Design Techniques for Ultra-low Voltage Sub-threshold Circuits and On-chip Reliability Monitoring

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

TAE-HYOUNG KIM

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

OCTOBER 2009

Acknowledgments

I would first like to express my deepest gratitude to my advisor, Professor Chris H. Kim. As a PhD student, it was my honor to meet Professor Chris H. Kim as an academic adviser. I am so grateful to Professor Chris H. Kim for his endless patience, encouragement and support for the last four years. Prof. Kim has taught me numerous things including circuit design, other technical issues and paper writing. It has been so fun to discuss ideas with him even though it became too hot sometimes. Prof. Kim has also shown what it is to be a professor and how to do good research with high ethical and technical standards. Even though I am older than him, his passion and thoroughness in research has never made me feel in that way. Professor Chris H. Kim, I really appreciate your help and advice over the years of my studying.

I also would like to thank my academic committee members, Professor Ramesh Harjani, Professor Sachin Sapatnekar, Professor Kiarash Bazargan, and Professor Antonia Zhai, for their generous but sharp advice on my research and dissertation. Their insightful suggestions helped me to execute my research successfully. Thank you for your great advice and service as committee members.

The two times of my internship at IBM Watson Research Center, NY and the internship at Broadcom, MN have made my graduate school life more exciting. I must really thank many people I met through my internship. Dr. Pong-Fei Lu, Dr. Kent Chuang, Dr. Jae-Joon Kim, Dr. Keith Jenkins, Dr. Saibal Mukhopadhyay, Dr. Shao-Yi Wang, Kevin LeClair, and all others, I appreciate all the feedback and corporation in my research.

All my colleagues in VLSI Research Laboratory have made my graduate school life more than a series of work or study. It is impossible to overstate how valuable it has been to have fellowship and work late together. I have to thank John Keane and Jie Gu who helped me a lot from research to paper writing. All nights we spent together are unforgettable. I am especially grateful to John for his patience and eagerness in correcting my writing. I also would like to thank Jonggab Kil, Hanyong Eom, Jason Liu, Paulo Butzen, Randy Persaud, Raghav Kamath, Nihar Mahtre, Kichul Chun, Pulkit Jain, Wei Zhang, Dong Zhao, Xiaofei Wang, Seunghwang Song, and Ayan Paul for all the supports they gave me and the happy environment they create in the lab.

I want to thank Professor Suki Kim, Professor Chulwoo Kim, Hyun-Geun Byun, Uk-Rae Cho, and all my previous colleagues at Samsung Electronics and ULSI Laboratory in Korea University. Even though they are in Korea, far from United States, their warm and continuous encouragements have sustained me over the years.

My successful life in Minnesota cannot be imagined without all the prayers and spiritual provisions flowing from Paul Mission members in Korea Presbyterian Church of Minnesota. I am also grateful to Pastor Kook-Jin Nam, Pastor Moon-Bong Kim, Pastor Sun-Woo Kang, and Pastor Jun-Hyuk Lim for their prayers and encouragements. Thank God for send those helpers to me.

Although all my achievements are logical, they have been possible due to the illogical love, caring, and support of my wife, Yunha Hwang. I cannot express enough appreciation to my wife. Yunha has always been my supporters providing the love and prayer that sustained me. I cannot forget the test chip implemented with the real help of my wife when I had my knee cap broken, couldn’t sit over fifteen minutes, but had

a project to be finished. Our marriage life has not been like a calm ocean, but has been blessed by God. It is my blessing to love my wife and to be loved by my wife who is so precious to me and my children.

[10]An excellent wife who can find?

She is far more precious than jewels.

[11]The heart of her husband trusts in her, and he will have no lack of gain.

(Proverbs 31:20-11)

I also thank my children, Shia, Aiden, and the to-be third, for their love to me. It was such a happiness to go home and play with my girl and son. Your love recovers me makes all disappointments from work relatively insignificant. You are my blessings and gifts from God. I am thankful to our family in Korea, Canada, and New Zealand. Their prayers and sacrifices have made it possible for me to finish my study successful.

Finally, I thank God for His grace, love, provision, and guidance through my life.

Abstract

Transistor scaling has driven the development of semiconductor industry over the last few decades. However, scaling has also generated numerous challenging problems over technology nodes such as power consumption and circuit variability. Power and circuit variability has continuously increased over technology generations, becoming significant concerns for circuit designers. Various circuit techniques have been developed to address these issues.

Recently, ultra-low power or energy systems are becoming more and more popular. These systems include implantable biomedical electronics, wireless sensor nodes, RFID tag, and many portable electronics. For these applications where minimal energy consumption is the primary design constraint, sub-threshold logic circuits are becoming increasingly accepted since they consume roughly an order of magnitude less power, compared with normal strong-inversion operation.

The first half of this thesis makes several contributions that facilitate reliable sub-threshold circuit design. First, we present a device-size optimization method for sub-threshold circuits utilizing reverse short-channel effect (RSCE) to achieve high drive current, low device capacitance, less sensitivity to random dopant fluctuations, better sub-threshold swing, and improved energy dissipation. Second, we apply the proposed sizing method to SRAMs and propose several circuit techniques for sub-threshold SRAMs that improve SRAM cell stability, writability, bitline sensing margin, and power reduction. By combining these proposed circuit techniques, we demonstrate two fully functional sub-threshold SRAMs in 130nm process technology.

Circuit variability is another big challenging issue in nano-scale technologies. Transistor aging is becoming one of the most pressing sources of circuit variations with each technology node. Transistor aging includes various mechanisms such and hot carrier injection (HCI), bias temperature instability (BTI), and time dependent dielectric breakdown (TDDB). One of the most dominant components among these challenges is NBTI, which is characterized by a positive shift in the absolute value of the PMOS threshold voltage.

In the second half of this thesis, we propose a fully-digital on-chip reliability monitor for high resolution frequency degradation measurements of digital circuits. The proposed technique measures the beat frequency of two ring oscillators; one stressed, the other unstressed; to achieve 50X higher delay sensing resolution than prior techniques. We also show ring oscillator based test structures that can separately measure the NBTI and PBTI degradation effects in digital circuits for high-k metal-gate devices. Finally, we present a test macro for fully-automated statistical measurements of SRAM V_{min} degradation induced by NBTI. An automated test sequence collects V_{min} data for statistical analysis and reduces measurement time. Various test strategies were proposed for V_{min} measurements to identify different SRAM fail metrics such as SNM failure and access time failure.

Content

List of Figures

List of Tables

Chapter 1 Introduction

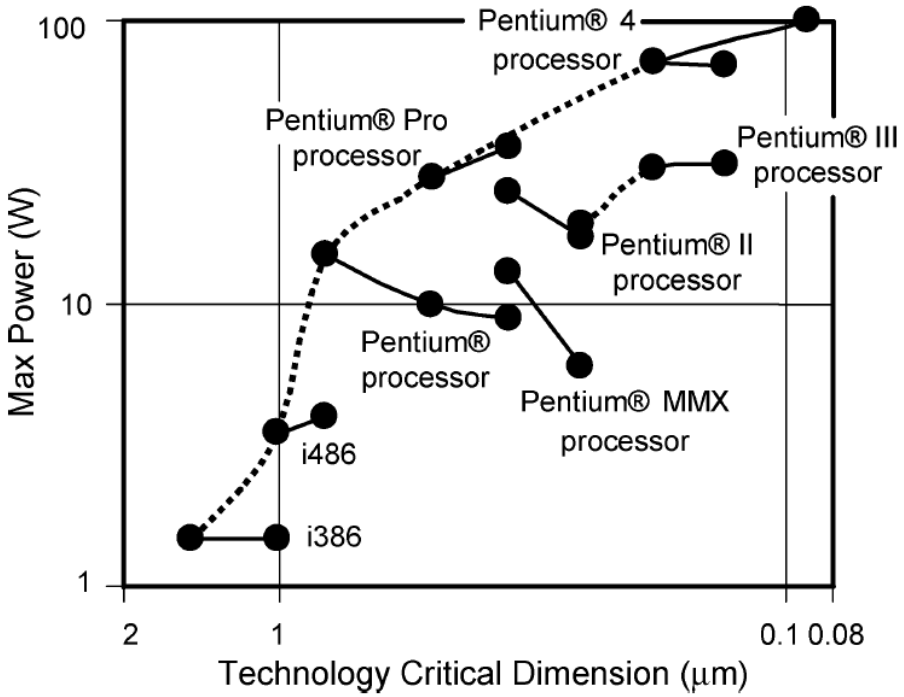
Transistor scaling has driven the development of semiconductor industry over the last few decades. However, scaling has also generated numerous challenging problems over technology nodes. The major problems in circuit design include power and variations.

As device size in integrated circuits (IC) continues to scale toward its fundamental physical limit, both power consumption and power density have kept increasing departing from the ideal scaling trend. Fig. 1.1(a) and (b) show the power consumption and power density on each generation of Intel microprocessor [1]. The ever-increasing power consumption is mainly due to the supply voltage that has not been scaled as transistor dimensions. Since there are trade-offs between power, performance, and device reliability in technology scaling, a technology alone cannot solve the power issue. Various circuit techniques have been developed to address this issue. Circuit schemes such as power gating, clock gating, sleep mode, and dynamic voltage frequency scaling (DVFS) have been popular for reducing power dissipation.

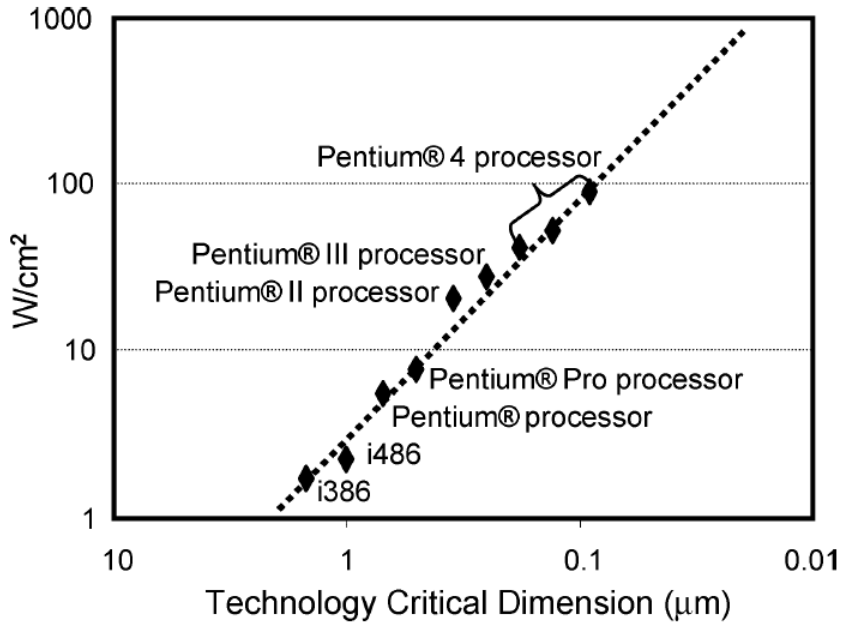
Circuit variability is another big challenging issue to both process and design in nano-scale technologies. Fig. 1.2 illustrates the leakage and operating frequency variations across one thousand Intel microprocessors [2]. Leakage variations of 500 % and operating frequency variations of 30% are observed because of process variations. Various methodologies have been exploited to understand, analyze, and reduce process variations. However, circuit variability due to transistor aging is becoming one of the most pressing sources of circuit variations in recent nano-scale technologies.

Transistor aging includes various mechanisms such and hot carrier injection (HCI), bias temperature instability (BTI), and time dependent dielectric breakdown (TDDB). One of the most dominant components among these challenges is BTI, which is characterized by a positive shift in the absolute value of the threshold voltage.

To address the above two challenging issues, this thesis will investigate (1) sub-threshold circuit optimization technique and (2) design of reliable sub-threshold SRAMs for power and energy efficient systems, and (3) on-chip reliability monitoring circuit for digital circuits and (4) statistical SRAM reliability monitoring.



(a)



(b)

Fig. 1.1 (a) Power consumption and (b) power density of Intel microprocessors over technology generations [1].

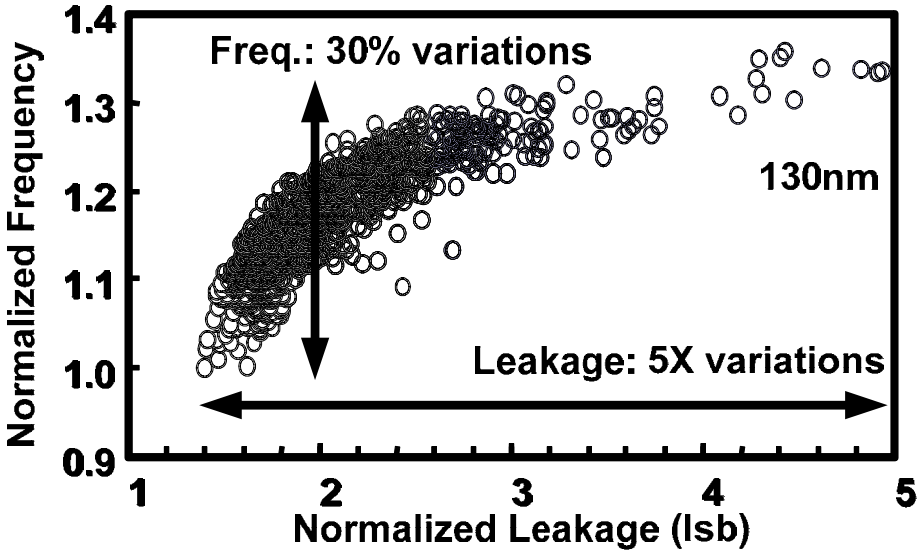


Fig. 1.2 Leakage and operating frequency variations in Intel microprocessors [2].

1.1 Sub-threshold Circuit Design

Recently, ultra-low power or energy systems are becoming more and more popular. These systems include implantable biomedical electronics, wireless sensor nodes, RFID tag, and many portable mobile electronics [3]-[9]. For these applications where minimal energy consumption is the primary design constraint, sub-threshold logic circuits are becoming increasingly accepted since they consume roughly an order of magnitude less power, compared with normal strong-inversion operation. Characteristics of MOS transistors in the sub-threshold region are significantly different from those in the strong-inversion region. The MOS saturation current, which is a near-linear function of the gate and threshold voltages in that region, becomes an exponential function of those values in the sub-threshold regime. This leads to an exponential increase in MOS current variability under Process-Voltage-Temperature (PVT) fluctuations.

A significant amount of research has been done dealing with sub-threshold circuits. Soeleman et al. analyzed various logic styles for sub-threshold operation [3]. The impact of PVT variations on sub-threshold circuits was investigated in [4] and [5]. Circuits such as analog voltage references, sub-threshold SRAMs, tiny-XOR circuits, and adaptive filters for hearing aid applications have been demonstrated [6]-[10]. New transistor scaling trends specifically for sub-threshold circuits have been suggested in [11].

1.2 Circuit Reliability

There are many sources generating circuit variability. Variations occur while transistors are under fabrication where short channel effect, random dopant fluctuation (RDF), and line edge roughness are major sources. Circuit variations also occur after fabrication steps due to voltage variations in power network, temperature variations, and transistor aging effect. Among these, the transistor aging is becoming a significant issue since it is getting larger and larger over technology scaling and leads to circuit reliability issues.

As CMOS process technology continues to follow an aggressive scaling roadmap, designing reliable circuits has become ever-more challenging with each technology node. Reliability issues such as Bias Temperature Instability (BTI) [12], [13], [14], [15], Hot Carrier Injection (HCI) [16], [17], and Time Dependent Dielectric Breakdown (TDDB) [18], [19] have become more prevalent as the electrical field continues to increase in nano-scale CMOS devices. BTI is the device aging occurring at the channel-oxide interface due to the interface traps when a transistor is held in the 'on' state. It is represented by the positive shift in the absolute value of threshold voltage. HCI is an aging happening when a transistor is switching. It usually happens in the channel near drain side where high electric field exists. Finally, TDDB is generated inside the oxide layer with high electric field. This is more like a catastrophic failure because defects pile up and a short circuit forms. Among these, BTI is becoming one of the dominant aging mechanisms, which is the other focus this thesis [12], [13], [14], [15].

1.3 Summary of Thesis Contributions

This thesis makes several contributions that facilitate reliable sub-threshold circuit design. First, we present a device-size optimization method for sub-threshold circuits utilizing reverse short-channel effect (RSCE) to achieve high drive current, low device capacitance, less sensitivity to random dopant fluctuations, better sub-threshold swing, and improved energy dissipation. Second, we apply the proposed sizing method to SRAMs and propose several circuit techniques for sub-threshold SRAMs that improve SRAM cell stability, writability, bitline sensing margin, and power reduction. By combining these proposed circuit techniques, we demonstrate two fully functional sub-threshold SRAMs in 130nm process technology. These works have been published in [20], [21], [22].

The second half of this thesis will research on-chip circuit reliability monitoring techniques. First, we proposed a fully-digital on-chip reliability monitor for high resolution frequency degradation measurements of digital circuits. The proposed technique measures the beat frequency of two ring oscillators; one stressed, the other unstressed; to achieve 50X higher delay sensing resolution than prior techniques. We also show ring oscillator based test structures that can separately measure the NBTI and PBTI degradation effects in digital circuits for high-k metal-gate devices. Finally, we present a test macro for fully-automated statistical measurements of SRAM V_{min} degradation induced by NBTI. An automated test sequence collects V_{min} data for statistical analysis and reduces measurement time. Various test strategies were proposed for V_{min} measurements to identify different SRAM fail metrics such as SNM failure and access time failure. These works have been published in [23][24].

The organization of this thesis is as follows: Chapter 2 describes the sub-threshold circuit optimization methodology utilizing Reverse Short Channel Effect (RSCE). Chapter 3 presents various circuits techniques for sub-threshold SRAMs. Two sub0threshold SRAMs are described. Chapter 4 discusses several on-chip reliability monitoring methods for digital circuits, high-k process technologies, and SRAMs. Chapter 5 concludes this thesis.

Chapter 2 Device-Size Optimization for Sub-threshold Circuits

2.1 Introduction

Short channel devices have been optimized for regular super-threshold circuits to meet various device objectives such as high mobility, reduced Drain-Induced-Barrier-Lowering (DIBL), low leakage current, and minimal V_{th} roll-off. However, a transistor that is optimized for super-threshold logic may not be optimal for achieving high performance and low power in the sub-threshold region where effects such as DIBL, V_{th} roll-off, and electron/hole tunneling are much less significant. For example, the reduced DIBL effect in the sub-threshold region, due to the low drain voltages, can eliminate the need for high doping in the channel which was traditionally used to overcome the Short Channel Effect (SCE) [25]. Although it would be ideal to have a dedicated process technology optimized for sub-threshold circuits, mainstream CMOS technology will continue to scale aiming at optimal performance in conventional super-threshold circuits. In order to design optimal sub-threshold circuits using CMOS devices that are targeted for super-threshold operation, it is crucial to develop techniques that can utilize the side effects that appear in this new regime. The main contribution of this research is utilizing one such mechanism-the pronounced Reverse Short Channel Effect (RSCE) to achieve optimal performance in sub-threshold circuits.

SCE (or V_{th} roll-off) is an undesirable phenomenon in short channel devices where V_{th} decreases as the channel length is reduced. Variation in critical device dimensions translate into a larger variation in the threshold voltage as SCE worsens with increasing DIBL [26]. Typically, non-uniform HALO doping is used to mitigate this problem by making the depletion widths narrow and hence reducing the DIBL effect [25]. As a byproduct of HALO, a short channel device shows RSCE behavior where the V_{th} decreases as the channel length is increased [27][28]. In sub-threshold circuits, the SCE mechanism is not as strong as in super-threshold circuits because the drain-to-source voltage is very small. On the other hand, RSCE is still significant enough to affect the sub-threshold performance. Moreover, current becomes an exponential function of V_{th} in this regime which makes it possible to use longer channel length devices that utilize RSCE for improving drive current. Unlike the case in super-threshold circuits, using a longer channel length in sub-threshold does not have a significant impact on the load capacitance. This is due to the reduced depletion capacitance under the gate.

2.2 Gate-Sizing Considerations

Conventional super-threshold logics require special modifications in order to achieve optimal performance and power consumption in sub-threshold operation. For example, the PMOS to NMOS width ratio (PN ratio) and stacked device sizing need to be reevaluated for sub-threshold operating voltages [29]. The optimal PN ratio for equal current drivability of PMOS and NMOS is roughly 2.5 in super-threshold logics, which comes from the mobility and threshold voltage difference. This ratio changes in the sub-threshold region because the weak-inversion current is an exponential function of threshold voltage, which differs in PMOS and NMOS devices. The weak-inversion current is also a function of the sub-threshold slope and is significantly affected by other secondary effects such as the narrow width effect, SCE, and RSCE. Fig. 2.1 shows the optimal PN ratio at different supply voltages. The significant reduction in the optimal PN ratio with a lower supply voltage can be attributed to the difference in V_{th} and sub-threshold slope. The mobility difference between electrons and holes remains the same as in the super-threshold region. Selection of the proper effective width of stacked transistors is also crucial for achieving optimal performance. The effective width of a transistor in a stack of n devices is roughly 1/n in the strong-inversion region. This means that in order for an n-stack to conduct the same amount of current as a single transistor, the devices in the stack must each be sized up by a factor of n. Simulation results indicate that stacks need to be sized up by a larger amount in the sub-threshold region due to the weak stack currents. For example, a single unit NMOS transistor is equivalent to a two-stack with transistor widths of 2.259 at 0.2V, 2.413 at 0.3V, and 1.6 at 1.2V in the 0.13μm process technology used

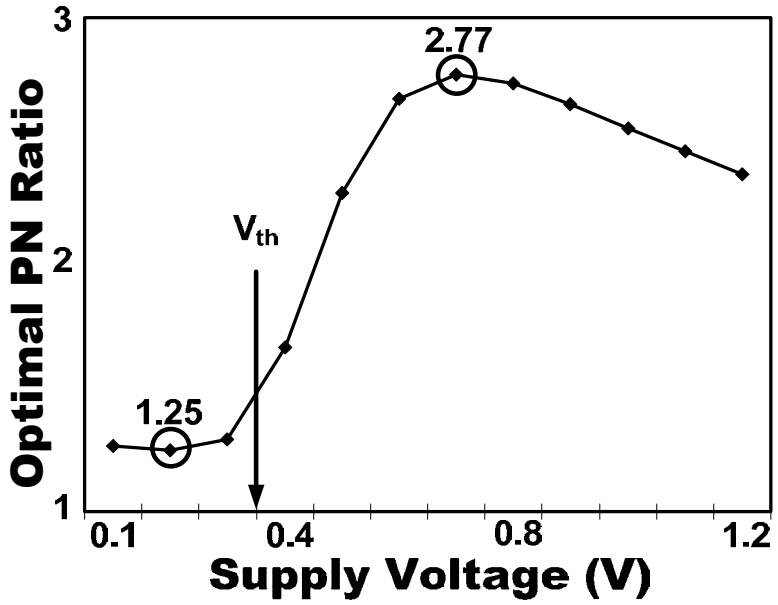


Fig. 2.1 PMOS to NMOS ratio as a function of supply voltage.

here. Consequently, the sizing methods that were used to obtain maximum performance in the super-threshold region must be reformulated in the sub-threshold region due to these different device characteristics.

Previous sizing methods for sub-threshold logics were based on the traditional assumption that the minimum channel length is still optimal for speed and power. This is true in the super-threshold region, but it does not hold true in sub-threshold logic since a device with a longer channel length and a fixed channel width can have higher on-current due to RSCE. The PN ratio will also have to be adjusted when we change the device channel lengths due to its dependency on NMOS and PMOS threshold voltages, which shift with those lengths. Therefore, a new sizing method suitable for sub-threshold circuits which considers the impact of RSCE on drive current, device capacitance, and sub-threshold slope is indispensable.

2.3 Transistor-Sizing Method Utilizing Reverse Short-Channel Effect

2.3.1 Reverse Short-Channel Effect (RSCE) Overview

Fig. 2.2 (top) shows the threshold voltages as a function of channel length at VDD=1.2V and VDD=0.2V. In the super-threshold region (1.2V), a strong V_{th} roll-off behavior is observed at the minimum channel length due to the high DIBL effect (point A in Fig. 2.2). To compensate the worsening V_{th} roll-off caused by DIBL in small dimensions, non-uniform p+ doping in the source-body and drain-body boundaries, called HALO implants, are used. These regions reduce the amount of control the drain has over the channel by making the depletion layer width narrow. HALO implants can also suppress the body punchthrough [25][30]. However, as a byproduct of using those implants, the threshold voltage decreases as the channel length increases. This phenomenon is known as the RSCE [27][28]. The larger distance between the highly doped HALO regions in longer channel devices decreases the surface doping level across the channel, which in turn causes the threshold voltage to decrease.

Fig. 2.3 (top) illustrates this trend by showing the effective surface doping in the longitudinal direction. RSCE becomes more significant with process scaling due to the higher HALO doping required to negate the aggravating V_{th} roll-off as shown in Fig. 2.2 (bottom). The combination of SCE and RSCE causes the V_{th} to peak at a channel length slightly longer than the minimum value in super-threshold devices. RSCE is not a major concern in conventional super-threshold designs since SCE is dominant in

minimum channel length devices in that region. However, in the sub-threshold region, only the RSCE effect is present due to the significantly reduced DIBL [25]. This causes the V_{th} to decrease monotonically, and operating current to increase exponentially, with longer channel length.

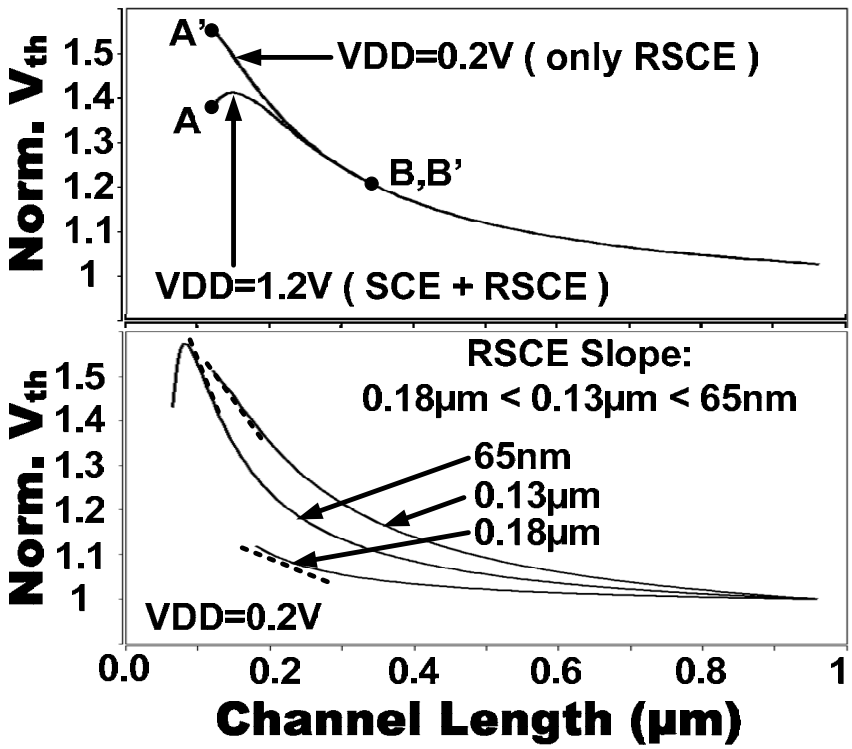


Fig. 2.2 Dependency of normalized V_{th} on channel length for VDD=1.2V and VDD=0.2V.

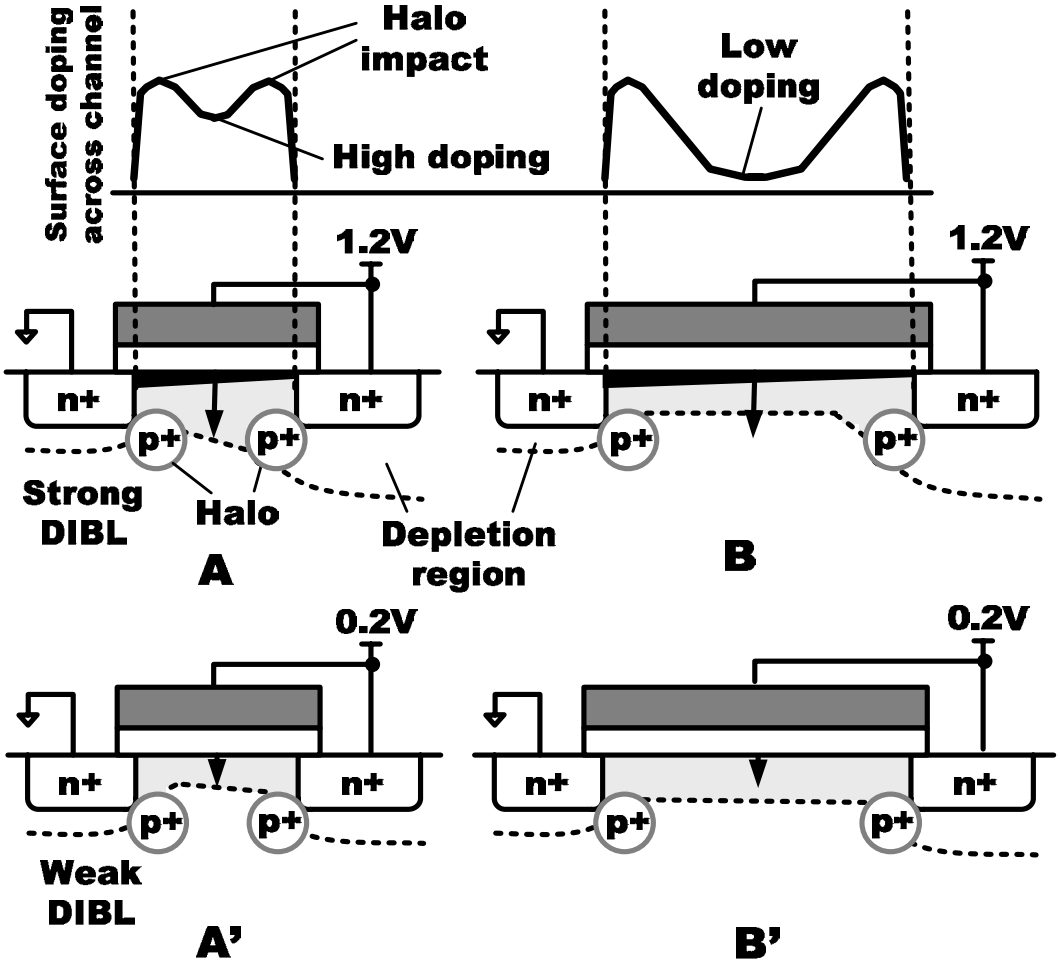
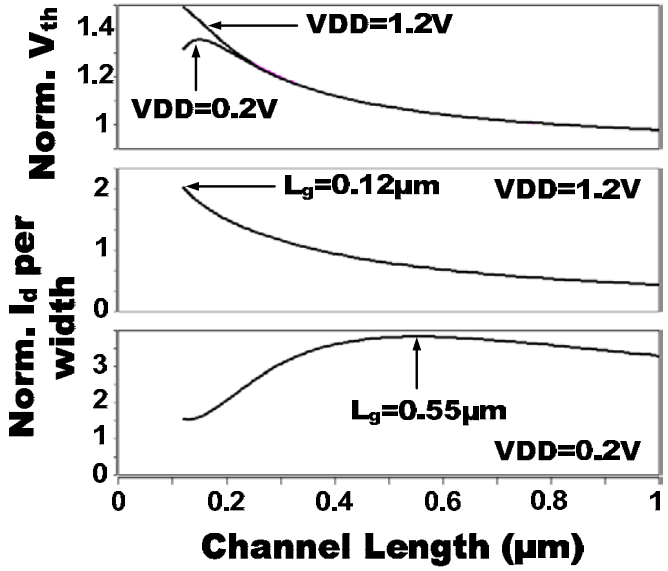


Fig. 2.3 Device cross sections corresponding to A, A', B, and B' in Fig. 2. 2. Surface doping across channel is shown to illustrate the RSCE.

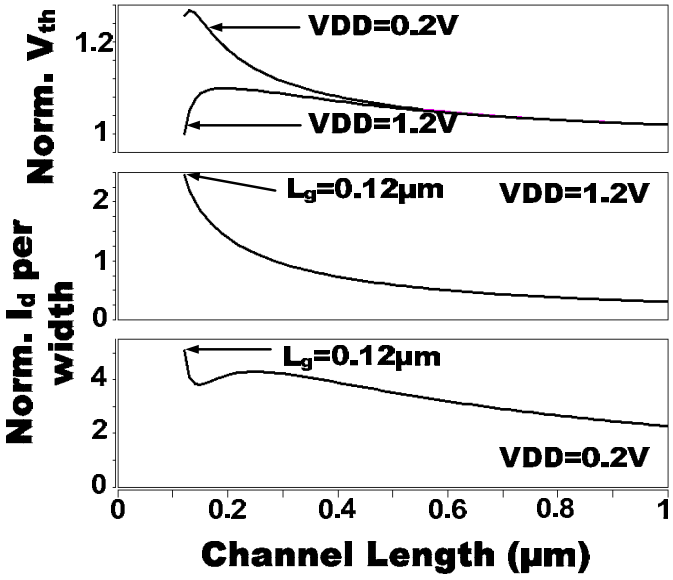
2.3.2 Optimal Channel Length fir Maximum Current Per Width

As the V_{th} behavior changes significantly in the sub-threshold region, the optimal channel length yielding maximum current-per-width changes accordingly. This is illustrated in Fig. 2.4, where V_{th} and current-per-width are plotted versus channel length in the sub-threshold and super-threshold regions. Maximum current-per-width is obtained at the minimum channel length (0.12μm) for VDD=1.2V because the

effect of maximized W/L is stronger than that of reduced threshold voltage on the current. However, the optimal channel length for an NMOS at VDD=0.2V increases to 0.55µm since the lower V_{th} caused by RSCE provides an exponential increase in current (see Fig. 2.4 (a)). Current is also proportional to W/L which makes it eventually decrease at channel lengths longer than the optimal. In this process technology, only NMOS device lengths are adjusted to utilize RSCE. The NMOS threshold voltage is reduced by 45% when changing the channel length from 0.12µm to 0.55µm. However, RSCE in PMOS devices is not strong enough in the given technology to provide current gain by increasing channel length as can be seen in Fig. 2.4 (b). The PMOS threshold voltage is reduced by only 23% which is around 50% of the NMOS threshold voltage change when applying the same channel length change. The effectiveness of our proposed sizing scheme depends on how strong the RSCE is. PMOS devices can also utilize the pronounced RSCE in future scaled process technologies where stronger RSCE effect is observed, as shown in Fig. 2.2 (bottom) [31].



(a)



(b)

Fig. 2.4 Dependency of normalized V_{th} and current-per-width on channel length: (a) NMOS, (b) PMOS.

Here we will derive the optimal channel length for maximum current-per-width in the sub-threshold region. The RSCE-affected threshold voltage can be expressed as

$$V_{th} = V_{th0} + K_1\left(\sqrt{1+\frac{K_2}{L_{eff}}} - 1\right)\sqrt{\Phi_S} \tag{1}$$

where V_{th0} is the zero-bias threshold voltage of a long channel device, K_1 and K_2 are technology parameters that are positive numbers, L_{eff} is the effective channel length, and Φ_s is the surface potential. DIBL effect is omitted because its effect is negligible in the sub-threshold region. Body effect is ignored for simplicity. The optimal channel length can be obtained by taking the derivative of the current equation.

$$I_D = I_{D0}\frac{W}{L_{eff}} e^{\frac{V_{GS}-V_{th}}{mV_t}}\left(1 - e^{\frac{-V_{DS}}{V_t}}\right) \tag{2}$$

$$\frac{\partial I_D}{\partial L_{eff}} = 0 \tag{3}$$

Here, m is a technology parameter and V_t is the thermal voltage. By solving equation (3), we can derive the optimal channel length for maximum current-per-width.

$$L_{eff}^{\ 2} + K_2 L_{eff} + K_3 = 0 \tag{4}$$

$$L_{eff} = \frac{-K_2 + \sqrt{K_2^{\ 2} - 4K_3}}{2} \tag{5}$$

$$K_3 = -\frac{K_1^{\ 2}\Phi_S}{m^2 V_t^{\ 2}} K_2 \tag{6}$$

The optimal channel length calculated using the analytical expression in (5) is 0.58μm which is very close to 0.55μm from simulation. We can also compare the current at the optimal channel length given by (5) with that at minimum channel length for validation. The maximum current-per-width is 2.5X larger than that at the minimum channel length in this process technology. However, using a longer channel length can have a negative impact on device capacitance which can affect the CV/I delay. In the following section, we derive the optimal channel length for maximum performance considering the RSCE and device capacitance behavior in the sub-threshold region.

2.3.3 Optimal Channel Length for Maximum Performance

We have shown that for sub-threshold circuits, the maximum current can be obtained at a channel length significantly longer than the minimum defined by the technology node. This phenomenon is attributed to the effect of RSCE on threshold voltage and current. Another factor to consider when increasing the channel length for optimal sub-threshold sizing is the increase in device capacitance. Delay and power consumption increases linearly with capacitance.

Fig. 2.5 shows the different components of device capacitance in the sub-threshold region. Each component can be described as follows:

$$C_{DEP} = \frac{\varepsilon_{si}}{W_{DEP}}, \qquad C_{OX} = \frac{\varepsilon_{ox}}{t_{ox}} \tag{7}$$

$$C_{GD} = C_{GS} = WC_{OV} \tag{8}$$

$$C_J = WC_j + (2W + L_j)C_{jsw} \tag{9}$$

where W_{DEP} is the depletion width, t_{ox} is the oxide thickness, W is the device width, C_{OV} is the overlap capacitance per width, C_j is the junction capacitance per width, L_j is junction length, and C_{jsw} is the junction sidewall capacitance.

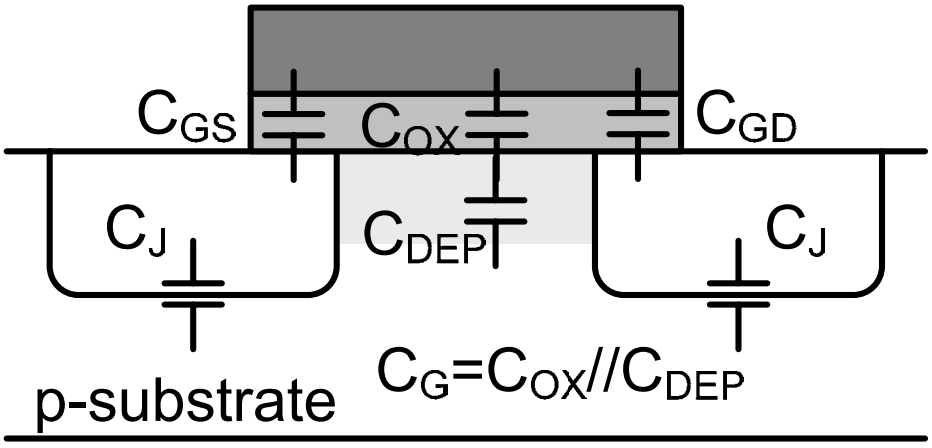


Fig. 2.5 Capacitance in sub-threshold MOS device.

In order to illustrate the effectiveness of increasing the channel length, the capacitances of a transistor having a constant current is plotted versus channel length in Fig. 2.6. Note that the device width can be reduced as the channel length is increased since RSCE lowers the V_{th} and exponentially increases the device current. This was not the case for super-threshold circuits where the decrease in W/L had a larger impact on current than the reduction in V_{th} due to RSCE. Increasing the channel length alone has no effect on junction capacitance (C_J) because C_J is only proportional to device width. However, since the device width is reduced simultaneously for constant current, the junction capacitance also goes down with a longer channel length as shown in Fig. 2.6. Simulation results showed that the junction capacitance can be reduced by 50%. The increase in gate capacitance (C_G) is moderate between channel lengths of 0.12µm and 0.36µm for two reasons. First, the reduction in width makes the

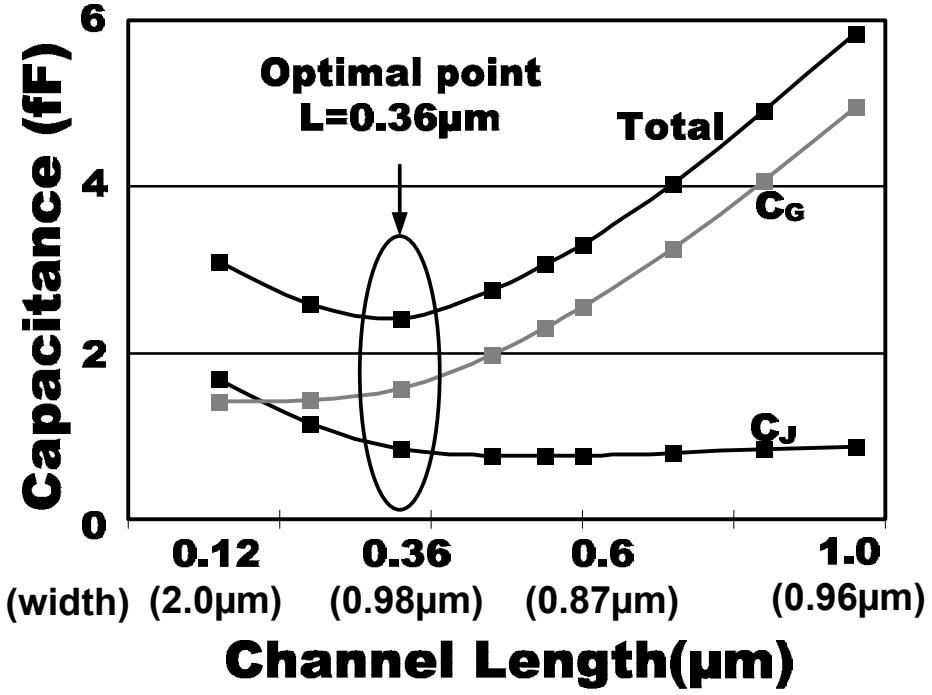


Fig. 2.6 Capacitance vs. channel length for constant current.

increase in gate area smaller. In this design, the gate area is increased by 50%. Second, the RSCE associated with longer channel length makes the depletion capacitance (C_{DEP}) smaller since the depletion layer width under gate increases as channel length increases, which is shown in Fig. 2.3 (bottom). At channel lengths longer than 0.36μm however, C_G increases rapidly since the RSCE becomes weaker, and gate area is increased to drive the same current. As a result, there exists a minimum point in total capacitance for iso-current at a channel length of 0.36μm. By using this optimal channel length, we can reduce delay and power consumption and therefore obtain maximum performance in sub-threshold circuits.

2.3.4 Effect of Supply Voltage on Optimal Channel Length

The drive current in the sub-threshold region is an exponential function of the supply voltage and threshold voltage. This is not the case in the moderate and strong-inversion regions where the current is governed by different equations. As can be seen in equation (5), optimal channel length for maximum current-per-width is independent of supply voltage in the sub-threshold region. It is a function only of process parameters. However, in the strong-inversion region, the gain in current obtained by utilizing the RSCE becomes smaller due to the reduced impact of V_{th} on current and the larger increase in device capacitance. As a result, the minimum channel length becomes the optimal channel length. This can be shown analytically as follows.

Current in strong-inversion can be modeled as:

$$I_{D\_STRONG} = K\frac{W}{L_{eff}}(V_{GS} - V_{th})^{\alpha} \tag{10}$$

where α and K are technology parameters, V_{GS} is the gate-to-source voltage, and V_{th} is the threshold voltage. Note that V_{th} is also a function of L_{eff} (see equation (1)). In short channel devices, α is between 1 and 1.5. Using equation (1), the derivative of the device current with respect to the channel length can be expressed as

$$\begin{aligned}\frac{\partial I_{D\_STRONG}}{\partial L_{eff}} &= \frac{\partial\left[K\frac{W}{L_{eff}}(V_{GS} - V_{th})^{\alpha}\right]}{\partial L_{eff}} \\ &= KW\frac{\frac{d(V_{gs} - V_{th})^{\alpha}}{dL_{eff}}L_{eff} - (V_{gs} - V_{th})^{\alpha}\frac{dL_{eff}}{dL_{eff}}}{{L_{eff}}^{2}} \\ &= C_1\left[-\frac{C_2 L_{eff}}{\sqrt{1 + K_2 / L_{eff}}} - (V_{gs} - V_{th})\right]\end{aligned} \tag{11}$$

where C_1 and C_2 are positive constants. K_2 was given in equation (1) as a positive constant and since V_{gs}-V_{th} is also positive, the following inequality holds true.

$$\frac{\partial I_{D\_STRONG}}{\partial L_{eff}} < 0 \tag{12}$$

Therefore, I_{D_STRONG} decreases monotonically as the channel length increases for a fixed channel width. Since the SCE was omitted in equation (1), it is important to note that the above derivation is only applicable for longer channel lengths where RSCE is dominant over SCE. However, it is trivial to show that a shorter channel length gives a higher current for channel lengths where SCE is stronger than RSCE; a

lower Vth and a higher W/L ratio at a shorter channel length together increases the device current. Hence, we can conclude that even with a strong RSCE, the minimum channel length is optimal for maximum current in the super-threshold region.

Fig. 2.7 shows simulation results on the optimal channel length for different supply voltages. Three different regions can be observed. In the sub-threshold and strong-inversion regions, the optimal channel length does not depend strongly on the supply voltage as expected from our derivations. Therefore, we can use the optimal channel length obtained from equation (5) which is independent of supply voltage in the deep sub-threshold region. In the moderate-inversion region however, the optimal channel length varies depending on the supply voltage. For the 0.13μm process technology used in this work, the optimal channel lengths in the sub-threshold and super-threshold region are 0.55μm and 0.12μm, respectively.

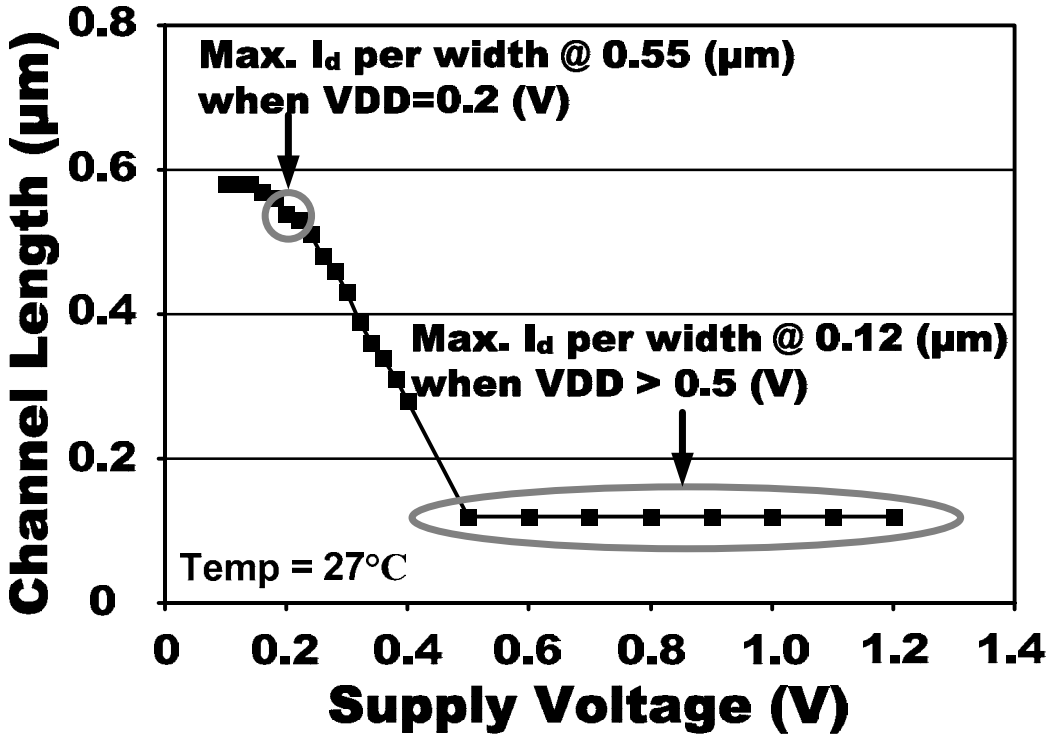
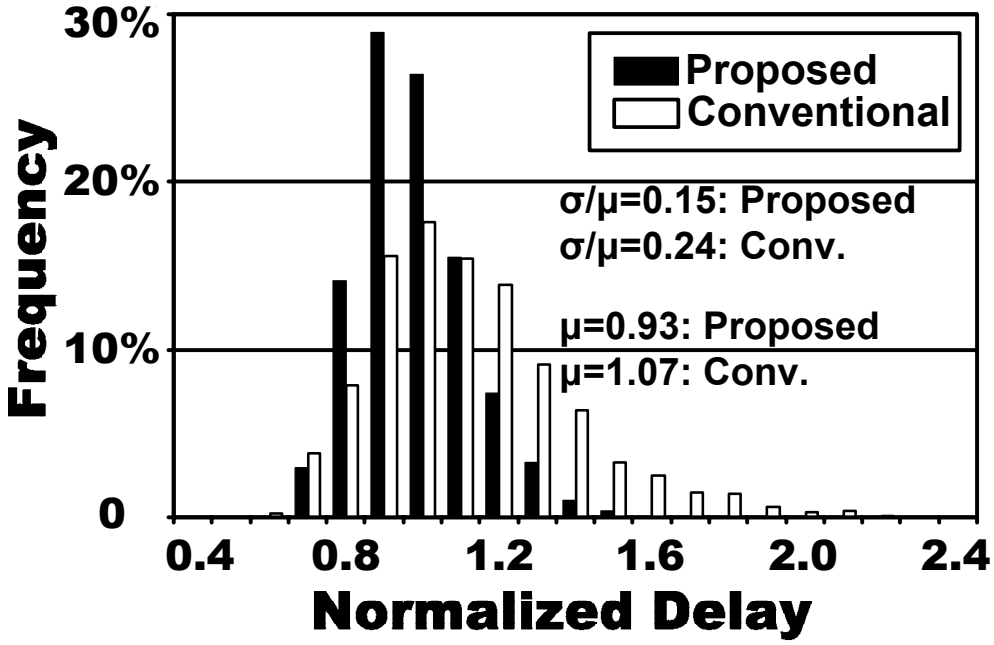


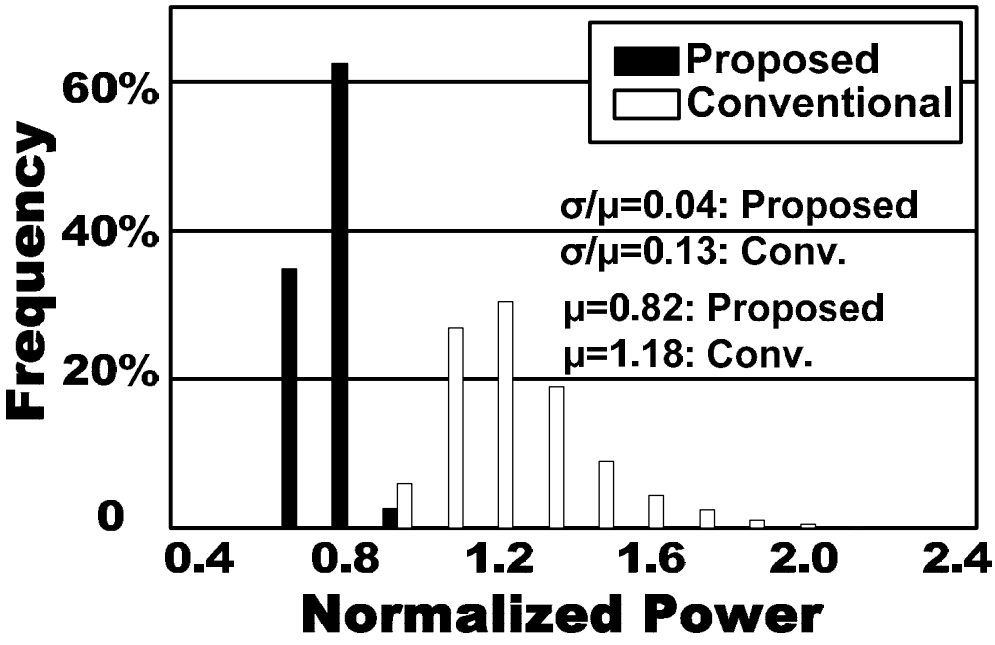
Fig. 2.7 The effect of supply voltage on the channel length providing maximum current per width.

2.3.5 Impact of Process Variation

Random Dopant Fluctuation (RDF) causes random parameter mismatches even between devices with identical layout in close proximity. The standard deviation (σ) of the threshold voltage distribution caused by RDF is proportional to $(WL)^{-1/2}$. Using the proposed sizing method, the sample gate area for optimal performance increases from $0.24\mu m^2$ ($=2\mu m \times 0.12\mu m$) to $0.35\mu m^2$ ($=0.98\mu m \times 0.36\mu m$) with identical current drivability and reduced device capacitance as shown in Fig. 2.6. This interesting characteristic leads to less threshold voltage variations for the proposed sizing scheme. To verify this, statistical studies were carried out using Monte Carlo simulation. Fig. 2.8 shows the delay and power consumption distribution of a static inverter chain designed using the proposed and conventional scheme. The 4 stage inverter chains are simulated at room temperature with the input switching at 100MHz. A supply voltage of 0.2V was used. Delays of the third stage inverters were measured. Power consumption was measured for the cycle time of the inverter chain implemented using the conventional sizing scheme and includes both the active and static leakage power components. The σ/μ ratios of the delay and power consumption distributions are reduced by 37.5% and 70%, respectively, resulting in a squeezed distribution for the proposed sizing scheme. Simulation results show a 13% improvement in average delay while simultaneously achieving a 31% reduction in average power dissipation.



(a)



(b)

Fig. 2.8 Statistical comparison of a static inverter chain: (a) delay distribution, (b) power consumption distribution.

2.3.6 Sub-threshold Swing and Ion-to-Ioff Ratio

Sub-threshold swing *(S)* is a critical parameter that determines the relationship between sub-threshold current and the gate voltage. It is defined as the amount of V_{GS} required to change the sub-threshold current by an order of magnitude. *S* has generally been considered a process-dependent parameter. A small *S* is preferred in order to achieve higher on-current for a given off-current value. Our proposed sizing scheme utilizes a longer channel length which reduces *S,* and therefore improves the Ion-to-Ioff ratio.

The sub-threshold swing can be represented as

$$S = m\frac{kT}{q}\ln 10 \quad (mV/dec) \tag{13}$$

where

$$m = 1 + \frac{C_{DEP}}{C_{OX}}, \qquad C_{OX} = \frac{\varepsilon_{ox}}{t_{ox}}, \qquad C_{DEP} = \frac{\varepsilon_{si}}{W_{DEP}} \tag{14}$$

and *kT/q* is the thermal voltage.

As we explained in section III-C, RSCE increases the depletion width underneath the channel and lowers the depletion capacitance, C_{DEP}, for long channel devices. This alters the value of *m* in (14) and reduces *S*. I-V characteristics of a conventional minimum channel device and an optimal longer channel device are shown in Fig. 2.9. The sub-threshold swing of the proposed method is 71mV/dec which is 16mV lower than that of the conventional minimum channel device. The improved sub-threshold slope reduces the off-current by 30% for the same on-current.

Improved Ion-to-Ioff ratio can be achieved by the reduced *S*. Ion-to-Ioff ratio is a critical factor in sub-threshold digital circuits and sub-threshold SRAMs [7]. The inherently small Ion-to-Ioff ratio limits the number of transistors connected per node. Fig. 2.10 shows the Ion-to-Ioff ratio for the conventional and proposed scheme at different supply voltages. At 0.2V, the Ion-to-Ioff ratio was 484 for the proposed scheme, a 2.5X improvement over the conventional minimum channel device.

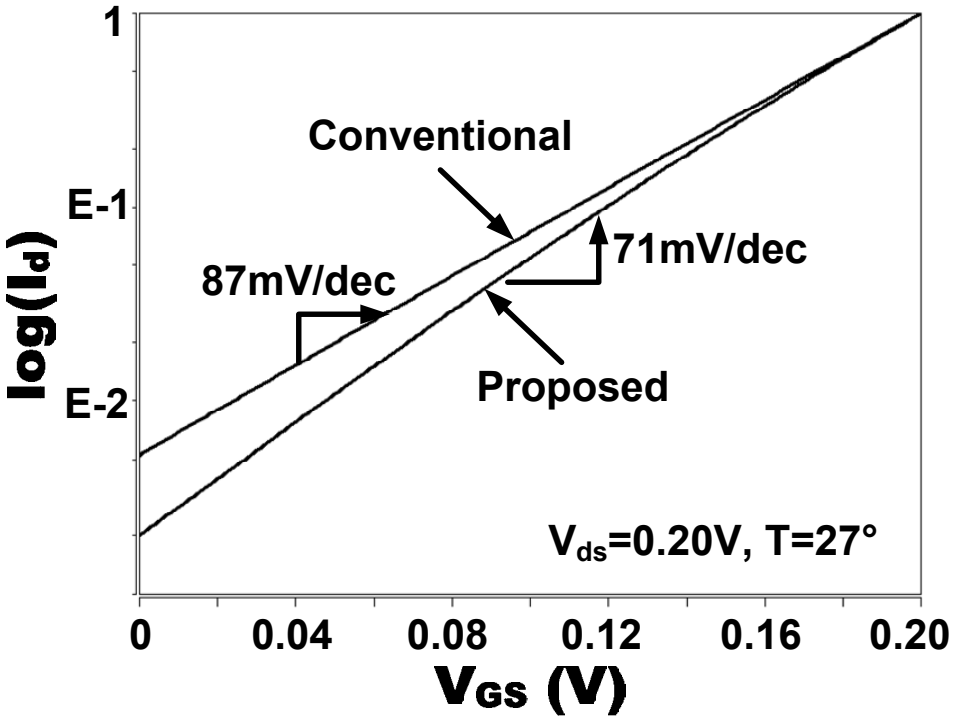


Fig. 2.9 Sub-threshold swing comparison for conventional and proposed sizing scheme.

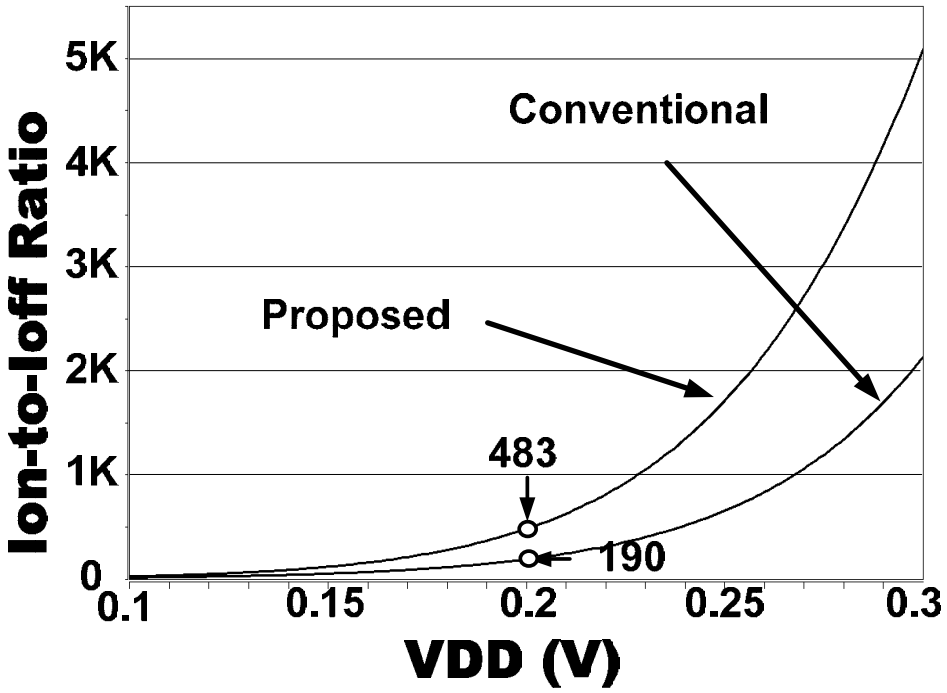


Fig. 2.10 Ion-to-Ioff ratio as a function of supply voltage.

2.3.7 Improvement in Delay, Power, and Energy

The proposed scheme offers a simultaneous improvement in circuit delay and power consumption which leads to a significant reduction in energy dissipation which is the product of delay and power consumption. Energy consumption is a critical metric in applications such as portable devices, medical instruments, and wireless sensor networks where sub-threshold circuits can be widely applied. In the super-threshold region, using a larger device for reducing the circuit delay always causes the power consumption to increase due to the increase in gate and junction capacitance. The energy dissipation would increase accordingly. The proposed scheme on the other hand reduces the junction capacitance without deteriorating the performance because a smaller device width can be used for the same current drivability. Since we

obtain a reduction in both delay and power consumption using the proposed scheme, a large improvement in energy consumption is achieved. Energy consumed in the sub-threshold region can be expressed as

$$
\begin{aligned}
E_{SWITCHING} &= \alpha \cdot C \cdot VDD^2 \\
E_{LEAK} &= VDD \cdot I_{LEAK} \cdot t_d \\
&= \beta \cdot VDD \cdot C e^{\frac{-V_{th}}{mVt}} \cdot \frac{VDD}{e^{\frac{VDD-V_{th}}{mVt}}} \\
&= \beta \cdot C \cdot VDD^2 e^{\frac{-VDD}{mVt}} \\
\therefore E_{TOTAL} &= E_{SWITCHING} + E_{LEAK} \\
&= \alpha \cdot C \cdot VDD^2 + \beta \cdot C \cdot VDD^2 \cdot e^{\frac{-VDD}{mVt}}
\end{aligned}
\tag{15}
$$

where α is the activity factor, β is a technology-related constant, C is the switching capacitance, VDD is the supply voltage, I_{leak} is the leakage current, and t_d is the propagation delay. At a fixed supply voltage, the total energy is a function of device capacitance. The proposed scheme reduces both C and t_d in equation (15). Simulations using ISCAS benchmark circuits show an energy reduction as large as 41.2%.

2.4 Test Chip Implementation and Experimental Results

A delay chain composed of inverters, 2-input NANDs and 2-input NORs was used in simulations to verify the effectiveness of the proposed sizing scheme. For accurate SPICE simulations, a post layout netlist was extracted including the RC parasitics. The layout of the sample delay chain is shown in Fig. 2.11. The conventionally sized gates have a taller layout than the gates sized using the proposed scheme. This is due to the "fat" devices in the proposed scheme have longer channel lengths and narrower widths. As mentioned in section III-B, minimum channel length is used for the PMOS devices since the strength of RSCE was not pronounced enough to provide current gain at a longer channel length for those transistors in this particular technology. In future technology nodes where RSCE is severe in both PMOS and NMOS devices [31], our proposed sizing scheme can be applied in general sub-threshold circuit design. The layout area of the proposed scheme is 18% smaller compared to that using conventional sizing in this delay chain. Fig. 2.12 shows simulated waveforms of the simple logic chain using corner parameters. It can be seen that the delay variation of the proposed scheme is 38.7% smaller than that of the conventional method. The reduction in power dissipation is shown in Fig. 2.13 for each process corner. The power savings range from 10% to 39%, mainly depending on the current of the conventional scheme which is sensitive to process variations.

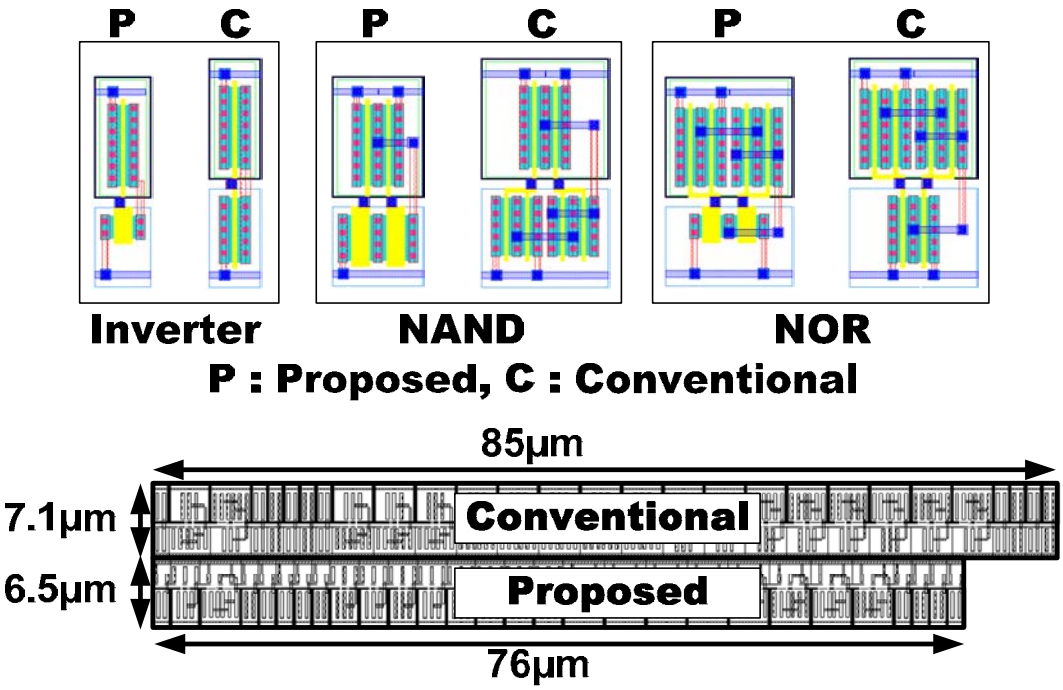


Fig. 2.11 Layout comparison for basic logic gates and sample delay chain.

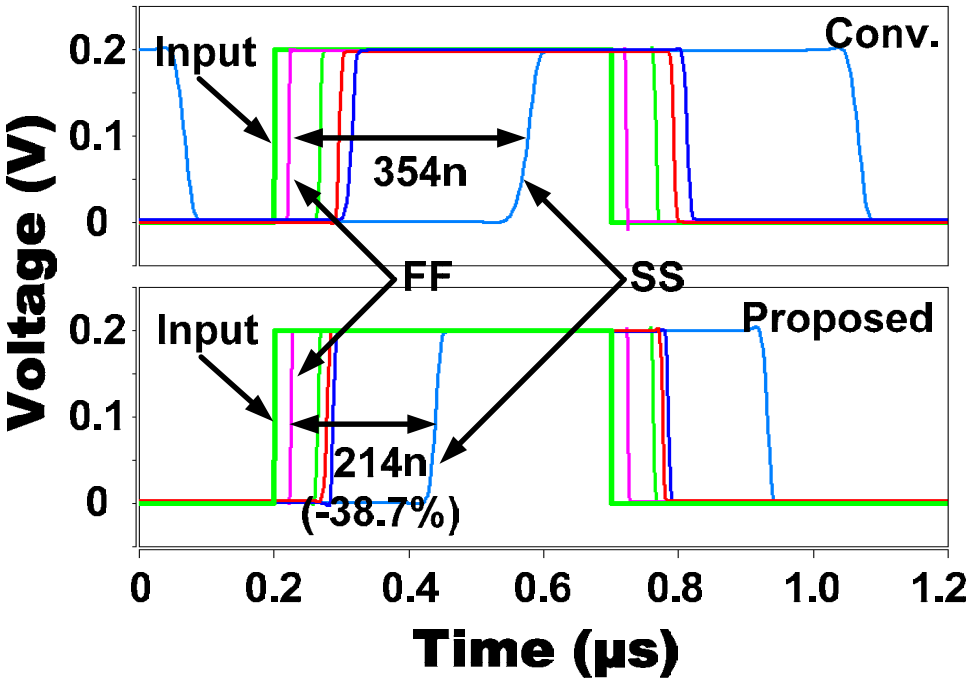


Fig. 2.12 Simulation waveforms using corner parameters showing improved tolerance to process variation using proposed scheme.

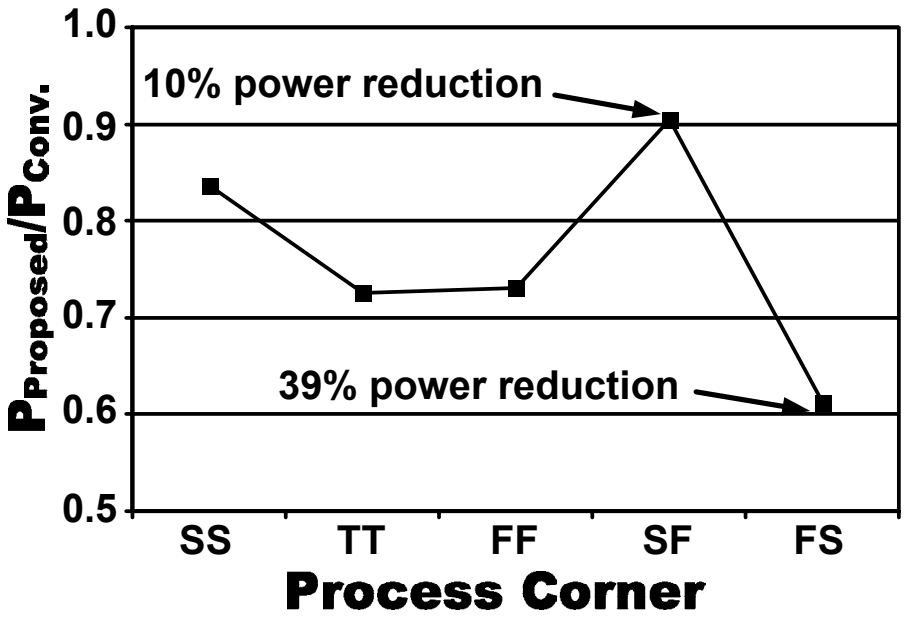


Fig. 2.13 Comparison of average power for corner parameters.

We tested our sizing method in more general logic paths by synthesizing a number of ISCAS benchmark circuits, as well as different component circuits used in that suite. Two cell libraries were created; a conventional library optimized for super-threshold operation and a new library based on our proposed sizing scheme. Each library contains inverters, two-input NANDs, and two-input NORs. Digital logic gates in the conventional library use the minimum channel length-0.12μm, in this process technology. The proposed library uses the optimized channel length of 0.36μm for NMOS devices and 0.12μm for PMOS devices. Critical path delays, power consumption, and energy consumption obtained from HSPICE simulations are compared in Tables 3.1, 3.2 and 3.3. Increasing the number of switching internal nodes and activity rate will result in more dynamic power savings compared to the leakage power savings. In this simulation, internal nodes which are connected to the

switching input signal contribute dynamic power savings and the static nodes contribute to leakage saving. Improvements in delay range from 7.8% to 10.4% depending on the type of logics used in the critical path. In addition, a simultaneous power reduction of 8.4% to 34.4% is achieved with the proposed scheme. As a result, reductions of energy ranging from 12.4% to 41.2% are obtained. Finally, the effect of activity rate on power savings is shown in Fig. 2.14 for the 4 stage inverter chains used in section III-E. The proposed sizing scheme reduces the leakage power and dynamic power simultaneously. Leakage power reduction from the improved sub-threshold slope is larger than the dynamic power reduction. Therefore, as the activity rate decreases, the power savings improves and converges to the leakage power savings as can be seen in Fig. 2.14.

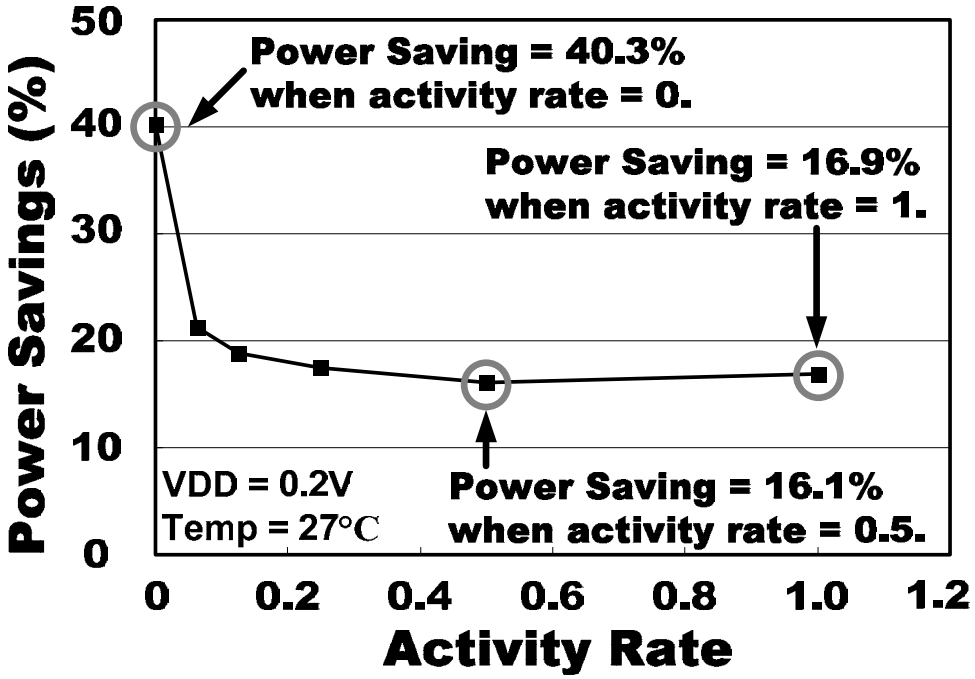


Fig. 2.14 Effect of activity rate on power savings in the 4 stage inverter chain used in section III-E.

Table 2.1

CRITICAL PATH DELAY COMPARISON FOR ISCAS BENCHMARK CIRCUITS

Circuit	0.2V, Temp = 27C°		
	Conv. (ns)	Prop.(ns)	Improvement (%)
C6288	119	107	10.1
C1355	443	397	10.4
74283	231	207	10.4
74L85	121	110	9.1
74182	103	95	7.8

Table 2.2

POWER COMPARISON FOR ISCAS BENCHMARK CIRCUITS

Circuit	0.2V, Temp = 27C°		
	Conv. (μW)	Prop. (μW)	Improvement (%)
C6288	2.24	1.50	33.0
C1355	0.64	0.42	34.4
74283	1.60	1.38	13.8
74L85	0.38	0.26	31.6
74182	0.40	0.38	8.4

Table 2.3

POWER COMPARISON FOR ISCAS BENCHMARK CIRCUITS

Circuit	0.2V, Temp = 27C°		
	Conv. (μW·nS)	Prop. (μW·nS)	Improvement (%)
C6288	266.56	160.50	39.8
C1355	283.52	166.74	41.2
74283	369.60	285.66	22.7
74L85	45.98	28.60	37.8
74182	41.20	36.10	12.4

2.5 Conclusions

As process technologies are scaled down, RSCE becomes stronger due to the increased HALO doping. RSCE is not a major concern in super-threshold designs since it does not affect the electrostatics of minimum channel length devices which are optimal for high performance and low power. Rather, DIBL and V_{th} roll-off were the main considerations for minimum channel length devices. However, in the sub-threshold region, where DIBL is reduced and current depends exponentially on threshold voltage, RSCE must be considered for optimal device sizing. In this work, we show that using minimum channel length is not optimal for sub-threshold circuits in the process technologies where strong RSCE effect can be observed. We propose a novel device size optimization scheme which can achieve high drive current, low device capacitance, and high Ion-to-Ioff ratio by utilizing the RSCE. Circuits using the proposed sizing scheme are more robust against RDF because of the increased gate area at the optimal performance point. The proposed sizing scheme reduces delay and power dissipation simultaneously, which is not possible using conventional sizing schemes. As a result, a significant improvement in energy is obtained. Average delay in ISCAS benchmark circuits was improved by 13% while average power dissipation and energy dissipation were reduced by 31% and 40%, respectively. The proposed scheme also offers a tighter delay and power consumption distribution by improving the σ/μ ratios by 37.5% and 70%, respectively.

Chapter 3 Design of Reliable Sub-threshold SRAMs

3.1 Introduction

SRAMs with a wide range of supply voltages are necessary for achieving high performance during normal modes while minimizing power consumption during low voltage modes [32]. For a reliable operation from the strong-inversion region down to the sub-threshold region, key memory design metrics such as noise margin, speed, and power consumption need to be examined across this range of supply voltages. Designing robust SRAM memory for sub-threshold systems is extremely challenging because of the reduced voltage margin and the increased device variability. Conventional 6-T SRAMs in the sub-threshold region fail to deliver the density and yield requirements due to the reduced Static Noise Margin (SNM), poor writability, limited number of cells per bitline, and reduced bitline sensing margin. Previously, 7-T, 8-T and 10-T SRAM cells have been proposed to improve the SNM by decoupling the SRAM cell nodes from the bitline and hence making the read mode SNM equal to the hold mode SNM [8][33][34]. Writability has been improved in prior designs by using a higher supply voltage for the write access transistors at the cost of generating and routing the extra supply voltage [8]. The maximum number of cells per bitline in previous sub-threshold SRAMs was limited to 256 at 0.3 V [8]. Robust high-density sub-threshold SRAMs are indispensable for the successful deployment of sub-threshold circuits in emerging ultra-low power applications.

3.2 Previous Sub-threshold SRAM Circuit Techniques

Designing sub-threshold SRAMs is challenging due to the degraded cell stability, small Ion-to-Ioff ratio, and large current variations [35][36][37][38][39]. In this section we will discuss several circuit techniques that have been proposed to mitigate these design issues associated with sub-threshold SRAMs.

Cell read stability is a critical design parameter in SRAMs. A decoupled cell is inevitable for sub-threshold SRAMs as it achieves the maximum read SNM at a given supply voltage for the same area constraint. Fig. 3.1 shows previous 8T and 10T SRAM cells with decoupled cell nodes [8][21][32][40]. Most of these cells use the 6T SRAM structure for data storage and write operation. All minimum sized devices can be used because the operation is no longer limited by the read stability problem, as the separate read port decouples the cell node from the read bitline. No disturbance current flows between the storage transistors and the read bitline making the read SNM equal to the ideal hold SNM.

Reliably sensing the read bitline voltage is another critical challenge for sub-threshold SRAMs. Verma et al. proposed using a redundant sense amplifier to improve the failure probability for bitline sensing [41]. Instead of using a single sense amplifier, two half-sized amplifiers were used to perform the single ended bitline sensing. It is claimed that the failure rate of the two half-sized sense amplifiers is smaller than that of a single sense amplifier when one out of the two half-sized sense amplifiers is selected through a start-up selection routine. However, this technique can only be used when the failure rate of the half-sized sense amplifier is low enough, which limits the minimum operational voltage. Zhai, et al. proposed using a

transmission gate as the access device [42]. However, the read disturbance problem will likely become worse for this SRAM cell due to the current through the access path.

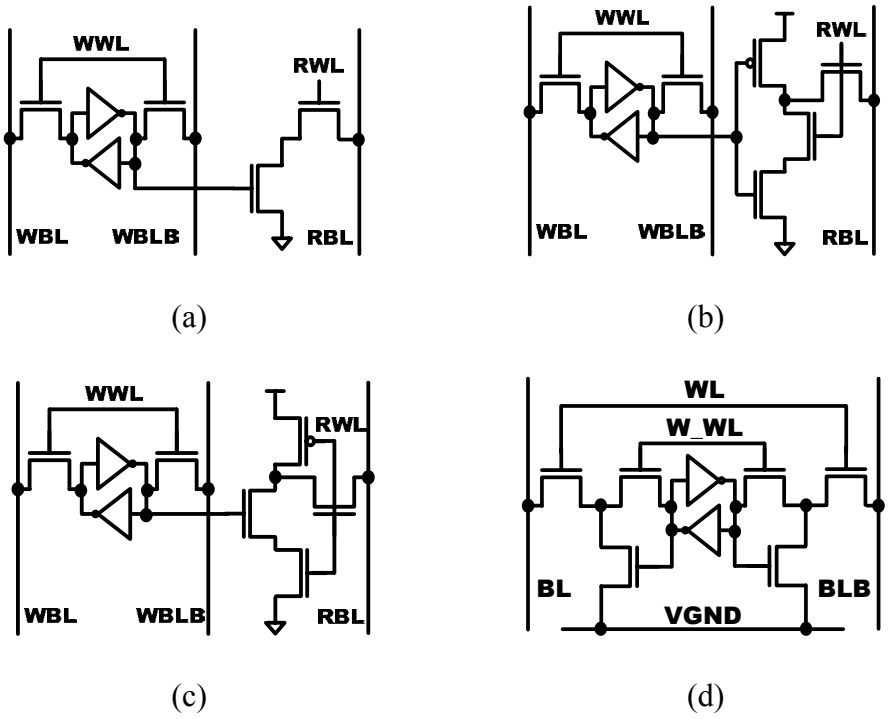


Figure 3.1 (a) Previous 8T SRAM cell [32]. (b)-(d) Previous 10T SRAM cells [8][21][40].

Finally, the write operation becomes problematic in sub-threshold SRAMs as the variation in current worsens due to its larger sensitivity to PVT parameters. In particular, weak write access transistors and strong pull-up PMOS transistors can cause a cell write failure to occur. To avoid this problem, write access transistors must be strong enough to overwrite the cell data even under the worst case PVT parameters. Various write margin improvement techniques have been proposed to make the access

devices stronger compared to the storage devices [8][41][43]. The collapsed supply rail scheme lowers the cell supply voltage during write operations weakening the current drivability of the PMOS storage transistors. However, the lowered supply voltage degrades the cell stability of SRAM cells in the pseudo-write mode (also known as the half-select mode) because the PMOS storage devices also become weaker due to the shared supply node. Furthermore, the SRAM cell stability is already close to the data retention limit in sub-threshold SRAMs making this scheme infeasible even with a column-by-column supply control. Alternatively, the boosted wordline scheme increases the drive strength of the write access transistors to improve cell write margin over that of pull-up PMOS transistors under process variations. However, this scheme cannot be used for column-muxed array architectures because it also increases the drive strength of the write access transistors in pseudo write mode. It also requires additional circuitry to generate and route the boosted supply voltage.

3.3 A 0.2V, 480 kb Sub-threshold SRAM with 1k Cells Per Bitline for Ultra-Low-Voltage Computing

3.3.1 Overview

This work introduces various circuit techniques for designing robust and high-density SRAMs in the sub-threshold regime. The following techniques are proposed to enable a fully functional 480kb SRAM operating at 0.2 V: (i) decoupled 10-T SRAM cell for read margin improvement, (ii) utilizing Reverse Short Channel Effect (RSCE) for write margin improvement, (iii) eliminating data-dependent bitline leakage to enable 1k cells per bitline, (iv) Virtual Ground (VGND) replica scheme for improved bitline sensing margin, and (v) writeback scheme for row data preservation in unselected columns during write. A 130 nm SRAM test chip was successfully measured and characterized.

3.3.2 10-T SRAM Bitcell Design

Fig. 3.2 shows the proposed 10-T SRAM cell and simulated SNM. The proposed SRAM cell consists of a cross-coupled inverter pair (M1, M2, M4, M5), write access devices (M3, M6), and decoupled read-out circuits (M7, M8, M9, M10). The write bitlines (WBL, WBLB) and the read bitline (RBL) are precharged to VDD before the cell is accessed. When read is enabled (RWL=1), RBL is conditionally discharged through pull-down transistors M7, M8, and M9 depending on the ‘QB’ value. The cell node is decoupled from the read bitline, retaining a hold mode SNM during the read operation. When read is disabled (RWL=0), node A is held to VDD by M10 making

the bitline leakage flow from node A to RBL, regardless of the data stored in the SRAM cell. This results in a bitline leakage independent of the cell data allowing a larger number of cells to be attached to a single bitline. Details on this topic will be described in section II-D. The proposed 10-T SRAM cell has an SNM of 82 mV at a supply voltage of 0.2 V and a temperature of 27 °C while the conventional 6-T SRAM cell SNM is 24 mV under these conditions (Fig. 3.2 (b)). The SNM of the proposed 10-T SRAM at a supply voltage of 0.2 V is equal to that of the conventional 6-T SRAM cell at 0.4 V (Fig. 3.2 (c)). In addition, the SNM normalized to supply voltage in Fig. 3.2 (d) shows that the variation of SNM in the proposed 10-T SRAM cell is smaller than that of the conventional 6-T SRAM cell, which is the result of reduced variation in the longer access transistor used in our design to utilize the short channel effect. Further details on this topic will be given in section II-B and II-C. Write operation is similar to 6-T SRAM cells where the write wordline is asserted (WWL=1) after new data is loaded onto the write bitlines (WBL, WBLB).

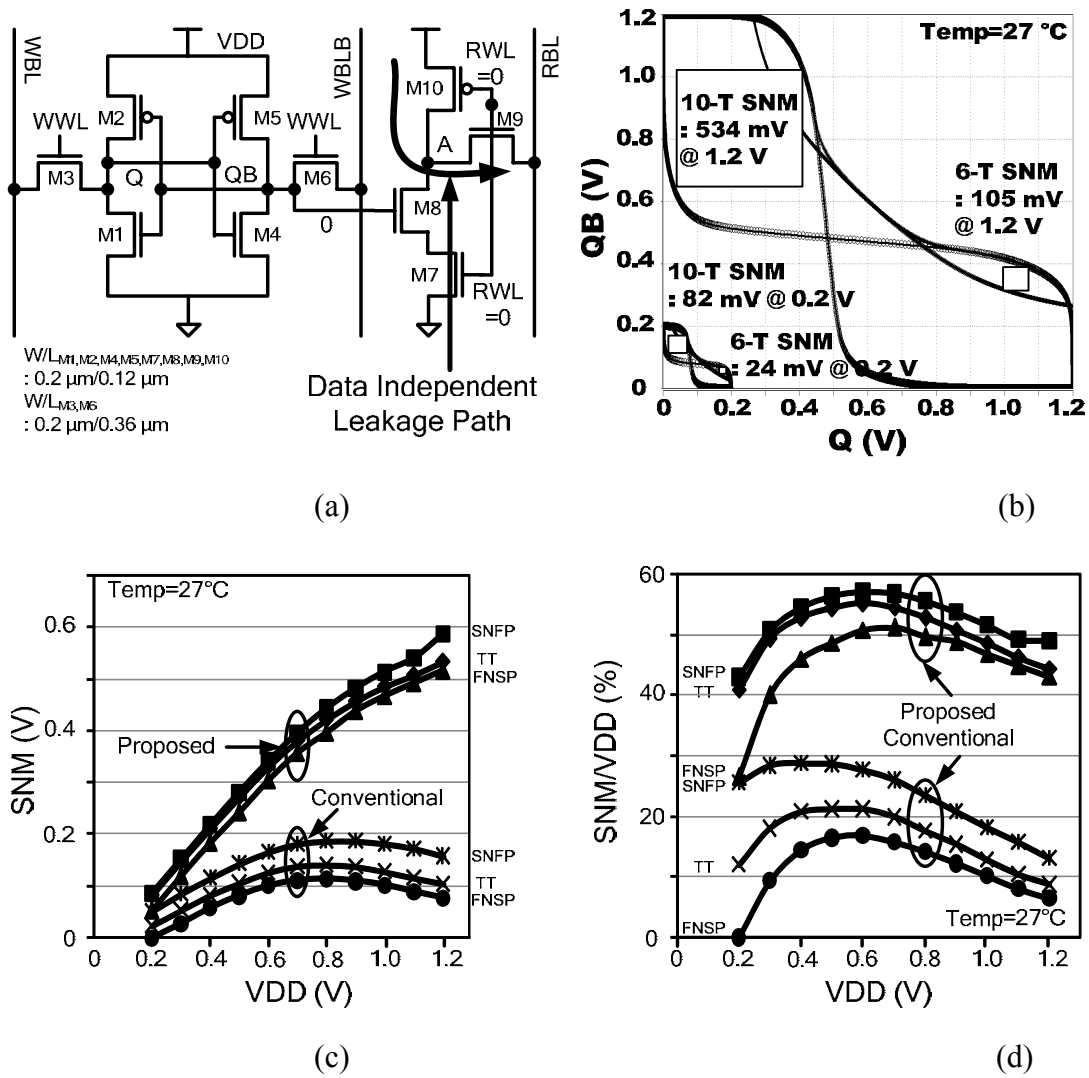
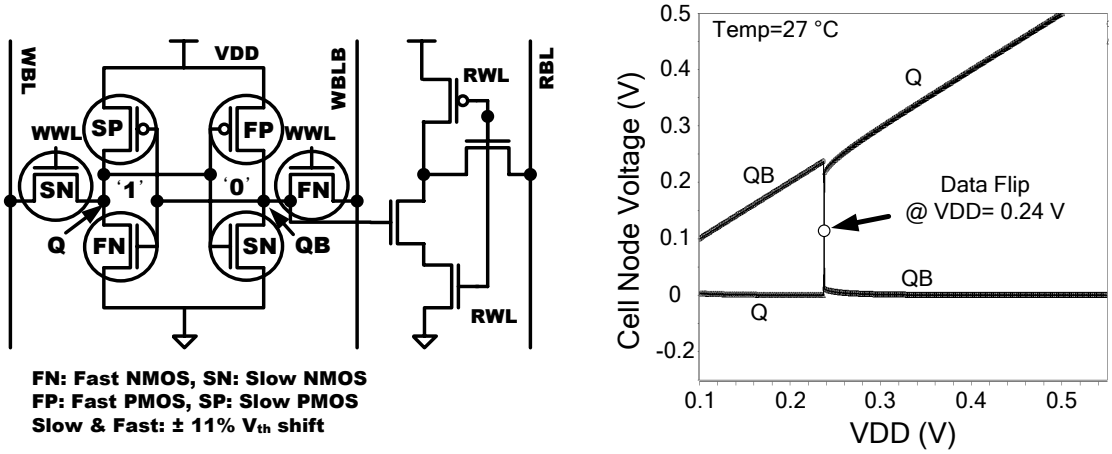


Figure 3.2. (a) Proposed 10-T SRAM cell with data independent leakage. (b) SNM comparison of conventional 6-T and proposed 10-T SRAM cell. (c) SNM comparison at different process corners and supply voltages. (d) SNM normalized to supply voltage for the results in (c).

Data retention voltage represents the minimum supply level below which an SRAM cell has a negative SNM. Global process variations and local device mismatches play major roles in determining this voltage. The worst case device corners for the data retention voltage simulation are illustrated in Fig. 3.3 (a). The weak pull-up device connected to Q and the strong pull-down device are the worst case for flipping the logic '1' at node Q. At the other side of the cross-coupled latch, strong pull-up device and weak pull-down device have the largest probability of flipping the logic '0'. The simulated waveforms (Fig. 3.3 (b)) indicate that the proposed 10-T SRAM cell has a data retention voltage of 0.24 V in this worst case scenario. The proposed SRAM has a positive SNM even at the supply voltage of 0.1V when only global process variation is considered.



(a) (b)

Figure 3.3 (a) Condition for worst case data retention voltage. (b) Simulated waveforms showing a minimum data retention voltage of 0.24 V.

3.3.3 Utilization of RSCE in SRAM Bitcell Design

Maintaining a sufficient write margin is challenging in sub-threshold SRAMs due to the small gate overdrive and large process variation in the write access devices (M3 and M6 in Fig. 3.2). Virtual supply rails have been used in previous work to improve cell writability [8]. In [8], the cell supply voltage of the selected column becomes floating during write operation. The virtual supply rails collapse making it easier for the write access devices to flip the cell value. However, this technique is not suitable in sub-threshold SRAMs as the virtual supply droop cannot be controlled accurately and the SNM is already close to the limitation. Another previous SRAM implementation used a wordline voltage which is higher than the cell voltage to increase the drive current of the write access transistors [8]. However, this technique requires an additional high VDD to be generated and routed.

In this work, we utilize the RSCE in the sub-threshold region to improve the cell writability without having to introduce a separate high VDD [20]. The cell writability in our SRAM design is improved by using write access transistors with a channel length that is 3X the minimum value to utilize the RSCE (Fig. 3.4 (a)). The stronger drive current enables a robust write operation, and hence lowers the minimum operating voltage. Unlike prior techniques, no additional supply voltage is required for our proposed technique. The bitline capacitance is the sum of the wire capacitance and the capacitance at the junction of the write access transistors. Since neither the junction nor the overlap capacitance change with the increased channel length, the bitline capacitance is not affected.

Simulation results in Fig. 3.4 (b) show that the write operation of the proposed

SRAM at 0.2 V is equivalent to that of a conventional scheme using a 0.27 V WWL voltage. Fig. 3.4 (c) and (d) show the write margin simulation results for different supply voltages. Fast PMOS and slow NMOS process parameters were used to represent the worst case write condition. All devices have a minimum channel width (200 nm). A negative write margin in Fig. 3.4 (c) indicates a write failure. Using a channel length of 0.36 μm for M3 and M6, the write margin of the proposed SRAM cell is improved from -90 mV to 70 mV at 0.2 V. Fig. 3.4 (d) illustrates the equivalent wordline boost normalized to the supply voltage by applying the proposed sizing. It can be seen that the normalized equivalent wordline boost increases at lower supply voltages, which illustrates the usefulness of the proposed technique in the deeper sub-threshold region.

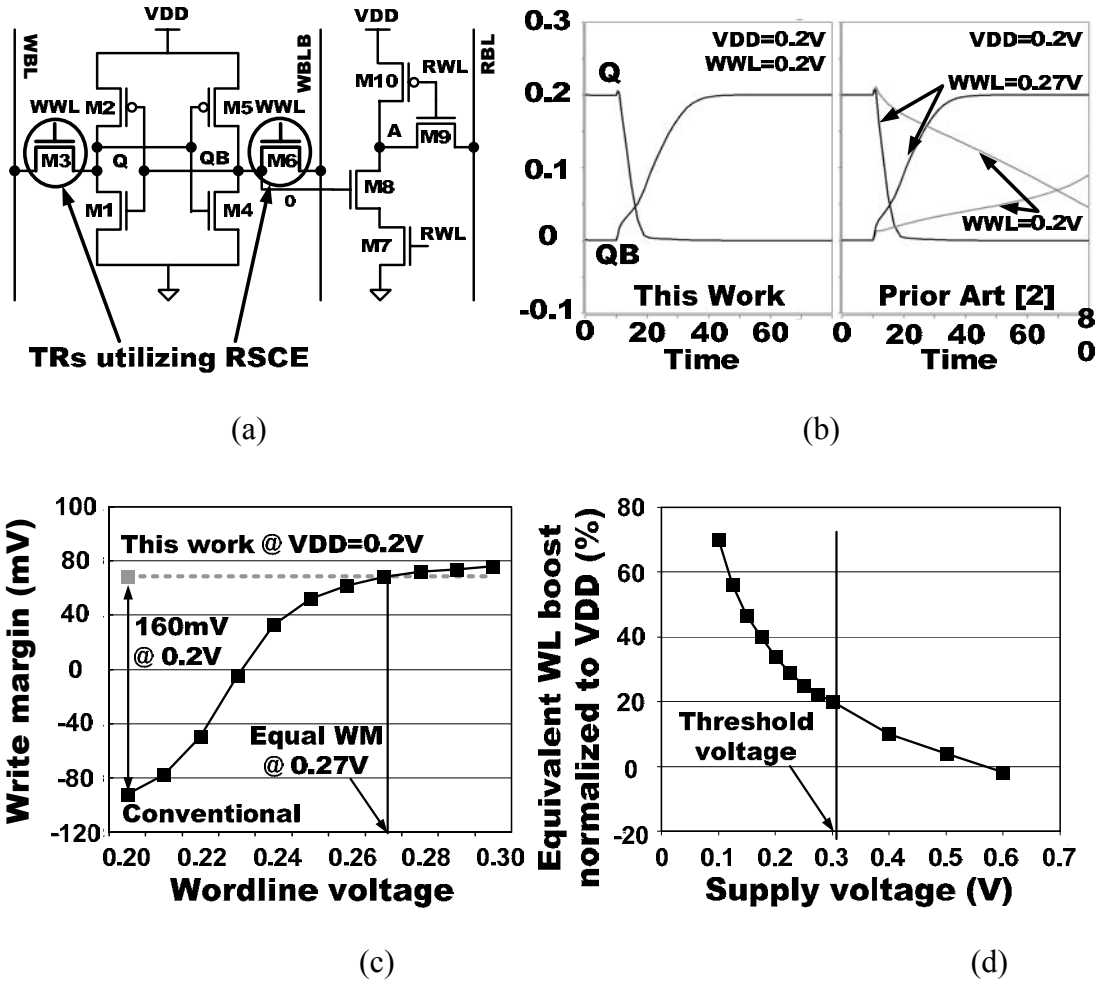
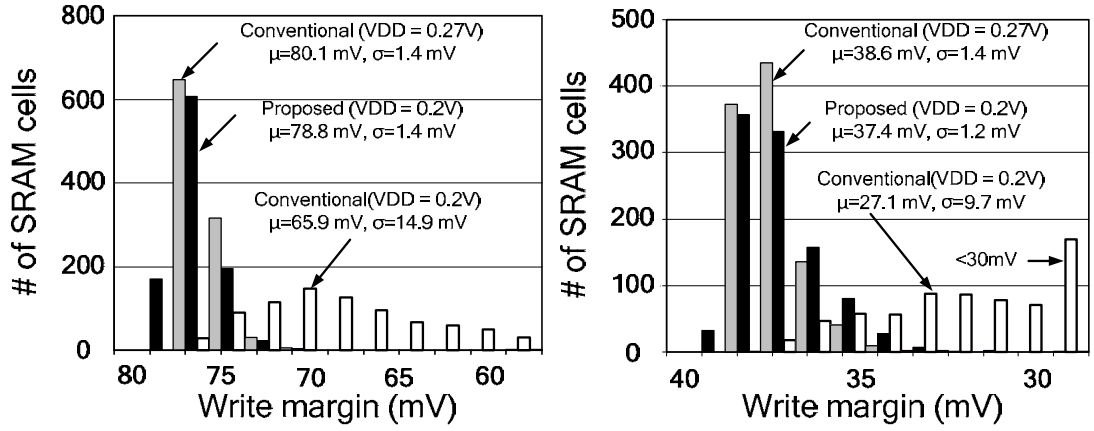


Figure 3.4 Reverse short channel effect is utilized for write margin improvement: (a) Proposed 10-T SRAM cell with long channel write access transistors to improve writability. (b) Simulation results showing improved write delay. (c) Write margin versus wordline voltage. (d) Equivalent wordline boost normalized to VDD.

Random Dopant Fluctuations (RDF) cause parameter mismatches even between devices with identical layout in close proximity [44]. The impact of RDF is more severe in the sub-threshold region due to the exponential relationship between the current and threshold voltage [5]. The standard deviation (σ) of the threshold voltage distribution is known to be proportional to $(WL)^{-1/2}$ [45] where W is the device width and L is the channel length. The gate area of the access transistors M3 and M6 utilizing RSCE is 0.072 μm^2 (=0.2 $\mu m \times 0.36$ μm) which is 2X larger than the minimum size access transistors in conventional 10-T SRAM cells. This translates into a 58% smaller standard deviation in the threshold voltage reducing the write margin variability in the proposed SRAM cell. Figures 3.5 (a) and (b) show write margin distributions using Monte Carlo simulation at two different supply levels. It is assumed that each device in the 10-T SRAM has independent threshold voltages which follow a normal distribution. Results are also shown for a 6-T SRAM cell using all minimum channel length devices at 0.2V and 0.27V. The average and the standard deviation of the proposed cell's write margin are 79 mV and 1.4 mV, respectively, which are much superior than those of the conventional cell (65 mV and 15 mV) at 0.2 V. The large improvement comes from the smaller random-dopant-fluctuation and the increased current drivability of the write access transistors in the proposed 10-T SRAM cell. In addition to the SRAM cells, longer channel length devices are used for the static CMOS gates in the SRAM row decoding path and peripheral read/write circuits to reduce the delay, power consumption, and circuit variability.



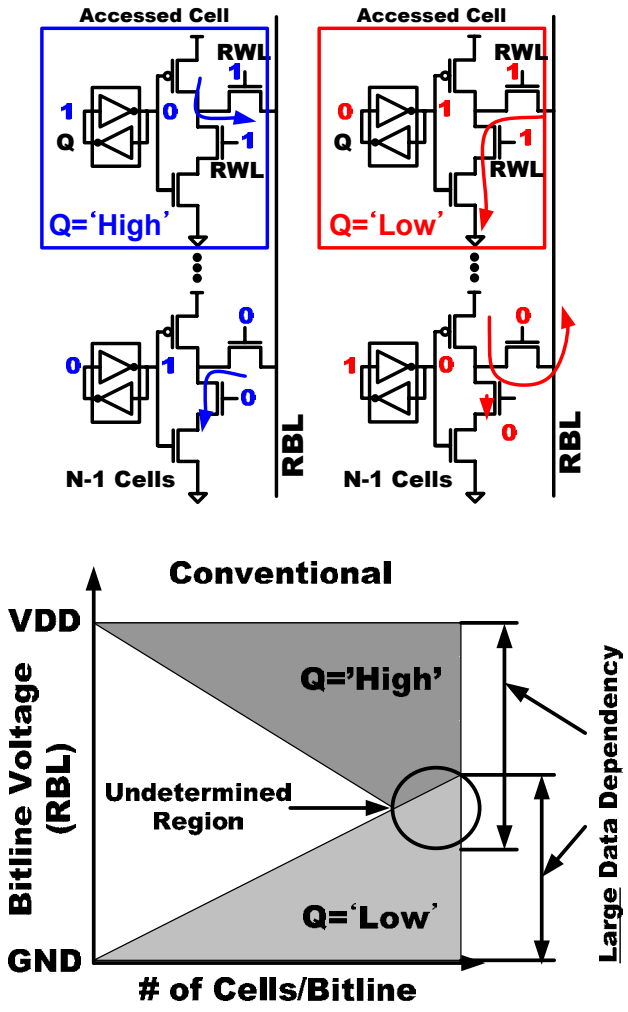
(a) (b)

Figure 3.5 Write margin distribution of proposed and conventional SRAM cell from 1000 Monte Carlo simulations: (a) VDD=0.2V (b) VDD=0.1V.

3.3.4 Data-Independent Bitline Leakage for High Density

The small I_{on}-to-I_{off} ratio in the sub-threshold region limits the number of cells per bitline and negatively impacts the SRAM density. As the number of cells in a bitline increases, bitline leakage from the unaccessed cells can rival the read current of the accessed cell making it difficult to distinguish between the bitline high and low levels. Previous techniques suffer from the data-dependent bitline leakage which can cause the RBL high level to droop or RBL low level to rise based on the data stored in the unaccessed cells of a bitline [8][46]. Fig. 3.6 (a) shows the simplified schematic of the bitline with data-dependent bitline leakage current [8]. For the sake of simplicity, only the cross-coupled inverters and read ports are shown. When reading a '1', the worst case read bitline (RBL) voltage is determined based on the contention between

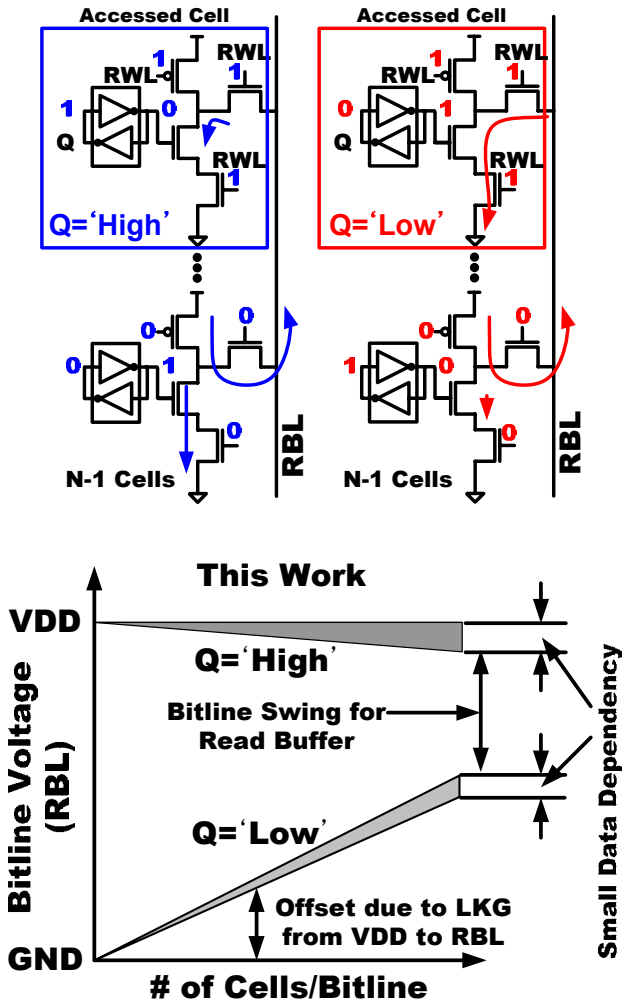
the pull up current from the accessed cell and the pull down bitline leakage currents from the unaccessed cells. Likewise, when reading a '0', the contention between the pull down current of the accessed cell and the pull up bitline leakage currents of the unaccessed cells decides the worst case RBL voltage. As the number of cells per bitline increases, the worst case RBL for data '1' decreases and that for data '0' increases due to the bitline leakage current. As a result, the bitline voltage for data '1' may be lower than that for data '0' under the worst case data patterns, which can cause the read buffer to generate an incorrect output as shown in Fig. 3.6 (b). A 0.3 V sub-threshold SRAM with 256 cells on a single bitline has been reported in [8]. Our simulations indicate that the maximum number of cells per bitline of the prior design quickly reduces to 16 at a supply voltage of 0.2 V due to the bitline leakage problem.



(a) (b)

Figure 3.6 Impact of data-dependent bitline leakage current on bitline voltage: (a) Simplified bitline schematic with data-dependent bitline leakage current. (b) Read bitline voltage dependency upon data pattern and number of cells per bitline.

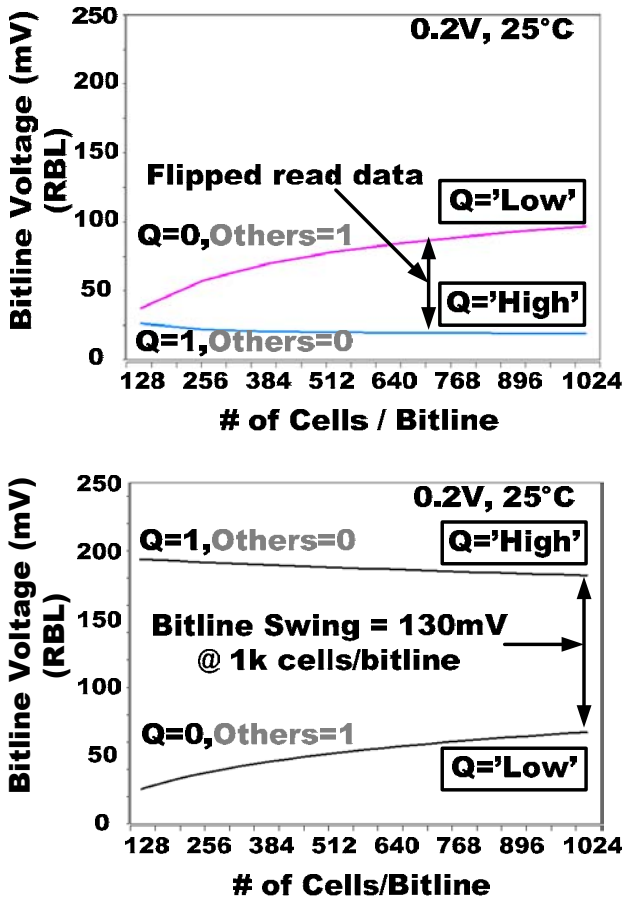
The proposed 10-T SRAM cell eliminates the data-dependent bitline leakage problem by turning on M10 in Fig. 3.2 (a) when the SRAM cell is unaccessed (RWL=0). The drain voltage of M10 therefore becomes VDD and forces the leakage current to flow from the cell into the bitline regardless of the data stored. Fig. 3.7 (a) shows the simplified schematic of the proposed bitline with data-independent bitline leakage current. The logic low level is decided by the balance between the pull up leakage current of unaccessed cells and the pull down read current of the accessed cell as shown in Fig. 3.7 (a). The logic high level is close to VDD because both bitline leakage current and cell current are pulling up the RBL. By doing so, RBL voltages for the different logic levels are pinned and are independent of the cell data pattern as described in Fig. 3.7 (b).



(a) (b)

Figure 3.7 Effect of data-independent bitline leakage current on bitline voltage: (a) Simplified bitline schematic with data-independent bitline leakage current. (b) Read bitline voltage independency upon data pattern.

Fig. 3.8 shows the worst case RBL voltages simulated using HSPICE. It can be seen that the RBL voltage for logic '1' is lower than that for logic '0' in previous scheme (Fig. 3.8 (a)) [8]. However, in this work, a bitline swing of 130 mV irrespective of the column data pattern is achieved at a 0.2 V supply voltage for a 1k cell bitline (Fig. 3.8 (b)).

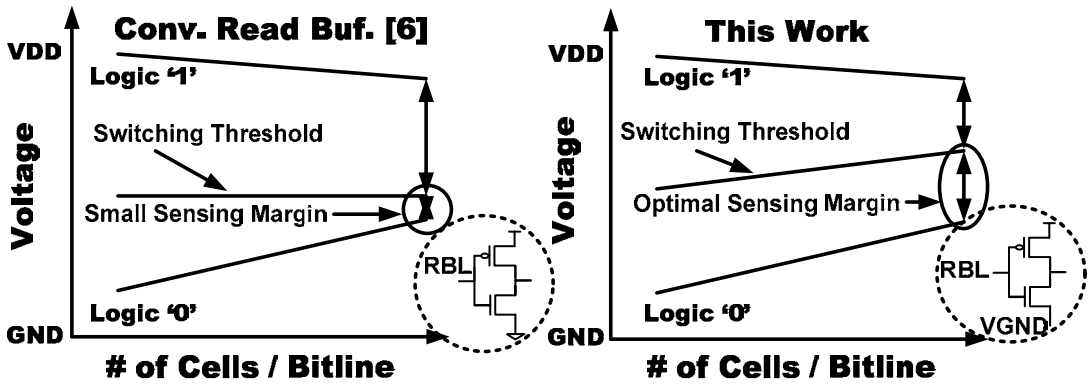


(a) (b)

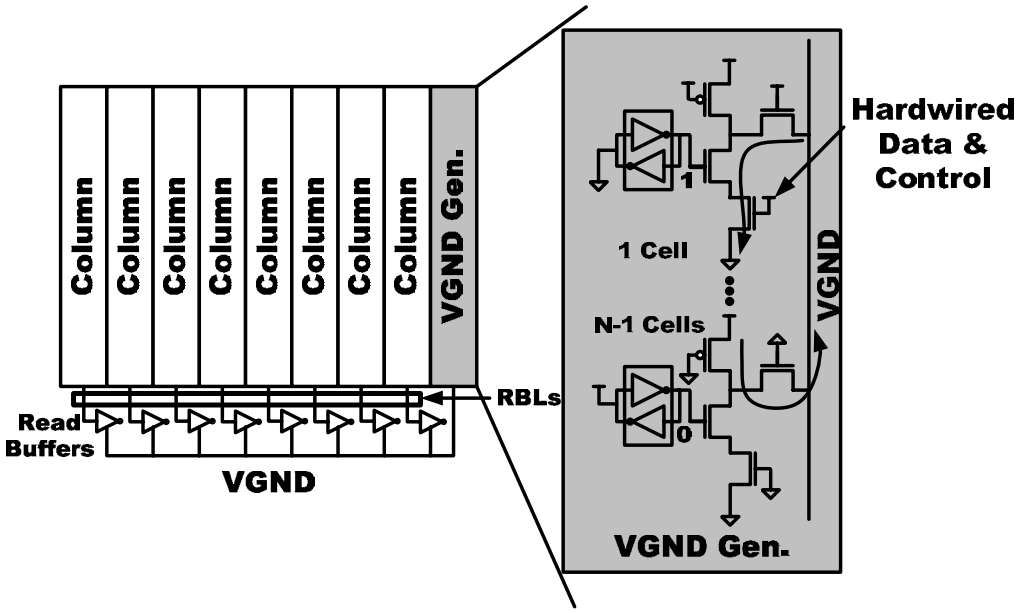
Figure 3.8 Simulation results of read bitline voltage with worst case data pattern using nominal process parameters: (a) Conventional scheme with data-dependent bitline leakage current. (b) Proposed scheme eliminating data-dependent bitline leakage current.

3.3.5 Virtual Ground (VGND) Replica Scheme for Improved Sensing Margin

In sub-threshold SRAMs, sense amplifiers are replaced with static inverter type read buffers because it is noise margin that is the key design concern and not the speed [7]. Therefore, these read buffers provide the maximum sensing margin for a given supply voltage due to the full swing in the bitlines. Based on the fact that the bitline logic levels are insensitive to the column data pattern in our design (Section II-D), a VGND replica scheme is devised to maximize the sensing margin of the read buffers. The proposed VGND replica scheme automatically tracks the optimal read buffer trip point to obtain the largest possible sensing margin. The trip point of the read buffer is set to the middle of the logic high and low levels by using the VGND level generated from a replica bitline as the ground level of the read buffer as shown in Fig. 3.9. Figures 3.9 (a) and (b) compare the sensing margin of the proposed scheme with a conventional scheme using a zero ground level. The sensing margin of the conventional scheme degrades significantly as the number of cells per bitline increases because the increased logic '0' level of RBL strengthens the pull down path. However, the trip point of the proposed scheme is always maintained at half the bitline swing because VGND tracks the logic '0' level balancing the strength of pull down device with pull up device. A replica bitline with hardwired data and control signals is used as VGND generator. The reading '0' condition is implemented to generate the logic low level, which is used as the ground level for the read buffers as shown in Fig. 3.9 (c). A single VGND is shared with multiple columns to reduce the area overhead of the replica bitline. Eight columns can share a single VGND generator without generating noise in VGND. VGND level is dependent upon the accessed cell current.



(a)



(b)

Figure 3.9 VGND replica scheme for ideal bitline sensing margin: (a) Bitline sensing margin comparison of read buffers. (b) VGND replica scheme using VGND generator with hardwired data and command.

Simulation result of VGND at various corner parameters shows a variation of 20 mV, which roughly translates into a trip point variation of 10 mV (Fig. 3.10). Due to this relatively small variation in trip point, the read buffer can generate robust output data even when the drive current of the devices in the read buffers differ by 5X.

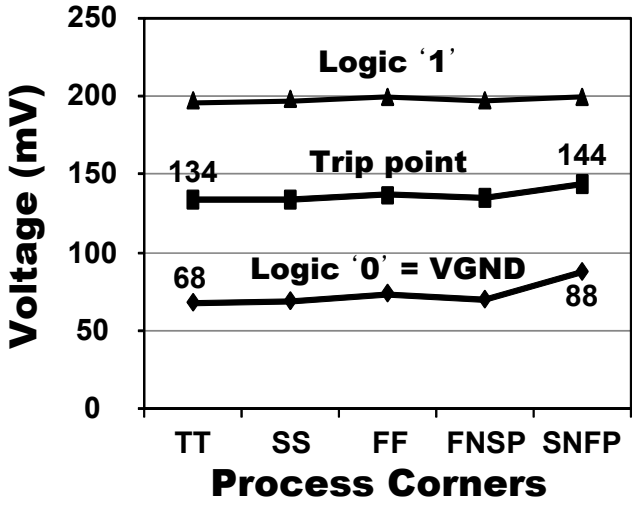


Figure 3.10 Simulation results of VGND and read buffer trip point at various corner parameters.

3.3.6 Writeback Scheme for Row Data Preservation

In a column muxed array, the write operation still has stability problems because the enabled write wordline is also shared by the unselected columns. This is also referred to as the pseudo-write (or pseudo-read) problem in conventional 6-T designs. Fig. 3.11 illustrates this issue where the unselected cells can undergo a write when the WWL signal is asserted while the write bitlines (WBL, WBLB) are precharged to VDD. This is exactly the same condition as the worst case read stability in conventional 6-T SRAMs.

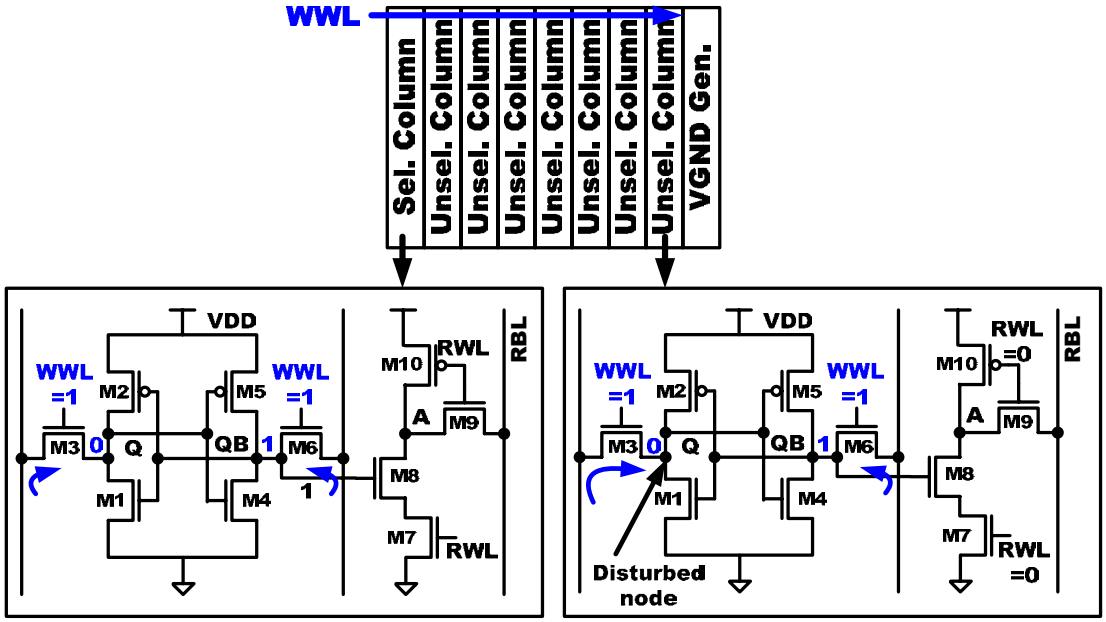


Fig. 3.11 Stability problem caused by pseudo-write in unselected SRAM cells.

A writeback scheme shown in Fig. 3.12 is applied to resolve the pseudo-write problem [47]. The write driver consists of a conventional write path and the writeback path. During write operation, read wordline (RWL) and write wordline (WWL) are enabled simultaneously. If the column is not selected for access (Y<i>=0), the write bitlines are kept to VDD and read operation is executed. The writeback signal (WB) is enabled from the rising edge of RWL with additional delay enabling the writeback path and the read data from the read buffer is transferred to D_INT and written back to WBL and WBLB. By rewriting the read data back to WBL and WBLB, there is no voltage difference between write bitlines (WBL, WBLB) and the cell nodes, eliminating the contention current.

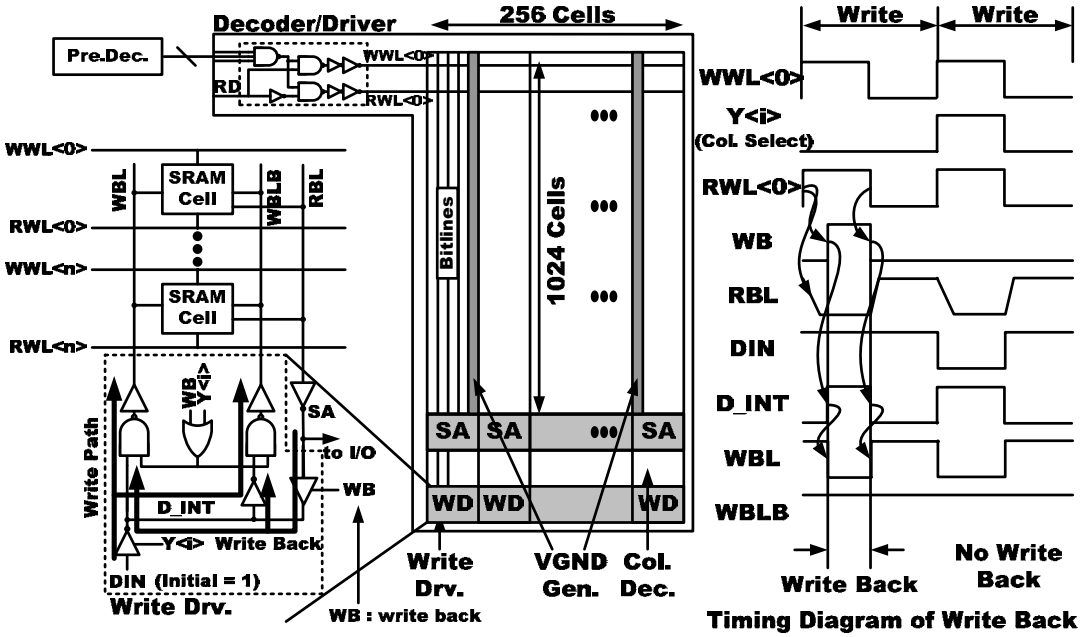


Fig. 3.12 Writeback scheme for preserving row data during write operation.

3.3.7 Test Chip Implementation and Experimental Results

A 1.5x4.1 mm^2 SRAM with 480kb cells was fabricated in a 130 nm, 8-metal CMOS technology. The cell size is 2.68x2.80 μm^2 using logic design rule. The threshold voltages of NMOS and PMOS are 0.32 V and -0.32 V, respectively. The nominal supply voltage for this process is 1.2 V. No standard IO circuit was used and the supply voltage for sub-threshold operation was directly applied to the power pads. The test chip microphotograph is shown in Fig. 3.13. The test chip contains four SRAM quadrants with different numbers of rows (128, 256, 512, and 1024) to demonstrate our proposed techniques on progressively longer bitlines. Each SRAM quadrant has 256 columns, which are divided by 32 sub-blocks. The size of sub-block with 1024 cells on a bitline is 42.9x3181 μm^2. To verify the effect of RSCE on circuit performance, a replica of the row decoding path was also implemented.

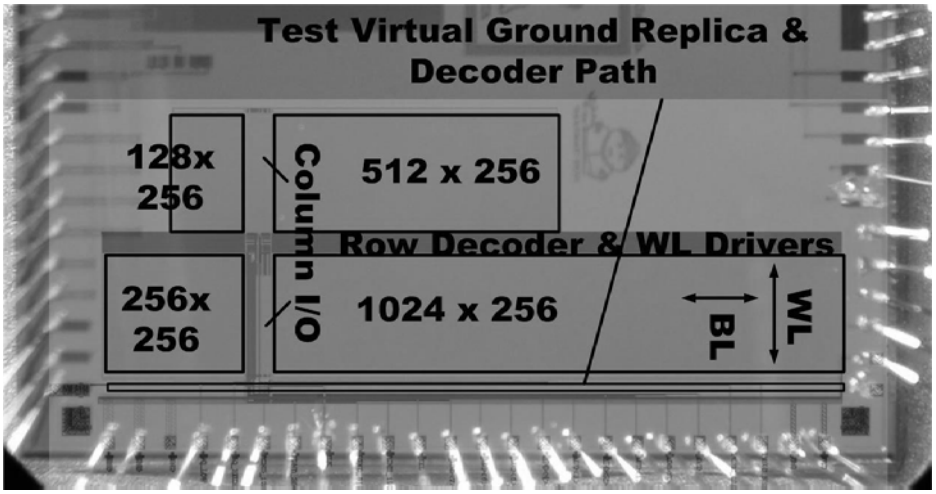
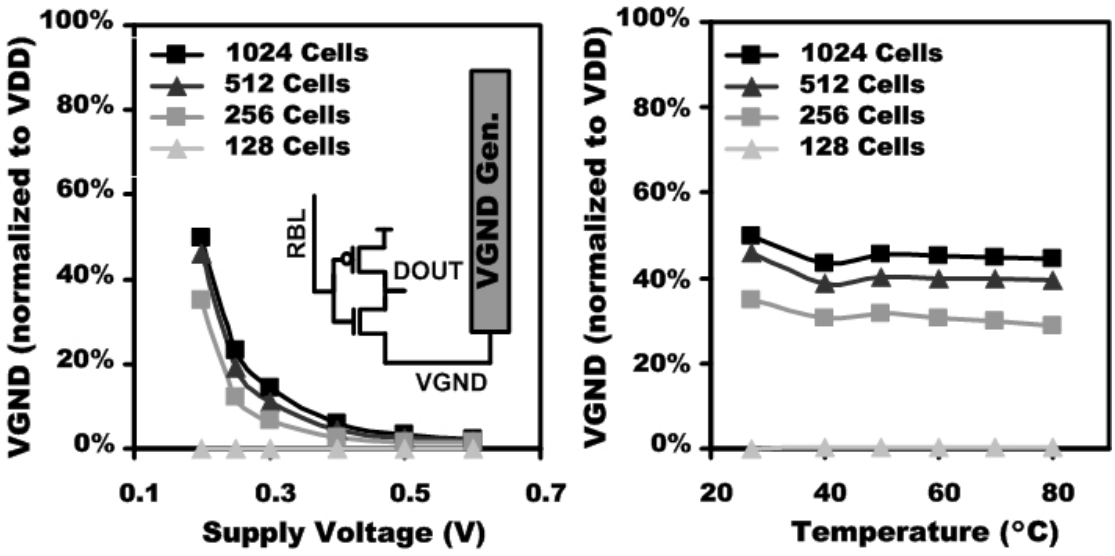


Fig. 3.13 Test chip microphotograph showing different sized quadrants.

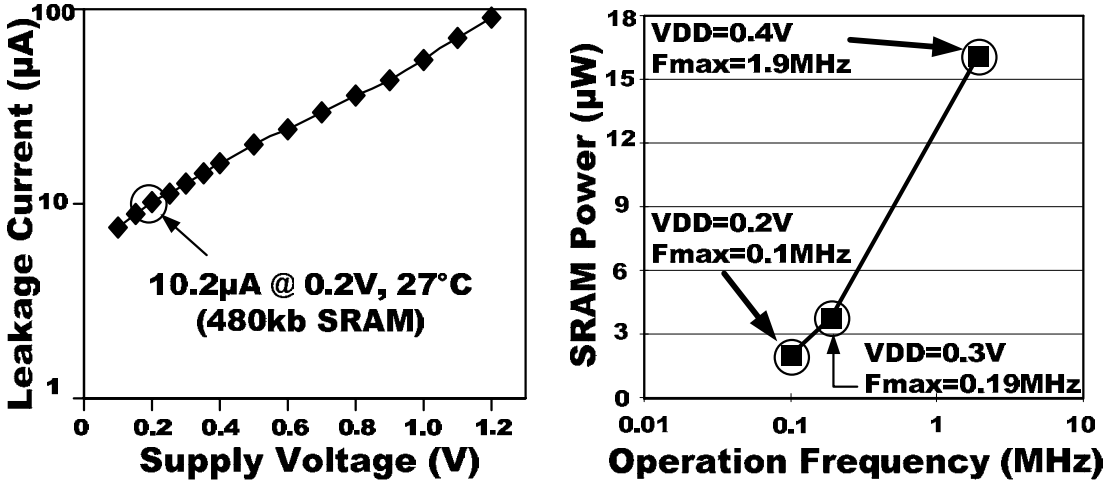
VGND from the replica bitline was measured to validate the proposed sensing scheme. The VGND level corresponds to the logic low level of the bitline. VGNDs of the four quadrants are measured from separate probing pads using a multi-meter. Fig. 3.14 shows the measurement data. The VGND level depends on the number of cells connected to a bitline and the supply voltage. As the number of cells increases, the amount of leakage current flowing from the unaccessed SRAM cells into the bitline also increases, causing a rise in the VGND level. The normalized VGND voltage also rises significantly as the supply voltage is reduced due to the decreased I_{on}-to-I_{off} ratio. This effect is shown in Fig. 3.14 (a) where VGND becomes as high as 50% of the supply voltage at 0.2 V for a bitline with 1k cells attached. Conventional read buffers will fail under these conditions due to the data-dependent bitline leakage, and the fixed trip point in the read buffers. Our proposed scheme tracks the logic low level using a replica bitline to provide the optimal read margin in the read buffers enabling 1k cells per bitline. The impact of temperature on the VGND level is small because the change in temperature causes a similar rate of change in both the bitline leakage and cell read current in the sub-threshold region, and VGND is determined by the balance between those currents. A 6% change in VGND was measured when varying the temperature from 27 °C to 80 °C at a supply voltage of 0.2 V (Fig. 3.14 (b)).



(a) (b)

Fig. 3.14 Measured VGND normalized to VDD: (a) Supply voltage dependency. (b) Temperature dependency.

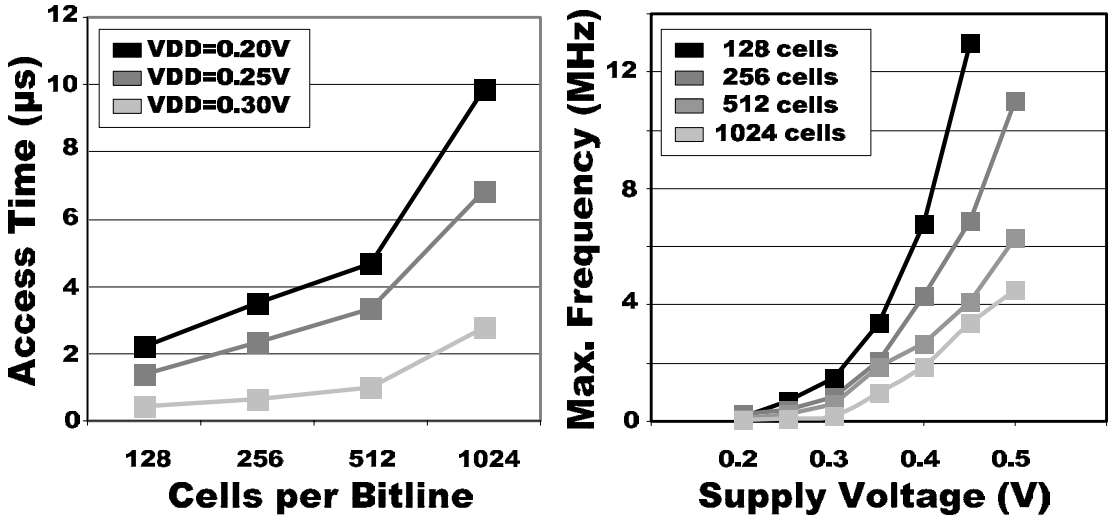
Leakage current and power consumption were measured and are summarized in Fig. 3.15. The leakage current of the 480k SRAM was 10 µA for a supply voltage of 0.2 V at 27 °C (Fig. 3.15 (a)). This current increases exponentially as the supply voltage increases. As seen in that figure, the leakage at a supply voltage of 0.2 V is 10% of that at 1.2 V. The total power consumption of the SRAM operating at the maximum frequency with a supply voltage of 0.2 V was 2 µW.



(a) (b)

Fig. 3.15 Leakage current and power measurements: (a) Measured SRAM leakage current versus supply voltage. (b) Measured SRAM power and maximum operating frequency versus supply voltage.

The access time and the maximum operation frequency of the four quadrants were measured. The maximum operating frequency was 100 kHz at 0.2 V and 27 °C for the quadrant with 1k cells per bitline (Fig. 3.16 (a,b)). The access time difference between the four quadrants was 4X. Operating frequency increases exponentially as the supply voltage is increased due to the sub-threshold MOS device behavior (Fig. 3.16 (b)).



(a) (b)

Fig. 3.16 Performance measurements: (a) Access time of four quadrants versus supply voltage. (b) Maximum operating frequency of four quadrants versus supply voltage.

The minimum supply voltage for proper read operation is shown in Fig. 3.17. The quadrants with 128 cells and 1k cells per bitline were readable at a supply voltage of 0.15 V and 0.17 V, respectively. This difference was caused by the VGND level, which limits the proper operation of the sense amplifier.

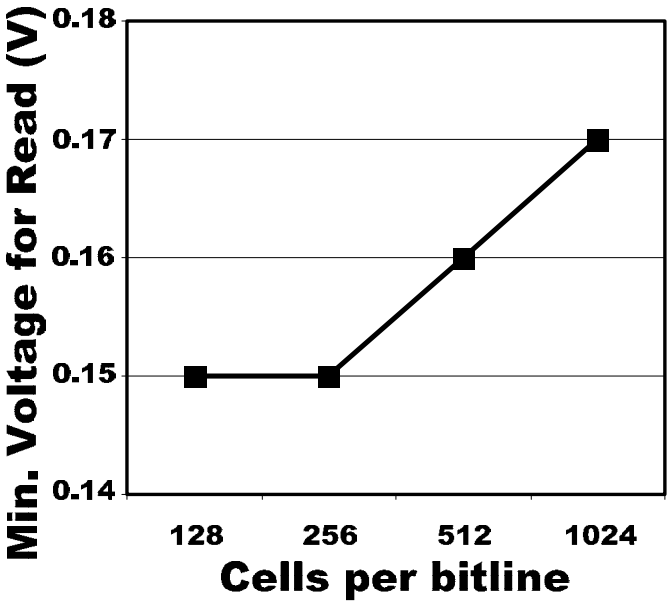
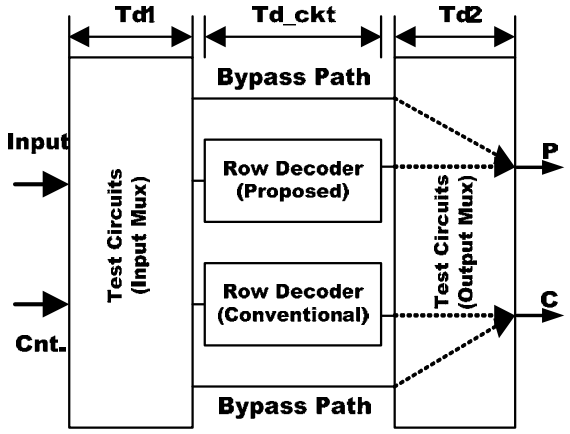
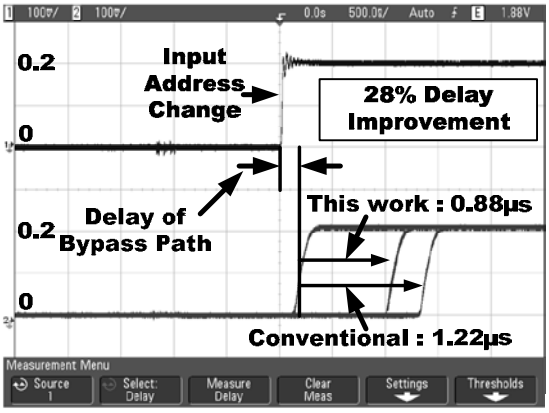


Fig. 3.17 Minimum supply voltage for proper read operation.

Measured waveforms from the replicated row decoding path are shown in Fig. 3.18 (b). For accurate on-chip delay measurements, a differential measurement technique was used where a dummy bypass path was included to cancel out the I/O path delay as shown in Fig. 3.18 (a). Measurement results indicate a 28% delay improvement by utilizing RSCE in the sub-threshold region. The devices with longer channel lengths offer a higher drive current per width which in turn is utilized to reduce the junction capacitance for higher performance.



(a)



(b)

Fig. 3.18 Measured performance improvement utilizing RSCE: (a) Block diagram for test circuit implemented. (b) Measured row decoding path delay improvement.

Fig. 3.19 (a) shows the read data output waveform at 0.17 V, which demonstrates a 100 kHz operation for the largest quadrant. The implemented SRAM is fully functional at 0.2 V for proper read and write operation and the key measured data is summarized in Table 3.1.

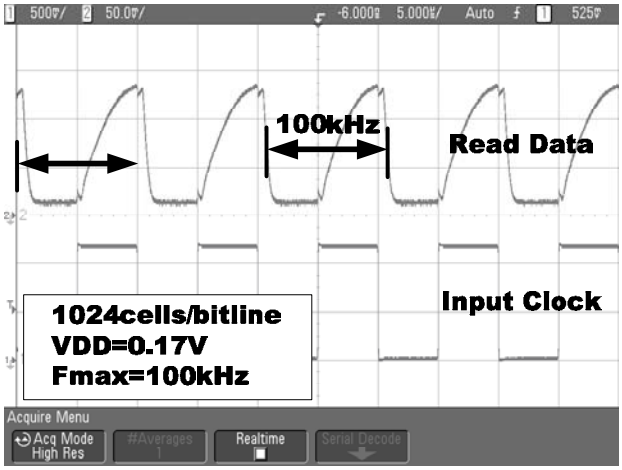


Fig. 3.19 Read data waveform at minimum supply voltage.

Table 3.1

Comparison between our design and previous sub-threshold SRAMs

	This work	[22]	[25]
Technology	130 nm CMOS	65 nm CMOS	130 nm CMOS
Density	480kb	256kb	512 × 13
Number of cells on a bitline	1024	256	16
SRAM cell type	10-T	10-T	7-T
Chip size	4.1 × 1.5 mm^2	1.89 × 1.12 mm^2	520 × 480 μm^2
VDD min	0.2 V @ 1024 cells per bitline, 27 °C	0.32 V for read, 0.38 V for write, 27 °C	0.19 V for read, 0.22 V for write, 27 °C
Performance	120 kHz @ 0.2 V, 27 °C	465 kHz @ 0.4 V, 27 °C	28 kHz @ 0.19 V, 27 °C
Power	2.04 μW	3.28 μW	1.197 μW @ 0.31 V

3.4 A Voltage Scalable 0.26V, 64 kb 8-T SRAM with V_{min} Lowering Techniques and Deep Sleep Mode

3.4.1 Overview

In this work, we demonstrate a voltage scalable 0.26V, 64kb SRAM with 512 cells per bitline using several circuit techniques that can be activated at ultra-low voltages to expand the operating range. Those novel techniques include the following: (i) 8T SRAM cell utilizing the Reverse Short Channel Effect (RSCE) for improved writability and read performance; (ii) Marginal Bitline Leakage Compensation (MBLC) scheme for improved read sensitivity and precharge elimination; (iii) floating Read BitLines (RBL) and Write BitLines (WBL) to minimize bitline leakage; (iv) deep sleep mode for reducing standby cell leakage; and (v) automatic read wordline pulse width control for improved bitline sensing margin and lower leakage power.

3.4.2 8-T SRAM Bitcell Design

Fig. 3.20. shows the schematic and layout of the proposed 8T SRAM cell. A minimum sized conventional 6T SRAM cell structure is used for data storage and write operation. Two NMOS devices are used for the read path with the cell node being isolated from the read bitline (RBL). The proposed 8T SRAM cell uses a 3X longer channel length in the write access devices and a 2X longer channel length in the read path devices (Fig. 3.20). The 3X longer channel length offers a 2.4X higher drive current (Fig. 3.21 (b)). However, the improved current drivability reduces the stability of the half-selected cells. Circuit techniques such as the write-back scheme that we proposed in [21] can be adopted to remove this issue. (The write-back scheme was not implemented in this test chip.) The 2X longer channel length in the read path devices improves the read speed without incurring additional cell area penalty. The proposed SRAM cell also has a smaller variation due to the larger device sizes [48]. The proposed 8T SRAM cell utilizing RSCE has an area overhead of 20% compared to a conventional all minimum sized device 8T cell (Fig. 3.20) [48].

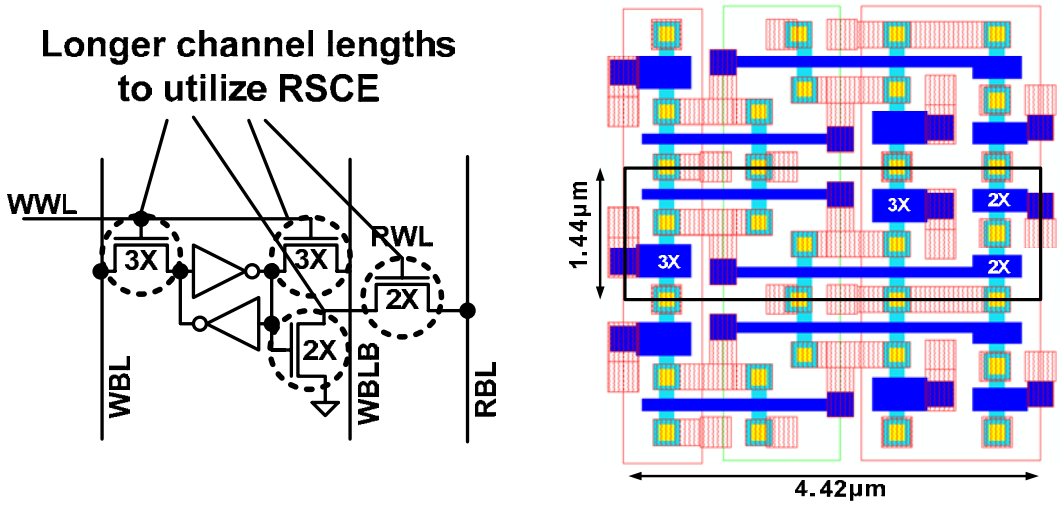
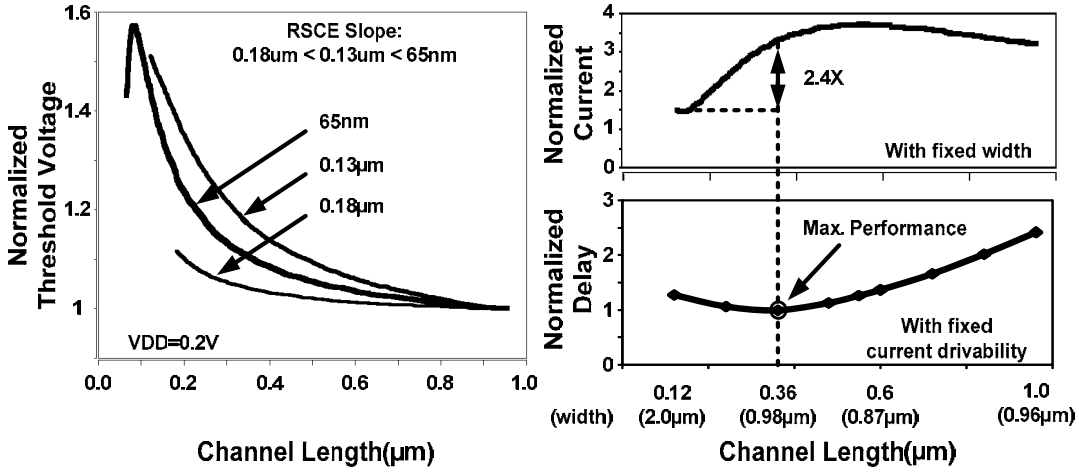


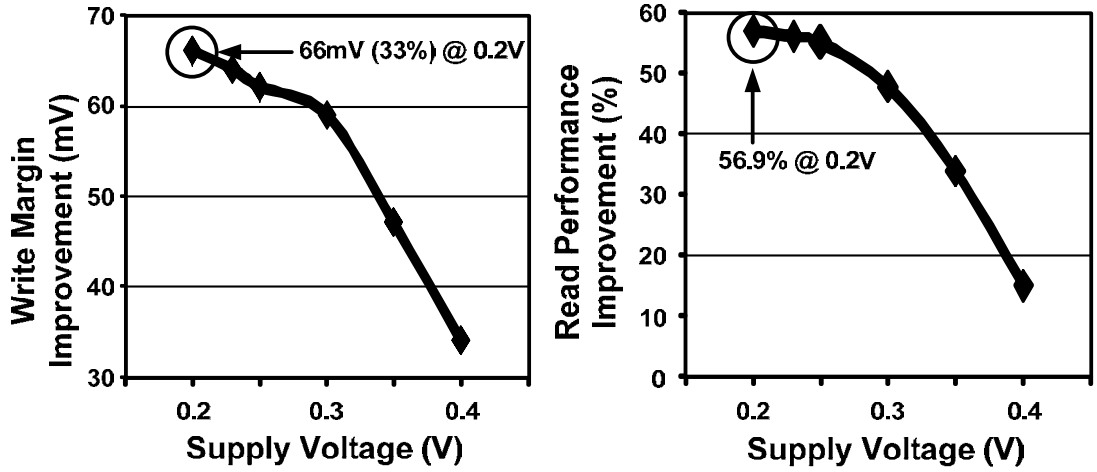
Figure 3.20 Schematic and layout of the proposed 8T SRAM cell utilizing RSCE.



(a) (b)

Figure 3.21 (a) Normalized V_{th} versus channel length shows that RSCE effect is more severe in scaled technologies. (b) Normalized current drivability and delay versus channel length.

Fig. 3.22 shows the simulated results of write margin improvement and read performance. Compared to previous 8T cells, the proposed cell improves write margin by 66mV (33%) and boosts read performance by 56.9% at 0.2V without any increase in the bitline capacitance or the need for additional peripheral circuitry. Utilization of RSCE for improving current drivability is effective when the supply voltage is around or below V_{th}. The improvement of write margin and read performance becomes more significant as the supply voltage decreases to these levels because of the stronger impact of RSCE on device current.



(a) (b)

Figure 3.22. (a) Write margin improvement at different supply voltages by utilizing RSCE. (b) Read performance improvement utilizing RSCE.

3.4.3 Marginal Bitline Leakage Compensation (MBLC) Scheme

At low supply voltages, transistor Ion-to-Ioff ratio decreases exponentially, which can cause the bitline leakage current to become significant compared to the SRAM cell read current. This makes it increasingly difficult to detect the cell data, as the inactive cells' leakage current can offset the read bitline voltage level. In addition, the amount of the bitline leakage is a function of the column data, which makes it even more challenging to distinguish the SRAM cell current from the bitline leakage current. To tackle this issue, Agawa et al. proposed a bitline leakage current compensation scheme using analog circuitry and MOS capacitors [49]. In this technique, the bitline leakage of each accessed column is measured during the precharge time using a PMOS diode. The diode voltage drop is stored in a capacitor and is used to inject an equal compensate current to the bitline when the read wordline signal is asserted. However, this technique cannot be used when the supply voltage is near or lower than the threshold voltage, as the voltage drop cannot be reliably sensed. In addition, the peripheral circuitry required for each bitline costs a significant area overhead for the SRAM. In this work, we propose a Marginal Bitline Leakage Compensation (MBLC) technique suitable for bitline leakage compensation in ultra-low voltage SRAMs.

The MBLC scheme shown in Fig. 3.23 compensates for the RBL leakage in the unaccessed cells using a replica bitline with dedicated control circuits. The RBL voltage is tuned to settle just above the Sense Amplifier (SA) trip point by turning on the marginal compensation devices, which is based on the replica bitline circuit. When a logic '0' is read, only a small swing is required to change the SA output, which is

beneficial when the cell current is comparable to the bitline leakage current. The logic level of RBL during read operation is decided by the static balance between the cell read current (I_{cell}), the pull-down leakage current (I_{bl_leak}), and this marginal compensation current (I_{cmp}) as shown in Fig. 3.23. The marginal compensation current should be large enough to produce logic '1' for the worst case pull-down leakage current, while still being small enough to produce logic '0' for the pull-down cell current and the smallest bitline leakage. The replica bitline generates the marginal compensation current to be used in an array (Fig. 3.23).

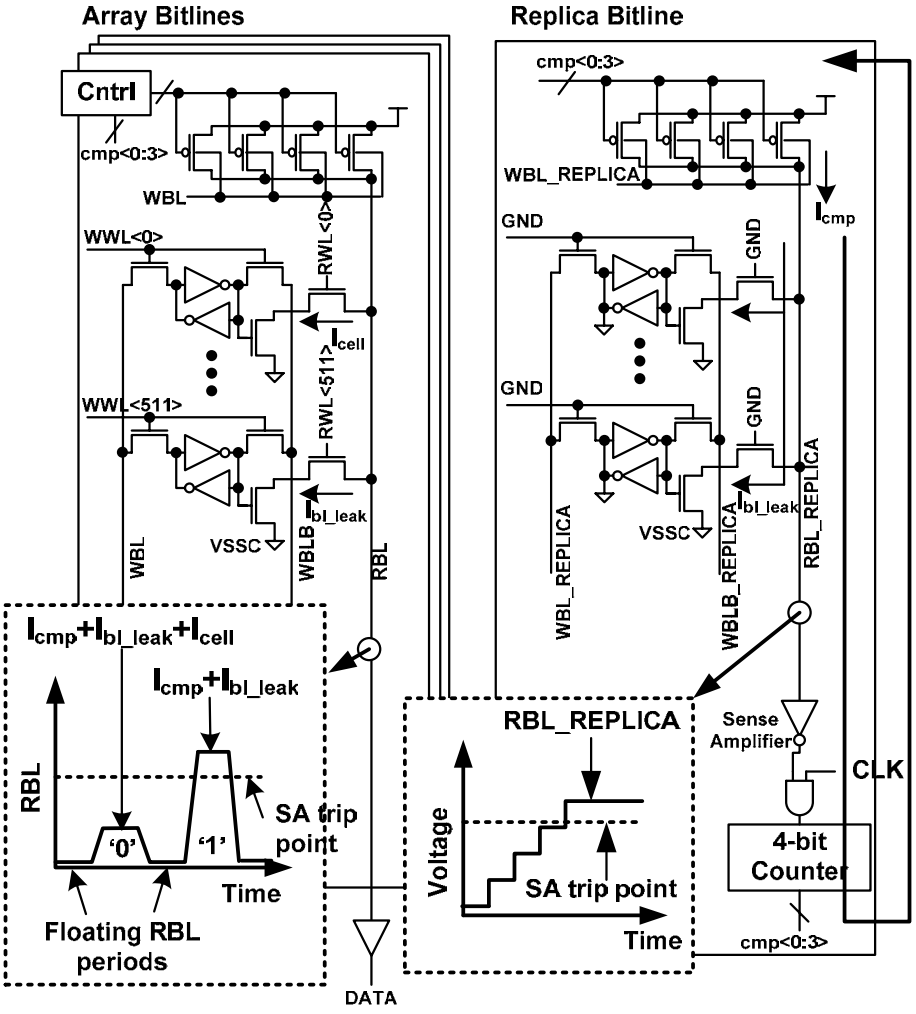


Figure 3.23 Marginal Bitline Leakage Compensation (MBLC) scheme.

A feedback loop controls the strength of the marginal compensation current charging RBL_REPLICA up to a point where the SA output switches to ‘1’ by progressively turning on the marginal compensation devices. Cell data in the replica bitline is hardwired to generate the maximum bitline leakage. This configuration was chosen to emulate the large bitline leakage current and small RBL sensing margin condition. Initially, the SA output is ‘0’ because bitline leakage current pulls down RBL_REPLICA and cmp<3:0> is initialized with ‘1’s, turning off all marginal compensation devices. An increasing number of compensation devices are then turned on raising the level of RBL_REPLICA until the SA output switches to ‘1’. The digital code from the replica bitline is used in array bitlines to generate the compensation current. The compensation devices are activated only during the short read windows because RBL voltage is determined by the static current balance. This is different from the conventional strong-inversion SRAM read operation where the device Ion-to-Ioff ratio is sufficiently large and bitline voltages are decided by the dynamic operation, discharging the precharged bitlines conditionally.

Additional margin for '1' can be built into the SAs by selectively turning on extra precharge devices in the accessed bitline and providing a more compensation current. This margin can be used to make all RBLs have a large enough compensation current to reliably generate data '1' without a pull-down cell current, accounting for within-die variations. The marginal precharging level can also be trimmed by changing the trip point of the SA. Fig. 3.24 shows the simplified schematic of the SA implemented in our design. By turning on additional devices here, we can change the SA trip point, which in turn adjusts the marginal precharging level.

However, a fixed compensation current can be problematic because the ideal compensation currents for the bitlines can be different from the replica bitline leakage due to the data dependant bitline leakage current. Section II.C describes the column data dependency of the compensation current and a circuit technique to deal with this issue.

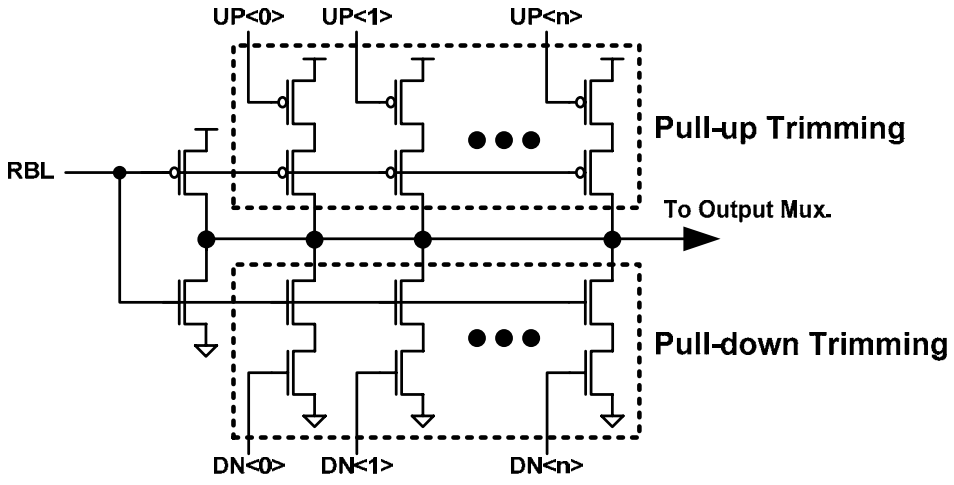


Figure 3.24. Schematic of sense amplifier with trip point trimming circuits

3.4.4 Column Data Dependency of MBLC Current

The optimal compensation current depends on the data pattern in a column because the amount of bitline leakage is also a function of this data. Since the replica bitline generates the marginal compensation current for the column data pattern resulting in the worst case bitline leakage, a method for incorporating column data dependency must be devised. In this work, data dependency was accounted for by connecting the body of the compensating PMOS devices to the floating WBL voltage, which is also determined by the data pattern stored in the SRAM column. The floating WBL is possible because this bitline does not need to be precharged during non-write operations as it does in conventional SRAMs.

The column data patterns of the replica bitline and the array bitlines are shown in Fig. 3.25. The best case bitline has the same data as the replica bitline. In this scenario, the compensation current will be identical to the bitline leakage current. On the other hand, the column data pattern giving rise to the minimum bitline leakage causes the worst-case discrepancy between the compensation current and the actual bitline leakage.

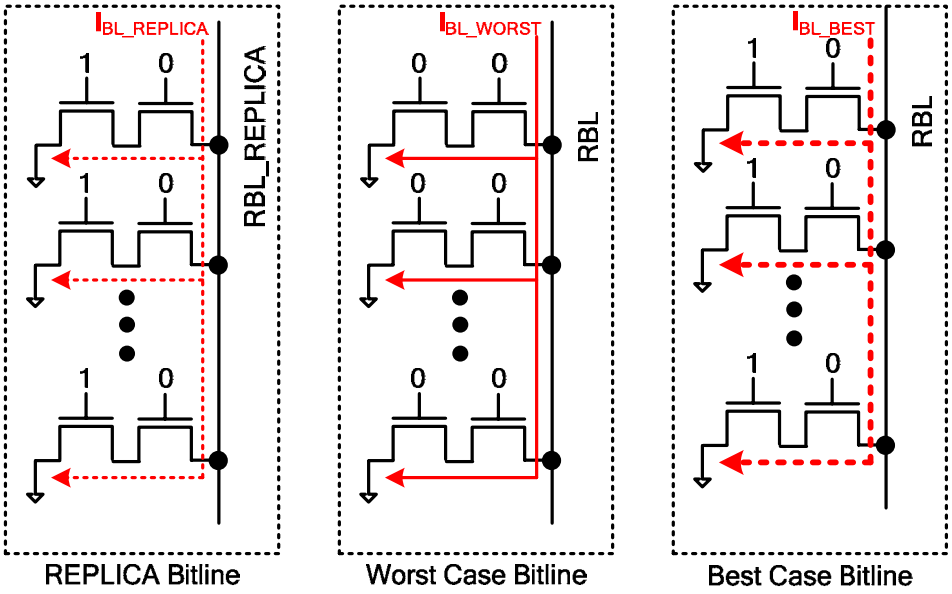


Figure 3.25 The best case sensing margin occurs when the accessed bitline and the replica bitline have identical leakage currents. Conversely, the sensing margin is worst for an all-‘0’ column which has the minimum bitline leakage.

Fig. 3.26 illustrates the change of RBL voltage due to column data patterns, and the principle of using body biasing to incorporate this dependency. The accessed column and replica column have the same RBL signal levels when they contain the same data (Fig. 3.26 (a)). However, the difference in the column data pattern will raise the RBL level due to the imbalance between the compensation current and the bitline leakage current, which degrades sensing margin (Fig. 3.26 (b)). This is inevitable as the replica bitline has to be hardwired with the data pattern generating the largest compensation current for reliable read operations with large bitline leakage current. To solve this problem, the floating WBL voltage which changes with the column data is used as the body bias of the marginal compensation devices.

The floating WBL voltage rises with more cells in the column storing data '1', which in turn decreases the amount of marginal compensation current by weakening the forward body bias in the PMOS compensation devices. The decreased compensation current cancels out the difference between the required bitline leakage current and the provided compensation current, which makes the RBL similar to that in the replica bitline (Fig. 3.26 (c)).

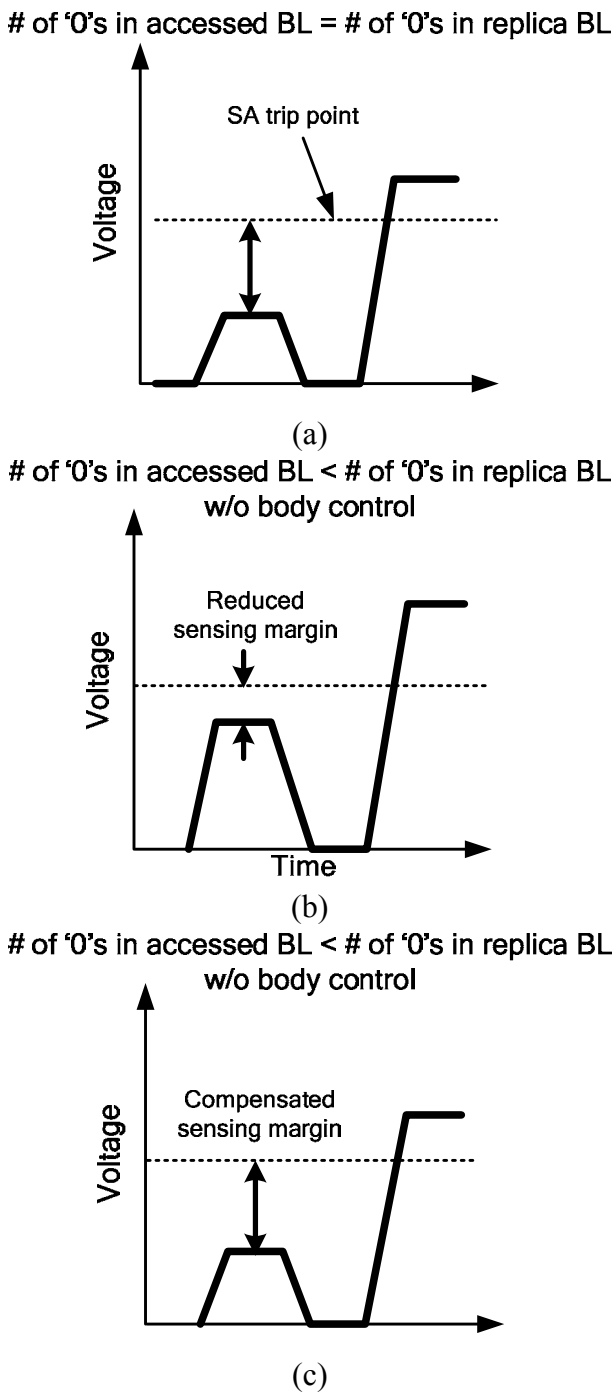
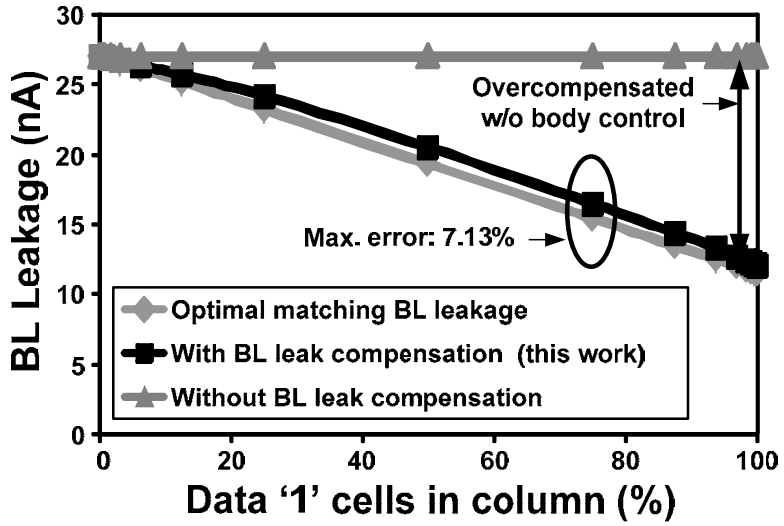
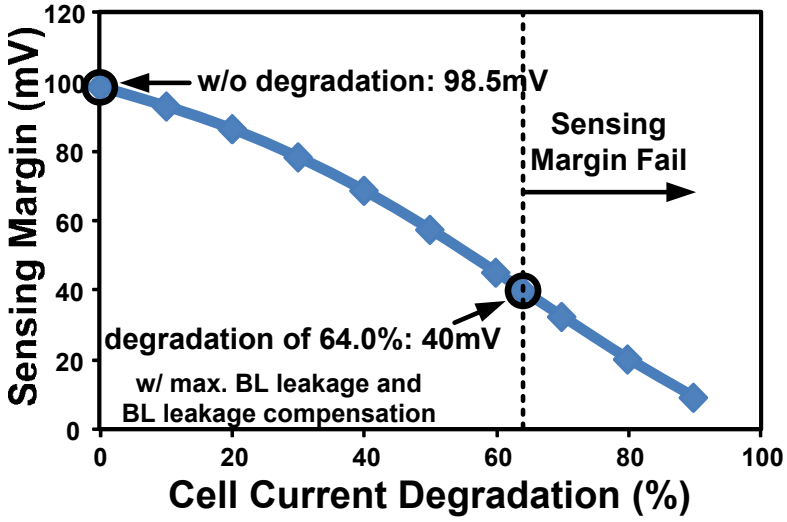


Figure 3.26. (a) RBL voltage when the accessed column has the same data as replica column. (b) RBL voltage with different column data. (c) RBL voltage with different column data after applying optimal body biasing (this work).

Simulation results for this compensation scheme are illustrated in Fig. 3.27 (a). As shown here, the body bias control using the floating WBL tracks the column data pattern and moves the compensation current close to the optimal matching bitline leakage. The maximum error in the compensation current was only 7.13% without considering within-die variations. A smaller cell read current due to within-die variations can increase the RBL level reducing the sensing margin for data '0'. Fig. 3.27 (b) shows the impact of cell current degradation on the sensing margin. These simulation results show that the MBLC scheme ensures a correct operation until the cell current is reduced by 64 %.



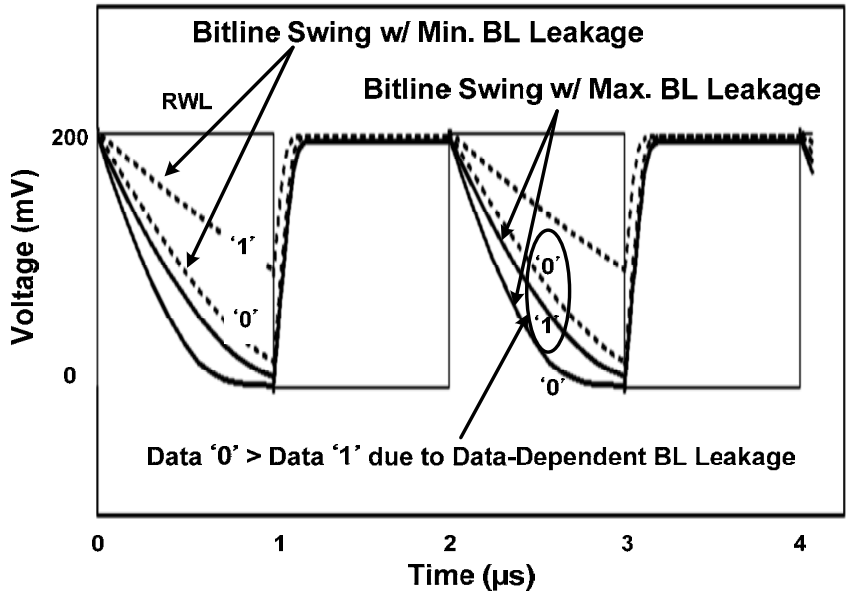
(a)



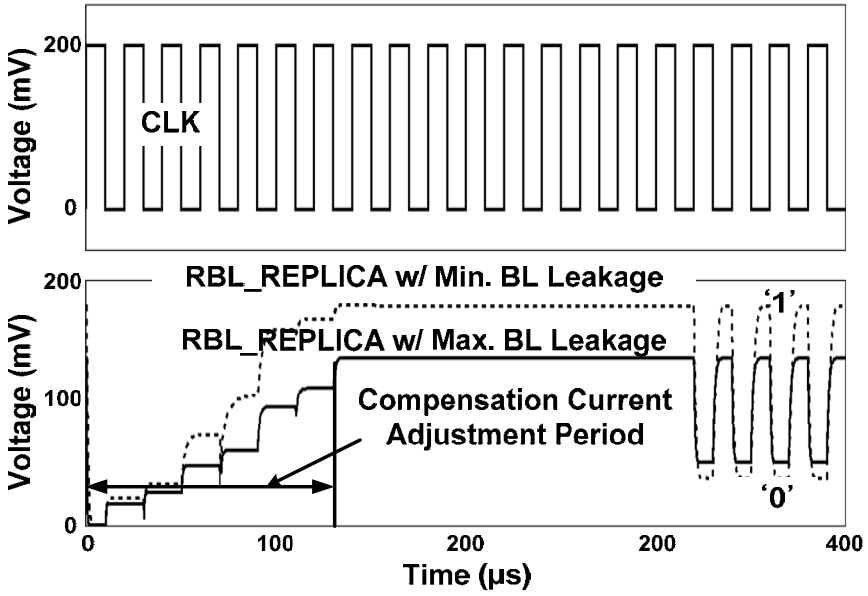
(b)

Figure 3.27 (a) Data dependent bitline leakage compensation using the floating write bitline voltage as the body bias. The nominal corner is used for simulation with the supply level of 0.2V at room temperature. (b) Impact of cell current degradation on sensing margin.

Fig. 3.28 compares the proposed MBLC scheme to the conventional precharged bitline scheme during read operations. In the conventional scheme (Fig. 3.28 (a)), the bitline leakage discharges RBL at a rate comparable to the cell current, which reduces bitline sensing margin. Furthermore, the RBL discharging speed of data '1' with the maximum bitline leakage is faster than that of data '0' with the minimum bitline leakage current. A sense amplifier cannot detect the read data correctly from a single ended bitline in this case. However, the proposed MBLC scheme generates a compensation current that tracks the column data and static bitline levels, making the bitline sensing margin constant over time. Fig. 3.28 (b) shows RBL_REPLICA waveforms with two different hardwired patterns. The body bias control of the compensation devices enhances the sensing margin of data '0' in the minimum bitline leakage condition. The change of RBL_REPLICA is shown as the MBLC control circuit adjusts the compensation current.



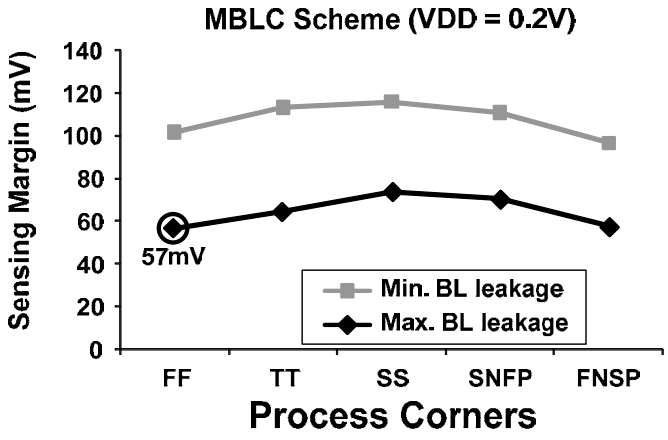
(a)



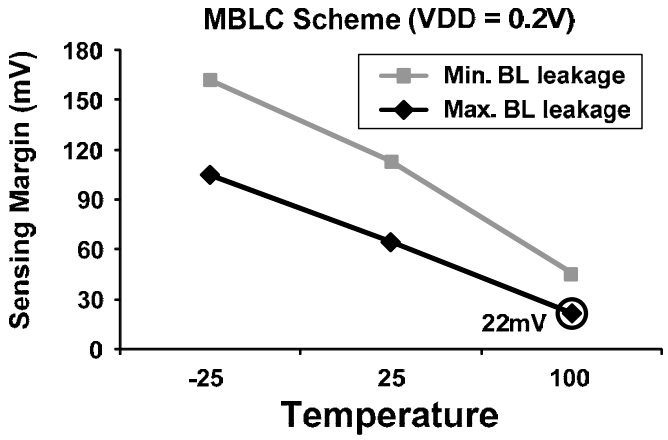
(b)

Figure 3.28 (a) RBL waveforms for a conventional precharged bitline. (b) RBL_REPLICA waveforms of the proposed MBLC scheme for maximum and minimum bitline leakage cases.

Simulated RBL sensing margins for different process corners and temperatures are illustrated in Fig. 3.29. The sensing margin decreases as temperature increases since the bitline leakage increases faster than the cell current.



(a)



(b)

Figure 3.29 The proposed MBLC scheme improves sensing margin compared with the conventional precharged bitline. The conventional precharged bitline fails in read operations. (a) Sensing margin of this work at different corners. (b) Sensing margin of this work at different temperatures.

3.4.5 Floating Read/Write Bitlines for Active Leakage Reduction

Leakage current in inactive memory cells accounts for most of the SRAM power consumption. Circuit techniques for leakage control are particularly critical for reducing the total power consumption in the sub-threshold region. RBL leakage is one of the most dominant leakage components and is inevitable in conventional memories where bitlines are precharged to VDD. In our design, the RBLs are left floating without being precharged whenever the Read WordLine (RWL) is low. This is possible because the RBL level is decided by the static current balance between the bitline leakage current, compensation current and cell read current. During the read operation, the MBLC scheme provides the compensation pull-up current to generate logic high or low levels in the RBL with large sensing margin. The static operation in deciding RBL makes the precharging operation unnecessary. During the non-read operation, however, the floating RBL level is determined by the strong pull-down leakage current formed by the read path in the SRAM cells and the negligible pull-up leakage current through the compensation devices. This makes RBL converge to GND, eliminating the leakage current from RBL in Fig. 3.30 (top).

Like the floating RBL, write bitlines (WBL and WBLB) are also left floating when WWL is low so that they will automatically settle to levels which minimize the leakage current as shown in Fig. 3.30 (middle). Forcing a specific voltage will break the balance of leakage current flowing through pull-up and pull-down devices and make one larger than the other. During a write operation, WBL is driven by the write driver. Therefore, precharging WBL is also redundant. The proposed scheme has no energy overhead during the write operation compared to the conventional scheme due

to the same voltage swing. This is based on the assumption that the probability of writing a data ‘1’ and a data ‘0’ are the same. WBL and WBLB are not at the same level. If WBL is higher than WBLB, writing a ‘0’ to WBL and a ‘1’ to WBLB will consume more energy than the conventional write operation. Writing the opposite data, however, will save the energy because both WBL and WBLB have smaller swings. Assuming the same probability of writing a ‘0’ and a ‘1’, the energy consumption in write operations can be calculated by the equations in Fig. 3.30 (middle). A leakage power reduction using the floating RBL and WBL is summarized in Fig. 3.30 (bottom). A total SRAM leakage reduction of 44% to 60% can be obtained by using the floating RBLs and WBLs. The variations in power reduction happen because the different column data pattern changes the floating WBL voltage, which also changes the leakage reduction.

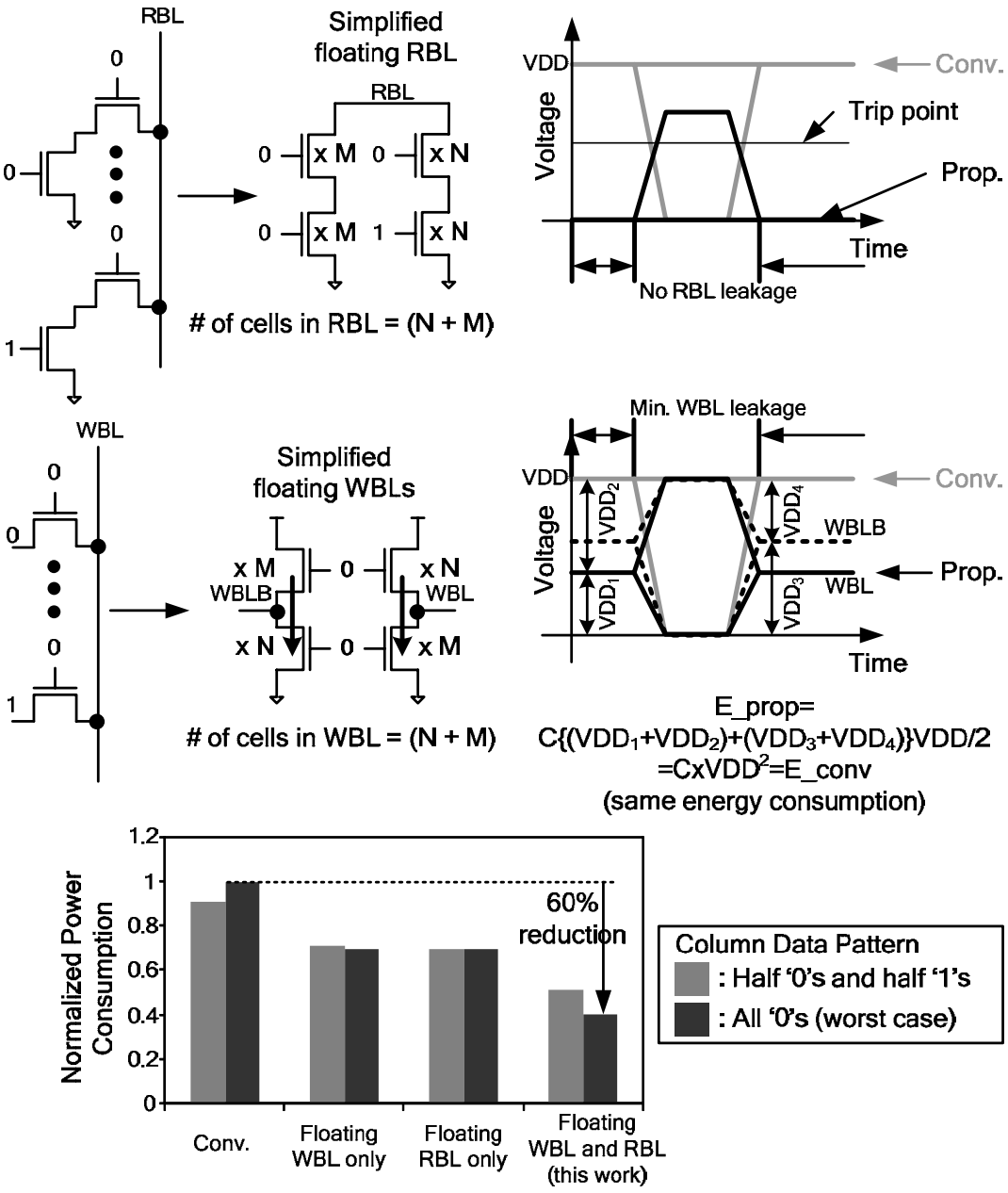


Figure 3.30 Power reduction using floating read and write bitlines. It is assumed that the probability of writing a '0' is equal to that of writing a '1'.

3.4.6 Deep Sleep Mode

Sleep transistors are popular for reducing SRAM leakage current in standby mode by collapsing the virtual supply rails [50][51]. However, due to the fact that the voltage margin is already close to the functionality limit, it is difficult to use conventional footer sleep transistors for sub-threshold SRAM designs. In this work, we propose a deep sleep mode illustrated in Fig. 3.31 (b) to reduce the standby leakage in sub-threshold memory designs. VDDC and VSSC represent virtual supply and virtual ground voltages of the SRAM array. Instead of collapsing VSSC for a sleep mode as shown in Fig. 3.31 (a), the proposed scheme raises both VDDC and VSSC while keeping the cell voltage, VDDC-VSSC, constant to reduce leakage while maintaining the same cell stability in the deep sleep mode. SRAM cell leakage is reduced due to the negative VGS in the write access transistors and the increased threshold voltage of the pull-down NMOS devices due to the reverse body bias. However, raising both VDDC and VSSC increases the floating write bitline voltages (WBL and WBLB) because they are decided by the column data pattern and the SRAM cell node voltages. If both VDDC and VSSC are raised excessively, the pull-up path in the interfacing circuit becomes leaky and a current starts to flow from the write bitlines to the virtual supply nodes.

Fig. 3.32 highlights the leaky current path and illustrates the normalized SRAM leakage reduction using the proposed deep sleep mode. A leakage current decreases as increasing VDDC and VSSC. By applying an optimal supply voltage (VDDC=0.83V, VSSC=0.60V), 87% reduction in the cell leakage was obtained during the deep sleep mode. Half ‘0’s and half ‘1’s are assumed in the simulation. Raising VDDC and VSSC beyond the optimal point increases the leakage current exponentially.

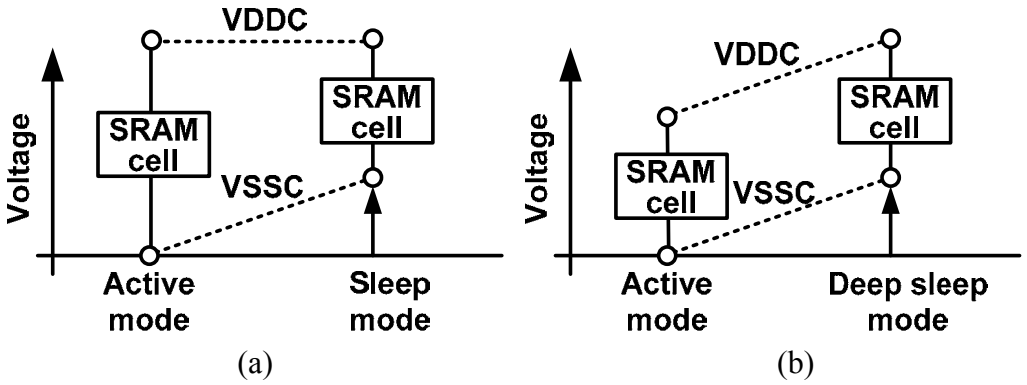


Figure 3.31 (a) Conventional sleep mode. (b) Proposed deep sleep mode.

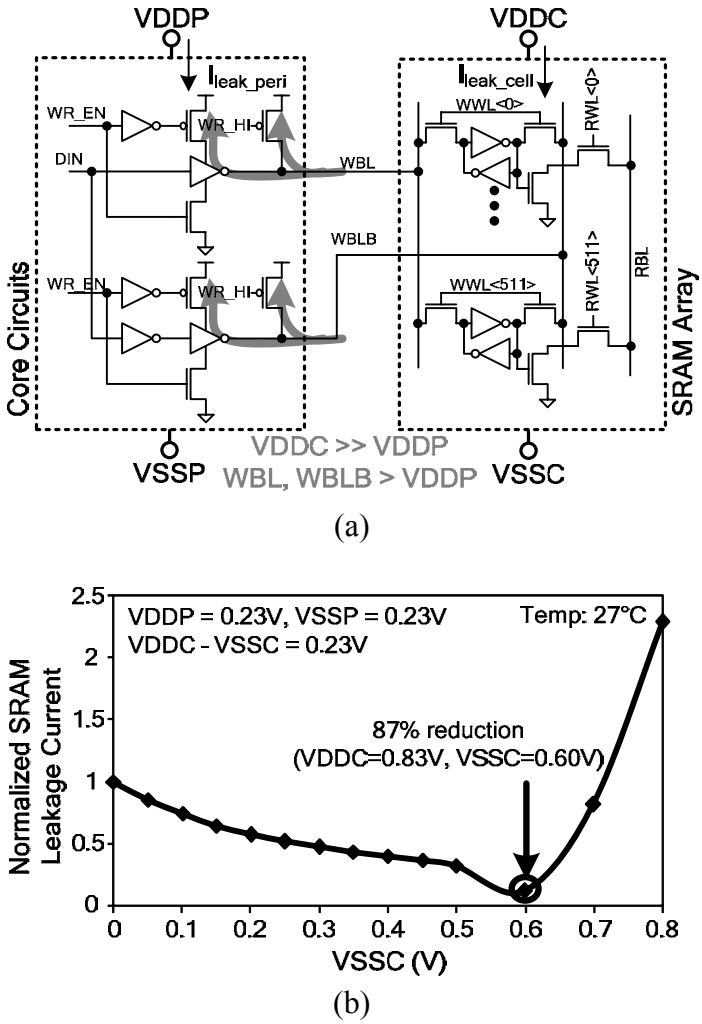


Figure 3.32 (a) Leaky current path at the interface circuit in deep sleep mode. (b) Simulated leakage reduction.

3.4.7 Automatic Wordline Pulse Width Control

The RWL activation time should be long enough for the sense amplifier to function reliably, but it should be turned off soon after the read operation is finished to cut off the marginal compensation current and reduce the power consumption. In order to address this tradeoff, we propose a scheme to automatically adjust the read wordline pulse width based on PVT variations (Fig. 3.33). A replica bitline generates the wordline pulse width needed for the SA to precisely capture the read data. The cell data in the replica bitline is hardwired so that an RBL_REPLICA pulse is generated for each read cycle. The delayed SA output RD_FIN from the replica bitline disables the read wordline and shuts off the marginal precharge devices. By doing so, the RWL is only enabled until the read operation is completed, saving the RBL leakage power. Another issue to be considered is the impact of within-die variations on the wordline pulse width. Fig. 3.34 shows a failure scenario where read data, D<i>, arrives later than the read data from the replica bitline due to within-die variations. To address this problem, an eight FO1 inverter delay chain is inserted in the replica bitline path to provide enough timing margin for correct a read operation. With this additional timing margin, the proposed SRAM is tolerant to a cell current variation of up to 50%.

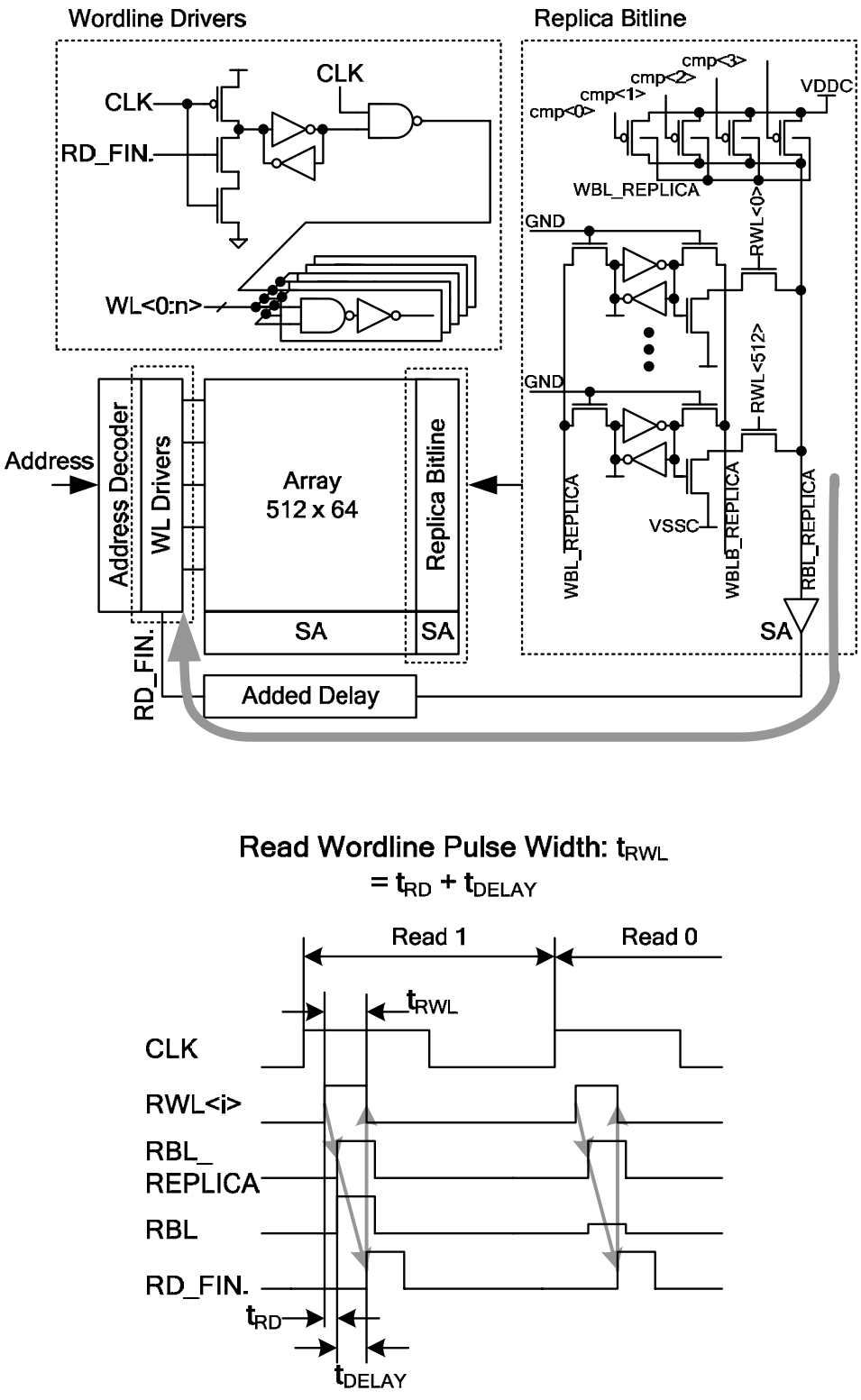


Figure 3.33 Read wordline pulse width control for PVT tracking.

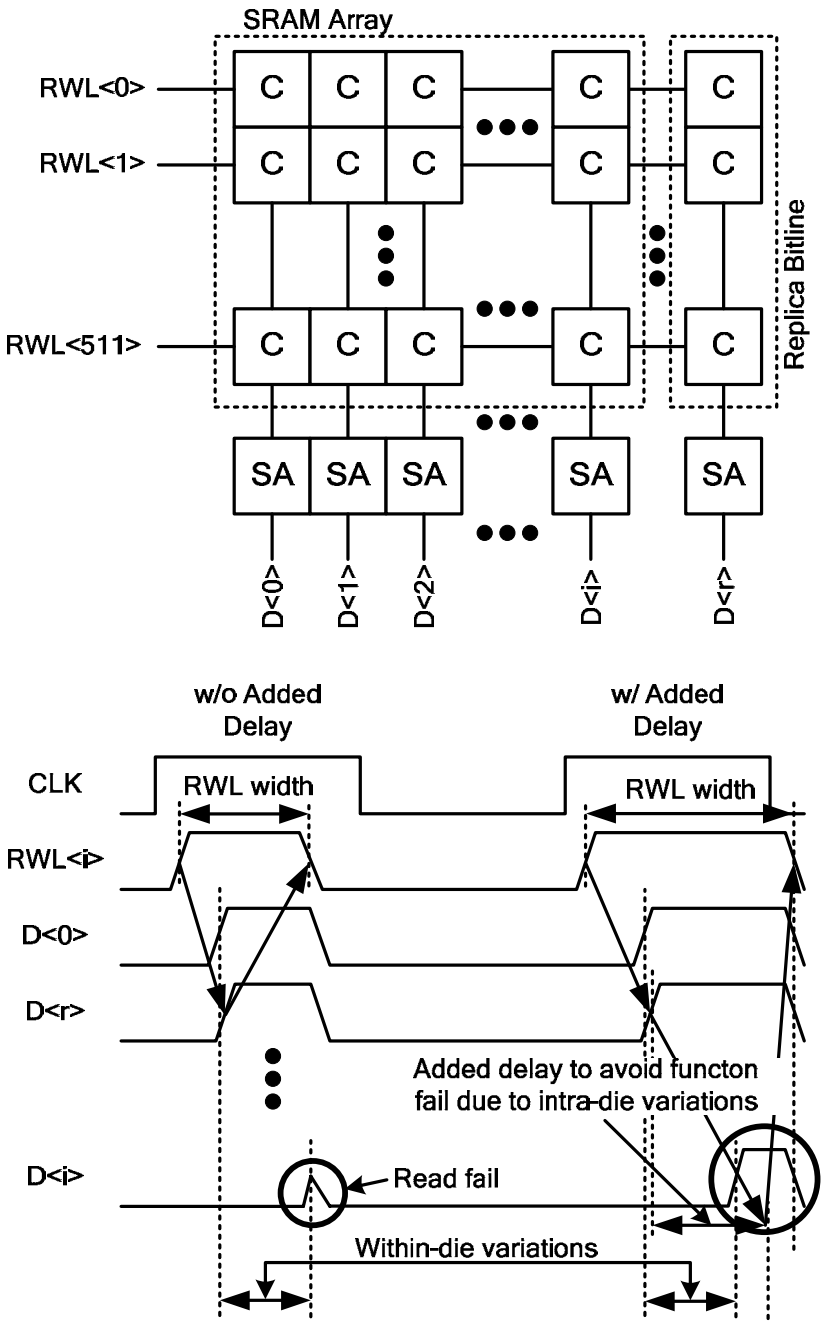


Figure 3.34 Within-die variation causes read failures when array bitlines are slower than the replica bitline. Failure rate is reduced by adding more delay to give enough timing margin under within-die variation.

3.4.8 Test Chip Implementation and Experimental Results

A 64kb SRAM was fabricated in a 130nm CMOS technology with a nominal supply voltage of 1.2V. Fig. 3.35 shows the architecture of the implemented SRAM. It consists of two SRAM cell arrays, each with 512 rows and 64 columns, 16 IOs, and replica bitlines and added delay for the proposed MBLC and wordline pulse width control. Each cell array is divided into eight sub-blocks generating one bit per sub-block, and a sub-block is composed of eight columns.

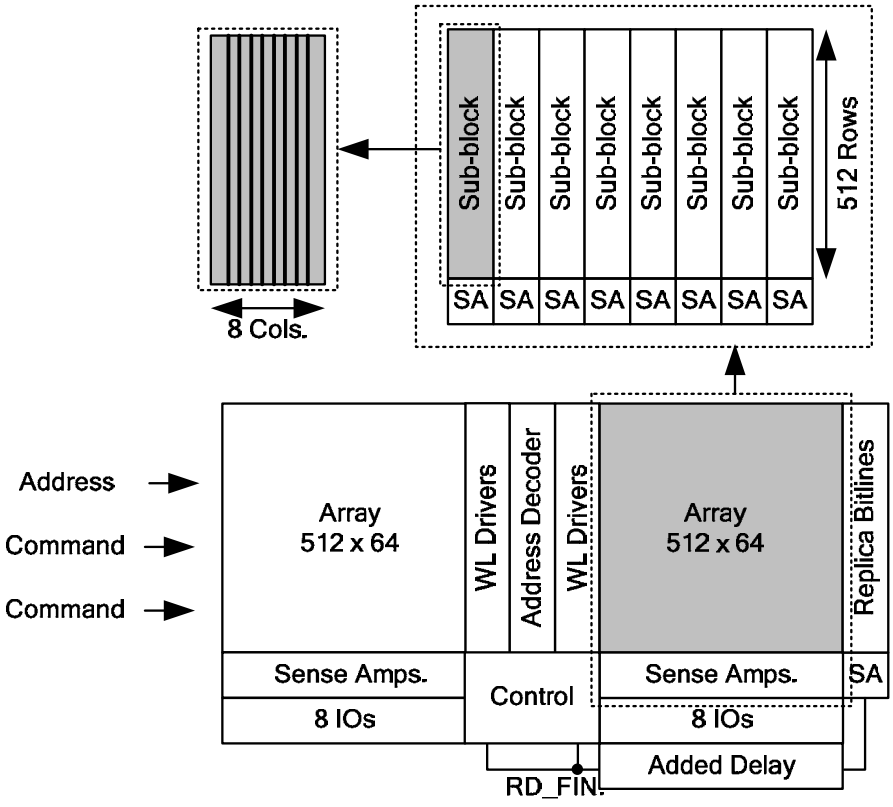


Figure 3.35 Test chip architecture.

Fig. 3.36 shows the measured power consumption and leakage current. We observed SRAM cells functional down to 0.23V running at 100 kHz and consuming 4.3μW (Fig. 3.36 (a)). At 0.4V, the operation frequency was 6.7MHz with a power consumption of 10.8μW. The measured SRAM leakage currents from different dies are shown in Fig. 3.36 (b). The leakage current in SRAM array is around 5X of that in peripheral circuits. Variation in leakage current was 2.0X at 0.3V due to its exponential dependency on device threshold voltage. The normalized leakage current measured at different temperatures is shown in Fig. 3.36 (c). The leakage current at 110°C is 3.4X larger than that at 27°C when the supply voltage is 0.23V.

Fig. 3.37 illustrates the normalized leakage current reduction achieved using the proposed deep sleep mode. The total SRAM leakage including the array and peripheral components was reduced by 69% in the deep sleep mode by raising the VSSC to 0.45V while maintaining the cell voltage of 0.23V. The initial leakage reduction is large when raising VSSC due to the strong negative Vgs effect in conjunct with the reverse body biasing effect. 58% leakage reduction was achievable using a VSSC of 0.2V during the deep sleep mode. The smaller offset in VDDC and VSSC improves the efficiency and area overhead of the charge pumps that can be used to generate the voltage on-chip [52]. In this test chip, we used an external supply for the higher supply voltages needed during the deep sleep mode.

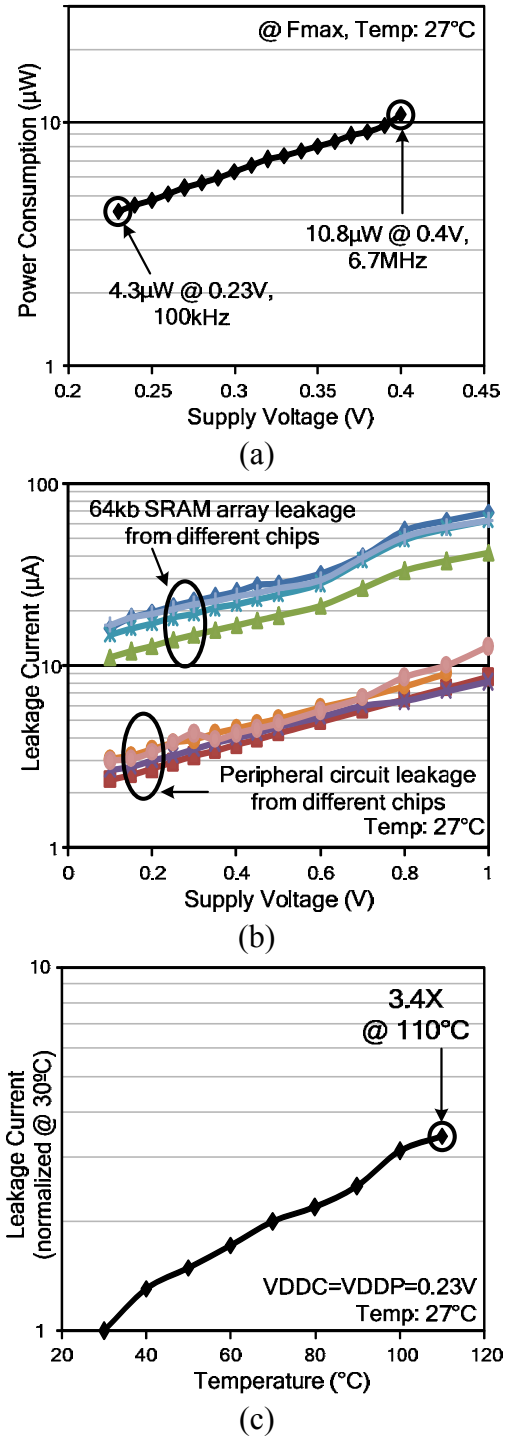


Figure 3.36 (a) Measured SRAM total power consumption. (b) SRAM leakage current varying supply voltage. (c) Normalized leakage current at different temperature. (d) Leakage current reduction in deep sleep mode.

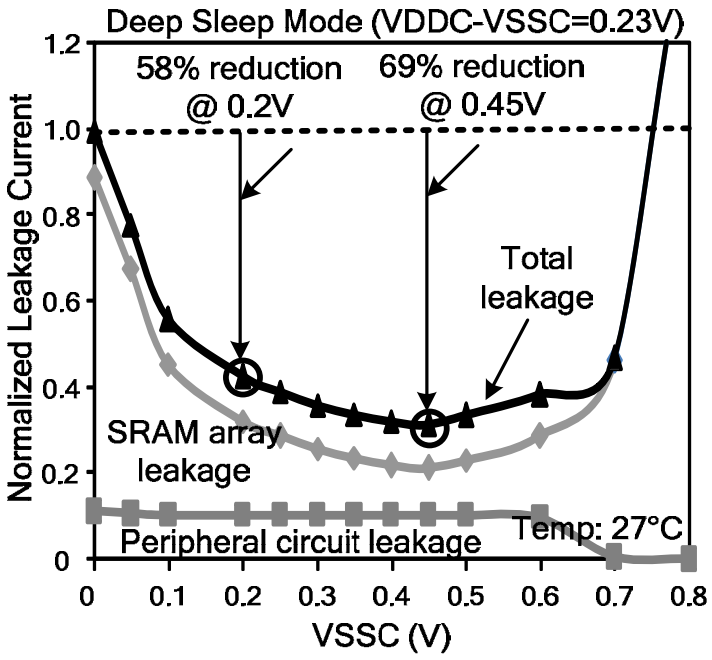


Figure 3.37 Leakage current reduction in deep sleep mode.

Fig. 3.38 shows the shmoo plot of a single SRAM cell when the proposed MBLC scheme is on and off. When the MBLC scheme is off, a conventional fixed precharge device is used. The V_{min} of the SRAM cell under test is improved from 0.28V to 0.23V by activating the MBLC scheme.

Fig. 3.39 illustrates the measured V_{min} of each SRAM cell for read and write operations from an 8-by-8 mini subarray. V_{min} for read operation ranges from 0.24V to 0.26V and V_{min} for write operation ranges from 0.18V to 0.20V. We have also tested the feedback control circuit for the MBLC scheme which compensates the bitline leakage on-the-fly. The 4 bit counter used in the MBLC requires up to 16 clock cycles to generate the optimal precharge strength.

Fig. 3.40 shows SA outputs with two different trip points to mimic two different compensation currents. It is shown that a SA with a higher trip point requires additional cycles to turn on more number of compensation devices. Similarly, more devices should be turned on for a larger bitline leakage current due to process variations.

The die photo and chip performance summary are given in Fig. 3.41. The proposed MBLC and read wordline pulse width control scheme incur an area overhead of 1.3%.

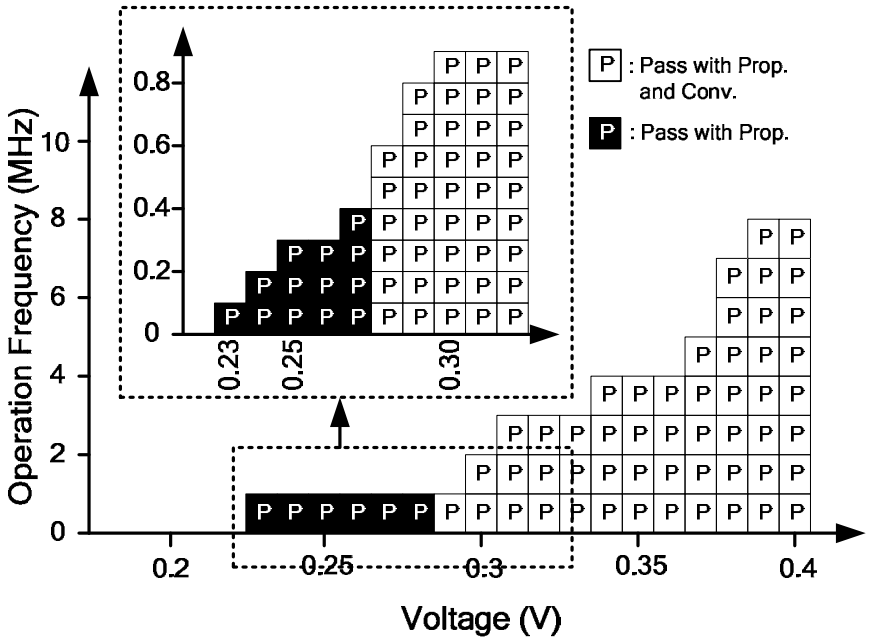


Figure 3.38 Shmoo plot for an SRAM cell with a 0.23V V_{min}.

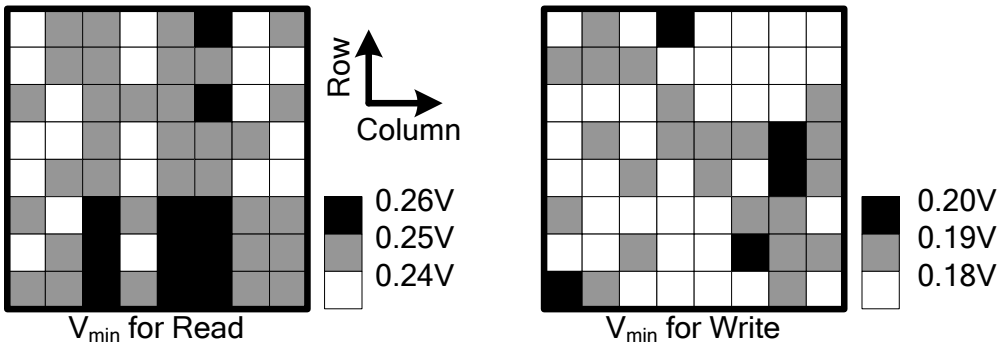


Figure 3.39 V_{min} for read and write from an 8-by-8 mini subarray.

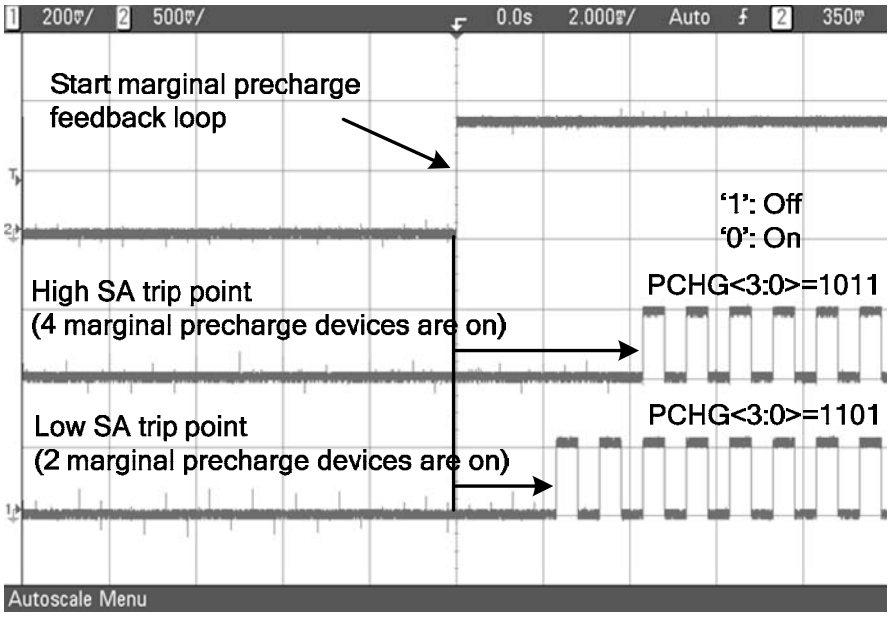
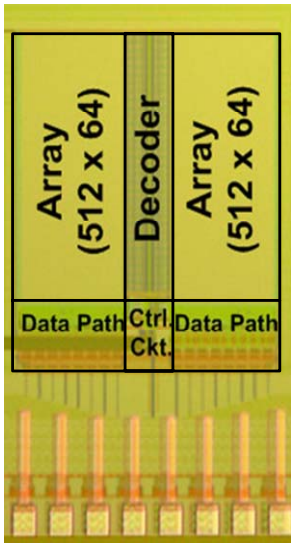


Figure 3.40 Output waveforms from marginal bitline leakage compensation control circuit.



Technology	130nm 8-metal CMOS
SRAM Area	0.72x0.85mm ^{2}
VCC	≥0.26V
Size	64kb
Performance	100kHz @ 0.23V, 27ºC 15MHz @ 0.60V, 27ºC
Power Consumption	**4.3µA @ 0.23V, 27ºC** **34µA @ 0.60V, 27ºC**
Deep Sleep Mode Leak.	**69% reduction**

Figure 3.41 Chip microphotograph and performance summary.

3.5 Conclusions

Sub-threshold SRAMs are becoming more important in applications where energy dissipation is the primary design constraint. This paper proposes various circuit techniques for enabling reliable sub-threshold SRAM design.

First, we implemented a 0.2 V 480kb sub-threshold SRAM in a 130-nm process technology. A 10-T SRAM cell is proposed to eliminate the read failure caused by data-dependent bitline leakage. A VGND replica scheme is proposed to track the logic 'low' level of the bitlines under PVT variations, which allows us to achieve the maximum read sensing margin. The strong RSCE in the sub-threshold region was utilized to improve cell writability, reduce power consumption, improve logic performance, and enhance circuit immunity to process variations. By combining these proposed circuit techniques, we were able to implement a fully functional sub-threshold SRAM with 1k cells per bitline operating at 0.2 V and 27 °C.

The second version of SRAM has also been fabricated in 130nm CMOS technology. Utilizing RSCE in the read and write ports of the SRAM cell improves write margin and read performance. The MBLC scheme lowers V_{min} by compensating bitline leakage and improving bitline sensing margin. The proposed floating bitline scheme and deep sleep mode improve the leakage current reduction during a normal operation and a standby mode. An automatic read wordline pulse width control scheme improves readability and reduces wasted read power by tracking the PVT variations. The 64kb SRAM with 512 cells per bitline verifies the V_{min} lowering and leakage reduction achieved by the proposed circuit techniques. These techniques

facilitate a superior minimum energy solution through improved leakage reduction and the enhanced SRAM performance.

Chapter 4 On-Chip Circuit Reliability Monitoring Techniques

4.1 Introduction

As CMOS process technology continues to follow an aggressive scaling roadmap, designing reliable circuits has become evermore challenging with each technology node. Reliability issues such as Bias Temperature Instability (BTI), Hot Carrier Injection (HCI), and Time Dependent Dielectric Breakdown (TDDB) has become more prevalent as the electrical field continues to increase in nanoscale CMOS devices. One of the most pressing of these challenges is Negative Bias Temperature Instability (NBTI) [12][13][14][15] caused by the trap generation in *Si-SiO₂* interface of PMOS transistor (Fig. 4.1). Structural mismatch at the *Si-SiO₂* interface causes dangling bonds, which act as interfacial traps. During the hydrogen passivation process that follows oxidation, dangling *Si* bonds are transformed into *Si-H* bonds. These bonds are weak enough to break during device operation, causing *H* atoms to diffuse into gate oxide, and the broken bonds that remain become traps, effectively degrading the drive current of PMOS transistors. NBTI is characterized by a positive shift in the absolute value of the PMOS threshold voltage (|Vtp|), which occurs when the device is stressed (Vgs = -VCC), and this effect is more pronounced at high temperatures. This degradation in Vtp has believed to exhibit a power–law dependency on time, and is an exponential function of the stress voltage level as well as temperature. When the stress conditions are removed (i.e., Vgs = 0), the device

enters a recovery or passivation phase, where *H* atoms diffuse back towards the *Si-SiO_2* interface and anneal the broken *Si-H* bonds, thereby reducing |Vtp| (Fig. 4.1(b) and (c)) [53]-[59].

To estimate the impact of NBTI on circuit performance and eventually design aging-tolerant circuits, accurate measurement of digital circuit reliability is imperative. Previous reliability measurements relied on device probing or on-chip ring oscillator frequency monitoring, which either require an extensive measurement setup or have limited sensing resolution [60][61]. Moreover, they were inefficient in collecting a statistically significant number of data points under various stress conditions, which is crucial in understanding the complexities of aging (e.g. statistical behavior, process and frequency dependencies, etc.).

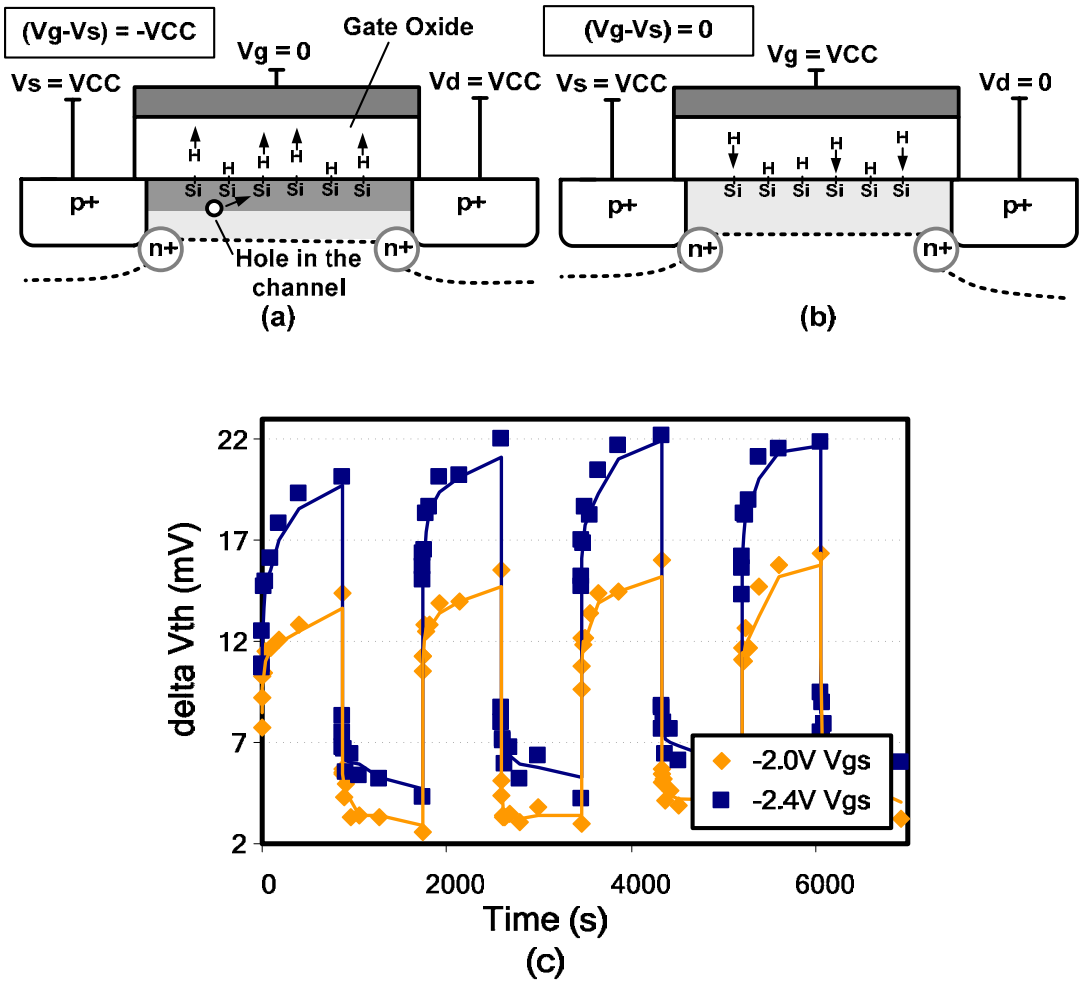


Figure 4.1 Cross section of PMOS device under (a) NBTI stress and in (b) recovery mode. (c) PMOS Vt degradation for alternating stress and recovery periods in 130nm CMOS [53].

4.2 Silicon Odometer: An On-Chip Reliability Monitor for Measuring Frequency Degradation of Digital Circuits

4.2.1 Overview

In this work, we propose an aging monitoring circuit which is capable of taking fast and precise degradation measurements by detecting the beat frequency of a pair of ring oscillators, where only one is placed under stress. This differential measurement method eliminates the effect of environmental variations that plague other approaches, such as changes in temperature and supply voltage. This implementation also facilitates the application of both DC and AC stress signals, allowing the effects of both types of phenomenon to be studied. No specialized measurement equipment is required for the proposed measurement circuit, as on-chip structures have been implemented which convert performance degradation into a simple digital code. The output of the proposed circuit can be used as a feedback to control system parameters such as supply voltage and clock frequency for preventing system failures coming from device aging.

4.2.2 Previous Reliability Monitoring Techniques

The typical approach used when measuring NBTI is to apply stress for a given duration, remove it, then perform an I-V measurement. To accurately measure the effects of NBTI, this measurement must be done quickly to avoid the effects of recovery, which has been reported to occur even between 1μs and 1ms [55]. On-the-fly techniques that minimize the recovery effect have been examined in [54][55][57][61][62].

Denais et al. proposed a measurement technique in which the stress voltage is kept quasi-constant while the linear drain current is measured to monitor device degradation. However, it still requires extra equipment for the accurate measuring of device current under test which limits its application for run-time NBTI monitoring in actual products. In [62], this on-the-fly technique was extended to characterize the recovery after stress conditions are removed.

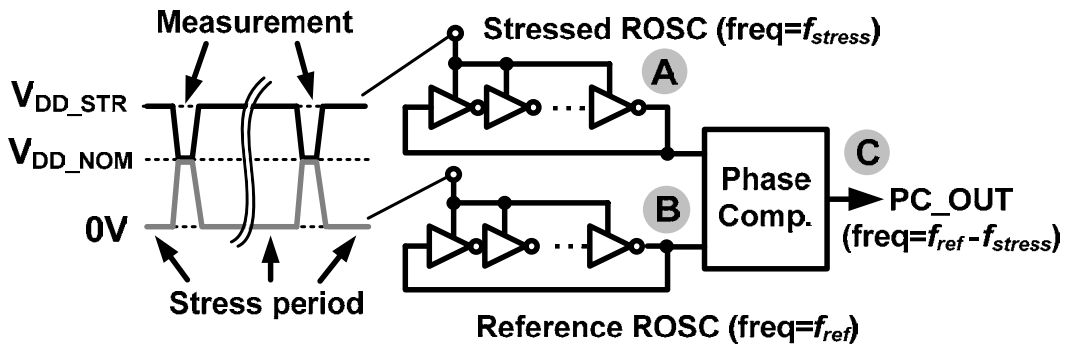
Fernández et al. proposed on-chip circuits for the characterization of device degradation due to AC NBTI stress, which is claimed to be viable up to the GHz range [63]. The authors assert that a high frequency stress signal can be reliably applied to the devices under test, and utilize this information to extract data regarding the frequency dependency of NBTI aging. A frequency degradation monitoring circuit was proposed in [57], where a ring oscillator is stressed and the difference of ring oscillator period before and after the stress is measured. However, this circuit has a low sensing resolution, which requires highly accurate and expensive test hardware, making it an invasive and intractable approach for run-time monitoring of NBTI. In addition, the measurement results are very sensitive to environmental variations, which make it difficult to determine what portion of device degradation is due solely to NBTI.

4.2.3 Beat Frequency Detection Scheme

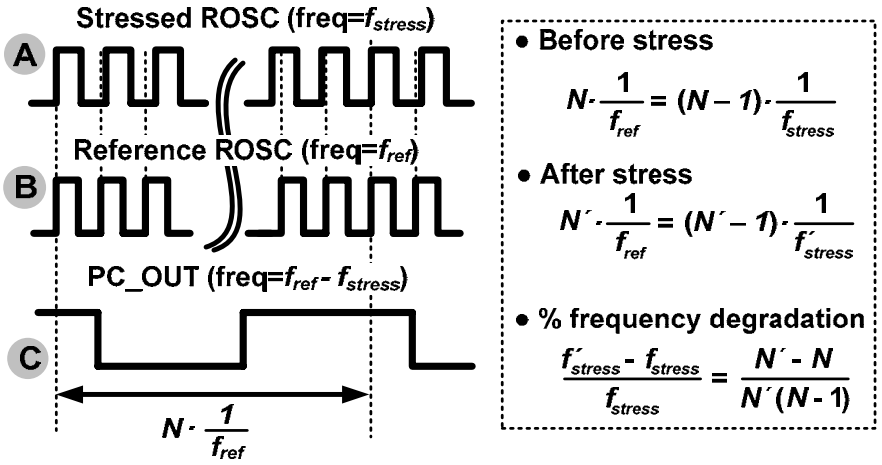
The core circuit for detecting frequency degradation consists of two free-running ring oscillators and a phase comparator as shown in Fig. 4.2 (a). During the stress period, one of the ring oscillators is stressed, while the other remains unstressed. The

supply voltage of the stressed ring oscillator is raised to $V_{DD\text{-}STR}$ during stress periods, and lowered to $V_{DD\text{-}NOM}$ during the periodic measurements, while the supply of the reference oscillator is lowered to 0V, and raised to $V_{DD\text{-}NOM}$ during the stress and measurement periods, respectively. The reference oscillator's supply voltage is grounded during the stress periods to prevent device aging. Once the measurement signal is triggered, a phase comparator uses the reference ring oscillator to sample of the output of it's stressed duplicate. The output of this phase comparator exhibits the beat frequency f_{stress}-f_{ref}, where f_{stress} is the stressed ring oscillator frequency and f_{ref} is the reference ring oscillator frequency. A counter which uses the reference ring oscillator signal as a clock measures the beat frequency. The counter's output N is measured after each stress period to calculate the percent frequency degradation, and the relationship between these two properties is shown in Fig. 4.2 (b). The period of the beat frequency is equal to the time when there is one clock difference between the number of reference and stress clock pulses, and the details of this beat frequency calculation is shown in Fig. 4.2 (b). Before stress, if the output of the counter is N, the number of clocks counted in the stressed ring oscillator is N-1. The period of beat frequency can be calculated by N/f_{ref} or $(N\text{-}1)/f_{stress}$. After stress, if the output of the counter is N', the number of clock pulses counted in the stressed ring oscillator is N'-1. Analogous to the calculation described above, the period of beat frequency is N'/f'_{ref} or $(N'\text{-}1)/f'_{stress}$. Using these two relations, the percentage of the frequency degradation can be obtained as illustrated in Fig. 4.2 (b). Previous measurement techniques that utilized only a single stressed ring oscillator [57] have a much more limited sensing resolution, as the counter output N is directly proportional to the frequency

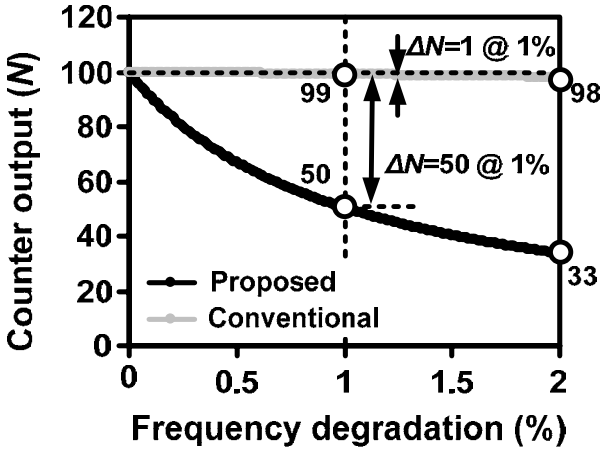
degradation. For example, in [57], 1% degradation in ring oscillator frequency translates into 1% change in counter output (Fig. 4.2 (c)). Using our proposed design, 1% degradation in ring oscillator frequency results in a 50% change in the counter output, offering 50X sensing resolution in the early stages of degradation. Increasing measurement sensitivity at the early stage of degradation generates less sensitivity when there is large frequency degradation. However, frequency degradation caused by device aging is usually less than 10% [57]. The proposed measurement circuit uses 90% of total code to detect 10% frequency change which has a higher sensing resolution compared to the previous scheme using 10% of total code [57]. The proposed silicon odometer circuit with a high sensing resolution can provide a number of benefits such as reduced test time, capability to study aging under various stress conditions, enabling non-accelerated stress measurements, etc. Note that the resolution of the proposed reliability monitor (i.e. $\Delta N/\Delta f_{stress}$) depends on the initial counter output N, which is set before the commencement of stress experiments. An initial N of 100 (or 256) allows a sensing resolution of 0.02% (or 0.0015%) at the early stage of degradation. The closer the frequencies of the two ring oscillators are brought together initially (i.e. the larger the initial N), the larger the change in counter output that can observed for the same degradation in ring oscillator frequency. Measurement accuracy can be easily programmed by changing the initial counter output using simple delay trimming circuits.



(a)



(b)



(c)

Figure 4.2 (a) Proposed beat frequency detection circuit for high resolution NBTI monitoring. (b) Principle of proposed beat frequency detection circuit. (c) Comparison of frequency sensing resolution between conventional and proposed techniques.

4.2.4 Silicon Odometer Circuit Design

A. Odometer System Architecture

The architecture of the silicon odometer test chip is illustrated in Fig. 4.3. The two 105 stage ring oscillators are identical structures with different control inputs. Process-Voltage-Temperature (PVT) variations that affect both structures equally will not alter the monitor output as the differential measurement approach cancels out this common-mode noise. Thick oxide I/O devices are used for the peripheral control circuits that are connected to the stress voltage. As described above, the phase comparator produces a digital signal representing the relationship between the frequencies of the reference and stressed ring oscillators. Bubbles (i.e. a lone '1' in a stream of '0's or a '0' in a stream of '1's) that may appear in the phase comparator output due to jitter and other circuit uncertainties can be eliminated by using a 5-bit majority voting circuit. The DETECT pulse generated by the beat frequency detector causes the register to sample the counter output and resets the counter for the next measurement cycle. For robust measurement results, multiple measurements are executed and the measured counter outputs are analyzed to calculate the frequency degradation. A parallel-to-serial register is used to scan out the measurement data.

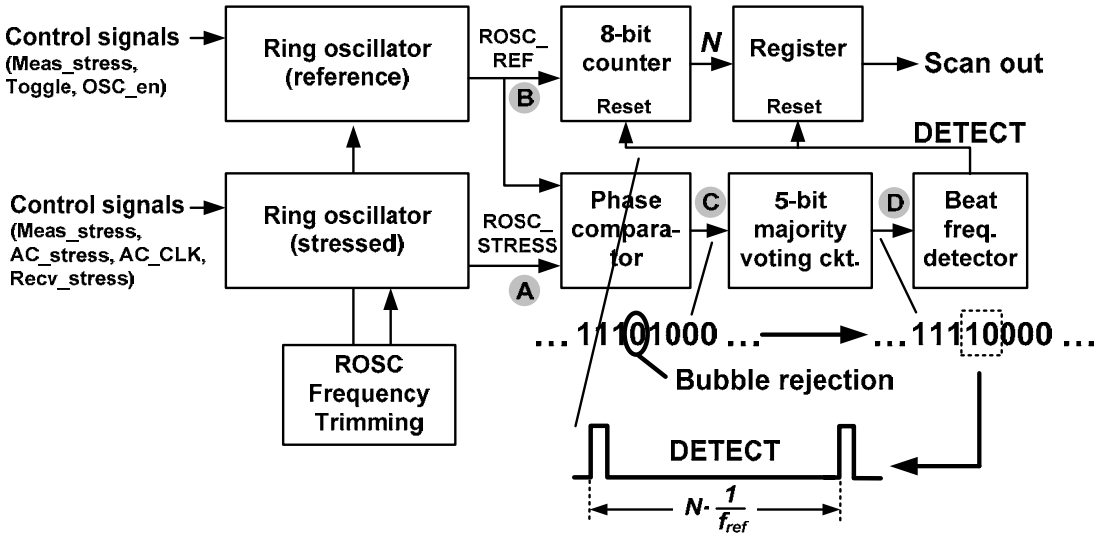


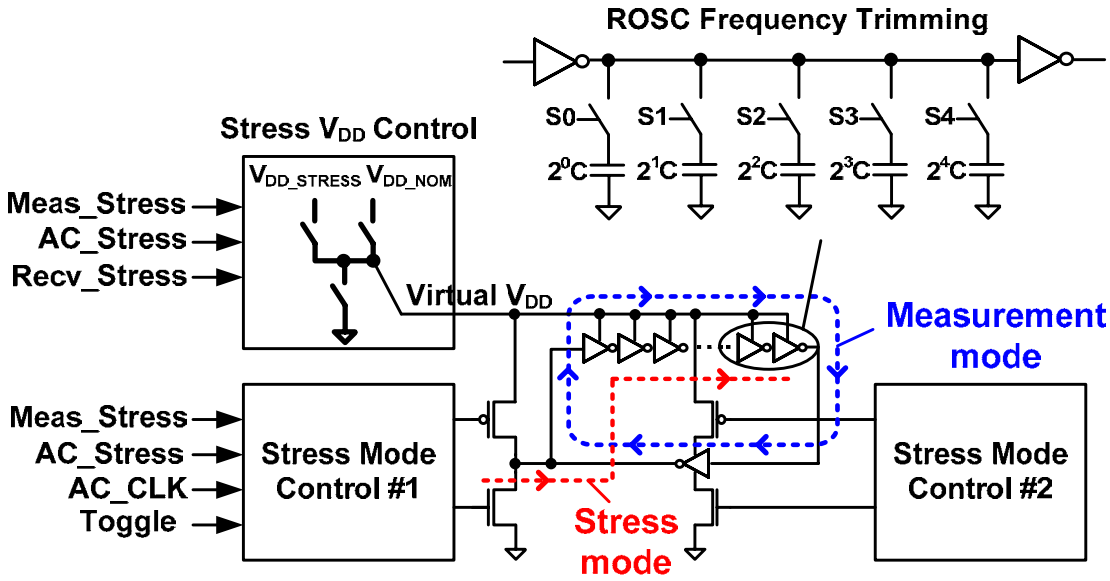
Figure 4.3 Reliability monitor test chip architecture.

B. Ring Oscillator Circuit

Fig. 4.4 shows a detailed schematic of the ring oscillator, as well as the various stress mode controls. The virtual V_{DD} can be switched to $V_{DD\text{-}STRESS}$, $V_{DD\text{-}NOM}$, and 0V to allow for stress, measurement, and recovery periods for the reference and stressed ring oscillators. During the stress period, the virtual V_{DD} in the stressed ring oscillator is connected to V_{DD_STRESS}, while that of the reference ring oscillator is connected to 0V to remove stress and keep its devices fresh. Only half of devices, which are turned on are stressed. In the measurement period, the virtual V_{DD} port in both ring oscillators is connected to V_{DD}, to allow measurement of the NBTI-induced degradation. Based on the values of the control signals, stress mode control #1 applies either AC or DC inputs. The AC_CLK signal utilized for AC stress is generated from an internal VCO. The ring oscillator input can also be toggled during each stress period to measure the circuit recovery with stress in alternating inverter stages. Stress mode control #2 disconnects the ring oscillator during stress mode to allow for various

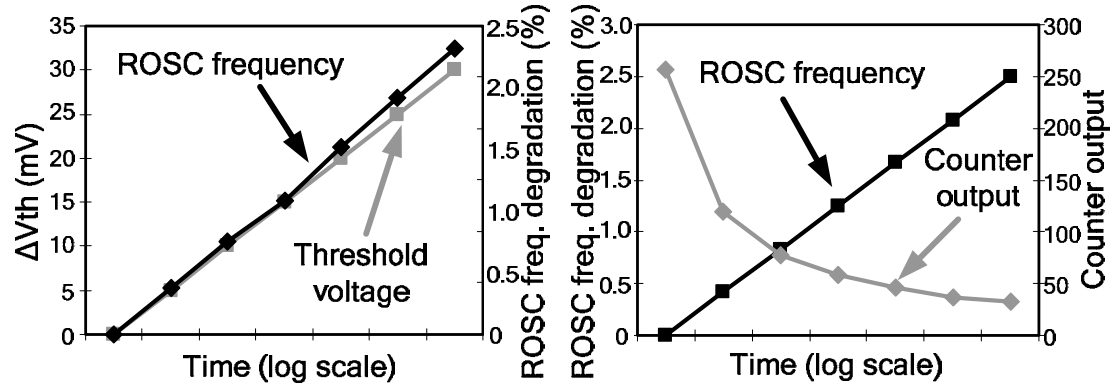
stress inputs to be applied. The table in Fig. 4.4 lists the control signals and corresponding measurement and stress modes. To achieve a high resolution frequency degradation measurement, the initial counter output should be large. The size of the initial counter output is highly sensitive to mismatches between two ring oscillators, so we have implemented a 5-bit binary-weighted switched-capacitor stage to allow adjustments to the initial ring oscillator frequencies. The desired counter output *N* is set prior to the stress experiments by scanning in control signals S0-S4. While in this work we have chosen to utilize an inverter chain-based ring oscillator for the test structure, other logic gates, such as NANDs, NORs, and pass gates, can also be utilized.

The effect of threshold voltage degradation on frequency degradation was simulated and is shown in Fig. 4.4 (b). It was assumed that all PMOS devices in the ring oscillator are stressed. The results of our simulations show that frequency and threshold voltage degradation are proportional to one another. It can be seen that a 30mV change in PMOS threshold voltage causes approximately a 2.79% degradation in performance. Fig. 4.4 (c) shows the change in counter output versus that of the simulated ring oscillator frequency degradation. As explained in section III-A, the initial small degradation in delay translates into a large change in the counter output. In simulation, the output code changed by 139 for a frequency degradation of 0.45% as shown in Fig. 4.4 (c). The threshold voltage before stress was 320mV, and an inverter chain-based ring oscillator was used for the simulation.



Mode	Stress condition	Meas_Stress	AC_Stress	Toggle
Measure	N/A	1	D/C	D/C
Stress	DC w/o toggle	0	0	0
	DC w/ toggle	0	0	1
	AC	0	1	D/C

(a)



(b) **(c)**

Figure 4.4 (a) Ring oscillator circuit and measurement/stress modes. (b) Simulation results of stress time versus PMOS threshold voltage and ring oscillator frequency. (c) Frequency and counter output as a function of stress time.

C. Phase Comparator Circuit

A phase comparator shown in Fig. 4.5 is used as a core circuit for detecting the beat frequency. A clock tapped out from the reference ring oscillator signal is used as the CLK input to control the operation modes of the phase comparator. When the CLK is low, the phase comparator is in pre-charge mode and resets the phase comparator output (PC_OUT). When the CLK becomes '1', the phase comparator switches to an evaluation mode and the PC_OUT is determined based on the arrival time of the two input signals, ROSC_REF and ROSC_STRESS. If there is an overlapped region between A′ and B, the pre-charged node is discharged and phase comparator goes high. No overlapped region will keep the pre-charged node high while giving a low output.

When the measurement begins, the rising edges of two input signals are aligned to each other and cause PC_OUT to go high. If the stressed ring oscillator has not been stressed, the frequency of two input signals will be identical, which makes PC_OUT always high. In this case, the maximum counter output is sent as read data. If the stressed ring oscillator has experienced the effects of aging, the frequency of the stressed ring oscillator decreases. As a result, the overlapped region between B and A' decreases in evaluation mode and the phase comparator output becomes low. The phase comparator will continuously generate a low output until there is a region of overlap between its two inputs. The data pattern at the phase comparator output repeats whenever there is one clock cycle difference between two input signals, which is used to measure the beat frequency.

In general, accurate measurement of phase differences requires a high resolution phase comparator. However, the proposed beat frequency detection scheme relaxes this design requirement. Any offset in the phase comparator simply shifts the start and end point of the measured time, without affecting the period. In addition, the measured period of the beat frequency is more sensitive to the degradation in ring oscillator frequency than the resolution of the phase comparator. For example, assume a jitter of 40ps in phase comparator, a period of 4ns in ring oscillator, and 1% frequency degradation in ring oscillator. 40ps of jitter represents 1% frequency error which is equal to the target frequency degradation, so a direct, non-differential frequency degradation measurement can have an error of 100%. However, in our beat frequency detection scheme, the time in which the measurement could be affected by phase comparator jitter is much smaller than the total measured period. Under the same assumptions discussed above and an initial counter output N of 100 before stress is applied, the 40ps jitter can only shift the clock count by one. By utilizing the equations in Fig. 4.2, the calculated frequency degradation including the error caused by the jitter in phase comparator becomes 1.05% or 0.97% while the true degradation is 1%. Our measurement technique has a 5% error which improves upon the direct measurement scheme by 20X.

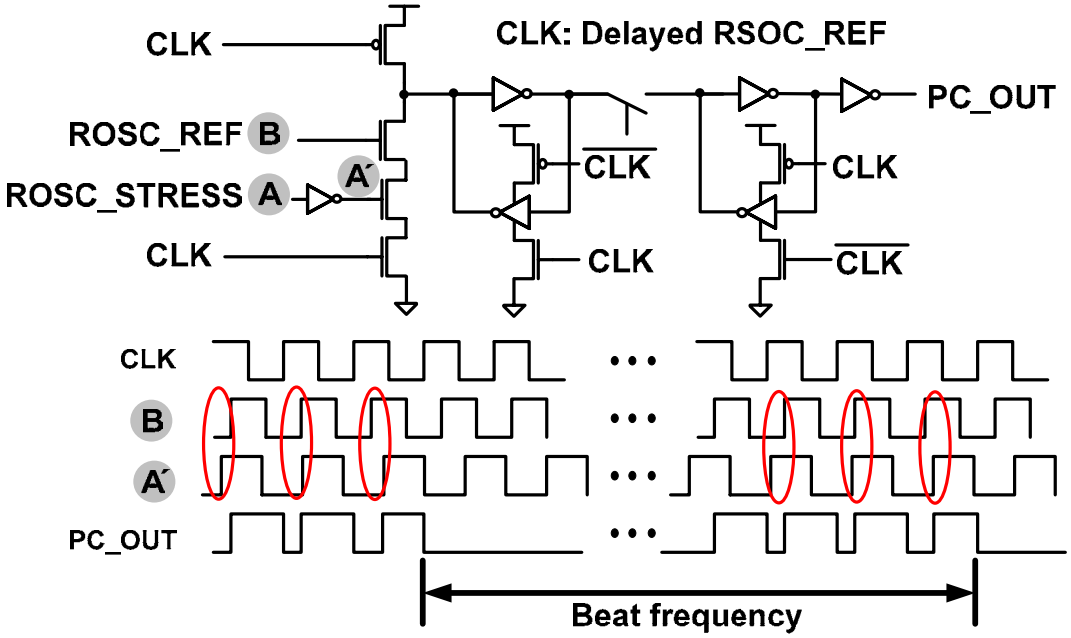


Figure 4.5 Phase comparator circuit.

D. Majority Voting Circuit

When two input signals are closely aligned, the power supply noise or other uncertainties in the phase comparator circuit can generate bubbles (i.e. lone '1' in a stream of '0's or a '0' in a stream of '1's) in the output. A 5-bit majority voting circuit was implemented to eliminate these bubbles. The implemented majority voting circuit can filter out two bubbles in a five bit data sequence. Fig. 4.6 shows a phase comparator outputs affected by bubbles, as well as the filtered data that is generated by the majority voting circuit.

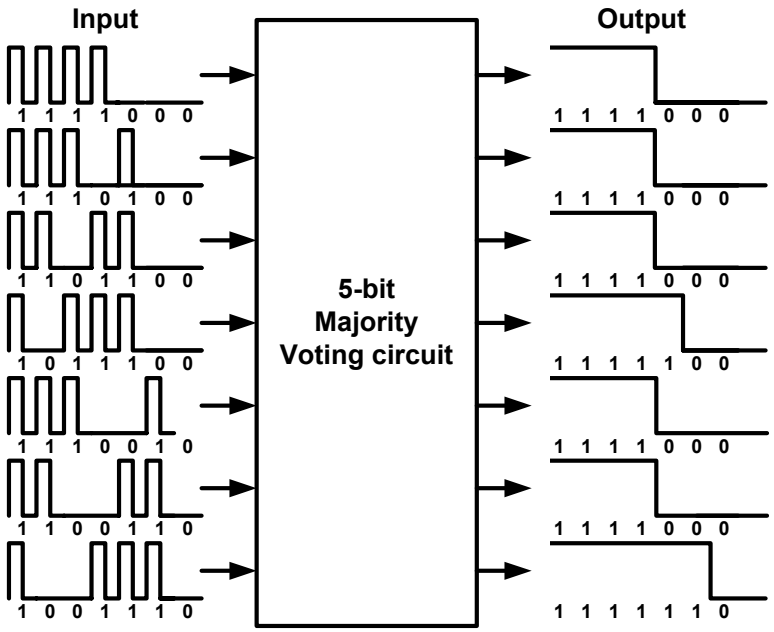


Figure 4.6 Operation of majority voting circuit.

E. Beat Frequency Detector

The output of the majority voting circuit is a signal with the beat frequency. The beat frequency detector generates a flag signal, DETECT, to read the counter output and reset the counter for the next measurement. The time interval between DETECT signals is the period of beat frequency. DETECT is used as a sampling clock in the register and reset signal in the counter. The rising edge of the majority voter output is detected by combinational logic using five received data points. Fig. 4.7 shows sample simulated waveforms. It can be seen that the period of PC_OUT and VOTE_OUT is identical to that of DETECT. There is a delay between VOTE_OUT and DETECT due to the data storing operation of the majority voting circuit and the latency of the beat frequency detector.

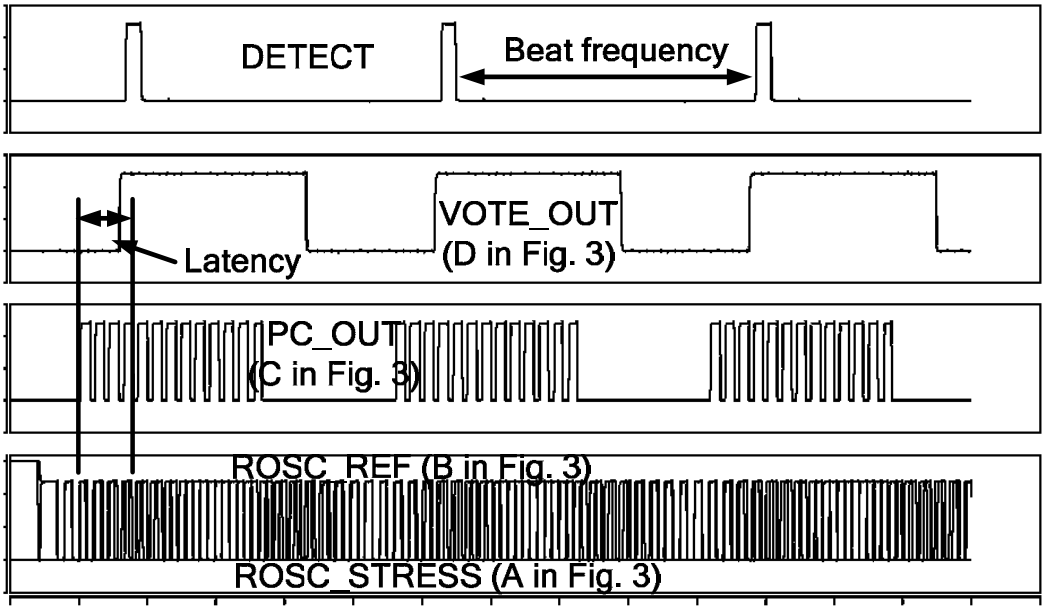
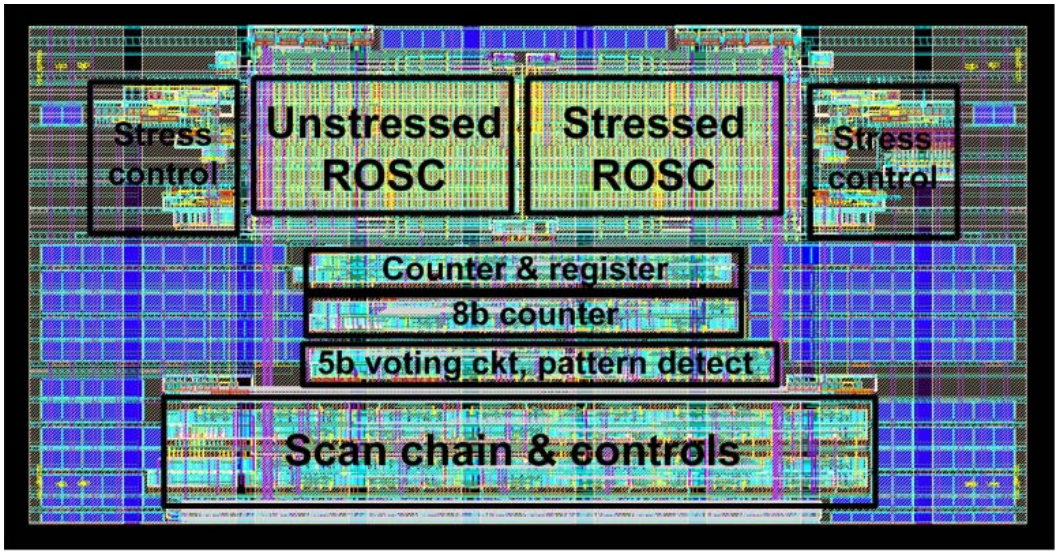


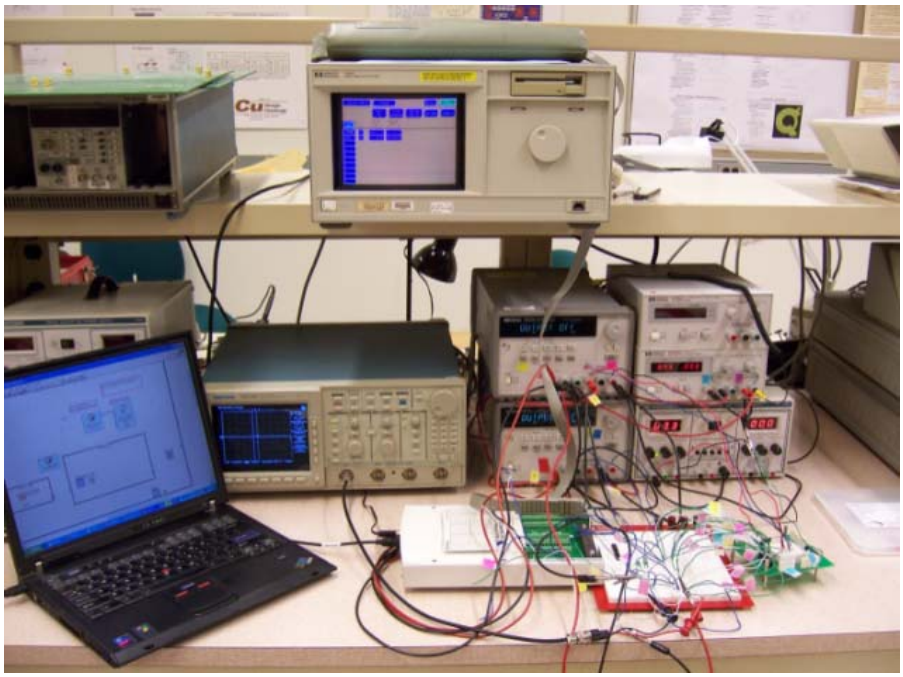
Figure 4.7 Simulated waveforms during measurement mode.

4.2.5 Test Chip Implementation and Experimental Results

A test chip was implemented in a 1.2V 130nm CMOS process technology to demonstrate the proposed silicon odometer circuit. Each 105 stage ring oscillators have a period of 4ns. To calibrate the ring oscillator frequency, we read out the counter output sweeping control signal S0-S4. The control signal generating the largest counter output is the optimum point. The target initial counter output (i.e. *N* in Fig. 4.2) was set to be 100 based on a target sensing resolution of 0.02% (or 0.8ps) so an 8 bit counter was used to allow for a counter output up to 256. This means 20X increase in measurement accuracy compared to the previous technique [57] where sensing resolution is 3.9%. In this measurement, the target initial counter output was limited by the noisy measurement environment. Fresh chips were used in each measurement as once stressed, circuits will not fully recover to its initial fresh state. An input signal with a frequency of up to 1GHz was applied to test for AC NBTI stress. The die area of the test circuit was 265x132μm^2 (Fig. 4.8 (a)). Fig. 4.8 (b) shows the laboratory setup for the test chip measurements.



(a)



(b)

Figure 4.8 (a) Layout of 130nm test chip occupying 265x132μm^2. (b) Laboratory setup for test chip NBTI measurements.

Fig. 4.9 (a) shows the measured counter output, while the corresponding frequency degradation is plotted in Fig. 4.9 (b) using the equation in Fig. 4.2. The supply voltage of the stressed ring oscillator was shut down during the recovery periods. The high sensing resolution of the proposed sensor enabled aging measurements using a nominal 1.2V supply voltage as the stress voltage. Such measurements done under a non-accelerated stress condition allows us to study the circuit aging effect during normal chip operation. Three measurement samples were taken at each measurement point of time and the error bar indicates the variation between the samples. The worst case error between the sampled data and the average point was only 0.022%. The ring oscillator frequency was reduced by 0.238% at the end of the first stress period of 1730 seconds when stressed at 1.2V and 30°C. Removing the stress voltage gave a 90.5% recovery of the performance loss by the end of the first recover period. Such large extent of recovery is typically seen in older processes with thicker oxides where most of the hydrogen atoms, the consequences of the broken *Si-H* bonds, remain in the oxide region and quickly anneal when the stress is removed. The frequency degradation dependency on temperature is illustrated in Fig. 4.9 (c). Measurements were done at 30°C and 130°C for comparison. It can be seen that higher temperature accelerates degradation faster. The device degradation is also a strong function of a stress voltage. Increasing stress voltage increases electric field and the degradation is exponentially dependent upon the electric field. The stress frequency effect on degradation is shown in Fig. 4.9 (d). It is said that degradation is highly related to the signal probability. In recursive RD models [64], it is also claimed that the

amount of degradation is proportional to the time assigned for stress. If a stress signal with a duty cycle of 50% is used for stress, the amount of degradation would be similar due to the same amount of time effectively used for stress. In Fig. 4.9 (d), DC stress shows higher degradation than AC stress and the effect of AC stress frequency on aging is small due to the constant duty cycle, which agrees with previous RD models. Finally, Fig. 4.10 shows the measured effect of stress voltage on degradation. Like threshold voltage, frequency degradation also has the same power-law dependency. Two power-law equations are obtained from fitted data. After 1730 seconds, the frequency degradation of 0.67% was observed when using 1.8V as a stress voltage. When a supply voltage of 1.2V was used, the measured frequency degradation was 0.24%.

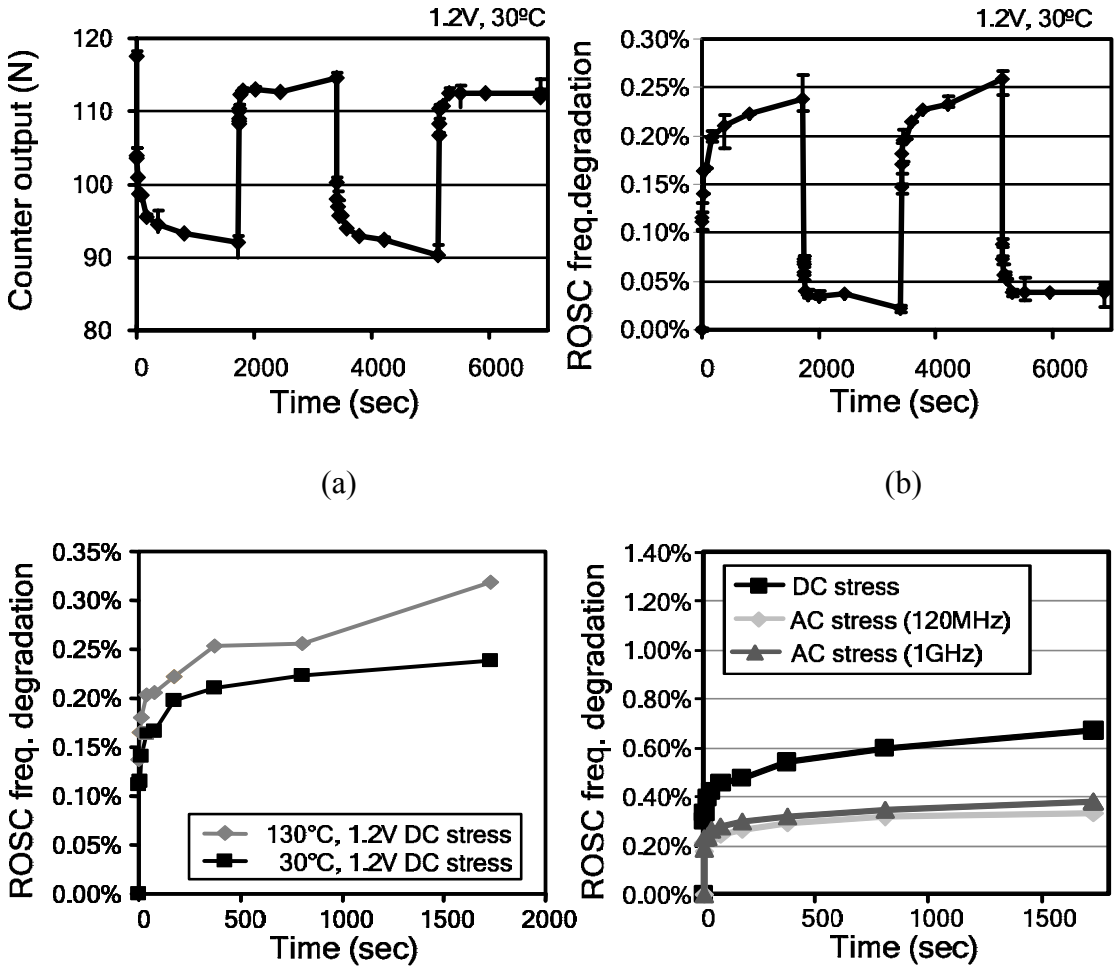


Figure 4.9 Measurement results: (a) Counter output. (b) Calculated frequency degradation for alternating stress and recovery periods. Error bars show the variation between the 3 sampled data taken at each measurement points. (c) Frequency degradation at different temperatures. (d) Frequency degradation under DC and AC stress.

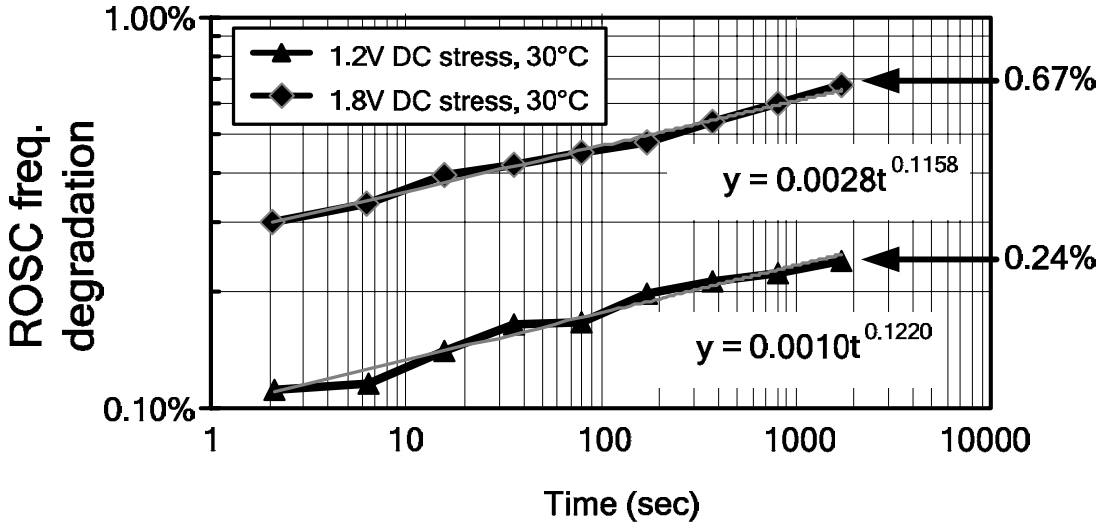
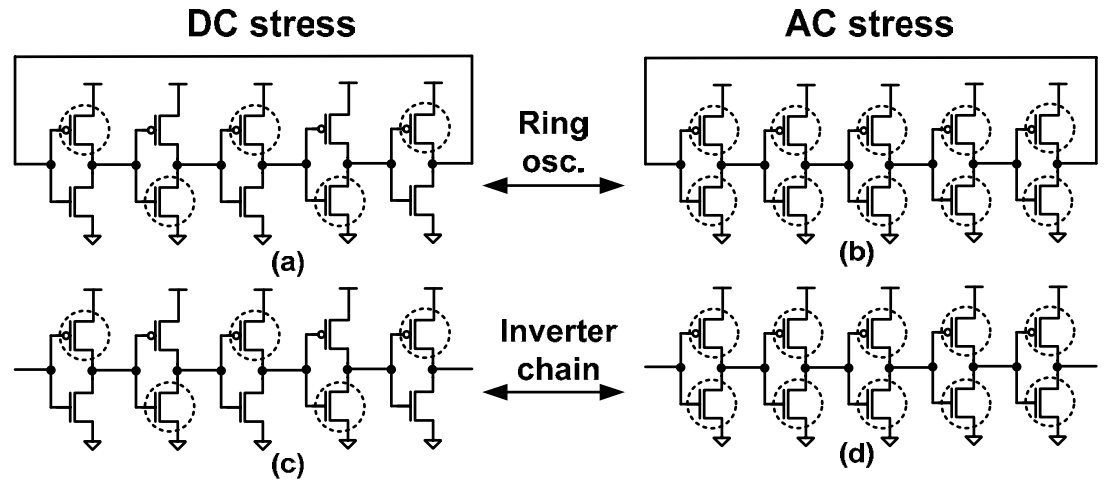


Figure 4.10 Frequency degradation for different stress voltages.

When applying DC stress to the ring oscillator, only half of the devices are under stress, so the period of the ring oscillator is decided by the summation of the delay from stressed path and that from unstressed path as shown in Fig. 4.11 (a). On the other hand, the worst-case frequency degradation of a true inverter path is determined by the delay of the stressed path only (Fig. 4.11 (b)). The relationship between the frequency degradation measured from ring oscillator, and that of our target, the true frequency degradation is given in Fig. 4.11. The true stressed inverter chain delay can be calculated by adding the stressed pull-up delay and stressed pull-down delay. By using these two expressions, the true frequency degradation can be represented as a function of ring oscillator frequency degradation. Our derivation shows that under DC stress, the degradation of the ring oscillator frequency is almost the half that of the true inverter chain. During periods of AC stress, all PMOS devices and NMOS devices are stressed equally, so the period of the ring oscillator is simply double that of the

inverter delay. As a result, the measured ring oscillator frequency degradation is equal to that experienced by the inverter chain. The true inverter chain frequency degradation from the DC stress measurement results is shown in Fig. 4.12 (a). Note that the amount of degradation shown in Fig. 4.12 (a) is twice as large as that in Fig. 4.9 (b). The true inverter frequency degradation from AC stress calculated using the equations in Fig. 4.11 is plotted in Fig. 4.12 (b). Note that the degradation of NMOS transistors is negligible when poly gate is used. Therefore, the measured degradations are mostly from NBTI.



$$T'_{rosc} = 1/f'_{rosc} \approx t'_{pu} \cdot \frac{N}{2} + t'_{pd} \cdot \frac{N}{2} + t_{pu} \cdot \frac{N}{2} + t_{pd} \cdot \frac{N}{2}$$

$$T'_{true} = 1/f'_{true} \approx t'_{pu} \cdot \frac{N}{2} + t'_{pd} \cdot \frac{N}{2}$$

$$\frac{f'_{rosc} - f_{rosc}}{f_{rosc}} = \frac{\frac{t_{pu} + t_{pd}}{t'_{pu} + t'_{pd}} - 1}{\frac{t_{pu} + t_{pd}}{t'_{pu} + t'_{pd}} + 1} = x$$

$$\frac{f'_{true} - f_{true}}{f_{true}} = \frac{t_{pu} + t_{pd}}{t'_{pu} + t'_{pd}} - 1 = y$$

$$\therefore\ y = \frac{2x}{1-x} \approx 2x$$

$$T'_{rosc} = 1/f'_{rosc} \approx t'_{pu} \cdot N + t'_{pd} \cdot N$$

$$T'_{true} = 1/f'_{true} \approx t'_{pu} \cdot \frac{N}{2} + t'_{pd} \cdot \frac{N}{2}$$

$$\frac{f'_{rosc} - f_{rosc}}{f_{rosc}} = \frac{t_{pu} + t_{pd}}{t'_{pu} + t'_{pd}} - 1 = x$$

$$\frac{f'_{true} - f_{true}}{f_{true}} = \frac{t_{pu} + t_{pd}}{t'_{pu} + t'_{pd}} - 1 = y$$

$$\therefore\ y = x$$

Figure 4.11 Relationship between the ring oscillator frequency degradation and the worst-case true inverter chain frequency degradation for DC and AC stress. Frequency degradation of a true inverter chain is twice that of the ring oscillator frequency degradation for the DC stress case. On the other hand, the two circuits observe the same amount of frequency degradation under AC stress.

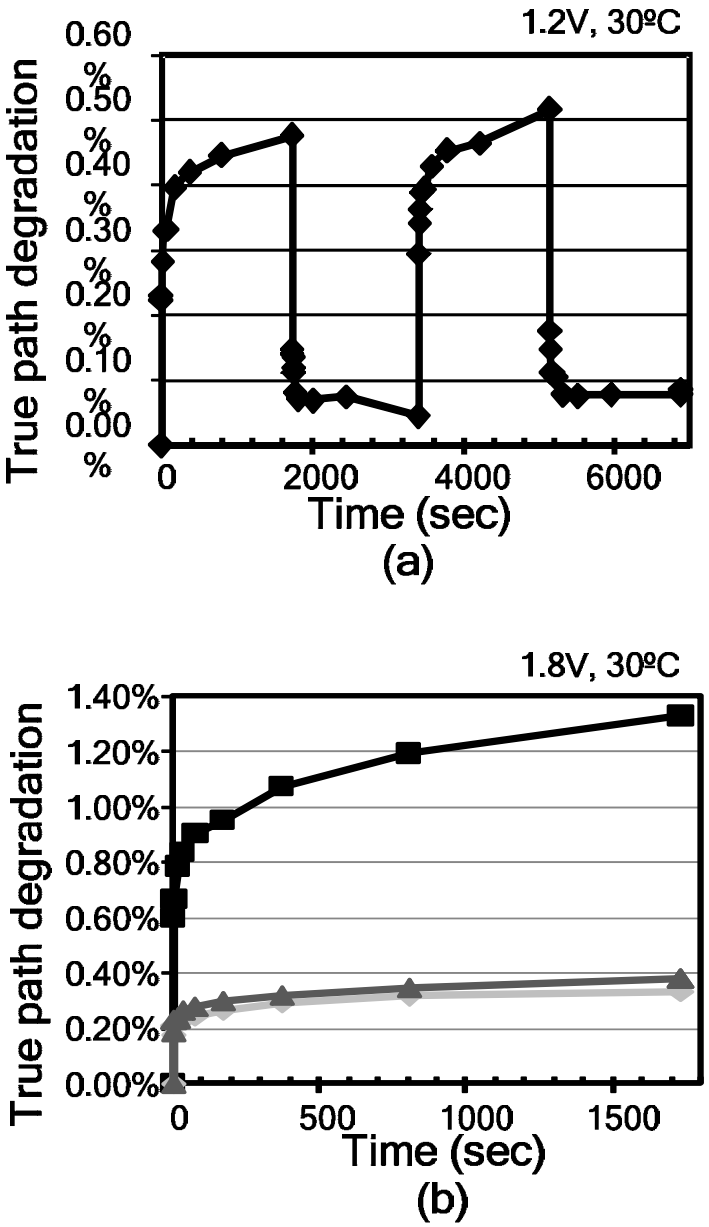


Figure 4.12 True frequency degradation of an inverter chain calculated from the measurement results in Fig. 4.9 (b). (b) True inverter chain frequency degradation calculated from the measurement results in Fig. 4.9 (d).

4.3 Isolated NBTI and PBTI Measurement Structures in 32nm High-k Metal-Gate CMOS

4.3.1 Overview

In poly-gate devices, NBTI in PMOS has been considered as a dominant reliability concern compared to the corresponding Positive Bias Temperature Instability (PBTI) in NMOS transistors. However, PBTI in NMOS devices is also becoming a reliability concern as high-k dielectric material and metal-gate are adopted for gate leakage reduction in sub-45nm CMOS technology nodes [65][66][67]. High-k dielectrics generate significant charge trapping compared to the conventional silicon dioxide in NMOS devices, but show the same NBTI results as those using conventional silicon dioxide in gate dielectric stacks [65][68]. Zafar, et al. presented that PBTI is more sensitive to the high-k dielectrics and gate material, and becomes a greater reliability issue than NBTI when HfO2 and NiSi are used as dielectric and gate material, respectively [65].

A number of previous works have presented the impact of NBTI on digital circuits [57][69][70]. Since PBTI has not been prominent before using high-k dielectrics and metal gate devices, most of these structures have used simple ring oscillators as NBTI monitors by measuring the frequency before and after stress. However, this will give a mixed result of NBTI and PBTI in high-k dielectric and metal gate CMOS technologies. Since the magnitude of NBTI and PBTI is different after a given stress time, the impact of each NBTI and PBTI on circuit performance should be estimated independently for better understanding of these factors on circuits.

Test structures facilitating the isolation of NBTI and PBTI, and their impacts on circuits are highly required for this purpose.

In this paper, we present on-chip test structures with isolated NBTI and PBTI stress capabilities which are applicable to 32nm high-k metal-gate devices. Both frequency degradation and threshold voltage degradation due to NBTI and PBTI are enabled with the proposed test structures. The separate measurements of degradation caused by NBTI and PBTI using the proposed test structure can precisely estimate the portion of degradation due to NBTI and PBTI in logic gates. The remainder of this paper is organized as follows. Section II reviews the previous NBTI/PBTI monitoring circuits. Section III is devoted to the design of the proposed NBTI/PBTI monitoring circuits. Two types of monitoring circuits are described for frequency degradation measurement and threshold voltage degradation measurement. Silicon odometer [23] is adopted for fast frequency degradation measurements. Section IV will address the test chip measurement results. The paper will be concluded with the summary in section V.

4.3.2 Previous NBTI/PBTI Measurement Structures

The conventional way of measuring NBTI effect on circuit is to use ring oscillators [57]. Fig. 4.13 shows the conventional ring oscillator for measuring frequency degradation due to NBTI. During the stress mode, supply voltage is raised to stress voltage (V_{str}) to accelerate the NBTI, and the input of the ring oscillator is connected to a fixed level, GND or VDD. Half of the PMOS and NMOS transistors in the ring oscillator are stressed. Since PBTI is insignificant in poly-gate devices, the measured result includes the degradation only in PMOS devices. However, in a ring oscillator using high-k dielectrics and metal gate devices, both NBTI and PBTI are prominent, and the measurement result after stress shows a mixed result of NBTI and PBTI. Therefore, the conventional ring oscillator cannot be used to monitor the NBTI and PBTI effect separately. Ketchen et al. proposed a ring oscillator based test structure for measuring NBTI in inverter-driven PMOS passgates [71]. By changing the gate voltage of PMOS passgates, the amount of threshold voltage degradation after stress is directly measured. However, it requires a negative voltage to stress the passgates under test, and should also be controlled accurately for controlling the ring oscillator frequency with high accuracy. In addition, the drain and source node voltages of the stressed passgates are biased to GND through the NMOS keeper devices which are turned off. It will lose the control of the drain and source bias voltage if the gate leakage becomes comparable to the leakage current in the keepers, which makes the measured result unreliable. Kim et al. presented a ring oscillator circuit structures for isolated NBTI/PBTI effects [72]. Additional devices are added to cut the device under test from the rest of circuits and bias the rest circuits free of stress.

The isolation of NBTI and PBTI during a stress mode makes these circuit structures applicable to high-k and metal gate CMOS technologies. However, the inserted switch is stacked with the device to be tested, which affects the operation of the circuit. The impact of the switch on the measurement can be reduced by increasing the size of the switch, but the measured result cannot represent the frequency degradation of the original circuit caused by NBTI or PBTI.

Our proposed circuit structures address many of the problems raised by the previous literature, including the isolation of NBTI and PBTI effects, no requirement of negative voltage, reliable circuit control ability, and estimation of the original circuit without discrepancy.

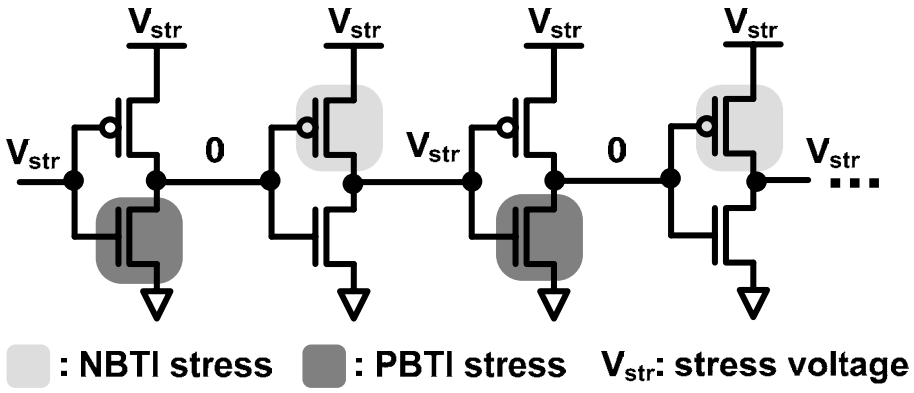


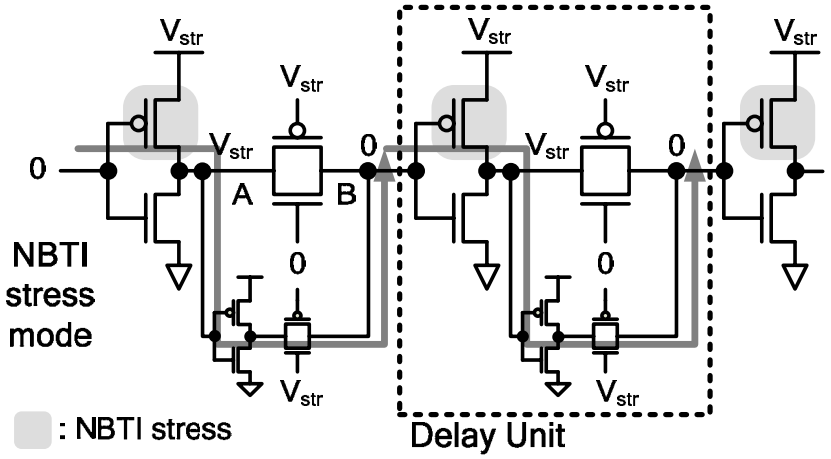
Figure 4.13 Conventional ring oscillator based NBTI monitor.

4.3.3 Isolated NBTI/PBTI Monitor: Frequency Measurements

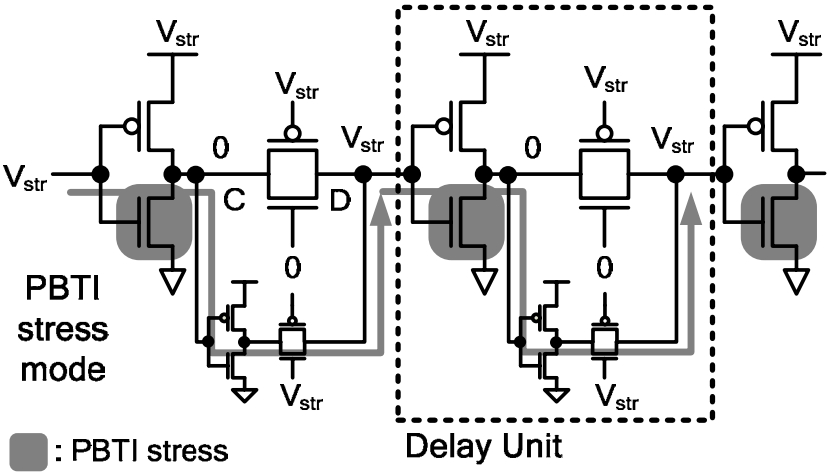
To achieve the isolation of NBTI and PBTI effects, the chain of delay units in a ring oscillator needs to be cut during a stress mode. Extra devices are inevitable for this purpose since all the delay units in the ring oscillator should have the same bias condition, stressing all devices under test and keeping the rest devices as fresh as possible. This is not possible in conventional ring oscillators composed of a chain of simple digital logic gates since the output of each delay unit alternates.

Fig. 4.14 shows the proposed test structure for separately measuring the NBTI and PBTI impact on inverter delay. The delay unit consists of two inverters with transmission gate load, forming two signal paths; a measurement path for frequency measurements and a control path for applying NBTI or PBTI stress to all devices under test simultaneously. During stress modes, a stress voltage V_{str} higher than the nominal supply voltage is applied to the test structure. The transmission gate in the measurement path is cut off while that in the control path is turned on. Signal A and C in Fig. 4.14 (a) and (b) are inverted and transferred to B and D, making the input of each inverter in the measurement path identical. For NBTI stress, the primary input of the ring oscillator is connected to ground in order to stress all PMOS transistors in the inverters. Likewise, the input of the ring oscillator is connected to V_{str} for PBTI stress. During stress periods, no other transistor in the measurement path is stressed except for the devices under test. Devices in the control path are stressed but their impact on measurement path delay is negligible because they are disconnected during the measurement modes. This test structure can be easily expanded to other types of

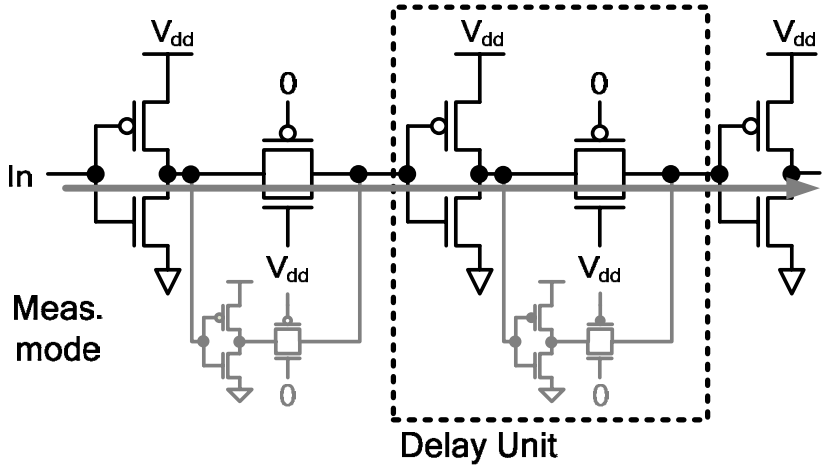
complex logic gates (e.g. NAND, NOR) by replacing the inverter in each stage with those logic gates.



(a)



(b)



(c)

Figure 4.14 Proposed ring oscillator for frequency measurements under isolated NBTI/PBTI stress. (a) NBTI stress mode. (b) PBTI stress mode. (c) Measurement mode

The proposed ring oscillator has additional devices compared to a conventional ring oscillator at the output of each stage, which will affect the delay of each delay unit. Here, the delay degradation relationship between the conventional ring oscillator and the proposed structure will be mathematically derived. Fig. 4.15 (a) shows the schematic of the conventional delay chain and simplified RC parameters. The period of the ring oscillator (T_1) can be expressed as

$$T_1 = \alpha \times n \times C_1\left(R_n + R_p\right) \tag{1}$$

where α is the constant, n is number of delay unit, C_1 is the capacitance at each node including gate capacitance and junction capacitance, and R_n and R_p are the resistance of NMOS and PMOS. The delay degradation due to NBTI and PBTI (ΔT_1) can be calculated by equation (1).

$$\Delta T_1 = \Delta T_{NBTI} + \Delta T_{PBTI} = \alpha \times \frac{n}{2} \times C_1 \left(\Delta R_n + \Delta R_p \right) \tag{2}$$

Here, ΔT_{NBTI} and ΔT_{PBTI} represent the delay degradation due to NBTI and PBTI, respectively, and ΔR_n and ΔR_p are the resistance degradations in NMOS and PMOS, correspondingly. Therefore, the frequency degradation is derived by dividing equation (2) by equation (1).

$$\frac{\Delta T_1}{T_1} = \frac{1}{2} \frac{\Delta R_n + \Delta R_p}{R_n + R_p} \tag{3}$$

Equation (3) shows that the frequency degradation of the conventional ring oscillator is only a function of resistance. Adding capacitive load at each node changes the absolute frequency value, but has no impact on frequency degradation.

Fig. 4.15 (b) illustrates the schematic of the proposed delay chain and RC parameters in measurement modes. The ring oscillator period with the proposed delay unit can be expressed as

$$\begin{aligned} T_P &= \alpha \times n \times \left\{ R_p C_2 + \left(R_p + R_g \right) C_3 + R_n C_2 + \left(R_n + R_g \right) C_3 \right\} \\ &= \alpha \times n \times \left\{ \left(R_p + R_n \right) \left(C_2 + C_3 \right) + 2 R_g C_3 \right\} \end{aligned} \tag{4}$$

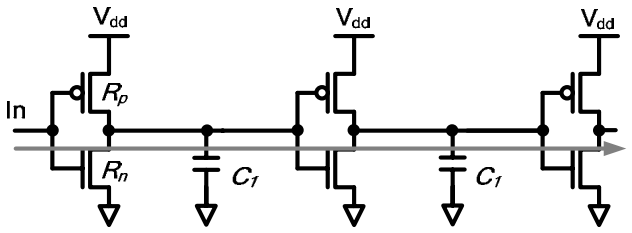
where C_2 and C_3 are the capacitance at the input and output of the inverters including gate capacitance and junction capacitance, and R_g is the resistance of the passgate. The delay degradation due to each NBTI and PBTI can be calculated similarly by using the equation (2).

$$
\begin{aligned}
\Delta T_{P\_NBTI} &= \alpha \times n \times \Delta R_p \left(C_2 + C_3\right) \\
\Delta T_{P\_PBTI} &= \alpha \times n \times \Delta R_n \left(C_2 + C_3\right) \\
\therefore \frac{\Delta T_{P\_NBTI}}{T_P} &= \frac{\alpha \times n \times \Delta R_p \left(C_2 + C_3\right)}{T_P} \\
\therefore \frac{\Delta T_{P\_PBTI}}{T_P} &= \frac{\alpha \times n \times \Delta R_n \left(C_2 + C_3\right)}{T_P}
\end{aligned} \tag{5}
$$

Finally, the delay relationship between the proposed structure and conventional ring oscillator is obtained from equation (3) and (5).

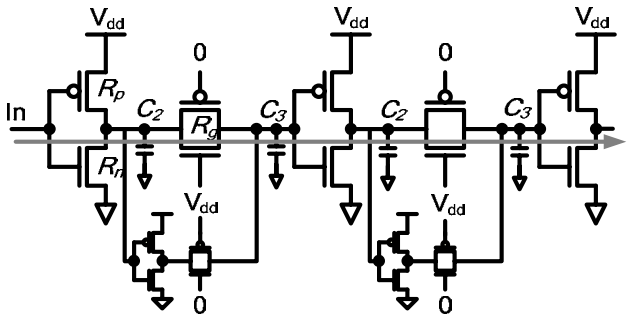
$$
\begin{aligned}
\frac{\Delta T_1}{T_1} &= \frac{1}{2}\left\{1 + \frac{2C_3 R_g}{\left(R_n + R_p\right)\left(C_2 + C_3\right)}\right\}\left(\frac{\Delta T_{P\_NBTI}}{T_P} + \frac{\Delta T_{P\_PBTI}}{T_P}\right) \\
&= \frac{1}{2}\left(1 + \beta\right)\left(\frac{\Delta T_{P\_NBTI}}{T_P} + \frac{\Delta T_{P\_PBTI}}{T_P}\right)
\end{aligned} \tag{6}
$$

Equation (6) indicates that they have a linear relationship, which eliminates the impact of the added devices on measurements. This allows us to straightforwardly estimate the portion of delay degradation caused by NBTI and PBTI using separate measured results.



$$\frac{\Delta T_1}{T_1} = \frac{n(R_n - R_P)}{2(R_n + R_P)}$$

(a)



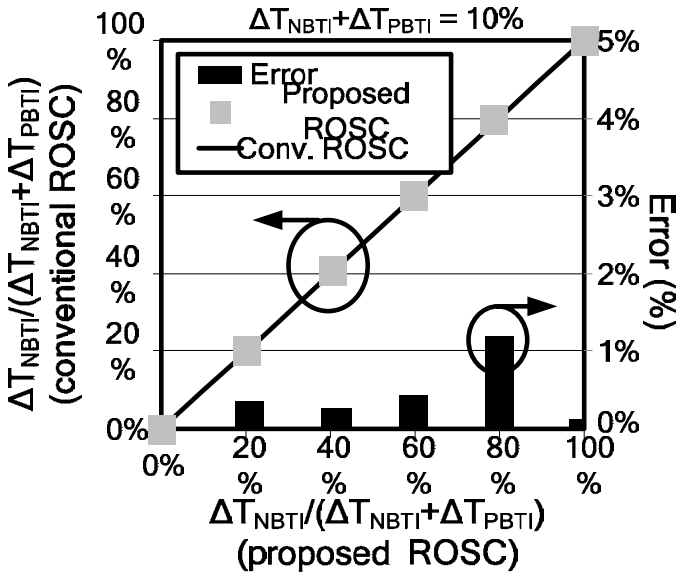
$$\frac{\Delta T_{P\_NBTI}}{T_P} = \frac{n(R_p \| C_2 \| C_3)}{T_P}$$

$$\frac{\Delta T_{P\_PBTI}}{T_P} = \frac{n(R_n \| C_2 \| C_3)}{T_P}$$

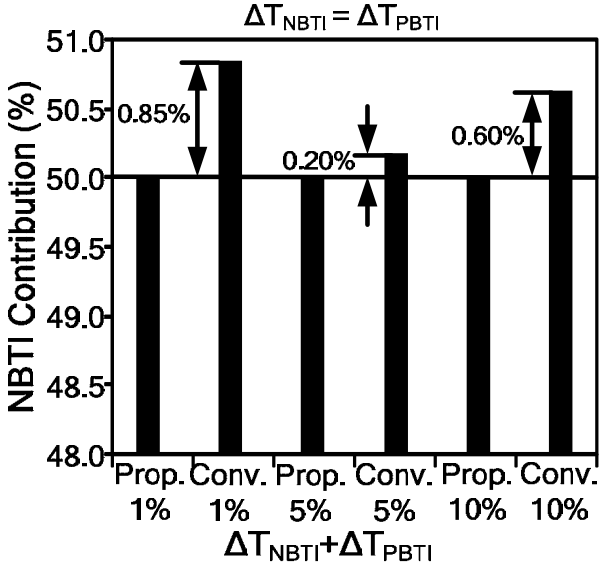
(b)

Figure 4.15 Measurement mode operation and delay relationships.

Fig. 4.16 (a) shows the percentage of delay degradation contributed by NBTI in the proposed structure and the conventional structure when the total delay degradation is 10%. Maximum error of the proposed scheme for estimating the NBTI and PBTI contributions in a conventional ring oscillator is only 1.2%. Fig. 4.16 (b) shows that the estimation error is less than 0.85% for different amounts of total delay degradations (1%, 5% and 10%) when NBTI and PBTI have equal contributions. Therefore, the proposed structure facilitates the independent measurement of frequency degradation due to NBTI and PBTI, which is impossible in the conventional ring oscillator structure.



(a)



(b)

Figure 4.16 Accuracy of proposed scheme in estimating NBTI/PBTI contributions.

4.3.4 Isolated NBTI/PBTI Monitor: Direct V_{th} Measurements

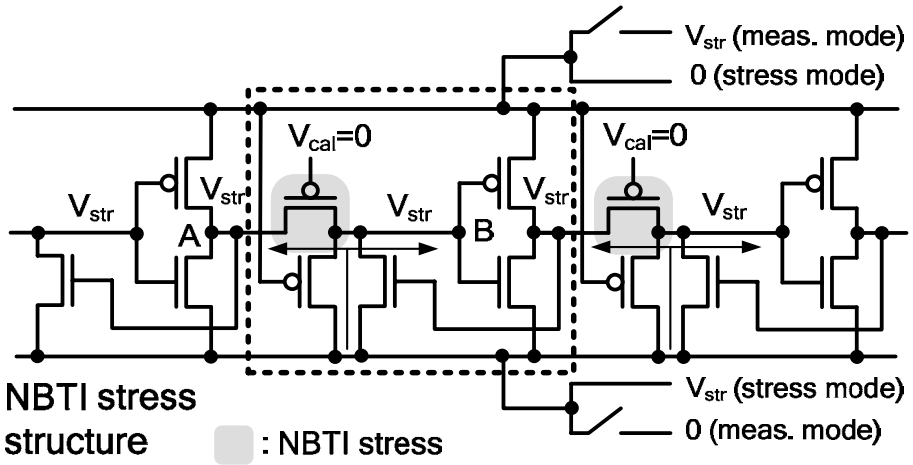
In addition to the frequency degradation measurement, the absolute value of threshold voltage degradation also needs to be measured. Previous techniques have been based upon discrete device probing, which takes long time to gather statistical data and suffer from NBTI recovery from relaxation mechanism.

Ring oscillator based test structures for directly measuring the threshold voltage degradation are shown in Fig. 4.17. They consist of a measurement path and stress-bias circuits for stressing either NMOS or PMOS pass gates during stress modes. Like the test structures for frequency degradation measurement, each delay unit should have the same bias condition, stressing the devices under test and avoiding stress in the rest circuits. This is achieved by forcing the inverters to perform as source followers. Fig. 4.17 (a) demonstrates the test structure for NBTI stress.

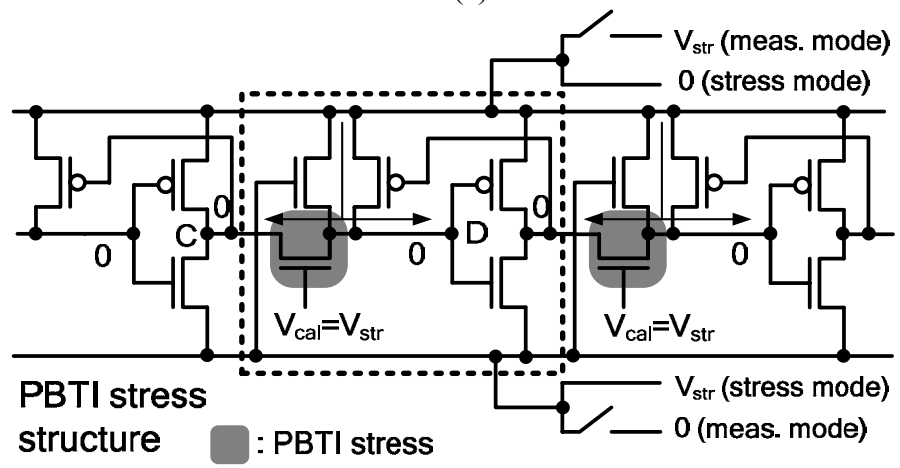
In stress modes, stress voltage (V_{str}) is applied to the ring oscillator input, the supply voltages are swapped, and the gate voltage of the PMOS pass gate is grounded. Since the PMOS is connected to the ground and the NMOS to the stress voltage (V_{str}), the first buffer weakly pulls up signal A, but the PMOS stress-bias transistor pulls signal B up to the stress voltage (V_{str}) which also drives signal A firmly to the stress voltage through the stressed PMOS pass gate. As a result all nodes are biased with the stress voltage (V_{str}), and all transistors in the measurement path are turned off preventing any unwanted aging.

During a measurement, the test structure reverts to the nominal supply condition, and the PMOS keepers are automatically turned off. The NMOS keepers recover the voltage drop across the PMOS pass gates while transferring logic ‘0’. The PBTI test

structure has the same structure and functionality except for the NMOS pass gates with PMOS keepers for restoring the signal levels and NMOS stress-bias transistor for pulling internal nodes down to ground during stress. Primary input of the test structure is connected to ground for the PBTI stress in order to bias the drains and sources of the NMOS pass gates to ground.



(a)



(b)

Figure 4.17 Proposed ring oscillator structure for direct V_{th} measurements under isolated NBTI/PBTI stress. (a) NBTI stress structure. (b) PBTI stress structure.

Circuit calibration is performed before applying stress where the relationship between the calibration voltage (V_{cal})and ring oscillator frequency is measured. As illustrated in Fig. 4.18, ΔV_{th} is directly proportional to ΔV_{cal} for an equivalent change in ring oscillator frequency. This equivalence allows us to later translate the measured frequency degradation into threshold voltage degradation (ΔV_{th}) [15].

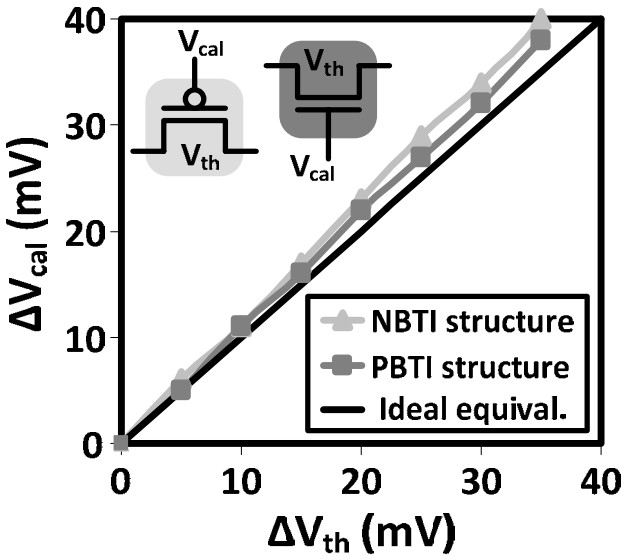


Figure 4.18 ΔV_{cal} vs. ΔV_{th} relationship for equivalent change in frequency.

4.3.5 Test Chip Implementation

The test chip architecture for the proposed NBTI and PBTI test structures is shown in Fig. 4.19. A beat frequency detection scheme [23] is used to achieve high resolution degradation measurements. Two identical ring oscillator sets, one for the reference ring oscillator and the other for the stressed ring oscillator, are implemented for the delay degradation measurement. Sixteen pairs of ring oscillators with different logic gates, sizes and stress types were designed using the proposed test structures. Phase comparators compare two input signals and generate a digital signal with beat frequency. To eliminate the effect of other circuits on measurements, each ring oscillator pair has a dedicated phase comparator forming a beat frequency detection unit, and a decoder and a 16-to-1 multiplexer is located outside of the beat frequency detection block.

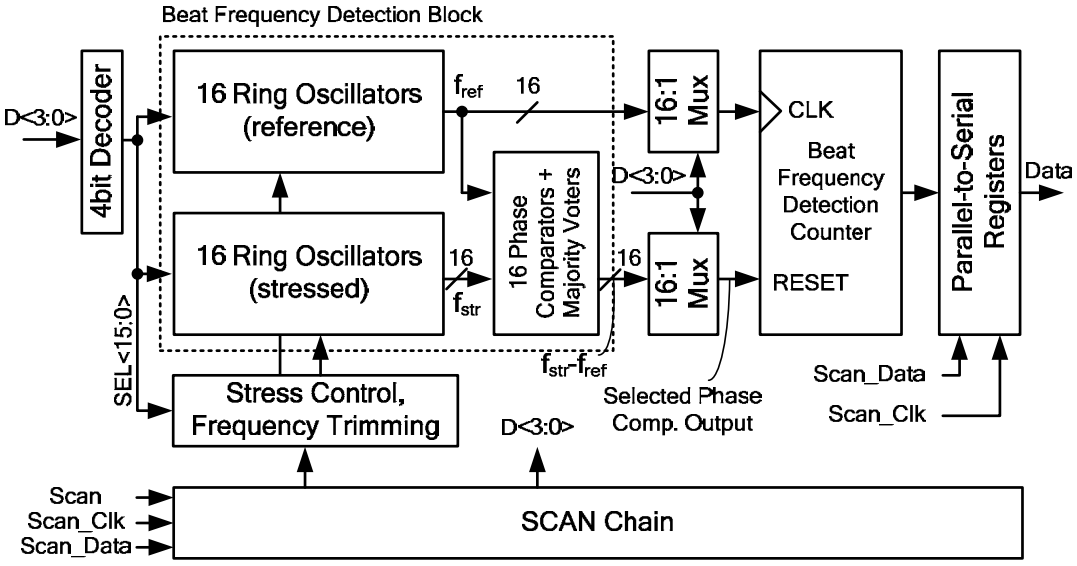


Figure 4.19 Test chip architecture based on beat frequency detection scheme.

Fig. 4.20 shows the test sequence for a frequency degradation measurement. Each measurement consists of three sub-sequences: an initialization sequence, a stress sequence, a measure sequence, and a scan read sequence. The initialization sequence writes control data into scan chain to select a pair of ring oscillators to be tested, control stress modes, and trim ring oscillator frequency. To avoid the unwanted stress due to the randomly generated select code (D<3:0>), stress voltage (VDD_STR) is grounded during the initialization. After the initialization, only the selected beat frequency unit is activated for each measurement using digital select codes D<3:0>. Stress is given by applying stress voltage at VDD_STR and logic 'high' to STRESS. The measurement sequence starts at the falling edge of STRESS and ends at the rising edge of STRESS. The measurement time for sampling a single beat frequency is less than 1μs to prevent any unwanted recovery from corrupting the aging data. To average out the effect of random noise and variations, three measurements are executed before 'Scan Read' operation. The measured results are stored in parallel-to-serial registers and read through 'Scan Read' operation. Fig. 4.21 illustrates the simulated waveforms showing the operation of the high-level architecture.

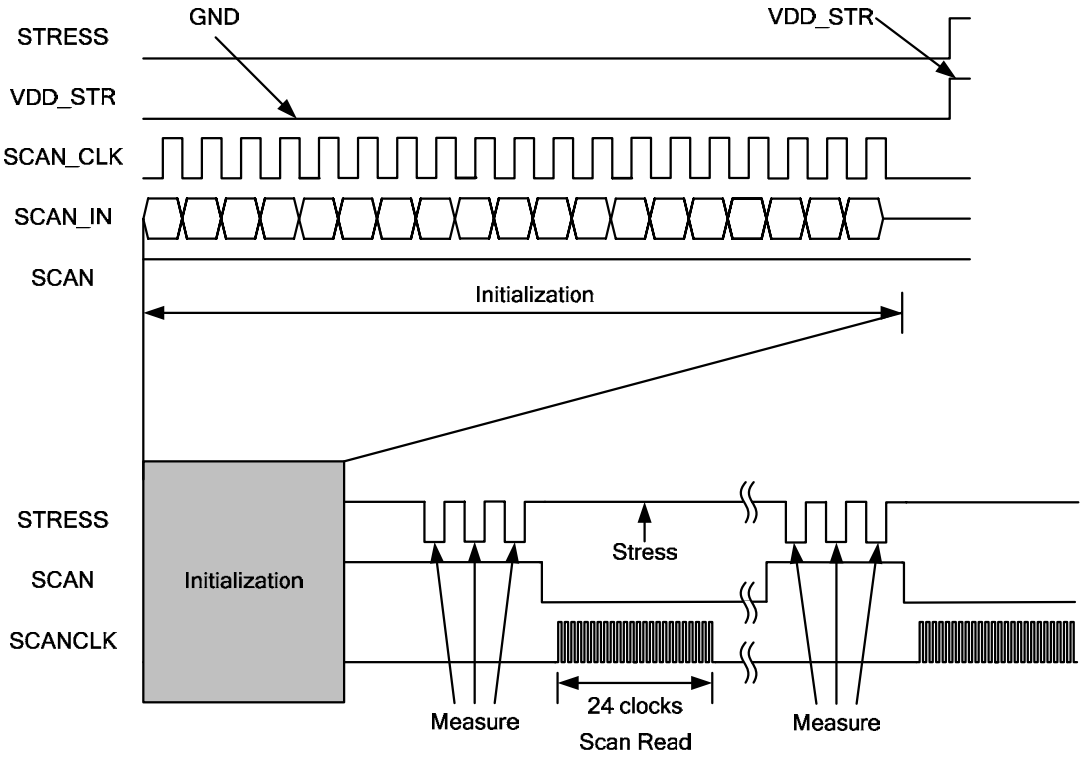


Figure 4.20 Input signal waveforms for frequency degradation measurements.

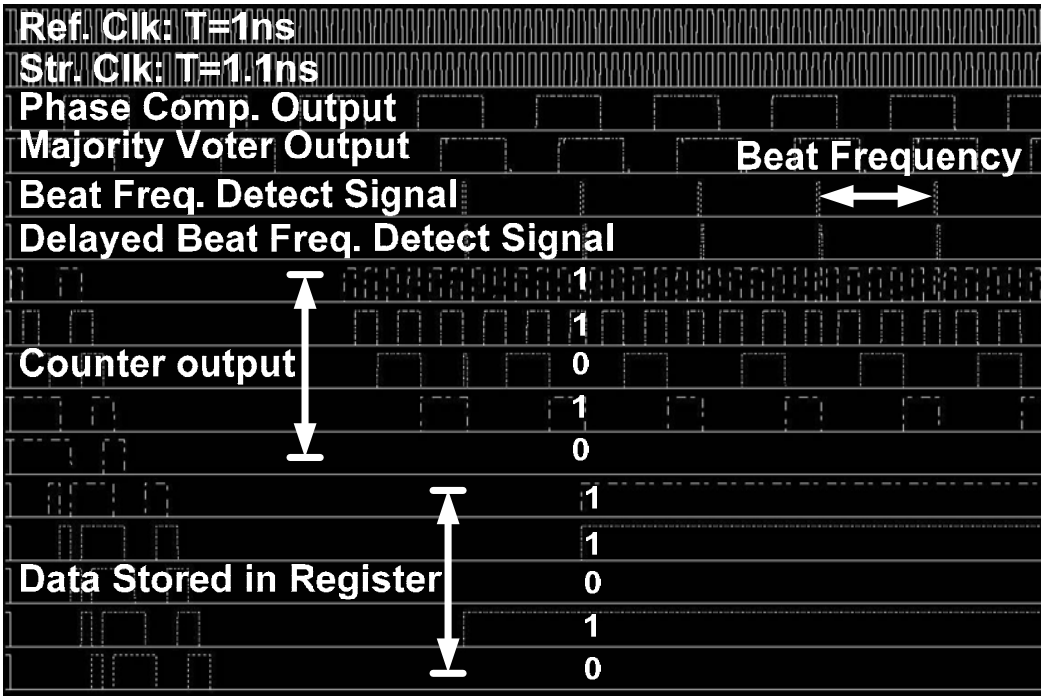
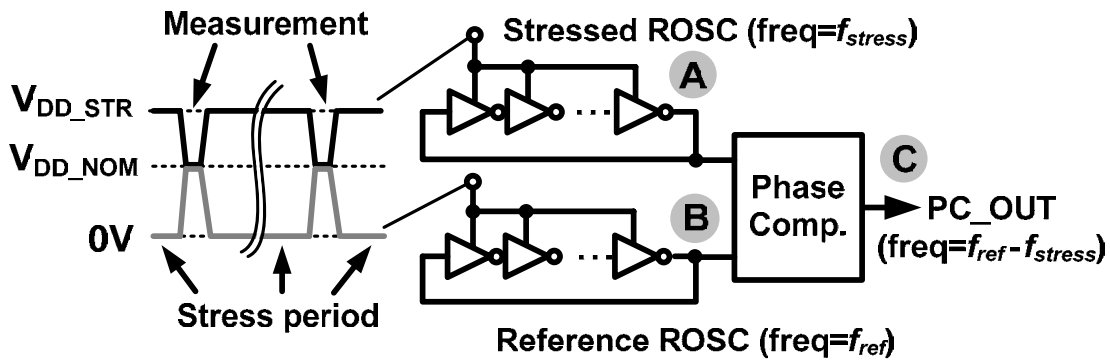
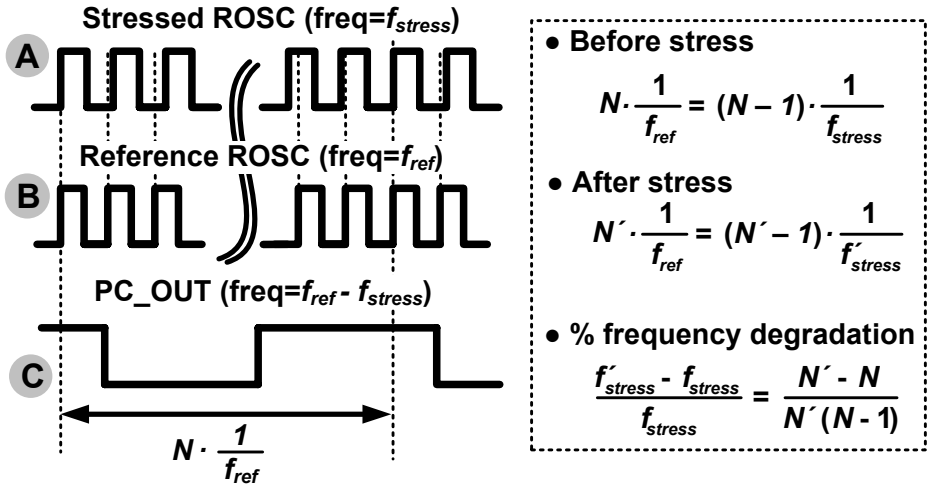


Figure 4.21 Test chip waveforms during measurement mode.

The beat frequency detection unit consists of two free-running ring oscillators and a phase comparator as shown in Fig. 4.22 (a). Instead of using a dynamic circuit with precharge phase in [23], a simple flip-flop is used to implement the phase comparator. This removes the return-to-zero transition in every clock cycle. During the stress period, the supply voltage of the stressed ring oscillator is raised to $V_{DD\text{-}STR}$ while the supply of the reference ring oscillator is lowered to 0V for preventing device aging. During measurement, the supply voltage of both the stressed ring oscillator and the reference ring oscillator becomes $V_{DD\text{-}NOM}$. Once the measurement signal is triggered by the rising edge of STRESS, a phase comparator uses the reference ring oscillator output to sample the output of its stressed duplicate. The phase comparator output exhibits the beat frequency f_{stress}-f_{ref}, where f_{stress} is the stressed ring oscillator frequency and f_{ref} is the reference ring oscillator frequency. The beat frequency detection counter in Fig. 4.19 measures the beat frequency using the reference ring oscillator signal as a clock. The counter's output N is read out through the parallel-to-serial registers to calculate the percent frequency degradation. The period of the beat frequency is equal to the time when there is one clock difference between the number of reference and stress clock pulses. The beat frequency calculation is summarized in Fig. 4.22 (b).



(a)



(b)

Figure 4.22 (a) Proposed beat frequency detection circuit for high resolution NBTI monitoring. (b) Principle of proposed beat frequency detection circuit.

The beat frequency detection enables us to achieve higher resolution for the early frequency degradation where we are interested. Fig. 4.23 shows that 56% of the counter output code is assigned to the early 1% frequency degradation achieving 72X higher resolution compared to a single ring oscillator type monitor for detecting 1% change in frequency. Fig. 4.24 demonstrates the layout of the test chip.

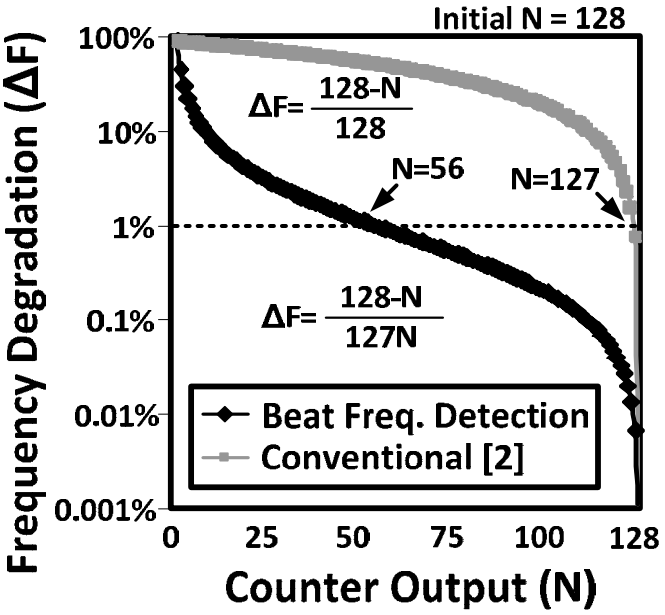


Figure 4.23 Counter output vs. frequency degradation.

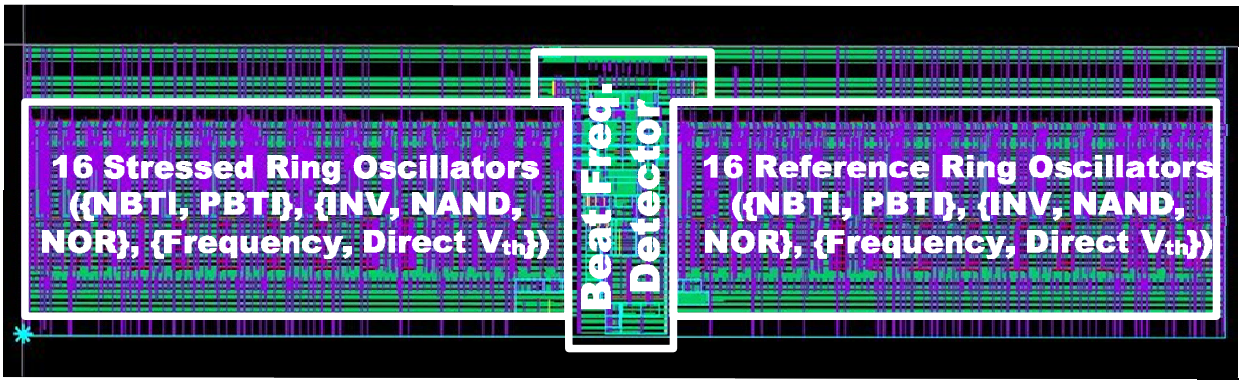


Figure 4.24 Layout of 0.7V, 32nm SOI test chip (372x90μm^2).

4.4 An SRAM Test Macro for Fully-Automated Statistical Measurements of V_{min} Degradation

4.4.1 Overview

Degradation in SRAM minimum operating voltage (V_{min}) is becoming a major reliability concern in deeply scaled process technologies. V_{min} of an SRAM cell is limited by several factors such as static noise margin (SNM), writability, and operation frequency that are determined by the relative strengths of all the devices in an SRAM cell. Negative Bias Temperature Instability (NBTI) and Time Dependent Dielectric Breakdown (TDDB), coupled with parametric device mismatch, have attracted attention as major reliability issues conspiring to worsen SRAM V_{min} [73][74][75][76].

In this paper, we analyze the impact of NBTI and TDDB on SRAM V_{min} degradation, and present a test macro for fully-automated statistical measurements of SRAM V_{min} degradation induced by NBTI. An automated test sequence collects V_{min} data for statistical analysis and reduces measurement time. Various test strategies were proposed for V_{min} measurements to identify different SRAM fail metrics such as SNM failure and access time failure. The proposed transient measurements also evaluate the speed of cell data flip affected by NBTI.

4.4.2 Previous Literatures about the Impact of NBTI on SRAM

A number of previous literatures have dealt with the impact of NBTI on SRAM cell stability and V_{min} [75][77-87]. Li et al. [75] presented the impact of NBTI on device lifetime and SRAM cell operation. It is claimed that NBTI has stronger impact on SRAM cell transition speed than on SNM. Krishnan et al. [77] showed that each transistor in an SRAM cell has different impact on V_{min} degradation. V_{min} sensitivity on NBTI is highest for a strong NMOS and weak PMOS combination. Rosa et al. [78] proposed a methodology for determining tolerable NBTI in PMOS for SRAM cell design. They measured the impact of NBTI on SRAM stability by using "N Curve" method [79]. Device parameters of an SRAM cell are influenced by the layout pattern. Fischer et al. [80] devised a test cell with slight modification from the conventional SRAM cell to minimize the layout pattern dependency of device parameters. The test cell facilitates the direct measurement of NBTI in PMOS loads with the marginal pattern dependency. An on-the-fly circuit technique for estimating threshold voltage degradation was proposed [81]. Measured results from individual stressed transistors were used to predict the median value of the threshold voltage shift due to NBTI. Kang et al. [82] proposed a methodology for the estimation of SRAM reliability by measuring leakage current. They developed a model predicting the NBTI of PMOS loads by measuring leakage current. Lin et al. [83] reported the additive impact of positive-bias temperature instability (PBTI) of NMOS drivers on SRAM cell stability. Using high-k metal-gate devices increases V_{min} even further due to the combined effect of NBTI and PBTI. A circuit technique for improving SRAM cell lifetime is described in [84]. Flipping SRAM cell data periodically removes stress and recovers

the degraded SNM. Ball et al. [85] developed a screening methodology for V_{min} drift in SRAM arrays. Measured results showed that NWELL body bias control increases the likelihood of fails coming from lowered SNM. Erratic fluctuation of SRAM V_{min} due to trapping and detrapping effects was reported by [86]. Experiment results showed that the erratic behavior follows 1/f noise spectral density. A modeling for projecting product failure was proposed in [87]. It presented that NBTI incurs systematic V_{min} shift following lognormal statistics, and TDDB generates a single bit error distributed randomly in an SRAM array following Weibull statistics.

Most of the previous works characterize V_{min} based upon simulation or single SRAM cell measurements. However, a single cell test through manual probing is impractical for obtaining statistical V_{min} data when a large number of cells have to be measured. Array-based structures and automated measurements are indispensible for efficiently and accurately gathering the V_{min} degradation statistics.

4.4.3 Impact of NBTI and TDDB on SRAM V_{min}

A. V_{min} Degradation due to NBTI and TDDB

Two major reliability concerns pertaining to a 6T SRAM cell, namely NBTI and TDDB, and the cell butterfly curves are shown in Fig. 4.25. The node storing '0' (Q) stresses the PMOS load (M5) causing NBTI that increases the threshold voltage. This lowers the trip point of the right inverter formed with M4 and M5, degrading the cell read stability. Similarly, the node storing '1' (QB) causes TDDB in the NMOS driver (M1) generating a current path through the gate dielectrics. This can be simply modeled by a resistor between the gate and the source as shown in Fig. 4.25. TDDB in M1 prevents QB from being raised up to VDD due to the pull-down current path. This also pushes down the transfer curve of the right inverter combined with the NMOS driver reducing the cell read stability. As the cell read stability declines, the cell data becomes more likely to flip which leads to an increase in supply voltage with stress.

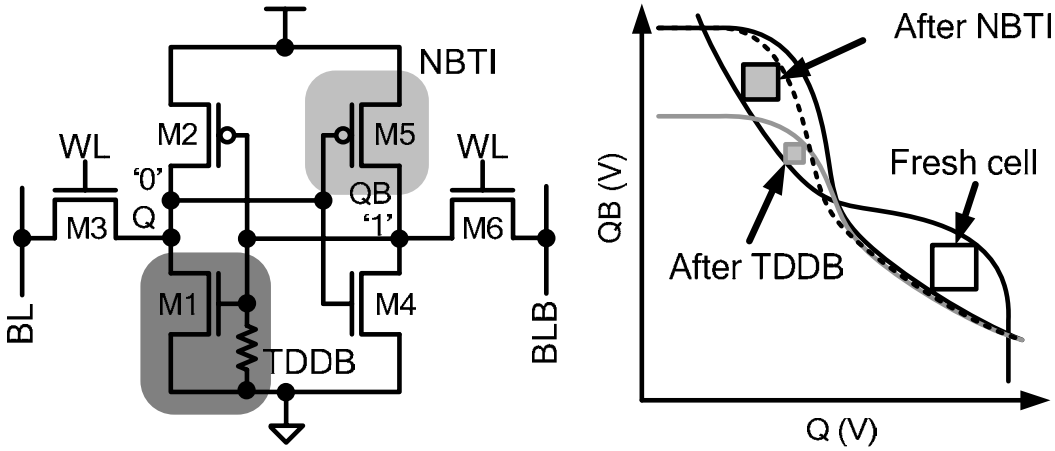


Figure 4.25 Impact of NBTI and TDDB on read static noise margin of a 6T SRAM cell

V_{min} is defined as the minimum supply voltage where an SRAM is operating without failure. Several components such as read stability, writability, and operational frequency restrict the V_{min} of an SRAM cell. V_{min} can be found when one of these becomes lower than a failure threshold. NBTI reduces the read stability, but improves the writability [88] since the writability is determined by the relative strength of access devices to that of PMOS loads. If the SRAM cell is limited by the read stability before stress, NBTI increases the V_{min}. However, if the writability is the limiter for V_{min}, NBTI will improve V_{min} until the Vmin for write operations becomes equal to that of read operations. This is due to the better writability after stress even though it still reduces the SNM. In DC analysis, the failure threshold is set to the point where the SNM becomes zero. However, at a higher clock frequency, V_{min} is more likely to be limited by the access time (T_{aa}) than the SNM failure. The NBTI in a PMOS load increases the access time and increases V_{min}. Therefore, the V_{min} is determined at the supply level where the access time requirement is met.

Fig. 4.26 shows the simulated SNM and SNM degradation due to NBTI at different supply levels. The SNM-limited V_{min} is found when the original SNM is equal to the SNM degradation leading to zero SNM. As the threshold voltage and SNM degradation becomes larger, V_{min} increases accordingly. In general, the SNM-limited V_{min} occurs at lower supply levels. If the supply voltage is high enough, NBTI will degrade the SNM, but not flip the cell data. To flip the cell data, the original SNM should be small enough to be zero or negative after the SNM degradation due to NBTI.

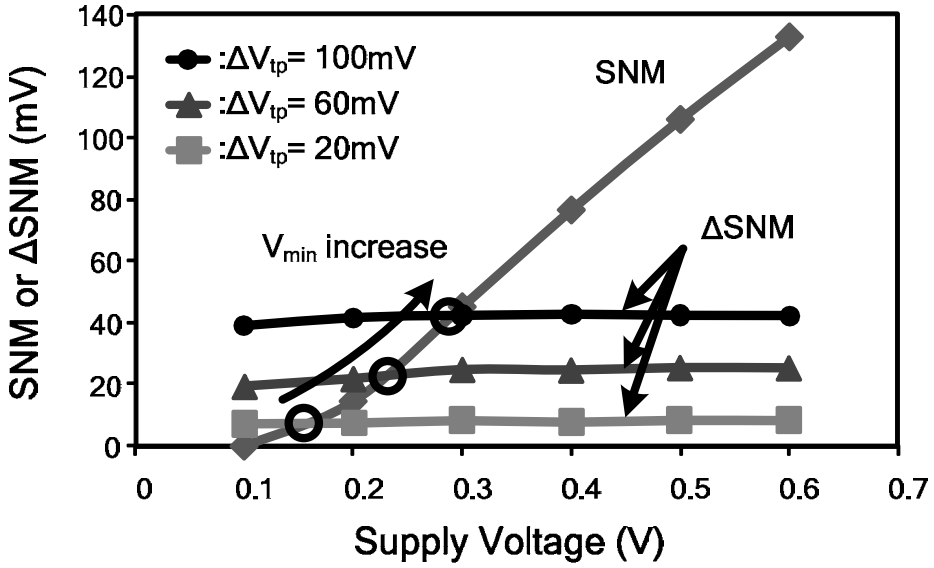


Figure 4.26 Impact of threshold voltage degradation on SNM and V_{min} degradation.

B. Other Factors Affecting V_{min}

V_{min} is determined by the combined parameters of all devices in an SRAM cell. In addition to the NBTI, the SRAM V_{min} is also subject to initial parametric mismatches in devices, and column data in a bitline. The device mismatches make an SRAM cell have two different V_{min} values, V_{min} for data '0' and V_{min} for data '1'. Stressing one PMOS load makes V_{min} for data '0' and that of data '1' move in the opposite direction. Fig. 4.27 explains the effect of NBTI on the butterfly curves. The NBTI in a PMOS load (M5) degrade the SNM of data '0', but improves the SNM of data '1'. This increases the V_{min} for data '0', but decreases the V_{min} for data '0'. Fig. 4.28 illustrates the mingled effect of device mismatches and selection of stressed device on the SRAM V_{min}. Before stress, the V_{min} for data '1' is larger than that of data '0' in both scenarios since the PMOS load at the left side is weaker. The SRAM cell in Fig. 4.28 (a) will experience the V_{min} degradation of data '0' and the improvement of data '1' since it is stressed with data '0'. Similarly, the SRAM cell in Fig 4.28 (b) will have the V_{min} degradation of data '1' after stress. As a result, the V_{min} of the SRAM cell in Fig, 4.28 (a) will improve until two V_{min} values are the same. The worst case V_{min} degradation of a single SRAM cell appears when the weaker device is stressed as shown in Fig. 4.28 (b).

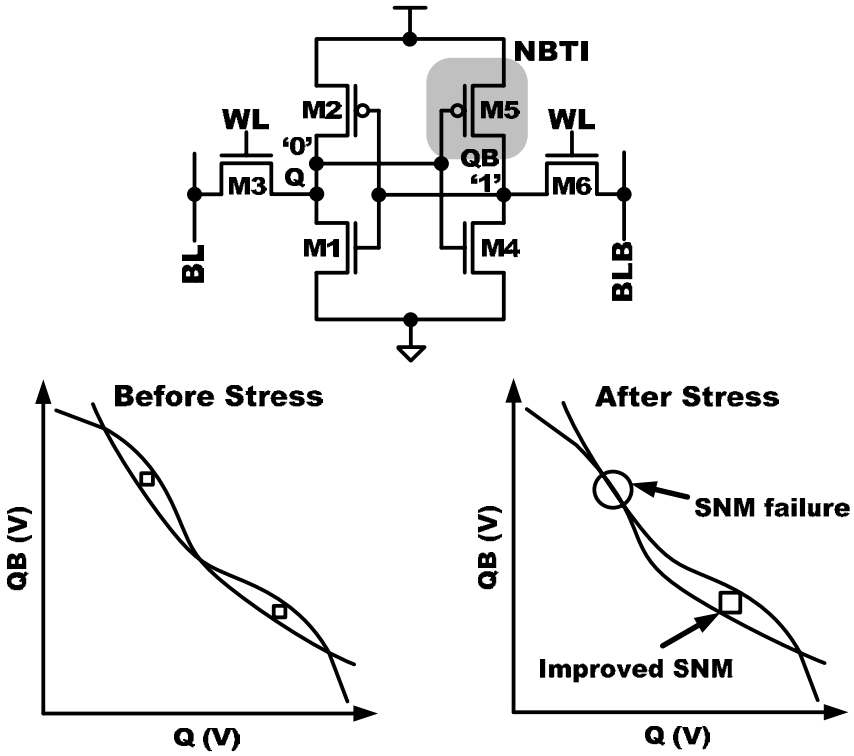
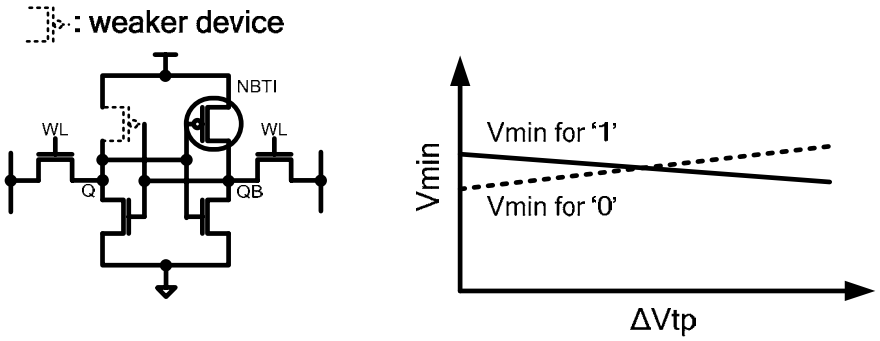
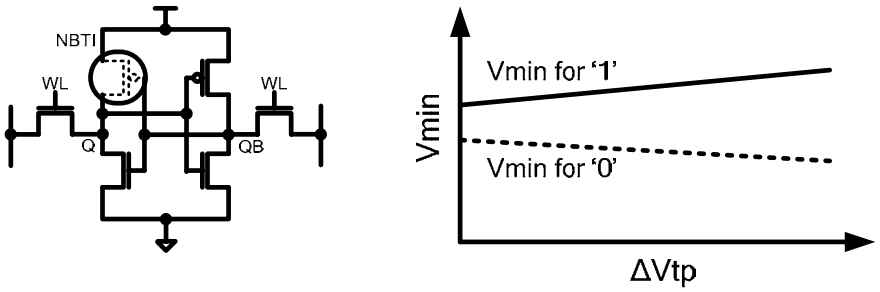


Figure 4.27 V_{min} affected by selection of stressed device. Stressing with data '0' degrades V_{min} for data '0', but improves V_{min} for data '1'.



(a)



(b)

Figure 4.28 V_{min} affected by the combined effect of initial device mismatch and selection of stressed device. (a) Weak '1' cell stressed with '0'. (b) Weak '1' cell stressed with '1'.

The SNM-limited V_{min} is observed when the supply voltage is around or below the transistor threshold voltage. In this region, SRAM read operations are highly affected by the bitline leakage current. The V_{min} measured from a stand-alone SRAM cell is different from that from a realistic bitline structure. To include the impact of the bitline leakage current in measurements, the bitline should be written with the worst case data pattern before the V_{min} measurements. If the accessed cell stores data '0', the rest cells in the same bitline should be written with data '1', and vice versa. Fig. 4.29

shows the simulated bitline waveforms influenced by the bitline leakage current. HSPICE simulation result shows that the bitline discharge speed and the bitline sensing margin are very sensitive to the column data pattern.

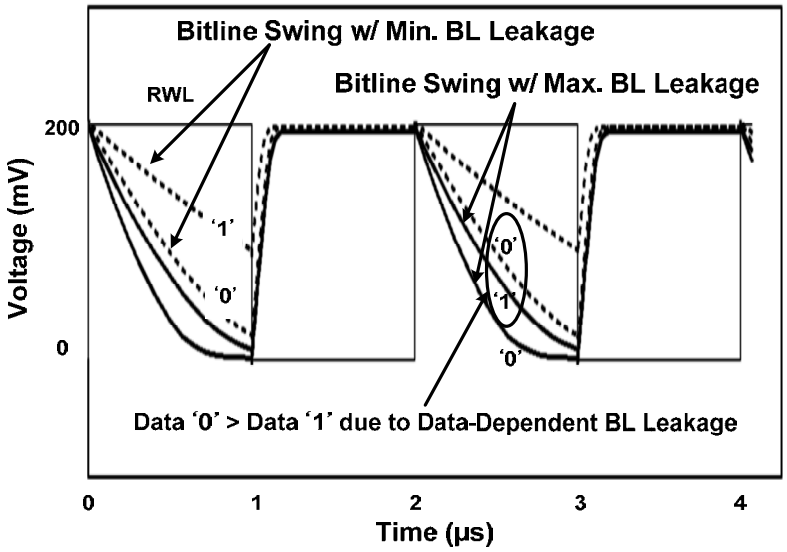
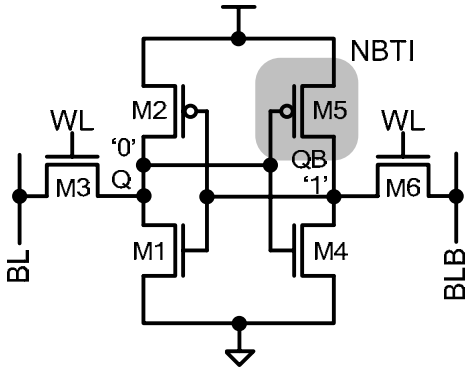


Figure 4.29 Simulated data dependency of RBL waveforms during read operations [22]. Supply voltage is 0.2V.

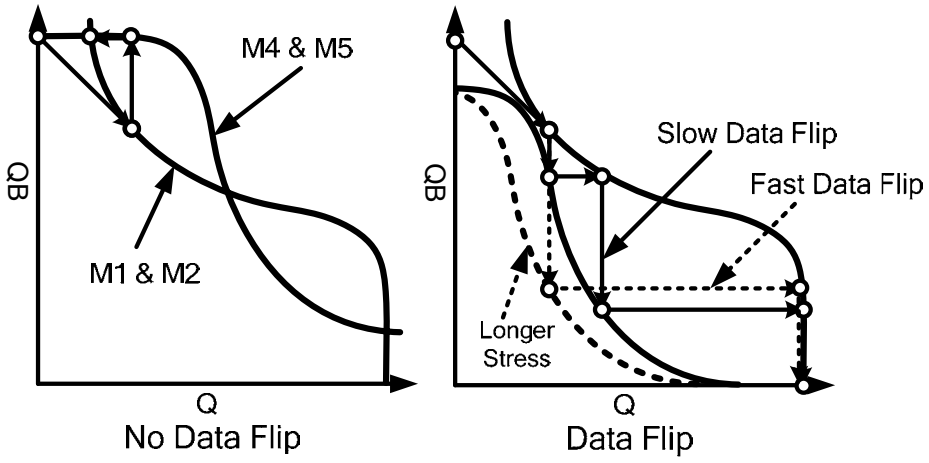
C. Speed of Cell Data Flip due to NBTI

The amount of NBTI degradation determines the speed of the cell data flip when the SNM is zero or below zero after stress. Fig. 4.30 illustrates the schematic of 6T SRAM cell and the butterfly curves of SRAM cells during a read operation. When a wordline (WL) is enabled, the cell node with data '0' (Q) rises and weakly turns on the NMOS driver (M4). The other cell node (QB) is generated by the relative strength ratio between the PMOS load (M5) and the NMOS driver (M4). The NBTI in the PMOS load (M5) pulls down the right inverter output. This is applied to the left

inverter changing the left cell node level (Q). The larger degradation in the PMOS load (M5) raises the right cell node (QB) more, making the cell data more likely to flip. If the SNM is positive after stress, the cell data is kept by the feedback operation as shown in Fig 6 (b) since the initial rise in Q is not high enough for the feedback operation to flip the data.



(a)



(b) (c)

Figure 4.30 (a) Schematic of 6T SRAM cell with NBTI in a PMOS load (M5). (b) No data flips occurs when SNM is positive. (c) Larger NBTI due to longer stress leads to faster data flip.

If the NBTI degradation is large enough to make the SNM zero or below zero, the feedback operation makes the cell data flip (Fig. 4.30 (c)). However, the larger NBTI degradation degrades the SNM more and reduces the number of feedback operation for the data flip. The dotted line shows a less number of feedback operations compared to that of the solid line. This represents that the data in the SRAM cell following the dotted line flips faster than that in the solid line. Fig. 4.31 demonstrates the simulated waveforms comparing the data flip speed and the bitline sensing margin degradation due to the NBTI in the PMOS load (M5). The larger NBTI degradation makes the bitline crossing point occur earlier reducing the bitline sensing margin. The time to data flip at different supply levels is summarized in Fig. 4.32. The speed of data flip is more sensitive to the NBTI degradation at lower supply voltage, and is easily observed accordingly.

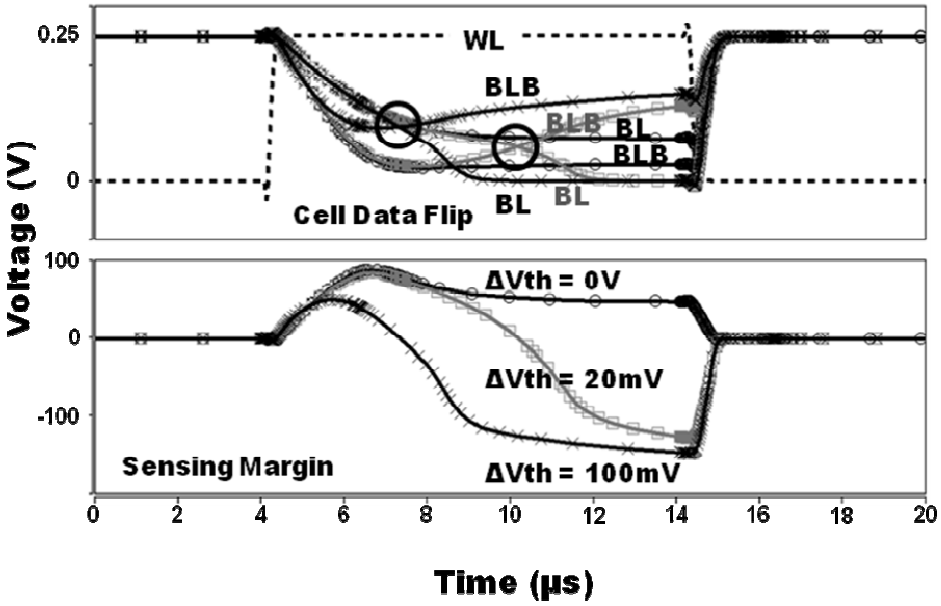


Figure 4.31 Simulated bitline waveforms with different threshold voltage degradations. Larger threshold voltage degradation shows faster cell data flip.

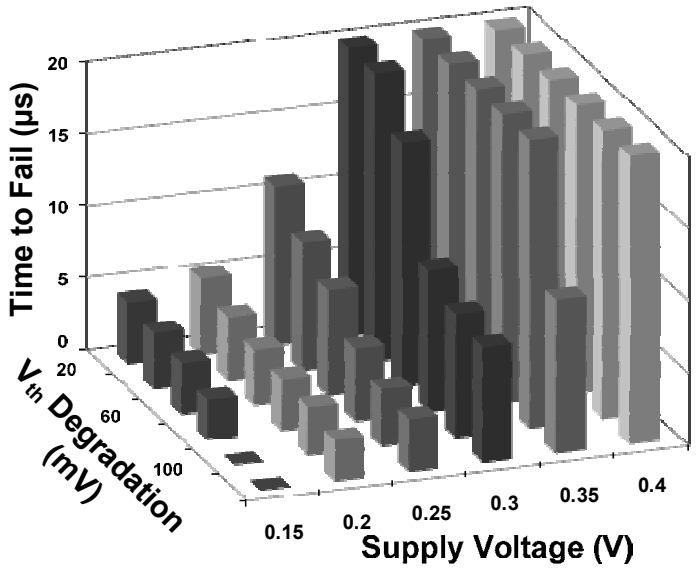


Figure 4.32 Simulated time to cell data flip due to NBTI varying supply voltage.

4.4.4 SRAM Test Macro Design

The simplified architecture of the SRAM test macro is illustrated in Fig. 4.33. The 2k bit SRAM array consists of 128 rows and 16 columns. The test chip is divided into four test units and each test unit is composed of a mini sub-array (128 rows × 4 columns), power switches, sense amplifiers, and write drivers. Four test units share control circuits, row circuits, output multiplexers, level converters, and output drivers.

Fig. 4.34 illustrates the core SRAM macro circuits for V_{min} degradation measurements. During a measurement, only the supply of the selected test unit is enabled for stress and a V_{min} test. SRAM cells in the rest sub-arrays are in a fresh mode where all supply levels are grounded to keep SRAM cells free of stress for later evaluations. In the selected array, stress voltage (V_{STRESS}) or measurement voltage (V_{MEAS}) is connected to the cell supply (V_{CELL}) based upon the SRAM V_{min} test modes. V_{STRESS} accelerates NBTI in the stress mode, and V_{MEAS} is used as cell supply voltage during the V_{min} measurement period. To avoid the impact of the supply change on other circuits, V_{MEAS} is only used in SRAM sub-arrays, last stages of wordline drivers, write drivers, and sense amplifiers. The rest SRAM circuits use the nominal supply voltage, V_{DD_NOM} for reliable operations. Level converters are used for interfacing V_{MEAS} and V_{DD_NOM}. PMOS devices are used as input transistors since they are more efficient than NMOS in dealing with a lower signal level.

A V_{min} measurement is executed through regular SRAM write and read operations, decrementing the cell supply voltage from the highest V_{MEAS} level (V_{MIN_HIGH}) to the lowest V_{MEAS} level (V_{MIN_LOW}). Since the V_{MEAS} is used in the SRAM sub-array, the last stage of wordline driver, the write drivers, and the sense

amplifiers, the measured V_{min} is also affected by these circuits. The measurement starts by writing data to the selected mini SRAM sub-array. After finishing write operations, stress is applied to accelerate the NBTI by raising the cell supply voltage. During V_{min} measurement interrupts, stress is removed and reading operations are performed with V_{MEAS} controlled by a test program whose sequences will be discussed in section II-B. V_{min} of an SRAM cell is found when the read data is different from the original write data.

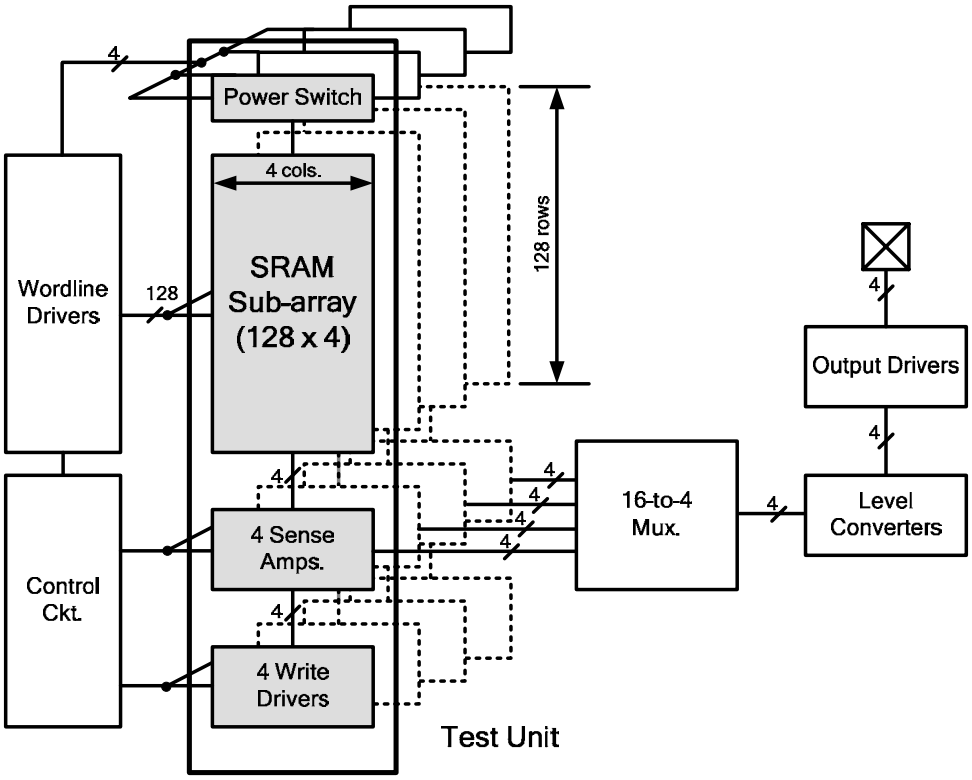


Figure 4.33 Test macro architecture.

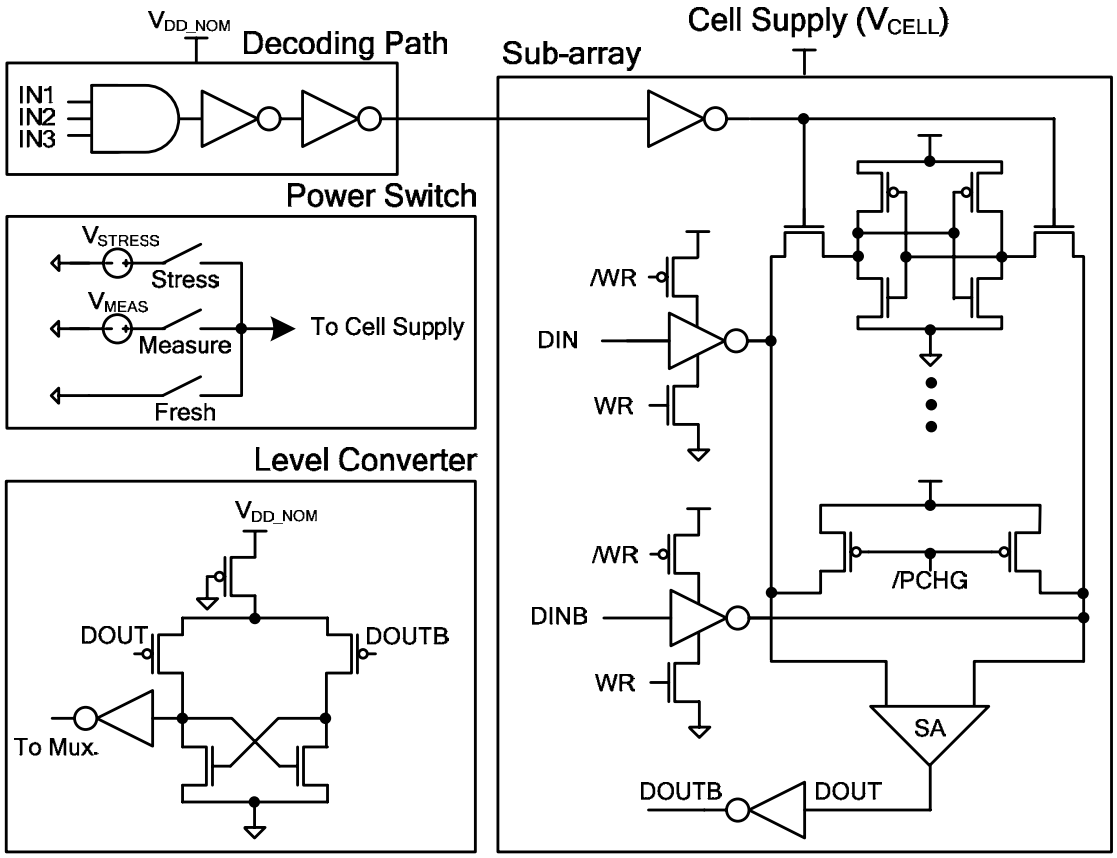


Figure 4.34 Core SRAM circuits and different power supply domains for fast transient V_{min} measurements.

4.4.5 Test Sequence for V_{min} Degradation Measurement

Automated measurements effectively gather statistical V_{min} data and reduce test time. Fig. 4.35 illustrates the fully-automated test sequence for a V_{min} degradation measurement. Each measurement sequence consists of three sub-sequences: an initialization sequence, a stress sequence, and a V_{min} measurement sequence. Before applying stress, write operations are performed as the initialization sequence. The initialization is required to select the devices to be stressed in SRAM cells. After the initialization, the power switches apply V_{STRESS} to the cell supply to stress the devices selected in the initialization sequence. Note that no stress period is needed at the initial fresh V_{min} measurement. After the stress sequence, the V_{min} degradation measurement starts setting the cell supply voltage to V_{MIN_HIGH}. Delay should be inserted after the stress and before the beginning of read operations to wait for the cell supply to settle down. After this delay, the read (or read after write to detect V_{min} for both data) operations are executed sweeping addresses until all the SRAM cells in the selected test unit are accessed. These read operations are repeated decrementing the cell supply voltage from V_{MIN_HIGH} to V_{MIN_LOW}.

Since stress is removed during the V_{min} measurement sequence, the total measurement time should be kept short for an accurate V_{min} degradation measurement without NBTI recovery. The number of read operations is calculated by multiplying the number of the cell supply levels and the number of addresses as shown in Fig. 4.35. Raising the clock frequency increases the number of cells that can be measured without NBTI recovery. However, SNM-limited V_{min} usually occurs at a lower supply voltage where the number of measurable cells at a time should be significantly

reduced compared to that in the measurement using a higher frequency clock. If V_{min} is too low enough to slow down SRAM operations below the minimum allowable frequency, the experimental results will include NBTI recovery. Utilizing this frequency dependency of V_{min}, the test macro allows V_{min} measurements for different SRAM failure modes.

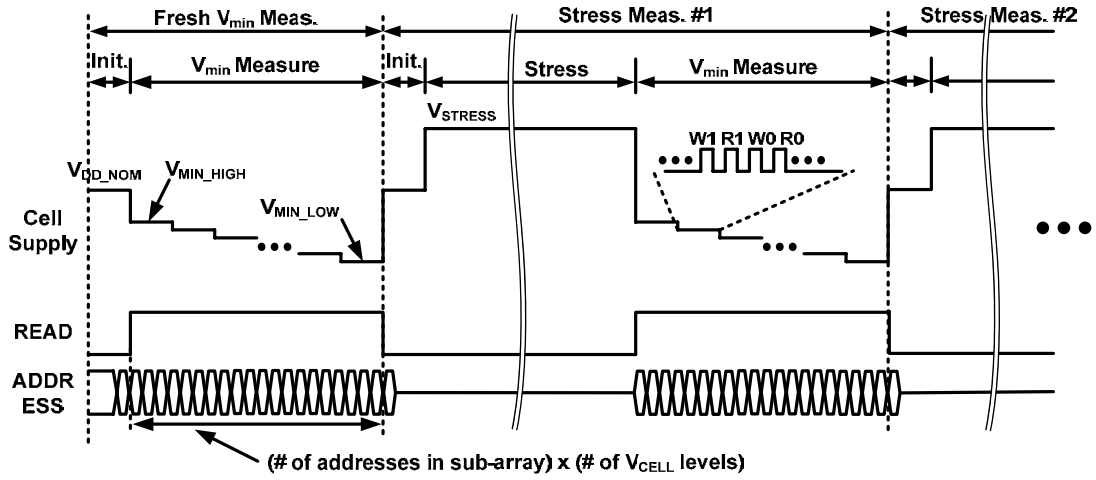
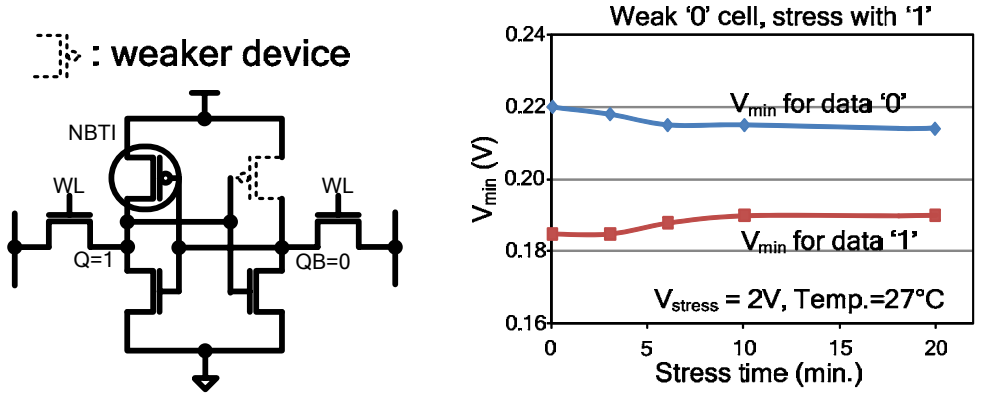


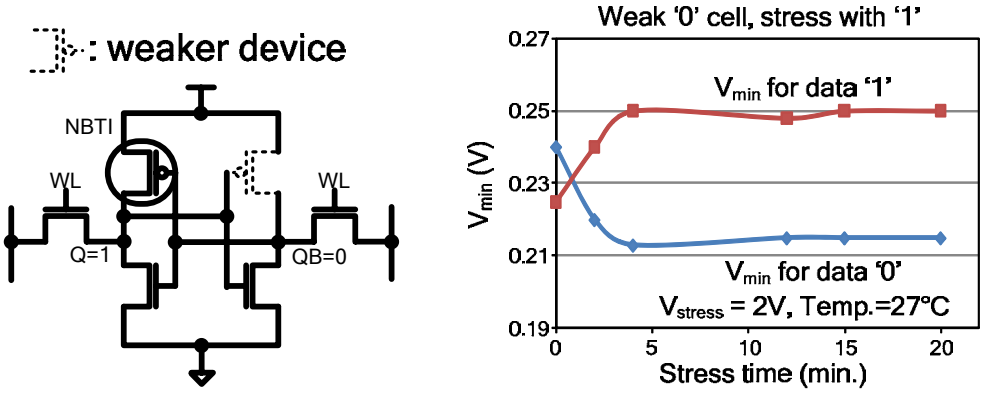
Figure 4.35 Automated test sequence for large-scale SRAM stress measurements.

4.4.6 V_{min} Degradation Measurements

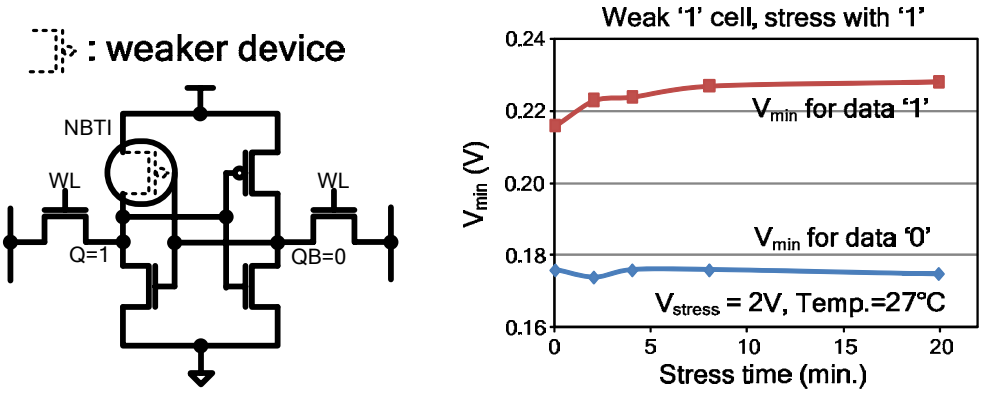
When measuring V_{min} degradation of a single SRAM cell, the initial device parametric mismatch should be considered in measurements. The initial parametric mismatch generates two cell types: a weak '0' cell which has a higher V_{min} for data '0' and a weak '1' cell which has a higher V_{min} for data '1'. In addition to the device mismatch, the direction of V_{min} change should also be deliberated. Stressing a cell storing data '1' increases the V_{min} for data '1' while that of data '0' declines since the weaker strength of the stressed device improves the SNM of data '0'. Fig. 4.36 demonstrates the measured V_{min} degradation of single SRAM cell with mismatch and data dependency. When stressing a weak '0' cell storing '1', V_{min} improves until the V_{min} values of both data become equal (Fig. 4.36 (a, b)). V_{min} worsens after this point (Fig. 4.36 (b)). Obtaining the worst case V_{min} degradation would require a weak '1' cell to be stressed while storing '1'and vice versa for a weak '0' cell (Fig. 4.36 (c)). For statistical measurements however, the above steps can be omitted since V_{min} distribution for data '1' will be identical to that of data '0' following a Gaussian distribution as device mismatches does. All SRAM cells are stressed with data '1' in our statistical measurements.



(a)



(b)



(c)

Figure 4.36 Single cell V_{min} degradation when stressed with data '1'. If a cell storing data '1' is stressed, V_{min} for data '1' worsens while V_{min} for data '0' improves. The change in V_{min} depends on the initial parametric mismatch as well as the stress mode data: (a) - (b) Weak '0' cells. (c) Weak '1'cell.

When stress is removed, V_{min} quickly drops due to the fast recovering nature of NBTI. Fig. 4.37 shows the measured V_{min} for repetitive stress and no stress periods at two different clock frequencies. It can be seen that V_{min} is more sensitive to NBTI at a lower supply voltage where transistor current follows an exponential function of device threshold voltage. V_{min} also depends on clock frequency; V_{min} at 1 MHz is higher than that at 10 kHz. At a higher clock frequency, access time failure happens before an SNM failure does. V_{min} of a single cell can be sampled within a few microseconds so the measurement results in Fig. 4.36 and 4.37 (top) are free of NBTI recovery.

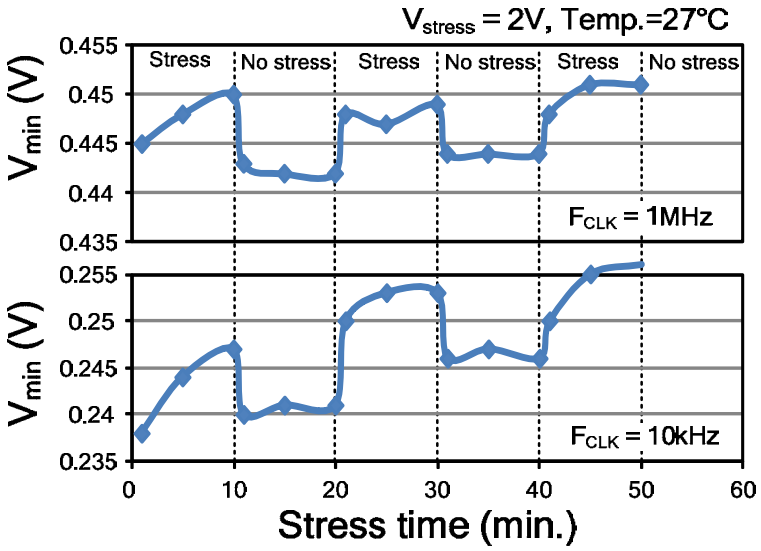


Figure 4.37 V_{min} for alternating stress and no stress periods showing NBTI recovery.

Fig. 4.38 illustrates the cumulative distribution function (CDF) of the measured SRAM failures. The V_{min} distribution using a 10 kHz clock frequency is 90 mV wider than that of 1MHz clock frequency since the SRAM is operating in the sub-threshold region where device current is an exponential function of threshold voltage. Fig. 4.39 shows the V_{min} degradation measured from 32 SRAM cells. For a 2.0 V stress voltage, a 30 mV degradation in V_{min} was measured after one hour of stress. The V_{min} spread is 80mV prior to stress and 100mV after stress.

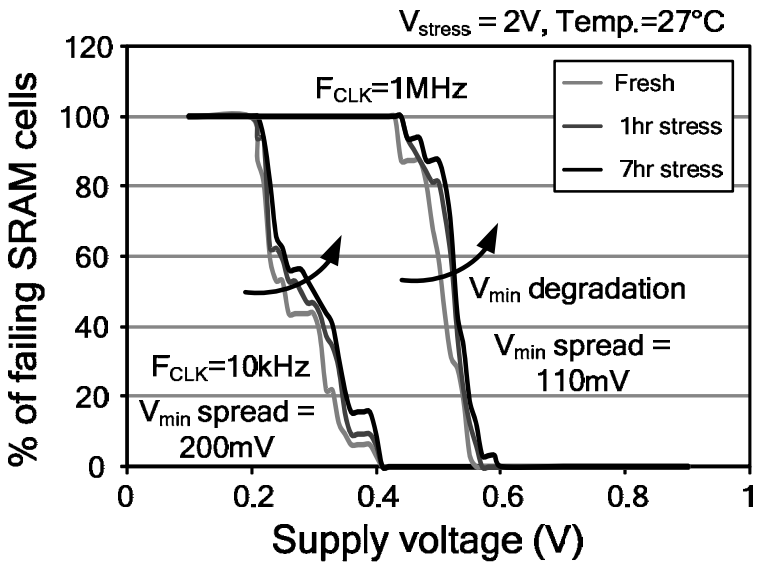


Figure 4.38 Measured cumulative V_{min} distribution for two clock frequencies.

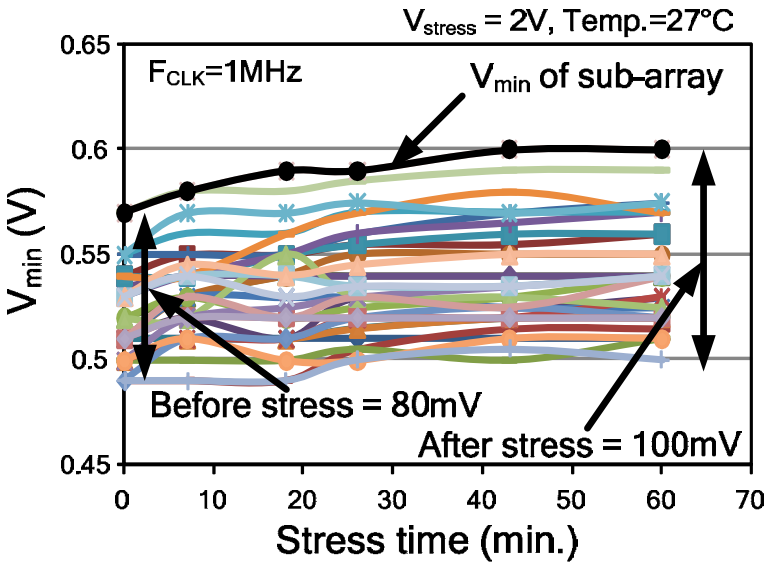


Figure 4.39 Measured V_{min} degradation versus stress time for multiple SRAM cells.

As explained in section II-B, V_{min} is also affected by the column data in a bitline. Fig. 4.40 presents the measured data dependency of V_{min}. Higher V_{min} is obtained when the accessed cell data is opposite to the majority of the column data. V_{min} improves as the number of cells storing the same data as the accessed cell increases. The measured V_{min} varies up to 23mV by the change in column data. Fig. 4.41 shows the frequency dependency of V_{min} measured from a sub-array. V_{min} of a cell shows constant value at a lower clock frequency since SNM is the limiter. V_{min} increases after a point where access time failure starts to occur. It shows an order of variations in clock frequency for the measured SRAM cells to have the same V_{min}.

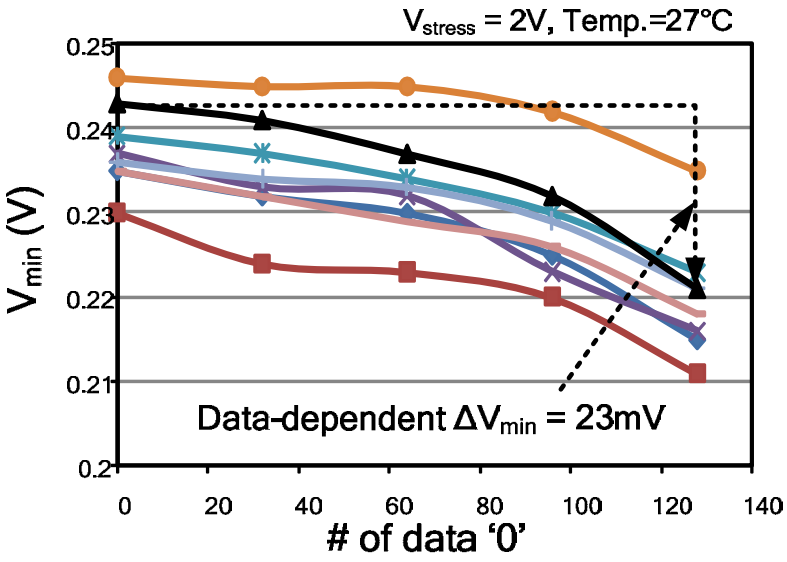


Figure 4.40 Measured V_{min} affected by the column data pattern.

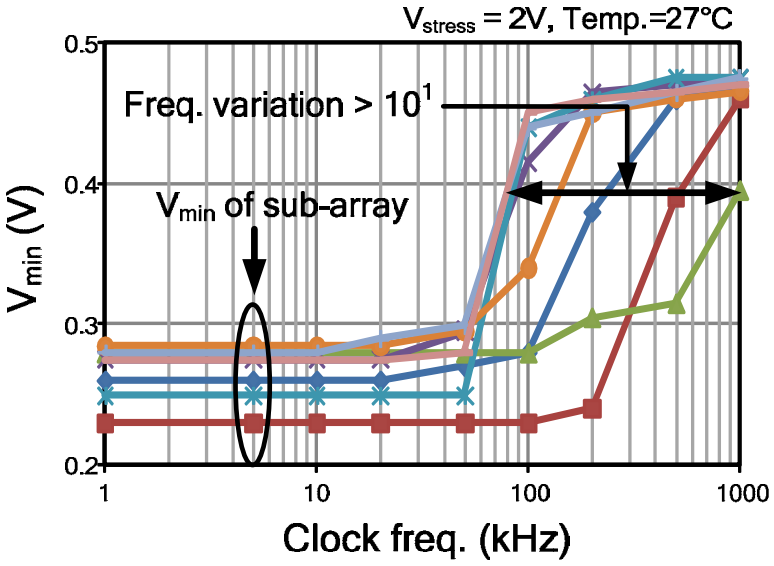


Figure 4.41 Measured V_{min} versus clock frequency.

The SNM failure is a dominant factor determining V_{min} at a low input clock frequency. Since the SNM failure flips the SRAM cell data, the original data does not recover by increasing the cell supply level. Fig. 4.42 (left) shows waveforms when the SNM failure occurs. However, access time can also limit V_{min}. In this case, cell data survives even though an SRAM read operation fails. The original cell data is observed by increasing the cell supply level (Fig. 4.42 (right)).

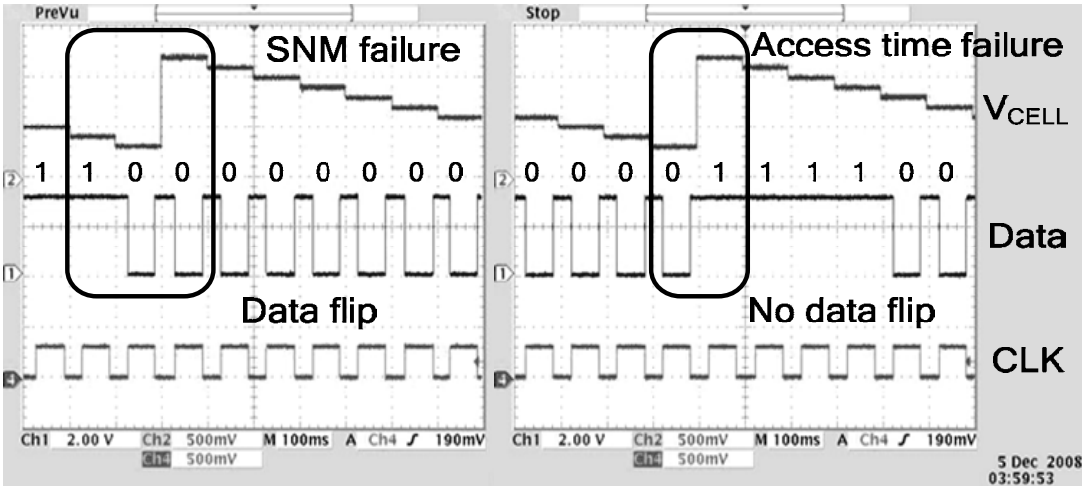


Figure 4.42 SNM failure scenario causes a cell data to flip (left). Access time failure scenario causes a transient fault (right).

Measuring the speed of data flip is informative in estimating the amount of NBTI. This information can also be used to prevent the data flip and improve V_{min} by controlling the wordline activation time [5]. Fig. 4.43 demonstrates measurement results of the speed of data flip affected by NBTI through a stress/measurement test. It demonstrates that the data flip points happens earlier as NBTI gets larger. The NBTI degradation explained by RD model is faster at an early stress period and slows down. The measured time to data flip also shows the larger change at the earlier stress/measurement test. The test chip microphotograph is shown in Fig. 4.44.

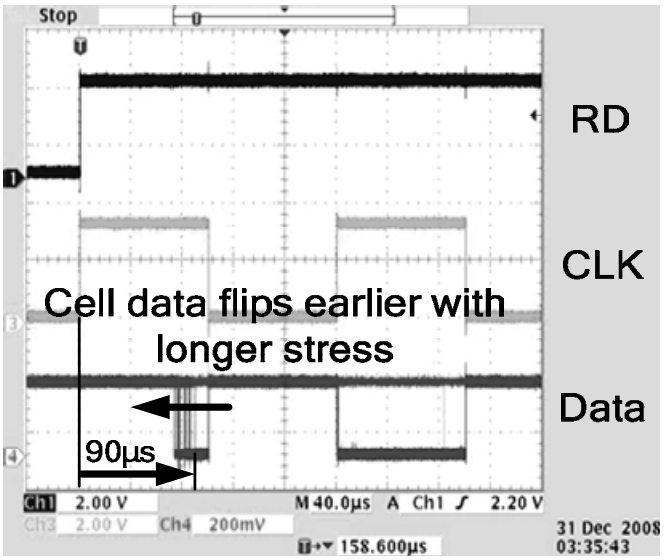


Figure 4.43 A longer stress time reduces the time for the cell data to flip which is caused by an SNM failure.

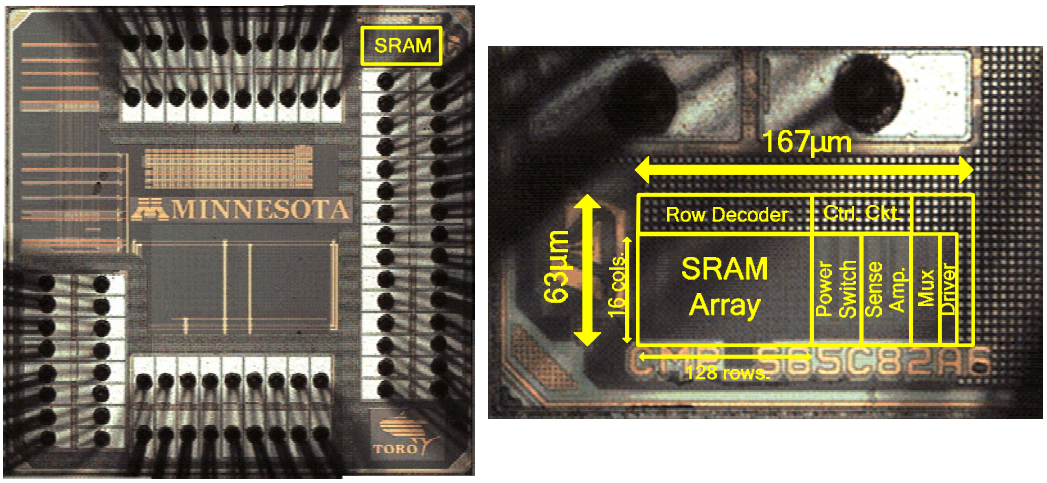


Figure 4.44 Microphotograph of the test chip

4.5 Conclusions

NBTI is a growing threat to circuit reliability due to its increased impact on circuit performance with each new technology node. By helping designers to monitor aging during circuit operation, on-chip NBTI monitoring structures can help them to better understand these effects, and ultimately assist them in building aging-tolerant systems.

First, we propose a silicon odometer based on a beat frequency detection scheme, which enabled the sensing resolution of 0.02% (or 0.8ps) with a 4ns period ring oscillator. The minimum time required for obtaining one data sample was 2μs which was short enough to avoid any noticeable recovery effects. In addition, the differential measurement approach minimized the effect of common mode environmental variations. The silicon odometer was implemented using fully digital circuits which only require a minimum of calibration. Various operation modes have been implemented and tested. Measurement results illustrate basic characteristics of NBTI that are in line with the effects described in previous works. Finally, the relationship between true inverter chain frequency degradation and ring oscillator frequency degradation was analyzed in both the DC as well as the AC stress cases. The measured ring oscillator frequency degradation during DC stress is 50% of that of the true inverter chain, and is equal to that of the true inverter chain frequency during AC stress.

We have also presented test structures for isolated NBTI and PBTI effects in digital circuits in high-k metal-gate CMOS technology. Two different types of structures are proposed: One is for measuring frequency degradations, and the other is

for measuring threshold voltage degradations. The proposed structures facilitate the precise estimation of the portion of degradation due to NBTI and PBTI in logic gates. This is also derived by calculations using simplified RC parameters. The proposed structures can also be easily extended to other logic gates. A beat frequency detection scheme achieves 72X higher frequency sensing resolution compared to single ring oscillator based monitors. A test chip was implemented in a 0.9V, 32nm high-k metal-gate SOI process technology.

Finally, we have presented an SRAM test macro for fully-automated characterization of V_{min} degradation due to NBTI. An automated test program facilitates large-scale measurements of V_{min} degradation and reduces test time. The proposed test structure can also measure V_{min} degradation from different SRAM failure modes: (a) the SNM-limited case and (b) the access-time-limited case. The impact of NBTI on the time to cell data flip has been measured, which can be used to model the temporal impact of NBTI on SRAM operation. A test chip was implemented in a 1.2V, 65nm CMOS process technology.

Chapter 5 Conclusions

Power has continuously increased over technology generations, becoming significant concerns for circuit designers. The ever-increasing power consumption is mainly due to the supply voltage that has not been scaled as transistor dimensions. Recently, energy efficient systems are becoming more and more popular where minimal energy consumption is the primary design constraint. Since minimum energy point can be achieved in sub-threshold region, circuit techniques for sub-threshold operations are highly required. Circuit variability is another major issue in nano-scale technologies. Transistor aging is becoming one of the most pressing sources of circuit variations with each technology node because of the continuously increased electric field inside a transistor. Therefore, accurate transistor aging monitoring is critical in aging-tolerant system design. This thesis explores circuit techniques for reliable sub-threshold circuit design and on-chip circuit reliability monitoring.

In chapter 2, we propose a device-size optimization method for sub-threshold circuits utilizing reverse short-channel effect (RSCE). Due to the combined effect of RSCE and the exponential dependency of current drivability on threshold voltage, the maximum performance is achieved at 3X channel. Higher drive current, low device capacitance, less sensitivity to random dopant fluctuations, better sub-threshold swing, and improved energy dissipation were achieved through the longer channel devices.

In chapter 3, two sub-threshold SRAMs implemented in 130nm process technology are described. We apply the proposed sizing method to SRAMs to improve writability and read performance. A 10-T SRAM cell with data-independent bitline leakage is proposed to improve cell stability and bitline sensing margin. An inverter-based sense amplifier with optimal noise margin is also designed with the aid of VGND replica scheme. The proposed 10-T SRAM is fully functional down to 0.2 V. We also present an 8-T sub-threshold SRAM with a bitline leakage compensation scheme. The proposed marginal bitline leakage compensation (MBLC) scheme combined with the floating write bitlines lowers the supply voltage from 0.28 V down to 0.23 V. The floating read bitlines reduces leakage current by 60%, and the deep sleep mode decreases the standby current by 58% when both supply rails are raised by 0.2 V.

In chapter 4, we present three on-chip circuit reliability monitoring techniques. First, a fully-digital on-chip reliability monitor for high resolution frequency degradation measurements of digital circuits is proposed. The proposed technique of the beat frequency detection of two ring oscillators achieves 50X higher delay sensing resolution than prior techniques. In addition, we show ring oscillator based test structures that can separately measure the NBTI and PBTI degradation effects for high-k metal-gate technologies. Finally, we present an SRAM test macro for statistical measurements of SRAM V_{min} degradation induced by NBTI. A fully-automated test sequence collects V_{min} data for statistical analysis and reduces measurement time. Various test strategies were proposed for V_{min} measurements to identify different SRAM fail metrics such as SNM failure and access time failure.

References

[1] K. L. Wong, T. Rahal-Arabi, M. Ma, et al., "Enhancing Microprocessor Immunity to Power Supply Noise with Clock/Data Compensation," *IEEE Journal of Solid-State Circuits,* vol. 41, no. 4, pp. 749-758, Apr. 2006.

[2] Pat Gelsinger, "The Era of Tera," Design Automation Conference, Keynote, 2004

[3] Soeleman, K. Roy, B. Paul, "Robust sub-threshold logic for ultra-low power operation", IEEE Transactions on VLSI Systems, Volume 9, Issue 1, pp. 90-99, Feb. 2001.

[4] A. Bryant, J. Brown, P. Cottrell, M. Ketchen, J. Ellis-Monaghan, E.J. Nowak, "Low-power CMOS at Vdd=4kT/q", Device Research Conference, pp. 22-23, 2001.

[5] B. Zhai, S. Hanson, D. Blaauw, D. Sylvester, "Analysis and mitigation of variability in sub-threshold design", International Symposium on Low Power Electronics and Design, pp. 20-25, Aug. 2005.

[6] E. Vittoz, J. Fellrath, "CMOS analog integrated circuits based on weak inversion operations", IEEE Journal of Solid-State Circuits, Volume 12, Issue 3, pp. 224-231, June 1977.

[7] A. Wang, A.P. Chandrakasan, "A 180-mV sub-threshold FFT processor using a minimum energy design methodology", IEEE Journal of Solid-State Circuits, Volume 40, Issue 1, pp. 310-319, Jan. 2005.

[8] B.H. Calhoun, and A. Chandrakasan, "A 256k Sub-threshold SRAM using 65nm CMOS," International Solid-State Circuits Conference, pp. 628-629, Feb. 2006.

[9] B. Calhoun, A. Chandrakasan, "Ultra-dynamic voltage scaling using sub-threshold operation and local voltage dithering in 90nm CMOS", International Solid-State Circuits Conference, pp. 300-301, Feb. 2005.

[10] C.H. Kim, H. Soeleman, K. Roy, "Ultra-low-power DLMS adaptive filter for hearing aid applications", IEEE Transactions on VLSI Systems, Volume 11, Issue 6, pp. 1058-1067, Dec. 2003.

[11] J.J. Kim, K. Roy, "Double gate-MOSFET sub-threshold circuit for ultra-low power applications", IEEE Transactions on Electron Devices, Volume 51, Issue 9, pp. 1468-1474, Sept. 2004.

[12] V. Huard, M. Denais, "Hole Trapping Effect on Methodology for DC and AC Negative Bias Temperature Instability Measurements in PMOS Transistors," IEEE International Reliability Physics Symposium, pp. 40-45, April 2004.

[13] M. Denais, V. Huard, C. Parthasarathy, et al., "New Perspectives on NBTI in Advanced Technologies: Modeling & Characterization," IEEE European Solid-State Device Research Conference, pp. 399-402, September 2005.

[14] R. Vattikonda, W. Wang, Y. Cao, "Modeling and Minimization of PMOS NBTI Effect for Robust Nanometer Design," IEEE Design Automation Conference, pp. 1047-1052, July 2006.

[15] M. Ershov, R. Lindley, S. Saxena, et al., "Transient Effects and Characterization Methodology of Negative Bias Temperature Instability in

PMOS Transistors," IEEE International Reliability Physics Symposium, pp. 606-607, April 2003.

[16] C. Lee, G. Yang, J. Park, et al., "A Unified Compact Model of the Gate Oxide Reliability for Complete Circuit Level Analysis," IEEE International Electron Devices Meeting, pp. 549-552, Dec. 2007.

[17] R. Wang, R. Huang, D. Kim, et al., "New Observations on the Hot Carrier and NBTI Reliability of Silicon Nanowire Transistors," IEEE International Electron Devices Meeting, pp. 821-824, Dec. 2007.

[18] A.M. Yassine, H.E. Nariman, M. McBride, et al., "Time dependent breakdown of ultrathin gate oxide," IEEE Transactions on Electron Devices, Volumn 47, Issue 7, pp. 1416-1420, July 2000.

[19] F. Chen, M. Shinosky, B. Li, et al., "Critical ultra low-k TDDB reliability issues for advanced CMOS technologies," IEEE International Reliability Physics Symposium, pp. 464-475, April 2009.

[20] T. Kim, J. Keane, H. Eom, C. Kim, "Utilizing Reverse Short-Channel Effect for Optimal Subthreshold Circuit Design," IEEE Transactions on VLSI Systems, Volume 15, no. 7, pp. 821-829, July 2007.

[21] T. Kim, J. Liu, J. Keane, C. Kim, "A 0.2V, 480kb Subthreshold SRAM With 1k Cells Per Bitline for Ultra-Low-Voltage Computing," IEEE J. of Solid-State Circuits, Volume 43, no. 2, pp. 518-529, Feb. 2008.

[22] T. Kim, J. Liu, C. Kim, "A Voltage Scalable 0.26 V, 64 kb 8T SRAM With V_{min} Lowering Techniques and Deep Sleep Mode," IEEE J. of Solid-State Circuits, Volume 44, no. 6, pp. 1785-1795, June 2009.

[23] T. Kim, R. Persaud, C. H. Kim, "Silicon Odometer: An On-Chip Reliability Monitor for Measuring Frequency Degradation of Digital Circuits," IEEE J. of Solid-State Circuits, Volume 43, no. 4, pp. 874-880, Apr. 2008.

[24] T. Kim, W. Zhang, C. H. Kim, "An SRAM Reliability Test Macro for Fully-Automated Statistical Measurements of V_{min} Degradation," IEEE Custom Integrated Circuits Conferences, Spt. 2009 (will appear).

[25] Bipul C. Paul, et al., "Device Optimization for Digital Sub-threshold Logic Operation," IEEE Transactions on Electron Devices", Volume 52, Issue 2, pp. 237-247, Feb. 2005.

[26] R. R. Troutman, "VLSI Limitations from drain-induced barrier lowering," IEEE Transactions on Electron Devices", Volume 26, Issue 4, pp. 461-469, Apr. 1979.

[27] C.Y. Lu, J. M. Sung, "Reverse short-channel effects on threshold voltage in submicrometer salicide devices", IEEE Electron Device Letters, Volume 10, Issue 10, pp. 446-448, Oct. 1989

[28] C. Subramanian, et al., "Reverse short channel effect and channel length dependence of boron penetration in PMOSFETs", International Electron Device Meeting, pp. 423-426, Dec. 1995.

[29] J. Keane, T. Kim, H. Eom, and C. Kim, "Sub-threshold logical effort: a systematic framework for optimal sub-threshold device sizing", Design Automation Conference, pp. 425-428, July 2006.

[30] Y. Taur, C. H. Wann, and D. J. Frank, "25nm CMOS Design Considerations", International Electron Devices Meeting, pp. 789-792, 1998

[31] M. Tohmason, J. Prasad, and J. De Greve, "Suppression of the reverse short channel effect in sub-micron CMOS devices", International Semiconductor Device Research Symposium, pp. 420-421, Dec. 2003.

[32] L. Chang, Y. Nakamura, R. K. Montoye, J. Sawada, et al., "A 5.3GHz 8T-SRAM with operation down to 0.41V in 65nm CMOS," in Proc. IEEE Symposium on VLSI Circuits, pp. 252-253, June 2007.

[33] L. Chang, D.M. Fried, J. Hergenrother, J.W. Sleight, et al., "Stable SRAM cell design for the 32 nm node and beyond," IEEE Symposium on VLSI Technology, pp. 128-129, June 2005.

[34] J. Chen, L. T. Clark, T. Chen, "An Ultra-Low-Power Memory With a Sub-threshold Power Supply Voltage," IEEE J. of Solid-State Circuits, Volume 41, pp. 2344-2353, Oct. 2006.

[35] M. Hwang, K. Roy, "A 135mV 0.13µW process tolerant 6T subthreshold DTMOS SRAM in 90nm technology," IEEE Custom Integrated Circuits Conference, pp. 419-422, Sept. 2008.

[36] M. Chang, W. Hwang, "A fully-differential subthreshold SRAM cell with auto-compensation," IEEE Asia Pacific Conference on Circuits and Systems, pp. 1771-1774, Dec. 2008.

[37] J. P. Kulkarni, K. Kim, K. Roy, "A 160mV Robust Schmitt Trigger Based Subthreshold SRAM," IEEE J. of Solid-State Circuits, Volume 42, no. 10, pp. 2303-2313, Oct. 2007.

[38] A. Raychodhury, S. Mukhopadhyay, K. Roy, "A feasible study of subthreshold SRAM across technology generations," IEEE International Conference on Computer Design, pp. 417-422, Oct. 2005.

[39] B. H. Calhoun, A. P. Chandrakasan, "Statis noise margin variation for sub-threshold SRAM in 65-nm CMOS," IEEE J. of Solid-State Circuits, Volume 41, no. 7, pp. 1673-1679, July 2006.

[40] I. Chang, J. Kim, S. Park, K. Roy, "A 32b 10T Subthreshold SRAM Array with Bit-Interleaving and Differential Read Scheme in 90nm CMOS," in Proc. IEEE International Solid-State Circuits Conference, pp. 388-389, Feb. 2008

[41] N. Verma, A. Chandrakasan, "A 256 kb 65 nm 8T Subthreshold SRAM Employing Sense-Amplifier Redundancy," IEEE J. of Solid-State Circuits, Volume 43, no. 1, pp. 141-149, Jan. 2008.

[42] B. Zhai, D. Blaauw, D. Sylvester, S. Hanson, "A Sub-200mV 6T SRAM in 0.13μm CMOS," in Proc. IEEE International Solid-State Circuits Conference, pp. 332-333, Feb. 2007.

[43] M. Yamaoka, N. Maeda, Y. Shinozaki, Y. Shimazaki, et al., "90-nm process-variation adaptive embedded SRAM modules with power-line-floating write technique," IEEE J. of Solid-State Circuits, Volume 41, no. 3, pp. 705-711, Mar. 2006

[44] R. Keyes, "The effect of randomness in the distribution of impurity atoms on FET threshold", Applied Physics A: Material Science Process., vol. 8, pp. 251–259, 1975.

[45] Y. Taur, T. Ning, *Fundamental of Modern VLSI Devices*, Cambridge University Press, 2002.

[46] J. Kim, K. Kim, C. Chuang, ”Back-gate controlled READ SRAM with improved stability,” IEEE International SOI Conference, pp. 211-212, Oct. 2005.

[47] M. Khellah, Y. Ye, N. Kim, D. Somasekhar, et al., “Wordline and bitline pulsing schemes for improving SRAM cell stability in low-Vcc 65 nm CMOS designs”, VLSI Circuits Symposium, pp. 9-10, June 2006.

[48] T. Kim, J. Liu, C. Kim, “An 8T Subthreshold SRAM Cell Utilizing Reverse Short Channel Effect for Write Margin and Read Performance Improvement,” in Proc. IEEE Custom Integrated Circuits Conference, pp. 241-244, Oct. 2007.

[49] K. Agawa, H. Hara, T. Dakayanagi, T. Kuroda, “A bitline leakage compensation scheme for low-voltage SRAMs,” IEEE J. of Solid-State Circuits, Volume 36, no. 5, pp. 726-734, May 2001

[50] K. Zhang, U. Bhattacharya, Z. Chen, F. Hamzaoglu, et al., “SRAM design on 65-nm CMOS technology with dynamic sleep transistor for leakage reduction,” IEEE J. of Solid-State Circuits, Volume 40, no. 4, pp. 895-901, Apr. 2005

[51] Y. Wang, H. Ahn, U. Bhattacharya, Z. Chen, et al., “A 1.1 GHz 12/Mb-Leakage SRAM Design in 65nm Ultra-Low-Power CMOS Technology With Integrated Leakage Reduction for Mobile Applications,” IEEE J. of Solid-State Circuits, Volume 43, no. 1, pp. 172-179, Jan. 2008

[52] H. Lee, P. Mok, "Switching Noise and Shoot-Through Current Reduction Techniques for Switched-Capacitor Voltage Doubler," IEEE J. of Solid-State Circuits, Volume 40, no. 5, pp. 1136-1146, May 2005

[53] J. Keane, T. Kim, C. H. Kim, "An On-Chip NBTI Sensor for Measuring PMOS Threshold Voltage Degradation," IEEE International Symposium on Low Power Electronics and Design, August 2007.

[54] S. Aota, S. Fujii, Z. W. Jin, et al., "A New Method for Precise Evaluation of Dynamic Recovery of Negative Bias Temperature Instability," IEEE International Conference on Microelectronic Test Structures, pp. 197-199, April 2005.

[55] C. Schlunder, W. Heinrigs, W. Gustin, et al., "On the Impact of the NBTI Recovery Phenomenon on Lifetime Prediction of Modern p-MOSFETs," IEEE International Integrated Reliability Workshop, pp. 1-4, October 2006.

[56] S. Zafar, "Statistical Mechanics Based Model for Negative Bias Temperature Instability," Journal of Applied Physics, Vol. 97, Issue 10, pp. 1-9, 2005.

[57] V. Reddy, A. Krishnan, A. Marshal, et al., "Impact of negative bias temperature instability on digital circuit reliability," IEEE International Reliability Physics Symposium, pp. 248-254, 2002.

[58] S. Rangan, N. Mielke, E. Yeh, "Universal recovery behavior of negative bias temperature instability," IEEE International Electron Devices Meeting, pp. 14.3.1-14.3.4, 2003.

[59] T. Yang, M. F. Li, C. Shen, et al., "Fast and Slow Dynamic NBTI components in p-MOSFET with SiON Dielectric and their Impact on Device Life-time and

Circuit Application," IEEE Symposium on VLSI Technology, pp. 92-93, June 2005.

[60] G. Chen, M. F. Li, C. H. Ang, et al., "Dynamic NBTI of p-MOS Transistors and its Impact on Device Lifetime," IEEE Electron Device Letters, pp. 734-736, December 2002.

[61] M. Denais, A. Bravaix, V. Huard, et al., "On-the-fly characterization of NBTI in ultra-thin gate oxide PMOSFET's," IEEE International Electron Devices Meeting, pp. 109-112, December 2004.

[62] M. Denais, A. Bravaix, V. Huard, et al., "Paradigm Shift for NBTI Characterization in Ultra-Scaled CMOS Technologies," IEEE International Reliability Physics Symposium, pp. 735-736, March 2006.

[63] R. Fernández, B. Kaczer, A. Nackaerts, et al., "AC NBTI Studied in the 1 Hz – 2 GHz Range on Dedicated On-Chip Circuits," IEEE International Electron Devices Meeting, pp. 337-340, December 2006.

[64] K.O. Jeppson, C. M. Svensson, "Negative bias of MOS devices at high electric fields and degradation of MNOS devices," Journal of Applied Physics, Vol. 48, Issue 5, pp. 2004-2014, 1977.

[65] S. Zafar, Y. H. Kim, V. Narayanan, et al., "A Comparative Study of NBTI and PBTI (Charge Trapping) in SiO2/HfO2 Stacks with FUSI, TiN, Re Gates," IEEE Symposium on VLSI Technology, pp. 23-25, June 2006.

[66] W. Wu, T. Chao, T. Chiu, et al., "Positive Bias Temperature Instability (PBTI) Characteristics of Contact-Etch-Stop-Layer-Induced Local-Tensile-Strained

HfO2 nMOSFET," IEEE Electron Device Letters, Vol. 29, no. 12, pp. 1340-1343, Dec. 2008.

[67] A. Kerber, K. Maitra, A. Majumdar, et al., "Characterization of Fast Relaxation During BTI Stress in Conventional and Advanced CMOS Devices With HfO2/TiN Gate Stacks," IEEE Transactions on Electron Devices, Vol. 55, no. 11, pp. 3175-3183, Nov. 2008.

[68] M. Khare, "High-K/Metal Gate Technology: A New Horizon," IEEE Custom Integrated Circuits Conference, pp. 417-420, Sept. 2007.

[69] B. C. Paul, K. Kang, H. Kufluoglu, et al., "Impact of NBTI on the temporal performance degradation of digital circuits," IEEE Electron Device Letters, Vol. 26, no. 8, pp. 560-562, Aug. 2005.

[70] S. V. Kumar, C. H. Kim, S. S. Sapatnekar, "NBTI-Aware Synthesis of Digital Circuits," IEEE Design Automation Conference, pp. 370-375, June 2007.

[71] M. Ketchen, M. Bhushan, R. Bolam, "Ring Oscillator Based Test Structure for NBTI Analysis," IEEE International Conference on Microelectronic Test Structures, pp. 42-47, Mar. 2007.

[72] J. J. Kim, R. Rao, S. Mukhopadhyay, C. T. Chuang, "Ring oscillator circuit structures for measurement of isolated NBTI/PBTI effects," IEEE International Conference on Integrated Circuit Design and Technology Tutorial, pp. 163-166, June 2008.

[73] A. Haggag, G. Anderson, S. Parihar, D. Burnett, et al., "Understanding SRAM High-Temperature-Operating-Life NBTI: Statistics and Permanent vs

Recoverable Damage," in Proc. IEEE International Reliability Physics Symposium, pp. 452-456, Apr. 2007.

[74] R. Kapre, K. Shakeri, H. Puchner, J. Tandigan, et al., "SRAM Variability and Supply Voltage Scaling Challenges," in Proc. IEEE International Reliability Physics Symposium, pp. 23-28, Apr. 2007.

[75] X. Li, J. Qin, B. Huang, X. Zhang, J. B. Bernstein, "SRAM circuit-failure modeling and reliability simulation with SPICE," IEEE Trans. on Device and Materials Reliability, Volume 6, no. 2, pp. 235-246, June 2006.

[76] A. Carlson, "Mechanism of Increase in SRAM V_{min} Due to Negative-Bias Temperature Instability," IEEE Trans. on Device and Materials Reliability, Volume 7, no. 7, pp. 473-478, Sept. 2007.

[77] A. T. Krishnan, V. Reddy, D. Aldrich, J. Raval, et al., "SRAM Cell Static Noise Margin and VMIN Sensitivity to Transistor Degradation," in Proc. IEEE International Electron Devices Meeting, pp. 1-4, Dec. 2006.

[78] G. L. Rosa, W. L. Ng, S. Rauch, R. Wong, J. Sudijono, "Impact of NBTI Induced Statistical Variation to SRAM Cell Stability," in Proc. IEEE International Reliability Physics Symposium, pp. 274-282, Mar. 2006.

[79] C. Wann, R. Wong, D. Frank, R. Mann, et al., "SRAM cell design for stability methodology," in Proc. IEEE VLSI-TSA International Symposium on VLSI Technology, pp. 21-22, Apr. 2005.

[80] T. Fischer, E. Amirante, K. Hofmann, M. Ostermayr, et al., "A 65nm test structure for the analysis of NBTI induced statistical variation in SRAM

transistors," in Proc. IEEE European Solid-State Device Research Conference, pp. 51-54, Sept. 2008.

[81] H. Aono, E. Murakami, K. Shiga, F. Fujita, et al., "A study of SRAM NBTI by OTF measurement," in Proc. IEEE International Reliability Physics Symposium, pp. 67-71, Apr. 2008.

[82] K. Kang, K. Kim, A. E. Islam, M. A. Alam, K. Roy, "Characterization and Estimation of Circuit Reliability Degradation under NBTI using On-Line IDDQ Measurement," in Proc. IEEE Design Automation Conference, pp. 358-363, June 2007.

[83] J. C. Lin, A. S. Oates, C. H. Yu, "Time Dependent Vccmin Degradation of SRAM Fabricated with High-k Gate Dielectrics," in Proc. IEEE Reliability Physics Symposium, pp. 439-444, Apr. 2007.

[84] S. V. Kumar, K. H. Kim, S. S. Sapatnekar, "Impact of NBTI on SRAM read stability and design for reliability," in Proc. IEEE International Symposium on Quality Electronic Design, pp. 27-29, Mar. 2006.

[85] M. Ball, J. Rosal, R. McKee, Wk. Loh, et al., "A Screening Methodology for VMIN Drift in SRAM Arrays with Application to Sub-65nm Nodes," in Proc. IEEE International Electron Devices Meeting, pp. 1-4, Dec. 2006.

[86] M. Agostinelli, J. Hicks, J. Xu, B. Woolery, et al., "Erratic fluctuations of sram cache vmin at the 90nm process technology node," in Proc. IEEE International Electron Devices Meeting, pp. 655-658, Dec. 2005.

[87] A. Haggag, M. Moosa, N. Liu, D. Burnett, et al., "Realistic Projections of Product Fails from NBTI and TDDB," in Proc. IEEE International Reliability Physics Symposium, pp. 541-544, Mar. 2006.

[88] V. Huard, C. Parthasarathy, C. Guerin, T. Valentin, et al., "NBTI Degradation: From Transistors to SRAM Arrays," in Proc. IEEE International Reliability Physics Symposium, pp. 289-300, Apr. 2008.