reducing the impact of the BL capacitances on the switching activity, hence significantly reducing both sensing delay and power consumption [265-266]. Similar to the alpha latch design, full swing voltages at nodes C and D are transferred to the DL differential voltage by the means of a pair of nMOS transistors, as shown in **Fig. 5-10(c).**

5.2.2. Operating principle of the proposed SA

The proposed SA, coupled with a simplified read-cycle-only memory system, is presented in **Fig. 5-11**. It consists of two sensing stages: local and global. The local sensing stage is formed by four pMOS (P3-P6) and three nMOS (N1, N2 and N7) transistors. While P3 and P4 act as a column switch, the rest of the transistors establish the local cross-coupled inverters, which are responsible for generating the BL differential currents and transferring them to the DLs. The global sensing stage consists of three pMOS (P7-P9) and five nMOS (N3-N6 and N8) transistors. In **Fig. 5-11**, two output inverters, which serve as buffers to drive the potentially large output loads to full CMOS logic output levels, are also included. The operation of the proposed SA is described as follows:

During the standby period, P3 and P4 are turned off to block any BL currents. The Column Select and Global Enable (CS and GEN) signals turn on N7 and N8 respectively to equalize nodes A, B and C, D to the same potential, respectively. Meanwhile, two pre-charge transistors N5 and N6 are turned on to pull both DLs to ground. At the same time, P9 is turned off to save power. Since P9 is off and the DLs are pre-charged to ground, C and D are also at a low potential (near V_{th}) during standby. The two output inverters are also cut-off by P9, as shown in **Fig. 5-11**. This topology ensures that the standby current of the circuit, and thus the power dissipation, are minimized.

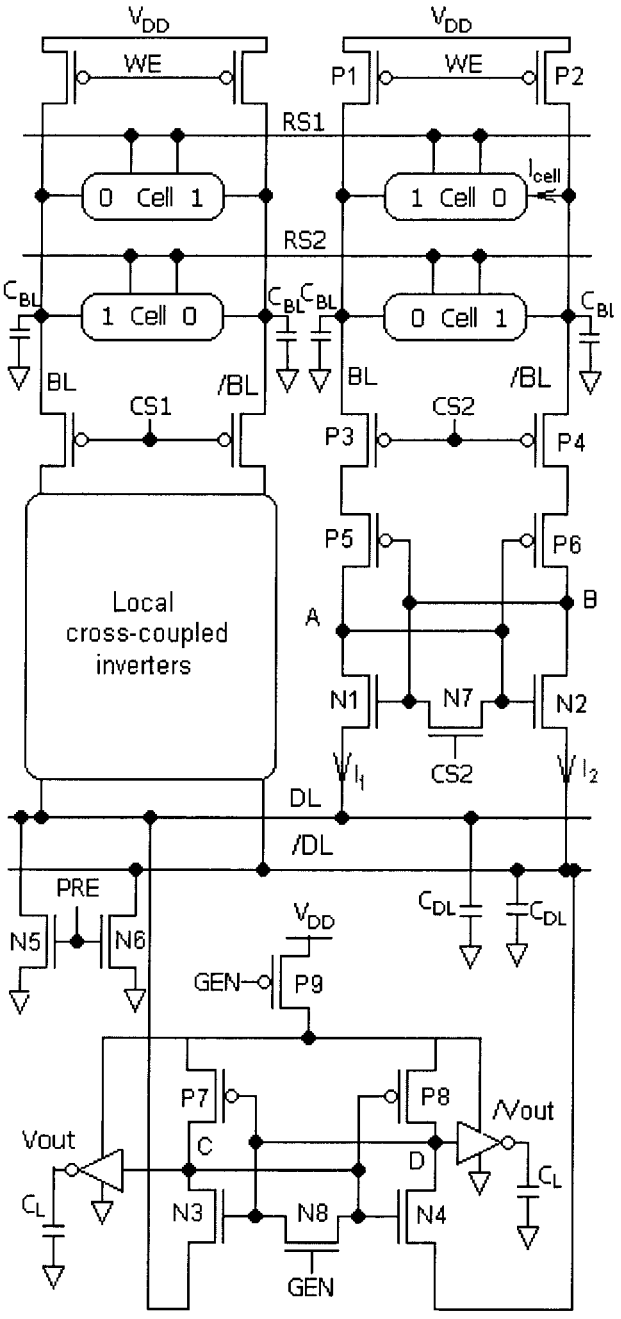


Figure 5-11 The proposed design coupled with a simplified read-cycle-only memory system.

Consider both RS1 and CS2 being activated during a read operation. The pre-charge signal (PRE) turns N5 and N6 off, allowing the DL voltages to change freely. The memory cell at the upper row and right column will be selected, resulting in a small cell current I_{cell} flowing from the /BL into the cell as shown in **Fig. 5-11** and discharges the /BL to a voltage level lower than that of the BL. As CS2 is triggered low, P3 and P4 are turned on to transfer the BL potentials and BL currents to the inputs of the local cross-coupled

inverters. At the same time, N7 is turned off to activate the local cross-coupled inverters. This building block senses the voltage and current difference at the source terminals of P5 and P6 and quickly finishes its latching process. Hence, node A is pulled to V_{DD} while node B is discharged to the same potential of the /DL, which is near ground, as shown in **Fig. 5-12** [265]. More importantly, during this latching process, the pulsing current flowing from N2 to /DL, i.e. I_2, is much higher than that from the N1 to the DL, i.e. I_1, as shown in **Fig. 5-12**. This phenomenon can be intuitively explained as follows: During standby, nodes A and B reside at a low potential near V_{th}. Once the SA is activated, both node potentials will slightly rise and then quickly start to deviate. For example, in **Fig. 5-12**, node A approaches near V_{DD} while node B plunges to near ground. Thus, transistor N1 is in cut-off most of the time. On the other hand, transistor N2 operates in the triode region and then moves to the saturation region, resulting in a much larger pulsing current when compared to that of N1. Integrating these two currents over time will yield the total charges flowing to DL and /DL, respectively.

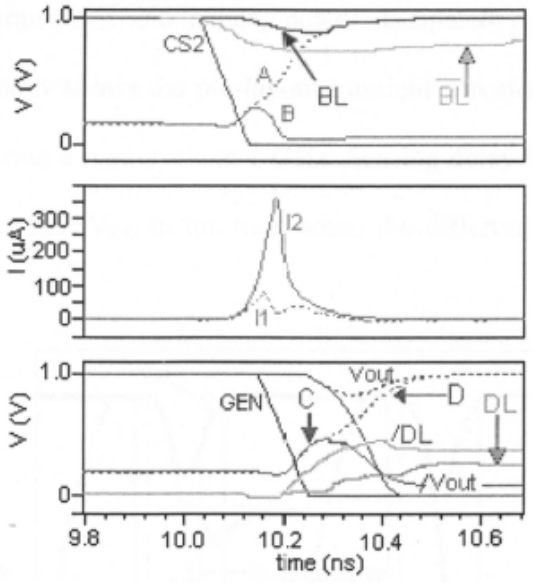


Figure 5-12 Waveforms at several nodes of the proposed SA during a read cycle.

These differential currents flow to the DLs and induce a voltage difference on the global DLs. Similarly, this voltage difference is amplified by the global sensing stage to the intermediate outputs V_C and V_D, also shown in **Fig. 5-12**. These two voltages are then fed

to the output buffers to get the full CMOS logic levels. It is worth mentioning that the global sensing stage can only be activated after the latching process of the local amplifier has completed. The waveforms of several nodes of the proposed SA during a read cycle are also shown in **Fig. 5-12**. This hierarchical two-level sensing scheme helps in reducing both power consumption and sensing delay imposed by the BLs and the DLs on high density SRAM designs. Furthermore, although nodes A and B have a near-full swing during a read operation, they are not to be tapped directly to the DLs. Otherwise, the total power consumption and sensing delay will be increased dramatically. As a result, a global sensing stage is required to amplify the small differential signal on the DLs to a full CMOS logic level at the output of the SRAM.

The total active power dissipated in the proposed SA is limited by the cell current flowing from one of the BLs to the node of the cell where a '0' is stored (which solely depends on the cell design) and the switching currents of the sensing stages. After latching, the cross-coupled configuration stays at one of its bi-stable stages and no additional current is consumed and hence, power dissipated on the BLs and DLs is optimized. **Fig. 5-13** below shows the pre-layout transient waveforms of several nodes of the proposed design during a read cycle at 1 GHz. Sensing delay is defined from the time when CS signal reaches half-V_{DD} to the time when the differential output reaches half-V_{DD}.

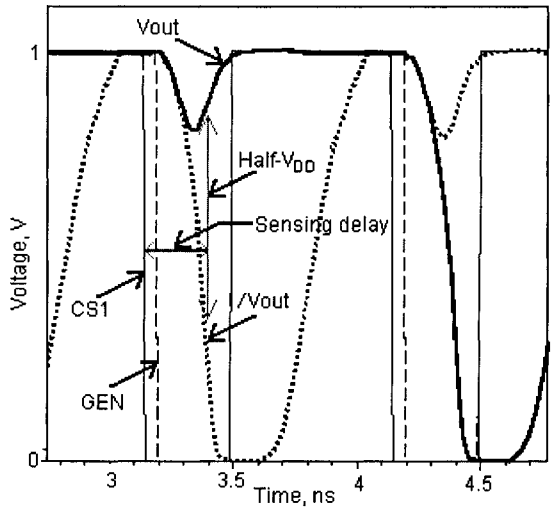


Figure 5-13 Output waveforms at 1 GHz.

Since the global DLs are shared among many columns, their parasitic capacitances are significant and have an important impact on the input margin of the global sensing stage. The voltage difference on the DLs must be larger than the input offset voltage of the global SA in order to perform a correct readout. Thus, the number of columns sharing the DLs should be considered carefully to maintain a reasonable input margin. It is determined by the size of the MOS transistors in the local SA (i.e. N1-N2 and P5-P6) and the layout dimension of the memory cell (as it affects the length of the DLs and hence their parasitic capacitances). This number does not depend on the technology as it can be adjusted by changing the size of the transistor in the local SA. Our analysis indicated that in order to maintain an input of at least 100 mV to the global SA at 1V voltage supply and 1.25 GHz operating frequency (as will be mentioned in Section VI. C), the number of columns sharing the DLs must not exceed 164.

5.2.3. Simulation and design methodology

5.2.1.1 Test structure

All the SAs in comparison, i.e. references [263], [264-266] and the proposed circuit have been extensively simulated using four identical 32×32-bit SRAM cores. Each column of the core has one local SA which transfers the signal to the DLs for global sensing. The order in which the memory cells are activated is identical for all four designs. Furthermore, lump-sum C_{BL} and C_{DL} are connected to the BLs and DLs to model additional parasitic capacitance in bigger SRAM macros. As a simple approximation, each row contributes 1 fF to the BL capacitance and each column contributes 1 fF to the DL. It means that if C_{BL} = 100 fF and C_{DL} = 150 fF, our structure is equivalent to a SRAM macro of 132 rows and 182 columns. This facilitates the need to vary both C_{BL} and C_{DL} for investigation. It also reduces the simulation time with reasonable accuracy. Detailed investigations for various C_{BL} and C_{DL} parasitic conditions and supply voltage V_{DD} have

also been performed to gauge the robustness of the designs. C_{DL} and C_{BL} are swept from 100 fF to 200 fF simultaneously while V_{DD} is swept from 0.2 V to 1 V. Besides the sensing delay and the average power consumption, Power-Delay Product (PDP) is used as the main performance indicator which takes both entities into consideration. The transistor sizes of different designs of SAs have also been fully optimized to achieve the minimum PDP.

5.2.1.2 Circuit optimization

All transistors in the read-out circuits of the four designs have a constant channel length of 65 nm and parameterized channel widths. Each circuit is then optimized using a systematic parameter sweeping methodology. To ensure the fairness of the comparison, transistor widths are set to obtain the minimum PDP at 1V supply voltage and C_{DL}= C_{BL} = 100 fF. Parasitic capacitances are extracted and back-annotated from the layout view to the schematic view to perform post-layout simulations. All results presented in **Fig. 5-14** to **Fig. 5-23** are based on post-layout simulation results.

5.2.1.3 Speed deviations

In digital and memory circuits, time matching is vital since it ensures that sufficient input voltage is available to be amplified. If the output signal of one stage is slowed down, the input of the next stage may be smaller than the input-offset voltage, resulting in an incorrect sensing. This issue is even more critical in highly compact SRAM macros, due to their heavily loaded bit- and DLs, which are likely to cause signal mismatches. Therefore, each sensing stage should have a very stable sensing delay to minimize the above-mentioned mismatches. Thus, speed deviations due to inter-die variations of the circuits in comparisons must be evaluated. These are done with the SA alone as well as in the context of 32×32-bit SRAM macro. Monte Carlo simulations are performed with inter-die variations to monitor the stability of the circuits and simulation results are presented in **Fig. 5-14** and **Fig. 5-15**. All circuits are simulated at a power supply of 1V, C_{DL} = 100 fF, C_{BL} = 100 fF, C_L = 20 fF and clock frequency of 250 MHz. The latching delay is defined as the interval from 0.5 V_{DD} of the enable signal of the SA to the time when the differential

output of the SA is 0.5 V_{DD}. The total sensing delay is measured from 0.5 V_{DD} of the CS signal to the time when the final differential output of the SRAM reaches 0.5 V_{DD}, as illustrated in **Fig. 5-13**.

5.2.1.4 BER consideration

In this work, we investigated the input-offset quality of the SA designs. Therefore, our BER investigations are only performed on the SAs alone. In this work, BER refers to the failure rate of the SA at some specific conditions, and not the memory cell. Since the input offset voltage is the main cause of read failure and is more critical to the cross-coupled based SAs, only three designs are investigated, namely the proposed, the decoupled latch and the alpha latch designs. The BL voltage is set to V_{DD}. The input voltage is defined as the difference between BL and /BL. All simulations are performed using Monte Carlo simulations, taking both process variations (inter-die) and device mismatches (intra-die) into considerations. Device variations are from foundry-given data with all parameters considered simultaneously (i.e. doping level, V_{th}, W, etc). Number of iterations is 35000. Simulation results are shown in **Fig. 5-16**.

5.2.1.5 Maximum operating frequencies at various supply voltages

As the supply voltage scales down, the maximum operating frequency of the SRAM also drops. For each supply voltage from 1V down to 0.2 V, we consider the maximum frequency at which the SAs are able to work correctly. Performance comparisons are also carried by monitoring the sensing delay and power consumption per MHz.

5.2.4. Sensitivity analysis

5.2.4.1 Process variations

As CMOS technology scales down, process variations are becoming predominant concerns in designing VLSI systems, especially in SRAM where device geometries are particularly small. It is therefore critical for an SA to work properly not only under power supply fluctuations but also process variations.

In this work, a detailed sensitivity analysis has been carried out to investigate the operation of the four designs using the process data from the foundry. While the latching delay analysis is only performed on the three cross-coupled based SAs, the total sensing delay analysis is carried out on all four designs, with the current-conveyor based high-speed SA used as a reference circuit.

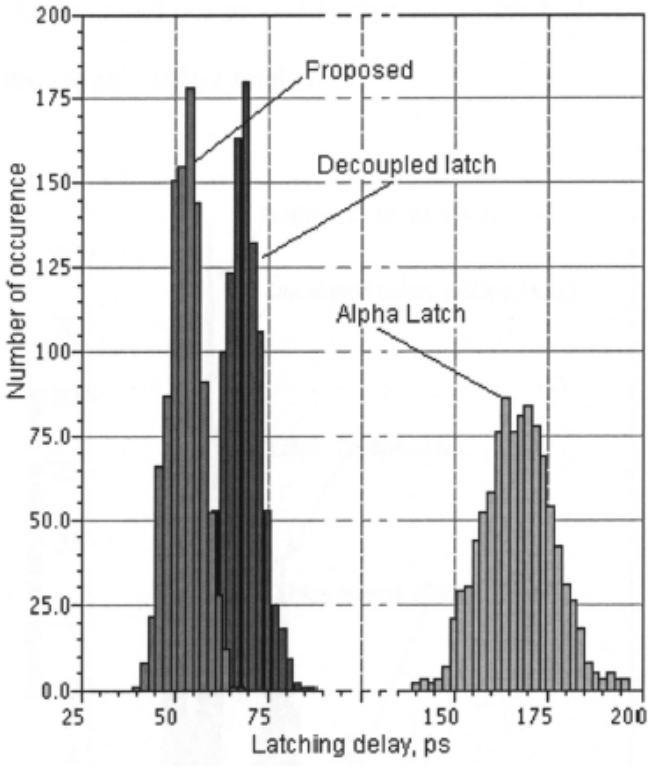


Figure 5-14 Latching delay distributions of the three designs using Monte Carlo simulation at room temperature, 1V supply voltage, 100 mV differential input. Number of iteration is 1000.

Fig. 5-14 shows the latching delay distribution of the proposed, the decoupled latch and the alpha latch designs. It is evident that the proposed design offers the best latching delay with the smallest mean value and a standard variation similar to that of the decoupled latch. This can be explained as the proposed design has the smallest capacitive load at the switching nodes (nodes A and B in **Fig. 5-11**) compared to those of the alpha latch (nodes A and B in **Fig. 5-10 (b)**) and the decoupled-latch (nodes C, D in **Fig. 5-10 (c)**). Furthermore, it contains the least number of transistors and hence, its variation is the smallest.

Fig. 5-15 illustrates the total sensing delay distribution of the three above-mentioned circuits with the high-speed design added as a reference. It is in accordance with the data shown in **Fig. 5-14** where the proposed and the decoupled designs offer the best performance. It is evident that all three cross-coupled based SAs are more reliable than the other designs with much smaller mean values and standard deviations, also shown in **Fig. 5-15**. For example, the proposed design is 3.6x faster than the high-speed design and its delay standard deviation is almost 10x smaller.

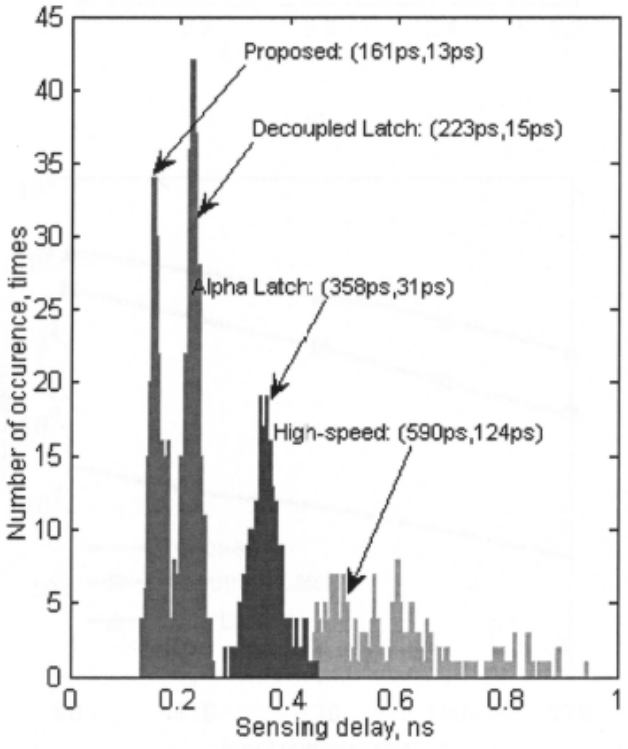
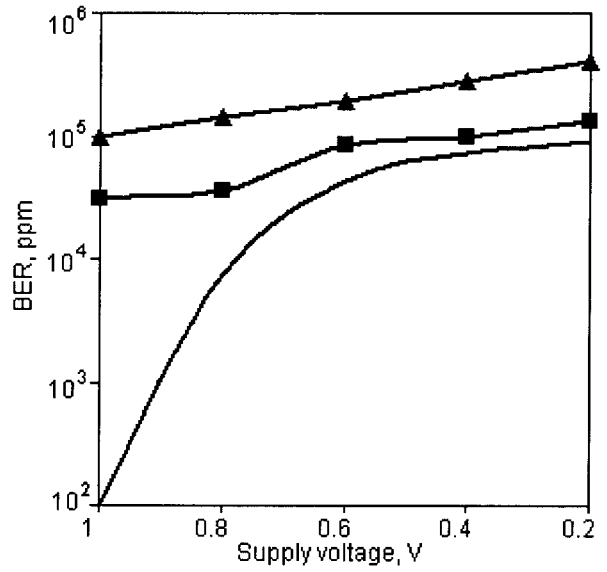


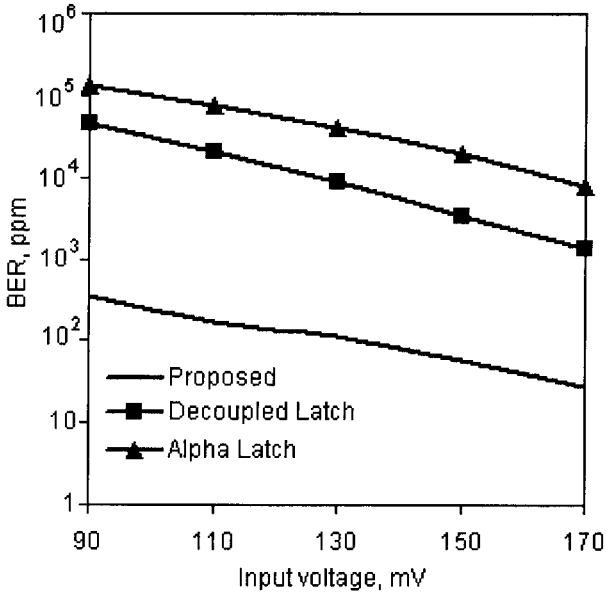
Figure 5-15 Total sensing delay distributions of the designs in comparison using Monte Carlo simulations at room temperature. Number of iterations is 200. The numbers in the brackets explain the mean and standard deviation in sensing delay of each design.

5.2.4.2 Device mismatches

Device mismatches refer to intra-die variations, which are caused by the local random variations during fabrication. In the sensing circuit, this issue is more critical than inter-die variations as it is the main cause of the input offset voltage which in turn leads to a wrong sensing if the input swing is smaller than the required offset value.



a)



b)

Figure 5-16 BER of the three cross-coupled based SA using Monte Carlo simulations with 35000 iterations. a) BER versus supply voltage, input voltage equals to 0.1 V_{DD}. b) BER versus input voltage at V_{DD} = 1V.

Fig. 5-16(a) and **Fig. 5-16(b)** show the BER of the three cross-coupled-based designs caused by the device mismatches in various supply and input conditions, respectively. Both figures show that the proposed circuit has a smaller BER at every condition. For example, at 1V voltage supply and 110 mV input, BER of the proposed, decoupled latch and alpha latch designs are 171, 20171 and 75532 part-per-million (ppm), respectively.

This is because the proposed design has the least transistor count (4 versus 6). Although the BER of the proposed design increases drastically when the supply voltage scales down (**Fig. 5-16(a)**), it is still smaller than the other two designs. Furthermore, this trend saturates when V_{DD} approaches 0.5V and still ensures a better performance than its counterparts down to 0.2 V supply voltage.

In contrast of **Fig. 5-16(a)**, **Fig. 5-16(b)** presents three parallel lines which indicate a predictable behavior of all three designs when the input voltage changes. At 1 V supply voltage, the BER of the proposed design is at least 50x better than the other designs. As the proposed design suffers less from the process variations (**Fig. 5-14, 5-15, 5-16**), it scales better with technologies. Therefore, it is reasonable to conclude that the proposed design is more reliable than the other latch-based topologies and hence more suitable for applications where reliability is of crucial concern.

5.2.5. Performance comparison

5.2.5.1 Power consumption and sensing delay

Performance indicators (sensing delay, power consumption and PDP) of the above-mentioned circuits are graphically presented in **Fig. 5-17** to **Fig. 5-19**. **Fig. 5-17** compares the sensing delay of the four designs with respect to C_{BL} and C_{DL} variations, respectively. It is apparent that all four designs are insensitive to both C_{BL} and C_{DL}, manifested by the almost-horizontal surfaces. This is because all switching nodes are isolated from the highly loaded BLs and DLs. However, DL capacitance has a greater impact on the performance of the circuits with a higher slope along the DL capacitance axis. This figure also demonstrates the superiority of the proposed design over the other circuits at 1 V supply voltage against C_{BL} and C_{DL} variations, respectively. For example, at C_{BL} = 100 fF, C_{DL} = 100 fF and C_L = 20 fF, its sensing delay is reduced to 21.3%, 72.8% and 27.6% of that of the of high-speed [263], decoupled latch [265] and alpha latch [264] designs, respectively. This observation is consistent over a wide range of parasitic conditions, also shown in **Fig. 5-17**.

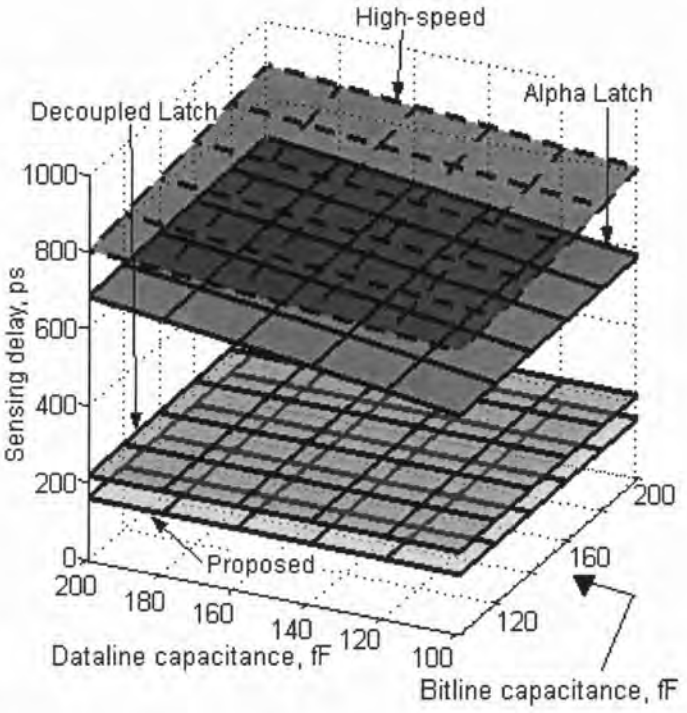


Figure 5-17 Sensing delay versus C_{BL} and C_{DL} variations for the circuits in comparison C_L = 20 fF.

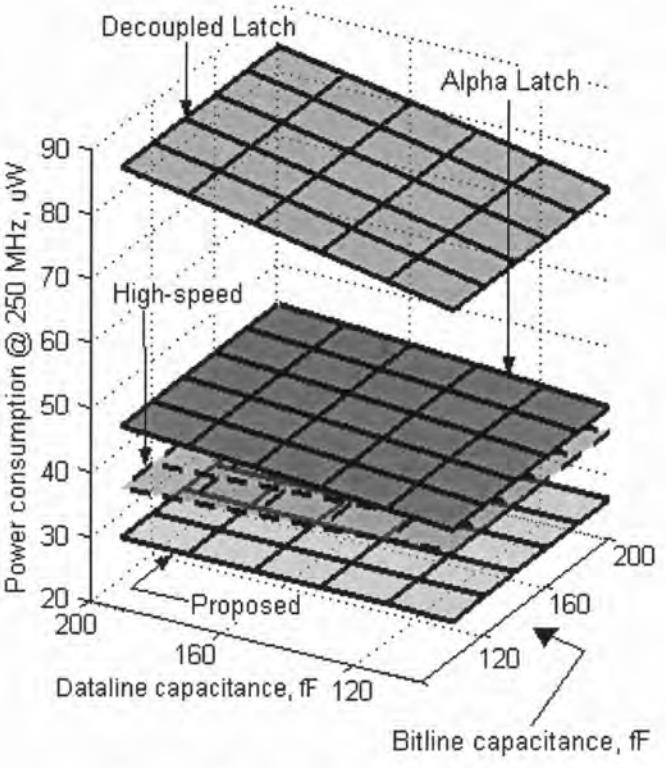


Figure 5-18 Power versus C_{BL} and C_{DL} variations for the circuits in comparison C_L = 20 fF.

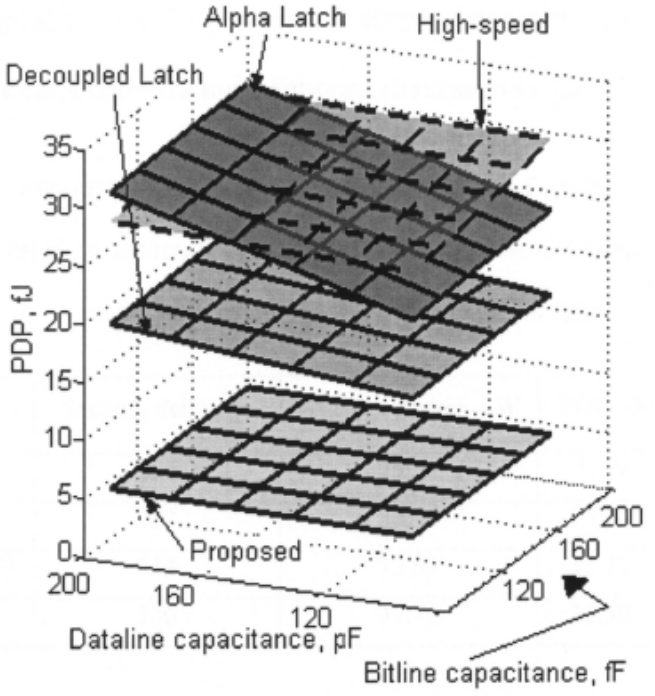


Figure 5-19 PDP versus C_{BL} and C_{DL} variations for the circuits in comparison C_L = 20 fF.

A similar observation can be seen in **Fig. 5-18**, regarding the power consumptions of the four circuits. For example, at the same working conditions as above (i.e. at V_{DD} = 1V, C_{BL} = 100 fF, C_{DL} = 100 fF and C_L = 20 fF), the power consumption of the new design is reduced to 70.2%, 34.7% and 64.3% of that of the high-speed [263], decoupled latch [265] and alpha latch [264] designs, respectively. This is because the output of the local sensing stage in our design has very low voltage swing and thus can be tapped directly to the DLs. Furthermore, after latching, no BL current flows from the BLs to the DLs. This is in contrast with the other designs in which at least one BL current flows from the BLs to the DLs. Thus, the PDP of the proposed design is more than 74% superior as compared to other designs, as shown in **Fig. 5-19**. In addition, the proposed circuit achieves the most stable behavior with a total change across the simulated regions (i.e. C_{DL} ranges from 100 fF to 200 fF and C_{BL} ranges from 100 fF to 200 fF) of 6.5% whereas that of high-speed [263], decoupled latch [265] and alpha latch [264] designs are 10.9%, 17.3% and 34.2%, respectively. Table IX summarizes the comparison of these four designs, including the layout area of each topology. As shown in Table IX and **Fig. 5-20**, the proposed local design occupies the smallest active area, which is only 79%, 67% and 64% of that of the

high-speed, decoupled latch and alpha latch designs, respectively. All transistor sizes are obtained from the circuit optimization mentioned in *subsection 5.2.3.2.*

TABLE IX. COMPARISON SUMMARY OF THREE CIRCUITS FOR C_L = 20 FF, C_{BL}= 100 FF, C_{DL} = 100 FF AT 65 NM CMOS TECHNOLOGY AND 250 MHZ FREQUENCY. ALL DESIGNS HAVE THE SAME LAYOUT WIDTH OF 1.6 μM TO FIT ONE COLUMN PITCH

	Sensing delay, ps	Average power, μW	PDP, fJ	Layout area, μm^2
Proposed	156	25.58	3..99	8.64
High-speed [263]	732	36.40	26.64	10.88
Decoupled latch [265]	214	73.69	15.77	12.80
Alpha latch [264]	566	39.76	22.50	13.44

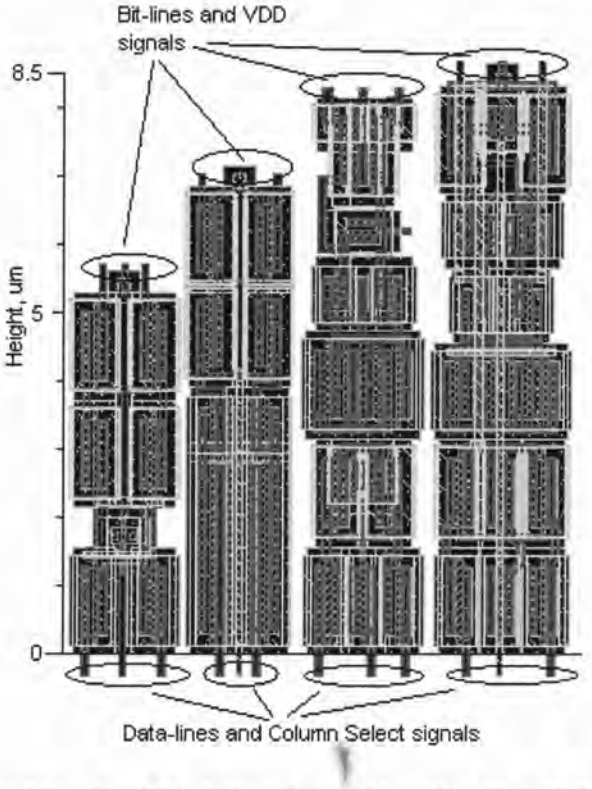


Figure 5-20 Layout of four local SA designs in consideration. From left to right: Proposed, high-speed, decoupled latch and alpha latch. V_{SS} signal runs horizontally and is not shown in this figure.

5.2.5.2 Leakage

Leakage currents of the four SAs are investigated at various operating temperatures using DC analysis. All four SAs (**Fig. 5-10** and **Fig. 5-11**) are turned off by setting their

control signals to either V_{DD} or 0 V. At the same time, V_{DD} is kept at 1 V and the temperature is swept from 0 ^{0}C to 125 ^{0}C, to adhere to the commercial standard range. Simulation results are shown in **Fig. 5-21**. As the proposed local design has only seven transistors cascaded into two branches (**Fig. 5-11**), it has the smallest leakage current, as illustrated by the black curves in **Fig. 5-21**. For example, at room temperature, leakage currents of the proposed, decoupled latch, alpha latch and high speed designs are 9 nA, 19 nA, 17 nA and 18 nA, respectively. Similarly, the proposed global SA also offers the least leakage although the difference between the four designs is not significant. The reason is because all four designs contain two pairs of output buffers which contribute a large portion to their total leakage. These two observations confirm that the proposed design consumes the least standby power and hence enabling a longer battery life for the system.

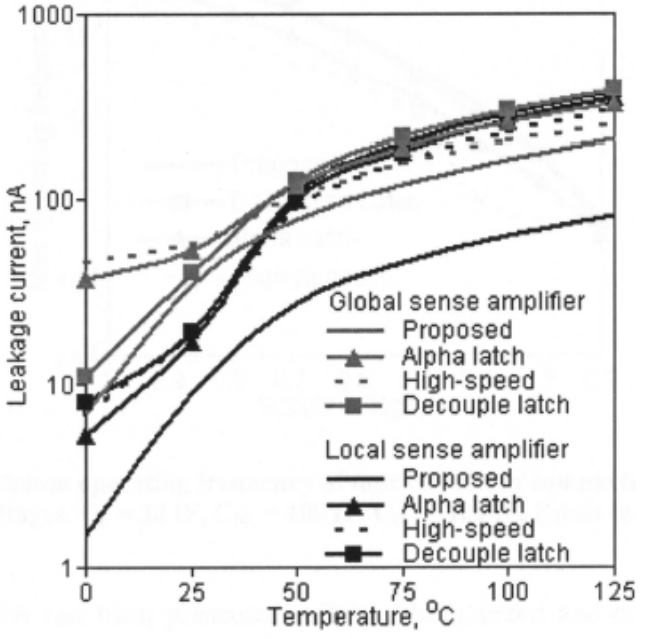


Figure 5-21 Leakage currents of the global and local SAs of four designs versus operating temperature.

5.2.5.3 Operating frequency

We aim to design a new SA that can work with a clock frequency higher than 1 GHz. Furthermore, we also study the maximum frequency of each design at several supply voltages, as shown in **Fig. 5-22**. It is noticeable that the high-speed design ceases to work

at a supply voltage of 0.3 V. As shown in **Fig. 5-22**, the proposed design and the decoupled-latch have similar maximum operating frequency at every supply voltage and are about 2x and 4x higher than those of the alpha latch and the high-speed circuits, respectively. This agrees with the data presented in **Fig. 5-23**, as the proposed design and the decoupled latch have similar sensing delay. However, the power consumption per MHz of the proposed design is smaller than that of the decoupled latch, which is even higher than that of the alpha latch. **Fig. 5-23** also clearly indicates that the current-conveyor-based high-speed SA has the largest sensing delay as well as power consumption. This conclusively proves the superiority of the proposed circuit when both stability and performance are of critical design specifications.

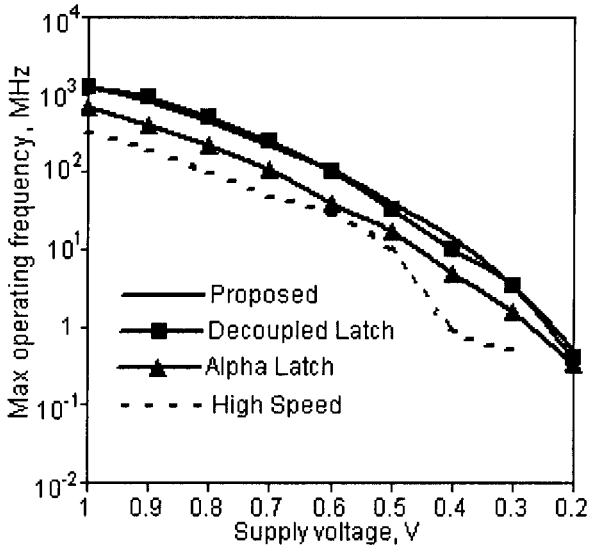


Figure 5-22 Maximum operating frequency of four circuits in comparison at different the supply voltages. C_L = 20 fF, C_{BL} = 100 fF, C_{DL} = 100 fF. Room temperature.

A latch-type SA has been presented, offering both speed and power improvements when compared to the existing circuit topologies. Furthermore, it can operate with clock frequency as high as 1.25 GHz, which is the highest among the circuits in consideration. The sensitivity analysis carried out across process corners has reaffirmed that the new design can tolerate excessive process variations with smallest performance fluctuations. It also provides better reliability with at least 50x BER at 1V supply voltage. In view of the

above, it can be concluded that the new SA is best suited for applications where low-voltage, low-power, high-speed and stability are of crucial design considerations.

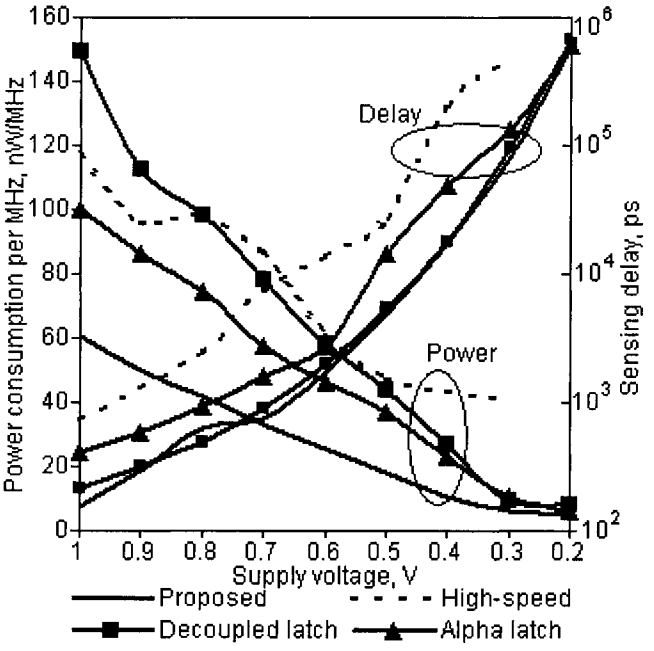


Figure 5-23 Power consumption and sensing delay of four circuits in comparison at different supply voltages. C_L = 20 fF, C_{BL} = 100 fF, C_{DL} = 100 fF. Room temperature.

5.3 Conclusion

In this chapter, the design, operation and analysis of two newly proposed circuits have been introduced. These designs achieve both high-speed and low-power and hence are suitable for low-power SRAM applications. Sensitivity analysis on these two designs also shows that they are highly reliable and can work at very low-supply voltage conditions. These designs hence can be incorporated into the SRAM macros introduced in **Chapter 3** and **Chapter 4** to further enhance the performance of the proposed SRAM cell designs.

CHAPTER 6 OFFSET ANALYSIS OF A LATCH-TYPE SA

6.1 Background

In a perfectly symmetric design, an infinitesimal initial differential input voltage V_{21} is sufficient to trigger a correct sensing of a latch-type SA as shown in **Fig. 6-1** [267-269]. However, with increasingly significant device mismatches due to the process variations, a larger initial input is required [269]. The minimum differential input voltage that is able to ignite a correct sensing is referred to as the input-offset voltage [267-269]. This implies that, during the read operation of an SRAM, the accessed memory cell must develop a differential input voltage larger than this offset voltage to ensure a correct readout [270]. Naturally, circuit designers must set some input margin by extending the delay time before enabling the SA or strengthening the bit-cell current. As a result, larger input voltage is required at the expense of additional sensing delay or power consumption on the BLs. Therefore, predicting an accurate offset value is essential not only to improve sensing speed and power consumption, but also to increase the yield of the memory [1, 19].

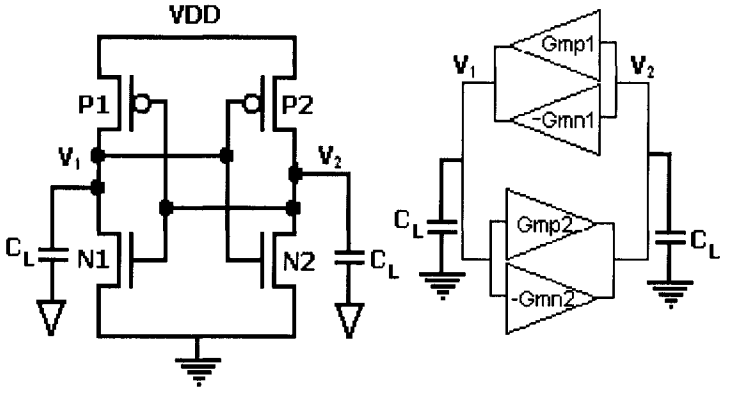


Figure 6-1 The latch-type SA. (a) Schematic (b) Small signal model

Several works in the literature have been dedicated to investigating the offset behavior of the cross-coupled inverters (henceforth, the two terms cross-coupled and latch-type will be used interchangeably) [267-269, 271-278]. Most of them only dealt with the threshold voltage mismatch, which is claimed to be the most influential factor of the input-offset

[273, 277]. Nonetheless, this approach is no longer suitable in the deep sub-micron or even nanometer technologies where other parameter mismatches are becoming increasingly significant [275, 279]. In other words, the effect of the threshold voltage, the trans-conductance and the parasitic capacitances must be taken into consideration simultaneously.

Sarpeshkar [268] has successfully analyzed the mismatch sensitivity of the latch-type SA and obtained a closed-form expression. Unfortunately, his method requires a very complicated algebraic computation and thus can hardly be generalized to other types of mismatches or be used to evaluate the impact of second-order components. Other than that, Nikoozadeh *et. al.* has recently detailed the effect of load capacitance mismatches on the input-offset [275]. However, the authors have failed to include the impact of both V_{th} and K mismatches.

Another method to evaluate the offset of the SA is by applying lump-sum voltage noise sources to the internal nodes of the cross-coupled structure [271]. However, it provides no insight into the causes of the input-offset voltage, hence giving no suggestion on how to reduce the offset. This method has also been used to determine the SNM of the 6T SRAM cell, which has a similar structure to the SA of interest. Several other works have modeled the SNM of the SRAM cell under the impact of intrinsic device fluctuations [128-129, 249, 254, 270]. Unfortunately, these approaches are not suitable for analyzing the offset of the SA. This is because SNM is a static property [280] and not directly related to input-offset voltage of the cross-coupled structure which is dynamic and changes with different input voltages.

In this chapter, we introduce a systematic and general method in deriving the input-offset voltage of the SA configuration as shown in **Fig. 6-1**. A similar approach to our model has been reported in [281] with a different SA configuration. However, it does not include the effect of the initial common-mode input voltage which has a significant impact

on the total offset. Furthermore, this method may require several iterations to obtain an accurate result [281].

Our work contributed in the following aspects:

1. A simple criterion is proposed to evaluate the offset voltage of the latch-type SA.

2. Analytical work is presented to provide a criterion to evaluate input-offset model of the SA due to threshold voltage, trans-conductance, and capacitance mismatches simultaneously. HSPICE simulations have been carried out to verify the model.

3. Weakness of the proposed method is discussed with suggested solutions.

6.2 Meta-stable states and criterion for correct sensing

6.2.1 Meta-stable states

We assume that the cross-coupled SA operates in the super-threshold region, i.e. $V_{DD} > |V_{tp}|+V_{tn}$. As a result, its loop gain at the meta-stable point is always larger than unity and any small disturbance is enough to drive it to a stable state. In general, due to device mismatches, the meta-stable potential of nodes V_1 and V_2 are not the same. For instance, **Fig. 6-2** depicts the transient waveforms of a SA where V_{S1} and V_{S2} represent the meta-stable values of V_1 and V_2, respectively and $V_{CM} = (V_1 + V_2)/2$, $V_{SCM} = (V_{S1}+ V_{S2})/2$. At $V_1 = V_{S1}$ and $V_2 = V_{S2}$, we have:

$$V_{S1} \times (G_{mp2} - G_{mn2}) = 0 \tag{6.1a}$$

$$V_{S2} \times (G_{mp1} - G_{mn1}) = 0 \tag{6.1b}$$

Where the trans-conductance values, Gm, are dependent on terminal voltages and intrinsic device parameters of the transistors.

Observation: At any point of time, if $V_1 > V_{S1}$, V_1 has the tendency to push V_2 to a potential smaller than V_{S2}. On the other hand, if $V_1 < V_{S1}$, it has the tendency to push V_2 to

a potential higher than V_{S2}. However, if $V_1 = V_{S1}$, it has the tendency to push V_2 back to V_{S2}. The same observation is applicable to V_2. These observations can be easily proved using the inverting property of the inverter.

Consider the case whereby $V_2 > V_1$ and the initial common-mode input voltage, V_{CM} is equal to meta-stable common-mode voltage, V_{SCM}. A correct sensing operation will eventually drive V_2 to V_{DD} and V_1 to 0. However, with significant device mismatches, there are three possibilities depending primarily on the values of input voltage, V_{in} (= $V_2 - V_1$) and V_S (= $V_{S2} - V_{S1}$): (1) $V_{in} < V_S$. In this case, the initial value of V_1 is higher than V_{S1} and the initial value of V_2 is lower than V_{S2}. According to the above-mentioned observations, V_1 will drive V_2 to a lower potential while V_2 will drive V_1 to a higher potential. With the assumption that the loop gain around the meta-stable points is higher than unity, the positive feed-back configuration of the circuit will finally make a wrong decision, as illustrated in **Fig. 6-2(a)**. (2) $V_{in} > V_S$. In this scenario, V_1 is lower than V_{S1} and V_2 is higher than V_{S2}. Using the same argument, we can conclude that the circuit makes a correct decision in this case **(Fig. 6-2(b))**. (3) $V_{in} = V_S$. The circuit falls into meta-stable state. From this observation, the positive feed-back configuration of the cross-coupled amplifier implies that if $V_{CM} = V_{SCM}$, the correctness of the sensing process can be decided at time t=0 when it is enabled. In a correct sensing cycle, V_{21} keeps increasing and hence its derivative is positive, even at the starting time (t = 0), as illustrated in **Fig. 6-2(b)**. This criterion can therefore be used to determine the offset voltage of the SA, which is the minimum required input voltage to realize a correct sensing (and is equal to V_S).

Until now, we have only discussed the "direction" of movement of the output nodes V_1 and V_2 but not the "speed" of movement. Since the lump-sum load capacitances (which include both the load and the parasitic capacitances) only affect the speed of charging/discharging of the output nodes, they have no impact on the correctness of the sensing process. However, this does not hold if $V_{CM} \neq V_{SCM}$.

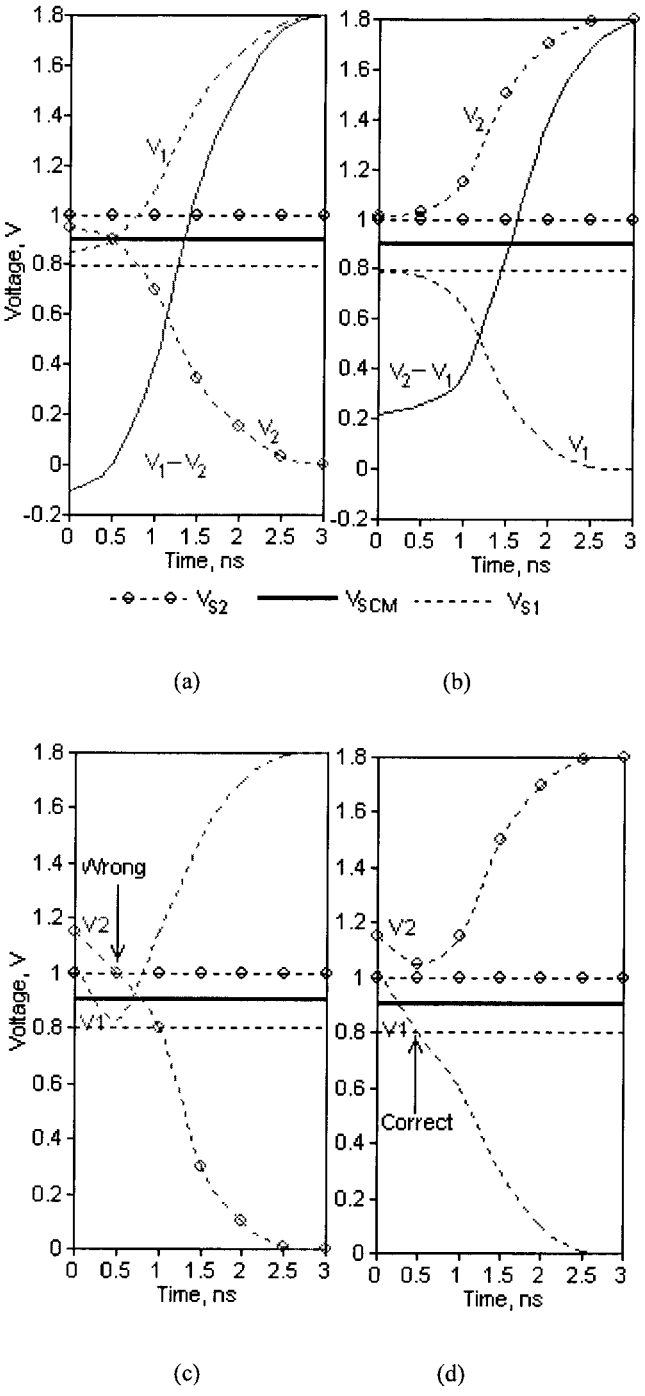


Figure 6-2 Switching waveforms of a latch-type SA. a) $V_{CM}=V_{SCM}$, wrong decision. b) $V_{CM}=V_{SCM}$, correct decision. c) $V_{CM}>V_{SCM}$, wrong decision. d) $V_{CM}>V_{SCM}$, correct decision

If $V_{CM} \neq V_{SCM}$, for the sake of simplicity, we only consider the case when $V_{CM} > V_{SCM}$. If $V_1 < V_{S1}$ and $V_2 > V_{S2}$, they will apparently lead to a correct sensing. However, if $V_1 > V_{S1}$ and $V_2 > V_{S2}$, using the same argument as above, it can be shown that both initial nodes will descend (**Fig. 6-2(c)** and **Fig. 6-2(d)**). When either V_1 or V_2 touches their

corresponding meta-stable value, i.e. V_{S1} and V_{S2}, the potential of the other node will decide the correctness of the sensing cycle. For instance, in **Fig. 6-2(c)**, when both nodes descend, V_2 approaches V_{S2} first when V_1 is still higher than V_{S1}. As a result, V_1 will "push" V_2 to a lower potential, i.e. lower than V_{S2}. In turn, V_2 pushes V_1 to a higher potential and the regenerative process continues and finally results in a wrong sensing. In contrast, if V_1 reaches V_{S1} when V_2 is still at a higher potential than V_{S2}, they lead to a correct sensing. In this second scenario where $V_{CM} \neq V_{SCM}$, discharging/charging speed at the output nodes is important. Thus, the capacitance mismatch comes into effect.

In the next Section, we will quantify the input-offset voltage due to the device parameter mismatches (V_{th} and K) with the assumption that $V_{CM} = V_{SCM}$. In **Section 6.2.4**, we extend our model to the case where $V_{CM} > V_{SCM}$ and include the effect of the capacitance mismatch.

It is worth mentioning that V_{DD} and temperature variations are not considered in the offset analysis since they are global parameters and have insignificant contribution to the total mismatch effect [19, 281]. The worst-case scenario is assumed to incorporate the effects of both V_{DD} and temperature. Consequently, only the threshold voltage, the trans-conductance and the capacitance mismatches contribute to the input-offset voltage.

6.2.2 Input-offset when $V_{CM} = V_{SCM}$

Suppose that the latch structure is enabled when V_1 and V_2 are at potentials close enough to V_{SCM}, the common-mode DC voltage (which again is assumed to be near $V_{DD}/2$). As a result, all four devices are in saturation and the MOSFET alpha law can be used (for 0.18 μm or larger CMOS technologies). By applying Kirchhoff's Current Law (KCL) at nodes V_1 and V_2, we have:

$$\frac{\partial v_2(t)}{dt} = \frac{1}{C_{TOT}}\left[\frac{K_{P2}}{2}(v_1 - V_{DD} + V_{TP2})^{\alpha} - \frac{K_{N2}}{2}(v_1 - V_{TN2})^{\alpha}\right] \tag{6.2}$$

$$\frac{\partial v_1(t)}{dt} = \frac{1}{C_{TOT}}\left[\frac{K_{P1}}{2}(v_2 - V_{DD} + V_{TP1})^{\alpha} - \frac{K_{N1}}{2}(v_2 - V_{TN1})^{\alpha}\right] \tag{6.3}$$

where

$C_{TOT} = C_L + C_{parasitic}$

$\alpha = 2$ for channel length greater or equal to 0.18 μm

K is the transconductance

V_T is the threshold voltage

And $K_{P1} = K_{P2} = K_P, K_{N1} = K_{N2} = K_N, V_{TP1} = V_{TP2} = V_{TP}, V_{TN1} = V_{TN2} = V_{TN}$ at the perfectly matched condition, assuming that all transistor parameters are fixed over time.

This model is based on representative I-V characteristic equations for generalization purpose, which in fact merely serves as an illustration. Circuit designers can then adopt the same technique to suit their specific process technologies with different I-V characteristics. A more accurate model (with $\alpha < 2$) is going to be used in latter section to predict the offset voltage in the 32 nm and 22 nm technologies.

Subtracting both sides of (6.2) by (6.3):

$$\begin{aligned}\frac{\partial v_{21}(t)}{dt} &= \frac{1}{C_{TOT}}\left[\frac{K_{P2}}{2}(v_1 - V_{DD} + V_{TP2})^2 - \frac{K_{N2}}{2}(v_1 - V_{TN2})^2\right] \\ &\quad - \frac{1}{C_{TOT}}\left[\frac{K_{P1}}{2}(v_2 - V_{DD} + V_{TP1})^2 - \frac{K_{N1}}{2}(v_2 - V_{TN1})^2\right] \\ &= F(u,t)\end{aligned} \tag{6.4}$$

Eq. (6.4) describes the response of the system with respect to a set of device parameters $u=(u_1,u_2,..u_8)$ along the time t. In this case, u= (V_{TP1}, V_{TP2}, V_{TN1}, V_{TN2}, K_{N1}, K_{N2}, K_{P1}, K_{P2}). Assuming that we know in closed-form the solution of (6.6) (or the exact value of the solution of (6.4) at all points of interest) with respect to the nominal parameter u_{NOM}. Thus, by applying Taylor's theorem with multiple variables, the response of the system to the parameter, $u = u_{NOM} + \delta u$,can be approximated as:

$$F(u,t) = F(u_{NOM}) + \frac{\delta u}{1!} \times \frac{\partial F}{\partial u}\Big|_{u=u_{NOM}} + \frac{(\delta u)^2}{2!} \times \frac{\partial^2 F}{\partial u^2}\Big|_{u=u_{NOM}} + ... \tag{6.5}$$

If $\delta u_i \ (i = \overline{1,m})$ are small, all components with the order higher than one can be ignored. Equation (6.4) can be rewritten explicitly as:

$$F(u,t) = F(u_{NOM}) + \sum_{i=1}^{m} \frac{\delta u_i}{1!} \times \frac{\partial F}{\partial u_i}\Big|_{u=u_{NOM}} \tag{6.6}$$

Where, by assuming that $v_1(0) \approx v_2(0) \approx V_{CM}$, we have:

$$F(u_{NOM}) = \frac{1}{C_{TOT}} v_{21}(t) \begin{bmatrix} (K_P(V_{DD} - V_{CM} - V_{TP}) + \\ K_N(V_{CM} - V_{TN}) \end{bmatrix} \tag{6.7}$$

The beauty of the model is that we only need to evaluate $\frac{\partial v_{21}}{\partial t}\Big|_{t=0}$. Eq. (6.7) clearly shows that an infinitesimally small $v_{21}(0)$ is sufficient for a correct sensing in a perfectly matched system since $\frac{\partial v_{21}}{\partial t}\Big|_{t=0} > 0$ as long as $v_{21}(0) > 0$. It is worth mentioning that this derivative can be evaluated at t = 0 because all devices are in saturation and ready to work at this time. In other words, there is no disruption at this point. If one worries about the existence of $\frac{\partial v_{21}}{\partial t}\Big|_{t=0}$, the designer can evaluate it at $t = 0^+$. By expanding Eq. (6.6) we have:

$$\begin{aligned} \frac{\partial v_{21}}{\partial t}(u_{NOM} + \delta u) &= F(u_{NOM}) + \\ &+ \left[\delta V_{TP1} \frac{dF}{dV_{TP1}} + \delta V_{TP2} \frac{dF}{dV_{TP2}} + \delta V_{TN1} \frac{dF}{dV_{TN1}} + \delta V_{TN2} \frac{dF}{dV_{TN2}} \right] + \\ &+ \left[\delta K_{P1} \frac{dF}{dK_{P1}} + \delta K_{P2} \frac{dF}{dK_{P2}} + \delta K_{N1} \frac{dF}{dK_{N1}} + \delta K_{N2} \frac{dF}{dK_{N2}} \right] \\ &= F(u_{NOM}) + Th + Tr \end{aligned} \tag{6.8}$$

where all mismatch parameters are set as in Table X to obtain the worst-case scenario.

TABLE X. MISMATCH PARAMETERS FOR THE DEVICES IN FIG. 6-1

P1	P2	N1	N2
$\delta V_{TP1} = -\delta V_{TP}$	$\delta V_{TP2} = +\delta V_{TP}$	$\delta V_{TN1} = +\delta V_{TN}$	$\delta V_{TN2} = -\delta V_{T}$
$\delta K_{P1} = +\delta K_{P}$	$\delta K_{P2} = -\delta K_{P}$	$\delta K_{N1} = -\delta K_{N}$	$\delta K_{N1} = +\delta K_{N}$

All derivatives are evaluated at the nominal condition without any perturbation, as shown below:

- $$\frac{\partial F}{\partial V_{TP1}}\Big|_{u_{NOM}} = \frac{-1}{C_{TOT}}\left[K_{P1}(v_2 - V_{DD} + V_{TP1})\right] \quad (6.9)$$

- $$\frac{\partial F}{\partial V_{TN1}}\Big|_{u_{NOM}} = \frac{-1}{C_{TOT}}\left[K_{N1}(v_2 - V_{TN1})\right] \quad (6.10)$$

- $$\frac{\partial F}{\partial V_{TP2}}\Big|_{u_{NOM}} = \frac{1}{C_{TOT}}\left[K_{P2}(v_1 - V_{DD} + V_{TP2})\right] \quad (6.11)$$

- $$\frac{\partial F}{\partial V_{TN2}}\Big|_{u_{NOM}} = \frac{1}{C_{TOT}}\left[K_{N2}(v_1 - V_{TN2})\right] \quad \ldots \quad (6.12)$$

$$\begin{aligned} Th &= \delta V_{TP1}\frac{\partial F}{\partial V_{TP1}} + \delta V_{TP2}\frac{\partial F}{\partial V_{TP2}} + \delta V_{TN1}\frac{\partial F}{\partial V_{TN1}} + \delta V_{TN2}\frac{\partial F}{\partial V_{TN2}} \\ \Rightarrow \quad &= \frac{1}{C_{TOT}}\left[K_P(v_1 + v_2 - 2V_{DD} + 2V_{TP})\delta V_{TP} - K_N(v_1 + v_2 - 2V_{TN})\delta V_{TN}\right] \quad (6.13) \\ &= \frac{-1}{C_{TOT}}\left[K_P(2V_{DD} - 2V_{TP} - 2V_{CM})\delta V_{TP} + K_N(2V_{CM} - 2V_{TN})\delta V_{TN}\right] \end{aligned}$$

Similarly with the trans-conductance mismatch, we have:

$$\begin{aligned} Tr &= \delta K_{P1}\frac{dF}{dK_{P1}} + \delta K_{P2}\frac{dF}{dK_{P2}} + \delta K_{N1}\frac{dF}{dK_{N1}} + \delta K_{N2}\frac{dF}{dK_{N2}} \\ &= \frac{-1}{2C_{TOT}}\left[\left[(v_1 - V_{DD} - V_{TP})^2 + (v_2 - V_{DD} - V_{TP})^2\right]\delta K_P + \left[(v_1 - V_{TN})^2 + (v_2 - V_{TN})^2\right]\delta K_N\right] \quad (6.14) \\ &= \frac{-1}{C_{TOT}}\left[\delta K_P(V_{DD} - V_{TP} - V_{CM})^2 + \delta K_N(V_{CM} - V_{TN})^2\right] \end{aligned}$$

By replacing (6.13) and (6.14) into (6.8) we have

$$\frac{\partial v_{21}}{\partial t}(u_{NOM} + \delta u) = F(u_{NOM}) + \frac{-2}{C_{TOT}}\begin{bmatrix} K_P(V_{DD} - V_{TP} - V_{CM})\delta V_{TP} + \\ K_N(V_{CM} - V_{TN})\delta V_{TN} \end{bmatrix} + \frac{-1}{C_{TOT}}\begin{bmatrix} \delta K_P(V_{DD} - V_{TP} - V_{CM})^2 + \\ \delta K_N(V_{CM} - V_{TN})^2 \end{bmatrix} \quad (6.15)$$

By replacing (6.7) into (6.15) we have:

$$\begin{aligned} &\frac{\partial v_{21}}{\partial t}(u_{NOM} + \delta u) = \\ &= \frac{1}{C_{TOT}}\left[\left(v_{21}(t)\begin{bmatrix}(K_P(V_{DD} - V_{CM} - V_{TP}) + \\ K_N(V_{CM} - V_{TN})\end{bmatrix}\right) - \begin{pmatrix} K_P(2V_{DD} - 2V_{TP} - 2V_{CM})\delta V_{TP} + \\ K_N(2V_{CM} - 2V_{TN})\delta V_{TN} + \\ \delta K_P(V_{CM} - V_{DD} + V_{TP})^2 + \\ \delta K_N(V_{CM} - V_{TN})^2 \end{pmatrix}\right] \end{aligned} \quad (6.16)$$

By rearranging the terms in (6.16), the criterion for a correct sensing is equivalent to:

$$v_{21}(0) > \frac{A\delta V_{TP} + B\delta V_{TN} + C\delta K_P + D\delta K_N}{E} = V_S \tag{6.17}$$

Where

$$A = 2K_P(V_{DD} - V_{TP} - V_{CM})$$

$$B = 2K_N(V_{CM} - V_{TN})$$

$$C = (V_{DD} - V_{TP} - V_{CM})^2$$

$$D = (V_{CM} - V_{TN})^2$$

$$E = \frac{A+B}{2}$$

V_S is the offset voltage of the SA.

6.2.3 Discussion

Eq. (6.17) describes the worst-case offset voltage of the latch-type SA. It also shows the impact of individual mismatch components, namely the threshold voltage and the trans-conductance.

6.2.3.1 Threshold voltage mismatch

The impact of the threshold voltage mismatch can be seen by assuming that the trans-conductances are perfectly matched. As a result, Eq. (6.19) is reduced to:

$$v_{21}(0) > \frac{\left[2K_P(V_{DD} - V_{TP} - V_{CM})\delta V_{TP} + 2K_N(V_{CM} - V_{TN})\delta V_{TN}\right]}{\left[(K_P(V_{DD} - V_{CM} - V_{TP}) + K_N(V_{CM} - V_{TN})\right]} \tag{6.18}$$

If the SA is symmetrical, i.e. $K_N = K_P$ and $2V_{SCM=}V_{DD+}V_{TN}-V_{TP}$, we have:

$$v_{21}(0) > \delta V_{TP} + \delta V_{TN} \tag{6.19}$$

Eq. (6.19) shows that the offset voltage is equal to the sum of the threshold voltage mismatches. Therefore, the impact of the threshold voltage mismatch is the most dominant component. This also agrees with the formula reported in [268] and [273].

6.2.3.2 Trans-conductance mismatch

The impact of the trans-conductance mismatch can be seen by assuming that the threshold voltages are perfectly matched. As a result, Eq. (6.17) is reduced to:

$$v_{21}(0) > \frac{\left[\delta K_P (v_{CM} - V_{DD} + V_{TP})^2 + \delta K_N (v_{CM} - V_{TN})^2\right]}{\left[(K_P (V_{DD} - V_{CM} - V_{TP}) + K_N (V_{CM} - V_{TN})\right]} \tag{6.20}$$

If the SA is symmetrical, i.e. $K_N = K_P$ and $2V_{CM}=V_{DD}+V_{TN}-V_{TP}$, we have:

$$v_{21}(0) > \frac{(\delta K_P + \delta K_N)}{(K_P + K_N)} \frac{(V_{DD} - V_{TN} - V_{TP})}{2} \tag{6.21}$$

Eq. (6.21) shows that the offset voltage is linearly related to the percentage trans-conductance mismatches and the power supply. Assuming that the threshold voltage is fixed for a given process, a higher power supply voltage leads to a higher input-offset caused by the trans-conductance mismatch. For example, in a 0.18 μm/1.8 V, V_{TP} is 0.37 V, V_{TN} is 0.29 V and hence $\frac{(V_{DD} - V_{TN} - V_{TP})}{2}$ equals to 0.57 V. Thus, 10% trans-conductance mismatch results in a 57 mV input-offset voltage, which is comparable to the threshold mismatch. Therefore, one of the suggestions to reduce the input-offset is by reducing the supply voltage.

6.2.3.3 Second-order approximation

The new criterion offers a very simple and straightforward procedure to calculate the combined input-offset voltage of the SA. Other than the analysis done in this work, one can choose his/her own mismatch parameters. For example, one can separate the trans-conductance into four terms W, L, C_{ox} and μ and replace them in Eq. (6.8):

$$\delta K = \frac{\partial K}{\partial L}\delta L + \frac{\partial K}{\partial W}\delta W + \frac{\partial K}{\partial C_{ox}}\delta C_{ox} + \frac{\partial K}{\partial \mu}\delta \mu \tag{6.22}$$

Furthermore, one can investigate the second-order approximation in the Taylor expansion with no difficulty. The reason is that the proposed method only requires the

estimation of $\frac{\partial v_{21}}{\partial t}$ at t=0. For instance, if we consider the interaction impact of the simultaneous trans-conductance and the threshold mismatches, the second-order approximation must be used. With the above-mentioned eight parameters, the estimated offset voltage is:

$$v_{os} = \frac{A\delta V_{TP} + B\delta V_{TN} + C\delta K_P + D\delta K_N}{E - 2(\delta V_{TP}\delta K_P + \delta V_{TN}\delta K_N)} \tag{6.23}$$

The impact of the second-order components on the offset voltage is insignificant if the mismatches are smaller than 5%. However, with larger mismatches, the second-order components need to be taken into consideration. For example, with 10% parameter mismatches, the second-order approximation is 4% larger than the first-order approximation and closer to the simulation results. This point is going to be verified in **Section 6.3**.

6.2.3.4 Suitability of the model in sub-45 nm technologies.

The precision of the approximation depends on the accuracy of the I-V characteristic equations used in Eqs. (6.2) and (6.3). As technology moves into the nanometers range, the square law used in Eqs. (6.2) and (6.3) is no longer valid. Thus, it must be adjusted to fit the operating characteristic of the devices. One simple approximation is to use α close to one, which is more accurate in describing the saturation current of the MOS device in nanometer technology. Thus, the above-mentioned parameters are computed as follows:

$$\begin{aligned}
F(u_{NOM}) &= \frac{1}{C_{TOT}} v_{21}(t) Q \\
Q &= \begin{Bmatrix} K_P[1 + \ln(V_{DD} - V_{CM} - V_{TP})(\alpha - 1)] + \\ K_N[1 + \ln(V_{CM} - V_{TN})(\alpha - 1)] \end{Bmatrix} \\
Th &= \frac{-1}{C_{TOT}} \alpha \begin{bmatrix} K_P(V_{DD} - V_{TP} - V_{CM})^{\alpha-1}\delta V_{TP} \\ + K_N(V_{CM} - V_{TN})^{\alpha-1}\delta V_{TN} \end{bmatrix} \\
Tr &= \frac{-1}{C_{TOT}} \left[\delta K_P(V_{CM} - V_{DD} + V_{TP})^{\alpha} + \delta K_N(V_{CM} - V_{TN})^{\alpha}\right]
\end{aligned} \tag{6.24}$$

In the case of symmetrical nominal conditions, i.e. $K_N = K_P$ and $2V_{SCM=}V_{DD+}V_{TN}-V_{TP}$,

we have:

$$v_{2I}(0) > (\delta V_{TP} + \delta V_{TN}) \frac{\alpha F^{\alpha-1}}{2[1+\ln(F)(\alpha-1)]} + \frac{(\delta K_P + \delta K_N)}{K_p + K_N} \frac{F^{\alpha}}{[1+\ln(F)(\alpha-1)]} \quad (6.25)$$

where $F = V_{DD} - V_{TP} - V_{CM} = V_{CM} - V_{TN}$

The improvement of (6.25) in comparison with (6.17) will be illustrated in **Section 6.3**.

6.2.3.5 Basic cross-coupled inverters' derivatives

Generally, the latch-type SA used in RAM read-out circuits includes one equalization device or a MOS tail device [93, 97, 246, 261, 271, 281]. Thus, its actual transient response is different from what is shown in Eq. (6.2) and Eq. (6.3). However, the criterion for a correct sensing still holds. In fact, the MOS tail device adjusts the voltage across the SA while the equalization device adds a parasitic capacitance [275]. The effect of V_{DD} has been included in the above while impact of the parasitic capacitance is going to be investigated in **Subsection 6.2.4**.

6.2.4 Input-offset when $V_{CM} > V_{SCM}$

As mentioned before, we only need to consider the case when $V_1(0) > V_{S1}$ and $V_2(0) > V_{S2}$ to find the minimum required input voltage. Similar to [275], the cross-coupled amplifier can be modeled using the following small-signal equations:

$$C_{TOT1} \frac{\partial V_1}{\partial t} = [V_2(t) - V_{S2}](G_{mp1} - G_{mn1}) \quad (6.26a)$$

$$C_{TOT2} \frac{\partial V_2}{\partial t} = [V_1(t) - V_{S1}](G_{mp1} - G_{mn1}) \quad (6.26b)$$

As both nodes descend, their derivatives are negative, i.e. $G_{mpi} < G_{mni}$. Assume that they descend at constant speeds until either of them reaches their corresponding meta-stable potential and the circuit makes decision. For a correct sensing, V_1 must reach V_{S1} first, after a time interval Δt:

$$V_{S1} = V_1(\Delta t) = V_1(0) + (V_2(0) - V_{S2})\frac{G_{m1}}{C_{TOT1}}\Delta t$$
$$\therefore \quad \Delta t = \frac{(V_{S1} - V_1(0))}{(V_2(0) - V_{S2})}\frac{C_{TOT1}}{G_{m1}} \tag{6.27}$$

And

$$V_{S2} < V_2(\Delta t) = V_2(0) + (V_1(0) - V_{S1})\frac{G_{m2}}{C_{TOT2}}\Delta t$$
$$= V_2(0) + (V_1(0) - V_{S1})\frac{G_{m2}}{C_{TOT2}}\frac{(V_{S1} - V_1(0))}{(V_2(0) - V_{S2})}\frac{C_{TOT1}}{G_{m1}} \tag{6.28}$$
$$\therefore V_{S2} - V_2(0) < (V_1(0) - V_{S1})\frac{G_{m2}}{C_{TOT2}}\frac{(V_{S1} - V_1(0))}{(V_2(0) - V_{S2})}\frac{C_{TOT1}}{G_{m1}}$$

$$\therefore [V_{S2} - V_2(0)]^2 > [V_{S1} - V_1(0)]^2 \frac{G_{m2}}{G_{m1}}\frac{C_{TOT1}}{C_{TOT2}}$$
$$\therefore [V_2(0) - V_{S2}] > [V_1(0) - V_{S1}]\sqrt{\frac{G_{m2}}{G_{m1}}\frac{C_{TOT1}}{C_{TOT2}}} \tag{6.29}$$

Where

$$G_m = G_{mp} - G_{mn}$$
$$= K_p(V_{DD} - V_G - V_{tp})(1 + \lambda_p V_{DSp}) -$$
$$- K_n(V_G - V_{tn})(1 + \lambda_n V_{DSn}) \tag{6.30}$$

It is important to note that the result obtained in Eq (6.29) agrees with what was reported in [275] by letting $\frac{G_{m2}}{G_{m1}} = 1$, although our model is much simpler and involves only linear equations.

As V_1 and V_2 decrease, magnitudes of both G_{m1} and G_{m2} decrease (Eq. (6.29)) but the ratio $\frac{G_{m2}}{G_{m1}}$ can be assumed to remain unchanged with good approximation if the initial potential of V_1 is close enough to V_{S1}. By assuming that the G_m and C mismatches are small, Eq. (6.29) becomes:

$$\begin{aligned}
&[V_2(0)-V_{S2}]>[V_1(0)-V_{S1}]\left(1+\frac{1}{2}\frac{\Delta_{GC}}{G_{m1}C_{TOT2}}\right)\\
&\Leftrightarrow [V_2(0)-V_{S2}]-[V_1(0)-V_{S1}]>[V_1(0)-V_{S1}]\frac{1}{2}\frac{\Delta_{GC}}{G_{m1}C_{TOT2}}\\
&\Leftrightarrow [V_2(0)-V_1(0)]>V_S+[V_1(0)-V_{S1}]\frac{1}{2}\frac{\Delta_{GC}}{G_{m1}C_{TOT2}}
\end{aligned} \tag{6.31}$$

where $\Delta_{GC}=G_{m2}C_{TOT1}-G_{m1}C_{TOT2}$

So the overall input-offset voltage of the SA is:

$$\begin{aligned}
V_{offset} &= V_S+[V_1(0)-V_{S1}]\frac{1}{2}\frac{\Delta_{GC}}{G_{m1}C_{TOT2}}\\
&= Intrinsic_offset + Extrinsic_offset
\end{aligned} \tag{6.32}$$

Eq. (6.32) shows that the overall offset voltage contains two components: Intrinsic offset and extrinsic offset. While the intrinsic offset is caused solely by the intrinsic device mismatches (V_{th}, K, etc), the extrinsic offset is caused by the initial input potentials and the parasitic mismatch. Several previously published results can be found by simplifying Eq. (6.31) to the corresponding special case. For instance, if the devices are matched (i.e. $V_S = 0$ and $G_{m1} = G_{m2}$), then Eq. (6.31) becomes:

$$V_{offset}=[V_1(0)-V_{S1}]\frac{1}{2}\frac{\Delta_C}{C_{TOT2}} \tag{6.33}$$

Where $\Delta_C=C_{TOT1}-C_{TOT2}$

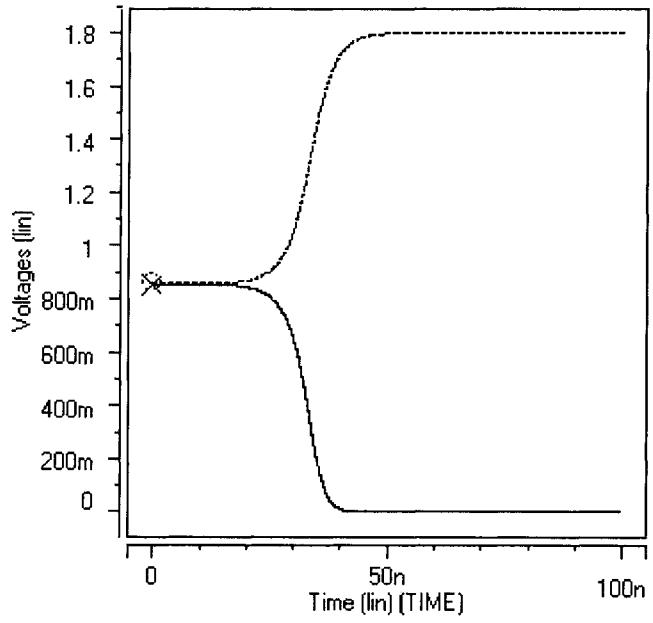


Figure 6-3 A correct sensing cycle with extreme capacitance mismatch and 0.6 mV input, all other device parameters are matched

This is exactly the same as the result reported in [275]. It is also worth mentioning that the capacitance mismatch itself does not cause the input-offset voltage when V_{CM}=V_{SCM}. **Fig. 6-3** confirms this by showing a correct sensing cycle with 50% capacitance mismatch and all other device parameters are matched. Simulation waveforms show that a matched SA makes a correct decision (V_{in}= 0.6 mV) even with extreme capacitance mismatch. However, once present, it is multiplied with both G_m mismatch and the difference between the initial input voltages and their meta-stable value (i.e. $V_I(0) - V_{SI}$). This point will be discussed in **Section 6.3**

6.3 Simulation results and analysis

6.3.1 Methodology

Extensive simulations using HSPICE simulator have been carried out to verify the accuracy of the proposed criterion. For the sake of simplicity and the clarity of the comparisons, several single-factor-at-a-time tests have been performed to evaluate the impact of individual mismatch type. All test corners have been measured using three

technology nodes: a 0.18 µm/ 1.8V standard CMOS process from GFS, a 32 nm and a 22nm Predictive Technology Model (PTM) developed by the Nanoscale Integration and Modeling (NIMO) Group at Arizona State University (ASU). Simulation results were plotted in comparison with the proposed approximated model. After that, combined mismatch simulations using multiple-factor-at-a-time approach have been carried out on the 0.18 µm/ 1.8V standard CMOS process from GFS. These simulations include all the mentioned mismatch types (i.e. δV_{TP}, δV_{TN}, δK_P, δK_N in Eq. (6.17) and Δ_{GC} in Eq. (6.32)) simultaneously. Simulation results are then compared with the first- and the second-order approximations to reaffirm the model.

The circuit set-up shown in **Fig. 6-1** was used to perform the simulations. The threshold voltage mismatch was injected by adjusting the V_{to} parameter in the model cards. The trans-conductance and capacitance mismatches were modeled by adjusting the W and L of the devices and C_L. The simulated offset voltage is defined as the largest differential input voltage that produces an incorrect sensing. For instance, **Fig. 6-4** shows a correct and an incorrect sense cycle of a 5% K-mismatch SA with a 31 mV input-offset.

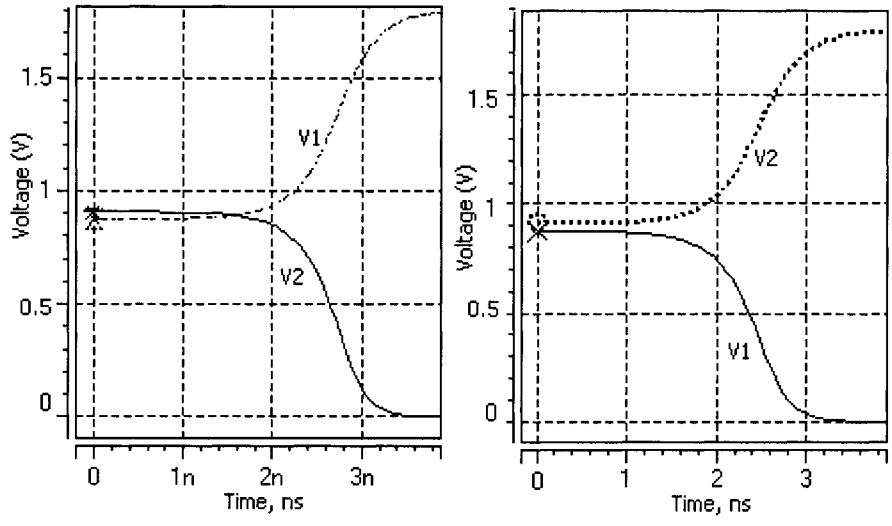


Figure 6-4 Output voltages at of a 5% K-mismatch SA using 0.18 µm/ 1.8 V standard CMOS process from GFS a) Wrong sensing with $v_{21}(0) = 31$ mV. b) Correct sensing with $v_{21}(0) = 32$ mV. $V_{CM} = V_{SCM}$

6.3.2 Results comparison

6.3.2.1 $V_{CM} = V_{SCM}$

Figs. 6-5 to **Fig. 6-8** present the input-offset voltage of the SA where $V_{CM} = V_{SCM}$. **Fig. 6-5** shows the first-order approximation of the input-offset voltage at 0.18 µm/ 1.8 V standard CMOS process from GFS. It confirms the accuracy of the model with about 95% precision. For example, at 10% V_{th} mismatch, the input-offset according to the simulation is 84 mV while the calculation results in 80 mV. Regarding the K mismatch, these quantities are 61.5 mV and 58.5 mV, respectively. **Fig. 6-6** also reaffirms previous works [268, 273] that the V_{th} mismatch has higher impact on the offset voltage with the offset voltages of 84 mV and 61.5 mV at 10% V_{th} and 10% K mismatches, respectively. However, they suggest that the impact of the K mismatch is significant and should not be ignored while calculating the overall offset of the SA.

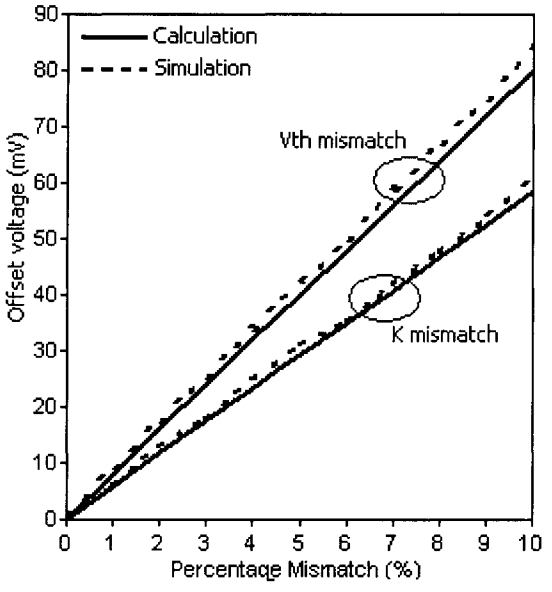


Figure 6-5 Input-offset voltage due to the threshold voltage and the trans-conductance mismatches individually at 0.18 µm/ 1.8 V standard CMOS process from GFS. Percentage V_{th} mismatch $= \frac{\delta V_{th}}{V_{th}} \times 100\%$ **. Percentage K mismatch** $= \frac{\delta K}{K} \times 100\%$ **.** $V_{CM} = V_{SCM}$

Fig. 6-6 illustrates the combined effects of the V_{th} and the K mismatches on the offset voltage. It includes both the first and the second-order approximations. As the threshold

and the trans-conductance mismatches interact, their combined effect results in an input-offset which is slightly higher than the sum of their individuals. Consequently, the second-order approximation is required to estimate the interaction between the two mismatches. As shown in **Fig. 6-6**, the second-order approximation gives better results with about 99% precision when compared to the 95% precision of the first-order approximation. This improvement becomes more profound as the mismatch levels increase. For example, at 10% V_{th} and K mismatches, the first-order, the second-order approximations and the simulation offset are 138 mV, 144 mV and 144.5 mV, respectively

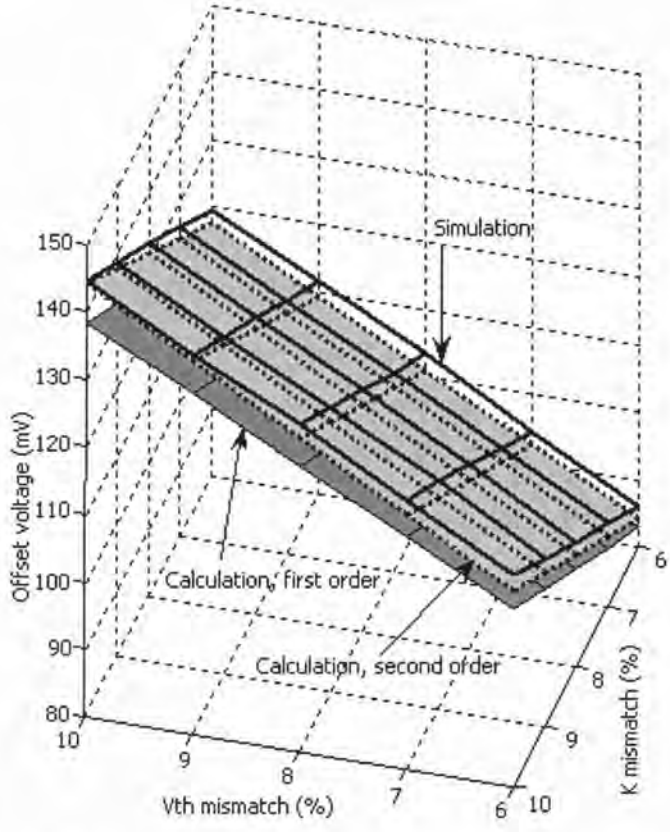
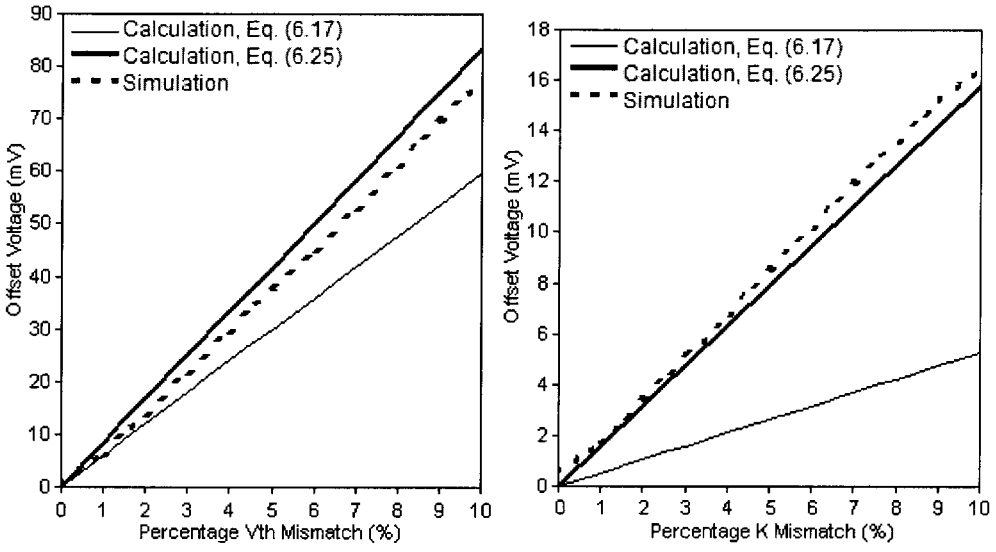


Figure 6-6 Input-offset voltage at 0.18 µm/1.8 V standard CMOS process from GFS due to the simultaneous mismatches. $V_{CM} = V_{SCM}$

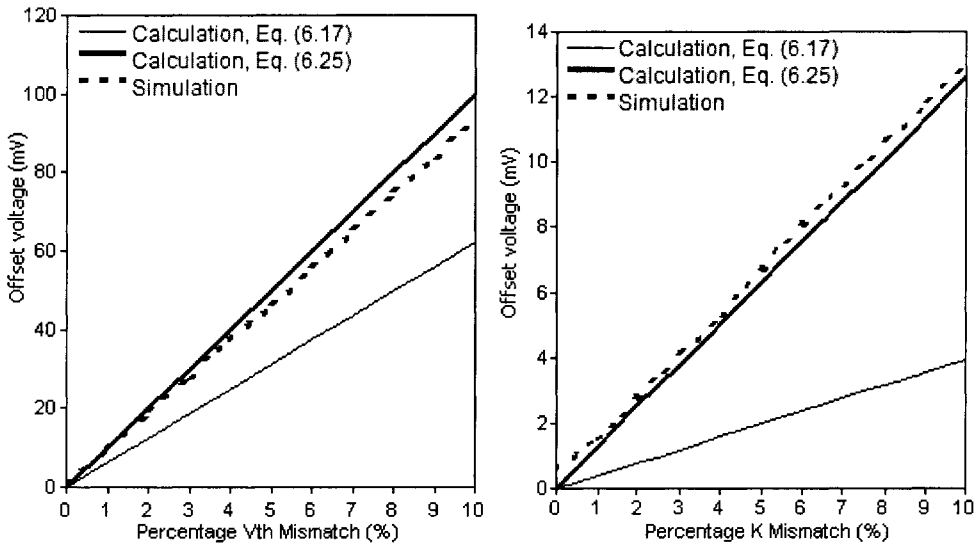
In order to verify the accuracy of the proposed model in the nanometer regimes, extensive simulations have been carried out with 32 nm and 22 nm predictive models. Using these technologies, the approximation using Eq. (6.17) reveals its weakness. As shown in **Figs. 6-7** and **Fig. 6-8**, the calculated offset using Eq. (6.17) lags far below the

simulation results, for both V_{th} and the K mismatches. This can be explained by the fact that the square law used in Eq. (6.2) and Eq. (6.3) is not accurate in the nanometer technologies. Therefore, a simple adjustment has been adopted by using lower values of α in Eq. (6.2) and Eq. (6.3). Thus, Eq. (6.25) was used to compute the offset voltage. Values of α were extracted from operating curves of the MOS devices. At 32 nm, α is equal to 1.15 whereas at 22 nm it is 1.09.



(a) (b)

Figure 6-7 Input-offset voltage at 32nm Predictive Technology Model, 0.8 V supply voltage, due to the device mismatches. α was extracted from the operating curves of the MOSs and is equal to 1.15. (a) threshold voltage mismatch. (b) trans-conductance mismatch. $V_{CM} = V_{SCM}$



(a) (b)

Figure 6-8 Input-offset voltage at 22nm Predictive Technology Model, 0.7 V supply voltage due to the device mismatches. α was extracted from the operating curves of the MOSs and is equal to 1.09. a) threshold voltage mismatch. b) trans-conductance mismatch. $V_{CM} = V_{SCM}$

Although this replacement does not describe the exact I-V characteristic equations of the MOS devices, it provides significant improvements, also shown in **Fig. 6-7** and **Fig. 6-8**. In this work, we focus on developing and verifying the proposed method and hence only simple adjustment has been used for the 32 nm and 22 nm characteristic equations. If higher precision is required, one must adopt more accurate models in Eq. (6.2) and Eq. (6.3). As long as these equations are correct, the proposed method offers very simple and straightforward steps to obtain the offset of the SA.

It is worth mentioning here that at 22 nm and 32 nm technology nodes, the offset voltage due to the K mismatch is significantly smaller than that of the V_{th} mismatch, as shown in **Fig. 6-7** and **Fig. 6-8**. The reason is that the supply voltages used for these technologies are much lower than those at 0.18 μm technology. Since the K-mismatch-related offset voltage heavily depends on the supply voltage (as shown in Eq. (6.17) and Eq. (6.25), reducing the supply voltage will reduce the corresponding input-offset component.

6.3.2.2 $V_{CM} > V_{SCM}$

Similar simulation steps have been carried out in this scenario. For this case, the input common-mode voltage is raised to a higher potential compared to its meta-stable value. Offset voltages were recorded and compared to the analytical model in Eq. (6.33). All results are presented in **Fig. 6-9** and **Fig. 6-10**.

Fig. 6-9 features the effect of the capacitance mismatch on the overall offset. It is noticeable that even when C_L mismatch is 0%, the simulation and computed offset is larger than the intrinsic offset (i.e. V_S = 144.5 mV). This can be easily explained using Eq. (6.32): although C_{TOT1} equals to C_{TOT2}, device mismatches causes G_m mismatch and hence Δ_{GC} is positive. In fact, 10% V_{th} and K mismatch can lead to 30% G_m mismatch in our calculation. The C_L mismatch, as shown in **Fig. 6-9**, increases the offset but its magnitude is very small. This observation strongly agrees with the results in [268, 273, 275].

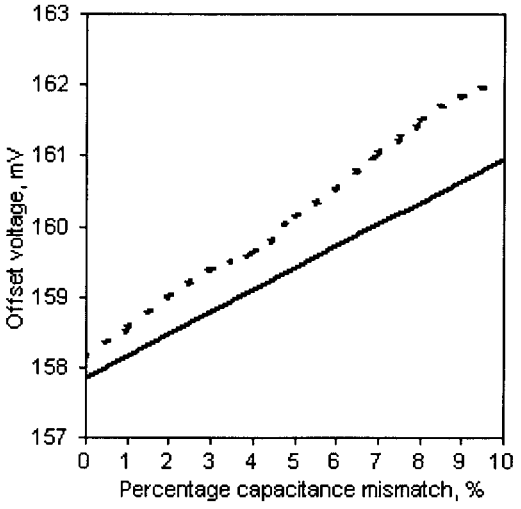


Figure 6-9 Input-offset voltage at 0.18 µm/1.8 V standard CMOS process from GFS due to 10% V_{th} and K mismatches against C_L mismatch values. $V_1(0) - V_{S1}$ = 100 mV. V_S = 144.5 mV

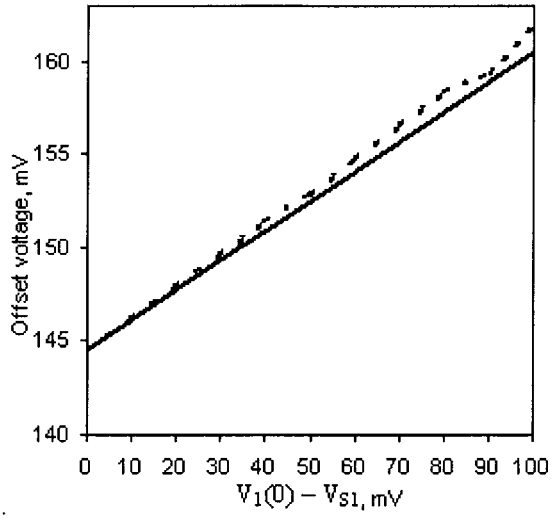


Figure 6-10 Input-offset voltage at 0.18 µm/1.8 V standard CMOS process from GFS due to 10% V_{th} and K and C_L mismatches against $V_1(0) - V_{S1}$. V_S = 144.5 mV

Fig. 6-10 depicts the impact of the difference between the initial input voltage V_1 and its meta-stable value V_{S1}. It shows that this factor has a greater impact on the input-offset voltage than the capacitance mismatch does. Therefore, it is recommended to enable the SA as close to the meta-stable value as possible to reduce the input-offset. It is also shown that when V_1 equals to V_{S1} (i.e. V_1-V_{S1} = 0), we have V_{offset} equals to V_S, as predicted in Eq. (6.33), even with 10% C_L mismatch. This result has reaffirmed the conclusion that C_L mismatch has no impact on offset voltage if the inputs are released at their optimum potential (i.e. close to meta-stable value) [268, 273].

6.4 Conclusion

A new approach and a new criterion have been proposed to calculate the input-offset voltage of the latch-type SA due to the device mismatches. The new criterion offers a simple yet accurate way to evaluate the correctness of a sensing cycle and hence the offset voltage. Since the proposed method only requires the evaluation of $\frac{\partial v_{21}}{\partial t}\Big|_{t=0}$, it can afford to include a large number of device parameters and various circuit setups. Furthermore, the simple calculation also allows designers to estimate the second-order approximation and consider multiple parameters simultaneously. Extensive simulations using HSPICE

simulator under three different technologies have reaffirmed the accuracy of the method. It is hence very useful in evaluating the offset voltage, which is directly related to the performance and the yield of the SA.

CHAPTER 7 CONCLUSION AND FUTURE WORKS

7.1 Conclusion

The main focus of this thesis is to minimize the power consumption and to improve reliability (due to the process variations) of SRAM designs in nanometer technologies. To attain efficient power savings of the SRAM we proposed two novel designs of the SRAM cell topologies and two innovative designs of the current-mode SAs which are capable of operating at very low supply voltages and are very robust against excessive process variations.

Our first proposal is a 10T fully differential SRAM cell design with separated read- and write- ports to improve both WTP and SNM. Furthermore, multiple-threshold transistors are used to **reduce the leakage of the cell by 90%.** As a result, our design offers a very low leakage and low active power consumptions. Besides, extensive statistical simulations have shown that the proposed design has a more reliable read operation with **2.5x higher SNM** and **10% higher** I_{cell}/I_{off} ratio at 1V. At 0.4 V supply voltage, these improvements becomes **3x and 80%**, respectively. However, this is achieved at the expense of **33% active silicon area overhead** and **23% decrease of the WTP** as compared to the conventional 6T. Nevertheless, it successfully delivered a very low power consumption, which is the niche of this design and hence the above-mentioned drawbacks are more than compensated for by the power advantage.

To solve the half-access issues of the 10T SRAM cell, an 8T SRAM cell is introduced using a column-based dynamic cell supply scheme. Its cell supply is raised to a higher voltage to improve the **SNM during read by 2.5x**, whereas during the write operation, the cell supply is maintained at V_{DD} to obtain a **20% write-ability improvement**. An inverter is added to the conventional 6T structure to block the WL signal from activating the unselected cells on the same row. **This results in two unique properties of the proposed 8T cell: 1) It can be bit-interleaved for efficient conventional Error Checking Code**

implementation for soft-error correction. 2) Only one bit among the interleaved bits is activated thus the power consumption is reduced by 54%. Extensive cell analysis showed that the 8T design has better performance when compared to the other 10T and 6T designs, although its area is marginally larger than the conventional 6T cell. Due to its robustness, the cell is able to operate properly even at sub-threshold supply voltages. In view of the above, our 8T is the best alternative to the conventional 6T in ultra low-power ultra low-supply SRAM designs.

Apart from SRAM cell designs, we also investigated and designed a new current-mode SA to further enhance the performance and reduce the power consumption of the SRAM macro. More specifically, our first SA proposal resize the transistors in the current path of the conventional current-mode SA. This results in **8x higher BL read current utilization** and **53% power reduction**. Next, we proposed a novel cross-coupled SA topology that also has ultra low-power consumption. Its power consumption is only **70%** of that of the other cross-coupled SAs. This is achievable because the BL currents are automatically cut off at the end of the local sensing stage. Furthermore, it is proven to be **27% faster** and has a lower sensitivity to process variations when compared to other contemporary SA designs.

Finally, an analytical work has been dedicated to quantify the input-offset voltage of the broadly used cross-coupled (or latch-type) SA. This analysis is critical for this type of SA designs as an input below its offset value is irreversible and will eventually lead to a wrong sensing output. **Our findings infer that the SA's offset voltage is not only dependent on the intrinsic device parameter variations but also affected by the common-mode DC voltage.** The total offset voltage of the SA is found to consist of its intrinsic and extrinsic components. To be more precise, threshold voltage mismatch is the main contributor to the intrinsic input offset, whereas common-mode voltage and load capacitor mismatches contribute significantly to the extrinsic offset. This observation can be extended to derive and approximate the input offset voltage. Our proposed criterion is also useful for other latch-based SAs as it is simple enough to be reused easily. By

appropriately changing the IV characteristic equations that describe the respective SA topology, our criterion can be applied with ease to compute the corresponding offset voltage.

7.2 Future works

In this project, we have focused on the power consumption and the reliability of the SRAM by designing the SA as well as the memory cell topology at the circuit level. Our future works will extent the scope of research to the architecture level of SRAM design.

Firstly, we will look at the possibility of developing CMOS SRAM based on multi-valued logic, which may offer a more compact layout and hence both active area and leakage power can be reduced. In a conventional memory cell, logic value is stored in terms of node voltage. A node voltage at ground level represents a "0" while at V_{DD} it represents a "1". In multi-valued logic, a cell can store more than one bit (for example, two bits) but its total number of transistor count must be less than twelve. As a result, it consumes less area than two equivalent single-bit 6T cells. Several related issues however must be addressed such as cell reliability at very low supply voltage and the sensing circuits.

Another direction to implement the multi-valued memory cell is to assign logic values to more than two voltage levels. For example, 0, 1 and 2 are assigned to voltage level of 0, $V_{DD}/2$ and V_{DD}, respectively. Alternatively, current-mode memory cell designs should also be explored. While this research direction promises to offer many advantages such as higher density, lower power consumption it may suffer a severe noise margin reduction. As CMOS technology scales down to sub 32-nm regime, the process variations are excessively high and therefore noise margin and data retention reliability must be carefully investigated.

Secondly, we will shift our research focus to the 3D integration of SRAM cache and microprocessor: Normal MOSFET devices are used to implement the memory while their parasitic BJTs are used to implement the computational units, or vice versa. Unlike the

conventional 3D concept, our research aims to use only one layer of CMOS devices to implement both memory and microprocessor. Both bulk-CMOS or SOI CMOS process can be studied. As a result, the microprocessor and the memory can communicate with each other directly and the corresponding delay will be substantially reduced. The projected advantages are lower power consumption, smaller area and most importantly higher operating speed. The main challenge is to control both normal MOSFETs operations as well as their parasitic BJTs. Furthermore, memory devices are normally identical and have minimum size while those used in the microprocessors are not. Nevertheless, if this revolutional idea can be implemented it will have a huge impact on the contemporary electronic design and transform it to a higher level of integration.

Thirdly, we may consider applying the concept of Probabilistic CMOS (PCMOS) to memory. In the nano-scale technologies, noise and process variations are unavoidable. While conventional design approach tries to eliminate these, PCMOS approach utilizes of them and makes intelligent trade off between performance and yield or reliability. The main idea of PCMOS design is to compromise a small percentage of reliability to obtain a much bigger percentage of power saving. PCMOS however is a new field of CMOS circuit design and a lot of design factors must be put under considerations. If successful, it may offer a considerable amount of power saving in high the density SRAM designs in nano-scale technologies.

Author's publications

Journal papers

1. **A. T. Do**, S. Jeremy Low Yung, Z. H. Kong, and K. S. Yeo, " Design and Sensitivity Analysis of a New Current-mode Sense Amplifier for Low-power SRAM" Accepted for publication on *IEEE Trans. VLSI Systems.*
2. **A. T. Do**, Z. H. Kong, and K. S. Yeo, "Criterion to evaluate input-offset voltage of a Latch-Type SA," Accepted for publication on *IEEE Transactions on Circuits and Systems I: RegularPaper.*
3. **A. T. Do**, Z. H. Kong, and K. S. Yeo, "Hybrid-Mode SRAM SAs: New Approach on Transistor Sizing," *Circuits and Systems II: Express Briefs, IEEE Transactions on,* vol. 55, pp. 986-990, 2008.
4. **A. T. Do**, Z. H. Kong, and K. S. Yeo, "0.9 V current-mode SA using concurrent bit- and data-line tracking and sensing techniques," Electronics Letters, vol. 43, pp. 1421-1422, 2007.

Workshop and Conferences

1. Z. H. Kong and **A. T. Do** " A 16Kb 10T-SRAM with 4x Read-Power Reduction", accepted for presentation at *The 2010 IEEE International Symposium on Circuits and Systems*, Paris, France 30 May - 2 June, 2010.
2. H. Fu, K. S. Yeo, Z. H. Kong and **A. T. Do** "Design and performance evaluation of a silent data-line SRAM sense amplifier" accepted for presentation at *The 12th International Symposium on Integrated Circuits, ISIC 2009,* Singapore 14-16 Dec. 2009.
3. **A. T. Do**, Jeremy Y. S. Low, Z. H. Kong, K. S. Yeo, and L. Joshua Low Yung, "A full current-mode SA for low-power SRAM applications," in *Circuits and Systems, 2008. APCCAS 2008. IEEE Asia Pacific Conference on*, 2008, pp. 1402-1405.
4. Z. H. Kong, **A. T. Do**, and K. S. Yeo, "Design of a current SA using a current distribution technique," in Int. Ph.D Student Workshop on SoC, pp. 41-43, Jul. 2007.

Submitted

1. **A. T. Do**, K. S. Yeo, Jeremy Y. S. Low, Joshua Y. L. Low and Z. H. Kong "An 8T differential SRAM with improved noise margin for bit-interleaving in 65 nm CMOS" Submitted for peer review on *IEEE TCAS-I* on 20-Jul-2010.
2. **A. T. Do**, K. S. Yeo, Jeremy Y. S. Low, Joshua Y. L. Low and Z. H. Kong " An 8T SRAM Cell With Column-based Dynamic Supply Voltage for Bit-interleaving" Submitted to *APPCAS 2010*, Malaysia
3. **A. T. Do**, S. S. Chen, K. S. Yeo and Z. H. Kong "Low IR Drop and Low Power Parallel CAM Design Using Gated Power Transistor Technique" Submitted to *APPCAS 2010*, Malaysia

Bibliography

[1] N. Verma and A. P. Chandrakasan, "A 256 kb 65 nm 8T Subthreshold SRAM Employing Sense-Amplifier Redundancy," *Solid-State Circuits, IEEE Journal of,* vol. 43, pp. 141-149, 2008.

[2] L. Zhiyu and V. Kursun, "Characterization of a Novel Nine-Transistor SRAM Cell," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on,* vol. 16, pp. 488-492, 2008.

[3] S. Okumura, *et al.*, "A 0.56-V 128kb 10T SRAM using column line assist (CLA) scheme," in *Quality of Electronic Design, 2009. ISQED 2009. Quality Electronic Design*, 2009, pp. 659-663.

[4] S. M. Jahinuzzaman, *et al.*, "A Soft Error Tolerant 10T SRAM Bit-Cell With Differential Read Capability," *Nuclear Science, IEEE Transactions on,* vol. 56, pp. 3768-3773, 2009.

[5] A. Seshadri and T. W. Houston, "The dynamic stability of a 10T SRAM compared to 6T SRAMs at the 32nm node using an accelerated Monte Carlo technique," in *Circuits and Systems Workshop: System-on-Chip - Design, Applications, Integration, and Software, 2008 IEEE Dallas*, 2008, pp. 1-4.

[6] H. Noguchi, *et al.*, "A 10T Non-Precharge Two-Port SRAM for 74% Power Reduction in Video Processing," in *VLSI, 2007. ISVLSI '07. IEEE Computer Society Annual Symposium on*, 2007, pp. 107-112.

[7] F. Moradi, *et al.*, "65NM sub-threshold 11T-SRAM for ultra low voltage applications," in *SOC Conference, 2008 IEEE International*, 2008, pp. 113-118.

[8] R. Tanabe, *et al.*, "Investigation of SNM with Random Dopant Fluctuations for FD SGSOI and FinFET 6T SOI SRAM Cell by Three-dimensional Device Simulation," in *Simulation of Semiconductor Processes and Devices, 2006 International Conference on*, 2006, pp. 103-106.

[9] L. Oniciuc and P. Andrei, "Sensitivity of static noise margins to random doping variations in 6T SRAM cells," in *Semiconductor Device Research Symposium, 2007 International*, 2007, pp. 1-2.

[10] R. Keerthi and C. i. H. Chen, "Stability and Static Noise Margin Analysis of Low-Power SRAM," in *Instrumentation and Measurement Technology Conference Proceedings, 2008. IMTC 2008. IEEE*, 2008, pp. 1681-1684.

[11] K. Itoh, *et al.*, "Memory at VLSI Circuits Symposium," *Solid-State Circuits, IEEE Journal of,* vol. 43, pp. 762-768, 2008.

[12] K. Itoh, "Low-voltage limitations and challenges of nano-scale CMOS VLSIs - A personal view of memory designer," in *Integrated Circuit Design and Technology and Tutorial, 2008. ICICDT 2008. IEEE International Conference on*, 2008, pp. 177-180.

[13] A. Wang, *et al.*, *Sub-threshold design for ultra low-power systems*. New York: Springer Science + Business Media, 2006.

[14] R. Kanj, *et al.*, "Statistical Evaluation of Split Gate Opportunities for Improved 8T/6T Column-Decoupled SRAM Cell Yield," in *Quality Electronic Design, 2008. ISQED 2008. 9th International Symposium on*, 2008, pp. 702-707.

[15] N. Hyunwoo, *et al.*, "Numerical Estimation of Yield in Sub-100-nm SRAM Design Using Monte Carlo Simulation," *Circuits and Systems II: Express Briefs, IEEE Transactions on,* vol. 55, pp. 907-911, 2008.

[16] R. M. Houle, "Simple Statistical Analysis Techniques to Determine Optimum Sense Amp Set Times," *Solid-State Circuits, IEEE Journal of,* vol. 43, pp. 1816-1825, 2008.

[17] B. Cheng, *et al.*, "Statistical variations in 32nm thin-body SOI devices and SRAM cells," in *Solid-State and Integrated-Circuit Technology, 2008. ICSICT 2008. 9th International Conference on*, 2008, pp. 389-392.

[18] M. H. Abu-Rahma, *et al.*, "A methodology for statistical estimation of read access yield in SRAMs," in *Design Automation Conference, 2008. DAC 2008. 45th ACM/IEEE*, 2008, pp. 205-210.

[19] E. Grossar, "Technolgy-aware design of SRAM memory circuits," PhD, Departement of Electronic, Katholieke Universiteit Leuven, 2007.

[20] W. Jiajing, *et al.*, "Statistical modeling for the minimum standby supply voltage of a full SRAM array," in *Solid State Circuits Conference, 2007. ESSCIRC 2007. 33rd European*, 2007, pp. 400-403.

[21] B. Giraud, *et al.*, "A Comparative Study of 6T and 4T SRAM Cells in Double-Gate CMOS with Statistical Variation," in *Circuits and Systems, 2007. ISCAS 2007. IEEE International Symposium on*, 2007, pp. 3022-3025.

[22] L. Zhiyu, *et al.*, "Statistical Data Stability and Leakage Evaluation of FinFET SRAM Cells with Dynamic Threshold Voltage Tuning under Process Parameter Fluctuations," in *Quality Electronic Design, 2008. ISQED 2008. 9th International Symposium on*, 2008, pp. 305-310.

[23] J. Singh, *et al.*, "Failure analysis for ultra low power nano-CMOS SRAM under process variations," in *SOC Conference, 2008 IEEE International*, 2008, pp. 251-254.

[24] T. S. Doorn, *et al.*, "Importance sampling Monte Carlo simulations for accurate estimation of SRAM yield," in *Solid-State Circuits Conference, 2008. ESSCIRC 2008. 34th European*, 2008, pp. 230-233.

[25] Q. Huifang, "Deep Sub-micron SRAM design for ultra-low leakage standby operation," PhD, Electrical Engineering and Computer Sciences, University of California, Berkeley, 2007.

[26] K. Itoh, *et al.*, *Ultra-low voltage nano-scale memories*. New York: Springer Science+Business Media, LLC, 2007.

[27] K. Nii, *et al.*, "A 90 nm low power 32 K-byte embedded SRAM with gate leakage suppression circuit for mobile applications," in *VLSI Circuits, 2003. Digest of Technical Papers. 2003 Symposium on*, 2003, pp. 247-250.

[28] A. Chandrakasan, *et al.*, *Digital integrated circuits: A design perspective*, 2nd ed. Upper Saddle River, New Jersey: Person Education, Inc, 2002.

[29] R. J. Baker, *CMOS: Circuit design, layout and simulation*. Hoboken, New Jersy: John Wiley &Son, Inc, 2007.

[30] F. Hamzaoglu, *et al.*, "A 3.8 GHz 153 Mb SRAM Design With Dynamic Stability Enhancement and Leakage Reduction in 45 nm High-k Metal Gate CMOS Technology," *Solid-State Circuits, IEEE Journal of,* vol. 44, pp. 148-154, 2009.

[31] V. Ramadurai, *et al.*, "An 8 Mb SRAM in 45 nm SOI Featuring a Two-Stage Sensing Scheme and Dynamic Power Management," *Solid-State Circuits, IEEE Journal of,* vol. 44, pp. 155-162, 2009.

[32] T. P. Haraszti, *CMOS memory circuits*. New York, NY 10036 Springer, 2000.

[33] K. Zhang, *et al.*, "Low-Power SRAMs in Nanoscale CMOS Technologies," *Electron Devices, IEEE Transactions on,* vol. 55, pp. 145-151, 2008.

[34] Y. H. Suh, *et al.*, "A 256MB synchronous-burst DDR SRAM with hierarchical bit-line architecture for mobile applications," in *Solid-State Circuits Conference, 2005. Digest of Technical Papers. ISSCC. 2005 IEEE International*, 2005, pp. 476-611 Vol. 1.

[35] Y. Byung-Do and K. Lee-Sup, "A low-power SRAM using hierarchical bit line and local sense amplifiers," *Solid-State Circuits, IEEE Journal of,* vol. 40, pp. 1366-1376, 2005.

[36] S. Ishikura, *et al.*, "A 45 nm 2-port 8T-SRAM Using Hierarchical Replica Bitline Technique With Immunity From Simultaneous R/W Access Issues," *Solid-State Circuits, IEEE Journal of,* vol. 43, pp. 938-945, 2008.

[37] Z. H. Kong, "CMOS VLSI subsystems for low-voltage low-power applications " PhD, Department of Circuits and Systems, Nanyang Technological University, Singapore, 2006.

[38] L. Yo-Sheng, *et al.*, "Leakage scaling in deep submicron CMOS for SoC," *Electron Devices, IEEE Transactions on,* vol. 49, pp. 1034-1041, 2002.

[39] S. Ishikura, *et al.*, "A 45nm 2port 8T-SRAM using hierarchical replica bitline technique with immunity from simultaneous R/W access issues," in *VLSI Circuits, 2007 IEEE Symposium on*, 2007, pp. 254-255.

[40] A. M. Fahim, *et al.*, "A low-power high-performance embedded SRAM macrocell," in *VLSI, 1998. Proceedings of the 8th Great Lakes Symposium on*, 1998, pp. 13-18.

[41] T. Hirose, *et al.*, "A 20 ns 4 Mb CMOS SRAM with hierarchical word decoding architecture," in *Solid-State Circuits Conference, 1990. Digest of Technical Papers. 37th ISSCC., 1990 IEEE International*, 1990, pp. 132-133.

[42] T. Hirose, *et al.*, "A 20-ns 4-Mb CMOS SRAM with hierarchical word decoding architecture," *Solid-State Circuits, IEEE Journal of,* vol. 25, pp. 1068-1074, 1990.

[43] A. Bellaouar and M. I. Elmasry, *Low-power digital VLSI design: Circuits and Systems*: The Netherlands: Kluwer Academic Publishers, 1995.

[44] K. S. Yeo and K. Roy, *Low Voltage, Low Power VLSI Subsystems*: McGraw-Hill Professional, 2005.

[45] Z. Zhu, *et al.*, "Low power bank-based multi-port SRAM design due to bank standby mode," in *Circuits and Systems, 2004. MWSCAS '04. The 2004 47th Midwest Symposium on*, 2004, pp. I-569-72 vol.1.

[46] E. Morifuji, *et al.*, "Supply and threshold-Voltage trends for scaled logic and SRAM MOSFETs," *Electron Devices, IEEE Transactions on,* vol. 53, pp. 1427-1432, 2006.

[47] N. Shibata, *et al.*, "A 0.5-V 25-MHz 1-mW 256-kb MTCMOS/SOI SRAM for solar-power-operated portable personal digital equipment - sure write operation by using step-down negatively overdriven bitline scheme," *Solid-State Circuits, IEEE Journal of,* vol. 41, pp. 728-742, 2006.

[48] J. P. Kulkarni, *et al.*, "A 160 mV Robust Schmitt Trigger Based Subthreshold SRAM," *Solid-State Circuits, IEEE Journal of,* vol. 42, pp. 2303-2313, 2007.

[49] E. Morifuji, *et al.*, "Power Optimization for SRAM and Its Scaling," *Electron Devices, IEEE Transactions on,* vol. 54, pp. 715-722, 2007.

[50] N. Shibata, *et al.*, "1-V 100-MHz embedded SRAM techniques for battery-operated MTCMOS/SIMOX ASICs," *Solid-State Circuits, IEEE Journal of,* vol. 35, pp. 1396-1407, 2000.

[51] N. Shibata, *et al.*, "A 1-V, 10-MHz, 3.5-mW, 1-Mb MTCMOS SRAM: with charge-recycling input/output buffers," *Solid-State Circuits, IEEE Journal of,* vol. 34, pp. 866-877, 1999.

[52] K. Tae-Hyoung, *et al.*, "A 0.2 V, 480 kb Subthreshold SRAM With 1 k Cells Per Bitline for Ultra-Low-Voltage Computing," *Solid-State Circuits, IEEE Journal of,* vol. 43, pp. 518-529, 2008.

[53] K. Itoh, "Low-voltage limitations and challenges of memory-rich nano-scale CMOS VLSIs," in *Solid-State and Integrated-Circuit Technology, 2008. ICSICT 2008. 9th International Conference on*, 2008, pp. 2111-2114.

[54] A. Kumar, *et al.*, "Fundamental Data Retention Limits in SRAM Standby Experimental Results," in *Quality Electronic Design, 2008. ISQED 2008. 9th International Symposium on*, 2008, pp. 92-97.

[55] S. A. Tawfik and V. Kursun, "Dynamic wordline voltage swing for low leakage and stable static memory banks," in *Circuits and Systems, 2008. ISCAS 2008. IEEE International Symposium on*, 2008, pp. 1894-1897.

[56] M. Shareef, *et al.*, "Energy Reduction in SRAM using Dynamic Voltage and Frequency Management," in *VLSI Design, 2008. VLSID 2008. 21st International Conference on*, 2008, pp. 503-508.

[57] S. Huang, *et al.*, "A novel SRAM structure for leakage power suppression in 45nm technology," in *Communications, Circuits and Systems, 2008. ICCCAS 2008. International Conference on*, 2008, pp. 1070-1074.

[58] G. Fukano, *et al.*, "A 65nm 1Mb SRAM Macro with Dynamic Voltage Scaling in Dual Power Supply Scheme for Low Power SoCs," in *Non-Volatile Semiconductor Memory Workshop, 2008 and 2008 International Conference on Memory Technology and Design. NVSMW/ICMTD 2008. Joint*, 2008, pp. 97-98.

[59] A. Bhavnagarwala, *et al.*, "A Sub-600mV, Fluctuation tolerant 65nm CMOS SRAM Array with Dynamic Cell Biasing," in *VLSI Circuits, 2007 IEEE Symposium on*, 2007, pp. 78-79.

[60] M. Khellah, *et al.*, "A 4.2GHz 0.3mm2 256kb Dual-Vcc SRAM Building Block in 65nm CMOS," in *Solid-State Circuits Conference, 2006. ISSCC 2006. Digest of Technical Papers. IEEE International*, 2006, pp. 2572-2581.

[61] J. Rajiv, *et al.*, "Statistical Exploration of the Dual Supply Voltage Space of a 65nm PD/SOI CMOS SRAM Cell," in *Solid-State Device Research Conference, 2006. ESSDERC 2006. Proceeding of the 36th European*, 2006, pp. 315-318.

[62] M. Khellah, *et al.*, "A 256-Kb Dual-VCCSRAM Building Block in 65-nm CMOS Process With Actively Clamped Sleep Transistor," *Solid-State Circuits, IEEE Journal of,* vol. 42, pp. 233-242, 2007.

[63] K. Zhang, *et al.*, "SRAM design on 65nm CMOS technology with integrated leakage reduction scheme," in *VLSI Circuits, 2004. Digest of Technical Papers. 2004 Symposium on*, 2004, pp. 294-295.

[64] J. B. Kuang, *et al.*, "A low-overhead virtual rail technique for SRAM leakage power reduction," in *Computer Design: VLSI in Computers and Processors, 2005. ICCD 2005. Proceedings. 2005 IEEE International Conference on*, 2005, pp. 574-579.

[65] K. Zhang, *et al.*, "SRAM design on 65-nm CMOS technology with dynamic sleep transistor for leakage reduction," *Solid-State Circuits, IEEE Journal of,* vol. 40, pp. 895-901, 2005.

[66] K. Ding-Ming, "Standby Current Reduction of Compilable SRAM Using Sleep Transistor and Source Line Self Bias," in *Solid-State Circuits Conference, 2006. ASSCC 2006. IEEE Asian*, 2006, pp. 23-26.

[67] A. Nourivand, *et al.*, "An Adaptive Sleep Transistor Biasing Scheme for Low Leakage SRAM," in *Circuits and Systems, 2007. ISCAS 2007. IEEE International Symposium on*, 2007, pp. 2790-2793.

[68] K. X. Zhang, "K-2 High-performance and low-Power SRAMs design in nano-scale CMOS technology," in *ASIC, 2007. ASICON '07. 7th International Conference on*, 2007, pp. 2-2.

[69] A. Agarwal, *et al.*, "A single-Vt low-leakage gated-ground cache for deep submicron," *Solid-State Circuits, IEEE Journal of,* vol. 38, pp. 319-328, 2003.

[70] A. J. Bhavnagarwala, *et al.*, "A pico-joule class, 1 GHz, 32 KByte×64 b DSP SRAM with self reverse bias," in *VLSI Circuits, 2003. Digest of Technical Papers. 2003 Symposium on*, 2003, pp. 251-252.

[71] M. Margala, "Low-power SRAM circuit design," in *Memory Technology, Design and Testing, 1999. Records of the 1999 IEEE International Workshop on*, 1999, pp. 115-122.

[72] K. W. Mai, *et al.*, "Low-power SRAM design using half-swing pulse-mode techniques," *Solid-State Circuits, IEEE Journal of,* vol. 33, pp. 1659-1671, 1998.

[73] E. A. John, *Digital Integrated Circuits: Analysis and Design*: CRC Press, 2004.

[74] W. Chua-Chin, *et al.*, "A 4-KB 500-MHz 4-T CMOS SRAM using low-Vthn bitline drivers and high-Vthp latches," in *ASIC, 2002. Proceedings. 2002 IEEE Asia-Pacific Conference on*, 2002, pp. 49-52.

[75] W. Chua-Chin, *et al.*, "6-T SRAM using dual threshold voltage transistors and low-power quenchers," in *Electronics, Circuits and Systems, 2002. 9th International Conference on*, 2002, pp. 827-830 vol.2.

[76] S. Singh, *et al.*, "Architecture and design of a high performance SRAM for SOC design," in *Design Automation Conference, 2002. Proceedings of ASP-DAC 2002.*

7th Asia and South Pacific and the 15th International Conference on VLSI Design. Proceedings., 2002, pp. 447-451.

[77] W. Chua-Chin, *et al.*, "An SRAM design using dual threshold voltage transistors and low-power quenchers," *Solid-State Circuits, IEEE Journal of,* vol. 38, pp. 1712-1720, 2003.

[78] P. Elakkumanan, *et al.*, "NC-SRAM - a low-leakage memory circuit for ultra deep submicron designs," in *SOC Conference, 2003. Proceedings. IEEE International [Systems-on-Chip]*, 2003, pp. 3-6.

[79] Y. Ye, *et al.*, "A 6-GHz 16-kB L1 cache in a 100-nm dual-Vt technology using a bitline leakage reduction (BLR) technique," *Solid-State Circuits, IEEE Journal of,* vol. 38, pp. 839-842, 2003.

[80] W. Chua-Chin, *et al.*, "A 4-kB 500-MHz 4-T CMOS SRAM using low-Vthn bitline drivers and high-Vthp latches," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on,* vol. 12, pp. 901-909, 2004.

[81] M. Mamidipaka, *et al.*, "Analytical models for leakage power estimation of memory array structures," in *Hardware/Software Codesign and System Synthesis, 2004. CODES + ISSS 2004. International Conference on*, 2004, pp. 146-151.

[82] B. Amelifard, *et al.*, "Low-leakage SRAM design with dual V/sub t/ transistors," in *Quality Electronic Design, 2006. ISQED '06. 7th International Symposium on*, 2006, pp. 6 pp.-734.

[83] L. Jungseob and A. Davoodi, "Comparison of Dual-Vt Configurations of SRAM Cell Considering Process-Induced Vt Variations," in *Circuits and Systems, 2007. ISCAS 2007. IEEE International Symposium on*, 2007, pp. 3018-3021.

[84] S. A. Tawfik and V. Kursun, "Low power and robust 7T dual-Vt SRAM circuit," in *Circuits and Systems, 2008. ISCAS 2008. IEEE International Symposium on*, 2008, pp. 1452-1455.

[85] T. Suzuki, *et al.*, "A Stable 2-Port SRAM Cell Design Against Simultaneously Read/Write-Disturbed Accesses," *Solid-State Circuits, IEEE Journal of,* vol. 43, pp. 2109-2119, 2008.

[86] I. Y. Hao, *et al.*, "Low-power floating bitline 8-T SRAM design with write assistant circuits," in *SOC Conference, 2008 IEEE International*, 2008, pp. 239-242.

[87] M. Yabuuchi, *et al.*, "A 45nm Low-Standby-Power Embedded SRAM with Improved Immunity Against Process and Temperature Variations," in *Solid-State Circuits Conference, 2007. ISSCC 2007. Digest of Technical Papers. IEEE International*, 2007, pp. 326-606.

[88] H. Pilo, *et al.*, "An SRAM Design in 65-nm Technology Node Featuring Read and Write-Assist Circuits to Expand Operating Voltage," *Solid-State Circuits, IEEE Journal of,* vol. 42, pp. 813-819, 2007.

[89] R. F. Hobson, "A New Single-Ended SRAM Cell With Write-Assist," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on,* vol. 15, pp. 173-181, 2007.

[90] H. Pilo, *et al.*, "An SRAM Design in 65nm and 45nm Technology Nodes Featuring Read and Write-Assist Circuits to Expand Operating Voltage," in *VLSI Circuits, 2006. Digest of Technical Papers. 2006 Symposium on*, 2006, pp. 15-16.

[91] K. Kanda, *et al.*, "90% write power-saving SRAM using sense-amplifying memory cell," *Solid-State Circuits, IEEE Journal of,* vol. 39, pp. 927-933, 2004.

[92] S. Hattori and T. Sakurai, "90% write power saving SRAM using sense-amplifying memory cell," in *VLSI Circuits Digest of Technical Papers, 2002. Symposium on*, 2002, pp. 46-47.

[93] A. Sil, *et al.*, "High speed single-ended pseudo differential current sense amplifier for SRAM cell," in *Circuits and Systems, 2008. ISCAS 2008. IEEE International Symposium on*, 2008, pp. 3330-3333.

[94] K. Kushida, *et al.*, "A 0.7V single-supply SRAM with 0.495um2 cell in 65nm technology utilizing self-write-back sense amplifier and cascaded bit line scheme," in *VLSI Circuits, 2008 IEEE Symposium on*, 2008, pp. 46-47.

[95] K. Itoh, "Low-voltage limitations and challenges of nano-scale CMOS LSIs - A personal view of memory designer," in *Integrated Circuit Design and Technology and Tutorial, 2008. ICICDT 2008. IEEE International Conference on*, 2008, pp. 177-180.

[96] A.-T. Do, *et al.*, "Hybrid-Mode SRAM Sense Amplifiers: New Approach on Transistor Sizing," *Circuits and Systems II: Express Briefs, IEEE Transactions on*, vol. 55, pp. 986-990, 2008.

[97] A.-T. Do, *et al.*, "A full current-mode sense amplifier for low-power SRAM applications," in *Circuits and Systems, 2008. APCCAS 2008. IEEE Asia Pacific Conference on*, 2008, pp. 1402-1405.

[98] S. Sundaram, *et al.*, "High speed robust current sense amplifier for nanoscale memories: a winner take all approach," in *VLSI Design, 2006. Held jointly with 5th International Conference on Embedded Systems and Design., 19th International Conference on*, 2006, pp. pp.569 - 574

[99] O. Thomas, *et al.*, "Ultra-low-voltage current-sense read circuits for CMOS SOI SRAMs," in *SOI Conference, 2005. Proceedings. 2005 IEEE International*, 2005, pp. 205-207.

[100] P. Tao, "How much mismatch should be simulated in the high density SRAM sense amplifier design," in *Reliability Physics Symposium, 2005. Proceedings. 43rd Annual. 2005 IEEE International*, 2005, pp. 672-673.

[101] M. Golden, *et al.*, "Sense amp design in SOI," in *SOI Conference, 2005. Proceedings. 2005 IEEE International*, 2005, pp. 118-120.

[102] S. Ardalan, *et al.*, "Current mode sense amplifier," in *Circuits and Systems, 2005. 48th Midwest Symposium on*, 2005, pp. 17-20 Vol. 1.

[103] R. E. Aly, *et al.*, "Dual sense amplified bit lines (DSABL) architecture for low-power SRAM design," in *Circuits and Systems, 2005. ISCAS 2005. IEEE International Symposium on*, 2005, pp. 1650-1653 Vol. 2.

[104] C. Hwang-Cherng and C. Shu-Hsien, "High performance sense amplifier circuit for low power SRAM applications," in *Circuits and Systems, 2004. ISCAS '04. Proceedings of the 2004 International Symposium on*, 2004, pp. II-741-4 Vol.2.

[105] M. Sinha, *et al.*, "High-performance and low-voltage sense-amplifier techniques for sub-90nm SRAM," in *SOC Conference, 2003. Proceedings. IEEE International [Systems-on-Chip]*, 2003, pp. 113-116.

[106] S. M. Wang and C. X. Wu, "Full current-mode techniques for high-speed CMOS SRAMs," in *Circuits and Systems, 2002. ISCAS 2002. IEEE International Symposium on*, 2002, pp. IV-580-IV-582 vol.4.

[107] A. Chrisanthopoulos, *et al.*, "SRAM oriented memory sense amplifier design in 0.18 μm CMOS technology," in *Circuits and Systems, 2002. ISCAS 2002. IEEE International Symposium on*, 2002, pp. V-145-V-148 vol.5.

[108] B. Wicht, *et al.*, "A simple low voltage current sense amplifier with switchable input transistor," in *Solid-State Circuits Conference, 2001. ESSCIRC 2001. Proceedings of the 27th European*, 2001, pp. 285-288.

[109] Y. Tsiatouhas, *et al.*, "New memory sense amplifier designs in CMOS technology," in *Electronics, Circuits and Systems, 2000. ICECS 2000. The 7th IEEE International Conference on*, 2000, pp. 19-22 vol.1.

[110] W. Jinn-Shyan and L. Hong-Yu, "A new current-mode sense amplifier for low-voltage low-power SRAM," in *ASIC Conference 1998. Proceedings. Eleventh Annual IEEE International*, 1998, pp. 163-167.

[111] M. Izumikawa and M. Yamashina, "A current direction sense technique for multi-port SRAMs," in *VLSI Circuits, 1995. Digest of Technical Papers., 1995 Symposium on*, 1995, pp. 23-24.

[112] T. Kobayashi, *et al.*, "A current-mode latch sense amplifier and a static power saving input buffer for low-power architecture," in *VLSI Circuits, 1992. Digest of Technical Papers., 1992 Symposium on*, 1992, pp. 28-29.

[113] S. K. Jain and P. Agarwal, "A low leakage and SNM free SRAM cell design in deep sub micron CMOS technology," in *VLSI Design, 2006. Held jointly with 5th International Conference on Embedded Systems and Design., 19th International Conference on*, 2006, p. 4

[114] N. Azizi, *et al.*, "Low-leakage asymmetric-cell SRAM," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on,* vol. 11, pp. 701-715, 2003.

[115] N. Azizi, *et al.*, "Low-leakage asymmetric-cell SRAM," in *Low Power Electronics and Design, 2002. ISLPED '02. Proceedings of the 2002 International Symposium on*, 2002, pp. 48-51.

[116] M. Wieckowski, *et al.*, "Portless SRAM-A High-Performance Alternative to the 6T Methodology," *Solid-State Circuits, IEEE Journal of,* vol. 42, pp. 2600-2610, 2007.

[117] B. Amelifard, *et al.*, "Leakage Minimization of SRAM Cells in a Dual-Vt and Dual-Tox Technology," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on,* vol. 16, pp. 851-860, 2008.

[118] C. Qikai, *et al.*, "Modeling and testing of SRAM for new failure mechanisms due to process variations in nanoscale CMOS," in *VLSI Test Symposium, 2005. Proceedings. 23rd IEEE*, 2005, pp. 292-297.

[119] E. Seevinck, *et al.*, "Static-noise margin analysis of MOS SRAM cells," *Solid-State Circuits, IEEE Journal of,* vol. 22, pp. 748-754, 1987.

[120] Standard Test Procedure for Noise Margin Measurements for Semiconductor Logic Gating Microcircuits [Online]. Available: http://www.jedec.org/download/search/JESD390a.pdf

[121] C. F. Hill, "Noise margin and noise immunity in logic circuits," *Micoelectron,* vol. 1, pp. 16-21, 1968.

[122] J. Lohstroh, "Static and dynamic noise margins of logic circuits," *Solid-State Circuits, IEEE Journal of,* vol. SC-14, 1979.

[123] J. Lohstroh, *et al.*, "Worst-case static noise margin criteria for logic circuits and their mathematical equivalent," *Solid-State Circuits, IEEE Journal of,* vol. Sc-18, pp. 803-807, 1983.

[124] E. Seevinck, "Application of the translinear principle in digital circuits," *Solid-State Circuits, IEEE Journal of,* vol. SC-13, pp. 528-530 1978.

[125] J. Lohstroh, "Calculation method to obtain worst-case static noise margins of logic circuits," *Electronics Letters,* vol. 16, pp. 273-274, Apr. 1980.

[126] J. Lohstroh, "The punch-through device as a passive exponential load in fast static bipolar RAM cells," *Solid-State Circuits, IEEE Journal of,* vol. SC-14, pp. 840-844, Oct. 1979.

[127] E. Seevinck, "Deriving stability criteria for non-linear circuits with application to worst-case noise margin for I2L," *Electronics Letters,* vol. 16, pp. 867-869, Nov. 1980.

[128] A. J. Bhavnagarwala, *et al.*, "The impact of intrinsic device fluctuations on CMOS SRAM cell stability," *Solid-State Circuits, IEEE Journal of,* vol. 36, pp. 658-665, 2001.

[129] E. Grossar, *et al.*, "Read Stability and Write-Ability Analysis of SRAM Cells for Nanometer Technologies," *Solid-State Circuits, IEEE Journal of,* vol. 41, pp. 2577-2588, 2006.

[130] T. Ichikawa and M. Sasaki, "A new analytical model of SRAM cell stability in low-voltage operation," *Electron Devices, IEEE Transactions on,* vol. 43, pp. 54-61, 1996.

[131] T. Sakurai and A. R. Newton, "Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulus," *Solid-State Circuits, IEEE Journal of,* vol. SCC-25, pp. 584-594, Apr. 1990.

[132] R. Heald and P. Wang, "Variability in sub-100nm SRAM designs," in *Computer Aided Design, 2004. ICCAD-2004. IEEE/ACM International Conference on*, 2004, pp. 347-352.

[133] M. Sharifkhani, *et al.*, "Dynamic Data Stability in Low-power SRAM Design," in *Custom Integrated Circuits Conference, 2007. CICC '07. IEEE*, 2007, pp. 237-240.

[134] D. E. Khalil, *et al.*, "Accurate Estimation of SRAM Dynamic Stability," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on,* vol. 16, pp. 1639-1647, 2008.

[135] W. Dong, *et al.*, "SRAM dynamic stability: Theory, variability and analysis," in *Computer-Aided Design, 2008. ICCAD 2008. IEEE/ACM International Conference on*, 2008, pp. 378-385.

[136] C. Wann, *et al.*, "SRAM cell design for stability methodology," in *VLSI Technology, 2005. (VLSI-TSA-Tech). 2005 IEEE VLSI-TSA International Symposium on*, 2005, pp. 21-22.

[137] M. Samson and M. B. Srinivas, "Analyzing N-Curve Metrics for Sub-Threshold 65nm CMOS SRAM," in *Nanotechnology, 2008. NANO '08. 8th IEEE Conference on*, 2008, pp. 25-28.

[138] C. Lage, *et al.*, "Advanced SRAM technology-the race between 4T and 6T cells," in *Electron Devices Meeting, 1996., International*, 1996, pp. 271-274.

[139] O. Kudoh, *et al.*, "A reduced size CMOS SRAM cell structure with two-level Al interconnection," in *Electron Devices Meeting, 1982 International*, 1982, pp. 474-477.

[140] C. E. Chen, *et al.*, "Stacked CMOS SRAM cell," *Electron Device Letters, IEEE,* vol. 4, pp. 272-274, 1983.

[141] R. Sundaresan, *et al.*, "A fully self-aligned stacked CMOS 64K SRAM," in *Electron Devices Meeting, 1984 International*, 1984, pp. 871-873.

[142] N. Okazaki, *et al.*, "A 16 ns 2K×8 bit full CMOS SRAM," *Solid-State Circuits, IEEE Journal of,* vol. 19, pp. 552-556, 1984.

[143] J. Miyamoto, *et al.*, "A 28ns CMOS SRAM with bipolar sense amplifiers," in *Solid-State Circuits Conference. Digest of Technical Papers. 1984 IEEE International*, 1984, pp. 224-225.

[144] O. Kudoh, *et al.*, "A new full CMOS SRAM cell structure," in *Electron Devices Meeting, 1984 International*, 1984, pp. 67-70.

[145] M. Isobe, *et al.*, "A 46ns 256K CMOS SRAM," in *Solid-State Circuits Conference. Digest of Technical Papers. 1984 IEEE International*, 1984, pp. 214-215.

[146] F. Bauer, *et al.*, "Layout options for stability tuning of SRAM cells in multi-gate-FET technologies," in *Solid State Circuits Conference, 2007. ESSCIRC 2007. 33rd European*, 2007, pp. 392-395.

[147] W. Dehaene, *et al.*, "Embedded SRAM design in deep deep submicron technologies," in *Solid State Circuits Conference, 2007. ESSCIRC 2007. 33rd European*, 2007, pp. 384-391.

[148] K. Johguchi, *et al.*, "A 0.6-Tbps, 16-port SRAM design with 2-stage- pipeline and multi-stage-sensing scheme," in *Solid State Circuits Conference, 2007. ESSCIRC 2007. 33rd European*, 2007, pp. 320-323.

[149] M. Yamaoka and T. Kawahara, "Operating-margin-improved SRAM with column-at-a-time body-bias control technique," in *Solid State Circuits Conference, 2007. ESSCIRC 2007. 33rd European*, 2007, pp. 396-399.

[150] B. Cheng, *et al.*, "The scalability of 8T-SRAM cells under the influence of intrinsic parameter fluctuations," in *Solid State Device Research Conference, 2007. ESSDERC 2007. 37th European*, 2007, pp. 93-96.

[151] T. Sugizaki, *et al.*, "Advantages of bulk over SOI in performance of thyristor-based SRAM cell with selective epitaxy anode," in *Solid State Device Research Conference, 2007. ESSDERC 2007. 37th European*, 2007, pp. 323-326.

[152] B. A. Chen, *et al.*, "0.25 μm low power CMOS devices and circuits from 8 inch SOI materials," in *Solid-State and Integrated Circuit Technology, 1995 4th International Conference on*, 1995, pp. 260-262.

[153] B. Cheng, *et al.*, "Impact of Intrinsic Parameter Fluctuations on SRAM Cell Design," in *Solid-State and Integrated Circuit Technology, 2006. ICSICT '06. 8th International Conference on*, 2006, pp. 1290-1292.

[154] Z. Feng, *et al.*, "Leakage Power Modeling Method for SRAM Considering Temperature, Supply Voltage and Bias Voltage," in *Solid-State and Integrated Circuit Technology, 2006. ICSICT '06. 8th International Conference on*, 2006, pp. 1180-1182.

[155] B. L. Ji, *et al.*, "On the Connection of SRAM Cell Stability with Switching History in Partially Depleted SOI Technology," in *Solid-State and Integrated Circuit Technology, 2006. ICSICT '06. 8th International Conference on*, 2006, pp. 788-791.

[156] Y. Jun-Jun and W. Peng-Jun, "Design of Adiabatic SRAM Based on CTGAL Circuit," in *Solid-State and Integrated Circuit Technology, 2006. ICSICT '06. 8th International Conference on*, 2006, pp. 2118-2120.

[157] S. V. Kosonocky, *et al.*, "Scalability options for future SRAM memories," in *Solid-State and Integrated Circuit Technology, 2006. ICSICT '06. 8th International Conference on*, 2006, pp. 689-692.

[158] A. Kumar, "Instabilities in deep submicron SRAM," in *Solid-State and Integrated Circuit Technology, 2006. ICSICT '06. 8th International Conference on*, 2006, pp. 792-795.

[159] H. Li, *et al.*, "Design of a Low Power Radiation Hardened 256K SRAM," in *Solid-State and Integrated Circuit Technology, 2006. ICSICT '06. 8th International Conference on*, 2006, pp. 1646-1648.

[160] Z. Zhitao and Z. Ganggang, "The Fast Simulation Model of SRAM," in *Solid-State and Integrated Circuit Technology, 2006. ICSICT '06. 8th International Conference on*, 2006, pp. 1333-1335.

[161] S. Natarajan, *et al.*, "Deep submicron embedded SRAM design issues," in *Solid-State and Integrated Circuits Technology, 2004. Proceedings. 7th International Conference on*, 2004, pp. 723-728 vol.1.

[162] P. Tao, *et al.*, "A design based yield and redundancy model for high density dualport SRAM on 90nm technology," in *Solid-State and Integrated Circuits Technology, 2004. Proceedings. 7th International Conference on*, 2004, pp. 729-731 vol.1.

[163] R. Wong, *et al.*, "Design and modeling of tapered LWL architecture for high density SRAM," in *Solid-State and Integrated Circuits Technology, 2004. Proceedings. 7th International Conference on*, 2004, pp. 732-734 vol.1.

[164] J. Wu, "CMOS transistor design challenges for mobile and digital consumer applications," in *Solid-State and Integrated Circuits Technology, 2004. Proceedings. 7th International Conference on*, 2004, pp. 90-95 vol.1.

[165] A. Azizi Mazreah, *et al.*, "A novel zero-aware read-static-noise-margin-free SRAM cell for high density and high speed cache application," in *Solid-State and Integrated-Circuit Technology, 2008. ICSICT 2008. 9th International Conference on*, 2008, pp. 876-879.

[166] S. Huang and W. Wong, "Analysis of contact resistance effect to SRAM performance in deep sub-micron technology," in *Solid-State and Integrated-Circuit Technology, 2008. ICSICT 2008. 9th International Conference on*, 2008, pp. 872-875.

[167] C. Qiang, *et al.*, "Critical current (I<inf>CRIT</inf>) based SPICE model extraction for SRAM cell," in *Solid-State and Integrated-Circuit Technology, 2008. ICSICT 2008. 9th International Conference on*, 2008, pp. 448-451.

[168] X. Xiaoyong, *et al.*, "Nonvolatile SRAM cell based on Cu<inf>x</inf>O," in *Solid-State and Integrated-Circuit Technology, 2008. ICSICT 2008. 9th International Conference on*, 2008, pp. 869-871.

[169] O. Xu, *et al.*, "Yield monitor for embedded-sige process optimization," in *Solid-State and Integrated-Circuit Technology, 2008. ICSICT 2008. 9th International Conference on*, 2008, pp. 1142-1145.

[170] M. Matsui, *et al.*, "An 8 ns 1 Mb ECL BiCMOS SRAM," in *Solid-State Circuits Conference, 1989. Digest of Technical Papers. 36th ISSCC., 1989 IEEE International*, 1989, pp. 38-39.

[171] F. Miyaji, *et al.*, "A 25 ns 4 Mb CMOS SRAM with dynamic bit line loads," in *Solid-State Circuits Conference, 1989. Digest of Technical Papers. 36th ISSCC., 1989 IEEE International*, 1989, pp. 250-251.

[172] K. Sasaki, *et al.*, "A 9 ns 1 Mb CMOS SRAM," in *Solid-State Circuits Conference, 1989. Digest of Technical Papers. 36th ISSCC., 1989 IEEE International*, 1989, pp. 34-35.

[173] T. Takahashi, *et al.*, "A 1.4 M-transistor CMOS gate array with 4 ns RAM," in *Solid-State Circuits Conference, 1989. Digest of Technical Papers. 36th ISSCC., 1989 IEEE International*, 1989, pp. 178-179.

[174] H. Tran, *et al.*, "An 8 ns BiCMOS 1 Mb ECL SRAM with a configurable memory array size," in *Solid-State Circuits Conference, 1989. Digest of Technical Papers. 36th ISSCC., 1989 IEEE International*, 1989, pp. 36-37.

[175] S. Aizaki, *et al.*, "A 15 ns 4 Mb CMOS SRAM," in *Solid-State Circuits Conference, 1990. Digest of Technical Papers. 37th ISSCC., 1990 IEEE International*, 1990, pp. 126-127.

[176] S. Flannagan, *et al.*, "8 ns CMOS 64 K*4 and 256 K*1 SRAMs," in *Solid-State Circuits Conference, 1990. Digest of Technical Papers. 37th ISSCC., 1990 IEEE International*, 1990, pp. 134-135.

[177] S. Hayakawa, *et al.*, "A 1 mu A retention 4 Mb SRAM with a thin-film-transistor load cell," in *Solid-State Circuits Conference, 1990. Digest of Technical Papers. 37th ISSCC., 1990 IEEE International*, 1990, pp. 128-129.

[178] Y. Maki, *et al.*, "A 6.5 ns 1 Mb BiCMOS ECL SRAM," in *Solid-State Circuits Conference, 1990. Digest of Technical Papers. 37th ISSCC., 1990 IEEE International*, 1990, pp. 136-137.

[179] K. Sasaki, *et al.*, "A 23 ns 4 Mb CMOS SRAM with 0.5 mu A standby current," in *Solid-State Circuits Conference, 1990. Digest of Technical Papers. 37th ISSCC., 1990 IEEE International*, 1990, pp. 130-131.

[180] K. Ishibashi, *et al.*, "A 1 V TFT-load SRAM using a two-step word-voltage method," in *Solid-State Circuits Conference, 1992. Digest of Technical Papers. 39th ISSCC, 1992 IEEE International*, 1992, pp. 206-207, 283.

[181] H. Kato, *et al.*, "A 9 ns 4 Mb BiCMOS SRAM with 3.3 V operation," in *Solid-State Circuits Conference, 1992. Digest of Technical Papers. 39th ISSCC, 1992 IEEE International*, 1992, pp. 210-211, 285.

[182] M. Matsumiya, *et al.*, "A 15 ns 16 Mb CMOS SRAM with reduced voltage amplitude data bus," in *Solid-State Circuits Conference, 1992. Digest of Technical Papers. 39th ISSCC, 1992 IEEE International*, 1992, pp. 214-215, 287.

[183] K. Nakamura, *et al.*, "A 6 ns 4 Mb ECL I/O BiCMOS SRAM with LV-TTL mask option," in *Solid-State Circuits Conference, 1992. Digest of Technical Papers. 39th ISSCC, 1992 IEEE International*, 1992, pp. 212-213, 286.

[184] K. Sasaki, *et al.*, "A 7 ns 140 mW 1 Mb CMOS SRAM with current sense amplifier," in *Solid-State Circuits Conference, 1992. Digest of Technical Papers. 39th ISSCC, 1992 IEEE International*, 1992, pp. 208-209, 284.

[185] K. Sasaki, *et al.*, "A 16 Mb CMOS SRAM with a 2.3 μm² single-bit-line memory cell," in *Solid-State Circuits Conference, 1993. Digest of Technical Papers. 40th ISSCC., 1993 IEEE International*, 1993, pp. 250-251, 297.

[186] G. Braceras, *et al.*, "A 200 MHz internal/66 MHz external 64 kB embedded virtual three-port cache SRAM," in *Solid-State Circuits Conference, 1994. Digest of Technical Papers. 41st ISSCC., 1994 IEEE International*, 1994, pp. 262-263.

[187] K. Ishibashi, *et al.*, "A 300 MHz 4-Mb wave-pipeline CMOS SRAM using a multi-phase PLL," in *Solid-State Circuits Conference, 1995. Digest of Technical Papers. 42nd ISSCC, 1995 IEEE International*, 1995, pp. 308-309, 386.

[188] M. Izumikawa, *et al.*, "A 0.9 V 100 MHz 4 mW 2 mm² 16 b DSP core," in *Solid-State Circuits Conference, 1995. Digest of Technical Papers. 42nd ISSCC, 1995 IEEE International*, 1995, pp. 84-85, 343.

[189] N. Kushiyama, *et al.*, "A 295 MHz CMOS 1 M (×256) embedded SRAM using bi-directional read/write shared sense amps and self-timed pulsed word-line drivers," in *Solid-State Circuits Conference, 1995. Digest of Technical Papers. 42nd ISSCC, 1995 IEEE International*, 1995, pp. 304-305, 385.

[190] K. Furumochi, *et al.*, "A 500 MHz 288 kb CMOS SRAM macro for on-chip cache," in *Solid-State Circuits Conference, 1996. Digest of Technical Papers. 42nd ISSCC., 1996 IEEE International*, 1996, pp. 156-157, 435.

[191] M. Krauss, *et al.*, "Fully-integrated 5 V CMOS system for a 20 M sample/s sampling oscilloscope," in *Solid-State Circuits Conference, 1996. Digest of Technical Papers. 42nd ISSCC., 1996 IEEE International*, 1996, pp. 384-385, 478.

[192] P. N. Glaskowsky, "MoSys explains 1T SRAM technology," 1999.

[193] R. F. Lyon and R. R. Schediwy, "CMOS static memory with a new four-transistor memory cell," Cambridge, MA, USA, 1987, pp. 111-32.

[194] K. Takeda, *et al.*, "A 16-Mb 400-MHz loadless CMOS four-transistor SRAM macro," *Solid-State Circuits, IEEE Journal of,* vol. 35, pp. 1631-1640, 2000.

[195] K. Takeda, *et al.*, "A 16 Mb 400 MHz loadless CMOS four-transistor SRAM macro," in *Solid-State Circuits Conference, 2000. Digest of Technical Papers. ISSCC. 2000 IEEE International*, 2000, pp. 264-265.

[196] A. A. Mazreah, *et al.*, "A Novel Zero-Aware Four-Transistor SRAM Cell for High Density and Low Power Cache Application," in *Advanced Computer Theory and Engineering, 2008. ICACTE '08. International Conference on*, 2008, pp. 571-575.

[197] H. Hong-Yi and S. Hsuan-Yi, "Low-power 2P2N SRAM with column hidden refresh," in *Circuits and Systems, 2002. ISCAS 2002. IEEE International Symposium on*, 2002, pp. IV-591-IV-594 vol.4.

[198] O. Semenov, *et al.*, "Sub-quarter micron SRAM cells stability in low-voltage operation: a comparative analysis," in *Integrated Reliability Workshop Final Report, 2002. IEEE International*, 2002, pp. 168-171.

[199] K. Noda, *et al.*, "A 1.9-μm² loadless CMOS four-transistor SRAM cell in a 0.18-μm logic technology," in *Electron Devices Meeting, 1998. IEDM '98 Technical Digest., International*, 1998, pp. 643-646.

[200] K. Noda, *et al.*, "An ultra-high-density high-speed loadless four-transistor SRAM macro with a dual-layered twisted bit-line and a triple-well shield," in *Custom Integrated Circuits Conference, 2000. CICC. Proceedings of the IEEE 2000*, 2000, pp. 283-286.

[201] S. Masuoka, *et al.*, "A 0.99-μm² loadless four-transistor SRAM cell in 0.13-μm generation CMOS technology," in *VLSI Technology, 2000. Digest of Technical Papers. 2000 Symposium on*, 2000, pp. 164-165.

[202] K. Noda, *et al.*, "An ultrahigh-density high-speed loadless four-transistor SRAM macro with twisted bitline architecture and triple-well shield," *Solid-State Circuits, IEEE Journal of,* vol. 36, pp. 510-515, 2001.

[203] K. Noda, *et al.*, "A loadless CMOS four-transistor SRAM cell in a 0.18-μm logic technology," *Electron Devices, IEEE Transactions on,* vol. 48, pp. 2851-2855, 2001.

[204] H. Hong-Yi and Y. Tzu-Sung, "A low-voltage loadless 4N SRAM with smart hidden refresh," in *SOC Conference, 2003. Proceedings. IEEE International [Systems-on-Chip]*, 2003, pp. 251-252.

[205] Y. Jinshen and C. Li, "A New Loadless 4-Transistor SRAM Cell with a 0.18 Âμm CMOS Technology," in *Electrical and Computer Engineering, 2007. CCECE 2007. Canadian Conference on*, 2007, pp. 538-541.

[206] B. Giraud, *et al.*, "In-depth Analysis of 4T SRAM Cells in Double-Gate CMOS," in *Integrated Circuit Design and Technology, 2007. ICICDT '07. IEEE International Conference on*, 2007, pp. 1-4.

[207] C. C. Wang, *et al.*, "A 4-kb Low-Power SRAM Design With Negative Word-Line Scheme," *Circuits and Systems I: Regular Papers, IEEE Transactions on,* vol. 54, pp. 1069-1076, 2007.

[208] T. S. Yang, *et al.*, "A 4-ns 4K×1-bit two-port BiCMOS SRAM," *Solid-State Circuits, IEEE Journal of,* vol. 23, pp. 1030-1040, 1988.

[209] H. Tran, "Demonstration of 5T SRAM and 6T dual-port RAM cell arrays," in *VLSI Circuits, 1996. Digest of Technical Papers., 1996 Symposium on*, 1996, pp. 68-69.

[210] I. Carlson, *et al.*, "A high density, low leakage, 5T SRAM for embedded caches," in *Solid-State Circuits Conference, 2004. ESSCIRC 2004. Proceeding of the 30th European*, 2004, pp. 215-218.

[211] M. Wieckowski and M. Margala, "A novel five-transistor (5T) sram cell for high performance cache," in *SOC Conference, 2005. Proceedings. IEEE International*, 2005, pp. 101-102.

[212] M. Wieckowski and M. Margala, "A portless SRAM Cell using stunted wordline drivers," in *Circuits and Systems, 2008. ISCAS 2008. IEEE International Symposium on*, 2008, pp. 584-587.

[213] L. You and H. Xiangqing, "A novel area-efficient and full current-mode dual-port SRAM," in *Communications, Circuits and Systems, 2008. ICCCAS 2008. International Conference on*, 2008, pp. 1079-1082.

[214] R. E. Aly and M. A. Bayoumi, "Low-Power Cache Design Using 7T SRAM Cell," *Circuits and Systems II: Express Briefs, IEEE Transactions on,* vol. 54, pp. 318-322, 2007.

[215] K. Takeda, *et al.*, "A read-static-noise-margin-free SRAM cell for low-VDD and high-speed applications," in *Solid-State Circuits Conference, 2005. Digest of Technical Papers. ISSCC. 2005 IEEE International*, 2005, pp. 478-611 Vol. 1.

[216] K. Takeda, *et al.*, "A read-static-noise-margin-free SRAM cell for low-VDD and high-speed applications," *Solid-State Circuits, IEEE Journal of,* vol. 41, pp. 113-121, 2006.

[217] C. Ching-Yun, *et al.*, "Power Gating Technique for Embedded Pseudo SRAM," in *VLSI Design, Automation and Test, 2007. VLSI-DAT 2007. International Symposium on*, 2007, pp. 1-4.

[218] C. Shin-Pao and H. Shi-Yu, "A low-power SRAM for Viterbi decoder in wireless communication," *Consumer Electronics, IEEE Transactions on,* vol. 54, pp. 290-295, 2008.

[219] B. Mohammad, *et al.*, "Cache Design for Low Power and High Yield," in *Quality Electronic Design, 2008. ISQED 2008. 9th International Symposium on*, 2008, pp. 103-107.

[220] K. Keejong, *et al.*, "A Low-Power SRAM Using Bit-Line Charge-Recycling," *Solid-State Circuits, IEEE Journal of,* vol. 43, pp. 446-459, 2008.

[221] L. Chang, *et al.*, "Stable SRAM cell design for the 32 nm node and beyond," in *VLSI Technology, 2005. Digest of Technical Papers. 2005 Symposium on*, 2005, pp. 128-129.

[222] A. Sil, *et al.*, "A Novel 90nm 8T SRAM Cell With Enhanced Stability," in *Integrated Circuit Design and Technology, 2007. ICICDT '07. IEEE International Conference on*, 2007, pp. 1-4.

[223] A. Sil, *et al.*, "A novel 8T SRAM cell with improved read-SNM," in *Circuits and Systems, 2007. NEWCAS 2007. IEEE Northeast Workshop on*, 2007, pp. 1289-1292.

[224] V. Ramadurai, *et al.*, "A Disturb Decoupled Column Select 8T SRAM Cell," in *Custom Integrated Circuits Conference, 2007. CICC '07. IEEE*, 2007, pp. 25-28.

[225] V. Naveen and A. P. Chandrakasan, "A 65nm 8T Sub-Vt SRAM Employing Sense-Amplifier Redundancy," in *Solid-State Circuits Conference, 2007. ISSCC 2007. Digest of Technical Papers. IEEE International*, 2007, pp. 328-606.

[226] C. Leland, *et al.*, "A 5.3GHz 8T-SRAM with Operation Down to 0.41V in 65nm CMOS," in *VLSI Circuits, 2007 IEEE Symposium on*, 2007, pp. 252-253.

[227] K. Young Bok, *et al.*, "Low power 8T SRAM using 32nm independent gate FinFET technology," in *SOC Conference, 2008 IEEE International*, 2008, pp. 247-250.

[228] K. Roy, *et al.*, "Process-Tolerant Ultralow Voltage Digital Subthreshold Design," in *Silicon Monolithic Integrated Circuits in RF Systems, 2008. SiRF 2008. IEEE Topical Meeting on*, 2008, pp. 42-45.

[229] H. Noguchi, *et al.*, "Which is the best dual-port SRAM in 45-nm process technology? — 8T, 10T single end, and 10T differential &#x2014," in *Integrated Circuit Design and Technology and Tutorial, 2008. ICICDT 2008. IEEE International Conference on*, 2008, pp. 55-58.

[230] K. Nii, *et al.*, "A 45-nm single-port and dual-port SRAM family with robust read/write stabilizing circuitry under DVFS environment," in *VLSI Circuits, 2008 IEEE Symposium on*, 2008, pp. 212-213.

[231] T.-H. Kim, *et al.*, "A voltage scalable 0.26V, 64kb 8T SRAM with V<inf>min</inf> lowering techniques and deep sleep mode," in *Custom Integrated Circuits Conference, 2008. CICC 2008. IEEE*, 2008, pp. 407-410.

[232] L. Chang, *et al.*, "An 8T-SRAM for Variability Tolerance and Low-Voltage Operation in High-Performance Caches," *Solid-State Circuits, IEEE Journal of*, vol. 43, pp. 956-963, 2008.

[233] K. Nii, *et al.*, "A 90nm dual-port SRAM with 2.04 μm² 8T-thin cell using dynamically-controlled column bias scheme," in *Solid-State Circuits Conference, 2004. Digest of Technical Papers. ISSCC. 2004 IEEE International*, 2004, pp. 508-543 Vol.1.

[234] K. Nii, *et al.*, "A 90-nm low-power 32-kB embedded SRAM with gate leakage suppression circuit for mobile applications," *Solid-State Circuits, IEEE Journal of*, vol. 39, pp. 684-693, 2004.

[235] K. Nii, *et al.*, "A 45-nm Bulk CMOS Embedded SRAM With Improved Immunity Against Process and Temperature Variations," *Solid-State Circuits, IEEE Journal of*, vol. 43, pp. 180-191, 2008.

[236] C. Yen-Jen, *et al.*, "Zero-aware asymmetric SRAM cell for reducing cache power in writing zero," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 12, pp. 827-836, 2004.

[237] Y. Morita, *et al.*, "An Area-Conscious Low-Voltage-Oriented 8T-SRAM Design under DVS Environment," in *VLSI Circuits, 2007 IEEE Symposium on*, 2007, pp. 256-257.

[238] L. Zhiyu and V. Kursun, "High Read Stability and Low Leakage Cache Memory Cell," in *Circuits and Systems, 2007. ISCAS 2007. IEEE International Symposium on*, 2007, pp. 2774-2777.

[239] K. Tae-Hyoung, *et al.*, "A High-Density Subthreshold SRAM with Data-Independent Bitline Leakage and Virtual Ground Replica Scheme," in *Solid-State Circuits Conference, 2007. ISSCC 2007. Digest of Technical Papers. IEEE International*, 2007, pp. 330-606.

[240] K. Tae-Hyoung, *et al.*, "Circuit techniques for ultra-low power subthreshold SRAMs," in *Circuits and Systems, 2008. ISCAS 2008. IEEE International Symposium on*, 2008, pp. 2574-2577.

[241] B. H. Calhoun and A. Chandrakasan, "A 256kb Sub-threshold SRAM in 65nm CMOS," in *Solid-State Circuits Conference, 2006. ISSCC 2006. Digest of Technical Papers. IEEE International*, 2006, pp. 2592-2601.

[242] C. Benton Highsmith and P. C. Anantha, "A 256-kb 65-nm Sub-threshold SRAM Design for Ultra-Low-Voltage Operation," *Solid-State Circuits, IEEE Journal of*, vol. 42, pp. 680-688, 2007.

[243] J. P. Kulkarni, *et al.*, "Process variation tolerant SRAM array for ultra low voltage applications," in *Design Automation Conference, 2008. DAC 2008. 45th ACM/IEEE*, 2008, pp. 108-113.

[244] S. A. Verkila, *et al.*, "A 100MHz to 1GHz, 0.35V to 1.5V Supply 256 x 64 SRAM Block Using Symmetrized 9T SRAM Cell with Controlled Read," in *VLSI Design, 2008. VLSID 2008. 21st International Conference on*, 2008, pp. 560-565.

[245] F. Frustaci, *et al.*, "Leakage energy reduction techniques in deep submicron cache memories: a comparative study," in *Circuits and Systems, 2006. ISCAS 2006. Proceedings. 2006 IEEE International Symposium on*, 2006, p. 4 pp.

[246] A. Chrisanthopoulos, *et al.*, "Comparative study of different current mode sense amplifiers in submicron CMOS technology," *Circuits, Devices and Systems, IEE Proceedings -*, vol. 149, pp. 154-158, 2002.

[247] P. Y. Chee, *et al.*, "High-speed hybrid current-mode sense amplifier for CMOS SRAMs," *Electronics Letters*, vol. 28, pp. 871-873, 1992.

[248] A. Vladimirescu, *et al.*, "Ultra-low-voltage robust design issues in deep-submicron CMOS," in *Circuits and Systems, 2004. NEWCAS 2004. The 2nd Annual IEEE Northeast Workshop on*, 2004, pp. 49-52.

[249] B. H. Calhoun and A. Chandrakasan, "Analyzing static noise margin for sub-threshold SRAM in 65nm CMOS," in *Solid-State Circuits Conference, 2005. ESSCIRC 2005. Proceedings of the 31st European*, 2005, pp. 363-366.

[250] Z. Bo, *et al.*, "A Variation-Tolerant Sub-200 mV 6-T Subthreshold SRAM," *Solid-State Circuits, IEEE Journal of*, vol. 43, pp. 2338-2348, 2008.

[251] C. Ik Joon, *et al.*, "A 32 kb 10T Sub-Threshold SRAM Array With Bit-Interleaving and Differential Read Scheme in 90 nm CMOS," *Solid-State Circuits, IEEE Journal of*, vol. 44, pp. 650-658, 2009.

[252] A. Sil, *et al.*, "A novel high write speed, low power, read-SNM-free 6T SRAM cell," in *Circuits and Systems, 2008. MWSCAS 2008. 51st Midwest Symposium on*, 2008, pp. 771-774.

[253] Q. Chen, *et al.*, "An Accurate Analytical SNM Modeling Technique for SRAMs Based on Butterworth Filter Function," in *VLSI Design, 2007. Held jointly with 6th International Conference on Embedded Systems., 20th International Conference on*, 2007, pp. 615-620.

[254] B. H. Calhoun and A. P. Chandrakasan, "Static noise margin variation for sub-threshold SRAM in 65-nm CMOS," *Solid-State Circuits, IEEE Journal of*, vol. 41, pp. 1673-1679, 2006.

[255] O. Chang-Bong, *et al.*, "Ultra low power 6T-SRAM chip with improved transistor performance and reliability by HfO₂-Al₂O₃ high-K gate dielectric process optimization," in *VLSI Technology, 2003. Digest of Technical Papers. 2003 Symposium on*, 2003, pp. 71-72.

[256] P. Andrei and M. S., "Soft Errors in SRAMs: Sources, Mechanisms and Mitigation Techniques," in *CMOS SRAM Circuit Design and Parametric Test in Nano-Scaled Technologies*. vol. 40, ed: Springer Netherlands, 2008.

[257] K. Zhang, *et al.*, "A 3-GHz 70MB SRAM in 65nm CMOS technology with integrated column-based dynamic power supply," in *Solid-State Circuits Conference, 2005. Digest of Technical Papers. ISSCC. 2005 IEEE International*, 2005, pp. 474-611 Vol. 1.

[258] E. Seevinck, "A current sense-amplifier for fast CMOS SRAMs," in *VLSI Circuits, 1990. Digest of Technical Papers., 1990 Symposium on*, 1990, pp. 71-72.

[259] K. S. Yeo, *et al.*, "High-performance low-power current sense amplifier using a cross-coupled current-mirror configuration," *Circuits, Devices and Systems, IEE Proceedings -*, vol. 149, pp. 308-314, 2002.

[260] Z. H. Kong, *et al.*, "An ultra-low power current-mode sense amplifier for SRAM applications," *J. Circuits Syst. Compt.*, vol. 14, pp. 939-951, 2005.

[261] S. Patil, *et al.*, "A Self-Biased Charge-Transfer Sense Amplifier," in *Circuits and Systems, 2007. ISCAS 2007. IEEE International Symposium on*, 2007, pp. 3030-3033.

[262] A. Hajimiri and R. Heald, "Design issues in cross-coupled inverter sense amplifier," presented at the IEEE Int. Symp. Circuits and Systems, 1998.

[263] K. S. Yeo, "High-performance, low-power current sense amplifier using a cross-coupled current mirror configuration," *IEE Proc. Circuits Dev. Syst.*, vol. 149, pp. 308-314, Oct/Dec. 2002.

[264] B. Witch, *et al.*, "Yield and speed optimization of a latch-type voltage sense amplifier," *IEEE J. Solid-State Circuit,* vol. 39, pp. 1148-1158, Jul. 2000.

[265] R. Singh and N. Baht, "An offset compensation technique for latch type sense amplifiers in high-speed low-power SRAMs," *IEEE Trans. VLSI System Transaction Briefs,* vol. 12, pp. 652-657, Jun. 2004.

[266] S. J. Lovett, *et al.*, "Yield and matching implications for static RAM memory array sense-amplifier design," *IEEE J. Solid-State Circuit,* vol. 35, pp. 1200-1204, Aug. 2000.

[267] B. Wicht, *Current sense amplifier for embedded SRAM in high-performance System-on-a-Chip designs*. New York: Springer-Verlag, 2003.

[268] R. Sarpeshkar, *et al.*, "Mismatch sensitivity of a simultaneously latched CMOS sense amplifier," *Solid-State Circuits, IEEE Journal of,* vol. 26, pp. 1413-1422, 1991.

[269] N. Hyunwoo, *et al.*, "Numerical estimation of yield in sub-100-nm SRAM design using Monte Carlo simulation," *IEEE Transactions on Circuits and Systems II: Express Briefs,* vol. 55, pp. 907-11, 2008.

[270] K. Agarwal and S. Nassif, "The Impact of Random Device Variation on SRAM Cell Stability in Sub-90-nm CMOS Technologies," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on,* vol. 16, pp. 86-97, 2008.

[271] A. Hajimiri and R. Heald, "Design issues in cross-coupled inverter sense amplifier," New York, NY, USA, 1998, pp. 149-52.

[272] T. Kawahara, *et al.*, "A high-speed, small-area, threshold-voltage-mismatch compensation sense amplifier for gigabit-scale DRAM arrays," *Solid-State Circuits, IEEE Journal of,* vol. 28, pp. 816-823, 1993.

[273] B. Wicht, *et al.*, "Yield and speed optimization of a latch-type voltage sense amplifier," *Solid-State Circuits, IEEE Journal of,* vol. 39, pp. 1148-1158, 2004.

[274] S. J. Lovett, *et al.*, "Yield and matching implications for static RAM memory array sense-amplifier design," *Solid-State Circuits, IEEE Journal of,* vol. 35, pp. 1200-1204, 2000.

[275] A. Nikoozadeh and B. Murmann, "An Analysis of Latch Comparator Offset Due to Load Capacitor Mismatch," *Circuits and Systems II: Express Briefs, IEEE Transactions on,* vol. 53, pp. 1398-1402, 2006.

[276] S. Rodrigues and M. S. Bhat, "Impact of Process Variation Induced Transistor Mismatch on Sense Amplifier Performance," in *Advanced Computing and Communications, 2006. ADCOM 2006. International Conference on*, 2006, pp. 497-502.

[277] J. F. Ryan and B. H. Calhoun, "Minimizing Offset for Latching Voltage-Mode Sense Amplifiers for Sub-Threshold Operation," in *Quality Electronic Design, 2008. ISQED 2008. 9th International Symposium on*, 2008, pp. 127-132.

[278] L. Pileggi, *et al.*, "Mismatch analysis and statistical design at 65 nm and below," in *Custom Integrated Circuits Conference, 2008. CICC 2008. IEEE*, 2008, pp. 9-12.

[279] A. Agarwal, *et al.*, "Process variation in embedded memories: failure analysis and variation aware architecture," *Solid-State Circuits, IEEE Journal of,* vol. 40, pp. 1804-1814, 2005.

[280] M. Sharifkhani and M. Sachdev, "SRAM Cell Stability: A Dynamic Perspective," *Solid-State Circuits, IEEE Journal of,* vol. 44, pp. 609-619, 2009.

[281] R. Singh and N. Bhat, "An offset compensation technique for latch type sense amplifiers in high-speed low-power SRAMs," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on,* vol. 12, pp. 652-657, 2004.