UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

INSTITUTO DE INFORMÁTICA

PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

PAULO FRANCISCO BUTZEN

Leakage Current Modeling in Sub-micrometer CMOS Complex Gates

Thesis presented in partial fulfillment of the requirements for the degree of Master of Computer Science

Prof. Dr. Renato Perez Ribas
Advisor

Porto Alegre, September 2007.

Livros Grátis

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

ACKNOWLEDGMENT

First, I would like to thank God for my wonderful life.

I would like to thank my family. My parents, Otavio and Gloria, that always look after me, for supporting, and for giving me the best of your heart. Thank you sister Paula, for being my roommate during my undergraduate and master studies, and especially for being my friend from the moment she was born.

Thank you Raquel, for being by my side with love and support since we first met.

I want to thank Professor Renato P. Ribas, my advisor, for the constant and unconditional support in my studies. More than my advisor, you are a good friend. I also want to express my gratitude to Professor Andre Reis, my co-advisor and friend, for all support since my first scholarship back in 2002, and to Professor Chris Kim, at the University of Minnesota, for have accepted me during my internship in his laboratory and for showing me other aspects of the research.

I would like to thank my colleagues from Nangate UFRGS Research Lab and University of Minnesota VLSI Research Lab for the good friendship and the excellent working environment.

Thank you my friends for giving me assistance to get through the difficulties. I want to thank you all for making my life so happy and fulfilled of emotion.

Finally, I acknowledge the CNPq for the scholarship during this Master.

TABLE OF CONTENTS

LIST OF ABREVIATIONS

BDD	Binary Diagram Decision
BPTM	Berkeley Predictive Technology Model
BTBT	Band-To-Band Tunneling
CAD	Computer-Aided Design
CMOS	Complementary Metal Oxide Semiconductor
CPU	Central Processor Unit
DIBL	Drain-Induced Barrier Lowering
DVTS	Dynamic V_{th} Scaling
ECB	Electron Conduction Band
EVB	Electron Valence Band
FBB	Forward Body Biasing
GIDL	Gate-Induced Drain Leakage
HVB	Hole Valence Band
MOS	Metal Oxide Semiconductor
MOSFET	Metal Oxide Semiconductor Field Effect Transistor
MTCMOS	Multi Threshold CMOS
RBB	Reverse Body Biasing
SCE	Short Channel Effects
SOI	Silicon-On-Insulator
VLSI	Very Large Scale Integration
VTCMOS	Variable Threshold CMOS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

To maintain performance at reduced power supply voltage, transistor threshold voltages and dimensions have been scaled down for decades. Scaling transistor into the sub-100nm technologies has resulted in a dramatic increase in leakage currents, which have become a significant portion of the total power consumption in scaled technologies, in many case achieving 30-50% of the overall power consumption under nominal operating conditions. For this condition, standby currents in CMOS logic gates represent an important challenge in nanometer technologies, leakage dissipation being a critical factor in low-power design. It means the static power dissipation should be considered as soon as possible in the integrated circuit design flow.

This thesis reviews the major leakage current mechanisms and several reduction techniques. It presents the development of a straightforward method for very fast estimation of subthreshold current in CMOS series-parallel logic gates. This estimation method is based on electrical conductivity association of series-parallel transistor arrangements. Combined with a gate oxide leakage model based on transistor bias condition, it is possible to provide a better prediction of total leakage consumption in transistor networks.

The previous estimation method is fast but it is not focused on accuracy. A new accurate subthreshold and gate leakage current estimation method is also developed based on simplified analytical leakage currents models. Instead of previous works focused on series-parallel device arrangements, this method evaluates the leakage in general transistor networks. The presence of on-switches in off-networks, ignored by previous works, is also considered in the proposed static current analysis. The new leakage model has been validated through electrical simulations, taking into account a 130nm and 90nm CMOS technology, with good correlation of the results, demonstrating the model accuracy.

Keywords: Leakage current, Low Power Circuits, CMOS

.

Modelagem de Corrente de Fugas em Portas Lógicas CMOS Submicrométricas

RESUMO

Para manter o desempenho a uma tensão de alimentação reduzida, a tensão de *threshold* e as dimensões dos transistores têm sido reduzidas por décadas. A miniaturização do transistor para tecnologias sub-100nm resulta em um expressivo incremento nas correntes de fuga, tornando-as parte significativa da potencia total, alcançando em muitos casos 30-50% de toda a potencia dissipada em condições normais de operação. Por estas condições, correntes estáticas em células CMOS representam um importante desafio em tecnologias nanométricas, tornando-se um fator crítico no design de circuitos de baixa potência. Isto significa que dissipação de potência estática deve ser considerada o quanto antes no fluxo de projetos de circuitos integrados.

Esta tese revisa os principais mecanismos de fuga e algumas técnicas de redução. Também é apresentado um modelo de estimativa rápida da corrente de *subthreshold* em células lógicas CMOS série - paralelo. Este método é baseado em associações de condutividade elétrica série – paralelo de transistores. Ao combinar com o modelo de estimativa da corrente de fuga de *gate* baseada nas condições estáticas dos transistores é possível fornecer uma melhor predição da corrente de fuga total em redes de transistores.

O modelo de estimativa anterior é rápido porem seu foco não esta na precisão. Um novo e preciso modelo para corrente de fuga de *subthreshold* e de *gate* é também apresentado baseado em modelos analíticos simplificados das correntes de fuga. Ao contrario do modelo anterior que era destinado a redes de transistores serie – paralelo, o novo método avalia as correntes de fuga em rede de transistores complexas. A presença de transistores conduzindo em redes de transistores não conduzindo, ignorados em trabalhos anteriores, é também avaliado no trabalho proposto. O novo modelo de corrente de fuga foi validado através de simulações elétricas, considerando processos CMOS 130nm e 90nm, com boa correlação dos resultados, demonstrando a precisão do modelo.

1 INTRODUCTION

In the past, the major concerns of the VLSI designers were performance and miniaturization. With the explosive growth in portable computing and wireless communication in the last few years, power dissipation has become a critical issue. Problems with heat removal and cooling are worsening because the magnitude of power dissipated per unit area is growing with scaling. Years ago, portable battery-powered applications were characterized by low computational requirement. Nowadays, these applications require the same computational performance as non-portable applications. It is important to prolong the battery life as much as possible. These are two reasons that power dissipation becomes a challenge for circuit designers and a critical factor in the future of microelectronics.

An integrated circuit is composed by sequential circuits, combinational circuits, memories blocks and I/O devices. Each one gives its own contribution to the total power dissipation in integrated circuits. Figure 1.1 shows the approximate power distribution of a microprocessor implemented in 0.5μm CMOS process (GRONOWSKI, 1998). Power consumption is concentrated in the logic circuits, 40 % in sequential blocks and 30 % in combinational blocks. Memory blocks and I/O device represent approximately 30% of the total power. Similar behavior is observed in different microprocessors (TAKAYANAGI, 2005) (LEON, 2007).
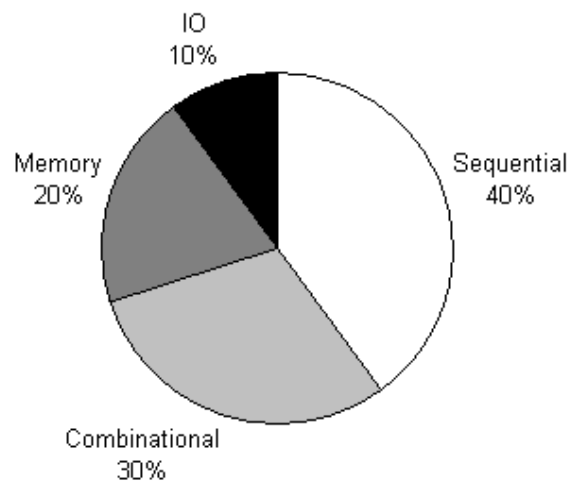
Figure 1.1: Power distribution of a 0.5μm CMOS microprocessor (GRONOWSKI, 1998).

There are four sources of power dissipation in digital CMOS circuits, as describe in equation 1.1.

$$P = P_{\text{dynamic}-switching} + P_{short-circuit} + P_{\text{static}-biasing} + P_{leakage} \tag{1.1}$$

where P is the total power dissipation, $P_{\text{dynamic-switching}}$ is the dynamic switching power, $P_{\text{short-circuit}}$ is the short-circuit power, $P_{\text{static-biasing}}$ is the static biasing power and P_{leakage} is the leakage power.

Dynamic switching power dissipation is caused by charging capacitances in the circuit. Considering C_L the model of routing and input gates capacitance, in a CMOS inverter, for instance, during each low-to-high output transition, C_L is charged through the PMOS transistor, and a certain amount of energy is drawn from the power supply. Part of this energy is dissipated in the PMOS device and part is stored on C_L. It is discharged during the high-to-low output transition, and the stored energy is dissipated in the NMOS transistor.
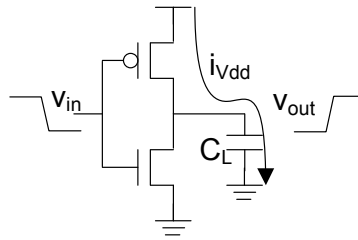
Figure 1.2: Dynamic switching power dissipation scheme in CMOS inverter.

Considering the CMOS inverter, shown in Figure 1.2, and assuming that the input waveform has zero rise and fall times, the energy consumption during low-to-high output transition can be derived by integrating the instantaneous power over the period of interest. Equation (1.2) shows that it draws $C_L.V_{dd}^2$ Joules from the power supply.

$$E_{V_{dd}} = \int_0^\infty i_{V_{dd}}(t)V_{dd}\,dt = V_{dd}\int_0^\infty C_L \frac{dv_{out}}{dt}\,dt = C_L V_{dd}\int_0^{V_{dd}} dv_{out} = C_L V_{dd}^2 \tag{1.2}$$

The charge stored on the load capacitor is equal to $C_L.V_{dd}^2/2$ by equation (1.3). This means that only half of the energy supplied by the power source is stored in C_L. The other half had been dissipated by the PMOS transistor. The high-to-low output transition dissipates the energy stored on the load capacitance into the NMOS transistor.

$$E_{C_L} = \int_0^\infty i_{V_{dd}}(t)v_{out}\,dt = \int_0^\infty C_L \frac{dv_{out}}{dt} v_{out}\,dt = C_L \int_0^{V_{dd}} v_{out}\,dv_{out} = \frac{C_L V_{dd}^2}{2} \tag{1.3}$$

To compute the power consumption, it is necessary to take into account how often the circuit is switched. Given a gate switching frequency f, the power drawn from the supply is given by:

$$P_{\text{dynamic}-switching} = C_L V_{dd}^2 f \tag{1.4}$$

The dynamic switching power dissipation was the dominant factor compared with the other components of power dissipation in digital CMOS circuits for technologies down to 0.18μm, where it is about 90% of total circuit dissipation (PARK, 2006).

14

Short–circuit power is the second source of total power dissipation described in equation (1.1). During a transient on input signal, there will be a period in which both NMOS and PMOS transistor will conduct simultaneously, causing a current flow through the direct path existing between power supply and ground terminals. This effect usually happens for very small intervals. In a static CMOS inverter this current flows as long as the input voltage is higher than a NMOS threshold voltage (V_{thn}) above ground and lower than a PMOS threshold voltage (V_{thp}) below the power supply, as shown in Figure 1.3. It is proportional to the input ramp, the output load, and the transistor size. It can be approximated by (VEENDRICK, 1984), according to equation (1.5)

$$P_{short-circuit} = K(V_{dd} - 2V_{th})^3.\tau.f \qquad (1.5)$$

where K is a constant that depends on the transistor sizes, as well as on the technology, V_{dd} is the supply voltage, V_{th} is the threshold voltage, τ is the rise or fall time of the input signal, and f is the clock frequency.
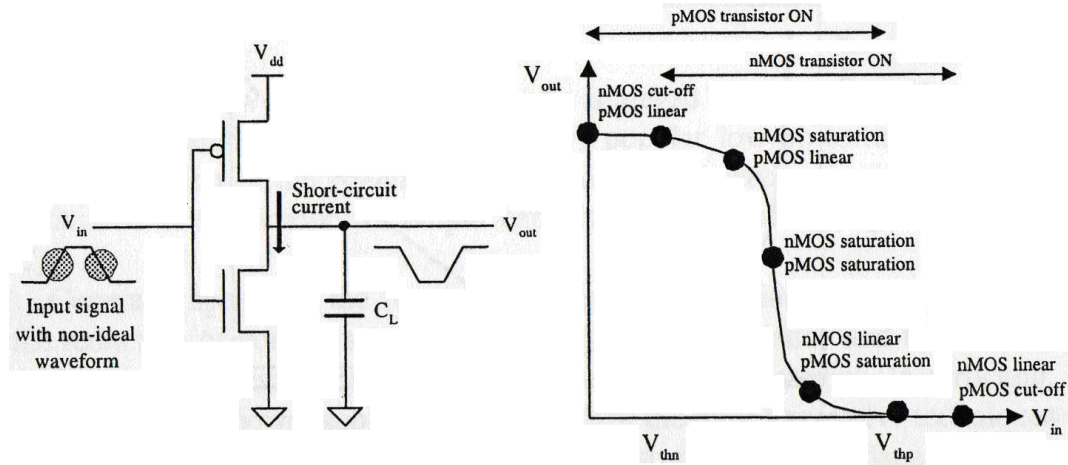
Figure 1.3: CMOS inverter short-circuit current (SOUNDRIS, 2002).

This component represents less than 20% of the dynamic switching power consumption if the NMOS and PMOS transistors are sized in order to balance the rise/fall signal slopes at input and output nodes (VEENDRICK, 1984).

Both of the above sources of power dissipation in CMOS circuits are related to transitions at gate terminals and for that reason are collective referred as dynamic power dissipation. On the other hand, the last two sources of power dissipation, static biasing and leakage, are related to the current that flows when the gate terminals are not changing, and are therefore collectively referred as static power dissipation.

Ideally, in the steady state of CMOS circuits there is no static power dissipation. This is the most attractive characteristic of CMOS technology. However, the actual operation of CMOS circuits is slightly different. Degraded voltage levels feeding CMOS gates and pseudo-NMOS logic family, present a current flow from the power supply to ground nodes. This flow is known as static biasing current.

In Figure 1.4, a NMOS pass-transistor drives an inverter. From basic CMOS circuit theory is known that the voltage in node A is degraded ($V_{dd}-V_{th}$). Since the inverter input is high ($V_{dd}-V_{th}$), its output should be low. However, the PMOS transistor will be weakly ON and, thus, present a static biasing current from power supply to ground nodes.
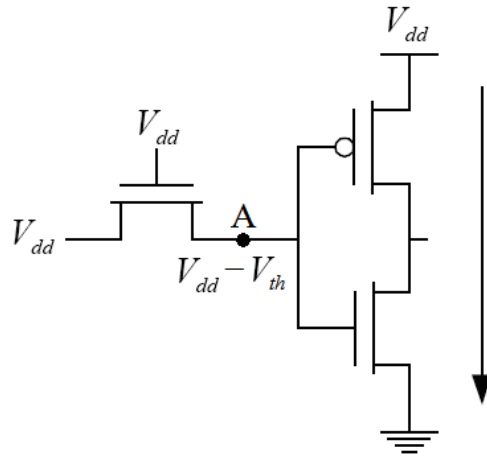
Figure 1.4: Degraded voltage level as input signal to an inverter results in static biasing power consumption.

Pseudo-NMOS logic gate consists of a single PMOS transistor, whose gate terminal is always grounded, and a NMOS pull-down network, which implements the boolean function. A pseudo-NMOS NAND gate is depicted in Figure 1.5. The main advantages of the pseudo-NMOS logic family are area and performance due to the inexistence of the PMOS pull-up network. However, when the NMOS pull-down network is conducting, there always exists a static biasing current from the power supply to ground, because the PMOS transistor is always ON.
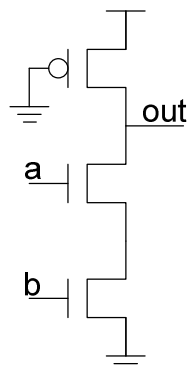
Figure 1.5: Pseudo-NMOS NAND2 gate.

The static biasing current only happen in specific conditions as reported above. Static current that flows from V_{dd} to ground nodes, without degraded inputs or in pseudo-NMOS logic family is known as leakage power. It is the main factor responsible for power dissipation during idle mode in standard CMOS gates. In past technologies the magnitude of leakage current was low and usually neglected. But as the devices have been being scaled to achieve higher density, performance, and lower dynamic power consumption, the leakage current in the nanometer regime is becoming a significant portion of power dissipation in CMOS circuits, as depicted in Figure 1.6.

16
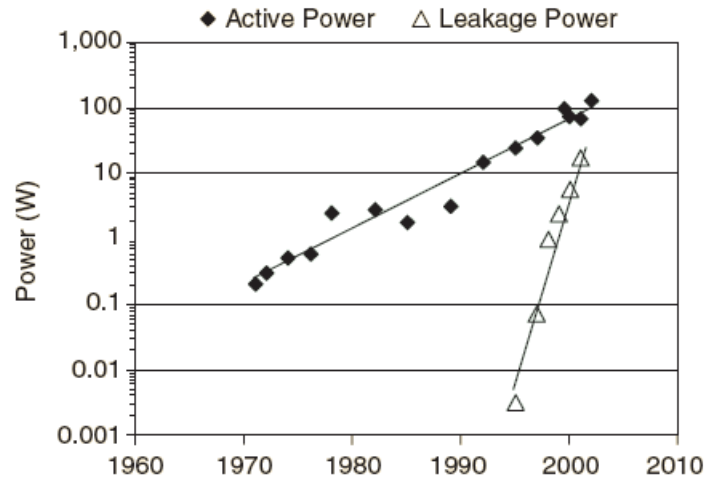
Figure 1.6: Active and leakage processor power (MOORE, 2003).

The power consumption reduction in digital systems involves optimization at all levels of design. This optimization includes the technology used to implement digital circuits, the design, the architecture, and the algorithm that are being implemented.

Optimization in technology level are related to materials used in fabrication process, like high-K gate dielectric and metal gates (SINGER, 2007), its dimension and concentration, like oxide thickness and substrate profile, and device structure, like "halo" doping (OGURA, 1980) and silicon-on-insulator (SOI) structures (DONAGHY, 2001). Design level involves optimization in physical and logic design. Place and route, and transistor sizing, are example of physical design optimization. Reduction in swing logic, logic minimization and technology mapping are example of logic design optimization techniques (CHANDRAKASAN, 1995). Architectural level typically presents solutions based on parallel or pipelined structures to achieve the same performance with a reduce supply voltage (ROY, 2000). Algorithm level explores the concurrency to be implemented in a parallel architecture and the minimization of the number of operations to reduce the switching activity, and consequently the dynamic consumption (CHANDRAKASAN, 1995).

When designing VLSI circuits, designers have to respect a power specification. Accurate and efficient power estimation during the design phase is required in order to meet the power specification without a costly redesign process. It is important to estimate both average and maximum power in CMOS circuits at different levels of design abstraction. The average power dissipation is important to determine battery life, while the maximum power demanded is related to circuit reliability issues.

This work tries to provide a deep understanding of the static power dissipation in CMOS circuits.

Different leakage mechanisms contribute to the total leakage current in MOS device. The three major types of leakage mechanisms are subthreshold, gate oxide and reverse-bias pn-junction leakage (BTBT – band-to-band tunneling). In addition to these three major leakage components, there are other ones like gate-induced drain leakage (GIDL) and punchthrough current. Those components can be neglected in normal modes of operation (AGARWAL, 2005). These leakage mechanisms are reported in Chapter 2.

The great majority of digital circuits are designed for the highest performance to satisfy the system frequency requirement. These circuits are typically composed by large gates, logic duplication and high parallel architectures. In this case the leakage power consumption is significant. However, not every application requires a fast circuit to operate at the highest level all the time. Some modules can be in idle mode often, and consequently, there is an opportunity to reduce the leakage power consumed. Chapter 2 also explores different circuit techniques to reduce leakage consumption.

The models to treat the leakage mechanisms (SHEU, 1987) (CAO, 2000) are still too complicated and they are hardly used by circuit designers. Precise simulators, such as HSPICE, can accurately account for leakage current, but they are only proper for small circuits due their solution convergence, explosion of memory and CPU time problems. Faster techniques to estimate the subthreshold and gate leakage current have been proposed in the literature (CHEN, 1998) (YANG, 2005). However, only basic series and parallel arrangements of transistor have been addressed. An improved subthreshold leakage model to be applied in general transistors networks is described in Chapter 3. Gate leakage model is presented in Chapter 4 and included in previous subthreshold model. Final conclusions are presented in Chapter 5.

2 STATIC CONSUMPTION

To achieve higher integration density and improved performance, CMOS devices have scaled down in each technology generation. However, static power dissipation has increased drastically with technology scaling and become a significant contributor to the total power dissipation in CMOS circuits. This chapter attempts to review the major leakage mechanisms and design techniques to reduce leakage power consumption in such technologies.

2.1 Leakage Current Mechanisms

For nanometer devices, leakage current is dominated by subthreshold leakage, gate-oxide tunneling leakage, and reverse-bias pn-junction leakage. Those three major leakage current mechanisms are illustrated in Figure 2.1. There are other leakage components, like gate induced drain leakage (GIDL) and punchthrough current, but those can be neglected in normal mode of operation (AGARWAL, 2005). In this section will be discussed those three major leakage mechanisms.
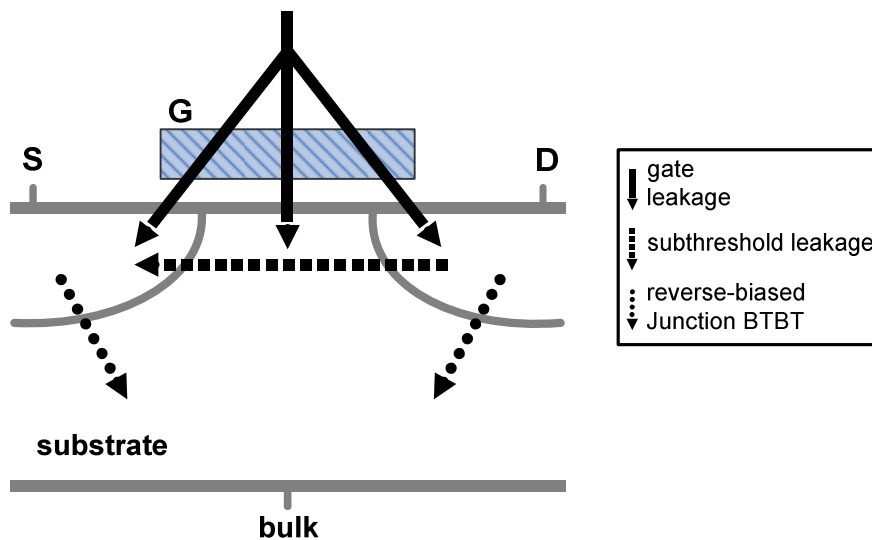
Figure 2.1: Major leakage mechanisms in MOS transistor.

2.1.1 Subthreshold Current

Supply voltage has been scaled down to keep dynamic power consumption under control. To maintain a high drive current capability, the threshold voltage has to be scaled too. However, the threshold voltage scaling results in increasing subthreshold leakage currents. Subthreshold current occurs between drain and source when transistor

is operating in weak inversion region, i.e., the gate voltage is below the threshold voltage.

The drain-to-source current is composed by drift current and diffusion current. The drift current is the dominant mechanism in strong inversion regime, when the gate-to-source voltage exceeds the threshold voltage. In weak inversion, the minority carrier concentration is almost zero, and the channel has no horizontal electric field, but a small longitudinal electric field appears due the drain-to-source voltage. In this situation, the carries move by diffusion between the source and the drain of MOS transistor. Therefore, the subthreshold current is dominated by diffusion current and it depends exponentially on both gate-to-source voltage and threshold voltage.

From the BSIM MOS transistor model (SHEU, 1987), the subthreshold leakage current for a MOSFET device can be expressed as:

$$I_{subthreshold} = I_0 e^{\frac{V_{gs} - V_{th}}{nV_T}} \left[1 - e^{-\frac{V_{ds}}{V_T}} \right] \tag{2.1}$$

where $I_0 = \dfrac{W\mu_0 C_{ox} V_T^2 e^{1.8}}{L}$, $V_T = \dfrac{KT}{q}$ is the thermal voltage, V_{th} is the threshold voltage, V_{ds} and V_{gs} are the drain-to-source and gate-to-source voltage respectively. W and L are the effective transistor width and length, respectively. C_{ox} is the gate oxide capacitance, μ_0 is the carrier mobility and n is the subthreshold swing coefficient.

In short channel devices, source and drain depletion regions penetrate significantly into the channel influencing the field and potential profile inside that. These are known as short channel effects (SCE). Such effects reduce transistor threshold voltage due to the channel length reduction (V_{th} roll-off) and the DIBL increasing. This results in significant subthreshold current in short channel devices.

2.1.2 Gate Tunneling Current

As mentioned before, the aggressive device scaling in nanometer regime increases short channel effects such as DIBL and V_{th} roll-off. To control the short channel effects, oxide thickness must also become thinner in each technology generation. Aggressive scaling of the oxide thickness, in turn, gives rise to high electric field, resulting in a high direct-tunneling current through transistor gate insulator.

Gate direct tunneling current is due to the tunneling of electrons (or holes) from the bulk and source/drain overlap region through the gate oxide potential barrier into the gate (or vice-versa). This phenomenon is related with the MOS capacitance concept. There are three major gate leakage mechanisms for a MOS structure. The first one is the electron conduction-band tunneling (ECB), which is due to the tunneling of electrons from conduction band of the substrate to the conduction band of the gate (or vice versa). The second one is the electron valence-band tunneling (EVB). It is due to the tunneling of electrons from the valence band of the substrate to the conduct band of the gate. The last one is known as hole valence-band (HVB) tunneling. It is due to the tunneling of holes from the valence band of the substrate to the valence band of the gate (or vice-versa). Figure 2.2 illustrates these three mechanisms.

20
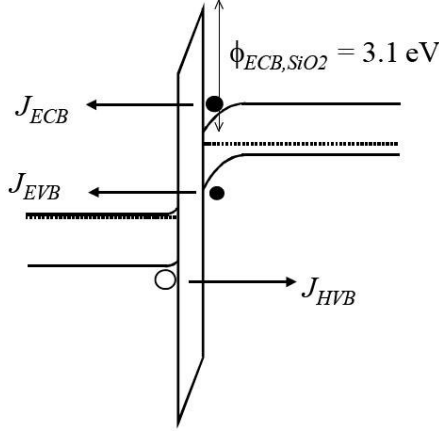
Figure 2.2: Three mechanisms of gate dielectric direct tunneling leakage (CAO, 2000).

Each mechanism is dominant or important in different regions of operation for NMOS and PMOS transistors. For each mechanism, gate leakage current can be modeled by (ROY, 2003):

$$I_{gate} = W.L.A\left(\frac{V_{ox}}{t_{ox}}\right)^2 \exp\left(\frac{-B\left[1-\left(1-\frac{V_{ox}}{\phi_{ox}}\right)^{3/2}\right]}{\frac{V_{ox}}{t_{ox}}}\right) \qquad (2.2)$$

where W and L are the effective transistor width and length, respectively, $A = q^3/16\pi^2 h\phi_{ox}$, $B = 4\pi\sqrt{2m_{ox}}\phi_{ox}^{3/2}/3hq$, m_{ox} is the effective mass of the tunneling particle, ϕ_{ox} is the tunneling barrier height, t_{ox} is the oxide thickness, h is $1/2\pi$ times Planck's constant and q is the electron charge.

2.1.3 Band-to-Band Tunneling Current

The MOS transistor has two pn junctions – drain and source to well junctions. These two pn junctions are typically reverse biased, causing a pn junction leakage current. This current is a function of junction area and doping concentration. When n and p regions are heavily doped, band-to-band tunneling (BTBT) leakage dominates the reverse biased pn junction leakage mechanism.

A high electric field across a reverse biased pn junction causes a current flow through the junction due to tunneling of electrons from the valence band of the p-region to the conduction band of the n-region, as shown in Figure 2.3.

Tunneling current occurs when the total voltage drop across the junction (applied reverse bias (V_{app}) + built-in voltage (ψ_{bi})) is larger than the band-gap. The tunneling current density through a silicon p-n junction is given by (ROY, 2003):

$$J_{BTBT} = A \frac{E V_{app}}{E_g^{1/2}} \exp\left(-B \frac{E_g^{3/2}}{E} \right) \qquad (2.3)$$

where $A = \sqrt{2m^* q^3} / 4\pi^3 h^2$, and $B = 4\sqrt{2m^*}/3hq$. m^* is the effective mass of electron; Eg is the energy-band gap; V_{app} is the applied reverse bias; E is the electric field at the junction; q is the electron charge; and h is $1/2\pi$ times the Planck's constant.
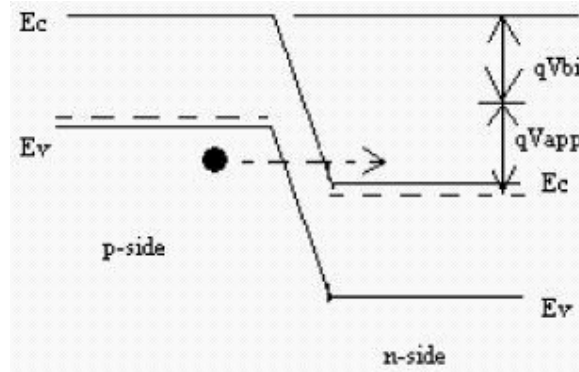
Figure 2.3: BTBT in reverse-biased pn junction (ROY, 2003).

Band-to-band tunneling leakage, negligible in current processes when compared to the subthreshold and gate leakages, starts to be taken into account in 25nm technologies (MUKHOPADHYAY, 2005).

The junction tunneling current depends exponentially on the junction doping and the reverse bias across the junction. Forward body bias can be used to reduce the band-to-band tunneling leakage.

2.2 Leakage Reduction Techniques

In CMOS circuit, the total power dissipation includes dynamic and static components during the active mode of operation. In the case of standby mode, the power dissipation is due to leakage currents. According to leakage mechanisms described in previous section, leakage power increases dramatically in the scaled devices. Particularly, with reduction of threshold voltage, to achieve high performance, leakage power becomes a significant component of total power consumption in both active and standby modes of operation.

To suppress power consumption in low-voltage circuits, it is necessary to reduce leakage power in both active and standby modes of operation. Reduction in leakage current can be achieved using both process and circuit level techniques. At process level, leakage reduction can be achieved by controlling the dimensions (length, oxide thickness, junction depth, etc.) and doping profile in transistor. At circuit level, several techniques to reduce leakage consumption have been proposed in the literature (ROY, 2003) (GUINDI, 2003) (PARK, 2006).

To reduce leakage currents, these techniques explore supply and threshold voltage leakage dependence, as well as concepts of stacking effect and body bias. Major focus of this section is to present several reduction techniques and the concepts explored in each one.

2.2.1 Dual Threshold CMOS

Dual threshold CMOS is a static technique that exploit the delay slack in non-critical paths to reduce leakage power. It provides both high and low threshold voltage transistors in a single chip that are used to deal with the leakage problem.

Fabrication process can achieve a different threshold voltage device by varying different parameters. Changing the channel doping profile, increasing the channel length, changing the body bias, and using a higher gate oxide thickness are examples of fabrications parameters that can be changed to achieve high threshold voltage transistor. Each parameter has its own trade-off in terms of process cost, effect on different leakage components, and short channel effects.

High V_{th} transistors suppress the subthreshold current, while low V_{th} transistors are used to achieve high performance. For a logic circuit, the transistors in non-critical paths can be assigned high threshold voltage to reduce subthreshold leakage current, while the performance is not sacrificed by using low V_{th} transistors in the critical paths (WEI, 1999). It has the same critical delay as the single low V_{th} CMOS circuits, while leakage power is saving in non-critical paths. Therefore, no additional control circuitry is required, and both high performance and low leakage power can be achieved simultaneously.

Dual threshold CMOS is effective in reducing leakage power during both standby and active modes without delay and area overhead. Researchers have proposed many others design techniques based on dual threshold CMOS. One considers upsizing a high V_{th} transistor to improve performance, but it is cause an area penalty (PANT, 1998). Figure 2.4 illustrates the basic idea of a dual V_{th} circuit.

critical path   noncritical path   low-with gate   high-with gate
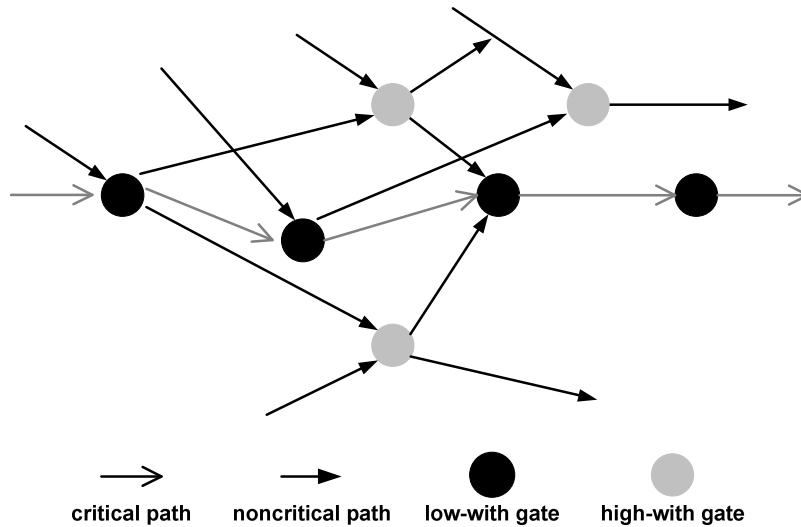
Figure 2.4: Dual V_{th} CMOS circuit (ROY, 2003).

With the increase in V_{th} variation and supply voltage scaling, it is becoming difficult to maintain sufficient gap among low V_{th}, high V_{th} and supply voltage required for dual V_{th} design. Furthermore, dual V_{th} design increases the number of critical paths in a die. It has been shown in (BOWMAN, 2002) that as the number of critical paths on a die increases, within-die delay variation causes both mean and standard deviation of the die frequency distribution to become smaller, resulting in reduced performance.

2.2.2 Supply Voltage Scaling

Supply voltage scaling is used to reduce dynamic and leakage power. It was originally developed for switching power reduction. It is an effective method of consumption reduction due to the quadratic dependence of the switching power in relation to supply voltage. Supply voltage scaling also provides leakage power savings.

Lowering supply voltage provides an exponential reduction in subthreshold current resulting from Drain-Induced Barrier Lowering (DIBL) effect. The DIBL effect tends to become more severe with process scaling to shorter gate lengths. For this reason, the achievable savings by this technique will increase with technology scaling.

Gate oxide leakage is also affected by this technique. Lowering V_{dd} will reduces gate leakage even faster than subthreshold leakage (KRISHNAMURTHY, 2002). Figure 2.5 shows how gate tunneling current reduces as V_{dd} decrease. Thus, this technique saves standby power by decreasing subthreshold and gate leakage currents.
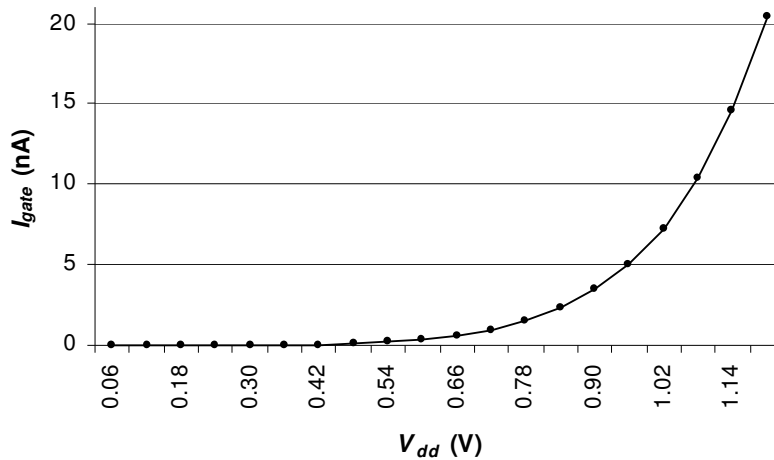
Figure 2.5: Gate oxide leakage current versus power supply.

In theory, the standby power supply for a circuit can decrease to zero, but the circuit will lose performance and all of it states. The optimal point for power savings using this technique is the lowest voltage for which the circuit retains state and does not compromising performance (WANG, 2006).

To achieve low-power benefits without compromising performance, two ways of lowering supply voltage can be employed: static supply scaling and dynamic supply scaling.

2.2.2.1   Static Supply Scaling

In static supply scaling, multiple supply voltages are used as shown in Figure 2.6. Critical and non-critical paths and/or units of the design are clustered and powered by higher and lower voltages, respectively (TAKAHASHI, 1998). In an extreme case the combinational logic in a circuit can fall all way to zero when the circuit is in idle mode because it does not need to hold state, increasing the power savings. Whenever an output from a low V_{dd} unit has to drive an input of a high V_{dd} unit, a level conversion is

needed at the interface. The secondary voltages may be generated off-chip (FUSE, 2001) or regulated on-die from the core supply (CARLEY, 1999).

Figure 2.6: Two-level multiple static supply voltage scheme.

2.2.2.2 *Dynamic Supply Scaling*

Dynamic supply scaling overrides the cost of using multiple supply voltages by adapting the single supply voltage to performance demand. When performance demand is low, supply voltage and clock frequency are lowered, delivering reduced performance with substantial power reduction (BURD, 2000).

As mentioned before, this technique gets rid of the cost of using multiple supply voltages. However, follow overheads are added when this technique is implemented:

- Circuit has to operate over a wide voltage range;

- Operating system to intelligently determine the processor speed;

- Regulator to generate the minimum voltage for specificspeed.

2.2.3 Transistor Stack Effect

Subthreshold leakage current flowing through a stack of series-connected transistors reduces when more than one transistor in the stack is turned off. This effect is known as the "stacking effect". It is best understood by considering a two transistor stack as illustrated in figure 2.6. When both transistor M1 and M2 are turned off, the voltage at the intermediate node (V_X) is positive due to small drain current. Positive potential at the intermediate node has three effects:

1) Due to the positive source potential V_X, gate to source voltage of transistor M1 (V_{gs1}) becomes negative; hence, the subthreshold current reduces substantially.

2) Due to $V_X > 0$, body to source potential (V_{bs1}) of transistor M1 becomes negative, resulting in an increase in the threshold voltage (larger body effect) of M1, and thus reducing the subthreshold leakage.

3) Due to $V_X > 0$, the drain to source potential (V_{ds1}) of transistor M1 decreases, resulting in an increase in the threshold voltage (less DIBL) of M1, and thus reducing the subthreshold leakage.
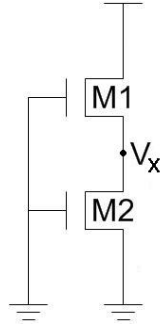
Figure 2.6: Subthreshold leakage current versus number of transistors off in stack.

The leakage of a two-transistor stack is about an order of magnitude less than the leakage in a single transistor. Figure 2.7 shows the subthreshold leakage current versus the number of off transistor in a stack.
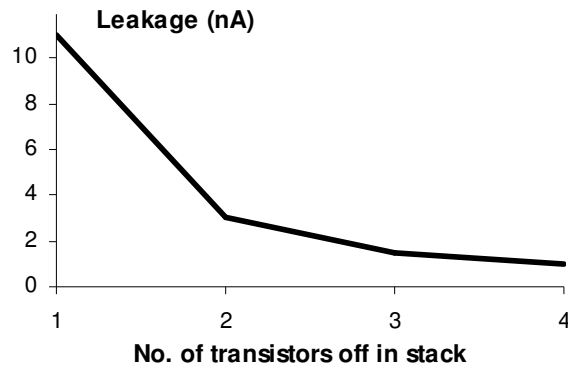
Figure 2.7: Subthreshold leakage current versus number of transistors off in stack.

2.2.3.1 Input Vector Dependence

Functional blocks such as NAND, NOR or other complex gates readily have a stack of transistors. Due to the stacking effect, the subthreshold leakage through a logic gate depends on the applied input vector. Maximizing the number of off transistors in a natural stack by applying proper input vectors can reduce the standby leakage of a functional block. Table 2.1 presents the input vector leakage dependence in a NAND gate for a 130nm process at 100 °C.

Table 2.1: Subthreshold leakage current for 2-input NAND gate.

Input Vector	Leakage current (nA)
00	3.94
01	15.25
10	13.65
11	4.57

Standby leakage power reduction due to the minimum leakage input vector is a very effective way of controlling the subthreshold current in the standby mode of circuit operation. The most straightforward way to find a low leakage input vector is to enumerate all input's combinations. For a circuit with n inputs, there are 2n input states combinations. Due to the exponential complexity with respect to the number of inputs, such an exhaustive method is limited to circuits with a small number of primary inputs. For large circuits, a random search-based technique can be used to find the best input vector.

Gate and band-to-band tunneling leakage are also important in scaled technologies, and can be a significant portion of total leakage. The input vector control technique using a stack of transistors needs to be reinvestigated to effectively reduce the total leakage.

Researchers have shown that with high gate leakage, the traditional way of using stacked transistors fails to reduce leakage and in the worst case might increase the overall leakage (MUKHOPADHYAY, 2003). In scaled technologies where gate leakage dominates the total leakage, using "10" might produce more savings in leakage as compared to "00". The gate leakage depends on the voltage drop across the transistor gate oxide. Applying "00" as the input to a two transistors stack reduces subthreshold leakage and does not change the gate leakage component. It has been shown that using "10" reduces the voltage drop across the terminals, where the gate leakage dominates, thereby lowering the gate leakage while offering marginal improvement in subthreshold leakage (MUKHOPADHYAY, 2003).

Band-to-band tunneling leakage is a weak function of input voltage and hence can be neglect it in this analysis (AGARWAL, 2006)

2.2.3.2 Stacking Single Switch

In CMOS complex gates, a certain number of transistor stacking (branches), between the supply voltage or ground nodes and the output node, can be observed. Such branches have usually different amounts of transistors. The basic idea of this technique is to duplicate transistors without increasing the longest transistor path or branch, expecting that the worst-case delay of the logic cell remains the same (BUTZEN, 2006). This procedure is applied to both pull-up (PMOS network) and pull-down (NMOS network) separately. Figure 2.8 (a) presents an original circuit and Figure 2.8 (b) illustrates the optimized circuit, resulted from the developed method described above.

Figure 2.8: Original (a) and leakage (b) optimized CMOS gate.

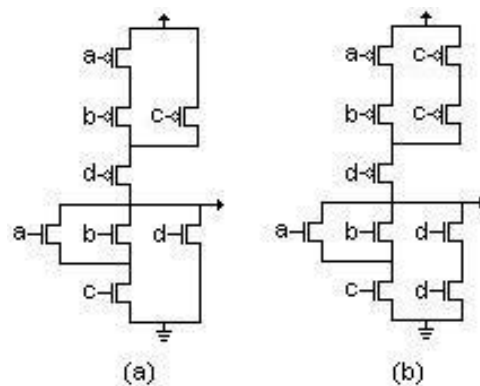Stacking single switch leakage reduction methodology has been tested for ISCAS85 benchmark c8 circuit gates. The c8 circuit was optimized and synthesized using Berkeley SIS tool and mapped into library 44-6.genlib. This results in a circuit with 490 transistors mapped in 40 gates. These gates are optimizing to leakage reduction through the CAD tool described in previous section. For leakage characterization, the 65nm BSIM4 Berkeley Predictive Technology Model (BPTM, 2007) has been considered. Original and leakage optimized gates are evaluated through a DC simulation using HSPICE. The experimental results are:

In the c8 mapped circuit, the design technique modified 13 gates of the 40 total gates. Considering the area issue, the benchmark circuit had a total of 490 transistors. The design methodology duplicated 22 transistors, 4.5% of total. This represents a leakage current reduction in 32.5% of gates of circuit, increasing only 4.5% of the circuit area. The design methodology also generates a delay penalty due to the transistor duplication. This penalty was considered not so significant, about 2%. Finally, the leakage reduction in the optimized gates ranged from 7% to 22%. The average leakage reduction is around 11%.

2.2.4 Power Gating

Power gating technique uses the power supply voltage as the primary source for reducing leakage current. It refers to using a MOSFET switch (sleep transistor) to cut off, or gate, a circuit from the power rails (V_{dd} and/or ground) during standby mode. The power gating switch typically is positioned as header between the circuit and the power supply or as footer between the circuit and the ground. During active operation, the power gating switch remains on, supplying the current that the circuit uses to operate. During standby mode, turning off the power gating switch reduces the current dissipated through the circuit.

Turning off the sleep transistor provides leakage reduction for two primary reasons. First, the width of the sleep transistor is usually less than total width of transistors being gated. The smaller width provides a linear reduction in the total current drawn from supply node during standby mode. Secondly, leakage currents diminish whenever stacks of transistors are off due to the source biasing effect.

During active mode, the same effects cause a degradation of circuit performance. Even though the on-resistance of the power gating switch is much less than its off-resistance, it still creates a small positive voltage at the virtual node. Again, these voltages reduce the drive capability and increase the threshold voltage of the NMOS devices through body biasing. Hence, this technique is typically used for paths that are non-critical.

2.2.4.1 MTCMOS

Multi-Threshold CMOS (MTCMOS) is a popular power gating approach that uses high V_{th} devices for power switches (MUTOH, 1995). Figure 2.9 shows the basic MTCMOS structure, where a low V_{th} computational block uses high V_{th} switches for power gating. The low V_{th} transistor in the logic gate allows them to provide a high performance operation. However, by introducing a series device to the power supplies, MTCMOS circuits incur a performance penalty compared to CMOS circuits. Subthreshold leakage reduction behavior of a MTCMOS circuit is characterized by the threshold voltage and width of sleep transistor and due to the stack effect.

In fact, only one type of high V_{th} transistor is sufficient for leakage reduction. The NMOS insertion scheme is preferable, since the NMOS on-resistance is smaller at the same width and hence it can be sized smaller than a corresponding PMOS (KAO, 1997).

However, MTCMOS can only reduce leakage power in standby mode and a large insertion of sleep transistors can increase significantly area and delay. Moreover, when data retention is required in standby mode, an extra high V_{th} memory circuit is needed to maintain the data (SHIGEMATSU, 1997).
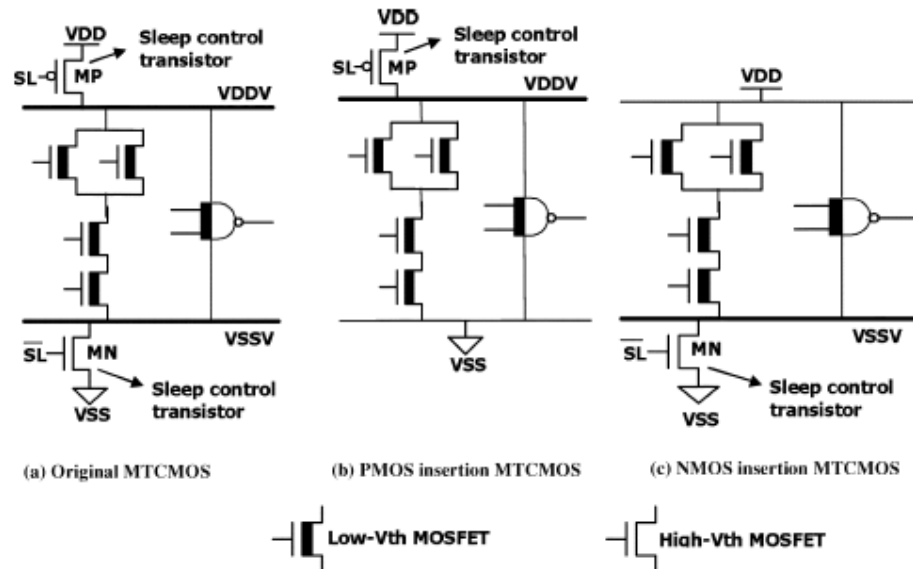
Figure 2.9: Schematic of MTCMOS circuit (ROY, 2003).

2.2.5 Body Biasing

Reverse body biasing (RBB) has been used in commercial memory chips since the 1970s, in order to mitigate the risk of memory data destruction. In logic chips, on the other hand, the substrate and wells are typically biased stably to the ground and power supply. However, since the 1990s, reverse body biasing has been applied in logic chips for a different reason: power reduction.

The original propose of the substrate biasing was utilized to reduce sub-threshold leakage in standby mode for portable applications. More recently, it has been employed to reduce the maximum power dissipation by lowering V_{th} (forward body biasing) in active mode, and by compensating V_{th} variations.

2.2.5.1 *Variable Threshold CMOS (VTCMOS)*

Variable threshold CMOS is a body biasing design technique (KURODA, 1996). Figure 2.10 shows the VTCMOS scheme. To achieve different threshold voltages, this scheme uses a self-substrate bias circuit to control the body bias.
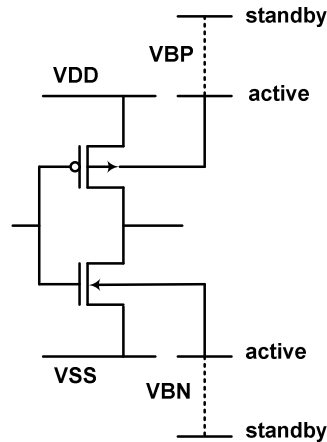
Figure 2.10: Schematic of VTCMOS technique.

In the active mode, VTCMOS technique applies a zero body bias (ZBB). As the subthreshold leakage current depends strongly on threshold voltage, in standby mode, a deep reverse body bias is applied to increase the V_{th} and save leakage power. However, this reduction technique has an overhead in chip area due to additional routing required to provide the body bias voltage.

Reverse body bias can reduce circuit leakage by three orders of magnitude in a 0.35μm CMOS technology (KESHAVARZI, 2001). However, more recent data shows that the effectiveness of RBB to lower I_{off} decreases as technology scales due to the exponential increase in band-to-band tunneling leakage at the source/substrate and drain/substrate pn junctions (KESHAVARZI, 2001). Moreover, smaller channel length with technology scaling and lower channel doping (to reduce V_{th}) worsen the short channel effect and diminish the body effect. This, in turns, weakens the V_{th} modulation capability of RBB.

For scaled technologies, recent design has been proposed using forward body biasing (FBB) to achieve better current drive with less short channel effect (NARENDRA, 2003). Circuit is designed using high V_{th} transistor (high channel doping) reducing leakage in standby mode, while FBB is applied in active mode to achieve high performance. Both high channel doping and FBB reduce short channel effect relaxing the scalability limit of channel length due to V_{th} roll off and DIBL. This result in higher I_{on} compared to low V_{th} design for same I_{off} worst case, improving performance. RBB can also be applied in standby mode together with FBB to further reduce the leakage current.

It has been shown that FBB and *high-V_{th}* along with RBB reduces leakage by 20X, as opposed to 3X for the RBB and *low-V_{th}* (NARENDRA, 2003). However, FBB devices has larger junction capacitance and body effect, which reduces the delay improvement mainly in stacked circuits.

2.2.5.2 *Dynamic Vth Scaling*

Not every application requires a fast circuit to operate at the highest performance level all the time. Active leakage techniques exploit this idea to intermittently slow down the fast circuitry and reduce the leakage power consumption as well as the dynamic power consumption when maximum performance is not required.

Dynamic V_{th} scaling (DVTS) scheme uses body biasing to adjust V_{th} based on the performance demand. (KIM, 2002) This technique uses the same concept of variable threshold CMOS (VTCMOS). While VTCMOS change V_{th} based on active and standby mode, DVTS modifies the V_{th} based on circuit performance demand, i.e., during active mode the threshold voltage can be changed depending on the system performance requirement.

The lowest V_{th} is delivered via zero body bias, if highest performance is required. When performance demand is low, clock frequency is lowered and V_{th} is raised via reverse body bias to reduce the run-time leakage power dissipation. In cases when there is no workload at all, the V_{th} can be increased to its upper limit to significantly reduce the standby leakage power. This scheme deliveries just enough throughput for the current workload by tracking the optimal V_{th}. It is considerably reduces leakage power by intermittently lowing down the circuit. A block diagram of the DVTS scheme and its feedback loop are presented in Figure 2.11.
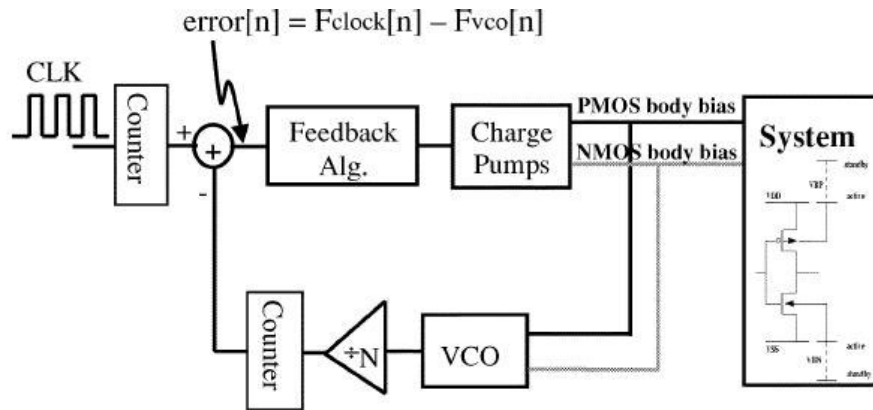

Figure 2.11: Schematic of DVTS hardware (KIM, 2002).

3 SUBTHRESHOLD LEAKAGE MODEL

This chapter reviews subthreshold leakage current models. A simplified model based on conductance association in series-parallel CMOS gates is demonstrated. Moreover, an improved analytical model based on physical parameters is proposed to general networks. It takes into account both drain induced barrier lowering (DIBL) and body effect, and can evaluate any complex gate. All analysis presented in this chapter use NMOS pull-down network. The same analysis is applicable to PMOS pull-up tree. First of all, the simple and straightforward method to fast subthreshold leakage prediction is presented. Next, three analytical subthreshold leakage models reported in the literature are discussed, and then, a detailed and complete analytical model for complex gates is described. Finally, at the end of the chapter, HSPICE simulations are used to validate the improved analytical model and to verify its accuracy.

3.1 Estimation Based on Conductance Association

The main objective of this simple method is to provide a normalized current value, related to a reference MOS off-switch, in order to use it as leakage cost of logic networks in technology mapping. It means the relative leakage prediction comparison and ordering of different off-networks is more important than the accuracy improvement in estimating the absolute leakage current values. Making so, the matching task during the mapping is able to take into account, among other design metrics, the static consumption of the logic gates identified in this process. It is crucial in library-free technology mapping (GAVRILOV,1997), where cells are not pre-characterized but automatically generated by software on-the-fly during the logic matching, in the concept of using virtual libraries in ASIC design.

Proposed method is based on the device electrical conductance association, that is, the conduction of parallel devices are summed while in series arrangements the equivalent conductance is inversely proportional to the number of devices. Being $G_{T[n]}$ the conductance of the n-index transistor in the arrangement, the equivalent conductance G_{eq} of parallel and series arrangements are given by:

- Parallel – $G_{eq} = G_{T[1]} + G_{T[2]} + \ldots + G_{T[n]}$

- Series – $G_{eq} = 1/(1/G_{T[1]} + 1/G_{T[2]} + \ldots + 1/G_{T[n]})$

At this moment, all transistors will be considered with equal size, and thus the individual device conductance $G_{T[i]}$ can be made unitary, normalized in respect to a reference transistor.

In more complex arrangements, the same principle is suitable for parallel and series associations. To exemplify this method, the transistor network illustrated in Figure 3.1 presents the following calculation:

$$G_{eq} = \cfrac{K}{\cfrac{1}{\cfrac{1}{\cfrac{1}{G_{T5}}+\cfrac{1}{G_{T6}}}+\cfrac{1}{\cfrac{1}{G_{T4}}+\cfrac{1}{G_{T3}+\cfrac{1}{\cfrac{1}{G_{T1}}+\cfrac{1}{G_{T2}}}}}}+\cfrac{1}{G_{T7}}+\cfrac{1}{G_{T8}+\cfrac{1}{\cfrac{1}{G_{T9}}+\cfrac{1}{G_{T10}}}}} \qquad (3.1)$$
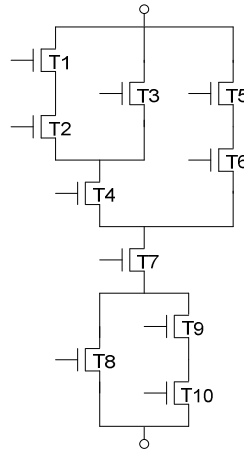
Figure 3.1: Transistor network with multi-level of series-parallel associations.

In the case of series transistor, the leakage reduction from a single off-device to two stacked off-transistors depends also on the fabrication process parameters (GU, 1996). As a result, a constant K must be included in the last step of the calculation procedure in order to calibrate the final result. This K value is obtained by relating the leakage current of two-stack and single off-device configurations. In this sense, two constants K_n and K_p may be derived according to NMOS and PMOS arrangements, respectively.

Furthermore, since NMOS and PMOS transistors present different subthreshold current behavior, such difference may be characterized to include in the same calculation both pull-up PMOS and pull-down NMOS planes in CMOS gates. Such relationship is represented by the constant K_{pn}.

Different CMOS logic gates were evaluated for all input signals combinations, resulting thus in a great variety of off-networks for subthreshold leakage estimation. These CMOS arrangements are shown in Figure 3.2 and Figure 3.3.

Table 3.1 and Figure 3.4 show results of the CMOS gates depicted in Fig. 3.3 (h), for each input vector. Note that the main goal of this simple method is to identify the less leakage consuming input vector, as well as to compare different CMOS arrangements in order to guide the technology mapping task in terms of static power dissipation. It also can be observed in Table 3.2 and Figure 3.5, where the worst-case and the average current values are given for the cells presented in Figure 3.2 and Figure 3.3.
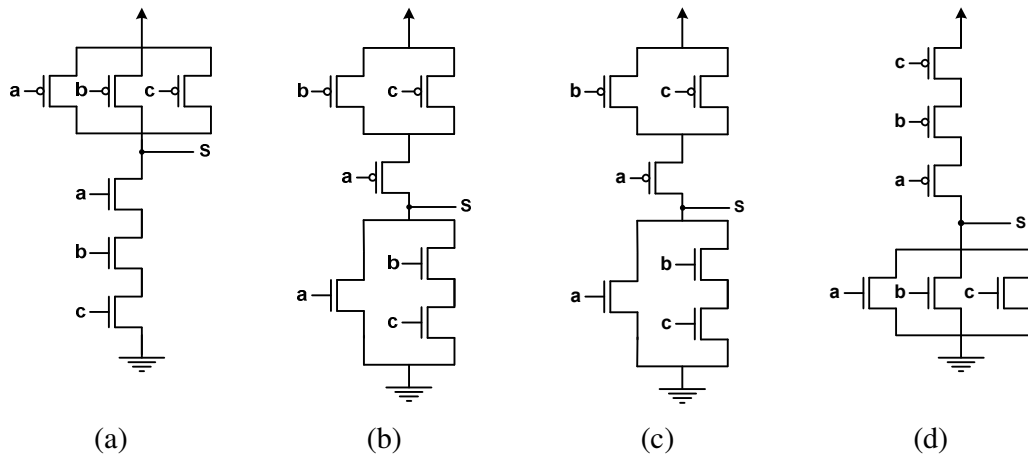
Figure 3.2: Different transistor arrangements: 3-inputs series-parallel CMOS gates.
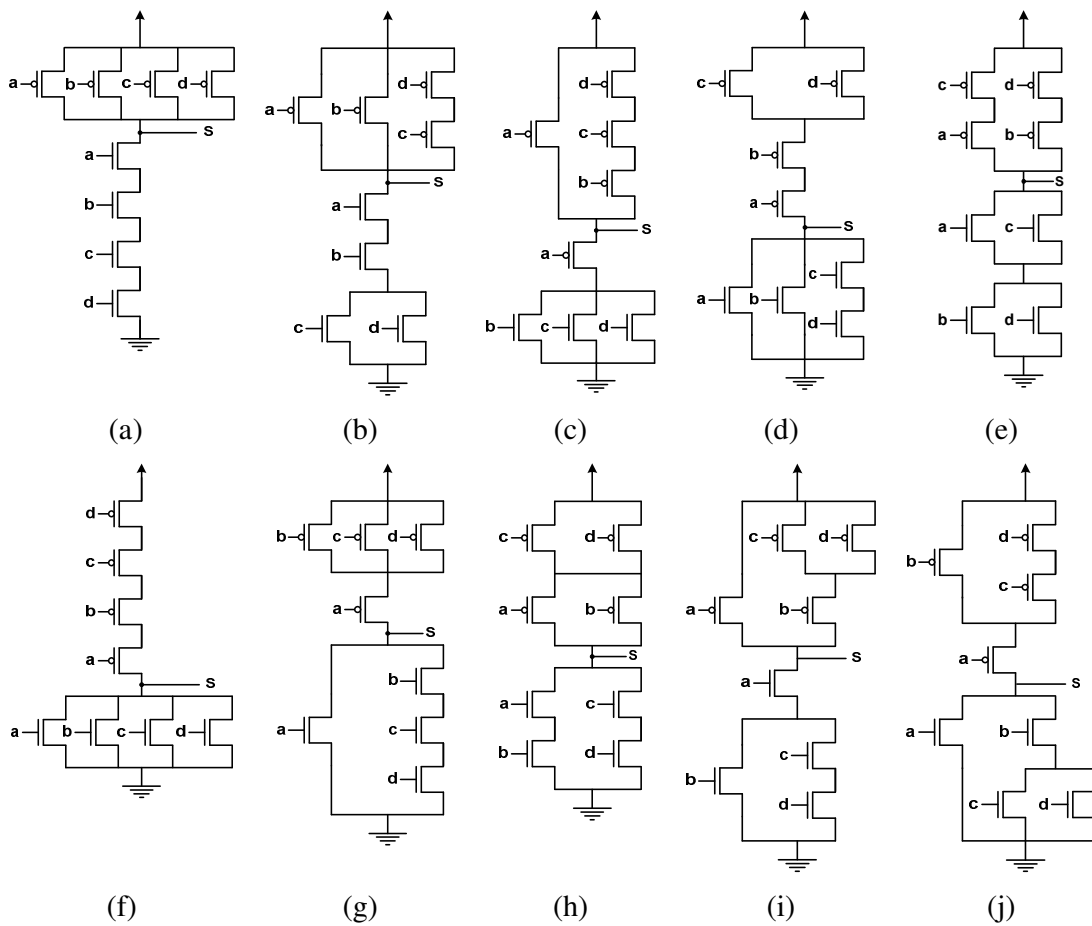
Figure 3.3: Different transistor arrangements: 4-inputs series-parallel CMOS gates.

Table 3.1: Normalized subthreshold leakage current in Figure 3.3 (h)

Input Vector	Hspice Simulation	Proposed Method	Diff (%)
0000	*0.49*	*0.52*	*6.12*
0001	*1.43*	*1.26*	*11.89*
0010	*1.10*	*1.26*	*14.55*
0011	*1.16*	*1.32*	*13.79*
0100	*1.43*	*1.26*	*11.89*
0101	*2.38*	*2.00*	*15.97*
0110	*2.05*	*2.00*	*2.44*
0111	*1.11*	*1.32*	*18.92*
1000	*1.10*	*1.26*	*14.55*
1001	*2.05*	*2.00*	*2.44*
1010	*1.72*	*2.00*	*16.28*
1011	*1.11*	*1.32*	*18.92*
1100	*1.53*	*1.32*	*13.73*
1101	*1.52*	*1.32*	*13.16*
1110	*1.52*	*1.32*	*13.16*
1111	*0.33*	*0.36*	*9.09*
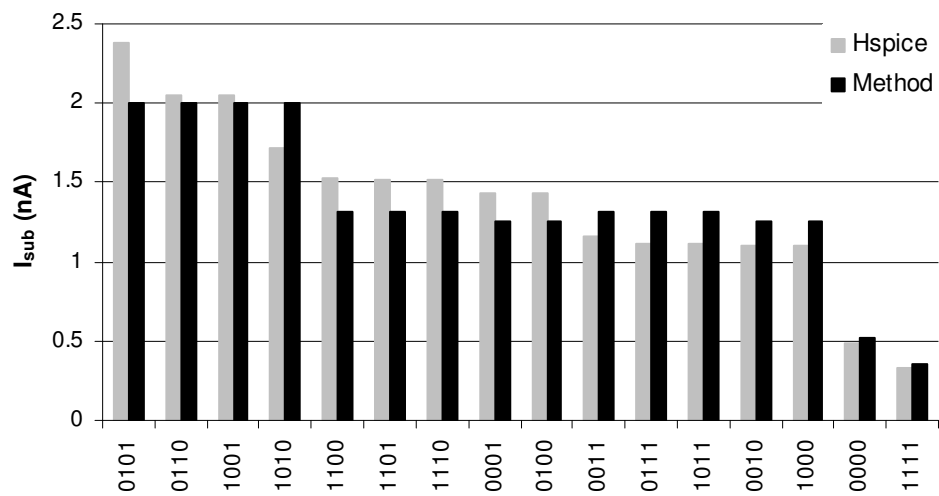
Figure 3.4: Subthreshold leakage currents in CMOS structure from Figure 3.3 (h), for each input vector.

Table 3.2: Correlation between empirical method and HSPICE for the worst-case and average leakage normalized values of different CMOS structures from Figure 3.2 and Figure 3.3

	Worst Case Leakage		Average Leakage	
	Hspice Simulation	Proposed Method	Hspice Simulation	Proposed Method
3.2 (a)	2.29	1.98	0.75	0.74
3.2 (b)	3.58	3.00	0.76	0.70
3.2 (c)	2.38	2.00	1.19	1.10
3.2 (d)	1.65	2.00	1.18	1.10
3.3 (a)	3.05	2.64	0.55	0.56
3.3 (b)	2.29	2.00	0.97	0.95
3.3 (c)	2.41	3.00	1.19	1.10
3.3 (d)	3.57	3.00	0.98	0.92
3.3 (e)	2.38	2.00	1.36	1.34
3.3 (f)	4.76	4.00	0.54	0.52
3.3 (g)	2.38	2.00	1.21	1.11
3.3 (h)	2.38	2.00	1.38	1.36
3.3 (i)	1.88	2.00	1.18	1.13
3.3 (j)	2.84	3.00	1.19	1.11
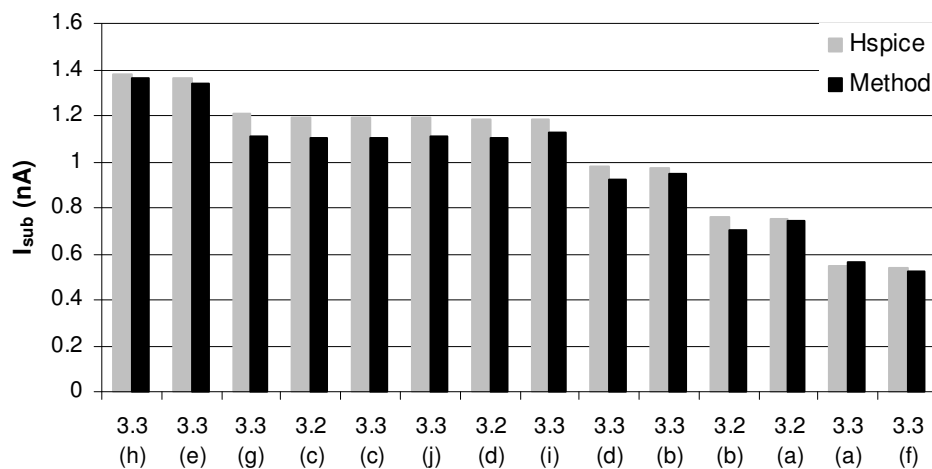
Figure 3.5: Average subthreshold leakage current in the different CMOS structures from Figure 3.2 and Figure 3.3.

3.2 Subthreshold Leakage Models

Several subthreshold leakage current models have been presented during the last decade. This work evaluates three of these models, presented by (NARENDRA, 2006), (GU, 1996) and (ROY, 2000). A brief analysis for a two transistor stack is presented for each model, as well as their advantages and limitations are reported. Based on this initial analysis, the model presented by (ROY, 2000) was selected to be reviewed and improved in order to evaluate CMOS complex gates.

3.2.1 (NARENDRA, 2006) Model

Subthreshold leakage current model reported by (NARENDRA, 2006) is given by:

$$I_S = W.I_1.10^{\frac{-1}{n}\left[\Delta V_{gs} + \eta \Delta V_{ds} + \gamma \Delta V_{bs}\right]} \tag{3.2}$$

where W is the effective transistor width, I_1 is the leakage of a single transistor of unit width in an OFF state with $V_{gs} = V_{bs} = 0\ V$ and $V_{ds} = V_{dd}$. ΔV_{gs}, ΔV_{bs} and ΔV_{ds} are respectively the gate-drive, body bias and drain-to-source voltage reduced based on above mentioned conditions. n is the subthreshold swing, η is the drain-induced barrier lowering and γ is the body effect coefficient. The above equation assumes that the resulting $V_{ds} > 3kT/q$.

In a two-transistor stack, as shown in Figure 3.6, the subthreshold leakage currents passing through the transistors is given by

$$I_{SM1} = W_1.I_1.10^{\frac{-V_2(1+\eta+\gamma)}{n}} \tag{3.3}$$

$$I_{SM2} = W_2.I_1.10^{\frac{-\eta(V_{dd}-V_2)}{n}} \tag{3.4}$$
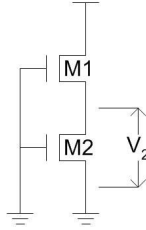
Figure 3.6: Two-transistor stack.

This two-transistor stack reaches its steady state condition when the leakage current in the upper and lower transistors are equal. Under this condition, the voltage V_2 can be expressed as

$$V_2 = \frac{\eta V_{dd} + n\log\left(\frac{W_1}{W_2}\right)}{1 + 2\eta + \gamma} \tag{3.5}$$

In order to confirm the model accuracy, HSPICE simulations were performed in 0.13μm CMOS process and compared to the theoretical results. The model parameters, $n=1.45$, $\eta=0.078$ and $\gamma=0.17$ are extracted by simulation by using $V_{dd} = 1.2V$,

temperature=100°C, minimum transistor width and channel length. Table 3.3 shows the numerical results and proves the accuracy of the theoretical model.

Although the model presents satisfactory accuracy, it is essential to point out that the model assumes the intermediate node voltage to be greater than *3kT/q*. This assumption invalidates the model when it is applied to three or more transistor stacks because occasionally the intermediate node voltage is not greater than *3kT/q*. This will be showed later.

3.2.2 (GU, 1996) Model

Subthreshold leakage current model reported by (GU, 1996) is given by:

$$I_S = I_0 W e^{\frac{V_{gs}-V_{th}}{nV_T}} \left[1 - e^{\frac{-V_{ds}}{V_T}} \right] \tag{3.6}$$

where $I_0 = \frac{\mu_0 C_{ox} V_T^2 e^{1.8}}{L}$, $V_T = \frac{kT}{q}$, W is the effective transistor width, L is the effective channel length, *n* is the subthreshold slope coefficient, C_{ox} is the gate oxide capacitance, μ_0 is the mobility, and V_{th} is the threshold voltage expressed by equation (3.7) .

$$V_{th} = V_{FB} + \Phi_S + K_1\sqrt{\Phi_S - V_{bs}} - K_2(\Phi_S - V_{bs}) - \eta V_{ds} \tag{3.7}$$

where V_{FB} is the flat-band voltage, Φ_S is the surface-inversion potential, K_1 and K_2 together model the body effect phenomenon, and η is the drain-induced barrier lowering coefficient.

In a two-transistor stack the subthreshold leakage currents passing through the transistors is given by

$$I_{SM1} = I_0 W_1 e^{\frac{V_{FB}+\phi_S-K_1\sqrt{\phi_S}+K_2\phi_S}{nV_T}} e^{\frac{\eta V_{dd}-(1+k_2-\eta)V_2-K_1\sqrt{V_2}}{nV_T}} \tag{3.8}$$

$$I_{SM2} = I_0 W_2 e^{\frac{V_{FB}+\phi_S-K_1\sqrt{\phi_S}+K_2\phi_S}{nV_T}} e^{\frac{\eta V_2}{nV_T}} \tag{3.9}$$

In order to simplify the analysis, the voltages $V_{ds1} = V_{dd} - V_2$ and $V_{ds2} = V_2$ were considered greater than *3kT/q* in equations (3.8) and (3.9). This assumption is true as verified in Table 3.3. The intermediate node voltage, V_2, can be derived by equating the two currents.

$$V_2 = \frac{\eta V_{dd} + nV_T \ln\left(\frac{W_1}{W_2}\right)}{1 + \frac{K_1}{2\sqrt{\Phi_S}} - K_2 + 2\eta} \tag{3.10}$$

This theoretical model was compared to HSPICE simulation to confirm the model accuracy. The model parameters $K_1 = 0.7$, $K_2 = 0.15$ and $\Phi_S = 0.9$ were extracted from

38

transistor model while *n = 1.45* and $\eta = 0.078$ were achieved through electrical simulation. Model presents good accuracy and the results are showed in Table 3.3.

Despite used in previous example, this model does not have the restriction mentioned in (NARENDRA, 2006) model and can evaluate gates where the intermediate voltage is smaller than *3kT/q*. However, the V_{th} definition is based on physical parameters which are not common to circuit designers.

3.2.3 (ROY, 2000) Model

Subthreshold leakage current model reported by (ROY, 2000) is given by:

$$I_S = I_0 W e^{\frac{V_{gs}-(V_{t0}-\eta V_{ds}-\gamma V_{bs})}{nV_T}}\left[1 - e^{\frac{-V_{ds}}{V_T}}\right] \tag{3.11}$$

where $I_0 = \frac{\mu_0 C_{ox} V_T^2 e^{1.8}}{L}$, $V_T = \frac{kT}{q}$, V_{t0} is the zero-bias threshold voltage, *W* is the effective transistor width, *L* is the effective channel length, *n* is the subthreshold slope coefficient, C_{ox} is the gate oxide capacitance, μ_0 is the mobility, η is the drain-induced barrier lowering coefficient and γ is the linearized body effect coefficient.

In a two-transistor stack the subthreshold leakage currents passing through the transistors is given by

$$I_{SM1} = I_0 W_1 e^{\frac{-V_2-[V_{t0}-\eta(V_{dd}-V_2)+\gamma V_2]}{nV_T}} \tag{3.12}$$

$$I_{SM2} = I_0 W_2 e^{\frac{-V_{t0}+\eta V_2}{nV_T}} \tag{3.13}$$

and the intermediate node voltage V_2 is expressed as

$$V_2 = \frac{\eta V_{dd} + nV_T \ln\left(\frac{W_1}{W_2}\right)}{1+2\eta+\gamma} \tag{3.14}$$

Table 3.3 lists the theoretical model estimation values for the previous example. Its accuracy is verified when compared to HSPICE simulation. The model parameters *n=1.45*, $\eta=0.078$, and $\gamma= 0.17$ were extracted from simulation.

Table 3.3: Proposed models accuracy for two stacked transistors

	HSPICE Simulation	(NARENDRA, 2006) model	(GU, 1996) Model	(ROY, 2000) Model
V_2 *(mV)*	*69.99*	*70.59*	*68.08*	*70.59*
I_S *(nA)*	*1.26*	*1.21*	*1.25*	*1.26*

In spite of the three previously presented models having good accuracy compared to HSPICE simulation, there are several restrictions. The model presented by (NARENDRA, 2006) cannot be used in three stacked transistors or in any gate where

V_{ds} cannot be considered greater than *3kT/q*. All those models do not present a solution for complex gates as the cell example in Figure 3.2. Additionally, none of the models consider the effect of ON-transistor in OFF-networks.

The model developed in this work evaluates all restrictions previously mentioned. The model presented by (ROY, 2000) is used as a reference because it shows the most familiar equation to circuit designers. In the next session, a detailed and complete model to CMOS complex gates is presented.

3.3 Modeling Subthreshold Leakage in CMOS Logic Gates

Standard CMOS logic gates are composed of series-parallel transistor networks. As mentioned previously, the total leakage dissipation results from the sum of the current in each branch of off-transistors between the supply voltage and ground node. To present the proposed method, the off-network illustrated in Figure 3.7 can be considered as the entire NMOS pull-down arrangement, or a branch from a more complex CMOS gate. The same analysis is applicable to a PMOS pull-up tree.

From the BSIM MOS transistor model (SHEU, 1987), the subthreshold current for a MOSFET device can be modeled as

$$I_S = I_0 W e^{\frac{V_{gs} - (V_{t0} - \eta V_{ds} - \gamma V_{bs})}{nV_T}} \left[1 - e^{\frac{-V_{ds}}{V_T}} \right] \qquad (3.15)$$

where $I_0 = \dfrac{\mu_0 C_{ox} V_T^2 e^{1.8}}{L}$ and $V_T = \dfrac{kT}{q}$. V_{gs}, V_{ds} and V_{bs} are the gate, drain and bulk voltage of the transistor, respectively. V_{t0} is the zero bias threshold voltage. W and L are the effective transistor width and length, respectively. γ is the body effect coefficient and η is the DIBL coefficient. C_{ox} is the gate oxide capacitance, μ_0 is the mobility and n is the subthreshold slope coefficient.
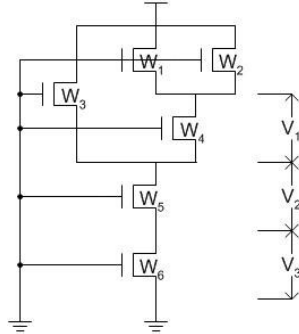

Figure 3.7: NMOS series-parallel network.

In Figure 3.7, the currents passing through the transistors is given by

$$I_{S1} = I_0 W_1 e^{\frac{-(V_1+V_2+V_3) - [V_{t0} - \eta(V_{dd}-V_1-V_2-V_3) + \gamma(V_1+V_2+V_3)]}{nV_T}} \qquad (3.16)$$

$$I_{S2} = I_0 W_2 e^{\frac{-(V_1+V_2+V_3) - [V_{t0} - \eta(V_{dd}-V_1-V_2-V_3) + \gamma(V_1+V_2+V_3)]}{nV_T}} \qquad (3.17)$$

$$I_{S3} = I_0 W_3 e^{\frac{-(V_2+V_3)-[V_{t0}-\eta(V_{dd}-V_2-V_3)+\gamma(V_2+V_3)]}{nV_T}} \tag{3.18}$$

$$I_{S4} = I_0 W_4 e^{\frac{-(V_2+V_3)-[V_{t0}-\eta V_1+\gamma(V_2+V_3)]}{nV_T}} \tag{3.19}$$

$$I_{S5} = I_0 W_5 e^{\frac{-V_3-[V_{t0}-\eta V_2+\gamma V_3]}{nV_T}} \tag{3.20}$$

$$I_{S6} = I_0 W_6 e^{\frac{-V_{t0}+\eta V_3}{nV_T}} \left[1 - e^{\frac{-V_3}{V_T}} \right] \tag{3.21}$$

The following derivation assumes that $V_1 \gg V_T$ and $V_2 \gg V_T$, which was confirmed through Hspice simulation. Thus, the term $[1 - e(-V_{ds}/V_T)]$ in equations (3.16), (3.17), (3.18), (3.19) and (3.20) has been ignored.

First of all, the currents across the first, the second and the fourth transistors are equalized. By solving the equation $I_{S1} + I_{S2} = I_{S4}$, then V_1 is given by

$$V_1 = \frac{\eta V_{dd} + nV_T \ln\left(\frac{W_1+W_2}{W_4}\right)}{1+2\eta+\gamma} \tag{3.22}$$

In next step, V_2 value is obtained by solving the equation $I_{S3} + I_{S4} = I_{S5}$. V_2 is given by

$$V_2 = \frac{nV_T \ln\left(\frac{W_3 e^{\frac{\eta V_{dd}}{nV_T}} + W_4 e^{\frac{\eta V_1}{nV_T}}}{W_5}\right)}{1+\eta+\gamma} \tag{3.23}$$

It is also assumed $V_3 < V_T$. As a consequence, the term $e(-V_3/V_T)$ in (3.21) can be expressed as $(1 - V_3/V_T)$. Introducing this assumption and making $I_{S5} = I_{S6}$, V_3 is then expressed by follow equation, which is accurately solved after some iteration,

$$\frac{1+\eta+\gamma}{n}\left(\frac{V_3}{V_T}\right) + \ln\left(\frac{V_3}{V_T}\right) = \frac{\eta V_2}{nV_T} + \ln\left(\frac{W_5}{W_6}\right) \tag{3.24}$$

3.3.1 General subthreshold leakage model

Based on previous calculation, the model can be generalized as following. The subthreshold current through the top devices, i.e. transistors connected to V_{dd}, can be expressed by equation (3.25):

$$I_{Si} = I_0 W_i e^{\frac{-\sum V_j-[V_{t0}-\eta(V_{dd}-\sum V_j)+\gamma\sum V_j]}{nV_T}} \tag{3.25}$$

Equation (3.25) considers W_i as the evaluated transistor width and V_j as the voltage across every transistor placed below of the top transistor in the stack.

Subthreshold current through the other transistors in the network is expressed by follow equation:

$$I_{Si} = I_0 W_i e^{\frac{-\sum V_j - \left[V_{t0} - \eta V_i + \gamma \sum V_j\right]}{nV_T}} \left[1 - e^{\frac{-V_i}{V_T}}\right] \tag{3.26}$$

The differences between both equations are observed in the η term (DIBL effect) and in the last term, which can be eliminated when $V_i \gg V_T$. Again, V_j represents the voltage across every transistor below the node in the stack, W_i is the evaluated transistor width and V_i is the voltage across the evaluated transistor.

Voltage across each transistor can be evaluated in three different situations, exemplified in the previous example. The subsequent analysis assumes that $V_{dd} \gg V_j$, which drop out all the V_j terms. It also considers the fact that $V_i \gg V_T$, so that the $[1 - e(-V_i/V_T)]$ term can be ignored.

First situation is represented by the voltage V_1 in Figure 3.7. For this condition, V_i is given by

$$V_i = \frac{\eta V_{dd} + nV_T \ln\left(\frac{W_{above}}{W_{below}}\right)}{1 + 2\eta + \gamma} \tag{3.27}$$

In this case, it is possible to associate every transistor connected in that node by series-parallel association. The terms W_{above} and W_{below}, in the equation (3.27), represent the width of the transistors above and below the node V_i, respectively.

The second situation, in turn, is represented by the voltage V_2 in Figure 3.7. In this case, it is not possible to make series-parallel associations between the transistors connected at i-index node. For this state, the voltage V_i is given by

$$V_i = \frac{nV_T \ln\left(\frac{\sum W_{above} e^{\frac{\eta V_{above}}{nV_T}}}{W_{below}}\right)}{1 + \eta + \gamma} \tag{3.28}$$

where V_{above} represents the voltage of the transistors above the node V_i.

Finally, the third situation is represented by the voltage V_3 in previous example. This case only happens at the bottom transistors and the analysis cannot assume $V_i \gg V_T$, so that the term $[1 - e(-V_i/V_T)]$ in equation (3.26) could not be ignored. To simplify the mathematic calculation, the expression $e(-V_i/V_T)$ can be replaced by $(1 - V_i/V_T)$. Then, V_i is obtained by follow equation, where $C = 1 + \eta + \gamma$:

$$\frac{C}{n}\left(\frac{V_i}{V_T}\right) + \ln\left(\frac{V_i}{V_T}\right) = \frac{\eta V_{above}}{nV_T} + \ln\left(\frac{W_{above}}{W_i}\right) + \ln\left(\frac{V_{above}}{V_T}\right) \tag{3.29}$$

3.3.2 Subthreshold leakage in non-series/parallel gates

Standard CMOS gates derived from logic equations are usually composed by series-parallel (SP) device arrangements. When a Wheatstone bridge configuration is presented at transistor level view, as observed in some BDD-based networks (YANG, 2002) (LINDGREN, 2001) (SHELAR, 2005), a non-series-parallel topology is identified, as depicted in Figure 3.8.

Proposed model, discussed above for series-parallel networks, can be used to calculate the voltage across each single transistor and estimate accurately the leakage current. When the model is applied in non-series-parallel (non-SP) configuration, sometimes is somewhat difficult to calculate the voltage across determined transistor, as occur in Figure 3.8 (b) for the transistor controlled by input "c". In this case, the transistor receiving signal "c" must be ignored until the voltage at one of its terminals is evaluated. For evaluating the other terminal, such device is then included. Similar procedure is suitable for any kind of non-SP networks.
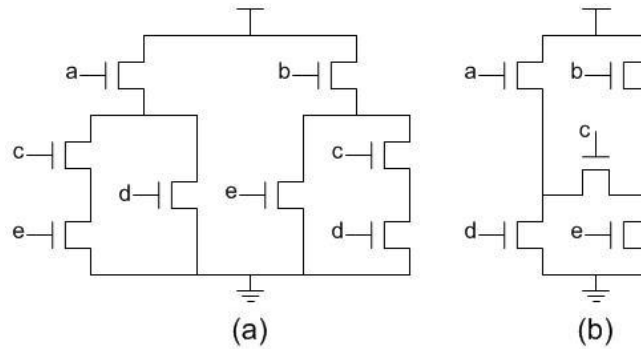

Figure 3.8: SP (a) and non-SP (b) transistor arrangements of the same logic function.

3.3.3 Influence of on-transistors in off-networks

Previous analysis considers only networks composed exclusively by transistors that are turned off. Usually, in the most cases, the transistors that are turned on could be treated as ideal short-circuits, since the drop voltage across such devices is some orders of magnitude smaller than the drop voltage across the off-transistors.

However, in the case of NMOS transistors switched on and connected to power supply V_{dd}, the drop voltage across them should be taken into account as illustrated in Figure 3.9. In the leakage current analysis, this voltage drop is really important when the transistor stack presents only one off-device at the bottom of the stack – Figure 3.9 (a) and (b). In stacks with more than one off-transistor in series configuration the on-devices could be considered as zero drop voltage short-circuit without a significant impact in the result accuracy, as depicted in Figure 3.9 (c) and (d).

Similar analysis is valid for PMOS transistors in off-networks when they are connected to the ground reference.

In the proposed model, the drop voltage across the transistor that is turned on is referenced V_{drop} and, to be consistent, the term $V_{dd} - \sum V_j$ in equation (3.25) must be replaced by $V_{dd} - V_{drop} - \sum V_j$ for all cases, including when the off stack has more than one transistor.
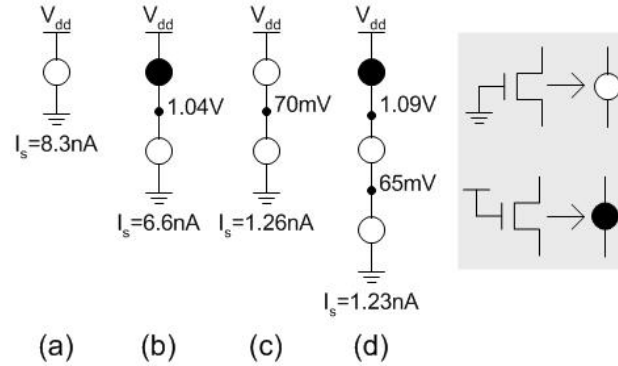
Figure 3.9: Influence of on-transistor in off-stack leakage current.

3.4 Experimental Results

In order to evaluate the model and validate this work, the results obtained from the proposed model were compared to Hspice simulation results, considering commercial 130nm CMOS process parameters, where subthreshold currents represents the main leakage mechanism, and operating temperature at 100°C. Table 3.4 presents the parameters used in the analytical modeling. In a first moment, transistors with equal sizing were applied to simplify the analysis, although the device size is a parameter in the model.

Table 3.4: Parameters used in the analytical model

Parameters	Values
V_{dd} (V)	1.2
V_{drop} (V)	0.14
I_0 (mA)	20.56
W (μm)	0.4
η	0.078
γ	0.17
n	1.45

Leakage current was calculated and correlated with Hspice results for several pull-down NMOS off-networks, depicted in Figure 3.10. The results presented in Table 3.5 show a good agreement between the analytical model and the simulation data, showing an absolute average error less than 10%. It is interesting to note that the static current in networks (h), (i), (j) and (k) from Figure 3.10, not treated by previous models, are accurately predicted. The main difference is observed for structures (d), (f) and (g), when three off-transistors are placed in series arrangement. This difference appears when the model assumes $V_i < V_T$ and the term $e(-V_i/V_T)$ in equation (3.26) is replaced by $(1 - V_i/V_T)$.
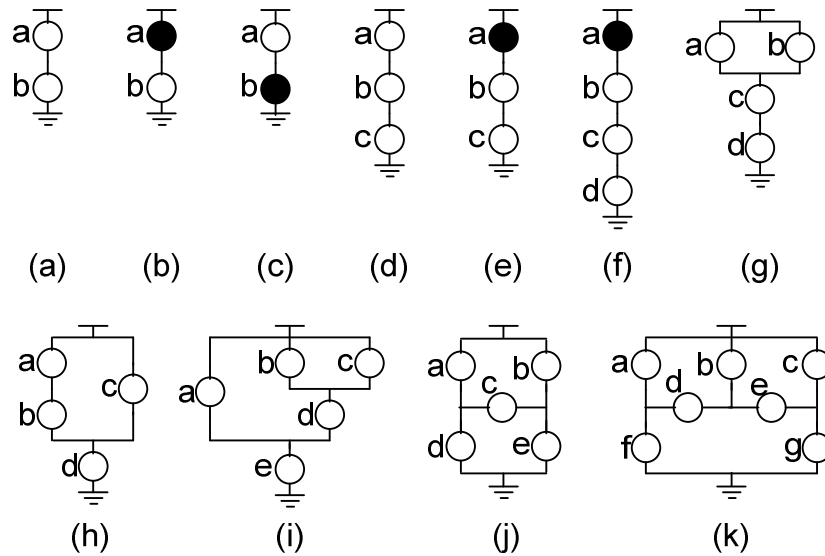
44

Figure 3.10: Pull-down NMOS networks.

Table 3.5: Subthreshold leakage current related to off-networks depicted in Figure 3.10

Network	Hspice results (nA)	Proposed Model (nA)	Diff(%)
(a)	1.26	1.26	-
(b)	6.58	6.60	0.3
(c)	8.34	8.34	-
(d)	0.69	0.75	8.7
(e)	1.23	1.24	0.8
(f)	0.68	0.74	8.8
(g)	0.72	0.77	6.9
(h)	1.29	1.28	0.8
(i)	1.29	1.28	0.8
(j)	2.52	2.53	0.4
(k)	2.56	2.54	0.8

Table 3.6, 3.7 and 3.8 correspond to the input leakage dependence related to NMOS trees in Figure 3.10 (h), (i) and (j), respectively. In some cases, different input vectors result in equivalent off-device arrangements. For that, the Hspice values are presented for minimum and maximum values obtained by applying the set of equivalent input states.

Moreover, the previous model presented in (ROY, 2000) was also calculated for such logic states. Note that, the first input vector in both cases, which represents the entire network composed by off-switches, is not treated by the model presented by (ROY, 2000).

When other input vector is applied, it results in a purely series-parallel off-network. Since both methods evaluate this kind of arrangements, different values are obtained when on-transistors are considered in the off-networks, providing thus more correlation with the electrical simulation results.

Table 3.6: Input dependence leakage estimation in logic network (h) from Figure 3.10 (pull-down NMOS tree)

Input-state (abcd)	Hspice results * (nA)	Proposed Model (nA)	Previous Model (ROY, 2000) (nA)
0000	1.29	1.28	-
0001	9.60	9.60	9.60
0010 [a]	6.30/6.70	6.60	8.34
0100	1.37	1.31	1.31
0101	16.67	16.69	16.69
1000	1.36	1.30	1.31
1001	14.91	14.94	16.69

* The HSPICE value is given for min./max. currents related to equivalent vectors.
[a] Equivalent vectors – 0110, 1010, 1100, 1110.

Table 3.7: Input dependence leakage estimation in logic network (i) from Figure 3.10 (pull-down NMOS tree).

Input-state (abcde)	Hspice results * (nA)	Proposed Model (nA)	Previous Model (ROY, 2000) (nA)
00000	1.29	1.28	-
00001	9.71	9.65	9.65
00010	1.43	1.34	1.34
00011	25.00	25.02	25.02
00100 [a]	1.36/1.37	1.30	1.31
00101 [b]	14.91/15.14	14.94	16.69
00110 [c]	6.30/6.73	6.60	8.34

* The HSPICE value is given for min./max. currents related to equivalent vectors.
[a] Equivalent vectors – 01000, 01100.
[b] Equivalent vectors – 01001, 01101.
[c] Equivalent vectors – 01010, 01110, 10000, 10010, 10100, 10110, 11000, 11010, 11100, 11110.

Table 3.8: Input dependence leakage estimation in logic network (j) from Figure 3.10 (pull-down NMOS tree).

Input-state (abcde)	Hspice results * (nA)	Proposed Model (nA)	Previous Model (ROY, 2000) (nA)
00000	2.52	2.53	-
00001[a]	10.54/10.54	10.76	10.76
00011[b]	16.68/16.68	16.68	16.68
00100	2.52	2.52	2.52
01000	7.90	7.90	7.91
01010[c]	21.05/21.05	21.54	24.02
01100[d]	12.55/13.15	13.20	16.68
10000	7.90	7.90	9.65

* The HSPICE value is given for min./max. currents related to equivalent vectors.
[a] Equivalent vectors – 00010.
[b] Equivalent vectors – 00101, 00110, 00111.
[c] Equivalent vectors – 10001.
[d] Equivalent vectors – 10100, 11000, 11100.

Figure 3.11 shows the subthreshold average leakage current related to the NMOS networks illustrated in Figure. 3.10 (h), (i), (j). As discussed before, the previous model from (ROY, 2000) cannot estimate the subthreshold current for the first input vector (all inputs at ground) for these pull-down networks. These values are not considered in the average static current calculation. Unlike the previous model, the proposed one presents results close to Hspice simulations data. The main reason for that is the influence of on-transistors in off-networks, neglected in previous works.
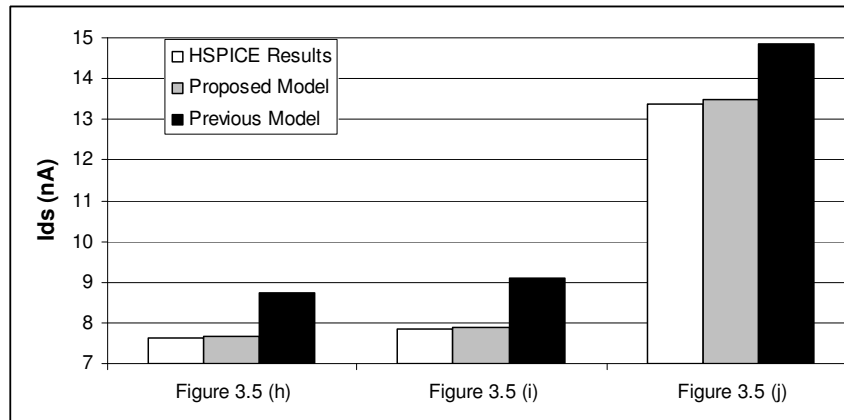

Figure 3.11: Average subthreshold leakage for Figure 3.10 (h), (i) and (j) pull-down networks.

In terms of combinational circuit static dissipation analysis, the technology mapping task divides the entire circuit in multiple logic gates. Thus, they can be treated separately for the leakage estimation, since the input state of each cell is known according to the primary input vector of the circuit. A complex CMOS logic gate,

whose transistors sizing were determined by considering the Logical Effort method (SUTHERLAND, 1999), is depicted in Figure 3.12. Table 3.9 presents the comparison between electrical simulation data and the proposed model calculation.
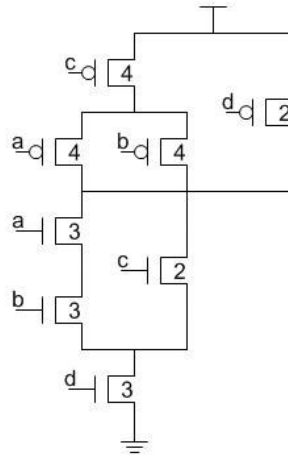

Figure 3.12: CMOS complex gate, with different transistor sizing, according to Logic Effort (SUTHERLAND, 1999).

Table 3.9: Subthreshold leakage current related to the CMOS complex gate depicted in Figure 3.9

Input-state (abcd)	Hspice results (nA)	Proposed Model (nA)	Diff(%)
0000	4.01	4.13	3.0
0001	20.67	20.68	0.0
0010	19.93	19.99	0.3
0011	44.52	43.27	2.8
0100	4.44	4.29	3.4
0101	42.34	42.37	0.1
0110	19.93	19.99	0.3
0111	43.38	43.27	0.3
1000	4.40	4.26	3.2
1001	36.81	36.50	0.8
1010	19.93	19.99	0.3
1011	43.38	43.27	0.3
1100	19.50	19.99	2.5
1101	96.67	96.92	0.3
1110	20.43	19.99	2.2
1111	20.48	20.21	1.3

Finally, the proposed model has been verified to the variation of power supply voltage and operating temperature, as depicted in Figure 3.13 and Figure 3.14, respectively. The influence of temperature variation in the predicted current shows good agreement with Hspice results. On the other hand, the difference between the subthreshold currents obtained from electrical simulation and analytical modeling to voltage variation can be justified by eventual inaccuracy in the parameter extraction listed in Table 3.4. Figure 3.15 shows the leakage current analysis in respect to the threshold voltage variation, validating the proposed method for this factor, critical in the most advanced CMOS processes.
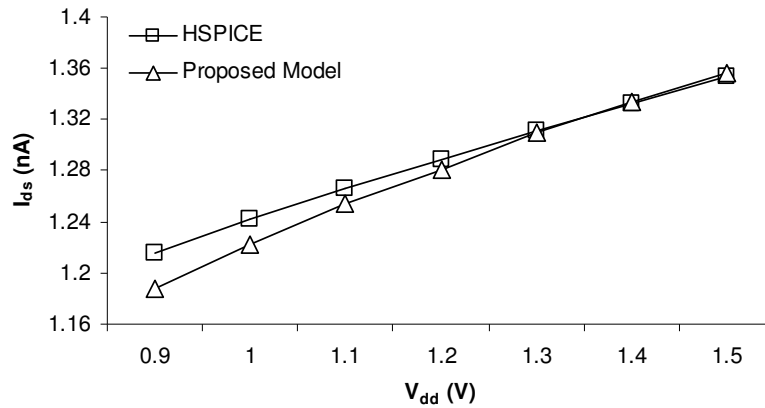
Figure 3.13: Variation of subthreshold leakage current in terms of power supply voltage variation.
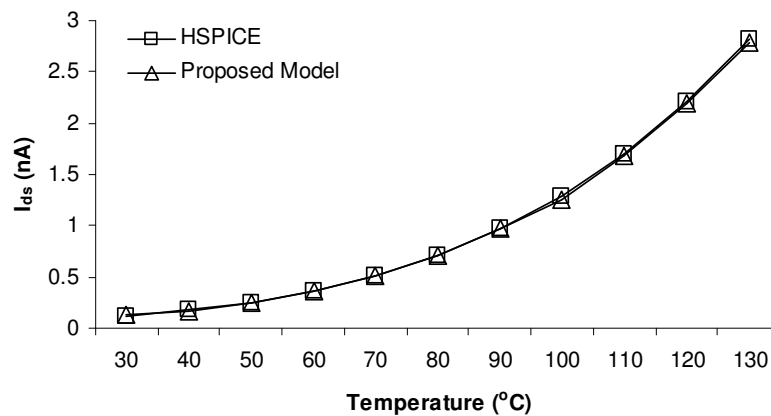

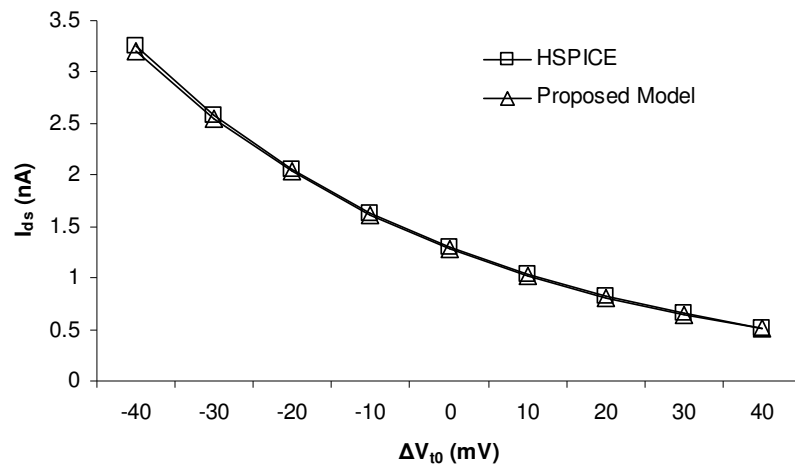Figure 3.14: Variation of subthreshold leakage current according to the operating temperature variation.

Figure 3.15: Variation of subthreshold leakage current according to the threshold voltage variation.

4 MODEL INCLUDING GATE OXIDE LEAKAGE

The reduction of vertical dimensions has been harder than horizontal ones. An aggressive scaling of gate oxide thickness is required to provide large current drive capability at reduced voltages supplies and to suppress short-channel effects, such as drain induced-barrier lowering. This scaling increases the field across the oxide. The high electric field coupled with the low oxide thickness results in gate tunneling leakage current from the gate to the inverter channel and source/drain overlap region, or from the source/drain overlap region to the gate. These mechanisms are depicted in Figure 4.1 (a) and (b), respectively.

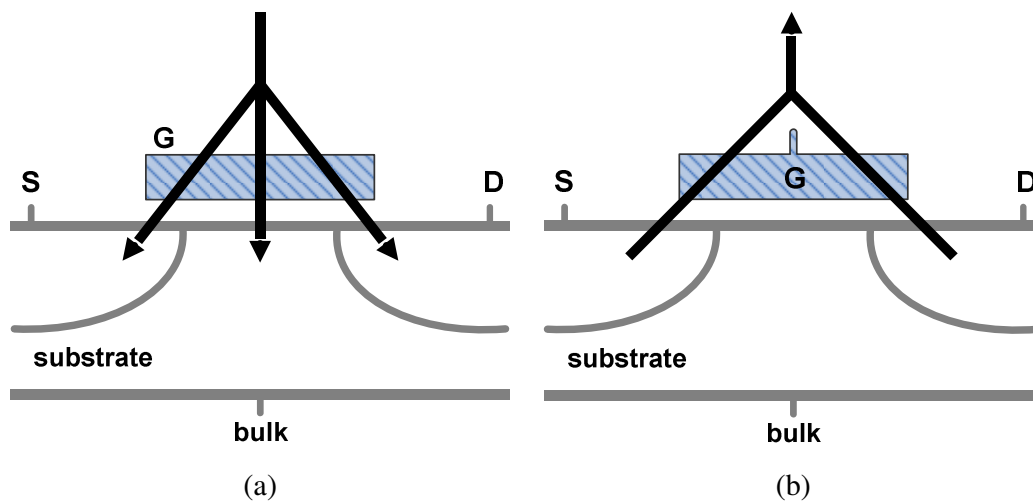(a)   (b)

Figure 4.1: Gate leakage current (a) from gate to channel and source/drain overlap region and (b) from source/drain overlap region to gate

Gate leakage current increases exponentially with decreasing oxide thickness. When the gate oxide thickness reaches 3nm and below, gate tunneling current comes into the order of the subthreshold leakage (YANG, 2005). It also increases exponentially with voltage across gate oxide. Figure 4.2 shows the density of gate leakage current (A/m^2) in a NMOS device versus potential drop across the oxide for several oxide thicknesses.
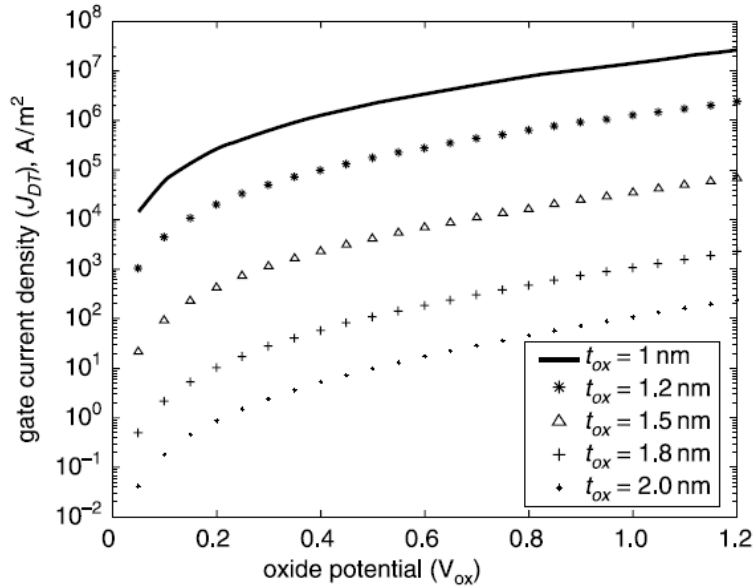
Figure 4.2: Variation of tunneling current density with potential drop across the oxide (AGARWAL, 2005).

This chapter reviews gate leakage behavior and presents models reported in literature. A new gate leakage model to general CMOS network is proposed and can be combined with previous subthreshold model to explore the interaction between both leakage mechanisms, estimating accurately the total leakage current in CMOS circuits. All analysis presented in this chapter use NMOS pull-down network. The same analysis is applicable to PMOS pull-up tree.

The first section explores the gate leakage behavior. Section 4.2 presents and evaluates gate leakage models reported in literature. After this analysis, Section 4.3 describes a detailed and complete gate leakage model to CMOS complex gates. Interactions between gate and subthreshold leakage are explored in Section 4.4. Finally, in order to validate the proposed model and to verify its accuracy, Hspice simulations are compared with model results at the end of the chapter.

4.1 Gate Leakage Behavior

As mentioned previously, gate leakage current is exponentially related with the voltage across gate oxide and the oxide thickness. Ignoring the variability in oxide thickness due to process variation, it is possible to consider only the voltage dependence in the gate leakage behavior analysis.

Subthreshold leakage is evaluated only when transistor is turned OFF. Gate leakage, on the other hand, occurs in both cases, when transistors are turned ON and OFF. Gate leakage current is independently in both, turned ON or OFF, transistor states. When transistor is turned OFF the current flows by the overlap source and drain regions. In the case where the transistor is turned ON, the current uses the overlap source/drain regions and the transistor channel. For these reasons, gate leakage is usually higher in such condition.

Considering previous statement, the easy method to investigate gate leakage current is evaluating the transistor bias conditions. Figure 4.3 presents all eight possible bias conditions for a NMOS transistor. Figure 4.3 (f) and (g) can be ignored because they represent transient states and does not occur in steady state. In Figure 4.3 (a) and (h) gate leakage is not present because all terminals have the same potential. In the other conditions gate leakage has to be computed.
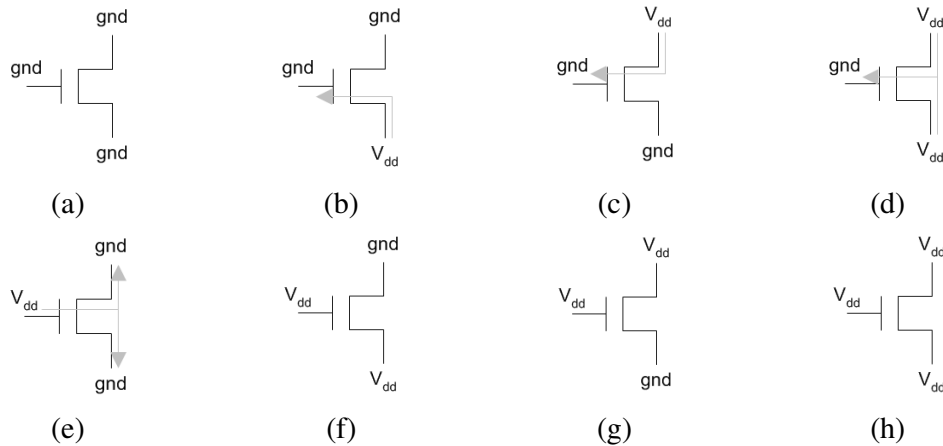

(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 4.3: Possible bias condition for NMOS transistors in CMOS logic circuits.

The effect of transistor stacking on circuit topology was first proposed and analyzed for subthreshold leakage, as discussed in Chapter 3. As mentioned, in CMOS technologies where subthreshold leakage is dominant, a stack of OFF transistor leaks less than a single device. The same behavior is found in process when gate leakage is becoming dominant, and the transistor stack is composed by purely OFF devices. However, ON transistors in the middle of stack introduce the gate leakage component (ignored in previous subthreshold leakage analysis). This new component increases the total leakage dissipation and changes some leakage statements as the leakage input vector dependence – discussed in Section 2.2. Figure 4.4 shows a CMOS gate with both subthreshold and gate leakage currents for a two specific input vectors.
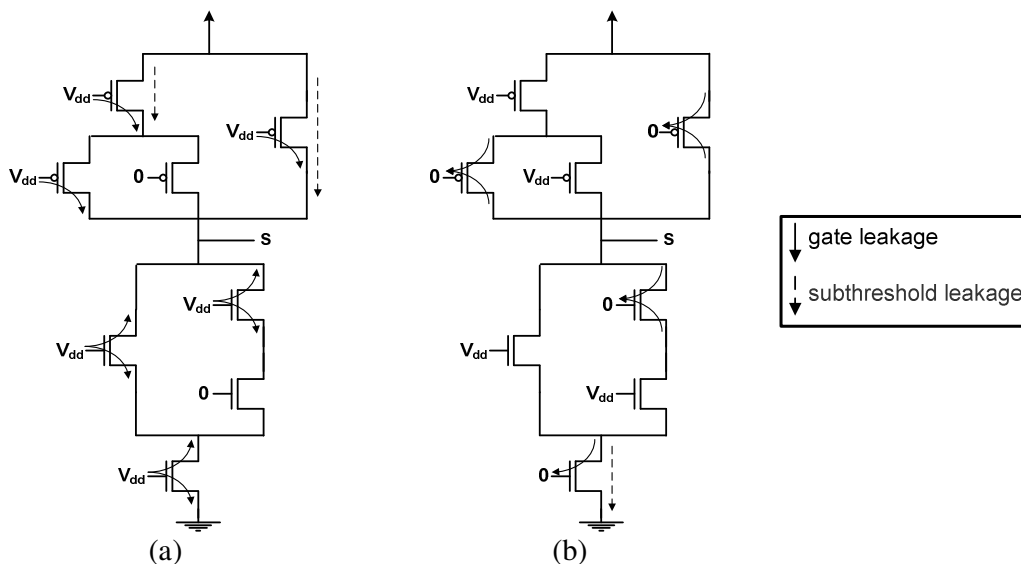

(a) (b)

Figure 4.4: Subthreshold and gate leakage current in a CMOS gate for tow specific input vectors.

4.2 Previous Gate Leakage Models

Recently, comprehensive analysis of gate leakage was carried out by several authors. The work presented by (LEE, 2003) estimates gate current based on transistor bias conditions, depicted in Figure 4.3. This is a good approach to evaluate gate leakage considering the relation with gate leakage behavior. However, the reverse tunneling current from source/drain to gate – Figure 4.3 (b) (c) (d) – is ignored and the proposed interaction with subthreshold leakage does not cover fully possible situations, as the simple three transistors stacks with the second transistor turned ON.

In the other hand, (MUKHOPADHYAY, 2003) presents a complete, but complex, analytical leakage estimation method, including Band-to-Band-Tunneling leakage. The methodology proposed in such work is too complex to provide fast leakage estimation. Another total leakage estimation method is presented by (XU, 2004). It is based on "table-lookup", which provides a fast estimation. However, the accuracy is restricted to common CMOS gates.

The relation between gate leakage and transistor bias condition is also explored by (RAO, 2003). It presents a fast technique for gate leakage estimation. It does not treat the subthreshold leakage and the interaction between both mechanisms. It is assumed that the internal nodes attain full logic levels (i.e., they are either at V_{DD} or V_{SS}) and in a transistor stack the entire voltage drops across the uppermost OFF device. These assumptions make the technique really fast, but the accuracy for complex gates is compromised.

Another work that uses the transistor bias condition to evaluate gate leakage is presented by (YANG, 2005). It explores the interaction between gate and subthreshold leakage mechanisms. The subthreshold leakage model used on that work is proposed by (GU, 1996) and the gate leakage is a simplification of BSIM4 gate leakage model (CAO, 2000). The assumptions proposed on that work are based on a technology process where gate leakage is at least two orders of magnitude superior to subthreshold leakage. This assumption simplify the analysis of interaction between both models in several situations, as the same described before, the simple three transistors stacks with the second transistor turned ON.

The models discussed above, in exception (MUKHOPADHYAY, 2003), present some assumptions that compromise their use and accuracy in general transistor networks. The method proposed by (MUKHOPADHYAY, 2003) can be used in such arrangements, but it is too complex for a fast estimation. The follow method avoids assumptions that can compromise its applicability in complex gates and, at the same time, it provides faster results than previous one.

4.3 Gate Leakage Model

High electric field coupled with reduced oxide thickness results in tunneling of electrons (holes) from the gate to the channel and source/drain overlap region, or from the source/drain overlap region to the gate, resulting in the gate oxide tunneling current. The tunneling current density is expressed as (ROY, 2003):

$$J_{TunnelingCurrentDensity} = W.L.A.\left(\frac{V_{ox}}{T_{ox}}\right)^2.e^{\left(\frac{-B.\left(1-\left(1-\frac{V_{ox}}{\phi_{ox}}\right)^{3/2}\right)}{\frac{V_{ox}}{T_{ox}}}\right)} \tag{4.1}$$

where W and L are the effective transistor width and length, respectively, $A = \frac{q^3}{16\pi^2\hbar\phi_{ox}}$

and $B = \frac{4\sqrt{2m*}\phi_{ox}^{3/2}}{3\hbar q}$. V_{ox} is the potential drop across the gate oxide, T_{ox} is the oxide thickness, ϕ_{ox} is the barrier height of the tunneling electron, $m*$ is the effective mass of an electron in the conduction band of silicon, q is the electronic charge and $\hbar$ is the reduced Plank constant.

By considering equation (4.1), it is easy to conclude that tunneling current increases exponentially with oxide thickness scaling and raising potential drop across the gate oxide. However, previous equation is complex and a simpler model to capture the dependence between gate leakage current and gate voltage is desirable for fast estimations.

Equation (4.2) explores this dependence and provides a good accuracy as shown in Figure 4.5. The gate oxide thickness dependence is suppressed considering estimations to only one technology node. In the case of estimation for different CMOS process, the oxide thickness influence has to be included in I_{gate0} term.

$$I_{gate} = I_{gate\_0}.W.e^{\frac{-K}{|V_{ox}|}} \tag{4.2}$$

where I_{gate_0} represents the gate leakage current for $V_{ox} = V_{dd}$. V_{ox} is the potential drop across the gate oxide, K is the calibration constant, extracted by simulation based on difference between gate leakage currents to $V_{ox} = V_{dd}$ and $V_{ox} = 0.9*V_{dd}$, and W is the transistor width.
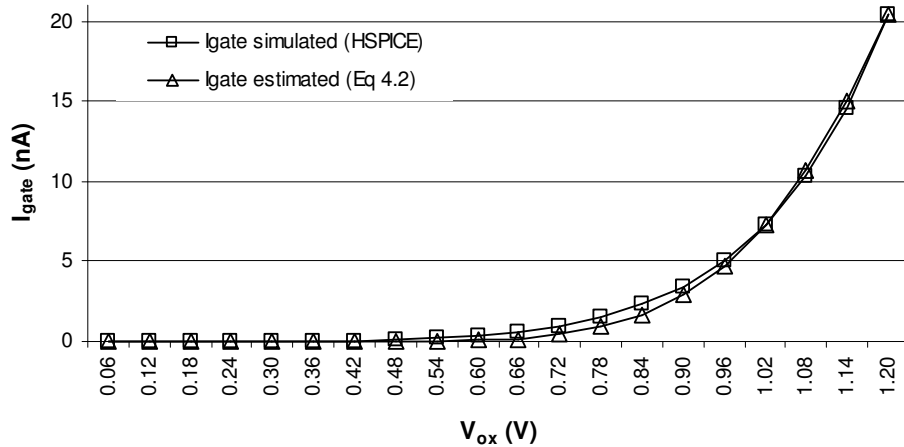

Figure 4.5: Gate leakage proposed model accuracy compared to HSPICE simulation.

All eight possible bias conditions seen for a NMOS device at steady state have been already showed in Figure 4.3. The same analysis is easily extended to PMOS devices. Figure 4.3 (f) and (g) represent transient states and can be ignored in this analysis. Figure 4.3 (a) and (h) do not present gate leakage because all terminals with the same potential. The others four cases, Figure 4.3 (b), (c), (d), (e) have to be analyzed in leakage current estimation. There is a relationship between gate leakage currents in Figure 4.3 (b), (c), (d), expressed by equations (4.3) and (4.4):

$$I_{gate\_(b)} = I_{gate\_(c)} \qquad (4.3)$$

$$I_{gate\_(d)} = I_{gate\_(b)} + I_{gate\_(c)} \qquad (4.4)$$

Considering previous analysis, the proposed gate leakage model needs evaluate the leakage current in two cases. The first one, named I_{gate_ON}, is illustrated in Figure 4.3 (e). It occurs when the transistor is turned on. The second one, named I_{gate_OFF}, is illustrated in Figure 4.3 (d) and occurs when the transistor is turned off. The cases illustrated in Figure 4.3 (b) and (c) are obtained from I_{gate_OFF} based on relations presented in equation (4.3) and (4.4), respectively. Equation (4.5) and (4.6) present both gate leakage case described above.

$$I_{gate\_ON} = I_{gate\_ON\_0}.W.e^{-\frac{K}{|V_{ox}|}} \qquad (4.5)$$

$$I_{gate\_OFF} = I_{gate\_OFF\_0}.W.e^{-\frac{K}{|V_{ox}|}} \qquad (4.6)$$

4.4 Subthreshold and Gate Oxide Leakage Iteration

Chapter 3 has presented an accurate subthreshold leakage model to CMOS complex gates. In the same way, previous section has presented gate leakage model. This section will explore the interaction between both leakage mechanisms. This analysis will provide an accurate leakage model to be used in CMOS process up to 50nm where these two mechanisms are dominants, as depicted in Figure 4.6.
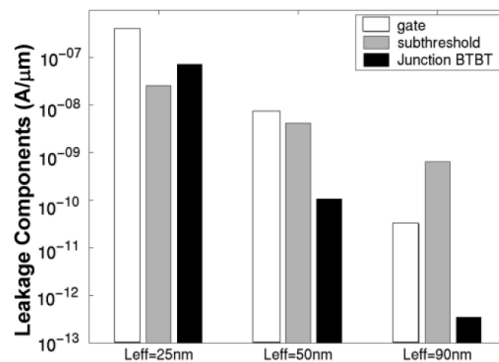

Figure 4.6: Contribution of different leakage components in NMOS devices at different technology generation (MUKHOPADHYAY, 2005).

Before evaluating the interaction between both mechanisms, it is important to provide a brief review of both mechanisms. Subthreshold leakage current only occurs when the transistor is turned OFF. This mechanism is modeled in Chapter 3 by equation (3.15). That equation is rewrite below.

$$I_S = I_0 W e^{\frac{V_{gs}-(V_{t0}-\eta V_{ds}-\gamma V_{bs})}{n V_T}}\left[1-e^{\frac{-V_{ds}}{V_T}}\right]$$

(4.7)

Gate leakage current occurs when devices are turned ON and turned OFF, as introduced in previous section. Equations (4.5) and (4.6) are used to describe both situations, respectively. To represent all three possible OFF states, equation (4.6) have to be combined with equations (4.3) and (4.4), as discussed before.

The simplest transistor arrangement that exemplify the interaction between both leakage mechanisms is a three transistor stack with the middle transistor turned ON. It is depicted in Figure 4.6. Most of previous models ignore the interaction between both mechanisms (LEE, 2003) or present a solution for specific technology process where one (gate) leakage mechanism is dominant and the other one (subthreshold) can be ignored (YANG, 2005). The follow analysis will evaluate the leakage mechanisms interaction for the arrangement illustrated in Figure 4.7. Before evaluates this interaction, it is important to identify the leakage mechanisms in each transistor. Figure 4.8 illustrates subthreshold and gate leakage currents in each transistor.
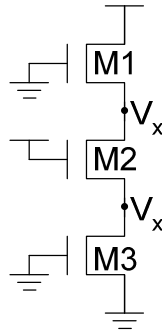

Figure 4.7: Three stack transistor arrangement.

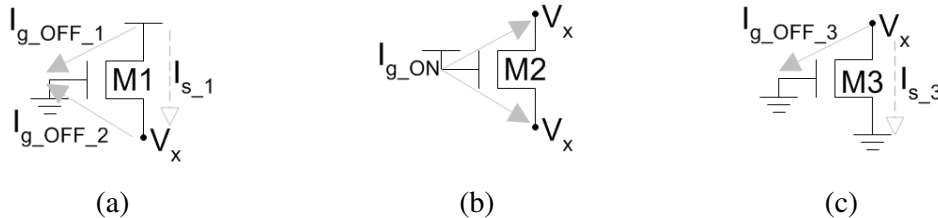(a)   (b)   (c)

Figure 4.8: Leakage currents in each transistor of arrangement depicted in Figure 4.7.

To evaluate accurately the gate and subthreshold current is necessary find the intermediate voltages. In this example there is only one unknown intermediate voltage, V_x, because transistor M2 is turned ON.

Intermediate voltage V_x can be calculated considering that the sum of currents flowing into V_x node is equal to the sum of currents leaving V_x node. Figure 4.9 illustrate currents in V_x node and equation (4.8) represents this behavior.
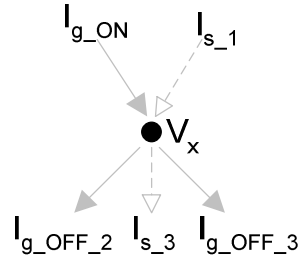
Figure 4.9: Currents in V_x node.

$$I_{S\_1} + I_{g\_ON} = I_{S\_3} + I_{g\_OFF\_2} + I_{g\_OFF\_3} \tag{4.8}$$

Equation (4.9) results from expanding terms in equation (4.8). It is accurately solved after some iteration, providing V_x value to calculate the total leakage current.

$$I_{S0}W_1 e^{\frac{-V_x-(V_{t0}-\eta(V_{dd}-V_x)+\gamma W_x)}{nV_T}}\left[1-e^{-\frac{V_{dd}-V_x}{V_T}}\right]+I_{g\_ON\_0}.W_2.e^{-\frac{K}{V_{dd}-V_x}} =$$

$$I_{S0}W_3 e^{\frac{\eta V_x}{nV_T}}\left[1-e^{-\frac{V_x}{V_T}}\right]+I_{g\_OFF\_0}.(W_1+W_3).e^{-\frac{K}{V_x}} \tag{4.9}$$

There are two ways to calculate the total leakage current. The first one is summing all currents flowing from V_{dd} node. The second one is summing all currents flowing into ground node. Proposed model have chosen the second one due the facility in calculate subthreshold current in transistors connected in ground nodes. Considering previous example, the total leakage current is given by:

$$I_{Leakage\_Total} = I_{S\_3} + I_{g\_OFF\_1} + I_{g\_OFF\_2} + I_{g\_OFF\_3} \tag{4.10}$$

This analysis was evaluated for simple NMOS arrangements, showing a good accuracy. It is easily extended to PMOS and CMOS arrangements.

4.5 Experimental Results

In order to validate this method, the results obtained from the proposed model were compared to Hspice simulation results, considering 90nm CMOS process parameters, where subthreshold and gate currents represent the major leakage mechanisms.

The model parameters were extracted by simulation considering operating temperature at 100°C. Table 4.1 presents these parameters used in the analytical model. In a first moment, transistors with equal sizing were applied to simplify the analysis, although the device size is a parameter in the model.

58

Table 4.1: Parameters used in the analytical model

Parameters	Values
V_{dd} (V)	1.2
W (μA)	0.4
Subthreshold Parameters	
V_{drop} (V)	0.25
I_{S_0} (nA)	4.95
η	0.058
γ	0.15
n	1.45
Gate Parameters	
$I_{g\_ON\_0}$(uA)	6.95
$I_{g\_OFF\_0}$(uA)	0.45
K	6.7

Subthreshold and gate leakage currents were calculated and correlated with Hspice results for several pull-down NMOS off-networks, depicted in Figure 4.10. The results presented in Table 4.2 show good agreement between the analytical model and the simulation data, showing an absolute average error less than 5%.
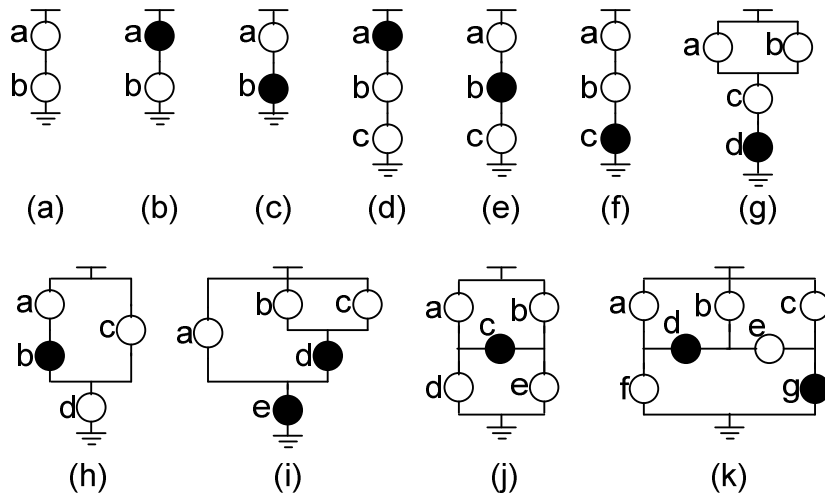
Figure 4.10: Pull-down NMOS networks.

Table 4.2: Total leakage current related to the off-networks depicted in Figure 4.10

Network	Hspice results (nA)	Proposed Model (nA)	Diff (%)
(a)	5.37	5.35	0.37
(b)	15.21	15.26	0.33
(c)	42.12	42.11	0.02
(d)	4.23	4.02	4.96
(e)	7.12	7.27	2.11
(f)	25.73	25.71	0.08
(g)	27.47	27.65	0.66
(h)	8.45	8.42	0.36
(i)	105.83	105.97	0.13
(j)	12.88	13.24	2.80
(k)	55.00	55.35	0.64

Previous results were performed in transistors with equal sizing. A complex CMOS logic gate, whose transistors sizing were determined by Logical Effort method (SUTHERLAND, 1999), is depicted in Figure 4.11. Table 4.3 presents the comparison between electrical simulation data and the proposed model calculation.
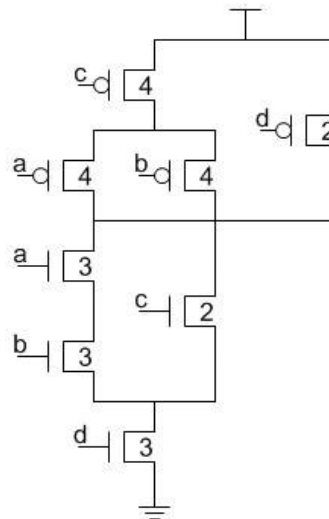
Figure 4.11: CMOS complex gate, with different transistor sizing, according to the Logic Effort (SUTHERLAND, 1999).

Table 4.3: Total leakage current related to the CMOS gate depicted in Figure 4.11

Input-state (abcd)	Hspice results (nA)	Proposed Model (nA)	Diff (%)
0000	18.97	19.35	2.01
0001	122.80	121.85	0.78
0010	50.66	50.35	0.61
0011	156.12	153.96	1.38
0100	24.58	24.87	1.18
0101	234.66	231.91	1.17
0110	50.58	50.22	0.72
0111	217.21	215.50	0.79
1000	17.25	17.18	0.41
1001	152.68	153.91	0.81
1010	48.29	48.95	1.38
1011	217.19	215.60	0.73
1100	48.12	48.79	1.39
1101	293.87	291.73	0.73
1110	46.64	47.43	1.69
1111	255.59	257.82	0.87

5 CONCLUSION

Power dissipation of electronic products has become an important issue with the massive growth in portable computing and wireless communication in the last few years. As power consumption is directly proportional to the square of the power supply voltage, MOS transistor has been scaled to maintain performance at reduced supply voltage. Transistor threshold voltage it also reduced to avoid short channel effect, resulting in a substantial increase in leakage currents when transistor scaling into nanometer dimensions. Standby current becomes a significant portion of the total IC power consumption, being a challenge for circuit designers and a critical factor in the future of low-power microelectronics design. It means the static power dissipation should be considered as soon as possible in the IC design flow. Thus, the main objectives of this research were:

- Review leakage mechanisms and reduction techniques, providing a minimum background to IC designers about leakage currents.

- Develop a leakage estimation method to general transistor networks, reducing restrictions on the previous methods presented in the literature.

Therefore, an analysis over leakage current mechanisms was the first task done in the research work. The information reviewed in this initial study has showed that subthreshold leakage was main leakage mechanism, but with the transistor scaling into sub-100nm sizing, gate leakage has achieved the same order of importance. Another important mechanism, Band-to-Band Tunneling leakage, that should be considered under 25nm CMOS process (MUKHOPADHYAY, 2005), was also reviewed.

Leakage reduction techniques were also reviewed to complete background related to leakage currents. Dual-threshold CMOS, which uses high V_{th} transistors in non critical path to achieve leakage reduction without performance penalties, was the first technique presented. Supply voltage scaling, usually used to reduce active leakage, is also a good alternative to leakage reduction. Leakage reduction techniques that explore staking effect were also explored, as well as techniques that based on power gating and body bias concept.

Fast leakage estimation is important to be used in library free technology mapping. A new subthreshold leakage estimation method based on conductance association was presented to suppress this demand. The method can be applied in series-parallel arrangements. It has been validated considering a 130nm CMOS technology, in which the subthreshold current is the most relevant leakage mechanism. In the case of sub-100nm processes where gate leakage becomes more significant, the present work should be combined with already published techniques which address fast gate leakage current estimation (RAO, 2003).

Accurate leakage prediction is a hard and complex task. This work provides an accurate estimation model to be used in general transistor networks. In a first moment, a subthreshold leakage model is proposed to complex CMOS gate. The presence of on-switches in off-networks, ignored by previous works in literature, is also considered on the proposed model. The new subthreshold leakage method has been validated through electrical simulations, taking into account a 130nm CMOS technology, with good correlation of the results, demonstrating the model accuracy.

Gate leakage becomes significant contributor of standby power in sub-100nm process. To provide a better leakage estimation solution to circuit designers, a model that evaluates iterations between subthreshold and gate leakage was proposed. All characteristics of subthreshold leakage model, presented in Chapter 3, are considered in this model that includes gate oxide leakage and the iteration between both mechanisms. The model has been validated, considering a 90nm CMOS technology, through electrical simulations. The model accuracy is verified by demonstrating the good results correlation.

The results presented in this work have been performed manually. A software is been developed motivated by the good model accuracy. Actually, the simple subthreshold leakage prediction model is already implemented, tested and validated. It has been used as a cost in technology mapping task. The analytical subthreshold leakage model is also already implemented. It is under tests to validate implementation and revalidate the model. Include the analytical gate leakage model in leakage estimation software is the next task to be developed. After finish the software implementation and model validation it is expected to provide a fast and accurate alternative to estimate leakage currents in CMOS circuits.

REFERENCES

AGARWAL, A. et al. Leakage Power Analysis and Reduction: Models, Estimation and Tools. **Proc. IEE**, v.152, n.3, p 353-368, May 2005

AGARWAL, A. et al. Leakage Power Analysis and Reduction for Nanoscale Circuits. **IEEE Micro**, Los Alamitos, v.26, n.2, p. 68-80, Mar. 2006

BOWMAN, K. A.; DUVALL, S. G.; MEINDL, J. D. Impact of Die-to-Die and Within Die Parameter Fluctuations on the Maximum Clock Frequency Distribution fo Gigascale Integration. **IEEE Journal of Solid State Circuits**, New York, v.37, n.2, p. 183-190, Feb. 2002.

BPTM: Berkeley Predictive Technology Model. Available at: <http://www.eas.asu.edu/~ptm/>. Visited on: Mar. 2007.

BURD, T. D. et al. A Dynamic Voltage Scaled Microprocessor System. **IEEE Journal of Solid State Circuits**, New York, v.35, n.11, p. 1571-1580, Nov. 2000.

BUTZEN, P. F.; MANCUSO, R.; SCHNEIDER, F. R.; ROSA JUNIOR, L. S. da; REIS, A. I.; RIBAS, R. P. Subthreshold Leakage Estimation in CMOS Complex Gates. In: SOUTH SYMPOSIUM ON MICROELECTRONICS, 22., 2007, Porto Alegre. **Proceedings…** Porto Alegre: SBC, 2007. p.47-50.

BUTZEN, P. F.; REIS, A. I.; KIM, C. H.; RIBAS, R. P. Modeling and estimating leakage current in series-parallel CMOS networks. In: GREAT LAKES SYMPOSIUM ON VLSI, 2007 Stresa-Lago Maggiore. **Proceedings...** New York: ACM, 2007. p.269-274.

BUTZEN, P. F.; REIS, A. I.; KIM, C. H.; RIBAS, R. P. Modeling subthreshold leakage current in general transistor networks. In: IEEE COMPUTER SOCIETY ANNUAL SYMPOSIUM ON VLSI, 2007, Porto Alegre. **Proceedings…** Los Alamitos: IEEE Computer Society, 2007. p.512-513.

BUTZEN, P. F.; REIS, A. I.; KIM, C. H.; RIBAS, R. P. Subthreshold Leakage Modeling and Estimation of General CMOS Complex Gates. In: INTERNATIONAL WORKSHOP ON POWER AND TIMING MODELING, OPTIMIZATION AND SIMULATION, 2007, Goteborg. **Proceedings…** Heidelberg: Springer, 2007. p. 474-484.

BUTZEN, P. F.; REIS, A. I.; RIBAS, R. P. Modeling and estimating leakage current in pass transistor logic networks. In: WORKSHOP IBERCHIP, 13., 2007, Lima. **XIII Workshop IBERCHIP**. Lima: Hazla S. R. L., 2007. p.295-298.

BUTZEN, P. F.; SCHNEIDER, F. R.; REIS, A. I.; RIBAS, R. P. Leakage reduction technique for CMOS complex gates. In: SOUTH SYMPOSIUM ON MICROELECTRONICS, 21., 2006, Porto Alegre. **Proceedings…** Porto Alegre: Instituto de Informática, UFRGS, 2006. p.111-114.

CAO, K.M. et al. BSIM4 Gate Leakage Model Including Source-Drain Partition.In: INTERNATIONAL ELECTRON DEVICES MEETING, 2000. **Digest of Technical Papers.** [S.l.: s.n.], 2000. p. 815-818.

CARLEY, L. R.; AGGARWAL, A. A Completely On-Chip Voltage Regulation Technique for Low Power Digital Circuits. In: INT. SYMP. LOW POWER ELECTRONIVS AND DESING, ISLPED, 1999. **Proceedings…** New York: ACM SIGDA, c2000. p.109-111.

CHANDRAKASAN, A. P.; BRODERSEN, R. W. Minimizing Power Consumption in Digital CMOS Circuits. **Proceedings of the IEEE**, New York, v.83, n.4, p.498-523, April 1995.

CHEN, Z. et al. Estimation of Standby Leakage Power in CMOS Circuits Considering Accurate Modeling of Transistor Stacks. In: INT. SYMP. LOW POWER ELECTRONIVS AND DESING, ISLPED, 1998. **Proceedings…** New York: ACM SIGDA, 1998. p.239-244.

DONAGHY, D.; BRACKENBURY, L.; HALL, S. A Simulation Study to Quantify the Advantages of Silicon-On-Insulator (SOI) Technology for Low Power. In: LOW POWER IC DESIGN SEMINAR, 2001. **Proceedings…** London: IEE, 2001. p. 11/1-11/6.

FUSE, T. et al. A 0.5 V Power-Supply Scheme for Low Power LSIs Using Multi-Vt SOI CMOS Technology. In: SYMPOSIUM ON VLSI CIRCUITS, 2001. **Digest of Technical Papers.** [S. l.]: IEEE, 2001. p. 219-220.

GAVRILOV, S. et al. Library-less Synthesis for Static CMOS Combinational Logic Circuits, In: IEEE/ACM INTERNATIONAL CONFERENCE ON COMPUTER-AIDED DESIGN, ICCAD, 1997. **Proceedings…** New York: ACM SIGDA, 1997. p. 658-662.

GRONOWSKI, P. E. et al. High-Performance Microprocessor Design. **IEEE Journal of Solid State Circuits**, New York, v.33, n.5, p. 676-686, May 1998.

GU, R. X.; ELMASRY, M.I. Power Dissipation Analysis and Optimization of Deep Submicron CMOS Digital Circuits. **IEEE Journal of Solid State Circuits**, New York, v.31, n.5, p. 707-713, May 1996.

GUINDI, R. S.; NAJM, F. N. Design techniques for gate-leakage reduction in CMOS circuits. In: INT. SYMP. QUALITY ELECTRONIC DESIGN, ISQED, 2003. **Proceedings…** [S.l.]: IEEE, 2003. p.61-65.

IMAN, S.; PEDRAM, M. **Logic Synthesis for Low Power VLSI Design**. Boston: Kluwer Academic, 1998. 236p.

KAO, J.; CHANDRAKASAN, A.; ANTONIADIS, D. Transistor Sizing Issues and Tool for Multi-Threshold CMOS Technology. In: ACM/IEEE DESIGN AUTOMATION CONFERENCE, 1997. **Proceedings...** [S.l.]: IEEE, 1997. p.409-414.

KESHAVARZI, A. et al. Effectiveness of Reverse Body Bias for Leakage Control in Scaled Dual Vt CMOS ICs. In: INT. SYMP. LOW POWER ELECTRONIVS AND DESING, ISLPED, 1998. **Proceedings...** New York: ACM SIGDA 2001. p.207-212.

KIM, C. H.; ROY, K. Dynamic V_{th} Scaling Scheme for Active Leakage Power Reduction. In: ACM/IEEE DESIGN AUTOMATION AND TEST IN EUROPE CONFERENCE, 2002. **Proceedings...** [S.l.]: IEEE, 2002. p. 163-167.

KRISHNARNURTHY, R. K. et al. High-Performance and Low-Power Challenges for Sub-70 nm Microprocessor Circuits. In: CUSTOM INTEGRATED CIRCUIT CONFERENCE, 2002. **Proceedings...** [S.l.]: IEEE, 2002. p. 125-128.

KURODA, T. et al. A 0.9-V, 150-MHz, 10-mW, 4 mm^2, 2-D Discrete Cosine Transform Core Processor with Variable Threshold-Voltage (VT) Scheme. **IEEE Journal of Solid State Circuits**, New York, v.31, n.11, p. 1770-1779, Nov. 1996.

LEE, D.; KWONG, W.; BLAAUW, D.; SYLVESTER, D. Analysis and Minimization Techniques for Total Leakage Considering Gate Oxide Leakage. In: ACM/IEEE DESIGN AUTOMATION CONFERENCE, 2003. **Proceedings...** [S.l.]: IEEE, 2003. p.175-180.

LEON, A. S. et al. A Power-Efficient High-Throughput 32-Thread SPARC Processor. **IEEE Journal of Solid State Circuits**, New York, v.42, n.1, p. 7-16, Jan. 2007.

LINDGREN, P. et al. Low power optimization technique for BDD mapped circuits. In: ACM/IEEE ASIAN DESIGN AUTOMATION CONFERENCE, 2001. **Proceedings...** [S.l.]: IEEE, 2001. p. 615-621.

MOORE, G. E. No exponential is forever: but "Forever" can be delayed! In: IEEE INT. CONF. SOLID STATE CIRCUITS, 2003. **Proceedings...** [S.l.]: IEEE, 2003. p. 20-23.

MUKHOPADHYAY, S. et al. Gate Leakage Reduction for Scaled Device Using Transistor Stacking. **IEEE Trans. on VLSI Systems**, New York, v.11, n.4, p. 716-730, Aug. 2003.

MUKHOPADHYAY, S.; RAYCHOWDHURY, A.; ROY, K. Accurate Estimation of Total Leakage in Nanometer-Scale Bulk CMOS Circuits Based on Device Geometry and Doping Profile. **IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems**, New York, v.24, n.3, p. 363-381, Mar. 2005.

MUTOH, S. et al. 1-V Power Supply High-speed Digital Circuit Technology with Multi-threshold Voltage CMOS. **IEEE Journal of Solid State Circuits**, New York, v.30, n.8, p. 847-854, Aug. 1995.

NARENDRA, S. et al. Forward Body Bias for Microprocessors in 130-nm Technology Generation and Beyond. **IEEE Journal of Solid State Circuits**, New York, v.38, n.5, p. 696-701, May 2003.

NARENDRA, S. G.; CHANDRAKASAN, A. **Leakage in Nanometer CMOS Technologies**. New York: Springer, 2006. 307 p.

OGURA, S. et al. Design and Characteristics of the Lightly Doped Drain-Source (LDD) Insulated Gate Field-Effect Transistor. **IEEE Journal of Solid State Circuits**, New York, v. SC-15, n.4, p. 424-432, Aug. 1980.

PANT, P. et al. Simultaneous power supply, threshold voltage, and transistor sizeoptimization for low-power operation of CMOS circuits. **IEEE Trans. on VLSI Systems**, New York, v.6, n.4, p. 538-545, Dec. 1998.

PARK, J. C.; MOONEY III, V. J. Sleepy Stack Leakage Reduction. **IEEE Transactions on VLSI Systems**, New York, v.14, n.11, p.1250-1262, Nov. 2006.

RAO, R. M. et al. Efficient techniques for gate leakage estimation. In: INT. SYMP. LOW POWER ELECTRONICS AND DESING, ISLPED, 2003. **Proceedings…** New York: ACM SIGDA, 2003. p.100-103.

ROY, K.; MUKHOPADHYAY, S.; MAHMOODI-MEIMAND, H. Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits. **Proceedings of the IEEE**, New York, v.91, n.2, p. 305-327, Feb. 2003

ROY, K.; PRASAD, S. C. **Low-Power CMOS VLSI Circuit Design**. New York: Wiley Interscience, 2000. 359 p.

SHELAR, R. S.; SAPATNEKAR, S. BDD decomposition for delay oriented pass transistor logic synthesis. **IEEE Trans. on VLSI**, New York, v. 13, n. 8, p. 957-970, Aug. 2005.

SHEU, B. J. et al. BSIM: Berkeley Short-Channel IGFET Model for MOS Transistors. **IEEE Journal of Solid State Circuits**, New York, v.SC-22, n.4, p. 558-566, Aug. 1987.

SHIGEMATSU, S. et al. 1-V high-speed MTCMOS circuit scheme for power-down application circuits. **IEEE Journal of Solid State Circuits**, New York, v.32, n.6, p. 861-869, June 1997.

SINGER, P. Intel and IBM Commit to High-K, Metal Gates. **Semiconductor International**. Available at: <http://www.reed-electronics.com/semiconductor/article/ CA6410945?spacedesc=news>. Visited on: Jan. 2007.

SOUDRIS, D.; PIGUET,C.; GOUTIS, C. **Designing CMOS Circuits for Low Power**. Boston: Kluwer Academic, 2002. 277 p.

SUTHERLAND, I. E.; SPROULL, R. F.; HARRIS, D. F. **Logical Effort:** Designing Fast CMOS Circuits. San Francisco: Morgan Kaufmann, 1999. 239 p.

TAKAHASI, M. et al. A 60-mW MPEG4 video codec using clustered voltage scaling with variable supply-voltage scheme. **IEEE Journal of Solid State Circuits**, New York, v.33, n.11, p. 1772-1780, Nov. 1998.

TAKAYANAGI, T. et al. A Dual-Core 64-bit UltraSPARC Microprocessor for Dense Server Applications. **IEEE Journal of Solid State Circuits**, New York, v.40, n.1, p. 7-17, Jan. 2005.

TAUR, Y.; NING, T. H. **Fundamentals of Modern VLSI Design**. New York: Cambridge University Press, 1998.

VEENDRICK, H.J.M. Short-Circuit Dissipation of Static CMOS Circuitry and its Impact on the Design of Buffer Circuits. **IEEE Journal of Solid State Circuits**, New York, v.SC-19, n.4, p. 468-473, Aug. 1984.

WANG, A.; CALHOUN, B.; CHANDRAKASAN, A. P. **Sub-Threshold Design for Ultra Low-Power Systems**. New York: Springer, 2006. 209 p.

WEI, L. et al. Design and Optimization of Dual-Threshold Circuits for Low-Voltage Low-Power Applications. **IEEE Trans. on VLSI Systems**, New York, v.7, n.1, p. 16-24, Mar. 1999.

YANG, C.; CIESIELSKI, M. BDS: a BDD-based logic optimization system. **IEEE Trans. on CAD**, New York, v.21, n.7, p. 866-876, July 2002.

YANG, S. et al. Accurate Stacking Effect Macro-modeling of Leakage Power in Sub-100nm Circuits. In: INTERNATIONAL CONFERENCE ON VLSI DESIGN, 18., 2005. **Proceedings…** Los Alamitos, CA: IEEE Computer Society, 2005. p. 165-170.

APPENDIX A PRESENTATION SLIDES
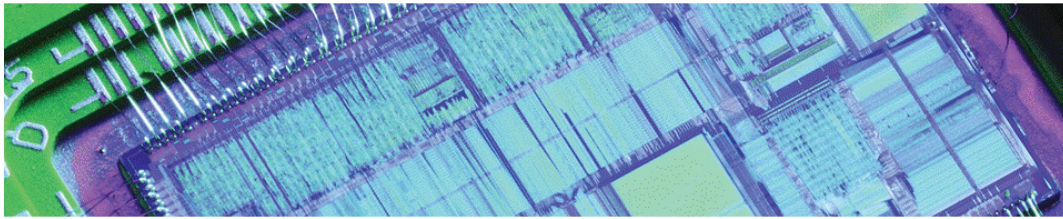
Outline

- Introduction

- Static Consumption Overview

- Motivation

- Subthreshold Leakage Model

- Gate Oxide Leakage Model

- Conclusions and Future Works

2/3/2008   Paulo F Butzen   2

Introduction
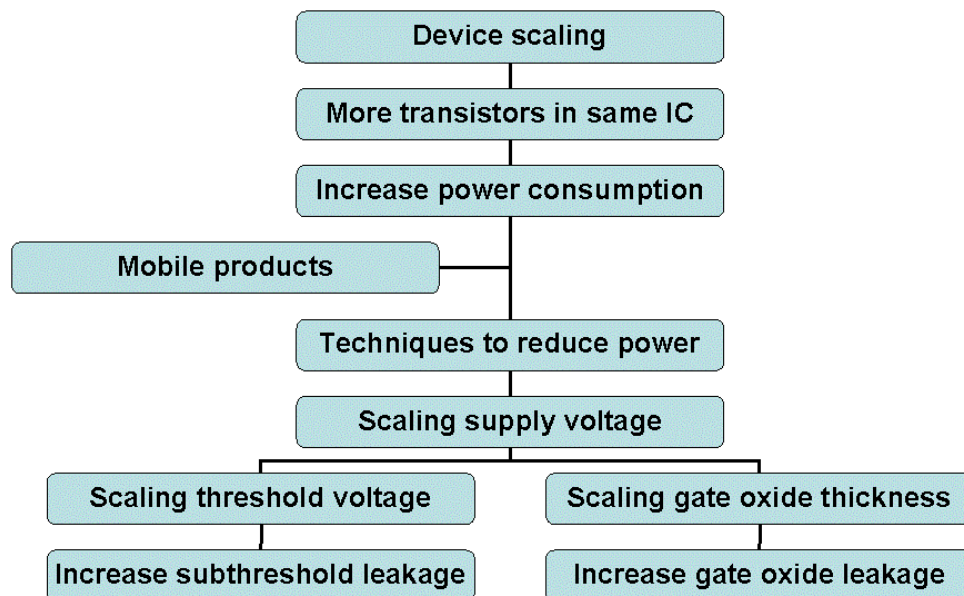
2/3/2008   Paulo F Butzen   3

70

Introduction

- Leakage currents are increasing significantly in advanced CMOS technologies due the threshold voltage and the gate oxide thickness scaling

(MOORE, 2003)
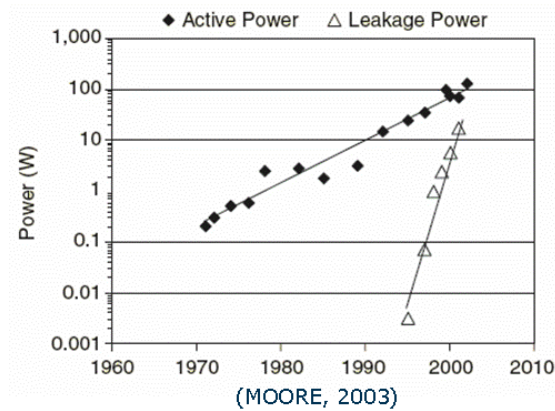
Leakage Mechanisms

- Subthreshold leakage
- Gate oxide leakage
- Band-to-band tunneling (BTBT) leakage
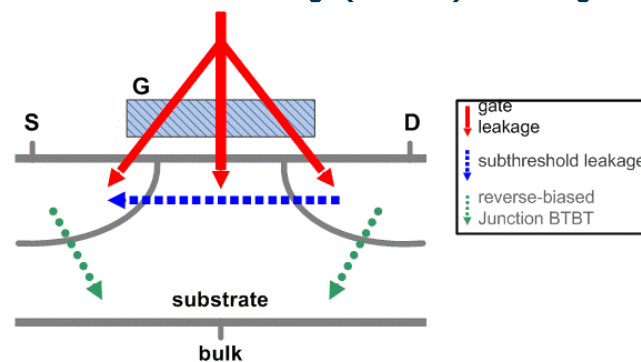
- BTBT starts to be taken into account in 25nm process (MUKHOPADHYAY, 2005)

Subthreshold Leakage

- Current between drain and source nodes when transistor is turned off
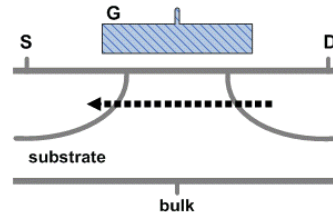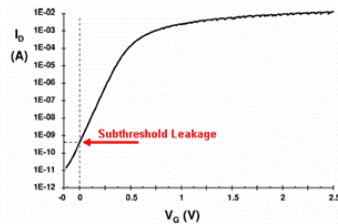
- Exponential dependent of both gate-to-source voltage (V_{gs}) and threshold voltage (V_{th})

$$I_{subthresho\,ld} = I_0 e^{\frac{V_{gs}-V_{th}}{nV_T}}\left[1 - e^{-\frac{V_{ds}}{V_T}}\right]$$

Gate Oxide Leakage

- Tunneling of electrons (or holes) from the bulk and source/drain overlap region through the gate oxide potential barrier into the gate (or vice-versa)
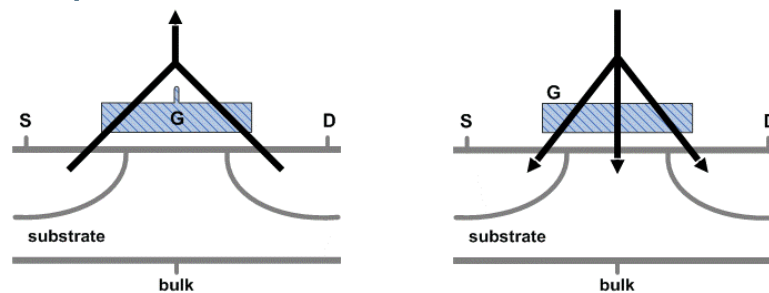
- Exponential dependent of both voltage across gate oxide and oxide thickness

Leakage Reduction Techniques

- Reduction in leakage current can be achieved using both process and circuit level techniques
- Process techniques
  - Dimensions control (oxide thickness, junction depth, …)
  - Doping profile
- Circuits techniques
  - Dual Threshold CMOS
  - Supply Voltage Scaling
  - Transistor Stack Effect
  - Power Gating
  - Body Biasing

Motivations and Goals

Develop analytical leakage current model to

evaluate CMOS complex gates

- Complex gates are widely used in library-free technology mapping
- Previous analytical models are not able to evaluate complex gates
- Analytical models are useful for guiding synthesis process

Subthreshold Leakage Model

- Subthreshold leakage behavior

- Estimation based on conductance association

- Previous analytical subthreshold leakage models

- Proposed subthreshold leakage model

- Experimental results

Subthreshold Leakage Behavior

- **Parallel transistors**
  - I_{total} is the sum of each transistor I_{sub}
  - $I_{total} = \Sigma \, I_{sub\_Mi}$

I_{sub_M1} $I_{sub\_M2} + I_{sub\_M3}$
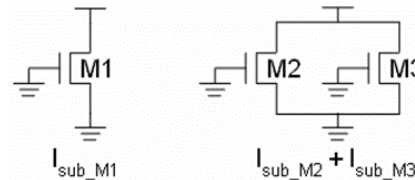
- **Series transistors**
  - Stack effect: series off-transistors present more leakage reduction than single equivalent transistor
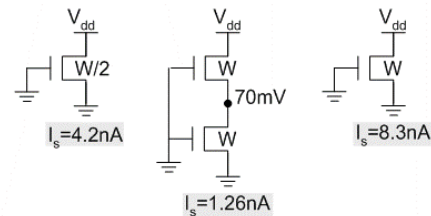
Stacking Effect

- Subthreshold leakage
  current flowing through a
  stack of series-connected
  transistors reduces when
  more than one transistor
  in the stack is turned off

Leakage (nA)

No. of transistors off in stack
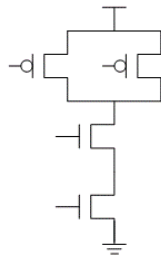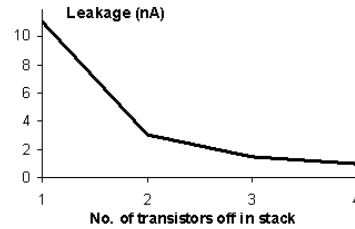
Subthreshold leakage current for 2-input NAND gate.

Input Vector	Leakage current (nA)
00	3.94
01	15.25
10	13.65
11	4.57

Estimation Based on Conductance Association

- Proposed model provides a normalized current
  value, related to a NMOS transistor
- Conductance of parallel devices are summed
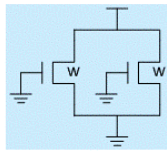
$$G_{eq} = G_{T[1]} + G_{T[2]} = 1 + 1 = 2$$

- Equivalent conductance of series arrangements
  is inversely proportional to the sum of the
  inverse device conductance
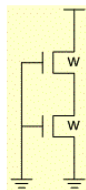
$$G_{eq} = \frac{1}{\dfrac{1}{G_{T[1]}} + \dfrac{1}{G_{T[2]}}} = \frac{1}{\dfrac{1}{1} + \dfrac{1}{1}} = \frac{1}{2}$$

Constant *K* – Stack Effect Modeling

- **Constant *K* included to calibrate the final result**
  - Obtained by relating the leakage current of two-stack and single off-device configurations

  - **K = 0.26** (in this example)

- **Series association**

I_{sub} >> $0.13\ I_{sub}$

$$G_{eq} = \frac{K}{\dfrac{1}{G_{T[1]}} + \dfrac{1}{G_{T[2]}}} = \frac{K}{\dfrac{1}{1} + \dfrac{1}{1}} = \frac{K}{2}$$

Example

$$G_{T[1]} + G_{T[2]}$$

Example

Example

Example

$$\cfrac{K}{\cfrac{1}{\cfrac{1}{\cfrac{1}{G_{T[1]}+G_{T[2]}}+\cfrac{1}{G_{T[4]}}}+G_{T[3]}}+\cfrac{1}{G_{T[5]}}}$$
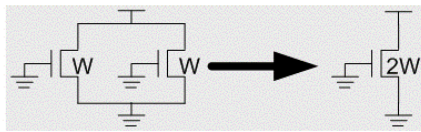
Experimental Results

78

Previous Analytical Subthreshold Leakage Models

- GU, 1996; ROY, 2000; NARENDRA, 2006
- Maximum two levels series/parallel off-transistors arrangements (NAND, NOR, AOI and OAI gates)
- Parallel transistors are collapsed and replaced by a single device with equivalent size

OAI gate (pull-down plane): NMOS off-network

- Stack effect: series off-transistors present more leakage reduction than single equivalent transistor

I_{sub} $>>$ $0.13\ I_{sub}$ $>$ $0.07\ I_{sub}$
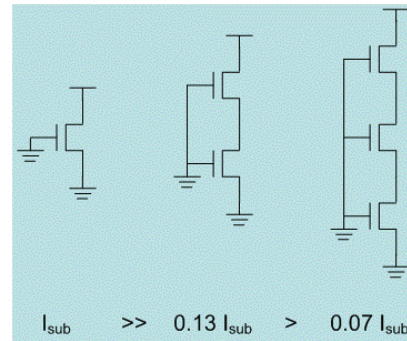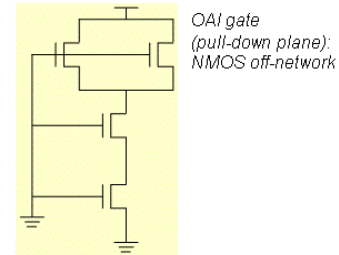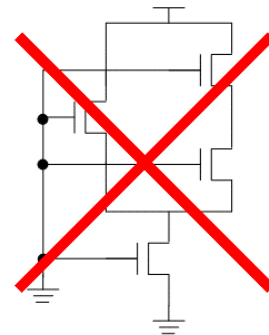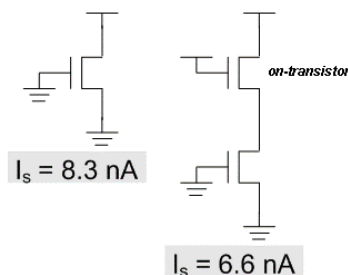
Previous Analytical Subthreshold Leakage Models

- Limitations
  - Complex gates with more than two logic depth levels cannot be treated
  - ON-transistors in OFF-networks are considered ideal short-circuits (on-switches)

on-transistor

$I_s = 8.3\ nA$

$I_s = 6.6\ nA$

Proposed Subthreshold Leakage Model

- BSIM subthershold current

$$I_S = I_0 W e^{\frac{V_{gs} + \eta V_{ds} + \gamma V_{bs}}{n V_T}} \left[1 - e^{-\frac{V_{ds}}{V_T}} \right]$$

- Approximation

$$V_{ds} > V_T \Rightarrow e^{-\frac{V_{ds}}{V_T}} \approx 0$$

$$V_{ds} < V_T \Rightarrow e^{-\frac{V_{ds}}{V_T}} \approx \left(1 - \frac{V_{ds}}{V_T} \right)$$
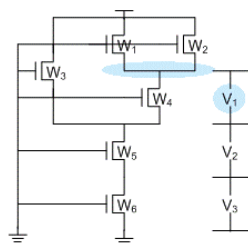
Example

- Subthreshold leakage current estimation in the following NMOS pull-down network

$$I_{S1} = I_0 W_1 e^{\frac{-(V_1 + V_2 + V_3) - [V_{t0} - \eta(V_{dd} - V_1 - V_2 - V_3) + \gamma(V_1 + V_2 + V_3)]}{n V_T}}$$

$$I_{S2} = I_0 W_2 e^{\frac{-(V_1 + V_2 + V_3) - [V_{t0} - \eta(V_{dd} - V_1 - V_2 - V_3) + \gamma(V_1 + V_2 + V_3)]}{n V_T}}$$

$$I_{S4} = I_0 W_4 e^{\frac{-(V_2 + V_3) - [V_{t0} - \eta V_1 + \gamma(V_2 + V_3)]}{n V_T}}$$

$$I_{S1} + I_{S2} = I_{S4}$$

$$V_1 = \frac{\eta V_{dd} + n V_T \ln\left(\frac{W_1 + W_2}{W_4}\right)}{1 + 2\eta + \gamma}$$

Example

$$I_{S3} = I_0 W_3 e^{\frac{-(V_2+V_3)-[V_{t0}-\eta(V_{dd}-V_2-V_3)+\gamma(V_2+V_3)]}{nV_T}}$$

$$I_{S3} + I_{S4} = I_{S5}$$

$$I_{S4} = I_0 W_4 e^{\frac{-(V_2+V_3)-[V_{t0}-\eta V_1+\gamma(V_2+V_3)]}{nV_T}}$$

$$I_{S5} = I_0 W_5 e^{\frac{-V_3-[V_{t0}-\eta V_2+\gamma V_3]}{nV_T}}$$

$$V_2 = \frac{nV_T \ln\left(\dfrac{W_3 e^{\frac{\eta V_{dd}}{nV_T}} + W_4 e^{\frac{\eta V_1}{nV_T}}}{W_5}\right)}{1+\eta+\gamma}$$

$$I_{S5} = I_{S6}$$

$$I_{S5} = I_0 W_5 e^{\frac{-V_3-[V_{t0}-\eta V_2+\gamma V_3]}{nV_T}}$$

$$e^{-\frac{V_3}{V_T}} \cong 1 - \left(\frac{V_3}{V_T}\right)$$

$$I_{S6} = I_0 W_6 e^{\frac{-V_{t0}+\eta V_3}{nV_T}}\left[1-e^{-\frac{V_3}{V_T}}\right]$$

$$\frac{1+\eta+\gamma}{n}\left(\frac{V_3}{V_T}\right) + \ln\left(\frac{V_3}{V_T}\right) = \frac{\eta V_2}{nV_T} + \ln\left(\frac{W_5}{W_6}\right)$$

General Subthreshold Leakage Model

1 - Evaluate nodes that have all transistors connected to output – Eq (a)

2 - Evaluate nodes that have some stack of two transistors until output – Eq (b)

3 - Evaluate nodes that have some stack with more than 2 transistors until output – Eq (c)

3.1 – If voltage > VT, evaluate node based on equation (b)

$$(a) \Rightarrow V_i = \frac{\eta V_{dd} + nV_T \ln\left(\dfrac{W_{above}}{W_{below}}\right)}{1+2\eta+\gamma}$$

$$(b) \Rightarrow V_i = \frac{nV_T \ln\left(\dfrac{\sum W_{above} e^{\frac{\eta V_{above}}{nV_T}}}{W_{below}}\right)}{1+\eta+\gamma}$$

$$(c) \Rightarrow \frac{C}{n}\left(\frac{V_i}{V_T}\right) + \ln\left(\frac{V_i}{V_T}\right) = \frac{\eta V_{above}}{nV_T} + \ln\left(\frac{W_{above}}{W_i}\right) + \ln\left(\frac{V_{above}}{V_T}\right)$$

ON-Transistors in OFF-Networks

- When NMOS ON-transistor is connected to V_{dd} (or PMOS ON-transistor to V_{ss}) the drop voltage (V_{drop}) across them should be taken into account

- Replace the term $V_{dd} - \Sigma V_j$ in subthreshold current equation by $V_{dd} - V_{drop} - \Sigma V_j$

ON-Transistors in OFF-Networks
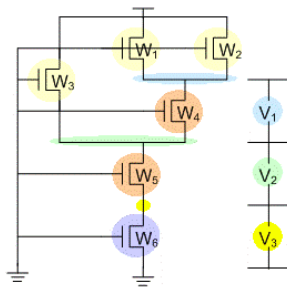
$$(a) \Rightarrow V_i = \frac{\eta V_{dd} + nV_T \ln\left(\frac{W_{above}}{W_{below}}\right)}{1 + 2\eta + \gamma}$$

$$(a) \Rightarrow V_i = \frac{\eta\left(V_{dd} - V_{drop}\right) + nV_T \ln\left(\frac{W_{above}}{W_{below}}\right)}{1 + 2\eta + \gamma}$$
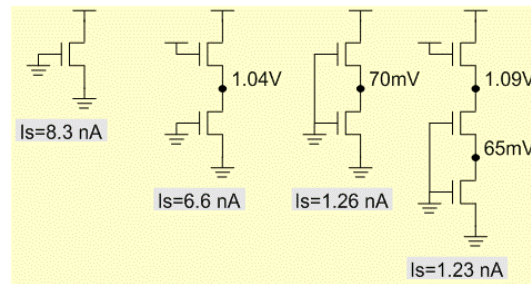
$$(b) \Rightarrow V_i = \frac{nV_T \ln\left(\frac{\sum W_{above} e^{\frac{\eta V_{above}}{nV_T}}}{W_{below}}\right)}{1 + \eta + \gamma}$$

82

Subthreshold leakage in H-like gate

- It is not possible to directly evaluate V_{ds} of all transistors
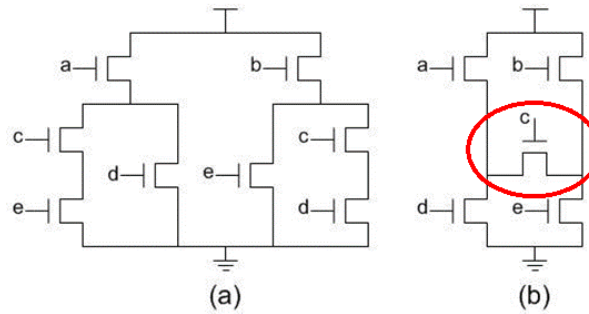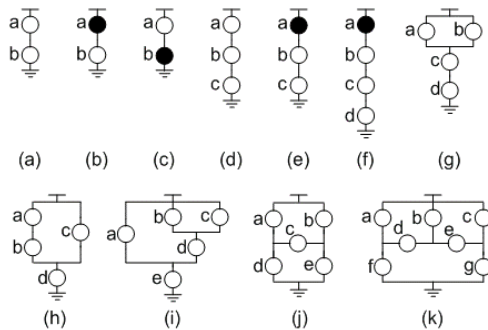- V_{ds} of specific transistor is dependent of voltage across transistors connected in its terminals
- Should be ignored to find first terminal voltage and computed when other terminal is evaluated

(a)   (b)

Experimental Results

(a)   (b)   (c)   (d)   (e)   (f)   (g)

(h)   (i)   (j)   (k)

Network	I_{sub} – HSPICE (nA)	I_{sub} – Model (nA)	Diff.(%)
(a)	1.26	1.26	-
(b)	6.58	6.60	0.3
(c)	8.34	8.34	-
(d)	0.69	0.75	8.7
(e)	1.23	1.24	0.8
(f)	0.68	0.74	8.8
(g)	0.72	0.77	6.9
(h)	1.29	1.28	0.8
(i)	1.29	1.28	0.8
(j)	2.52	2.53	0.4
(k)	2.56	2.54	0.8

- Networks (h) - (k), not treated by previous models, are accurately predicted

$$V_{ds} < V_T \Rightarrow e^{\frac{V_{ds}}{V_T}} \approx \left(1 - \frac{V_{ds}}{V_T}\right)$$

Experimental Results – Input Dependence

(h)   (i)
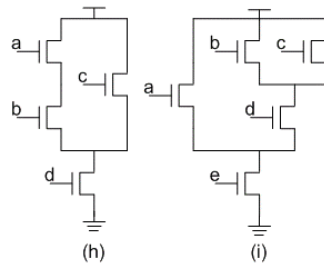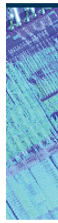
Input vector dependence in logic network (h)

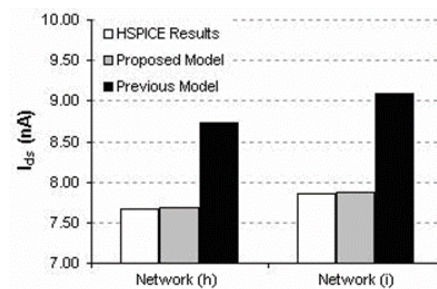| Input-state (abcd)	HSPICE Results (nA)	Proposed model (nA)	Previous model (nA)
0000	1.29	1.28	-
0001	9.60	9.60	9.60
0010 *	6.30/6.70	6.60	8.34
0100	1.37	1.31	1.31
0101	16.67	16.69	16.69
1000	1.36	1.30	1.31
1001	14.91	14.94	16.69

* Values given for min./max. currents related to such equivalent vectors.

Input vector dependence in logic network (i)

Input-state (abcd)	HSPICE Results (nA)	Proposed model (nA)	Previous model (nA)
00000	1.29	1.28	-
00001	9.71	9.65	9.65
00010	1.43	1.34	1.34
00011	25.00	25.02	25.02
00100 *	1.36/1.37	1.30	1.31
00101 *	14.91/15.14	14.94	16.69
00110 *	6.30/6.73	6.60	8.34

* Values given for min./max. currents related to such equivalent vectors.

Experimental Results – Sized gate

Input state (abcd)	HSPICE results (nA)	Proposed model (nA)	Diff(%)
0000	4.01	4.13	3.0
0001	20.67	20.68	0.0
0010	19.93	19.99	0.3
0011	44.52	43.27	2.8
0100	4.44	4.29	3.4
0101	42.34	42.37	0.1
0110	19.93	19.99	0.3
0111	43.38	43.27	0.3
1000	4.40	4.26	3.2
1001	36.81	36.50	0.8
1010	19.93	19.99	0.3
1011	43.38	43.27	0.3
1100	19.50	19.99	2.5
1101	96.67	96.92	0.3
1110	20.43	19.99	2.2
1111	20.48	20.21	1.3

Gate Oxide Leakage Model

- Gate oxide leakage behavior

- Previous gate oxide leakage models

- Proposed gate oxide leakage model

- Subthreshold and gate oxide leakage iteration

- Experimental results

Gate Oxide Leakage Behavior

- Exponential related with the voltage across gate oxide and the oxide thickness
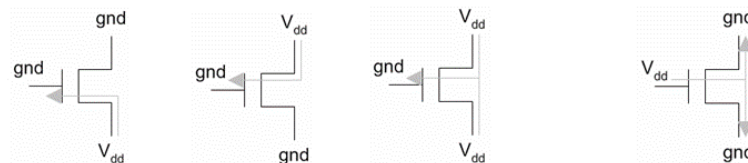
- Occurs when transistors are turned ON and turned OFF

- Dependent of transistor bias conditions

Previous Gate Oxide Leakage Models

- Based on transistor bias conditions

Device	Bias(GDS) and Control Conditions	90nm(nA)	65nm(nA)	45nm(nA)
NMOS	001	74.498	578.0	1585.7
	010	74.498	578.0	1585.7
	011	148.99	1155.4	3171.3
	100	171.17	1659.5	4032.0
PMOS	110	7.739	212.2	1240.5
	101	7.739	212.2	1240.5
	100	15.478	424.5	2480.7
	011	12.594	347.2	1991.5

(YANG, 2005)

- Internal nodes attain full logic levels
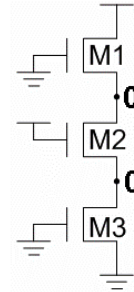
Proposed Gate Oxide Leakage Model

- Tunneling current density is expressed as (ROY, 2003):

$$J_{Tunneling\,Current\,Density} = W\,L.A.\left(\frac{V_{ox}}{T_{ox}}\right)^2.e^{\left(\dfrac{-B.\left(1-\left(1-\dfrac{V_{ox}}{\phi_{ox}}\right)^{3/2}\right)}{\dfrac{V_{ox}}{T_{ox}}}\right)}$$

- A simpler model to capture the dependence between gate leakage current and gate voltage is desirable for fast estimations

Proposed Gate Oxide Leakage Model

- Considering estimations to only one technology node, gate oxide thickness dependence can be suppressed

$$I_{gate} = I_{gate\_0}.W.e^{\frac{-K}{|V_{ox}|}}$$

K is the calibration constant, extracted by simulation based on difference between gate leakage currents to $V_{ox} = V_{dd}$ and $V_{ox} = 0.9*V_{dd}$
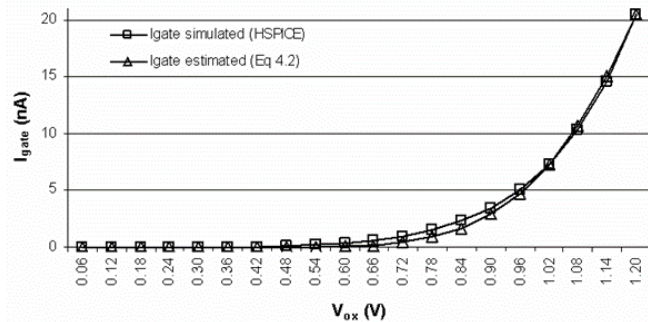
Proposed Gate Oxide Leakage Model

- Gate current is dependent on transistor bias conditions
- Can be grouped in OFF bias condition and ON bias condition
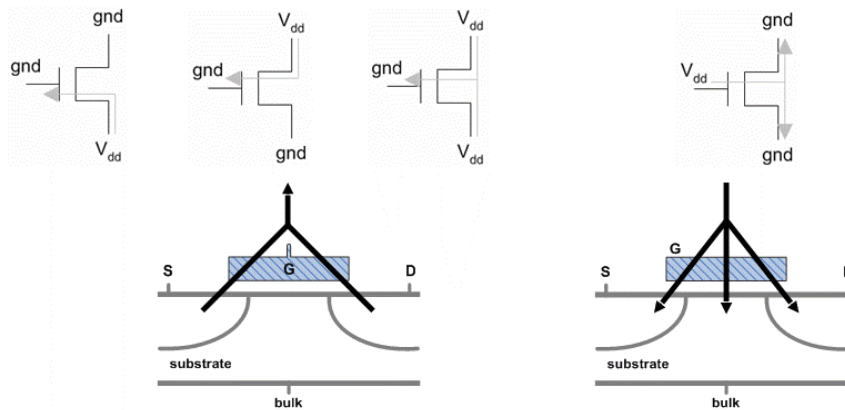
$$I_{gate\_OFF} = I_{gate\_OFF\_0}.W.e^{-\frac{K}{|V_{ox}|}}$$

$$I_{gate\_ON} = I_{gate\_ON\_0}.W.e^{-\frac{K}{|V_{ox}|}}$$

Proposed Gate Oxide Leakage Model

$$I_{gate\_OFF} = I_{gate\_OFF\_0}.W.e^{-\frac{K}{|V_{ox}|}}$$
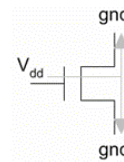
$$I_{gate\_ON} = I_{gate\_ON\_0}.W.e^{-\frac{K}{|V_{ox}|}}$$

- Drain and source voltages are crucial to a accurate estimation
- Iteration between subthreshold and gate leakage is required
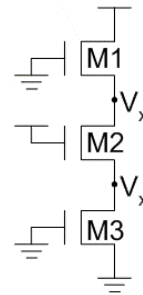
Subthreshold and Gate Leakage Iteration

- YANG, 2005 and LEE, 2003 present works that compute both subthreshold and gate leakages.

- In both works subthreshold and gate currents are treated independently and summed.

$$I_{leak\_total} = I_{gate\_total} + I_{sub\_total}$$

- It can work well when one of leakage mechanisms is dominant

88

Subthreshold and Gate Leakage Iteration

- To evaluate accurately gate and subthreshold currents is necessary find the intermediate voltage (V_x)

Subthreshold and Gate Leakage Iteration

$$I_{g\_OFF} = I_{gate\_OFF\_0}.W.e^{-\frac{K}{|V_{ox}|}}$$

$$I_{g\_ON} = I_{gate\_ON\_0}.W.e^{-\frac{K}{|V_{ox}|}}$$

$$I_S = I_0 We^{\frac{V_{gs}+\eta V_{ds}+\gamma V_{bs}}{nV_T}}\left[1-e^{-\frac{V_{ds}}{V_T}}\right]$$

$$I_{S\_1} + I_{g\_ON} = I_{S\_3} + I_{g\_OFF\_2} + I_{g\_OFF\_3}$$

$$I_{S0}W_1 e^{\frac{-V_x-(V_{t0}-\eta(V_{dd}-V_x)+\gamma V_x)}{nV_T}}\left[1-e^{-\frac{V_{dd}-V_x}{V_T}}\right] + I_{g\_ON\_0}.W_2.e^{-\frac{K}{V_{dd}-V_x}} =$$

$$I_{S0}W_3 e^{\frac{\eta V_x}{nV_T}}\left[1-e^{-\frac{V_x}{V_T}}\right] + I_{g\_OFF\_0}.(W_1+W_3).e^{-\frac{K}{V_x}}$$

Example

$$V_1 = V_2$$
$$V_3 = V_{ss}$$

I_{g_OFF}

I_{g_ON}

I_{sub}

V$_{ss}$ M2 V$_{ss}$ M3

V$_{ss}$ M1

V$_1$

V$_{dd}$ M4

V$_2$

V$_{ss}$ M5

V$_3$

V$_{dd}$ M6

Example

$$I_{g\_ON} = I_{gate\_ON\_0}.W.e^{-\frac{K}{|V_{ox}|}}$$

$$I_{g\_OFF} = I_{gate\_OFF\_0}.W.e^{-\frac{K}{|V_{ox}|}}$$

$$I_S = I_0 W e^{\frac{V_{gs}+\eta V_{ds}+\gamma V_{bs}}{nV_T}}\left[1 - e^{-\frac{V_{ds}}{V_T}}\right]$$

V$_{ss}$ M2 V$_{ss}$ M3
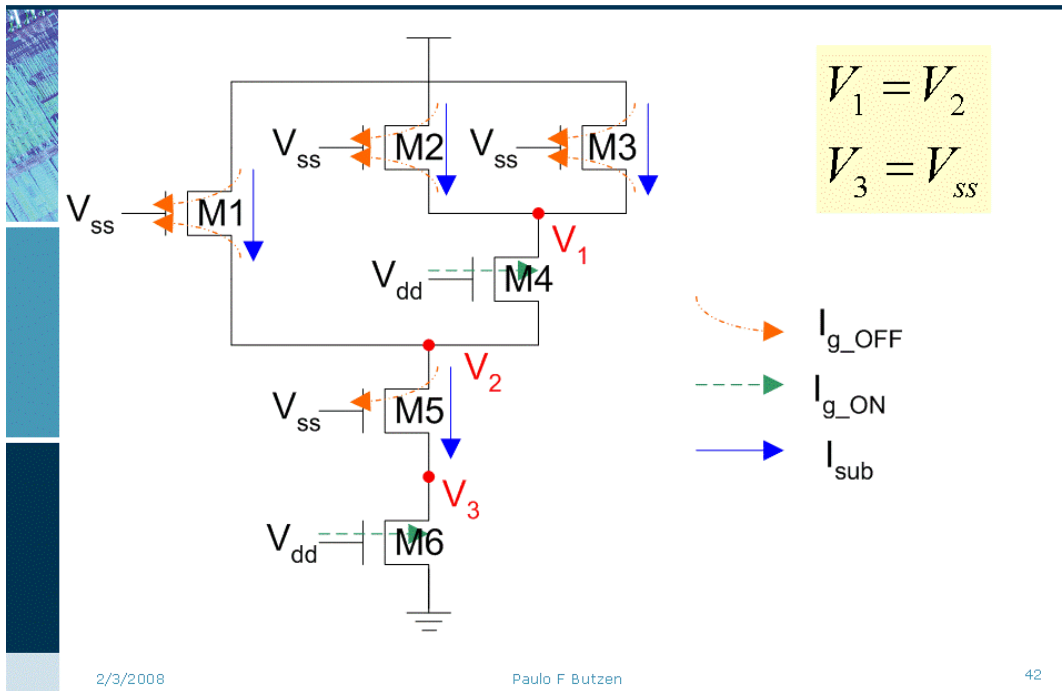
V$_{ss}$ M1

V$_1$

V$_{dd}$ M4

V$_1$

V$_{ss}$ M5

V$_{ss}$

V$_{dd}$ M6

I_{g_OFF}

I_{g_ON}

I_{sub}

$$I_{S\_1} + I_{S\_2} + I_{S\_3} + I_{g\_ON\_4} =$$
$$I_{S\_5} + I_{g\_OFF\_1} + I_{g\_OFF\_2} + I_{g\_OFF\_3} + I_{g\_OFF\_5}$$

90

Experimental Results

Network	I_{sub} − HSPICE (nA)	I_{sub} − Model (nA)	Diff.(%)
(a)	5.37	5.35	0.4
(b)	15.21	15.26	0.3
(c)	42.12	42.11	0.0
(d)	4.23	4.02	5.0
(e)	7.12	7.27	2.1
(f)	25.73	25.71	0.1
(g)	27.47	27.65	0.7
(h)	8.45	8.42	0.4
(i)	105.83	105.97	0.1
(j)	12.88	13.24	2.8
(k)	55.00	55.35	0.6

Experimental Results – Sized gate

Input state (abcd)	HSPICE results (nA)	Proposed model (nA)	Diff(%)
0000	18.97	19.35	2.0
0001	122.80	121.85	0.8
0010	50.66	50.35	0.6
0011	156.12	153.96	1.4
0100	24.58	24.87	1.2
0101	234.66	231.91	1.2
0110	50.58	50.22	0.7
0111	217.21	215.50	0.8
1000	17.25	17.18	0.4
1001	152.68	153.91	0.8
1010	48.29	48.95	1.4
1011	217.19	215.60	0.7
1100	48.12	48.79	1.4
1101	293.87	291.73	0.7
1110	46.64	47.43	1.7
1111	255.59	257.82	0.9

Conclusions and Future Works

- **Main objective of this research were achieved:**
  - Develop an analytical leakage estimation method to general transistor networks
- **Review leakage mechanisms and reduction techniques, providing a minimum background to IC designers about leakage currents**
- **Simple leakage estimation method was presented**
- **Future Works**
  - Implement the methods
  - Develop BTBT leakage model to evaluate future technologies

Publications

- Paulo F Butzen, André I. Reis, Chris H. Kim, Renato P. Ribas, "Subthreshold Leakage Modeling and Estimation of General CMOS Complex Gates", PATMOS'07
- Paulo F Butzen, André I. Reis, Chris H. Kim, Renato P. Ribas, "Modeling and Estimating Leakage Current in Series-Parallel CMOS Networks", GLSVLSI'07
- P. F. Butzen, R. Mancuso, F. R. Schneider, L. S. Rosa Jr, A. Reis, R. P. Ribas, "Leakage Behavior in CMOS and PTL Logic Styles for Logic Synthesis Orientation", IWLS'07
- Paulo F Butzen, André I. Reis, Chris H. Kim, Renato P. Ribas, "Modeling Subthreshold Leakage Current in General Transistor Networks", ISVLSI'07
- Paulo F Butzen, André I. Reis, Renato P. Ribas, "Modeling and Estimating Leakage Current in Pass Transistor Logic Networks", IBERCHIP'07.
- P. F. Butzen, R. Mancuso, F. R. Schneider, L. S. Rosa Jr, A. Reis, R. P. Ribas, "Subthreshold Leakage Estimation in CMOS Complex Gates", SIM'07
- Paulo F Butzen, Felipe R. Schneider, André I. Reis, Renato P. Ribas, "Leakage Reduction Technique for CMOS Complex gates", SIM'06

APPENDIX B MODELAGEM DA CORRENTE DE FUGA EM CÉLULAS COMPLEXAS SUB-MICROMÉTICAS

Resumo da Dissertação em Português

As tecnologias utilizadas na concepção de circuitos integrados estão em constante miniaturização durante as ultimas décadas. Efeitos antes ignorados na análise de circuitos integrados, como as correntes de fuga, devido a sua magnitude reduzida passam a ser considerados quando avaliados em tecnologias sub-micrométricas. Além disso, a potência dissipada pelos dispositivos integrados passou a ser um importante critério durante o design dos circuitos devido ao advento dos aparelhos portáteis e da comunicação wireless.

A redução da tensão de alimentação é uma forma natural de diminuir a potência consumida pelos dispositivos integrados. Para não comprometer o desempenho destes circuitos, a tensão de *threshold* também é reduzida, provocando um incremento significativo na magnitude da corrente de fuga de subthreshold. A corrente de tunelamento através da porta do transistor também sofre um aumento significativo a cada novo processo de fabricação. Este aumento se deve a necessidade de reduzir da espessura do óxido de porta para evitar o agravamento de efeitos de canal curto como o efeito de corpo e DIBL (Drain Induced Barrier Lowering). Outra medida utilizada para manter sob controle os efeitos de canal curto é o aumento da dopagem do substrato e das regiões de dreno e source do transistor. Esse aumento da dopagem faz com que as correntes reversas nas junções pn do transistor também passem a ser consideradas. Essas correntes de fuga podem representar de 30-50% de toda a potência dissipada em um circuito em condições normais de operação.

Este trabalho revisa os principais mecanismos de fuga e técnicas de redução, e apresenta um método analítico de estimativa das correntes de fuga válido para qualquer rede transistores.

Apesar de existirem diversos mecanismos de fuga nos transistores de tecnologias sub-micrométricas, pode-se considerar que os três principais são: A corrente de fuga de subthreshold, descrita na equação (1), exponencialmente dependente da tensão de theshold e da temperatura; A corrente de tunelamento através da porta do transistor, descrita na equação (2), exponencialmente dependente da tensão aplicada à porta e da espessura do óxido; A corrente de fuga na junção pn polarizada reversamente, descrita na equação (3), exponencialmente dependente da temperatura e da dopagem do substrato e das regiões de dreno e source dos transistores.

$$I_{subthresho\,ld} = I_0 e^{\frac{V_{gs}-V_{th}}{nV_T}}\left[1-e^{-\frac{V_{ds}}{V_T}}\right]$$

(1)

onde $I_0 = \dfrac{W\mu_0 C_{ox}V_T^2 e^{1.8}}{L}$, $V_T = \dfrac{KT}{q}$ é a tensão termica, V_{th} é a tensão de threshold, V_{ds} e V_{gs} são as tensões dreno-source e porta-source respectivamente. W e L são respectivamente a largura e comprimento efetivos do transistor. C_{ox} é a capacitancia do oxido de porta, μ_0 é a mobilidade das cargas e n é o coeficiente de subthreshold swing.

$$I_{gate} = W.L.A\left(\frac{V_{ox}}{t_{ox}}\right)^2 \exp\left(\frac{-B\left(1-\left(1-\frac{V_{ox}}{\phi_{ox}}\right)^{3/2}\right)}{\frac{V_{ox}}{t_{ox}}}\right)$$

(2)

onde W e L são respectivamente a largura e comprimento efetivos do transistor, $A = q^3/16\pi^2 h\phi_{ox}$, $B = 4\pi\sqrt{2m_{ox}}\phi_{ox}^{3/2}/3hq$, m_{ox} é a massa efetiva de tunelamento de uma partícula através do oxido de porta, ϕ_{ox} é a altura da barreira de tunelamento, t_{ox} é a espessura do óxido, h é ½*π vezes a constante de Planck e q é a carga do electron.

$$J_{BTBT} = A\frac{EV_{app}}{E_g^{1/2}}\exp\left(-B\frac{E_g^{3/2}}{E}\right)$$

(3)

onde $A = \sqrt{2m*q^3}/4\pi^3 h^2$, e $B = 4\sqrt{2m*}/3hq$. $m*$ é a massa efetiva do eletron; Eg é o energy-band gap; V_{app} é a tensão reversa aplicada à junção; E é o campo eletrico na junção; q é a carga do eletron; e h é ½*π vezes a constante de Planck.

Diversas técnicas de redução da corrente de fuga são encontradas na literatura. Esta redução pode ser alcançada tanto em nível de processo, através das dimensões e dopagem do dispositivo, quanto em nível de circuito, onde é explorada a dependência com a tensão de alimentação e a tensão de threshold, o conceito de "stack effect" e polarização do substrato.

A técnica denominada "Dual threshold CMOS" explora a diferença de atraso nos caminhos não críticos para reduzir as correntes de fuga. Esta técnica exige que o processo de fabricação forneça transistores com duas tensões de threshold diferentes. Transistores "high V_{th}" são utilizados para reduzir a corrente de fuga nos caminhos não críticos do circuito enquanto transistores "low V_{th}" garantem o desempenho dos caminhos críticos.

A redução da tensão de alimentação é uma técnica comumente utilizada para reduzir o consumo de potência de dinâmica. Esta mesma técnica é válida para as correntes de fuga, visto que todos os mecanismos citados anteriormente são exponencialmente dependentes da tensão de alimentação. Esta redução da tensão de alimentação irá provocar uma perda de desempenho do circuito. A tensão ótima para a redução do consumo utilizando esta técnica é a menor tensão de alimentação que não comprometa o desempenho do circuito.

Uma pilha de dois ou mais transistores não conduzindo provocam uma redução significativa na corrente de fuga de subthreshold. Isso se deve ao fato de existir uma pequena tensão nos nodos intermediários da pilha de transistores. A Figura 1 ilustra o comportamento da corrente de fuga de subthreshold para pilhas com diferentes números de transistores não conduzindo. Este efeito é largamente explorado pela técnica de controle dos vetores de entrada do circuito. Esta técnica aplica um determinado vetor quando o circuito esta em modo idle. Este vetor maximiza a quantidade de pilhas de transistores não conduzindo do circuito, e conseqüentemente reduz a corrente de fuga de subthreshold.
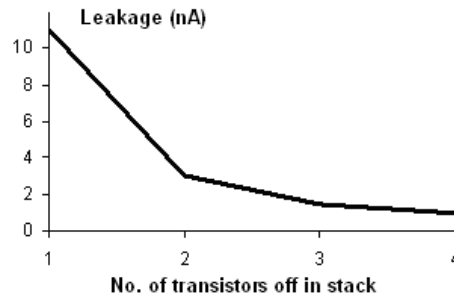

Figura 1: Corrente de fuga de subthreshold versus pilha com diferente número de transistores não conduzindo.

Outra técnica de redução da corrente de fuga quando o circuito esta em modo idle explora o conceito de "Power Gating". Nesta técnica, denominada de MTCMOS e ilustrada na Figura 2, um transistor (denominado "*switch*") é inserido entre as linhas de alimentação e a célula lógica propriamente dita. Quando o circuito esta operando normalmente, os transistores "*switch*" conduzem, fornecendo a corrente necessária para a operação do circuito. Quando o circuito entra em modo idle, os transistores "*switch*" deixam de conduzir, reduzindo a corrente de fuga por isolar as células lógicas das linhas de alimentação. Para reduzir ainda mais a corrente de fuga, os transistores "*switch*" podem ser dispositivos "*high V_{th}*" se os mesmos estiverem presentes no na tecnologia alvo.
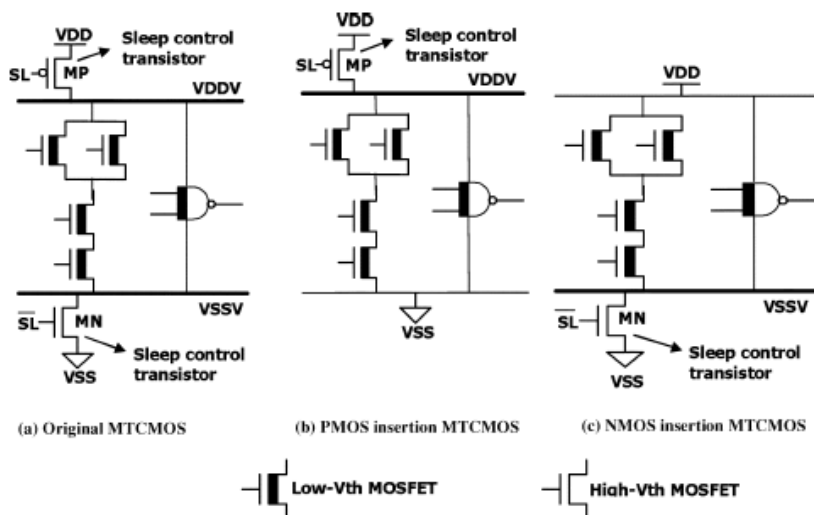

Figura 2: Esquemático de um circuito MTCMOS (ROY, 2003).

A última técnica de redução apresentada neste trabalho explora a influencia da polarização do substrato na tensão de threshold e conseqüentemente na corrente de fuga. Em circuitos lógicos, o substrato é normalmente polarizado em "*ground*" ou pela tensão de alimentação. Quando o substrato é polarizado reversamente, a tensão de threshold dos transistores aumenta e provocando uma redução na corrente de fuga de subthreshold. Esta polarização ocorre quando o circuito esta em modo idle. Uma variante que explora o mesmo conceito, implementa o circuito com transistores "*high V_{th}*", que possuem corrente de fuga e efeitos de canal curto reduzidos, mas também são mais lentos. Neste caso, a polarização do substrato ocorre quando o circuito esta em operação e ela é da forma direta (FBB – forward body biasing), provocando a redução da tensão de threshold dos transistores e acelerando o processamento.

Modelos analíticos e rápidos estimadores da corrente de fuga em células lógicas são indispensáveis no processo de concepção dos circuitos integrados. Estes estimadores fornecem custos a diferentes etapas do projeto sem que exista a necessidade da realização de custosas simulações elétricas. A seguir são apresentados modelos simplificados para as correntes de fuga de subthreshold e de tunelamento através do oxido de porta do transistor e um modo eficiente de estimar a corrente de fuga total de qualquer célula lógica independente da família lógica a qual pertence. A corrente de fuga através das junções pn polarizadas reversamente não são abordadas nesta análise, pois não possuem magnitude significativa nas tecnologias atuais e não estão presentes nos modelos elétricos atuais como o BSIM 4 utilizado neste trabalho. Contudo, podem ser facilmente agregadas visto que o método de estimativa que utiliza o conceito de soma das correntes associadas aos nodos das células.

As correntes de fuga em uma célula dependem diretamente da tensão nos terminais, tanto internos quanto externos, da mesma. A maioria dos estimadores das correntes de fuga em células considera as tensões nos terminais das mesmas para então fornecerem os valores das correntes de fuga. Os terminais externos das células são facilmente definidos a partir dos vetores de entrada (terminais de entrada) e de sua função lógica (terminal de saída). Aos terminais internos são normalmente atribuídos valores referentes às tensões de alimentação. Esta aproximação compromete a precisão da estimativa. A proposta deste trabalho faz uso da lei das correntes de Kirchoff, que diz que a soma das correntes entrando em um nodo é igual à soma das correntes deixando esse nodo. Dessa forma, todos os terminais internos de uma célula têm sua tensão definida baseado na premissa acima. As figuras 3, 4 e 5 exemplificam a metodologia descrita acima para definir a tensão no terminal V_x. A equação (4) é reflete a lei das correntes de kirchoff no terminal V_x. Resolvendo esta equação se obtém a tensão desconhecida. Após todas as tensões serem definidas, a corrente de fuga total de uma célula é definida como toda a corrente que sai da alimentação ou vai para o "ground". Esta é calculada utilizando as equações (5) e (6) e as tensões nos terminais.
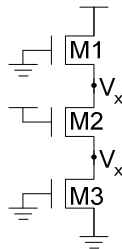

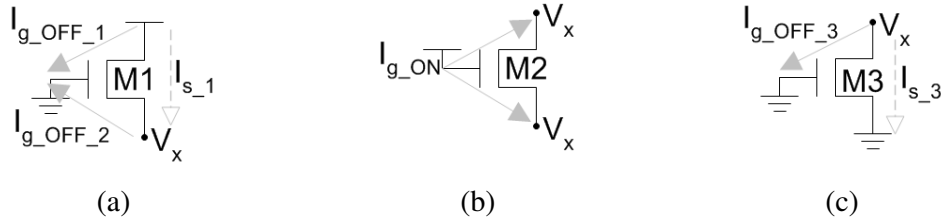Figura 3: Pilha de três transistores NMOS.

(a)　　　　　　　　(b)　　　　　　　　(c)

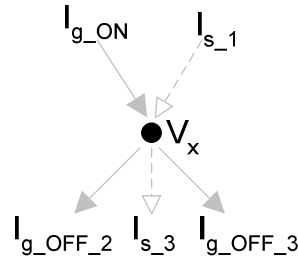Figura 4: Correntes de fuga associadas a cada transistor da Figura 3.

Figura 5: Correntes associadas ao nodo V_x da Figura 3.

$$I_{S\_1} + I_{g\_ON} = I_{S\_3} + I_{g\_OFF\_2} + I_{g\_OFF\_3} \qquad (4)$$

Os termos da equação (4) podem ser expandidos a partir das equações (5) e (6), onde a equação (5) representa a tradicional equação da corrente de fuga de subthreshold apresentada pelo modelo BSIM3 e a equação (6) é uma simplificação da equação da corrente de tunelamento através da porta do transistor. A Figura 6 ilustra um comparativo entre a equação (6) e a corrente estimada pelo HSPICE, apresentando uma grande precisão. Todos os parâmetros das equações (5) e (6) são extraídos previamente através de simulações elétricas.

$$I_S = I_0 W e^{\frac{V_{gs} - (V_{t0} - \eta V_{ds} - \gamma V_{bs})}{n V_T}} \left[1 - e^{\frac{-V_{ds}}{V_T}} \right] \qquad (5)$$

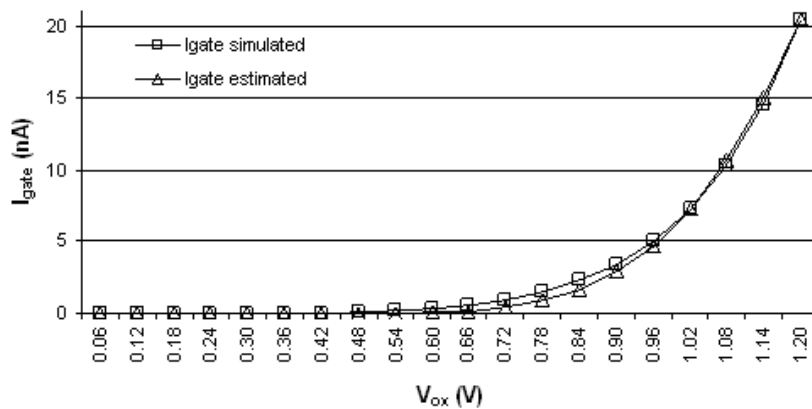$$I_{gate} = I_{gate\_0}.W.e^{\frac{-K}{|V_{ox}|}} \qquad (6)$$

Figura 6: Corrente de fuga de tunelamento através do oxido.

O método proposto foi validado para determinadas configurações de redes de transistores do plano "*pull-down*". A mesma análise é válida para transistores PMOS do plano "*pull-up*". A Figura 7 apresenta uma variedade de redes, onde bolas brancas representam transistores NMOS cortados e bolas pretas representam transistores NMOS conduzindo. A tabela 1 apresenta os valores estimados pelo método descrito acima e compara com os valores simulados eletricamente, apresentando uma boa correlação.
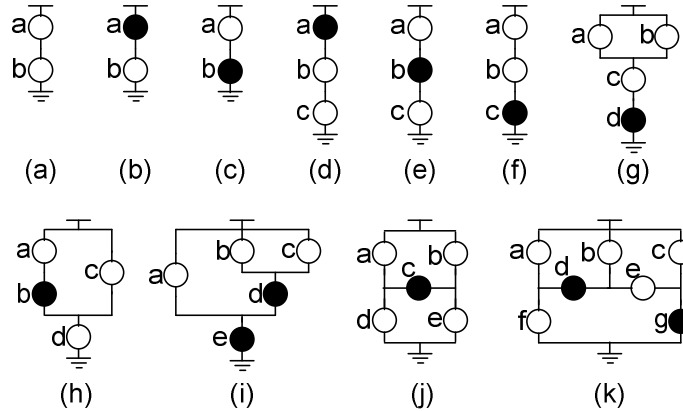

Figura 7: Planos "*pull-down*".

Tabela 1: Corrente de fuga total para redes da figura 7.

Rede	Hspice (nA)	Método Proposto (nA)	Dif (%)
(a)	5.37	5.35	0.37
(b)	15.21	15.26	0.33
(c)	42.12	42.11	0.02
(d)	4.23	4.02	4.96
(e)	7.12	7.27	2.11
(f)	25.73	25.71	0.08
(g)	27.47	27.65	0.66
(h)	8.45	8.42	0.36
(i)	105.83	105.97	0.13
(j)	12.88	13.24	2.80
(k)	55.00	55.35	0.64

Este trabalho revisou os principais mecanismos responsáveis pelas correntes de fuga em transistores MOS e as principais técnicas de redução em nível de circuito. Um novo método de estimativa foi proposto apresentando uma excelente precisão quando comparado com simulações elétricas.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Administração
Baixar livros de Agronomia
Baixar livros de Arquitetura
Baixar livros de Artes
Baixar livros de Astronomia
Baixar livros de Biologia Geral
Baixar livros de Ciência da Computação
Baixar livros de Ciência da Informação
Baixar livros de Ciência Política
Baixar livros de Ciências da Saúde
Baixar livros de Comunicação
Baixar livros do Conselho Nacional de Educação - CNE
Baixar livros de Defesa civil
Baixar livros de Direito
Baixar livros de Direitos humanos
Baixar livros de Economia
Baixar livros de Economia Doméstica
Baixar livros de Educação
Baixar livros de Educação - Trânsito
Baixar livros de Educação Física
Baixar livros de Engenharia Aeroespacial
Baixar livros de Farmácia
Baixar livros de Filosofia
Baixar livros de Física
Baixar livros de Geociências
Baixar livros de Geografia
Baixar livros de História
Baixar livros de Línguas