國立交通大學

電子工程學系 電子研究所碩士班

碩 士 論 文

利用動態基體偏壓與電源閘技術之低功率設計

Dynamic Body-Biasing and Power-Gating Techniques for Low Power Design

研 究 生：鄭東栓

指導教授：黃 威 教授

中 華 民 國 九 十 三 年 六 月

利用動態基體偏壓與電源閘技術之低功率設計
Dynamic Body-Biasing and Power-Gating Techniques for Low Power Design

研究生：鄭東栓　　Student：Tung-Shuan Cheng

指導教授：黃　威　　Advisor：Wei Hwang

國立交通大學
電子工程學系電子研究所
碩士論文

A Thesis
Submitted to Department of Electronics Engineering & Institute of Electronics
College of Electrical Engineering and Computer Science
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
Master

利用動態基體偏壓與電源閘技術之低功率設計

學生：鄭東栓　　　　　　　　　　　　　　　　　　指導教授：黃 威

國立交通大學電子工程學系電子研究所碩士班

摘　　要

本論文使用動態基體偏壓與電源閘技術來實現低功率之電路設計。基於系統晶片之使用彈性與重複使用特性，一個可產生多種輸出電壓的基體電壓產生器被提出且用 TSMC 100nm CMOS 技術設計。此電路可經由改變輸入信號設定來得到不同的輸出電壓。另外，一個可產生雙電壓的基體電壓產生器被用在靜態隨機存取記憶體陣列的設計，藉此觀察基體偏壓對於漏電流抑制的有效性。電路模擬和佈局是用 TSMC 0.13um CMOS 技術實現。模擬結果顯示 64 字元的記憶體單元減少 75% 的淨功率消耗，32 字元則是 64%。

一個利用行解碼器與列解碼器來控制電源閘的靜態隨機存取記憶體陣列被提出，且利用 TSMC 0.13um CMOS 技術來實現電路設計與佈局。同一條字線上的字元被分成數個區塊，每一個區塊擁有各自的電源閘控制元件。模擬結果顯示可以減少大量的靜態和動態的功率消耗，而且功率-延遲乘積說明對於速度變慢的影響極小。使用 8 字元的區塊，佈局面積將增加 20.7%，而使用 16 字元區塊會增加 12.1%的面積。若使用 32 字元區塊，則面積增加 8.1%。此技術可以應用在靜態隨機存取記憶體，暫存器，內容定址記憶體，動態隨機存取記憶體，快閃記憶體，快取記憶體，或是其他類似之記憶體與邏輯電路。

Dynamic Body-Biasing and Power-Gating Techniques for Low Power Design

student：Tung-Shuan Cheng Advisors：Dr. Wei Hwang

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

ABSTRACT

The low-power circuit designs using dynamic body-biasing and power-gating techniques are realized in this thesis. For the flexibility and reusability in System-on-Chip designs, an on-chip configurable body-bias generator that produces various voltage levels is proposed and simulated in TSMC 100nm technology. The output voltage can be controlled through digital input signals. A dual-level on-chip body-bias generator is presented and combined with SRAM cell arrays to observe the effectiveness in leakage suppression. Simulation results in TSMC 0.13um technology show that 75% and 64% net cell leakage reductions are achieved for 64-bit and 32-bit wordlines, respectively. The physical layout is implemented in TSMC 0.13um technology and triple-well structure is necessary for separating body nodes of transistors.

A column/row co-controlled SRAM cell arrays scheme is also proposed and simulations and layout are implemented in TSMC 0.13um technology. The cells on the same wordline are divided into blocks and each block has a dedicated gating device. The gating devices are controlled by signals from both column and row decoders. Simulation results show a great amount of active and standby power saving and power-delay product demonstrates that the induced performance overhead is insignificant. Moreover, the area overheads for 8-bit block and 16-bit block conditions are 20.7% and 12.1%, respectively, and only 8.1% is for 32-bit block condition. This technique can be applied to SRAM, register file, CAM, DRAM, flash memory, cache, or other similar memory and logic circuits.

Acknowledgements

Many people have contributed to the success of this thesis. First, I would like to thank my advisor, Prof. Wei Hwang, for his wise advices and suggestions. With his support, I have the confidence and energy to stride forward. Next, I would also like to thank Hong-Ren Liao, manager of Taiwan Semiconductor Manufacturing Company (TSMC), who gave me the chance to work at TSMC and inspired me to accomplish many interesting ideals. Finally, I would like to thank all the graduate students at LPSOC Lab., and they helped me a lot in both school courses and my thesis research.

Moreover, I would also like to thank my family and my girlfriend for their support. Especially my girlfriend, Lans, who always encourages me when I fall disappointed. Her understanding, devotion, and patience for the past years are vital to my life and education during this period. Thank you all.

Contents

List of Figures

List of Tables

Chapter 1
Introduction

In recent years, portable devices and wireless applications are getting popular, such as cell phones, PDAs, and portable computers. It's emergent to increase battery life and make chips consume as less energy as possible. For the future integrated-circuit (IC) and System-on-Chip (SoC) designs, high-performance operation and low power consumption must be achieved at the same time. Not only the increase of battery life, low-power designs diminish the risks of hot spots and thermal runaway problems. The giga-Hertz operating frequency dramatically increases the temperature and thus degrades the performance. Moreover, the unbalanced temperature distribution across the whole chip causes performance and power fluctuations [1.1], [1.2].

As the technology scales down to deep-submicron and nano-scale eras, both supply voltage (V_{DD}) and threshold voltage (Vt) are reduced for high-performance and low-power designs. The dynamic power is well known the dominant power component in digital circuits, and the reduction of supply voltage reaches a significant percentage of total power saving. However, the standby leakage cannot be ignored in deep-submicron and nano-scale technologies. In deep-submicron technologies, subthreshold leakage is the critical component among all the leakage currents. Subthreshold leakage increases due to the reduction of threshold voltage with the scaling of technology. In order to compensate the performance degradation of descending supply voltage, threshold voltage is scaled down to satisfy speed requirement. Therefore, the influence of leakage power is becoming significant if threshold voltage keeps on scaling down. One more leakage source called gate-tunneling leakage (or gate leakage) is becoming important in nano-scale technologies. The increase of gate-tunneling leakage is mainly due to the scaling of thickness of gate oxide. Many predictions show that gate leakage has the potential to exceed subthreshold leakage and dominate the standby leakage current in the future [1.3], [1.4]. More seriously, the total standby leakage current may even exceed the dynamic power and dominate the total power consumption. Therefore, this thesis focuses on the techniques for power reduction and leakage currents suppression and applies to SRAM cell array design.

The principles of low-power designs and leakage currents suppression are described in Chapter 2. The detail characteristics and scaling trend of leakage currents in future technologies are discussed. The impacts of parameter variations on performance and power distribution are also mentioned in this chapter. Body biasing is a popular technique to control and eliminate the influences on performance and power fluctuations. Forward body-bias is applied to low-performance circuits since threshold voltage is reduced. On the other hand, reversed body-bias that raises threshold voltage can be applied to leaky circuits for leakage suppression. Some circuit-level and system-level leakage control techniques will be discussed in the rest of Chapter 2, including the role of power management unit (PMU) and the concepts of Voltage Island techniques.

On-chip voltage generators and configurable body-bias generator design are discussed in Chapter 3. In VLSI and SoC designs, various voltage levels are required for distinct functional blocks. This chapter focuses on the designs of body-bias generators that can reversely bias the substrate of transistors. Charge pumps are popular circuits for generating voltages below GND for NMOS and voltages beyond V_{DD} for PMOS. Moreover, a configurable scheme that can produce various voltage levels according to control signals is proposed.

In Chapter 4, a dual-level body-biasing generator is proposed and applied to the design of SRAM cell arrays. Many similar designs have been realized before and will be compared in this chapter. However, almost all of these prior designs used external voltage sources instead of on-chip voltage generators. Therefore, the design of SRAM cell arrays adopting on-chip body-bias generators is implemented in this chapter. Finally, a time-out-policy controller for body-bias generator is also presented.

The design of SRAM cell arrays using power-gating technique is realized in Chapter 5. Power gating is an effective technique to suppress leakage current in standby mode by isolating power lines and idle circuits. Attention must be paid to power-gated SRAM cell array designs since data must be retained. Some prior power-gated SRAM architectures are introduced in this chapter. Moreover, a novel architecture that achieves both active and standby power savings is proposed and realized. This architecture induces performance and area overheads, but the power-delay product comparison demonstrates that these overheads are insignificant. Finally, the overall investigation results will be presented in Chapter 6.

Chapter 2

Overview of Low Power Design and Leakage Control Techniques

In modern digital CMOS integrated circuits, power consumption can be classified into three different components: dynamic, short circuit, and leakage power. Dynamic power is the dominant component of power consumption and results from the charging and discharging of capacitances. Short circuit currents occur when both NMOS and PMOS devices are ON during switching transients. The third component is leakage power. Leakage power is small in comparison with dynamic power, but it's becoming important in the future deep-submicron and nano-scale technologies and it has the potential to exceed dynamic power [2.1].

In Sec. 2.1, the three components of power consumption will be described briefly. Next, the impacts on performance and power distribution due to process, supply voltage, and temperature variations are discussed in Sec. 2.2. The effects of applying body-bias are described in Sec. 2.2 as well. Besides, some threshold-controlled techniques and novel circuits are presented in Sec. 2.3. Finally, the concepts of Voltage Islands and system-level power control issues are described in Sec. 2.4.

2.1 Power Sources in Digital CMOS Circuits

Dynamic, short circuit, and leakage power are the three major components of digital CMOS integrated circuits.

2.1.1 Dynamic Power

Among the three components of power sources, dynamic power is the dominant component and results from the charging and discharging of capacitances. Dynamic power is given by

$$P_{dynamic} = C_{switched} V_{DD}{}^{2} f_{clk} \tag{2.1}$$

where $C_{switched}$ is the total effective switched capacitance, V_{DD} is the supply voltage, and f_{clk} is the switching frequency. It's obvious that to scale down V_{DD} is the most efficient way to reduce dynamic power. However, V_{DD} cannot be scaled down forever since the operating frequency degrades with low supply voltage.

2.1.2 Short-Circuit Power

The second component of power consumption is short-circuit power, which results from non-zero rise and fall times of the input waveforms. The non-zero input rise and fall times cause a direct path between V_{DD} and GND for a short time period during switching. Short-circuit power can be expressed as

$$P_{short-circuit} = t_{sc} V_{DD} I_{peak} f_{clk} = C_{sc} V_{DD}{}^{2} f_{clk} \qquad (2.2)$$

where t_{sc} represents the time that the direct path is conducting. Short-circuit power is proportional to the switching activity, as similar to the dynamic power.

2.1.3 Leakage Power

The third component is leakage power. Leakage current can be divided into many sources, such as subthreshold, band-to-band tunneling, gate tunneling, pn junction reverse bias, DIBL, GIDL, and punchthrough leakage. Fig. 2.1 illustrates these leakage sources in a MOS device.

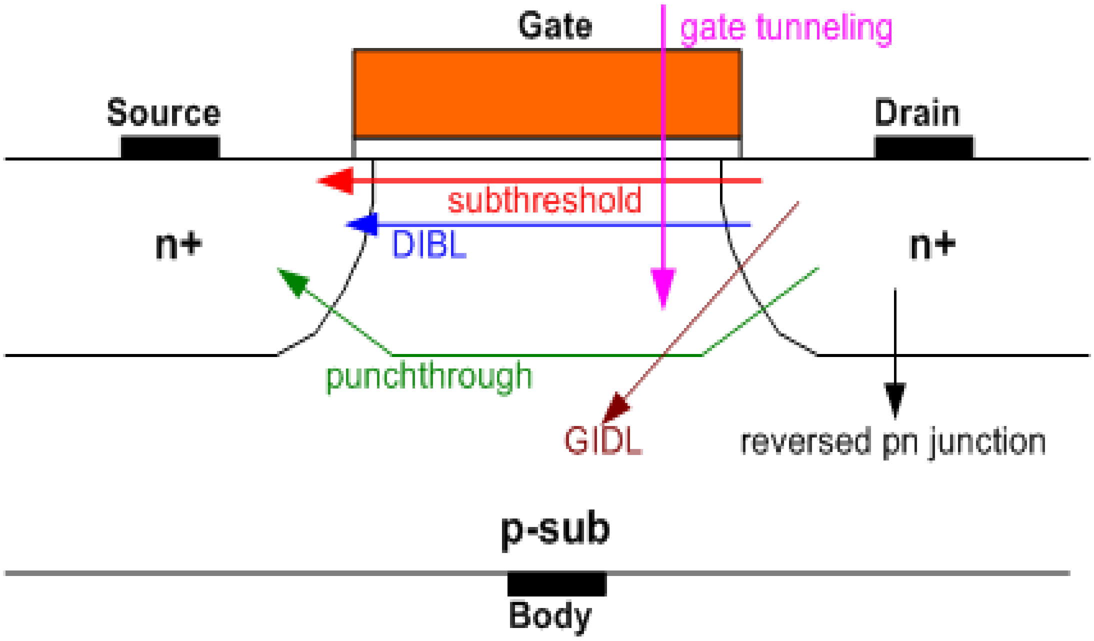


Fig. 2.1 Leakage sources in a MOS device.

Among the various leakage sources, subthreshold leakage is the dominant component that is given as

$$I_{leakage} = I_0 \exp(\frac{V_G - Vs - V_{T0} - \gamma Vs + \eta V_{DS}}{nV_{thermal}}) \bullet (1 - \exp\frac{-V_{DS}}{V_{thermal}}) \qquad (2.3)$$

where V$_{\mathrm{thermal}}$ is the thermal voltage, n is the subthreshold swing coefficient constant, γ is the linearized body effect coefficient, and η is the DIBL coefficient. Assuming that V$_{\mathrm{DS}}$ >> V$_{\mathrm{thermal}}$ and (2.2) can be simplified to the expression

$$I_{leakage} = I_0 \cdot 10^{(VGS - Vt)/nV_{thermal} \ln 10} \qquad (2.4)$$

Eq. (2.4) implies that subthreshold leakage is smaller with higher threshold voltage, Vt, and this component is becoming important since threshold voltage is scaled down with the progress of CMOS technology.

Subthreshold leakage is becoming the dominant component of power consumption in deep-submicron technologies. However, as the process steps into the region of nano-scale technologies, gate leakage has the potential to dominate the leakage current or even exceed the level of dynamic power. Gate leakage is given as

$$J_{DT} = AE_{ox}^{\;2} \exp\{-\frac{B[1 - (1 - \frac{V_{ox}}{\phi_{ox}})^{3/2}]}{E_{ox}}\} \qquad (2.5)$$

$$A = \frac{q^3}{16\pi^2\hbar\phi_{ox}}, \; B = \frac{4\sqrt{2m^*}\phi_{ox}^{\;3/2}}{3\hbar q} \qquad (2.6)$$

where V_{ox} is the voltage drop across the oxide, ϕ_{ox} is the barrier height in the conduction band, and E_{ox} is the field across the oxide. In detail, the gate leakage is composed of I$_{\mathrm{gd}}$, the gate leakage between gate and drain, I$_{\mathrm{gb}}$ between gate and body, and I$_{\mathrm{gs}}$ between gate and source. Fig. 2.2 predicts that gate leakage is indispensable in nano-scale technologies and the amount of gate leakage is far beyond the standby power constraint. Gate leakage becomes critical due to the decrease of thickness of gate oxide.

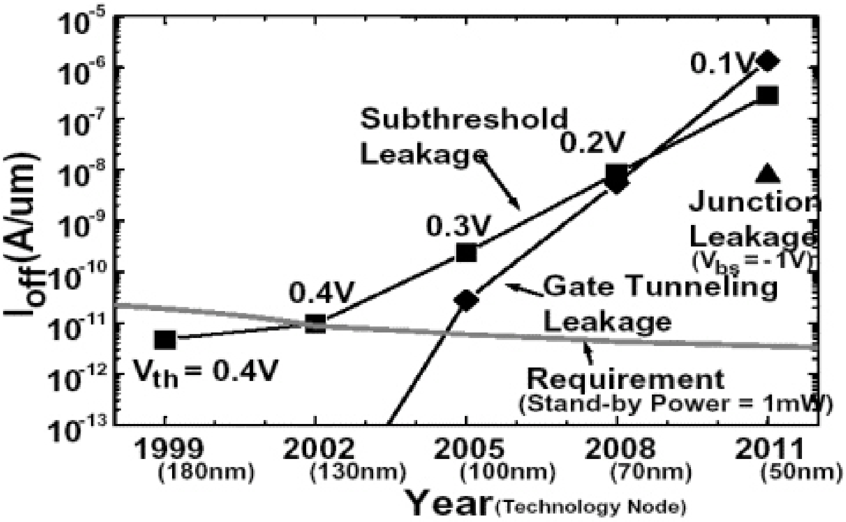


Fig. 2.2 The trend of standby current of MOSFETs.

Fig. 2.3 plots the I_{ON}/I_{OFF} ratio and Vt for both low and high Vt NMOS transistors for 130nm, 100nm, and 70nm technologies [2.2], where I_{ON} means driving current in active mode and I_{OFF} stands for leakage current in standby mode. Fig. 2.3 reveals that the leakage current is increasing by 3-5x per generation and becoming comparable with active driving current. The extra leakage current wastes a significant amount of power and causes thermal hot spots and thermal run away problems.

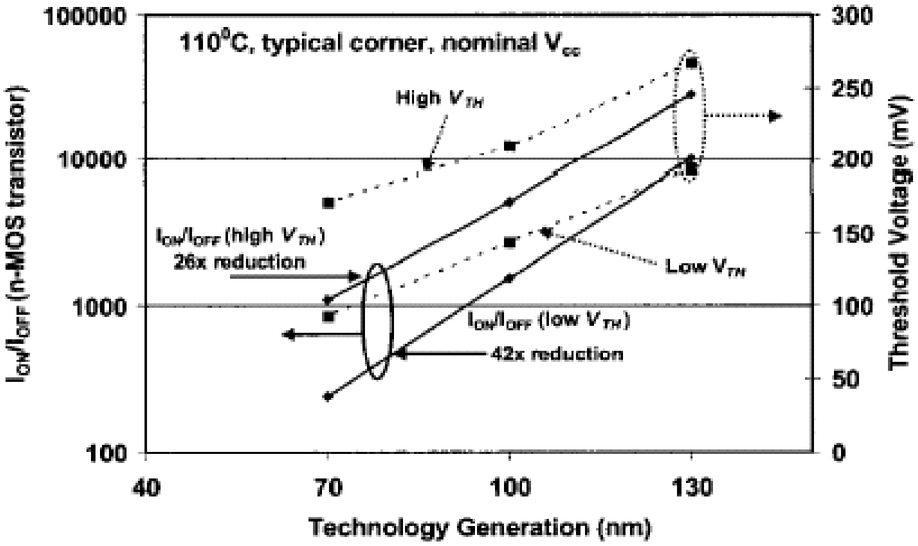


Fig. 2.3 I_{ON}/I_{OFF} and Vt scaling for sub-130nm generations.

2.2 Parameter Variations

As CMOS technology scales down to deep-submicron region, parameter variations in process, supply voltage, and temperature (P, V, T) are becoming a major challenge in designing future high-performance processors. In the past, CMOS technology variations are mainly due to imperfect process control. However, in present and future devices, intrinsic atomistic variations are becoming very important and cause uncertainty in I-V curves, in timing, and in power dissipation.

Process variations impact the frequency and leakage contribution of chips, causing die-to-die and within-die performance and power fluctuations. In addition, the demand for low power and low supply voltage making voltage variation a significant influence. Above-mentioned variations make some dies on a single wafer cannot achieve the target frequency, while some others fail to satisfy the leakage power constraint.

2.2.1 Process Variations

Fig. 2.3 plots the frequency and leakage distributions of dies on a signal wafer [2.3]. Due to both die-to-die and within-die parameter variations, significant variation exists in frequency and leakage power. At the same time, accepted dies must meet the frequency and leakage constraints. Notice that most of the ultra-high speed dies consume too large leakage power and they must be discarded. The ultra-low speed dies that have reasonably high leakage must be discarded as well since they cannot achieve the performance requirement.

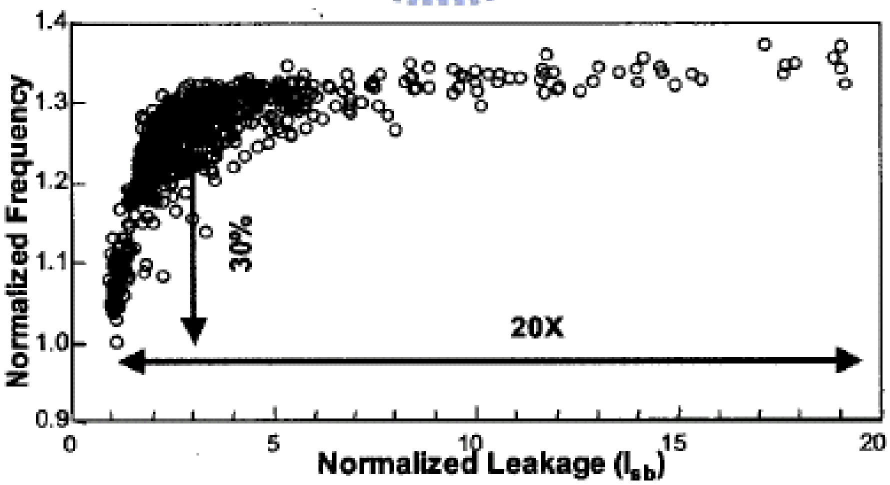


Fig. 2.4 Fluctuation of frequency and leakage for dies on a single wafer.

The wide leakage or standby current distribution comes from channel length and threshold voltage (Vt) variations, as illustrated in Fig. 2.2. The fluctuation of Vt among the dies results in wide spread of leakage current.

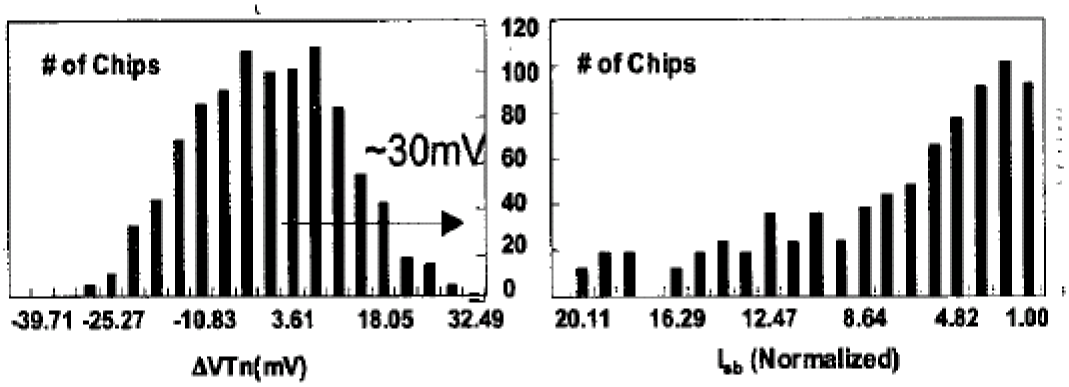


Fig. 2.5 Die-to-die Vt and standby leakage variations.

2.2.2 Supply Voltage Variations

Differences of switching activity and logic circuits across the die cause uneven power dissipation in the die. Thus, uneven supply voltage distribution and temperature hot spots occur and lead to variation of subthreshold leakage across the die. The scaling of supply voltage due to technology progress degrades this effect since the impact on supply voltage is relatively larger.

2.2.3 Temperature Variations

As described previously, differences of switching activity and types of logic across the die cause uneven power dissipation in the die. Therefore, different thermal distributions appear in the distinct parts of the die and result in variations of leakage. The measurement results in [2.4] demonstrate that the standby leakage current increases with the increase of temperature. Not only the increase of standby leakage current, a high temperature further degrades the performance of devices.

2.2.4 Applying Body-Bias for Reducing Parameter Variations

The leakage and frequency on a single die can be controlled through body bias. The leakage current can be significantly reduced by applying reversed body-bias (RBB) due to the increase of threshold voltage. On the other hand, by applying forward body-bias (FBB) the threshold voltage is lowered down and thus the speed is improved.

2.2.4.1 Reversed Body-Bias for Reducing Leakage

Fig. 2.6 shows that the leakage current decreases with applied reversed boy-bias due to the increase of threshold voltage [2.5]. RBB can be applied to the dies on a single wafer that are too leaky for suppressing leakage current. Besides, RBB can be applied to circuit blocks on a single die that are too leaky for compensating within-die leakage fluctuation. However, the performance degradation due to RBB must be taken into account.

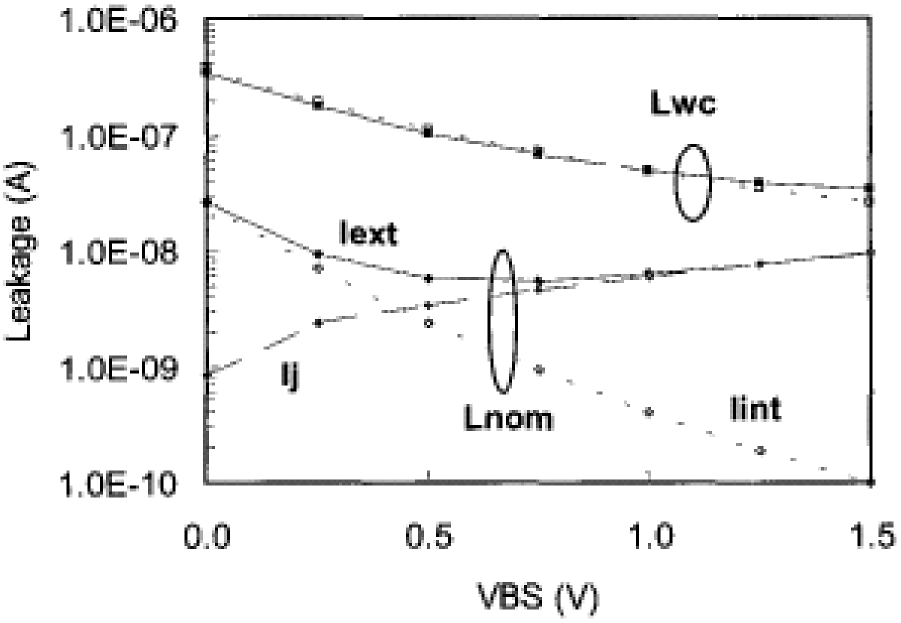


Fig. 2.6 Die-to-die Vt and standby leakage variations.

2.2.4.2 Forward Body-Bias for Improving Performance

By applying forward body-bias the operating speed is improved due to the decrease of threshold voltage, as shown in Fig. 2.7 [2.6]. Since the threshold voltage is decreased, thus the active driving current increases to speedup the operation. Therefore, FBB is beneficial to dies on a single wafer or circuit blocks on a single die that fail to achieve required performance. However, the induced extra leakage current is an issue that one must pay attention to it.

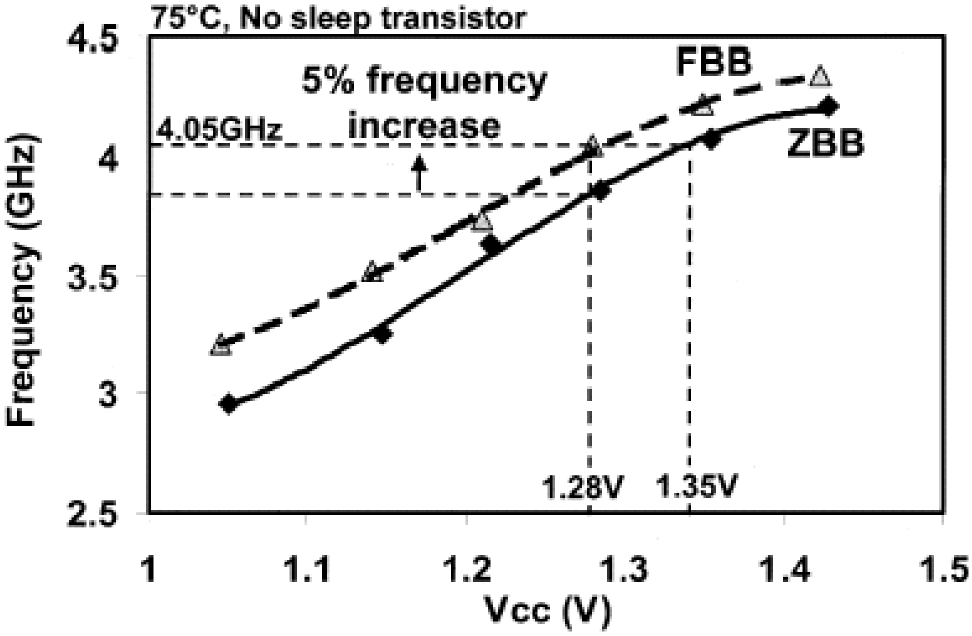


Fig. 2.7 Operating speed increases with forward body bias.

2.2.5 Effectiveness and Optimum Value of Reversed Body-Bias

From the previous description it has been seen that applying RBB is an effective and widely used technique to reduce leakage current. Unfortunately, the effectiveness degrades in advanced technologies. [2.5] shows that the intrinsic leakage current increases with the decrease of channel length. In addition, the effectiveness of applying RBB at nominal transistor channel lengths (Lnom) is better than shorter channel lengths condition (Lwc). Because of worsening short channel effect (SCE), effectiveness of RBB diminishes with technology scaling. This means that to keep SCE under control becomes more important as the technology scales down. Another reasons is the growing gate leakage, which is immune to RBB.

Many researches and measurements have shown that an optimum RBB value exists which is different from different technologies [2.7], [2.8], [2.9]. Biasing in the optimum RBB condition a least leakage power is consumed, and leakage power increases when the applied RBB exceeds the optimum value. One reason is that the band-to-band tunneling leakage increases due to RBB [2.10]. Fig. 2.8 illustrates the characteristics of leakage sources with RBB [2.2] and shows that GIDL increases with RBB. Obviously, an optimum RBB value exists and a least leakage power is achieved in this condition.

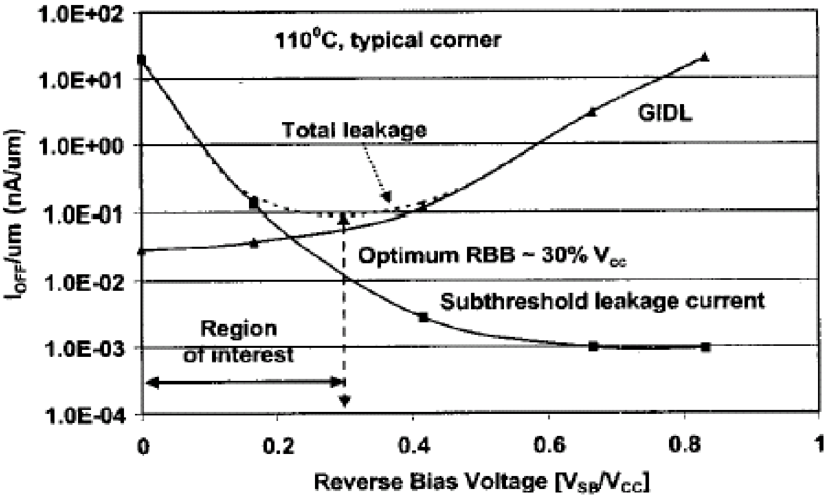


Fig. 2.8 Characteristics of leakage sources with reversed body-bias.

2.3 Design Techniques for Controlling Threshold Voltage

The most popular technique for controlling threshold voltage is applying body bias. Threshold voltage is raised with RBB, while threshold voltage is lowered with FBB. In this section, many designs and techniques for controlling threshold voltage are introduced, including RBB and FBB control methodologies. Beside, one more matching circuit for balancing threshold voltage of NMOS and PMOS is also described.

2.3.1 Adaptive Variable Threshold Voltage Techniques

Here some circuit techniques for dynamically varying threshold voltage are introduced. The threshold voltage is adaptively adjusted in real time according to operating speed or leakage current.

2.3.1.1 Self-Adjusting Threshold-Voltage Scheme

A circuit technique called Self-Adjusting Threshold-Voltage Scheme (SATS) is to reduce the Vt fluctuation by using self-substrate-biasing [2.11]. The SATS comprises a leakage sensor and a self-substrate-bias circuit (SSB). The leakage sensor detects the leakage current of the design and controls SSB.

Fig. 2.9 shows the block diagram of the SATS to reduce the Vt fluctuation. A leakage sensor detects leakage current of a representative MOS transistor and sends a control signal, V_{cont}, to SSB. V_{cont} triggers SSB only when the leakage is higher than a predetermined level. That is, Vt is set to the lowest value that satisfies the power specification. However, the tradeoff between leakage current and operating speed must be considered carefully.

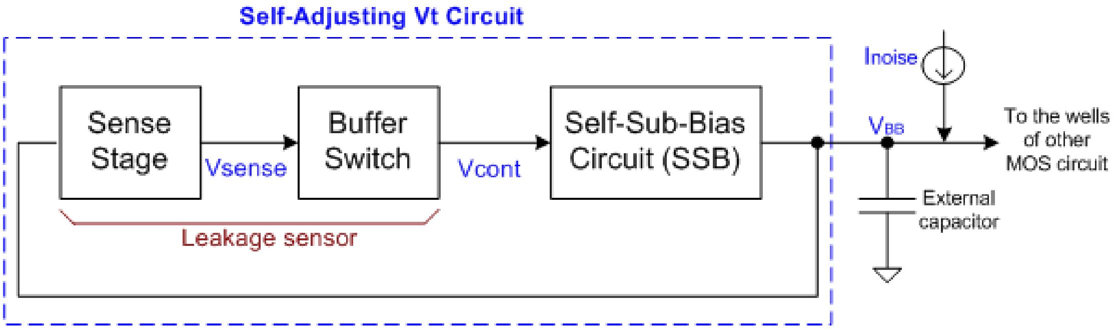

Self-Adjusting Vt Circuit
Sense Stage
Vsense
Buffer Switch
Vcont
Self-Sub-Bias Circuit (SSB)
Leakage sensor
Inoise
VBB
To the wells of other MOS circuit
External capacitor

Fig. 2.9 Block diagram of Self-Adjusting Vt Scheme (SATS).

Fig. 2.10 shows a leakage sensor for NMOS transistor. The size of N1 should be large enough to detect the leakage current that flows in N1. The leakage current is amplified by the load so that the load circuit must be sensitive enough to amplify the small leakage current. V_G is generated by dividing the supply voltage and is set to a small value that is necessary to enhance the leakage current. If the leakage current of N1, I_{leak}, is getting large and V_{sense} is low enough to set V_{cont} to high, the SSB is triggered and applies proper body bias to N1.

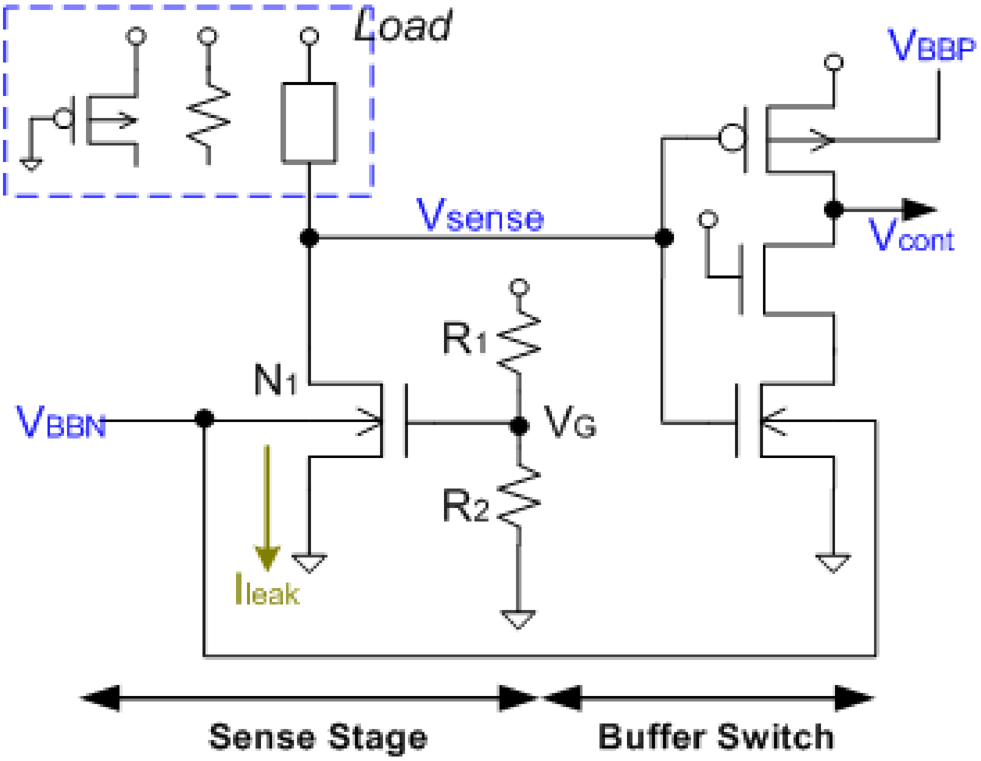


Fig. 2.10 Leakage sensor for detecting leakage current of a NMOS transistor.

2.3.1.2 Variable Threshold-Voltage Scheme

Fig. 2.11 illustrates the Variable Threshold-voltage (VT) scheme [2.12] and the threshold voltage of a transistor is controlled through a Variable Threshold-voltage circuit (VT circuit). The VT circuit controls V_{BB} to compensate the Vt fluctuation in the active mode. In the standby mode, the VT circuit applies deeper RBB to increase Vt and thus reduce leakage current.

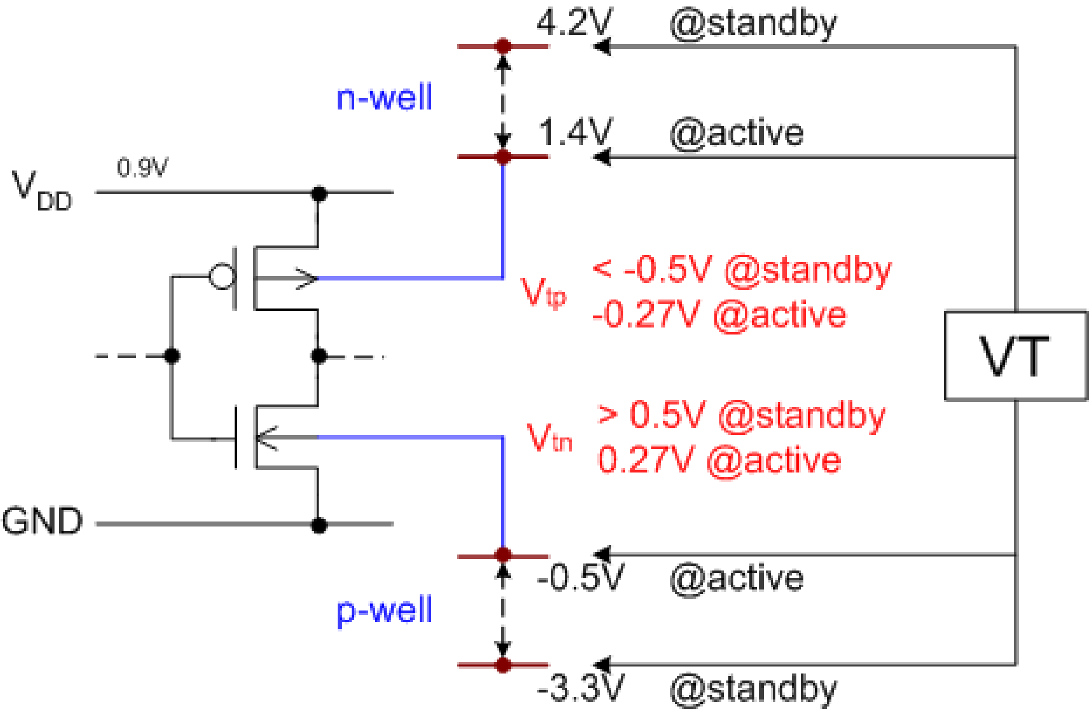


Fig. 2.11 Variable threshold-voltage (VT) scheme.

As depicted in Fig. 2.12, the VT circuit consists of four leakage current monitors (LCM's), a self-substrate bias circuit (SSB), and a substrate charge injector (SCI). The SSB is used to lower the voltage of V_{BB}, while the SCI injects charges into the substrate to raise V_{BB}. The monitors that detect the level of V_{BB} control the operations of both the SSB and SCI.

Fig. 2.13 shows the schematic of LCM and the connections between SSB and DCT core. The LCM detects the leakage current of DCT core, $I_{leak,\ DCT}$, through the transistor M4 that shares the same substrate with the DCT. If $I_{leak,\ LCM}$ is larger than a predetermined level, the node N_1 goes low and force N_{out} to go high to activate SSB. Consequently, the V_{BB} becomes deeper and $I_{leak,\ LCM}$ and $I_{leak,\ DCT}$ become smaller. It's undoubtedly that the SSB is disabled when $I_{leak,\ LCM}$ is still under the predetermined level.

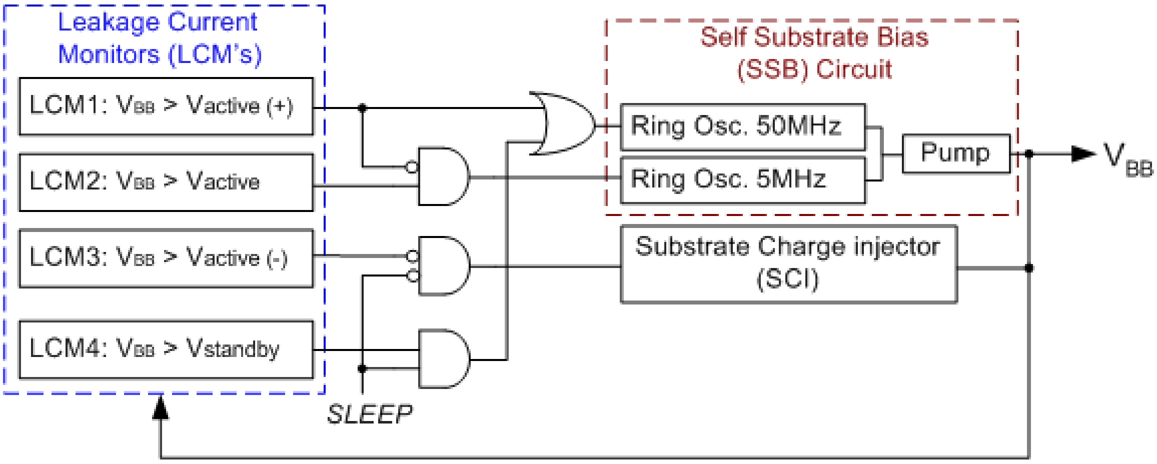


Fig. 2.12 Block diagram of VT circuit.

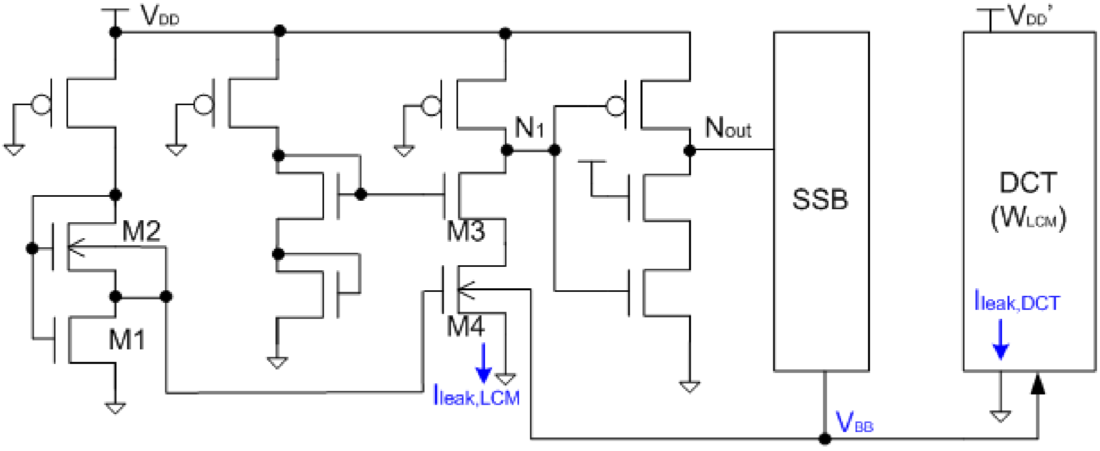


Fig. 2.13 Schematic of leakage current monitor.

2.3.1.3 Speed-Adaptive Threshold-Voltage Scheme

A speed adaptive threshold-voltage (SA-Vt) scheme is illustrated in Fig. 2.14 [2.13]. It consists of a delay line, a delay comparator, a decoder, and V_{BB} generators. The propagation delay of delay line is controlled by varying the substrate voltage of delay line. If the speed of the delay line changes, the comparator detects and recognizes whether the speed is slower or faster. Then, the comparator sends signals to the decoder, and the decoder controls the V_{BB} generators so that proper V_{BN} and V_{BP} are produced.

The delay line is simply an inverter chain, and the comparator can be implemented with a phase detector. Basically, the concept of SA-Vt scheme is like phase-locked loop (PLL) or delay-locked loop (DLL), but the difference is that SA-Vt scheme dynamically varies the substrate bias of delay line.

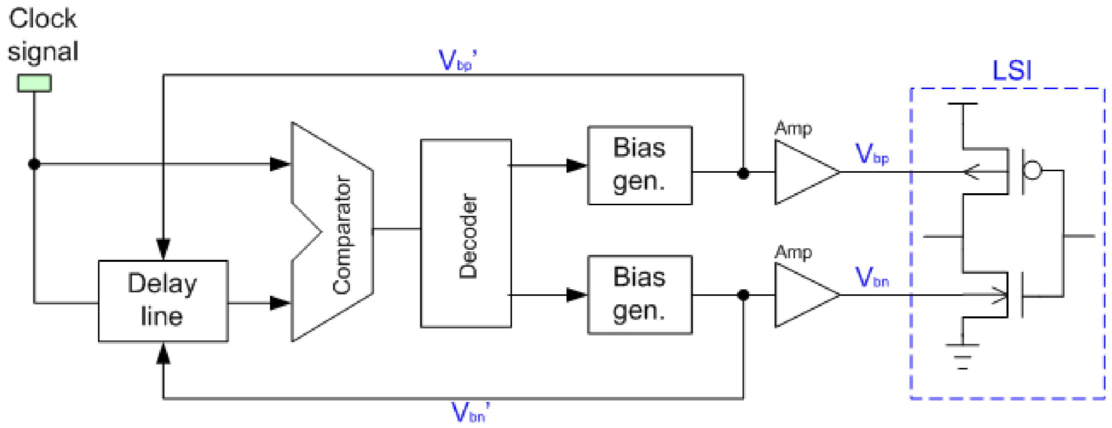


Fig. 2.14 Concept of speed adaptive Vt scheme.

2.3.1.4 Software-Controlled Vt-Hopping Scheme

Fig. 2.15 shows the Vt-hopping scheme that the substrate bias is dynamically controlled by software [2.14]. The power control block generates select signals, Vt_low_enable, and Vt_high_enable, according to the control signal, CONT. Note that CONT comes from the processor. The threshold voltage of the processor is higher if Vt_high_enable is asserted, while the threshold voltage is lower if Vt_low_enable is asserted.

Signal CONT is also used to control the operating frequency of the processor. When Vt_low_enable is high, the frequency controller feeds f_{clk} to the processor. On the other hand, when Vt_high_enable is asserted, the frequency controller generates f_{clk} /2 to the processor. Besides, more than two sets of frequency and threshold voltage can be extended if necessary.

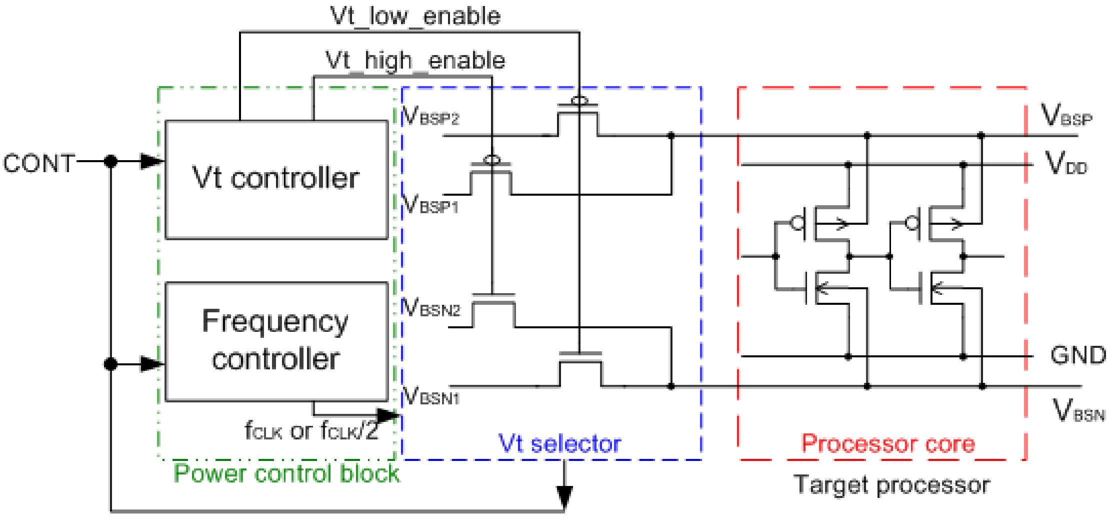


Fig. 2.15 Schematic diagram of Vt-hopping scheme.

The required maximum operating frequency determines the value of the lower threshold voltage. The processor must achieve f_{clk} with the lower threshold voltage, while f_{clk} /2 must be achieved with the higher threshold voltage.

2.3.1.5 Automatic Supply Voltage and Body Bias Scheme

An adaptive supply voltage and body bias (ASB) scheme has been developed to exploit the optimum operating condition that least active power is consumed [2.15]. Fig. 2.16 illustrates the schematic diagram of ASB, which is based on DLL and critical path replica. The ASB scheme uses a DLL to adjust the body bias values until the speed of the critical path replica equals the target frequency exactly. The critical path replica must closely track the performance of the actual critical path of a design as supply and body voltages vary.

The authors intend to exploit the optimum V_{DD}-Vt operating point by using ASB circuit. At first, they found out the minimum supply voltage to achieve a specific frequency without body bias applied. Next, lowering the supply voltage and applying proper amount of FBB concurrently to maintain the operating frequency. There are many possible V_{DD}-Vt operating points that can achieve the same performance. The authors discovered that there is a definite operating point that minimizes the total power consumption.

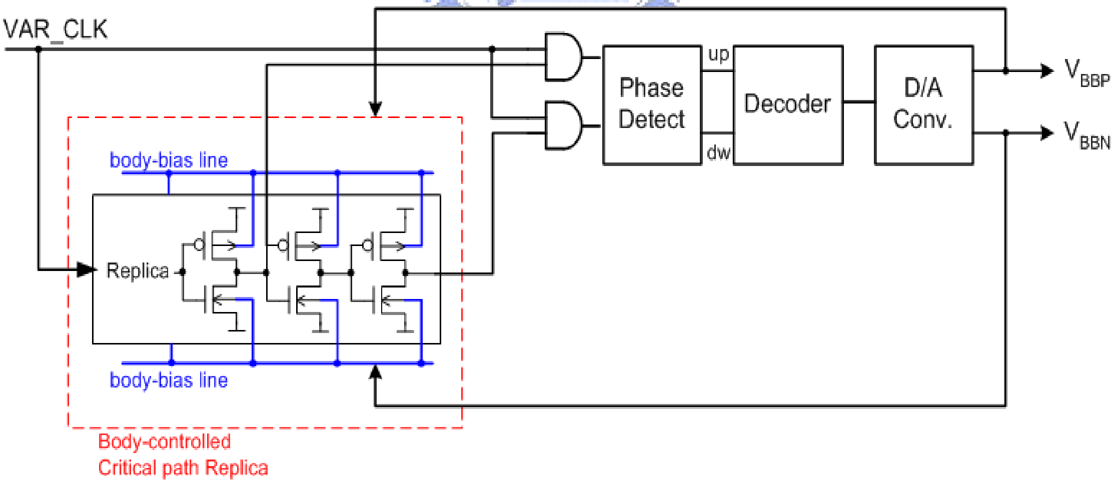


Fig. 2.16 Auto body bias generator based on DLL and critical path replica.

2.3.1.6 Digital Control in Adaptive Body Bias

A bidirectional adaptive body bias (ABB) generator has been developed to apply RBB for low leakage and FBB for high performance, as in Fig. 2.17 [2.16]. The authors used ABB to reduce the impacts of die-to-die and within-die parameter variations. That is, proper FBB is applied to the dies that are below the frequency constraint, and proper RBB is applied to the dies that consume too much leakage power.

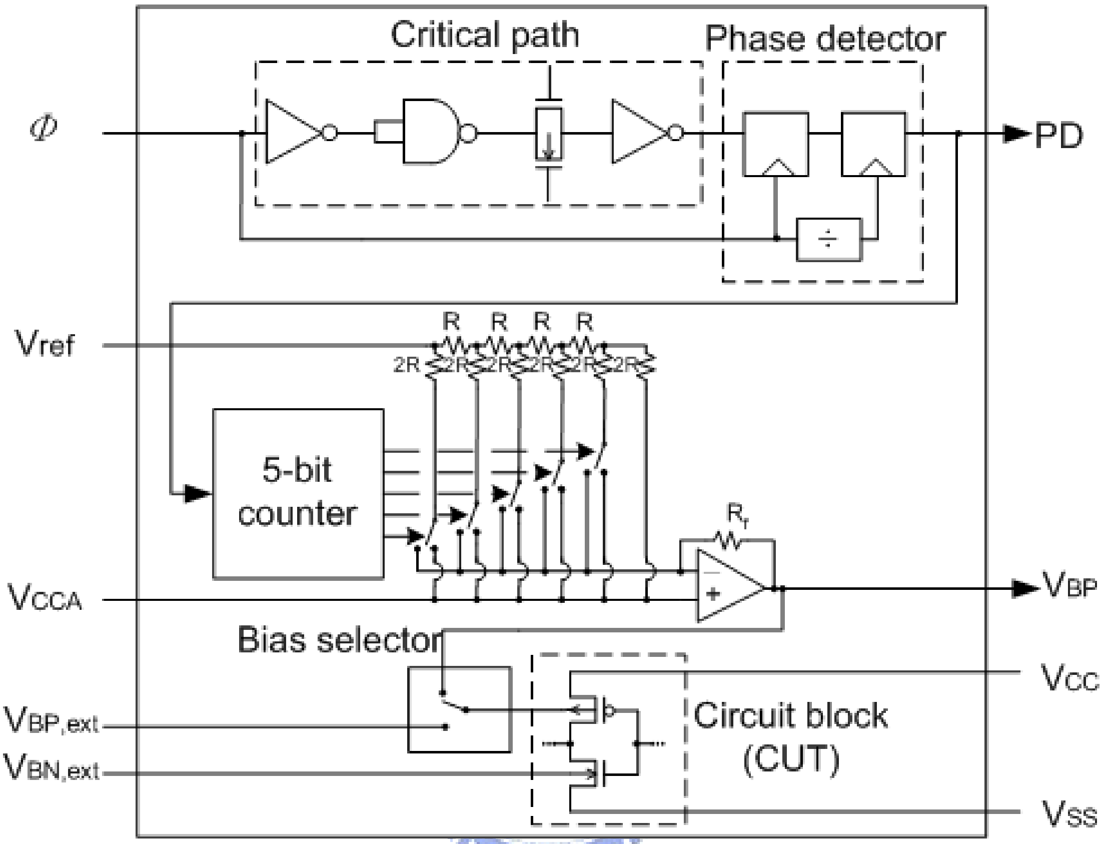


Fig. 2.17 Schematic diagram of adaptive body bias generator and the target design.

ABB circuit comprises a critical path replica, a phase detector, a 5-bit counter, and a resistor network with an amplifier. The critical path replica can be constructed by a ring-oscillator structure for measuring the operating frequency. The phase detector compares the critical path delay with the target clock period and sends control signals to the counter. The counter receives signals from the phase detector and controls the resistor network. The resistor network and the amplifier can be seen as a D/A converter that produces body bias to the substrate of PMOS. Therefore, the body bias voltage can be configured digitally through the counter.

2.3.2 Dynamic Threshold Voltage Techniques

In contrast to the adaptive variable threshold-voltage techniques described previously, dynamic threshold-voltage techniques here without sensors and adjust threshold voltage dynamically according to mode control signals. In addition, their

FBB or RBB values are fixed and one bias level is for one operating mode.

2.3.2.1 Standby Power Reduction Circuit

Fig. 2.18 depicts a standby power reduction (SPR) scheme that switches the voltages of n-well and p-well in different operating modes [2.17]. The SPR circuit consists of a level shifter and a voltage switch.

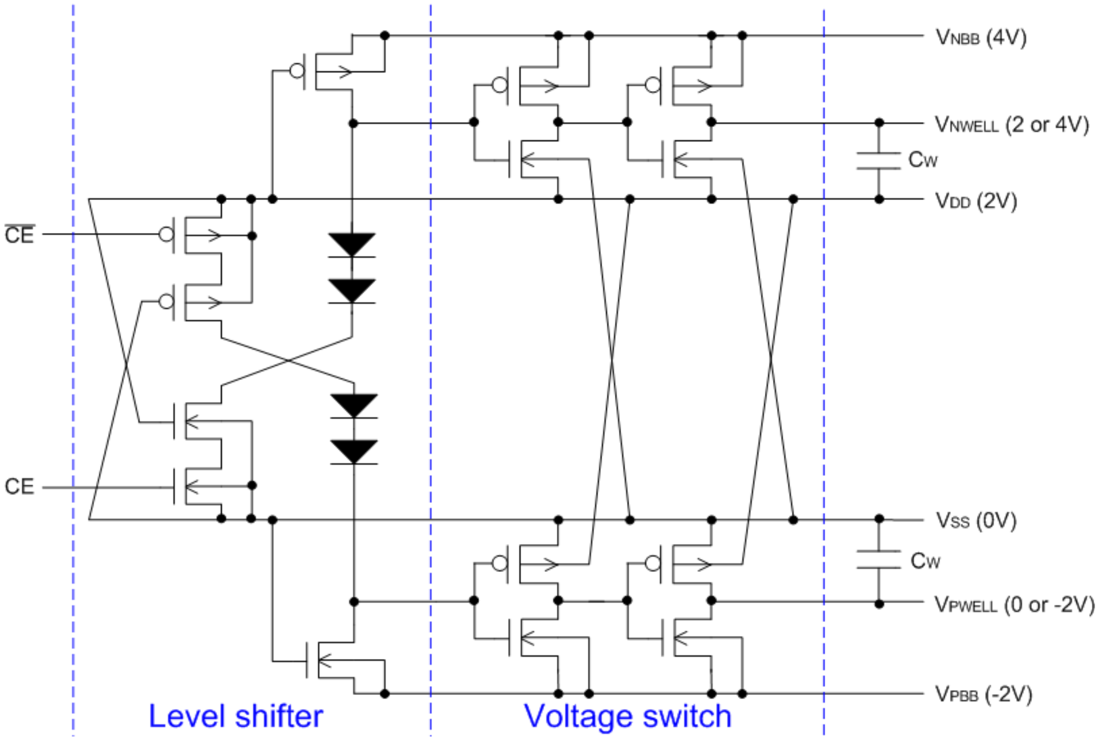


Fig. 2.18 Schematic diagram of SPR circuit.

In the active mode CE is asserted, and V_{PWELL} and V_{NWELL} are switched to 0V and 2V, respectively. In the active mode zero body bias (ZBB) is applied and high performance is maintained. When CE is pulled low in standby mode, V_{PWELL} and V_{NWELL} are switched to -2V and 4V, respectively. In the standby mode RBB is applied and standby leakage is significantly reduced.

It’s remarkable that four external power sources are needed, V_{NBB}, V_{DD}, V_{SS}, and V_{PBB}. However, the operations of this scheme are questionable if V_{NBB} and V_{PBB} are generated by charge pump circuits. Assume that V_{NBB} and V_{PBB} come from charge pump circuits and V_{SS} and V_{DD} are external power sources, the voltage levels of V_{NBB} and V_{PBB} will be influenced during the switching activities. This is because of the weak driving capabilities of V_{NBB} and V_{PBB}.

2.3.2.2 Self-Adjusted Forward Body Bias

Fig. 2.19 shows a technique called self-adjusted forward body bias (SAFBB) that adopts current sources to bias the substrate nodes [2.11]. The purpose of this scheme is to achieve high performance with a low supply voltage and thus low active power.

The signals C_{bn} and C_{bp} control the body bias depending on the condition of the target design. C_{bn} is pulled low and C_{bp} is pulled high in the standby mode, thus ZBB is applied. When in the active mode, C_{bn} is pulled high and C_{bp} is pulled low and FBB is applied to the substrate nodes. The amount of applied FBB can be varied by carefully designing the current sources.

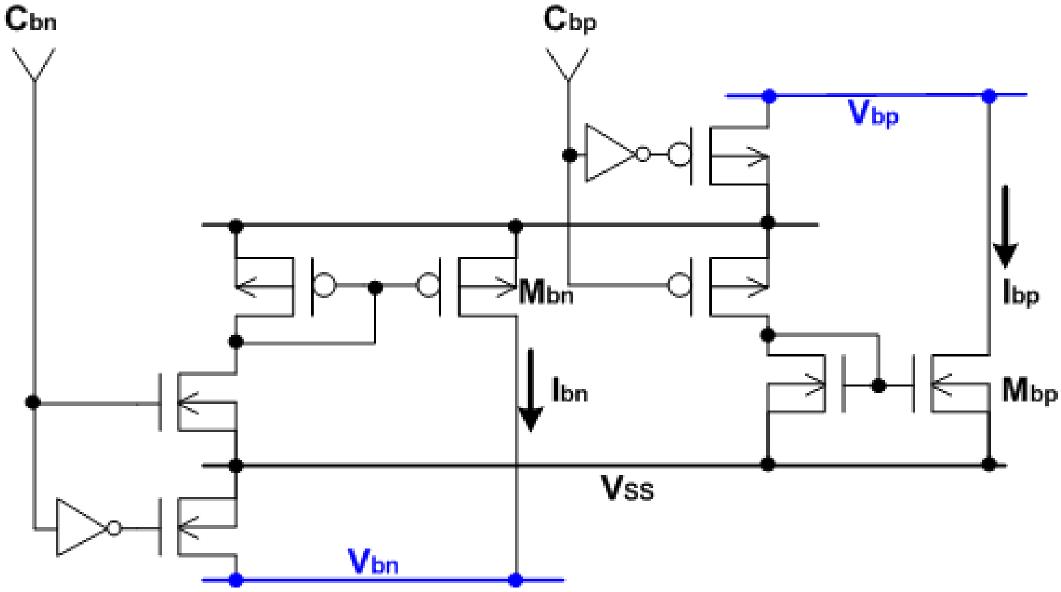


Fig. 2.19 Schematic diagram of self-adjusted FBB scheme.

2.3.2.3 Central and Local Body Bias Generators

A design that uses central and local body bias generators is depicted in Fig. 2.20 [2.18], which applies FBB to the PMOS transistors of the target core circuit. Note that only one central bias generator (CBG) is constructed, but many local bias generators (LBG) are necessary to drive many macros.

The CBG uses a scaled bandgap circuit to generate a process, voltage, and temperature-invariant reference voltage, and then this voltage is routed to all of the LBGs. The reference voltage is about 450mV below the bandgap supply V_{CCA}, and this means that the amount of FBB applied is about 450mV. The LBG is used to translate the reference voltage to a body bias voltage that is referenced to the local supply V_{CC}.

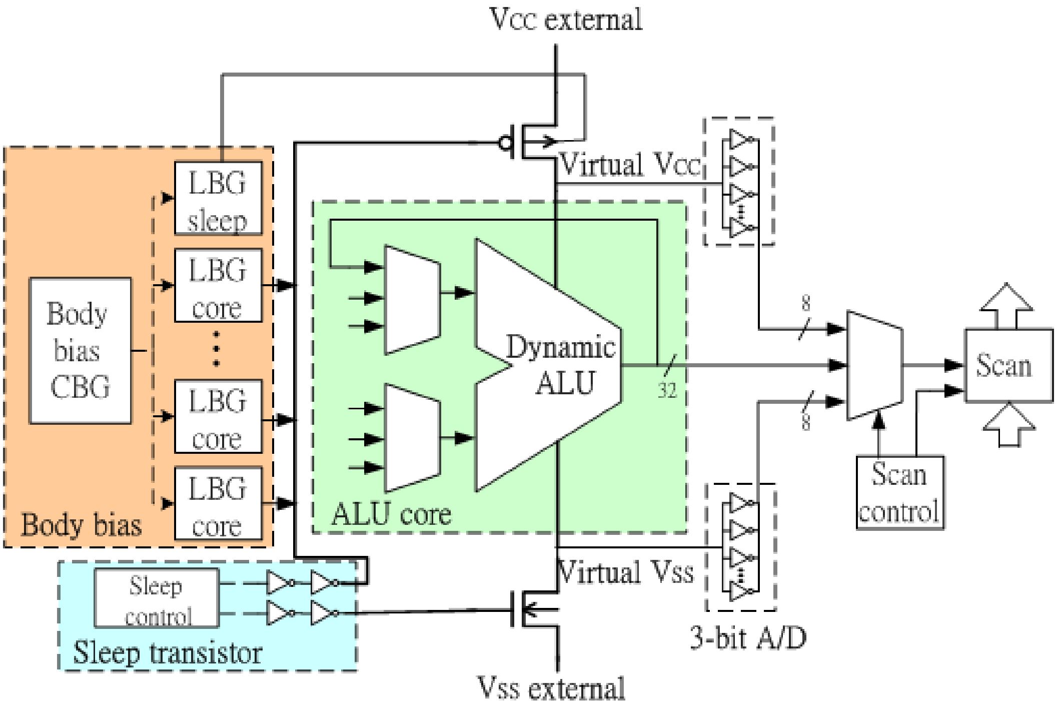


Fig. 2.20 The design that using central and local body bias generators to apply FBB.

Fig. 2.21 shows the schematic of CBG and LBG. The usage of separate CBG and LBG ensures that any variations in the local supply voltage V_{CC} will be tracked by the body voltage and thus a constant FBB of 450mV is maintained. Translation of the reference voltage is realized through a current mirror followed by a voltage driver, which is to drive the final n-well load. FBB is applied to the target design in the active mode, while ZBB is applied in the standby mode by switching on the zero-bias switch.

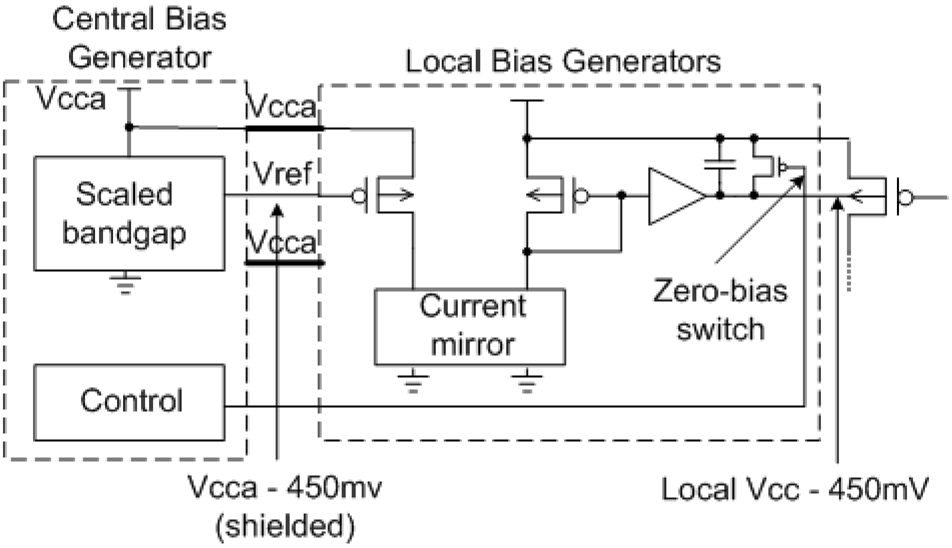


Fig. 2.21 Schematic diagram of CBG and LBG.

2.3.3 Threshold Voltage Matching Scheme

For low-power and low-voltage system, the fluctuation of performance is becoming a critical issue. The variations of process, supply voltage, and temperature are major sources of fluctuations of performance. As the technologies scale down, the supply voltage and threshold voltage are lower so that the variations of supply voltage and threshold voltage are becoming more serious. Moreover, the Vt difference between PMOS and NMOS degrades the performance and operating margin of chips. Since Vt is getting smaller in advanced technologies, a small variation of Vt can significantly impact the performance and reliability.

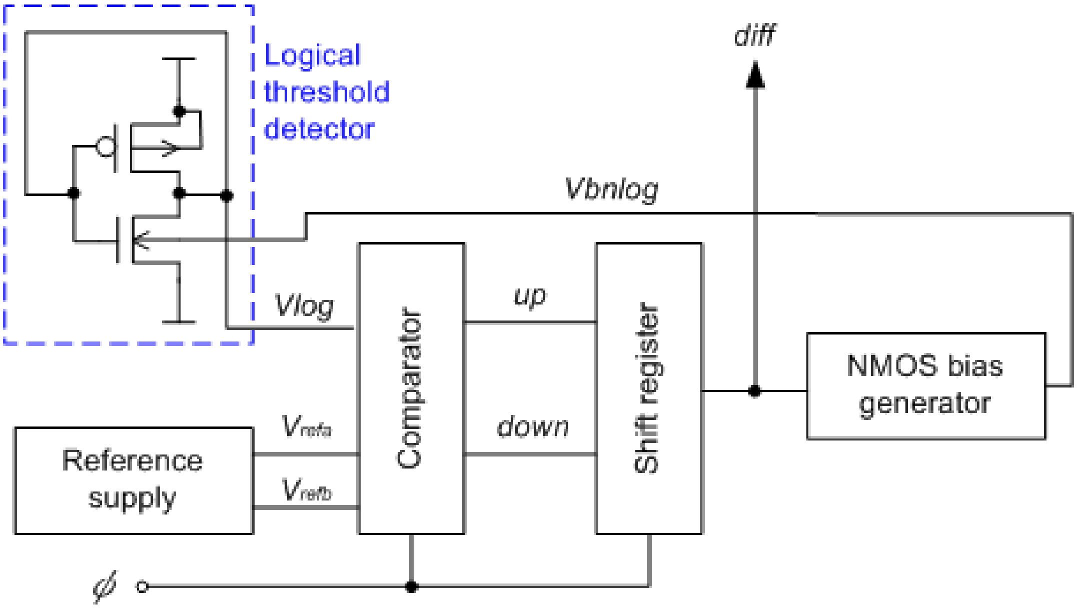


Fig. 2.22 PMOS/NMOS Vt matching scheme.

Fig. 2.22 shows a Vt matching scheme that consists of a logical threshold detector, a comparator, a shift register, a NMOS bias generator, and a reference supply [2.19]. The logical threshold detector is composed of a CMOS inverter chain that the input and the output is connected. The logical threshold detector is used to detect Vt unbalance. The comparator detects whether the V_{log} level is between V_{refa} and V_{refb} and outputs the diff signal to control the NMOS bias generator. The NMOS bias generator varies the bias voltage V_{bnlog} if the V_{log} level is above or below the region determined by V_{refa} and V_{refb}.

Fig. 2.23 shows the operating waveforms of Vt matching scheme. As first, the V_{log} is beyond the predetermined region due to Vt unbalance. The comparator detects the unbalance and shift register forces the NMOS bias generator to supply larger V_{bnlog}. The NMOS bias generator is stopped until the V_{log} is within the predetermined region.

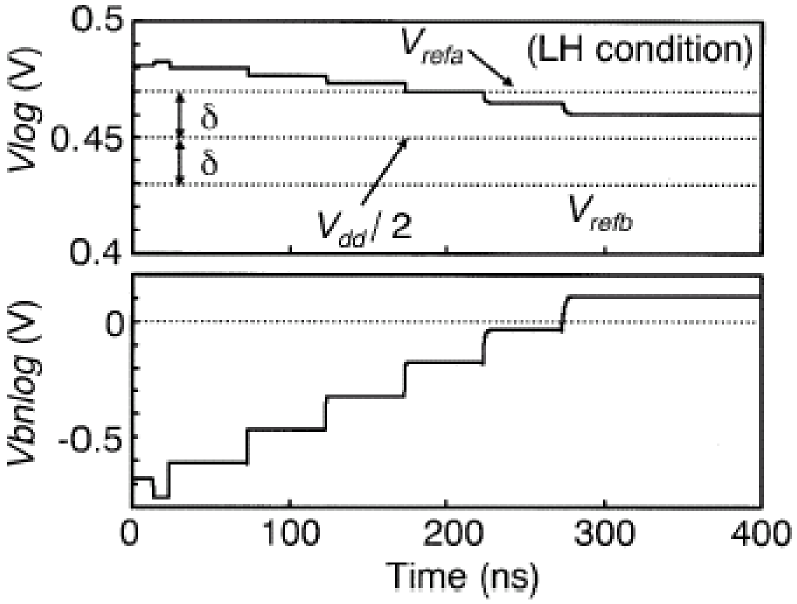


Fig. 2.23 Operating waveforms of Vt matching scheme.

2.4 System-Level Power Control

The previous sections described many circuit-level threshold-voltage control techniques. However, with the progress of semiconductor technology and the trend of System-on-Chip (SoC), a system-level power and performance control methodology is more efficient. In this section, a power management unit (PMU) and its functionalities are introduced. Besides, the concept of Voltage Islands and a power control technique using status table is described.

2.4.1 Overview of PMU

Power management is a real-time technique to dynamically monitor and control power distribution and performance of a chip. Fig. 2.24 shows the block diagram of a possible PMU solution. It's mainly consists of supply and body-bias voltage generators, device performance and thermal monitors, control logics, clock generators, and state machines. The voltage generators internally supply various voltages to serve as supply voltage or body bias; the monitors observe and detect the device performance and the temperature to keep the functionality and performance. Control logics are used to control the operation modes of functional blocks, and the power management state machine keeps tract of the mode transitions of functional blocks.

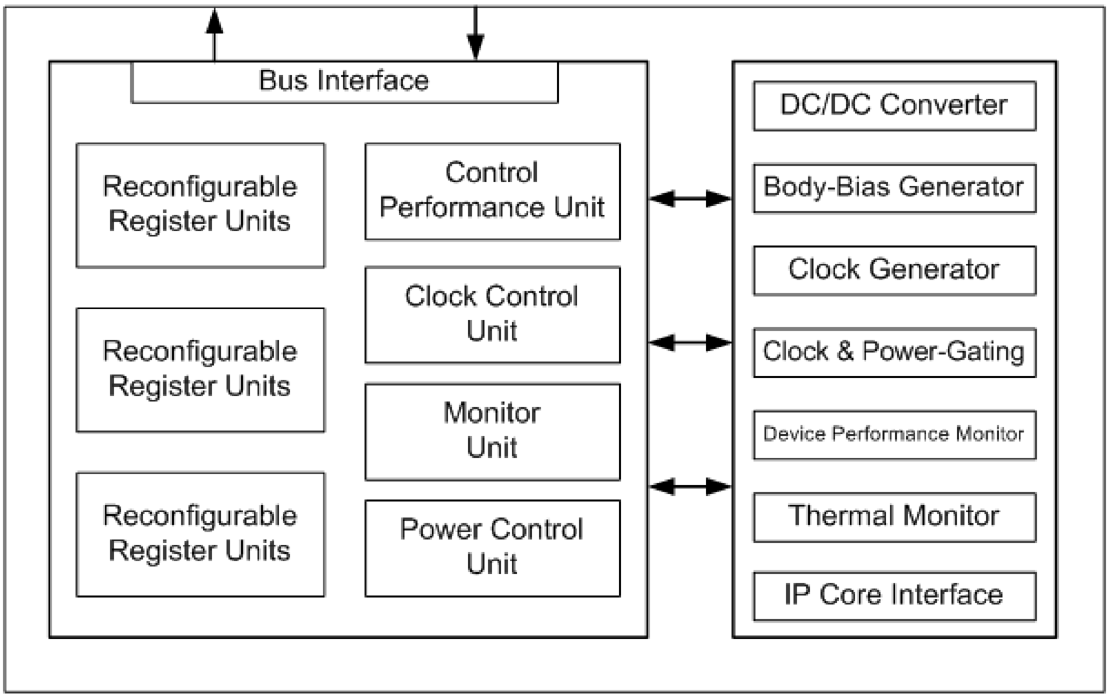


Fig. 2.24 A possible solution of PMU.

PMU monitors the power and performance conditions of all functional blocks, and executes operating mode transitions according to the activities. That is, if a functional block is in its high power and performance state but without any task to do, PMU sends control signals to change the power state. In addition, if the thermal detector observes that the temperature of a functional block goes too high, PMU enforces the IP to slow down the operating speed.

2.4.2 Concept of Voltage Islands

Voltage Islands are areas (logic and/or memory) on the same chip that are supplied by different voltage sources [2.20]. As discussed before, the various voltages may come from DC/DC converters. Voltage Islands restore the concept of individual voltage optimization of functional blocks to SoC design. Individual functional blocks of the SoC design can have different power characteristics from the rest of the design, and can be optimized accordingly. For example, the most performance-critical element of the design, such as a processor core, requires the highest voltage to maintain the required high performance. On the other hand, such as memory cells or control logics may not require this level of voltage. Therefore, significant power can be saved if they can run at lower voltages.

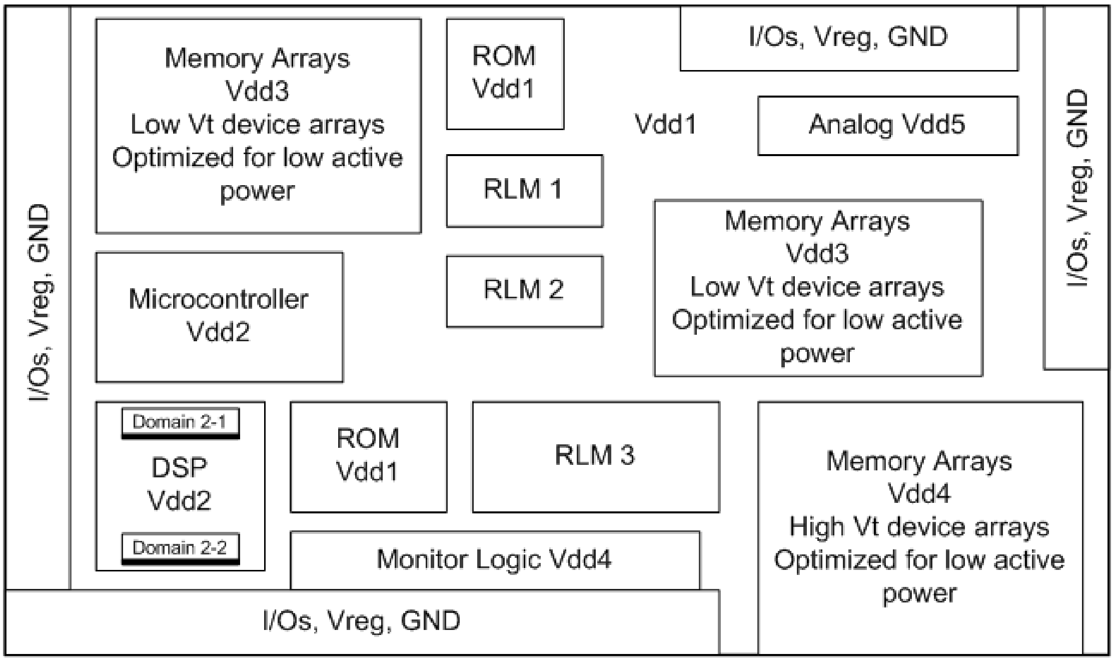


Fig. 2.25 Multiple on-chip Voltage Islands.

Fig. 2.25 shows the multiple on-chip Voltage Islands, which are operated under different supply voltages. In general, the circuits in the same Voltage Island have similar operation characteristics. As in Fig. 2.25, for instance, the Voltage Island 2 is a DSP (digital signal processing) processor and it can be fully shut down if we know that there are no DSP operations needed. An effective way is to add gating devices between the processor and the supply voltage or the ground.

Power domains are areas within an Island supplied by the same power supply but have distinct gating devices. According to the operating characteristics, part of an Island can be power gated but others are still power on.

2.4.3 Managing Threshold-Voltage Through a Status Table

A Vt management scheme using programmable status table is illustrated in Fig. 2.26 [2.21]. It consists of an instruction decoder, status monitor, programmable registers, and some logics. The decoder decodes an instructions and the requirement table identifies the function units that are required to execute this instruction. The status table stores the present power status of each function units. The logic observes whether the power requirement and power status are in agreement. If an instruction requires one function unit that is in lower power state, the execute logic sends control signals to change the power state to a higher level. A higher power state means both larger supply voltage and higher operating frequency. On the other hand, if another one function unit that is in higher power state is not required, the execute logic lowers the supply voltage and frequency or even shuts down the function unit.

The existence of override register allows direct control of power and speed by application software. In other words, the override register can be programmed to directly control the power and speed of function units regardless of the internal managing scheme. For example, a reset signal from software can initialize the tables and put all the function units to the lowest power state.

The power latency table, as the broken-lined square in Fig. 2.26, is included but it does not appear in [2.21]. The power latency table contains the information of time periods required for power state transitions. The time latency of power state transitions must be taken into account since input data should be stalled until the required power state is ready. A false operation results from the execution under incorrect power state.

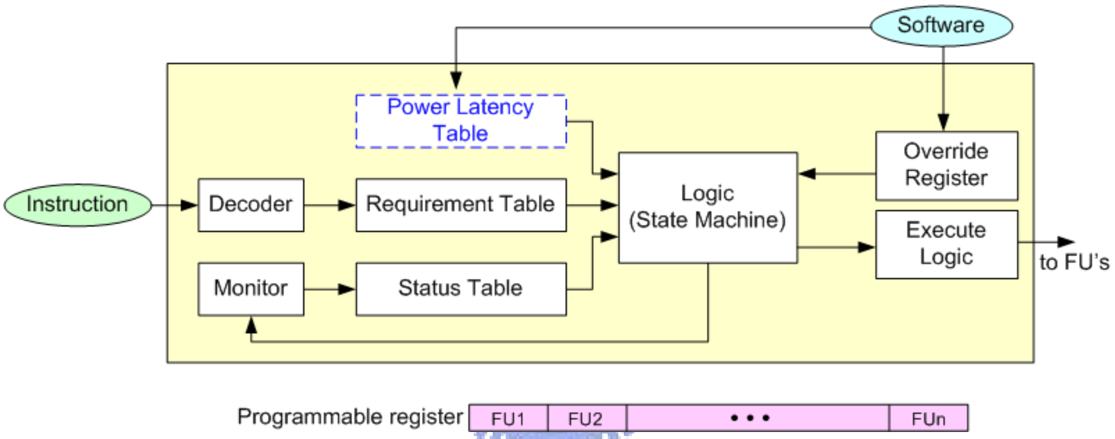


Fig. 2.26 A Vt management scheme using programmable status table.

2.4.4 Execution Circuit for Power and Performance Transition

Fig. 2.27 shows an execution circuit to adjust the voltage and frequency. Assume that each function unit has dedicated DC/DC converter and body bias generator, which can be independently controlled by PMU. The voltage generators can be shut down if the function unit is inactive. Moreover, the supply voltage and body bias can be dynamically adjusted if the DC/DC converter and body-bias generator are configurable.

Each function unit has a dedicated frequency divider to vary the operating frequency. According to the control signals from PMU, the frequency of function unit can be adjusted through a frequency divider. Undoubtedly, the clock signal fed into function unit can be gated for further power saving.

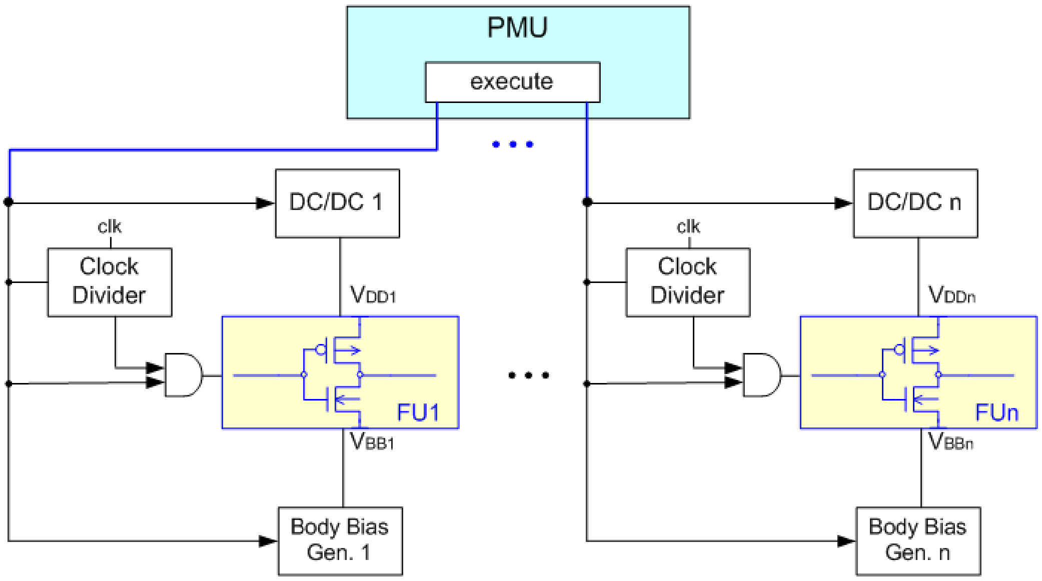


Fig. 2.27 Execution circuit to control power and frequency.

2.5 Conclusion

The most important leakage sources in deep-submicron and nano-scale CMOS devices are described in Sec. 2.1. Among them, the subthreshold leakage is the most critical portion and can be significantly reduced by raising threshold voltage, Vt. A popular technique for adjusting threshold voltage is to apply body bias. The influences on leakage and performance of applying body bias are also discussed. In Sec. 2.2, the influence of die-to-die and within-die parameter variations are discussed.

Some techniques and circuits for controlling threshold voltage are introduced in Sec. 2.3. Some of them detect and control threshold voltage in real time, and others switch body voltage depending on operating mode. Finally, the concept of PMU and Voltage Islands are described in Sec. 2.4, and a threshold-voltage managing scheme using a status table is introduced.

By applying RBB a significant amount of leakage current can be reduced, while applying FBB can achieve higher performance. However, the extra power and area induced by body-bias generators cannot be ignored. Generally speaking, the power overhead must be less than saved leakage power. In addition, the time latency of charging and discharging substrate nodes is another issue that influences the performance.

Voltage Islands is a system architecture and chip implementation methodology that can dynamically manage power and performance for SoC designs. An Island means the area that is fed by the same supply voltage. The process of functional partitioning identifies the optimal supply voltage of each functional component that minimized active power at the required performance.

Chapter 3
On-Chip Voltage Generators and Configurable Body-Bias Generator Design

In the past decades, on-chip voltage generators have been widely used in commercial memory chips such as DRAMs and Flash memory [3.1], [3.2], [3.3]. As shown in Fig. 3.1 [3.4], For example, DRAM chips need various kinds of power-supply voltages, which have been generated internally by using single external power supply. This feature is getting more emergent in recent low-power and SoC (System-on-Chip) systems.

Not only the advantages in memory designs, voltage generators are beneficial in other digital ICs as well. In the future low-voltage CMOS IC designs, internally generated voltages will be indispensable to reduce subthreshold current, which exponentially increases with decreasing threshold voltage. Many logic-oriented circuits have adopted this technique to control their threshold voltages [3.5], [3.6], [3.7]. It's commonly used to dynamically adjust threshold voltage by utilizing body bias in a triple-well technology, as illustrated in Fig. 3.2[3.8]. Obviously, various voltages must be generated internally to meet the requirements, and therefore some on-chip voltage generators are necessary.

As for body bias generators (V_{BB} generators), charge pump based structure is quite popular, and they can generate voltages that higher than supply voltage (V_{DD}) or lower than ground (GND). The basic principles of positive-pumping and negative-pumping are described in Sec. 3.1 and Sec. 3.2, respectively. A digital-controlled configurable scheme is discussed in Sec. 3.3 and Sec. 3.4 shows the simulation results. Finally, some conclusions are made in Sec. 3.5.

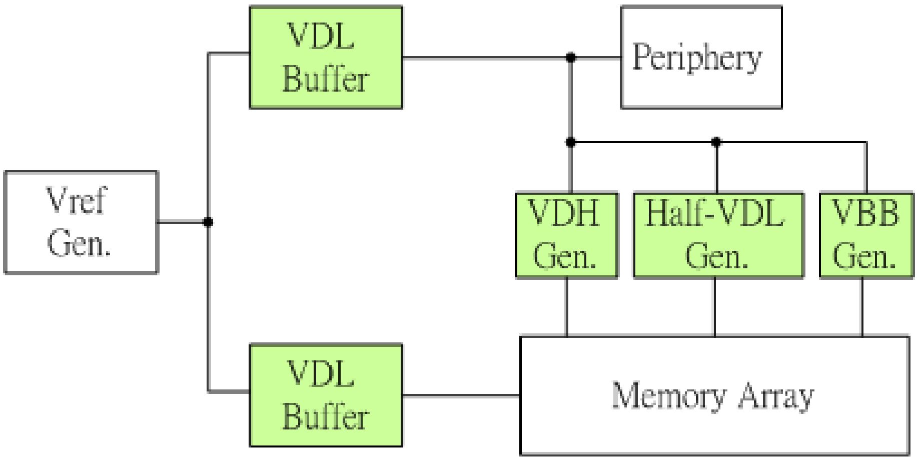


Fig. 3.1 Internal supply voltages for modern DRAM.

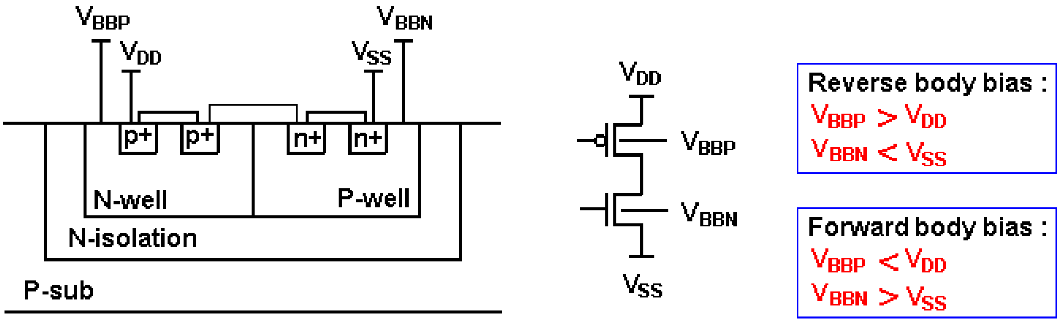


Fig. 3.2 Cross-section view of triple-well technology.

3.1 Positive-Pumping Circuits

Positive charge pumps are circuits that can pump charges upward to produce voltages higher than the common supply voltage. Those circuits are widely used in non-volatile memories such as EEPROM and Flash memories [3.9]. This section introduces the basic operations of positive-pumping circuits and some advanced positive charge pumps.

3.1.1 Dickson Charge Pump

Most charge pumps are based on the circuit proposed by Dickson [3.10], and the circuit is called “Dickson charge pump”.

3.1.1.1 Overview of Dickson Charge Pump

Fig. 3.3 shows the Dickson charge pump and the MOS transistors act as diodes, so the charges can only be pushed in one way. The circuit is composed of diode-connected MOS transistors and pumping capacitors, C_p. Generally the pumping capacitors can be replaced with MOS capacitors. The two pumping signals, clk and $\overline{\text{clk}}$ are out of phase and their peak-to-peak swings are both V_{DD}.

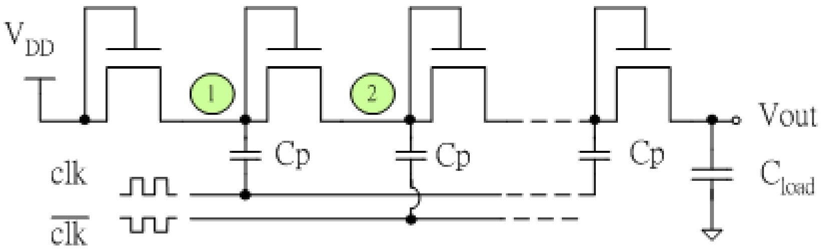


Fig. 3.3 Dickson charge pump.

3.1.1.2 Operation of Dickson Charge Pump

With the pumping capacitors, the two clocks push the charged nodes upward through the transistors. Each time when the clock signal goes from low to high, the voltage difference (denoted as △V) at internal node can be expressed as [3.11]

$$\Delta V = V_{DD} \cdot \frac{C_p}{C_p + C_s} - \frac{I_0}{f \cdot (C_p + C_s)} \tag{3.1}$$

where C_s is the parasitic capacitance at each node, f is the pumping frequency, and I_0 is the output current loading. When *clk* goes from low to high and $\overline{\text{clk}}$ goes from high to low, the voltage at node 1 is pumped to V_1 +△V, and the voltage at node 2 is settled to V_2, where V_1 and V_2 are defined as the steady-state lower voltage at node 1 and node 2, respectively. The voltage pumping gain for second pumping stage is defined as the difference between V_2 and V_1,

$$V_2 - V_1 = \Delta V - V_{tn2} \tag{3.2}$$

where V_{tn2} is the threshold voltage of the second transistor. Therefore, the necessary condition for the circuit to function is that△V must greater than the threshold voltage. For an ideal charge pump, the output voltage goes toward [3.12]

$$V_{out} = (V_{DD} - V_t) \cdot N + V_{DD} \tag{3.3}$$

where N is the number of stages.

3.1.1.3 Limitation of Dickson Charge Pump

From (3.3), it’s obvious that the voltage gain per stage of Dickson charge pump suffers from the threshold voltage losses. Unfortunately, the threshold voltage increases due to body effect, especially at the high-voltage nodes near the output. Therefore, the output voltage of Dickson charge pump cannot be a linear function of the number of stages. Moreover, the pumping efficiency degrades as the number of stages increases. Fig. 3.4 depicts the operation behavior abstractly.

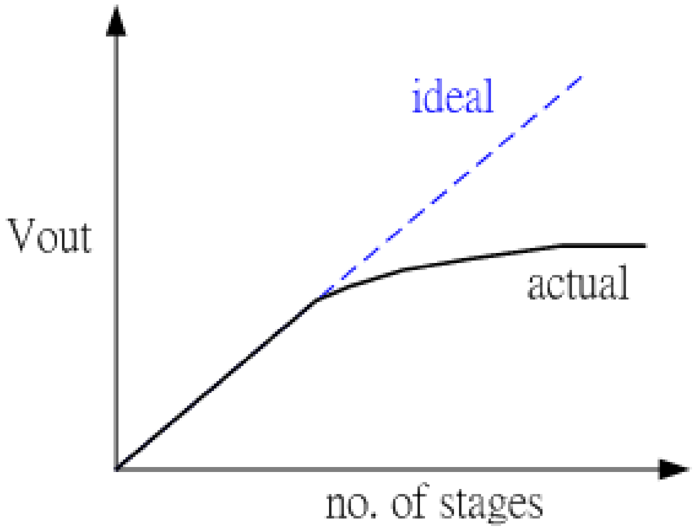


Fig. 3.4 Abstract behavior of Dickson charge pump.

3.1.2 Improvement of Voltage Gain

Due to the threshold voltage loss and the influence of body effect, a large number of researches attempt to alleviate this problem. In order to diminish the threshold voltage loss of conventional charge pump, one replaces most of the NMOSFET's with PMOSFET's [3.13]. This circuit achieves high efficiency and pumping speed, but some bootstrapped clock generators are needed. Moreover, it requires four pumping signals so that increases the complexity. On the other hand, using floating-well to eliminate the body effect is proposed [3.14]. However, the substrate currents generated may still reduce the efficiency.

3.1.3 Charge Pump Without Body Effect

From the discussions in the sections above, the main obstacles of charge pumps are the influence of body effect, and the operations under low supply voltage. One proposed a scheme that uses two auxiliary MOSFET's to control the body bias [3.12]. As in Fig. 3.5, each charge transfer block is composed of three PMOS transistors, where M_T is the charge-transfer transistor, and M_S and M_D are the two auxiliary transistors. When M_T is ON, the charges are transferred through it. Meanwhile, M_S is ON and M_D is OFF, the body and source of M_T are connected through M_S. on the other hand, when M_T is OFF, thus M_S is OFF and M_D is ON. In this condition the source and the body of M_T are still connected through M_D.

In summary, the two auxiliary transistors supply two paths to join the source and the body of M_D, one for *clk* is high and the other for *clk* is low. In this scheme since the charge-transfer transistor has zero source-body voltage, therefore it suffers no body effect and achieves higher voltage gain.

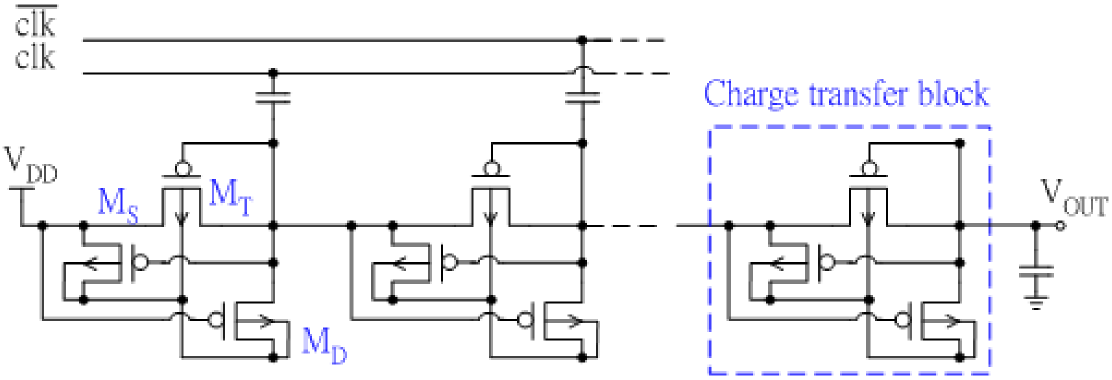


Fig. 3.5 The charge pump without body effect.

3.1.4 Charge Pump for Low-Voltage Operation

As discussed before, one major limitation of charge pump is the functional ability under low supply voltage. Fig. 3.6 shows a charge pump using dynamic charge transfer switch and backward control [3.11]. This circuit is suitable for operating under low supply voltages.

In this scheme the single-stage voltage pumping gain is

$$G_V = G_{V2} = V_2 - V_1 = \Delta V \tag{3.4}$$

When ϕ_1 is high and ϕ_2 is low, both the voltages at node 1 and node 2 are V2, and the voltage at node 3 is $2\Delta V$. In order to function well, the following expression must be satisfied

$$2\Delta V > V_{tp} \text{ and } 2\Delta V > V_{tn}(V_2) \tag{3.5}$$

On the other hand, when ϕ_1 is low and ϕ_2 is high, the voltage at node 1 is V1, both the voltages at node 2 and node 3 are $2\Delta V$. It must satisfies

$$2\Delta V > V_{tn}(V1) \tag{3.6}$$

In comparison with expressions (3.2), this circuit can achieve the required conditions more easily under low supply voltage.

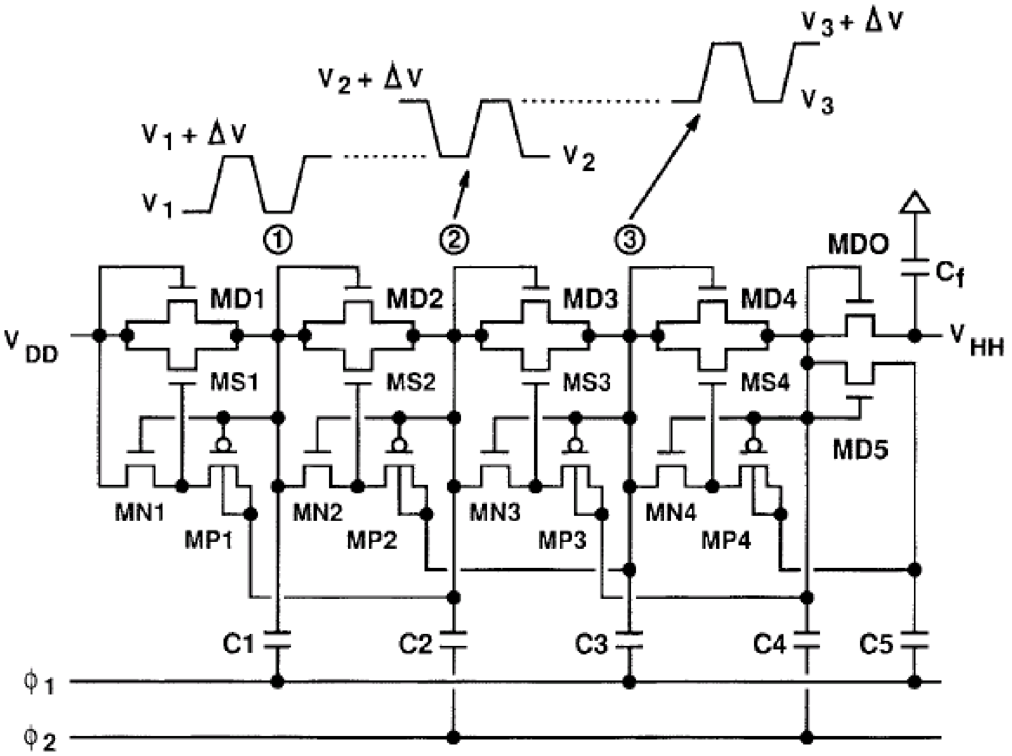


Fig. 3.6 A four-stage charge pump.

3.1.5 Voltage Doubler

A large amount of voltage multipliers are based on Dickson charge pump, as discussed in the sections above. However, a high voltage can also be achieved by cascading several voltage doublers [3.15]. A voltage doubler can generate twice the magnitude of input voltage. Fig. 3.7 shows the popular cross-coupled structure voltage doubler proposed in [3.16]. This circuit needs a series switch to output the doubled DC voltage.

A scheme uses two charge pump blocks, one for the supply and the other to bias the body of the switch, as shown in Fig. 3.8 [3.17]. Since there are no junction bias between the body and the output, thus no substrate current exists. The disadvantages of Fig. 3.8 are that it requires two equivalent blocks and the body of *P2* is still unbiased.

To solve these problems, another design uses a dual series switch and the principle of bulk switching, as depicted in Fig. 3.9 [3.18]. M3 and M4 are series switches, and M5 and M6 switch to the highest voltage. For M3 and M4, their body and the output node and the chip substrate compose of vertical PnP bipolar transistor. Since M5 and M6 switch the bodies of M3 and M4 to the highest voltage, the circuit is latch-up immune.

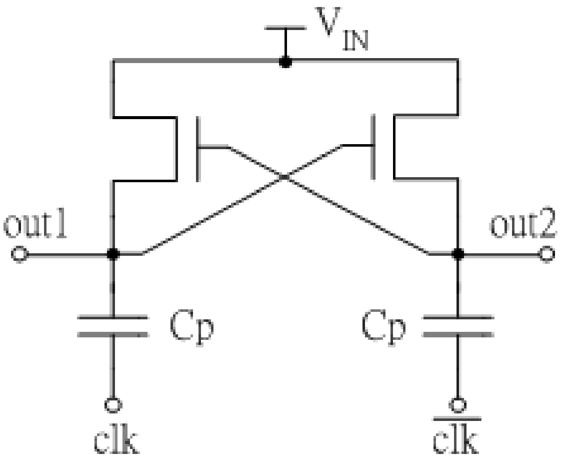


Fig. 3.7 Cross-coupled voltage doubler.

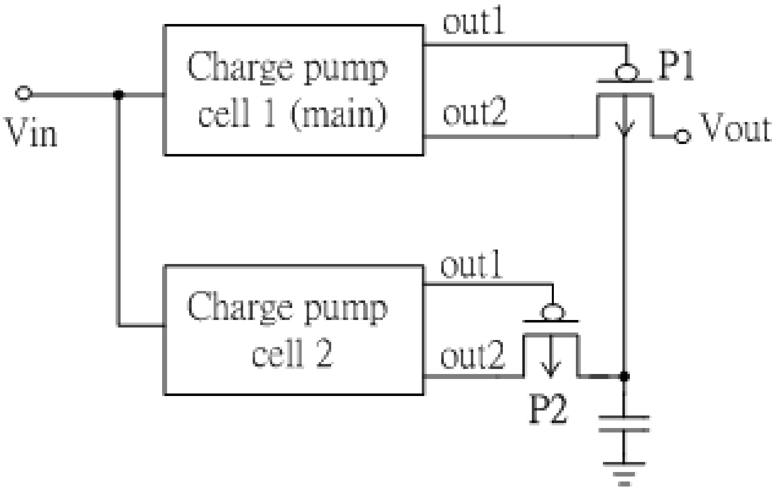


Fig. 3.8 Charge pump with PMOS bias.

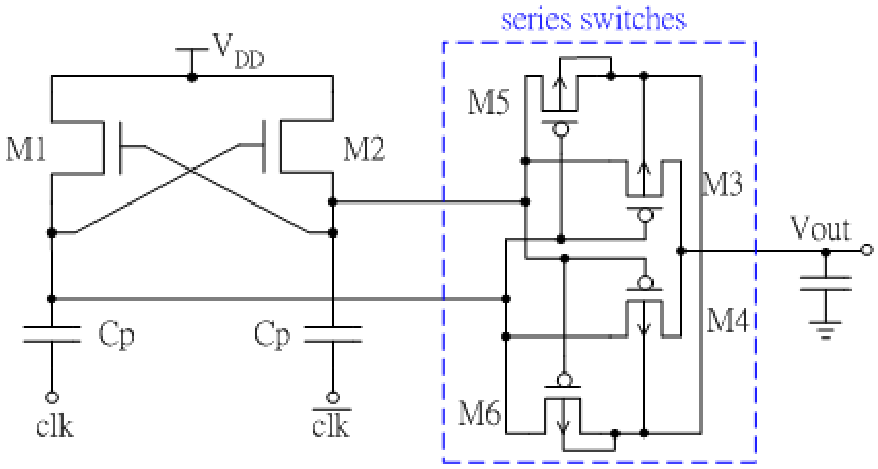


Fig. 3.9 Voltage doubler with series switches.

3.2 Negative-pumping circuits

In contrast to positive-pumping circuits, negative-pumping circuits generate voltages lower than ground (potential = 0). The most common usage of negative-pumping circuits is to reversely bias the pn junction between source and body of NMOS transistors and suppress subthreshold leakage current. The body-bias voltage generators have been adopted in memory for a long time. They can stabilize the operations of memory cells and peripheral circuits [3.4]. Not only in memories, they are getting more important in logic designs, especially in nano-scale and SoC eras. In this section, the principles of negative-pumping circuits will be addressed.

3.2.1 Basic Principles of Negative-Pumping Circuits

Fig. 3.10 (a) shows the conventional negative-pumping circuit [3.19]. The circuit comprises two diode-connected NMOS transistors and one capacitor. We can call it NMOS system here. When *clk* is high, the internal node n1 is pushed upward to V_{tn}, the threshold voltage of NMOS. When *clk* goes low, node n1 is pulled to $(-V_{DD} + V_{tn1})$ and the output node V_{BB} is

$$V_{BB} = -V_{DD} + V_{tn1} + V_{tn2} \tag{3.7}$$

It's easily understood that the NMOS system suffers from body effect seriously, since V_{BB} becomes shallower with the increases of V_{tn1} and V_{tn2} due to body effect. Besides, the minority carrier injection occurs when node n1 goes low, because the n+ regions of Q1 and Q2 are forward biased against the p-well.

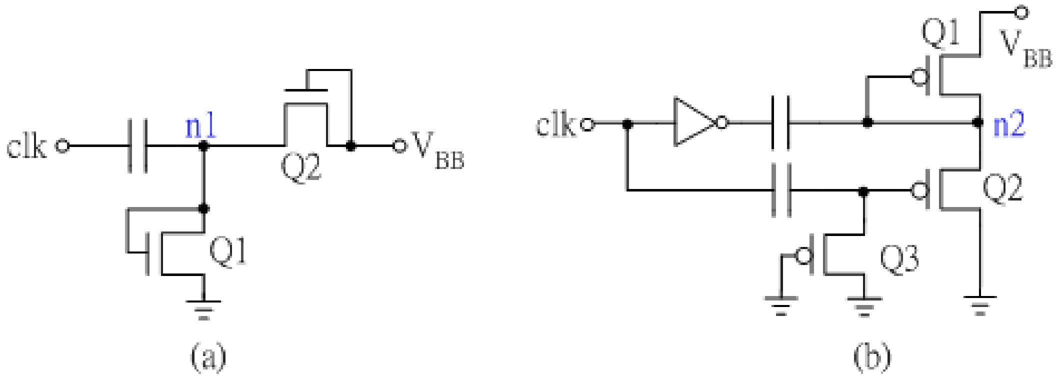


Fig. 3.10 (a) NMOS system and (b) PMOS system.

A charge pump that is composed of PMOS transistors called PMOS system is shown in Fig. 3.10 (b). Since node n2 can achieve $-V_{DD}$, thus output node V_{BB}

reaches

$$V_{BB} = -V_{DD} + \left|V_{tp}\right| \tag{3.8}$$

where the V_{tp} is due to Q1. In comparison with equation (3.8) and (3.7), the PMOS system generates deeper voltage than NMOS system but still slightly suffers from threshold voltage loss.

3.2.2 No Vt-Loss Pumping Circuits

In contrast to the previous two circuits, in the following two high-performance circuits without Vt-loss are described.

3.2.2.1 Hybrid Pumping Circuit

The hybrid pumping circuit (HPC) uses both NMOS and PMOS, as in Fig. 3.11 [3.20]. The most important feature of HPC is the replacement of Q1. When *clk* is low, node n3 reaches ($-V_{DD} + |V_{tp}|$) and node n4 is grounded through Q2. When *clk* goes high, node n4 is pulled down to $-V_{DD}$. Meanwhile, the high voltage at node n3 turns on Q1, and pulls V_{BB} down to $-V_{DD}$.

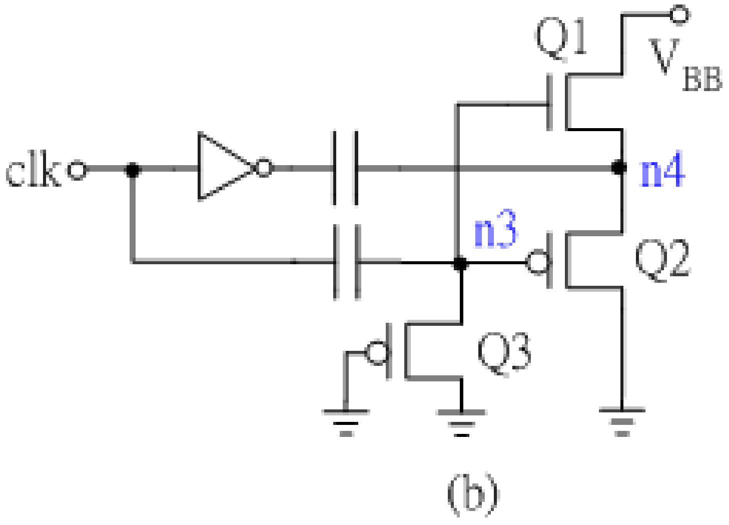


Fig. 3.11 Hybrid pumping circuit (HPC).

3.2.2.2 Cross-Coupled Hybrid Pumping Circuit

Another high performance pumping circuit uses cross-coupled structure to achieve high-speed [3.21], as shown in Fig. 3.12. Because nodes n5 and n6 are cross-coupled, their low voltages are enough to fully turn on MP1 and MP2, respectively.

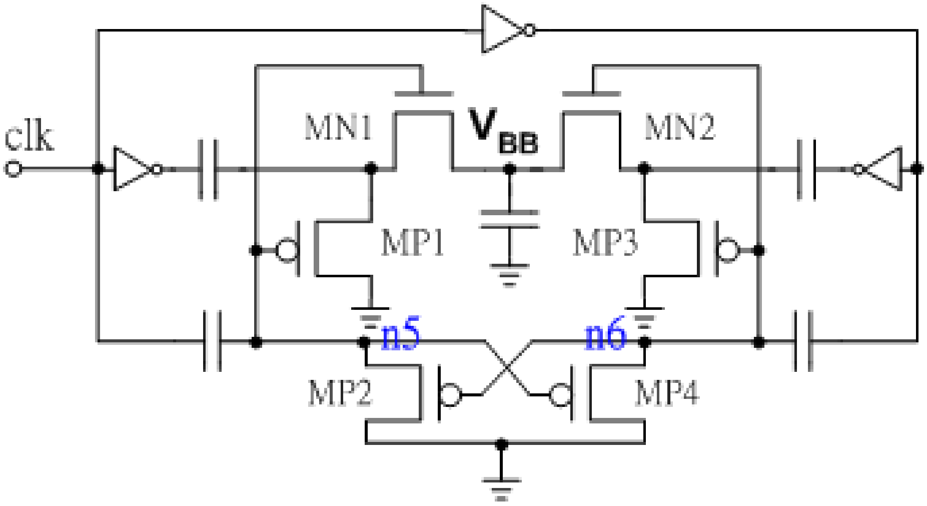


Fig. 3.12 Cross-coupled hybrid pumping circuit.

3.2.2.3 Simulated Output Waveforms

Fig. 3.13 shows the simulated output waveforms for the previous charge pumps. As described above, NMOS system suffers from two Vt losses and PMOS system suffers from one. On thee other hand, both HPC and cross-coupled HPC are free from body effect so that they can closely reach $-V_{DD}$.

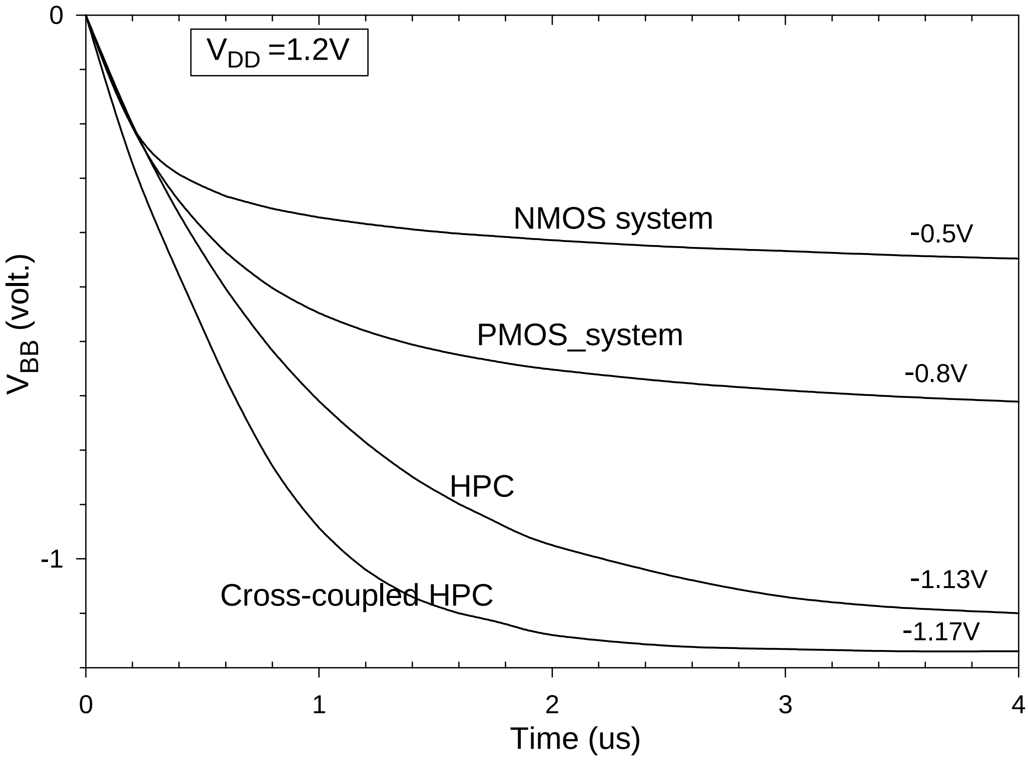


Fig. 3.13 Simulated waveforms for the charge pumps.

3.2.3 Pumping Current (I_{CP}) and Substrate Current (I_{BB})

The pumping current I_{CP} is proportional to $(C \cdot V_{DD} \cdot f)$ and decreases with increase of $|V_{BB}|$, where C and f are pumping (or kicking) capacitor and pumping frequency, respectively. On the other hand, the substrate current is generated at the drain with high electric field. For a fixed V_{DD}, the electric field near the drain strengthens as the MOSFET is scaled down. Consequently, electrons flowing from the source to the drain obtain a high energy from the high electric field (also called “hot-carrier”), and generate electron-hole pairs as the result of impact ionization at the drain. As shown in Fig. 3.14, some electrons flow into the drain, and the others are injected into the gate insulator as the gate current (I_G) and are trapped there. The trapped electrons cause a gradual change in V_T and a decrease in the transconductance of the MOSFET. On the other hand, some holes of the pairs flow into the substrate and result in substrate current I_{BB}. Therefore, I_{BB} is the hole component of the impact ionization current, which is subject to shallow V_{BB}. In summary, the V_{BB} level is settled by both charge pumping current (I_{CP}) of the voltage generator and substrate current (I_{BB}).

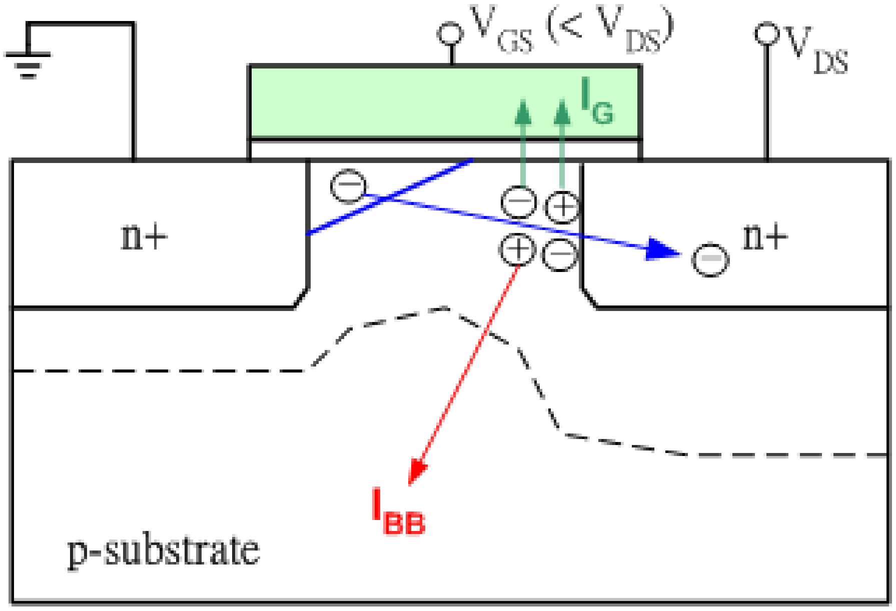


Fig. 3.14 The mechanism of hot-carrier injection.

3.3 Configurable Multiple-Voltage Generators

The previous sections describe both negative and positive pumping circuits. In general, all of the circuits generate only one voltage level and therefore their applications are limited. As being a body bias generator to reversely bias the body terminal of a chip, it seems so inflexible since many researches and measurements have demonstrated that an optimal reverse bias point exists. Moreover, the optimal bias points are different for different technologies and temperatures.

With the trend of SoC (System-on-Chip), a chip needs several supply voltage levels and several body-bias voltage levels as well. If we use conventional voltage generators, however, several distinct voltage generators are needed. It’s a big cumbrance to deal with so many kinds of voltage generators. For these reasons, a configurable voltage generator that produces several voltage levels is quite beneficial. In the following the configurable voltage generators that generate multiple voltage levels will be discussed in detail, including positive and negative pumping structures.

3.3.1 Overview of Configurable Scheme

Fig. 3.15 shows the configurable multiple-level voltage generator. It’s comprises a ring oscillator, a code converter, two D/A converters and initial controller, a charge pump, and a recovery circuit. The detail functionalities of the functional blocks will be described later.

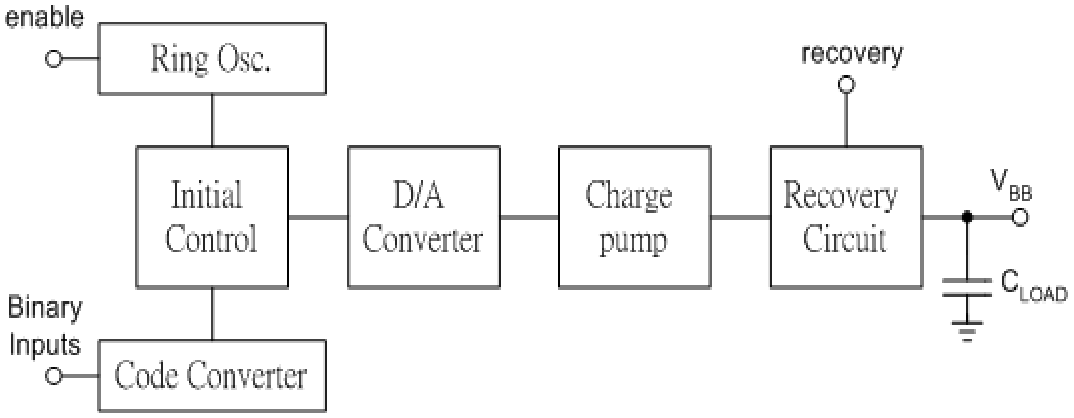


Fig. 3.15 Configurable scheme for multi-voltage generator.

3.3.1.1 Ring Oscillator

The ring oscillator supplies the required pumping signals internally. It’s a basic inverter chain with odd number of inverters with an enable control, as shown in Fig. 3.16. As for the whole system, the higher the pumping frequency is, the faster the output node achieves the steady-state voltage level. However, more power is consumed for higher pumping frequency. Therefore, the frequency of the ring oscillator is a compromise between power consumption and settling speed.

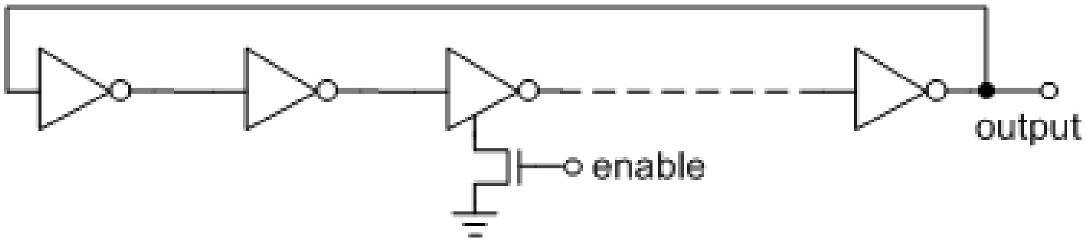


Fig. 3.16 The ring oscillator with enable control.

3.3.1.2 D/A Converter

The D/A converters are used to generate various clocking signals whose swings are between GND and V_{DD}. The input full-swing pumping signals come from the ring oscillator, and the D/A converters truncate the peak-to-peak voltages. Base on the input settings of the D/A converters, the output clocking signals have different values of swings. Usually, a D/A converter receives digital input signals and generates analog output voltages. It's a little bit tricky that the input signals here are clocking signals and therefore the outputs are like reshaped clocking signals.

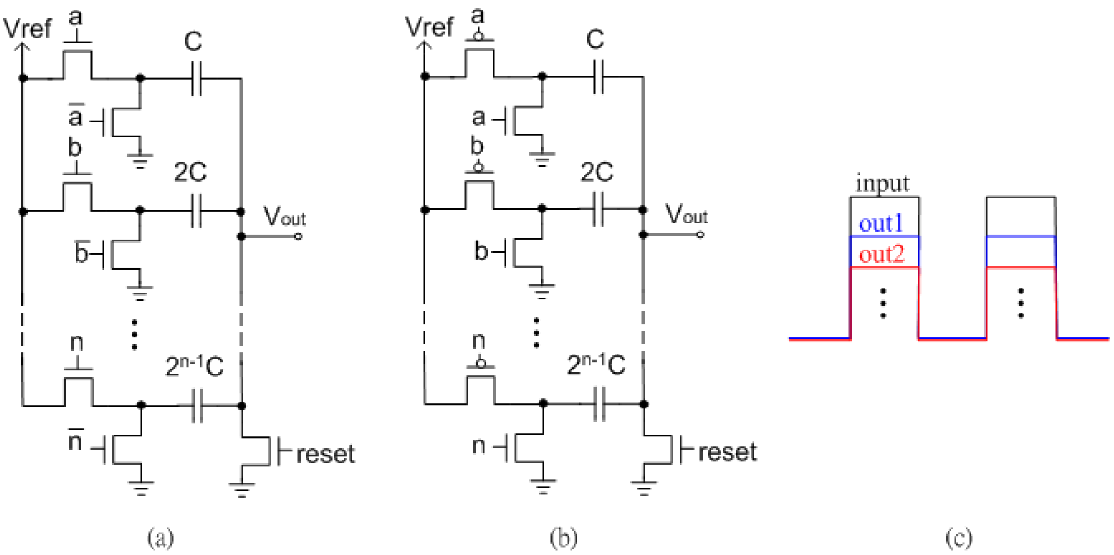
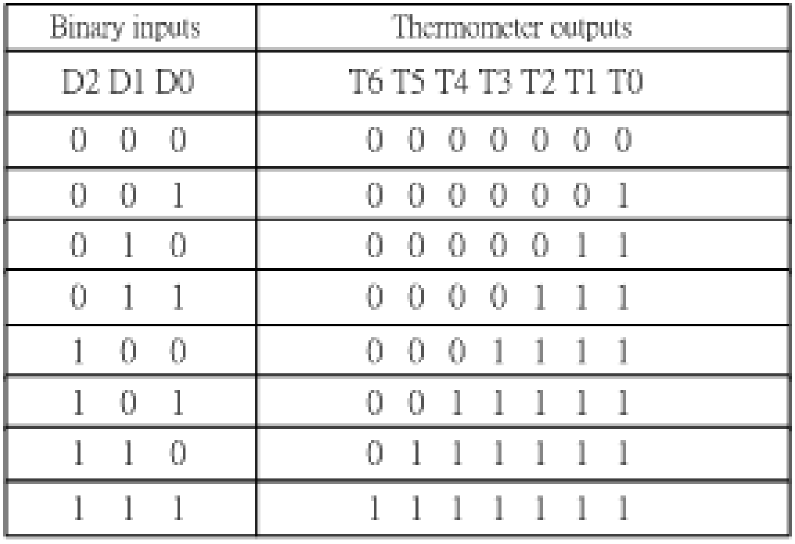


Fig. 3.17 (a) The charge-redistribution D/A converter in [3.26] (b) The modified circuit to prevent from Vt loss. (c) The output waveforms.

Generally, all kinds of D/A converters are usable in this scheme, and here we use charge-redistribution D/A converter as an example. Fig. 3.17 shows the well-known charge-redistribution D/A converter that comprises some MOS switches and capacitors [3.22]. There are two major classifications of charge-redistribution D/A converters according to the capacitors scaling. If the sizes of the capacitors are in power of two, the D/A converter is binary-weighted. On the other hand, it's unary-weighted if the capacitors are identical. In contrast to the unary-weighted version, the binary-weighted version has fewer inputs but worse linearity. Fig. 17 (a) is the original circuit in [3.22]. Since V_{ref} always equals V_{DD}, the NMOS pass transistors connected to V_{ref} are replaced with PMOS ones to prevent from Vt loss. Besides, an unexpected advantage is that no inversed input control signals are needed.

3.3.1.3 Code Converter

As mentioned above, the linearity of a unary-weighted D/A converter is better than a binary-weighted one, but unary-weighted version has more inputs than binary-weighted version. In order to have the both advantages of the two versions at the same time, a code converter is included. The code converter simply converts binary inputs to thermometer codes. A 3-bit binary-to-thermometer conversion is described in Table 3.1. The code converter may occupy a large area compared to other functional blocks if it's constructed by basic NAND and NOR gates. Therefore, this block can be omitted to reduce area if the number of input pins is not a problem.



Binary inputs	Thermometer outputs
D2 D1 D0	T6 T5 T4 T3 T2 T1 T0
0 0 0	0 0 0 0 0 0 0
0 0 1	0 0 0 0 0 0 1
0 1 0	0 0 0 0 0 1 1
0 1 1	0 0 0 0 1 1 1
1 0 0	0 0 0 1 1 1 1
1 0 1	0 0 1 1 1 1 1
1 1 0	0 1 1 1 1 1 1
1 1 1	1 1 1 1 1 1 1

Table 3.1 A 3-bit binary-to-thermometer conversion.

3.3.1.4 Initial Control

This block is used to reset all the internal nodes of D/A converters in the initial state. At first, the internal nodes inside the D/A converters contain some charges and they will degrade the linearity and operating speed of the D/A converters. With the use of initial control, all the internal nodes are pulled to ground before the operation of the D/A converter, whatever the input signals are. Fig. 3.18 depicts the circuit and it consists of basic gates.

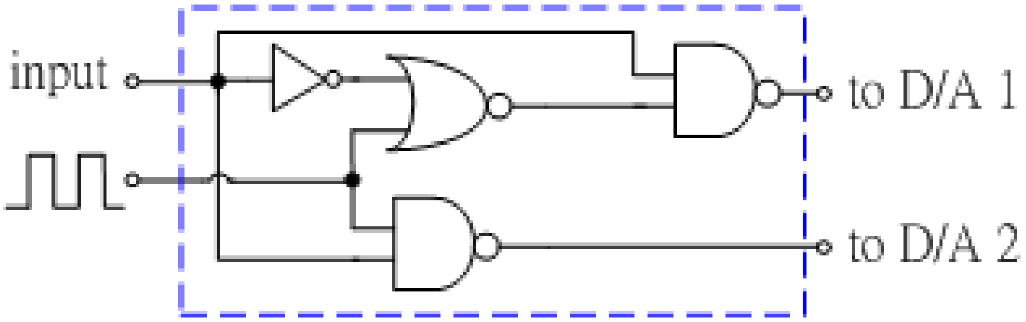


Fig. 3.18 The initial control to initialize the D/A converters.

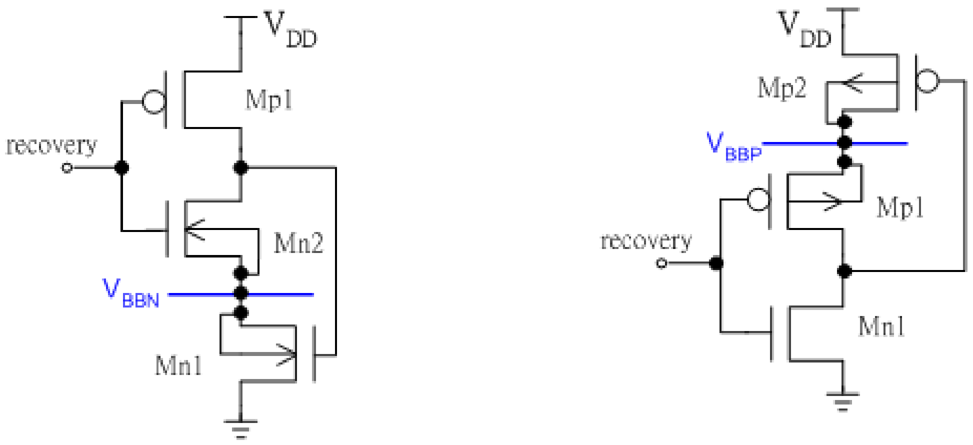
3.3.1.5 Charge Pump

This block can be either negative or positive pumping circuits, and negative circuits are for biasing V_{BBN} (V_{BB} for NMOS) and positive circuits for V_{BBP} (V_{BB} for PMOS). As for negative pumping circuits, Vt-loss-free charge pumps are preferable since they have wider scaling space. In order not to induce too much band-to-band tunneling leakage, the maximum value of RBB (Reversed Body Bias) had better to be limited to $|V_{DD}|$. Therefore, a voltage doubler is an excellent choice for biasing V_{BBP}.

3.3.1.6 Recovery Circuit

V_{BB} generators reversely bias the substrate of inactive circuits to reduce subthreshold leakage current. However, V_{BB} generators must have the ability to return V_{BBN} and V_{BBP} to their nominal values, GND and V_{DD} respectively. Recovery circuits in this proposed scheme are responsible for the recovering operations of V_{BBN} and V_{BBP}.

Fig. 3.19 shows the recovery circuits for V_{BBN} and V_{BBP}. V_{BBN} is set to GND when recovery signal is low, and the circuit has no influence on V_{BBN} when recovery signal is high. In contrast, V_{BBP} is set to V_{DD} when recovery signal is high and the circuit has no influence on V_{BBP} when the recovery signal is low.



	V_{BBN}
recovery = 0	**0 (recovery)**
recovery = 1	**Pump down**

(a)

	V_{BBP}
recovery = 0	**Pump up**
recovery = 1	**V_{DD} (recovery)**

(b)

Fig. 3.19 Recovery circuits (a) for V_{BBN} and (b) for V_{BBP}.

3.3.2 Output Voltage Equations

The output voltage equations can be easily derived since only the D/A converters have influence on output voltage. The equation below describes the relation between V_{BBN} and input codes.

$$\boxed{V_{BBN} = -V_{DD} * \left(\frac{input\ \ value}{2^{(no.\ \ of\ \ inputs)}-1}\right)} \tag{3.9}$$

where the term ‘input’ means the binary inputs of D/A converters. If input code is {011}, for example, the terms ‘input value’ and ‘no. of inputs’ in (3.9) are 6 and 3, respectively. The equation for V_{BBP} can be derived as below

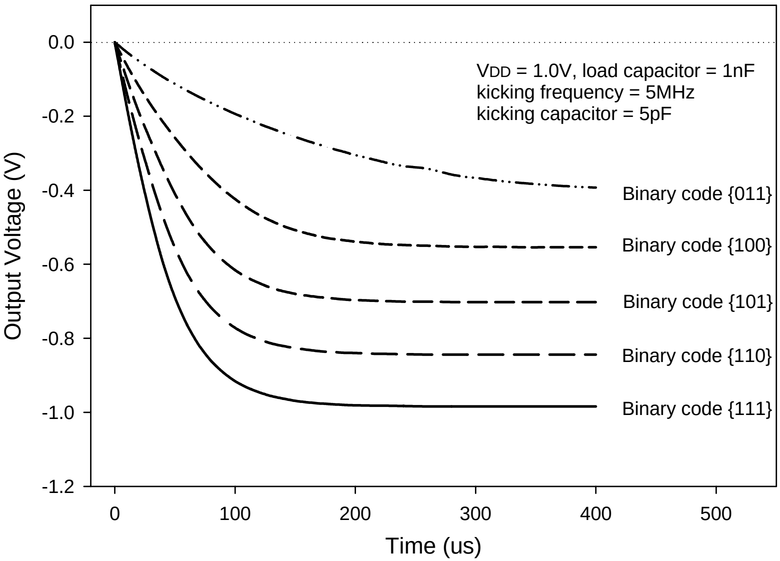
$$\boxed{V_{BBP} = V_{DD} + V_{DD} * \left(\frac{input\ \ value}{2^{(no.\ \ of\ \ inputs)}-1}\right)} \tag{3.10}$$

3.4 Simulation Results

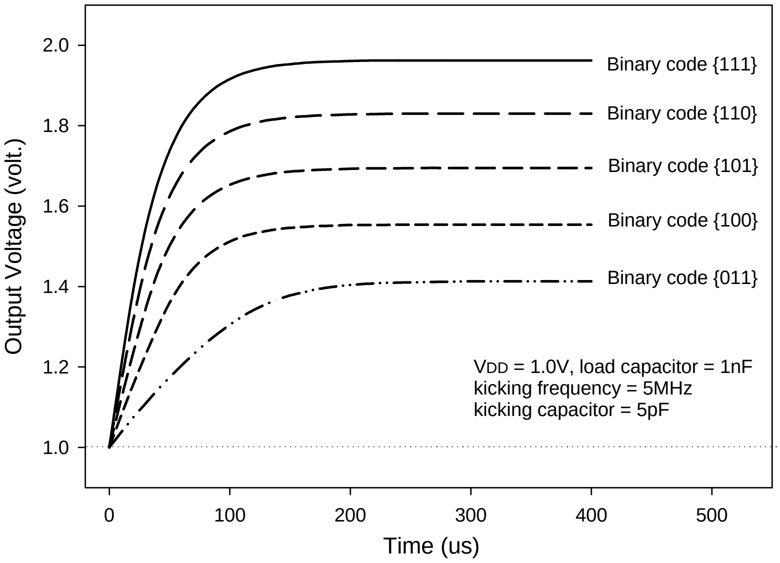
The proposed configurable scheme is simulated by HSPICE with the spice parameter of TSMC 100nm CMOS technology.

3.4.1 Transient Waveforms

Figure 3.20 shows the output transient waveforms of the V_{BB} generators with different values of binary inputs. Note that the number of bits of the binary inputs is 3. Fig. 3.20 (a) is for V_{BBN} and Fig. 3.20 (b) is for V_{BBP}. Both the figures not only illustrate the flexibility of the proposed architecture, but also demonstrate the feature of configurability. Various voltage levels can be achieved through binary inputs.



(a)



(b)

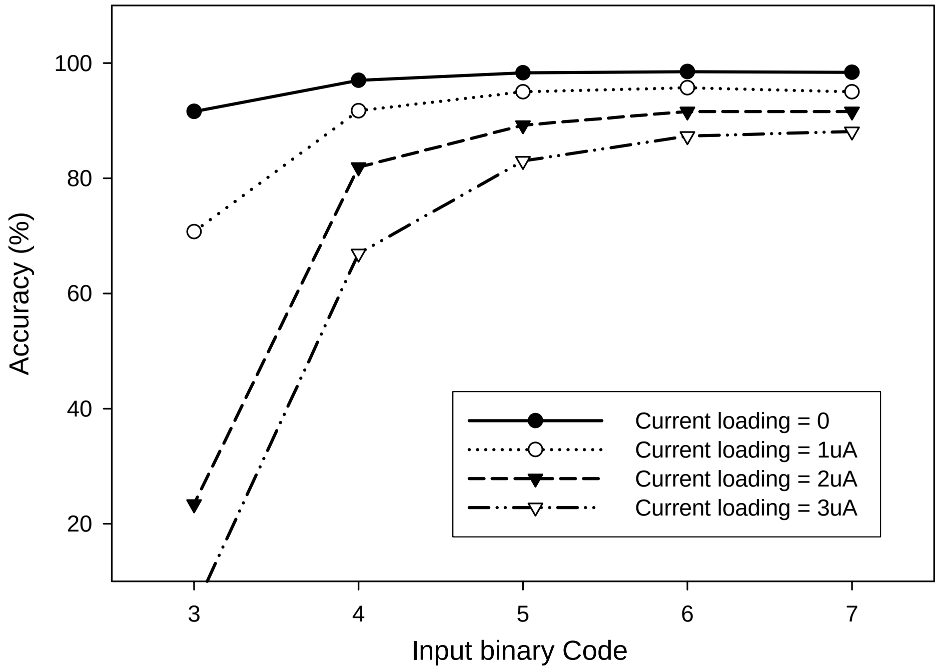
Fig. 3.20 Output transient waveforms of the V_{BB} generators (a) for V_{BBN} and (b) for V_{BBP}.

3.4.2 Accuracy Versus Current Loading

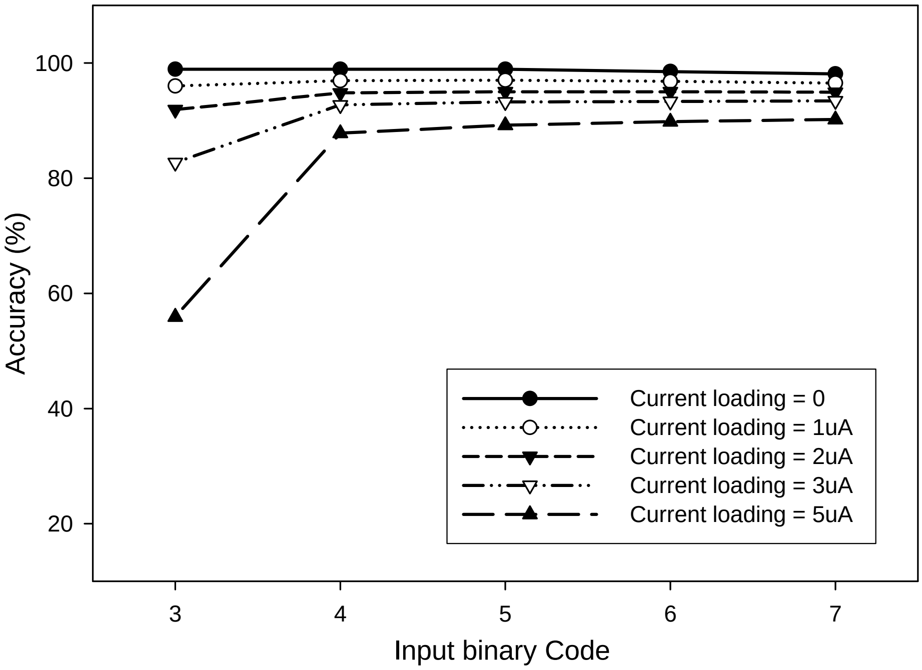
Fig. 3.21 plots the accuracy curves that with the influence of current loading. The accuracy is defined as the ratio of simulated output voltage to ideal output voltage. The output accuracy is higher than 90% without current loading. The applied current loading is served as the substrate current flows in the substrate of a chip, and it degrades the pumping efficiency and accuracy.

In both of the situations in Fig. 3.21, the accuracy starts to degrade severely at the binary input code {011}, and the curves maintain flat with input codes larger than {011}. The output voltage levels are settled by the charge pumping current and the substrate current loading. The absolute value of output voltage is larger with larger pumping current, which is proportional to the voltage swing of the pumping signals.

Fig. 3.21 indicates the weakness of this scheme. It would not be useful for applications with large load currents. Therefore, two or more V_{BB} generators are necessary for that kind of applications. For example, a 16-bit multiplier comprises thousands of gates and the substrate is an extremely large capacitance. Moreover, the substrate current might be large enough to destroy the functionality of V_{BB} generator. Therefore, the huge substrate capacitance must be divided into sections and one V_{BB} generator is responsible for one section. In this way, each V_{BB} generator will not face too huge capacitance and substrate current so that they can maintain good functionalities.



(a)

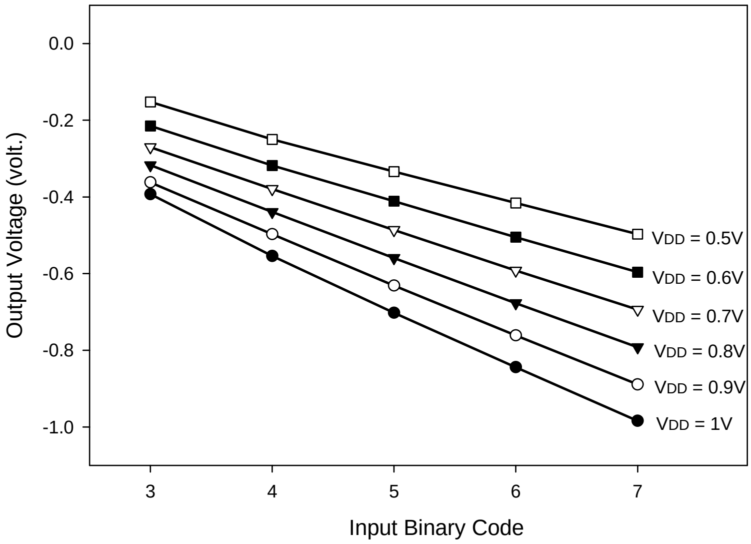


(b)

Fig. 3.21 Accuracy versus current loading (a) for V_{BBN} and (b) for V_{BBP}.

3.4.3 Operation at Low Supply Voltages

Fig. 3.22 shows the V_{BB} generators operate under different supply voltages. They work well at higher input codes even when the supply voltage is down to 0.5V. At lower supply voltages, however, the values of V_{BB} fail to achieve the targets as described in (3.9) and (3.10) due to small pumping currents.



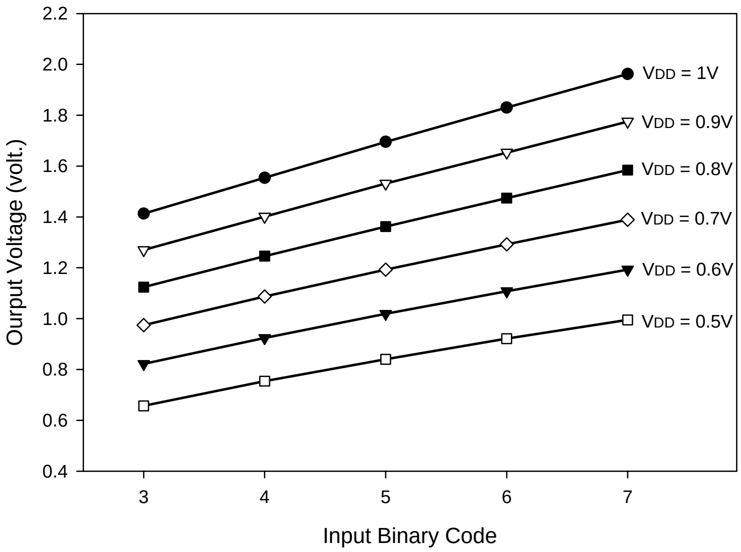


Fig. 3.22 Operations under different supply voltages.

3.5 Conclusion

In this section, the principles and operations of charge pumps are discussed. Both positive and negative pumping circuits are addressed, and some high-performance and advanced charge pumps are introduced. Besides, a novel scheme that generates multiple voltage levels through configurable control signals is realized.

Applying reversed body bias (RBB) is a popular technique to reduce subthreshold current, and voltages lower than GND and higher than V_{DD} should be produced. Using charge pump is a useful and simple method to realize that. But the performance of charge pump degrades severely for high load-current applications.

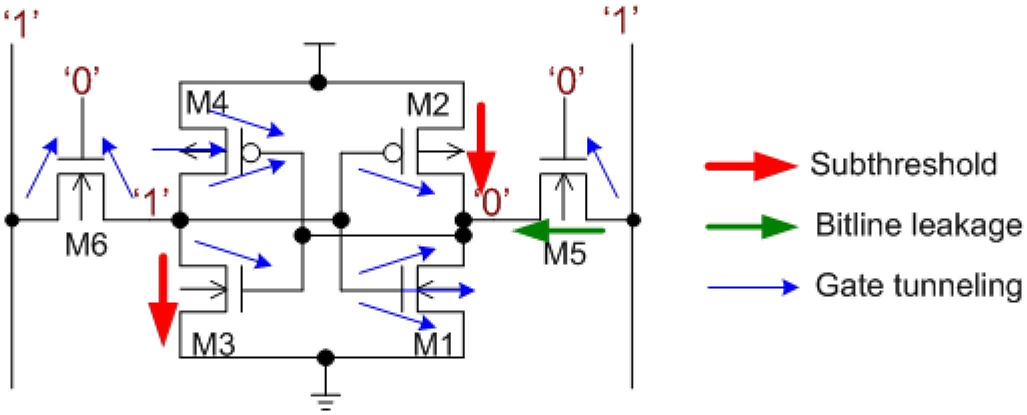
The configurable scheme generates various voltages with different input settings. This feature is especially useful in SoC designs, which need several different supply voltages and body-bias voltages. Instead of dealing with kinds of charge pump circuits, those voltages can be generated by using the same circuits with different control signals.

Chapter 4
Variable-Threshold CMOS (VTCMOS) SRAM Cell Arrays With On-Chip Body-Bias Generators

Power consumption is becoming a critical issue in designing processors, memories, and other logic circuits. By scaling down the supply voltages and threshold voltages, the active power consumptions of logic circuits are reduced dramatically and high performance is maintained. However, as the technology scales down, the leakage current in standby mode cannot be ignored anymore.

For almost all the VLSI or SoC (System-on-Chip) chips, various kinds of memories occupy most of the area of the chips. Therefore, the requirements of low power and low voltage memories are emergent. Since the memories cost a large fraction of chip area, their power consumptions play an important role in the whole chips. As above, active power can be significantly reduced by scaling down the supply voltage, but leakage current is increasing with the scaling of technologies.

Fig. 4.1 shows the leakage paths and standby leakage equations in a SRAM cell. Since subthreshold leakage is the dominant part of leakage currents in deep-submicron and nano-scale technologies, many variable-threshold CMOS (VTCMOS) SRAM have been proposed to reduce subthreshold current by dynamically applying reversed body-bias (RBB).



$$I_{off} = I_{sub\_cell} + I_{gate\_cell} + I_{bitline}$$

$$I_{sub\_cell} \sim I_{sub\_M2} + I_{sub\_M3}$$

$$I_{gate} \sim I_{gate\_M1}$$

$$I_{bitline} \sim I_{DIBL\_M5}$$

Fig. 4.1 Leakage currents and standby currents equations in a SRAM cell.

In Sec. 4.1, some representative VTCMOS SRAM architectures are introduced. An on-chip dual-level body-bias (V_{BB}) generator is presented in Sec. 4.2, and this circuit is applied to SRAM cell arrays to observe the effectiveness in saving leakage power. In Sec. 4.3, a time-out-policy controller for V_{BB} generator is described, and finally some conclusions are addressed in Sec. 4.4.

4.1 Variable-Threshold CMOS SRAM

In this section several VTCMOS SRAM circuits are reviewed, they dynamically adjust the body-bias to reduce subthreshold leakage current. Their operations and disadvantages are also discussed.

4.1.1 Dynamic Leakage Cut-off Scheme

Fig. 4.2 (a) shows the schematic diagram of dynamic leakage cut-off (DLC) SRAM, and Fig. 4.2 (b) is the operating waveforms and Fig. 4.2 (c) and (d) are the well bias drivers [4.2]. The n- and p-well bias voltages are V_{DD} and GND respectively for selected rows, while the unselected rows are $2V_{DD}$ and $–V_{DD}$, respectively. Through the well bias drivers, the n- and p-well bias voltages can be dynamically adjusted. In this way, the selected memory cells maintain high performance while the unselected memory cells perform low subthreshold leakage.

However, there are some questions about this scheme. First, Fig. 4.2 (b) depicts that V_{PWELL} and V_{NWELL} return to GND and V_{DD} respectively before V_{WL} rises. There might be some extra logic circuits to detect or predict the rises of V_{WL}. Second, the substrate is a large capacitive load and it takes a long time to charge and discharge it. Before V_{PWELL} and V_{NWELL} go back to the nominal values, V_{WL} and input signals must be delayed to avoid incorrect operations. Finally, no any V_{BB} generators are adopted in this scheme, it means that the voltages $-V_{DD}$ and $2V_{DD}$ are external voltage sources. This scheme seems so impractical since two more external voltage sources are needed.

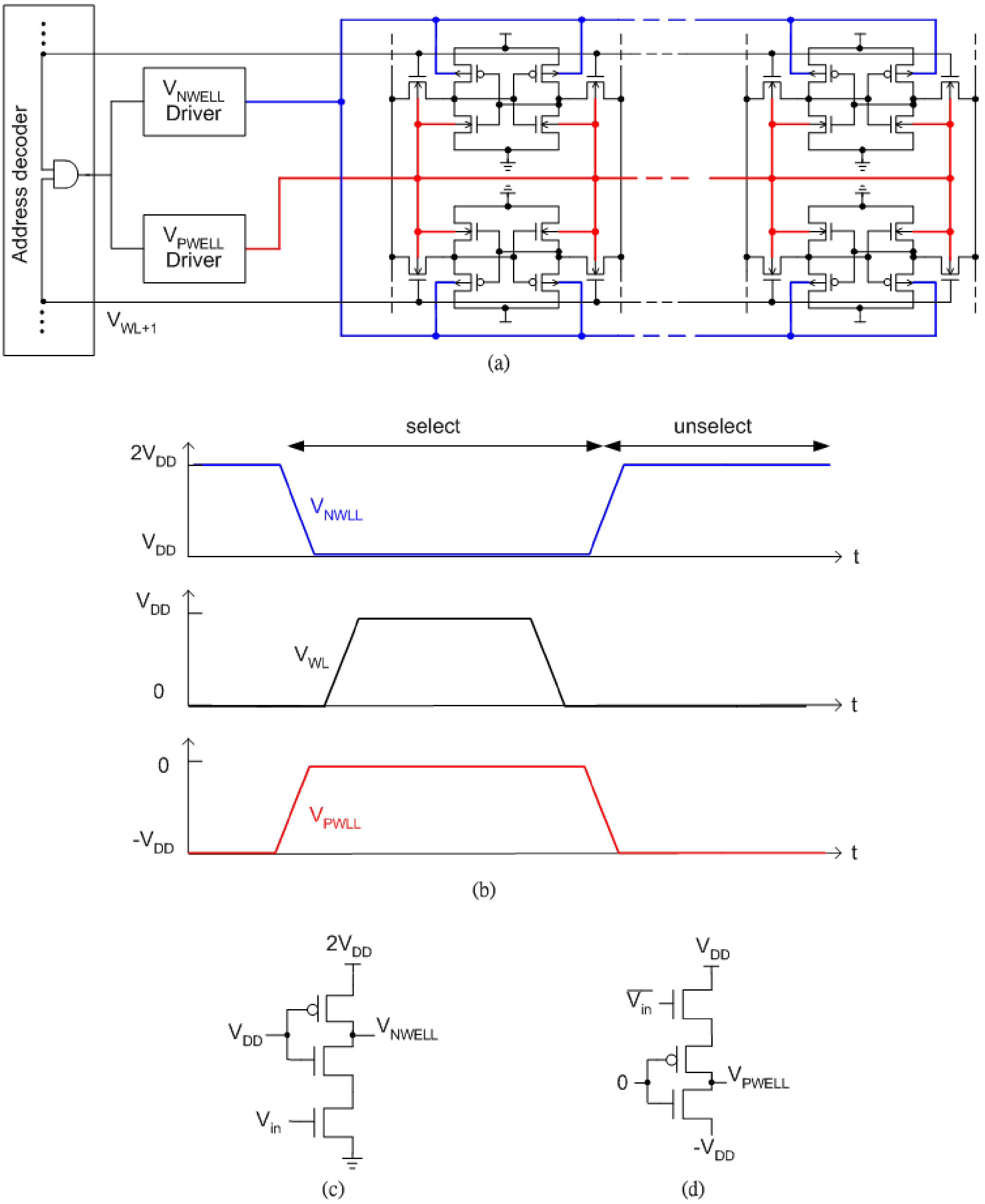


Fig. 4.2 (a) Dynamic leakage cut-off SRAM, (b) operating waveforms, well bias drivers for (c) n-well and (d) for p-well.

4.1.2 Preactivating Mechanism for VTCMOS Cache

Fig. 4.3 [4.2] uses address prediction to solve the problem of DLC circuit. It uses three address lines and two extra address decoders to predict the activity of wordline. Moreover, a reservation counter is included and it indicates the number of reservations for line accesses. This method concerns about the processor architecture and Fig. 4.4 [4.2] shows the processor architecture with a preactivating DLC cache.

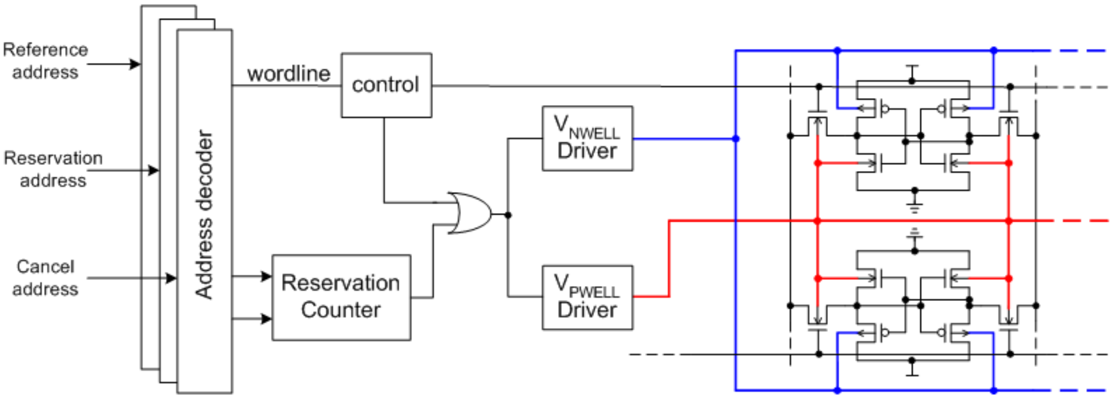


Fig. 4.3 Preactivating mechanism for a VTCMOS cache.

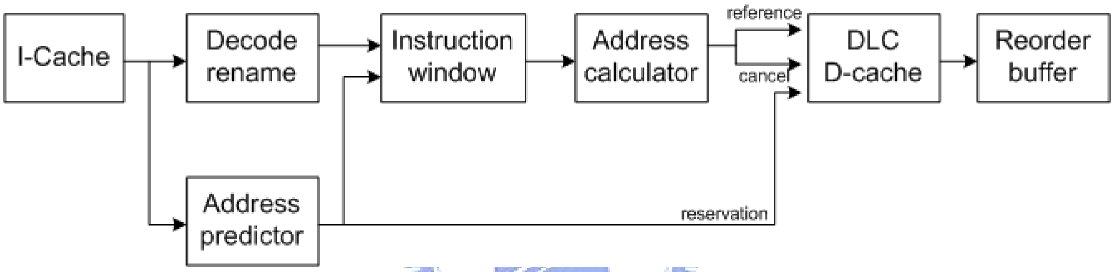


Fig. 4.4 Processor organization with a preactivating DLC cache.

4.1.3 Auto-Backgate-Controlled MT-CMOS

Fig. 4.5 shows the concept of the Auto-Backgate-Controlled MT-CMOS (ABC-MT-CMOS) circuit that uses two distinct external voltage sources (V_{DD1} and V_{DD2}) in different operating modes [4.3]. Q1 and Q2 here are high-Vt transistors, and low-Vt transistors are used for the internal circuits. While the circuit is operating (active mode), Q1 and Q2 are turned on and therefore the virtual source line VVDD and the virtual ground line VGND are 1V and 0, respectively.

In the sleep mode, Q1 and Q2 are turned off and the other voltage source V_{DD2} (3.3V) supplies the memory cells. The VVDD is connected to V_{DD2} through diode D1, while VGND is connected to ground through diode D2. Note that each of D1 and D2 consists two diodes and the forward bias of one diode is 0.5V. Hence, the VVDD and VGND are 2.3V and 1V respectively in the sleep mode.

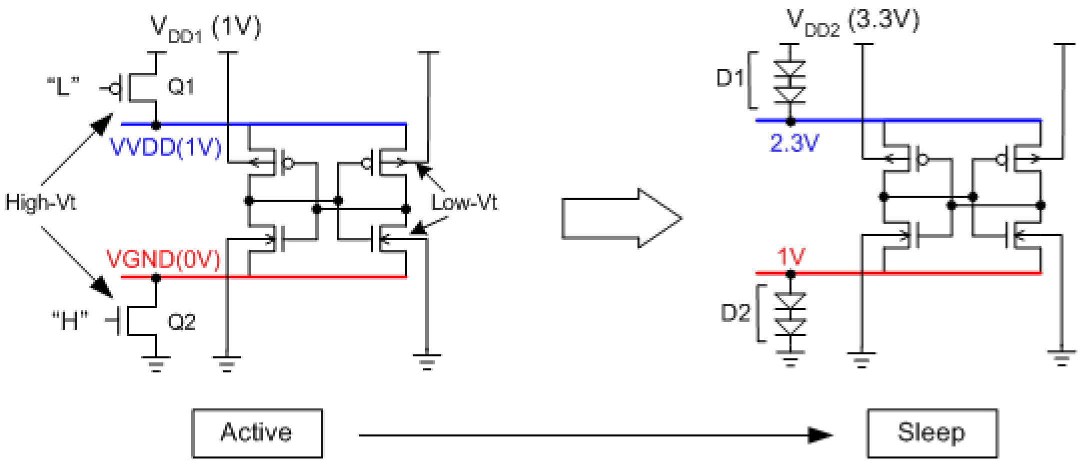


Fig. 4.5 Concept of ABC-MT-CMOS.

The static leakage current consumed by V_{DD2} is significantly reduced compared with that in the active mode because the threshold voltage of the internal transistors increases by the reversed body-source voltage. From Fig. 4.5 it can be easily understood that a 1V reversed body-source voltage is applied to the internal circuits.

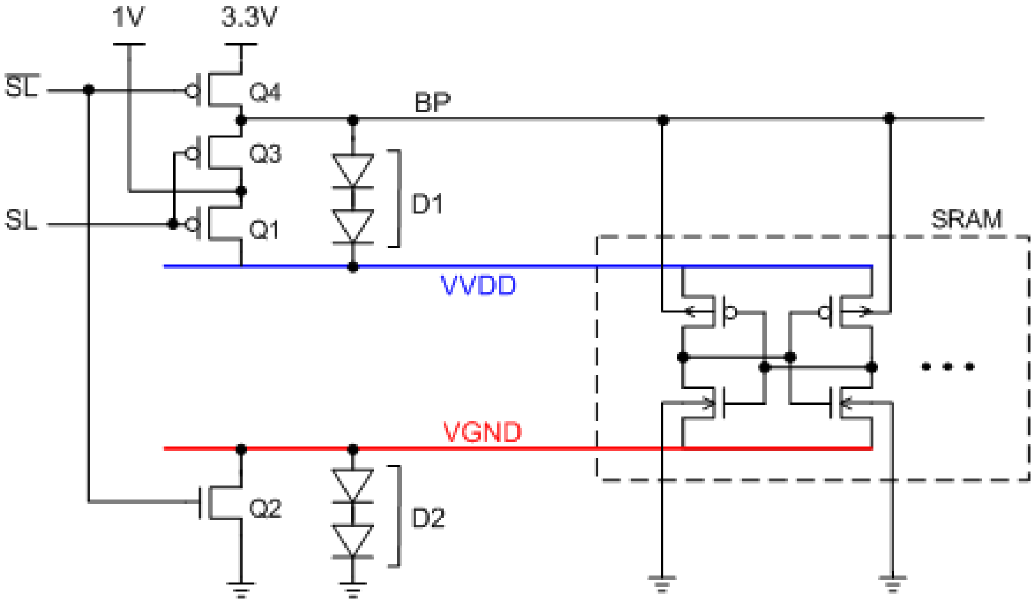


Fig. 4.6 Configuration of ABC-MT-CMOS.

Fig. 4.6 shows the actual configuration of the ABC-MT-CMOS circuit with two additional high-Vt transistors Q3 and Q4. In the active mode, *SL* is low and $\overline{SL}$ is high and thus Q1, Q2, and Q3 are turned on. Hence, both VVDD and substrate bias BP are 1V. On the other hand, in the sleep mode *SL* is high and $\overline{SL}$ is low and thus only Q4 turns on and BP becomes 3.3V. The operations of Fig. 4.6 and Fig. 4.5 are equivalent.

However, this scheme needs a voltage regulator or converter to transform 3.3V to 1V, if 1V is internally generated. The regulator or converter induces extra power and area overheads. Besides, the nodes VVDD and VGND are large capacitive nodes and they probably cost a great amount of time to charge and discharge. Therefore, the sizes of Q1-Q4, D1, and D2 are indispensably large to diminish charging and discharging time. The area overhead is hence significant and the extra power to charge and discharge the virtual source lines is another power overhead.

4.1.4 Dynamic-Vt SRAM

Fig. 4.7 shows a dynamic Vt SRAM to reduce subthreshold leakage current [4.4]. The two NMOS transistors serve as voltage switches to dynamically adjust the voltage of substrate in different operating modes. The substrate is switched to 0V in active mode for high performance, while it's switched to Vbs (a negative value) in sleep mode for saving leakage power.

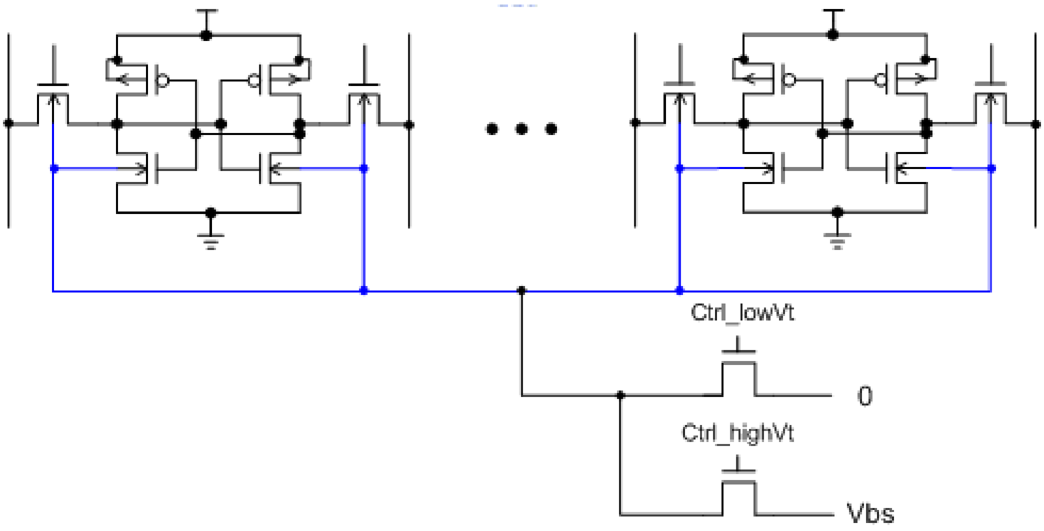


Fig. 4.7 Schematic of a dynamic Vt SRAM set.

A time-based capacitor-discharging scheme for Vt-control is shown in Fig. 4.8 [4.4]. The circuit consists of an RC decay circuit, a level converter, and V_{sub} switches. When the data line is accessed, V_{cap} is charged by WL and immediately switches V_{sub} to 0V. V_{cap} starts to discharge slowly as long as WL is pulled low, and it'd recharged whenever WL is accessed again. After a sufficient idle period, V_{cap} is low enough to switch V_{sub} to –1.0V. Fig. 4.9 depicts the operating waveforms for the nodes.

There are some questionable problems about the operation. First, Fig. 4.8 shows that the Vt control circuit needs 1.5V, 1.0V, and –1.0V three supply voltages. It' difficult and impractical to have so many external voltage sources, and some voltage converters or charge pump circuits are indispensable if only one external voltage source is available. However, no any voltage converters or charge pumps are mentioned in this scheme.

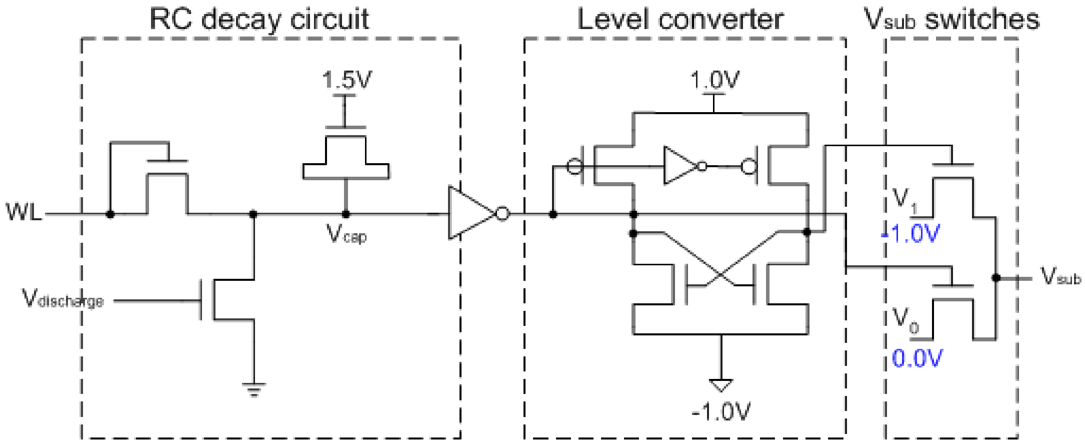


Fig. 4.8 Schematic of the Vt control circuit using capacitor-discharging scheme.

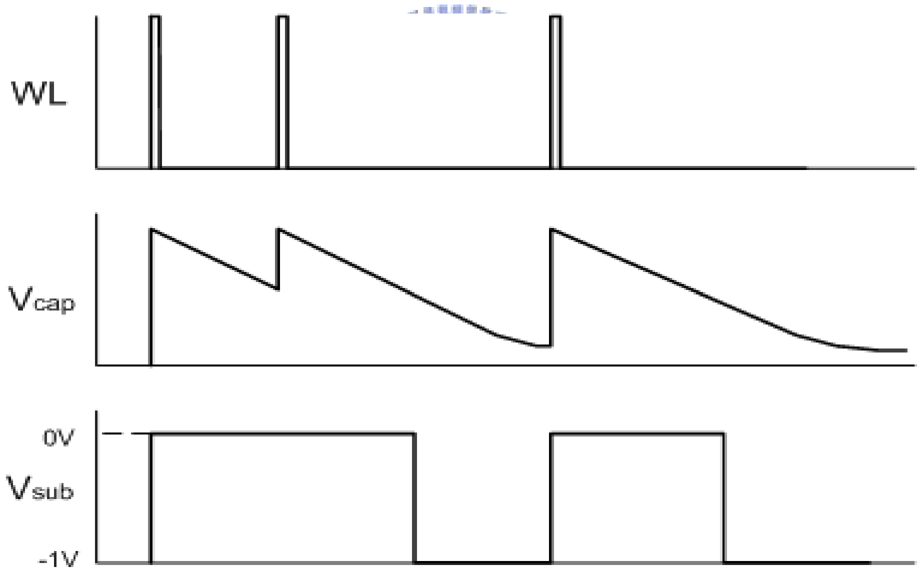


Fig. 4.9 Operating waveforms for Vt control circuit.

Second, the V_{sub} switches in Fig. 4.8 are not robust if V_1 is generated by a charge pump instead of an ideal external source. When the switch that is connected to V_1 turns on, voltage –1.0V passes to V_{sub} through the switch. Unfortunately, the charges at V_1 redistribute between V_1 and V_{sub} since V_1 is connected to a charge pump. Therefore, in the steady state V_{sub} and V_1 are both between –1.0V and 0V due to charge redistribution. Using an external voltage source V1 can solve this problem, but generally for logic chips, no negative supply voltages are available.

Finally, the operation waveforms in Fig. 4.9 do not concern about the loading effect. The substrate is a large capacitive load and it takes a lot of time to charge and discharge. Hence, the waveforms in Fig. 4.9 is too ideal and actual situations are much more complicated.

4.1.5 Forward Body-Biased SRAM

In this subsection, a forward body-biased (FBB) SRAM scheme is described. In contrast to the previous schemes, a FBB SRAM intends to achieve high-speed operation instead of suppressing standby leakage. However, this scheme uses super high Vt devices to reduce subthreshold leakage in both active and standby modes. The performance degradation due to super high Vt devices is diminished by forward biasing the body-source junctions.

Fig. 4.10 shows the schematic diagram of FBB SRAM scheme with body bias drivers M1-M3 [4.5]. The SUBSL signal is generated by the decoder circuit and each subarray has a dedicated SUBSL signal. When the subarray is accessed, the SUBSL is pulled high and the switches M1 and M2 and turned on. Therefore, the p-well of the selected subarray is charged to 0.5V, increasing the active current and achieving a fast operation. On the other hand, the p-well voltage of unselected subarrays is switched to 0V through M3.

Fig. 4.11 shows the operating waveforms of the control signals. The scheme uses extra decoder circuits to decode the most significant address bits, ensuring the SUBSL signal is pulled high before the wordline signal. As in Fig. 4.11, the SUBSL signal goes to high before the coming of the wordline signal, and V_{PWELL} is switched to 0.5V before the wordline arrives as well.

The operating waveforms in Fig. 4.11 seem so perfect but some problems exist. First, a voltage converter is necessary to generate 0.5V and this circuit induces power and area overhead. Next, due to the extra decoder circuits for generating SUBSL signal before the wordline, another power and area overhead is included. Finally, it seems so difficult to switch V_{PWELL} to the FBB level before the arrival of wordline signal. Fig. 4.10 shows that a subarray contains 1024 cells and the capacitance at V_{PWELL} probably exceeds the order of pico-farad. The time period between wordline and SUBSL is about the order of nano-second. Therefore, in comparison with the two parameters, correct operations of this scheme seem so questionable.

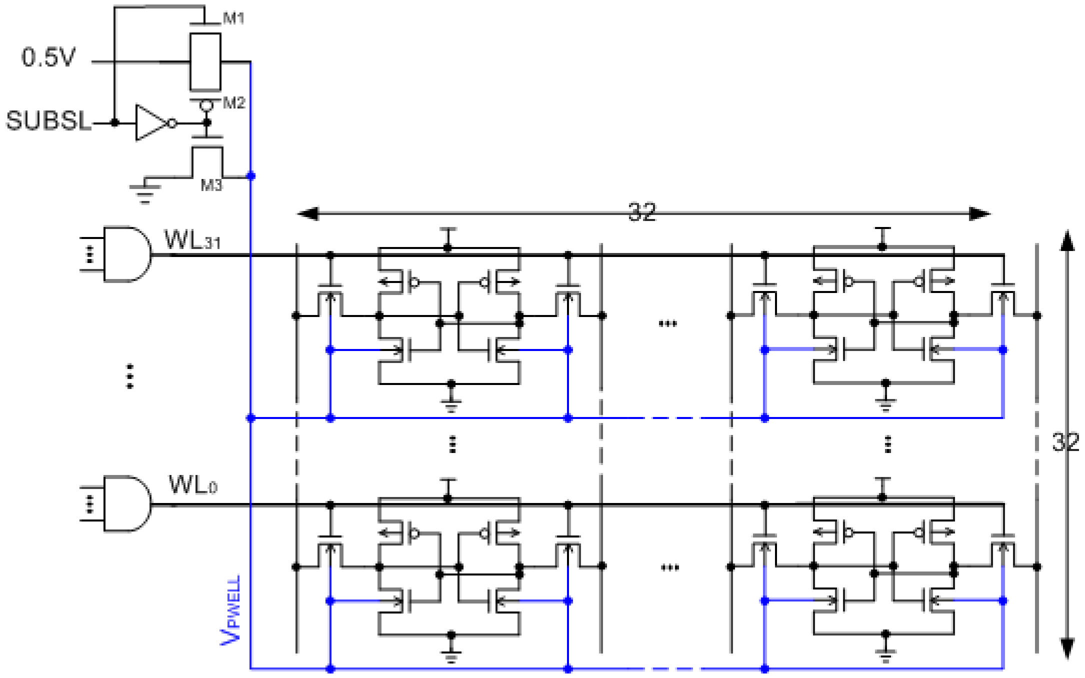


Fig. 4.10 Schematic diagram of forward body-biased SRAM.

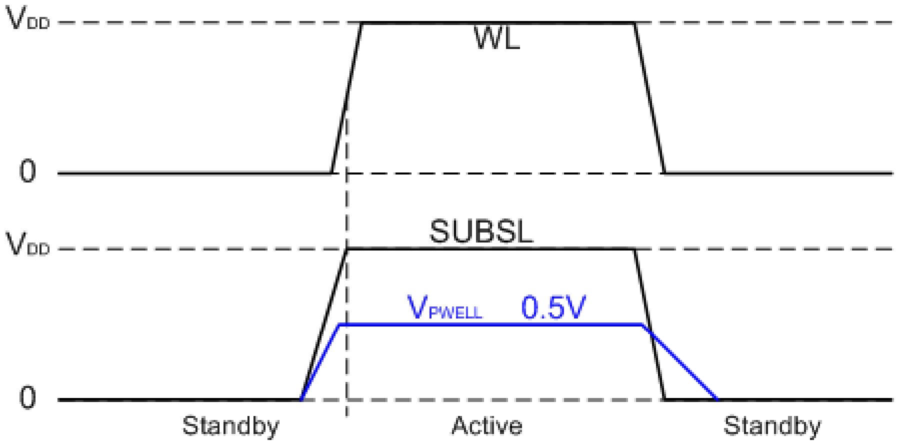


Fig. 4.11 Operating waveforms of FBB SRAM.

4.2 SRAM Cell Arrays With On-Chip V_{BB} Generators

Section 4.1 introduces several VTCMOS SRAM schemes that both RBB and FBB schemes are included. However, almost all of them use external voltage sources instead of on-chip voltage generators. For general digital circuits, it seems so impractical to have so many external voltage sources so that on-chip voltage generators are necessary. In this section, an on-chip dual-level body-bias generator is presented and applied to the design of SRAM cell arrays. The power overhead of on-chip voltage generator is also taken into account.

4.2.1 On-Chip Dual-Level V_{BB} Generator

Fig. 4.12 shows the schematic diagram of SRAM cells with the proposed V_{BB} generator, which comprises two substrate bias generators, two recovery circuits, and a high/low control circuit. The substrate bias generators have been introduced in Chapter 3, and the recovery circuits are used to return V_{BBN} and V_{BBP} to their original values. High/low control circuit controls input pumping signals and thus the output voltages.

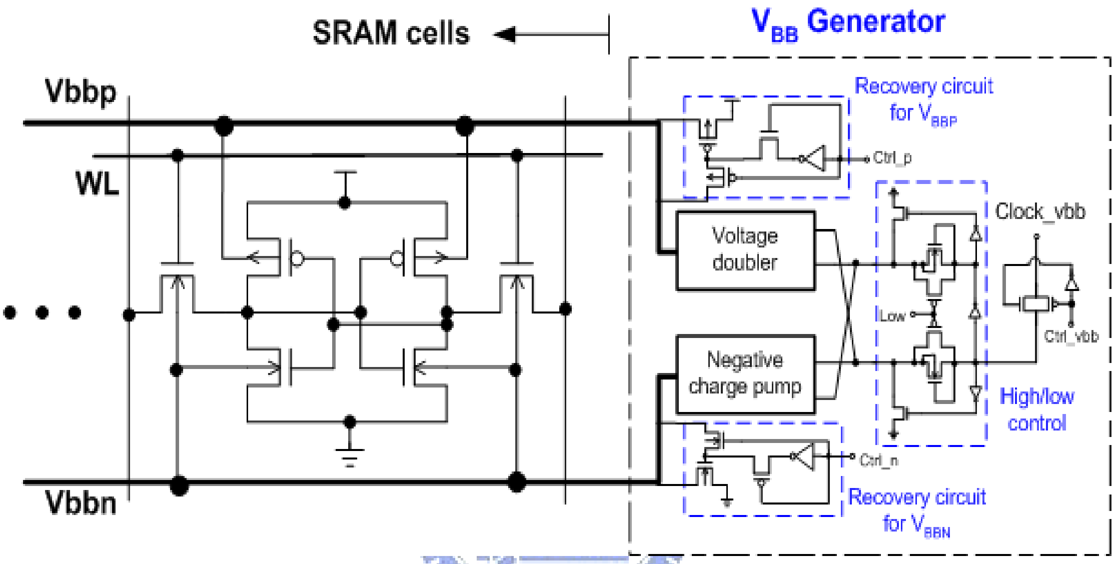


Fig. 4.12 Schematic diagram of SRAM cells with on-chip dual-level V_{BB} generator.

4.2.1.1 Substrate Bias Generator

There are two substrate bias generators that the voltage doubler is for V_{BBP} and the negative charge pump is for V_{BBN}. Please refer to Chapter 3 for detail schematics and operations.

4.2.1.2 High/Low Control

Fig. 4.13 shows the schematic and operating waveforms of high/low control circuit. A clocking signal is fed into the circuit and Low signal is used to control the swing of output signals, Vout0 and Vout1. When Low is pulled low, both the PMOS transistors are turned on and input clocking signals directly pass through the transistors. When Low signal is pulled high, the PMOS transistors are off and the NMOS are on. Consequently, the swing of output signals is smaller than input clocking signal by an amount of Vt. The operating waveforms clearly illustrate the operations.

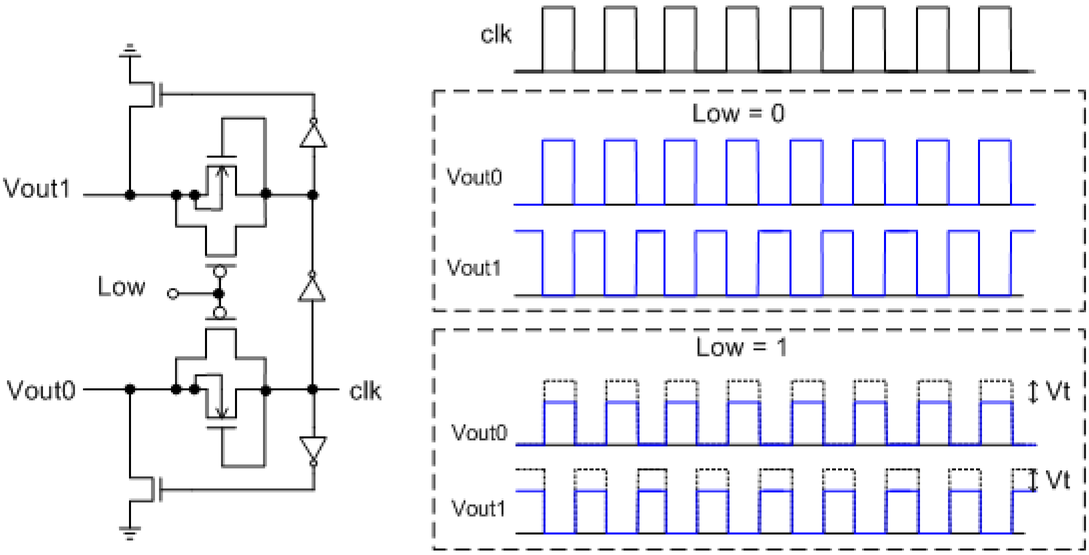
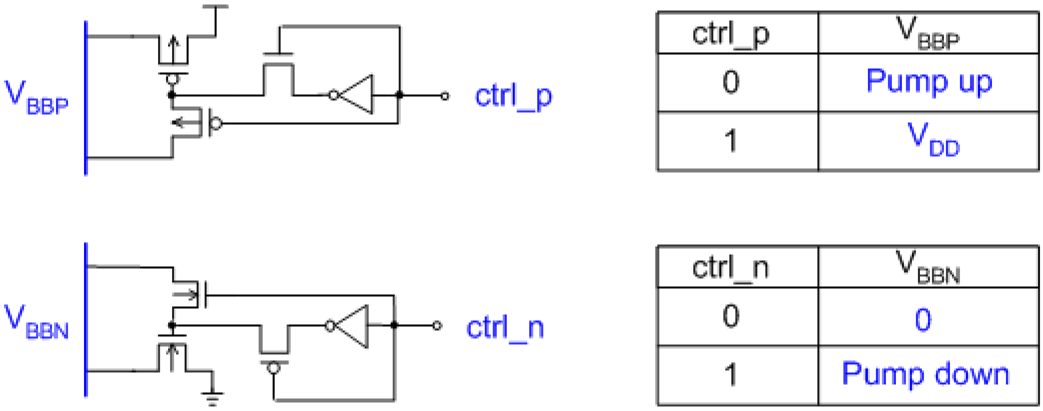


Fig. 4.13 Schematic of high/low control and operating waveforms.

4.2.1.3 Recovery Circuits

Reversed body-bias is applied to unselected rows to reduce subthreshold leakage. Once the rows are selected some mechanisms must be done to cancel the body-bias. Fig. 4.14 shows the recovery circuits for both V_{BBN} and V_{BBP} and their operating tables.



ctrl_p	V_{BBP}
0	Pump up
1	V_{DD}

ctrl_n	V_{BBN}
0	0
1	Pump down

Fig. 4.14 Recovery circuits for V_{BBN} and V_{BBP} and operation tables.

4.2.2 Simulation Results

The SRAM cell arrays with on-chip V_{BB} generators are simulated in TSMC 0.13um technology. Some power consumption and power saving information are discussed below.

4.2.2.1 Waveform of V_{BB} Generator

Fig. 4.15 shows the simulated waveforms of the V_{BB} generator for both high and low conditions. In high condition V_{BBN} reaches –1.15V and V_{BBP} reaches 2.35V, where V_{DD} is 1.2V. On the other hand, V_{BBN} reaches –0.92V and V_{BBP} reaches 2.12V in low condition. Note that the pumping frequency and output loading are 5KHz and 10pF, respectively.

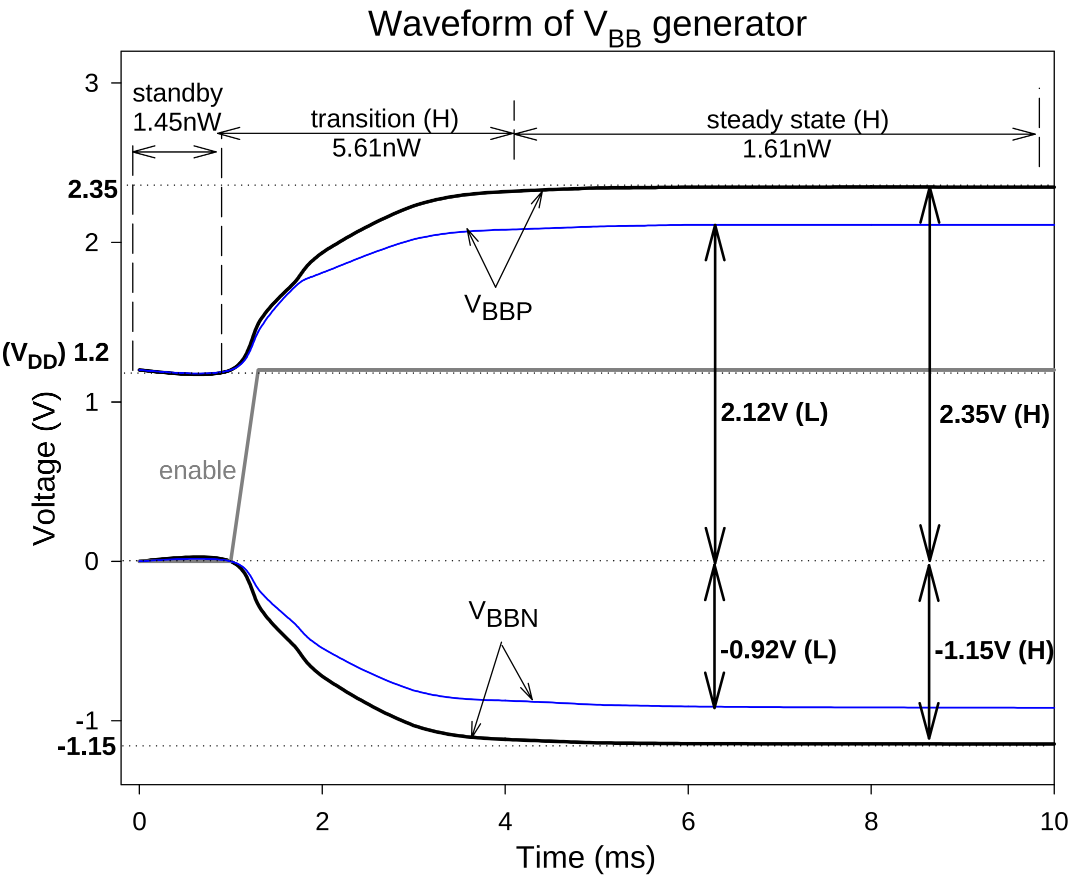


Fig. 4.15 Simulated waveforms of V_{BB} generator.

4.2.2.2 Average Power of V_{BB} Generator

Fig. 4.16 shows the average power of V_{BB} generator and it reveals that the average power converges with time. Fig. 4.16 clearly illustrates this feature that the power consumption in steady state is less than transition state. That is, for a row of SRAM with sufficient time period in standby, the average power overhead of V_{BB} generator converges to about 1.6nW. Therefore, the factor of time period in standby mode must be taken into account when evaluating the net power saving.

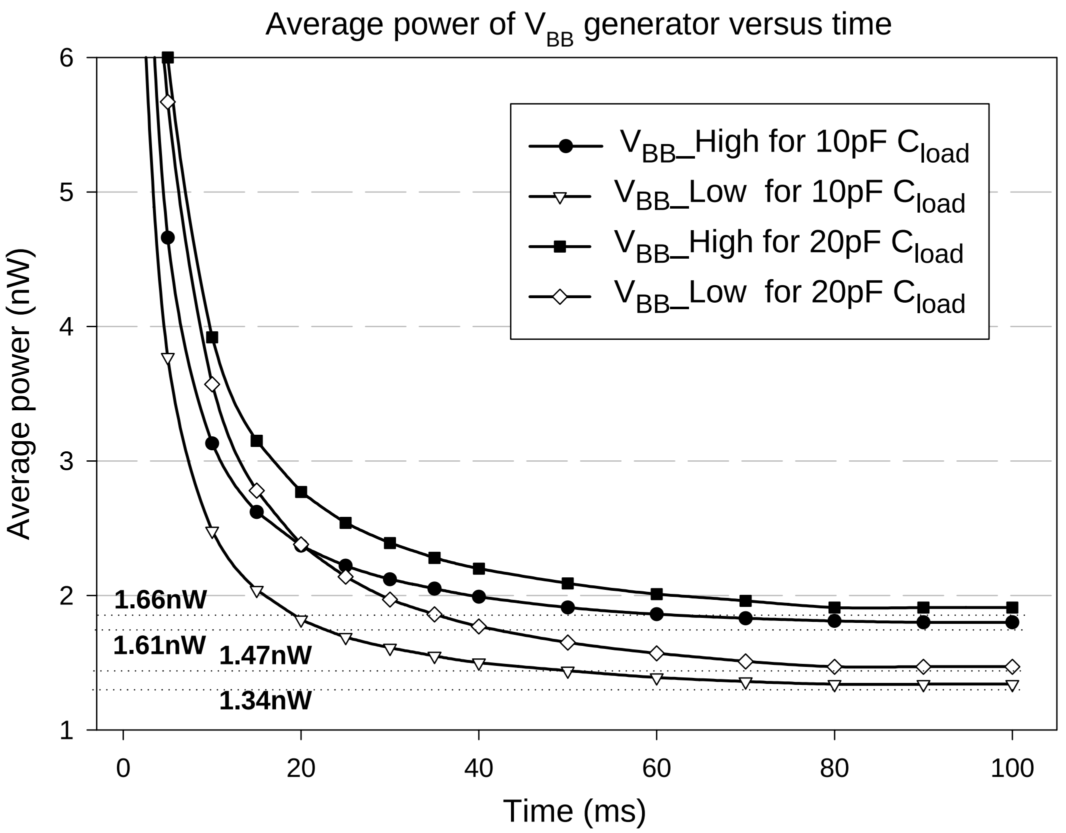


Fig. 4.16 Average power of V_{BB} generator versus time.

4.2.2.3 Net Power Saving of SRAM

Fig. 4.17 shows the net power saving of SRAM versus time period in standby mode. The net power saving is defined as the original SRAM leakage power minus the remaining part and the power overhead of V_{BB} generators. Fig. 4.17 illustrates that the net power saving increases with the increase of time. This is because the power overhead of V_{BB} generator decreases with time.

Fig. 4.17 also shows the curves of different wordline lengths. It can be seen that wider wordline lengths achieve larger net power savings and reach the break-even points in less time. Break-even point means the point that the saved leakage power is equivalent to the power overhead of V_{BB} generator.

Fig. 4.17 strongly proves the statement mentioned above that the power saving is time dependent. If the time period in standby is 3 milliseconds, for example, a 64-bit row obtains positive net power saving but negative ones are achieved for 32-bit and 16-bit rows. This means that for 32-bit and 16-bit rows, the saved leakage power within 3 milliseconds is not enough to compensate the power overhead.

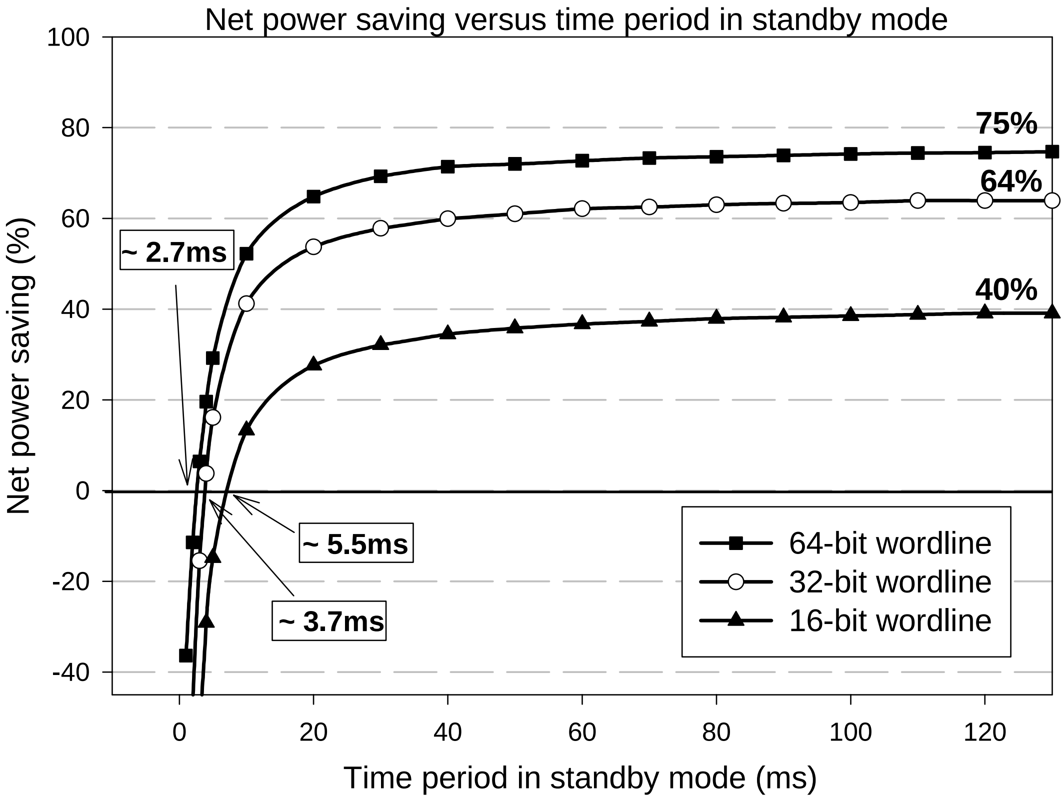


Fig. 4.17 Net power saving of SRAM versus time period in standby mode.

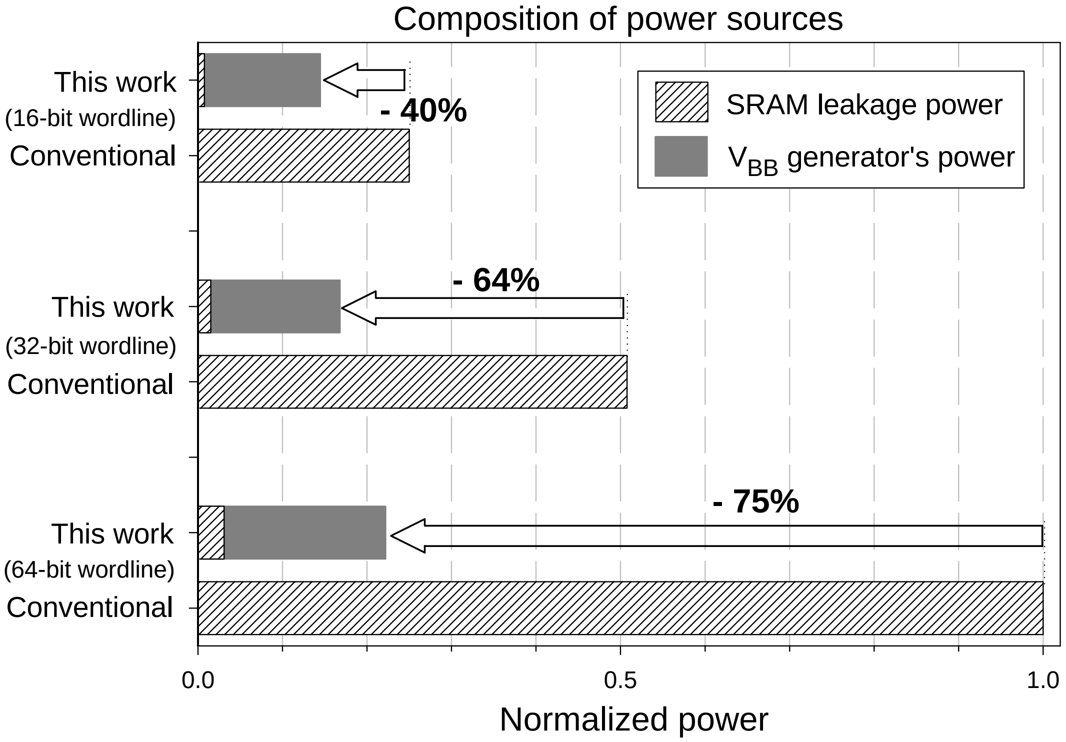


Fig. 4.18 Composition of power sources.

4.2.2.4 Composition of Power Sources

Fig. 4.18 depicts the composition of power sources of a wordline with V_{BB} generators. The leakage power of SRAM is proportional to wordline length and the power consumption of V_{BB} generator increases slightly with the increase of wordline length. Wider wordlines save much more leakage power and with a small fraction of increased overhead. Therefore, more net power saving is achieved when wordline length increases. Fig. 4.19 further shows that the fraction of power overhead is getting relatively smaller with the increase of wordline length.

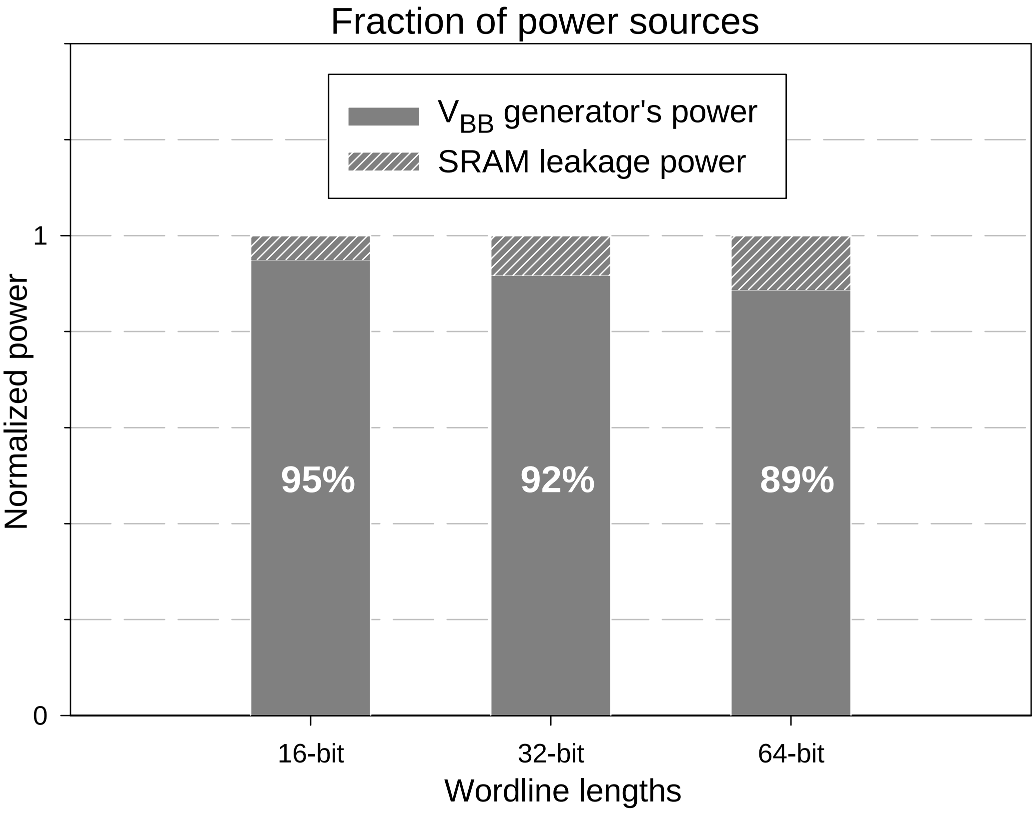


Fig. 4.19 Fraction of power overhead for different wordline lengths.

4.2.2.5 RBB V_{BBN} or V_{BBP} Alone

Fig. 4.20 and 4.21 show the power information of using RBB V_{BBN} or V_{BBP} alone for a 32-bit wordline. The net power saving of applying both V_{BBN} and V_{BBP} is about 64%, while the net power saving of applying V_{BBN} alone is about 64.5%. This result demonstrates that V_{BBP} generator has less significant effect on leakage saving. The information of applying V_{BBP} alone shown in Fig. 4.20 and Fig. 4.21 supports this result.

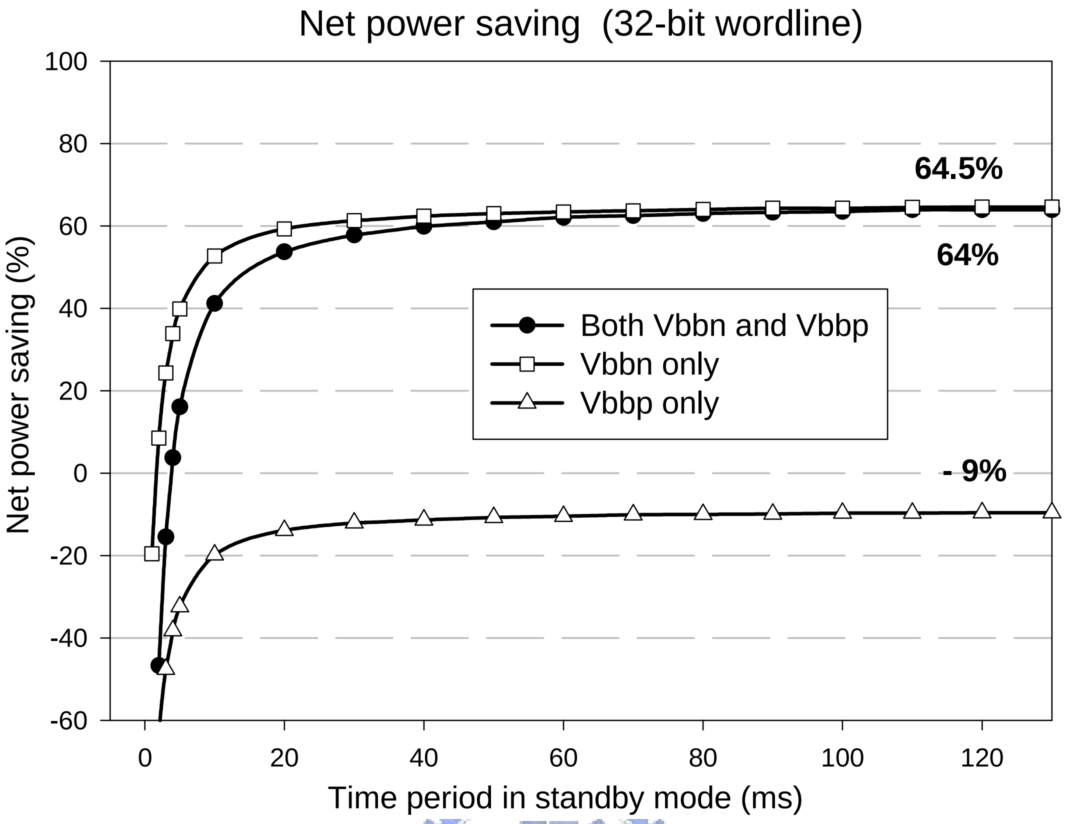


Fig. 4.20 Effectiveness of net power saving for V_{BBN} or V_{BBP} alone.

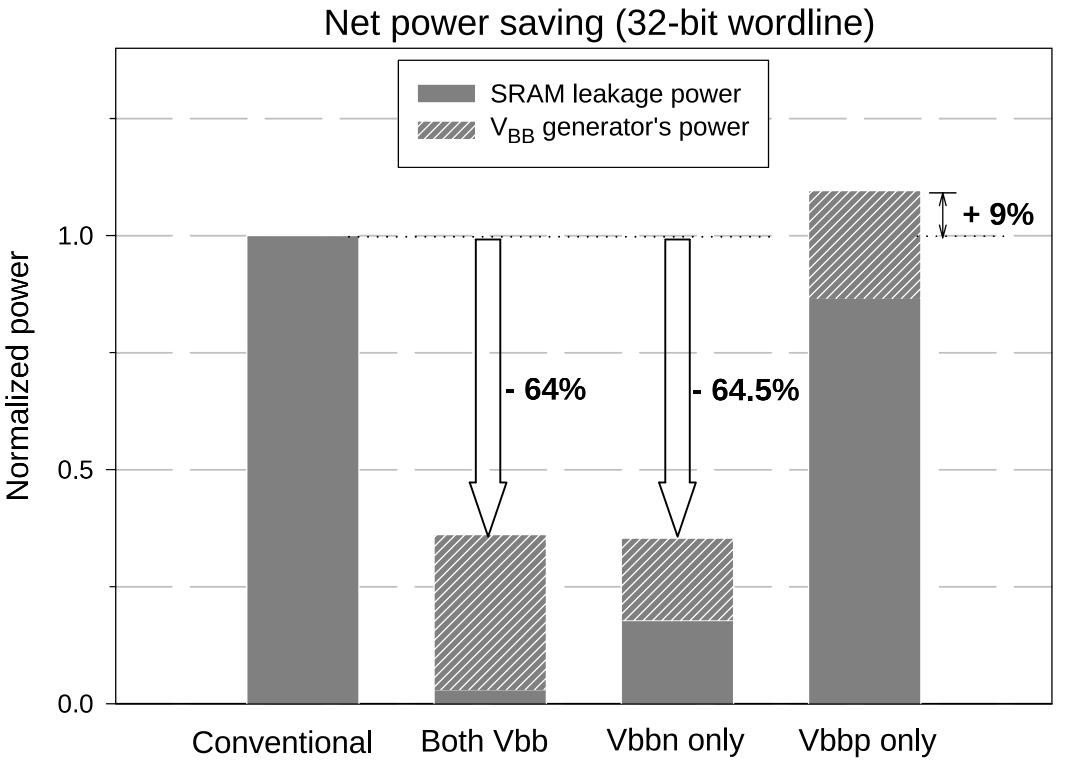


Fig. 4.21 Power information for V_{BBN} or V_{BBP} alone.

Fig. 4.20 and Fig. 4.21 show that no positive net power saving is possible if RBB V_{BBP} is applied alone, due to the remaining leakage power plus power overhead exceed the nominal leakage power. In comparison with the conditions of using both V_{BB} generators and V_{BBN} generator alone, the prior condition obtains more leakage power saving but more power overhead induced by V_{BB} generators. Therefore, it’s another solution to apply RBB V_{BBN} alone and the net power saving is slightly larger.

4.2.3 Triple-Well Layout for SRAM Cells and V_{BB} Generator

Fig. 4.22 shows the layout of conventional and triple-well SRAM cells, and the cross-sectional views depict the difference. Triple-well structure uses an n-well ring and deep n-well to form a p-well region, which serves as the substrate of NMOS transistors. Voltages can be easily applied to p-well and n-well through well contacts, as shown in Fig. 4.22. Fig. 4.23 further shows the layout and configuration of two 64-bit rows with a V_{BB} generator. The two p-well regions for NMOS are connected together through well contacts and supplied by V_{BB} generator. Likewise, the two n-well regions for PMOS are connected together through well contacts and supplied by V_{BB} generator.

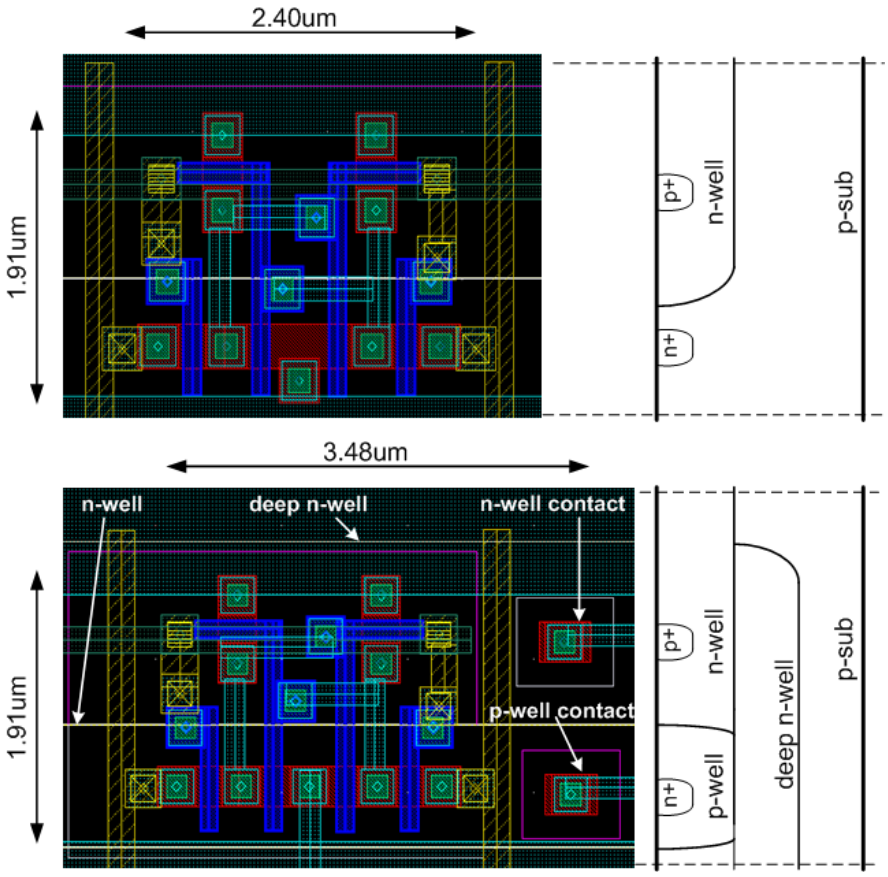


Fig. 4.22 Layout of conventional and triple-well SRAM cells.

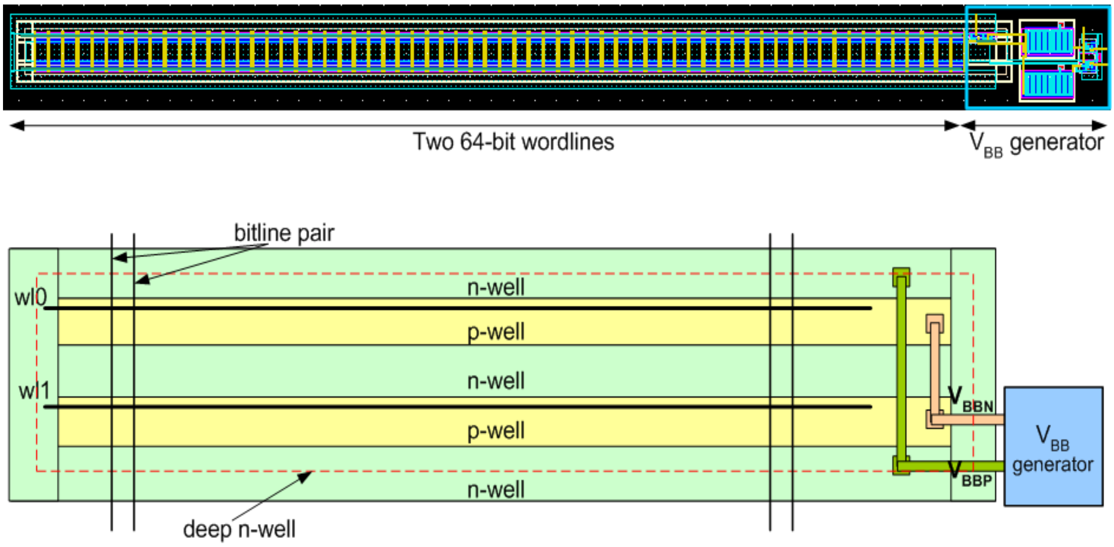


Fig. 4.23 Layout and configuration of triple-well SRAM rows with a V_{BB} generator.

4.3 Time-Out-Policy V_{BB} Generator Controller

In this section a V_{BB} generator controller is presented. It adopts the concept of time-out policy to determine the operation of V_{BB} generator. Time-out policy is a commonly used technique for controlling operating modes in software. Here, a hardware and circuit-level implementation is realized to control one wordline (row) in SRAM.Moreover, a data-retention latch is included so that the supply voltage of most the circuits can be shut down for more power saving.

4.3.1 Schematic Diagram

Fig. 4.24 shows the schematic diagram of V_{BB} generator controller, which mainly comprises a pulse generator, a state machine, and a data-retention latch. The pulse generator detects transitions of clk_vbb and produces pulse signals, and these pulse signals are fed into the state machine. The state machine controls the operating modes of V_{BB} generator according to the pulse signals and WL (wordline signal). The data-retention latch is used to retain the output control signal. Once the signal Enable_Vbb is pulled high, the power supply of pulse generator and state machine is gated to save standby power.

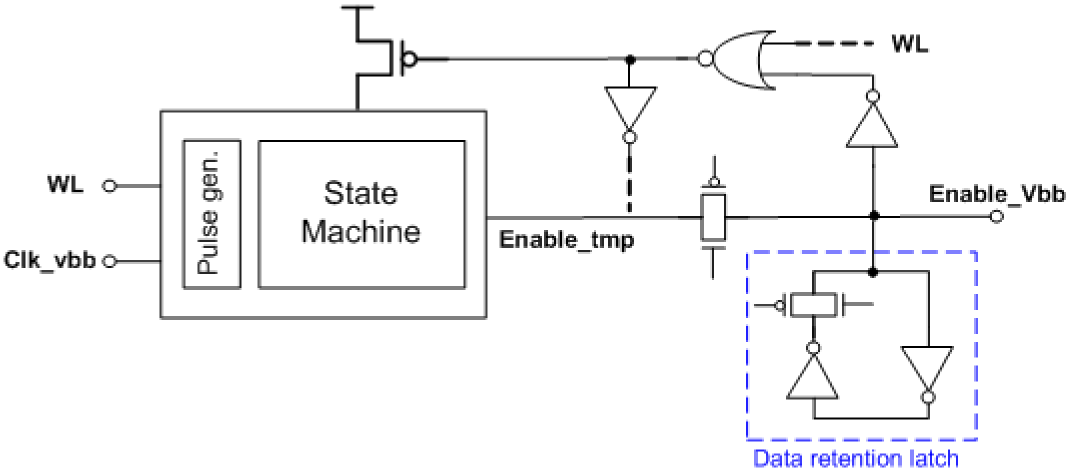


Fig. 4.24 Schematic diagram of V_{BB} generator controller.

4.3.2 Pulse Generator

The pulse generator comprises a delay element and an exclusive-OR (XOR) gate, as shown in Fig. 4.25. A pulse signal is generated once a transition of clk_vbb is detected. Fig. 4.25 also illustrates the operating waveforms.

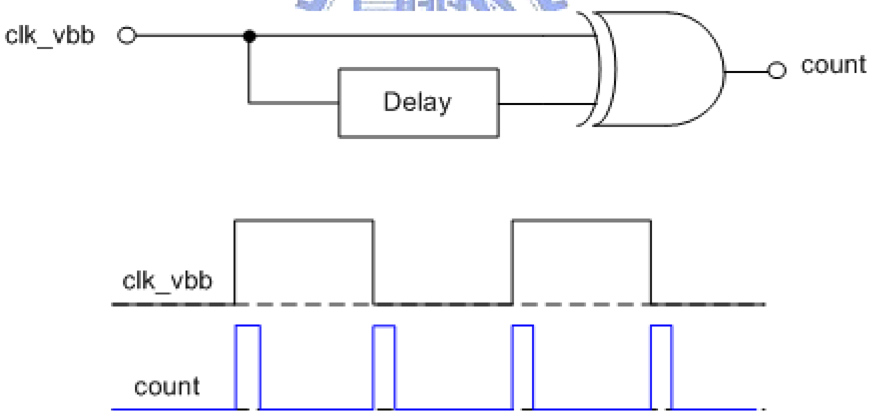


Fig. 4.25 Pulse generator and operating waveforms.

4.3.3 State Machine

Fig. 4.26 shows the schematic of state machine that consists of two flip-flops. The flip-flops are triggered by pulse signal from pulse generator and wordline signal from address decoder. Besides, the flip-flop is reset to zero whenever the wordline signal WL is high (WLb is low). Fig. 4.27 shows the details of the flip-flop.

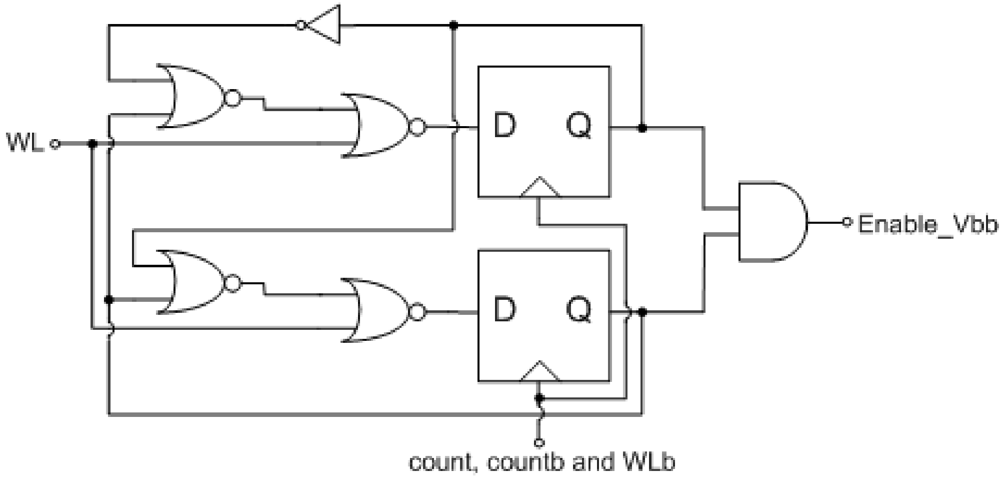


Fig. 4.26 Schematic diagram of state machine.

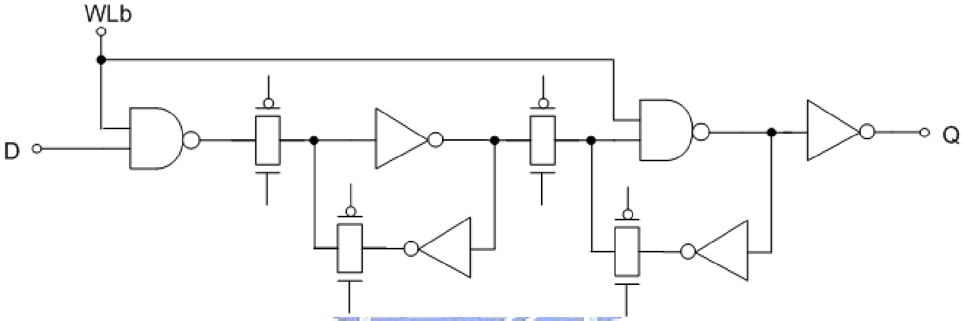


Fig. 4.27 Schematic of flip-flop with reset signal WLb.

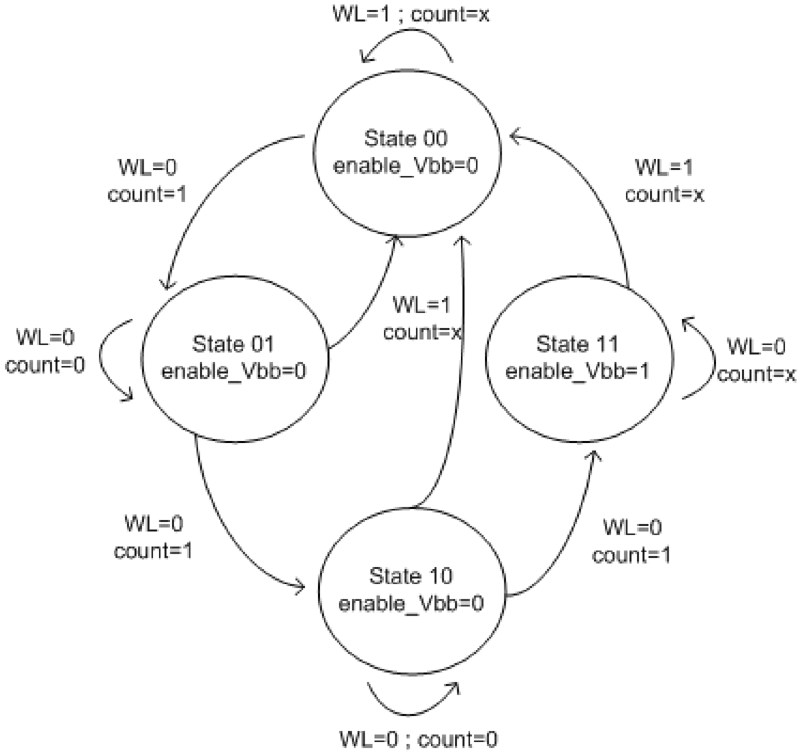


Fig. 4.28 State graph of state machine.

Fig. 4.28 illustrates the state graph of state machine. The state registers (flip-flops) remain at the state “00” when WL keeps high. When WL is pulled low the state registers count upward when count signal is triggered. The state registers go back to “00” once WL goes high regardless of count signal. If the state reaches “11”, the output signal Enable_Vbb goes high and enables V_{BB} generator to produce reverse-biased V_{BBN} and V_{BBP}.

4.3.4 Time-Out Value

Time-out value is the time period between the last “High” of WL and the time V_{BB} generator starts to function. Fig. 4.29 shows the two extreme conditions for defining the time-out value, and it depends on the position of clk_vbb when the last “High” WL occurs. Fig. 4.29 (a) shows the first condition that a pulse comes right after the falling edge of the last “High” WL. The time-out value is **one** cycle time of clk_vbb in this condition. Alternately, the other extreme condition occurs when the last “High” WL and one pulse happen concurrently, as shown in Fig. 4.29 (b). Therefore, the time-out value is **one and half** cycle time of clk_vbb. If clk_vbb is 5KHz, for example, the time-out value is between 200us and 300us.

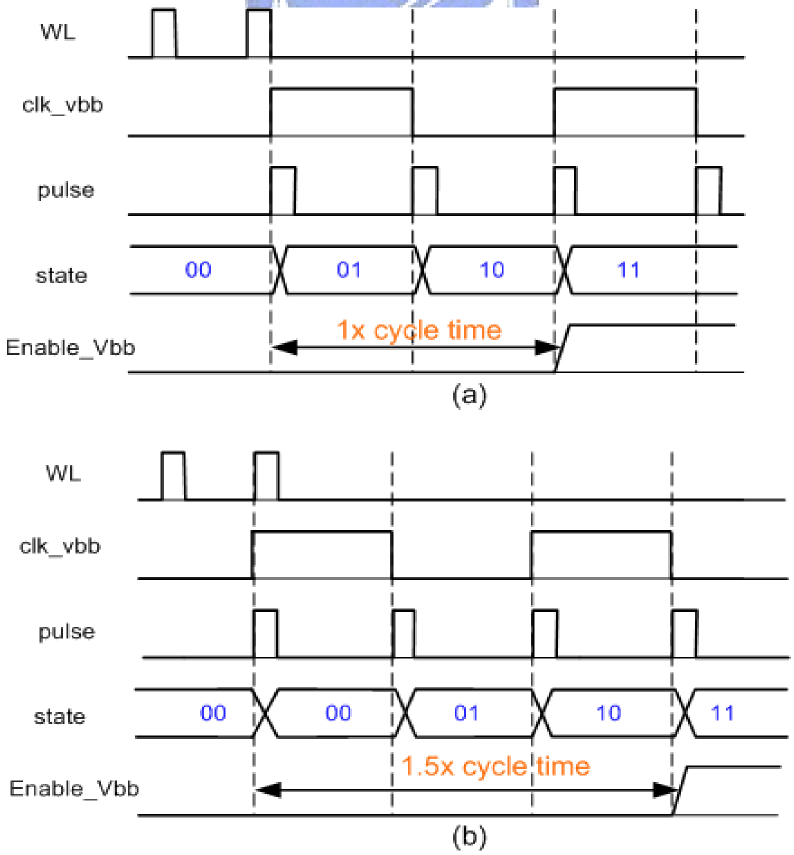


Fig. 4.29 Two extreme conditions for defining the time-out value.

4.3.5 Simulated Waveforms

Fig. 4.30 shows the simulated waveform of V_{BB} generator with power-down scheme. When the state registers reach “11” then the power supply of the state machine is turned off. However, the power supply of the data-retention latch is still on so that it can retain output signal.

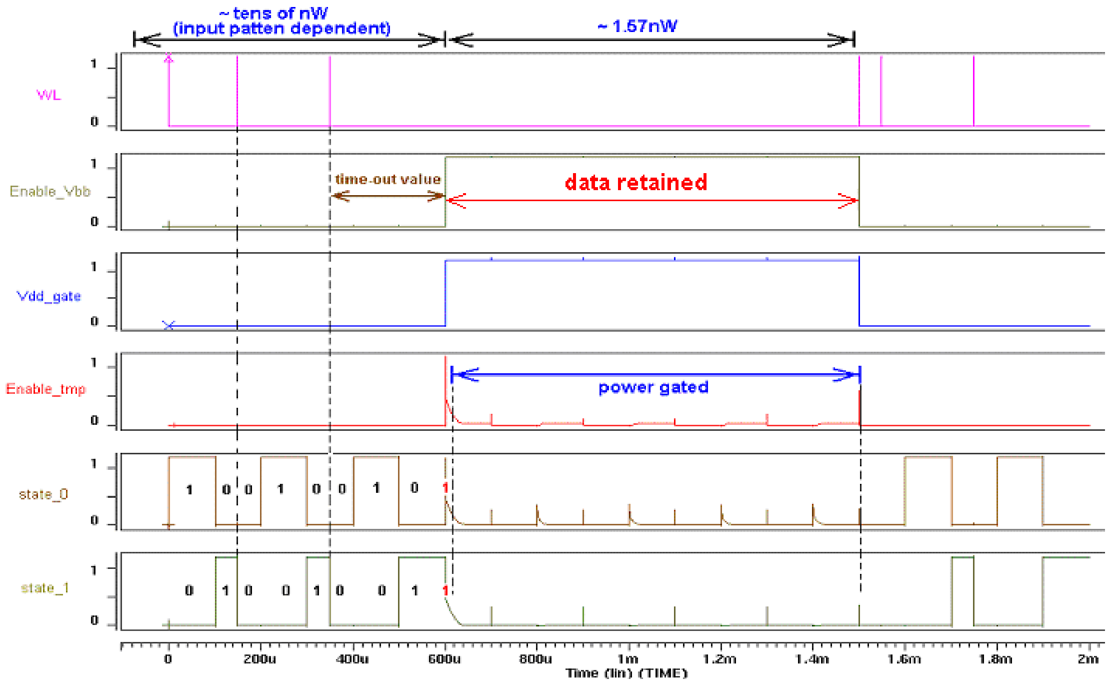


Fig. 4.30 Simulated waveforms of V_{BB} generator controller.

4.3.6 Power Reduction Due to Power-Gating

The controller without power-gating scheme consumes tens of nano-Watt active and standby power. The term ‘active’ means that the state machine is evaluating and ‘standby’ means the output signal is kept high.

Fig. 4.31 shows the comparison of power consumption between with and without power-gating scheme. Although some extra gates are added for power-down control, they are inactive except in standby mode and therefore consume almost no energy in active mode (no active power overhead). Eventually, the standby power is reduced to 6% if power-gating mechanism is adopted.

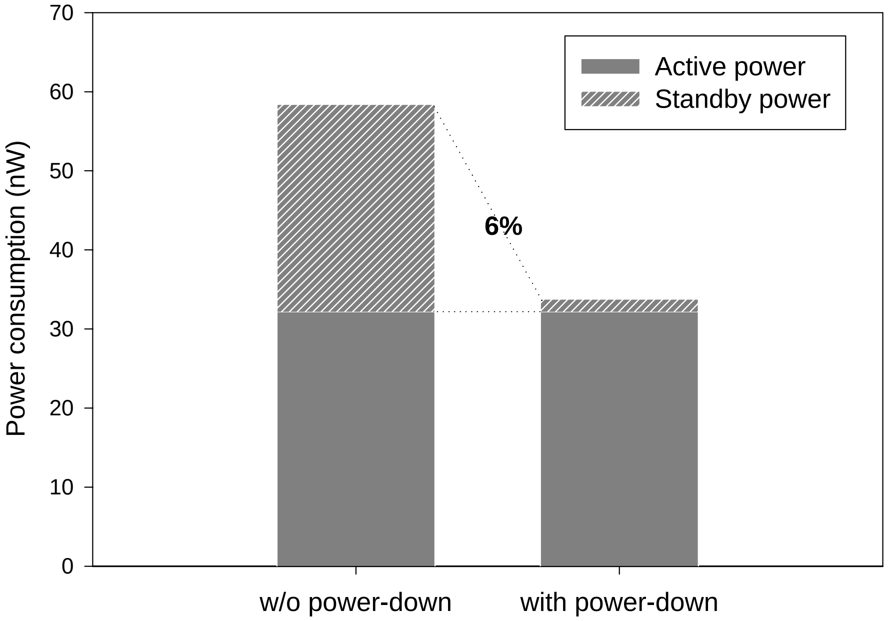


Fig. 4.31 Power comparison between with and without power-down scheme.

4.4 Conclusion

In this section, some VTCMOS SRAM designs are presented. These designs dynamically adjust the threshold voltage of transistors whether to achieve low standby power or high performance.

Besides, a VTCMOS SRAM scheme with on-chip V_{BB} generators is proposed and discussed. The V_{BB} generator generates two voltage levels and consumes nano-Watt order of power consumption. Simulation results show that this scheme can significantly reduce standby power of SRAM. Since the V_{BB} generator consumes insignificant power, a great amount of net power saving is obtained. Simulation results show that about 75% net power saving is achieved for 64-bit wordline and 64% for 32-bit wordline. These results show that a significant power saving is achieved even the power overhead of V_{BB} generator is included.

A time-out-policy controller for V_{BB} generator that adopts a data-retention latch and power-gating mechanism is also presented. Once the output of the controller enables V_{BB} generator, the most part of the controller is power-gated and about 94% power saving is achieved.

Chapter 5

Power-Gating Technique In Ultra-Low Power SRAM Cell Array Design

Power gating is a popular low-power technique to reduce leakage current in standby mode and it has been widely used in logic circuits [5.1]. However, several SRAM architectures adopting power gating are proposed in recent years.

In Sec. 5.1 the principles of stacking effect is described. Design issues of power-gated (or gated-V_{DD}) SRAM cells are discussed in Sec. 5.2, 5.3, and 5.4. Two power-gated SRAM architectures, column-controlled and row-controlled schemes are introduced in Sec. 5.5. In Sec. 5.6, a column/row co-controlled scheme is realized and comparisons with other schemes are shown in Sec. 5.7. Furthermore, the layout is implemented in 0.13um CMOS technology and area comparison is done. Finally, some conclusions and discussions are addressed in Sec. 5.8.

5.1 Stacking Effect

It has been observed that the stacking of two off transistors has significantly smaller subthreshold leakage current than one off transistor [5.2], [5.3]. This is called stacking effect and it is due to self-reverse biasing of stacked transistors.

5.1.1 Self-Reverse Biasing

Fig. 5.1 explains the phenomenon of self-reverse biasing. On the left is an off NMOS transistor with leakage current I_1, which is mainly composed of subthreshold leakage. On the right are two stacked off NMOS transistors and the leakage current is I_2. In the steady state, the voltage at node Vx is slightly higher than ground and thus transistor M_{21} has a negative V_{gs} (gate-to-source voltage) to make the pn junction reversely biased. Therefore, leakage current I_2 is smaller than I_1 due to the reversed bias of transistor M_{21}.

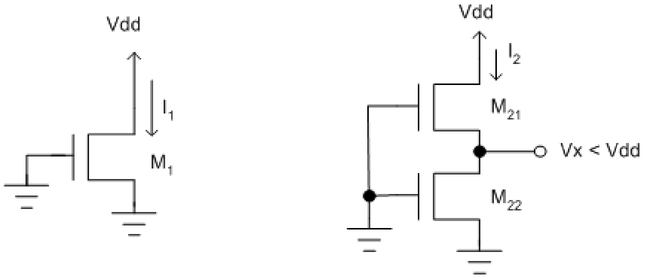


Fig. 5.1 Stacking effect due to self-reverse biasing of transistor M_{21}.

5.1.2 Tradeoff Between Delay and Leakage

As mentioned above, two staked off transistors have smaller subthreshold current than one off transistor. However, due to the stacked devices the drive current is smaller and results in increased delay. Fig. 5.2 shows the circuits used to observe the tradeoff of delay and leakage. In the middle of Fig. 5.2 is a normal inverter that the channel widths for NMOS and PMOS are *W* and 2 *W*, respectively. On the rightmost shows a modified inverter that the NMOS is split into two half-sized NMOS transistors whose channel widths are *W/2*. Fig. 5.3 is the simulated result that shows the delay-leakage tradeoff, and it clearly shows that smaller leakage current with larger delay. Therefore, paths that are faster than required can adopt this effect to slow down and reduce leakage current.

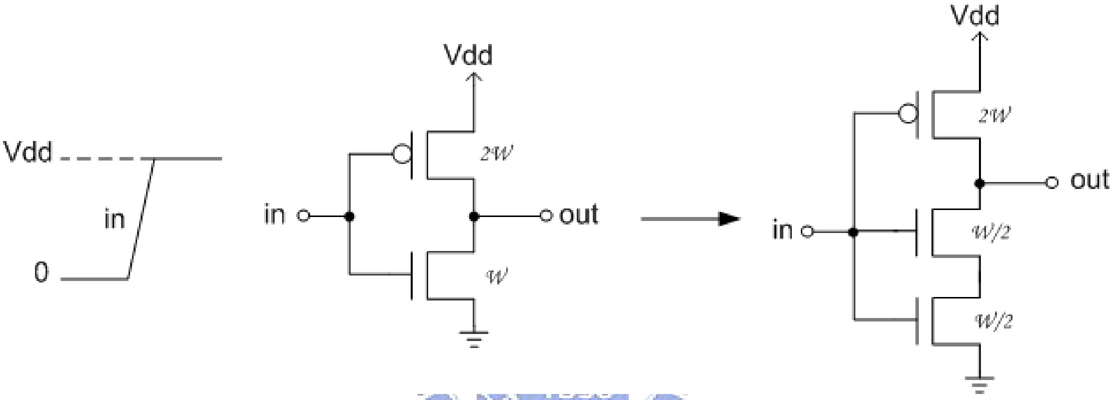


Fig. 5.2 Using a stacked inverter to observe the tradeoff of delay and leakage current.

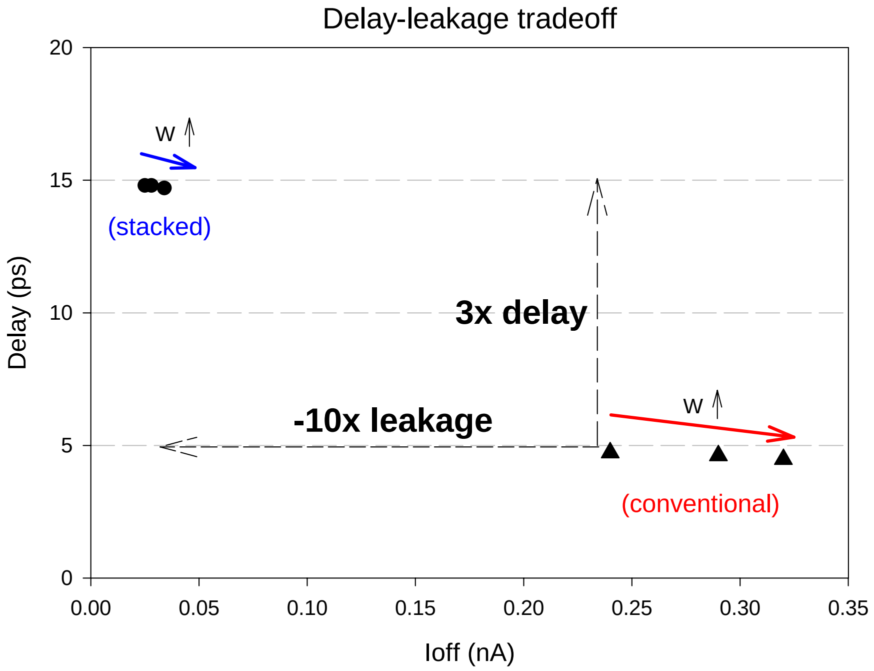


Fig. 5.3 Delay-leakage tradeoff of stacking effect.

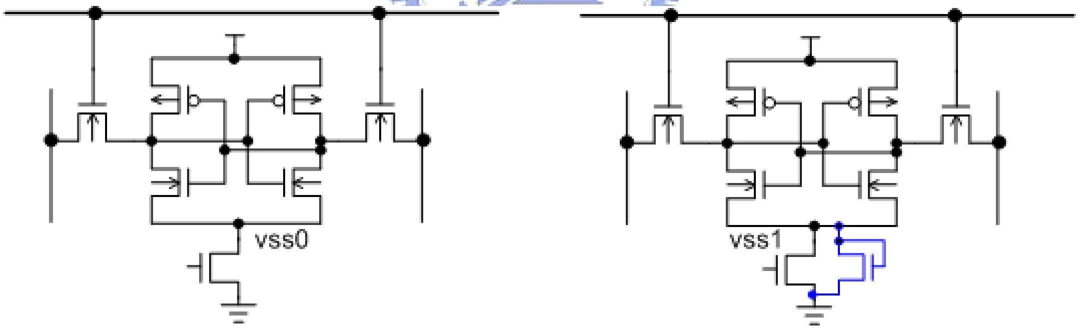
5.2 Gated-V_{DD} SRAM Cell

Gated-V_{DD} SRAM is just similar to power gating technique used in logic circuits. Adding a PMOS transistor between virtual V_{DD} and actual V_{DD}, or adding a NMOS transistor between virtual GND and actual GND can reduce standby leakage.

5.2.1 Virtual GND Node Fluctuation

However, the situations of SRAM and normal logic circuits are different because the internally stored data in SRAM will disappear gradually after the gating transistor is turned off. Fig. 5.4 (a) shows a SRAM cell with a gating device between virtual GND and actual GND. The virtual node vss0 is floating and charged by cell leakage current when the gating device is turned off. Therefore, cell leakage current charges vss0 and forces the potential to increase. Fig. 5.5 is the simulated result shows that vss0 probably exceeds 1/2 V_{DD} (600mV for 1.2V V_{DD}) and thus influences the stored data.

The voltage at virtual GND node can be limited to a small value by adding a diode-connected NMOS transistor between virtual GND node vss1 and actual GND node [5.4], as in Fig. 5.4 (b). Fig. 5.5 shows that vss1 is limited to about 100mV.



(a) (b)

Fig. 5.4 Gated-V_{DD} SRAM cells (a) without diode and (b) with diode.

5.2.2 Stability/Static Noise Margin

The stability of data in SRAM or register file is a critical factor for satisfied yield and low cost. Static noise is DC disturbance such as mismatches and offsets due to processing and variations in operating conditions. The static noise margin (SNM) is the maximum amount of DC disturbances that SRAM cell can tolerate that the stored data is not flipped [5.5]. Fig. 5.6 (a) shows a latch that comprises two inverters and two static noise sources, and Fig. 5.6 (b) shows the graphical view of SNM.

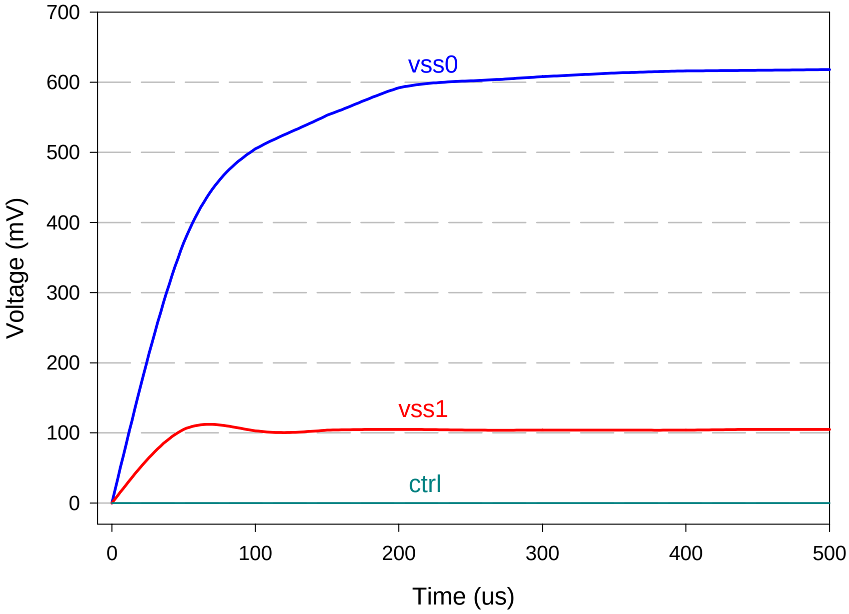
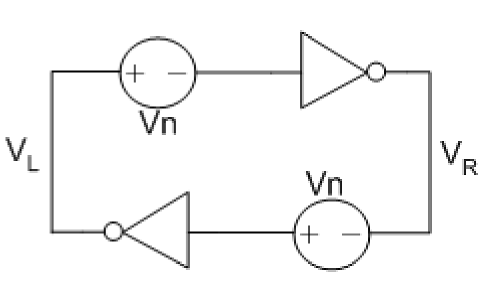
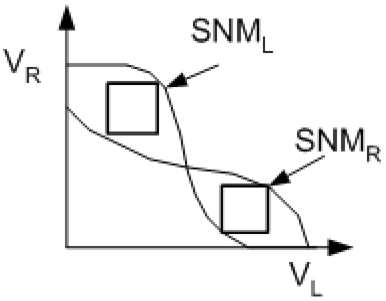


Fig. 5.5 Voltage of virtual GND increases after turning off gating device.



(a)



(b)

Fig. 5.6 (a) A latch with static noise sources and (b) Static noise margin.

5.2.3 SNM Issue of Gated-V_{DD} SRAM Cell

Fig. 5.7 shows the circuit used to discuss the SNM of gated-V_{DD} SRAM cell. The two inverters are connected to a gating transistor and a diode-connected transistor, and both the two transistors have channel width “*Wg*. The channel widths of NMOS and PMOS of the inverters are *n*Wg* and *2*n*Wg*, respectively. The term *n* is a positive number and we can observe the behavior of SNM by scaling the value of *n*. If *n* is 1, for example, it means that the gating device is the same size as the NMOS transistors of the inverters. Fig. 5.8 is the simulated result that shows SNM is a function of scale factor *n*, and the optimum value of *n* is about 2. This means that the maximum SNM occurs when the width of gating device is about half of the NMOS transistors of inverters.

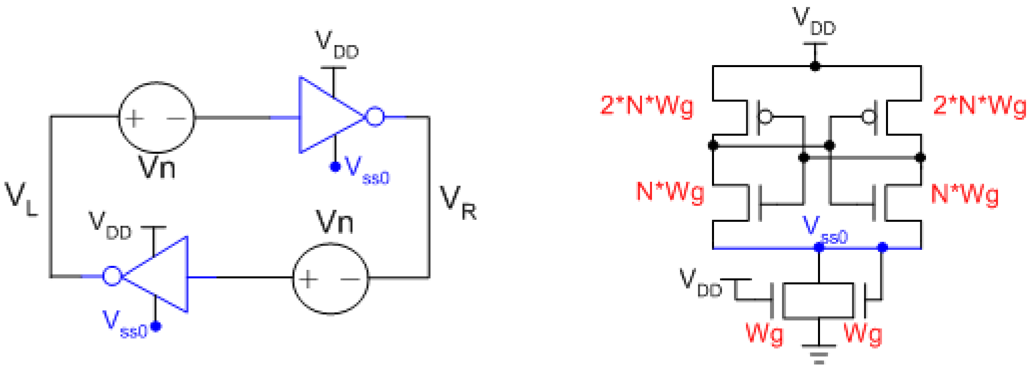


Fig. 5.7 Circuit used to observe the SNM of gated-V_{DD} SRAM cell.

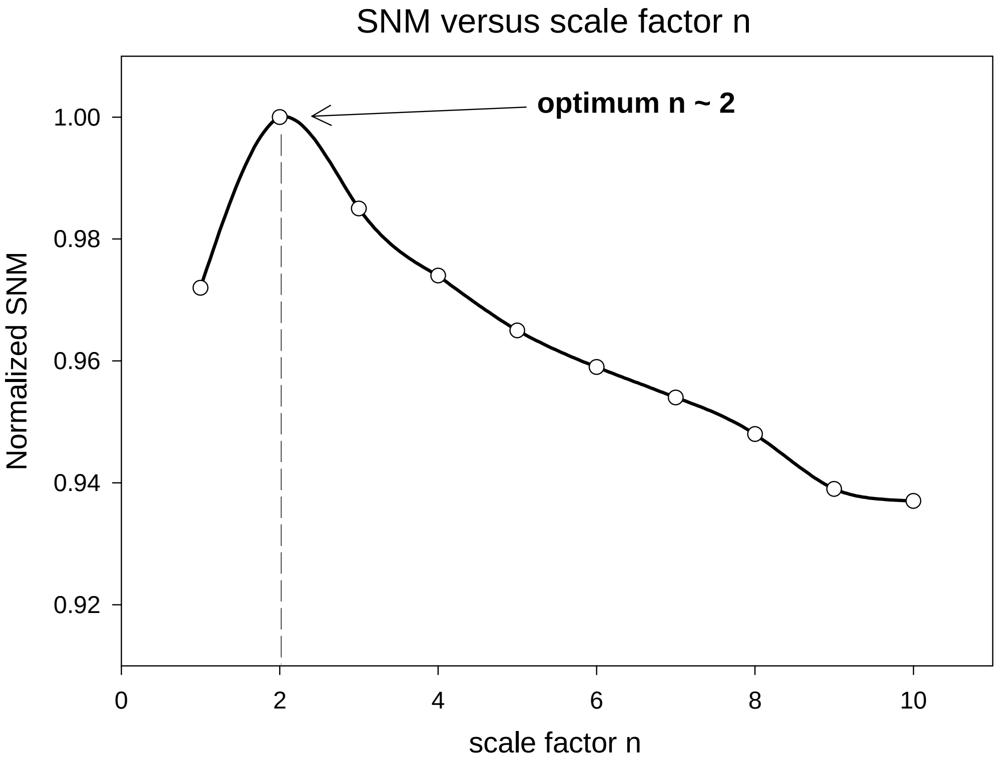


Fig. 5.8 SNM versus scale factor *n*.

5.3 Gate Leakage Reduction

As mentioned above, using an additional stacking device can reduce subthreshold leakage because of stacking effect. Moreover, gate leakage current decreases as well due to the positive voltage of virtual GND node. Gate leakage current increases exponentially with decrease in oxide thickness and increase in voltage across oxide [5.6], and it is shown that gate leakage through PMOS is smaller than NMOS. Fig. 5.9 shows the components of gate leakage in a gated-V_{DD} SRAM cell while in standby mode. Since gate leakage is a function of voltage across oxide, the gate leakage currents through M5 and M6 depend on bitline precharged voltage. Therefore, the dominant component is through M1 because it includes gate-to-source, gate-to-drain, and gate-to-substrate gate leakage currents.

After adopting a gating device and a diode-connected transistor between virtual GND node and actual GND, most of the gate leakage currents mentioned above are reduced because of the rising of virtual GND node. Note that the voltage at node storing '0' equals to the virtual GND node. It's clearly that the gate leakage currents through M1, M2, M3, and M4 can be reduced when the voltage of virtual GND node rises to a positive value. As shown in Fig. 5.9, the dotted lines represent the extra leakage currents induced by the two additional devices. However, these two components are negligible since both the V_{DS} and V_{GS} voltages of them are small.

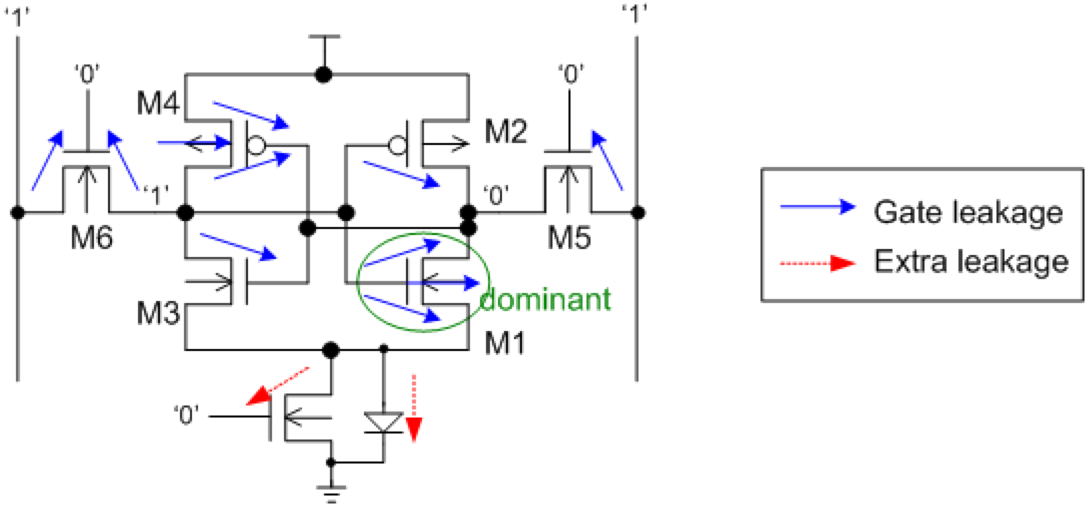
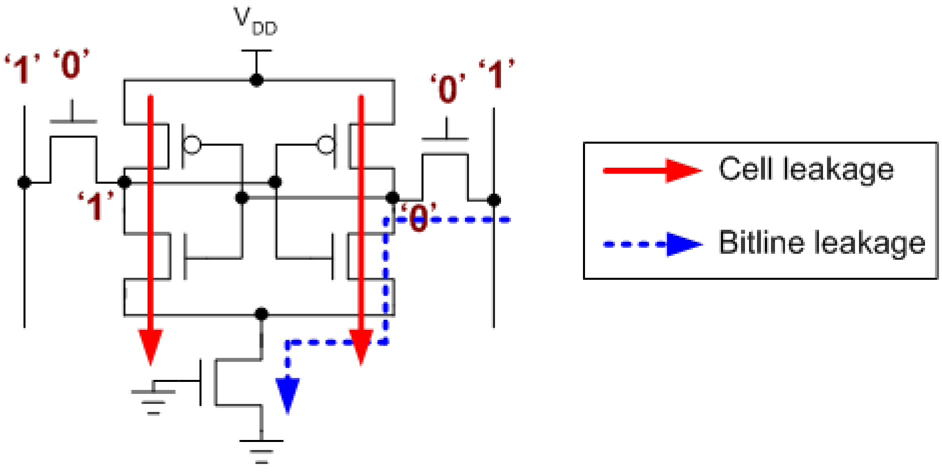


Fig. 5.9 Gate leakage components and extra leakage currents in gated-V_{DD} SRAM cell.

5.4 NMOS or PMOS Gating Device

Fig. 5.10 depicts the dominant leakage sources of an inactive SRAM cell. The bitlines are precharged to V_{DD} and the gating device is turned off. The solid lines index the subthreshold leakage paths and the dotted line represents the bitline leakage path. Using NMOS or PMOS gating devices can reduce the subthreshold leakage currents. But a PMOS gating device, however, does not create the isolation between bitlines and the ground as an NMOS gating device does. Therefore, using an NMOS gating device can save more standby power as a PMOS device can do, since an NMOS gating device isolates the bitline leakage path.



$$I_{off} = I_{sub\_cell} + I_{gate\_cell} + I_{bitline}$$
$$I_{sub\_cell} \sim I_{sub\_M1} + I_{sub\_M4}$$
$$I_{gate} \sim I_{gate\_M1}$$
$$I_{bitline} \sim I_{DIBL\_M5}$$

Fig. 5.10 Dominant leakage sources in an inactive SRAM cell.

5.5 Gated-V_{DD} SRAM Architectures

Several gated-V_{DD} SRAM architectures adopting power-gating technique have been proposed in recent years. In this section, two of them are introduced and discussed in detail about their features and behaviors.

5.5.1 Row (Wordline)-Controlled Architecture

Fig. 5.11 shows a gated-V_{DD} (or gated-ground) SRAM that row decoder controls the gating devices [5.7]. In this architecture, all the SRAM cells on the same wordline share a common gating device. All the cells on the same wordline are turned on when the row is selected, and other unselected rows are in standby mode to save leakage power. Note that no any diode device is included between virtual GND and actual GND. The authors carefully sized the gating devices to maintain the data stability.

Since the row-controlled scheme share a common gating device per wordline, the capacitance at the virtual GND node is quite large and it may take a large amount of time to discharge this node. Consequently, the maximum operating clock frequency is limited due to the extra time to discharge the virtual node. Moreover, not all the cells are necessary for each read/write operation. Therefore, turning on the unnecessary cells just wastes a great mount of active power.

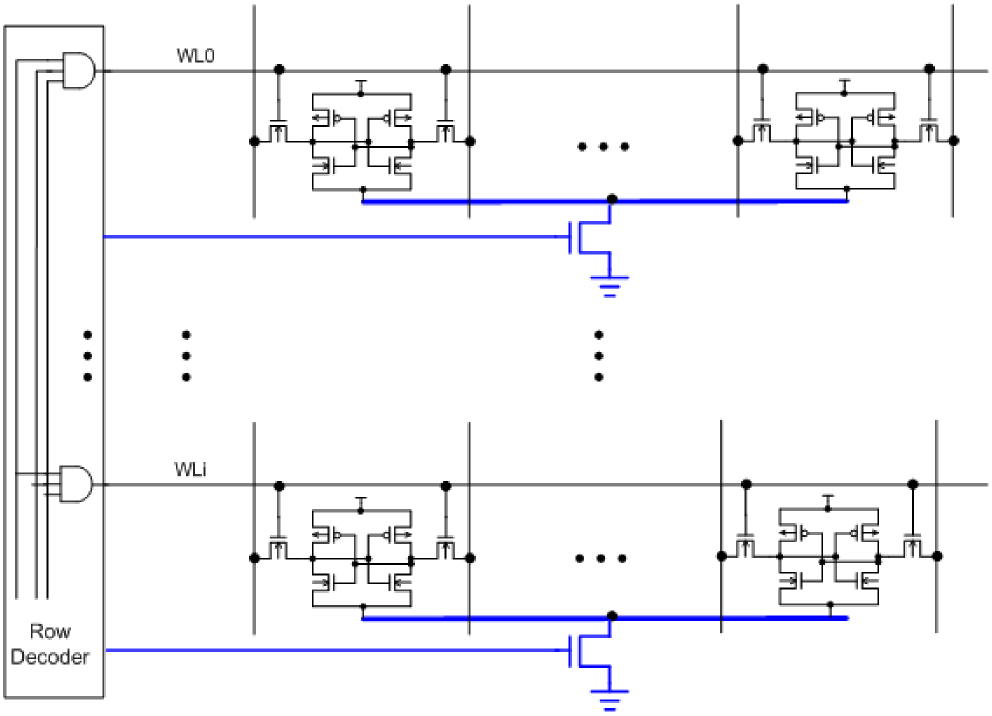


Fig. 5.11 Row-controlled SRAM architecture that row decoder controls the gating devices.

5.5.2 Column-Controlled Architecture

In contrast to row-controlled scheme, column-controlled scheme controls the gating devices by column decoder. Fig. 5.12 shows the schematic diagram of column-controlled SRAM architecture [5.8]. All the cells on the same bitline share a common gating device, and only the cells of the selected bitline are turned on. The virtual GND node is also a large capacitive node and the value bases on the number of wordlines.

As in column-controlled scheme, all the cells on one selected bitline are turned on but only one of them is selected by wordline. Consequently, less power saving is obtained due to the cells in active mode but unnecessary.

This scheme and the previous one have the same drawbacks, one is that they turn on many unnecessary cells for each read/write operation, and the other is that the virtual nodes are large capacitive nodes. The first one drawback makes power saving less and the other limits the maximum operating frequency due to the discharging of the large capacitive nodes.

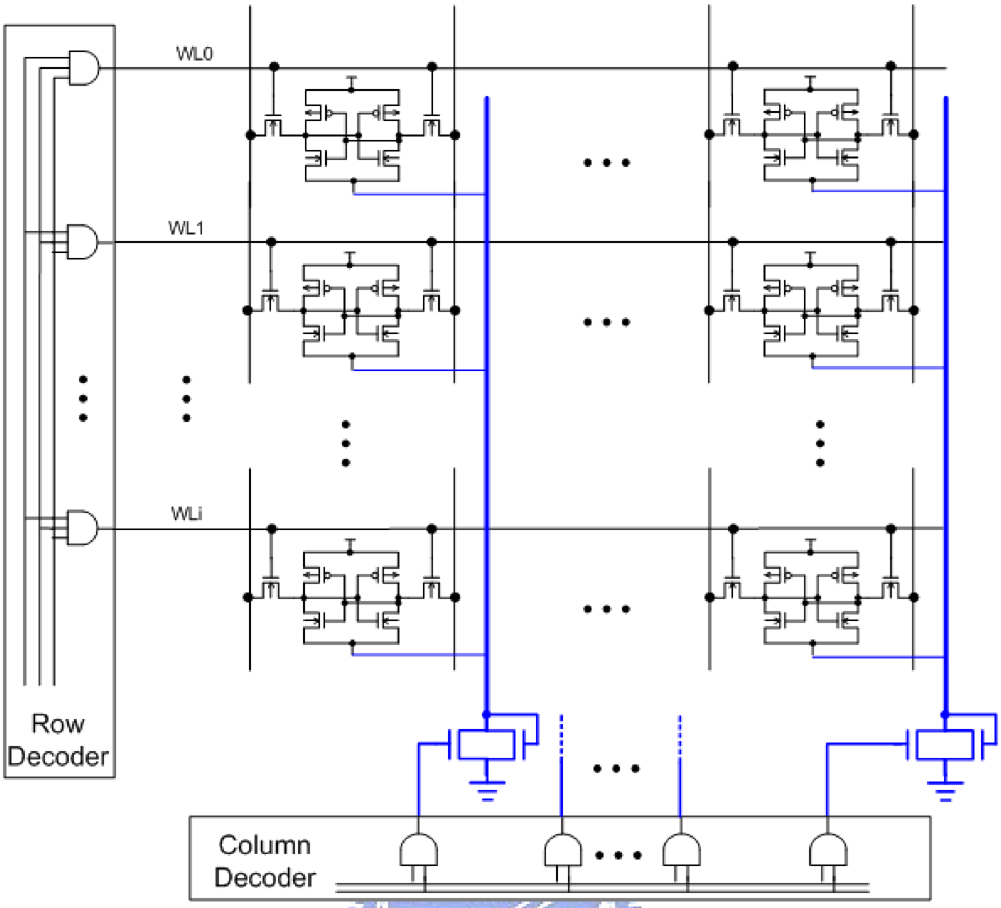


Fig. 5.12 Column-controlled SRAM architecture that column decoder controls the gating devices.

5.6 Column/Row Co-Controlled Architecture

A new gated-V_{DD} SRAM architecture is proposed in this section, and this scheme conquers the two drawbacks of the two previous schemes.

5.6.1 Schematic Diagram

Fig. 5.13 shows the schematic diagram of proposed SRAM scheme. In contrast to the previous two schemes, this scheme controls the gating devices with signals from both row and column decoders. The cells on the same wordline are grouped in blocks, and the block size depends on the number of I/O pins. Fig. 5.13 is an example of 32-bit wordline and 8-bit I/O and the 32 cells are divided into four blocks. Note that the wordline signals right from row decoder are not directly connected to the cells but through AND gates. The AND gates receive signals from row and column decoders and generate control signals to serve as 'local' wordlines and gating devices control signals.

Fig. 5.13 shows that each block is turned on only when both wordline and selection signals (*sel0*, *sel1*, and so on) are pulled high. The reason for this scheme is that for an 8-bit I/O SRAM core, only 8-bit data are either read from the SRAM or written into the SRAM per operation. That's why the block size depends on the lengths of I/O. It's straightforward to realize that the active power of this example is about 25% of row-controlled scheme.

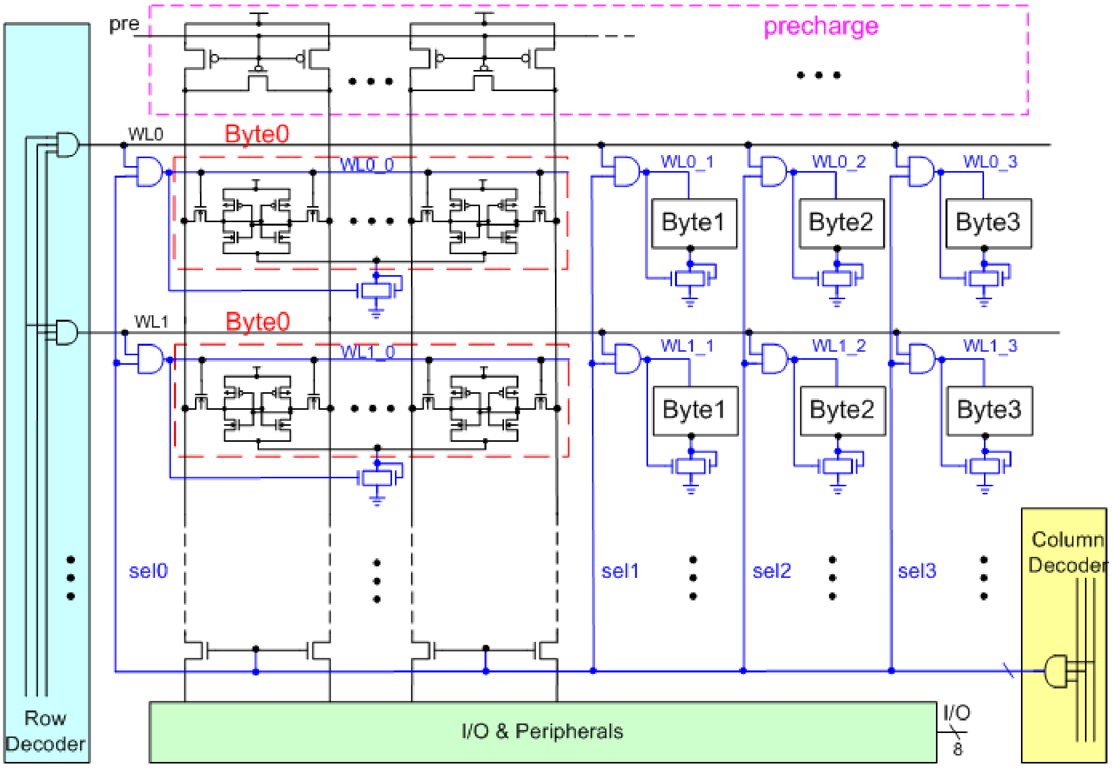


Fig. 5.13 Proposed column/row co-controlled SRAM scheme.

5.6.2 Effectiveness of Proposed Scheme

Fig. 5.14 is the test circuit used here to observe the cell current in standby and active modes. It comprises eight SRAM cells and a gating device and a diode-connected transistor. Table 5.1 shows the total cell current of the circuit in standby and active modes. Table 5.1 reveals that the cell current in active mode is greater than standby mode by more than 1000 times. Therefore, more unnecessary cells are power gated, more power saving is predictable.

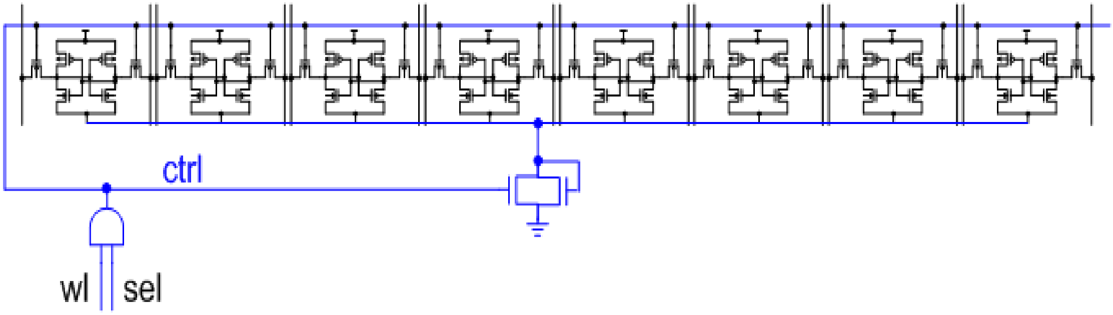


Fig. 5.14 Test circuit to observe the cell current in standby and active modes.

Table 5.1 Cell current in standby and active modes.

Standby/Active	Icell (8-bit)
ctrl=0 (standby)	**1.08nA**
ctrl=1 (active)	**1.22uA**

5.7 Three Typical Schemes for Comparison

In order to judge the effectiveness and usefulness, three test circuits are constructed and Fig. 5.15 shows the three different SRAM architectures. Note that all of them contain 32 SRAM cells per wordline. In the following the performance and power consumption of these three circuits will be compared.

In Fig. 5.15, the first one is conventional SRAM, which contains no any gating device. The second one is row-controlled scheme with only one gating device for all the cells. Besides, the gating device is controlled directly by wordline signal. The last one is proposed column/row co-controlled scheme, which contains four 8-bit blocks. The four 8-bit blocks have separate gating devices and one for each block. The control signals come from both column and row decoders. Any block is active only when both the wordline and byte selection (*sel0*, *sel1*, and so on) signal is high. It's noticeable that the gating device in (2) is four-time larger than any one in (3).

5.7.1 Read-Out Delay

Fig. 5.16 shows the simulated waveform of data read-out delay. The curve 'Dout (1)' is the output for conventional scheme, 'Dout (2)' is for row-controlled scheme, and 'Dout (3)' is for proposed scheme. Undoubtedly, conventional scheme is the fastest one to read out data. From Fig. 5.16, the read-out delay for row-controlled scheme is slightly larger than conventional. There are at least two reasons. First, row-controlled scheme has smaller active current due to stacked gating device. Second, row-controlled scheme needs extra time to discharge virtual GND node.

The read-out delay of proposed scheme is obviously larger than the other two schemes. This is mainly due to the gate delay of the 'AND' gate used to generate control signal for gating device and cells. According to Fig. 5.16, the delay is larger than conventional by about 27%. Note that this value is measured from wordline-to-output time. Therefore, the overhead would be smaller if we measure it from the whole time of one complete read/write operation (clock-to-output).

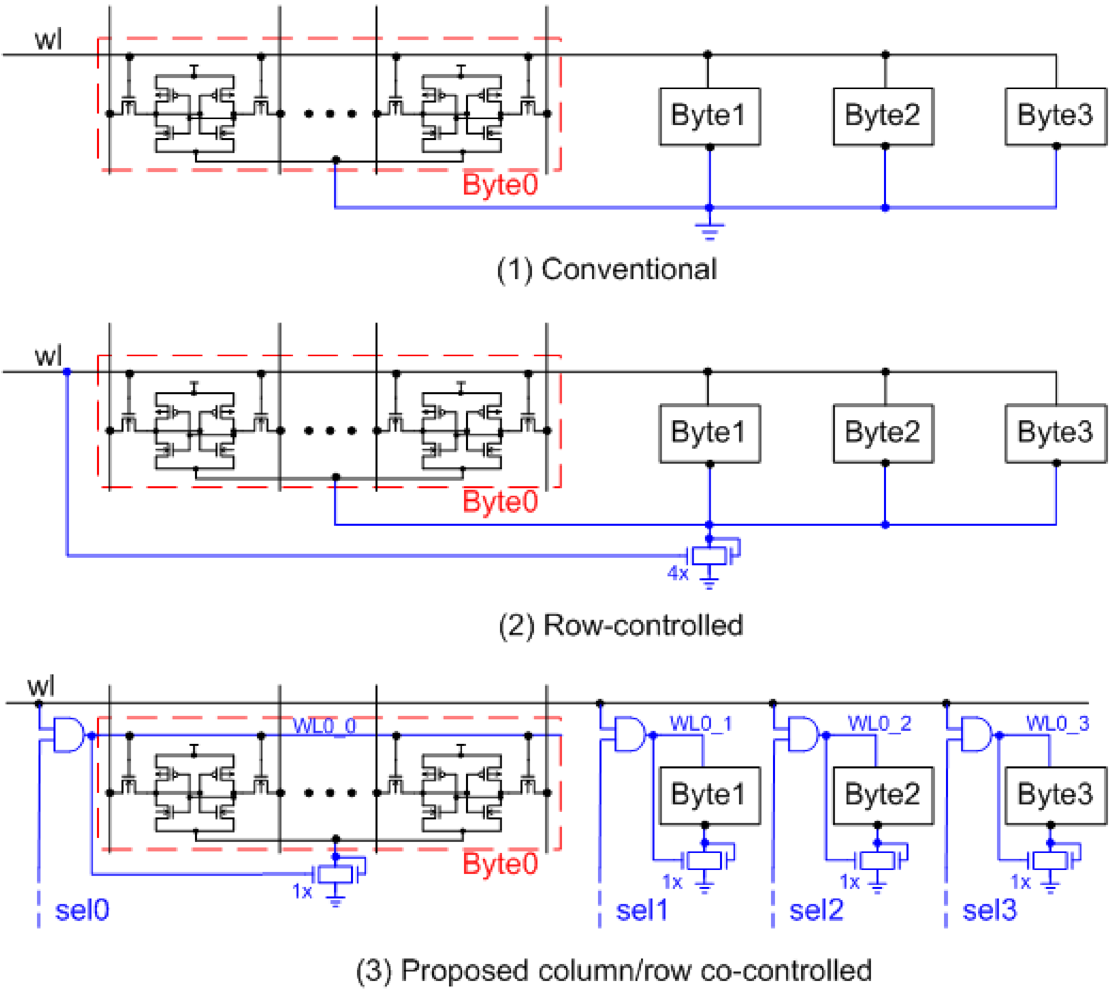


Fig. 5.15 Three SRAM test circuits to compare their performance and power consumption.

5.7.2 Cell Standby Power

Fig. 5.17 shows the comparison of cell standby power of the three test circuits. Obviously, conventional scheme consumes most cell standby power since no gating device adopted. Row-controlled and proposed schemes almost have the same cell standby power consumption. These two schemes are equivalent in standby mode because they both turn off all the cells, and about 60% cell standby power is reduced.

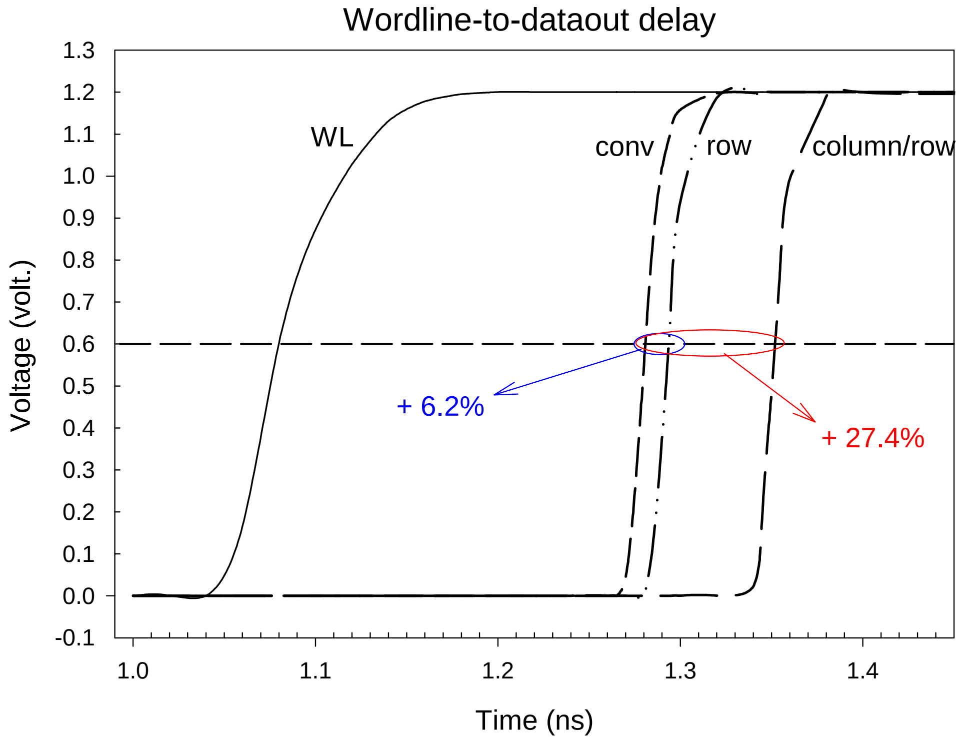


Fig. 5.16 Simulated read-out delay for the three test circuits.

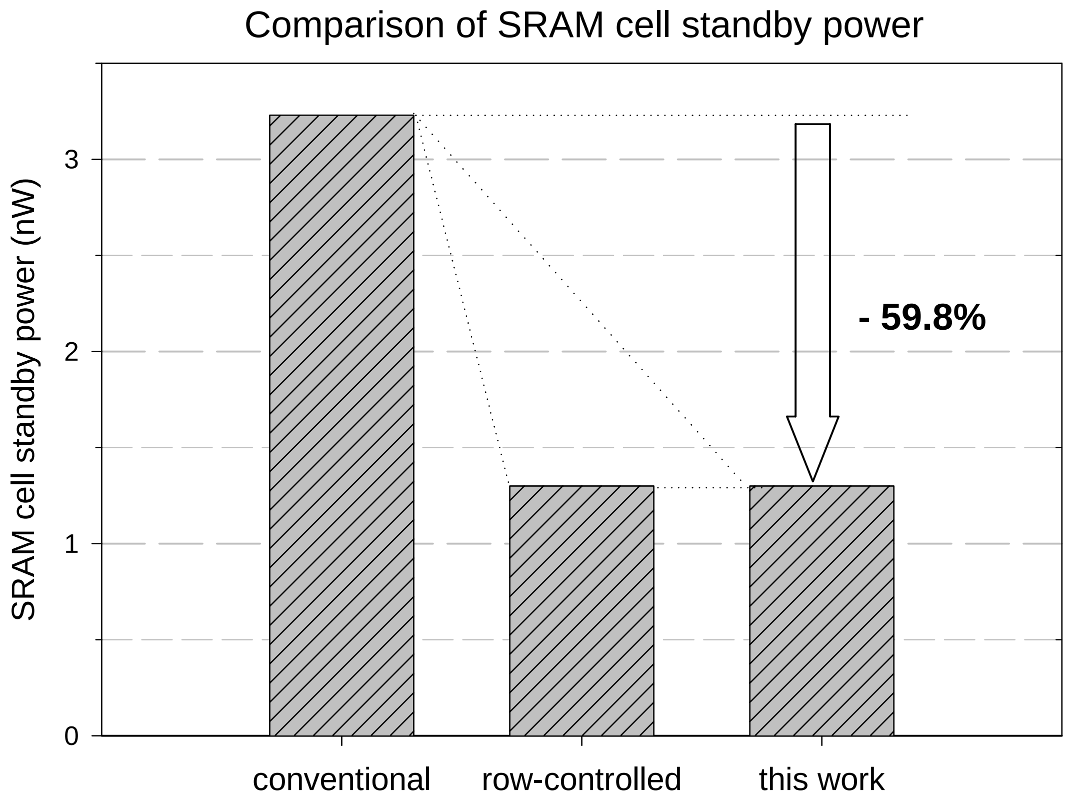


Fig. 5.17 Cell standby power comparison.

5.7.3 Active Power

Fig. 5.18 shows the comparison of normalized cell active power of the three SRAM architectures. The cell active power of row-controlled scheme is slightly smaller than conventional scheme. Due to the stacked gating device, row-controlled scheme has smaller active current and thus smaller active power. As mentioned before, this smaller current makes larger read-out delay.

In comparison with row-controlled and proposed schemes, the active power of the later one achieves 77% power saving. It seems so straightforward that proposed scheme just turns on the gating device of one block and the other three blocks are remained power-gated. No doubt about three-fourth power saving is obtained.

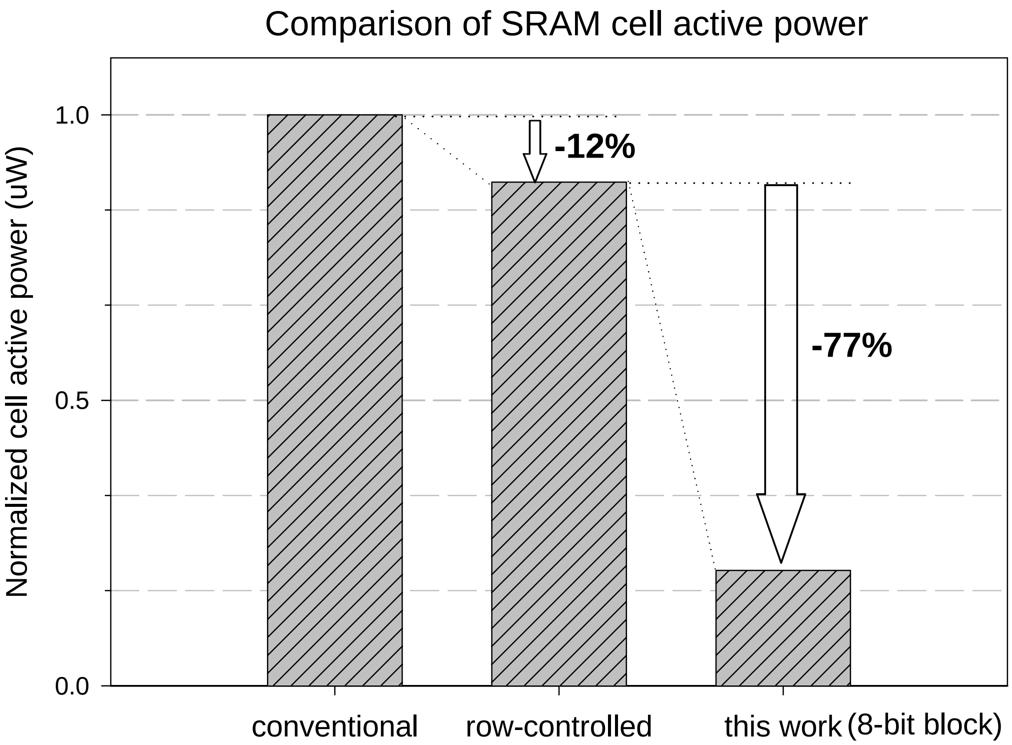


Fig. 5.18 Cell active power comparison.

5.7.4 Wordline Length, Block Size, and Cell Active Power

Fig. 5.19 depicts cell active power versus various wordline lengths and block sizes. The figure reveals that the cell active power of conventional and row-controlled schemes is proportional to wordline lengths, since both of them turn on all the cells for each operation. As for proposed scheme, however, the active power is almost a constant for a fixed block size, regarding of wordline length. This is because that only one block of cells is active at the same time no matter the length of wordline length is.

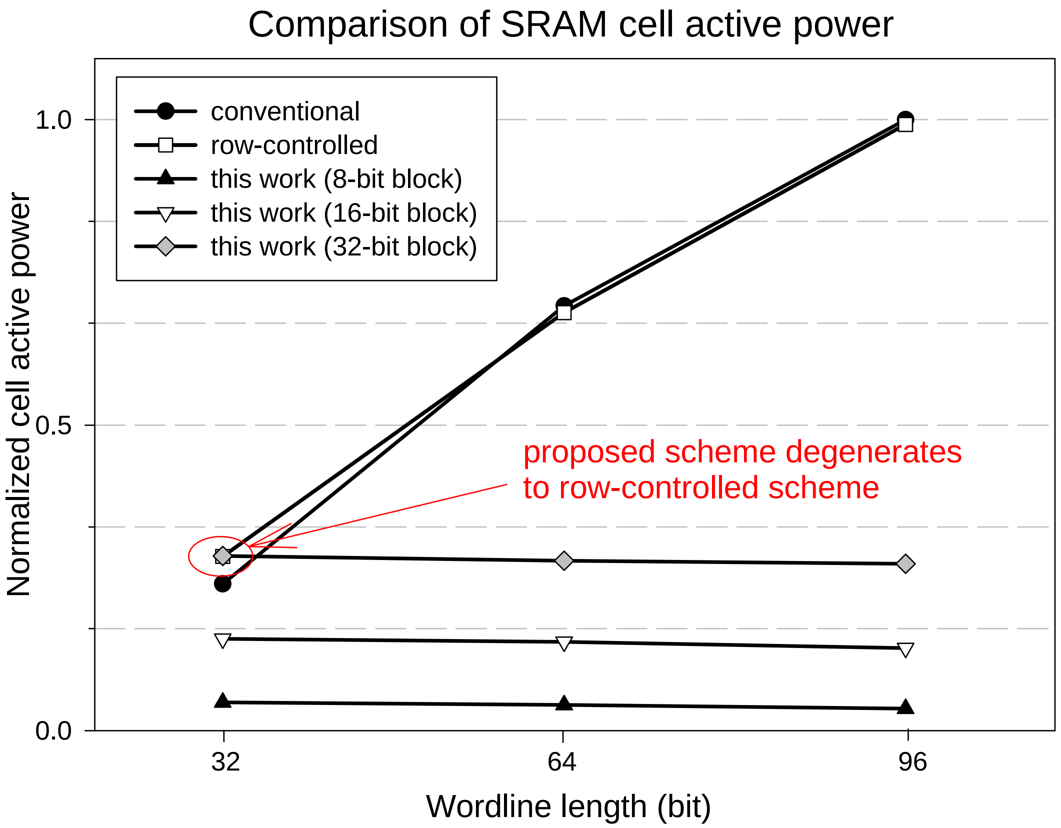


Fig. 5.19 Cell active power versus different wordline lengths and block sizes.

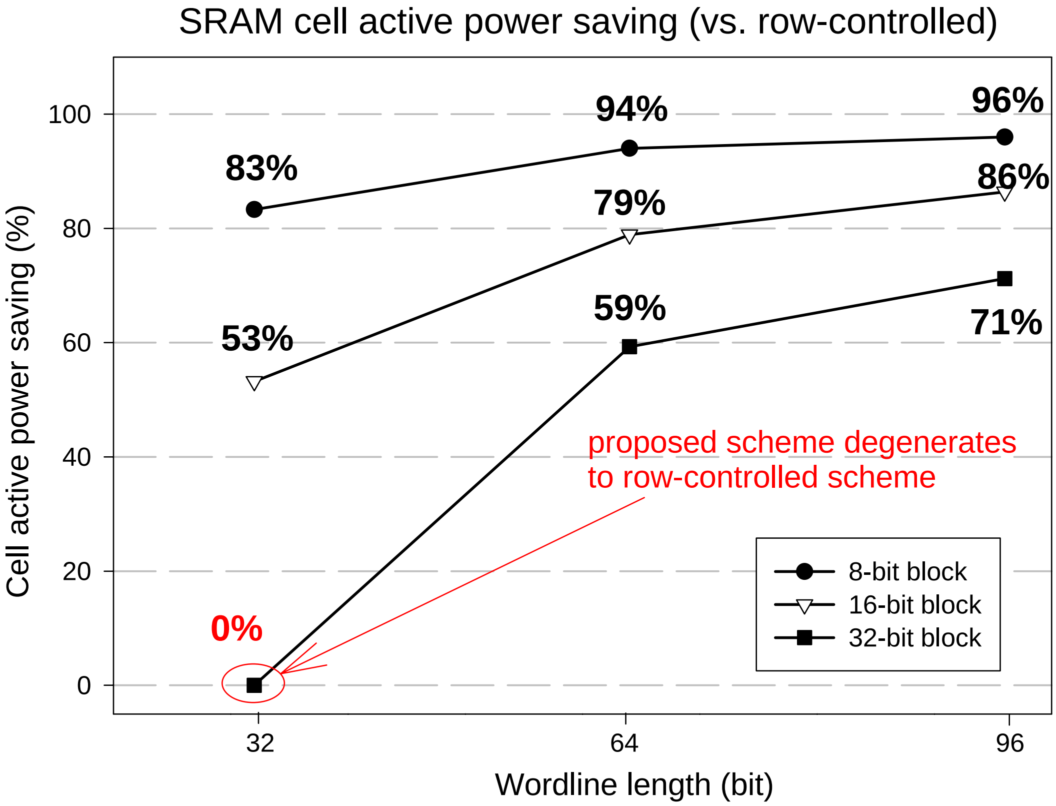


Fig. 5.20 Cell active power saving versus different wordlines and block sizes.

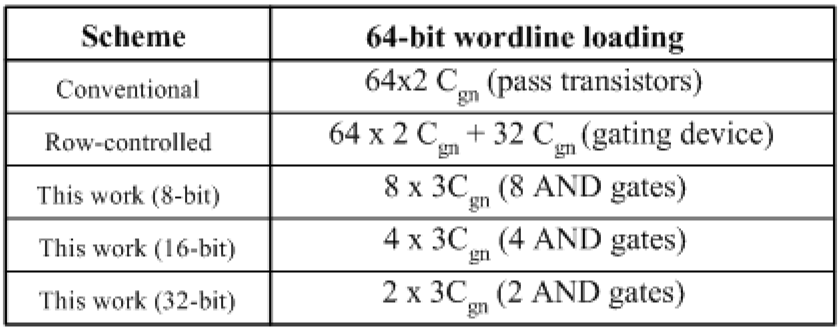
Fig. 5.19 also shows that less cell power consumption is obtained with smaller block size. It' clearly understood that smaller block size means fewer active cells for each operation and thus less cell active power consumption. However, the block size depends on the number of I/O pins that is usually fixed.

Fig. 5.20 shows the cell active power saving of proposed scheme with various wordline lengths and block sizes. It's obvious that most cell active power saving is achieved while the block size is the smallest. The circled points in Fig. 5.19 and Fig. 5.20 are interesting that the wordline length and block size both are 32 bits. In this situation, proposed scheme degenerates to row-controlled scheme since all the cells on the same wordline are turned on at the same time. Therefore, proposed and row-controlled schemes consume the same amount of cell active power, as shown in Fig. 5.18. Undoubtedly, the circled point in Fig. 5.20 shows a 0% power saving in this condition.

5.7.5 Reduction of Wordline Loading

Although proposed scheme induces an extra gate delay and increases wordline-to-output delay, this overhead is diminished due to the division of wordline loading. As in Fig. 5.15, the wordline of conventional scheme connects to 64 NMOS pass transistors of the 32 cells. Besides the 64 pass transistors above, the wordline of row-controlled scheme connects to one more gating transistor. As for proposed scheme, however, the wordline just connects to four AND gates, including four NMOS and four PMOS transistors. Table 5.2 lists the statistic of wordline loading of 64-bit wordline for different schemes, and Fig. 5.21 shows the simulated waveform. Note that the term C_{gn} in Table 5.2 means the gate capacitance of a unit NMOS transistor. Fig. 5.21 clearly shows that the rising time for row-controlled scheme is larger than conventional due to the gating device. Furthermore, the wordline signals of proposed scheme rise with faster speed.

Table 5.2 Statistic of 64-bit wordline loading for different schemes.



Scheme	64-bit wordline loading
Conventional	64x2 C_{gn} (pass transistors)
Row-controlled	64 x 2 C_{gn} + 32 C_{gn} (gating device)
This work (8-bit)	8 x 3C_{gn} (8 AND gates)
This work (16-bit)	4 x 3C_{gn} (4 AND gates)
This work (32-bit)	2 x 3C_{gn} (2 AND gates)

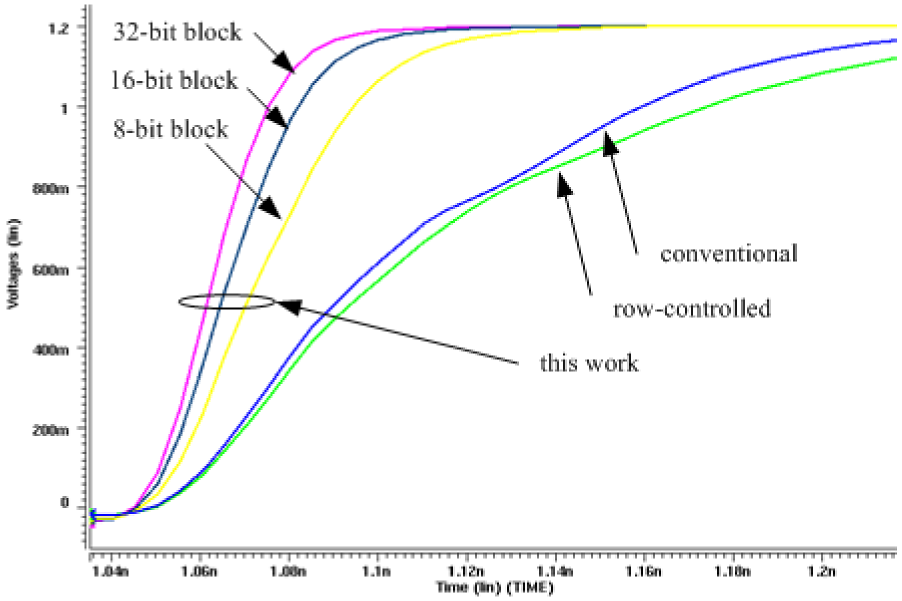


Fig. 5.21 Simulated waveforms of wordline curves.

5.7.6 Power-Delay Product

Fig. 5.22 shows the power-delay product (64-bit wordline) normalized with respect to the conventional scheme. From the graph, it's observed that the proposed scheme achieves significant reductions in power-delay product. A reduction of 93% is obtained for 8-bit block scheme, 75% for 16-bit block, and 49% for 32-bit block.

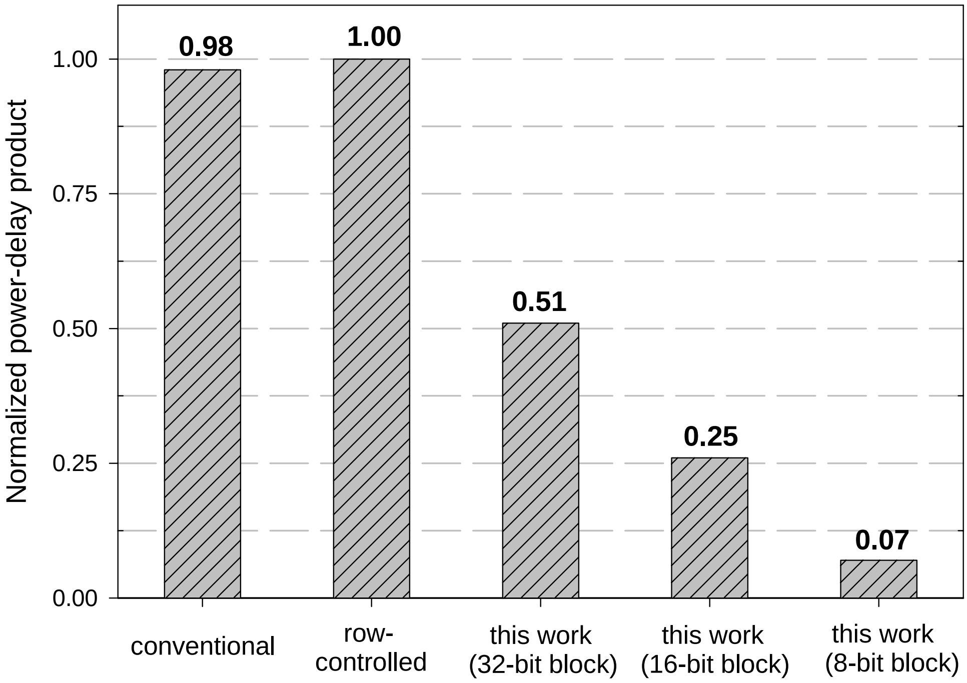


Fig. 5.22 Power-delay product.

5.7.7 Signal Routing and Device Allocation for SRAM Layout

Fig. 5.23 (a) shows the layout of two SRAM cells with an AND gate and a gating device, which are implemented in 0.13um CMOS technology. For an 8-bit block wordline, the extra AND gate and gating device each occupies the area that is nearly equivalent to the area of a SRAM cell. This implies that the area overhead is limited and it’s inversely proportional to block size. Since a larger driving current flows through a larger block of cells, thus a wider gating device is required to sustain such amount of current.

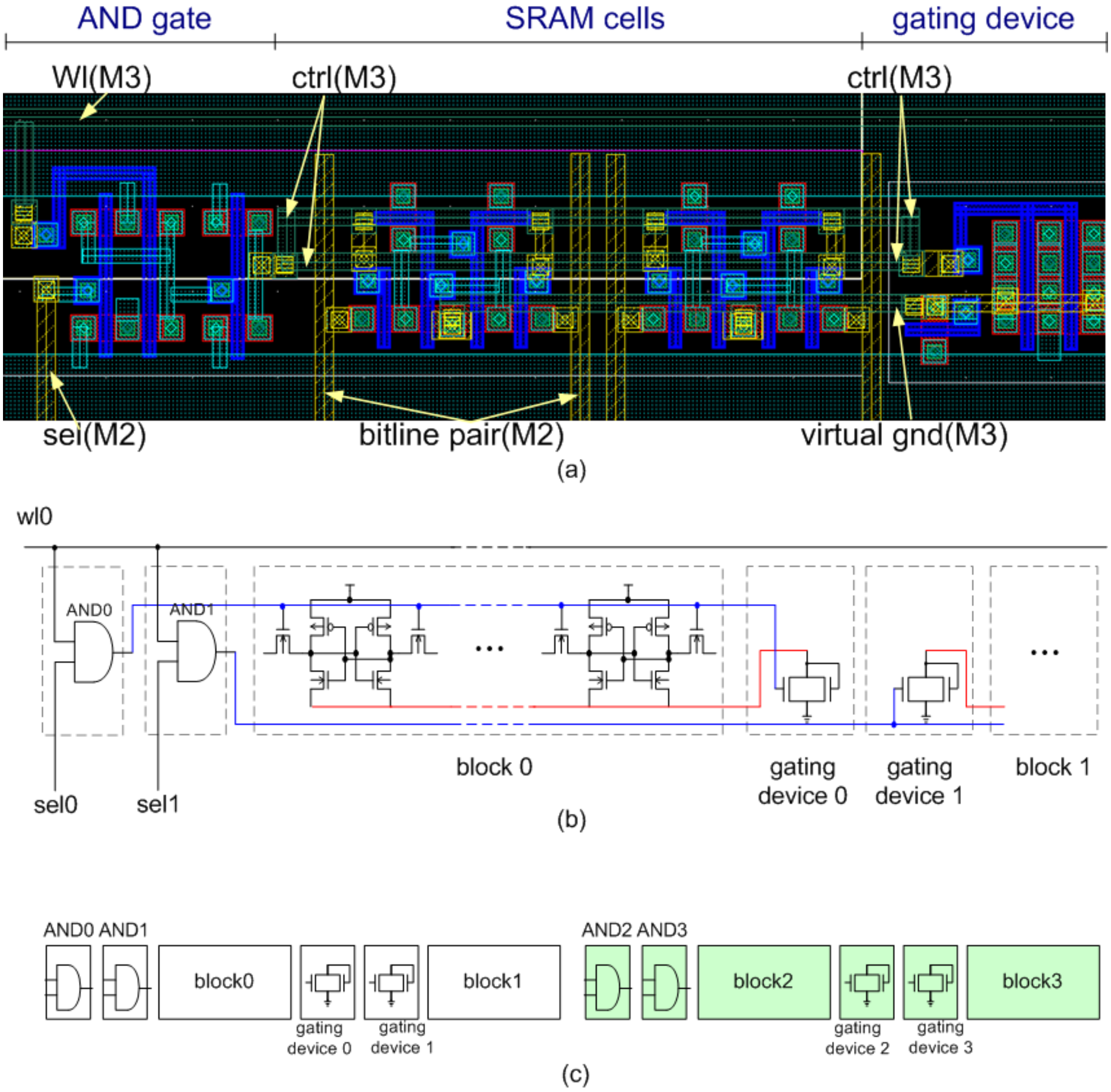


Fig. 5.23 (a) Layout and (b) allocation of AND gates and gating devices and (c) a wordline with four blocks.

Fig. 5.23 (b) illustrates the allocation of AND gates and gating devices. The two AND gates of two blocks are implemented together and so are their gating devices. However, the AND gates are in the forefront of the wordline and the gating devices are between the two blocks of cells. Because a gating device consists of two NMOS transistors, the layout area can be reduced if more gating devices are implemented together. On the other hand, the gating device should be adjacent to the block in order not to induce a significant IR drop. Therefore, only two gating devices are grouped together so that the two gating devices are adjacent to the blocks.

Fig. 5.23 (c) further depicts the allocation of AND gates and gating devices for a wordline with four blocks. It clearly shows that no any block suffers too much IR drop and the virtual GND nodes can be completely pulled to zero potential. Fig. 5.24 shows the layout and floorplan of 8Kb cell arrays that the block size is 16 bits.

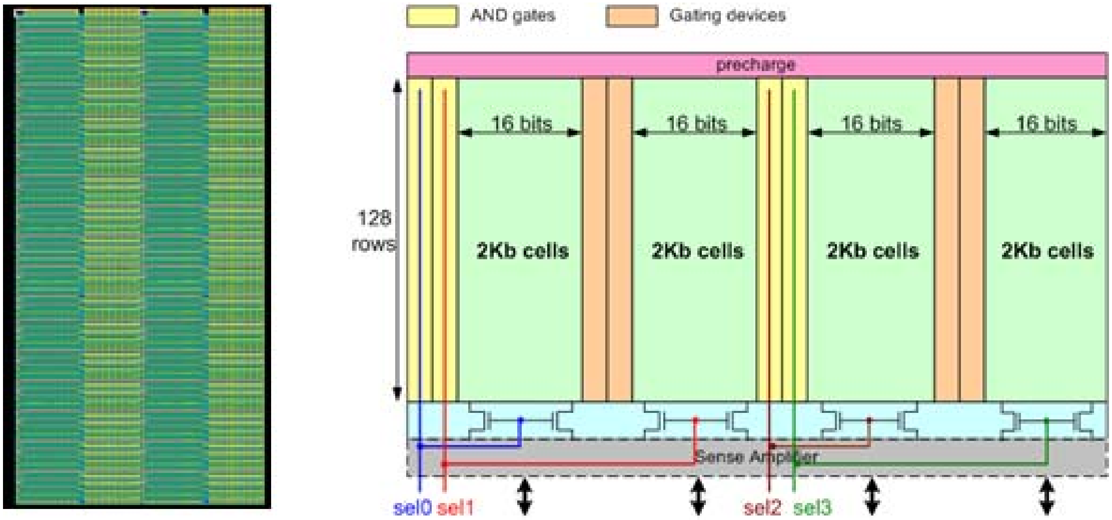


Fig. 5.24 Layout and signal routing of 8Kb SRAM cell arrays.

5.7.8 Area Overhead Due to Extra Devices

Since the new scheme includes some extra devices, the area overhead must be taken into account. Fig. 5.25 compares the area of conventional and new scheme for several block sizes. Fig. 5.25 shows that the area for 8-bit block scheme is much larger than the others because of more AND gates and gating devices. Besides, the overhead percentage keeps constant with various lengths of wordlines.

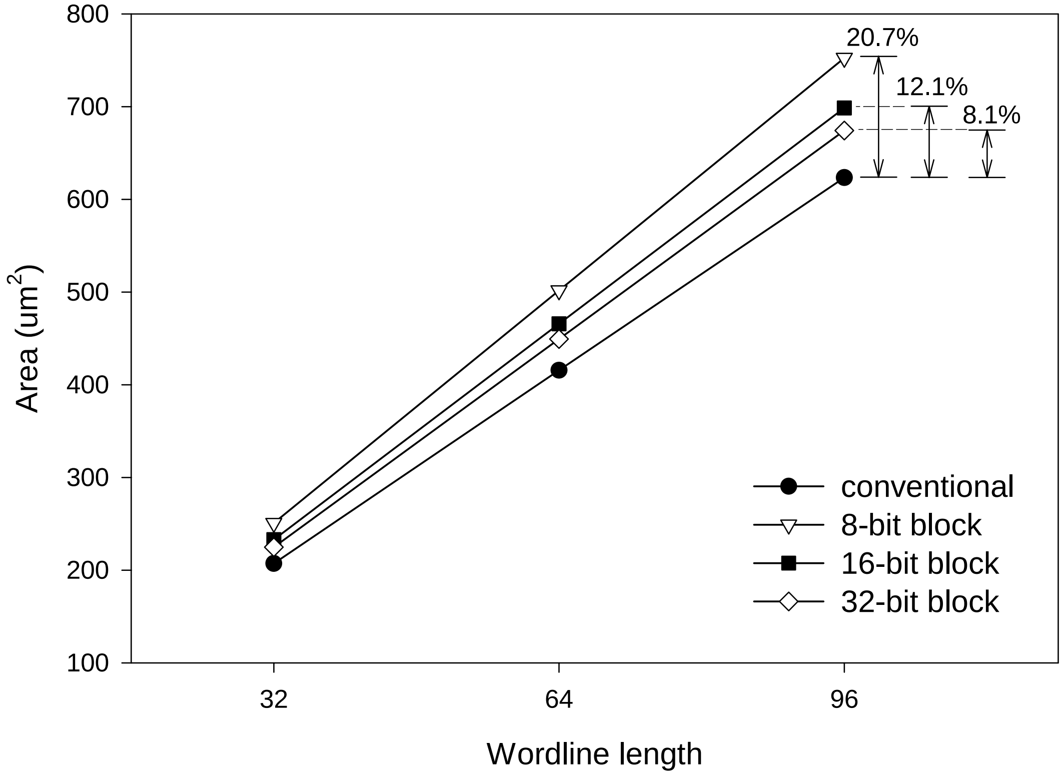


Fig. 5.25 Area comparison of conventional and new scheme with different block sizes.

Fig. 5.26 summarizes the area overhead for different block sizes. For an 8-bit block wordline, about 54% area overhead comes from the AND gates since in this situation the gating devices are smaller, and totally a 20.7% area increase is presented. For a 16-bit wordline, however, a total 12.1% area increase is obtained and about 44% of it comes from the AND gates. Finally, for a 32-bit block wordline, a total 8.1% area overhead is induced and only 34% of it comes from the AND gates. Fig. Although the overall area overhead is getting less severe with larger block size, the influence of gating devices is getting severe and the AND gates are becoming less significant. This is because that the number of AND gates is becoming fewer while the sizes of gating devices are becoming larger.

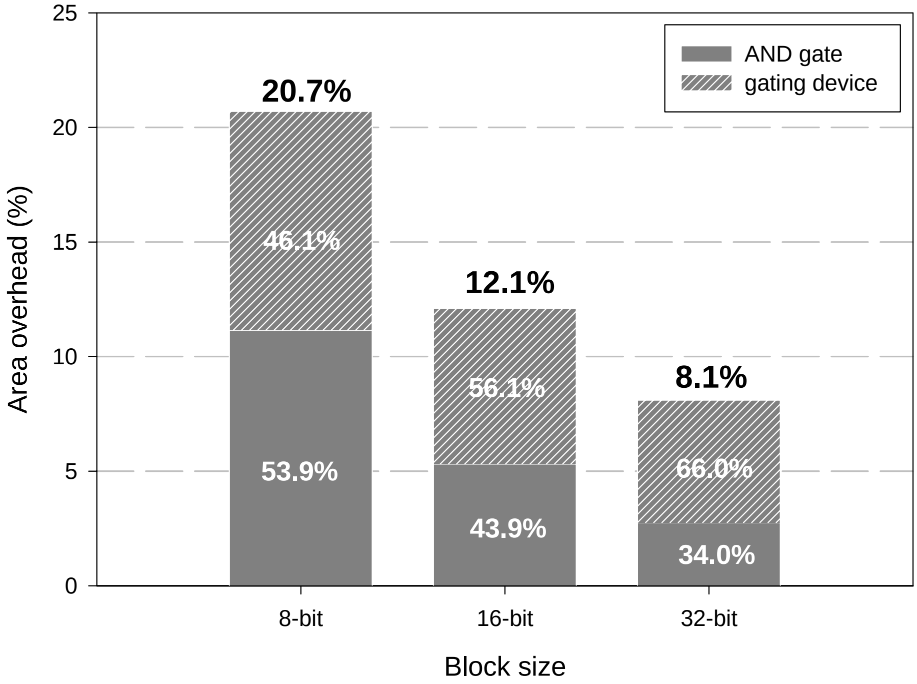


Fig. 5.26 Summary of area overhead.

5.8 Conclusion

The design issues of gated-V_{DD} SRAM are discussed and some prior designs are introduced in this chapter. Moreover, an ultra-low active-power gated-V_{DD} SRAM architecture is realized.

The column/row co-controlled scheme divides the SRAM cells on the same wordline into blocks, and one gating device is responsible for one block. The cell active power of the scheme is much smaller than the other two schemes compared. Simulation also shows that the scheme achieves a significant reduction of power-delay product. This demonstrates the effectiveness of this new scheme in reducing active power consumption with insignificant performance degradation.

The layout is implemented with 0.13um CMOS technology, and a SRAM cell occupies an area of about 2.40um x 1.91um. The new scheme has larger silicon area because of the extra AND gates and gating devices. About 20.7% and 12.1% area increases are obtained for 8-bit block and 16-bit block situations, respectively. Besides, only 8.1% area overhead is induced for 32-bit block condition.

Chapter 6
Conclusions

As technology continues to scale down, subthreshold leakage is becoming worse and it has the potential to dominate the whole power consumption of a chip. Many researches and predictions have showed that leakage power is becoming comparable with dynamic power in nano-scale technologies. In the past, leakage currents are ignored because they are insignificant compared to dynamic currents, since threshold voltage is high. In the deep-submicron and nano-scale technologies, however, IC designers must pay much more attention to leakage currents and some circuit techniques and design considerations must be developed to control leakage currents in both active and standby modes.

The dominant leakage source, subthreshold current, increases exponentially with the scaling of threshold voltage. Therefore, many techniques that dynamically adjust threshold voltage by applying body bias are developed in recent decades. Threshold voltage is raised with reversed body bias, while it’s lowered with the application of forward body bias. Power gating is another popular technique to suppress leakage current in standby mode. It inserts sleep transistors between virtual and real power lines and those transistors are turned on in active mode and turned off in standby mode. In standby mode, the leakage currents are reduced due to the stacking of off transistors. All the low power design techniques have been reviewed in Chapter 2.

In this thesis dynamic body-biasing and power-gating techniques are investigated and applied to SRAM cell arrays to observe the effectiveness in suppressing leakage. Notice that this thesis focuses on applying reversed body bias and the reduction of leakage currents. For the purpose of flexibility and reusability in SoC systems, a configurable scheme for generating multi-level body-bias voltages is presented in Chapter 3 and various voltage levels can be produced through control signals. For SoC designs, several supply and body-bias voltages are necessary for biasing different parts of the designs. Therefore, an on-chip configurable voltage generator is quite useful since no other kinds of voltage generators are required. Besides, another dual-level body bias generator is also constructed in Chapter 4 and it produces two voltage levels according to control signal. The dual-level body bias generator is applied to SRAM cell arrays and a great amount of leakage saving in standby mode is observed. Moreover, the net standby power saving is significant even the power overhead of body bias generator is included. The average power consumption of body bias generator converges with time so that the net power saving is time dependent. Simulation results show that about 75% net power saving is achieved for 64-bit wordline and 64% for 32-bit wordline. A time-out-policy controller that monitors the activities of wordline signals controls the body bias generator. The controller detects the activities of wordline signals and filters out short standby periods. Body bias generators are enabled if no active wordline occurs within predetermined time period.

In the rest of this thesis, power-gating technique is applied to SRAM cell array designs, and a low active and standby power SRAM scheme is presented. The gating devices are controlled by signals from both column and row decoders and only selected cells are power-on. The SRAM cells on the same wordline are divided into blocks and each block shares a common gating devices. For each read/write operation, only selected block is power-on and others are power-gated. Simulation results show that for 64-bit wordlines, 59% active power saving is achieved for 32-bit block, 79% for 16-bit block, and 94% for 8-bit block conditions. However, this scheme induces area overhead since some extra AND gates and gating devices are added. About 20.7% and 12.1% area increases are obtained for 8-bit block and 16-bit block conditions, respectively. Besides, only 8.1% area overhead is induced for 32-bit block condition. Although the AND gates induce performance overhead, power-delay produces demonstrate that the influences are insignificant.

The configurable body-bias generator in Chapter 3 is simulated in TSMC 100nm technology, while the simulations and physical layout in Chapter 4 and 5 are based on TSMC 0.13um technology.

References

References of Chapter 1

[1.1] K. Bowman, S. Duvall, and J. Meindl, “Impact of Die-to-Die and Within-Die Parameter Fluctuations on the Maximum Clock Frequency Distribution for Gigascale Integration,” *IEEE Journal of Solid-State Circuits*, vol. 37, pp. 183-190, Feb. 2002.

[1.2] J. Tschanz, S. Narendra, R. Nair, V. De, “Effectiveness of Adaptive Supply Voltage and Body Bias for Reducing Impact of Parameter Variations in Low Power and High Performance Microprocessors,” *IEEE Journal of Solid-State Circuits*, vol. 38, pp. 826-829, .May 2003.

[1.3] http://www-device.eecs.berkeley.edu: BSIM 100nm and 70nm predictive technology process files.

[1.4] International Technology Roadmap for Semiconductors 2001 edition, Semiconductor Industry Association, http://public.itrs.net.

References of Chapter 2

[2.1] T. Inukai, M. Takamiya, K. Nose, H. Kawaguchi, T. Hiramoto, and T. Sakurai, "Boosted Gate MOS (BGMOS): Device/Circuit Cooperation Scheme to Achieve Leakage-Free Giga-Scale Integration," *Custom Integrated Circuits Conference*, pp. 409-412, 2000.

[2.2] B. Chatterjee, M. Sachdev, S. Hsu, R. Krishnamurthy, and S. Borkar, "Effectiveness and Scaling Trends of Leakage Control Techniques for Sub-130nm CMOS Technologies," *International Symposium on Low Power Electronics and Design*, pp. 122-127, 2003.

[2.3] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De, "Parameter Variations and Impact on Circuits and Microarchitecture," *Design Automation Conference*, pp. 338-342, June 2003.

[2.4] Y. Liu, C. Wu, C. Chang, R. Yang, W. Chen, J. Liaw, and C. Diaz, "Leakage Scaling in Deep Submicron CMOS for SoC", *IEEE Transaction on Electron Devices*, vol. 49, pp. 1034-1041, June 2002.

[2.5] A. Keshavarzi, S. Ma, S. Narendra, B. Bloechel, K. Mistry, T. Ghani, S. Borkar, and V. De, "Effectiveness of Reverse Body Bias for Leakage Control in Scaled Dual Vt CMOS ICs," *International Symposium on Low Power Electronics and Design*, pp. 207-212, 2001.

[2.6] J. Tschanz, S. Narendra, Y. Ye, B. Bloechel, S. Borkar, and V. De, "Dynamic Sleep Transistor and Body Bias for Active Leakage Power Control of Microprocessors," *IEEE Journal of Solid State Circuits*, vol. 38, pp. 1838-1845, Nov. 2003.

[2.7] C. Neau, and K. Roy, "Optimal Body Bias Selection for Leakage Improvement and Process Compensation Over Different Technology Generations," *International Symposium on Low Power Electronics and Design*, pp. 116-121, 2003.

[2.8] X. Liu, and S. Mourad, "Performance of Submicron CMOS Devices and Gates With Substrate Biasing," *International Symposium on Circuits and Systems*, pp. IV9-IV12, 2000.

[2.9] A. Keshavarzi, S. Narendra, S. Borkar, C. Hawkins, K. Roy, and V. De, "Technology Scaling Behavior of Optimum Reverse Body Bias for Standby Leakage Power Reduction in CMOS IC's," *International Symposium on Low Power Electronics and Design*, pp. 252-254, 1999.

[2.10] M. Chen, H. Huang, C. Hou, and K. Yang, "Back-Gate Bias Enhanced Band-to-Band Tunneling Leakage in Scaled MOSFET's," *IEEE Electron Device Lett.*, vol. 19, pp. 134-136, 1998.

[2.11] T. Kobayashi, and T. Sakurai, "Self-Adjusting Threshold-Voltage Scheme (SATS) for Low-Voltage High-Speed Operation," *Custom Integrated Circuits Conference*, pp. 271-274, 1994

[2.12] T. Kuroda, T. Fujita, S. Mita, T. Nagamatsu, S. Yoshioka, K. Suzuki, F. Sano, M. Norishima, M. Murota, M. Kako, M. Kinugawa, M. Kakumu, and T. Sakurai, "A 0.9V, 150-MHz, 10-mW, 4mm^2, 2-D Discrete Cosine Transform Core Processor With Variable Threshold-Voltage (VT) Scheme," *IEEE Journal of Solid-State Circuits*, vol. 31, pp. 1770-1779, Nov. 1996.

[2.13] M. Miyazaki, G. Ono, and K. Ishibashi, "A 1.2-GIPS/W Microprocessor Using Speed-Adaptive Threshold-Voltage CMOS With Forward Bias," *IEEE Journal of Solid-State Circuits*, vol. 37, pp. 210-217, Feb. 2002.

[2.14] K. Nose, M. Hirabayashi, H. Kawaguchi, S. Lee, and T. Sakurai, "Vth-Hopping Scheme to Reduce Subthreshold Leakage for Low-Power Processors," *IEEE Journal of Solid-State Circuits*, vol. 37, pp. 413-419, March 2002.

[2.15] J. Kao, M. Miyazaki, and A. Chandrakasan, "A 175-mV Multiply-Accumulate Unit Using an Adaptive Supply Voltage and Body Bias Architecture," *IEEE Journal of Solid State Circuits*, vol. 37, pp. 1545-1554, Nov. 2002.

[2.16] J. Tschanz, J. Kao, S. Narendra, R. Nair, D. Antoniadis, A. Chandrakasan, "Adaptive Body Bias for Reducing Impacts of Die-to-Die and Within-Die Parameter Variations on Microprocessor Frequency and Leakage," *IEEE Journal of Solid State Circuits*, vol. 37, pp. 1396-1402, Nov. 2002.

[2.17] K. Seta, H. Hara, T. Kuroda, M. Kakumu, and T. Sakurai, "50% Active-Power Saving Without Speed Degradation Using Standby Power Reduction (SPR) Circuit," *International Solid-State Circuits Conference Dig. Tech. Papers*, pp. 318-319, 1995.

[2.18] J. Tschanz, S. Narendra, Y. Ye, B. Bloechel, S. Borkar, and V. De, "Dynamic Sleep Transistor and Body Bias for Active Leakage Power Control of Microprocessors," *IEEE Journal of Solid State Circuits*, vol. 38, pp. 1838-1845, Nov. 2003.

[2.19] G. Ono, and M. Miyazaki, "Threshold-Voltage Balance for Minimum Supply Operation," *IEEE Journal of Solid State Circuits*, vol. 38, pp. 830-833, May 2003.

[2.20] D. Lackey, P. Zuchowski, T. Bednar, D. Stout, S. Gould, and J. Cohn, "Managing Power and Performance for System-on-Chip Designs using Voltage Islands," *IEEE/ACM International Conference on Computer-Aided Design*, pp. 195-202, No. 2002.

[2.21] C. Bertin, A. Dean, K. Goodnow, S. Gould, W. Pricer, W. Tonti, and S. Ventrone, "Managing Vt for Reduced Power Using a Status Table," USA patent 6345362, Feb. 2002.

References of Chapter 3

[3.1] K. Itoh, K. Sasaki, and Y. Nakagome, “Trends in Low-Power RAM Circuit Technologies,” *Proceeding of the IEEE*, vol. 83, No. 4, April 1995.

[3.2] K. Itoh, Y. Nakagome, S. Kimura, and T. Watanabe, “Limitations and Challenges of Multigigabit DRAM Chip Design,” *IEEE Journal of Solid-State Circuits*, vol. 32, No. 5, pp. 624-634, May 1997.

[3.3] H. Masuda, R. Hori, Y. Kamigaki, K. Itoh, H. Kawamoto, and H. Katto, “A 5 V-Only 64K Dynamic RAM Based on High S/N Design,” *IEEE Journal of Solid-State Circuits*, vol. SC-15, No. 5, pp. 846-854, Oct. 1980.

[3.4] K. Itoh, "VLSI Memory Chip Design," Springer Verlag, 2001.

[3.5] T. Kuroda, T. Fujita, S. Mita, T. Nagamatsu, S. Yoshioka, K. Suzuki, F. Sano, M. Norishima, M. Murota, M. Kako, M. Kinugawa, M. Kakumu, and T. Sakurai, “A 0.9V, 150-MHz, 10-mW, 4mm^2, 2-D Discrete Cosine Transform Core Processor with Variable Threshold-Voltage (VT) Scheme,” *IEEE Journal of Solid-State Circuits*, vol. 31, pp. 1770-1779, Nov. 1996.

[3.6] H. Mizuno, K. Ishibashi, T. Shimura, T. Hattori, S. Narita, K. Shiozawa, S. Ikeda, and K. Uchiyama, “An 18-A Standby Current 1.8-V, 200-MHz Microprocessor with Self-Substrate-Biased Data-Retention Mode,” *IEEE Journal of Solid-State Circuits*, vol. 34, pp. 1492-1500, Nov. 1999.

[3.7] M. Miyazaki, G. Ono, T. Hattori, K. Shiozawa, K. Uchiyama, and K. Ishibashi, “A 1000-MIPS/W Microprocessor Using Speed-Adaptive Threshold-Voltage CMOS with Forward Bias,” *International Solid-State Circuits Conference Dig. Tech. Papers*, pp. 420–421, Feb. 2000.

[3.8] J. Kao, M. Miyazaki, and A. Chandrakasan, “A 175-mV Multiply-Accumulate Unit Using an Adaptive Supply Voltage and Body Bias Architecture,” *IEEE Journal of Solid State Circuits*, vol. 37, pp. 1545-1554, Nov. 2002.

[3.9] T. Tanzawa, T. Tanaka, K. Takeuchi, and H. Nakamura, “Circuit Technologies for a Single-1.8 V Flash Memory,” *Symposium of VLSI Circuits Dig. Tech. Papers*, pp. 63–64, June 1997.

[3.10] J. Dickson, “On-Chip High-Voltage Generation in NMOS Integrated Circuits Using an Improved Voltage Multiplier Technique,” *IEEE Journal of Solid-State Circuits*, vol. 11, pp. 374–378, June 1976.

[3.11] J. Wu, and K. Chang, “MOS Charge Pumps for Low-Voltage Operation,” *IEEE Journal of Solid-State Circuits*, vol. 33, pp. 592–597, April 1998.

[3.12] J. Shin, I. Y. Chung, Y. J. Park, and H. S. Min, “A New Charge Pump Without

Degradation in Threshold Voltage Due to Body Effect," *IEEE Journal of Solid-State Circuits*, vol. 35, pp. 1227–1230, Aug. 2000.

[3.13] K. Sawada, Y. Sugawara, and S. Masui, "An On-Chip Voltage Generator Circuit for EEPROM's with a Power-Supply Voltage below 2 V," *Symposium of VLSI Circuits Dig. Tech. Papers*, 1995, pp. 75–76.

[3.14] K. Choi, J. Park, J. Kim, T. Jung, and K. Suh, "Floating Well Charge Pump Circuits for Sub-2.0 V Single Power Supply Flash Memories," *Symposium of VLSI Circuits Dig. Tech. Papers*, pp. 61–62, June 1997.

[3.15] J. Starzyk, Y. Jan, and F. Qiu, "A DC-DC Charge Pump Design Based on Voltage Doublers," *IEEE Transaction on Circuits System I*, vol. 48, pp. 350–358, March 2001.

[3.16] Y. Nakagome, H. Tanaka, K. Takeuchi, E. Kume, Y. Watanabe, T. Kaga, Y. Kawamoto, F. Murai, R. Izawa, D. Hisamoto, T. Kisu, T. Nishita, E. Takeda, and K. Itoh, "An experimental 1.5 V 64Mb DRAM," *IEEE Journal of Solid-State Circuits*, vol. 26, pp. 465–472, Apr. 1991.

[3.17] T. Cho, and P. Gray, "A 10-bit, 20 MS/s, 35 mW Pipeline A/D Converter," *IEEE Custom Integrated Circuits Conference*, pp. 499–502, 1994.

[3.18] P. Favrat, P. Deval, and M. Declercq, "A High-Efficiency CMOS Voltage Doubler," *IEEE Journal of Solid-State Circuits*, vol. 33, pp. 410–416, Apr. 1998.

[3.19] K. Sato, H. Kawamoto, K. Yanagisawa, T. Matsumoto, and S. Shimizu, "A 20ns Static Column 1Mb DRAM in CMOS Technology," *International Solid-state Circuits Conference Dig. Tech. Papers*, pp. 254–255, Feb. 1985.

[3.20] Y. Tsukikawa, T. Kajimoto, Y. Okasaka, Y. Morooka, K. Furntani, H. Miyamoto, and H. Ozaki, "An Efficient Back-Bias Generator with Hybrid Pumping Circuit for 1.5 V DRAM's," *IEEE Journal of Solid-State Circuits*, vol. 29, pp. 534–538, Apr. 1994.

[3.21] K. Min, and J. Chung, "A Fast Pump-Down V_{BB} Generator for Sub-1.5V DRAMs," *IEEE Journal of Solid-State Circuits*, vol. 36, pp. 1154–1157, July 2001.

[3.22] Behzad Razavi, Principles of data conversion system design, IEEE Press, New York, 1995.

References of Chapter 4

[4.1] H. Kawaguchi, Y. Itaka, and T. Sakurai, "Dynamic Leakage Cut-off Scheme for Low-Voltage SRAM's," *Symposium on VLSI Circuits Dig. Tech. Papers*, pp. 140–141, 1998.

[4.2] R. Fujioka, K. Katayama, R. Kobayashi, H. Ando, and T. Shimada, "A Preactivating Mechanism for a VT-CMOS Cache using Address Prediction," *International Symposium on Low Power Electronics and Design*, pp. 247-250, 2002.

[4.3] K. Nii, H. Makino, Y. Tujihashi, C. Morishima, Y. Hayakawa, H. Nunogami, T. Arakawa, and H. Hamano, "A Low Power SRAM using Auto-Backgate-Controlled MT-CMOS," *International Symposium on Low Power Electronics and Design*, pp. 293-298, 1998.

[4.4] C. Kim, K. Roy, "Dynamic Vt SRAM: A Leakage Tolerant Cache Memory for Low Voltage Microprocessors," *International Symposium on Low Power Electronics and Design*, pp. 251-254, 2002.

[4.5] C. Kim, J. Kim, S. Mukhopadhyay, and K. Roy, "A Forward Body-Biased Low-Leakage SRAM Cache: Device and Architecture Considerations," *International Symposium on Low Power Electronics and Design*, pp. 6-9, 2003.

References of Chapter 5

[5.1] J. Kao and A. Chandrakasan, “Dual-Threshold Voltage Techniques for Low-Power Digital Circuits,” *IEEE Journal of Solid-State Circuits*, vol. 35, pp. 1009-1018, July 2000.

[5.2] T. Kawahara, M. Horiguchi, Y. Kawajiri, G. Kitsukawa, T. Kure, M. Aoki, “Subthreshold Current Reduction for Decoded-Driver by Self-Reverse Biasing,” *IEEE Journal of Solid-State Circuits*, vol. 28, pp. 1136-1144, Nov. 1993.

[5.3] Z. Chen, M. Johnson, L. Wei, and K. Roy, “Estimation of Standby Leakage Power in CMOS Circuits Considering Accurate Modeling of Transistors Stacks,” *International Symposium on Low Power Electronics and Design*, pp. 239-244, 1998.

[5.4] A. Agarwal, and Kaushik Roy, “A Noise Tolerant Cache Design to Reduce Gate and Sub-threshold Leakage in the Nanometer Regime,” *International Symposium on Low Power Electronics and Design*, pp. 18-21, Aug. 2003.

[5.5] E. Seevinck, F. J. List, and J. Lohstroh, “Static-Noise Margin Analysis of MOS SRAM Cells,” *IEEE Journal of Solid-State Circuits*, vol. 22, pp. 748-754, Oct. 1987.

[5.6] K. Cao, W. C. Lee, W. Liu, X. Jin, P. Su, S. Fung, J. An, B. Yu, and C. Hu, “BSIM4 Gate Leakage Model Including Source Drain Partition,” *IEEE International Electron Device Meeting Tech. Dig.*, pp. 815-818, 2000.

[5.7] A. Agarwal, H. Li, and K. Roy, “A Single-V_t Low-Leakage Gated-Ground Cache for Deep Submicron,” *IEEE Journal of Solid-State Circuits*, vol. 38, pp. 319-328, Feb. 2003.

[5.8] K. Nii, Y. Tsukamoto, T. Yoshizawa, S. Imaoka, and H. Makino, “A 90nm Dual-Port SRAM with 2.04um^2 8T-Thin Cell Using Dynamically-Controlled Column Bias Scheme,” *International Solid-State Circuits Conference Dig. Tech. Papers*, pp. 508-509, Feb. 2004.

Vita

PERSONAL INFORMATION

Birthdate:	September 19, 1980
Birthplace:	Tainan, Taiwan, R.O.C.
Address:	Department of Electronics Engineering
	National Chiao Tung University
	1001 Ta-Hsueh Rd.
	Hsinchu, Taiwan, 30050, R.O.C.
E-mail address:	cruiser.ee91g@nctu.edu.tw

EDUCATION

B.S. [2002] Department of Electrical Engineering, National Central University.

M.A. [2004] Institute of Electronics, National Chiao-Tung University.