Subthreshold SRAM Design for Energy Efficient Applications in Nanometric CMOS Technologies

by

Morteza Nabavi

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2018

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Bruce Cockburn Professor, Dept. of Electrical and Computer Engineering, University of Alberta
Supervisor(s):	Manoj Sachdev Professor, Dept. of Electrical and Computer Engineering, University of Waterloo
Internal Member:	David Nairn Associate Professor, Dept. of Electrical and Computer Engineering, University of Waterloo
Internal Member:	Peter Levine Assistant Professor, Dept. of Electrical and Computer Engineering, University of Waterloo
Internal-External Member:	James Martin Associate Professor, Dept. of Physics and Astronomy, University of Waterloo

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Embedded SRAM circuits are vital components in a modern system on chip (SOC) that can occupy up to 90% of the total area. Therefore, SRAM circuits heavily affect SOC performance, reliability, and yield. In addition, most of the SRAM bitcells are in standby mode and significantly contribute to the total leakage current and leakage power consumption. The aggressive demand in portable devices and billions of connected sensor networks requires long battery life. Therefore, careful design of SRAM circuits with minimal power consumption is in high demand. Reducing the power consumption is mainly achieved by reducing the power supply voltage in the idle mode. However, simply reducing the supply voltage imposes practical limitations on SRAM circuits such as reduced static noise margin, poor write margin, reduced number of cells per bitline, and reduced bitline sensing margin that might cause read/write failures. In addition, the SRAM bitcell has contradictory requirements for read stability and writability. Improving the read stability can cause difficulties in a write operation or vice versa.

In this thesis, various techniques for designing subthreshold energy-efficient SRAM circuits are proposed. The proposed techniques include improvement in read margin and write margin, speed improvement, energy consumption reduction, new bitcell architecture and utilizing programmable wordline boosting. A programmable wordline boosting technique is exploited on a conventional 6T SRAM bitcell to improve the operational speed. In addition, wordline boosting can reduce the supply voltage while maintaining the operational frequency. The reduction of the supply voltage allows the memory macro to operate with reduced power consumption. To verify the design, a 16-kb SRAM was fabricated using the TSMC 65 nm CMOS technology. Measurement results show that the maximum operational frequency increases up to 33.3% when wordline boosting is applied. Besides, the supply voltage can be reduced while maintaining the same frequency. This allows reducing the energy consumption to be reduced by 22.2%. The minimum energy consumption achieved is 0.536 fJ/b at 400 mV. Moreover, to improve the read margin, a 6T bitcell SRAM with a PMOS access transistor is proposed. Utilizing a PMOS access transistor results in lower zero level degradation, and hence higher read stability. In addition, the access transistor connected to the internal node holding V_{DD} acts as a stabilizer and counterbalances the effect of zero level degradation. In order to improve the writability, wordline boosting is exploited. Wordline boosting also helps to compensate for the lower speed of the PMOS access transistor compared to a NMOS transistor. To verify our design, a 2kb SRAM is fabricated in the TSMC 65 nm CMOS technology. Measurement results show that the maximum operating frequency of the test chip is at 3.34 MHz at 290 mV. The minimum energy consumption is measured as 1.1 fJ/b at 400 mV.

Acknowledgements

I would like to take this opportunity to express my extreme gratitude to my Ph.D. research supervisor Professor Manoj Sachdev for providing me with his technical knowledge and moral support. At many stages in my program, I benefited from his advice and positive feedback that inspired confidence in me. Without his patience, I would not be standing at this point.

I would also like to express my deepest respect to my father, Professor Abdolreza Nabavi, and my mother, Masoumeh Mirzaee, for their unconditional love, encouragement, having my back, and enduring all the hardship they went through since I was born.

I would like to thank Dr. Mohammad Sharifkhani and Dr. Roghaye Saeidi for their generous and unsparing help and comments. I should acknowledge Dr. Adam Neal's help especially at the beginning of my Ph.D. journey when he shared his experiences with me. The other valuable members of our research group also never withheld their support whenever I needed it. I would like to particularly thank Dr. Derek Wright, Dr. Jaspal Singh, Sunil Sanjeevi, and Dhruv Patel.

Table of Contents

List of Tables

List of Figures

Abbreviations

6T Six Transistors 5, 6, 8, 9, 11, 12, 18–22, 31, 49, 50, 76, 77, 80

6TNA 6T Bitcell with NMOS Access Transistor 53

6TPA 6T Bitcell with PMOS Access Transistor 53, 77

BL Bitline 3–6, 9–15, 18–21, 25, 29, 50, 80, 81

BLB Bitline-Bar 18–20, 29, 50, 61, 80

DBA Delta-Boosted Array Voltage 3, 4

DIBL Drain-Induced Barrier Lowering 2, 31, 54

DRV Data Retention Voltage 2

FFT Fast Fourier Transform 11

FOM Figure of Merit 79, 80

GIDL Gate-Induced Barrier Lowering 3

MIM Metal Insulator Metal 70

MS Mode Select 63

OEB Output Enable Bar 22, 23

PD Pull Down 54–56, 59

RD Rectangular Diffusion 3

RDF Random Dopant Fluctuation 11, 12

SA Sense Amplifier 16, 25, 29, 30, 79

SAE Sense Amplifier Enable 29

SEC-DED Single-Error Correction and Double-Error Detection 49

SNM Static Noise Margin 3–5, 11, 12, 20, 21, 48–50, 61, 76, 79, 80

SOC System on Chip 31, 48

V-Boost Boost Voltage 63

VTC Voltage Transfer Characteristic 20–22

WL Wordline 3–6, 12, 18, 19, 25, 29, 32, 35–38, 46, 47, 49, 50, 58, 59, 61–63, 65–72, 76, 78–80, 92

WM Write Margin 3–5, 13, 21, 22, 48, 49, 61, 64, 65, 76, 79, 80

ZLD Zero-Level Degradation 19, 50, 55, 58, 59, 61, 80

List of Symbols

$\mathbf{C_{ox}}$ Oxide Capacitance of a MOSFET 54

$\mathbf{C_R}$ Cell ratio 50, 55

$\mathbf{\Delta BL}$ Differential voltage between BL and BLB 70

$\mathbf{D_R}$ Driving strength ratio of the pull-down transistor to the access transistor 54, 55

η DIBL coefficient of a MOSFET 54

$\mathbf{\Gamma}$ called the subthreshold cell ratio modification factor 55

$\mathbf{I_{cell}}$ The cell read current 6

$\mathbf{I_{Read}}$ Read Current 72

λ The body effect coefficient of a MOSFET 54, 55

$\mathbf{L_a}$ Length of the access transistor 54

$\mathbf{L_n}$ Length of the NMOS driver transistor 54

$\mu_\mathbf{n}$ Mobility of an NMOS transistor 54

$\mu_\mathbf{p}$ Mobility of an PMOS transistor 54

μ Charge carrier mobility of a MOSFET; Micro; Mean value 54

$\nu_\mathbf{T}$ Thermal voltage 54

$\mathbf{V_{BS}}$ Body-Source voltage of a MOSFET 54

$\mathbf{V_{DS}}$ Drain-Source voltage of a MOSFET 54

$\mathbf{V_{GS}}$ Gate-Source Voltage 6, 54, 58, 61

$\mathbf{V_{min}}$ Minimum operational supply voltage 9–11, 76

$\mathbf{V_{n}}$ Noise Voltage Source 20, 21

$\mathbf{V_{QB}}$ Voltage of node QB 20

$\mathbf{V_{SS}}$ GND 3

$\mathbf{V_{t0}}$ Zero-biased threshold voltage of a MOSFET 54

$\mathbf{V_{tn}}$ Threshold voltage of the NMOS transistors 54

$\mathbf{V_{tp}}$ Threshold voltage of the PMOS transistors 54, 61

$\mathbf{W_{a}}$ Width of the access transistor 54

$\mathbf{W_{n}}$ Width of the NMOS driver transistor 54

β The size of the pull down transistor to the size of the access transistor 5, 55

$\mu\mathbf{m^2}$ Micro meter square 9

$\mathbf{V_{DD}}$ Supply Voltage 2, 3, 13, 18–21, 23, 25, 29, 50, 53, 58, 59, 61, 63, 65, 67, 80

Chapter 1

Introduction

1.1 Motivation and Problem Statement

In today's portable device market, SRAM circuits can significantly contribute to the total power consumption especially in the standby mode. The energy budget for portable devices is typically one lithium-ion battery of about 3000 mWH (1000 mAH). In addition to the limited battery budget, the peak active power must be held under 1W to manage the effect of temperature variation. The standby power of smart-phones including RF amplifier, the LCD display, and the baseband system should not consume more than 0.5 to 1.0 mW [1]. In addition to the portable devices, the main challenge, that the billions of nodes constructing the internet of things pose, is energy efficiency. Therefore, designing SRAM circuits consuming low power/energy is in high demanded [2].

There are several challenges in reducing the power/energy consumption of SRAM circuits including reduced static noise margin, poor write margin, reduced $\frac{I_{on}}{I_{off}}$ ratio (limited number of cells per bitline), and reduced bitline sensing margin [3]. In this thesis, various circuit techniques for designing subthreshold energy-efficient SRAM circuits are proposed. These techniques, in particular, include improving the read and write margins, increasing the number of bitcells per column, adopting a new bitcell architecture, and utilizing programmable wordline boosting.

1.2 Literature Review

In the following section, previously reported techniques of power and energy reduction are presented. These techniques include:

- Supply Voltage and Source-Line Manipulation
- Read/Write Assist Circuitry and Bitline and Wordline Signal Manipulation
- Bitline Leakage Reduction
- Transistor-Level Techniques
- Subthreshold Bitcell Design
- Application-Specific Techniques

1.2.1 Supply Voltage and Source-Line Manipulation

To reduce power consumption, several researchers have suggested reducing the power supply voltage [4–6]. This is due to the fact that the power consumption is proportional to the supply voltage and total current consumption. By reducing the supply voltage the current consuption also reduces. In [4], micro-architectural techniques are explored to implement data caches operating in the sleep mode. It is shown that by simple micro-architectural techniques, about 80% of the data cache lines can be maintained in a drowsy state (reverse back bias) with a negligible performance loss. Researchers in [5] have investigated the leakage power by reducing the standby supply voltage to a limit called the Data Retention Voltage (DRV). The impact of process variations, chip temperature, and transistor sizing on DRV are analyzed. An analytical model for DRV as a function of these parameters is also presented. It is shown that the DRV is a strong function of process variation. This model is verified by measurement results in 130 nm CMOS technology. The measurement results show that the SRAM module is capable of preserving data at sub-300 mV where 90% leakage-power reduction can be achieved. The authors in [6] show that the leakage-power can be reduced by reducing the Drain-Induced Barrier Lowering (DIBL) effect. The supply voltage of non-accessed cells is dynamically dropped row-by-row. A negative voltage is also applied to the non-accessed wordlines to decrease the leakage current of the bitlines through the access transistors. To match PMOS and NMOS leakage currents, N-well biasing and reduced $\mathbf{V_{DD}}$ are used in addition to negatively biasing the unselected wordlines. Measurement results show about 90% leakage-current reduction. A transient negative Bitline (BL) voltage is also proposed in [7] to improve the WM of the bitcell. A coupling capacitance is used to generate the required negative voltage.

In [8], two supply voltages are exploited. During a read operation, the higher supply voltage is chosen to create a positive differential voltage between the cell and WL to increase the read stability or Static Noise Margin (SNM). During a write operation, the lower supply voltage is chosen to create a negative differential voltage between the cell and WL to improve the WM and to make the cell data easier to flip.

In [9], the supply voltage of each column is connected to the global supply voltage by a power switch. This strategy improves the WM and eliminates the half-selected issue. This technique can also decrease the minimum supply voltage.

Another alternative to power supply scaling is to increase the ground level ($\mathbf{V_{SS}}$). In [10], a charge-recycle offset-source driving scheme is proposed. The simulation results show a reduction in power consumption by one-fourteenth compared to [11]. The source line of the SRAM bitcells in [11] are set to a negative and high-impedance voltage (floating) during read and write operations, respectively. This technique results in an improved access time. Another similar approach using a virtual-GND along the bitlines are presented in [12]. The source lines are shared by the cells in the same column. This technique significantly increases the power consumption of the read operation.

In [13], the BL voltages are reduced from 1.5 V to 1 V and the $\mathbf{V_{SS}}$ is raised from 0 to 0.5 V. This voltage scheme reduces the gate tunnel leakage, and the Gate-Induced Barrier Lowering (GIDL) leakage by about 90%.

The impact of reverse-biased transistors is explored in [1]. The technique proposed in this paper uses device back-bias to reduce the subthreshold current. The $\mathbf{V_{SS}}$ of the n-channel devices is raised while the substrate is kept at 0. At the same time, the $\mathbf{V_{DD}}$ of the p-channel devices is reduced while the substrate is kept at $\mathbf{V_{DD}}$. This technique leads to a 16× reduction in standby leakage current for a 2 MB array.

1.2.2 Read/Write Assist Circuitry and Bitline and Wordline Signal Manipulation

The SRAM array in [14] utilizes a Rectangular Diffusion (RD) cell and a Delta-Boosted Array Voltage (DBA). Utilizing a rectangular-diffusion cell decreases the pattern fluctuation that mitigates the impact of process variations which is one of the main barriers in low-voltage operation. To have a proper SNM, the cell ratio is usually set to around 1.5.

The rectangular-diffusion cell results in a cell ratio of 1.0 which in turn deteriorates the SNM of the bitcell. The DBA scheme is exploited to compensate for the deteriorated SNM. However, the DBA scheme reduces the WM of the bitcell. To compensate for the WM, pull-up transistors with higher threshold voltage are used in the SRAM bitcell.

The read assist circuit used in [15] provides full BL amplification to half-selected columns to write back the original data. This scheme requires a sense amplifier per column. In addition, a lower power supply voltage is provided to the write-only columns during a write operation to increase the WM.

A hierarchical BL and local sense amplifier scheme is used in [16]. This scheme reduces both the capacitance and write swing voltage of bitlines resulting in lower write power consumption without noise margin degradation. Simulation results illustrate 34% power savings compared to the conventional scheme. The fabricated SRAM test chip operates at 2.5 V running at 200 MHz. The test chip consumes 26 mW of read power and 28 mW of write power.

A replica technique on the bitlines is used in [17] to produce a reference voltage to track the delay of the bitlines. This technique reduces the impact of process variation. In addition, the WL pulse width is minimized to the minimum required amount. This, in turn, reduces the BL swing and reduces the power consumption.

In order to improve the WM, a power-line-floating technique during the write operation is presented in [18]. This technique also reduces the minimum supply voltage. A process-variation-adaptive write replica circuit is also exploited to decrease the leakage current. The floating technique is only applied to the selected columns, and the replica circuit saves power on the non-selected columns.

The authors in [19], show that large signal sensing is also a viable option as opposed to small signal sensing in the deep sub-micron regime. The new scheme creates a small signal swing on the local BLs and creates a large signal swing on the global BLs with reduced capacitance.

The authors in [20] propose pulsed-BL and pulsed-WL techniques to improve SRAM cell stability in single-V_{CC} microprocessors. In the pulsed-BL scheme, the BLs are discharged to a value of 100-300 V lower than the nominal supply voltage. This scheme decreases the cell current but increases the SNM. To compensate for the reduction of WM, a read-modify-write scheme is incorporated into the design. These techniques are made programmable to adapt to process and temperature variations. The pulsed-WL technique improves the cell failure rate by 15×. Simulation results show that utilizing both the pulsed-WL and pulsed-BL techniques with the read-modify-write scheme provides 26× read stability with an area overhead of 4-8%.

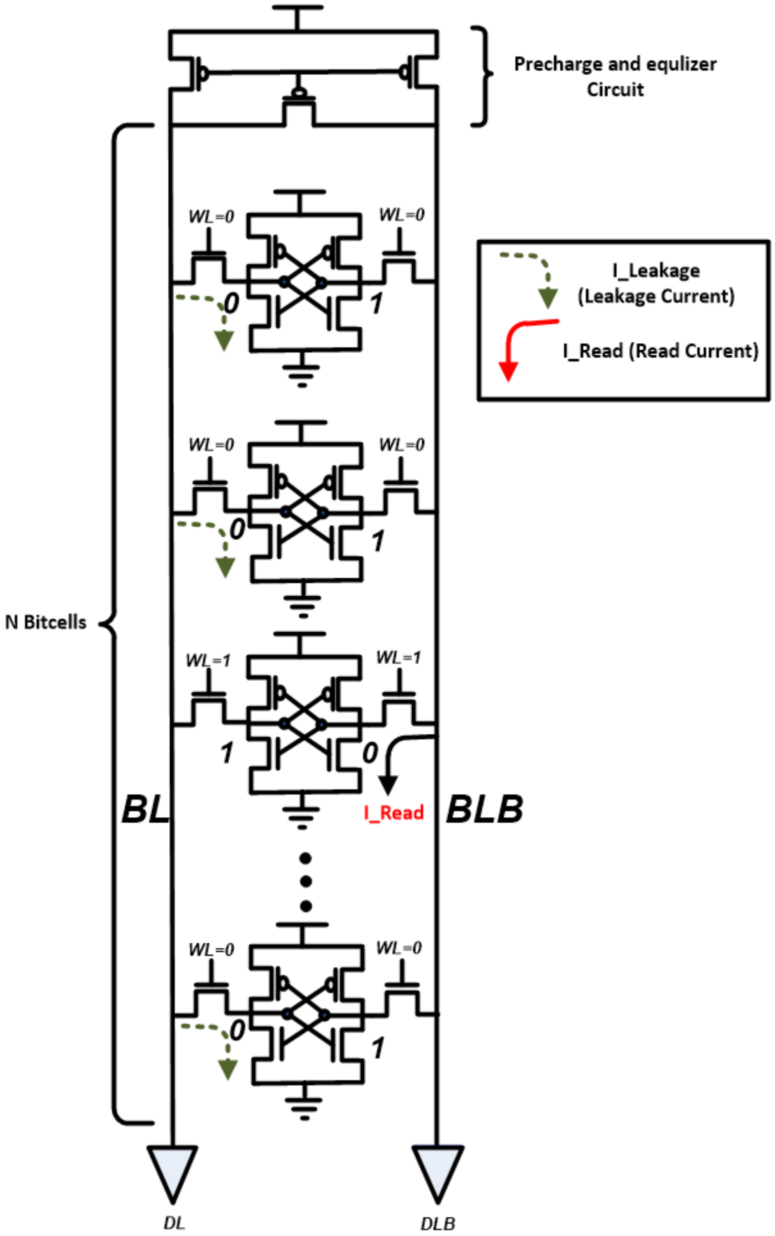
A variability-tolerant 6T SRAM cell that improves both the SNM and WM is presented in [21]. To mitigate the impact of process variations, the β ratio of the bitcells is chosen to be equal to 1. In addition, a read-assist circuit is used to reduce the voltage level of WL compared to the nominal supply voltage. This improves the read stability. Moreover, a capacitive write assist circuit is used to improve the WM. However, this scheme is prone to process variation. The WL is pulled down by multiple NMOS transistors and as a result their threshold voltage is dependent on process and temperature variations. The proposed circuitry in [22] overcomes these problems. The NMOS transistors are placed at the source of the WL driver with resistance elements using N^+ polysilicon gate. The write assist circuit utilizes the capacitive ratio between the local and global supply rail. The supply voltage to each SRAM bitcell decreases based on this ratio. Simulation results show improved immunity against process variations.

A hierarchical SRAM architecture with multi-step WL scheme is presented in [23]. The divided BL scheme used in this architecture reduces the capacitance on the bitlines by a factor of four which in turn reduces the power consumption and increases the read stability by decreasing the amount of charge flow to the selected bitcells. Moreover, it is shown that both SNM and BL speed are improved by the use of local sense amplifiers. In order to improve the WM, a slow transition of the WL is considered in addition to the WL boosting scheme. Simulation results show the superiority of this scheme against process variations. The slow transition of the WL boosting adds an extra delay to the total delay and increases the complexity of the timing signals.

The WL boosting technique is also implemented in [24] to improve the WM and reduce the impact of process variations. In the proposed WL boosting technique a Miller capacitance is used for each WL. A large area is required to provide one large capacitance for each WL and this makes such an approach inefficient.

A single-power-supply 6T SRAM exploiting read and write circuitry operating at 0.7 V and 1 GHz is presented in [25]. Both the WM and the cell current are improved using a β ratio of 1. To enhance cell stability and the SNM, a fine-grained BL segmentation scheme as well as a reduction in the number of cells per column, are implemented.

One issue in the write operation is to avoid the unnecessary BL swing and hence reduce extra power consumption. One example that extends the concept described above is to add an extra NMOS in the series with the VSS rail in the 6T bitcell [26]. During a write operation, this NMOS turns on and the VSS node of the 6T transistor floats. Therefore, the two back-to-back inverters get weak and can easily flip the state by a smaller differential supply voltage on the bitlines. Consequently, this approach reduces the write power consumption by 90%.



(a)

Figure 1.1: Schematic of a column with N bitcells.

1.2.3 Bitline Leakage Reduction

Figure 1.1 shows a column with N bitcells. The read current and the leakage currents are also shown in this figure. BL leakage creates several problems in SRAM memories. In the standby mode, it increases the leakage power and temperature. The worst-case leakage happens when all the non-accessed cells hold the complement of the data in the accessed cell. During a read operation, the BL leakage might be opposed to the read current ($\mathbf{I_{cell}}$) and create an extra delay or create an error in the cell. The leakage current in the BL imposes a delay in the read operation, or it might result in a false read.

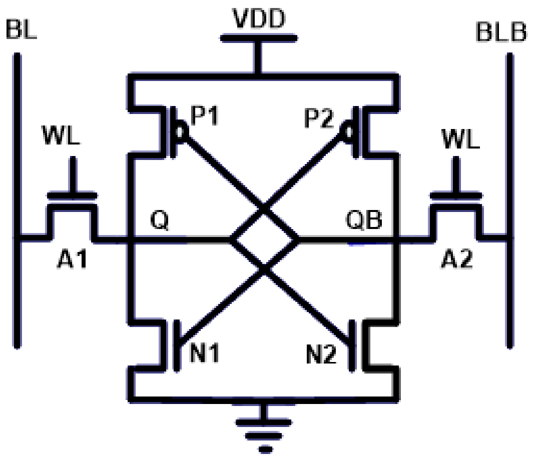
Reducing the voltage of the non-accessed WLs to a negative value is proposed in [6]. This reduces the subthreshold leakage of the non-accessed cells by creating a negative $\mathbf{V_{GS}}$ on their access transistors, but it requires extra circuitry to create a negative voltage. In [27] a BL leakage reduction technique is proposed to eliminate the impact of BL leakage on performance and noise margin with a minimal area overhead. In this technique, high threshold-voltage transistors are used for the access transistors. A negative WL voltage is also used for non-accessed transistors, and the voltage of BL and bitcells are reduced from the nominal supply voltage to decrease the leakage currents of the bitlines. The results show a 23% improvement in BL delay as compared to the best conventional design, thus enabling 6-GHz operation at a 15% higher energy consumption. However, there is a reliability issue due to the exploitation of multiple supply voltages.

Another relatively complicated approach to BL leakage is to measure the actual leakage current and then compensate accordingly [28]. This approach adds an extra delay by measuring and injecting the compensation currents.

A simpler approach uses two extra transistors in the 6T cell to equalize the BL leakage [29]. This scheme imposes the worst-case leakage not only on one BL, but also, on both. However, it ensures the same leakage on both bitlines. By using this technique, the BL differential development time is decreased by around 80%. Moreover, even this bitcell itself is 40% larger; the resulting SRAM memory is 6% smaller in the area due to the integration of 256 rows per column rather than only 16 [29].

1.2.4 Transistor-Level Techniques

In [2], the channel length is increased to decrease the leakage current. However, this comes at the cost of performance in high-voltage design. In some CMOS technologies, such as the 90 nm CMOS technology [30], increasing the channel length improves the performance in the subthreshold region. Therefore, this technique is beneficial in low-voltage applications.



(a)

Figure 1.2: Schematic of the 6T bitcell.

A new logic gate that reduces the input gate signal swing is presented in [31]. This logic gate reduces the signal swing on high capacitive lines in the SRAM circuit to reduce the power consumption. A SRAM circuit fabricated in the 250 nm CMOS technology using this new logic gate dissipates 0.9 mW at 1 V while running at 100 MHz. The half-swing pulse-mode logic gate with self-resetting techniques used in this architecture show significant power savings without loss of performance. The main disadvantage of this technique is the need for level conversion. Another drawback is the reduced noise margin.

1.2.5 Subthreshold Bitcell Design

As mentioned earlier, the conventional 6T SRAM bitcell (shown in Figure 1.2) faces challenges operating at low voltages. SRAM parameters such as noise margin severely degrade at voltages lower than 0.7 V [9]. This is mainly because the read and write operations share a common access transistor within the conventional 6T SRAM bitcell. Extra transistors are introduced to the conventional 6T SRAM to enable read and write operations through different access transistors. Table 1.1, 1.2, and 1.3 summarizes the basic features of the proposed bitcells.

A significant time and resource consuming challenge in designing subthreshold bitcells is the amount of Monte Carlo simulations required to predict the stability of a bitcell during the read and write operations. This concern is addressed in [47] by providing a fast analytical method to estimate the failure probability of a SRAM cell due to parameter variations.

Table 1.1: Summary of New Bitcell Designs

Design	**Technology (nm)**	**Transistor Count**	**Size**	**E min**	**Frequency**	**V $_{\textbf{min}}$**	**Bitcell per BL**	**Area (μm^2)**	$I_{Leakage}(\mu A)$
JSSC 08 [32]	130	6T	2kb	0.78 pJ	21.5KHz	210mV	16	4.0068	N.A
JSSC 16 [33]	65	8T	32kb	1pJ	2.5us @ 0.2V	200mV	256	1.352	1.1
JSSC 08 [34]	65	8T	256 kb	136pJ	25KHz	350mV	256		6.28
JSSC 13 [35]	65	9T	2kb	0.57pJ	T_{acc} = 4.55μs (0.3 V)	220mV	64	2.93	0.05
JSSC 14 [36]	65	8T	128kb	17.6 pJ	N.A	370mV	256	1.53	N.A
JSSC 11 [37]	90	7T	8kb, 256 x 32	1.74pJ	13MHz (0.4V) 1MHZ (0.25V), 3.5MHz (0.3V)	250mV	256	2.62	
JSSC 13 [38]	65	L-Shaped 7T	32kb (256 x 128)	5.6 pJ	T_{acc} = 551 ns (0.26 V)	260mV	256	1.15×	
JSSC 06 [39]	90	7T	64kb	N.A	20 ns	440 mV	8	13% more than 6T	N.A

Table 1.2: Summary of New Bitcell Designs

Design	Technology (nm)	Transistor Count	Size	E min	Frequency	V$_{\text{min}}$	Bitcell per BL	Area (μm^2)	$I_{Leakage}(\mu A)$
JSSC 06 [26]	350	7T	64kb	13.6 mW	100 MHz @ 1.5 V	N.A	256		N.A
TCAS 14 [40]	40	12T	4 kb	1.91pJ	11.5MHz (3MHz write)	350 mV	16	4.42	58
JSSC 11 [41]	65	9T	4 kb	N.A	$T_{acc} =$ 500ns (0.25V)	250mV	256	1.4	300
A-SSCC 12 [42]	65	9T	16 Kb	2.07 pJ	0.85μs (0.26V)	260mV	256	1.89	1.4
JSSC 15 [43]	40	9T	72 kb	0.267 pJ/bit	600KHz (0.325)	325mV	32	0.8446	14.43
JSSC 05 [44]	180	10T	16kb	N.A	164Hz	180 mV	Hierarchical	N.A	N.A
JSSC 06 [45]	65	10T	256kb	1.75 pJ	400KHz	380mV	256	N.A	6.66
JSSC 07 [3]	130	10T	480kb	N.A	120KHz	200mV	1024	6.15	10.2

Table 1.3: Summary of New Bitcell Designs

Design	**Technology (nm)**	**Transistor Count**	**Size**	**E min**	**Frequency**	**V$_{\text{min}}$**	**Bitcell per BL**	**Area (μm^2)**	$I_{Leakage}(\mu A)$
JSSC 2007 [46]	130	10T	480kb	0.235pJ	600 kHz @ 400mV	160 mV	256	9.18	10.2μA
This Work	65	6T	2kb	1.1 fJ/b	9.2MHz @ 350mV	290mV	32	2.15	4.25 nA/b
This Work	65	6T	16kb	0.536 fJ/b	6 MHz @ 400mV	340mV (write) 360mV (Read)	128	1.38	3.43 nA/b

An accurate closed-form solution for the SNM of SRAM bitcell in the near/subthreshold region is derived in order to address this challenge.

A first attempt to enable low voltage operation of SRAM circuits is introduced in [44]. A Fast Fourier Transform (FFT) processor with SRAM subsystem is designed to operate at 180 mV at 164 Hz with a power consumption of 90 nW. The authors show the difficulty of both read and write operations of the 6T SRAM bitcell at voltages below 500 mV due to the susceptibility of the bitcell to process variation. To mitigate the problem of process variation, they utilize a multiplexer-tree based decoder to decrease the number of cells connected to the bitlines. However, this approach creates a significant area overhead and has an unacceptable performance for commercial applications [32] [48].

A single-ended 6T SRAM design with a gated-feedback write-assist is presented in [32]. This bitcell is fabricated in the 130 nm CMOS technology and shows robust operation at below 200-mV. Measurements of the fabricated test chip illustrate 36% improvement in energy consumption over the previously proposed multiplexer-based subthreshold SRAM design [44] while occupying half of the area. In the subthreshold region, the main component attributing to process variation is Random Dopant Fluctuation (RDF). In this design, to mitigate the effect of RDF, a single-ended cell with a gated-feedback write-assist is exploited in addition to transistor upsizing. It is shown that the transistor sizes must be increased by 6.5× at 0.3 V to reduce the noise margin variation acceptably.

A 7T read-SNM free SRAM cell is developed to overcome the speed limits of conventional SRAMs [39]. In this new bitcell, the threshold voltage of the NMOS transistors is reduced to the threshold voltage of logic gates to enable both high-speed and low-voltage operations. By adding the 7th transistor, the SNM of the bitcell is significantly improved during the read operation, and this new transistor also eliminates the half-selected issue at the write operation. In addition to the new transistor, the voltage level of the WL is also decreased during the read operation to improve the cell stability and SNM. However, the area overhead of this bitcell is 11% more than the conventional 6T transistor. Another drawback of this bitcell is its limited performance below 0.5 V. Due to the reduced performance, the number of bitcells connected to the BLs is reduced to 8.

Another 7T SRAM bitcell is provided in [26]. An NMOS transistor is introduced to the VSS node of the 6T bitcell. This reduces the BL swing to $V_{DD}/6$ and leads to 90% write power reduction.

The authors in [45] [49] propose a 10T bitcell that significantly improves the read-SNM by buffering the stored data during a read access. Therefore, the worst-case read-SNM is equal to 6T hold-SNM. The area overhead of this bitcell is 66% more compared to the conventional 6T bitcell. This architecture uses a full-swing single-ended read. One advantage of this bitcell is its reduced leakage-power, as compared to the 6T bitcell. Simulation results show 2.25× less leakage power at 0.6 V. In order to improve the impact of process variation, the level of WL voltage is boosted by 100 mV above the nominal supply voltage. To achieve write operation in the subthreshold region, the cell supply voltage is floated during the write operation. Measurement results present both read and write operations at below 400 mV while consuming 3.28 μW and running at 475 kHz.

A novel 10T SRAM bitcell with improved bitcell stability is proposed in [46]. This new bitcell uses a Schmitt-trigger technique to create a built-in feedback mechanism to assuage the effect of process variation. This new bitcell shows a 1.56× SNM improvement, as compared to the conventional 6T bitcell. Simulation results show that using a feedback mechanism can be more effective than transistor upsizing in a conventional 6T bitcell. A fabricated test chip in the 130 nm CMOS technology shows robust functionality at 160 mV of the supply voltage.

Kim et al. [3] propose a combination of several techniques to overcome the challenges of the conventional 6T bitcell operating at low voltage. To decouple the read path, four extra transistors are added to the 6T bitcell and the reverse short channel effect is exploited for WM improvement. Moreover, a virtual ground replica scheme for improved BL sensing margin is proposed. In addition, the BL leakage is independent of the data stored in the bitcell resulting in a high number of bitcells in each column. Measurement results show that 1024 cells on a BL is functional at 0.20 V running at 120 KHz (27C).

A subthreshold multi-threshold 9T bitcell is presented in [35]. The design of this bitcell allows the retention nodes to be disconnected from the BL during the read operation. To enhance the stability and reduce power consumption, the length of the back-to-back transistors are increased. To guarantee that the samples don't fail due to BL leakage, the number of bitcells per column is limited to 64 and 16. PMOS transistors are used as the access transistors since, as the simulation results show, PMOS transistors are less susceptible to process variations. For the 64 × 32 blocks, the minimum energy per operation occurs in the range from 0.30 V to 0.35 V, from 529 fJ to 620 fJ.

A differential 10T bitcell that effectively separates read and write operations is proposed in [50]. With the column-wise write access control, the proposed 10T SRAM cell allows bit-interleaving. This bitcell also allows a differential read path. To reduce the leakage current, the GND of the bitcell is virtually forced to $\mathbf{V_{DD}}$ during the hold mode and the virtual GND is forced back to 0 during the read operation. Measurement results show successful operation below 300 mV. With aggressive word line boosting, the supply voltage can be scaled down to 160 mV. This 10T bitcell is also exploited in [51], [52] where the leakage is measured as 1.83 pW/bit at 250 mV at 25 °C.

The authors in [38] propose an L-shaped 7T SRAM bitcell and a read-BL swing expansion scheme to minimize the area and supply voltage. This bitcell provides a decoupled 1T read port capable of providing a wide space for WM improvement. The read-BL swing expansion scheme utilizes a boosted BL to secure the sensing margins. The fabricated 65 nm 256-row 32-Kb L7T SRAM macro achieves a 260 mV minimum supply voltage.

A 12T subthreshold SRAM with data-aware-power-cutoff write assist is proposed in [40]. The data-aware-power-cutoff write assist scheme eliminates read disturb half-select issue. A 4-kb SRAM macro implemented in 40 nm general-purpose CMOS technology shows VDD_{min} for the read operation at 350 mV. The write operation can be performed at 300 mV. The maximum frequency is reported as 11.5 MHz with total power consumption of 22 μW at 350 mV. The minimum energy per operation is achieved as 1.6 pJ at 450 mV.

A symmetrical and differential 8T bitcell is proposed in [41]. This bitcell uses a zigzag shape layout to achieve a compact area and fully symmetric device placement for a litho-friendly layout. Due to the differential sensing, this bitcell can operate at a higher access speed compared to the conventional 8T bitcell [34]. In addition, for the same supply voltage, the proposed bitcell reduces the cell area by 15% compared to the conventional 8T bitcell. The measured minimum supply voltage for the 256-row 32-Kb macro and a 32-

row 4-Kb macro fabricated in 65 nm CMOS technology is 430 mV and 250 mV, respectively. The measured minimum supply voltage for a 256-row 64-Kb macro fabricated in the 90 nm CMOS technology is 230 mV.

Do et al. [33] propose a system-level approach to reduce the SRAM supply voltage for image and video-specific applications. In order to avoid the worst-case read scenario, the stored data in columns are randomized to make the distribution of the 0 and 1 s close to 50%. They show that the 8T bitcell in [34] can operate at 200 mV when utilizing data randomization.

A 9T SRAM bitcell with BL leakage equalization and Content-Addressable-Memory-assisted performance boosting techniques is presented in [42]. To improve the write performance, a CAM-assisted boosting technique is developed. The inserted tiny CAM conceals the slow data development after data flipping. This, in turn, improves the overall operating frequency. The fabricated 16-Kb SRAM in the 65 nm CMOS technology consumes a minimum energy of 0.33 pJ at 0.4V.

A single-ended 8T bitcell is presented in [34] that is capable of operating as low as 350 mV. This design suffers from low-speed single-ended sensing and is not able to assimilate half-selected cells. To overcome the BL leakage issue, a write assist technique is proposed.

A two-port disturb-free 9T subthreshold SRAM cell with independent single-ended read BL and write BL is presented in [43]. To enhance the writability of the proposed bitcell, variation-tolerant line-up write-assist scheme is employed. The 72-kb chip SRAM fabricated in 40 nm CMOS technology performs at 260 MHz (450 kHz) at 1.1 V (0.32 V) at 25 C.

1.2.6 Application-Specific SRAMs

In addition to all the techniques discussed above, there is additional room for improvement in energy consumption when exploiting the specific features of applications such as image processing. While designing SRAMs, these considerations can result in extra savings in terms of energy consumption, in addition to the savings already achieved through supply voltage scaling. These savings can be attained at the algorithm and architectural levels.

An embedded subthreshold SRAM for a quality-scalable and high-profile video decoder IP are presented in [37]. In addition to utilizing the conventional 7T bitcell, power-gating techniques and multi-output dynamic circuits are developed for achieving low energy, a small area overhead, and higher operating speed. The power/ground-gating techniques, as well as the conventional 7T bitcell, are exploited to reduce VDD_{min} with a small area overhead. The multi-output dynamic circuits are exploited to construct the address decoder for improving the operating speed. The SRAM circuit is fabricated in the 90 nm CMOS technology based on the techniques proposed in this paper. The SRAM provide an energy-efficient scalable video decoding of 42.8 pJ/cycle for QCIF, 78 pJ/cycle for CIF, and 235 pJ/cycle for HD720 at 0.3, 0.4, and 0.7 V, respectively.

The authors in [36] present a new optimization technique for applications where the data is highly correlated such as in video and imaging applications. A new bitcell topology is proposed that uses bit-wise prediction to reduce BL switching activity. Each row represents one word, and no half-selected cells are utilized. Also, a column multiplexing ratio of one is used, with a sense amplifier is assigned to each column. During a read operation if a correct prediction is performed, no voltage difference is introduced across the read buffer connected to the BL. Hence, with correct prediction, none of the BLs are discharged, and the switching activity on the BLs is prevented. To achieve further improvement, a statistically gated sense amplifier approach is developed. This approach takes advantage of the biased transition probabilities on the bitlines. These techniques reduce the energy/access consumption by up to 1.9×, as compared with the traditional 8T bitcell.

Chapter 2

SRAM Architecture and Circuit Implementation

This chapter presents the basics of the CMOS SRAM architecture and circuit implementation. Section 2.1 explains the main architecture and basic blocks that are used to construct an SRAM circuit. Sections 2.2, 2.3, 2.5, 2.6, and 2.7 explain each block in more detail.

2.1 SRAM Circuit Architecture

The SRAM architecture shown in Figure 2.1 is composed of the following blocks:

- Address buffers
- Row decoder
- SRAM array consisting of bitcells
- Read/Write column decoder
- Sense Amplifier (SA) array
- Input/Output data buffers

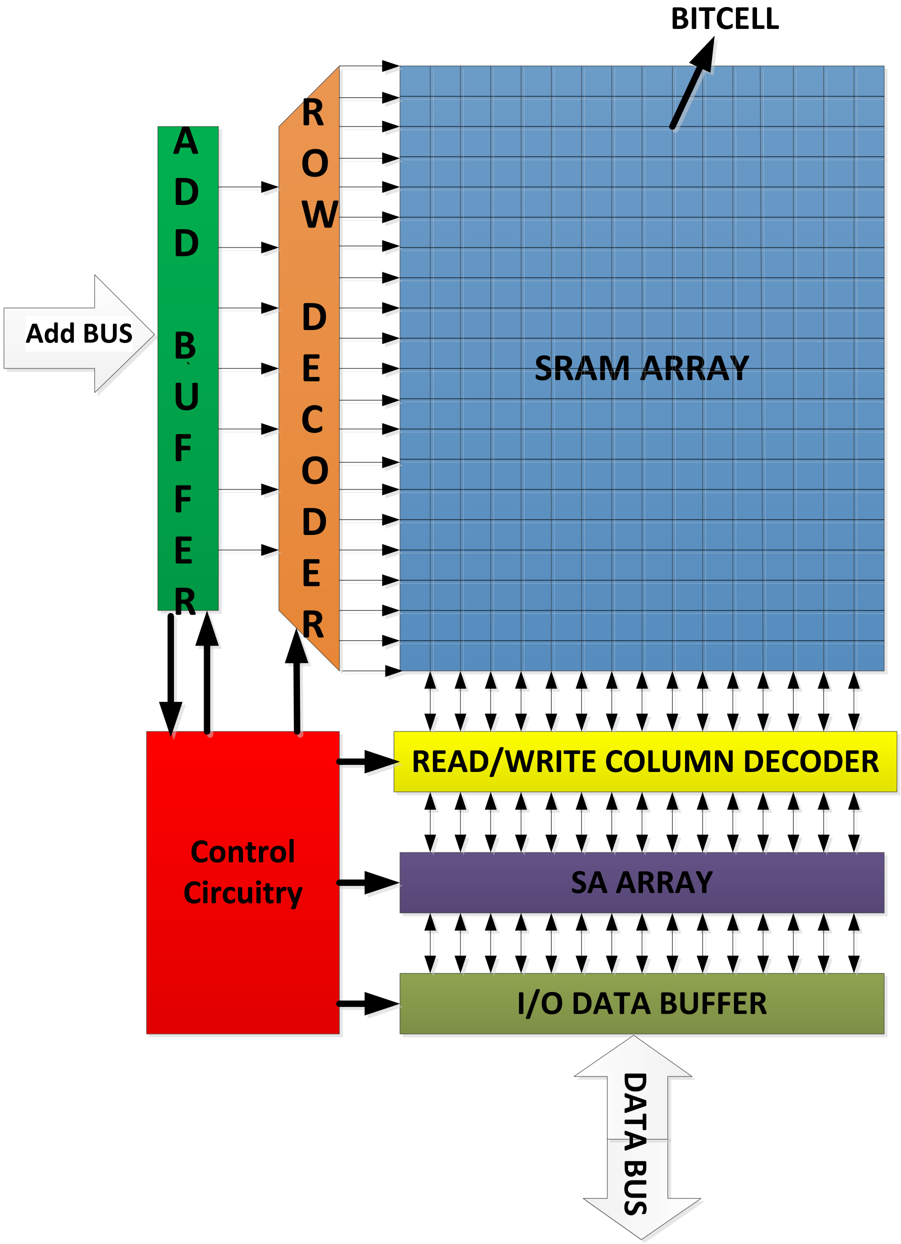
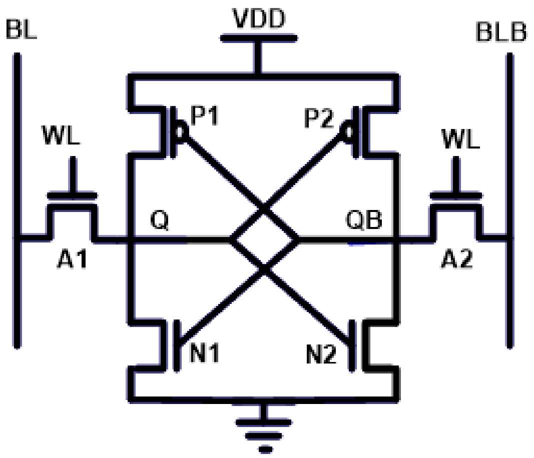


Figure 2.1: Diagram of a SRAM architecture.



(a)

Figure 2.2: Schematic of the 6T bitcell.

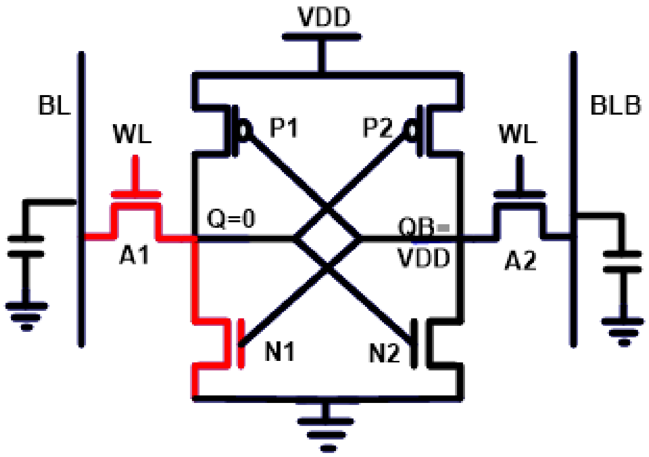
2.2 SRAM Bitcell and Array Design

A SRAM array is composed of multiple rows and columns of SRAM bitcells. All bitcells in the same column share the same BL and Bitline-Bar (BLB). The bitcells on each row share the same WL. A conventional SRAM bitcell with 6T is shown in Figure 2.2. The SRAM bitcell comprises of two back-to-back inverters (P1, N1, P2, N2) forming a latch to hold the data, and two access transistors (A1, A2). The data is stored at nodes Q and QB. A SRAM bitcell has three modes of operation as described below:

- Retention Mode: A SRAM cell retains the data indefinitely as long as it is powered.
- Read Operation: The data of the bitcell is read during a read operation while the data should remain stable.
- Write Operation: The data of the bitcell is set to a certain value regardless of its original value.

2.2.1 Read Operation

Figure 2.3 shows the 6T bitcell during the read operation. During a read operation, initially, the BLs are precharged to the high voltage level (typically $\mathbf{V_{DD}}$). A read operation is



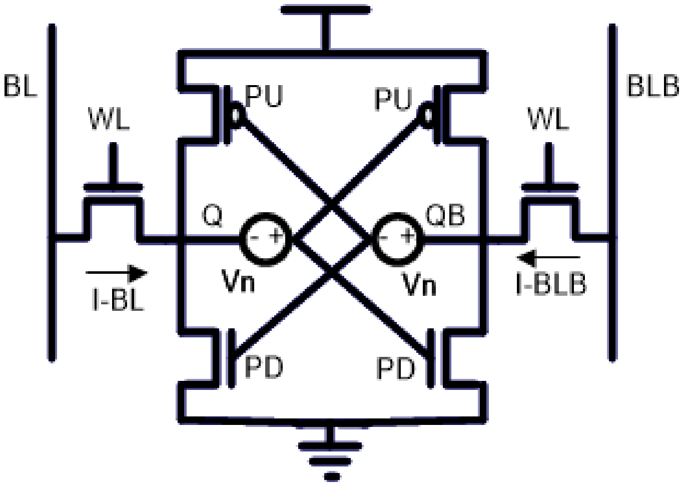
(a)

Figure 2.3: 6T SRAM bitcell during read operation.

initiated upon the activation of the WL signal. The WL signal turns the access transistors ON, and a discharging path is created from the BL capacitance through the access (A1) and the driver transistor (N1) to GND. This path is shown in red in Figure 2.3. BLB remains at $\mathbf{V_{DD}}$ while the BL discharges. During this process, node Q acquires a potential higher than zero known as ZLD. A larger ZLD can adversely affect the read stability of a SRAM bitcell. Therefore, it is desirable to keep the ZLD close to the GND level. This is usually done by keeping the width of the driver transistor larger than the access transistor. The read operation finishes when the sense amplifier is enabled after the differential voltage between the BL and BLB is sufficiently developed. The sense amplifier amplifies the small developed differential voltage (usually about 100 mV) to full swing at its outputs.

2.2.2 Write Operation

A write operation is initiated by activating the write driver to discharge either the BL or BLB to 0 and activating the WL. Once the WL is activated, the BLs force the data in the internal nodes (Q and QB) to flip if necessary. The positive feedback mechanism of the back-to-back inverters accelerates the voltage-level degradation and enhances the data flip speed. It is worth mentioning that during a write operation, the WL of all bitcells on the same row is activated. However, only those bitcells located on the selected columns undergo a write operation. The bitcells located on the non-selected columns, known as half-selected cells, perform a normal read operation called read access where the BLs



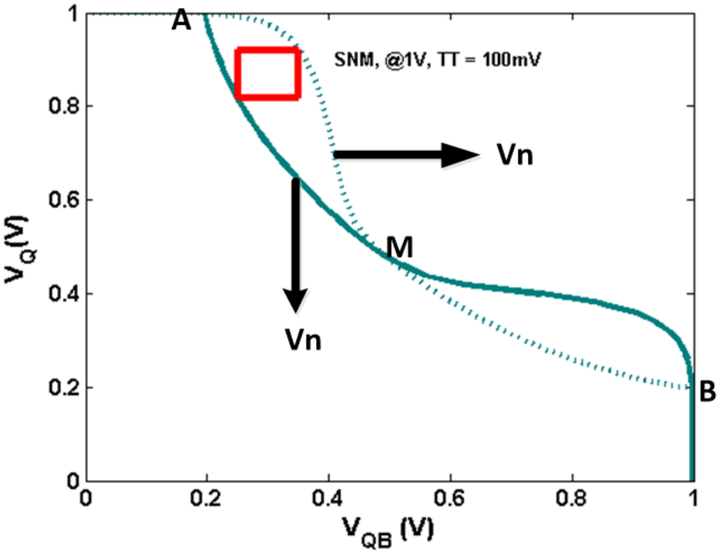
(a)

Figure 2.4: 6T bitcell with two differential noises.

develop the differential voltage, but the sense amplifiers are disabled. The read and write operations in a conventional 6T SRAM cell have contradicting requirements. A successful read operation requires large driver transistors (N1 and N2 in Figure 2.3) and weak access transistors (A1 and A2 in Figure 2.3), whereas a successful write operation requires strong access transistors and weak load transistors (P1 and P2 in Figure 2.3). Additionally, the data retention operation requires a reasonably strong driver and load transistors. As such, a delicate device sizing approach must be adopted to ensure a stable and functional SRAM cell with sufficient read, write and retention noise margins.

2.2.3 Static Noise Margin During Read Operation

The SNM is the maximum amount of voltage noise that can be introduced at the internal nodes of the two inverters such that the cell still retains its data. Figure 2.4 shows a conceptual setup for modelling the SNM [53]. Noise sources with value $\mathbf{V_n}$ are introduced at each of the internal nodes in the bitcell. As $\mathbf{V_n}$ increases, the stability of the bitcell reduces. To plot the butterfly curves the BL and BLB are connected to $\mathbf{V_{DD}}$ and both access transistors are active. As explained in [53], the Voltage Transfer Characteristic (VTC) and inverse VTC (VTC^{-1}) are plotted. To plot the VTC, we plot $\mathbf{V_{QB}}$ versus V_Q by sweeping V_Q and for plotting the VTC^{-1}, we plot V_Q versus $\mathbf{V_{QB}}$ by sweeping $\mathbf{V_{QB}}$. The resulting two-lobed curve shown in Figure 2.5 is called a "butterfly curve" and is used to determine the SNM. The SNM is defined as the length of the side of the



(a)

Figure 2.5: Butterfly curves of the 6T bitcell during read operation.

largest square that can be embedded inside the lobes of the butterfly curve [53]. Butterfly curves have two stable points (A and B) and one meta-stable point (M). To have a better understanding, consider the case when the value of $\mathbf{V_n}$ increases from 0. On the plot, this causes the VTC^{-1} to shift downward and the VTC to shift to the right. As $\mathbf{V_n}$ increases, the metastable point moves closer to one of the stable points in the plot (point B in this example). Once both curves move by the SNM value, the metastable point coincides with one of the stable points, and the curves meet at only two points. Any further noise flips the cell data.

2.2.4 Write Margin

During the write access mode, the cell WM defines the voltage limit required to flip the cell data. This can be accomplished by reducing either the BL voltage or the cell's supply voltage $\mathbf{V_{DD}}$. In other words, the WM is defined as the lowest voltage level required to flip the cell data. Graphically, the WM can be quantified by calculating the length of the maximum square that can be embedded between the read and write VTC curves, as shown in Figure 2.6. During a successful write operation, there are no lobes on the butterfly curve.

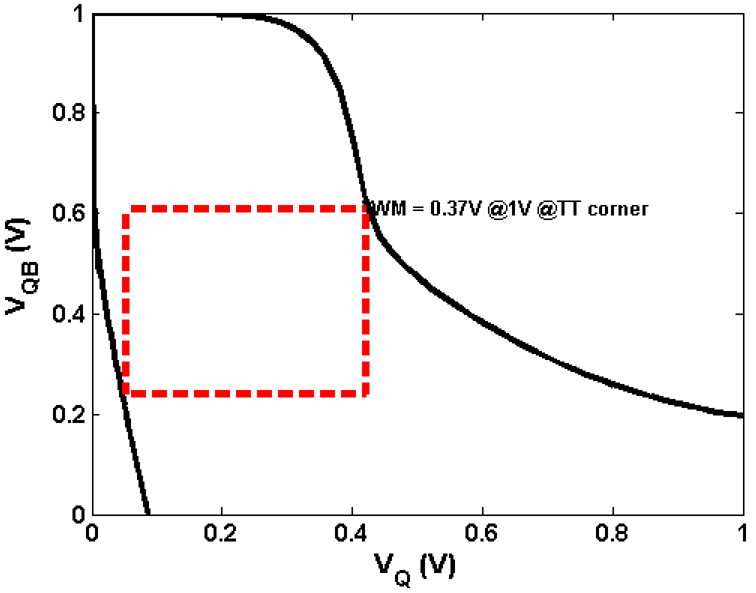


Figure 2.6: Write margin of the 6T bitcell at TT corner at 1 V

If the VTC and VTC^{-1} curves on the plot shift by an amount equal to the WM, then the cell will regain bistability.

2.3 Address and Data Buffers

In order to perform correct read and write operations, it is necessary to avoid any changes in the address and input data during the read and write operations. This is done by using latches that store the address and data signals and are disconnected from any changes from outside of the chip with a control signal. For this purpose, a D-latch is used, for each signal, as shown in Figure 2.7. When the control signal (CTL) is high, any change on the input propagates to the output. However, when the CTL signal is deactivated, the pass-gate (PG1) disconnects the input from the rest of the circuit, and the data is stored by the loop created by INV1, INV2, and PG2. The output data buffer is also followed by a tri-state buffer to avoid connecting two outputs to the bus at the same time. Figure 2.8 shows the implementation of a tri-state buffer and a tri-state inverter. In Figure 2.8(a), depending on the state of the Output Enable Bar (OEB), the output may enter the high-impedance mode. When the OEB signal is low, the output signal goes into

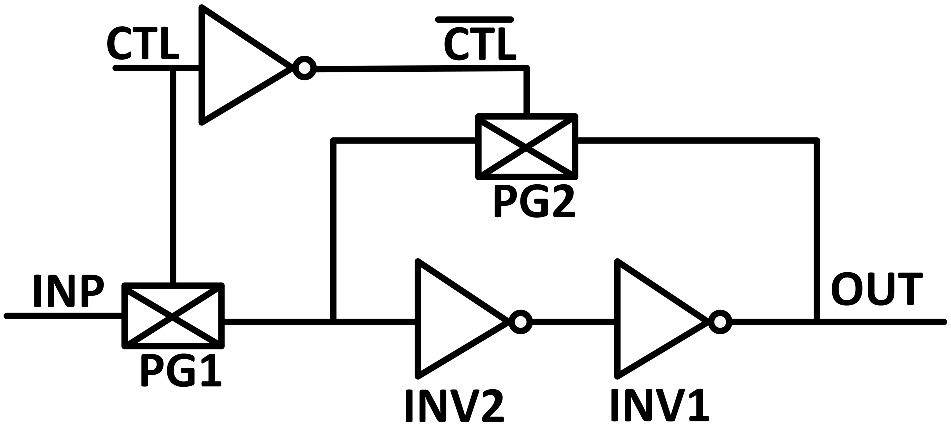
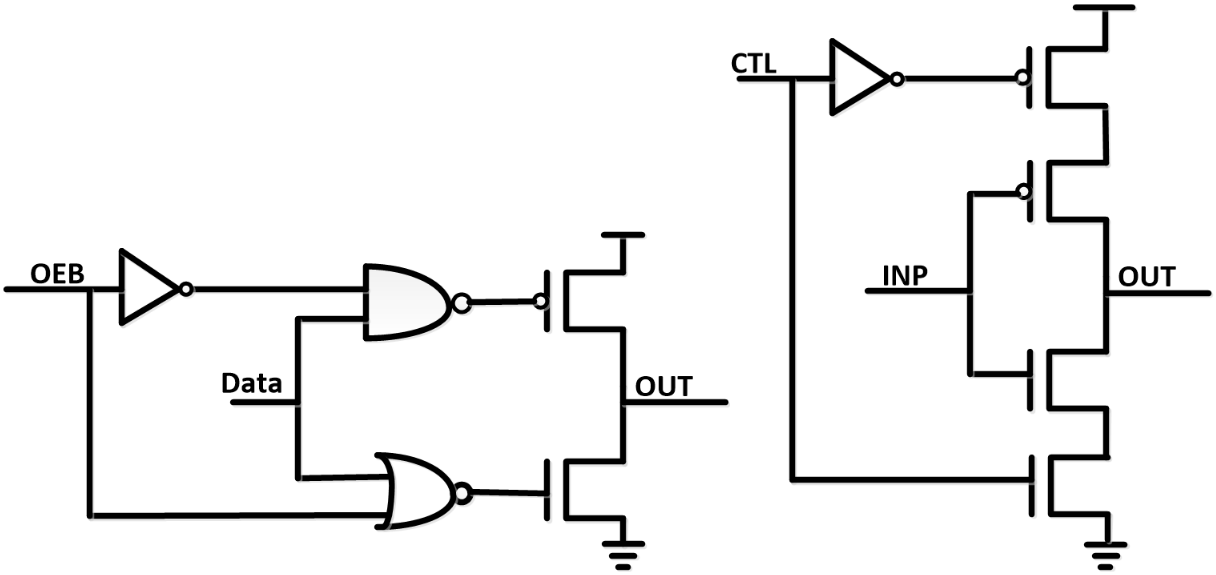


Figure 2.7: D-latch implementation.



(a)

(b)

Figure 2.8: Implementation of (a) tri-state buffer and (b) tri-state inverter.

the high-impedance mode, and when the OEB signal is at $\mathbf{V_{DD}}$, the DATA signal is copied to the output. Figure 2.8(b) shows another implementation of the tri-state buffer. When the CTL signal is high, the inverted input is propagated to the output. Otherwise, the output remains in a high-impedance mode.

2.4 Row Decoder Design

A row address decoder is used to activate one out of N rows in the memory array. Decoders are designed in two stages: pre-decoder and post-decoder. The outputs of the pre-decoder are combined to create the outputs of the post-decoder. In the decoder design, six main parameters characterize the longest path, speed, and power consumption [54]. These six parameters are listed below followed by a detailed explanation:

1. Choice of logic gates in each decoding stage
2. Logic depth
3. Fan-in of each decoding stage
4. Fan-out of each decoding stage
5. Geometries and resistivity of wires driven by each decoding stage
6. Device sizes within pull-up and pull-down networks in each stage along the decode path

Choice of logic gates: The logic gates used to implement the decoders vary from dynamic logic to static logic to pulsed and self-resetting logic. Clocked decoding is also used as another alternative to CMOS gates. Most decoders that are implemented using CMOS gates use NAND gate followed by an inverter.

Logic Depth: The logic depth is determined by the number of WLs to be decoded as well as the average fan-in of the logic (NAND, INV) gates along the decode path.

Fan-in: A fan-in of two minimizes the decoder delay [55]. Increasing the fan-in of each NAND gate increases the fan-out of internal nodes. The gates connected to higher fan-outs are required to be sized-up proportionally and that translates into a larger area. Moreover, increasing the fan-in increases the gate delay.

Fan-out and wire length: The fan-out of each decoder stage and the maximum wire-lengths driven by each stage are determined by the architecture of the decoder.

Device sizes within pull-up and pull-down networks: Different sizing techniques, such as logical effort, can be used to optimize the total delay along the decode path [56]. Optimal device widths depend on the logic, fan-in, and fan-out of the gate used and the parasitic wiring being driven by each gate.

Figure 2.9 shows a 7-to-128 row decoder. All the outputs of the decoder have to be deactivated before the control signal (CLK-EN) is set. The CLK-EN signal activates the enable signals (En1 and En2) and allows one of the outputs associated with the input address of the decoder to be activated. The timing of the CLK-EN signal is set by the control circuitry.

2.5 Read/Write Column Decoder and Write Driver

A read column decoder in a SRAM uses a 2^K-input multiplexer where the inputs are the BLs, and the output is the SA inputs. A read column decoder allows several columns to be connected to a single SA and thereby, relaxes the area constraints on the SA design. An example of a read column decoder is shown in Figure 2.10. As shown in this figure, a SA is assigned to two columns. The R0 and R1 signals chose between the two columns, and the corresponding BLs are provided to the SA inputs. The SAE0 and SAE1 signals choose which SA is be activated and its output to be connected to the output bus. Therefore, in each read operation, one out four columns are read.

The read operation starts after the precharge phase in which the BLs are precharged to $\mathbf{V_{DD}}$. When the WL is activated, the BLs start to develop the differential voltage. The differential voltage is transferred to the corresponding SA inputs after one of the R0 and R1 signals is activated. The read operation finishes after one SA is activated by activating one of the enable signals (SA0 or SA1).

During a write operation, the W0, W1, WriteEnable0, and WriteEnable1 signals connect the input data and its complement to the BLs of one column out of four. The write operation completes by activating the WL causing the data on the BLs flip the data in the bitcells. The write driver consists of two NAND gates. The NAND gates are sized such that they are strong enough to discharge the BL capacitance to 0.

2.6 Sense Amplifier Design

The primary function of the SA in the SRAMs is to amplify a small analog differential voltage to a full-swing digital output signal. This avoids a full-swing discharge on the high capacitive BLs, and therefore a significant amount of power consumption is saved.

Special attention is given to the SA area in SRAM circuits. Architectures that do not use column multiplexing are required the SA to fit within in a column pitch. How-

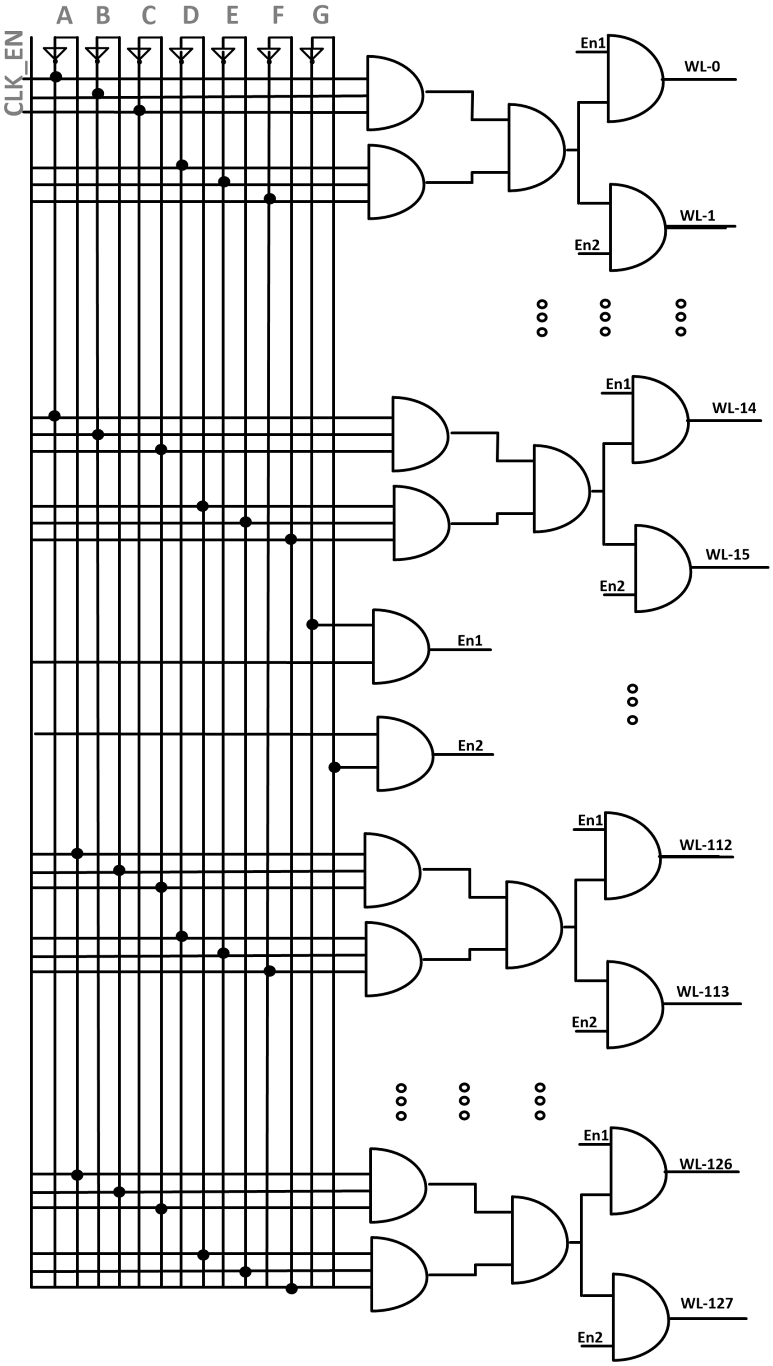


Figure 2.9: 7-to-128 row decoder.

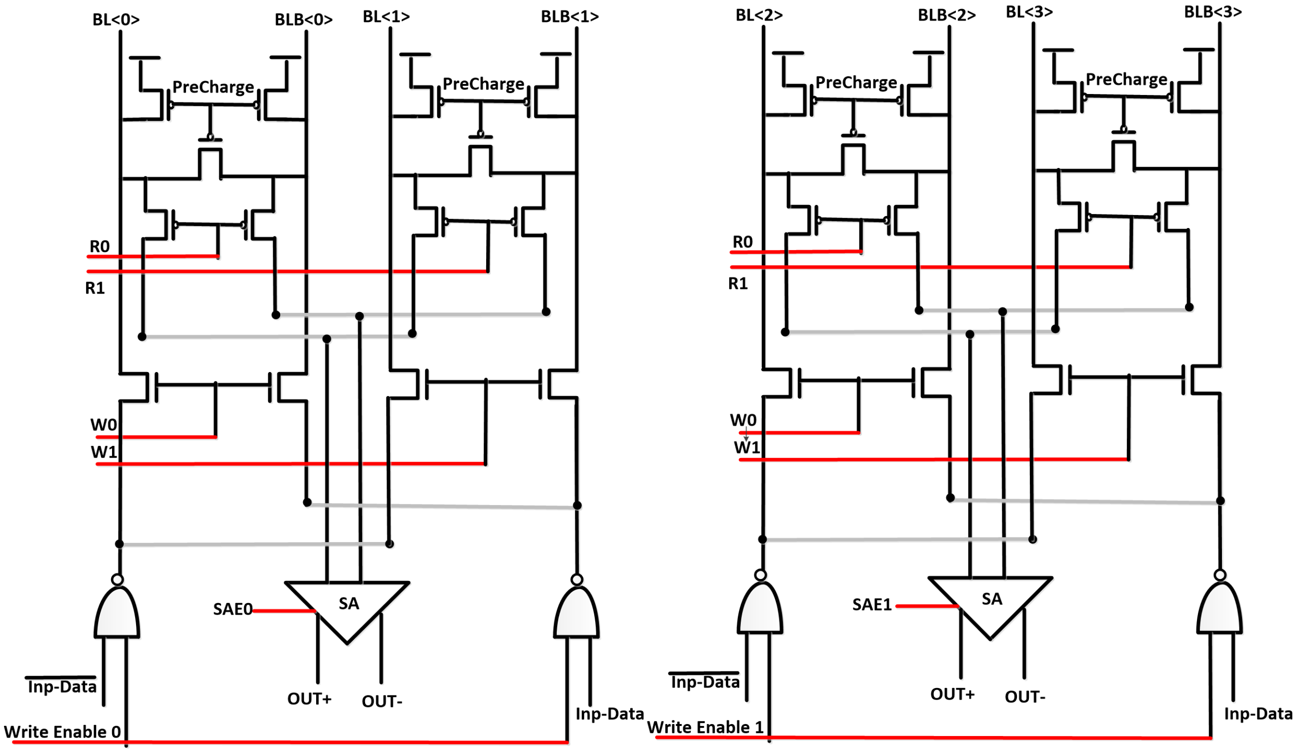


Figure 2.10: Read and Write decoder and write driver.

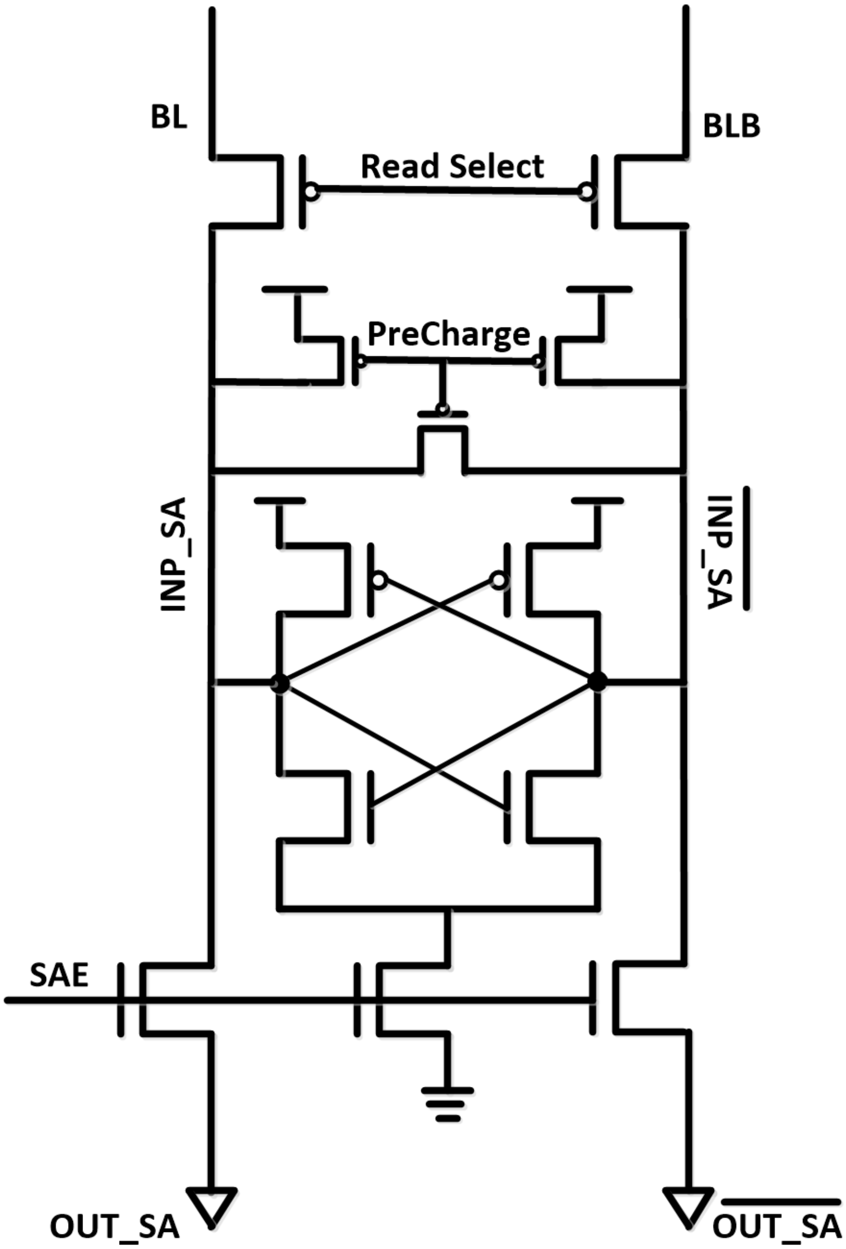


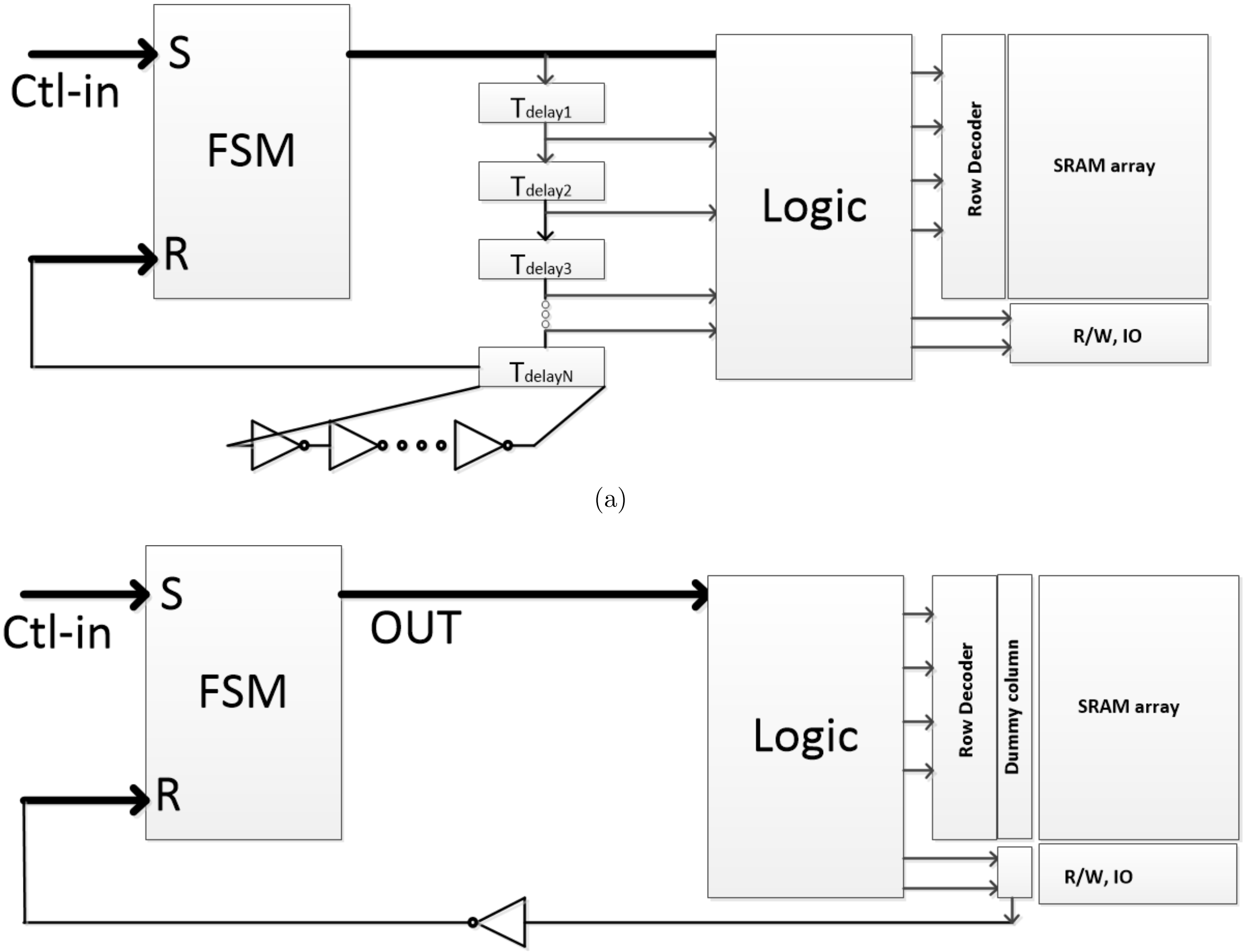
Figure 2.11: Sense amplifier schematic

ever, utilizing column multiplexing relieves this constraint by assigning each SA to multiple columns. High sensitivity to process variations in the subthreshold region can inject common-mode noise to both SA inputs. In designing SAs operating in the subthreshold region, differential sensing reduces in the impact of the common-mode noise that may present on both BLs. Figure 2.11 shows the schematic of a common SA that is used in SRAM architecture. The sensing operation begins with setting the SA operation point by precharging and equalization of both inputs of the SA to the identical precharge voltage level ($\mathbf{V_{DD}}$). Next, the decoded WL of a read-accessed cell is activated starting the build-up of the differential voltage on the BL and BLB. The Sense Amplifier Enable (SAE) signal is issued after a sufficient differential voltage is developed on the inputs. As a consequence, the amplification of the small signal to full swing output is performed, and the output data becomes available on the data bus.

2.7 Control Circuitry

The timing control circuitry provides the timing of the precharge, row-decoder enable, SAE, and write-enable signals, and ensures a correct read and write operation. The two main methods used for implementing the control circuitry are based on delay-line timing control [57] and asynchronous replica timing techniques [55]. The schematic of the delay-line timing loop is shown in Figure 2.12(a). A control signal, which is usually the main clock signal, sets the FSM. The total timing is defined by the total delay elements (Tdelay1-TdelayN) in the FSM reset path. The delay elements are usually constructed by a chain of logic circuits (INV, NAND, NOR). The delay time can be extended by using non-minimal length devices in the delay chain. The timing intervals constructed by the delay elements are used to generate the control signals for the read/write control signals. The drawback of this method is that the delay of the delay loop may not track the delay variations of the SRAM bitlines caused by the process variations in modern nano-scaled technologies.

The asynchronous replica timing circuit provides a tighter tracking of the bit line discharge delay and alleviates the effects of process variations. The schematic of this timing method is presented in Figure 2.12(b). A replica (dummy) column is used to track the same number of SRAM cells in each column as the reference delay element. The replica signal path mimics the capacitive loads on the BLs and the associated delays of the real signal path. Therefore, it can provide more precise timing signals. Similarly to the delay-line based method, control signal (Ctl-in) sets the FSM. The output signal initiates the word lines both in the row decoder and in the dummy row. The dummy column provides a reset signal to the FSM after its BL is discharged. By resetting the FSM, the SRAM



(a)

(b)

Figure 2.12: (a) Delay-line timing loop (b) Asynchronous replica timing circuit.

enters into the precharge phase and the SA completes it operation by driving the data on the data bus.

Chapter 3

A 16kb SRAM with Programmable Wordline Boost for Energy Efficient Applications

Embedded SRAMs are essential parts of a modern System on Chip (SOC) as they significantly affect the SOC's performance, energy consumption, reliability, and yield. The aggressive demand in portable devices and billions of connected sensor networks requires long battery life. Therefore, there is a critical need for the design of SRAM circuits that entail minimal energy consumption with little or no performance cost.

Several architectural approaches have promisingly demonstrated energy reduction in SRAM circuits. In [4], the authors show that by simple micro-architectural techniques, the leakage energy consumption can be reduced by 75%. In [6], the leakage power is decreased by reducing the DIBL effect. Measurement results show about 10% leakage current reduction. A hierarchical bitline and local sense amplifier scheme are presented in [16]. This scheme reduces both the capacitance and write swing voltage of bitlines resulting in lower write power consumption without noise margin degradation. The authors in [19], show that large signal sensing is also a viable option as opposed to small signal sensing in the deep sub-micron regime. The new scheme creates a small signal swing on the local BLs and creates a large signal swing on the global BLs with reduced capacitance.

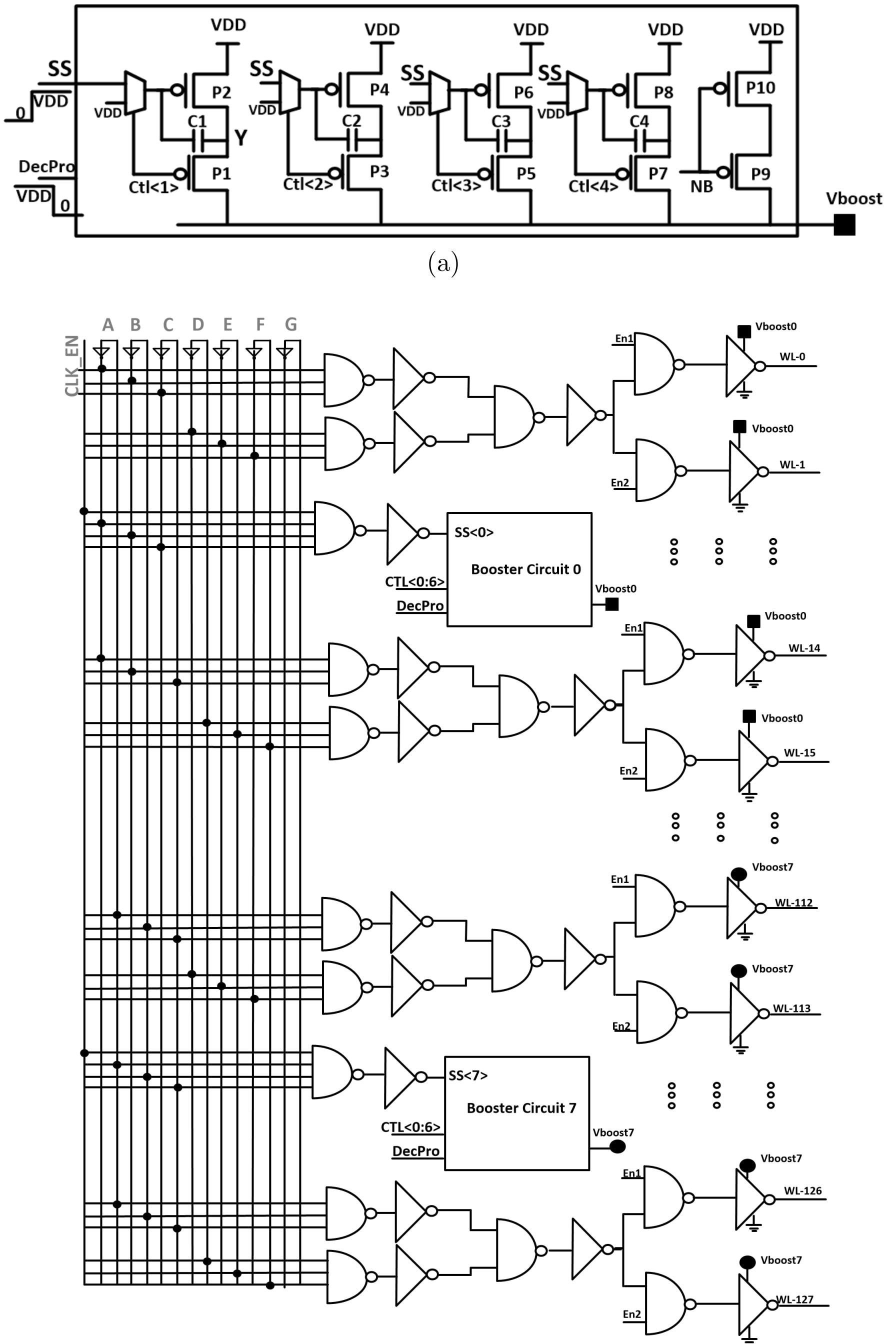
Another prevalent approach to reduce the energy consumption of SRAM circuits is to reduce the power supply into the near or subthreshold region [58]. Nevertheless, reducing the power supply voltage in SRAM requires careful consideration owing to its data stability during the read operation and write margins. The conventional 6T SRAM bitcell has contradictory requirements for read stability and writability. This contradiction becomes even more challenging in the subthreshold region. To estimate the failure probability of the 6T bitcell's stability, a fast analytical closed-form solution in the subthreshold region is provided in [47]. Another limitation of SRAM blocks operating in the near-threshold and subthreshold regions is that their low-energy requirements necessitates the development of a near or subthreshold circuit operation with an acceptable performance, to perform complex tasks [59].

The design in [23] utilizes a two-step WL boosting to overcome this conflict and improve the frequency of the operation. The divided bitline scheme used in this architecture reduces the capacitance on the bitlines by a factor of four which, in turn, reduces the power consumption and increases the read stability by decreasing the amount of charge flow to the selected bitcells. The designs proposed in [32] and [34] have reduced the supply voltage and improved both read and write margins. The design in [32] uses two back-to-back inverters and a pass-gate as an access transistor. The bitcell is significantly over-sized to make the design variation-tolerant in the subthreshold region. The 8T bitcell designed in [34] uses a separate path for the read operation, providing improved data stability during the read operation. The single-ended sensing of both designs in [32] and [34] does not allow the incorporation of half-selected cells.

It is shown in [60] that utilizing WL boosting results in a 28.5% improvement in the developed bitline differential voltage and a 39% reduction in cell leakage current. A selective WL boosting is proposed in [61]. This approach shows a 80% reduction in yield losses. In [62], the design employs a boosted WL technique for improving both read performance and writeability. An adaptive voltage detector (AVD) with a binary boosting control is used to mitigate gate electric over-stress.

In this chapter, a four-level programmable WL boosting technique is proposed that can further improve the above mentioned contradictory requirements of the 6T bitcell. Incorporating programmability enables a process-tolerant design; and optimization of the read and write margins independently. Moreover, the 6T bitcell does not have to be over-designed for low-voltage operation. The measurement results on a 16-kb SRAM shows that the WL boosting reduces the minimum supply voltage for write operation down to 330 mV at a speed of 6 MHz.

The rest of the chapter is organized as follows. The booster circuit implementation is discussed in Section 3.1. The effect of WL boosting on the propagation delay is analytically investigated in Section 3.2. In Section 3.3, the effect of the temperature on WL boosting is investigated. Section 3.4 presents the measurement results. Finally, conclusions are drawn in Section 3.5.



(a)

(b)

Figure 3.1: a) Booster circuit b) An implementation of 7-to-128 bit decoder with booster circuit.

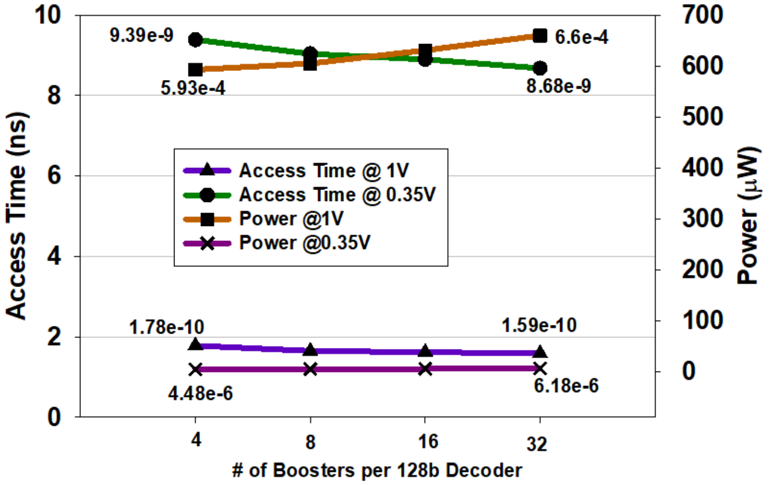


Figure 3.2: Access time and power consumption versus different number of boosters at 1 V and 0.35 V.

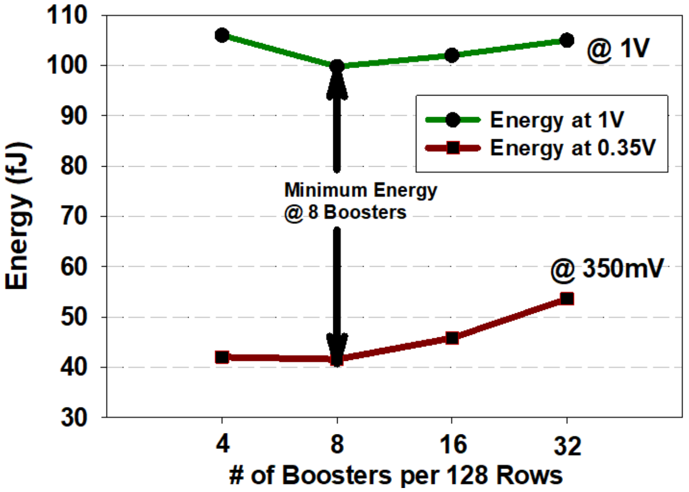


Figure 3.3: Energy consumption versus number of boosters at 1 V and 0.35 V. Minimum energy occurs when the number of boosters is 8.

3.1 Decoder and Booster Design

Figure 3.1(a) shows the proposed booster circuit and Figure 3.1(b) shows a 1128-row decoder with a booster circuit. The amount of the boosted voltage is proportional to the number of boosters per 128 WL drivers. Increasing the number of boosters increases the level of the boosted voltage. This, in turn, decreases the access time and increases the total power consumption. Fig. 3.2 shows the access time and power consumption of the 7-to-128 row decoder versus the number of boosters at 1 V and 350 mV. The access time is measured when the BLs develop 100 mV. As shown in Fig. 3.2, the minimum access time is achieved when the number of boosters is equal to 32. This number of boosters also gives the maximum amount of power consumption. Fig. 3.3 shows that the minimum energy consumption is achieved when the number of boosters is equal to 8.

The booster circuit shown in Fig. 3.1(a) consists of four Miller capacitances (C1= 200 fF, C2= 300 fF, C3= 400 fF, and C4= 500 fF) corresponding to four-levels of boosted voltage. These four levels are controlled by four control signals (CTL<1:4>) that are externally programmable. When any of the boosting controls are active (CTL<1>-CTL<4>), the Vboost is boosted to a value higher than the supply voltage (V_{DD}) and when CTL<5> is active, Vboost is equal to V_{DD}. it is assumed, without loss of generality, that CTL<1> is active. The select signal (SS) is initially at 0 and node Y is at V_{DD}. When the SS signal makes a transition to V_{DD}, the transistor P2 turns off, and due to the Miller capacitance (C1), the voltage of node Y goes higher than V_{DD} and this voltage is conveyed to the node Vboost through P1.

Fig. 3.4 shows the transient simulation of the WL and the corresponding BLs when the four levels of boosting are applied. The voltage-level of the WL increases and the corresponding BL discharges faster as the level of boosting increases. As shown in this figure, the access time is reduced by 28%, 34%, 37%, and 39% when level 1, level 2, level 3, and level 4 of boosting are applied, respectively.

A foundry provided metal-insulator-metal (MIM) capacitor is utilized for the Miller capacitance. The MIM capacitors are constructed with the top layer metals. As such, they are capable of being positioned on top of the decoder with no area overhead. Unlike the MIM capacitor, the MOS capacitor used in [60] is constructed with low-level metals and cannot be positioned on top of the array or decoder. Therefore, utilizing the MOS capacitor increases the decoder area by 9%.

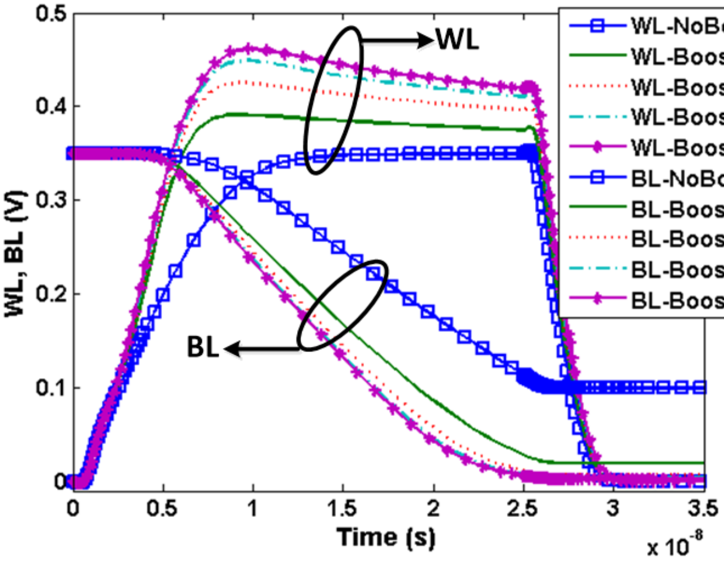


Figure 3.4: Simulated timing of WLs and BLs for boosted and non-boosted options at 350 mV.

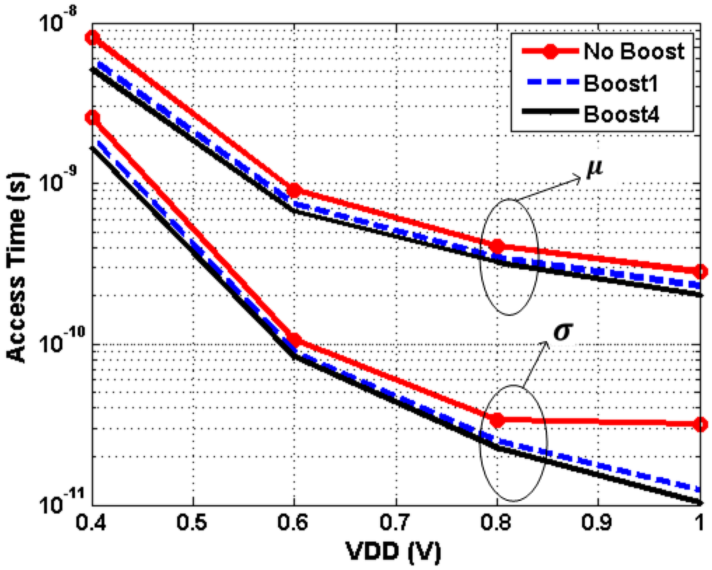


Figure 3.5: Monte Carlo simulation results (μ and σ) of access time versus supply voltage with different levels of WL boosting.

3.2 Analysis of the Effect of WL Boosting on Propagation Delay

As explained in [63], the subthreshold current of a MOSFET transistor has a log-normal distribution (LogN (μ,σ^2)) and its mean and variance values are defined as

$$\mathbf{E[I] = I_0 e^{\frac{(V_{GS} - \mu(V_{th}))}{nU} + \frac{\sigma^2(V_{th})}{2(nU)^2}}} \tag{3.1}$$

$$\mathbf{VAR[I] = (e^{\frac{\sigma^2(V_{th})}{2(nU)^2}} - 1)(E[I])^2} \tag{3.2}$$

The propagation delay of a logic gate can be calculated as [64]:

$$\mathbf{t_p = \frac{CV_{DD}}{I}} \tag{3.3}$$

The read access time of an SRAM bitcell can be calculated by Equation 3.3 where C is the BL capacitance and I is the current through access (or driver) transistor. Since t_p is inversely proportional to the current I, it has a log-normal distribution with a mean of -μ and a variance of σ^2 (i.e., LogN(-μ,σ^2). Therefore, the mean value and the variance of the propagation delay can be calculated as

$$\mathbf{E[t_p] = CV_{DD}\frac{1}{I_0} e^{\frac{(-V_{GS} + \mu(V_{th}))}{nU} + \frac{\sigma^2(V_{th})}{2(nU)^2}}} \tag{3.4}$$

$$\mathbf{VAR[t_p] = CV_{DD}(e^{\frac{\sigma^2(V_{th})}{2(nU)^2}} - 1)(E[t_p])^2} \tag{3.5}$$

As shown in Equation 3.4 and Equation 3.5, WL boosting (i.e., increasing the V_{GS}) decreases the mean value and the variance of the propagation delay. Fig. 3.5 plots the μ and σ of the access time versus V_{DD} with no boost and two levels of boosting. As shown in this figure, by increasing the supply voltage and also increasing the boosted voltage, both μ and σ decrease.

3.3 Temperature Effect on WL Boosting

Fig. 3.6 shows the boost voltage variation with respect to the temperature. This figure shows that the boost voltage decreases by 12% and 7% at 350 mV and 1 V, respectively. Fig. 3.7 shows the access time versus temperature at 350 mV and 1 V. As shown in this figure, the access time decreases at 350 mV while it increases at 1 V when the temperature increases. To further analyze the opposite behavior of the access time versus temperature at different supply voltages, the threshold voltages of the NMOS and PMOS transistors at these two voltages are plotted in Fig. 3.8. Fig. 3.8 shows that the threshold voltage of both the NMOS and PMOS transistors decreases at 350 mV and 1 V as a function of the temperature. The current of the PMOS and NMOS transistors in the 65 nm CMOS technologies versus temperature is depicted in Fig. 3.9. As shown in this figure, the current of the PMOS and NMOS transistors increases while the temperature increases at 350 mV. However, the current of the NMOS and PMOS transistors decreases with temperature at 1 V. The transistor mobility decreases by increasing the temperature as explained in [64] and [65] ($\mu \propto T^{-2.4}$). At 1 V (super-threshold region), where the MOSFET current is linearly proportional to the threshold voltage, the effect of the mobility on the current dominates the effect of the threshold voltage on the current. However, at 350 mV (i.e., in the subthreshold region), where the MOSFET current is exponentially proportional to the threshold voltage, the effect of threshold voltage dominates the effect of the mobility. Therefore, the MOSFET current has an opposite behavior with respect to the temperature in the subthreshold region versus the superthreshold region. Considering Equation 3.3, since the MOSFET current (I) has a more dominant effect on the access time, as compared to the small effect (7 to 12 %) of WL boosting, by increasing the temperature, the access time decreases in the subthreshold region and increases in the super-threshold region.

3.4 Measurement

A test chip with a 16-kb SRAM was designed and fabricated using the TSMC 65 nm GP CMOS technology. The I/Os in this technology operate at 2.5 V and are capable of interfacing with the core logic at 1.0 V. The level shifters are capable of shifting a 200 mV input to 1.0 V, and vice versa. The sizing of the 6T bitcell and its layout are shown in Fig. 3.10(a-b). The die photo is shown in Fig. 3.11.

Figure 3.12(a) shows the measured maximum operational frequency versus supply voltage when different levels of boosting are exploited. As shown in this graph, the frequency increases when the boost voltage increases.

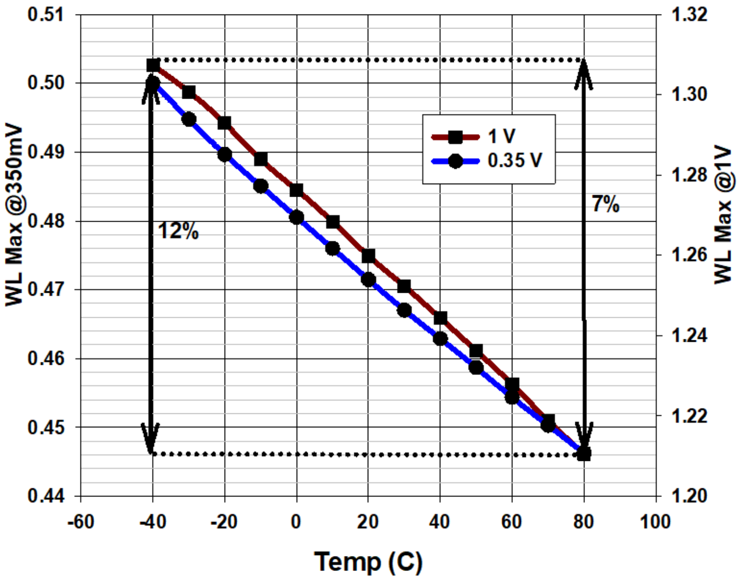


Figure 3.6: Maximum WL voltage at Boost4 versus temperature at 350 mV and 1 V.

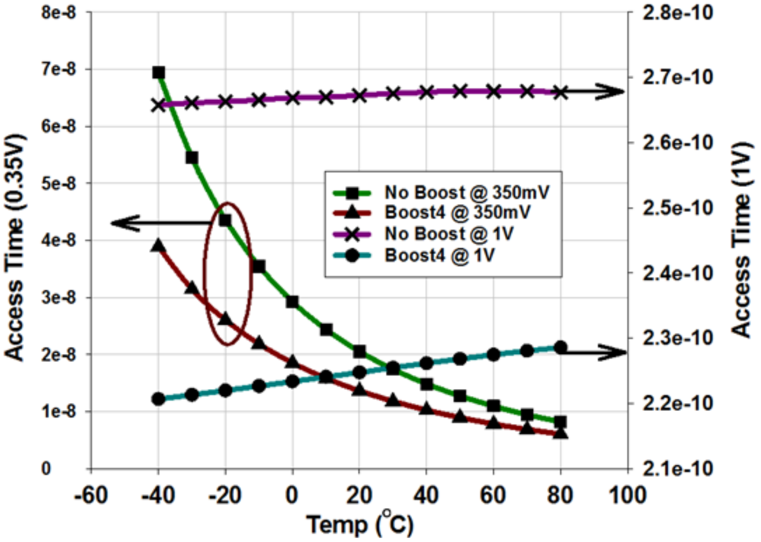


Figure 3.7: Access time versus temperature at 350 mV and 1 V.

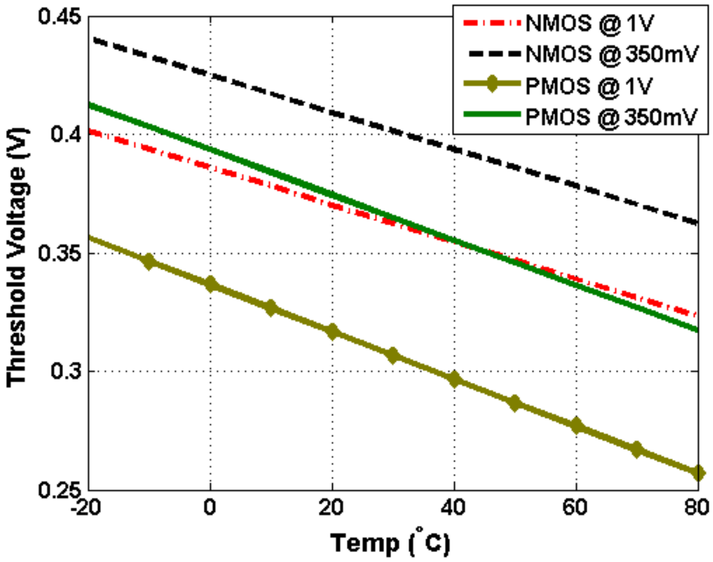


Figure 3.8: Threshold voltage versus temperature for the NMOS and PMOS transistors at 350 mV and 1 V.

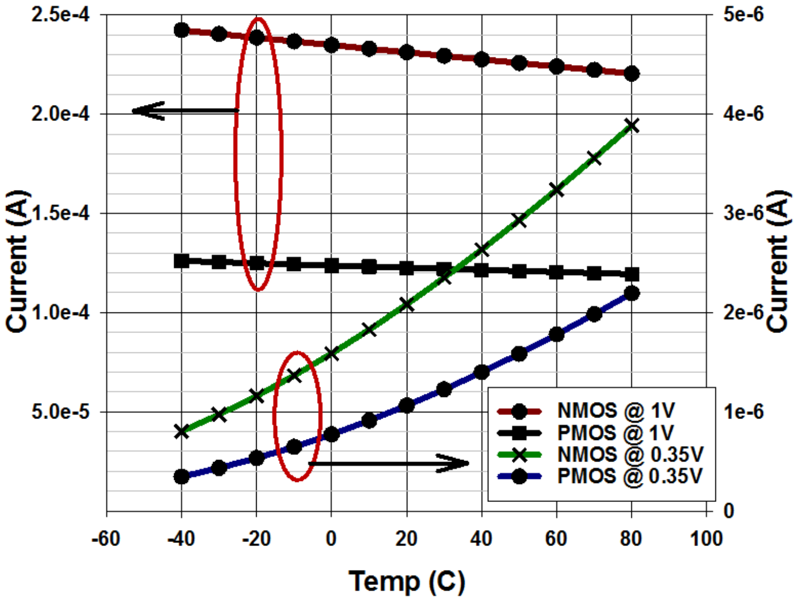
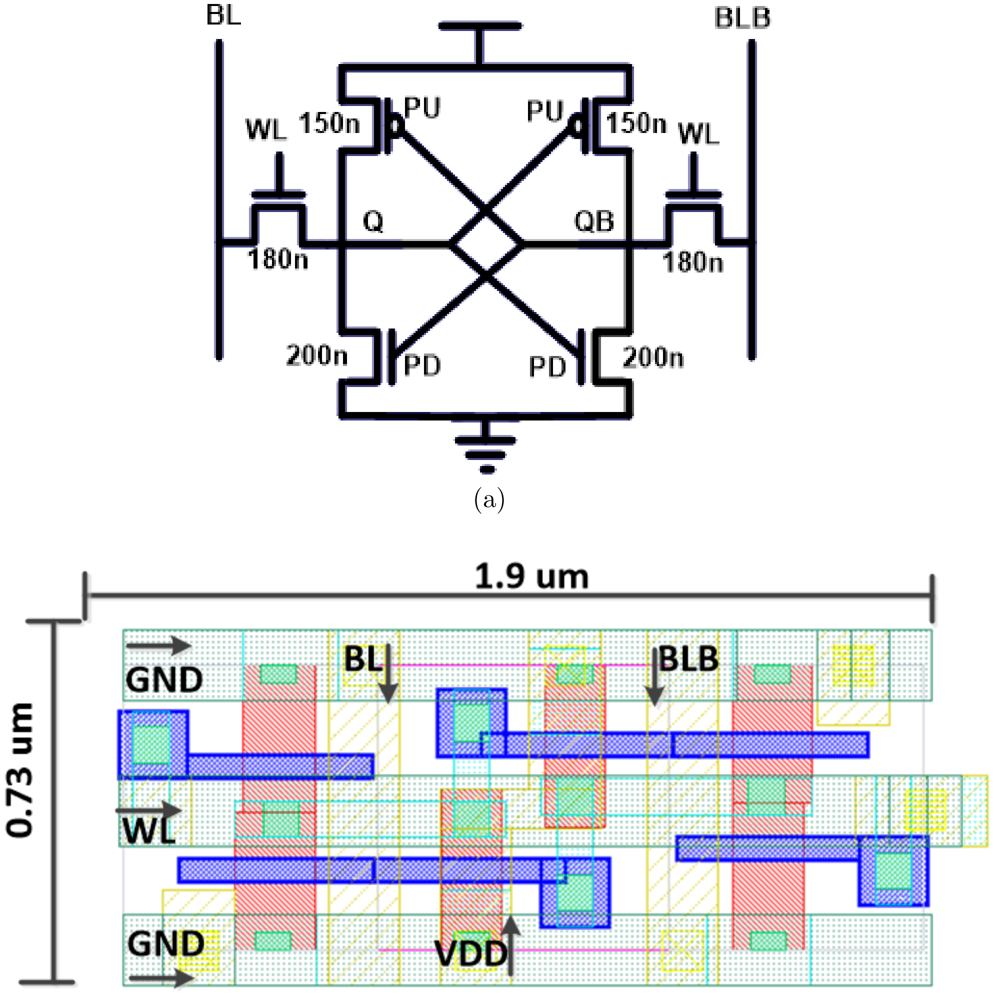


Figure 3.9: NMOS and PMOS current versus the temperature at 1 V and 0.35 V.



(a)

(b)

Figure 3.10: a) Sizing of the 6T bitcell. b) Layout of the bitcell.

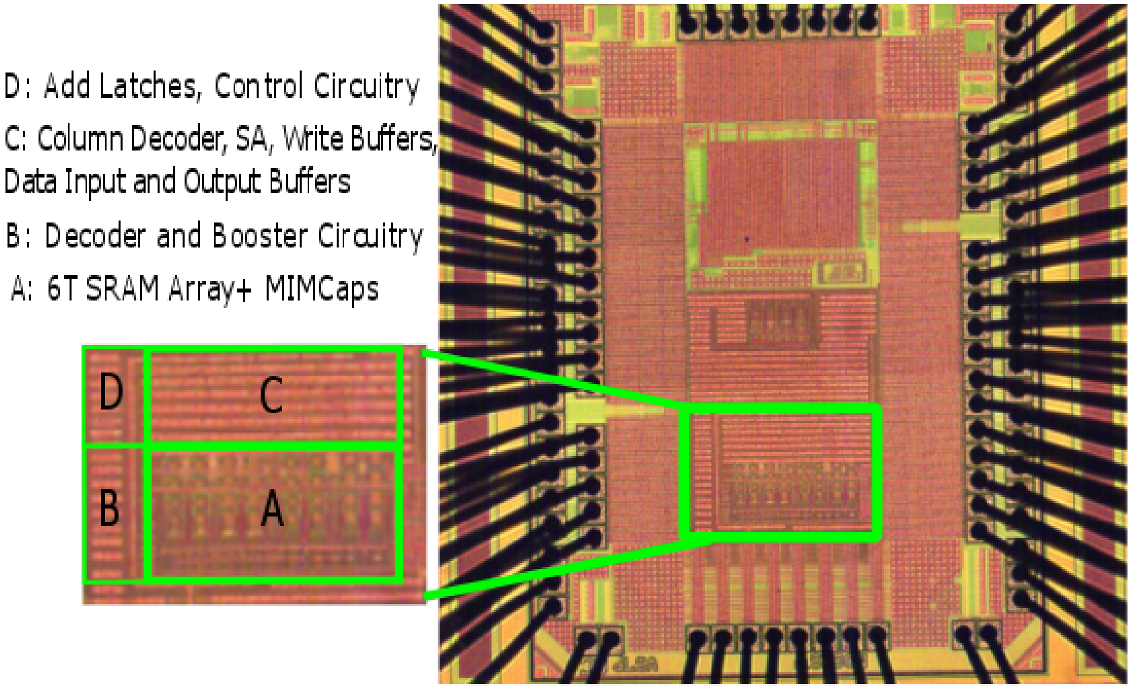
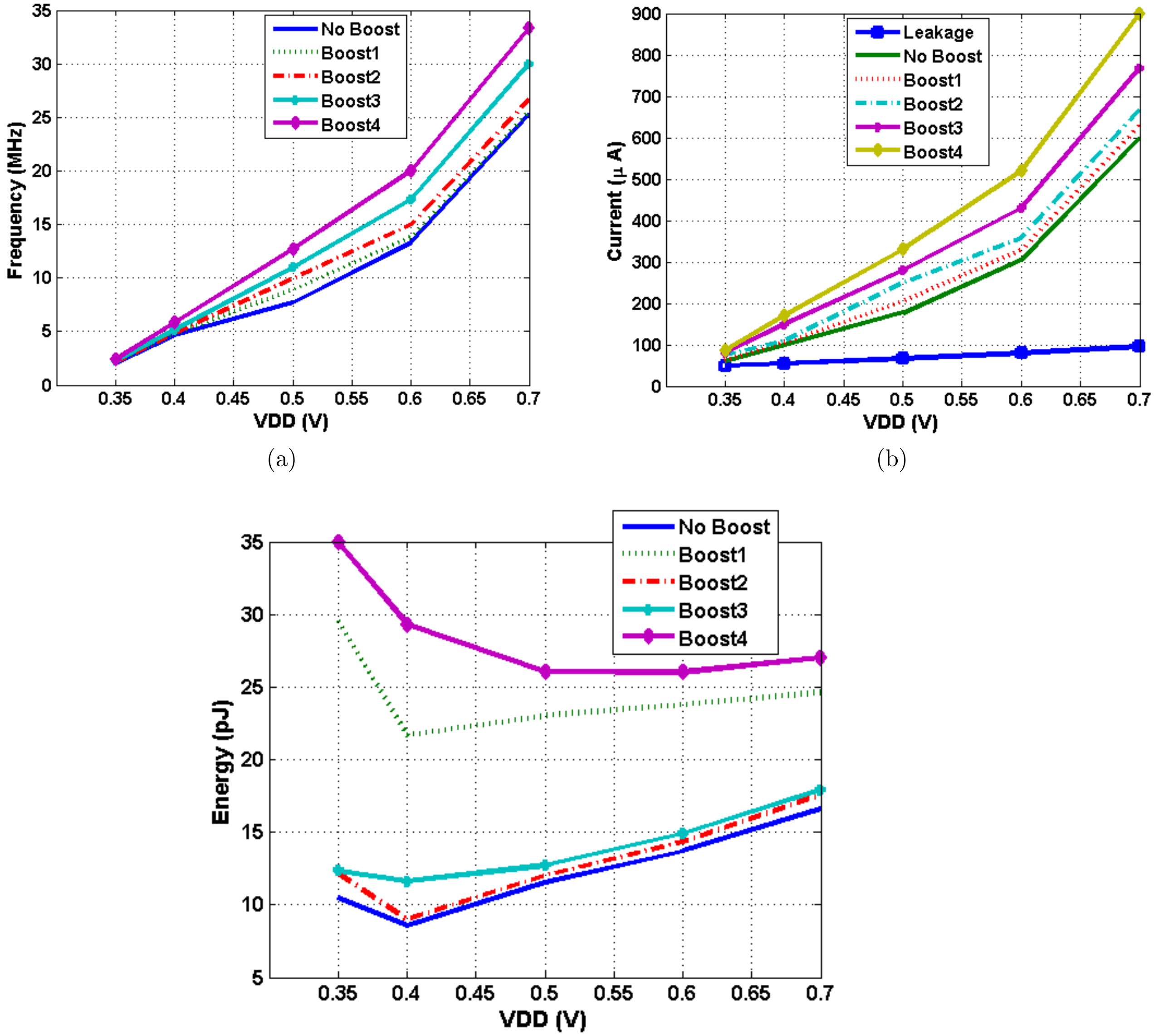

D: Add Latches, Control Circuitry
C: Column Decoder, SA, Write Buffers,
Data Input and Output Buffers
B: Decoder and Booster Circuitry
A: 6T SRAM Array+ MIMCaps
D
C
B
A

Figure 3.11: Micro-graphic image of the fabricated chip



(a)

(b)

(c)

Figure 3.12: a) Measured frequency of operation with respect to the supply voltage; b) Measured total current and leakage current with respect to the supply voltage; c) Total energy and leakage energy with respect to the supply voltage.

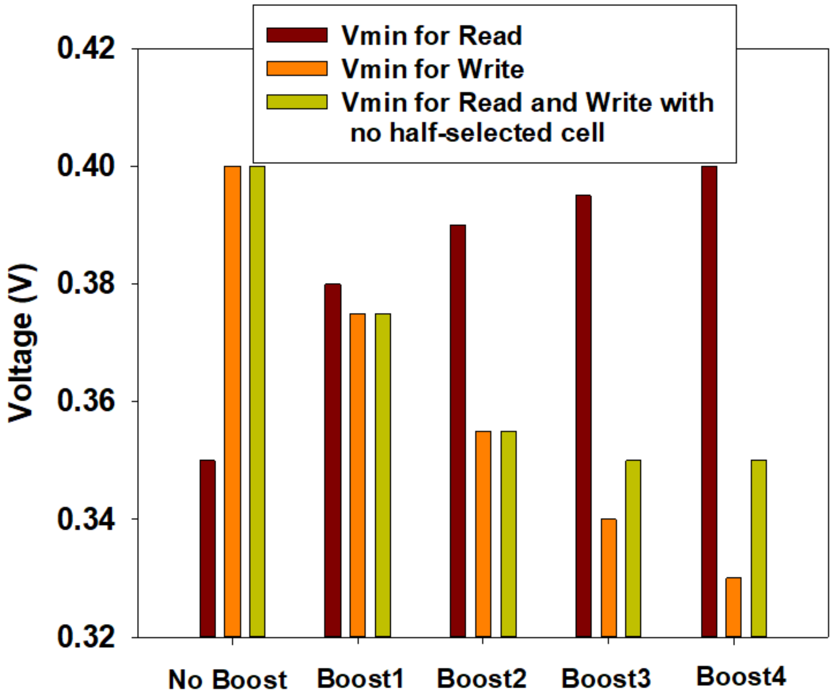


Figure 3.13: Measured minimum read and write voltages versus different levels of boosting.

Table 3.1: Comparison with Chosen Previous Subthreshold SRAMs.

Design	**Technology**	**Transistor Count**	**Size**	**E min (fJ/b)**	**Speed @ 400 mV**	V_{min}	**Bitcell per BL**	**Area** (μm^2)	EDP ($\times 10^{-}21$ fJ/b) @ 400 mV	$I_{Leakage}$ **(nA/b)**
JSSC 08 [32]	130 nm	6T	2-kb	0.55	1 MHz	210mV	16	4	0.625	119
JSSC 06 [39]	90 nm	7T	64-kb	N.A	N.A	440mV	8	13% more than 6T	N.A	N.A
JSSC 08 [34]	65 nm	8T	256-kb	0.5	30KHz	350mV	256	N.A	11.7	6.28
TCAS 14 [40]	40 nm	12T	4-kb	0.47	3MHz	350mV	16	4.42	0.133	14.5
This Work	65 nm	6T	16-kb	0.536	6MHz	330mV (write) 350mV (Read)	128	1.38	0.089	3.43

Figure 3.12(b) illustrates the measured total and leakage current. The total current, was measured while performing successive write and read operations at different addresses. The average of this current is shown in this figure. The leakage current is measured while the macro was inactive. The total current when no boosting is applied, and the leakage current are measured as 100 and 55 μA, respectively, at 400 mV.

The energy consumption can be computed by dividing the power consumption by the maximum frequency. The total energy consumption is shown in Figure 3.12(c) when different levels of WL boosting are applied. The minimum total energy is calculated as 0.536 fJ/bit at 400 mV.

As shown in Fig. 3.12(a), the frequency at which the memory can operate in when there is no WL boosting and the supply voltage is at 500 mV can be achieved when Boost2 is applied and the supply voltage is at 450 mV. Therefore, by reducing the supply voltage while maintaining the frequency of operation, the energy consumption is reduced by 22.2%.

Fig. 3.13 shows the minimum supply voltage that produces 100% yield when different levels of boosting are applied for read and write operations. Increasing the level of the WL boosting increases the read failure. Therefore, the minimum voltage that allows correct read operation with the desired yield increases. This is while, increasing the level of the WL boosting decreases the write failure, and consequently, the minimum supply voltage at which the write operation can be performed, with the desired yield, decreases. The minimum supply voltage to perform a read operation is shown as 350 mV in Fig. 3.13 when no boosting is applied. By utilizing different levels of boosting, the minimum supply voltage for the read operation increases to 380 mV, 390 mV, 395 mV, and 400 mV. For the write operation, the minimum supply voltage, when no boosting is observed is at 400 mV. Utilizing the WL boosting decreases the minimum supply voltage for the write operation. As shown in Fig. 3.13, the minimum supply voltage for the write operation decreases to 375 mV, 355 mV, 340 mV, and 330 mV.

Fig. 3.13 also shows the minimum supply voltage when there are no half-selected cells. The minimum supply voltage is limited by the write operation when no-boosting, Boost1, and Boost2 options are exploited. However, when Boost3 and Boost4 are applied, the minimum supply voltage is limited by the read operation. In this case with no half-selected cells the minimum supply voltage decreases to 350 mV.

Table 3.1 summarizes and compares the key features of our design with previous SRAMs that include the 6T [32], 7T [39], 8T [34], and 12T [40] bitcells. As this table shows, utilizing different levels of WL boosting enables us to reduce the supply voltage to 330 mV for the write operation. The 6T design in [32] reduces the Vmin close to 210 mV at the cost of significant additional bitcell area. This design also utilizes single-ended read sensing which reduces the speed of the read operation. In addition, utilizing wide pass-transistors to access the data in each bitcell creates a significant leakage on the BLs from the unaccessed bitcells in each column. Therefore, the number of the bitcells in each column is limited to 16.

To perform a comparison of the speed of these designs, the speed of all the memory macros are reported at 400 mV. The comparison shows that our design can operate at a relatively higher speed due to the WL boosting, as compared to the designs in [32], [39], [34], and [40].

The over-sized bitcells in [32] and [40] significantly add to the total leakage per bit. Our proposed design has the minimum bitcell area and lowest leakage current per bit among other designs in Table 3.1.

For the sake of reliability at low supply voltages, drivers and peripheral circuits are over-designed. As a consequence, a slight increase is observed in the leakage current and minimum energy consumption of our design.

To provide a fair figure of merit that compares both the delay and the energy consumption, the energy-delay-product (EDP) per bit of all designs are evaluated. The comparison shows that our design has the lowest EDP per bit amongst all.

3.5 Conclusion

SRAM circuits significantly affect the SOC's performance, energy consumption, reliability, and yield. There is critical need to reduce the energy consumption of the SRAM circuits for portable devices and billions of connected sensor networks that require long battery life. In this chapter, we have presented a 4-level programmable WL boosting technique in order to reduce the supply voltage, and provide a process-tolerant design. A 16-kb SRAM memory is fabricated in the 65 nm TSMC GP CMOS technology. Measurement results show that the operational frequency improves up to 33.3% when the WL boosting is applied. By utilizing the WL boosting, the supply voltage can be decreased by 50 mV while maintaining the same operational frequency. This, in turn, allows a reduction in the energy consumption by 22.2%.

Chapter 4

A 290-mV, 3.34-MHz, 6T SRAM with PMOS Access Transistors and Boosted Word Line in 65-nm CMOS Technology

4.1 Introduction

Ultra-low[1] power applications such as sensor networks, pacemakers, and many portable devices require extreme energy constraints for a longer battery life. It is shown that very low energy operation is achieved when the supply voltage is in the near, or subthreshold region [58]. By reducing the supply voltage of a SOC, the dynamic energy is decreased quadratically at the expense of increased delay. As the clock cycle period is reduced to accommodate the increased delay, leakage power and energy contributions become significant [48]. One of the approaches to reduce this component is to shut down the macro after completing the task [48]. Unfortunately, SRAM power cannot be switched off without losing its data. Even reducing its power supply voltage requires careful consideration, owing to its data stability, SNM, and WM. Therefore, SRAM blocks are the main bottleneck to reduce the operating supply voltage of the SOCs [67]. Another challenge of SRAM blocks is their low speed in the subthreshold region, due to the reduced supply voltage and stability issue. In addition to the stability challenge of SRAMs, the low speed of subthreshold

[1] Note that most of this chapter has been published in [66]

circuits and specifically SRAM arrays, limits the complexity of the tasks that these circuits can perform. It is also required to develop subthreshold circuits operating at higher speeds that can perform more complex tasks [59].

The conventional 6T SRAM bitcell has contradictory requirements for read stability and writability. For example, decreasing the access transistor width improves the read stability, while it decreases the WM. This conflict becomes even more emphasized in the subthreshold region. The design in [23] utilizes a two-step WL boosting to overcome this conflict. The designs proposed in [32] and [34] have improved both SNM and WM at the expense of increased bitcell area, reduced speed, removing half-selected cells and not being able to utilize differential sensing. The main drawback of single-ended sensing versus differential sensing is its slow sensing speed and not being immune to common-mode noise. In addition, not incorporating half-selected cells requires higher area and more complexity for the extra needed sense amplifiers and peripheral circuitry [68]. Moreover, since they do not have bit-interleaving, Single-Error Correction and Double-Error Detection (SEC-DED) schemes may not be adequate in mitigating soft errors [69].

A 6T bitcell operating in the subthreshold region is reported in [32]. This asymmetrical and single-ended 6T bitcell uses one pass-gate instead of two NMOS access transistors; and in order to overcome the small sensing window and vulnerability to process variation, they significantly increase the sizes of each transistor in each bitcell. One main weakness of this design is its relatively low-speed operation. Several 65 nm designs have proposed bitcells with an extra number of transistors. For example, in [34], a single-ended 8T bitcell is fabricated that is capable of operating as low as 350 mV. This design suffers from low-speed single-ended sensing and is not able to tolerate half-selected cells. The proposed bitcell in [35] utilizes nine transistors to enable differential sensing. They also show that utilizing PMOS access transistors makes their bitcell less susceptible to the process variation effect. This design operates at a speed of 200 KHz at 350 mV. The authors in [33] utilize a system level approach to reduce the SRAM supply voltage for image and video specific applications. In order to avoid the worst-case read scenario, the stored data in columns are randomized to make the distribution of the 0s and 1s close to 50%. They show that the 8T bitcell in [34] can operate at 200 mV when utilizing data randomization. Researchers also have designed the SRAM cell with PMOS access transistors in an ECL-CMOS process [70]. With the PMOS access transistor, the authors claim that they can reduce the power supply voltage by an additional 0.5 V, as compared to the NMOS access transistor.

In this Chapter [66], a 6T bitcell optimized for low voltage applications is proposed. In order to improve the read stability of the bitcell during the read operation, the PMOS access transistors are utilized as they can provide a better read stability compared to the NMOS transistors. In addition, the access transistor connected to the node that holds

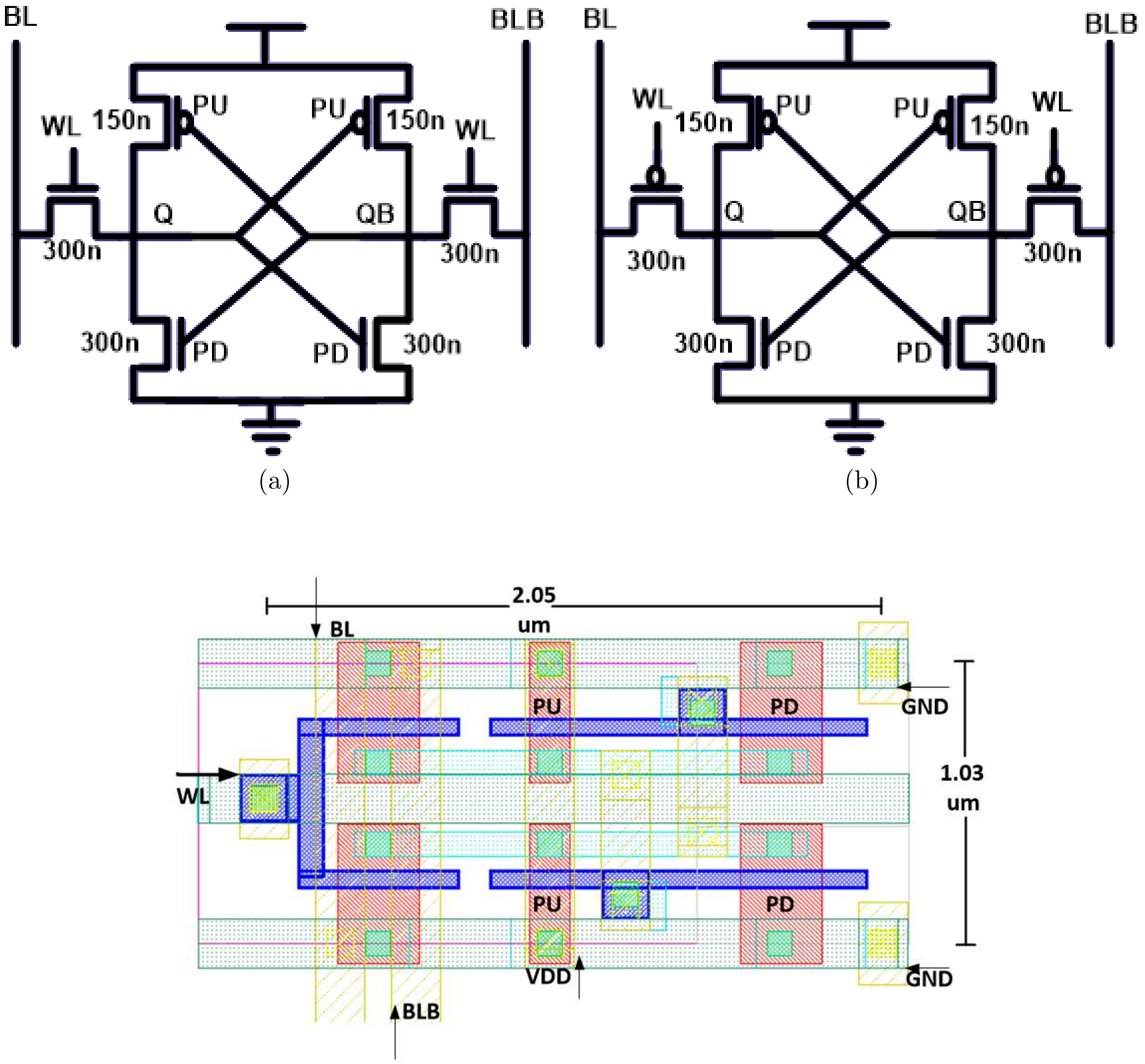
$\mathbf{V_{DD}}$ in the proposed bitcell, unlike the conventional 6T bitcell, is fully on and mitigates the ZLD. Moreover, to overcome the weak writability of the new bitcell, the WL boosting is exploited. Even though the WL boosting emphasizes the ZLD, unlike the conventional 6T bitcell, the access transistor connected to the internal node with high voltage also increases its robustness against the ZLD. Moreover, the WL boosting also shows more than a 3× speed improvement in the subthreshold region. In addition, differential sensing is exploited in our design.

The rest of the chapter is organized as follows. In Section 4.2, the read stability of the 6T bitcell with the PMOS access transistor is investigated through simulations and analytical analysis. In Section 4.3, the improvement of the writability utilizing WL boosting is described. The boosted circuit implementation is discussed in Section 4.4. In Section 4.5, the read and leakage current of the new bitcell are compared with that of the conventional 6T bitcell. Measurement results and comparison with previously published results are provided in Section 4.6. Finally, in Section 4.7 conclusions are drawn.

4.2 Read Stability of the 6T SRAM Bitcell with PMOS Access Transistors

The 6T bitcells with the NMOS access transistor (6T-NA) and the PMOS access transistor (6T-PA) are shown in Figure 4.1(a-b). The layout of the 6T-PA is also shown in Figure 4.1(c). The read butterfly curves of the 6T-NA and 6T-PA are shown in Figure 4.2. This figure shows that the 6T-PA has a higher SNM compared to the 6T-NA at 1 V and the SNM is almost the same at 500 mV and 300 mV. To compare the read stability of both bitcells, a 1k Monte Carlo simulation of both bitcells at the same condition is performed. Figure 4.3 shows the behavior of node QB of both bitcells. As shown in this figure, for the 6T-NA, data-flip occurs 105 times (i.e., yield = 89.5%), while only 1 data-flip occurs for the 6T-PA (i.e., yield = 99.9%). Assuming, without loss of generality, that the node QB in both 6T bitcells is high; the BLB remains high while BL starts discharging. In this process, node Q acquires a non-zero potential known as ZLD. A larger ZLD can adversely affect the read stability of a SRAM bitcell. Since a PMOS transistor has lower mobility, for the iso-area the $\mathbf{C_R}$ in the superthreshold region is increased by a factor of $\frac{\mu_n}{\mu_p}(= 2.5)$ as follows [56]:

$$C_R = \frac{\mu_n W_n / L_n}{\mu_p W_a / L_a} \tag{4.1}$$

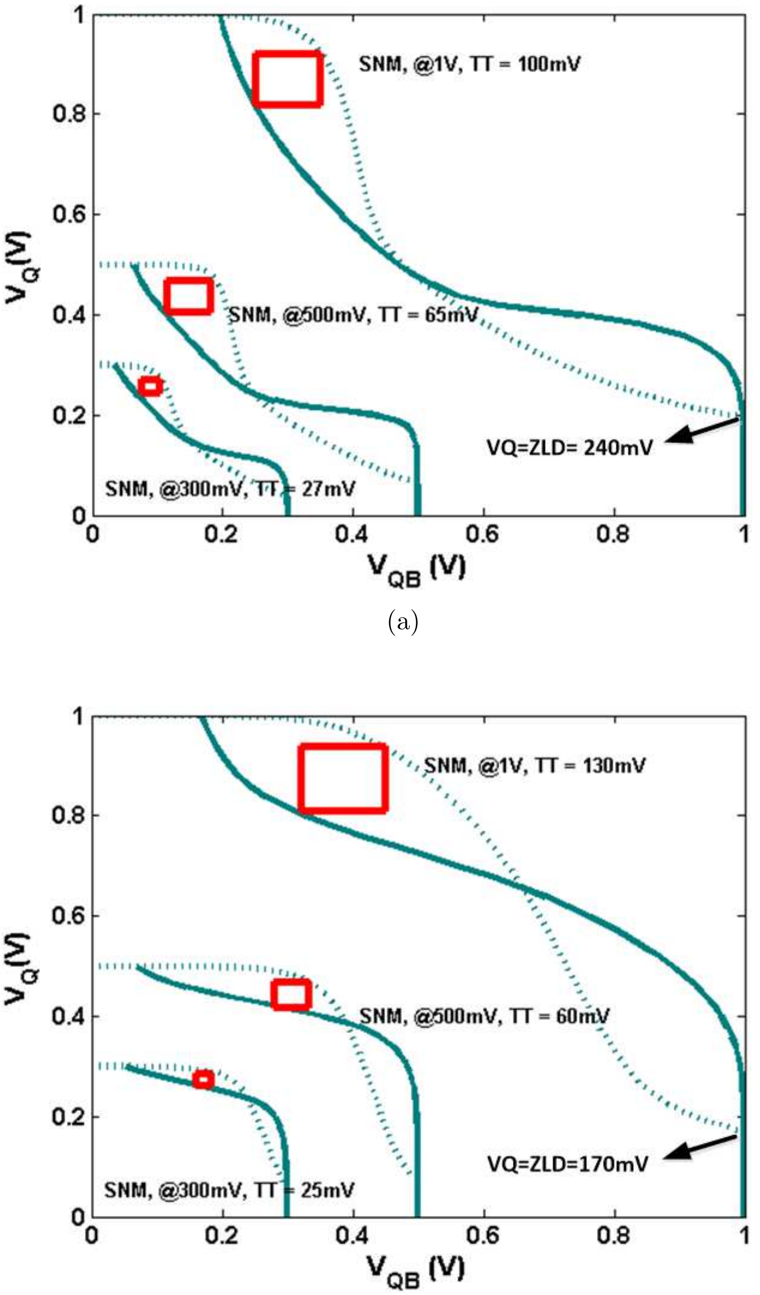


(a)

(b)

(c)

Figure 4.1: (a) 6T-NA bitcell b) 6T-PA bitcell c) Layout of 6T-PA.



(a)

(b)

Figure 4.2: Read butterfly curves at the TT corner for a) 6T-NA, b) 6T-PA, (T= 25°C).

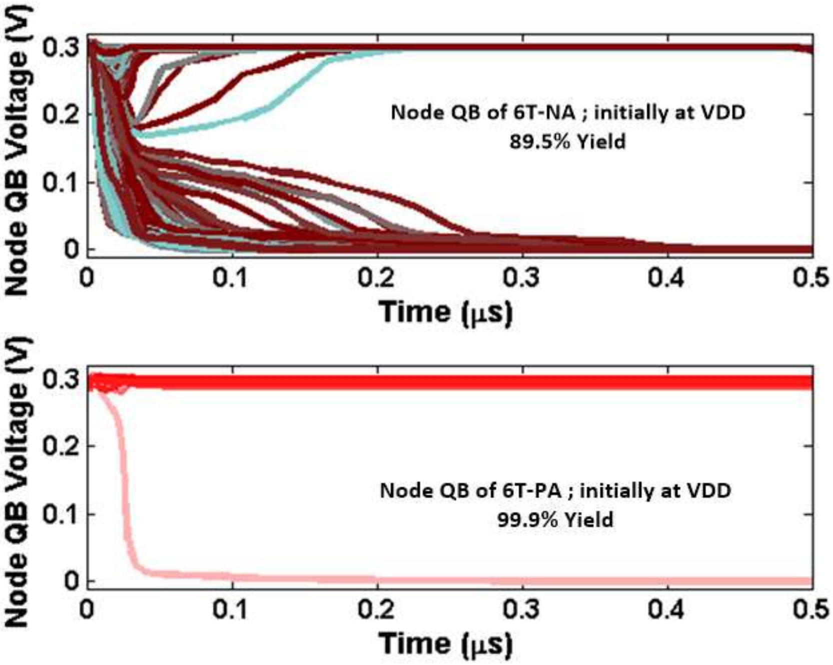


Figure 4.3: 1k Monte Carlo read simulation for the 6TPA and 6TNA bitcells at 300 mV. A data flip occurs when node QB makes a transition from $\mathbf{V_{DD}}$ to 0.

Table 4.1: NMOS/PMOS Transistor Parameters in the 65 nm CMOS Technology.

Transistor Type	V_{t0}	λ	η
NMOS	400 mV	99 m	90 m
PMOS	370 mV	110 m	133.2 m

where, $\mathbf{W_n}(\mathbf{L_n})$ and $\mathbf{W_a}(\mathbf{L_a})$ are the width (length) of the Pull Down (PD) and access transistors, and $\mu_\mathbf{n}$ and $\mu_\mathbf{p}$ are the mobility of the NMOS and PMOS transistors, respectively.

The subthreshold current can be expressed by [71] [72]:

$$I_{sub} = \mu C_{ox}\frac{W}{L}(n-1)\nu_T^2 e^{\frac{(V_{GS}-V_{th})}{n\nu_T}}(1-e^{\frac{-V_{DS}}{\nu_T}}) \tag{4.2}$$

$$V_{th} = V_{t0} - \lambda V_{BS} - \eta V_{DS} \tag{4.3}$$

where μ is the charge carrier mobility, $\mathbf{C_{ox}}$ is the gate-oxide capacitance, $\nu_\mathbf{T}$ is the thermal voltage, $\mathbf{V_{GS}}$ is the MOSFET's gate-source voltage, and n is the subthreshold slope factor. $\mathbf{V_{t0}}$ represents the zero-biased threshold voltage of a MOSFET. Parameters λ and η represent the body effect coefficient and DIBL coefficient of a MOSFET, respectively. The parameters $\mathbf{V_{t0}}$, λ, and η for the NMOS and PMOS transistors in the 65 nm CMOS technology are presented in Table 4.1. The body effect and DIBL coefficient multiplied by the $\mathbf{V_{DS}}$ and $\mathbf{V_{BS}}$, respectively, can be assumed to be negligible compared to the zero-biased threshold voltage. Although n varies between 1.3 to 1.5, for convenience, it can be assumed to be equal for the NMOS and PMOS transistors in the subthreshold region [64].

$\mathbf{D_R}$ in the subthreshold region is defined as the driving strength ratio of the PD transistor to the access transistor. Considering $\mathbf{V_{tp}}$ and $\mathbf{V_{tn}}$ as the zero-biased threshold voltages of PMOS and NMOS transistors, respectively, and assuming

$$\alpha_n = e^{\frac{-V_{tn}}{n\nu_T}}, \alpha_p = e^{\frac{-|V_{tp}|}{n\nu_T}} \tag{4.4}$$

$$\Gamma = \frac{\alpha_n}{\alpha_p}, \Lambda = \frac{\mu_n}{\mu_p}, \beta = \frac{W_n/L_n}{W_p/L_p} \tag{4.5}$$

and the $\mathbf{D_R}$ in the subthreshold region can be expressed as

$$D_R = \Gamma.\Lambda.\beta \tag{4.6}$$

The difference of the $\mathbf{D_R}$ ratio with the $\mathbf{C_R}$ is the $\mathbf{\Gamma}$ factor which is called the subthreshold $\mathbf{C_R}$ modification factor. This parameter is exponentially dependent upon the difference of the zero-biased threshold voltages of PMOS and NMOS transistors ($V_{tp} - V_{tn}$). Figure 4.4 exhibits that the variation of this factor in 1k Monte Carlo samples at the supply voltage of 0.3 V is between 0.66 to 0.44. Since, $\Lambda = 2.48$, for β equal to 1, the $\mathbf{D_R}$ value varies from 1.1 to 1.5, and still provides a higher driving strength of the PD transistor compared to the access transistor for lower ZLD. For the 6T-NA, λ is equal to 1, and the threshold voltage mismatch between access and the PD transistor causes variation in $\mathbf{\Gamma}$. The variation in $\mathbf{\Gamma}$ due to threshold voltage mismatch is between 0.84 to 1.34. Based on the results, the following comments can be made. For the iso-area (i.e., for the same area and channel lengths), the $\mathbf{D_R}$ value of the 6T-PA is greater than that of 6T-NA in the subthreshold region. To make the $\mathbf{D_R}$ of the 6T-NA greater than 1.1 (minimum $\mathbf{D_R}$ of 6T-PA), the width of the PD transistor has to be 30% larger than the access transistor to alleviate the variation of $\mathbf{\Gamma}$. The most suitable technologies for providing stable 6T-PA bitcells in the subthreshold operation are those with $|V_{tp}| > V_{tn}$ (i.e., $\Gamma > 1$). The optimum 6T-PA bitcells implemented in these technologies are smaller and, hence, consume lower amounts of energy.

In the following, the ZLD of both bitcells are calculated analytically. The subthreshold current of the access transistor of the 6T-NA is given in Equation 4.7.

$$I_A = \mu_n C_{ox} \frac{W_A}{L_A}(n-1)\nu_T^2 e^{(\frac{V_{DD}-V_Q-V_{tn}+\lambda V_{BS}+\eta V_{DS}}{n\nu_T})}(1 - e^{-\frac{V_{DD}-V_Q}{\nu_T}}) \tag{4.7}$$

Subtituting V_{BS} by V_Q and V_{DS} by $V_{DD} - V_Q$, Equation 4.7 becomes

$$I_A = \mu_n C_{ox} \frac{W_A}{L_A}(n-1)\nu_T^2 e^{(\frac{V_{DD}-V_Q-V_{tn}-\lambda V_Q+\eta(V_{DD}-V_Q)}{n\nu_T})}(1 - e^{-\frac{V_{DD}-V_Q}{\nu_T}}) \tag{4.8}$$

Similarly, the subthreshold current of the PD transistor of the 6T-NA is given in Equation 4.9.

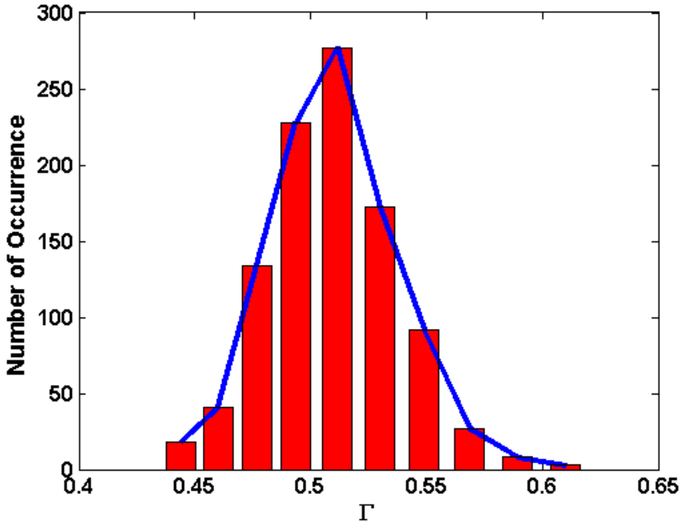


Figure 4.4: Γ variation at 1k Monte Carlo simulations at 0.3 V.

$$I_D = \mu_n C_{ox} \frac{W_D}{L_D}(n-1)\nu_T^2 e^{(\frac{V_{DD}-V_{tn}+\eta V_Q}{n\nu_T})}(1-e^{\frac{-V_Q}{\nu_T}}) \tag{4.9}$$

$$I_D = \mu_n C_{ox} \frac{W_D}{L_D}(n-1)\nu_T^2 e^{(\frac{V_{DD}-V_{tn}}{n\nu_T})} e^{(\frac{\eta V_Q}{n\nu_T})}(1-e^{\frac{-V_Q}{\nu_T}}) \tag{4.10}$$

Assuming that the current through the pull up transistor is negligible, the current flowing through the access transistor is equal to that of the PD transistor ($i.e., I_A = I_D$). Therefore,

$$\frac{W_A}{L_A} e^{\frac{V_{DD}-V_{tn}}{n\nu_T}} e^{\frac{\eta V_{DD}}{n\nu_T}} e^{\frac{(-1-\eta-\lambda)V_Q}{n\nu_T}}(1-e^{\frac{V_Q-V_{DD}}{\nu_T}}) = \frac{W_D}{L_D} e^{\frac{V_{DD}-V_{tn}}{n\nu_T}} e^{\frac{\eta V_Q}{n\nu_T}}(1-e^{\frac{-V_Q}{\nu_T}}) \tag{4.11}$$

$$\frac{W_A}{L_A} e^{\frac{\eta V_{DD}}{n\nu_T}} e^{\frac{(-1-\eta-\lambda)V_Q}{n\nu_T}}(1-e^{\frac{V_Q-V_{DD}}{\nu_T}}) = \frac{W_D}{L_D} e^{\frac{\eta V_Q}{n\nu_T}}(1-e^{\frac{-V_Q}{\nu_T}}) \tag{4.12}$$

Considering

$$\eta = 0.091, \lambda = 0.099, n = 1.5, V_{DD} = 0.3, e^{\frac{\eta V_{DD}}{n\nu_T}} = 2 \tag{4.13}$$

and assuming

$$X = e^{(\frac{-V_Q}{n\nu_T})}, \beta = \frac{W_D/L_D}{W_A/L_A} \tag{4.14}$$

Equation 4.12 can be simplified to

$$2\frac{W_A}{L_A}X^{1.2}(1 - e^{\frac{-V_{DD}}{\nu_T}}X^{-n}) = \frac{W_D}{L_D}X^{-\eta}(1 - X^n) \tag{4.15}$$

$$2\frac{W_A}{L_A}X^{1.2}(X^n - e^{\frac{-V_{DD}}{\nu_T}}) = \frac{W_D}{L_D}X^{n-\eta}(1 - X^n) \tag{4.16}$$

$$2\frac{W_A}{L_A}(X^{1.2+n} - e^{\frac{-V_{DD}}{\nu_T}}X^{1.2}) = \frac{W_D}{L_D}(X^{n-\eta} - X^{2n-\eta}) \tag{4.17}$$

$$X^{2.7} = \frac{\beta}{2}(X^{1.4} - X^{2.9}) \tag{4.18}$$

By calculating X from Equation 4.18, the V_Q can be calculated by

$$V_Q = -n\nu_T ln(X) \tag{4.19}$$

Similar to the 6T-NA, the subthreshold current of the access transistor and the PD transistor of the 6T-PA are presented in Equation 4.20 and 4.21, respectively.

$$I_A = \mu_p C_{ox}\frac{W_A}{L_A}(n-1)\nu_T^2 e^{(\frac{V_{DD}-V_{tp}}{n\nu_T})}(1 - e^{-\frac{V_{DD}-V_Q}{\nu_T}}) \tag{4.20}$$

$$I_D = \mu_n C_{ox}\frac{W_D}{L_D}(n-1)\nu_T^2 e^{(\frac{V_{DD}-V_{tn}}{n\nu_T})}(1 - e^{\frac{-V_Q}{\nu_T}}) \tag{4.21}$$

Equalizing the access transistor and the PD transistor currents ($I_A = I_D$) results in

$$\mu_p\frac{W_A}{L_A}e^{(\frac{-V_{tp}}{n\nu_T})}(1 - e^{-\frac{V_{DD}-V_Q}{\nu_T}}) = \mu_n\frac{W_D}{L_D}e^{(\frac{-V_{tn}}{n\nu_T})}(1 - e^{\frac{-V_Q}{\nu_T}}) \tag{4.22}$$

$$\frac{\mu_p}{\mu_n}e^{(\frac{V_{tn}-V_{tp}}{n\nu_T})}\frac{W_A/L_A}{W_D/L_D}(1 - e^{-\frac{V_{DD}}{\nu_T}}e^{\frac{V_Q}{\nu_T}}) = (1 - e^{\frac{-V_Q}{\nu_T}}) \tag{4.23}$$

By assuming

$$X = e^{\frac{V_Q}{\nu_T}}, \beta = \frac{W_D/L_D}{W_A/L_A} \tag{4.24}$$

Equation 4.23 can be simplified to

$$\big(-e^{-\frac{V_{DD}}{\nu_T}}\big)\big(\frac{\frac{\mu_p}{\mu_n}e^{(\frac{V_{tn}-V_{tp}}{n\nu_T})}}{\beta}\big)X^2 + \big(\frac{\big(\frac{\mu_p}{\mu_n}e^{(\frac{V_{tn}-V_{tp}}{n\nu_T})}\big)}{\beta} - 1\big)X + 1 = 0 \tag{4.25}$$

where X and V_Q can be obtained as

$$X \approx \frac{-1}{\frac{\frac{\mu_p}{\mu_n}e^{(\frac{V_{tn}-V_{tp}}{n\nu_T})}}{\beta} - 1} \tag{4.26}$$

$$V_Q = \nu_T \times ln\big(\frac{-1}{\frac{\frac{\mu_p}{\mu_n}e^{(\frac{V_{tn}-V_{tp}}{n\nu_T})}}{\beta} - 1}\big) \tag{4.27}$$

Figure 4.5 illustrates the V_Q obtained analytically from Equation 4.19 and 4.27 and from the simulation for both 6T-NA and 6T-PA at 300 mV. This figure shows that the 6T-PA suffers less from the ZLD.

As mentioned before, unlike the 6T-NA, the 6T-PA provides better read stability partly owing to the access transistor connected to the internal node with high voltage $\mathbf{V_{DD}}$. To further investigate this behavior, a single-ended positive noise source (Figure 4.6(a)) to both cells at node retaining logic 0 is applied. Single-ended noise mimics the read disturb behavior of the cell and can be correlated to cell stability during the read operation. As shown in Figure 4.6(b), when a pulse of 150 mV is applied to the node Q of the 6T-NA bitcell, the node QB decreases down to 146 mV. However, the node QB in the 6T-PA discharges down to 246 mV. Figure 4.6(c) illustrates the simulation results of a single-ended voltage noise source applied on the bitcells in worst-case corners as a function of supply voltage. As shown in this figure, the 6T-PA can tolerate much higher single-ended noise compared to the 6T-NA. For example, at 0.3 V, the 6T-PA can tolerate 215 mV of single-ended noise whereas the 6T-NA tolerates 135 mV. By applying the WL boosting the $\mathbf{V_{GS}}$ of the right access transistor increases and this causes the right access transistor to become more resistive in holding the node QB at $\mathbf{V_{DD}}$. In other words, the right access

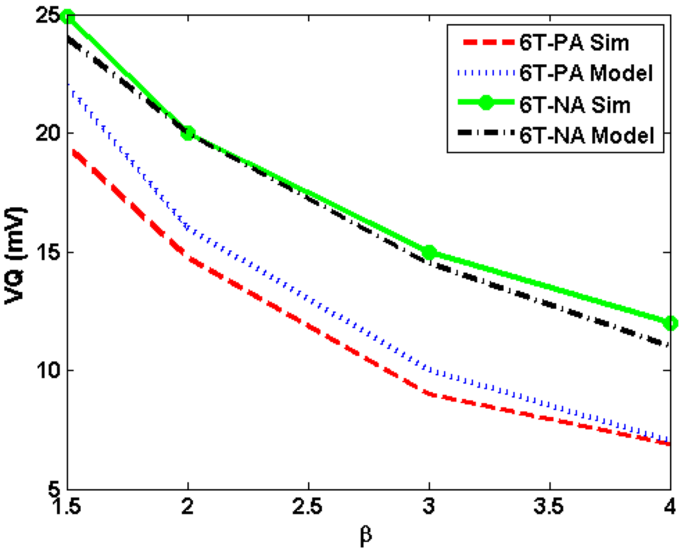
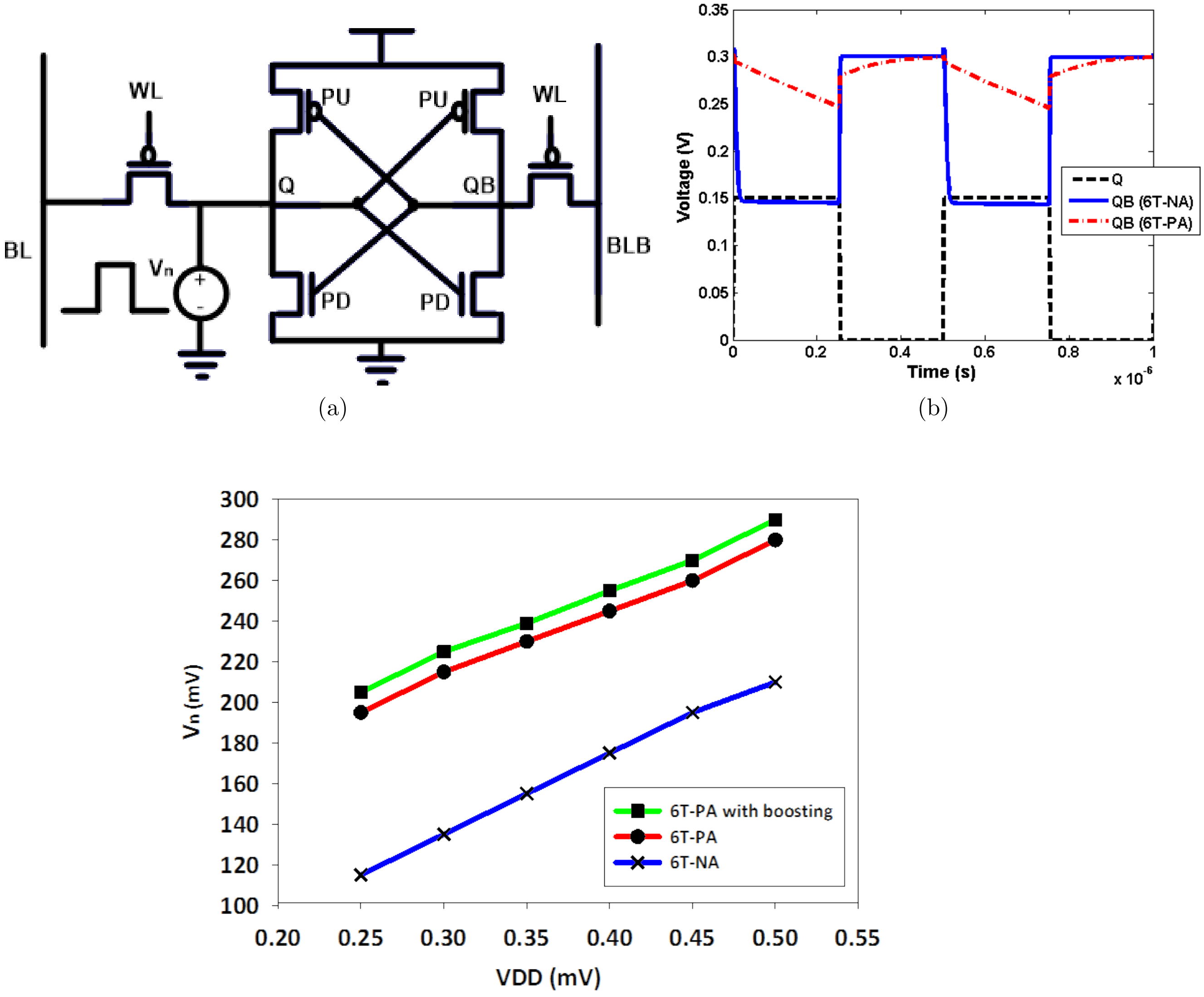


Figure 4.5: Analytical and simulated ZLD versus β for both 6T-NA and 6T-PA at 290 mV, TT corner, 25°C.

transistor partially offsets the effect of the ZLD. Therefore, the 6T-PA can tolerate up to 225 mV of single-ended noise when -65 mV of WL boosting is applied at 0.3 V (shown in Figure 4.6(c)).

The stability of the 6T-PA is also compared with the 6T-NA when two differential noise sources are incorporated in the bitcells as shown in Figure 4.7(a-b) [53]. Figure 4.8(a-c) shows the transient behavior of node Q and QB during a read operation when a differential noise of 25 mV is applied on the 6T-NA, 6T-PA, and 6T-PA with the WL boosting. As shown in this figure, a data loss occurs for the 6T-NA and data remains stable for both cases of 6T-PA. Moreover, when WL boosting is applied on the 6T-PA, the node QB remains close to $\mathbf{V_{DD}}$, and the node Q of the 6T-PA shows a higher ZLD. In total, the 6T-PA with boosting shows less stability compared to when the WL boosting is not available. Figure 4.9 shows the maximum differential noise tolerated by the 6T-NA, 6T-PA with and without boosting as a function of $\mathbf{V_{DD}}$.

The proposed sizing of the 6T-PA shown in Figure 4.1(a) achieves a read yield of 99.99%. The yield is obtained by counting the number of correct read operations in 10k Monte Carlo simulations. Monte Carlo simulation results show that to achieve the same read stability of the 6T-PA bitcell, the PD transistors of the 6T-NA bitcell have to be sized 60% larger, which results in a 20% larger bitcell area.



(a)

(b)

(c)

Figure 4.6: a) Schematic for simulating read stability of the 6T-PA cell with single-ended noise. b) Transient simulation of node QB for 6T-NA at FS corner and 6T-PA at SF corner when a single-ended noise of 150 mV is applied on node Q at $V_{DD} = 300$ mV. c) Maximum tolerable single-ended noise during read operation at FS corner for 6T-NA and SF corner for 6T-PA with and without boosting, T= 25°C.

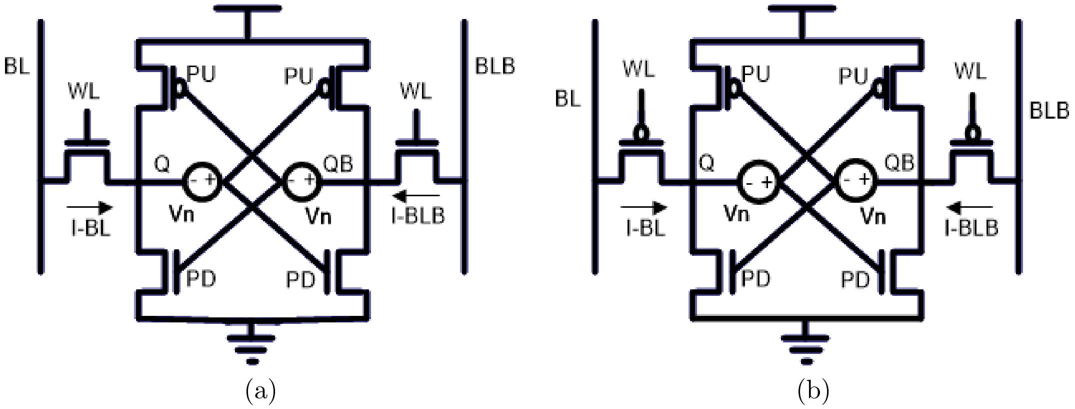


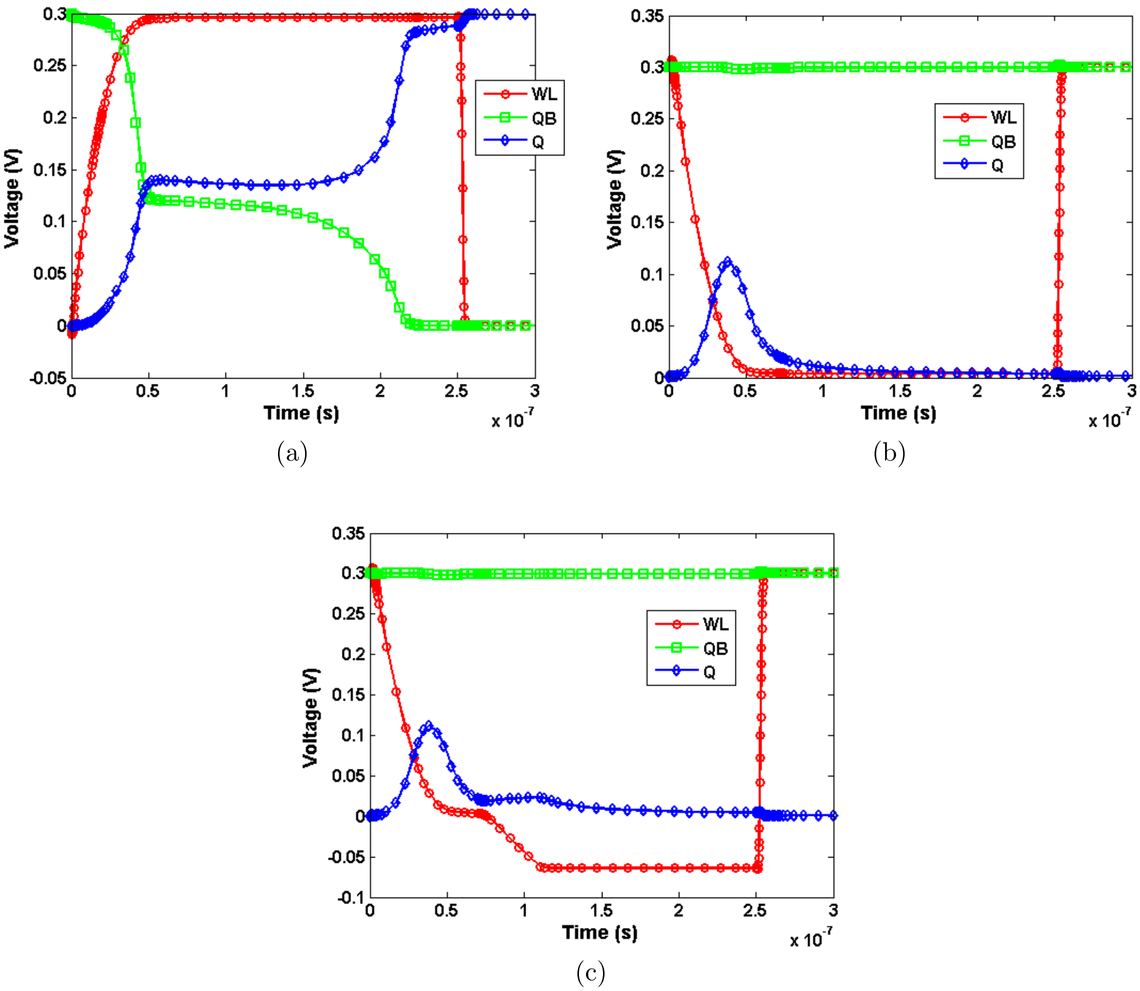
Figure 4.7: Test set up with two differential noise sources for a) 6T-NA and b) 6T-PA.

4.3 Writability Analysis

As described in Section 4.2, the 6T-PA has an improved SNM compared to the 6T-NA; consequently, the 6T-PA has a lower WM compared to the 6T-NA.

Figure 4.10 shows the butterfly curves of a write operation for 6T-NA and 6T-PA for their worst corners. The worst corner for writing into the 6T-NA is the SF corner (NMOS slow, PMOS fast) and the worst corner for writing into the 6T-PA is the FS corner. For example, the WM of the 6T-PA and the 6T-NA is equal to 12 mV and 27 mV, respectively, at 300 mV. Figure 4.11(a) shows the WM of both bitcells at worst corners versus supply voltage ($\mathbf{V_{DD}}$). As shown in these figures, the 6T-PA has a lower WM compared to the 6T-NA. Assuming both bitcells have logic zero initially in Figure 4.1, the right access transistor of the 6T-NA is fully on ($V_{GS} = V_{DS} = V_{DD}$) and starts to discharge the QB node. At the same time, the left-access transistor is also fully on ($V_{GS} = V_{DS} = V_{DD}$) and helps in writing by raising the voltage of node Q to ZLD level. For the 6T-PA, the left access transistor is fully on similar to that of the 6T-NA. However, as opposed to the 6T-NA, since the BLB and the WL are both at 0, the $\mathbf{V_{GS}}$ is constructed between the WL and node QB. During the write process, where the node QB starts discharging, the right-access transistor starts getting weaker as the $\mathbf{V_{GS}}$ decreases, and it turns OFF when the node QB goes below the threshold voltage $\mathbf{V_{tp}}$. Therefore, the 6T-PA bitcell has reduced writability compared to the 6T-NA.

Figure 4.11(b) depicts the write-yield percentage of the write operation of both 6T-NA and 6T-PA bitcells at 250 mV at worst corners. The write yield is achieved by counting the successful write operations in 10k Monte Carlo simulations at the worst corner. As shown



(a)

(b)

(c)

Figure 4.8: Transient behaviour of internal nodes at 300 mV when a differential noise of +/- 25 mV is applied on a) 6T-NA at FS corner, b) 6T-PA at SF corner, and c) 6T-PA with -65 mV of WL boosting at SF corner at T= 25°C. Data flips in 6T-NA while 6T-PA and 6T-PA with WL boost remain stable.

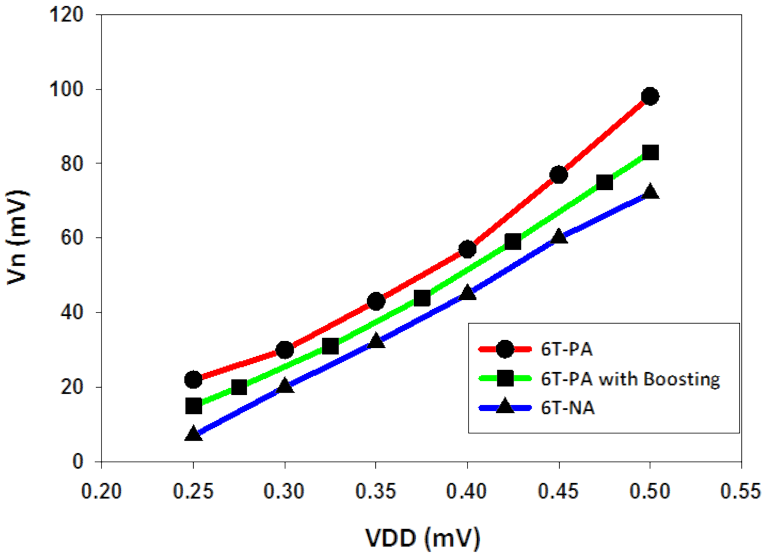
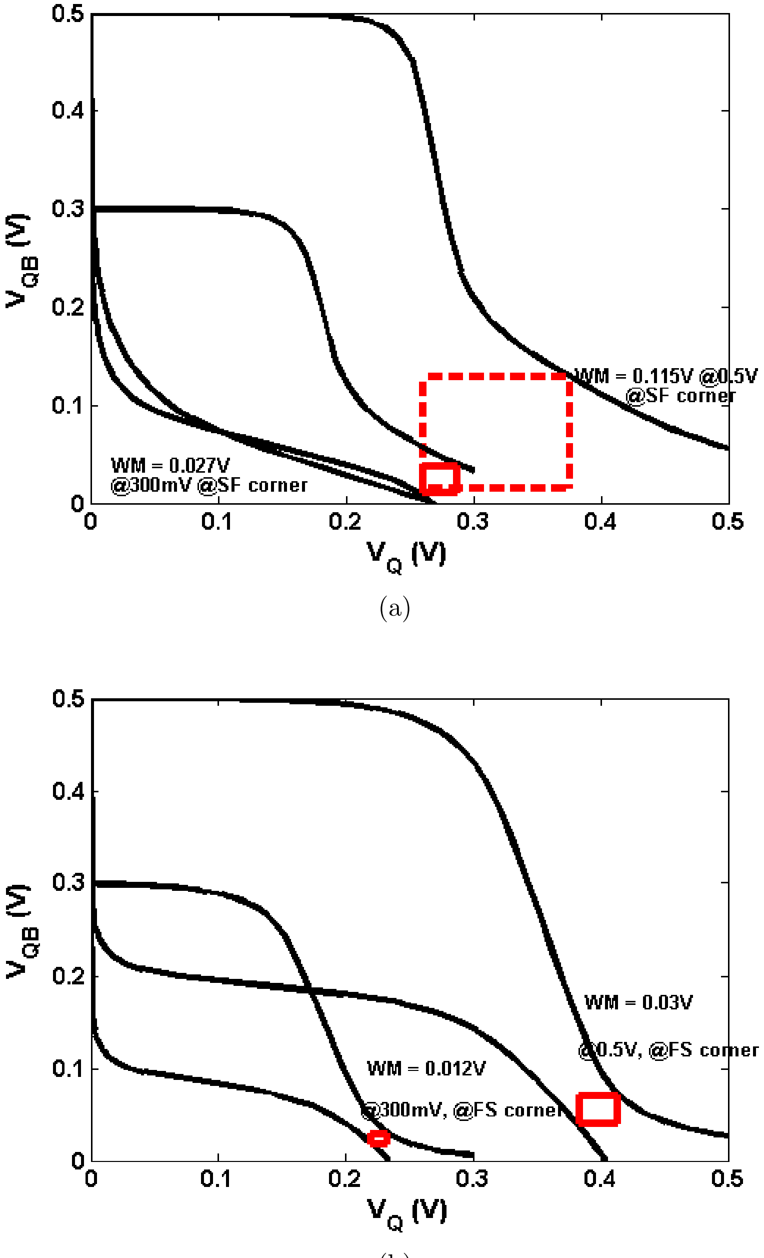


Figure 4.9: Maximum tolerable differential noise during read operation versus $\mathbf{V_{DD}}$ at the FS corner for 6T-NA and at the SF corner for 6T-PA with and without boosting, T= 25°C.

in this figure, the yield of the 6T-PA is 22% less than the 6T-NA. To overcome the weak writability of the 6T-PA, negative WL boosting is utilized. As shown in Figure 4.11(b), by applying 40 mV of negative WL boosting on the 6T-PA bitcell, the yield percentage increases up to 99.99%. The boosting circuitry and the permitted range are explained in Section 4.4.

4.4 Wordline Boosting Circuit Implementation

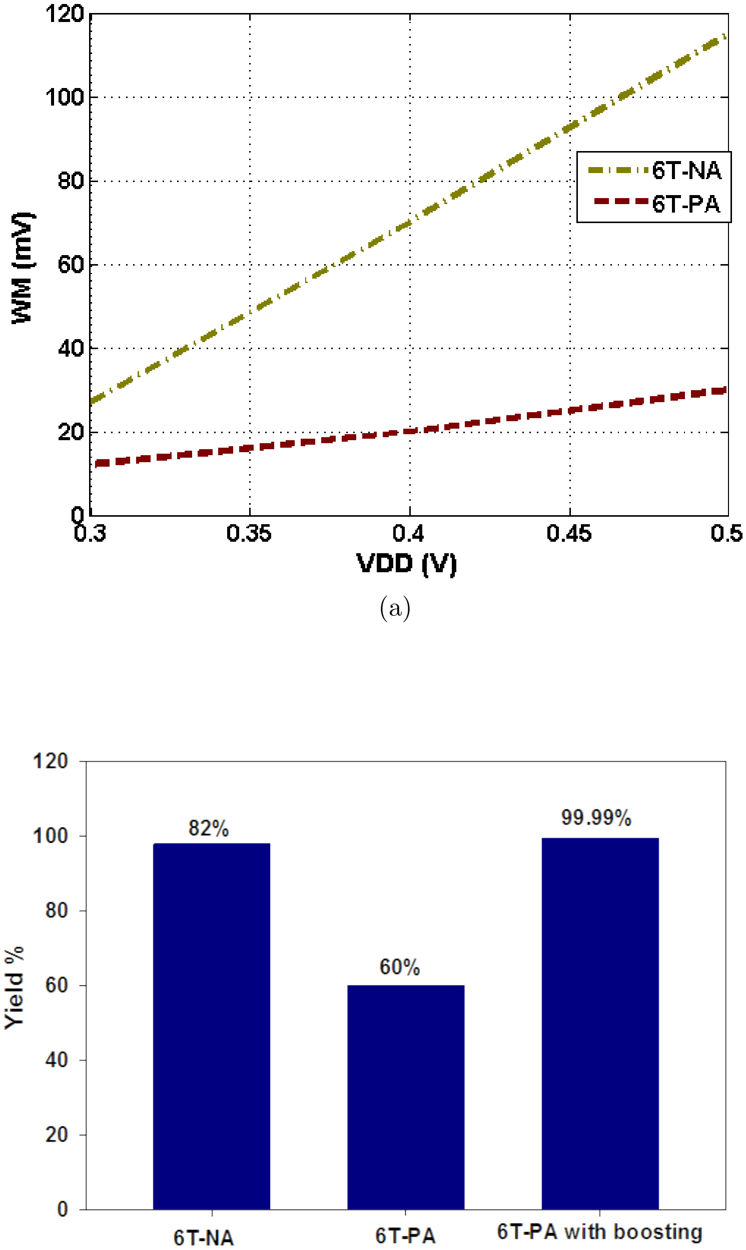
Figure 4.12 illustrates a 5-to-32 row decoder with two booster circuits and the corresponding control block. The booster circuit is externally programmable to provide the WL-boost and no-boost options. The boosting option is selected when the Mode Select (MS) signal is asserted high. Together with the CLK-EN signal, the MSB address bit, A, choose one of the two booster circuits. When both of these signals make a positive transition, the output of the corresponding NAND gate goes low switching off N1. The Miller capacitance between the gate and the drain of N1 makes its drain voltage negative. Since the N2 transistor is on, the Vboost goes to the negative voltage, and this will negatively boost the selected WL in the decoder. Figure 4.13(a) shows the read and write yield of the proposed 6T-PA bitcell versus Boost Voltage (V-Boost) at 300 mV. The yield percentage is achieved by



(a)

(b)

Figure 4.10: WM butterfly curves at $V_{DD} = 0.3$ V and $V_{DD} = 0.5$ V for a) 6T-NA at the SF corner and b) 6T-PA at the FS corner, T= 25°C.



(a)

(b)

Figure 4.11: a) WM versus $\mathbf{V_{DD}}$ for 6T-NA at the SF corner and 6T-PA at the FS corner, T= 25°C, b) Write yield percentage of the 6T-NA, 6T-PA, and 6T-PA with negative WL boosting at 250 mV.

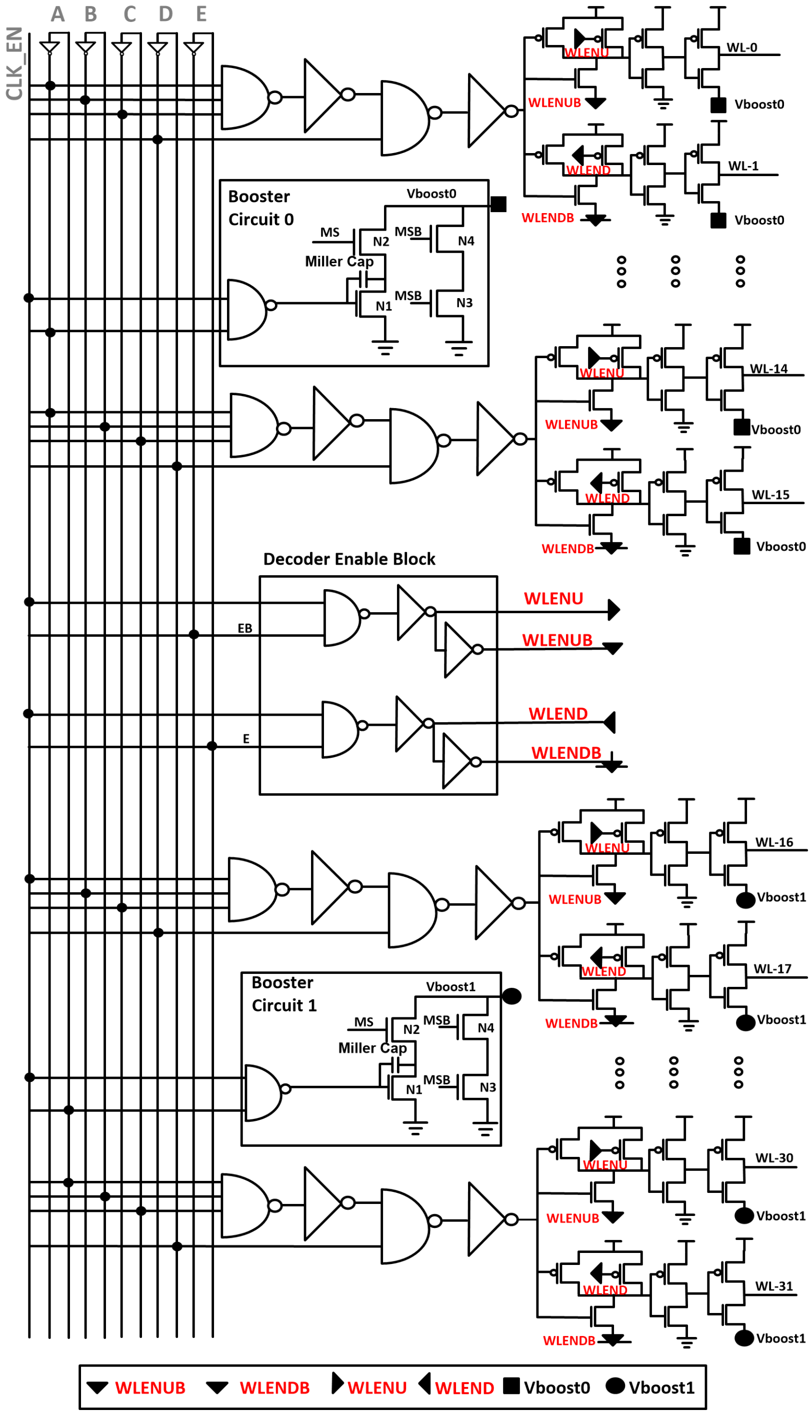
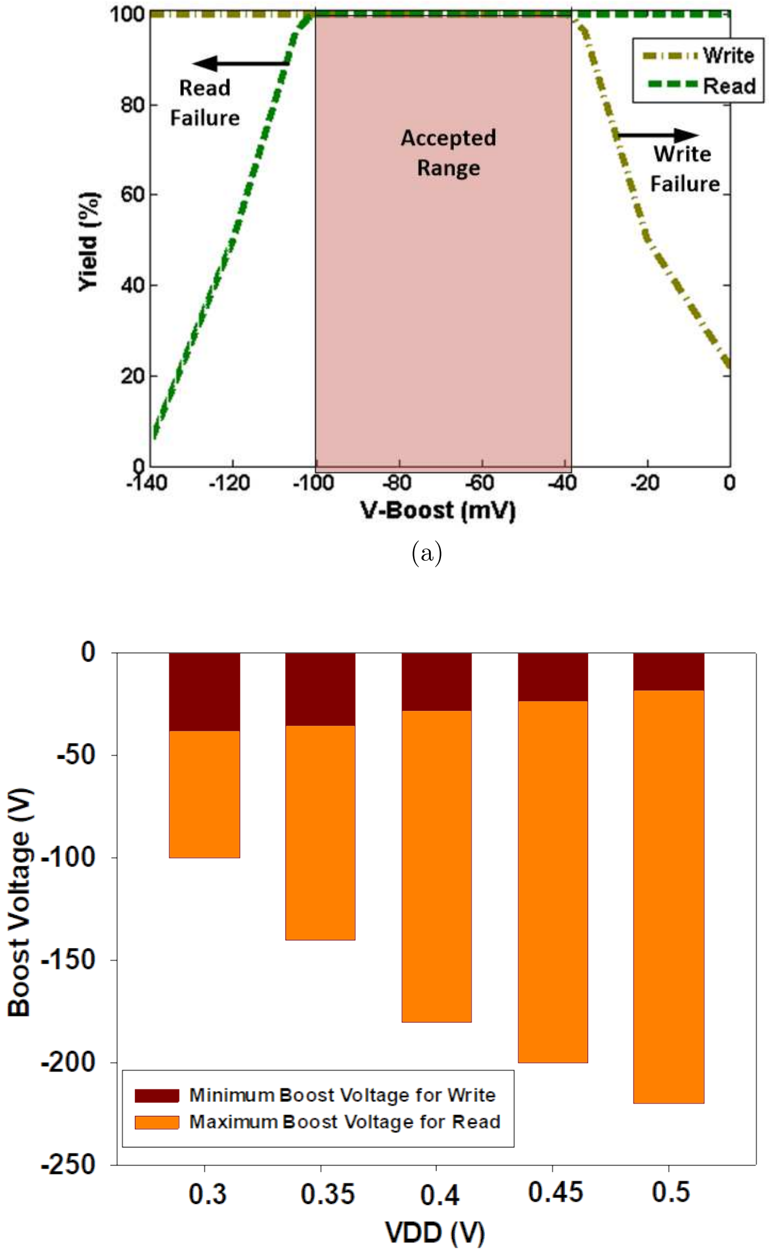


Figure 4.12: An implementation of a 5-bit row decoder with a negative WL booster circuit.



(a)

(b)

Figure 4.13: a) Write and read yield versus boosted WL voltage at 300 mV. The colored area shows the accepted range of WL boosting. b) The permitted range of the WL boosting voltage versus $\mathbf{V_{DD}}$.

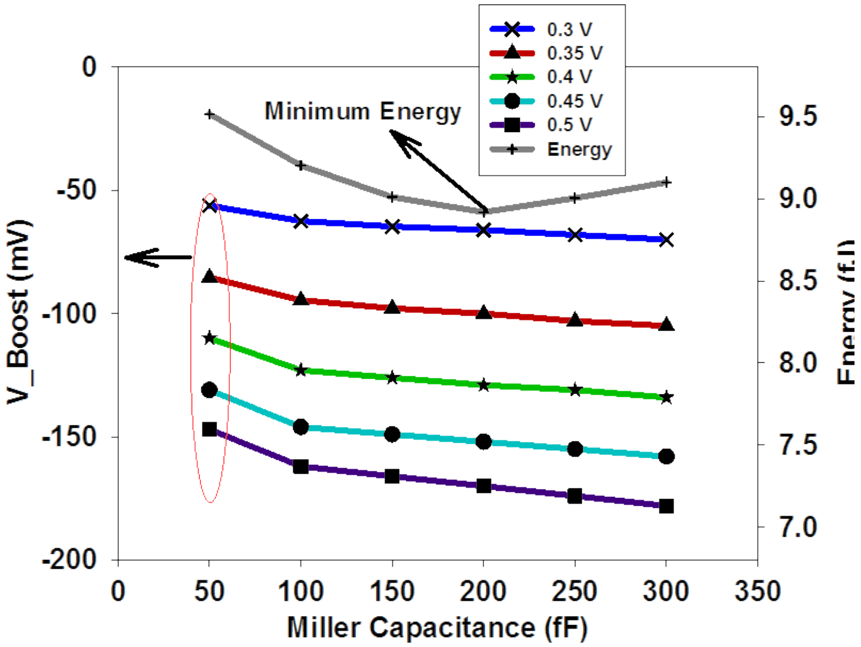


Figure 4.14: Boost voltage of the WL versus Miller capacitance at different supply voltages, TT corner, 25°C and energy consumption of the 5-bit row decoder versus the Miller capacitance at 300 mV, TT corner, 25°C.

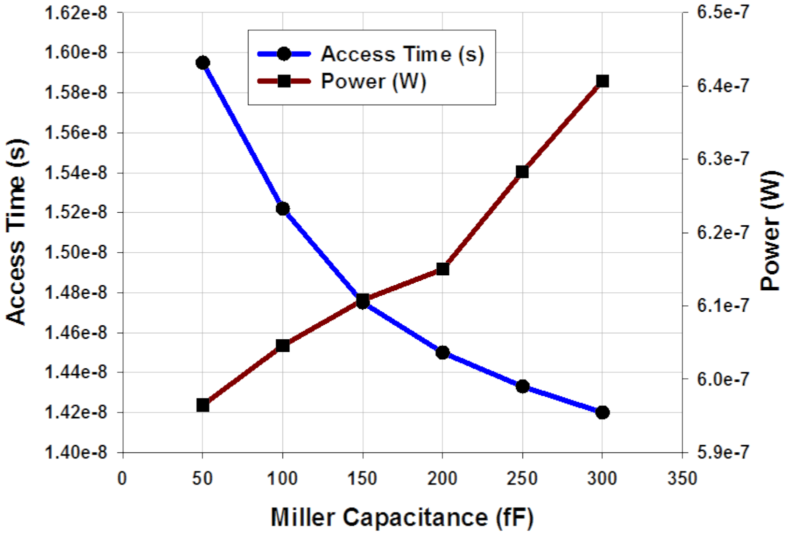


Figure 4.15: Access time and power consumption of the 5-bit row decoder versus the Miller capacitance at 300 mV, TT corner, 25°C.

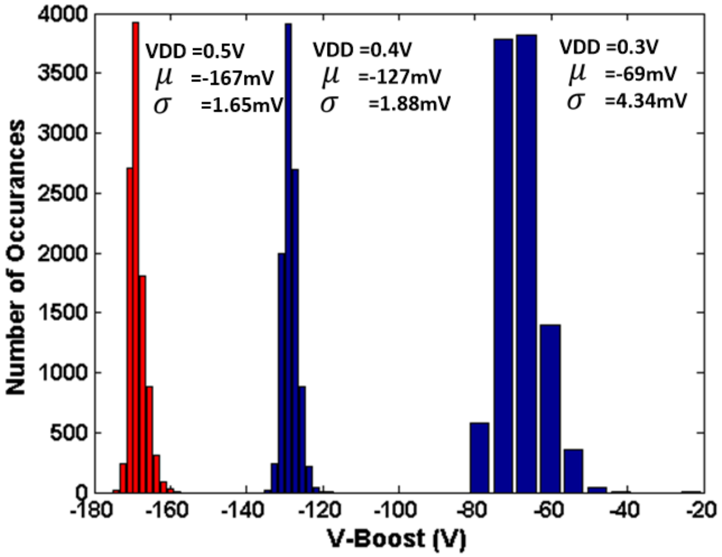


Figure 4.16: 10k Monte Carlo simulation of the boosted WL voltage at 0.3 V, 0.4 V, and 0.5 V.

counting the number of successful read (write) operations in 10k Monte Carlo read (write) operations. As shown in this figure, the minimum boosting voltage required to achieve a 100% write yield is -40 mV. Moreover, the read failure starts happening when -100 mV of the WL boosting is applied. Therefore, the permitted range of WL boosting is between -40 mV and -100 mV at 300 mV of supply voltage. Figure 4.13(b) shows the permitted range of WL boosting, the minimum required WL boosting voltage for the write operation, and the maximum level of WL boosting for the read operation at different supply voltages. As shown in this figure, the permitted range of WL boosting increases by increasing the supply voltage.

The boosted voltage is a function of the Miller capacitance, the capacitance of the Vboost node shown in Figure 4.12, and the supply voltage. Figure 4.14 shows the boost voltage versus the Miller capacitance at different supply voltages. The negative boost value increases by increasing the Miller capacitance and the supply voltage.

Figure 4.15 shows the access time and power consumption of the 5-bit decoder with the booster circuit connected to the memory array, versus the Miller capacitance. As shown in this figure, by increasing the Miller capacitance, the access time decreases while the power consumption increases. The energy consumption versus the Miller capacitance shown in Figure 4.14 is calculated by multiplying the access time by the power consumption. As shown in this figure, the minimum energy consumption occurs when a 200 fF is utilized for the Miller capacitance.

The Miller capacitance in the booster circuit is implemented with the Metal Insulator Metal (MIM) capacitor provided by the foundry, as top-level metals can be utilized, thereby reducing the area overhead for the implementation. Since the MIM capacitors are constructed using top metal layers, they are positioned on top of the decoder with no area overhead. However, since low-level metals are utilized in constructing MOS capacitances, the decoder area increases by 11%. In addition, for the subthreshold operation, MIM capacitors provide a reliable alternative to the MOS based capacitors. The MOS gate capacitance is inherently non-linear, and also has leakage associated with it. Simulation results show that a 200 fF capacitance realized through gate oxide is impacted by process variation in the subthreshold voltage regime, which leads to 30 mV variation in the boost voltage at 0.5 V. Figure 4.16 shows a 10k Monte Carlo simulation of the boost voltage at different supply voltages. As shown in this figure, the variation of the boosted voltage is about 9.9 mV at 0.5 V.

Figure 4.17 depicts the transient simulation of the WL with and without boosting. When the WL is negatively boosted, the time required to develop 100 mV of differential voltage (**ΔBL**) is reduced by 10 ns. In addition, simulation results show that activating

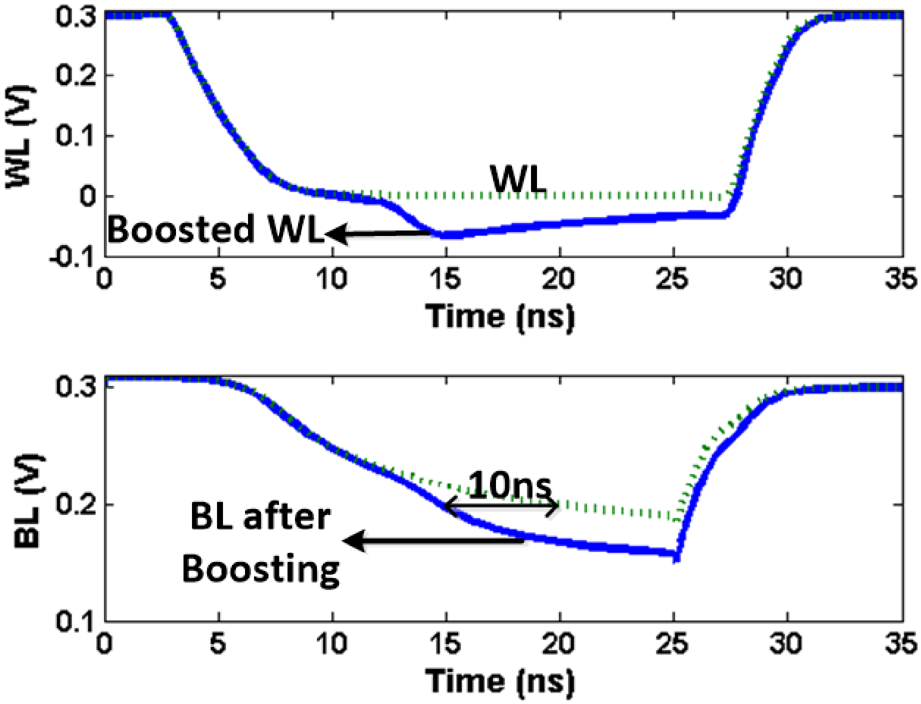


Figure 4.17: Simulated timing of WL and BLs for boosted and non-boosted options at 300 mV, TT corner, 25°C.

the booster circuitry increases the average consumed total current by 2.6%.

4.5 Read and Leakage Current

Amongst other factors, the SRAM read current ($\mathbf{I_{Read}}$) determines its operational speed. In particular, the $\mathbf{I_{Read}}$ can be constrained either by the driver transistor or access transistor. For example, for the conventional 6T-NA cell, the saturated access transistor limits the read current. The driver transistor is typically designed to be stronger to ensure read stability and is capable of sinking a larger current. The situation is similar for the 6T-PA where the saturated PMOS access transistor limits the cell current. However, owing to its small mobility, the $\mathbf{I_{Read}}$ is substantially smaller. Figure 4.18 illustrates the $\mathbf{I_{Read}}$ of both bitcells. As shown in this figure, for the iso-area, the $\mathbf{I_{Read}}$ of the 6T-NA is higher than that of the 6T-PA bitcell. For example, the $\mathbf{I_{Read}}$ of the 6T-NA and the 6T-PA at 290 mV is 180 nA and 36 nA, respectively. Negative WL boosting enhances the $\mathbf{I_{Read}}$ of the 6T-PA substantially, specifically in the subthreshold region. For example, a negative WL boost of 65 mV at VDD of 290 mV increases the read current to 140 nA.

Figure 4.18 also shows the leakage current (I_{Lakage}) of the 6T-NA and 6T-PA bitcells. The leakage current for the 6T-NA and the 6T-PA at 290 mV is 0.44 nA and 0.22 nA, respectively. Therefore, a SRAM array with 6T-PA cell has the potential to reduce its leakage current.

A sense amplifier requires sufficient differential voltage to make a reliable decision which necessitates not only a high cell read current, but also as low as possible leakage current, through unselected cells in the column. Consequently, the ratio of I_{Read} $/I_{Lakage}$ is an important parameter that restricts the number of cells in a column. Figure 4.18 illustrates this ratio for 6T-NA and 6T-PA cells. As expected, the 6T-NA cell is substantially better compared to the 6T-PA. However, a negative WL voltage boost significantly improves this ratio, specifically for sub-350 mV operation.

4.6 Test Chip Measurement and implementation

A test chip with 2 kb SRAM was designed and fabricated in the TSMC 65 nm GP CMOS technology. The I/Os in this technology operate at 2.5 V and are capable of interfacing with the core logic at 1 V. Level shifters capable of shifting 200 mV inputs to 1 V, and vice versa are designed for this test chip. The die photo is shown in Figure 4.19. To test the

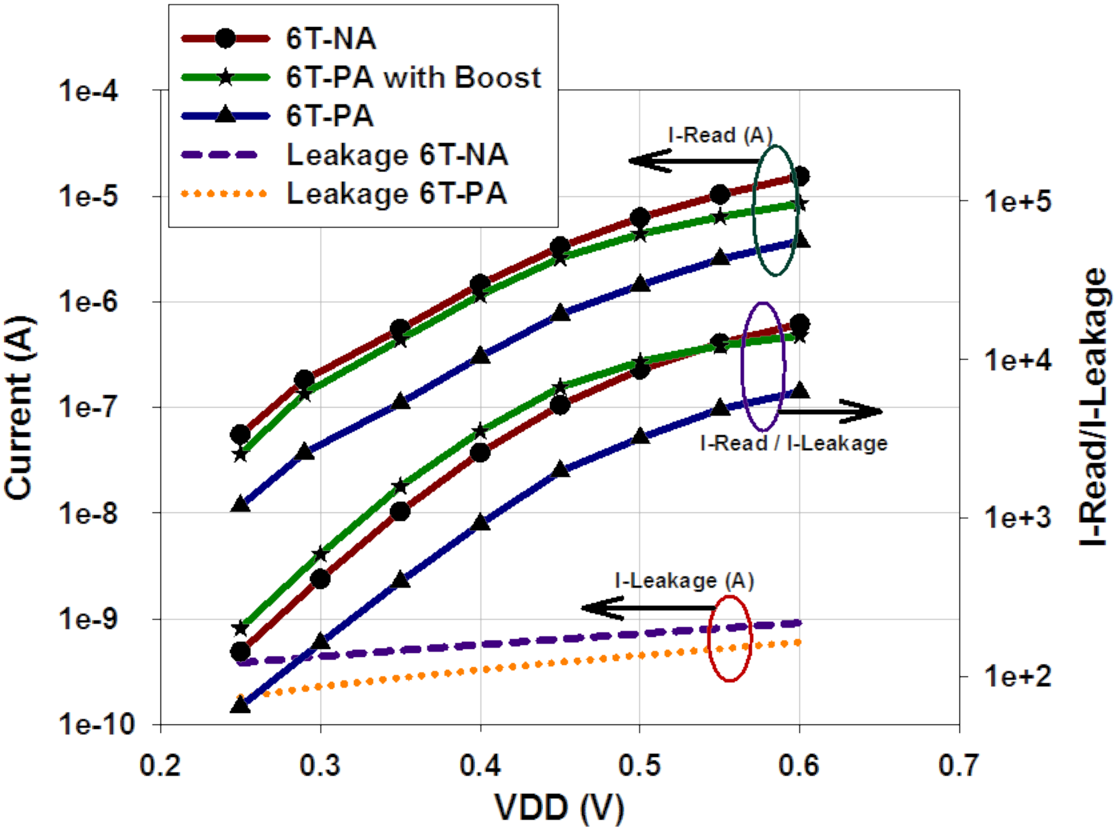


Figure 4.18: Read current, leakage current, and read current to leakage current ratio of the 6T-NA and 6T-PA bitcells versus the supply voltage, at the TT corner, 25°C.

functionality of each die, write and read accesses are performed with random data. A total of 10 dies were measured and found to meet functional requirements. Within these samples all were able to operate at 310 mV, nine of the dies were able to operate at 300 mV, and two at 290 mV.

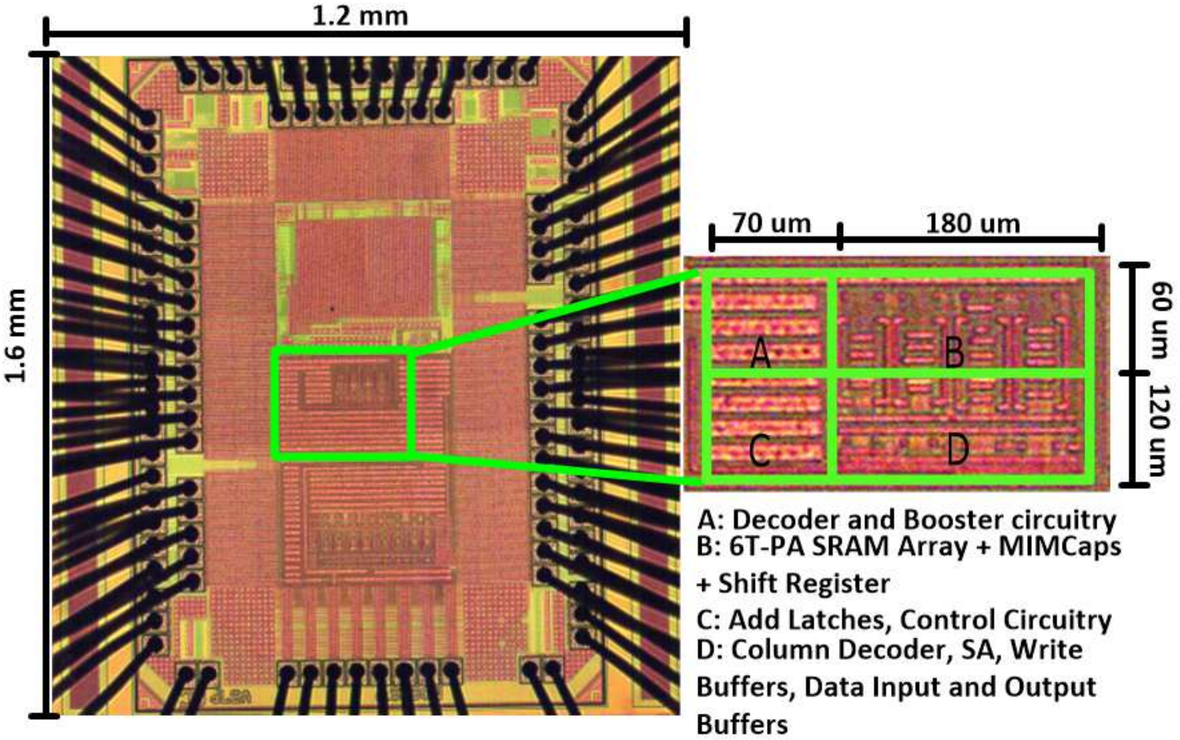
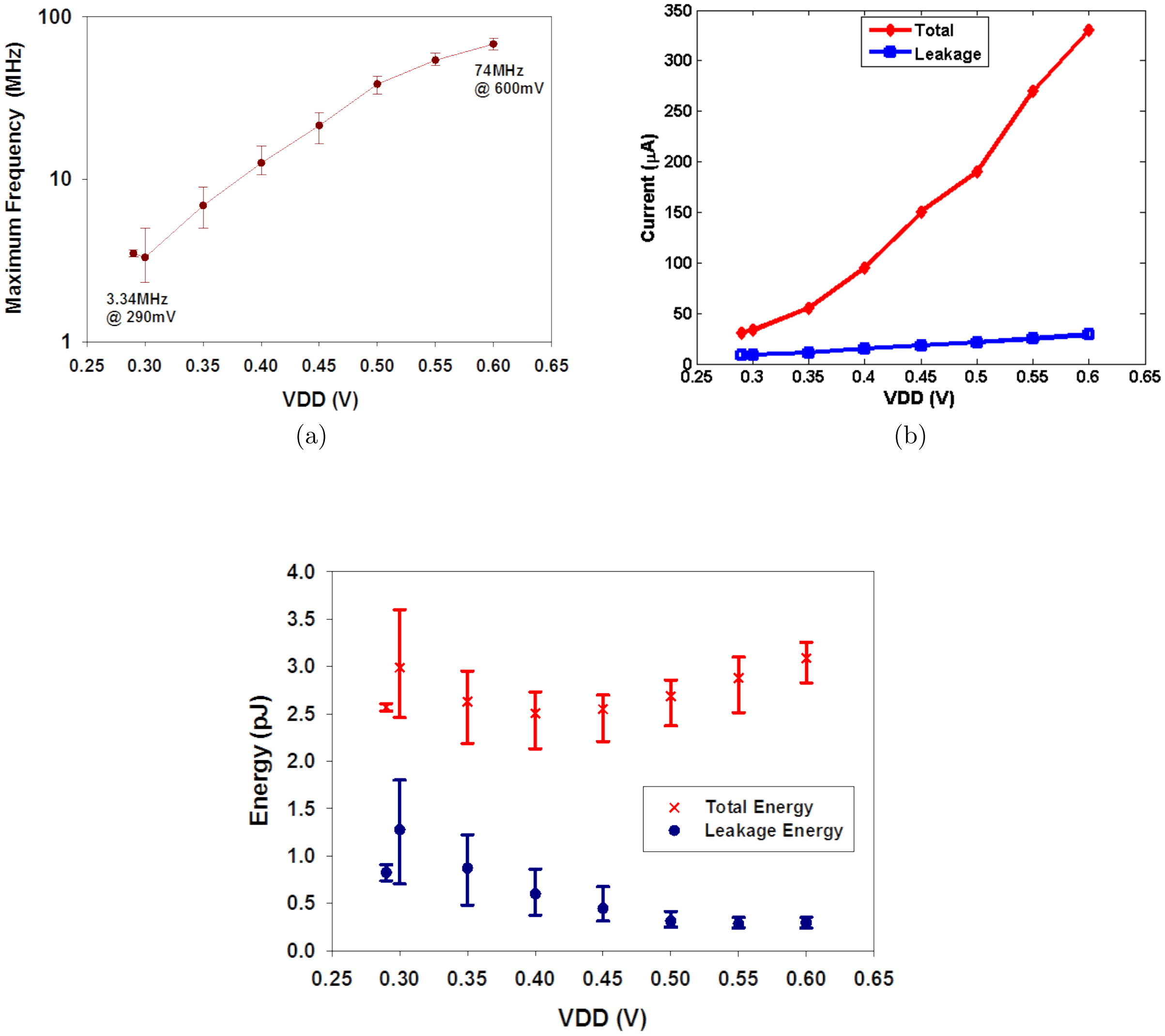

1.2 mm
1.6 mm
70 um
180 um
60 um
120 um
A
B
C
D
A: Decoder and Booster circuitry
B: 6T-PA SRAM Array + MIMCaps + Shift Register
C: Add Latches, Control Circuitry
D: Column Decoder, SA, Write Buffers, Data Input and Output Buffers

Figure 4.19: Micro-graphic image of the fabricated chip in the 65 nm CMOS technology.

Figure 4.20(a) shows the measured maximum operational frequency versus the supply voltage. Each vertical bar shows the maximum, minimum, and the average measured data. The maximum frequency is achieved as high as 3.34 MHz at 290 mV. At 0.6 V the maximum frequency achieved is 74 MHz.

Figure 4.20(b) illustrates the measured total and leakage current. The total current is measured while performing successive write and read operations at different addresses. The average of this current is shown in the figure. The leakage current at different supply voltages are measured while the macro is inactive. The total and leakage currents are measured as 30 and 8.5 μA, respectively, at 290 mV. Measurement results show that the total average current increases by 3% when the booster circuit is activated.



(a)

(b)

(c)

Figure 4.20: a) Measured frequency of operation with respect to the supply voltage; b) Measured total current and leakage current with respect to the supply voltage; c) Total energy and leakage energy with respect to the supply voltage. T= 25°.

The energy consumption can be computed by dividing the power consumption by the maximum frequency. The total and leakage energy consumption is shown in Figure 4.20(c). The minimum total energy is calculated as 1.1 fJ/bit at 400 mV and the leakage energy is calculated as 0.37 fJ/b at 290 mV.

Table 4.2 summarizes and compares the key features of our design with previous subthreshold SRAMs that include the 6T [32], 7T [38], 8T [34] [33], and 9T [35] bitcells. Comparing our design with the designs shown in Table 4.2 reveals that utilizing 6T-PA bitcell and incorporating the WL boosting enables us to reduce the supply voltage to 290 mV, which is lower than the $\mathbf{V_{min}}$ reported for the 8T in [34]. The designs in [32] and [35] was able to further reduce the $\mathbf{V_{min}}$ close to 200 mV at the cost of significant additional bitcell area. The design in [33] has reduced the $\mathbf{V_{min}}$ of 350 mV of the 8T in [34] by manipulating the stored data at the system level that eliminates the worst-case data distribution in each column. This design reveals how system-level approaches can improve the key parameters of application-specific SRAMs.

To perform a comparison on the speed of these designs, the speed of all the memory macros are reported at 350 mV. The comparison shows that our design can operate at a higher speed due to the combination of the differential sensing and negative WL boosting, as compared to the designs in [32–34, 38]. The 9T bitcell in [35] utilizes differential sensing, however, non-minimum length transistors and high threshold voltage (low speed) transistors are used in their bitcell.

To enable reliable low power operations, drivers and peripheral circuits, especially the circuits connected to the IOs such as level shifters, latches, buffers, and flip-flops are over designed which, in turn increases the leakage current and the energy consumption.

Finally, the energy-delay-product (EDP) per bit of all the designs are compared in Table 4.2. The comparison shows that except for the design in [33], which is optimized for video specific applications, our design has the lowest EDP per bit.

4.7 Conclusion

For the subthreshold operation, the conventional NMOS access 6T SRAM cell suffers from poor SNM and WM. In this chapter, a 6T SRAM bitcell with PMOS access transistors and enhanced SNM and WM operating in the subthreshold region is proposed. The PMOS access transistors are utilized to increase the stability of the bitcell during the read operation. This is verified by simulation and analytical analysis. To overcome weak writability of the proposed 6T bitcell, WL boosting is incorporated in this design.

Table 4.2: Comparison with Chosen Previous Subthreshold SRAMs.

Design	**Technology (nm)**	**Cell Type**	**Size (kb)**	**V_{min} (mV)**	**$I_{Leakage}$ (nA/b)**	**Speed (MHz) @ 0.35 V** [a]	**E_{min} (fJ/b)**	**EDP ($\times 10^{-21} \frac{J.s}{b}$) @ 0.35 V**	**Cell Area (μm^2)**
6TPA Array [66]	65	6T	2	290	4.25	9.2	1.1	0.125	2.15
JSSC08 [32]	130	6T	2	210	119	0.45	0.55	1.22	4
JSSC13 [38]	65	7T	32	260	N.A	8.5	0.175 [c]	N.A [b]	N.A [b]
JSSC08 [34]	65	8T	256	350	0.024	0.025	0.5	20	N.A[b]
JSSC16 [33]	65	8T	32	200	0.034	2	0.031	0.015	1.35
JSSC13 [35]	65	9T	2	220	N.A [b]	1.2	0.3	15.16	4.6

[a] The speed is extracted based on the given data for each reference.

[b] Not Available.

[c] The energy data is reported only for 260 mV.

The negative WL boosting also helps to compensate for the loss of speed of the PMOS access transistors. A 2kb SRAM is fabricated in the 65 nm TSMC technology. The measurement results show 3.34 MHz of speed and 8.5 μA of leakage current at 290 mV. The minimum energy is observed as 1.1 fJ/bit at 400 mV.

Chapter 5

Conclusions and Future Work

SRAM circuits contribute significantly to the total power consumption of mobile devices, especially in the standby mode. Therefore, designing SRAM circuits with low power consumption is in great demand [2]. The main approach to achieve this goal is by reducing the supply voltage. However, a straight forward reduction of the supply voltage of SRAM circuits impose critical challenges such as a reduced WM, SNM, and speed when the bitcells are being accessed. In this thesis, system-level, architectural-level, and transistor-level techniques are proposed to mitigate SRAM limitations that occur when operating in the subthreshold region. These proposed techniques are also backed by theory and analyzed analytically. To further improve the above mentioned contradictory requirements of the conventional 6T bitcell, a 4-level programmable WL boosting technique is exploited. Incorporating programmability enables independent optimization of read and write margins. Moreover, the 6T bitcell does not have to be over-designed for low-voltage operation. The measurement results on a 16-kb SRAM silicon prototype show that the WL boosting technique reduces the minimum supply voltage for write operation down to 340 mV at the speed of 6 MHz.

Chapter 1 presents the motivation and problem statement. A detailed explanation of previous research performed in the area of the low voltage SRAM design is also provided.

Chapter 2 investigates the architecture of the SRAM circuit. The circuit implementation of the main sub-blocks of the SRAM architecture such as the address buffers, the row decoder, the SRAM bitcell, the read/write column decoder, the SA array, and the input/output data buffers are described in each subsection. Different operational modes (read, write, and hold) of a SRAM circuit are presented. The design challenges and Figure of Merit (FOM), such as WM and SNM, in each mode of operation, as well as the required considerations to overcome these challenges, are provided.

In Chapter 3, a programmable WL boosting technique with four levels of boosting to improve the contradictory requirements of the 6T bitcell is presented. Incorporating programmability enables a process-tolerant design. A 16-kb SRAM memory is fabricated in the 65 nm TSMC GP CMOS technology. Measurement results show that the operational frequency improves by up to 33.3% when the WL boosting technique is applied. By using the WL boosting, the supply voltage can be decreased while maintaining the same operational frequency. This, in turn, allows the energy consumption to be reduced by 22.2%.

In Chapter 4, a 6T bitcell with improved read and write margin, optimized for low voltage applications, is proposed. The read stability is improved by exploiting the PMOS access transistors. The PMOS access transistor has a lower mobility and therefore the 6T bitcell provides a higher cell ratio, thus giving higher read stability. In addition, utilizing the PMOS access transistors can improve the resistivity of the back-to-back inverters to hold the data and alleviate the effect of the ZLD. Moreover, to overcome the weak writability of the new bitcell, a negative WL boosting is exploited. The negative WL boosting shows up to a 3 times improvement in the WM of the proposed bitcell compared to the conventional bitcell. A 2-kb fabricated SRAM using the 65 nm CMOS technology shows 3.34 MHz of speed and 8.5 μA of leakage current at 290 mV. The ZLD of both the conventional bitcell and the proposed bitcell is also formulated for in the subthreshold region.

5.1 Future Work

The notion of the SNM calculated based on the butterfly curves has been used as the main FOM in designing digital circuits for four decades. However, the main drawback of this FOM is that it is based on the DC criteria of digital circuits. In addition, this method is a voltage-based method where the voltage of one of the internal nodes is forced by a DC supply voltage. In a DC simulation, the WL is completely turned ON and the BL is fixed at $\mathbf{V_{DD}}$. However, in a real transient read simulation, the BL (or BLB) starts discharging as soon as the WL starts turning on and the internal nodes adjust their voltage in correspondence to the noise being injected on the internal nodes. It is necessary to explore a current-based method where a current source can dynamically simulate the effect of a noise injection during a real read simulation. By applying a current noise on the bitcell, unlike the DC-voltage-based butterfly curves, the voltage of internal nodes is not fixed and could be dynamically adjusted, based on the current noise value being injected on the bitcell.

One of the main obstacles to design more compact-area SRAMs is the BL leakage. The BL leakage limits the number of bitcells in each column. The BL leakage can be reduced if the total number of zeros or ones stored in each column is known. By knowing the total number of zeros and ones, the required leakage current proportional to the number of zeros can be compensated on the BL or BLB. This, in turn, could significantly reduce the effect of the BL or BLB leakage. Therefore, the worst-case timing scenario would be avoided and a greater number of bitcells could be incorporated into a column for a more compact SRAM area. Saving and updating the number of zeros and ones during each write operation is the main challenge of this idea.

Even though dozens of new bitcells have been proposed over the last decade to reduce the supply voltage below the subthreshold region, there is still room to create new bitcells and further reduce the minimum operating supply voltage. Our primary simulations show that combining the 6TPA bitcell provided by this research with the conventional 6T bitcell and providing a new 8T bitcell can lead to significant reduction in the supply voltage.

References

[1] R. Islam, A. Brand, and D. Lippincott, "Low Power SRAM Techniques for Handheld Products," in *Proceedings of the 2005 International Symposium on Low Power Electronics and Design, 2005. ISLPED'05.* IEEE, 2005, pp. 198–202. 1, 3

[2] L. T. Clark, N. Deutscher, S. Demmons, and F. Ricci, "Standby Power Management for a 0.18 μm Microprocessor," in *Proceedings of the 2002 international symposium on Low power electronics and design.* ACM, 2002, pp. 7–12. 1, 8, 79

[3] T.-H. Kim, J. Liu, J. Keane, and C. Kim, "A High-Density Subthreshold SRAM with Data-Independent Bitline Leakage and Virtual Ground Replica Scheme," in *IEEE International Solid-State Circuits Conference. Digest of Technical Papers.*, 2007, pp. 330–606. 1, 10, 12

[4] N. S. Kim, K. Flautner, D. Blaauw, and T. Mudge, "Circuit and Microarchitectural Techniques for Reducing Cache Leakage Power," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 12, no. 2, pp. 167–184, 2004. 2, 31

[5] H. Qin, Y. Cao, D. Markovic, A. Vladimirescu, and J. Rabaey, "SRAM Leakage Suppression by Minimizing Standby Supply Voltage," in *5th International Symposium on Quality Electronic Design.* IEEE, 2004, pp. 55–60. 2

[6] K. Kanda, T. Miyazaki, M. K. Sik, H. Kawaguchi, and T. Sakurai, "Two Orders of Magnitude Leakage Power Reduction of Low Voltage SRAMs by Row-By-Row Dynamic Vdd Control (RRDV) Scheme," in *15th Annual IEEE International ASIC/SOC Conference.* IEEE, 2002, pp. 381–385. 2, 6, 31

[7] S. Mukhopadhyay, R. Rao, J.-J. Kim, and C.-T. Chuang, "SRAM Write-Ability Improvement with Transient Negative Bit-Line Voltage," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 19, no. 1, pp. 24–32, Jan 2011. 3

[8] K. Zhang, U. Bhattacharya, Z. Chen, F. Hamzaoglu, D. Murray, N. Vallepalli, Y. Wang, B. Zheng, and M. Bohr, "A 3-GHz 70Mb SRAM in 65nm CMOS Technology with Integrated Column-Based Dynamic Power Supply," *IEEE Journal of Solid-State Circuits*, vol. 41, no. 1, pp. 146–151, 2006. 3

[9] M. Yamaoka, N. Maeda, Y. Shinozaki, Y. Shimazaki, K. Nii, S. Shimada, K. Yanagisawa, and T. Kawahara, "Low-Power Embedded SRAM Modules with Expanded Margins for Writing," *IEEE Intl Solid-State Circuits Conf. Digest of Technical Papers*, pp. 480–481, 2005. 3, 8

[10] H. Yamauchi, T. Iwata, H. Akamatsu, and A. Matsuzawa, "A 0.8 V/100 MHz/Sub-5 mW-Operated Mega-Bit SRAM Cell Architecture with Charge-Recycle Offset-Source Driving (OSD) Scheme," in *Symposium on VLSI Circuits. Digest of Technical Papers.* IEEE, 1996, pp. 126–127. 3

[11] H. Mizuno and T. Nagano, "Driving Source-Line Cell Architecture for Sub-l-V High-Speed Low-Power Applications," in *Symposium on VLSI Circuits, 1995. Digest of Technical Papers.* IEEE, 1995, pp. 25–26. 3

[12] N. Shibata, "A Switched Virtual-GND Level Technique for Fast and Low Power SRAM's," *IEICE Trans. Electron*, vol. E80-C, no. 8, p. 15981607, 1997. 3

[13] K. Osada, Y. Saitoh, E. Ibe, and K. Ishibashi, "16.7-fA/cell Tunnel-Leakage-Suppressed 16-Mb SRAM for Handling Cosmic-Ray-Induced Multierrors," *IEEE Journal of Solid-State Circuits*, vol. 38, no. 11, pp. 1952–1957, 2003. 3

[14] M. Yamaoka, K. Osada, and K. Ishibashi, "0.4-V Logic-Library-Friendly SRAM Array using Rectangular-Diffusion Cell and Delta-Boosted-Array Voltage Scheme," *IEEE Journal of Solid-State Circuits,*, vol. 39, no. 6, pp. 934–940, 2004. 3

[15] H. Pilo, J. Barwin, G. Braceras, C. Browning, S. Burns, J. Gabric, S. Lamphier, M. Miller, A. Roberts, and F. Towler, "An SRAM Design in 65nm and 45nm Technology Nodes Featuring Read and Write-Assist Circuits to Expand Operating Voltage," in *Symposium on VLSI Circuits, 2006. Digest of Technical Papers. 2006.* IEEE, 2006, pp. 15–16. 4

[16] B.-D. Yang and L.-S. Kim, "A Low-Power SRAM using Hierarchical Bit Line and Local Sense Amplifier," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 6, pp. 1366–1376, 2005. 4, 31

[17] B. S. Amrutur and M. A. Horowitz, “A Replica Technique for Wordline and Sense Control in Low-Power SRAMs,” *IEEE Journal of Solid-State Circuits,*, vol. 33, no. 8, pp. 1208–1219, 1998. 4

[18] M. Yamaoka, N. Maeda, Y. Shinozaki, Y. Shimazaki, K. Nii, S. Shimada, K. Yanagisawa, and T. Kawahara, “90-nm Process-Variation Adaptive Embedded SRAM Modules with Power-Line-Floating Write Technique,” *Solid-State Circuits, IEEE Journal of*, vol. 41, no. 3, pp. 705–711, March 2006. 4

[19] K. Zhang, K. Hose, V. De, and B. Senyk, “The Scaling of Data Sensing Schemes for High Speed Cache Design in Sub-0.18 μm Technologies,” in *Symposium on VLSI Circuits, 2000. Digest of Technical Papers. 2000.* IEEE, 2000, pp. 226–227. 4, 31

[20] M. Khellah, Y. Ye, N. S. Kim, D. Somasekhar, G. Pandya, A. Farhang, K. Zhang, C. Webb, and V. De, “Wordline Amp; Bitline Pulsing Schemes for Improving SRAM Cell Stability in Low-Vcc 65nm CMOS Designs,” in *VLSI Circuits, 2006. Digest of Technical Papers. 2006 Symposium on*, 2006, pp. 9–10. 4

[21] S. Ohbayashi, M. Yabuuchi, K. Nii, Y. Tsukamoto, S. Imaoka, Y. Oda, M. Igarashi, M. Takeuchi, H. Kawashima, H. Makino *et al.*, “A 65 nm SoC Embedded 6T-SRAM Design for Manufacturing with Read and Write Cell Stabilizing Circuits,” in *Symposium on VLSI Circuits, 2006. Digest of Technical Papers. 2006.* IEEE, 2006, pp. 17–18. 5

[22] M. Yabuuchi, K. Nii, Y. Tsukamoto, S. Ohbayashi, S. Imaoka, H. Makino, Y. Yamagami, S. Ishikura, T. Terano, T. Oashi *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.* IEEE, 2007, pp. 326–606. 5

[23] K. Takeda, T. Saito, S. Asayama, Y. Aimoto, H. Kobatake, S. Ito, T. Takahashi, M. Nomura, K. Takeuchi, and Y. Hayashi, “Multi-step Word-line Control Technology in Hierarchical Cell Architecture for Scaled-down High-density SRAMs,” *Solid-State Circuits, IEEE Journal of*, vol. 46, no. 4, pp. 806–814, 2011. 5, 32, 49

[24] M. Iijima, K. Seto, M. Numa, A. Tada, and T. Ipposhi, “Low Power SRAM with Boost Driver Generating Pulsed Wordline Voltage for Sub-1V Operation,” *Journal of Computers*, vol. 3, no. 5, pp. 34–40, 2008. 5

[25] A. Kawasumi, T. Yabe, Y. Takeyama, O. Hirabayashi, K. Kushida, A. Tohata, T. Sasaki, A. Katayama, G. Fukano, Y. Fujimura, and N. Otsuka, "A Single-Power-Supply 0.7V 1GHz 45nm SRAM with an Asymmetrical Unit-β-ratio Memory Cell," in *IEEE International Digest of Technical Papers in Solid-State Circuits Conference, 2008. ISSCC 2008.* IEEE, 2008, pp. 382–622. 5

[26] K. Kanda, H. Sadaaki, and T. Sakurai, "90% Write Power-Saving SRAM using Sense-Amplifying Memory Cell," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 6, pp. 927–933, 2004. 5, 10, 12

[27] Y. Ye, M. Khellah, D. Somasekhar, A. Farhang, and V. De, "A 6-GHz 16-kB L1 Cache in a 100-nm Dual-VT Technology using a Bitline Leakage Reduction (BLR) Technique," *IEEE Journal of Solid-State Circuits*, vol. 38, no. 5, pp. 839–842, 2003. 6

[28] K. Agawa, H. Hara, T. Takayanagi, and T. Kuroda, "A Bitline Leakage Compensation Scheme for Low-Voltage SRAMs," *IEEE Journal of Solid-State Circuits*, vol. 36, no. 5, pp. 726–734, 2001. 6

[29] A. Alvandpour, D. Somasekhar, R. Krishnamurthy, V. De, S. Borkar, and C. Svensson, "Bitline Leakage Equalization for Sub-100nm Caches," in *Proceedings of the 29th European Solid-State Circuits Conference, 2003. ESSCIRC'03.* IEEE, 2003, pp. 401–404. 6

[30] F. Ramezankhani, *Designing Faster CMOS Sub-threshold Circuits Utilizing Channel Length Manipulation.* Carleton University, 2012. 8

[31] K. W. Mai, T. Mori, B. S. Amrutur, R. Ho, B. Wilburn, M. A. Horowitz, I. Fukushi, T. Izawa, and S. Mitarai, "Low-Power SRAM Design using Half-Swing Pulse-Mode Techniques," *IEEE Journal of Solid-State Circuits*, vol. 33, no. 11, pp. 1659–1671, 1998. 8

[32] B. Zhai, S. Hanson, D. Blaauw, and D. Sylvester, "A Variation-Tolerant Sub-200 mV 6-T Subthreshold SRAM," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 10, pp. 2338–2348, 2008. 9, 11, 32, 45, 46, 47, 49, 76, 77

[33] A. T. Do and Z. C. Lee and B. Wang and I. J. Chang and X. Liu and T. T. H. Kim, "0.2 V 8T SRAM With PVT-Aware Bitline Sensing and Column-Based Data Randomization," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 6, pp. 1487–1498, June 2016. 9, 14, 49, 76, 77

[34] N. Verma and A. P. Chandrakasan, "A 256-kb 65-nm 8T Subthreshold SRAM Employing Sense-Amplifier Redundancy," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 1, pp. 141–149, Jan 2008. 9, 13, 14, 32, 45, 46, 47, 49, 76, 77

[35] S. Lutkemeier, T. Jungeblut, H. K. O. Berge, S. Aunet, M. Porrmann, and U. Ruckert, "A 65 nm 32-b Subthreshold Processor With 9T Multi-Vt SRAM and Adaptive Supply Voltage Control," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 1, pp. 8–19, Jan 2013. 9, 13, 49, 76, 77

[36] M. E. Sinangil and A. P. Chandrakasan, "Application-specific SRAM Design using Output Prediction to Reduce Bit-Line Switching Activity and Statistically Gated Sense Amplifiers for Up to 1.9 Lower Energy/Access," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 1, pp. 107–117, 2014. 9, 15

[37] J. S. Wang, P. Y. Chang, T. S. Tang, J. W. Chen, and J. I. Guo, "Design of Subthreshold SRAMs for Energy-Efficient Quality-Scalable Video Applications," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 1, no. 2, pp. 183–192, June 2011. 9, 14

[38] M.-F. Chang, M.-P. Chen, L.-F. Chen, S.-M. Yang, Y.-J. Kuo, J.-J. Wu, H.-Y. Su, Y.-H. Chu, W.-C. Wu, T.-Y. Yang *et al.*, "A Sub-0.3 V Area-Efficient L-Shaped 7T SRAM With Read Bitline Swing Expansion Schemes Based on Boosted Read-Bitline, Asymmetric-V Read-Port, and Offset Cell VDD Biasing Techniques," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 10, pp. 2558–2569, 2013. 9, 13, 76, 77

[39] K. Takeda, Y. Hagihara, Y. Aimoto, M. Nomura, Y. Nakazawa, T. Ishii, and H. Kobatake, "A Read-Static-Noise-Margin-Free SRAM Cell for Low-VDD and High-Speed Applications," *IEEE Journal of Solid-State Circuits*, vol. 41, no. 1, pp. 113–121, 2006. 9, 12, 45, 46, 47

[40] Y.-W. Chiu, Y.-H. Hu, M.-H. Tu, J.-K. Zhao, Y.-H. Chu, S.-J. Jou, and C.-T. Chuang, "40-nm Bit-Interleaving 12T Subthreshold SRAM with Data-Aware Write-Assist," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 61, no. 9, pp. 2578–2585, 2014. 10, 13, 45, 46, 47

[41] J. J. Wu, Y. H. Chen, M. F. Chang, P. W. Chou, C. Y. Chen, H. J. Liao, M. B. Chen, Y. H. Chu, W. C. Wu, and H. Yamauchi, "A Large $\sigma V_{TH}/V_{DD}$ Tolerant Zigzag 8T SRAM With Area-Efficient Decoupled Differential Sensing and Fast Write-Back Scheme," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 4, pp. 815–827, April 2011. 10, 13

[42] B. Wang, T. Q. Nguyen, A. T. Do, J. Zhou, M. Je, and T. T. Kim, "A 0.2 V 16-Kb 9T SRAM with Bitline Leakage Equalization and CAM-Assisted Write Performance Boosting for Improving Energy Efficiency," in *Solid State Circuits Conference (A-SSCC), 2012 IEEE Asian*. IEEE, 2012, pp. 73–76. 10, 14

[43] C.-Y. Lu, C.-T. Chuang, S.-J. Jou, M.-H. Tu, Y.-P. Wu, C.-P. Huang, P.-S. Kan, H.-S. Huang, K.-D. Lee, and Y.-S. Kao, "A 0.325 V, 600-kHz, 40-nm 72-kb 9T Subthreshold SRAM with Aligned Boosted Write Wordline and Negative Write Bitline Write-Assist," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 23, no. 5, pp. 958–962, 2015. 10, 14

[44] A. Wang and A. Chandrakasan, "A 180-mV Subthreshold FFT Processor using a Minimum Energy Design Methodology," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 1, pp. 310–319, 2005. 10, 11

[45] B. H. Calhoun and A. P. Chandrakasan, "A 256-kb 65-nm Sub-threshold SRAM Design for Ultra-Low-Voltage Operation," *IEEE Journal of Solid-State Circuits*, vol. 42, no. 3, pp. 680–688, 2007. 10, 12

[46] J. P. Kulkarni, K. Kim, and K. Roy, "A 160 mV Robust Schmitt Trigger Based Subthreshold SRAM," *IEEE Journal of Solid-State Circuits*, vol. 42, no. 10, pp. 2303–2313, 2007. 11, 12

[47] R. Saeidi, M. Sharifkhani, and K. Hajsadeghi, "Statistical Analysis of Read Static Noise Margin for Near/Sub-Threshold SRAM Cell," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 61, no. 12, pp. 3386–3393, 2014. 9, 32

[48] M. Alioto, "Ultra-Low Power VLSI Circuit Design Demystified and Explained: A Tutorial," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 59, no. 1, pp. 3–29, 2012. 11, 48

[49] B. 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*, Feb 2006, pp. 2592–2601. 12

[50] I. J. Chang, J. J. Kim, S. P. Park, and K. Roy, "A 32 kb 10T Sub-Threshold SRAM Array With Bit-Interleaving and Differential Read Scheme in 90 nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 2, pp. 650–658, Feb 2009. 13

[51] J. Myers, A. Savanth, R. Gaddh, D. Howard, P. Prabhat, and D. Flynn, "A Subthreshold ARM Cortex-M0+ Subsystem in 65 nm CMOS for WSN Applications with

14 Power Domains, 10T SRAM, and Integrated Voltage Regulator," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 1, pp. 31–44, 2016. 13

[52] J. Myers, A. Savanth, D. Howard, R. Gaddh, P. Prabhat, and D. Flynn, "An 80nW Retention 11.7 pJ/Cycle Active Subthreshold ARM Cortex-M0+ Subsystem in 65nm CMOS for WSN Applications," in *2015 IEEE International Solid-State Circuits Conference-(ISSCC) Digest of Technical Papers*. IEEE, 2015, pp. 1–3. 13

[53] E. Seevinck, F. J. List, and J. Lohstroh, "Static-Noise Margin Analysis of MOS SRAM Cells," *IEEE Journal of Solid-State Circuits*, vol. 22, no. 5, pp. 748–754, Oct 1987. 20, 21, 59

[54] A. J. Bhavnagarwala, S. Kosonocky, and J. D. Meindl, "Interconnect-Centric Array Architectures for Minimum SRAM Access Time," in *30th International Conference on Computer Design (ICCD)*. IEEE Computer Society, 2001, pp. 0400–0400. 24

[55] B. S. Amrutur and M. A. Horowitz, "Speed and Power Scaling of SRAM's," *IEEE Journal of Solid-State Circuits*, vol. 35, no. 2, pp. 175–185, 2000. 24, 29

[56] J. Rabaey, A. Chandrakasan, and B. Nikolic, *Digital Integrated Circuits: A Design Perspective*. NJ: Prentice Hall/Pearson Education, 2003. 24, 50

[57] S. Schuster, B. Chappell, R. Franch, P. Greier, S. Klepner, F. Lai, P. Cook, R. Lipa, R. Perry, W. Pokorny *et al.*, "A 15-ns CMOS 64K RAM," *IEEE journal of solid-state circuits*, vol. 21, no. 5, pp. 704–712, 1986. 29

[58] D. Liu and C. Svensson, "Trading Speed for Low Power by Choice of Supply and Threshold Voltages," *IEEE Journal of Solid-State Circuits*, vol. 28, no. 1, pp. 10–17, 1993. 31, 48

[59] S. Hanson, B. Zhai, M. Seok, B. Cline, K. Zhou, M. Singhal, M. Minuth, J. Olson, L. Nazhandali, T. Austin, D. Sylvester, and D. Blaauw, "Exploring Variability and Performance in a Sub-200-mV Processor," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 4, pp. 881–891, 2008. 32, 49

[60] T. Shakir and M. Sachdev, "A Word-line Boost Driver Design for Low Operating Voltage 6T-SRAMs," in *Circuits and Systems (MWSCAS), 2012 IEEE 55th International Midwest Symposium on*. IEEE, 2012, pp. 33–36. 32, 35

[61] Y. Pan, J. Kong, S. Ozdemir, G. Memik, and S. W. Chung, "Selective Wordline Voltage Boosting for Caches to Manage Yield Under Process Variations," in *Proceedings of the 46th Annual Design Automation Conference*. ACM, 2009, pp. 57–62. 32

[62] N.-C. Lien, L.-W. Chu, C.-H. Chen, H.-I. Yang, M.-H. Tu, P.-S. Kan, Y.-J. Hu, C.-T. Chuang, S.-J. Jou, and W. Hwang, "A 40-nm 512-kb Cross-point 8T pipeline SRAM with Binary Word-Line Boosting Control, Ripple Bit-line and Adaptive Data-Aware Write-Assist," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 61, no. 12, pp. 3416–3425, 2014. 32

[63] B. Liu, H. R. Pourshaghaghi, S. M. Londono, and J. P. de Gyvez, "Process Variation Reduction for CMOS Logic Operating at Sub-threshold Supply Voltage," in *Digital System Design (DSD), 2011 14th Euromicro Conference on*. IEEE, 2011, pp. 135–139. 37

[64] M. Nabavi, F. Ramezankhani, and M. Shams, "Optimum pMOS-to-nMOS Width Ratio for Efficient Subthreshold CMOS Circuits," *IEEE Transactions on Electron Devices*, vol. 63, no. 3, pp. 916–924, 2016. 37, 38, 54

[65] S. M. Sze and K. K. Ng, *Physics of Semiconductor Devices*. John wiley & sons, 2006. 38

[66] M. Nabavi and M. Sachdev, "A 290-mV, 3.34-MHz, 6T SRAM With pMOS Access Transistors and Boosted Wordline in 65-nm CMOS Technology," *IEEE Journal of Solid-State Circuits*, vol. PP, no. 99, pp. 1–12, 2017. 48, 49, 77

[67] Y.-H. Chen, W.-M. Chan, W.-C. Wu, H.-J. Liao, K.-H. Pan, J.-J. Liaw, T.-H. Chung, Q. Li, C.-Y. Lin, M.-C. Chiang *et al.*, "A 16 nm 128 Mb SRAM in High-k metal-gate FinFET Technology with Write-Assist Circuitry for Low-VMIN Applications," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 1, pp. 170–177, 2015. 48

[68] S. Nalam and B. H. Calhoun, "5T SRAM with Asymmetric Sizing for Improved Read Stability," *IEEE journal of solid-state circuits*, vol. 46, no. 10, pp. 2431–2442, 2011. 49

[69] M. Sharifkhani and M. Sachdev, "Segmented Virtual Ground Architecture for Low-Power Embedded SRAM," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 15, no. 2, pp. 196–205, 2007. 49

[70] H. Okamura, H. Toyoshima, K. Takeda, T. Oguri, S. Nakamura, M. Takada, K. Imai, Y. Kinoshita, H. Yoshida, and T. Yamazaki, "A 1 ns, 1 W, 2.5 V, 32 Kb NTL-CMOS SRAM Macro using a Memory Cell with PMOS Access Transistors," *IEEE Journal of Solid-State Circuits*, vol. 30, no. 11, pp. 1196–1202, 1995. 49

[71] Y. Tsividis and C. McAndrew, *Operation and Modeling of the MOS Transistor*. Oxford Univ. Press, 2011. 54

[72] Y. Cheng and C. Hu, *MOSFET Modeling & BSIM3 User's Guide*. Springer Science & Business Media, 1999. 54

APPENDICES

Appendix A

Publications

- **Morteza Nabavi** and M. Sachdev, ”A 290-mV, 3.34-MHz, 6T SRAM With pMOS Access Transistors and Boosted WL in 65 nm CMOS Technology,” in IEEE Journal of Solid-State Circuits, vol. PP, no. 99, pp. 1-12, August 2017.