

國立交通大學

電子工程學系 電子研究所碩士班

碩士論文

具有寫入輔助電路的穩健低功率靜態隨機存取記憶體



A Robust Low Power SRAM Design with Write Assist

Circuits

研究生：賴思詠

指導教授：黃 威 教授

中華民國九十七年八月

具有寫入輔助電路的穩健低功率靜態隨機存取記憶體
電路設計

A Robust Low Power SRAM Design with Write Assist
Circuits

研究生：賴思詠

Student : Ssu-Yun Lai

指導教授：黃 威 教授

Advisor : Prof. Wei Hwang



A Thesis

Submitted to Department of Electronics Engineering & Institute of Electronics
College of Electrical Engineering and Computer Engineering

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

in

Electronics Engineering

August 2008

Hsinchu, Taiwan, Republic of China

中華民國九十七年八月

具有寫入輔助電路的穩健低功率靜態隨機存取記憶體 電路設計

學生：賴思詠

指導教授：黃 威 教授

國立交通大學電子工程學系電子研究所

摘 要

本論文提出一個具有寫入輔助電路的靜態隨機存取記憶體。此寫入輔助電路可用來解決而嚴重的 Write Half-Select 的寫入問題，而且經過模擬的結果指出，此類型的寫入輔助電路在 65nm 和 45nm 之類的先進製程下，仍然可以維持作用良好的情況。除此之外，更額外設計了讀取與寫入的複製電路用來控制精準的時序。根據模擬結果指出，這個靜態隨機存取記憶體大量的降低了功率的消耗便且擁有非常大的電壓操作範圍，它可在電壓為 1V 時，操作頻率高達 1GHz，而電壓為 0.5V 時，操作頻率則可達 200MHz，所消耗的功率分別為 9mW 與 826uW，極有利於行動裝置的使用。

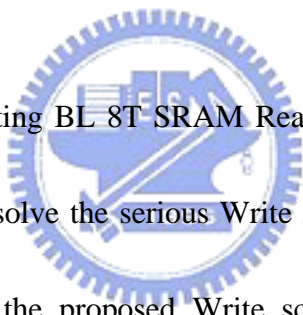
A Robust Low Power SRAM Design with Write Assist Circuits

Student : Ssu-Yun Lai

Advisors : Prof. Wei Hwang

Department of Electronics Engineering & Institute of Electronics
National Chiao-Tung University

ABSTRACT



This paper presents a floating BL 8T SRAM Read/Write scheme. A Write assist scheme is also proposed to resolve the serious Write half-select disturb problem, and simulation results show that the proposed Write scheme can work well in more advanced technology nodes, such as 65nm and 45nm. Furthermore, Read/Write replica circuits are designed to control access timing. Moreover, a 32-Kb 8T SRAM subarray is implemented in UMC 90nm CMOS technology. According to simulation results, the proposed 8T SRAM shows its benefits on low power access operations and wide-operating voltage range. It can operate at 1GHz when V_{DD} is 1V and at 200MHz when V_{DD} is 0.5V. So it is suitable to be adopted in portable devices.

致謝

感謝許多人的幫助，讓我完成了這一篇論文。

首先，我要感謝我的指導教授黃威，在他的指導下讓我對自己研究的領域有更深入的了解，建立了研究的興趣與自信心。黃教授提供了一個常優良的研究環境與充足的研究資源，讓我能夠充分發揮自己的能力完成這一篇論文。

感謝實驗室的成員，在過去的生活上與研究上對我的幫助。感謝黃柏蒼、張銘宏和楊皓義三位學長對於我在研究上的幫忙，特別是楊皓義學長，讓我能夠有更加優良的研究成果。

最後我要感謝我的家人對我在生活上的關心與幫助，讓我能夠順利的完成碩士的論文研究。



Content

Chapter 1 Introduction.....	1
1.1 Background.....	1
1.2 Motivation.....	2
1.3 Thesis Organization	2
Chapter 2 Overview of Low Power CMOS Circuit and Staitc Random Access Memory	3
2.1 Introduction.....	3
2.2 Power Dissipation	3
2.2.1 Dynamic Dissipation.....	3
2.2.2 Leakage Dissipation.....	5
2.2.3 Short Circuit Dissipation.....	9
2.2.4 Putting It All Together.....	10
2.3 Low Power Circuit Techniques.....	10
2.3.1 Supply Voltage Scaling	10
2.3.2 Transistor Stacking.....	14
2.3.3 Multiple Threshold Designs.....	22
2.4 Overview of SRAM Operation.....	28
2.4.1 6T SRAM Cell	29
2.4.2 SRAM Cell Stability.....	31
2.4.2.1 Hold Stability	31
2.4.2.2 Read Stability.....	32
2.4.2.3 Write Stability	33
2.4.3 Column Circuitry	34
2.5 Single Read Bitline 8T SRAM Cell.....	35
2.6 Summary	36
Chapter 3 Low Power Write Assist Scheme.....	38
3.1 Introduction.....	38
3.2 Low Power Write Assist Scheme.....	39
3.3 Summary	43
Chapter 4 Low Voltage Replica Wordline Timing Controller	44
4.1 Introduction.....	44
4.2 Clock Matching.....	44
4.3 Low Voltage Replica Wordline Timing Controller	49
4.4 Summary	50
Chapter 5 A Robust Low Power SRAM Design with Write Assist Circuits.....	51

5.1 Introduction.....	51
5.2 System Architecture	52
5.3 Circuit Description.....	54
5.3.1 Cell Array.....	54
5.3.2 Read/Write Wordline Replica Circuitry	56
5.3.3 Wordline Driver	56
5.4 Design Implementation.....	58
5.5 Simulation Result.....	58
Chapter 6 Conclusions.....	61
6.1 Conclusions.....	61
Bibliography	62



List of Figures

1.1	Power density versus gate length.....	1
2.1	A CMOS inverter	4
2.2	Leakage current components in an NMOS transistor	5
2.3	Components of tunneling current.....	7
2.4	Gate leakage current versus gate oxide thickness.....	7
2.5	Gate leakage current versus gate voltage.....	8
2.6	A CMOS inverter chain	11
2.7	Power versus supply voltage.....	12
2.8	Time delay versus supply voltage	12
2.9	PDP versus supply voltage.....	13
2.10	Noise margin versus supply voltage	13
2.11	Two-input NAND gate stacking effect illustration	15
2.12	NMOS footer array power gating devices	16
2.13	PMOS header array power gating devices.....	16
2.14	Inverter chain with footer power gating.....	17
2.15	Inverter chain with header power gating	17
2.16	Standby power comparisons when applying footer power gating	18
2.17	Standby power comparisons when applying header power gating.....	18
2.18	Time delay comparisons when applying footer power gating	19
2.19	Time delay comparisons when applying header power gating	19
2.20	Active power comparisons when applying footer power gating	20
2.21	Active power comparisons when applying header power gating	20
2.22	Time delay comparisons between footer and header. One footer/header is applied on four inverters	21
2.23	Standby power comparisons between footer and header. One footer/header is applied on four inverters	21
2.24	Dual threshold CMOS circuit	23
2.25	MVT CMOS scheme	23
2.26	Footer insertion MTCMOS circuit.....	24
2.27	Header insertion MTCMOS circuit	24
2.28	MTCMOS inverter chain with footer power gating	24
2.29	MTCMOS inverter chain with header power gating	25
2.30	Standby power comparisons when applying footer insertion MTCMOS circuit.....	25
2.31	Standby power comparisons when applying header insertion MTCMOS circuit	26

2.32	Time delay comparisons when applying footer insertion MTCMOS circuit	26
2.33	Time delay comparisons when applying header insertion MTCMOS circuit	27
2.34	Active power comparisons when applying footer insertion MTCMOS circuit	27
2.35	Active power comparisons when applying header insertion MTCMOS circuit	28
2.36	SRAM organization	29
2.37	Conventional 6T SRAM Cell.....	30
2.38	Read example of 6T SRAM.....	30
2.39	Write example of 6T SRAM.....	30
2.40	Standard setup for finding the Hold SNM	31
2.41	Butterfly curve	32
2.42	Standard setup for finding Read SNM.....	32
2.43	Example butterfly curve plots for hold SNM and Read SNM.....	33
2.44	Setup for finding WTP.....	34
2.45	Write margin of a SRAM cell, determined by WTP.....	34
2.46	An SRAM column	35
2.47	Single read bitline (SRBL) 8T cell	36
3.1	(a)Conventional WWL and WBL pair signals;(b)Proposed WWL and WBL pair signals	38
3.2	Monte-Carlo simulation results (32,000 nodes) for noise margin distributions with different initial WBL voltage levels. (Rising edge = 160ps, BL = 32-bit, $V_{DD} = 1.0v$).	39
3.3	Write selected bit and Write half-select disturb.	40
3.4	Monte-Carlo simulation results (32,000 nodes) for noise margin distributions with different rising edge. (Initial WBL voltage = 0V, BL = 32-bit, $V_{DD} = 1.0v$).	41
3.5	Monte-Carlo simulation results (32,000 nodes) for noise margin distributions with different BL length. (Initial WBL voltage = 0V, rising edges = 160ps, $V_{DD} = 1.0v$).	42
3.6	Monte-Carlo simulation results (32,000 nodes) for noise margin distributions with different rising edge in different technology nodes: (a) PTM 65nm; (b) PTM 45nm. (Initial WBL voltage = 0V, BL = 32-bit, $V_{DD} = 1.0v$)	43
4.1	Common sense clock generation techniques.	46

4.2	Delay matching between the bitline delay to generate 120mV and two delay elements, one based on an inverter chain and the other on a replica cell-bitline combination..	48
4.3	A 2x2 memory array with replica cell and circuits	49
5.1	Block diagram of the proposed SRAM.	53
5.2	1024x32 SRAM symbol.	53
5.3	The proposed 8T SRAM array	55
5.4	Read/write replica circuits	56
5.5	The wordline driver (WLD) structure	57
5.6	Layout view of the proposed SRAM	58



List of Tables

5.1	Signal description.....	54
5.2	Command truth table.....	54
5.3	Summary of the SRAM features.....	59
5.4	Process corner simulation (@500mV ; 25°C).....	60
5.5	Voltage variation simulation (@TT corner ; 25°C).....	60
5.6	Temperature variation simulation (@TT corner ; 500mV).....	60



Chapter 1

Introduction

1.1 Background

Device miniaturization and the rapidly growing demand for mobile or power-aware systems have resulted in the urgent need for low power circuit design [1]. In modern CMOS technology, active power (dynamic power) and passive power (leakage power) are equally significant. This trend is shown in Figure 1.1 [2]. Therefore, to achieve low power operation, both active and passive power needs to be considered seriously.

In emerging system on chip (SoC) designs, an indispensable component is the on-chip memory module. As device density increases, a larger fraction of chip area is devoted to the memory block to enable more complex functionality and higher performance[3][4]. As a result, power of memory blocks often dominates the total power consumption. Memory power consumption has thus been a major challenge and design consideration in future SoC.

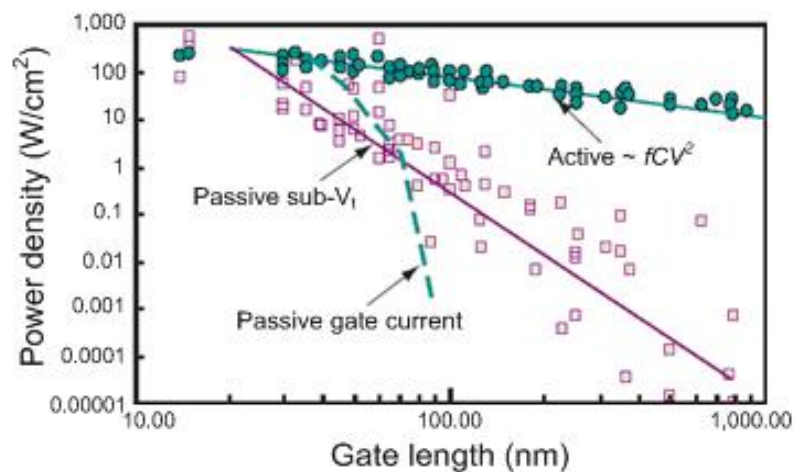


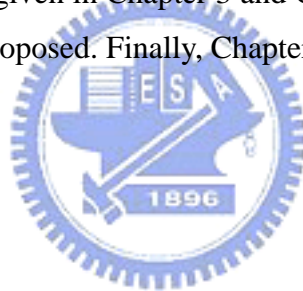
Figure 1.1: Power density versus gate length.

1.2 Motivation

As system integration growing, more SRAM are integrated on a chip or in a package. According to ITRS2005 predictions, SRAM will occupy over 90% area of a chip in the future 10 years. Low power SRAM design has become critical. Furthermore, In order to extend portable devices' working time, operating frequencies and voltages of these devices should be adjustable according to applications' demands. Moreover, devices' standby voltage can be set at a much lower level to reduce leakages. Consequently, wide-operating-voltage range SRAM is needed.

1.3 Thesis Organization

The rest of this thesis is organized as following. Chapter 2 reviews CMOS circuit power sources and gives possible solutions to reduce power consumption. Low power SRAM design techniques are given in Chapter 3 and Chapter 4. In Chapter 5, a robust low power SRAM design is proposed. Finally, Chapter 6 concludes this work.



Chapter 2

Overview of Low Power CMOS Circuit and Static Random Access Memory

2.1 Introduction

This chapter begins with a study of power dissipation of CMOS circuit and circuit technique for power reduction. Power dissipation, including *dynamic dissipation*, *leakage dissipation*, and *short circuit dissipation*, is presented in Section 2.2. Low power circuit techniques, including *supply voltage scaling*, *transistor stacking*, *multiple threshold voltage design*, is presented in Section 2.3. Summary of this chapter is presented in Section 2.4.

2.2 Power Dissipation

2.2.1 Dynamic Dissipation

For a CMOS inverter, shown in Figure 2.1, the average dynamic power dissipation can be obtained by summing the average dynamic power in the NMOS transistor and the PMOS transistor. Assuming that the input V_{in} is a square wave having a period T and that the rise and fall times of the input are much less than the repetition period, the dynamic power is given by

$$P_D = \frac{1}{T} \int_0^{T/2} i_N(t) V_{out} dt + \frac{1}{T} \int_{T/2}^T i_P(t) (V_{DD} - V_{out}) dt \quad (2.1)$$

Since $i_N(t) = C_L \frac{dV_{out}}{dt}$ and $i_P(t) = C_L \frac{d(V_{DD} - V_{out})}{dt}$,

$$P_D = \frac{C_L}{T} \int_0^{V_{DD}} V_{out} dV_{out} + \frac{C_L}{T} \int_{V_{DD}}^0 (V_{DD} - V_{out}) d(V_{DD} - V_{out}) = \frac{C_L V_{DD}^2}{T} \quad (2.2)$$

Where C_L is the load capacitance, $\frac{1}{T} = f$, f is the operating frequency. Therefore

$$P_D = f C_L V_{DD}^2 \quad (2.3)$$

Moreover, power dissipation is data dependent, i.e. power dissipation depends on the switching probability α , thus, dynamic power can be expressed as

$$P_D = \alpha f C_L V_{DD}^2 \quad (2.4)$$

By (2.4), dynamic power dissipation of CMOS logic gate is proportional to switching frequency, load, capacitance, square of the supply voltage, and operation frequency.

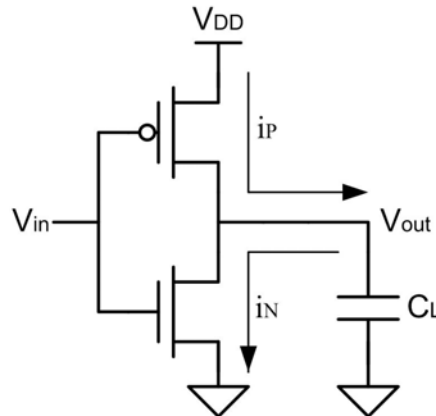


Figure 2.1: A CMOS inverter.

2.2.2 Leakage Dissipation

There are four main sources of leakage current in a CMOS transistor as illustrated in Figure 2.2 [14][15][16]. They are *reverse-biased junction leakage current* (I_{REV}), *gate induced drain leakage* (I_{GIDL}), *gate direct-tunneling leakage* (I_G), and *subthreshold leakage* (I_{SUB}). Each source of leakage current will be further described in the followings.

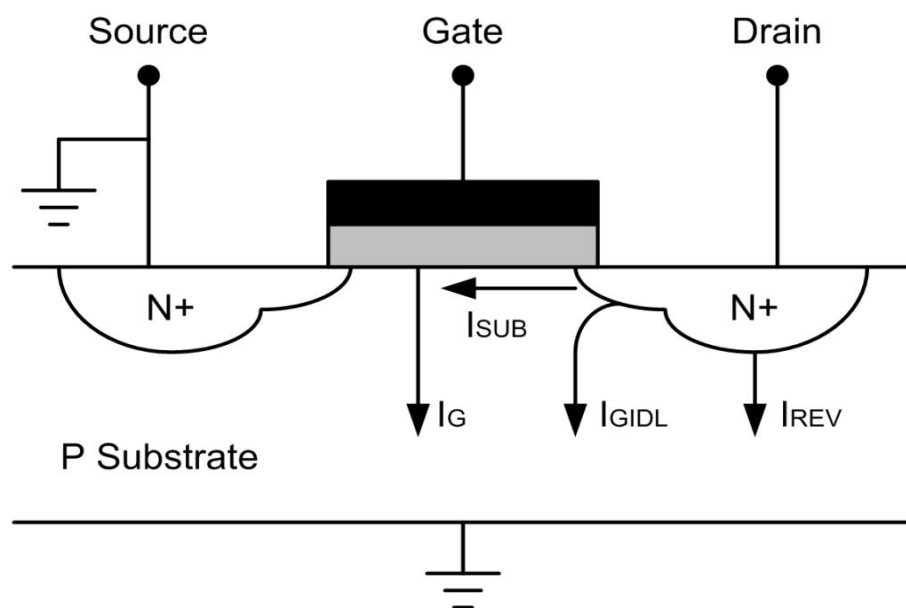


Figure 2.2: Leakage current components in an NMOS transistor.

Junction Leakage

The junction leakage occurs from the source/drain to the substrate through the reverse-biased diodes when the transistor is off, indicated as I_{REV} in Figure 2.2. A reverse-biased pn junction leakage has two major components: one is minority carrier diffusion/drift near the edge of the depletion region; the other is due to electron-hole pair generation in the depletion region of the reverse-biased junction.

Junction leakage current depends on the area of the drain diffusion and the leakage current density, which is in turn determined by the doping concentration. Junction leakage components from both the source-drain diodes and the well diodes

are generally negligible with respect to the other three leakage components.

Gate-Induced Drain Leakage

Gate-induced drain leakage (GIDL), indicated as I_{GIDL} in Figure 2.2, arises in the high electric field under the gate/drain overlap region. GIDL occurs at large V_{DB} and generates carriers into the substrate and drain from surface traps or band-to-band tunneling. Thinner oxide, higher supply voltage, and lightly doped drain structures increase GIDL current.

Gate Direct Tunneling Leakage

Gate direct tunneling current is due to the tunneling of an electron/hole from the bulk silicon through the gate oxide potential barrier into the gate [17][18]. Reduction of gate oxide thickness results in the increase in the field across the oxide. The high electric field coupled with low oxide thickness results in tunneling of electrons from substrate to gate and also from gate to substrate through the gate oxide, resulting in the gate leakage. In nanometer-scale CMOS technologies, where ultra-thin gate oxide thickness takes place for effective gate control, gate leakage becomes appreciable and dominates the total leakage dissipation [19].

Figure 2.3 shows the components of tunneling current in a scaled NMOS transistor.

They are classified in to three categories:

1. Edge direct tunneling (EDT) components between the gate and the source-drain extension (SDE) overlap region (I_{gso} and I_{gdo}).
2. Gate-to-channel current (I_{gc}), part of which goes to the source (I_{gcs}), and the rest goes to the drain (I_{gcd}).
3. Gate-to-substrate leakage current (I_{gb}).

Therefore, the gate leakage (I_G) can be divided into three major components:

1. Gate-to-source ($I_{gs} = I_{gso} + I_{gcs}$).
2. Gate-to-drain ($I_{gd} = I_{gdo} + I_{gcd}$).
3. Gate-to-substrate (I_{gb}).

The magnitude of the gate leakage current increases exponentially with the gate oxide thickness T_{OX} and the gate-to-source voltage V_{GS} , as shown in Figure 2.4 and Figure 2.5, respectively [20].

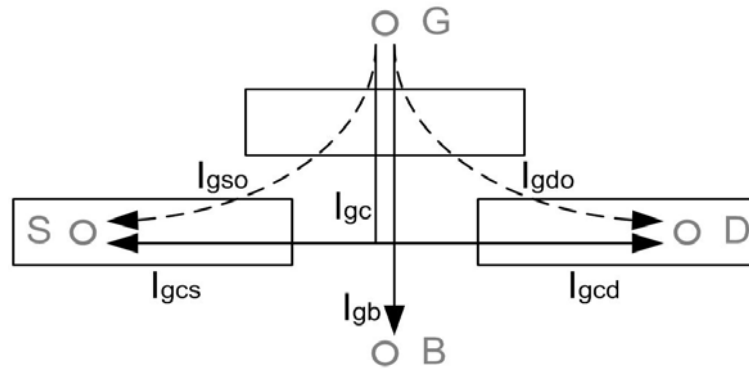


Figure 2.3: Components of tunneling current.

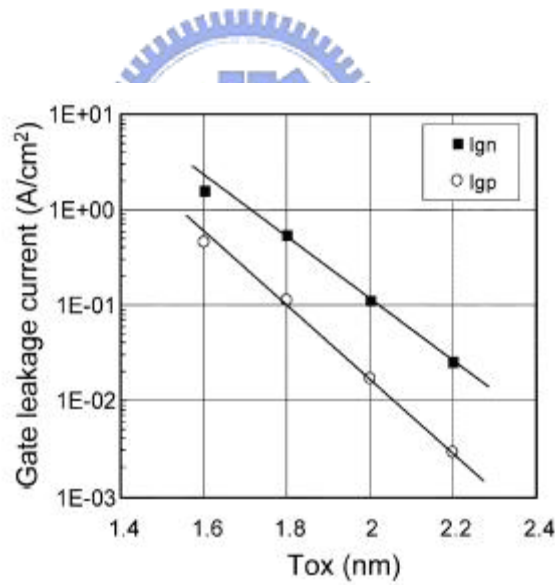


Figure 2.4: Gate leakage current versus gate oxide thickness.

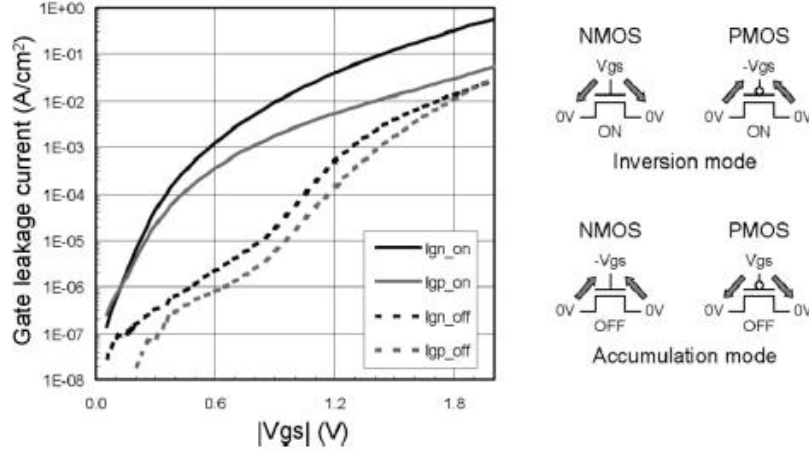


Figure 2.5: Gate Leakage current versus gate voltage.

Subthreshold Leakage

Subthreshold or weak inversion conduction current between source and drain of an MOS transistor occurs when gate voltage is below the threshold voltage level. Unlike the strong inversion region in which the drift current dominates, the subthreshold conduction is due to the diffusion current of the minority carriers in the channel for a MOS device. For instance, in an inverter with a low input voltage and high output voltage, for the NMOS transistor, even V_{GS} is 0V, there is still a current flowing in the channel of the off NMOS transistor due to the V_{DD} potential of the V_{DS} .

Subthreshold leakage current (I_{SUB}) becomes apparent as CMOS technologies enter the submicron era [21]. I_{SUB} can be expressed based on the following:

$$I_{SUB} = \frac{W}{L} \mu v_{th}^2 C_{sth} e^{\frac{V_{GS} - V_T + \eta V_{DS}}{nV_{th}}} \left(1 - e^{-\frac{V_{DS}}{V_{th}}}\right) \quad (2.5)$$

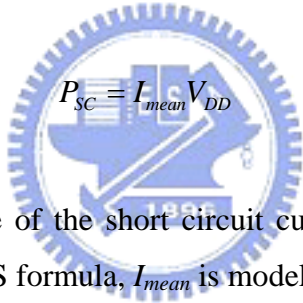
where W and L denote the transistor width and length, μ denotes the carrier mobility, $V_{th} = kT/q$ denotes the thermal voltage at temperature T , $C_{sth} = C_{dep} + C_{it}$ denotes the summation of the depletion region capacitance and the interface trap capacitance both per unit area of the MOS gate, and η is the drain-induced barrier lowering (DIBL) coefficient. n is the slope shape factor and is calculated as:

$$n = 1 + \frac{C_{sth}}{C_{ox}} \quad (2.6)$$

where C_{ox} denotes the gate input capacitance per unit area of the MOS gate. Thus, the magnitude of the subthreshold leakage current is a function of the temperature, supply voltage, device size, and the process parameters out of which the threshold voltage plays a dominant role.

2.2.3 Short Circuit Dissipation

The short circuit power dissipation results due to a direct path current flowing from the power supply to the ground during the switching of a static CMOS gate. Short circuit dissipation can be expressed as:

$$P_{SC} = I_{mean} V_{DD} \quad (2.7)$$


where I_{mean} is the mean value of the short circuit current. Assuming a symmetrical inverter and using simple MOS formula, I_{mean} is modeled as [22]:

$$I_{mean} = \frac{1}{12} \frac{\beta}{V_{DD}} (V_{DD} - 2V_T)^3 \frac{\tau}{T} \quad (2.8)$$

where β is the gain factor of a MOS transistor, τ is the input rise/fall time.

From (2.7) and (2.8), short circuit dissipation of a CMOS inverter without load is derived as:

$$P_{SC} = \frac{\beta}{12} (V_{DD} - 2V_T)^3 \frac{\tau}{T} \quad (2.9)$$

Although this is a simplified model, it reveals the fact that short circuit dissipation is affected by supply voltage, threshold voltage, rise/fall time, and operation frequency.

Therefore, it is effective to minimize short-circuit power by lowering supply voltage, increasing threshold voltage, and minimizing input rise/fall time.

2.2.4 Putting It All Together

The total power consumption of a digital CMOS circuit can be expressed as the sum of its three components:

$$P_{Total} = P_D + P_{Leak} + P_{SC} = \alpha f C_L V_{DD}^2 + I_{Leak} V_{DD} + I_{SC} V_{DD} \quad (2.10)$$

Clearly, supply voltage has a major dominance over power consumption. In the next section, several circuit techniques for power control and reduction are presented, including *supply voltage scaling*, *transistor stacking*, and *multiple threshold voltage design*. Both Active and standby power reduction are considered.

2.3 Low Power Circuit Techniques



2.3.1 Supply Voltage Scaling

In a given technology, supply voltage reduction is the key to low power operation [23][24]. When lowering the supply voltage, there are two issues that must be considered:

1. Impact on delay: Since both capacitance and threshold voltage are constant, the speed of the basic gates will also decrease with the voltage scaling, where the relation between time delay T_d and supply voltage V_{DD} can be modeled by using a quadratic model:

$$T_d = k \frac{C_L V_{DD}}{(V_{DD} - V_T)^2} \quad (2.11)$$

2. Impact on stability: Low supply voltage circuits are very sensitive to both manufacturing variations and operating point changes, which leads to less stable and less robust operation.

Following is an example of supply voltage scaling. Figure 2.6 shows an inverter chain composed of four inverters. Figure 2.7 shows the relation between power and supply voltage; Figure 2.8 shows the relation between time delay and supply voltage. It is revealed that as supply voltage drops, power consumption is reduced, but the time delay is increased. A common vector for finding the optimal supply voltage is the *power delay product (PDP)*, which is the product of power and time delay, as shown in Figure 2.9. Another strategy is to find the worst case critical time delay and choose the minimum supply voltage that is capable of performing the expected operation speed.

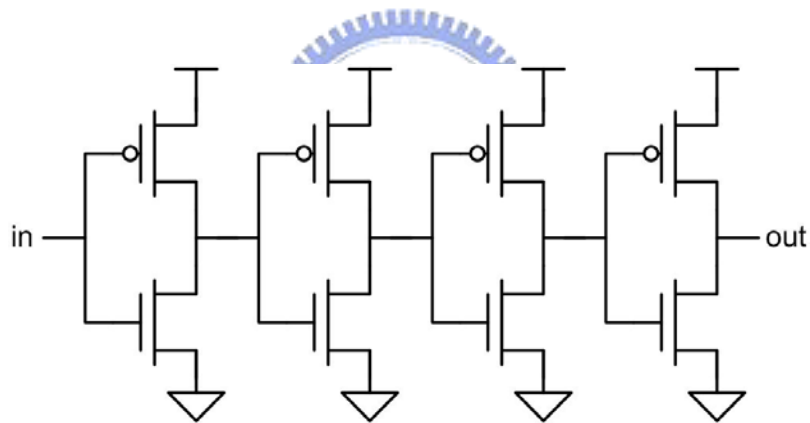


Figure 2.6: A CMOS inverter chain.

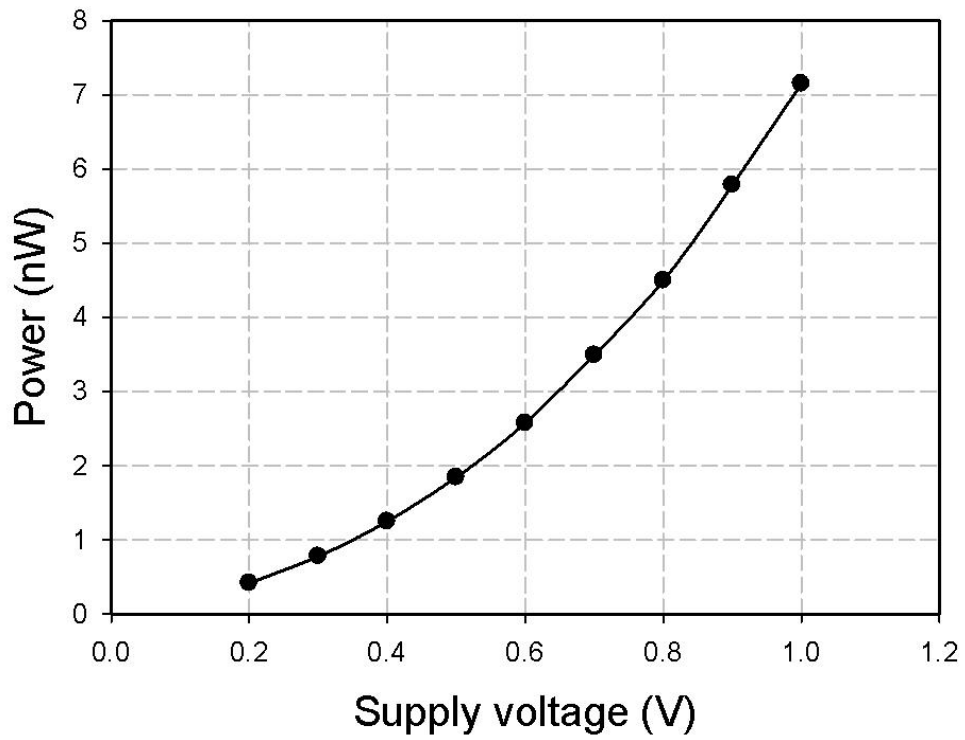


Figure 2.7: Power versus supply voltage.

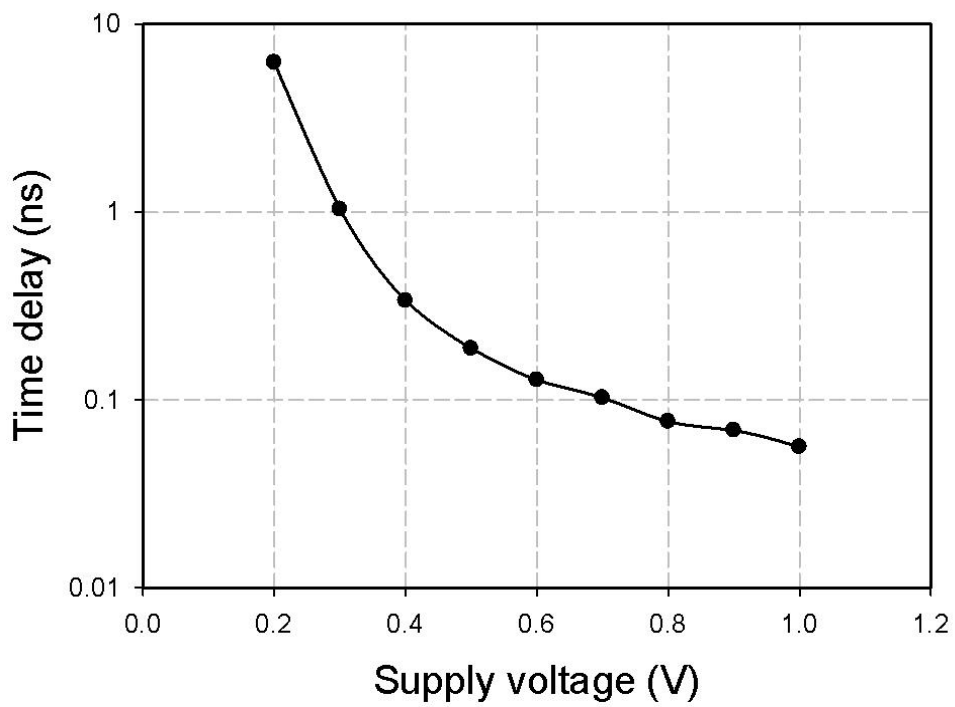


Figure 2.8: Time delay versus supply voltage.

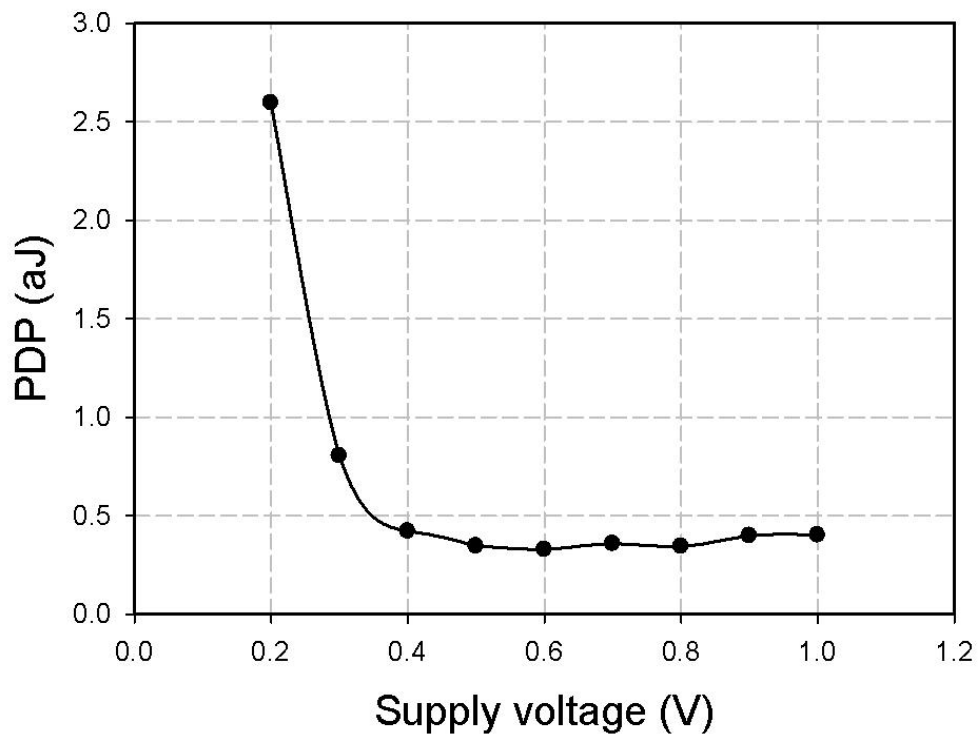


Figure 2.9:PDP versus supply voltage.

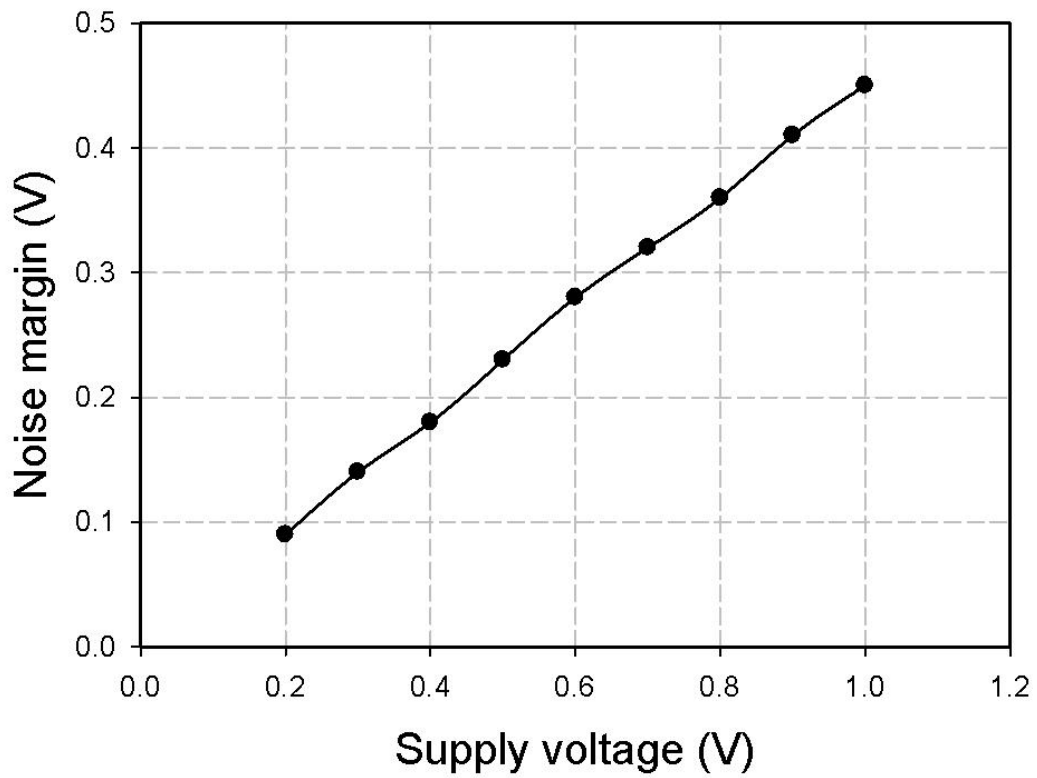


Figure 2.10: Noise margin versus supply voltage.

Relation between noise margin [25] and supply voltage is shown in Figure 2.10. As shown, noise margin decreases as supply voltage drops. Noise margin issue is especially important in low voltage and subthreshold circuit designs [26].

2.3.2 Transistor Stacking

Transistor stacking is an effective technique to reduce subthreshold and gate leakage current[27][28]. Leakage current flowing through a stack of series-connected transistors reduces if more than one transistor in the stack is off, which is known as the stacking effect. The stacking effect can be understood by considering a two-input NAND gate, as shown in Figure 2.11. When both MN1 and MN2 are off, the voltage at the intermediate node (V_M) raises to a positive value due to a small drain current. Positive potential at the intermediate node leads to three effects:

1. Gate-to-source voltage of MN1 becomes negative.
2. Negative body-to-source potential of MN1 causes more body effect. The body effect describes how the potential difference between source and body affects the threshold voltage, which can be modeled as:

$$V_T = V_{T0} + \gamma(\sqrt{\phi_S + V_{SB}} - \sqrt{\phi_S}) \quad (2.12)$$

3. Drain-to-source potential of MN1 decreases, resulting in less drain-induced barrier lowering.

As a result, negative gate-to-source voltage, higher threshold voltage due to the body effect, and less drain-induced barrier lowering due to the reduction of drain-to-source voltage, leakage current is reduced.

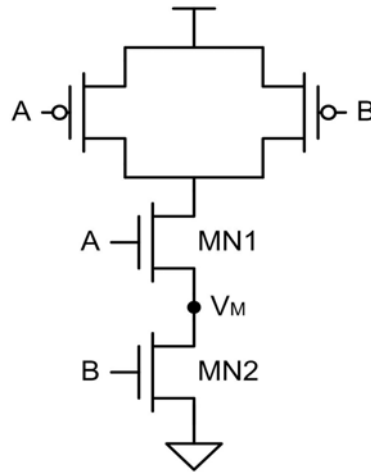


Figure 2.11: Two-input NAND gate stacking effect illustration.

Transistor stacking for low power can be referred to power gating. Power gating devices can be classified into two main categories: footer and header devices. Footer is by inserting NMOS sleep transistors between real GND and virtual GND, while header is by inserting PMOS sleep transistors between real V_{DD} and virtual V_{DD} , as shown in Figure 2.12 and Figure 2.13, respectively.

Figure 2.14 and Figure 2.15 are testing examples of footer and header. The effectiveness of standby power saving by footer and header are shown in Figure 2.16 and Figure 2.17. Time delay comparisons are shown in Figure 2.18 and Figure 2.19. As shown, by sacrificing operation speed, a circuit with power gating devices has significant standby power (leakage power) reduction. Trade off between power and speed is also illustrated. For a circuit with power gating, the less power gating are inserted, the more power is saved, and the more power gating are inserted, the less time delay it performs. Adding power gating devices usually contributes very slight active power overhead, which is revealed in Figure 2.20 and Figure 2.21. Another interesting thing worth notice is the comparison between footer and header, which is demonstrated in Figure 2.22 and Figure 2.23. NMOS has stronger driving ability than PMOS, resulting in smaller time delay when applying footer power gating. On the other hand, as shown in Figure 2.5, PMOS has smaller leakage current than NMOS, resulting in smaller power consumption when applying header power gating.

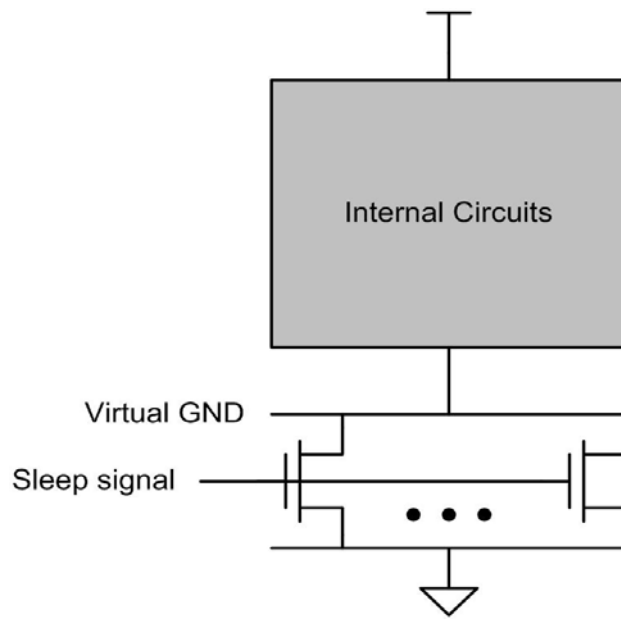


Figure 2.12: NMOS footer array power gating devices.

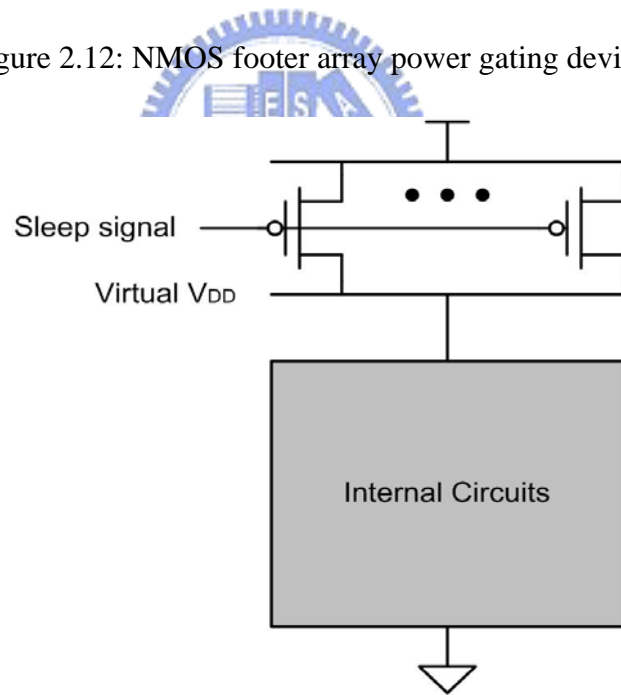


Figure 2.13: PMOS header array power gating devices.

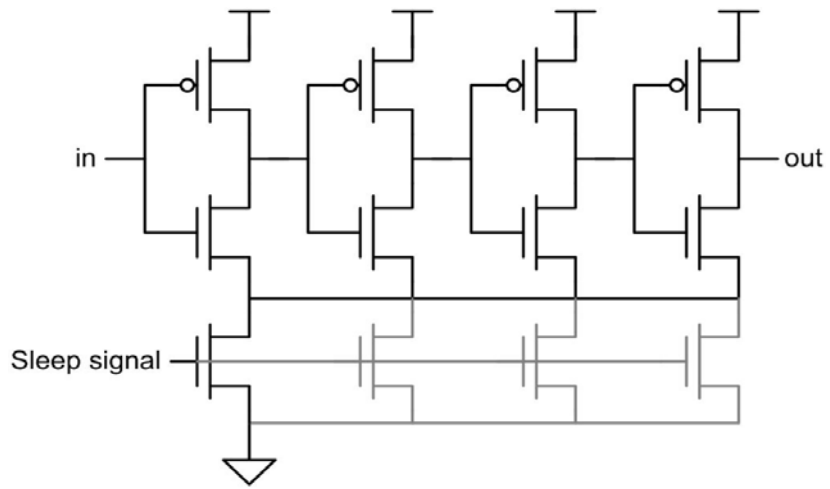


Figure 2.14: Inverter chain with footer power gating.

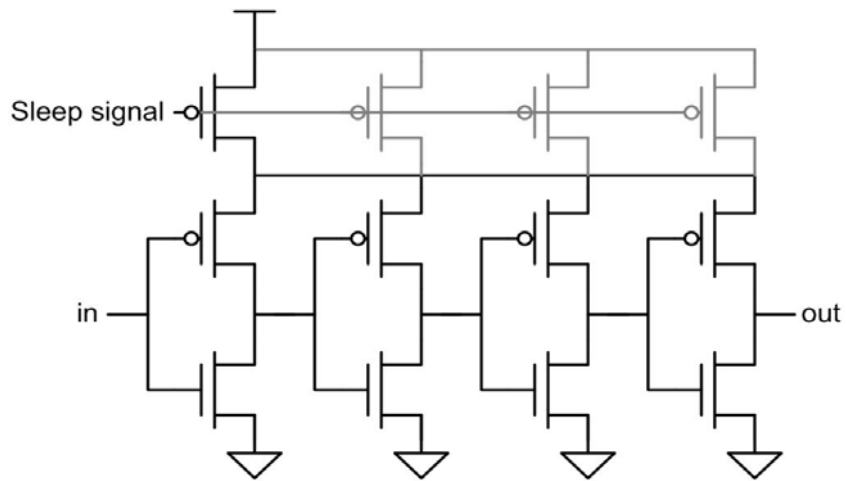


Figure 2.15: Inverter chain with header power gating.

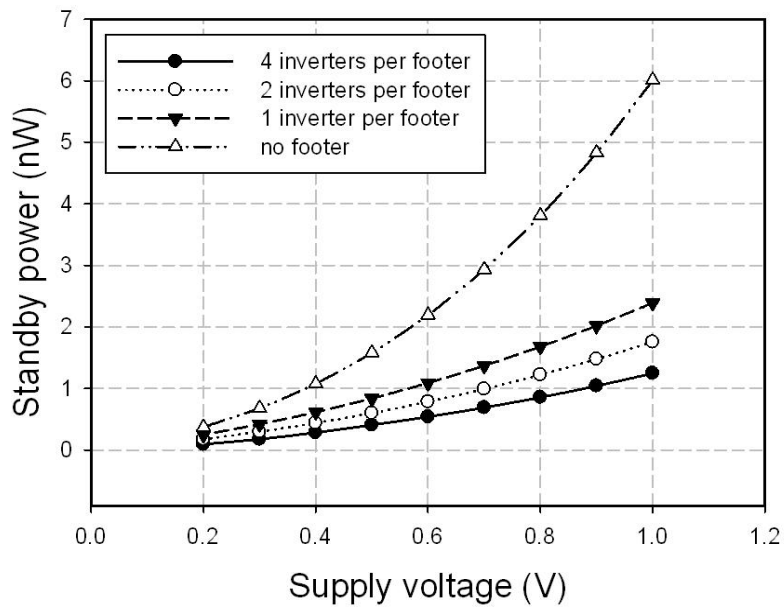


Figure 2.16: Standby power comparisons when applying footer power gating.

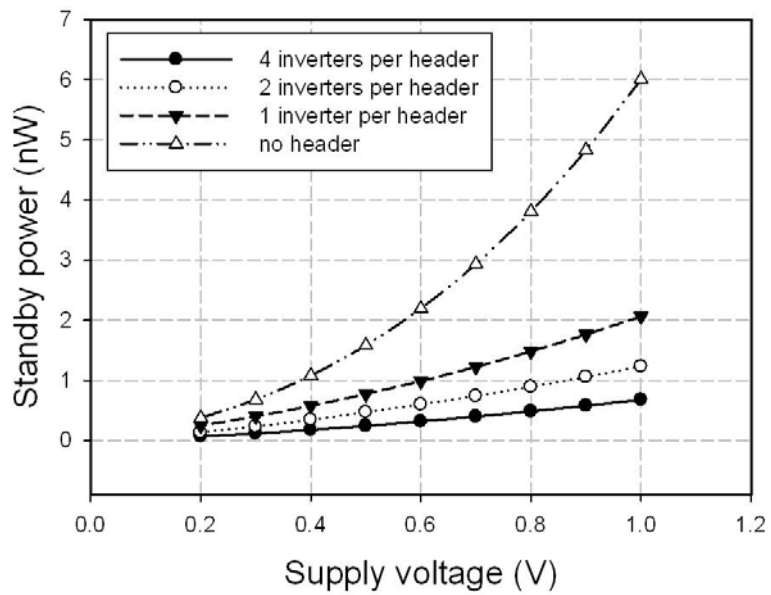


Figure 2.17: Standby power comparisons when applying header power gating.

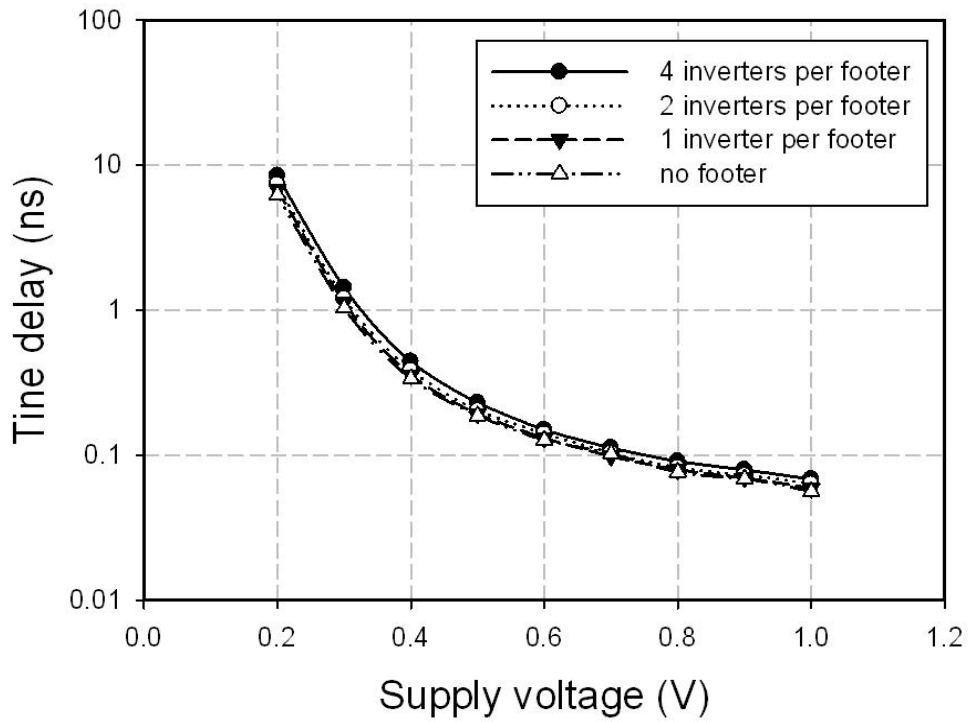


Figure 2.18: Time delay comparisons when applying footer power gating.

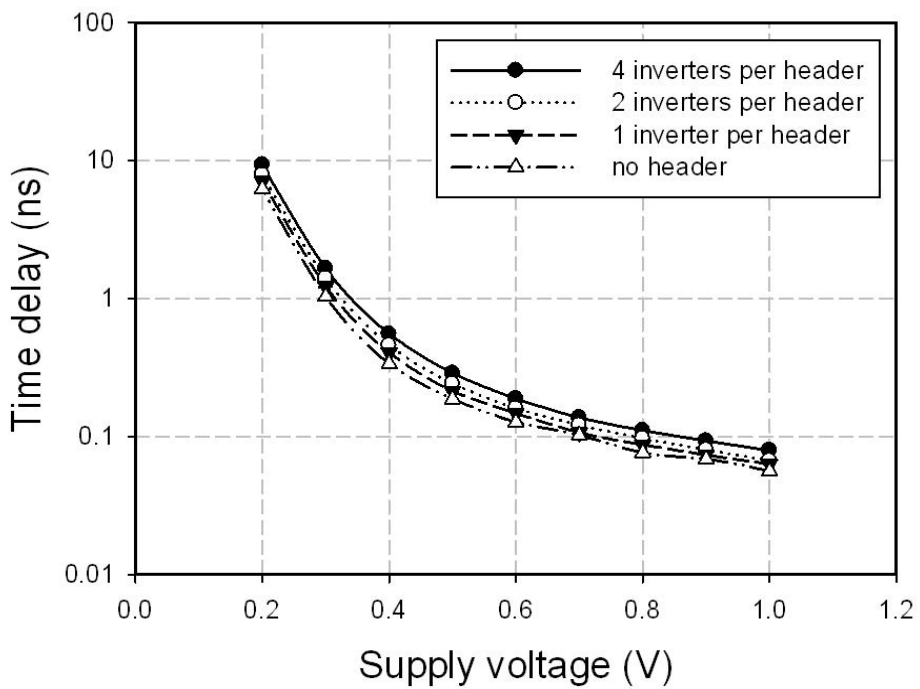


Figure 2.19: Time delay comparisons when applying header power gating.

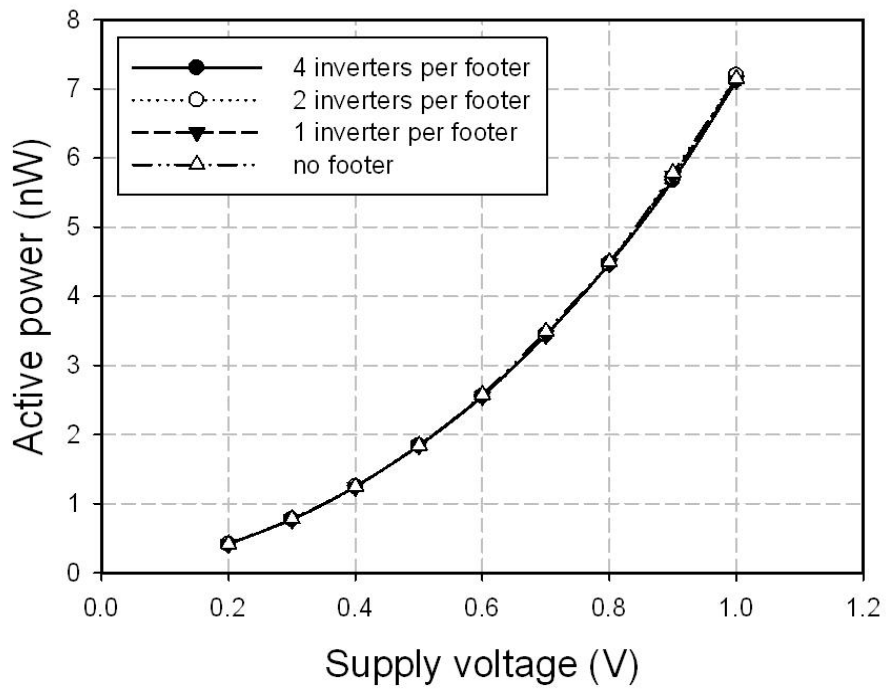


Figure 2.20: Active power comparisons when applying footer power gating.

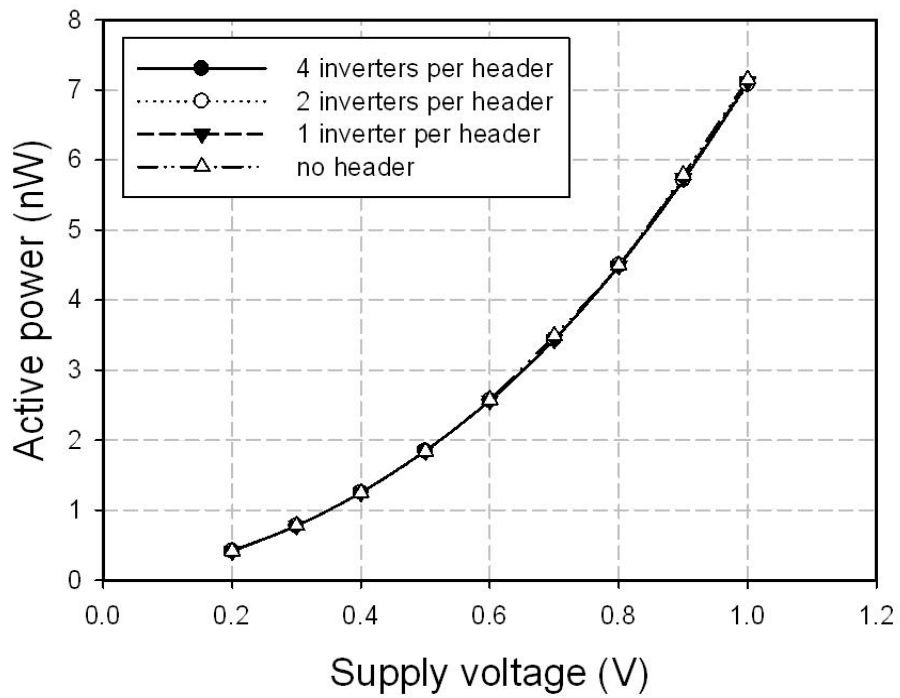


Figure 2.21: Active power comparisons when applying header power gating.

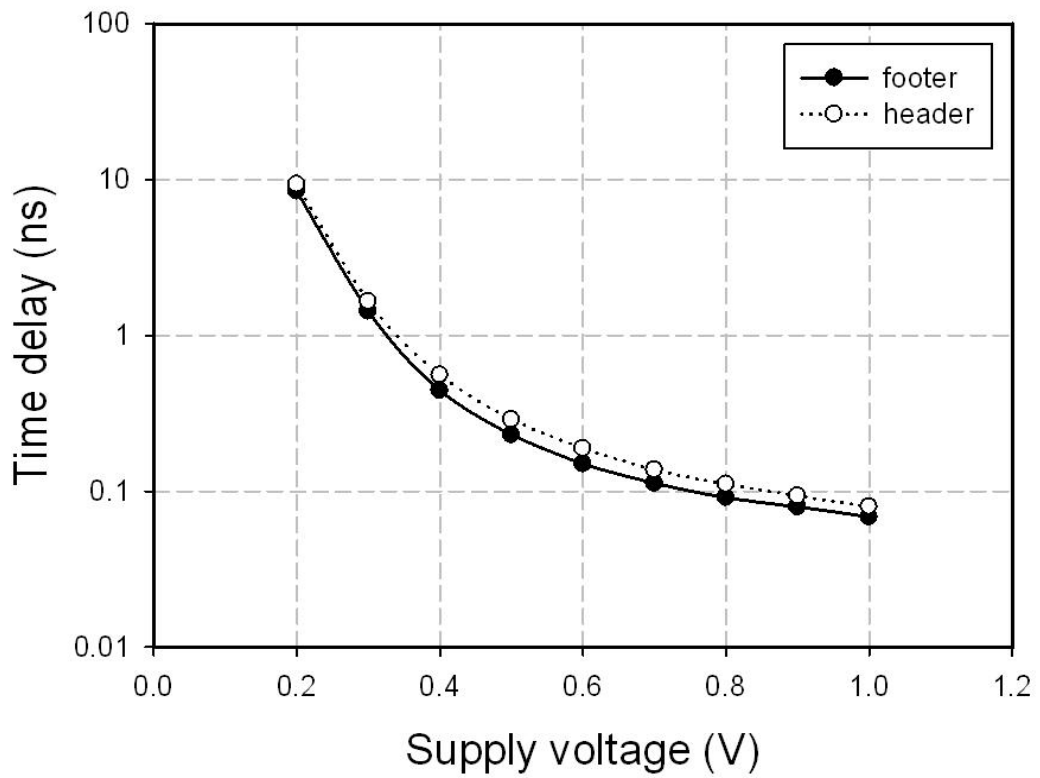


Figure 2.22: Time delay comparisons between footer and header. One footer/header is applied on four inverters.

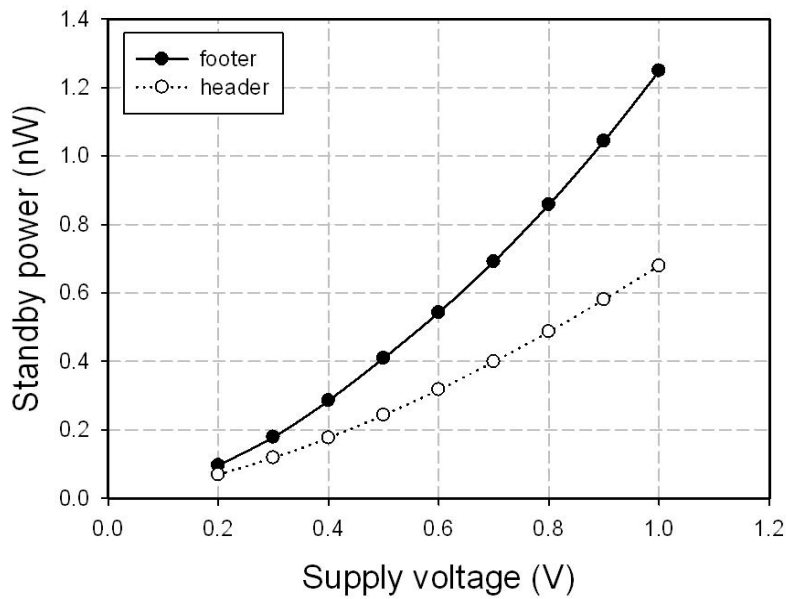


Figure 2.23: Standby power comparisons between footer and header. One footer/header is applied on four inverters.

2.3.3 Multiple Threshold Designs

Multiple threshold CMOS (MTCMOS) circuit has transistors with different threshold voltage. In general, there are regular threshold (regular- V_T) transistors, low threshold (low- V_T) transistors, and high threshold (high- V_T) transistors. Low- V_T transistors has larger driving ability, and can be used to achieve high performance, but it has the largest leakage current among the three types of transistors. High- V_T transistors has the least leakage current, but its performance is the slowest among the three types of transistors. The performance of regular- V_T transistors is in between low- V_T and high- V_T transistors. Following are three multiple threshold technologies:

1. Dual threshold CMOS: In a logic circuit, if a logic gate is in the critical path, the gate is implemented by low- V_T transistors to maintain performance; if a logic gate is in a non-critical path, the gate is implemented by high- V_T transistors for leakage power reduction [29]. This technique is demonstrated in Figure 2.24.
2. Mixed- V_T (MVT) CMOS technique: Unlike dual threshold CMOS technique, MVT CMOS design technique allows different thresholds within a logic gate, placing high- V_T transistors in non-critical paths to reduce leakage power, and placing low- V_T transistors in critical path(s) to maintain performance [30][31]. Figure 2.25 is an example of MVT CMOS logic gate. Suppose that the transistors in squares are the transistors in the critical paths, thus, assigning low- V_T . For the other transistors, high- V_T are assigned for leakage power reduction without degrading performance. Both dual threshold CMOS and MVT CMOS technique can achieve power reduction without delay and area overhead.
3. Multithreshold-voltage CMOS: Multithreshold-voltage CMOS (MTCMOS) technique is based on transistor stacking technique, but utilizes low- V_T transistors for logic gates and apply high- V_T transistors to power gating [32]. Examples are shown in Figure 2.26 and Figure 2.27. Assigning high- V_T to power gating devices can further improve leakage cut off efficiency, while the delay overhead can be compensated by low- V_T logic gates. Figure 2.28 and Figure 2.29 are testing examples of MTCMOS

circuit. Figure 2.30 and Figure 2.31 show the standby power comparison between inverter chain with and without MTCMOS technique. It is obvious MTCMOS technique significantly reduces standby power. Figure 2.32 and Figure 2.33 show the time delay comparison between inverter chain with and without MTCMOS technique. High- V_T has smaller driving current, thus resulting delay overhead. Delay overhead can be reduced by replacing regular- V_T transistors with low- V_T transistors. Figure 2.34 and Figure 2.35 show the active power comparison between inverter chain with and without MTCMOS technique. Active power reduction by MTCMOS is not apparent in this case, since the gate count under simulation is very limited.

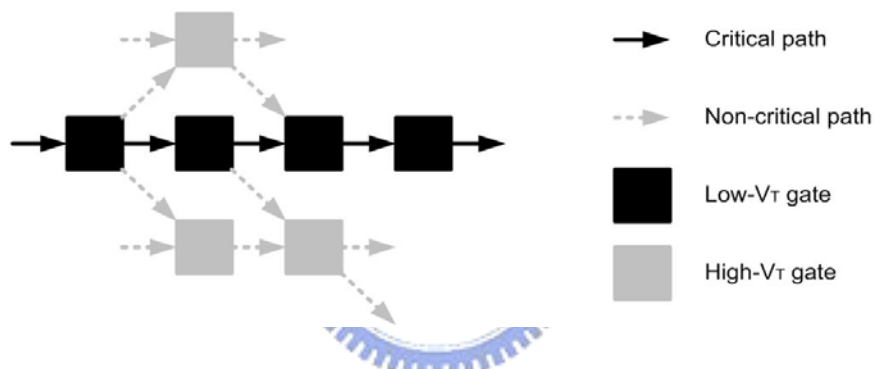


Figure 2.24: Dual threshold CMOS circuit.

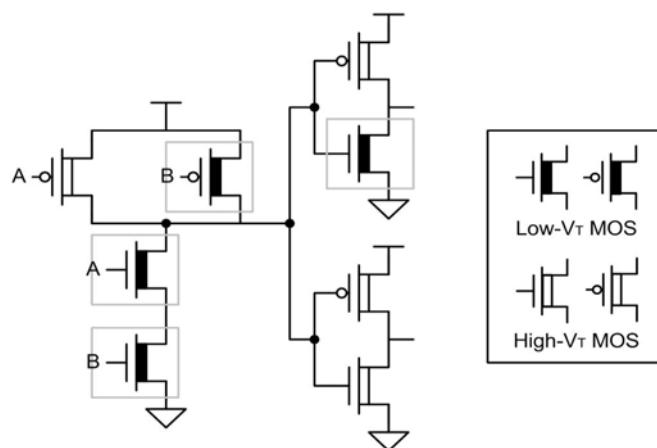


Figure 2.25: MVT CMOS scheme.

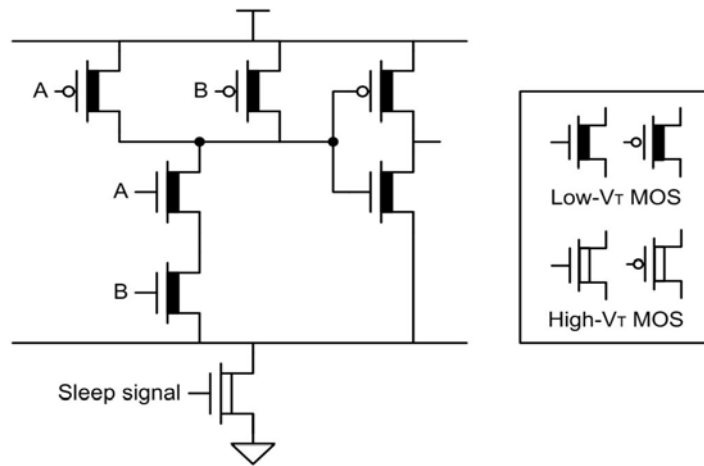


Figure 2.26: Footer insertion MTCMOS circuit.

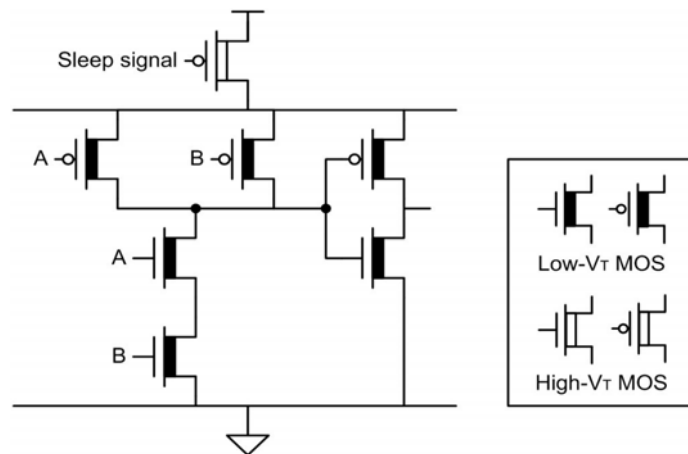


Figure 2.27: Header insertion MTCMOS circuit.

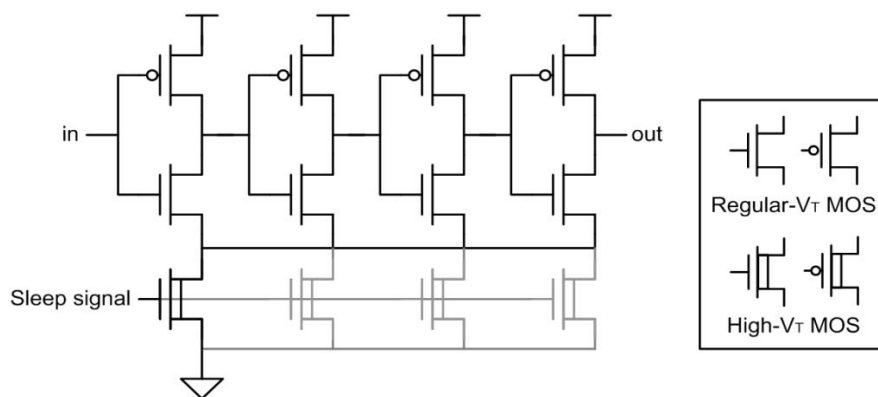


Figure 2.28: MTCMOS inverter chain with footer power gating.

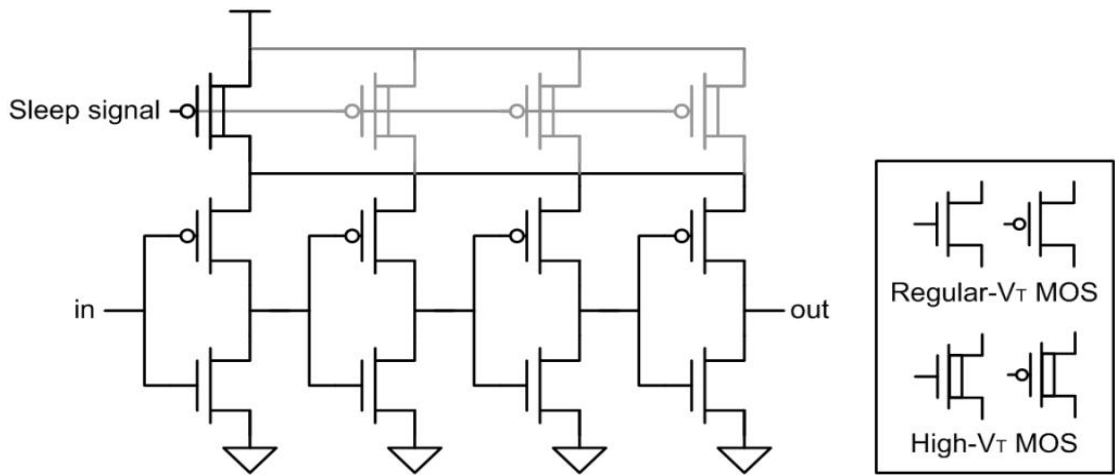


Figure 2.29: MTCMOS inverter chain with header power gating.

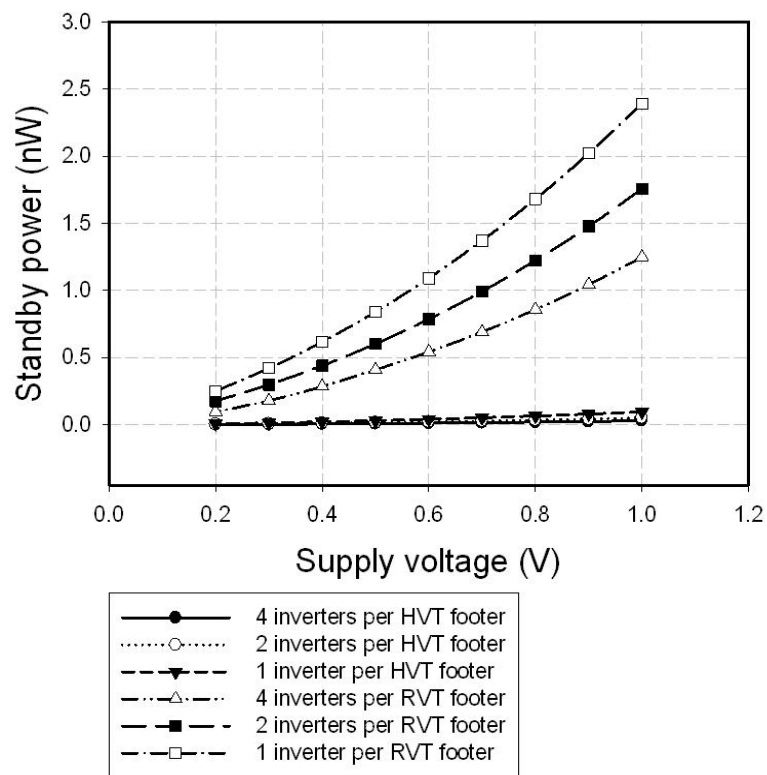


Figure 2.30: Standby power comparisons when applying footer insertion MTCMOS circuit.

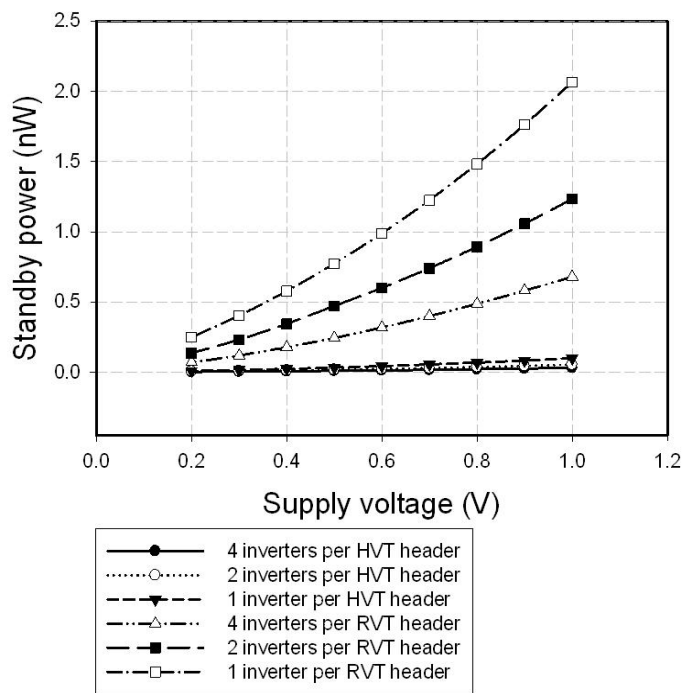


Figure 2.31: Standby power comparisons when applying header insertion MTCMOS circuit.

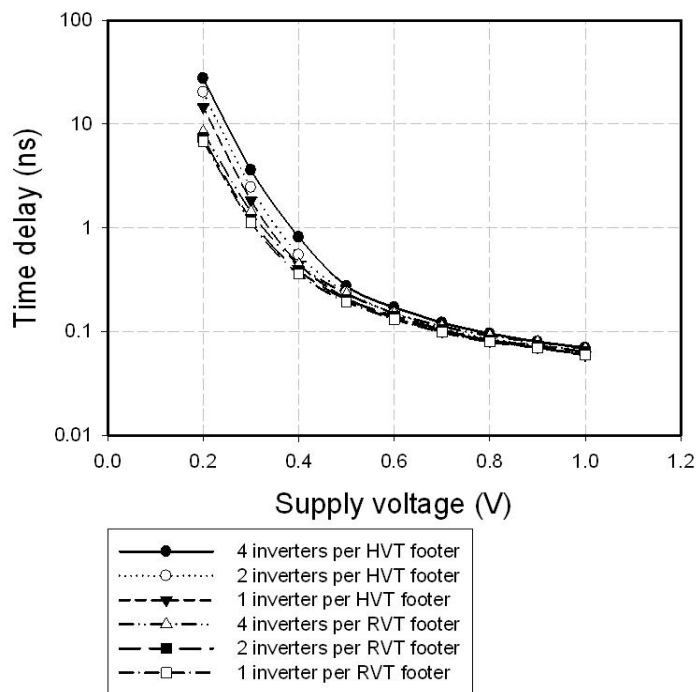


Figure 2.32: Time delay comparisons when applying footer insertion MTCMOS circuit.

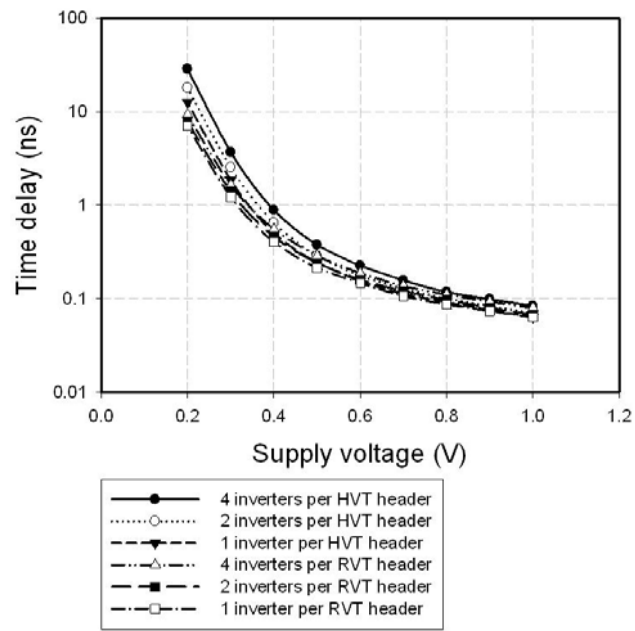


Figure 2.33: Time delay comparisons when applying header insertion MTCMOS circuit.

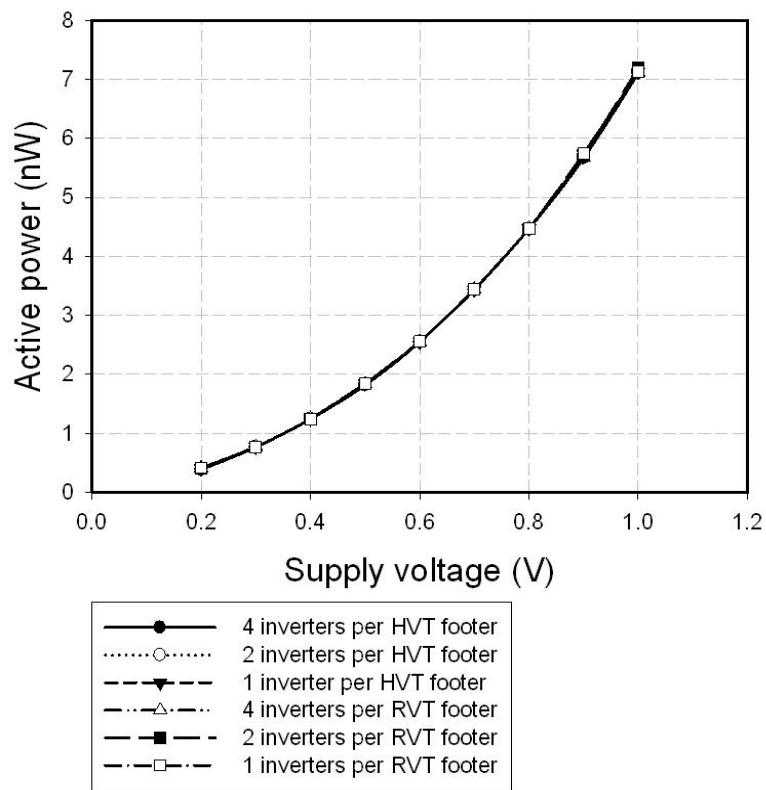


Figure 2.34: Active power comparisons when applying footer insertion MTCMOS circuit.

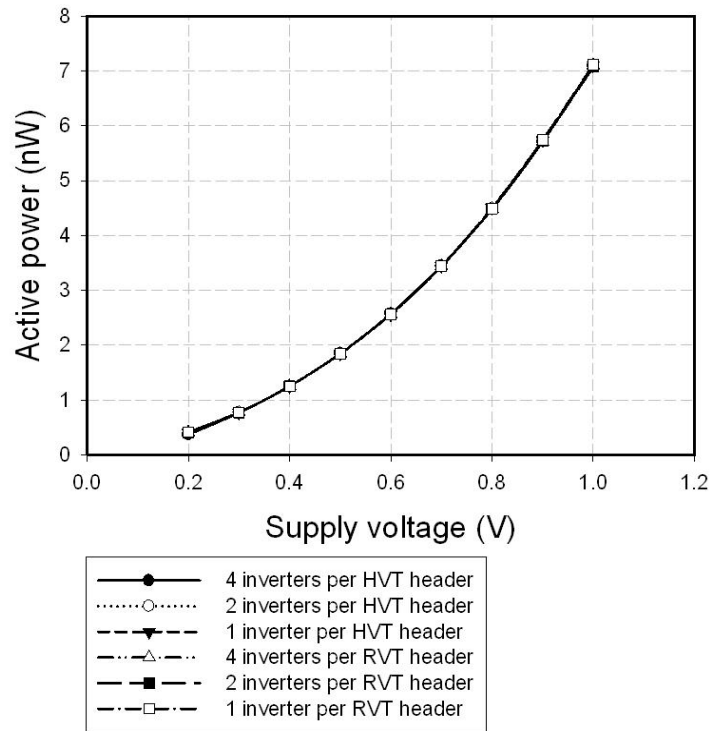


Figure 2.35: Active power comparisons when applying header insertion MTCMOS circuit.

2.4 Overview of SRAM Operation

Embedded memory typically occupies the largest portion of SoC die area, and has the largest influence on cost, power, performance, and reliability. It is predicted that over 90% of the future chip area is occupied by memory circuits [33]. Thus, robust low power memory design is a key for low power systems.

The most widely used form of embedded memory is the *static random access memory* (SRAM). Figure 2.36 [34] is a typical SRAM organization. It includes storage cells, row and column decoder for appropriate word selection, sense amplifiers to amplify bitline swing, read/write circuitry for proper read/write control and data buffer.

2.4.1 6T SRAM Cell

Figure 2.37 shows the schematic of the 6T SRAM cell commonly used in practice. The cell uses a single wordline and both true and complementary bitlines. The cell contains a pair of cross-coupled inverters for data storage and an access transistor for each bitline.

For read operation, bitlines are first precharged to high. The wordline is then activated, and one of the bitlines will be pulled down by the cell. For example, in Figure 2.38, $Q=0$ and $Q_b=1$, BL will therefore be pulled down by transistors MAL-MNL, while BL_b stays high. A differential signal is generated on the bitline pair, and the sense amplifier at the read output end will detect this small signal and transforms it into full swing voltage.

For write operation, one bitline is driven high and the other low. The wordline is then turned on, and data on bitlines will overpower the cell content with the new value. For example, in Figure 2.39, $Q=0$, $Q_b=1$, $BL=1$, and $BL_b=0$, Q_b will be forced to low, and Q will rise high.

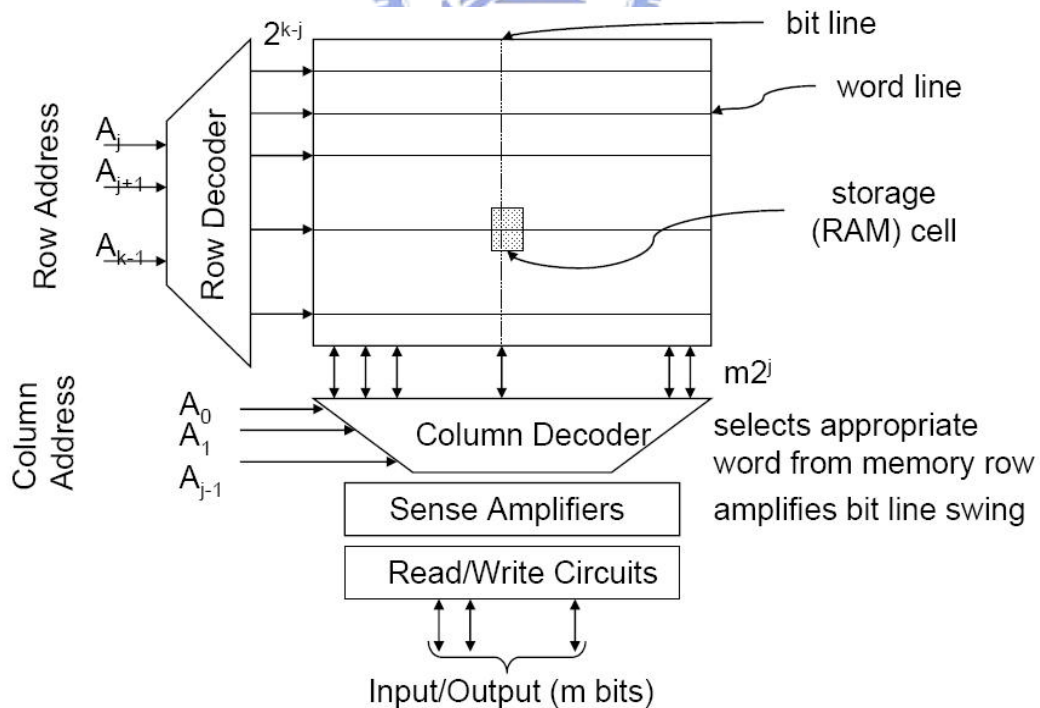


Figure 2.36: SRAM organization.

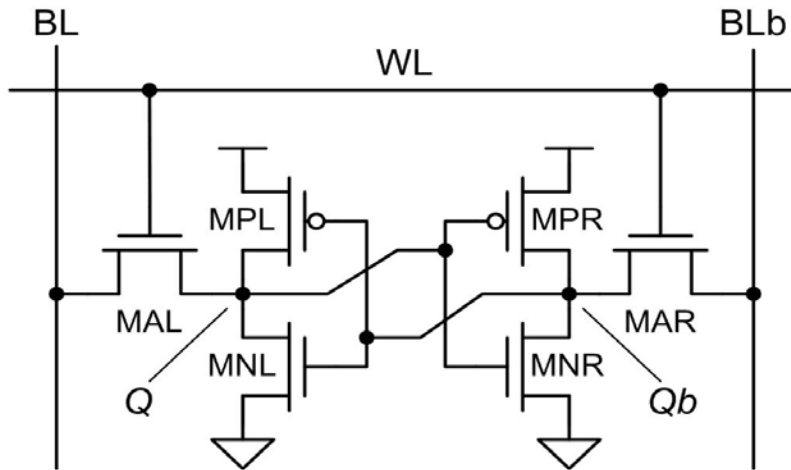


Figure 2.37: Conventional 6T SRAM cell.

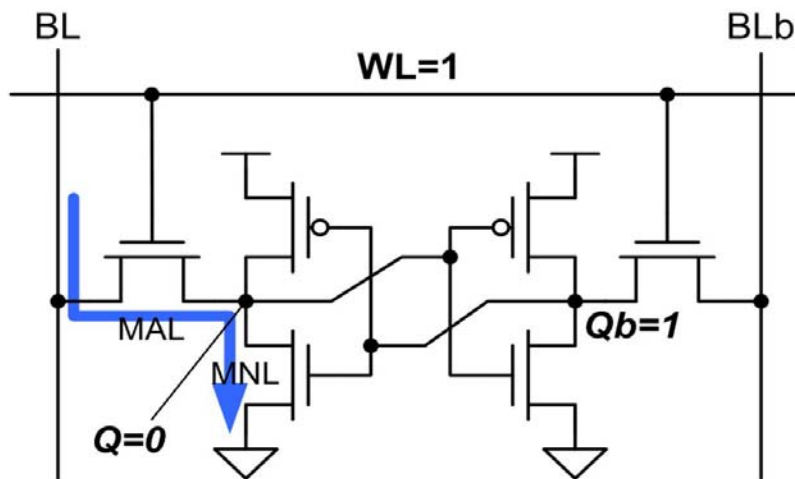


Figure 2.38: Read example of 6T SRAM.

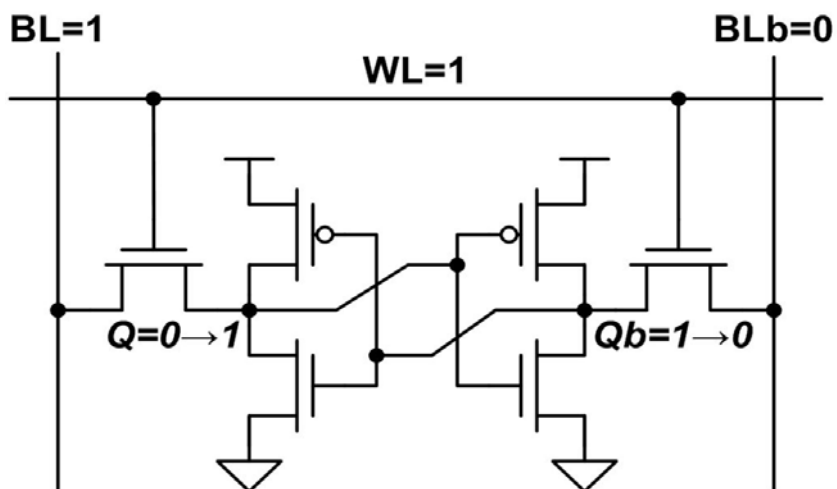


Figure 2.39: Write example of 6T SRAM.

2.4.2 SRAM Cell Stability

2.4.2.1 Hold Stability

Figure 2.37 is a conventional 6T SRAM bitcell. When the bitcell is holding data, the wordline (WL) is low so that NMOS access transistors (MAL and MAR) are off. The cross-coupled inverters must maintain bi-stable operating points in order to properly hold data. The best measure of the ability of the cross-coupled inverters to maintain their state is the static noise margin (SNM) [37]. The Hold SNM is defined as the maximum value of DC voltage noise that can be tolerated by the SRAM cell without changing the stored bit when the access transistors are off. Figure 2.40 shows the standard setup for modeling Hold SNM. DC noise sources V_N are introduced at each of the internal nodes in the bitcell. Cell stability changes as V_N increases. Figure 2.41 [36], known as the butterfly curve, is the most common way of representing the SNM graphically. The butterfly curve plots the voltage transfer characteristic (VTC) of Inverter R and the inverse VTC of Inverter L. Inverter R and Inverter L are shown in Figure 2.42. The SNM is defined as the length of the side of the largest square that can be embedded inside the lobes of the butterfly curve. When the value of V_N increases, the VTCs move horizontally and/or vertically. When the value of V_N is equal to the value of SNM, the VTCs meet at only two points. Further noise flips the cell content.

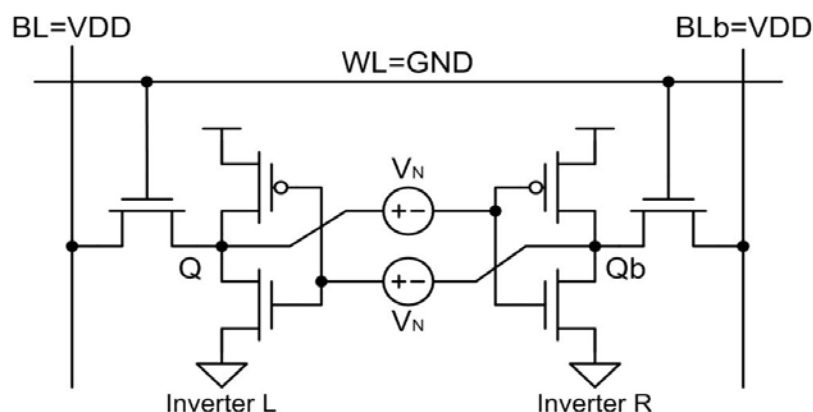


Figure 2.40: Standard setup for finding the Hold SNM.

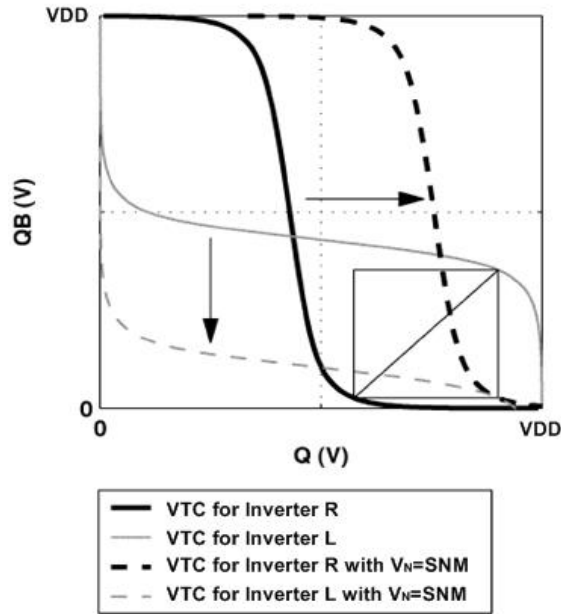


Figure 2.41: Butterfly curve plots for representing SNM. The VTCs of the cross-coupled inverters are represented by the solid curves. The length of the side of the largest embedded square in the butterfly curve is the SNM. When the worst case static noise is applied (e.g., $V_N = \text{SNM}$), the bitcell is mono-stable, thus losing its data.

2.4.2.2 Read Stability

The most common method to measure read stability is the Read SNM. SNM is defined in the previous subsection, but the setup for Read SNM is different from Hold SNM. Figure 2.42 shows the standard setup for modeling Read SNM. WL is on for read access; BL and BL_b are set to V_{DD} to indicate the initial value of bitlines are precharged to high.

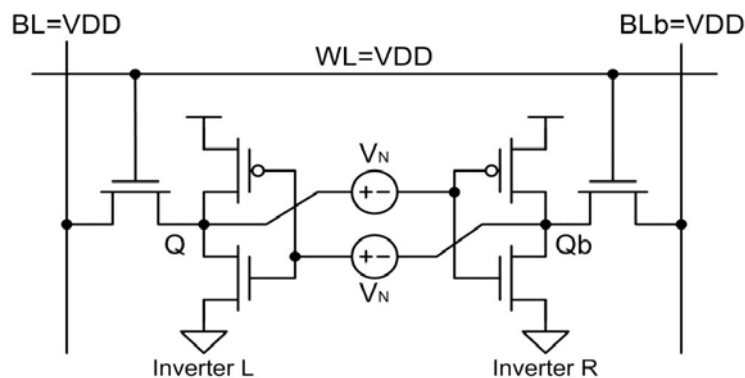


Figure 2.42: Standard setup for finding the Read SNM.

In a conventional 6T cell, Read SNM is worse than Hold SNM. During read, the cell begins with the wordline being turned on, with the bitlines initially high. This causes the low node within the cell to rise due to the voltage dividing effect across the access transistors and the pull down transistors. If this node voltage becomes close to the threshold of the pull down devices, process variations combined with noise coupling may flip the state of the cell. Figure 2.43 [36] shows example of butterfly curves during hold and read, revealing the degradation in SNM during read.

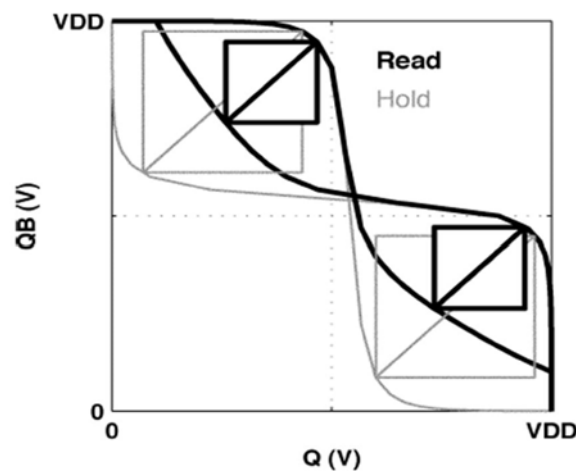


Figure 2.43: Example butterfly curve plots for hold SNM and Read SNM.

2.4.2.3 Write Ability

A common way to characterize write ability is the write margin (WM) or write trip point (WTP) [38][39]. WTP defines the maximum voltage on the bitline needed to flip the cell content. Figure 2.44 shows the conceptual setup to measure WTP of 6T SRAM cell. Figure 2.45 [35] shows a corresponding example of finding WTP. As the bitline voltage is lowered to a certain level, the cell content is flipped, indicating a successful write. Larger WTP means smaller voltage must be lowered below V_{DD} for successful write, indicating it is easier to write into the cell. If the WTP becomes negative, it means that it is not possible to write into the cell. To sum up, a higher WTP represents better write ability.

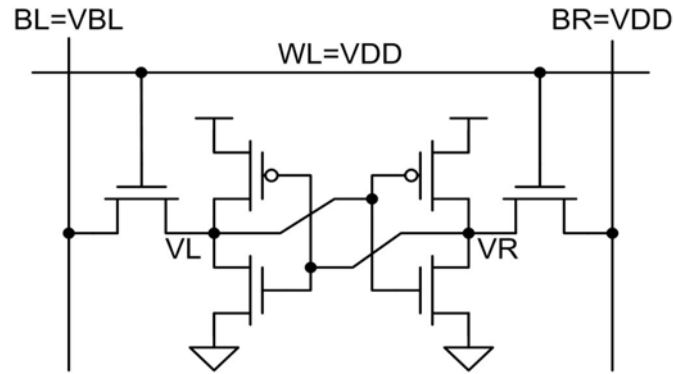


Figure 2.44: Setup for finding WTP.

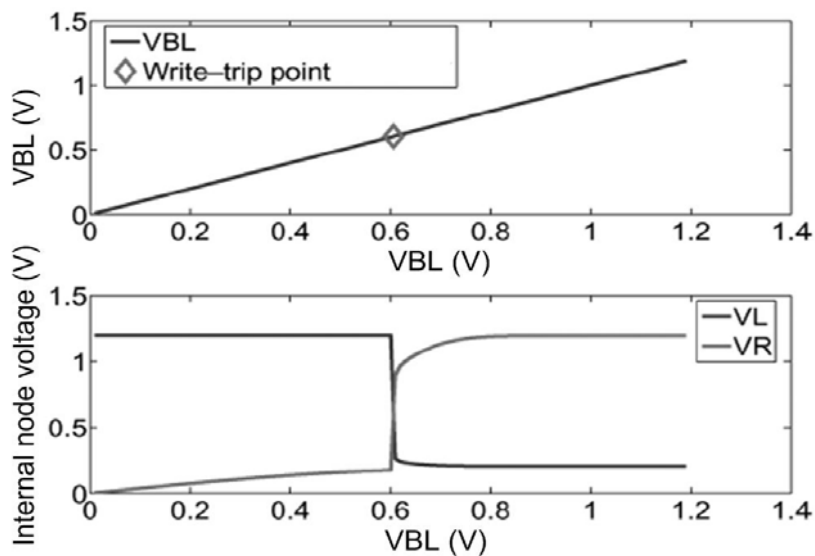


Figure 2.45: Write margin of a SRAM cell, determined by WTP.

2.4.3 Column Circuitry

Figure 2.46 shows a SRAM column configuration. The precharge circuit is used to precharge the bitlines high and equalize bitline pair before operation. Each column must also contain write drivers and read sensing circuits. Write drivers pull the bitline or its complement low during write operation. The sense amplifier shown is a commonly used latch type sense amplifier. When the sense amplifier is activated, the cross-coupled inverter pair pulls one output low and the other high through regenerative feedback.

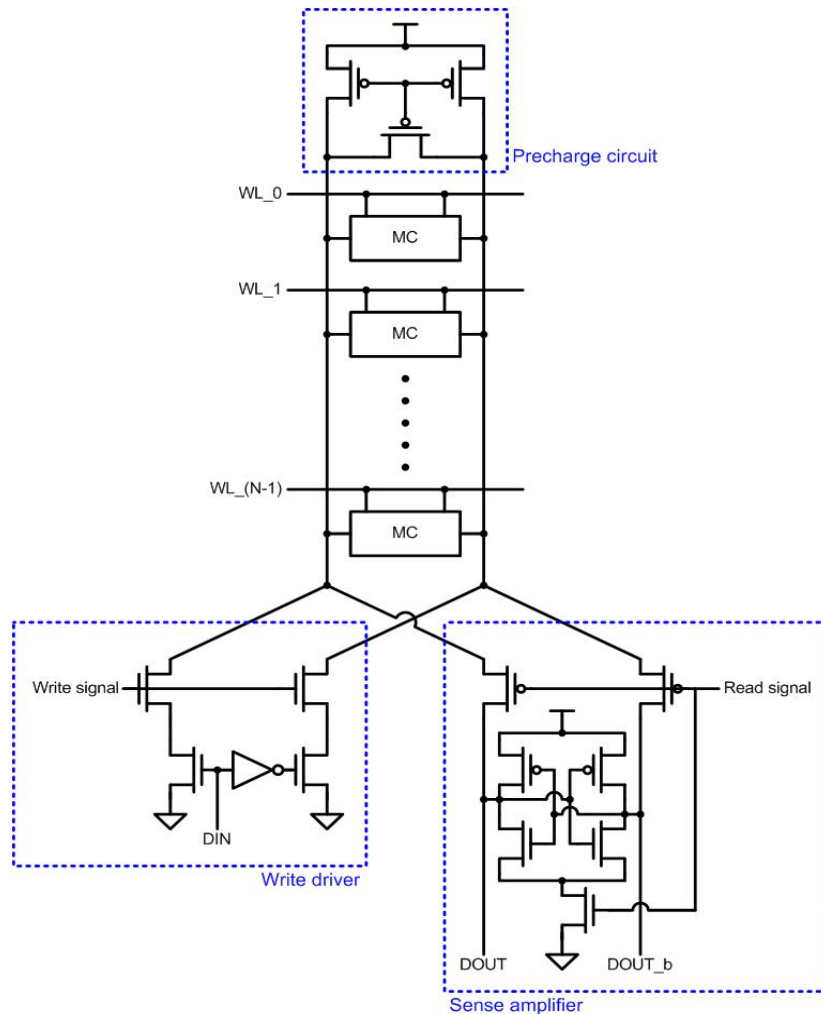


Figure 2.46: An SRAM column.

2.5 Single Read Bit Line 8T SRAM Cell

As variability concerns mount in future CMOS technologies, SRAM cell stability, which depends on delicately balanced transistor characteristics, becomes a significant concern. So there are a few of problems within 6T Cell. At the cell level, transistor strength ratios must be chosen such that cell static noise margin and write margin are both maintained, which presents conflicting constraints on the cell transistor strengths. For cell stability during a read, it is desirable to strengthen the storage inverters and weaken the pass-gates. The opposite is desired for cell writeability: a weak storage inverter and strong pass-gates. This delicate balance of transistor strength ratios can be severely impacted by device variation, which dramatically degrades stability and write margins, especially in scaled technologies. Low supply voltages further

exacerbate the problem as threshold voltage variation consumes a larger fraction of these voltage margins. Variability can thus limit the minimum operating voltage of SRAM.

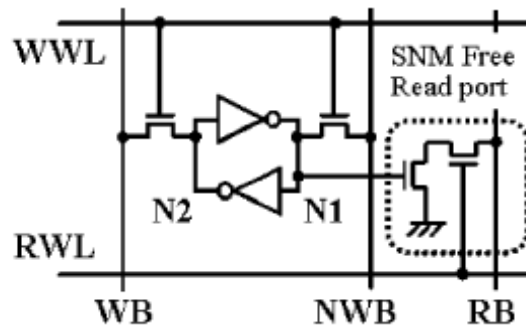


Figure 2.47: Single read bitline (SRBL) 8T cell.

In the single read bitline (SRBL) 8T cell, shown Figure 2.47, two transistors are added to create a disturb-free read mechanism. Since read and write are controlled by separate devices within the cell, the two are entirely decoupled—a level that 6T cells can never reach even with dual or dynamic voltage techniques. This widens the cell optimization space to achieve sufficient stability and writeability margins without the need for secondary or dynamic voltage supplies. In addition, like register files, separation of read and write ports in the cell itself enables 1R1W dual-port operation, which provides significant systems performance advantages.

2.6 Summary

In this chapter, power dissipation is first reviewed, including dynamic dissipation, leakage dissipation, and short circuit dissipation. After analyzing power dissipation sources, some useful low power techniques are presented, including supply voltage scaling, transistor stacking, and multiple threshold design. Testing examples and simulation results are demonstrated, which shows the effectiveness of applying these low power techniques. All simulations done in this chapter is based on UMC 90nm

CMOS technology.



Chapter 3

Low Power Write Assistant Scheme

3.1 Introduction

Conventionally, BL power and WL power are two main factors of SRAM active power. BL power can be reduced by decreasing its capacitance or voltage swing, such as reference. Besides, WL power can also be reduced by shortening the WL length.

In conventional 8T SRAM, WBL pairs are pre-charged first during write cycles, and the selected WWL is asserted rapidly. Figure 3.1(a) shows these signals. In addition, there are usually several words sharing a WWL. If one of these words is selected, others would face Write half-select disturb, and it would lead to high power waste.

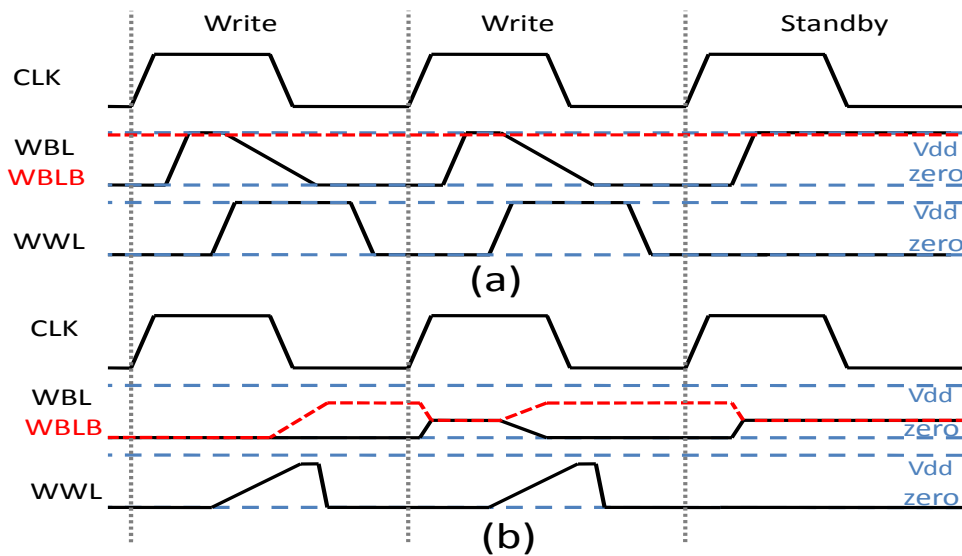


Figure 3.1: (a) Conventional WWL and WBL pair signals;
(b) Proposed WWL and WBL pair signals.

3.2 Low Power Write Assist Scheme

Write margin (WM) of SRAM cells can be defined as:

$$WM = V_{dd} - [V(WWL) - V(WBL)]_{trip}; \quad (1)$$

Where $[V(WWL) - V(WBL)]_{trip}$ represents the voltage difference between the selected WWL and WBL when SRAM cells are at meta-stable point during Write cycles. The Eq. (1) implies that data can be written into a SRAM cell by only turning on a WWL if one of WBL pairs has been connected to ground without WBL pairs pre-charged. However, Write half-select problem is more serious when initial WBL pairs voltage level is lower than $V_{dd} - V_{tn}$ during Write cycles, as Figure 3.2 shown. This phenomenon makes floating WBL scheme impractical if WWL is operated conventionally.

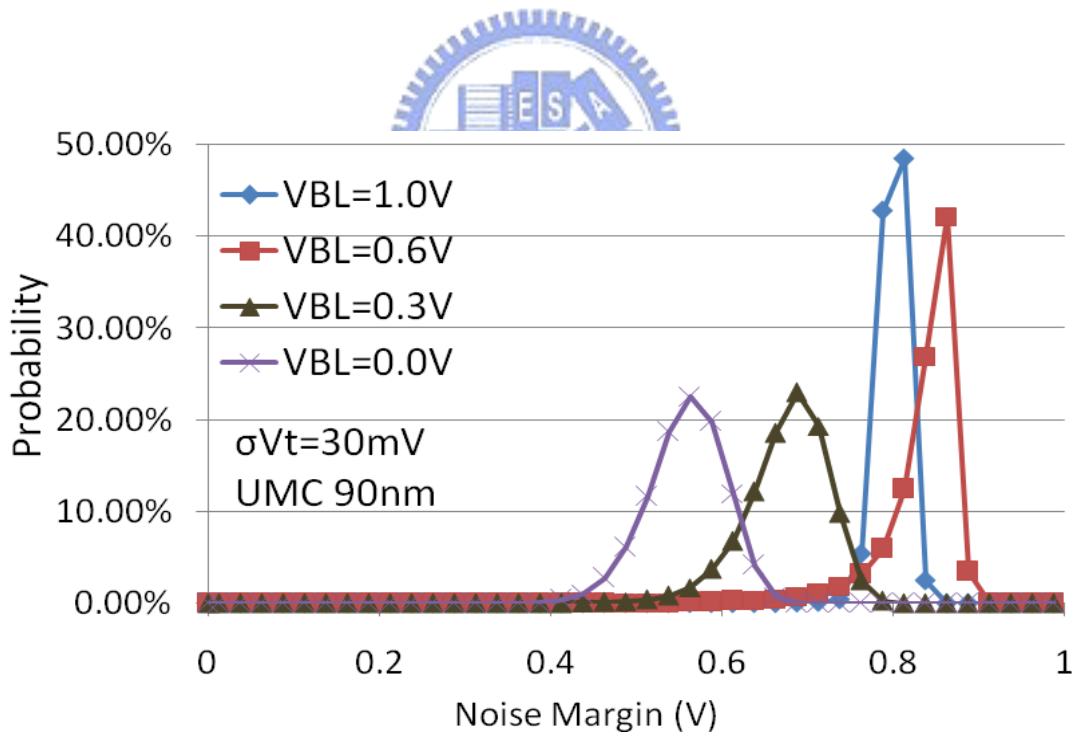


Figure 3.2: Monte-Carlo simulation results (32,000 nodes) for noise margin distributions with different initial WBL voltage levels. (Rising edge = 160ps, BL = 32-bit, $V_{dd} = 1.0v$).

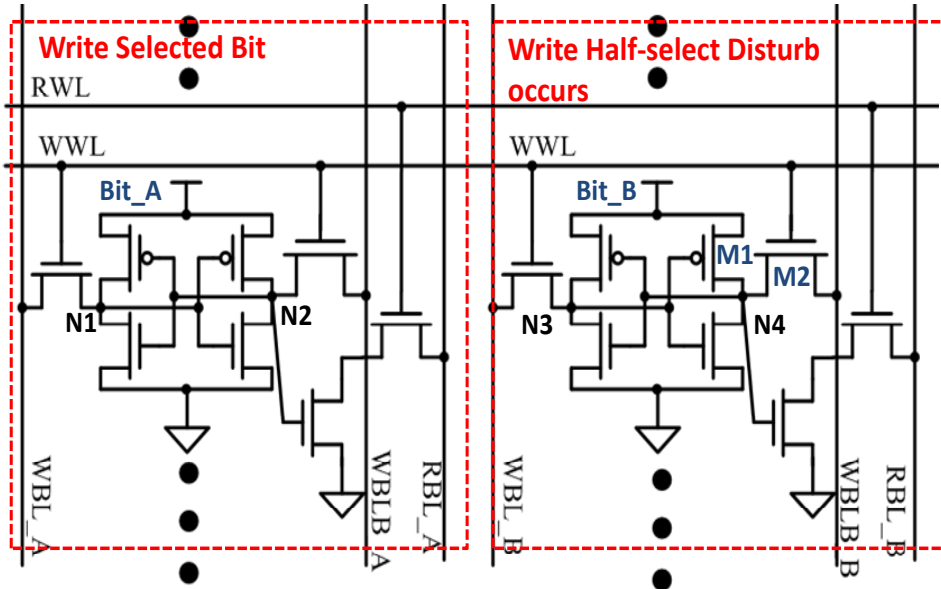


Figure 3.3: Write selected bit and Write half-select disturb.

An example is made to explain the mechanism of Write half-select disturb. Figure 3.3 shows two SRAM cells sharing a WWL. Initially, N1 and N3 are zero, and N2 and N4 are Vdd. It is also assumed that Bit_A is selected and Bit_B is half-selected during Write cycles. Thus, WBL_B and WBLB_B are floating. Moreover, M1 and M2 form a voltage divider after WWL turned on. If equivalent resistance of M2 (R2) is too small, the voltage of N4 would drop too low, and Bit_B data may be violated. Because R2 is determined by the voltage difference between WWL and WBLX_B, it can be kept as a larger value by controlling WL rising slowly or rising WBLB_B voltage. Figure 3.4 shows that noise margin (NM) of a Write-half-selected cell becomes better with longer WWL rising edge when WBL pairs are initially zero.

Otherwise, NM of a Write-half-selected cell is affected by WBL pair length. Figure 3.5 shows that NM of a Write-half-selected cell becomes worse with longer WBL pairs. Figure 3.6 is used to explain this phenomenon. After WWL turned on, Bit_B starts to charge WBLB_B. If WBLB_B is longer, WBLB_B voltage level would rise slower. As a result, equivalent resistance of M2 would be smaller, and NM of Bit_B becomes worse. Therefore, WWL should be turned on more slowly with longer WBL pairs, but it leads to SRAM performance degradation.

In our proposed Write scheme, WBL pairs are designed hierarchical to decrease

local WBL pair length. In addition, Pre-charged circuits are eliminated, but equalizers are still reserved. During Write cycles, voltage of WBL pairs are set equivalent first. Then, the selected WWL starts to rise. After detecting Write operation success, Write replica circuit sends a signal to turn off WWL driver instantly. These signals are shown in Figure 3.1(b). In the proposed scheme, WWL drivers are designed with weaker PMOSs, and the rising edge of WWL is longer than conventional designs. By this way, Write half-select disturb would be eased a lot, and WWL driver area is also decreased. Another advantage of this design is that WWL swing is reduced. Data can be written into a cell when WWL voltage is larger than $[V(WWL)-V(WBL)]_{trip}$. Therefore, WWL don't rise to V_{DD} . Otherwise, floating side of WBL pairs is charged by the Write half-select cell. These charges can help to ease Write half-select disturb in this Write cycle and they can also be reused in the following Write cycles.

This proposed scheme is also implemented in 65nm and 45nm technology nodes by PTM models. The simulation results are shown in Figure 3.6. It clearly shows that the proposed Write assist scheme can still work well in these advanced technology nodes.

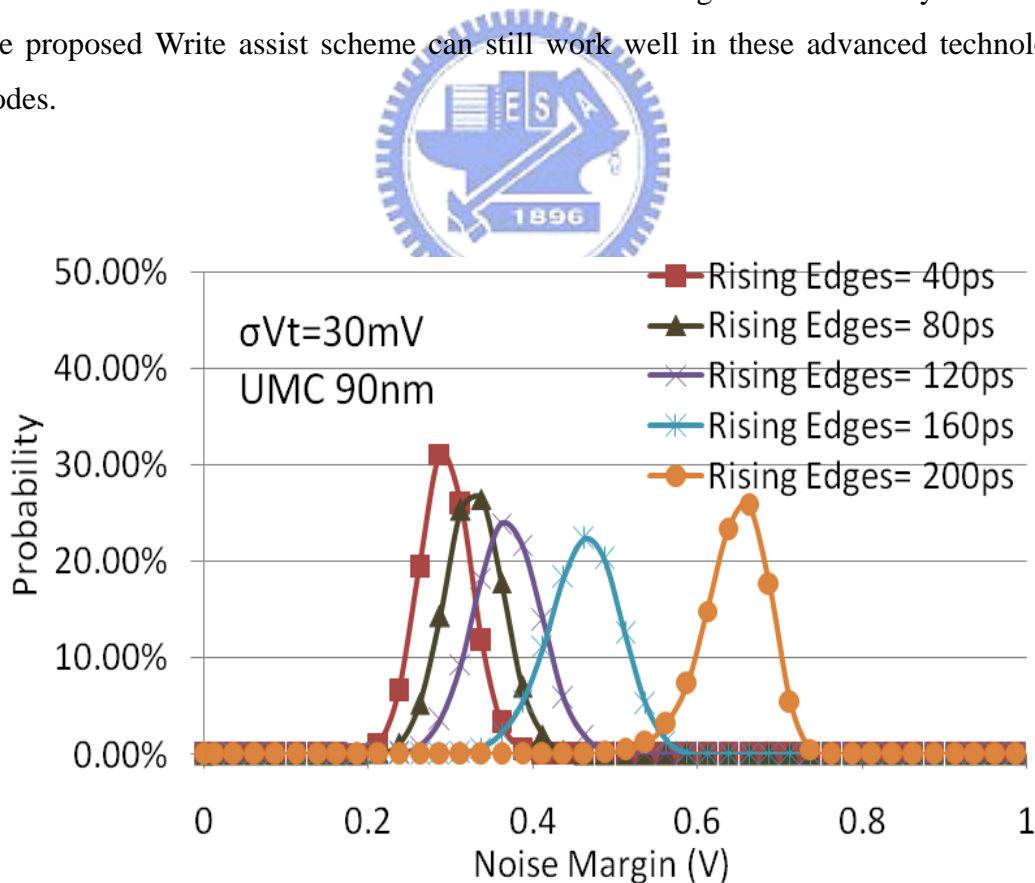


Figure 3.4: Monte-Carlo simulation results (32,000 nodes) for noise margin distributions with different rising edge. (Initial WBL voltage = 0V, BL = 32-bit, V_{DD} = 1.0v)

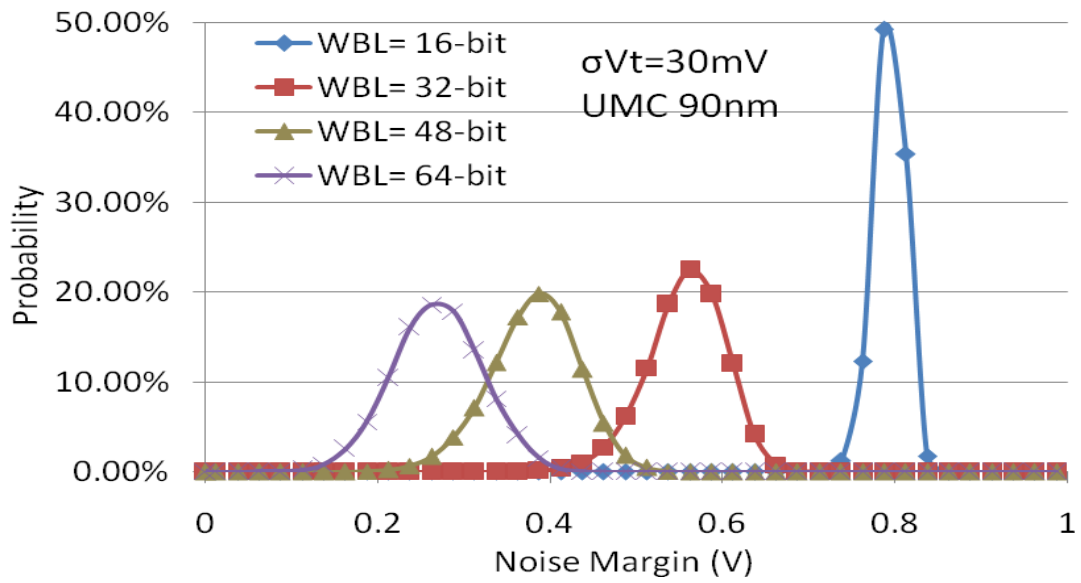
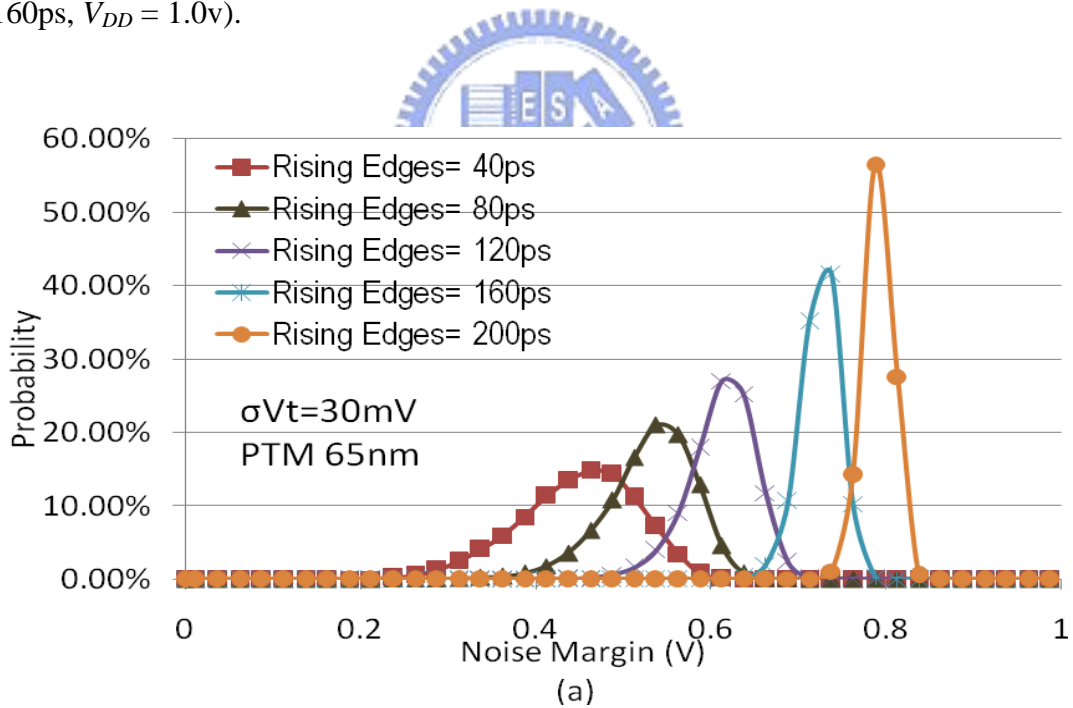


Figure 3.5: Monte-Carlo simulation results (32,000 nodes) for noise margin distributions with different BL length. (Initial WBL voltage = 0V, rising edges = 160ps, $V_{DD} = 1.0\text{v}$).



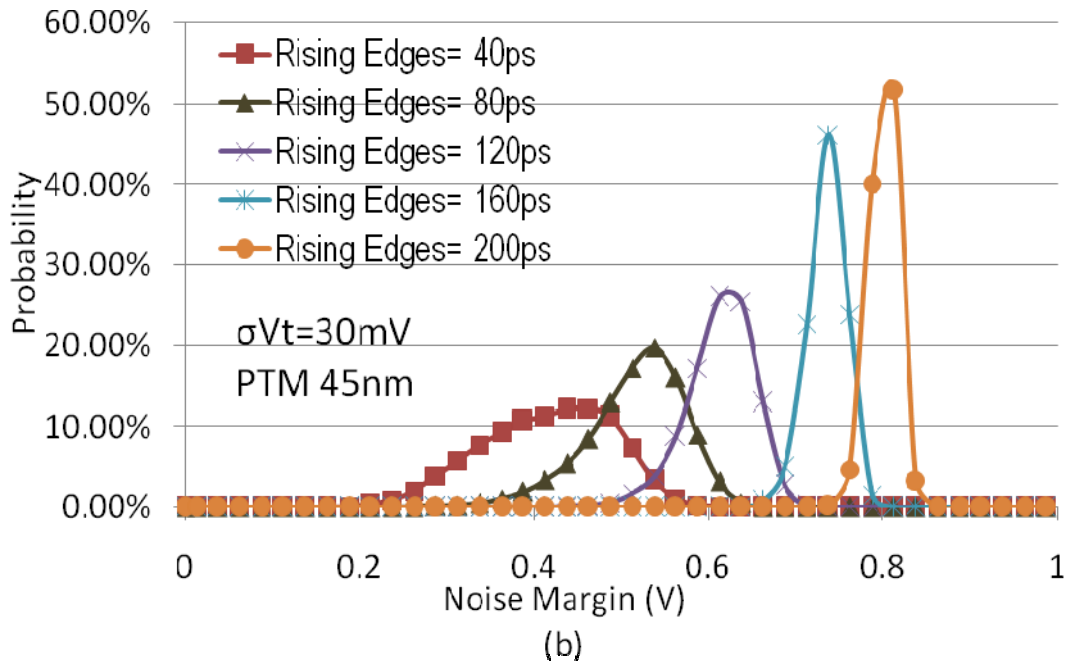


Figure 3.6: Monte-Carlo simulation results (32,000 nodes) for noise margin distributions with different rising edge in different technology nodes: (a) PTM 65nm; (b) PTM 45nm. (Initial WBL voltage = 0V, BL = 32-bit, $V_{DD} = 1.0\text{v}$)



3.3 Summary

In this chapter, a write assist scheme is proposed to resolve the serious write half-select disturb problem, and simulation results show that the proposed Write scheme can work well in more advanced technology nodes, such as 65nm and 45nm.

Chapter 4

Low Voltage Replica Wordline Timing Controller

4.1 Introduction

With the migration toward low supply voltages in low-power SRAM designs, threshold and supply voltage fluctuations will begin to have larger impacts on the speed and power specifications of SRAM's. Furthermore, with the fluctuations in the threshold voltages also not expected to decrease in future submicron devices, the delay variability of low-power circuits across process corners will increase in the future. The large delay spreads across process corners will necessitate bigger margins in the design of the bitline path in an SRAM, and also will result in larger bitline power dissipation and loss of speed. This problem can be mitigated by using a self-timed approach to designing the bitline path, based on delay generators which track the bitline delays across operating conditions.

The replica circuits can minimize the effect of operating conditions' variability on the speed and power. Replica memory cells and bitlines are used to create a reference signal whose delay tracks that of the bitlines. This signal is used to generate the sense clock with minimal slack time and control wordline pulsewidths to limit bitline swings. And this type of bitline swing control can be achieved by a precise pulse generator that can match the bitline delay.

4.2 Clock Matching

The prevalent technique to generate the timing signals within the array core essentially uses an inverter chain. This can take one of two forms—the first kind relies on a clock phase to do the timing [Figure. 4.1(a)], and the second kind uses a delay chain within the accessed block, and is triggered by the block select signal

[Figure 4.1(b)] or a local wordline. The main problem in these approaches is that the inverter delay does not track the delay of the memory cell over all process and environment conditions. The tracking issue becomes more severe for low-power SRAM's operating at low voltages due to enhanced impact of threshold and supply voltage fluctuations on delays as described by

$$\frac{\sigma_T^2}{T^2} \propto \frac{\sigma_{V_{dd}}^2 + \sigma_{V_t}^2}{(V_{dd} - V_t)^2} \quad (1)$$

which shows that delay variations are inversely proportional to the gate overdrive. Figure 4.2 plots the ratio of bitline delay to obtain a bitline swing of 120 mV from a 1.2V supply and the delay of two different delay elements for various operating conditions. One delay element is based on an inverter chain with a fan-out of four loading (diamonds), and the other is based on a replica structure consisting of a replica memory cell and a dummy bitline. The process and temperature are encoded as *XYZ* where *X* represents the nMOS type (*S* = slow, *F* = fast, *T* = typical), *Y* represents the pMOS type (one of *S*, *F*, *T*), and *Z* is the temperature (*H* for 115 C and for 25 C). The *S* and *F* transistors have a 2-sigma threshold variation unless suffixed by a 3, in which case they represent 3-sigma threshold variations. The process used is a typical 0.25-um CMOS process, and simulations are done for a bitline spanning 64 rows. We can observe that the bitline delay to inverter delay ratio can vary by a factor of two over these conditions, the primary reason being that, while the memory cell delay is mainly affected by the nMOS thresholds, the inverter chain delay is affected by both nMOS and pMOS thresholds. The worst case matching for the inverter delay chain occurs for process corners where the nMOS and pMOS thresholds move in the opposite direction. In the above simulations, it is assumed that they move independently, while in reality, there will be some correlation between them which would make the mismatch for the inverter delay chain less pronounced, but still worse than that of the replica element.

The delay element is designed to match the delay of a nominal memory cell in a block. But in an actual block of cells, there will be variations in the cell currents across the cells in the block. Figure 4.3 displays the ratio of delays for the bitline and the delay elements for varying amounts of threshold mismatch in the access device of the memory cell compared to the nominal cell. The graph is shown only for the case

of the accessed cell being weaker than the nominal cell as this would result in a lower bitline swing. The curves for the inverter chain delay element (hatched) and the replica delay element (solid) are shown with error bars for the worst case fluctuations across process corners. The variation of the delay ratio across process corners in the case of the inverter chain delay element is large even with zero offset in the accessed cell, and grows further as the offsets increase. In the case of the replica delay element, the variation across the process corners is negligible at zero offsets, and starts growing with increasing offsets in the accessed cell. This is mainly due to the adverse impact of the higher nMOS thresholds in the accessed cell under slow nMOS conditions. It can be noted that the tracking of the replica delay element is better than that of the inverter chain delay element across process corners, even with offsets in the accessed memory cell.

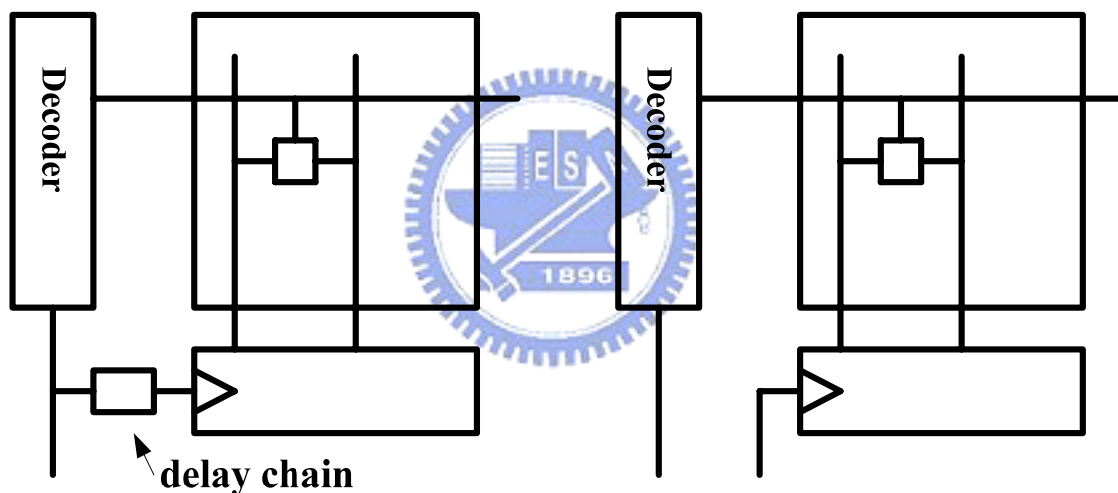


Figure 4.1: Common sense clock generation techniques.

There are two more sources of variations that are not included in the graphs above and make the inverter matching even worse. The minimum sized transistors used in memory cells are more vulnerable to ΔW variations than the nonminimum sized devices used typically in the delay chain. Furthermore, accurate characterization of the bitline capacitance is also required to enable a proper delay chain design. These two sources of variations would make the matching even worse for the inverter chain delay element.

All of the sources of variations have to be taken into account in determining the

speed and power specifications for the part. To guarantee functionality, the delay chain has to be designed for worst case conditions, which means that the clock circuit must be padded in the nominal case, degrading performance. Replica-based delay elements, by virtue of their good tracking, offer the possibility of designing SRAM's with tight specifications across all process corners. Two ways of creating and using these replica structures are explained in the following sections.



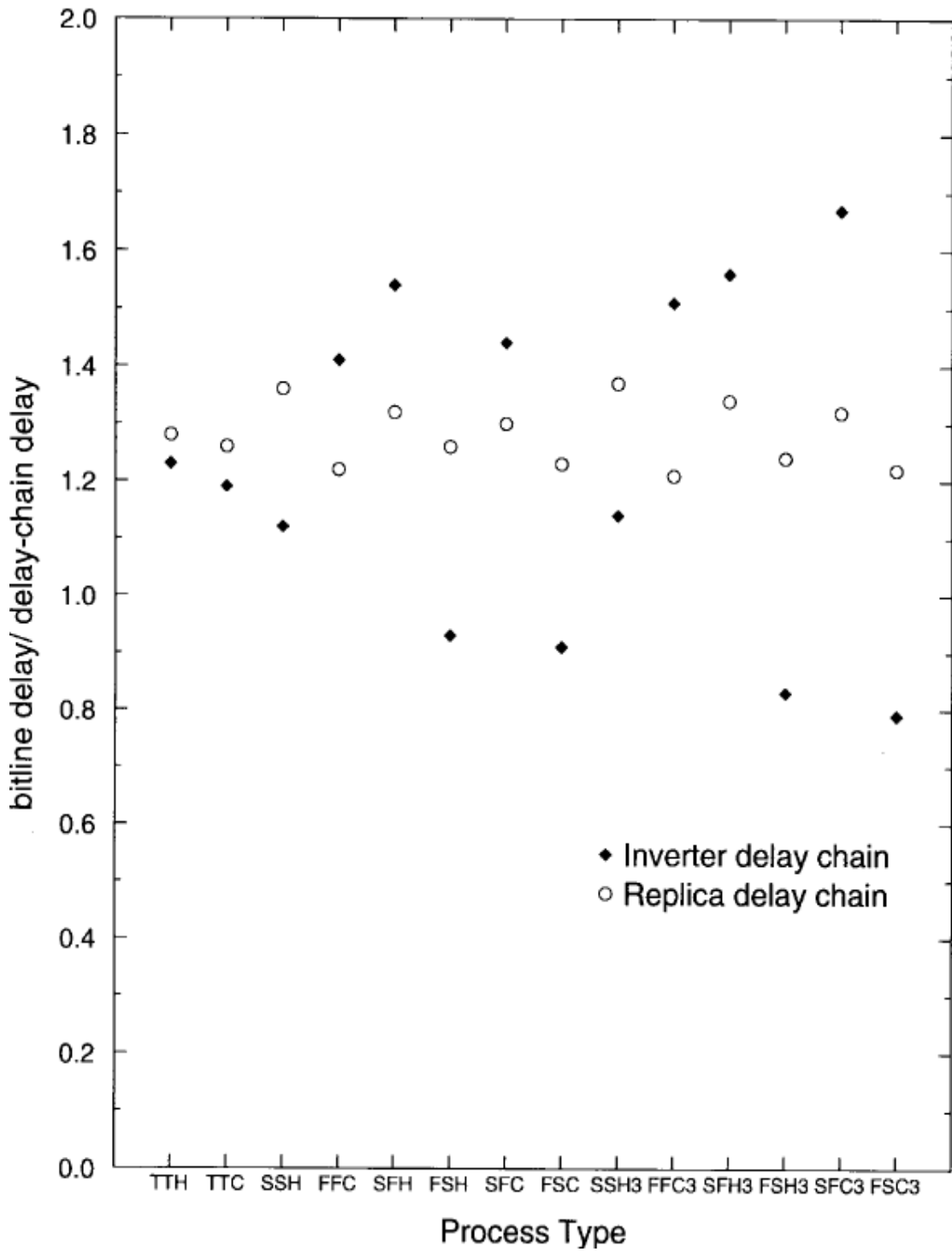


Figure 4.2: Delay matching between the bitline delay to generate 120mV and two delay elements, one based on an inverter chain and the other on a replica cell-bitline combination.

4.3 Low Voltage Replica Wordline Timing Controller

We take a 2x2 SRAM memory cell, called MC, array for an example. In the Figure 4.3, there are one column and one row replica memory cell, called RMC, around the memory array. So there is a 3x3 cell array in total. Besides, there are still column drivers, wordline drivers, and data detecting circuits.

The write wordline replica operation is discussed. First, we get a signal from decoder to activate the wordline driver and select the correct wordline. When the selected data is written by way of controlling column driver, the RMC in the red rectangle is written at the same time. Because the RMC is the same as common MC, the timing for data changing is equal. We can observe the data changing situation directly and sending a signal to data detecting circuit to decide what time to close the wordline driver to ensure the correct timing and less power consumption.

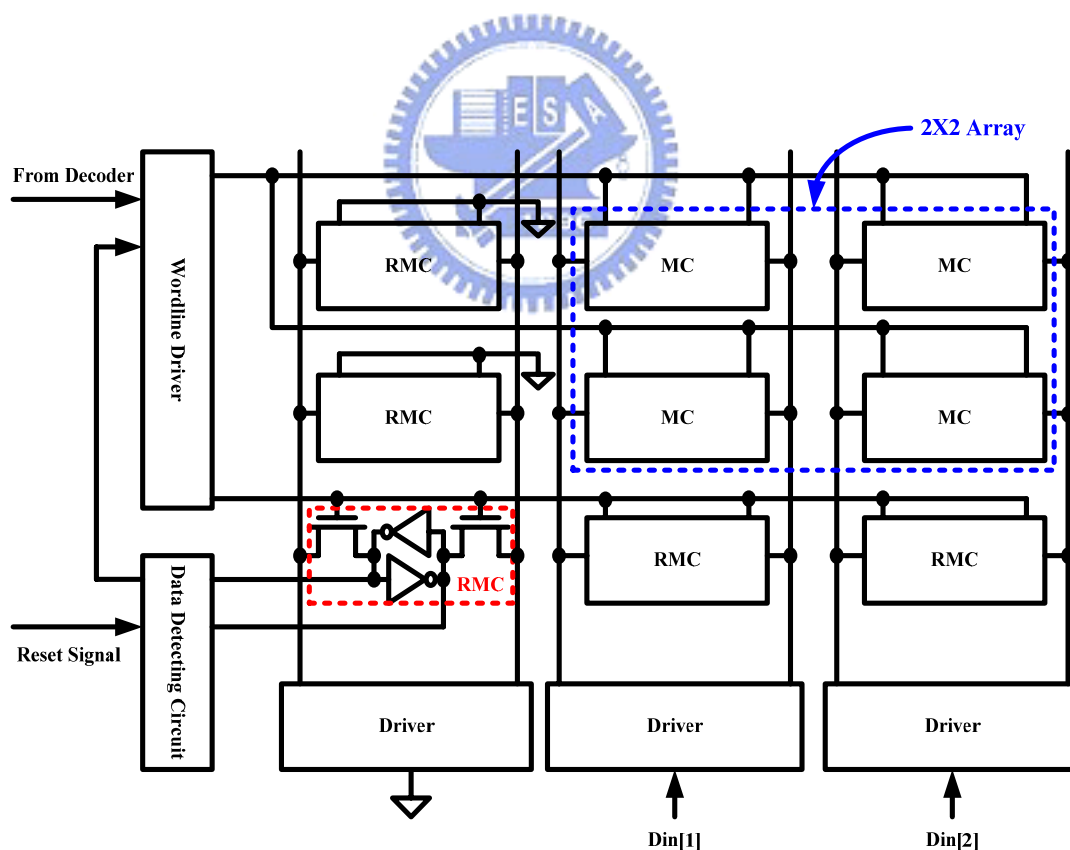


Figure 4.3: A 2x2 memory array with replica cell and circuits.

4.4 Summary

In this chapter, the basic concepts of wordline replica circuit are narrated carefully, and the detailed write/read wordline replica operation will be discussed at Section 5.3.2. And, A 8T low power SRAM including write assist circuit and write/read wordline replica circuit is introduced in chapter 5.



Chapter 5

A Robust Low Power SRAM Design with Write Assist Circuits

5.1 Introduction

SRAM is a key component of various modern SoC applications. Growing demand for multimedia-rich applications in handheld portable devices continues to drive the need for large and high-speed embedded SRAM to enhance system performance. In this chapter, a low power embedded SRAM with wide supply voltage ($V_{DD} = 0.5-1.0V$) has been designed with UMC 90nm CMOS technology model. A write assist scheme is proposed to resolve the serious write half-select disturb problem. It not only reduces the power dissipation, but also maintains the data accuracy. Furthermore, read/write replica circuits are designed to control access timing.

This chapter is organized as follows. The whole system architecture of the proposed robust and low power SRAM is carefully discussed in Section 5.2. Section 5.3 presents the circuit implementation to fulfill the intention of the proposed low power write assist scheme which is presented in Section 4.4 and the low power timing controller which is presented in Section 4.5. The completed design implementation is shown in Section 5.4. The simulation results of the proposed SRAM implemented with UMC 90nm CMOS technology model are shown in Section 5.5.

5.2 System Architecture

The complete proposed 8T low power SRAM, shown in Figure 5.1, consists of four major blocks; they are SRAM cell array including read/write circuitry and read/write wordline replica circuitry, control circuitry, read/write wordline driver, low power pre-decoder circuitry. The equivalent SRAM symbol of size 1024-word by 32-bit is shown in Figure 5.2 The signal description is shown in Table 5.1, and the command truth table is further shown in Table 5.2.

When the Control Circuit is activated ($CE = '1'$), it sends signals to decoder (Row Decoder and Column Decoder) to turn on the wordline by the wordline driver (RWL Driver or WWL Driver). If the operation is in the write operation condition ($WE = '1'$), input data ($D[31:0]$) is previously fed into the SRAM array. When the write wordline opening, the content of the word is overwritten. After the content is overwritten, the write wordline replica circuits in the SRAM array send a signal to turn off the WWL Driver to save power and to the accuracy of the data. On the other hand, if the operation is read operation ($WE = '0'$), the local read amplifier in the SRAM array starts to sense the content of the word then sending it to the global reading bitline. Resembling to the condition of the WWL, the read wordline (RWL) is activated at the same time. After global reading scheme achieving the sensing, it sends a signal to turn off the RWL Driver. Last, the output data ($Q[31:0]$) is read from the global reading scheme.

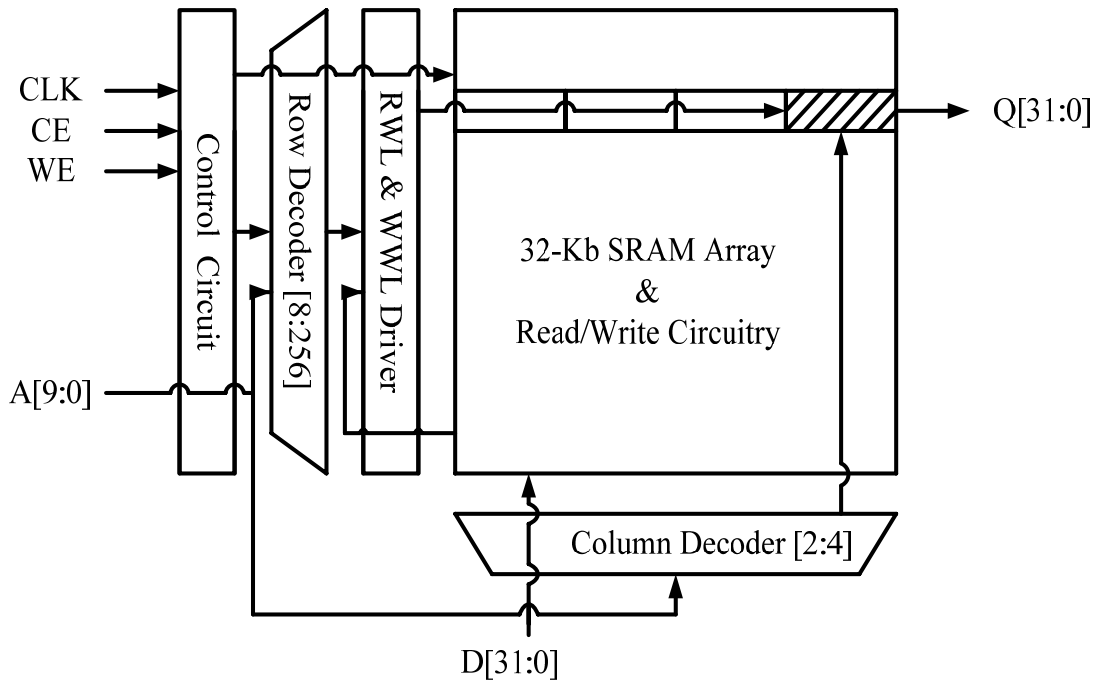


Figure 5.1: Block diagram of the proposed SRAM.

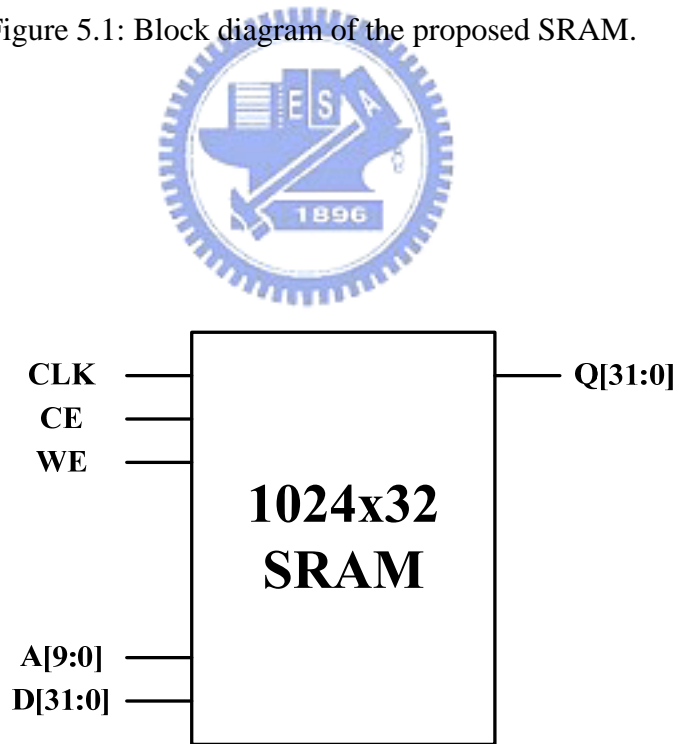


Figure 5.2: 1024x32 SRAM symbol.

Table 5.1: Signal description.

Input	CLK	clock frequency
	CE	chip enable
	WE	read/write enable
	A[9:0]	word address
	D[31:0]	32-bit input data
Output	Q[31:0]	32-bit output data

Table 5.2: Command truth table.

CE	WE	Operation
Low	X	chip disable
High	Low	read operation
High	High	write operation

5.3 Circuit Description

5.3.1 Cell Array

The access path in the proposed 8T low power SRAM is shown in Figure 5.3. There are 32 bits on a local Column, and each local Column has its dedicated local write driver (LW_Driver). In addition, local Columns are also organized interleaving, and four local Columns are combined to form a Section. Two Sections share a local read amplifier (LR_AMP), and eight Sections form a Block. There are 32 Blocks in this 32-Kb SRAM. All Sections in the same Block share an output circuit and a global

write driver (GLW_Driver). In order to simply timing controller designs, global read/write path are static.

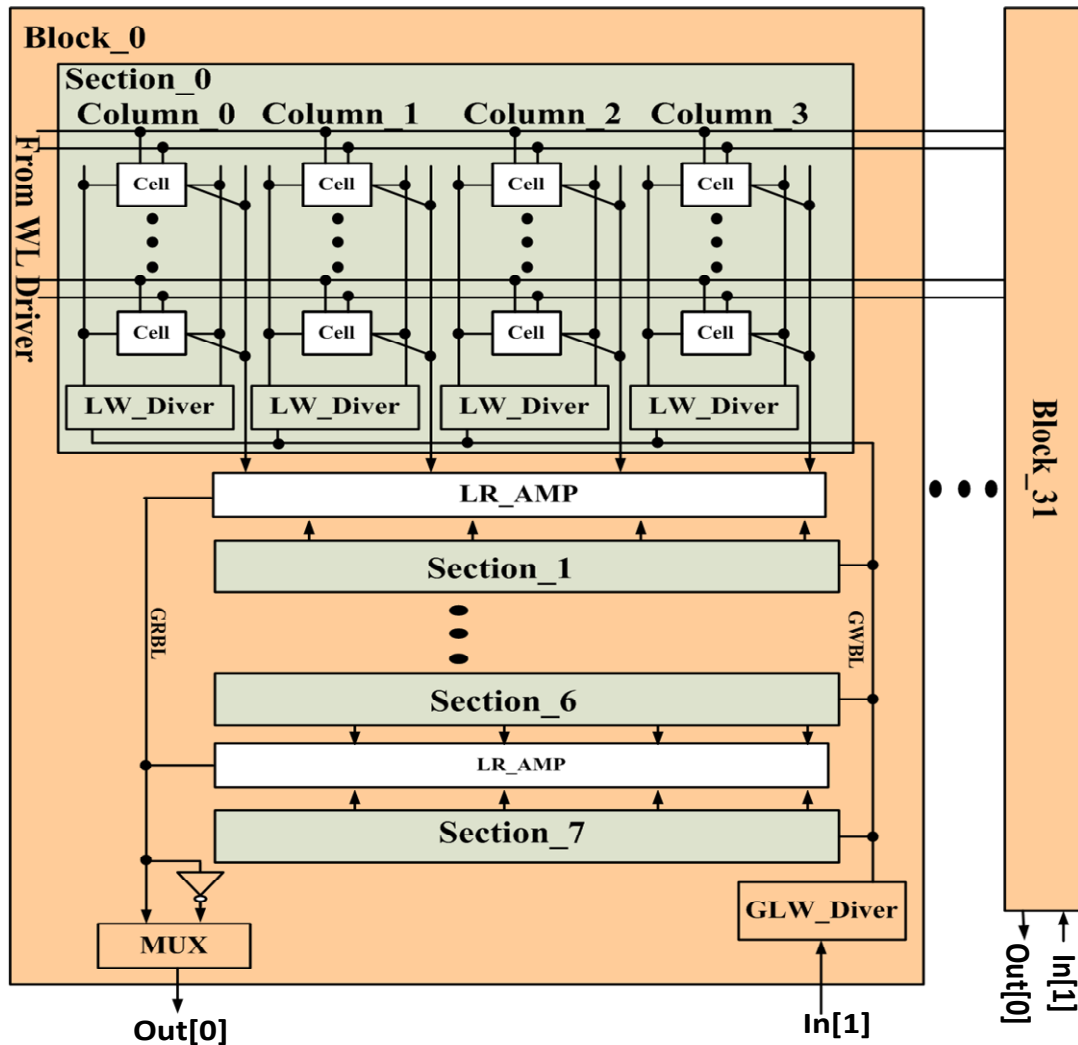


Figure 5.3: The proposed 8T SRAM array.

Because LR_AMP is shared by two sections, its sensing results of LR_AMP are complement with different section inputs. Therefore, the global output circuit contains an inverter and a multiplexer. By selecting outputs directly from GBLs or inverters, output signals would be correct.

5.3.2 Read/Write Wordline Replica Circuitry

Read and Write replica circuits are also designed in the proposed SRAM. They are shown in Figure 5.4. When selected cells are at trip points, Write replica sends a signal to turn off the selected WWL as soon as possible. Thus, WWL swing can be reduced. On the other hand, replica RBL is pre-charged to $V_{dd}-V_{tn}$ at beginning of Read operations. The reason is that RBL voltage level would be charged to $V_{dd}-V_{tn}$ if successive “logic 1” are read. By this special design, a local read replica circuit can monitor the worst Read case. After detecting zero on replica RBL, a Read replica circuit sends a signal to turn off pass transistors of LR_AMPs. Then, LR_AMPs finish sensing local RBLs.

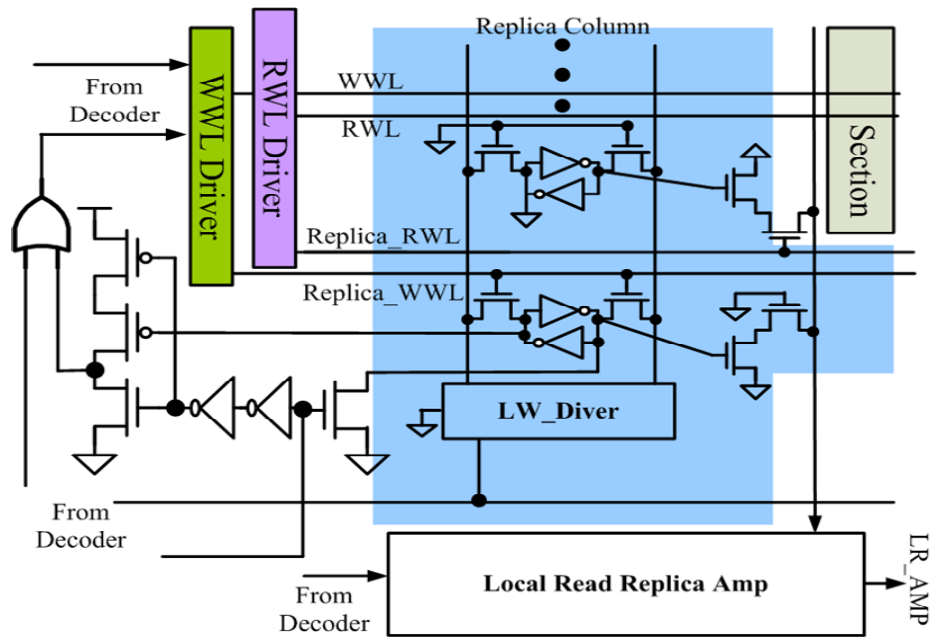


Figure 5.4: Read/write replica circuits.

5.3.3 Wordline Driver

The wordline driver (WLD) is analogous to the two input NOR gate in nature. The Figure 5.5 shows the structure of the wordline driver. In order to match the structure of the SRAM architecture, there are 32 WLDs in a block. The first kind of

the inputs are EWL[31:0] from the row decoder to decide which wordline in some Block of the SRAM array is activated whether in the write operation or in the read operation. In order to reduce the power consumption, when the write or read operation is finished, the other input called WWL_rep from the replica circuitry is fed into the WLDs to discharge the wordline.

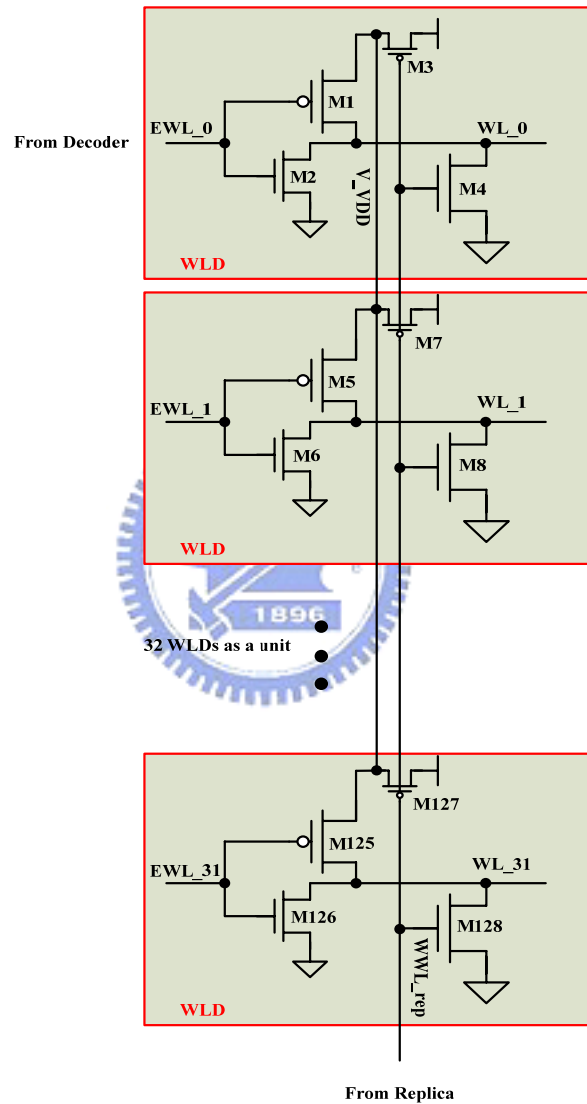


Figure 5.5 : The wordline driver (WLD) structure.

There are two kinds of the WLD. One is the write wordline driver (WWL Driver), and the other is the read wordline driver (RWL Driver). They are completely equal in the appearance, but different in the size tuning. To prevent the write-half-select

condition appearing in the write operation, we narrow the width of the total PMOS to lower the speed of the WWL rising.

5.4 Design Implementation

A 1024-word by 32-bit robust low power 8T SRAM design with write assist circuits is implemented in UMC 90nm CMOS technology with 1V supply voltage. This sizing configuration is suitable for being a basic SRAM block. The target application is the portable devices, where chip design is required to be highly integrated in a tiny area, and data needed highly stability. The proposed design is fully functional in all process corners. The layout view of the SRAM block is shown in Figure 5.6.

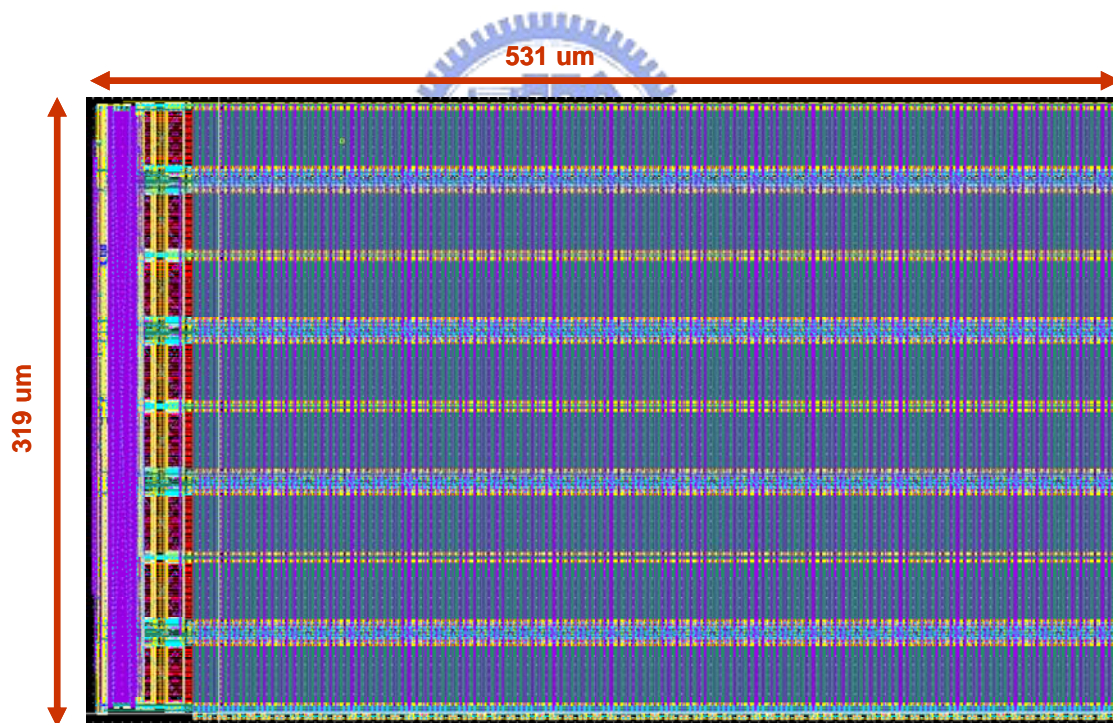


Figure 5.6 : Layout view of the proposed SRAM.

5.5 Simulations Result

As shown in Table 5.3, SRAM size, system frequency, supply voltage, read and

write power are given by the system speculation. The proposed robust low power SRAM is simulated in UMC 90nm CMOS technology. According to simulation results, this 32-Kb SRAM can operate at 1GHz when $V_{DD} = 1.0V$. It also can operate at 200MHz when $V_{DD} = 0.5V$. When this SRAM works at 1GHz, it takes 7.73mW and 9.22mW power consumption during read and write operations respectively. When it works at 200MHz, it takes 754uW and 826uW power consumption during read and write operations respectively. When supply voltage is 1.0V, WWL rising edge is 0.3ns and its voltage swing is 850mV. When supply voltage is 0.5V, WWL rising edge is 1.18ns and its voltage swing is 330mV.

Table 5.3: Summary of the SRAM features.

Technology	UMC 90nm CMOS
SRAM Configuration	1024-word by 32bit (32-kb)
Area	531um X 319um
Supply Voltage	1.0V ~ 0.5V
Frequency	1GHz ~ 200MHz
Read Power	7.73mW ~ 754uW
Write Power	9.22mW ~ 826uW

Process, voltage, and temperature variations are simulated. The proposed design is fully functional within +/-% voltage variation, 0°C to 100°C temperature variation, and all process corner. Simulations results are summarized in Table 5.4, Table 5.5, Table 5.6.

Table 5.4: Process corner simulation (@500mV ; 25 °C).

	Read Power (uW)	Read Power (uW)
TT	754	826
SS	500	496
FF	2840	2901
FNSP	1089	1239
SNFP	585	632

Table 5.5: Voltage variation simulation (@TT corner ; 25 °C).

	Read Power (uW)	Write Power (uW)
450mV	602	750
500mV	754	826
550mV	835	923

Table 5.6: Temperature variation simulation (@TT corner ; 500mV).

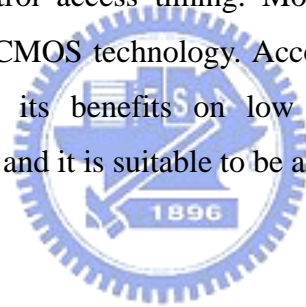
	Read Power (uW)	Write Power (uW)
0°C	560	624
25°C	754	826
100°C	2459	2533

Chapter 6

Conclusions

6.1 Conclusions

A novel 200MHz - 1GHz low power 8T SRAM is proposed. A low power write assist scheme is also proposed to resolve the serious write half-select disturb problem, and the simulation results show that the proposed write scheme can work well in more advanced technology nodes, such as 65nm and 45nm. Furthermore, read/write replica circuits are designed to control access timing. Moreover, a 32-Kb 8T SRAM is implemented in UMC 90nm CMOS technology. According to simulation results, the proposed 8T SRAM shows its benefits on low power access operations and wide-operating voltage range, and it is suitable to be adopted in portable devices.



Bibliography

- [1] S. Hanson, B. Zhai, K. Bernstein, D. Blaauw, A. Brant, L. Chang, K. K. Das, W. Haensch, E. J. Nowak, and D. M. Sylvester, "Ultralow-Voltage, Minimum-Energy CMOS," *IBM Journal of Research and Development*, vol. 50, no. 4/5, pp. 469-490, July/September 2006.
- [2] W. Haensch, E. J. Nowak, R. H. Dennard, P. M. Solomon, A. Bryant, O. H. Doku-maci, A. Kumar, X. Wang, J. B. Johnson, and M. V. Fischetti, "Silicon CMOS devices beyond scaling," *IBM Journal of Research and Development*, vol. 50, no. 4/5, pp. 339-361, July/September, 2006.
- [3] Y. Nakagome, M. Horiguchi, T. Kawahara, and K. Itoh, "Review and Future Prospects of Low-Voltage RAM Circuits," *IBM Journal of Research and Development*, vol. 47, no. 5/6, pp. 525-552, Septempber/November, 2003.
- [4] H. Qin, "Deep Sub-Micron SRAM Design for Ultra-Low Leakage Standby Operation," Ph.D. dissertation, University of California, Berkeley, 2007.
- [5] T. Norgall T, R. Schmidt, T. von der Grun, "Body Area Network, a Key Infrastructure Element for Patient-Centered Medical applications," *Biomed. Tech (Berl)*, 47:365-368, 2002.
- [6] A. Wang, B. H. Calhoun, and A. P. Chandrakasan, *Sub-threshold Design for Ultra Low-Power Systems*. Springer US, 2006, ch. 1, pp. 1-6.
- [7] S. Roundy, P. Wright, and J. Rabaey, "A Study of Low Level Vibrations as a Power Source for Wireless Sensor Nodes," *Computer Communications*, vol. 26, no. 11, pp.1131-1144, 2003.
- [8] S. Roundy, D. Steingart, L. Frechette, P. Wright, and J. Rabaey, *Power Sources for Wireless Sensor Networks*. Springer-Verlag Berlin Heidelberg, 2004.
- [9] R. Weinstien, "RFID: A Technical Overview and Its Application to the Enterprise," *IT Professional*, vol. 7, no. 3, pp. 27-33, May-June 2005.
- [10] C. C. Chang, D. Marculescu, "Design and Analysis of a VLIW DSP Core,"

Proc. Emerging VLSI Technologies and Architectures, March 2006.

- [11] M. Nakai, S. Akui, K. Seno, T. Meguro, T. Seki, T. Kondo, A. Hashiguchi, H. Kawahara, K. Kumano, and M. Shimura, "Dynamic Voltage and Frequency Management for a Low-Power Embedded Microprocessor," *IEEE J. Solid-State Circuits*, vol. 40, no. 1, pp. 28-35, January 2005.
- [12] K. Romer and F. Mattern, "The Design Space of Wireless Sensor Networks," *IEEE Wireless Communications*, vol. 11, no. 6, pp. 54-61, December 2004.
- [13] I.F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless Sensor Networks: A Survey," *IEEE Computer Networks*, vol. 38, no. 4, pp. 393-422, March 2002.
- [14] F. Fallah and M. Pedram, "Standby and Active Leakage Current Control and Minimization in CMOS VLSI Circuits," *IEICE Trans. Electron*, vol. E88-C, no. 4, pp. 509-519, April 2005.
- [15] K. Roy, S. Mukhopadhyay, and H. Mahomoodi-Meimand, "Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits," *Proceedings of the IEEE*, vol. 91, no. 2, pp. 305-327, February 2003.
- [16] K. Roy and S. C. Prasad, *Low-Power CMOS VLSI Circuit Design*. New York: Wiley, 2000, ch. 5, pp. 214-222.
- [17] K. M. Cao, W. C. Lee, W. Liu, X. Jin, P. Su, S. K. Fung, J. X. An, B. Yu, C. Hu, "BSIM4 Gate Leakage Model Including Source-Drain Partition," in *IEDM Technical Digest*, December 2000, pp. 815-818.
- [18] S. Mukhopadhyay, C. Neau, R. T. Cakici, A. Agarwal, C. H. Kim, and K. Roy, "Gate Leakage Reduction for Scaled Devices Using Transistor Stacking," *IEEE Trans. VLSI System*, vol. 11, no. 4, pp. 716-730, August 2003.
- [19] N. Yang, W. K. Henson, and J. Wortman, "A Comparative Study of Gate Direct Tunneling and Drain Leakage Currents in N-MOSFETS with Sub-2100-nm Gate Oxides," *IEEE Trans. Electron Devices*, vol. 47, pp. 1636-1644, August 2000.

- [20] K. Nii, Y. Tsukamoto, T. Yoshizawa, S. Imaoka, Y. Yamagami, T. Suzuki, A. Shibayama, H. Makino, and S. Iwade, "A 90-nm Low-Power 32-kB Embedded SRAM With Gate Leakage Suppression Circuit for Mobile Applications," *IEEE J. Solid-State Circuits*, vol. 39, no. 4, pp. 684-693, April 2004.
- [21] Semiconductor Industry Association, *International Technology Roadmap for Semi-conductors*, 2003 ed., <http://public.itrs.net>.
- [22] H. J. M. Veendrick, "Short-Circuit Dissipation of Static CMOS Circuitry and Its Impact on the Design of Buffer Circuits," *IEEE J. Solid-State Circuits*, vol. sc-19, no. 4, pp. 468-473, August 1984.
- [23] A. P. Chandrakasan, S. Sheng, and R. W. Brodersen, "Low-Power CMOS Digital Design," *IEEE J. Solid-State Circuits*, vol. 27, no. 4, pp. 473-484, April 1992.
- [24] M. Horowitz, T. Indermaur, and R. Gonzalez, "Low-Power Digital Design," in *IEEE Symp. Low Power Electronics Technical Digest*, October 1994, pp. 8-11.
- [25] N. H. E. Weste and D. Harris, *CMOS VLSI Design*, 3rd ed., New York: Addison Wesley, 2005, ch. 2, pp. 98-99.
- [26] J. Chen, L. T. Clark, and Y. Cao, "Ultra-low Voltage Circuit Design in the Presence of Variations," *IEEE Circuit and Devices Magazine*, vol. 21, no. 6, pp. 12-20, November/December 2005.
- [27] T. Kawahara, M. Horiguchi, Y. Kawajiri, G. Kitsukawa, T. Kure, and M. Aoki, "Sub-threshold Current Reduction for Decoded-Driver by Self-Reverse Biasing," *IEEE J. Solid-State Circuits*, vol. 28, no. 11, pp. 1136-1144, November 1993.
- [28] S. Mukhopadhyay, C. Neau, R. T. Cakici, A. Agarwal, C. H. Kim, and K. Roy, "Gate Leakage Reduction for Scaled Devices Using Transistor Stacking," *IEEE Trans. VLSI Systems*, vol. 11, no. 4, pp. 716-730, August 2003.
- [29] L. Wei, Z. Chen, M. Johnson, K. Roy, Y. Ye, and V. De, "Design and Optimization of Dual Threshold Voltage Circuits for Low Voltage Low Power

- Applications," IEEE Trans. VLSI Systems, vol. 7, no. 1, pp. 16-24, March 1999.
- [30] L. Wei, Z. Chen, K. Roy, Y. Ye, and V. De, "Mixed-Vth (MVT) CMOS Circuit Design Methodology for Low Power Applications", in IEEE Proc. DAC, 1999, pp. 430-435.
- [31] J. Y. Lin, L. R. Wang, C. L. Hu, and S. J. Jou, "Mixed-Vth (MVT) CMOS Circuit Design for Low Power Cell Libraries," in IEEE Proc. SOCC, 2007, pp. 181-184.
- [32] S. Shigematsu, S. Mutoh, Y. Matsuya, Y. Tanabe, and J. Yamada, "A 1-V High Speed MTCMOS Circuit Scheme for Power-Down Applications," IEEE J. Solid-State Circuits, vol. 32, No. 6, pp. 861-869, June 1997.
- [33] T. Sakurai, "Perspectives on Power-Aware Electronics," in ISSCC Dig. Tech. Papers, February 2003, pp. 26-29.
- [34] W. Hwang, (2008), "Embedded Memory Design", Lecture/Class, National Chiao Tung University.
- [35] E. Grossar, M. Stucchi, K. Maex, and W. Dehaene, "Read Stability and Write-Ability Analysis of SRAM Cells for Nanometer Technologies," IEEE J. Solid-State Circuits, vol. 41, no. 11, pp. 2577-2588, November 2006.
- [36] B. H. Calhoun and A. P. Chandrakasan, "Static Noise Margin Variation for Sub-threshold SRAM in 65-nm CMOS," IEEE J. Solid-State Circuits, vol. 41, no. 7, pp.1673-1679, July 2006.
- [37] E. Seevinck, F. List, and J. Lohstroh, "Static Noise Margin Analysis of MOS SRAM Cells," IEEE J. Solid-State Circuits, vol. SC-22, no. 5, pp. 748-754, October 1987.
- [38] A. Rychowdhury, S. Mukhopadhyay, and K. Roy, "A Feasibility Study of Sub-threshold SRAM Across Technology Generations," in IEEE Proc. ICCD, October 2005, pp. 417-412.
- [39] R. Heald and P. Wang, "Variability in Sub-100nm SRAM Designs," in

IEEE/ACM Proc. ICCAD, November 2004, pp. 347-352.

- [40] N. Verma and A. P. Chandrakasan, "A 256kb 65nm 8T Subthreshold SRAM Employing Sense-Amplifier Redundancy," *IEEE J. Solid-State Circuits*, vol. 41, no. 1, pp. 141-149, January 2008.
- [41] J. P. Kulkarni, K. Kim, and K. Roy, "A 160 mV Robust Schmitt Trigger Based Subthreshold SRAM," *IEEE J. Solid-State Circuits*, vol. 42, no. 10, pp. 303-2313, October 2007.
- [42] I. J. Chang, J. J. Kim, S. P. Park, and K. Roy, "A 32kb 10T Subthreshold SRAM Array with Bit-Interleaving and Differential Read Scheme in 90nm CMOS," in *ISSCC Dig. Tech. Papers*, February 2008, pp. 388-389.
- [43] L. Chang, D. M. Fried, J. Hergenrother, W. Sleight, R. H. Dennard, R. K. Montoye, L. Sekaric, S. J. McNab, A. W. Topol, C. D. Adams, K. W. Guarini, and W. Haensch, "Stable SRAM Cell Design for the 32nm Node and Beyond," in *Symposium on VLSI Dig. Tech. Papers*, June 2005, pp. 128-129.
- [44] B. H. Calhoun, and A. P. Chandrakasan, "A 256kb 65-nm Sub-threshold SRAM Design for Ultra-Low Voltage Operation," *IEEE J. Solid-State Circuits*, vol. 42, no. 3, pp. 680-688, March 2007.
- [45] T. H. Kim, J. Liu, J. Keane, and C. H. Kim, "A High-Density Subthreshold SRAM with Data-Independent Bitline Leakage and Virtual Ground Replica Scheme," in *ISSCC Dig. Tech. Papers*, February 2007, pp. 330-331.
- [46] J. Y. Yu, W. C. Liao, and C. Y. Lee, "An MT-CDMA Based Wireless Body Area Network for Ubiquitous Healthcare Monitoring," in *IEEE BioCAS*, November 2006.
- [47] J. Y. Yu, C. C. Chung, W. C. Liao, and C. Y. Lee, "A sub-mW Multi-Tone CDMA Baseband Transceiver Chipset for Wireless Body Area Network Applications," in *ISSCC Dig. Tech. Papers*, February 2007, pp. 364-365.
- [48] N. Shibata, M. Watanabe, and Y. Tanabe, "A Current-Sensed High-Speed and Low-Power First-In First-Out Memory Using a Wordline/Bitline-Swapped Dual-Port SRAM Cell," *IEEE J. Solid-State Circuits*, vol. 37, no. 6, pp.

735-750, June 2002.

- [49] G. Gerosa, S. Gary, C. Dietz, P. Dac, K. Hoover, J. Alvarez, H. Sanchez, P. Ippolito, N. Tai, S. Litch, J. Eno, J. Golab, N. Vanderschaaf, J. Kahle, "A 2.2 W, 80 MHz Superscalar RISC Microprocessor," *IEEE J. Solid-State Circuits*, vol. 29, no. 12, pp. 1440-1454, December 1994.
- [50] V. Stojanovic and V. Oklobdzija, "Comparative Analysis of Master-Slave Latches and Flip-Flops for High-Performance and Low-Power Systems," *IEEE J. Solid-State Circuits*, vol. 34, no. 4, pp. 536-548, April 1999.
- [51] H. Partovi, R. Burd, U. Salim, F. Weber, L. DiGregorio, D. Draper, "Flow-through Latch and Edge-Triggered Flip-Flop," in *ISSCC Dig. Tech. Papers*, February 1996, pp. 138-139.
- [52] J. Montanaro, R. T. Witek, K. Anne, A. J. Black, E. M. Cooper, D. W. Dobberpuhl, P. M. Donahue, J. Eno, W. Hoepfner, D. Kruckemyer, T. H. Lee, P. C. M. Lin, L. Madden, D. Murray, M. H. Pearce, S. Santhanam, K. J. Snyder, R. Stehpany, S. C. Thierauf, "A 160-MHz, 32-b, 0.5-W CMOS RISC Microprocessor," *IEEE J. Solid-State Circuits*, vol. 31, no. 11, pp. 1703-1714, November 1996.
- [53] B. Fu and P. Ampadu, "Comparative Analysis of Ultra-Low Voltage Flip-Flops for Energy Efficiency," in *IEEE Proc. ISCAS*, May 2007, pp. 1173-1176.
- [54] Hao-I Yang, Ming-Hung Chang, Ssu-Yun Lai, Hsiang-Fei Wang, and Wei Hwang, "A Low-Power Low-Swing Single-Ended Multi-Port SRAM," in *IEEE VLSI-DAT*, April 2007.
- [55] C. A. Otto, E. Jovanov, and A. Milenkovic, "A WBAN-based System for Health Monitoring at Home," in *IEEE-EMBS Proc. International Summer School and Sym. Medical Devices and Biosensors*, September 2006, pp. 20-23.
- [56] E. Jovanov, A. Milenkovic, C. A. Otto, P. C. de Groen, "A wireless body area network of intelligent motion sensors for computer assisted physical rehabilitation," *J. NeuroEngineering and Rehabilitation*, 2:6, March 2005.
- [57] K. Takeda, H. Ikeda, Y. Hagihara, M. Nomura, and H. Kobatake, "Rede_nition

of Write Margin for Next-Generation SRAM and Write-Margin Monitoring Circuit," in ISSCC Dig. Tech. Papers, February 2005, pp. 478-479.

- [58] Keejong Kim, Hamid Mahmoodi, and Kaushik Roy, "A Low-Power SRAM Using Bit-line Charge-Recycling," IEEE J. Solid-State Circuits, Vol. 43, No. 2, pp.446-459, Feb. 2008.
- [59] Shin-Pao Cheng and Shi-Yu Huang, "A Low-Power SRAM Design Using Quiet-Bitline Architecture," IEEE MTDT, pp. 135-139, Aug. 2005.
- [60] Byung-Do Yang and Lee-Sup Kim, "A Low-Power SRAM Using Hierarchical Bit Line and Local Sense Amplifiers," IEEE J. Solid-State Circuits, Vol. 40, No. 6, pp. 1366-1376, Jun. 2005.
- [61] Koichi Takeda, Hidetoshi Ikeda, Yasuhiko Hagihara, Masahiro, and Hiroyuki Kobatake, "Redefinition of Write Margin for Next-Generation SRAM and Write-Margin Monitoring Circuit," ISSCC, pp. 2602-2611, Feb. 2006.



Vita

PERSONAL INFORMATION

Birth Date: Augst. 12, 1982

Birth Place: Chiayi, Taiwan, R.O.C.

Address: Department of Electronics Engineering
National Chiao Tung University
1001 Ta-Hsueh Road
Hsin-chu, Taiwan 30010, R.O.C.

E-Mail Address: nice.ee94g@nctu.edu.tw

EDUCATION

B.S. [2005] Department of Electrical Engineering, Fu-Jen University.

M.A. [2008] Institute of Electronics, National Chiao-Tung University.

