Design of a Robust, Low-Leakage Register File for Sub-130nm Technologies

by

Kar Ting Christine Kwong

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Master of Applied Science
in
Electrical Engineering

Waterloo, Ontario, Canada, 2004

©Kar Ting Christine Kwong 2004

I hereby declare than I am the sole author of this thesis.

I authorize the University of Waterloo to lend this thesis to other institutions or individuals for the purpose of scholarly research.

Kar Ting Christine Kwong

I further authorize the University of Waterloo to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Kar Ting Christine Kwong

Abstract

The speed of microprocessors has increased dramatically over the past years. Register files (RF) are used in processors as temporary storage devices. In order to provide efficient read access to the execution cores, the size and number of register files in a microprocessor are also rising. Since a large percentage of circuit blocks in a RF are idle as leakage devices, leakage energy also increases with the trend. Furthermore, the I_{ON} to I_{OFF} ratio is decreasing with each new process generation, hence leakage energy is becoming the dominant form of energy consumption as technology scales. Therefore, a great deal of research has gone into leakage current reduction in circuit design.

In this thesis a low-leakage, 3-port (2-read, 1-write), 32-word by 32-bit RF is designed in 0.13μm CMOS technology. Various leakage control techniques, including non-minimum channel length, multi-threshold devices, and self-reverse body bias have been examined. Their implementation possibilities and tradeoffs are examined in the RF system level. In order to demonstrate the effectiveness of these leakage current techniques in larger, as well as future process generations, projections of energy consumption in 256-word by 64-bit RF and scaling trend studies with 90nm and 70nm were done. Simulations showed that the low-leakage implementation has 25% leakage reduction while RF performance degrades by 5%.

Acknowledgements

I would like to thank my supervisor **Dr. Manoj Sachdev**, for his guidance and support throughout my graduate studies. His stimulating ideas and in-depth knowledge in the subject have been most invaluable to the project.

I would also like to thank my mentor **Bhaskar Chatterjee**, who has been most helpful in solving the numerous problems encountered. His expertise in the area has been instrumental to the successful completion of the project. His guidance and help made graduate work enjoyable.

Special thanks to **Dr. Siva Sivoththaman** and **Dr. Cathy Gebotys** for making themselves available from their busy schedules to review my thesis. Their effort and time are greatly appreciated.

Finally, and most importantly, I thank my family and friends. My family played an important role and it is an honor to dedicate this work to them. They have always been supportive for every aspect of my life, especially for academic choices. My friends have also helped me in many ways, and have brought much happiness to me throughout all these years. The memories we share are so delectable that I would never forget.

Contents

1 Introduction **1**
1.1 Motivations . . . 1
1.2 Logic Styles Employed . . . 2
1.3 Transistor Sizing . . . 3
1.4 Research Contributions . . . 4
1.5 Thesis Organization . . . 5

2 Leakage Mechanisms **6**
2.1 PN Junction Reverse-Bias Current . . . 7
2.2 Subthreshold Leakage . . . 8
2.2.1 Weak Inversion Current . . . 9
2.2.2 Conduction due to DIBL . . . 9
2.3 Gate-Induced Drain Leakage (GIDL) . . . 10
2.4 Punch-Through . . . 11
2.5 Gate Oxide Tunneling . . . 11

3 Circuit Level Leakage Control Techniques **12**
3.1 Self-Reverse Body Biasing . . . 13
3.2 Multi-Threshold Voltage CMOS (MTCMOS) . . . 14
3.2.1 MTCMOS Sleep Transistor . . . 15
3.2.2 Dual-Threshold CMOS . . . 16
3.3 Reverse Body Biasing (RBB) . . . 17
3.4 Non-Minimum Channel Length Devices . . . 18

4 RF Structure and Design **19**
4.1 Address Decoders . . . 20
4.1.1 Decoder Comparisons . . . 21
4.2 Word Line Drivers . . . 23
4.3 Memory Cells . . . 24
4.4 Local Bit Line . . . 25
4.4.1 Modelling of Wide-Domino Gates . . . 26
4.5 NAND Gate . . . 33
4.6 Global Bit Line . . . 34
4.7 Discussions . . . 34

5 Low-Leakage Register File Circuit Blocks Design **35**
5.1 Address Decoder 36
5.1.1 Dual-Threshold Technique 36
5.1.2 Non-Minimum Channel Length 38
5.1.3 Gated Leakage Path with Split Decoder Signals 38
5.2 Word Line Driver 40
5.2.1 Dual-threshold Technique 40
5.2.2 Non-Minimum Channel Length 42
5.2.3 Gated V_{DD} for the Second Stage 42
5.3 Local Bit Line 44
5.3.1 Concept of Unity Gain Noise Margin 44
5.3.2 Dual-threshold Technique 44
5.3.3 Non-Minimum Channel Length 45
5.3.4 Footer Transistor 46
5.4 Single-Ended Interrupted Feedback Memory Cell 47
5.4.1 Static Noise Margin 48

6 Low-Leakage RF System and Scaling Trends **55**
6.1 Selection Methodology for Leakage Control Techniques 55
6.1.1 Decoders 56
6.1.2 Word Line Drivers 56
6.1.3 Local Bit Lines 57
6.1.4 Memory Cells 57
6.2 RF Systems and Scaling Trends 58

7 Register File Test Chip **60**
7.1 Test Chip Circuit Blocks and Layout 61
7.2 Test Chip Performance 65

8 Conclusions **67**

A Transistor Width Calculation **69**
A.1 Single Stage 5:32 Decoder 69
A.2 1:2/4:16 Split Decoder 70
A.3 2:4/3:8 Split Decoder 71
A.4 3:8/2:4 Split Decoder 72
A.5 4:16/1:2 Split Decoder 72

B Derivation of W_{MN}^{opt} **74**

List of Figures

1.1 (a) Static logic organization, (b) Static complex logic gate 3
1.2 (a) Dynamic logic organization, (b) Dynamic NAND gate 4

2.1 Leakage current mechanisms of deep-submicron transistors 7
2.2 Band-to-band tunneling of reverse biased junction 8
2.3 Subthreshold leakage and S_t for NMOS transistor 9
2.4 DIBL due to channel length reduction and increase of V_D 10
2.5 GIDL current from drain to substrate 11

3.1 Self-reverse biasing in 2-input static NAND gate 13
3.2 MTCMOS circuit structure with virtual supply lines 15
3.3 Circuit configuration for dual-V_{th} domino logic 17
3.4 Leakage current (Ioff) as a function of Body bias [17] 18
3.5 V_{th} roll-off with decreasing channel length [16] 18

4.1 Organization of a 32-word by 32-bit Register File 19
4.2 Block diagrams of a (a) single-stage and (b) two-stage 5:32 decoder 22
4.3 Layout for single stage 5:32 decoder 23
4.4 Layout for 2:4/3:8 split decoder 23
4.5 Illustration of a word line driver circuit 23
4.6 Circuit diagram of a (a) single bit line RF memory cell and (b) standard 6-T SRAM cell 24
4.7 Wide domino structure as a LBL organization 26
4.8 Normalized transistor current and capacitances 28
4.9 Wide domino gate dynamic node capacitance plot 29
4.10 Dynamic node voltage and current transients 31
4.11 Energy-delay plots for wide domino gate of n=8 33

5.1 RF system leakage energy breakup 35
5.2 Static (a) 3-input NAND and (b) 2-input NOR gates 37
5.3 1 bit circuit of a 2-stage 5:32 decoder with gated leakage path scheme . . . 39
5.4 Leakage energy distribution among the WL driver stages 41
5.5 Leakage path for a deselected WL driver in evaluation phase. 42
5.6 Gated supply for second stage WL drivers 42
5.7 UGNM measurement setup for LBL circuits 45
5.8 UGNM measurement waveforms 45

5.9 LBL leakage control footer implementation . . . 46
5.10 Schematic of (a) SEIFMC and (b) SEMC . . . 48
5.11 Illustration of two different SNM interpretation . . . 49

6.1 Energy Savings for 32-word by 32-bit RF . . . 58
6.2 Leakage energy in RF systems for various technologies . . . 59

7.1 Layout of RF test chip . . . 61
7.2 Layout of an address bit input buffer and flip-flop . . . 62
7.3 Layout of a read address decoder . . . 63
7.4 Layout of (a) Stage 1 and 2 (b) Stage 3 and 4 of WL Drivers . . . 63
7.5 Layout of (a) SEMC (b) SEIFMC . . . 64
7.6 Layout of LBL pulldown paths and NAND . . . 64
7.7 Layout of GBL and output latch . . . 65
7.8 Read operation output waveform . . . 66

A.1 Single stage 5:32 decoder . . . 70
A.2 Second stage circuit configuration of the 1:2/4:16 decoder scheme . . . 71
A.3 Circuit configuration for 2:4/3:8 decoder . . . 72
A.4 Circuit configuration for 3:8/2:4 decoder . . . 73

List of Tables

3.1 Leakage reduction by 2-, 3- and 4- transistor stacks 14

4.1 Transistor Width for different 5:32 decoder configuration 21

5.1 Truth table for 3-input NAND Gate 37
5.2 Efficiency of the 3 address decoder leakage control techniques 40
5.3 Efficiency of the 3 WL drivers leakage control techniques 43
5.4 Efficiency of the 3 LBL leakage control techniques 47
5.5 Comparison of SNM simulation and model data for SEMC 51
5.6 Comparison of SNM simulation and model data for SEIFMC 53

Chapter 1

Introduction

A register file (RF) is a memory structure that serves as a temporary data storage in a microprocessor. RF stores a wide range of data, including the operands, intermediate, as well as final results computed by an Arithmetic Logic Unit (ALU), information from the cache used by any execution units, and memory addresses of data accessed by the processor. RF may also be used to store data for specialized hardware operations. Very often in modern microprocessors, there are multiple RFs to accommodate the different types of execution units. For example, integer execution units execute integer instructions only and the corresponding RF stores data with integer format, while floating point execution units administrate floating point instructions and require bigger RF to store floating point data. Depending on the architecture, processor speed can be limited by the RF performance, which is determined by data access time. Being such an integral part in a processor, a great deal of research has been performed to improve the design of a RF.

1.1 Motivations

As notebook computers and portable devices, such as Palm-Pilots and wireless-phones, are becoming more and more popular as daily activity commodities, the market for low-power and low-cost commercial product escalates. One of the most critical issue with portable device design is battery lifetime, therefore it is of utmost importance to limit

energy consumption whenever possible. Furthermore, the functionality and performance of microprocessors are improving dramatically over the years. Processor speed is doubling every eighteen months, the number of size, read or write ports and size of register files are also proliferating. Studies have shown that register file dominate power consumption in modern microprocessors. The RF in Motorola's M.CORE architecture consume 16% of the total processor power and 42% of the data path power [1]. In addition, with device feature sizes continue to reduce, leakage mechanisms in devices began to have an impact on device behavior. Leakage energy is becoming a significant contributor to the total energy of a system. The organization of many memory systems, including RF, consist blocks made up of large number of repeated simple logic, where only a few circuits operate simultaneously. This kind of structures result in leakage energy dominant systems. As a result, it is vital to introduce effective leakage control mechanisms to RF design while maintaining its performance for applications in low-power, high-speed microprocessors.

1.2 Logic Styles Employed

There are many logic styles available to implement a logic function, each of them has their advantages and disadvantages. The RF structure studied in this project is composed of static complementary metal oxide semiconductor (CMOS) and dynamic logic styles. It is important to understand their organization and operation in order to appreciate their applications in the system.

A static CMOS gate consists of two circuit blocks, a pull-up (PUN) and a pull-down network (PDN), Figure 1.1(a) illustrate the organization for such logic. Figure 1.1 (b) shows a static complex gate of logic $F = (A' + B')C'$. Depending on the input combinations, the output node is connected to either the power supply through PUN, or ground through the PDN at steady state. During transients, there exist a short circuit path between V_{DD} and ground which is caused by non-ideal input transition times. The PUN and PDN are composed of PMOS and NMOS respectively, with them being a dual of each other, i.e. series connection for PDN translates to parallel connection in the PUN, vice versa [2]. Since, the output evaluates once the inputs change states, there is no timing control in this

logic family. Static CMOS has high noise margin, however when the fan-in becomes large, more than 3, the circuit becomes relatively slow. Therefore, dynamic logic is employed for performance critical circuit blocks.

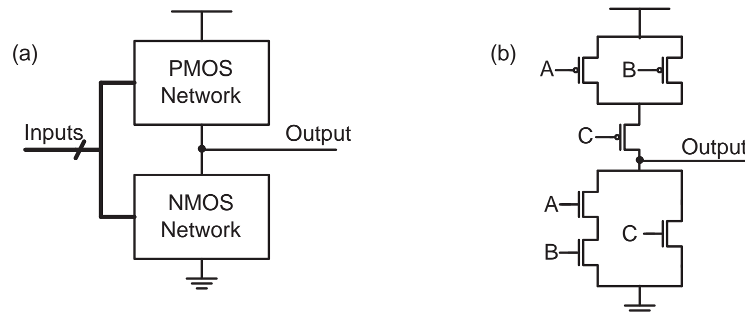


Figure 1.1: (a) Static logic organization, (b) Static complex logic gate

A dynamic logic gate is notable for its speed performance and circuit conciseness [3]. Figure 1.2(a) shows the organization of dynamic logic while Figure 1.2(b) is a dynamic NAND gate. MP1 and MN1 are controlled by a clock signal. During the precharge phase, the clock signal is low and the output is precharged to V_{DD}. When the clock changes from low to high, the circuit enters the evaluation phase. The pull-up PMOS is disabled, and the output node is discharged to ground if the NMOS logic is true. Since there is a clock signal to initiate the evaluation phase, dynamic logic is a type of synchronous logic. Due to the fact that the entire PMOS network of the static logic is replaced by a single precharge clock transistor, the capacitance at the output node is greatly reduced, resulting in a faster logic family. In addition, when the inputs to the NMOS logic tree is domino compatible, meaning the inputs are always zero during precharge phase, the footer clock transistor, MN1, can be removed to reduce clock load.

1.3 Transistor Sizing

The choice of transistor dimension has direct influence on a circuit's performance, area and power dissipation. Determining the exact appropriate transistor size for a circuit is a complex process. It involves the manipulations of transistor current-voltage equations,

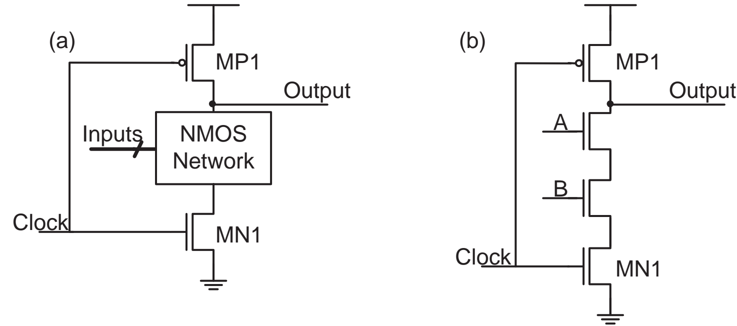


Figure 1.2: (a) Dynamic logic organization, (b) Dynamic NAND gate

capacitance at the output node, and/or capacitance at the intermediate nodes. Performing such analysis for a system like the RF is impractical. A more general approach using the concept of stage ratio is described in this section.

Stage ratio is used to size a chain of gates which drives a large capacitive load. Each successive gate is made larger than the previous one until the last inverter in the chain can drive the large load at the output of the chain. Stage ratio is the ratio of a stage, in terms of size, is increased by. The stage ratio for energy-delay optimized circuit is usually around 3, however a stage ratio of 1.5 is used for high-performance circuit [4]. A smaller stage ratio indicates that a bigger device is used to drive the load.

For symmetric rise and fall times, channel width of PMOS devices (W_P) should be made 2.5 times larger than that of NMOS devices (W_N) due to the difference in carrier mobilities, therefore $W_P = 2.5W_N$. For example, in order to calculate the correct size of an inverter with an output load of a 6μm device, the load is divided by the stage ratio 1.5, 6μm/1.5 = 4μm. This is the total width for the NMOS and PMOS, $W_P + W_N = 3.5W_N = 4$, therefore W_N=1.14μm while W_P=2.86μm. Similar approach can be used to size other circuit blocks.

1.4 Research Contributions

This work has the following contributions to the design of a register file.

1. This thesis examined various leakage control techniques, namely multi-threshold technique, non-minimum channel length and self-reverse body bias. There applications

in various major RF circuit blocks are studied, effectiveness on leakage control and switching energy reduction, as well as, degree of performance degradation are compared among the techniques.

2. Wide-domino gates are used as local bit lines and global bit lines in a RF. Special attention is required for these gates because they suffer from self-loading effect due to the large number of fan-in. In addition, they contribute significantly to overall RF delay and energy consumption. Therefore, a simple yet accurate mathematical model is developed to aid the design of energy-delay optimized wide domino gates.

3. A single-ended interrupted feedback memory cell is proposed. It is compared to the conventional single-ended memory cell in terms of static noise margin (SNM), leakage and energy consumption. A SNM mathematical model is developed for the analysis.

4. A combination of leakage control techniques is selected from the ones studied to construct a low-power, low-leakage, high-performance RF. The RF was designed, laid out and fully tested. A test chip was also completed and sent to fabrication.

1.5 Thesis Organization

The remainder of the thesis is organized as follows. Chapter 2 provides an overview on the various leakage mechanisms in a short-channel device. Chapter 3 outlines the several proposed leakage control techniques in the circuit level. Chapter 4 discusses the structure and design of a conventional RF. Chapter 5 examines the applications of leakage control techniques in individual RF circuit blocks. Chapter 6 presents a low-leakage RF design that has incorporated the most effective technique for each block examined in Chapter 5. Chapter 7 discusses the layout of the completed test chip. Lastly, Chapter 8 provides a summary and concludes the thesis.

Chapter 2

Leakage Mechanisms in Short Channel Devices

Since the dynamic power consumption of CMOS circuit is quadratically dependent on the supply voltage (V_{DD}), low-voltage operation has become very popular for low-power design. However, voltage supply reduction alone can result in performance degradation, thus transistors with lower threshold voltage are employed in low-voltage design to maintain performance. The trend to develop circuits with higher performance, and density while having lower power consumption has motivated the scaling of CMOS for the past years. As threshold voltage, channel length and gate oxide thickness are reduced, the leakage current increased drastically in the deep-submicron regimes. According to the International Technology Roadmap for Semiconductors (ITRS) static power of a CMOS circuit caused by leakage currents will dominate over dynamic power consumption unless measures are taken to reduce leakage currents [5].

The total power consumption of a CMOS circuit has two components: dynamic power and static power. Dynamic power prevails during active mode of operation. It consists of the switching power due to charging and discharging of load capacitances, and the short circuit power due to non-ideal input/output transition time [6]. Leakage currents determine the amount of static power dissipation in standby mode. Besides from increasing static power consumption, particularly for systems that have large percentage of idle circuits,

leakage currents also severely degrade noise margin. Thereby, it compromises the stability of designs, especially that of memory devices. In deep-submicron processes, there are five main types of leakage currents: PN junction reverse-bias current, subthreshold leakage, Gate-Induced Drain Leakage (GIDL), punch-through and gate oxide tunneling. These components are illustrated in Figure 2.1 [7].

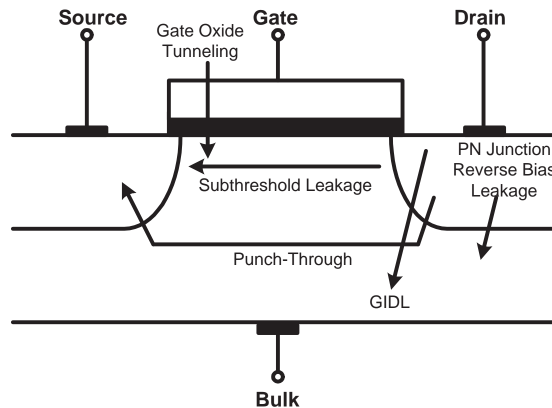


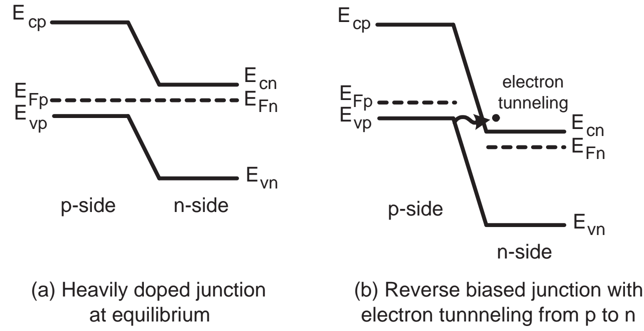
Figure 2.1: Leakage current mechanisms of deep-submicron transistors [7]

2.1 PN Junction Reverse-Bias Current

The PN junction leakage currents exist because the drain and source to bulk junctions of a MOSFET are usually reverse biased. Such reverse biased currents have two components: drifting of minority carrier near the edge of the junction depletion region and electron-hole pair generation in the depletion region. The leakage current of this nature is a function of doping concentration and junction area. In modern CMOS process, junctions are heavily doped and the reverse-bias leakage is dominated by band-to-band tunneling [8].

When a PN junction is heavily doped, the conduction band of the N-side is very close to the valance band of the P-side as shown in the PN junction band diagram in Figure 2.2. When the junction is in reverse bias, the barrier separation can become small enough for electrons to tunnel from the P-side valance band to the conduction band, causing a reverse current [9]. However, the voltage required for this breakdown to occur is usually of a few volts, PN junction reverse-bias current is negligible when compared to the overall

transistor leakage current.



Note: p. 164 streetman

Figure 2.2: Band-to-band tunneling of reverse biased junction [9]

2.2 Subthreshold Leakage

Subthreshold leakage (I_{subth}) increases exponentially with the reduction of threshold voltage (V_{th}), therefore it is the most significant contributor to leakage currents for deep-submicron processes. For low-voltage circuit design, MOS devices tend to have lower threshold voltages; the effect of subthreshold leakage is even more prominent in these applications. Subthreshold leakage region is the linear region on a semilog plot of drain current (I_D) vs. gate voltage (V_G) as shown in Figure 2.3. The inverse of the slope at the region is called subthreshold slope (S_t), it indicates how well a device can be turned off and is measured in mV per decade of current. Theoretically, the most ideal S_t is about 60mV/decade, however, the typical value for actual MOSFETS ranges from 70 to 120mV/decade [8]. I_{subth} can be modelled to by the following equations:

$$I_{subth} = A \times \exp \frac{1}{nv_T}(V_G - V_S - V_{TH0} - \gamma' \times V_S + \eta \times V_{DS}) \times (1 - \exp \frac{-V_{DS}}{v_T}) \quad (2.1)$$

with

$$A = \mu_0 C_{OX}{\prime} \frac{W}{L_{eff}} {v_T}^2 \exp 1.8 \exp \frac{-\triangle V_{TH}}{\eta v_T} \quad (2.2)$$

V_{TH0} is the zero bias threshold voltage, $v_T = kT/q$ is the thermal voltage, γ', η and n are the linearized body effect, DIBL and body effect coefficient, C_{OX} is the gate oxide capacitance, andμ_0 is the zero bias mobility. $\triangle V_{TH}$ is used to account for transistor-to-transistor leakage variations due to threshold voltage fluctuation [7]. I_{subth} consists of two major components: weak inversion current and conduction due to Drain-Induced Barrier Lowering (DIBL).

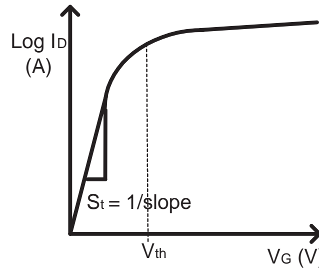


Figure 2.3: Subthreshold leakage and S_t for NMOS transistor

2.2.1 Weak Inversion Current

In the case of weak inversion, a device is driven by a V_G below V_{th} and a small voltage difference exists between drain and source. In such condition, the number of mobile carriers and the magnitude of electric field are small and current conduction is dominated by diffusion [8]. Weak inversion current is the diffusion of carriers between source and drain in a transistor when V_G is below V_{th} [6].

2.2.2 Conduction due to Drain-Induced Barrier Lowering (DIBL)

DIBL is the reduction of threshold voltage when a drain voltage (V_D) is applied. This effect is observed in short channel devices when the source and drain junction depletion regions are closely located to each other. The increase in V_D lowers the source to channel potential barrier, which results in the injection of carriers from the source into the channel independent of the gate voltage. The mechanism is illustrated in Figure 2.4, in comparison of a short channel device to a long channel one [10]. Ideally, DIBL does not affect S_t, but causes V_{th} reduction and increase in subthreshold leakage.

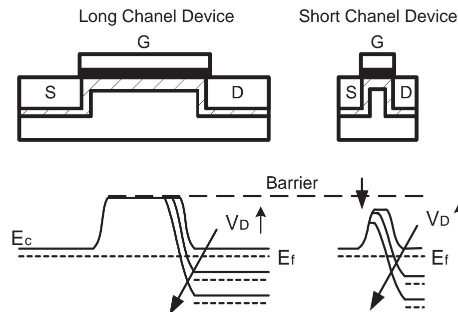


Figure 2.4: DIBL due to channel length reduction and increase of V_D [10]

2.3 Gate-Induced Drain Leakage (GIDL)

Gate-Induced Drain Leakage (GIDL) is caused by the high electric field under the gate and drain overlap region. This occurs when V_G is negative or zero while V_D maintains high. Under this condition, a NMOS transistor is biased in the accumulation mode while holes from the p-substrate accumulates under the gate at the silicon surface. This higher concentration of carriers causes the depletion width near the silicon surface to be thinner, resulting in an increase in local electric field.

With a high V_D, the overlap region can become depleted or even inverted resulting a more intense electric field. Band-to-band tunneling and avalanche breakdown are the consequences of high electric field. Minority carriers are generated and emitted into the channel, which are then swept towards the substrate that has a lower potential, thus completing the leakage path [8]. Figure 2.5 illustrates the GIDL mechanism. GIDL becomes more important as device dimensions scale down, it is because thinner gate oxide will result in even higher electric field, and aggravating the effect. GIDL contribution to the overall leakage current is small in nominal operating condition. However, it plays a significant role when V_D is raised close to burn-in voltage [7]. Therefore the effect of GIDL is particularly evident during the testing phase of a device.

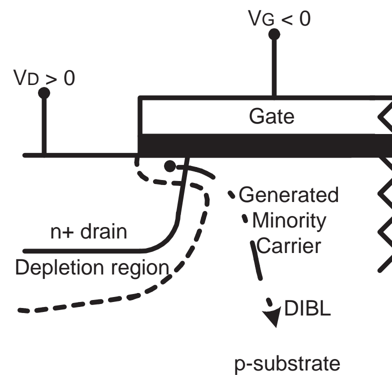


Figure 2.5: GIDL current from drain to substrate [8]

2.4 Punch-Through

In short channel devices, the drain and source junctions are in close proximity. Punch-through occurs when V_D is increased to such a point (punch-through voltage) that the depletion region extends through the channel and reaches the source junction. V_D lowers the potential barrier for the majority carriers in the source; causing more carriers to be emitted into the substrate and collected by the drain. This results in an increase in sub-threshold current and degradation of the subthreshold slope [8]. This leakage mechanism is also known as the subsurface version of DIBL.

2.5 Gate Oxide Tunneling

Gate oxide tunneling is caused by a high electric field in the gate oxide and can affect the reliability of a device. It contains two tunneling mechanisms: Fowler-Nordheim (FN) tunneling through the oxide bands and direct tunneling through the gate [7]. FN tunneling is negligible for the normal operation since the field strength required is very high. Direct tunneling on the other hand can become significant for thin oxides of with 3-4nm. It is the tunneling of electrons from the silicon surface into the gate through the forbidden energy gap of silicon oxide. Gate oxide tunneling is not a major issue in technology nowadays, however it will dominate over weak inversion and DIBL as oxide becomes thinner in future process generations [7].

Chapter 3

Circuit Level Leakage Control Techniques

For a block of logic gates which are made up of transistors connected in series and/or in parallel, the total leakage power can be estimated as the sum of each gate. The leakage power for a logic gate can be expressed as

$$P_{leak} = V_{DD} * I_{leak} \tag{3.1}$$

with V_{DD} being the supply voltage and I_{leak} is the leakage current. It has been shown that overall leakage is highly dependent on the input vectors. This is because input vectors determine the transistor configurations, and the number of "OFF" transistor determines the amount of leakage current drawn [8]. In general, low leakage input vectors can turn off as many transistors as possible in each leakage path. When a circuit is in standby mode, the "OFF" transistors draw the leakage current while the "ON" transistors complete the conduction path from power supply or a "HIGH" node to ground. Apart from leakage control using input vector activation, several other circuit techniques have been proposed. Some of these will be addressed in this chapter: (1) self-reverse body biasing; (2) Multi-threshold process; (3) reverse body biasing; (4) non-minimum channel length devices. Since subthreshold leakage is the most prominent leakage mechanism among the different types previously discussed, the thesis from here onwards will focus on subthreshold leakage reduction techniques.

3.1 Self-Reverse Body Biasing

Self-reverse body biasing is also know as transistor stack effect. It refers to the reduction of leakage in a transistor stack when more than one serially connected transistors is turned off [8]. Consider the NMOS stack in a 2-input NAND gate, as shown in Figure 3.1.

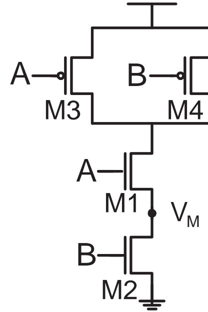


Figure 3.1: Self-reverse biasing in 2-input static NAND gate

When input A and B are both at ground, M_1 and M_2 are "OFF", and the voltage at the intermediate node (V_M) is positive. The following effects can be observed [12]:

1. A positive V_M at the source of M1 makes the gate-to-source voltage, V_{GS1}, negative. As modelled in Equation 2.1, subthreshold leakage is exponentially dependent on V_{GS}. With a negative V_{GS}, the subthreshold leakage is reduced significantly.

2. Threshold voltage can be expressed as

$$V_{th} = V_{fb} + 2\Psi_B + \frac{\sqrt{2\xi_{si} q N_a (2\Psi_B + V_{sb})}}{C_{OX}} \quad (3.2)$$

   with V_{fb} as the flat-band voltage which is equal to the difference of work function between the gate metal and the substrate. There is no net charges in the semiconductor, and no voltage drop at this point. Ψ_B is the Fermi potential difference between the intrinsic silicon and the substrate, ξ_{si} is the permittivity of silicon, V_{sb} is the source-to-body voltage, and C_{OX} is the gate capacitance [8]. When the body-to-source junction of a MOS is reverse biased, V_{sb} becomes positive and V_{th} increases. With V_M being positive, threshold voltage of M1 increases, and subthreshold leakage reduces.

3. With a positive V_M, The drain-to-source voltage of M1 reduces. DIBL effect is diminished and thus also decreases the subthreshold leakage [12].

It has been found that the leakage current for a 2-stack "OFF" transistor is an order of magnitude (10 times) smaller than that of a single "OFF" transistor. If a stack of 3 or 4 transistors is used, the leakage reduction obtained can be up to 20 to 30 times as shown in Table 3.1 for 0.1 μm process at 30° as presented in [12].

Table 3.1: Leakage reduction by 2-, 3- and 4- transistor stacks

	High V_{th}	Low V_{th}
2 NMOS	10.7×	9.96×
3 NMOS	21.1×	18.8×
4 NMOS	31.5×	26.7×
2 PMOS	8.6×	7.9×
3 PMOS	16.1×	13.7×
4 PMOS	23.1×	18.7×

Base on the above observations, one can see that transistors can be inserted in the paths where it is not possible to have more than one "OFF" transistor to reduce leakage. However, when the path is to be evaluated (all inputs are high for NMOS stack), the output will have longer delay time due to reduction in current drive [13]. Therefore, insertion of stack transistors needs to be discrete and should minimize its affect on a design's performance. In general, transistor insertion for leakage control should only be utilized in data paths that are non-critical.

3.2 Multi-Threshold Voltage CMOS (MTCMOS)

Many variations of Multi-threshold technologies have been proposed in recent years for leakage reduction. The primary methodology is to use both high V_{th} and low V_{th} in a single design. High V_{th} MOS are used as leakage control devices, while low V_{th} MOS

are used in the critical paths for speed performance [8]. In this section, MTCMOS sleep transistor and dual-threshold CMOS design techniques will be discussed.

3.2.1 MTCMOS Sleep Transistor

In this circuit scheme, high V_{th} devices are used as sleep transistors to control the power supplies of low V_{th} logic gates, as illustrated in Figure 3.2 [14]. When in active mode, the Sleep-Control (SC) signal is low which turns on both MS1 and MS2 and the logic gates composed of low V_{th} are evaluated through the virtual supplies. The addition of sleep transistors in series with the logic usually result in performance degradation, therefore the sleep transistor devices are typically very large and there is a silicon-area tradeoff associated with this technique. When in standby mode, SC becomes high and the gating transistors are off. Since an extra high V_{th} "OFF" transistor now exist between the logic and the power lines, subthreshold leakage becomes really small [14].

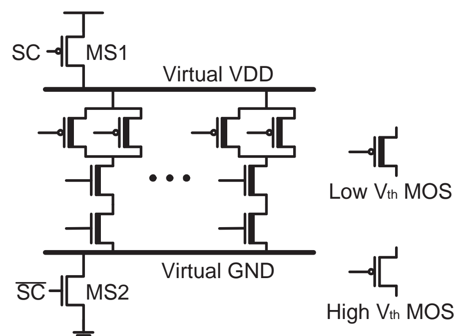


Figure 3.2: MTCMOS circuit structure with virtual supply lines

The circuit scheme depicted in Figure 3.2 shows two gating transistor, in actual fact, only one is required for static logic. Some proposed that NMOS sleep transistor is a better choice because it has lower "ON" resistance. As such, NMOS sleep transistors can be sized smaller compare to the PMOS counterpart [8]. One can also insert sleep transistor according to logic configuration. For example, logic with PMOS stack (such as a NOR gate) can use a NMOS sleep transistor, while logic with NMOS stack (such as a NAND gate) can use a PMOS sleep transistor. This is because most leakage occurs through single transistors, and by gating these paths, a large percentage of the leakage is reduced.

In conclusion, MTCMOS sleep transistor is a very attractive leakage control technique for combinational logic. It can be easily incorporated into existing design by inserting appropriately sized sleep transistors. These transistors are usually large and will increase area consumption. Furthermore, extra memory circuitry is required for data retention function in standby mode when the virtual supplies are disconnected [14]. Due to this property, MTCMOS in sequential circuit is not as easily implemented as combinational logic.

3.2.2 Dual-Threshold CMOS

Dual-threshold CMOS technique differs from MTCMOS as it does not require the additional sleep transistors, which are in series with the logic unit and can cause extra delay. Dual-threshold CMOS uses dual V_{th} transistors in an existing logic, leakage reduction is achieved by placing high V_{th} devices in non-critical paths, while low V_{th} transistors are used in the critical paths. Since the critical paths are still consist of low V_{th} devices, as in the case for single V_{th} design, the performance of the entire circuit is unaffected [8]. However, for bigger systems, the determination of criticals can be very complex.

Embedded dual-V_{th} design for domino circuits make use of the single transition properties of domino gates for low V_{th} devices assignment. The logic tree, which will be evaluated during the evaluation phase and is performance critical, are composed of low V_{th} devices. The precharge devices and the keeper use high V_{th} transistors as shown in Figure 3.3 [15]. Since precharge phase is non-performance critical and can tolerate a slower transition time, the usage of high V_{th} for those devices has no performance implications. Standby mode is achieved when all the inputs to NMOS tree are at high to anchor the dynamic node to ground and avoid short circuit current through the subsequent inverter. More importantly, the precharge devices are turned “OFF” so that leakage paths are through the high V_{th} transistors [15].

Dual-V_{th} design can maintain the speed performance of single V_{th} design, but with longer precharge time. When compare to MTCMOS, it is much easier to implement without the performance and area overhead.

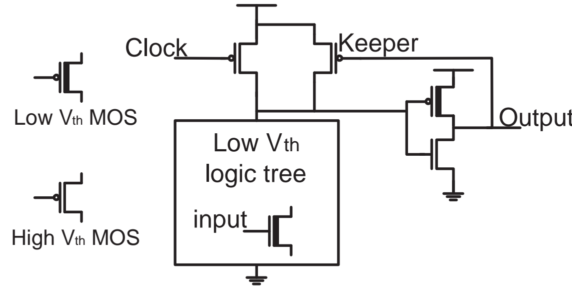


Figure 3.3: Circuit configuration for dual-V_{th} domino logic

3.3 Reverse Body Biasing (RBB)

A decrease in V_{th} can cause exponential increase in subthreshold leakage current. With this relationship in mind, leakage reduction can be achieved by modulating the V_{th} through applying a negative (positive) voltage to the body of NMOS (PMOS). This technique is called reverse body biasing (RBB). The relationship between V_{th} and source-to-body voltage (V_{sb}) can be observed in Equation 3.2, V_{th} increases along with V_{sb}. It has been demonstrated that an optimal RBB voltage exits, at which the circuit can achieve maximum leakage reduction. Beyond this optimal point, leakage current remains stable and may even increase with RBB voltage. This phenomenon is illustrated in Figure 3.4, adopted from [17].

RBB can modulate V_{th} to the advantage of leakage control due to body effect; however, its effectiveness is highly dependent on the degree of short channel effects (SCE). As discussed in the previous chapter, transistors with shorter channel length have a lower V_{th} and higher leakage due to DIBL and V_{th} roll-off. As transistors become smaller, SCE becomes more prominent and the body effect weakens, which diminishes the efficiency of RBB [18]. Therefore, one can conclude that RBB leakage control technique does not have a good scalability as its efficiency decreases as device dimension becomes smaller. Furthermore, its implementation poses great challenge since extra power supplies are required and additional power rails are to be routed throughout a chip, this technique is has a large overhead.

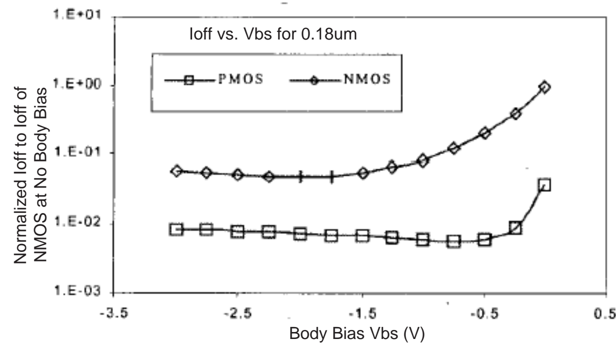


Figure 3.4: Leakage current (Ioff) as a function of Body bias [17]

3.4 Non-Minimum Channel Length Devices

Leakage control through device channel length adjustment is base on V_{th} roll-off SCE. Figure 3.5 is adopted from [16], it illustrates the relationship of feature size and V_{th}. It can be observe that, V_{th} increases almost linearly with channel length in the roll-off region, and thus achieving leakage reduction. Since leakage current is exponentially dependent on V_{th}, this technique is very effective. However, an increase in channel length, will result in higher gate capacitance, which have a negative effect on circuit performance and energy consumption [8]. Therefore, non-minimum feature size devices, for leakage control purposes, should be applied discretely to avoid performance degradation and unacceptable energy increase.

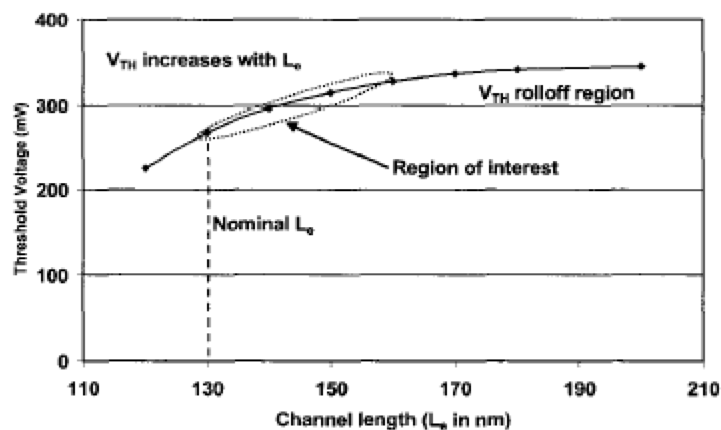


Figure 3.5: V_{th} roll-off with decreasing channel length [16]

Chapter 4

Register Files Structure and Design

The register file structure under investigation has a capacity of 32-word by 32-bit, and it can be accessed by 2-read ports and 1-write port. Figure 4.1 illustrates the organization and various functional blocks in the RF. It contains mainly of 3 sections, the read port, the write port and the memory array bit slices. The read and write functions are explained in details below.

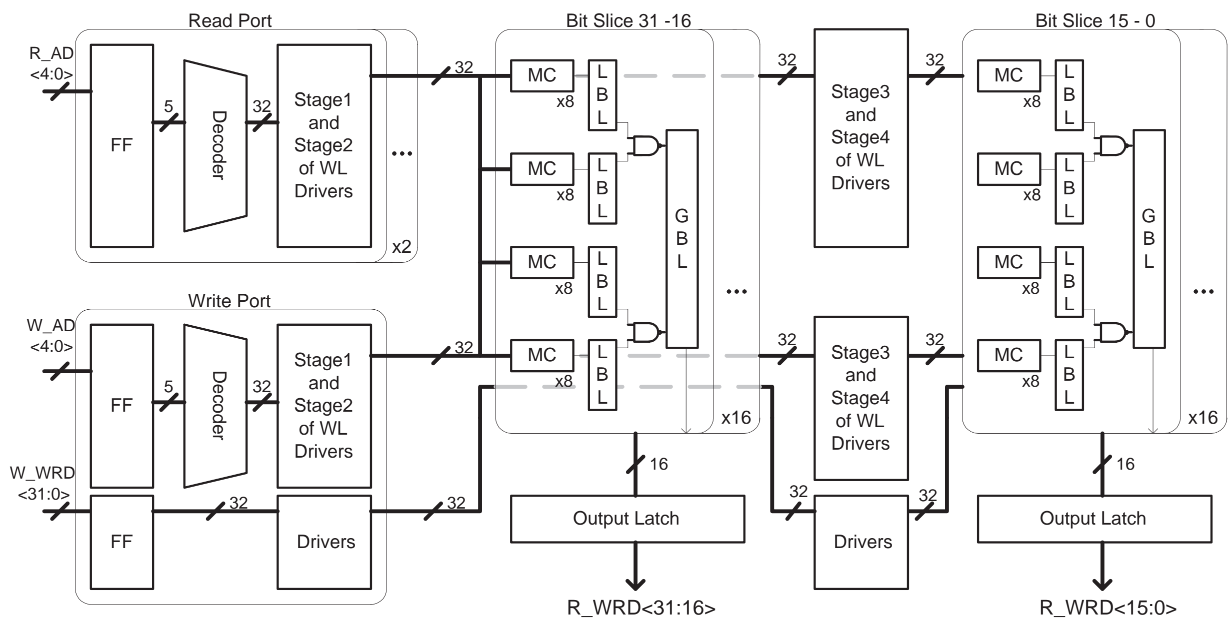


Figure 4.1: Organization of a 32-word by 32-bit Register File

The read operation is the performance critical function, it determines the maximum

operating frequency of the RF. The system is in read mode when the "Read Enable" control is active. The address signals ($R_AD < 4:0 >$) are latched into the input flip-flops, where the output is fed into the read address decoder. The decoder then determines which of the 32 word lines will be activated. The outputs are then past onto the word line drivers which propagate the signals through the 32 bit slices. Only data stored in the memory cells of the active word line is evaluated through the local bit lines, static NAND gates and global bit lines. Finally, the global bit line outputs are captured by the output latches and available to other circuits in the microprocessors as ($R_WRD < 31:0 >$). The two read ports can operate simultaneously, reading from the same or different word.

The write operation begins with the "Write Enable" being active. The write address signals ($W_AD < 4:0 >$) are decoded and the selected word line signal is propagated into the memory array similar to the read operation. In the meantime, write data ($W_WRD < 31:0 >$) are also driven into the memory array by buffer stages. The data is stored into the memory location at which the word line is active. Since a write operation goes through fewer circuit stages than a read operation, the speed of the system is determined by the read paths.

The following subsections detail the operation and design process of each major circuit block, mainly address decoder, word line drivers, memory cells, local bit lines and global bit lines. The discussion here focuses on the design of a high-speed, energy efficient base model, it includes design choices and the reasons to support it. Leakage control mechanisms are incorporated with the base model for a low-leakage design and is discussed in the next chapter.

4.1 Address Decoders

A 32-word by 32-bit RF requires a 5:32 decoder for each of the read or write port. Only one of the 32 decoder outputs is selected by any input combination and has a logic value of high. The decoded signals are used to enable or disable the read or write operation for each of the thirty-two words. The timing constraints for an address decoder is relatively relax, this is because it has the entire half cycle (when the clock is low) to evaluate. Therefore, it

can be implemented using static CMOS logic. The circuit remains the same for all the bit positions, the only difference is the input combination. For example, word 0 is enabled by the address A4́A3́A2́A1́A0́, while the address representation for Word 31 is A4A3A2A1A0. The decoder can be of single-stage or two-stage designs, called split decoder scheme [19] as 1:2/4:16, 2:4/3:8, 3:8/2:4, or 4:16/1:2. Taking the 2:4/3:8 decoder as an example, the two most significant address bits are first decoded by a 2:4 decoder to form four enable signals for the subsequent four 3:8 decoders. All of the outputs of the second stage decoders constitute the 5:32 decoder outputs.

4.1.1 Comparisons Across Different Architecture

Due to the presence of various configuration alternatives, a study is done to investigate which structure is the most favorable. Firstly, the total transistor width is compared across all architectures. In order to obtain iso-performance, a constant load and overall sizing stage ratio is used. Details of sizing and transistor width calculation of each of the configuration can be found in Appendix A. It has been demonstrated that the total transistor width, including input buffer widths, varies among the configurations as illustrated in Table 4.1.

Table 4.1: Transistor Width for different 5:32 decoder configuration

	Transistor Width (μm)
5:32	7825
1:2/4:16	9500
2:4/3:8	7822
3:8/2:4	7813
1:2/4:16	9500

By comparing the total transistor width, one can determine which configuration is the most energy efficient. When there are less transistors or when device features are smaller, the switching capacitance for each evaluation decreases. Since switching energy is

proportional to the switching capacitance, lower capacitance will result in smaller switching energy. Similarly, the total width for "OFF" transistors will be smaller, which can translate to a smaller leakage current. It can be observed that among the split decoders, 2:4/3:8 and 3:8/2:4 decoder have the smallest total transistor width while 1:2/4:16 or 4:16/1:2 split decoders are the least favorable with 24% more when normalized to the 2:4/3:8 decoder.

All configurations listed in 4.1 are of two-stage split decoder format except for the single stage 5:32 decoder. A study was carried out to compare the single-stage decoder to a split decoder, namely the 2:4/3:8 configuration. Figure 4.2 shows the block diagrams of the two configuration. Four parameters, including performance, switching energy, leakage energy and layout area requirement are compared. Simulation results show that the worst-case switching energy and delay differences are negligible (within 1%), and 2-stage configuration has a slightly smaller leakage current (2%). Two separate layouts are completed for area consumption comparison. Single stage decoder has a dimension of 320μm by 55μm, whereas that for a 2:4/3:8 split decoder is 292μm by 64μm. The layouts are included in Figure 4.3, 4.4 respectively. The 2:4/3:8 decoder consumes 8% more silicon area since it has two tracks of input signals and it has unused area due to un-match pitch between the first and second stage decoders. Despite the area disadvantage, the two-stage decoder is preferred in the RF design. It is because the split decoder signals (outputs of the first stage) can be used as leakage control signals, details will discussed in the following chapter.

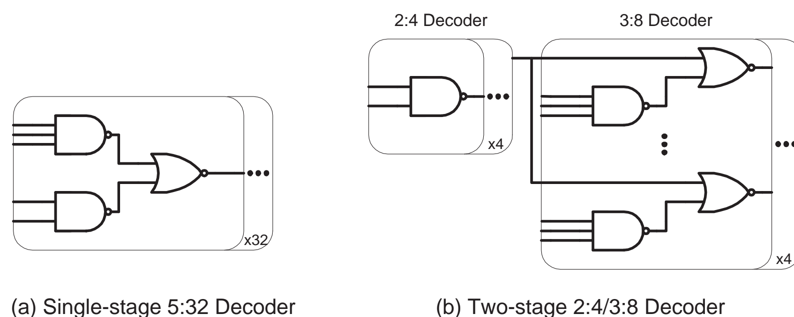


Figure 4.2: Block diagrams of a (a) single-stage and (b) two-stage 5:32 decoder

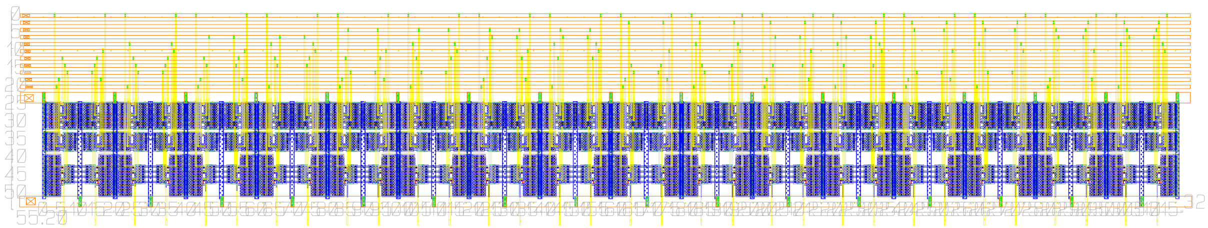


Figure 4.3: Layout for single stage 5:32 decoder

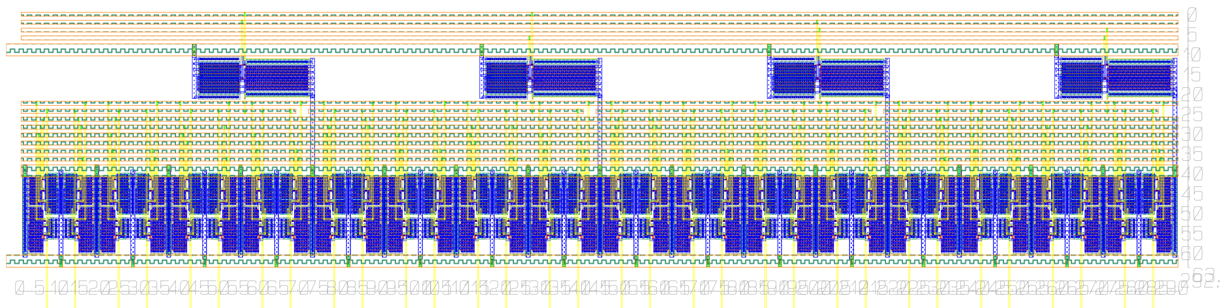


Figure 4.4: Layout for 2:4/3:8 split decoder

4.2 Word Line Drivers

Word line (WL) drivers are the first circuit stage to evaluate when the clock signal changes from low to high. These drivers are needed to drive the decoder outputs to all the bit slices of a word. Each WL driver consists of four stages of large inverters sized to drive the large load presented by the local bit lines of the read paths or memory cells of the write path. Due to the presence of a large load, the drivers are segmented into two sections. Stage 3 and stage 4 are responsible to drive the load incurred by bit 15 to bit 0, while stage 1 and stage 2 see a load presented by bit 31 to bit 16, as well as the input capacitance of driver inverter stage 3. The organization of a WL driver is shown in Figure 4.5.

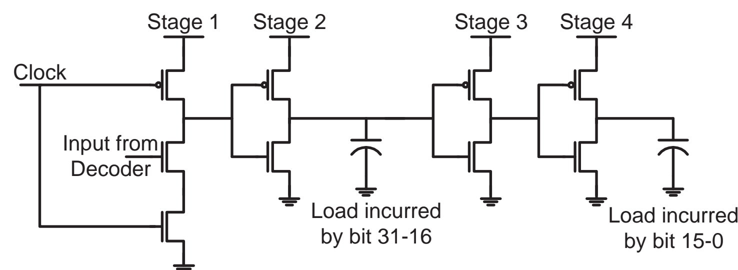


Figure 4.5: Illustration of a word line driver circuit

Since only one word line among the 32 is selected, 31 of the drivers are deselected and leaking. Because of large feature sizes and high percentages of inactive circuits, leakage contribution by WL drivers are high. Therefore, it is one of the circuit block to focus on for leakage control application.

4.3 Memory Cells

The basic requirements of memory cells are small area consumption for higher density and high static noise margin. Conventional RF uses single bit line memory cells, which is similar to a 6-T SRAM memory cell, where data retention is achieved through the two back-to-back inverters. The major difference is that SRAM uses pass transistor to control the access [2], while RF cells uses Local Bit line (LBL) circuits instead. The output of memory cells are used to control the conduction of LBL circuit paths. Furthermore, single-ended memory cells only require single ended data, while 6-T SRAM cell needs both the true and complement data signal, which demands a higher wiring complexity. Figure 4.6 illustrate the circuit of a single bit line RF cell and a 6-T SRAM cell.

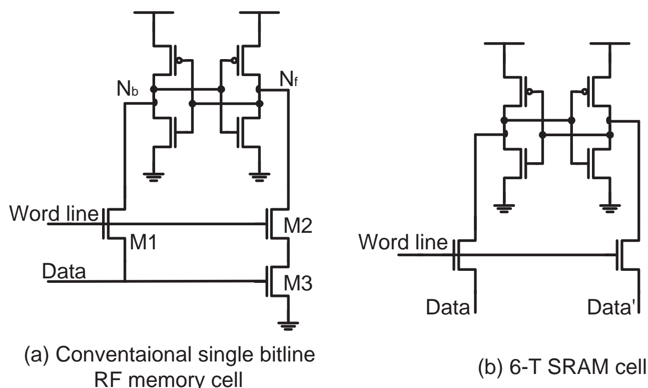


Figure 4.6: Circuit diagram of a (a) single bit line RF memory cell and (b) standard 6-T SRAM cell

When the word line for the particular word entry asserts a high value, the two word line NMOS (M1 and M2) are enabled and passes the data value into the two back-to-back inverters. When the data is low, the output node of the forward inverter (N_f) is driven to a high, which turns on the NMOS of the backward inverter driving (N_b) to a low. On

the other hand, when the data is high, M1 will not be sufficient to invert the forward inverter since there is a V_{th} drop across NMOS when passing a high due to body effect [2]. Therefore, M3 is there to pass a strong zero through ground to discharge N_f and invert the backward inverter.

Leakage energy dominates in the memory core during a read operation, since all memory cells are quiescent. Even in a write operation, only one of the 32-word entries will be switching as only one word line will be active, 31×32 memory cells continue to leak. Therefore, it is imperative to design low-leakage memory cells for a low-leakage RF. A novel memory cell will be presented in the subsequent chapter.

4.4 Local Bit Line

Local bit line (LBL) circuits are wide-domino logic gates. The pulldown logic consists of parallel 2-NMOS transistor stacks. Each bit slice has four LBL circuits, therefore each circuit will have eight parallel paths of 2 transistor in series. The gate of the upper transistors are connected to word line signals, while the gate of the bottom transistors are connected to memory cells. Figure 4.7 illustrates a LBL circuit. The state of the bottom transistors are determined by the data stored in the memory cell, which is available ahead of the LBL evaluation time. The word line signals are from the WL drivers, which is valid after the clock signal turns high and after the driver chains have been evaluated. Since only one WL is active, and providing that the corresponding data is high, only one of the pulldown path in each bit slice will be conducting at any times, therefore a LBL circuit has a function of a multiplexer. Furthermore, as illustrated in 4.7, the footer transistor of a domino gate is missing. This is because the first stage of WL drivers are domino inverters, WL signals generated are domino compatible. As a result, the clocked footer transistor of a LBL can be removed [20] without any direct short circuit current from the supply line to ground. This is beneficial in terms of clock load reduction.

Since at most only one path of a LBL structure will evaluate in a bit slice and the remaining thirty-one paths are quiescent and leaking, LBL's are a significant contributor to leakage energy. A keeper is required to compensate for the charge leaking through the

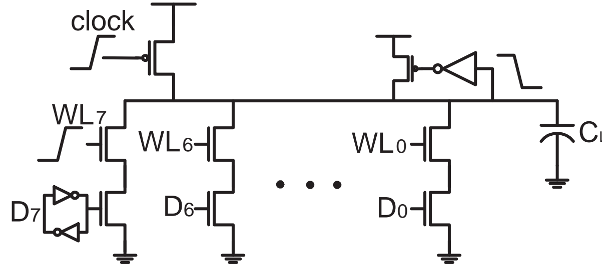


Figure 4.7: Wide domino structure as a LBL organization

"OFF" paths. It ensures that the DC robustness is under acceptable region. However, by introducing a strong keeper, the contention during evaluation causes a short-circuit current through the PMOS keeper to ground. To minimize this effect, it is best to reduce leakage currents to allow the use of a smaller keeper.

LBL's are often characterized by its high dynamic node capacitance and degraded evaluation fall time. For this reason, apart from leakage energy, these circuits contribute significantly to the overall switching energy and delay of a RF system. The following section presents a model, suitable for hand calculation, to design a energy-delay optimized wide domino circuit of a LBL structure.

4.4.1 Modelling of Energy-Delay Optimized Wide-Domino Logic Gates

A model for wide-domino logic gate is needed because of the high dynamic node capacitance and low driving capability of LBL. The optimal fan-out for optimized energy-delay product (EDP) differ from a conventional domino gate. Domino logic with high fan-in suffers from self-loading effect because the multiple parallel pulldown paths contribute to the dynamic node capacitance (C_{DYN}) through their gate-drain capacitances (C_{GD}). Upsizing the pulldown transistors is an ineffective approach to improve performance since only one pulldown path is enabled for LBL structures. The overall C_{GD} increases more than proportionately when compare to the driving capability of the upsized pulldown transistors [21]. The upsized transistors which are supposed to enhance the evaluation performance,

ended up loading the dynamic node and increasing C_{DYN}. This effect does not exist in conventional dynamic gates because they do not have high fan-in.

4.4.1.1 Transistor Parameter Extraction

Transistor parameter extractions is needed to develop a relationship between transistor width and the following parameters, C_{GD}, C_G and saturation current (I_{DSAT}). Figure 4.8 shows the plot for the above parameters normalized to data of 0.5μm transistor width . It can be observe that there exist a linear relationship, the transistor width increases along with the current and capacitances. These parameters can be modelled by the following equations:

$$I_{DSAT} = K_I^N W_{MN} \tag{4.1}$$

$$C_{GD} = K_{GD}^N W_{MN} \tag{4.2}$$

$$C_G = K_G^N W_{MN} \tag{4.3}$$

with K_I^N, K_{GD}^N and K_G^N be the constants if proportionality and W_{MN} be the transistor width. The absolute values of these constants are technology dependent, for the particular technology used, they have the following values: K_I^N=67μA/μm, K_{GD}^N=0.84fF/μm and K_G^N=0.83fF/μm. Although the above equations are developed for NMOS transistors, similar approach can be used to characterize PMOS devices.

4.4.1.2 Dynamic Node Capacitance Model

The energy and delay of a dynamic gate is closely related to the total capacitance at the dynamic node (C_{DYN}). It consists of two major components: C_{GD} which is contributed by the pull down paths, PMOS keeper and the precharge clock transistor, and C_G of the subsequent circuit stage and the keeper inverter. These relationship can be expressed by the following equations:

$$C_{DYN} = C_{GD}^{TOT} + C_G^{TOT} \tag{4.4}$$

$$C_{GD}^{TOT} = C_{GD}^{MP1} + C_{GD}^{MPK} + \sum_{i=1}^{n} C_{GD}^{MNi} \tag{4.5}$$

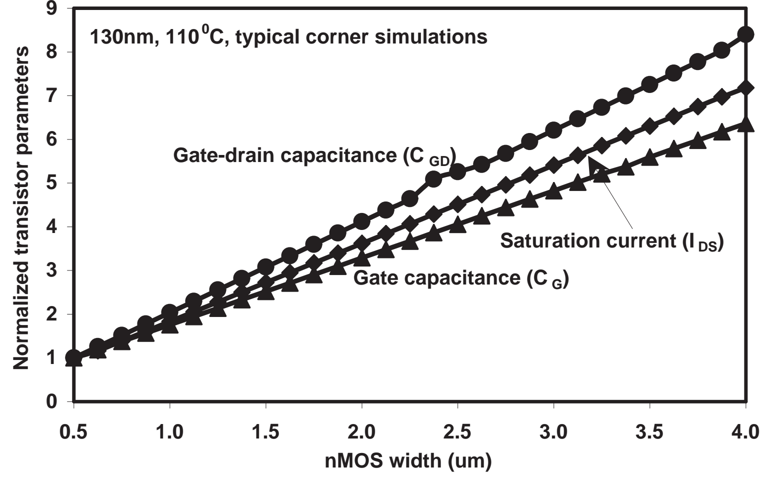


Figure 4.8: Normalized transistor current and capacitances

$$C_G^{TOT} = C_G^{kpr_inv} + C_L \tag{4.6}$$

the superscripts correspond to transistors as specified in Figure 4.7, "n" represents the total number of parallel paths. n=8 for the eight parallel pulldown paths of the LBL circuit under investigation. A C_{DYN} expression in terms of transistor width can be obtained if Equation 4.1 to Equation 4.6 is combined.

$$C_{DYN} = K_{GD}^P(W_{MP}^{MP1} + W_{MP}^{MPK}) + nK_{GD}^N W_{MN} + K_G^P W_{MP}^{kpr_inv} + K_G^N W_{MN}^{kpr_inv} + C_L \tag{4.7}$$

where the subscript MN and MP correspond to NMOS and PMOS transistor width respectively. Generally, the precharge transistor and the keeper inverter are significantly smaller than the effective pulldown width, they can be lumped with the load capacitance to form a lumped capacitive load (C_L'). Furthermore, the keeper size has a fix relation to the total pulldown transistor width (3% in this particular design) to maintain a targeted robustness, Equation 4.7 can be further simplified.

$$C_{DYN} \approx (0.03K_{GD}^P + K_{GD}^N)nW_{MN} + C_L' \tag{4.8}$$

Figure 4.9 compares C_{DYN} obtained from SPICE simulations with the model discussed above for two loading conditions as a function of pulldown transistor sizes. It is shown that the model closely tracks the SPICE simulation with only 3% error.

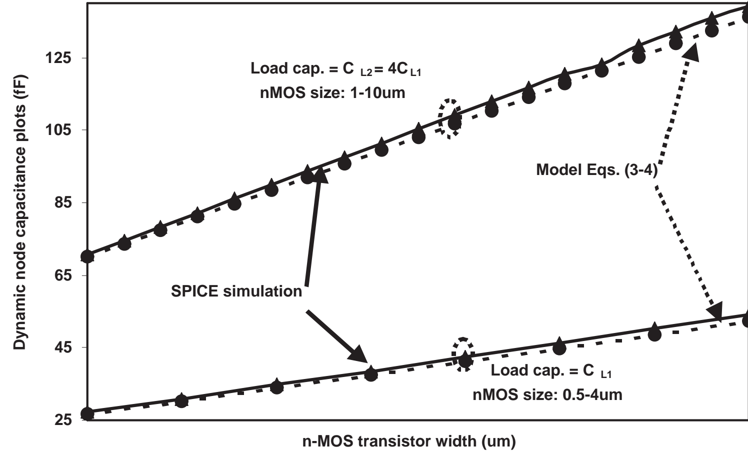


Figure 4.9: Wide domino gate dynamic node capacitance plot

4.4.1.3 Energy Model

The total energy (E_{TOT}) per transition of a logic gate consists of three different types of energy, namely switching, short circuit and leakage energy. A model is developed for each of them and are presented as follows:

1. Switching Energy

   Switching energy (E_{SW}) is governed by the total switching capacitance during a transition. The capacitances that switches are the gate of the enabled pulldown NMOS transistor (C_G^{MN}), the dynamic node capacitance (C_{DYN}), and the capacitance associated with the output node of keeper inverter, including C_G^{MPK} and $C_{GD}^{kpr_inv}$. E_{SW} can be expressed as:

   $$E_{SW} = C_{TOT}V_{DD}^2 = (C_G^{MN} + C_{DYN} + C_{GD}^{kpr_inv} + C_G^{MPK})V_{DD}^2 \qquad (4.9)$$

   Using Equation 4.1 to 4.3, E_{SW} can be represented as a function of transistor width.

2. Short Circuit Energy

   The degraded fall time of a wide domino gate causes a short circuit current to flow through the PMOS keeper and the enabled NMOS pulldown transistor during evaluation. According to short circuit model in [22], it can be modelled as:

   $$E_{SC} = \frac{I_{SC}t_{SC}}{2}V_{DD} \qquad (4.10)$$

where I_{SC} is the short circuit current and t_{SC} is the short circuit interval. Figure 4.10 shows the voltage and current transient waveforms when the circuit just begins to evaluate with the WL driver signal turns high. I_{SC} can be approximated by the PMOS keeper saturation current as $K_I^P W_{MP}^{MPK}$. Since the keeper bears a 3% relationship to the total pulldown width, it can be expressed in terms of pulldeown transistor width, $0.03nK_I^P W_{MN}$. Simulation results show that t_{SC} begins when the input WL_7 reaches V_{th} of NMOS, and ends when the keeper inverter output reaches $V_{DD} + V_{th}$ of PMOS. t_{SC} can be approximate by the follow:

$$t_{SC} = t_{50\%}^{kpr_inv} - t_{50\%}^{WL_7} + \frac{1.25}{V_{DD}}[t_r^{kpr_inv} V_{thN} - t_r^{WL_7}(V_{DD} + V_{thP})] \quad (4.11)$$

where $t_{50\%}^{kpr_inv}$, $t_{50\%}^{WL_7}$ are the 50% V_{DD} crossover point, while $t_r^{kpr_inv}$ and $t_r^{WL_7}$ are the rise time of the inverter and input signal respectively. The above expression is obtained by linearizing the rise time to model the signal timing behavior. It should be noted that the keeper inverter are sized according to the keeper feature size, therefore inverter delay remains approximately the same. WL drivers are also sized with a constant stage ratio, therefore its rise time also remains approximately invariable. As a result, the value of t_{SC} is very close to a constant across a wide range of load conditions and pulldown transistor width.

3. Leakage Energy

The total leakage current of a wide domino circuit is determined by the number of "OFF" pulldown paths and the size of the pulldown transistors. For a worst-case analysis, assume all the data in the memory cell is storing a high value, as such leakage currents are only leaking through the top transistors of the stacks and only one path is enabled, meaning (n-1) paths are leaking. Once the leakage parameter, I_{off}/μm is obtained, the total leakage current at $110°C$ can be expressed as:

$$I_{Leakage}^{TOT} = (n-1)K_I^N W_{MN} \frac{I_{off}}{I_{on}}^{110°C} \quad (4.12)$$

The leakage energy is dependent on the integration interval (t_{int}) which usually represents the switching interval. Therefore, leakage energy can be modelled as:

$$E_{Leakage} = I_{Leakage}^{TOT} V_{DD} t_{int} \quad (4.13)$$

With all of the energy component appropriately modelled, the mathematical representation of E_{TOT} becomes:

$$E_{TOT} = E_{SW} + E_{SC} + E_{Leakage} = C_{TOT}V_{DD}^2 + \frac{I_{SC}t_{SC}}{2}V_{DD} + I_{Leakage}^{TOT}V_{DD}t_{int} \quad (4.14)$$

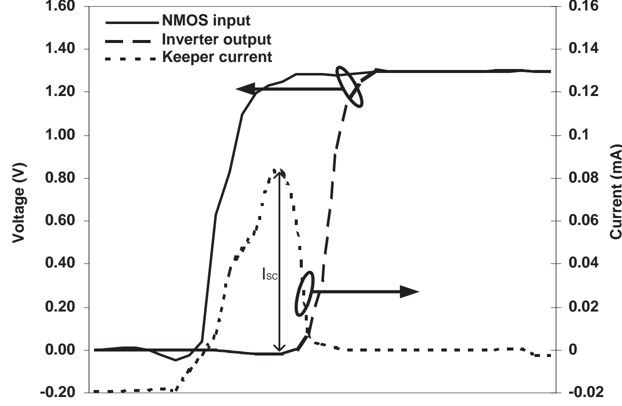


Figure 4.10: Dynamic node voltage and current transients

4.4.1.4 Delay Model

The input signal to the pulldown stack is driven by a WL driver, which is upsized to provide a fast transition input to the LBL pulldown network. With fast input transition, one can model the evaluation delay, of a gate by the following equation derived in [22]:

$$t_{pHL} = (\frac{1}{2} - \frac{1 - \frac{V_{th}}{V_{DD}}}{1 + \alpha})t_T + \frac{C_L V_{DD}}{2I_{D0}} \quad (4.15)$$

where C_L is the load capacitance at the output of the gate, α is the velocity saturation coefficient, and t_T is the input waveform transition time. t_T is used to approximate the actual waveform by a linear ramp, it can be expressed as $t_T = \frac{t_{0.9} - t_{0.1}}{0.8}$. Equation 4.15 accounts for 2 components of the gate delay. The first term signifies the input slope contribution whereas the second term is the time required to discharge the output capacitance. C_L should be replaced by C_{TOT} when applying Equation 4.15 to the LBL model.

4.4.1.5 Derivation of Optimal Transistor Width

With the energy (Equation 4.14) and delay (Equation 4.15) models derived as a function of dynamic node capacitance, and which can be expressed in terms of pulldown transistor width (W_{MN}) as shown in Equation 4.8, one can obtain the optimal W_{MN} for optimal EDP operation. Optimal EDP is achieved when the following condition is true:

$$\frac{\partial E_{TOT}}{\partial t_{pHL}} = \frac{\frac{\partial E_{TOT}}{\partial W_{MN}}}{\frac{\partial t_{pHL}}{\partial W_{MN}}} = -1 \tag{4.16}$$

By substituting the equations for energy, delay and capacitances, and taking the partial derivatives with respect to W_{MN}, one can obtain the following expressions:

$$\frac{\partial E_{TOT}}{\partial W_{MN}} = (nK_1 + K_2 + nK_3)V_{DD}^2 + K_3 t_{SC} V_{DD} \tag{4.17}$$

$$\frac{\partial t_{pHL}}{\partial W_{MN}} = \frac{C_L'}{K_5 K_6 + nK_1} \tag{4.18}$$

where K_1-K_6 are expressions that relates to C_L and the constants of proportionality K_I, K_{GD} and K_G. One can then derive the optimal W_{MN} expression by substituting Equation 4.17 and Equation 4.18 into Equation 4.16. The resulted expression is fairly complex, however, one can safely assume that $C_{DYN} \gg C_L$, due to self loading effect, to obtain the following simple relationship:

$$W_{MN}^{opt} = \frac{C_L'}{(\frac{1}{2} - \frac{1-\frac{V_{th}^N}{V_{DD}}}{1+\alpha})t_T \cdot \frac{2K_I^N}{V_{DD}} + (0.03K_{GD}^P + K_{GD}^N)n} \tag{4.19}$$

The derivation of the above expressions is detailed in Appendix B. From the above formulation, one can see that the optimal transistor width is inversely proportional to the number of pulldown paths.

4.4.1.6 Energy-Delay Comparisons

A comparison between the simulation results and the analytical models is performed over a wide range of energy (5x), delay (3x) and load (4x). Figure 4.11 illustrates the findings. The error concerning energy has a range of 6% while that for delay is 5%. Point P and Q in Figure 4.11 are the optimal operating region, corresponding to a fan-out of 2.3-2.7.

Simulation results indicated that if wide domino gates are designed with the optimal EDP fan-out for conventional domino gates (usually between 2 to 1 for high-performance circuit), they are in suboptimal conditions. Using the fan-out obtained, one can achieve upto 10% reduction in EDP.

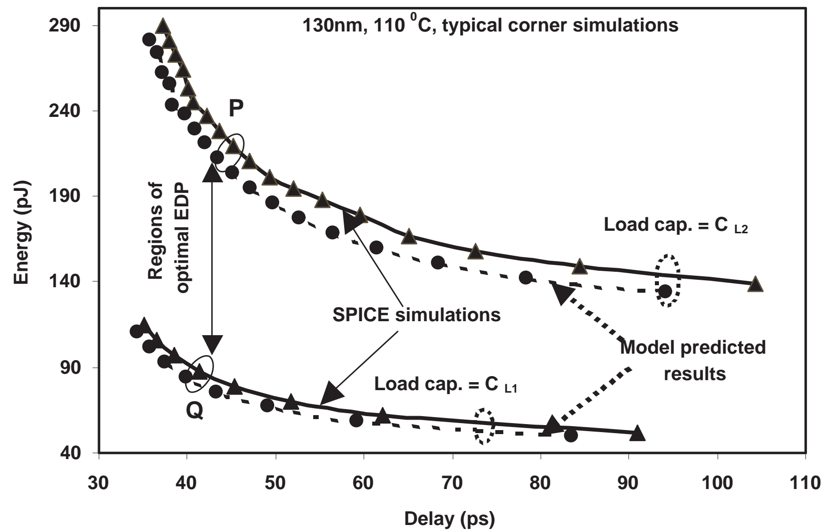


Figure 4.11: Energy-delay plots for wide domino gate of n=8

Although the models developed are simple and suitable for hand calculations, they do not account for the bias dependence of switching capacitance, stack effect of series transistors or short circuit current through the keeper inverter. Should these issues be addressed, the models will be of much higher complexity, but accuracy and percentage error will improve.

4.5 NAND Gate

The organization of a static NAND gate following the LBL circuits form a compound domino logic style. The NAND gate replaces the inverter at the output of a domino gate, to increase functionality of this stage in the RF system level. Its primary function is to combine two LBL outputs (for a two input NAND gate) and setup the signals for the subsequent global bit line stage. Since it lies in the critical path of the read operation,

transistors should be sized carefully.

4.6 Global Bit Line

The global bit line (GBL) organization is similar to that of a LBL. It is a domino gate with the gate of the pulldown transistors connected to the output of static NAND gates. The fan-in of a GBL utilized in a 32-word RF is very small, with only two pulldown paths. As a result, leakage is not a concern in this architecture. This is because a 32-word RF has four LBL circuits, and there are only two NAND gates for each bit slice. However, for RF of bigger and more realistic size, say 256-word by 64-bit, the number of fan-in will increase drastically. Design of such high fan-in domino gates should follow the EDP optimized model presented earlier in the LBL section.

4.7 Discussions

After understanding the operation of a register file, one can see that a RF system is leakage dominant. Firstly, at most two of the thirty-two outputs of an address decoder switch, as one output evaluates to high and the previous active output evaluates to a low. Moreover, only one of the thirty-two word line drivers is active, and at most one path of thirty-two LBL pulldown paths is conductive for each bit slice. For register files of more realistic and practical size, say 256-word by 64 bit, leakage is even more prominent. This is because no matter how deep a register file is, only one word line is active for each read or write port. The rest of the words are idle and contribute to leakage energy. Due to the above reasons, leakage current reduction is essential for a energy efficient RF, especially for large RF and future technologies where leakage energy surpasses switching energy and becomes dominant.

Chapter 5

Low-Leakage Register File Circuit Blocks Design

As aforementioned in the previous chapter, a large number of circuit components in a register file is leakage dominant. The leakage energy distribution is illustrated in Figure 5.1 for a conventional 2-read, 1-write 32-word, 32-bit RF in 0.13μm technology at 1.3V. The corresponding leakage percentage are listed in parenthesis. It can be observed that aside from clock buffer, which is made up of large transistors to drive the system clock tree, word line drivers, local bit lines, address decoders and the memory core are the largest contributors.

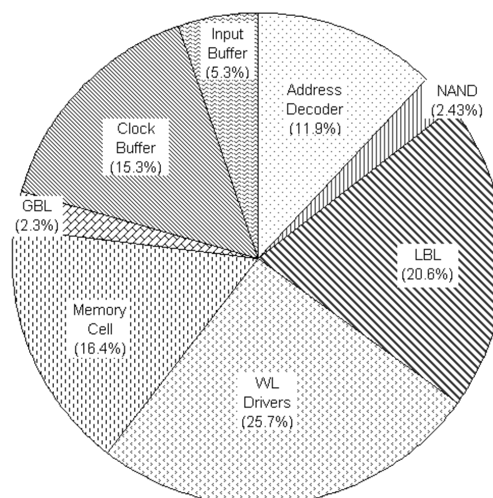


Figure 5.1: RF system leakage energy breakup

Leakage control mechanisms are specifically implemented to target these components to minimize the system's leakage energy. In this chapter, the possibilities and effectiveness in applying multi-threshold, non-minimum channel length transistors and stack effect are discussed. Furthermore, a novel static single bit line memory cell is introduced and a mathematical static noise margin model is presented.

5.1 Address Decoder

A 3-port RF has three address decoders, this is because each port can operate simultaneously, the address inputs can be different and are independent of one another. Since address decoders operate in the negative phase of a clock cycle and is not performance critical, they are implemented using static CMOS logic. In the previous chapter, a study comparing different architectures of decoders is done. It has been concluded that a 2-stage 2:4/3:8 decoder is favorable from the leakage control perspective, since it has the relatively small transistor width among all the 2-stage decoders and the split decoder signals from the first stage is readily available. This section presents various leakage techniques that are applicable to the decoder structures.

5.1.1 Dual-Threshold Technique

The principle of dual-threshold CMOS is to use high V_{th} devices in non-critical paths to minimize leakage flow. This idea is adopted to individual transistors in gate level. The 2-stage decoder structure is shown in Figure 4.2(b), one can see that it is made up of 2- or 3-input NAND gates and 2-input NOR gates. Figure 5.2 illustrates the schematic for a 3-input NAND and 2-input NOR gate.

Base on the property of a decoder, only one line is active for any given inputs. For example, a 3:8 decoder is made up of eight 3-input NAND gates, the input to these gates iterates all the combination in the truth table as shown in Table 5.1. Only one input combination will result in conduction through the serially connected transistors. Seven of the eight 3-input NAND gates have “OFF” transistor(s) in the NMOS stacks. As per

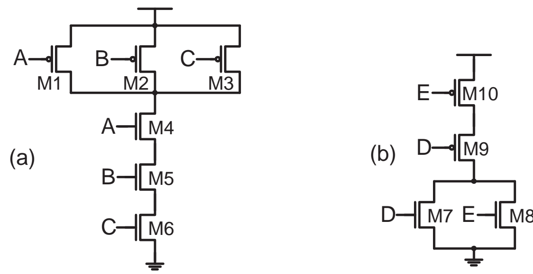


Figure 5.2: Static (a) 3-input NAND and (b) 2-input NOR gates

mentioned in Section 3.1, due to self-reverse body biasing, leakage through a stack (device M4-M6 in Figure 5.2) can be 10-30 times smaller than a single device. Therefore, leakage through the stack is limited and special attention should be given to the parallel PMOS devices. Apart from the input combination A'B'C', all inputs will have at least one PMOS device in "OFF" state. However, only with input ABC that there will be leakage through these transistors, this is because for the other six inputs, the output node is at the same voltage as V_{DD} by the conducting PMOS. Since both the DIBL conduction and weak inversion subthreshold leakage mechanism require a potential difference between the drain and the source, no current flows through those devices.

Table 5.1: Truth table for 3-input NAND Gate

Input combination	Output
0 0 0	1
0 0 1	1
0 1 0	1
0 1 1	1
1 0 0	1
1 0 1	1
1 1 0	1
1 1 1	0

With M1, M2 and M3 identified as the leaking devices, it is wise to use high V_{th} devices for the PMOS network. Similar reasoning is applied to 2-input NOR gates to reduce leakage. Simulation results indicate that through selectively changing devices to high V_{th} transistor, the decoder is capable to achieve a leakage reduction of 25%, with a performance

penalty of 4.7%. Since the total energy (switching energy and leakage energy) of a decoder is leakage dominate, reducing leakage current also reduces the total energy by 9.6%.

5.1.2 Non-Minimum Channel Length

Non-minimum channel length transistors have a higher V_{th}, however it increases the gate capacitance of a device. Therefore, they should not be used in the entire logic, but rather discretely at high leakage paths. Identical analysis as discussed in the previous section can be used to identify leakage devices and determine which transistors should be replaced by non-minimum channel length devices to achieve the most efficient leakage reduction.

Simulation results indicate that through changing parallel single-stacked devices to non-minimum channel length transistor, the decoder is capable to achieve a leakage reduction of 27.5%, with a performance penalty of 4.2% and total energy reduced by 11.5%. When comparing to the multi-threshold technique, the total energy reduction is less because the increase in capacitance causes an increase in switching energy.

5.1.3 Gated Leakage Path with Split Decoder Signals

The organization of a 2-stage decoder is such that the each bit of the second stage consists of a 3-input NAND and a 2-input NOR gates. As shown in Figure 4.2(b), the NAND gate of the second stage decoder evaluates independent of the first stage output. One can conclude that evaluation of the second stage is futile if the first stage decoder output is true, i.e. the fisrt stage NAND gate is not selected. In order to rectify that, the split decoder signal can be used to gate the second NAND logic, by inserting a PMOS between the pullup and pulldown network. For a 2:4 decoder, there are four 2-input NAND gates, each of its output is connect to a 3:8 decoder, which consists of eight 3-input NAND and eight 2-input NOR gates. The circuit path configuration for one output bit is shown in Figure 5.3.

The operation details can be explained as follows:

1. When the two most significant input combination is not true, the 2:4 decoder NAND gate output is high. The split decoder signal turns “OFF” the transistor MG of the

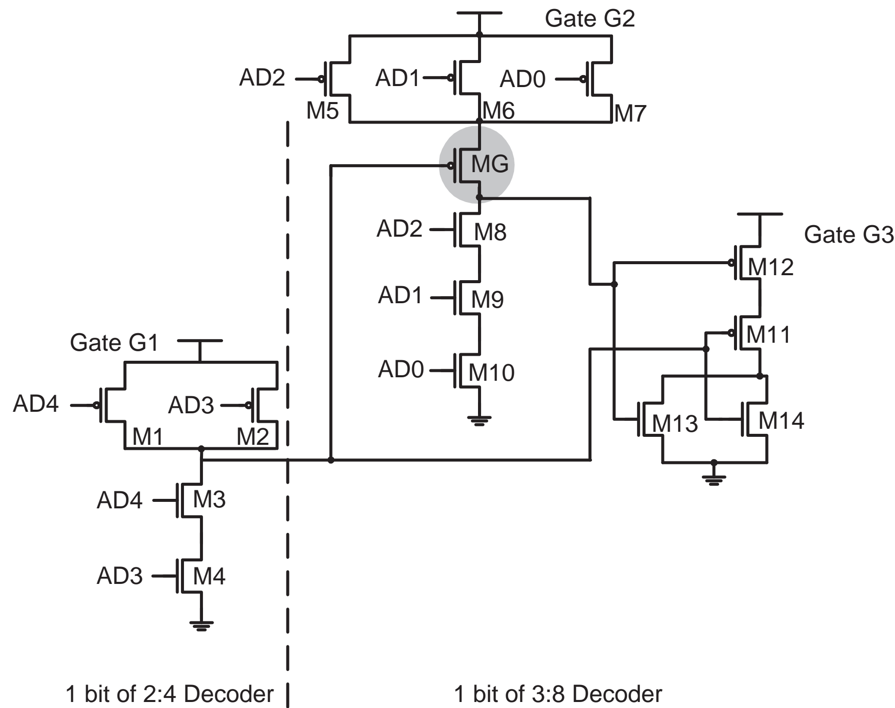


Figure 5.3: 1 bit circuit of a 2-stage 5:32 decoder with gated leakage path scheme

second stage 3-input NAND gates. This extra "OFF" PMOS induces stack effect on the entire PMOS network. The leakage paths change from single transistors to two serially connected devices. Due to self-reverse body biasing leakage current is reduced substantially. When MG is "OFF", the output node is cutoff from the PMOS network, the static output node is floating if all of AD2-AD0 are not high to enable to pulldown stack. However, due to the presence of the NOR gate, this floating node does not affect the robustness of the decoder. This is because only the split decoder signal is sufficient to enable the pulldown NMOS, M14, of the NOR gate and generate a "LOW" output.

2. When the split decoder signal is low, the second stage decoder operate as usual. The MG transistor turns on and connects the PMOS network to the output node.

With this implementation, leakage paths for three quarters of the second stage 3:8 decoders become stacked. This is because given any input combination, only one split decoder signal is "LOW" and the rest are "HIGH". Simulation results indicate that there is a 18.6%

reduction in leakage, 25% reduction in total energy with a performance hit of 4%.

Table 5.2 summarizes simulation results for the various techniques on an address decoder. It can be seen that non-minimum channel length is the most effective mechanism with the smallest performance hit.

Table 5.2: Efficiency of the 3 address decoder leakage control techniques

	Leakage Red.	Total Energy Red.	Performance Hit
Multi-threshold	25%	9.6%	4.7%
Non-min channel length	27.5%	11.5%	4.2%
Gated PMOS	18.6%	25%	4%

5.2 Word Line Driver

The word line drivers in this RF architecture is divided into two sections which drive the segmented memory core, the organization has been previously discussed in Section 4.2. At any point of time only one WL is ever selected for a read/write port. Furthermore, due to the large load presented to the drivers, the transistor sizes are fairly large. Hence, it is immensely important to implement leakage control for the deselected WLs.

5.2.1 Dual-threshold Technique

WL drivers are in the read critical path, they are the first stage to operate in the positive clock phase. Therefore, extra attention is paid to minimize the performance hit introduced by leakage reduction implementation. An obvious choice of leakage reduction is to use high V_{th} transistors on all devices, however, it would impede the performance greatly. Simulation results indicate that a leakage reduction of 77% is accompanied by an unacceptable 32% increase in delay. An alternative is required to apply high V_{th} transistors. WL driver operation is studied in details, only devices that are leaking the most will be replaced by high V_{th} transistors in order to minimize performance penalty.

All four stages of drivers, shown in Figure 4.5 are differ in size with stage 2 and stage 4 being the largest. Therefore the leakage current contribution is not evenly distributed,

the pie chart shown in Figure 5.4 illustrate the distribution. Stage 2 and 4 consumes the largest percentage of leakage energy with 67.4% and 21.3% respectively.

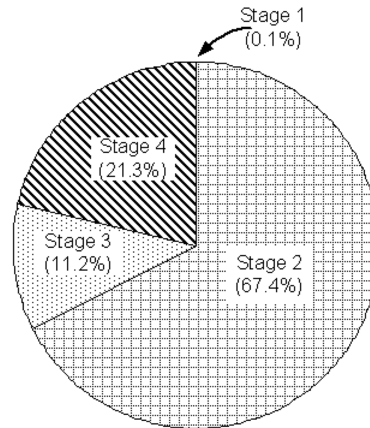


Figure 5.4: Leakage energy distribution among the WL driver stages

Furthermore, during evaluation phase, when a WL is deselected the "OFF" transistors alternate between the NMOS and PMOS between stages as shown in Figure 4.7. Although the PMOS of stage 1 is "OFF" as the clock is high, there is no leakage current through it. This is because the output of stage one is charged to supply voltage V_{DD} during precharge phase. The lack of voltage difference prevent leakage from taking place. The first stage leaks through the NMOS input transistor, however since there is a NMOS stack formed by the bottom clock transistor, leakage is minimal. The second stage PMOS has a large transistor size and it is the leakage device, therefore it should be replaced by a high V_{th} PMOS. The third stage is leaking through the NMOS, while the fourth stage is through another large PMOS. Similar to the reasoning applied to the second stage, it should be replaced by a high V_{th} transistor as well.

In conclusion, by replacing the second and fourth stage PMOS transistor with high V_{th} devices, leakage energy is reduced by a significant amount with very little speed degradation. Simulation results indicate that there is a 33.6% reduction in leakage, 21.2% reduction in total energy and 11.2% increase in delay. The total energy decreases because WL drivers, similar to address decoders, are leakage energy dominant circuits. Only one of the thirty-two word lines is switching and the rest are idle in a cycle.

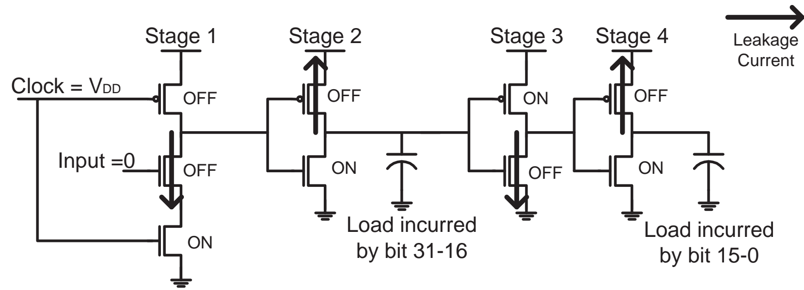


Figure 5.5: Leakage path for a deselected WL driver in evaluation phase.

5.2.2 Non-Minimum Channel Length

Similarly, transistors identified to be the leakage elements in the multi-threshold analysis can be replaced by non-minimum channel length transistors. By increasing the channel length for PMOS in the second and fourth stage, leakage and total energy reduced by 35% and 14.6% respectively, with delay increased by 8.2%.

5.2.3 Gated V_{DD} for the Second Stage

This technique is targeted to limit leakage on the second stage of WL drivers. It uses the decoded signal as a control to a newly added PMOS, MG, in Figure 5.6. MG introduces stack effect into the circuit which has substantial effect on leakage current reduction. Split decoder signals are the output of the first stage 2:4 decoder. WL of 8 bit increments are connected to the same split decoder signal, for example, WL 0-7 are connected to split decoder signal 0, while WL 8-15 are to split decoder signal 1, etc. Three situations can

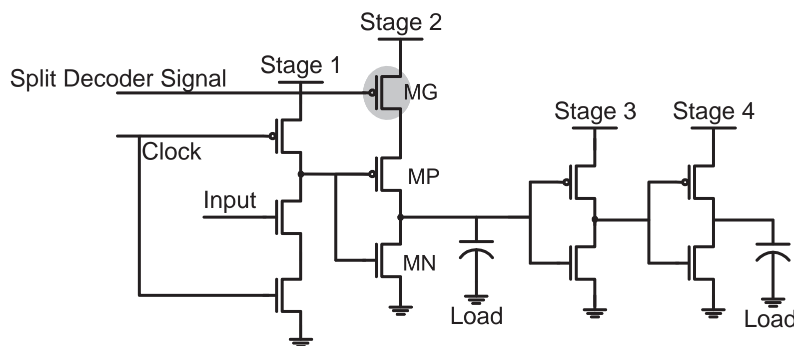


Figure 5.6: Gated supply for second stage WL drivers

arise during the operation, it can be explained as follows:

1. For the WL connected to a high split decoder signal, i.e. they are not selected. MG transistor is turned off, and the input to the WL drivers are 0. The large size second stage PMOS, MP, is leaking through a stack of "OFF" devices, thereby reducing leakage. MN anchors the output node to ground allowing the WL drivers to operate normally.

2. Seven of the eight WL drivers connected to a low split decoder signal will have an value of low input connected to the first driver stages. For these WLs, leakage reduction due to stack effect is reduced because the stack is make up of an "ON" MG and "OFF" MP devices. However, due to body effect on MP, it can still reduce the leakage current by a small amount.

3. The selected WL will have a conducting MG and MP devices. The sizing of MG should be increased to avoid causing significant performance penalty. This is of great importance because stack transistors have reduced current driving capability.

Simulation results indicate that when gating transistor is upsized to 2x of MP, there is a performance hit of 1.2%. Total Energy increases by 8.8% and leakage reduction is 34.3%.

Table 5.3 tabulate the simulation results for the techniques discussed above for WL drivers. Gated V_{DD} has a higher switching energy, this is because the split decoder signals are now driving more capacitance when they are connected to upsized MG transistors. It is also be observed that since extra devices are required, gated V_{DD} technique will have an area overhead. It is not the most desirable among the three proposed leakage control schemes.

Table 5.3: Efficiency of the 3 WL drivers leakage control techniques

	Leakage Red.	Total Energy Red.	Performance Hit
Multi-threshold	33.6%	21.2%	11.2%
Non-min channel length	35%	14.6%	8.2%
Gated V_{DD}	34.3%	-8.8%	1.2%

5.3 Local Bit Line

As discussed in the previous chapter, local bit line has an extremely long fall time, due to the large capacitance at the output node. Each bit slice has four LBL circuits, and at most only one of them switches while the other remains idle. For a 32-word by 32-bit RF, there are 128 LBL circuits and make up for 20% of the system leakage. It is imperative to implement leakage control mechanism in this stage. Furthermore, wide domino gates are especially susceptible to leakage induced false evaluations, improving the leakage tolerance will improve the unity gain noise margin (UGNM) for such gates. In this section, the concept of UGNM is explained, following with leakage control mechanisms.

5.3.1 Concept of Unity Gain Noise Margin

UGNM is a figure of merit used to determine the DC robustness of a gate, the analysis method is outlined in [20]. For a wide domino circuit followed by a static gate, in this case a LBL followed by a NAND gate as shown in Figure 5.7, UGNM is measured with respect to the output of the NAND gate (OP). For worst-case leakage analysis, the gate of the bottom transistors are connected to a high ($D_7 - D_0$), and the input to the top transistors ($WL_7 - WL_6$)are subjected to input noise. DC noise is simulated by using a very slow rising ramp. The point when the output of the static NAND gate toggles and crosses over with the input noise ramp is the UGNM. The waveform diagrams are shown in Figure 5.8. It has been shown that when leakage control technique is implemented into a wide domino circuit, UGNM improves and the PMOS keeper can be downsized to retain iso-robustness [20]. It can be seen that when keeper is reduced in size, the short circuit energy through the keeper and conduction path during transient is reduced, resulting in overall switching energy reduction.

5.3.2 Dual-threshold Technique

The operation of a LBL is such that the data signals connected to the bottom transistors of the NMOS stacks are available before the clock signal changes high. The performance

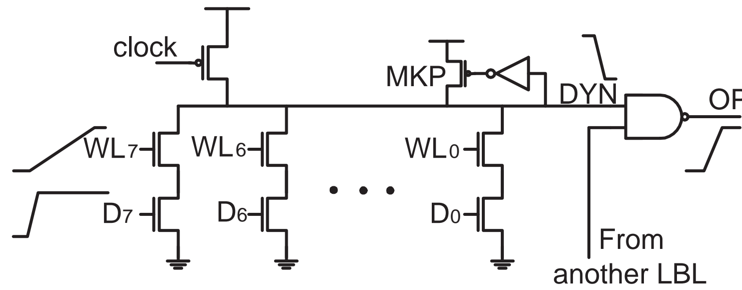


Figure 5.7: UGNM measurement setup for LBL circuits

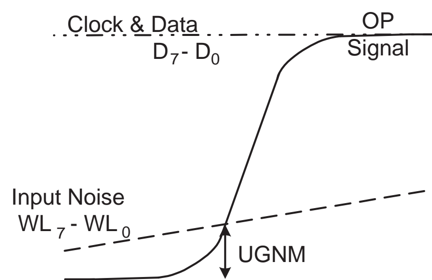


Figure 5.8: UGNM measurement waveforms

critical WL signals usually drives long interconnect and arrive after the evaluation phase has begun. These signals are subjected to input noise, and can cause the circuit to falsely evaluate [20]. For a dual-threshold implementation, the top stack transistors (WL_7 to WL_0) are replaced with high V_{th} devices while the bottom ones remain low V_{th}.

Simulation results indicate that this technique can reduce leakage substantially by 73%, while total energy has a saving of 24.3%. There is a 10 ps performance degradation incurred, which translate to a 10% increase when compare to a LBL of all low V_{th} devices.

5.3.3 Non-Minimum Channel Length

Similar to the dual-threshold implementation, the top transistors of the NMOS stacks are replaced by non-minimum channel length devices, 10nm longer than the minimum. There is a leakage and total energy reduction of 77% and 44.3% respectively. However, there is also a 9.3% increase in delay.

5.3.4 Footer Transistor

The concept of footer transistor as leakage control device is similar to that of MTCMOS Sleep transistor as discussed in Section 3.2.1. Instead of using high V_{th} device as sleep transistor, the footer implementation use a low V_{th} NMOS. Figure 5.9 illustrate the circuit organization. Since each bit slice has four LBL circuits and the 2:4 decoder has four

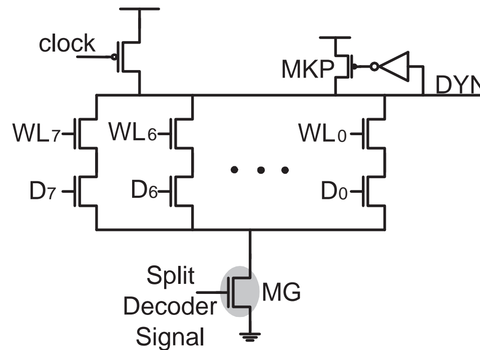


Figure 5.9: LBL leakage control footer implementation

outputs, the obvious choice for the footer transistors control signals is the split decoder outputs. The split decoder signal which enables the first 3:8 decoder is also connected to the first LBL circuit, while the second decoder signal is connected to the second LBL circuits, and so-forth. The operation of the footer LBL circuits are as follows:

1. When the split decoder signal is high, none of the WL's connected to the particular LBL is selected. Therefore, the LBL will not evaluate and its output should remain high. Under this circumstances, the LBL circuit is leaking through the NMOS network. With the extra "OFF" footer transistor leakage current is greatly reduced since the footer induces self-reverse body bias effect.

2. One of the four split decoder signals is low, indicating one of the eight WL connecting to the LBL circuit will be enabled. In this case, the footer transistor needs to provide a path to ground for the circuit to evaluate. The drain of MG, node VG, act as a virtual ground. When the WL and data signals are both high, the dynamic node evaluate and discharges through the VG to ground.

Since the LBL circuits are a few stages down the data path when compare to the decoder, the split decoder signals require a couple buffer stages to aid the propagation. These buffers add to the overhead of this technique. Simulation results indicate that, when not considering buffers, there is a reduction of 80.3% in leakage and 25% in total energy. However, when the buffer leakage current and switching energy is taken into considerations, leakage reduction becomes 68.1% and LBL total energy increased by 2x. This dramatic increase in energy is accounted for by the switching energy of the buffers. The performance penalty introduced by the footer is 7%. The performance hit is reasonably small because the split decoder signal is available in the negative phase of a clock cycle, node VG is at ground before the circuit evaluates and thus introduces minimal delay to the organization.

A summary of simulation results for the various leakage control implementation discussed in this section is tabulated in Table 5.4. It is clear that non-minimum channel length has the most leakage and energy reduction, with a slightly higher performance penalty than that of dual-threshold implementation.

Table 5.4: Efficiency of the 3 LBL leakage control techniques

	Leakage Red.	Total Energy Red.	Performance Hit
Dual-threshold	73%	24.3%	10%
Non-min channel length	77%	44.3%	9.3%
Footer transistor	68.1%	-100%	7%

5.4 Single-Ended Interrupted Feedback Memory Cell

Single-ended memory cells are preferred over dual-ended memory cells in superscalar register files because the parasitic capacitance associated with the data bit lines can be drastically reduced and thus reducing the switching energy. A comparative analysis of the proposed 6-transistor single-ended interrupted feedback memory cell (SEIFMC) (Figure 5.10(a)) and a conventional 7-transistor single-ended memory cell (SEMC) (Figure 5.10(b)) is performed.

The operation of SEMC has been discussed in the previous chapter, in section 4.3. The

following describe the operation of a SEIFMC:

1. The WL signal is the word line signal from the write port, which is only enabled during the write operation and when the particular word is selected. When WL is high and WL' is low, M2 is "OFF" and disabled the feedback inverter (MB1 and MB2). This interrupted feedback characteristic eliminates the contention between the back-to-back inverters when a new data bit is being written into the cell, allowing a reduction in write energy and write delay.

2. During the data retention phase, WL' is always at V_{DD} and the feedback inverter is operating like that of SEMC.

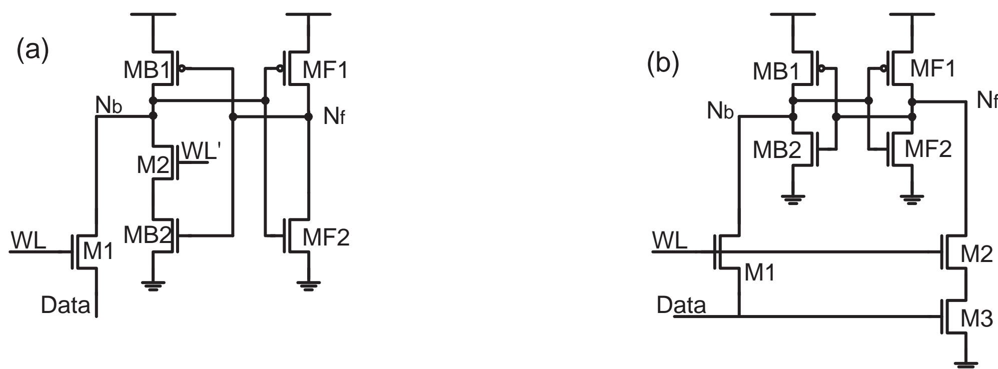


Figure 5.10: Schematic of (a) SEIFMC and (b) SEMC

The SEIFMC requires an extra inverter to generate the WL' signal, however the energy associated with it can be absorbed when it is shared among a few cells of the same word.

The fundamental function of memory cells is data retention, therefore it is of utmost importance for the proposed circuit to maintain static noise margin (SNM). A SNM mathematical representation is derived for SEIFMC and a comparison to the SNM of SEMC is presented in the following section .

5.4.1 Static Noise Margin

SNM is a parameter used to measure the data retention integrity of a memory cell. The SNM is defined as the maximum amount of noise that the back-to-back inverters can tol-

erate before changing stages [23]. A graphical representation of SNM is obtained from the flip-flop voltage-transfer characteristic curves. SNM is the diagonal of the maximum possible square between the two curves, as shown in Figure 5.11(a) [23]. Another interpretation of SNM is based on the assumption that (V_{OL}, V_{IH}) and (V_{OH}, V_{IL}) points on the transfer curves are points where the slopes are equals to -1. Then SNM is the diagonal of the rectangle with apices at those points, as shown in Figure 5.11(b) [24]. The later representation is adopted in this study since it is more suitable for mathematical analysis.

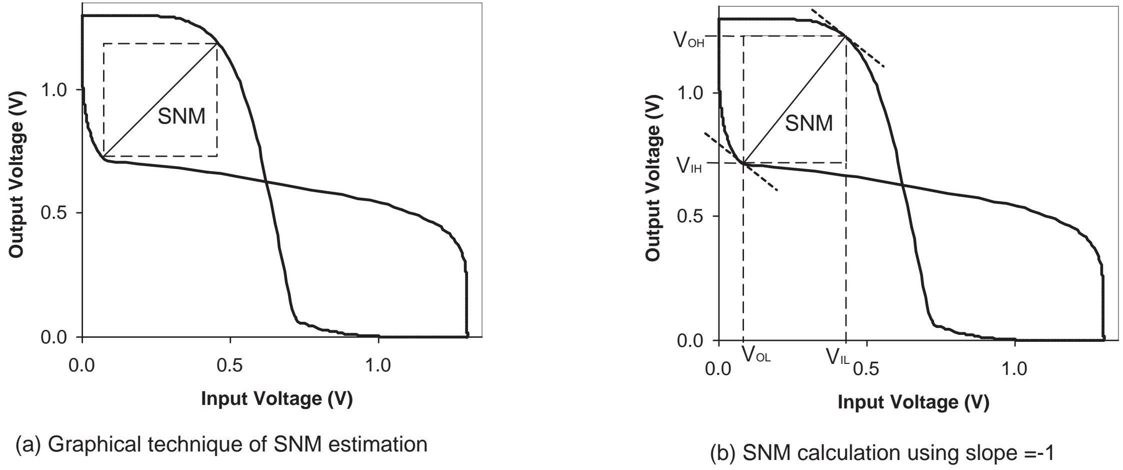


Figure 5.11: Illustration of two different SNM interpretation

The α-power law model for short-channel MOSFET proposed by [25] is used in the following two subsections which focus on the SNM modelling for both SEMC and SEIFMC. The voltage and current models for short-channel NMOS are as follows:

$$I_D = \begin{cases} 0, & \text{cutoff region} \\ K_l(V_{GS} - V_{th})^{\alpha/2} V_{DS}, & \text{linear region} \\ K_s(V_{GS} - V_{th})^{\alpha}, & \text{saturation region} \end{cases} \tag{5.1}$$

where $K_l = \frac{I_{D0}}{V_{D0}(V_{DD}-V_{th})^{\alpha/2}}$ and $K_s = \frac{I_{D0}}{(V_{DD}-V_{th})^{\alpha}}$. I_{D0} is the drain current at $V_{GS} = V_{DS} = V_{DD}$, and V_{D0} is the drain saturation voltage at $V_{GS} = V_{DD}$ and α is the velocity saturation coefficient. The models are also applicable to short-channel PMOS devices when the voltage polarities are reversed.

5.4.1.1 SNM Modelling for Single-Ended Memory Cell

The SNM definition adopted required the knowledge of four voltage points on the transfer curve, namely V_{IH},V_{OL},V_{IL} and V_{OH}. These points can be derived from the current equations. When the input of an inverter is at V_{IH} and output is at V_{OL}, the pulldown NMOS is in linear mode while the PMOS in in saturation mode. According to Equation 5.1 and substituting the appropriate unknown variables, the current equation becomes:

$$I_{D,l}^{N} = K_l^N (V_{IH} - V_{th}^N)^{\alpha_l^N/2} V_{OL} \tag{5.2}$$

$$I_{D,s}^{P} = K_s^P (V_{DD} - V_{IH} + V_{th}^P)^{\alpha_s^P} \tag{5.3}$$

Kirchhoff's Current Law state that the sum of current entering a node should be zero, therefore $I_{D,l}^{N}$=$I_{D,s}^{P}$. Furthermore, the slope of characteristic curves at the point of interest is equal to -1, therefore $\frac{dV_{OL}}{dV_{IH}} = -1$. One can obtain the V_{IH} and V_{OL} expressions by manipulating Equation 5.2 and 5.3 according to the above two conditions.

$$V_{IH} = (\frac{2K_l^N}{\alpha_l^N K_s^P})^{-\frac{2}{\alpha_l^N+2}} (V_{DD} + V_{th}^P)^{\frac{2\alpha_s^P}{\alpha_l^N+2}} + V_{th}^N \tag{5.4}$$

$$V_{OL} = \frac{K_s^P (V_{DD} - V_{IH} + V_{th}^P)^{\alpha_s^P}}{K_l^N (V_{IH} - V_{th}^N)^{\frac{\alpha_l^N}{2}}} \tag{5.5}$$

When the circuit is at the V_{IL} and V_{OH} point, the PMOS is in the linear region while NMOS is in saturation. Their voltage and current behavior is governed by the following:

$$I_{D,l}^{P} = K_l^P (V_{DD} - V_{IL} + V_{th}^P)^{\alpha_l^P/2} V_{OH} \tag{5.6}$$

$$I_{D,s}^{N} = K_s^P (V_{IL} - V_{th}^N)^{\alpha_s^N} \tag{5.7}$$

Similar to the above analysis, by applying Kirchhoff's Current Law and $\frac{dV_{OH}}{dV_{IL}} = -1$, one can deduce the following V_{IL} and V_{OH} formulations:

$$V_{IL} = (\frac{K_l^P}{\alpha_s^N K_s^N})^{\frac{1}{\alpha_s^N-1}} (V_{DD} + V_{th}^P)^{\frac{\alpha_l^P}{2(\alpha_s^N-1)}} + V_{th}^N \tag{5.8}$$

$$V_{OH} = V_{DD} - \frac{K_s^N (V_{IL} - V_{th}^N)^{\alpha_s^N}}{K_l^P (V_{DD} - V_{IL} + Vth^P)^{\frac{\alpha_l^P}{2}}} \tag{5.9}$$

With the apices defined by the above equations, the SNM can be found by simply calculating $SNM = \sqrt{(V_{IL} - V_{OL})^2 + (V_{OH} - V_{IH})^2}$. Table 5.5 illustrate the simulation result for the four characteristic voltages and SNM. The model developed tracks the HSPICE simulation closely, with 14.2% error.

Table 5.5: Comparison of SNM simulation and model data for SEMC

	Simulation Data	Mathematical Model
V_{OL}	0.07V	0.07V
V_{IL}	0.43V	0.37V
V_{IH}	0.72V	0.78V
V_{OH}	1.24V	1.23V
SNM	0.63V	0.54V

5.4.1.2 SNM Modelling for Single-Ended Interrupted Feedback Memory Cell

The methodology for calculating the four voltage points and SNM for SEIFMC is similar to that of SEMC. However, one has to consider the intermediate mode voltage between NMOS transistor M2 and MB2 in the backward inverter. Figure 5.10 shows that the forward inverter composed of transistor MF1 and MF2 are identical in structure for both the SEIFMC and SEMC. V_{IL} and V_{OH} are determined by the forward inverter, their expression for SEIFMC are identical to Equation 5.8 and 5.9. The derivations of V_{IH} and V_{OL} are more complex and the intermediate node (V_X) current and voltage play a part in the derivation. At the point where the transfer characteristic slope equals to -1, the PMOS MB1 is in saturation mode, while M2 and MB2 are in linear mode. The current equations for the three transistors are:

$$I_{D,s}^{P,MB1} = K_s^P (V_{DD} - V_{IH} - V_{th}^P)^{\alpha_s^P} \tag{5.10}$$

$$I_{D,l}^{N,M2} = K_l^N (V_{DD} - V_X - V_{th}^N)^{\frac{\alpha_l^N}{2}} (V_{OL} - V_X) \tag{5.11}$$

$$I_{D,l}^{N,MB2} = K_l^N (V_{IH} - V_{th}^N)^{\frac{\alpha_l^N}{2}} V_X \tag{5.12}$$

Due to Kirchhoff's current law, current through the MB1 is equivalent to current through M2 and which in turn is equal to the current in MB2. V_{OL} can be expressed by equating

$I_{D,l}^{N,MB2}$ and $I_{D,l}^{N,M2}$, while V_X is obtained from $I_{D,s}^{P,MB1}$ and $I_{D,l}^{N,M2}$.

$$V_{OL} = \frac{K_l^N (V_{IH} - V_{th})^{\frac{\alpha_l^N}{2}} V_X}{K_l^N (V_{DD} - V_X - V_{th}^N)^{\frac{\alpha_l^N}{2}}} + V_X \tag{5.13}$$

$$V_X = \frac{K_s^P (V_{DD} - V_{IH} - V_{th}^P)^{\alpha_s^P}}{K_l^N (V_{IH} - V_{th}^N)^{\frac{\alpha_l^N}{2}}} \tag{5.14}$$

Both V_{OL} and V_X are both defined in terms of V_{IH}, and which can be obtained by equating $I_{D,s}^{P,MB1}$ and $I_{D,l}^{N,MB2}$ and $\frac{dV_{OL}}{dV_{IH}} = -1$. Mathematically, the conditions can be combined and express as:

$$\frac{dI_{D,s}^{P,MB1}}{dV_{IH}} - \frac{\partial I_{D,l}^{N,M2}}{\partial V_{IH}} = -\frac{\partial I_{D,l}^{N,M2}}{\partial V_{OL}} \tag{5.15}$$

with

$$\frac{dI_{D,s}^{P,MB1}}{dV_{IH}} = -\alpha_s^P K_s^P (V_{DD} - V_{IH} - V_{th}^P)^{\alpha_s^P - 1} \tag{5.16}$$

$$\frac{\partial I_{D,l}^{N,M2}}{\partial V_{OL}} = K_l^N (V_{DD} - V_X - V_{th}^N)^{\frac{\alpha_l^N}{2}} \tag{5.17}$$

$$\frac{\partial I_{D,l}^{N,M2}}{\partial V_{IH}} = -K_l^N [\frac{\alpha_l^N}{2} (V_{DD} - V_X - V_{th}^N)^{\alpha_l^N - 1} + (V_{DD} - V_X - V_{th}^N)^{\frac{\alpha_l^N}{2}}] \frac{dV_X}{dV_{IH}} \tag{5.18}$$

and

$$\frac{dV_X}{dV_{IH}} = \alpha_s^P K_s^P (V_{DD} - V_{IH} - V_{th}^P)^{\alpha_s^P - 1} K_l^N (V_{IH} - V_{th}^N)^{\frac{\alpha_l^N}{2}}$$
$$- K_s^P (V_{DD} - V_{IH} - V_{th}^P)^{\alpha_s^P} \alpha_1^N K_l^N (V_{IH} - V_{th}^N)^{\frac{\alpha_l^N}{2} - 1} \tag{5.19}$$

$\frac{dV_X}{dV_{IH}}$ is the derivative of V_X taken with respect to V_{IH}. The value of $\frac{dV_X}{dV_{IH}}$ is very small, since K_l and K_s are in the order of 10^{-4}. This makes $\frac{\partial I_{D,l}^{N,M2}}{\partial V_{IH}} \approx 0$ and can be neglected from Equation . The expression obtained after substituting Equation 5.4.1.2, 5.17, 5.18 can be solved by performing a Taylor approximation at $V_{IH} = h$ and $V_X = x$. The result is an equation in quadratic form, $Ay^2 + By + C$ where $y = (V_{IH} - h)$, and can be readily solved. With the help of MathCad tool, the constants A, B and C are found to be as follows:

$$A = \frac{\alpha_l^N}{2} (h - V_{th}^N)^{\frac{\alpha_l^N}{2} - 1} \times \frac{\alpha_s^P K_s^P}{K_s^N} (V_{DD} - h - V_{th}^P)^{\alpha_s^P - 2}$$

$$B = \frac{\alpha_s^P K_s^P}{K_s^N}(V_{DD} - h - V_{th}^P)^{\alpha_s^P - 2} \times (h - V_{th}^N)^{\frac{\alpha_l^N}{2}} - \alpha_l^P (V_{DD} - h - V_{th}^P)^{\alpha_s^P - 1}$$
$$- \frac{\alpha_l^N}{2}(h - V_{th}^N)^{\frac{\alpha_l^N}{2} - 1} \times \frac{\alpha_s^P K_s^P}{K_s^N}(V_{DD} - h - V_{th}^P)^{\alpha_s^P - 1}$$
$$+ \frac{\alpha_l^N}{2}(h - V_{th}^N)^{\frac{\alpha_l^N}{2} - 1} \times (V_{DD} - x - V_{th}^N)^{\frac{\alpha_l^N}{2}}$$
$$+ \frac{\alpha_l^N}{2}(h - V_{th}^N)^{\frac{\alpha_l^N}{2} - 1} \times \frac{\alpha_l^N}{2}(V_{DD} - x - V_{th}^N)^{\frac{\alpha_l^N}{2} - 1} \times x$$
$$C = -\frac{\alpha_l^N}{2} \frac{K_s^P}{K_l^N}(V_{DD} - x - V_{th}^N)^{\frac{\alpha_l^N}{2} - 1} \times (V_{DD} - h - V_{th}^P)^{\alpha_s^P}$$
$$+ (h - V_{th}^N)^{\frac{\alpha_l^N}{2}} [-\frac{\alpha_s^P K_s^P}{K_s^N}(V_{DD} - h - V_{th}^P)^{\alpha_s^P - 1} + (V_{DD} - x - V_{th}^N)^{\frac{\alpha_l^N}{2}}$$
$$+ \frac{\alpha_l^N}{2}(V_{DD} - x - V_{th}^N)^{\frac{\alpha_l^N}{2} - 1} \times x]$$

With the apices defined by the above equations, the SNM can be obtained from $SNM = \sqrt{(V_{IL} - V_{OL})^2 + (V_{OH} - V_{IH})^2}$. Table 5.6 illustrate the simulation and model results for the four characteristic voltages and SNM. Taylor expansion was done at h=0.9 and x=0.07, one can see that the model tracks the HSPICE simulation closely, with 13.1% error.

Table 5.6: Comparison of SNM simulation and model data for SEIFMC

	Simulation Data	Mathematical Model
V_{OL}	0.08V	0.12V
V_{IL}	0.43V	0.37V
V_{IH}	0.75V	0.78V
V_{OH}	1.24V	1.24V
SNM	0.60V	0.52V

By examining the data in Table 5.5 and 5.6, one can see that SNM of SEIFMC and SEMC is different by 30 mV. The difference translate to less than 5% reduction in SNM. Furthermore, the architecture of SEMC provides four leakage paths, while SEIFMC only has three leakage paths. Due to this reason, the leakage current of SEIFMC is 2.7% less. Although the speed of a write operation is not performance critical, it is interesting to note that since SEIFMC disabled the feedback loop when the word is selected in a write operation, the write energy and delay shows a 43% and 21% reduction respectively. One

major drawback with SEIFMC is that, it uses WL' signal which requires extra buffer and adds to the wiring complexity. However, there are increasing number of metal layers in newer technologies, the extra wiring will have minimal affect on the overall memory core area.

Chapter 6

Low-Leakage RF System and Scaling Trends

In the previous chapter, each leakage reduction technique for the major functional blocks is discussed individually, along with their pros and cons. In this chapter, their effectiveness is examined in the system level. Possibility of combining leakage control techniques for each block is investigated with the objective of leakage reduction and minimal performance penalty infliction. Furthermore, a study on scaling trends for a 32-word by 32-bit RF for 90nm and 70nm technologies, as well as projections for bigger and more realistically sized RF is performed.

6.1 Selection Methodology for Leakage Control Techniques

Each leakage control technique presented yields leakage and/or energy reduction, however there are also different degree of performance tradeoff associated with them. One can see that when all the probable techniques are combined, the leakage reduction is maximized, however, it also results in maximum delay penalty. RF is highly performance critical and a single cycle system, therefore any leakage reduction means should not compromise its performance to a great extend. The techniques selected for a low-leakage, high-performance RF design are listed below:

1. Address Decoders

   Combination of non-minimum channel length and gated PMOS

2. Word Line Drivers

   Non-minimum channel length only

3. Local Bit Lines

   Non-minimum channel length only

4. Memory Cells

   Single-ended interrupted feedback memory cells

6.1.1 Decoders

Table 5.2 in the previous section listed the leakage, switching energy reduction and performance penalty for the three leakage control mechanisms investigated. Between multi-threshold and non-minimum channel length, the former one incurs more performance hit and has less leakage and energy reduction. Therefore, non-minimum channel length is a better choice among the two. Gated PMOS technique can be combine with non-minimum channel length parallel transistors. Simulation results show that the delay introduced is 4.2% with 51% and 30% reduction in leakage and energy respectively. In the system level, leakage reduction of 5% is achieved with the combination. There is no performance penalty since the decoder is not in the critical path, and there exists enough timing slack in the negative clock phase to absorb the delay.

6.1.2 Word Line Drivers

The possibility of combining non-minimum channel length with gated V_{DD} has been investigated. Simulation on the circuit block shows that there are only 5% gain in leakage reduction when compare to employing non-minimum channel length alone. The energy reduction is a total of 40% while energy and delay increased by 45% and 15% respectively. Since WL drivers are in the critical path, a 15% increase in delay is unaccepted, especially when the gain in leakage is merely a few percentages more. Therefore, it is more beneficial to employ non-minimum channel length only. In the system level, this implementation reduces leakage by 9.6% with a performance penalty of 3.8%.

6.1.3 Local Bit Lines

The local bit line contributes the most to the RF system leakage. Study of footer transistor technique using split decoder signal was very promising at first. However, once the energy and leakage for the split decoder signal buffers are included in the comparison, the result is not as gratifying. The leakage reduction achieved by merely using non-minimum channel length surpasses that of footer transistor technique. It is concluded that non-minimum channel length is the most effective without area and wiring overhead coming from the decoder signal. In the RF system level, this technique induces a 8ps delay, which is 2% of the system performance, and a saving of 14.2% in leakage. Its influences on switching energy of the entire RF system is negligible.

6.1.4 Memory Cells

The memory core of a 32-word by 32-bit RF consists of 1024 memory cells. It has been shown that an individual SEIFMC has 11.7% reduction in leakage energy, however when extra buffers are inserted to drive the WL' signals, the leakage induced by the buffers balance the savings. Therefore, if extra buffers are required, SEIFMC is not more favorable than the conventional SEMC. However, when examining its application in the system level, there are WL' signals readily available. Similar to the read port structure, the write port also has word line drivers that deliver the decoder outputs to the memory core. These WL drivers are made up of stages of inverters, as such the WL' signal is available in the existing circuitries. The outputs of the first and third stage are suitable for this application, therefore, no extra inverters are required. Since the transistors in a memory cell is close to minimum width, the additional load presented to these drivers is very small, only about four μm extra. With this implementation, the memory core has a leakage saving of 2.8%. Since the RF studied is very small, with only 1024 memory cells, the observable saving is insignificant. However, the saving would be more meaningful for larger RF when there are more memory cells in the systems.

6.2 RF Systems and Scaling Trends

With all the aforementioned techniques implemented on one RF structure, the circuit becomes a robust low-leakage design. When compare to the base model, the read operation delay has a 5% increase, from 354ps to 373ps. Switching energy reduces by 4.2%, whereas the total leakage current has a 25% reduction from 1028μA to 676μA. Despite the tremendous leakage current reduction, when examining the total energy, which includes switching and leakage energy, a very small amount of saving (4.2%) is observed. This is because there are not enough leakage component in a 32-word by 32-bit RF. Although, 31 of the 32 word lines are "OFF" and leaking, the total leakage current only makes up 1.5% of the total energy. Therefore, the 25% savings in leakage energy is not observable from examining the total energy.

As scaling continues, the amount of leakage current in the system increases. Base on the Berkeley Predictive Model for 90nm and 70nm technologies, the I_{ON}/I_{OFF} ratio degrades by 26x for high V_{th} devices and 42x for low V_{th} transistors [20]. The amount of lekage in a system increase and savings introduced by there techniques have a more significant role. Figure 6.1 illustrates the amount of reduction introduces for leakage, switching and total energy in different technologies using the Berkeley Predictive Model.

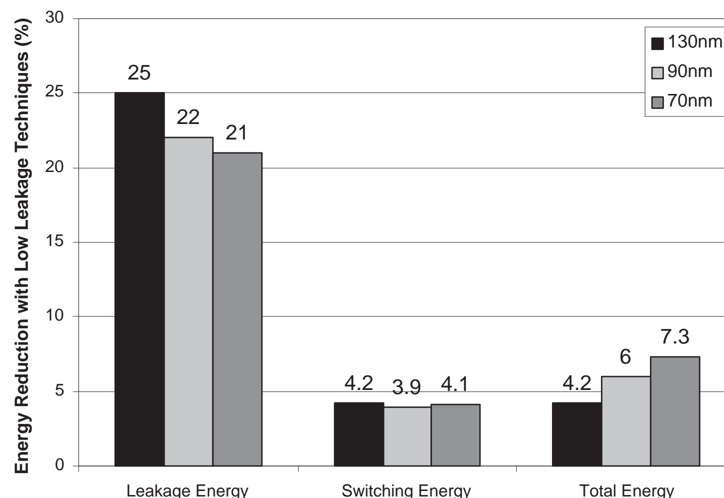


Figure 6.1: Energy Savings for 32-word by 32-bit RF

Furthermore, when one examines a larger RF system, say a 256-word by 64-bit 2-

read, 1-write RF, the percentages of leakage of total energy increase tremendously. This is because there is only one word line enable independent of the overall RF size. The percentages of leakage components increases in a larger RF system. With the limitation of computation power, a schematic simulation on a 256-word by 64-bit 2-read, 1-write RF is impractical. In turn, a projection on energy and leakage consumption is made base on the measurement from a 32-word by 32-bit RF. The leakage energy for each circuit block is calculated by dividing the measurement data with the correct weighing factor. For example, the leakage current for LBL structure is 171μA. There are four LBL circuits in each bit slice, as such there are 128 LBL circuits in the system. Each LBL structure consumes $171/128 = 1.34\mu A$. For a 256-word by 64-bit RF, each bit slice has 32 8-bit wide LBLs, there are a total of 2048 LBL circuits. The leakage current pertaining to LBLs in the system is $1.34\times2048 = 2744.3\mu A$. Similar calculations are performed for the other circuits blocks, and the total leakage current becomes 43.5mA. Employing the same methodology for total energy projection, a 789pJ per transition is found. Therefore, leakage energy makes up 7.7% of total energy in a 1 ns integration cycle. Figure 6.2 shows the amount of leakage energy in a 32-word by 32-bit RF and 256-word by 64-bit RF for 130nm, 90nm and 70nm technologies. Leakage energy becomes much more dominant in large systems for sub-130nm technologies. For 90nm and 70nm large RF systems, it is eminent that leakage control is very important, since leakage energy makes up 34.9% and 55% of total energy.

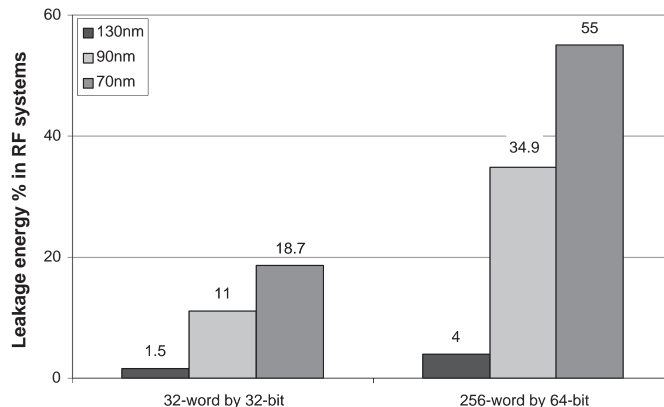


Figure 6.2: Leakage energy in RF systems for various technologies

Chapter 7

Register File Test Chip

A test chip which contains the essential circuits of a low leakage RF has been laid out and sent to fabricate through the Canadian Microelectronics Corporation (CMC) on run 0402CF. The device models used in schematic simulations discussed in the previous chapters employs the Berkeley Predictive Models, which does not have a physical layout tool. Furthermore, CMC does not provide fabrication service for 0.13μm technology till the very end of the research phase. With no means to fabricate in a leaking process, 0.13μm and beyond, the test chip was scaled back to 0.18μm and fabricated using a TSMC process. The chip serves as a medium to measure performance of the innovations on the conventional design when implemented on silicon. Furthermore, it is also a valuable exercise to tape out a circuit for fabrication with custom layouts and top level floor planning. The silicon area granted for the tape out is of size 1 mm by 1 mm, which can accommodate 28 pins. 9 of the 28 pins are used by another test circuit, therefore 19 pins were available to the RF circuits.

The TSMC 0.18μm process has 6 metal layers. Throughout the layout, lower metal layers were used to route signals within a circuit block, while upper metals were used as power and ground. The base model and the low leakage RF should both be fabricated on silicon in order to compare one sets of measurements to the other. However, the area granted was not enough to have two RF layouts, therefore the two models are merged into one structure. Circuits pertaining to word 0-15 were the base model, whereas circuits for word 16-31 include the innovation discussed. Moreover, since RF is of repetitive structure, all the bit slice have identical circuit block, only one bit slice was completely laid out for

simplification. Figure 7.1 illustrates the overall test chip layout. The numbers on the diagram correspond to the circuit blocks discussed in the following section.

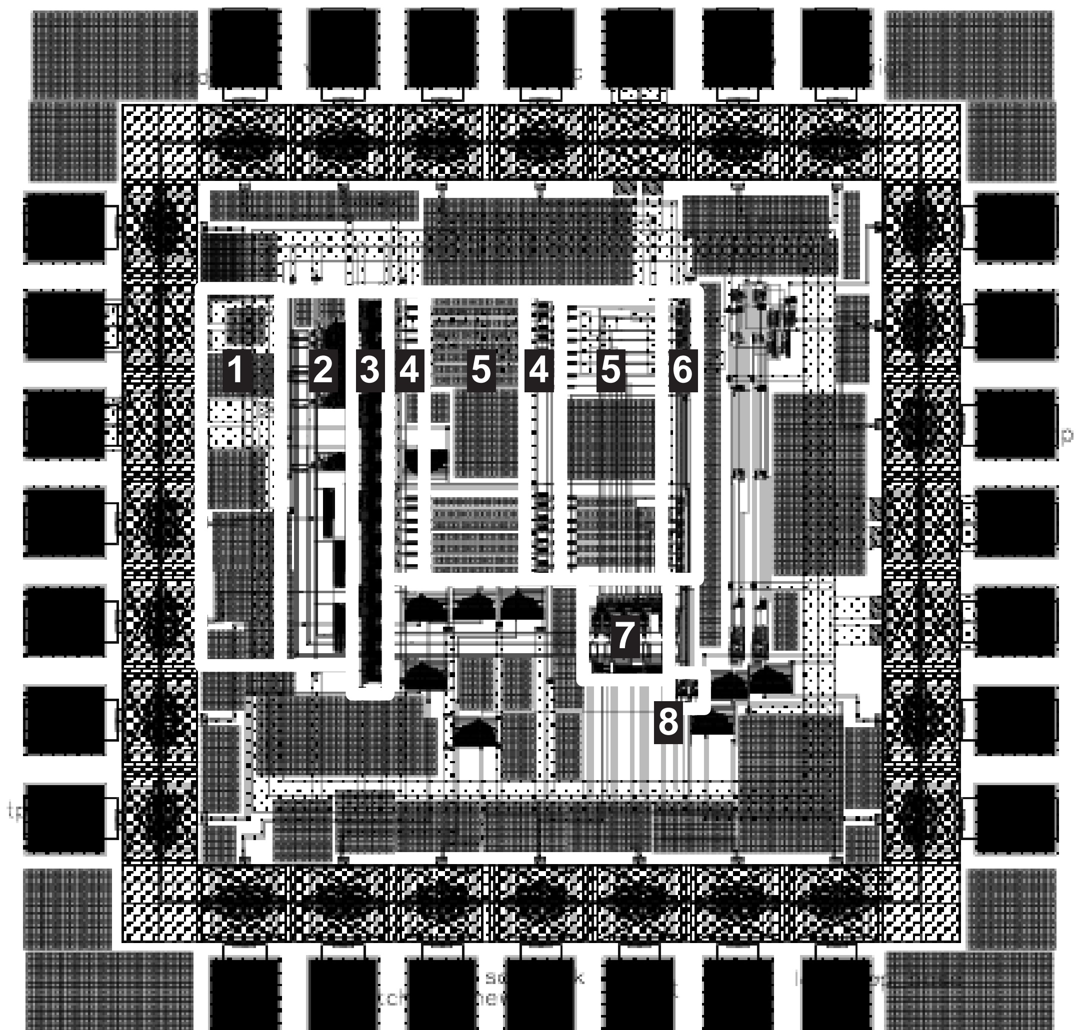


Figure 7.1: Layout of RF test chip

7.1 Test Chip Circuit Blocks and Layout

1. Input Scan Chain Flip-Flops

   This scan chain was used due to the fact that the number of pins available were not enough to accommodate all the input signals required. There were altogether 12 inputs, including read/write address bits, write data bit, read and write control, all

of them were tied to the scan chain, except for the clock signal. Since the scan chain is not part of the design, a standard cell from the CMC library was use.

2. Input Buffers and Input Flip-Flops

   The circuit of interest began at the input buffers, their function was to prepare the input signals with the required driving capability. The input flip-flops were of the master-slave logic. The master latches were transparent during the positive clock phase, and the input signal is latched in during this time. The slave latches then release the inputs to the RF structure during the negative clock phase. Although circuit organization remains the same for all inputs, their sizes differ according to the load at their output. A layout of a input buffer and flip-flop pair for an address input bit is shown in Figure 7.2.

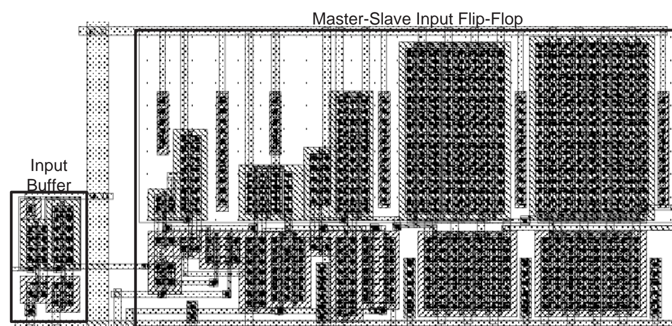


Figure 7.2: Layout of an address bit input buffer and flip-flop

3. Read Address Decoder

   The read address decoder was of 2:4/3:8 split decoder structure. Since only one address decoder was on the test chip, due to area restriction, the conventional model without any leakage control was used. The decoder had its own power supply pin for energy consumption measurements. Furthermore, one of its output is connected to a buffer chain, which drove an output pin. It is used to measure the propagation delay for the decoder individually. The layout is shown in Figure 7.3, the pitch for each bit matched that of a WL driver. The height of the decoder defines the height of the RF system.

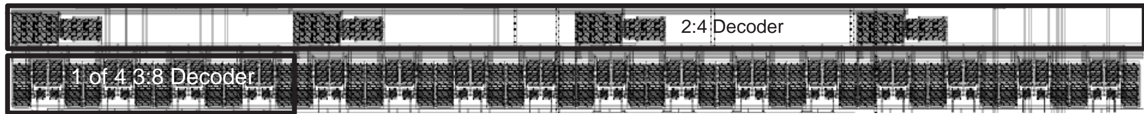


Figure 7.3: Layout of a read address decoder

4. Word Line Drivers

   The WL drivers for word 0-15 were the baseline model, whereas the ones for word 16-31 were the low-leakage design. Non-minimum channel length transistors were used for the PMOS in stage 2 and 4. As discussed previously, the four stages WL drivers are segmented into two sections. Stage 1 and 2 driving bit slice 16-31, while stage 3 and 4 drives bit slice 0-15. The layouts shown in Figure 7.4 are of the low-leakage design. The poly for PMOS in second and fourth stage was 10nm wider than the others.

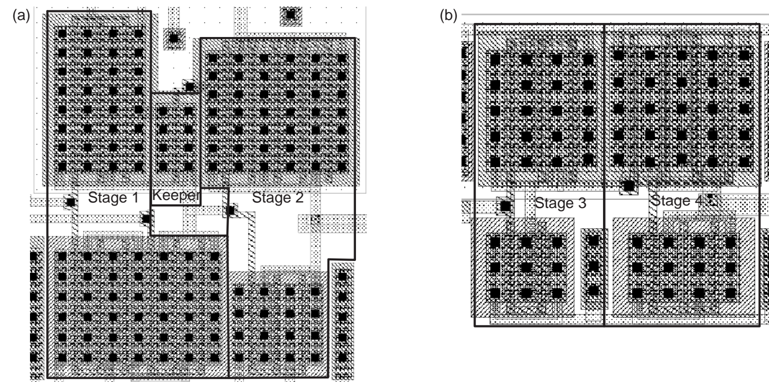


Figure 7.4: Layout of (a) Stage 1 and 2 (b) Stage 3 and 4 of WL Drivers

5. Memory cells

   Two types of memory cells were included in the test chip. The most significant 15 words used the low leakage SEIFMC, with a cell area of 5.56μm by 4.23μm. The rest of the words employed the conventional SEMC with an area of 6.79μm by 4.23μm. For the SEIFMC, it was possible to route the extra WL' for SEIFMC in the same metal layer as WL signal. Therefore, no extra metal layer was required. The layouts are illustrated in Figure 7.5.

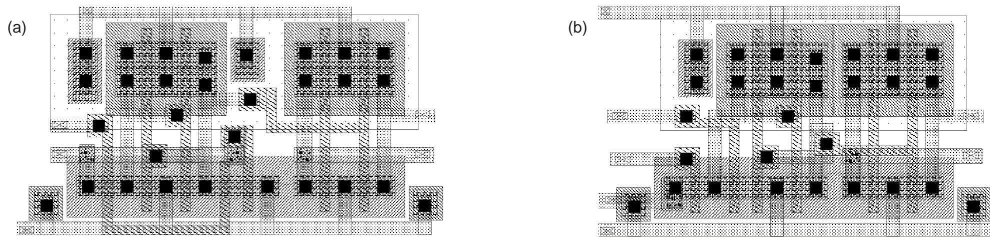


Figure 7.5: Layout of (a) SEMC (b) SEIFMC

6. Local Bit lines, NAND

   The LBL is a distributed structure, a 8-pulldown LBL spans the height of eight word lines. This is because the bottom transistors of the LBL pulldowns are connected to the output of the memory cells. As such, the LBL structures on the test chip were laid out in a distributive way. The LBL for word 0-15 is the baseline model, whereas that of word 16-31 had non-minimum channel length transistors for the WL pulldown transistors. The output of LBL was connected to a static NAND gate. For clarity, Figure 7.6 only illustrate two of the pulldown path of an LBL circuit with it's output connected to a NAND gate as in compound domino logic.

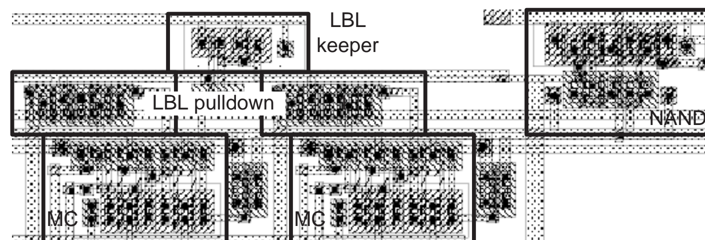


Figure 7.6: Layout of LBL pulldown paths and NAND

7. Global Bit Line and Output Latch

   The GBL for a 32-word by 32-bit structure is a simple domino gate with 2 pulldown paths connected to the static NAND gate outputs. Each output latch was sized to drive a load equivalent to a 20μm transistor. These two circuits were placed near the edge of silicon for easy routing to the output pin. Figure 7.7 shows the layout for such circuits. The RF test chip only has two output pins, corresponding to the two output of the completed bit slice. One of them was for the baseline model, the other

for the low-leakage RF.

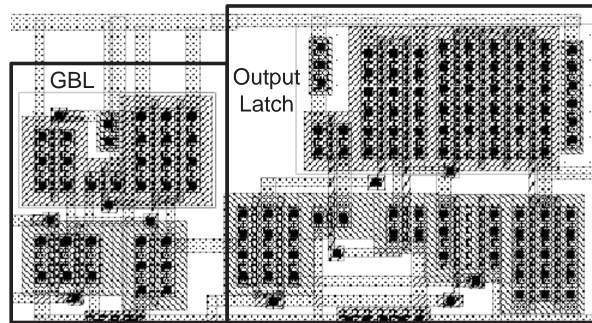


Figure 7.7: Layout of GBL and output latch

8. Write Circuitries

   The write circuitries includes input buffer and flip-flops for write address bits, and data. Since the write operation is not performance critical and not the focus of this study, the write circuits are scaled down significantly. The primary function is to setup the memory cell data for testing the read operations. The write circuits on the test chip did not reflect the write circuits in a full scale RF system, it is only designed to write to one bit slice.

7.2 Test Chip Performance

After the layout was completed with all design rules violation eliminated, a Layout vs. Schematic (LVS) was performed to ensure the layout reflects the schematic without any discrepancies. An extraction with parasitic capacitances was done and its functionality was tested. A read operation to word line 16 was performed after writing a "1" to the memory location. The resulting waveform is illustrated in Figure 7.8. The delay from the clock raising edge to the output rising edge is 553ps in the TSMC 0.18μm technology.

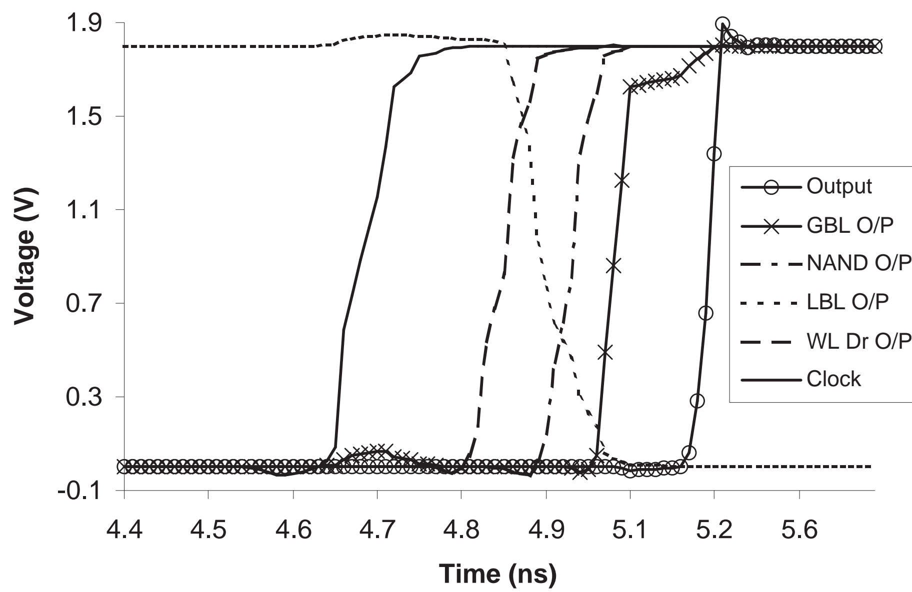


Figure 7.8: Read operation output waveform

Chapter 8

Conclusions

As the drive for better performance and higher density for integrated circuits moves forward, also as larger and more RF are used in microprocessors, the amount of leakage energy in a system proliferates. It has been shown that leakage energy may become the most dominant form of energy consumption, even surpasses that of switching energy. The need for effectual leakage control technique which imposes little performance penalty to a circuit is immense.

In this thesis, the operation of each essential RF circuit block is studied in detail. The possibilities of various leakage control implementations are examined. Effective leakage reduction techniques are carefully selected and implemented into the RF design with the condition of minimal performance penalty in mind. Finally, the 32-word by 32-bit RF was implemented in schematic level, and fully tested. A test chip was also designed, laid-out and simulations were performed to verify its functionality and performance.

The major contribution of this thesis have been:

1. Various low-leakage implementation have been studied and applied to the RF design. The design methodologies aim to have the most leakage reduction while maintaining high-performance.

2. A model has been developed to design energy-optimized wide-domino gates. These gates appear as LBL and GBL in an RF system.

3. A low-leakage, low-power single ended interrupted feedback memory cell is studied. A mathematical model has also been developed for SNM analysis.

4. The architecture is ported to the 90nm and 70nm technologies to study the effectiveness of the selected leakage control techniques in future generations. Also, the energy consumption for 256-word by 32-bit RF is projected from the 32-word by 32-bit RF simulation results. The results indicated that leakage energy is the dominant energy consumption channel in future processes and large RF systems.

Appendix A

Total Transistor Width Calculations for 5:32 Decoder Architectures

This appendix is dedicated to explain the calculation procedures in finding total transistor width along with the required input buffer for the different 5:32 decoder architectures as stated in Table 4.1. The static CMOS circuits are sized with a constant stage ratio of 1.5, equivalent output load of a 10μm transistor and a P/N ratio of 2.5. The input buffers are consist of 3 stages of inverter chains. Input signals to the decoder are connected to the input of the first inverter, the outputs of the second inverter are the true input values and are fed to the decoder circuit. Input complements are available at the output of the third inverters. Second stage inverters are the largest in size, this is because they drive the input capacitance of the decoder bit slices, as well as the one of the third stage inverters. In general, these buffers are sized according to the input capacitance of the decoder. Since all the bit slice of a decoder iterates all the input combination, each true or complement signal is connected to half of the total bit slices. Therefore these buffers are fairly large in size.

A.1 Single Stage 5:32 Decoder

The circuit structure for a single stage 5:32 decoder is illustrated in Figure 4.2(a), and it is stated in Figure A.1 with transistor sizing indicated beside each logic gate. The total transistor width for the decoder circuit is obtained by summing all transistor widths and

can calculated by the following equation:

$$32\times[3(7.7+9.3)+2(7.7+6.2)+3(14.3+1.9)] = 4074.9\mu m \tag{A.1}$$

The input buffers width varies between input A,B,C and D,E. This is because the input capacitance of a 3-input NAND gate is different from that of a 2-input NAND gate. Stage 3 input buffers for signal A,B,and C see a load of $16\times(7.7+9.3) = 272\mu m$, whereas that for input D and E see a load of $16\times(7.7+6.2) = 211.2\mu m$. Using a stage ratio of 1.5, all of these stages are sized accordingly. The total width for the input buffer stages is $3749.8\mu m$. Furthermore, the buffer for read/wirte enable signal is sized to a load of $32\times(14.28+1.9)\mu m$ and the width of buffer is $575.7\mu m$. Therefore, the total width for this configuration becomes $(4079.9+3749.8+575.7)=7824.7\mu m$.

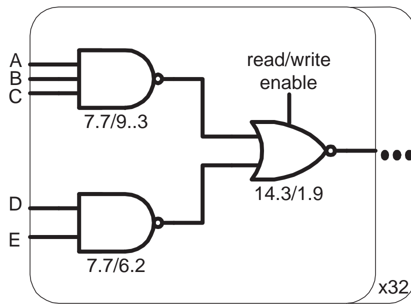


Figure A.1: Single stage 5:32 decoder

A.2 1:2/4:16 Split Decoder

The first stage of this configuration is simply a large inverter for the most significant address signal to generate split decoder signal A and A'. The third stage inverter in the input buffer chain can be used as the first stage circuitry. Different from other structures, the critical path of this configuration has 3 stages, as shown in Figure A.4. In order to have a fair comparison for equal delay across different architectures, the devices are sized larger to have better current drive. The stage ratio employed is 1.3. After 3 stages, the overall remaining stage ratio becomes $1.3^3 = 2.2$. When compared to the overall stage ratio of $1.5^2 = 2.25$ for 2 stage circuit, it is reasonably close. The total transistor width

for the second stage circuits can be calcuted as:

$$32\times[4(15.2+3)+2(8.9+7.1)+3(16.5+2.2)] = 5148.4\mu m \tag{A.2}$$

The buffer chain width for the most significant address bit is $753.4\mu m$, whereas that for the other four bits is $4\times(733.3) = 2933.2\mu m$. The enable signal buffer size is $664.7\mu m$. Therefore, the total width for this configuration is $(5148.4+753.4+2933.2+664.7)=9499.7\mu m$.

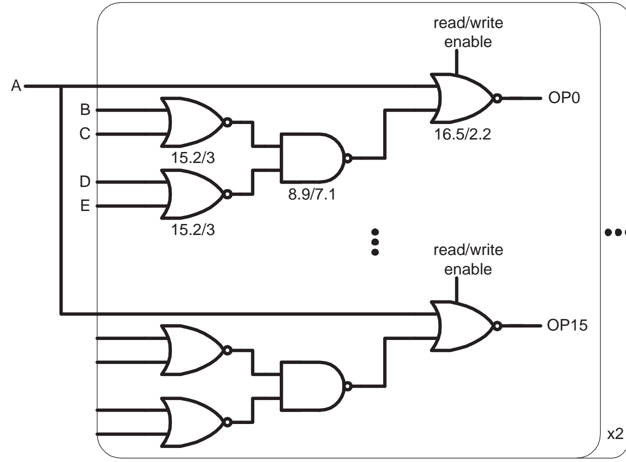


Figure A.2: Second stage circuit configuration of the 1:2/4:16 decoder scheme

A.3 2:4/3:8 Split Decoder

The circuit organization has been stated previously in Figure 4.2(b), it is included below in Figure A.3 along with the device feature sizes. Similar to the calculation procedure described above, the transistor width for the first stage decoder is

$$4\times[2(61.6+49.3)] = 887.2\mu m \tag{A.3}$$

Furthermore, transistor width for the second stage which consists of four 3:8 decoders is all the second stage bit slices width combined.

$$32\times[3(14.3+1.9)+3(7.7+9.3)] = 3187.2\mu m \tag{A.4}$$

The two most significant buffer chains have a total size of $2(558.6) = 1117.2\mu m$, whereas the other input butter sizes are $3(684.9) = 2054.7\mu m$. The read/write enable buffer size is

$575.7\mu m$. Therefore, the total transistor width for this configuration is $(887.2 + 3187.2 + 1117.2 + 2054.7 + 575.7) = 7822\mu m$.

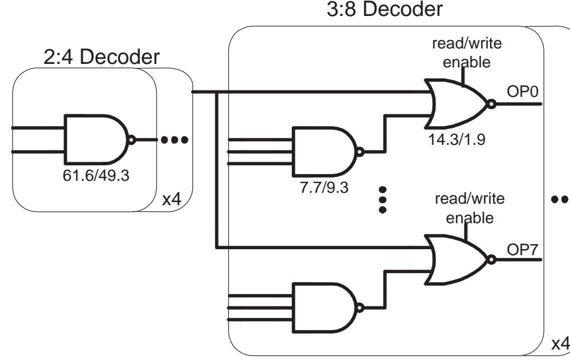


Figure A.3: Circuit configuration for 2:4/3:8 decoder

A.4 3:8/2:4 Split Decoder

The circuit structure of a 3:8/2:4 split decoder is similar to that of 2:4/3:8 decoder. The difference is that the first stage decoder is replaced by a 3:8 decoder which consists of 8 3-input NAND gates. The second stage is made up of eight 2:4 decoders. The total width for the first stage decoder is

$$8\times[3(30.8 + 37)] = 1627.2\mu m \qquad (A.5)$$

where as the width for the second stage is

$$32\times[2(7.7 + 6.2) + 3(14.3 + 1.9)] = 2442.9\mu m \qquad (A.6)$$

The buffer chains for the three most significant bits has a width of $3(682.4) = 2047.2\mu m$, whereas that for the remaining bits is $2(559.9) = 1119.8\mu m$. The read/write enable buffer size is $575.7\mu m$. Therefore, the total transistor width for this configuration is $(1627.2 + 2442.9 + 2047.2 + 1119.8 + 575.7) = 7812.8\mu m$.

A.5 4:16/1:2 Split Decoder

The circuit organization of this configuration is identical to that of 1:2/4:16 decoder. The only difference is that the 1:2 decoder becomes the least significant bit rather than the most

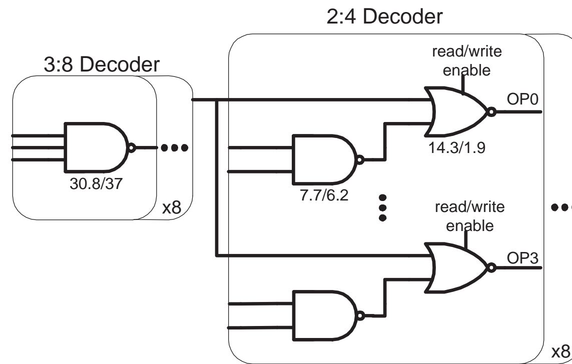


Figure A.4: Circuit configuration for 3:8/2:4 decoder

significant bit. As such, the total transistor width is identical to the formerly discussed 1:2/4:16 decoder, which is 9499.7μm.

Appendix B

Derivation of W_{MN}^{opt}

Capacitance at the output of a wide domino gate can be expressed as:

$$C_{DYN} \approx (0.03K_{GD}^{P} + K_{GD}^{N})nW_{MN} + C_{L}' \tag{B.1}$$

The total capacitance that switches during a transition includes C_{DYN}, the gate capacitance of the enabled pulldown device, C_{G}^{MN}, and the gate-drain capacitance of the keeper transistor C_{G}^{MPK}. Notice the gate-drain capacitance of the keeper inverter is neglected. Then, the total capacitance can be represented by:

$$C_{TOT} = C_{DYN} + C_{G}^{MN} + C_{G}^{MPK} \tag{B.2}$$

$$= (0.03K_{GD}^{P} + K_{GD}^{N})nW_{MN} + C_{L}' + K_{G}^{N}W_{MN} + K_{G}^{P}0.03nW_{MN} \tag{B.3}$$

For conciseness constants K_1-K_3 are used to represent the proportionality constants in front of the variable of interest, W_{MN}. Equation B.4 becomes:

$$C_{TOT} = K_1 nW_{MN} + C_{L}' + K_2 W_{MN} + K_3 nW_{MN} \tag{B.4}$$

Transistor models used in the study are the Berkeley Predictive Technology models. Simulation results indicate that the worst case I_{ON}/I_{OFF} ratio for a transistor is in the range of $\sim 10^4$ to 10^3. Hence for the 8-wide domino gate, the worst case leakage is limited to $< 1\%$. The energy model can be simplified by disregarding the leakage energy, and considering switching and short circuit energy only.

$$E_{TOT} \approx C_{TOT}V_{DD}^{2} + \frac{I_{SC}t_{SC}}{2}V_{DD} \tag{B.5}$$

$$= [C'_L + K_1 n W_{MN} + K_2 W_{MN} + K_3 n W_{MN}] V_{DD}^2 + K_4 W_{MN} t_{SC} V_{DD} \tag{B.6}$$

The delay model can be represented by:

$$t_{pHL} = (\frac{1}{2} - \frac{1 - \frac{V_{th}}{V_{DD}}}{1+\alpha}) t_T + \frac{C_{DYN} V_{DD}}{2 I_{D0}} \tag{B.7}$$

$$= K_6 + \frac{K_1 n W_{MN} + C'_L}{K_5 W_{MN}} \tag{B.8}$$

The optimal W_{MN} is obtained when the following condition is met:

$$\frac{\partial E_{TOT}}{\partial t_{pHL}} = \frac{\frac{\partial E_{TOT}}{\partial W_{MN}}}{\frac{\partial t_{pHL}}{\partial W_{MN}}} = -\frac{E_{TOT}}{t_{pHL}} \tag{B.9}$$

From Equation B.6 and B.8, the following is obtained:

$$\frac{\partial E_{TOT}}{\partial W_{MN}} = V_{DD}^2 (K_1 n + K_2 + K_3 n) + K_4 t_{SC} V_{DD} \tag{B.10}$$

$$\frac{\partial t_{pHL}}{\partial W_{MN}} = -\frac{C'_L}{K_5 W_{MN}} \tag{B.11}$$

Substituting Equation B.6 and B.8, as well as, Equation B.10 and B.11 into Equation . Also, make the assumption that C'_L is much smaller than the lumped C_{GD} of the pulldown transistors, and solve for W_{MN}:

$$W_{MN} = \frac{C'_L}{K_6 K_5 + K_1 n} \tag{B.12}$$

Therefore

$$W_{MN}^{opt} = \frac{C'_L}{(\frac{1}{2} - \frac{1 - \frac{V_{th}^N}{V_{DD}}}{1+\alpha}) t_T \cdot \frac{2K_I^N}{V_{DD}} + (0.03 K_{GD}^P + K_{GD}^N) n} \tag{B.13}$$

The relationship that W_{MN}^{opt} decreases while the number of paths, n, increases can be observed.

Bibliography

[1] J. Scott, L.H. Lee, J. Arends, B. Moyer, "Designing the Low-Power M.CORE Architecture", *IEEE Power Driven Microarchitecture Workshop*, pp. 145-150, 28 June 1998.

[2] J. M. Rabaey, *Digital Integrated Circuits, A Design Perspective*, Upper Saddle River, New Jersey: Prentice Hall, 1996.

[3] J. B. Kuo, *CMOS Digital IC*, Taiwan: McGraw Hill International Enterprise, 1996.

[4] D. De Venuto, *CMOS-Gate Transistor Sizing*, http://www-dee.poliba.it/dee-web/Personale/Devenuto/Didactical%20Activities/ProgettazioneAutomatica/Testo%20Appunti/lesson_22.PDF, 15 Jul. 2004

[5] *International Technology Roadmap for Semiconductors 2002 Update*, http://public.itrs.net/Files/2002Update/2002Update.pdf, 19 Jul. 2004.

[6] L. Wei, K. Roy, V. De, "Low Voltage Low Power CMOS Design Techniques for Deep Submicron ICs", *Thirteenth International Conference on VLSI Design 2000*, pp. 24-29, 3-7 Jan. 2000.

[7] A. P. Chandrakasan, F. Fox, W. J. Bowhill, *Design of High-Performance Microprocessor Circuits*, New York: IEEE Press, 2001.

[8] K. Roy, S. Mukhopadhyay, H. Manmoodi-Meimand, "Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits", *IEEE Transactions on Circuits and Systems I: Regular Papers*, Volume 51, pp.495-503, 26-29 March 2004.

[9] B. G. Streetman, *Solid State Electronic Devices*, Englewood Cliffs, New Jersey: Prentice Hall, 1995.

[10] J. Kao, S. Narendra, A. Chandrakasan, "Subthreshold Leakage Modeling and Reduction Techniques", *IEEE/ACM International Conference on Computer Aided Design*, pp. 141-148, 10-14 Nov. 2002.

[11] B. Van Zeghbroeck, *Principles of Semiconductor Devices*, http://ece-www.colorado.edu/∼bart/book/book/title.htm, 10 Jul. 2004.

[12] Y. Ye, S. Borkar, V. De, "A New Technique for Standby Leakage Reduction in High-Performance Circuits", *1998 Symposium on VLSI Circuits Digest of Technical Papers*, pp. 40-41, 11-13 Jun. 1998.

[13] S. Narendra et al, "Monocycle Shapes for Ultra Wideband System", *International Symposium on Low Power Electronics and Design 2001*, pp. 195-200, 6-7 Aug. 2001.

[14] S. Mutoh et al, "1-V Power Supply High-Speed Digital Circuit Technology with Multithreshold-Voltage CMOS", *IEEE Journal of Solid-State Circuits*, Volume 30, no. 8 pp. 847-854, Aug. 1995.

[15] J. Kao, A. Chandrakasan, "Dual-Threshold Voltage Techniques for Low-Power Digital Circuits", *IEEE Journal of Solida-State Circuits*, Volume 35, no. 7 pp. 1009-2000, Jul. 2000.

[16] B. Chatterjee, M. Sachdev, S. Hsu, R. Krishnamurthy, S. Borkar, "Effectiveness and Scaling Trends of Leakage Control Techniques for Sub-130nm CMOS Technologies", *International Symposium on Low Power Electronics and Design 2003*, pp. 122-127, 25-27 Aug. 2003.

[17] X. Liu, S. Mourad, "Performance of Submicron CMOS Devices and Gates with Substrate Biasing", *IEEE International Symposium on Circuits and Systems 2000*, Volume 4, pp. 9-12, 28-31 May 2000.

[18] A. Keshavarzi et al, "Effectiveness of Reverse Body Bias for Leakage Control in Scaled Dual Vt CMOS ICs", *International Symposium on Low Power Electronics and Design 2001*, pp. 207-212, 6-7 Aug, 2001.

[19] S. Hsu et al, "A 90nm 6.5GHz 256x64b Dual Supply Register File with Split Decoder Scheme", *Symposium on VLSI Circuits 2003*, pp. 237-238, 12-14 Jun. 2003 .

[20] B. Chatterjee, M. Sachdev, R. Krishnamurthy, "Leakage Control Techniques for Designing Robuts, Low Power Wide-OR Domino Logic for Sub 130-nm CMOS Technologies", *International Symposium on Quality Electronic Design 2004*, pp. 415-420, 22-24 Mar. 2004.

[21] C. Kwong, B. Chatterjee, M. Sachdev, "Modeling and Designing Energy-Delay Optimized Wide Domino Logic", *International Symposium on Circuits and Systems 2004*, Volume 1, Volume II, pp. 921-924, 23-27 May 2004.

[22] T. Sakurai, A. Newton, "Alpha-Power Law MOSFET Model and its Applications to CMOS Inverter Delay and Other Formulas", *IEEE Journal of Solid-State Circuits*, Volume 25, no. 2, pp. 584-594, Apr. 1990.

[23] E. Seevinck, F. List, J. Lohstroh, "Static-Noise Margin Analysis of MOS SRAM Cells", *IEEE Journal of Solid-State Circuits*, Volume SC-22, no. 5, pp. 748-754, Oct. 1987.

[24] O. Semenov, A. Pavlov, M. Sachdev, "Sub-quarter Micron SRAM Cells Stability in Low-Voltage Operation: A comparative Analysis", *IEEE International Integrated Reliability Workshop Final Report, 2002*, pp.597-600, 21-24 Oct 2002.

[25] T. Sakurai, A. Newton, "A Simple MOSFET Model for Circuit Analysis", *IEEE Transactions on Electron Devices*, Volume 38, no. 4 pp. 887-894, Apr. 1991.

[26] *Berkeley preidctive model*, http://www-device.eecs.berkeley.edu/~ptm/, 15 Jul. 2004