The Scaling Of Submicron Cmos Devices.

Order Number 9109586

The scaling of submicron CMOS devices

Tran, Conn-Luân, Ph.D.

Lehigh University, 1990

U·M·I

300 N. Zeeb Rd.
Ann Arbor, MI 48106

The Scaling of Submicron CMOS Devices

by
Conn-Luân Tran

A Dissertation
Presented to the Graduate Committee
of Lehigh University
in Candidacy for the Degree of
Doctor of Philosophy

in

Electrical Engineering
Lehigh University
1990

Certificate of Approval

This dissertation is approved and recommended for acceptance in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

September 26, 1990
Date

Prof. M. H. White
Dissertation Advisor

Dissertation Committee

Prof. M. H. White (Chairman)

Prof. M. K. Hatalis

Dr. D. S. Yaney

Prof. D. R. Young

ACKNOWLEDGEMENTS

I wish to express my grateful thanks to Prof. Marvin H. White, for his guidance, suggestions and tolerance during the course of this research. Special thanks are due to the committee members, Prof. Miltiadis Hatalis, Dr. Dave Yaney and Prof. Donald Young for recommendation to focus on the dissertation topics. The late Prof. Frank Feigl, who served as a committee member during my general examination, is respectfully remembered. The encouragement and support from the management at AT&T Bell Laboratories, Drs. Pete Panousis, Dinesh Mehta, Dan McGillis, Morgan Thoma, Jorge Agraz-Guerena are deeply appreciated. A special acknowledgement is due to Dr. Robert Ashton, my long time officemate, for his encouragement, allowing me to use of his several useful measurement subroutines, and timely proof-reading chapters of this dissertation. Collaboration with Technology CAD colleagues in device simulation, B. Meinerzhagen, J. Prendergast, W. Lui, and H. Dirks are greatly appreciated. Technical discussions, encouragement and sometimes a push by my colleagues and friends at Lehigh University and AT&T are fondly remembered. Among them, Cris Lawrence, Nancy Minor, Jane Swiderski, Felicia Herring, Daniel Chesire, Mike Kelly, Richard Booth, Tom Krutsick, Richard Siergiej. Members of the BiCMOS Groups, Tom Ham, Sue Vitkavage, John Osenbach, and Don Dennis gave supporting boost during the final weeks of this writing. Thanks are due to my brothers, Luan Q. and Luong, who did the figures during the various phases of the manuscript preparation. My dear friend, Dr. BichVan Phan, whose daily encouragement from a long distance is heartedly appreciated. Support from my family members and loved ones has been the main thrust for my endeavor, and to that, this work is especially dedicated to my parents and grandparents.

I am greatly appreciate to the AT&T Bell Laboratories Graduate Study Program for the financial support of this doctoral program.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

List of Symbols

Symbols	Description	Unit
A	device Area	cm^2
α	depletion width coefficient	$cmV^{-1/2}$
α_{it}	emperical surface reduction factor	cm^2
β_o	MOS Transistor gain	A V^{-2}
C_{ox}	gate oxide capacitance (per unit area)	F/cm^2
C_{it}	interface trap capacitance	F/cm^2
C_n	n-ch inversion layer capacitance	F/cm^2
C_p	p-ch inversion layer capacitance	F/cm^2
C_D	depletion layer capacitance	F/cm^2
C_s	semiconductor layer capacitance	F/cm^2
C_G	MOS gate capacitance	F/cm^2
δ	drain effect term on threshold voltage	
D_{it}	interface state density	$eV^{-1}cm^{-2}$
Δg_m	change in transconductance g_m	S
δL	total change in L_m due to lateral diffusion and gate etching	cm
ΔL	pinch-off length in saturation	cm
ΔL_d	damaged length from the drain due to hot carrier	cm
ΔV_t	change in threshold voltage V_t	V
ε_o	vacuum dielectric constant	8.85E-14 F/cm
ε_{ox}	oxide dielectric constant	$3.9\varepsilon_o$
ε_s	silicon dielectric constant	$11.8\varepsilon_o$
E_c	silicon conduction band-edge energy	eV
E_t	interface trap energy in the bandgap	eV
E_v	silicon valence band-edge energy	eV
E_i	silicon intrinsic Fermi level in the bulk	eV
E_{is}	silicon intrinsic Fermi level at the surface	eV
E_{Fn}	Fermi level of electrons	eV

E_{Fp}	Fermi level of holes	eV
f_t	interface trap occupancy function	
γ	band bending over $2\phi_F$	V
g_m	MOS transistor transconductance	S
g_{mo}	initial transconductance	S
I_B, I_{sub}	bulk or substrate current	A
I_D, I_{DS}	drain current	A
I_G	gate current	A
I_S	source current	A
I_{off}	drain current at V_{GS}=0 and V_{DS} = max. op. voltage	A
I_{on}	drain current at $V_{GS} = V_{DS}$ = max. op. voltage	A
λ	body effect coefficient	V^{-1}
L	channel length between s/d metallurgical junctions	cm
κ_{ox}	relative dielectric constant of oxide	3.9
κ_s	relative dielectric constant of silicon	11.8
L_e, L_{eff}	electrical or effective channel length	cm
L_g	gate electrode length	cm
L_m	coded (designed mask) gate length	cm
m	normalized depletion capacitance to C_{ox}	
n	normalized total capacitance to oxide capacitance	
n'	exponent of D_{it} vs stress time	
n_c	density of electrons in the conduction band	cm^{-3}
n_i	intrinsic electron concentration	cm^{-3}
n_s	electron surface concentration	cm^{-3}
N_A	acceptor doping density (p-type)	cm^{-3}
N_D	donnor doping density (n-type)	cm^{-3}
N_{ch}	channel doping density	cm^{-3}
N_B	bulk concentration, (acceptor or donor)	cm^{-3}
ϕ	quasi Fermi level	V
ϕ_{Fn}	Fermi potential from midgap to E_{Fn}	V
ϕ_{Fp}	Fermi potential from midgap to E_{Fp}	V
ϕ_n	quasi Fermi level for electrons	V
ϕ_p	quasi Fermi level for holes	V

ϕ_s	surface potential, referenced to the bulk E_i	V
ψ	electrostatic potential	V
ψ_s	surface potential, referenced to the source	V
p_s	hole surface concentration	cm^{-3}
p_v	density of holes in the valence band	cm^{-3}
q	electron charge	C
Q_B	bulk silicon charge	C/cm^2
Q_c	junction capacitance charge integrated over voltage	C/cm^2
Q_D	depletion layer charge	C/cm^2
Q_f	oxide fixed charge	C/cm^2
Q_{it}	interface trap charge	C/cm^2
Q_{ot}	oxide bulk trapped charge	C/cm^2
Q_n	n-ch inversion layer charge	C/cm^2
Q_p	p-ch inversion layer charge	C/cm^2
Q_t	total of oxide and interface trapped charge	C/cm^2
Q_s	silicon surface charge	C/cm^2
R_{mn}	normalized current ratio to calculate m,n	
R_S, R_D	source, drain series resistance	Ω
s	slope for device lifetime extrapolation	
τ	aging lifetime for 10% g_m degradation	minutes
θ_s	high vertical field surface mobility reduction factor	V^{-1}
t_{ox}	gate oxide thickness	cm
T	absolute temperature	°K
$\mu_{n,eff}$	effective n-channel mobility	$cm^2V^{-1}s^{-1}$
$\mu_{p,eff}$	effective p-channel mobility	$cm^2V^{-1}s^{-1}$
μ_o	low field channel mobility	$cm^2V^{-1}s^{-1}$
v_{th}	thermal velocity of carriers	$cm \cdot s^{-1}$
V_D	drain voltage	V
V_G	gate voltage	V
V_S	source voltage	V
V_{BS}	bulk to source voltage	V
V_{DD}	operating voltage	V
V_{DS}	drain to source voltage	V

V_{GS}	gate to source voltage	V
V_{FB}	flat band voltage	V
V_{tn}	n-ch threshold voltage	V
V_{tp}	p-ch threshold voltage	V
V_T	thermal voltage (kT/q)	V
W, W_{eff}	effective channel width	cm
x_{jn}	n-ch junction depth	cm
x_{jp}	p-ch junction depth	cm

ABSTRACT

In the continuing scaling of the CMOS (Complementary Metal Oxide Semiconductor) transistors to achieve higher circuit performance, density and functional complexity, the device physicists and process technologists are faced with the challenges of the high field phenomena in these devices due to the shear reduction in device dimensions, both laterally and vertically. The shorter channel length in an MOS device causes the device to be susceptible to subthreshold leakage, or in the extreme case, punch-through, if the device structure is not designed properly by increasing the doping concentration in the channel to suppress the bulk and surface conduction paths. At the same time, the source and drain junction depths must be reduced. The higher concentration of the doping impurity in the channel will lead to mobility degradation in the inversion layer, lower junction breakdown, and increased hot carrier generation. The hot carrier generation in MOSFET transistors can cause degradation in performance due to interface states built-up at the drain and a reduction in transconductance and a shift in threshold voltage. The trapped charge in the gate oxide can lead to permanent oxide breakdown and drain junction leakage.

This dissertation aims at the detailed treatments of device and process design for submicron CMOS structures, with a strong emphasis on hot-carrier effects. The process and device modeling using numerical simulators is used to analyze and optimize device structures before the fabrication process. The devices are fabricated with a state-of-the-art processing technology. Experimental results from submicron devices are presented to verify the analysis and modeling. The gate oxide thickness of the devices is in the range of 125 Å to 210 Å, and the effective channel length is in the range of 0.4 to 1.0 μm. A comprehensive device and process characterization methodology is presented. New phenomena in small devices, such as reverse subthreshold swing and reverse short channel effects, are reported. An experimental procedure to determine the optimum off current with respect to channel length and leakage requirements at different temperatures is presented.

For hot carrier considerations, the NMOS LDD drain was designed with an optimized dose and a spacer width, and verified with substrate current and aging results. The work on the hot carrier generation process in the MOS device leads to an improved model to predict substrate current as a function of the applied drain voltage. Therefore, the lifetime prediction is improved for a MOS device, which operates at a lower drain voltage than the accelerated stress voltage. The physical understanding of the interface state built-up during device aging is analyzed using the method, based on the physics of subthreshold conduction and the extension of drain depletion region, to extract the spatial distribution of interface state density along the channel from the drain end. The distribution of interface states as a function of time and space uncovers the dynamics of charges built-up during device stressing. This method can be implemented to routinely monitor device aging during development and in manufacturing. Predicting device lifetime using substrate current extrapolation for different channel length devices can lead to errors. The device lifetime prediction using substrate current for the same channel length devices is recommended for a more consistent in channel electric field of the devices. A new model to relate peak substrate current to the drain voltage can also be used to predict the device lifetime at operating voltage, using an inverse of the square-root of the drain voltage. The substrate current measured at operating voltage is also used to detect process variation conditions in etching the gate sidewall spacer in manufacturing. However, the ultimate objective of the study of device degradation is to apply these predictions to actual circuits. In this light, a simple analysis of substrate current generated during switching in an CMOS inverter is used to predict the lifetime.

The uniqueness of this work is in the treatments of device design, process integration, device reliability and circuit design techniques as an integral system, in which, each area is interrelated to each other. A combined solution from the above specialized areas for the high field effects in small devices should be explored for submicron technologies. The work in this dissertation also serves to bridge the gap between device physics theory and its use in the design and characterization of the CMOS devices in the semiconductor industry.

Chapter 1

INTRODUCTION

With the rapid progress in the Silicon VLSI and ULSI (Very and Ultra Large Scale Integration) technology and its applications in the past 3 decades, it is instructive to review several key breakthroughs during this period in the areas of device invention, technology advancements, and the circuit and system applications. This chapter starts with a review of the historical events which have led to the present progress in silicon integrated circuits. Special emphasis is devoted to modern CMOS technology since it is the driving force behind device scaling to achieve higher level of circuit performance, density, and more complex functions. However, the emerging BiCMOS technology in which bipolar and CMOS devices are integrated on the same silicon wafer promises to be the trend for the future.

1.1 HISTORICAL REVIEW

Let us first review the major scientific inventions in the areas of semiconductor device physics, material and processing techniques, and apply this silicon technology to the development of complex integrated circuits. We shall divide the semiconductor revolution into three periods: The first period spanned the early discoveries and inventions in device structures, basic material processes, and the fundamental device physics theory. The second period began with the invention of the integrated circuit in which active devices (transistors) and passive components (resistors and capacitors) are integrated on a monolithic semiconductor substrate. It was in this period that the thrust for miniaturization or scaling to smaller devices and interconnect dimensions became prominent. This period included many processing breakthroughs and characterization techniques that overcame earlier difficulty in device fabrication. The third period focused on applications, and started with the development of the 1-transistor DRAM (Dynamic Random Access Memory) integrated circuits, followed by the 256-bit SRAM (Static RAM) and a 4-bit microprocessor at Intel Corp.. We are now over 2 decades into this period, and the thrust for scaling feature sizes is continuing at a very rapid pace. This trend in silicon technology will take us well into the information age of the 21^{th} century, and may only be supplanted by fields such as

bioelectronics, artificial intelligence, and neural networks.

Let's step back and see how all this began.

1.1.1 Major Inventions And Breakthroughs in Device Structures and Theory

The major inventions that led to the present state-of-the-art in MOS technology can be traced back to 1928 with the patent disclosure by Lilienfeld[1] of the field-effect modulation of the conductance of Copper Sulfide (CuS) by an Al plate on top of an Al_2O_3 insulator. This is now considered the first field effect device or Field Effect Transistor (FET). The idea was not implemented until 1960, when D. Kahng and M. Atalla of Bell Telephone Laboratories (BTL) demonstrated the enhancement-mode inversion-layer Metal Oxide Semiconductor Field Effect Transistor (MOSFET) using SiO_2 as the insulator and Si as the semiconductor material.[2] Between these two events, several other important inventions took place in areas of device physics, processing technology and materials that all contributed to the MOS and bipolar integrated circuit evolution.

1.1.1.1 The Bipolar Transistor

The most famous invention that revolutionized the electronic industry, the modern style of living, and the information age, is the discovery of the transistor (transfer-resistor) at Bell Laboratories in December 1947 by Bardeen and Brattain.[3] The first transistor was in the form of a point-contact. This structure was built from a plate of n-type germanium and two line-contacts of gold supported on a mylar wedge. The n-type Ge material is referred to as "base" in this structure, and the emitter and collector are referred to as the emitting and collecting functions of their respective terminals. According to W. Shockley,[4] the birth of the transistor has been accepted as the day before Christmas Eve (December 23) of 1947, within the Bell Labs research community, although the rest of the world did not know until the announcement and first public demonstration of the discovery on June 30, 1948. Shockley felt some frustration by not being one of the inventors despite the efforts he had started eight years earlier. This motivated Shockley to produce more than 90 patents related to the transistor. Among these works was the concept of the junction transistor recorded on January 23, 1948 in Shockley's notebook, for which the patent was filed in June, 1948.[5] 1948 was an important year in device physics and

structure for the junction transistor. Shockley introduced many concepts that are still being used. These concepts and physical understanding include IMREF (or quasi-Fermi level), heterojunctions with wide energy band gap to increase emitter efficiency, and multi-layer structures for modulation. These last two important concepts are now employed in GaAs technology for HBT (Heterojunction Bipolar Transistor) and HEMT (High Electron Mobility Transistor) devices. The idea of minority carrier injection was a major milestone in understanding the physics of transistor action. The theory of p-n junctions and transistors was published in a classic paper in 1949.[6] During the early part of the 1950's, when the discrete transistor technology was licensed to the industry, more device concepts and theory were developed at Bell Labs. In 1952, Shockley, Read and Hall developed the recombination theory which greatly helped in understanding the generation-recombination processes of electron-hole pairs in semiconductor bulk and surface. This theory has had a direct effect in understanding the defect centers and the processes which control the reverse junction leakage current in p-n junction, bipolar and MOS devices.

1.1.1.2 The Field Effect Transistor Concepts

With the advancement in theory and technology know-how during the early days of transistor, the 1953 JFET (Junction FET) was built at Bell Laboratories by Dacey and Ross.[7] The JFET operates on the principle of bulk conduction modulation by varying the depletion width of the 2 junctions biased in the reverse mode. The application of this device is limited to operational ampliers and other analog circuits. In this same year, Brown[8] at Bell Labs studied the inversion layer, or channel, which conducts the leakage current in the base surface of a germanium *npn* transistor. This work served as the basic foundation for understanding the MOS device operation. It was noted that the conductance between the emitter and collector of an n-p-n bipolar device was much larger than expected with the floating base. Brown[8] proposed a model and experimentally verified the existence of a channel layer bridging the 2 p-n junctions of the emitter and collector. More important to the inversion layer observation, Ross[9] in 1955 proposed that the inversion layer can be induced electrostatically by a control electrode placed in the vicinity of the channel region. Ferroelectric material was used to separate the electrode and

the floating base. Ross further proposed the channel could be controllably turned on and off as the potential on the gate electrode is applied or removed. An impedance meter connected between collector (or drain) and emitter (or source) can monitor whether the channel is formed. This is the first form of Read Only Memory (ROM) cell. The use of a ferroelectric as a gate dielectric material, which proved impractical at the time, was not pursued.

In 1959, Atalla[10] suggested that thermally grown SiO_2 on Si single crystal substrate be used as the gate insulator. This concept led to the first successful fabrication of modern Metal Oxide Semiconductor Field Effect Transistor (MOSFET) as shown in cross-section in Fig. 1-1.[11] It was this work and the outgrowth of the SiO_2 works headed by Atalla [12] at Bell Labs that paved the way for further studies in silicon dioxide for MOS device applications.

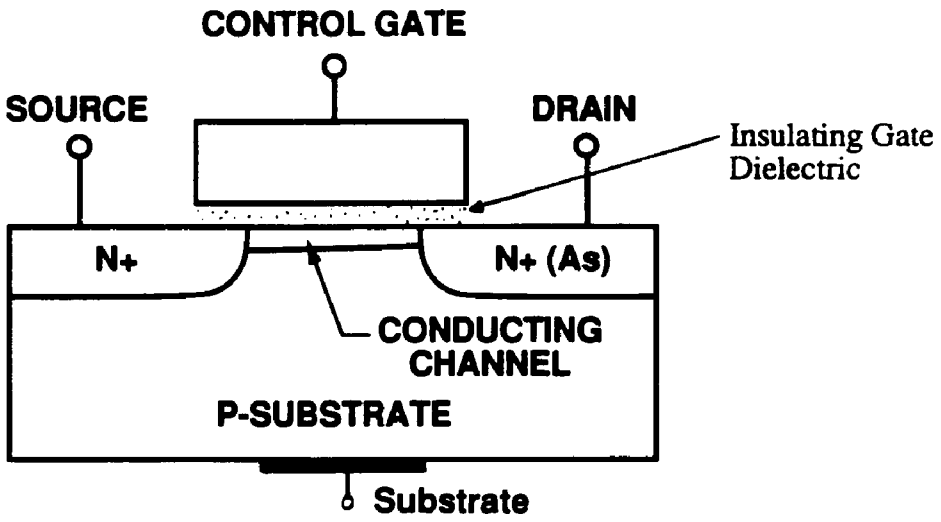


Figure 1-1. Cross-Section of a MOS Transistor

1.1.1.3 Integrated Circuits Inventions

All of the inventions and developments described above took place at Bell Laboratories. However, the IC industry did not take off until the late 1950's with two major inventions that took place at Texas Instruments by Kilby (1958)[13] and Noyce at Fairchild Semiconductor (1959).[14] (Incidentally, Fairchild was spun off by Noyce, Moore and others from Shockley Transistor, Inc. founded by Shockley after his leaving the Bell Labs in 1957).

The initial motivation was to integrate active and passive discrete components together on a ceramic substrate when Kilby was with the CentraLab.[13] It was not until Kilby joined Texas Instruments in 1958 that the idea of integrating components on a single semiconductor material was invented. In Kilby's invention, the passive components were realized using the doped portion of semiconductor bulk for the resistors and p-n junctions for capacitors. The idea was first demonstrated by assembling the separate silicon components together. Subsequently, a phase-shift oscillator was integrated on a 0.4-in-square germanium wafer. A Digital flip-flop was also built using the same techniques. One year later, a planar silicon bipolar transistor technology was developed by Noyce at Fairchild, using similar integrated techniques. Although the inventor of integrated circuits often is a subject for controversial debate, both Kilby and Noyce are generally considered the forefathers of ICs for their discoveries.

To summarize this period, the three major inventions were the bipolar junction transistor, the demonstration of the silicon MOS transistor, and the integrated circuits. Although these are considered the major milestones, other basic processing technologies were developed along the way. Without these developments, the semiconductor technology would not have been in its present state.

1.1.2 Technological Breakthroughs

The late 1940's and early 1950's was also a period of advancement in crystal growth and doping techniques, techniques needed to realize practical transistors. In 1950, a good p-n junction was formed by Teal, Sparks, and Buehler[15] by changing the doping of a melt as a crystal was grown. In 1952, Teal, Theuerer and Pfann developed the zone melting and crystal refining technique,[16] [17] which was essential for growing a purified and low intrinsic defect density material. Along with his other inventions, Shockley also pioneered the idea of using ion implantation to selectively dope a region of a semiconductor wafer.[18] This technique has become the standard doping method for fabricating VLSI circuits.

As mentioned earlier, the efforts in silicon dioxide growth led to the fabrication of the MOSFET transistor. The use of silicon dioxide as a diffusion mask by Frosch and Derrick,[19] in 1957 was also considered a technological breakthrough. This technique has been used to

selectively dope a silicon area by a patterned oxide layer, which blocks the dopant where it is not needed and allows the dopant to enter areas in which a diffused dopant is desired. The Epitaxial Layer used for bipolar junction transistor was proposed in 1959 by Theuerer.[20] The epitaxial layer, used as part of the collector, is grown on a patterned highly doped buried layer.[21] The buried layer reduces the collector series resistance, improving the transistor speed.

In order to understand the property of the Si-SiO_2 interface which was very difficult to control in the early days of MOS device fabrication, Terman [22] at Stanford Uninversity developed the Capacitance-Voltage (C-V) Technique in 1961. (This work was suggested and guided by J. Moll.)[23] This method has aided in the understanding of the charges present in a MOS system. This technique, and its derivatives, together with silicon oxidation techniques[24] have resulted in the Si-SiO_2 system being the most extensively studied interface in solid-state science. Another problem associated with silicon dioxide is contamination with sodium ions which drift in the oxide bulk and cause a shift in the flat-band voltage of an MOS capacitor and translates into a threshold voltage shift in the MOS transistor. This problem was studied by Snow et al.[25] at Fairchild. Kerr[26] at IBM in 1964 used a phosphorus-silicate-glass to stabilize the sodium drift in the oxide and to getter heavy impurities in silicon material, reducing junction leakage.

Kooi[27] at Philips contributed to the understanding of interface traps and most importantly the Local Oxidation (LOCOS) technique for device isolation, in which a layer of silicon nitride is used as an oxidation mask during the growth of field oxide. The invention of the poly-silicon gate by Kerwin et al [28] at Bell Labs led to the self-aligned gate-to-source-drain MOSFET structure. This structure reduces the overlap capacitance and avoids the stringent alignment requirements of an aluminum gate MOS transistor. The LOCOS isolation technique and polysilicon gate structures have remained the mainstream technology for fabricating MOS devices.

On the technology integration side, in 1963 Wanlass and Sah[29] of Fairchild Semiconductor were the first who suggested the use of n and p channel MOSFET devices in low power logic circuits such as inverter, NOR gate, and flip-flop. However, the NMOS and PMOS

devices were fabricated on two different types (p and n respectively) of silicon wafers by a planar diffusion process. White and Cricchi were the first who successfully fabricated CMOS devices on a single substrate epitaxial wafer.[30] Kahng and Sze[31] in 1967 proposed MNOS (Metal Nitride Oxide Semiconductor) device structures that can change the threshold voltage of an MOS device upon applying high enough potential onto the gate to induce charge injection. This structure has led to the field of non-volatile memory, presently known as EPROM (Electrically Programmable Read Only Memory), and EEPROM (Electrically Erasable PROM). The present mainstream applications in integrated circuits stemmed from the 2 circuits developments in late 1960's, which will be discussed in the next section.

1.1.3 Circuit and System Applications Driven Technology

From the invention of integrated circuits until late 1960's, most applications focused on linear applications such as operational amplifiers, oscillators, comparators. During this period, digital applications included logic gates, flip-flops, counters, decoders , adders, etc.. This period constitutes an era known as SSI (Small Scale Integration) and MSI (Medium Scale Integration), when the transistor count on a given chip was in the range of tens to hundreds.

The two circuit concepts that fueled the drive for device scaling are the 1-Transistor DRAM (Dynamic Random Access Memory) cell proposed by Dennard[32] of IBM, and the 6-Transistor (or 4-T with resistor load) SRAM cell. The DRAM cell is comprised of one MOS transistor serving as an access device to a storage capacitor used as a memory element. This basic circuit configuration is a foundation for the modern day high density DRAM chip. The LSI era dawned with products like the 16K DRAM, the 8-bit microprocessors and gate array where the transistor count was in the range of few thousand to 100 thousand. Since the early days of integrated circuits application, the growth of the semiconductor industry has been astronomical. From the Small Scale level of integration (SSI) of TTL logic gates with dozens of transistors and features sizes greater than 10 μm in the early 1960's, [33] the level of integration has reached the level at this writing of integrating more than 36 million device components on a 16 megabit DRAM chip with minimum feature sizes of 0.5 micrometer.[34] An experimental 64Mb DRAM chip designed with 0.3μm feature size using e-beam lithography, operating at 1.5V supply voltage has been

successfully demonstrated.[35]

The progress of down scaling in device geometry and minimum feature size can be monitored by the introduction of high density DRAM's and SRAM's. Memory is usually the first product to benefit from device scaling. It is estimated that DRAM density quadruples every 2.5 years (Figure 1-2).

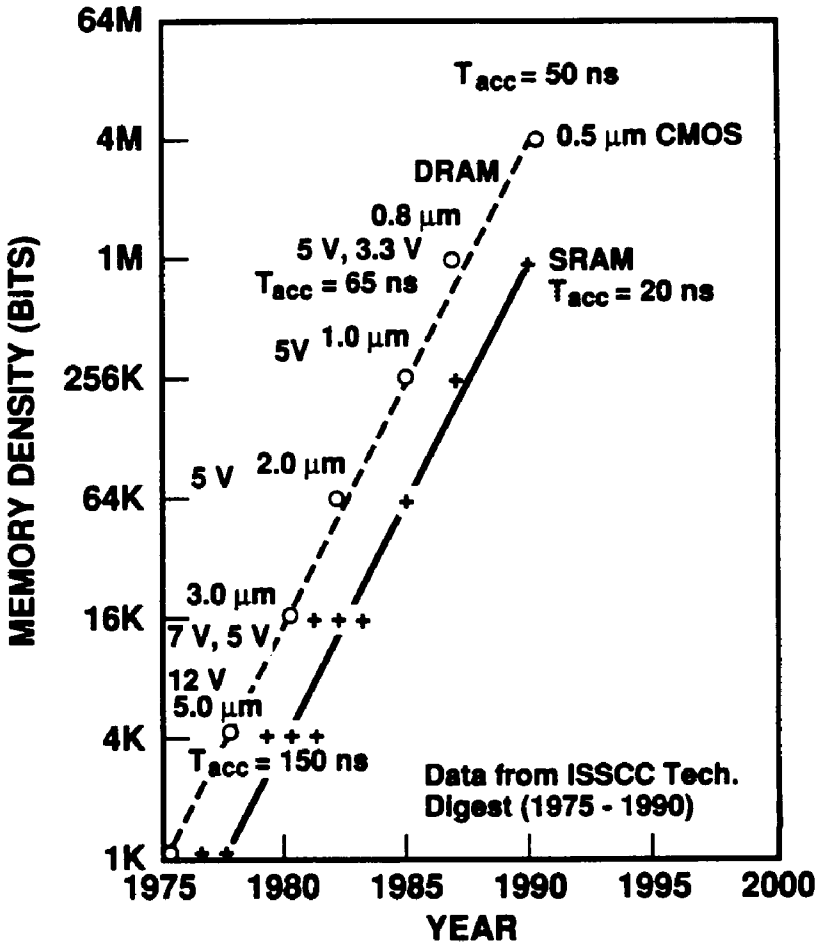


Figure 1-2. DRAM and SRAM density vs the year of introduction.

The trend is the same for Static Random Access Memory (SRAM) but the memory capacity of a state of the art SRAM is typically one fourth of the DRAM capacity. Memory products are considered technology drivers because circuit regularity and ease of failure mode analysis help the technologist 'fine-tune' the fabrication process.

Other applications, such as microprocessors and custom digital circuits, typically follow the lead of memory technology, even though DRAM technology is not directly compatible with logic circuit applications. Power delay products of logic gates have dropped by more than 6 decades to reach 10^{-1} pJ in 1990 as a result of device scaling as shown in Figure 1-3. The design of logic custom integrated circuits also drives the active software development for CAD (Computer Aided Design) tools that help to integrate millions of transistors on a logic chip.[36]

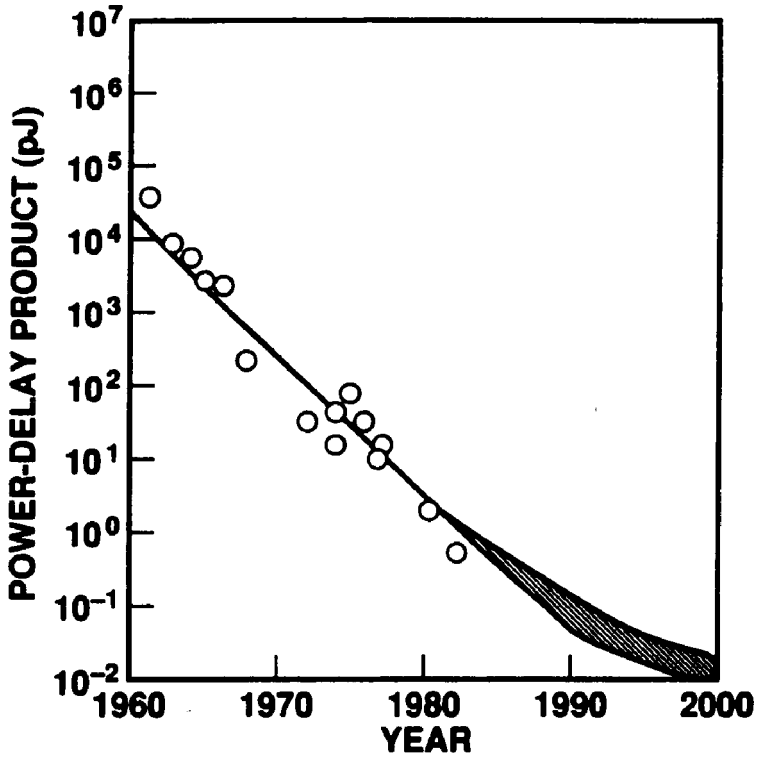


Figure 1-3. Power-Delay Product per inverter gate versus year.

In all of these achievements, the MOS technology is the driving force to dominate the high density, high functionality circuits in both commercial and special purpose applications. During the period from 1959 to 1975 the minimum feature size decreased at a rate of approximately 11% a year resulting in a halving every six years.[37] This rate of reduction has continued, reaching a production feature size of 0.8 μm in 1990 in state-of-the-art VLSI products (Fig. 1-4). Figure 1-4 shows the component count per chip, for DRAM MOS memory, is growing at the same time as the feature size is shrunk. This trend does not seem to slow as predicted previously by G.

Moore.[37] However, the delay between matured process development to manufacture tends to be longer due to the multitude of technology transfer problems.

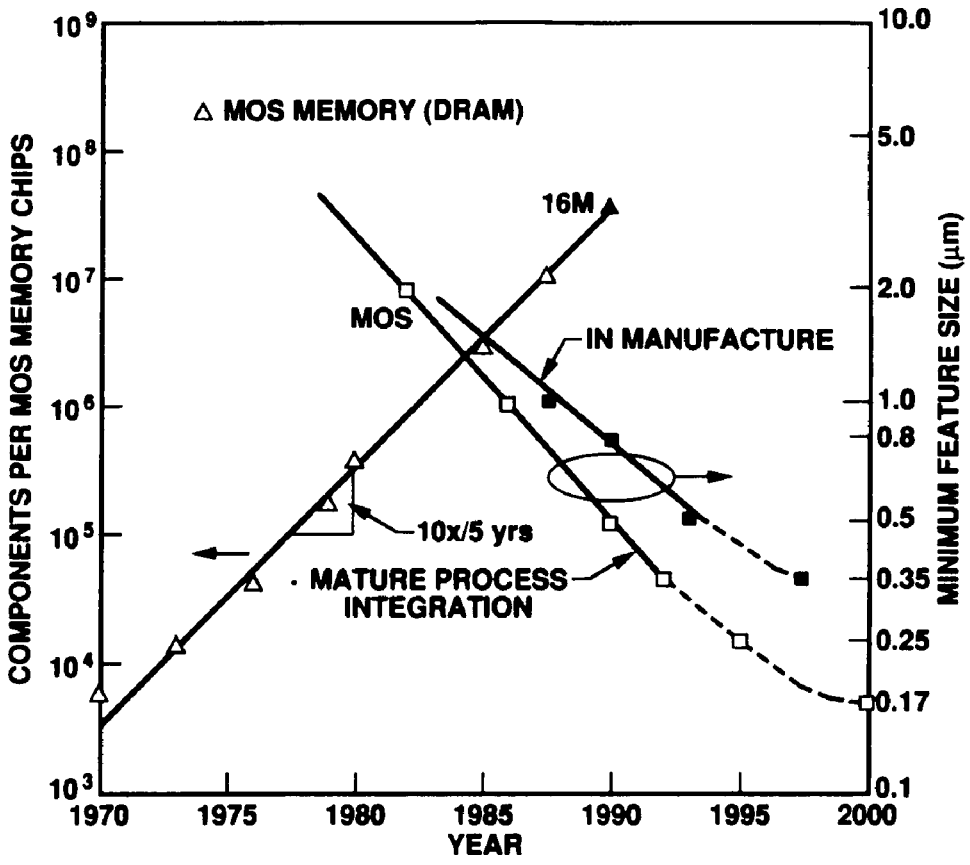


Figure 1-4. MOS memory component growth and feature size reduction vs calendar year

Continuing scaling to submicron features will be faced with several obstacles: process difficulties in fabricating submicron geometries, noticeably lithography and etching; the physical limitations of devices and materials; and the economic factors in capital investments to fabricate the products.

1.2 DEVELOPMENTS IN MOS LSI

As indicated above, the drive for miniaturization is derived from the circuit and system requirements to ever increasing the performance. To this end, device and interconnect scaling is inevitable. The art of designing a full integrated circuit technology is known as process

integration, whereby many individual process steps are integrated in a logical sequence to fabricate the desired device structures and circuit interconnects. CMOS technology is a combination of PMOS and NMOS technologies. The following section reviews the development of PMOS and NMOS processes and the resulting CMOS technology.

1.2.1 PMOS and NMOS Technologies

In the late 1960's, PMOS was the dominant technology since control of threshold voltage was easier and the surface inversion due to positive fixed oxide charge was not a problem for device isolation. PMOS devices, however, suffer from low current drive, because the mobility of the holes which carry the current is typically one third of the mobility of electrons which serve as the current carrier in NMOS devices.

The NMOS transistor became popular as the mainstream technology from mid 1970's until the first half of 1980's. Several technology problems were overcome, mainly the ability to control the fixed charge at the Si-SiO_2 interface of an n-channel device, the use of ion implantation to obtain the desirable threshold voltages, (both enhancement and depletion modes), and to increase the field threshold voltage for device isolation. NMOS technology remained the workhorse for a ten year period, due to its virtues of high speed and simple processing which made it a low cost technology. Device scaling laws and problems associated with high fields and hot carriers were learned during this period. The high performance products built with this technology include a 32-bit Floating-Point Digital Signal Processor (DSP),[38] the HP microprocessor[39] and most of the 256K DRAM memories. The most serious drawback of NMOS technology is the power consumption. The DSP, and HP microprocessor chips dissipated 3 W and 7W of power, respectively. The device junction temperature at the operating condition was as high as 150°C. This was a major problem for reliability, ambient temperature control in electronic systems, and packaging considerations. The device scaling for NMOS went to a design rule of 1.5μm, with some companies managing to go down to about 1.0μm.

1.2.2 CMOS Technology

CMOS technology was also developed in the 1960's at Fairchild and in the early 1970's RCA was the first company to fabricate low power CMOS logic families in SSI and MSI

integration levels. It was not until the early 1980's that CMOS emerged from an aerospace technology to become a mainstream technology due to the drawbacks of NMOS devices described in the previous section. CMOS has changed a great deal in structure and process sequence depending on the application. The following is a review of the development in CMOS devices.

CMOS technology has become a major driving force for VLSI integrated circuits. The advantages for converting to CMOS stem from the availability of n and p channel complementary devices, thus, easing circuit design, and eliminating bootstrap circuitry often found in NMOS. In digital circuits, only one device type conducts for a given state, hence the dc power dissipation is virtually eliminated in the standby mode. CMOS logic circuits are mainly static in nature, therefore the time needed for precharge in dynamic NMOS circuits is eliminated and the design effort is reduced. A CMOS circuit can achieve higher speed than its NMOS counterpart because the operating temperature of CMOS IC's is significantly lower, hence the current drive penalty due to higher temperature, especially in n-channel devices, is less severe.

CMOS technology begins with either a N-well or P-well structure depending on the application or prior experience. Parrillo et al.[40] were the first to introduce the twin-tub (or twin-well, as will be used interchangeably throughout the dissertation) approach. CMOS circuits built on bulk silicon or epitaxial Si on Si-substrate starting material are often referred to as bulk CMOS. CMOS circuits built on Silicon On Saphire (SOS) or, more recently, Silicon On Insulator (SOI) substrate constitute a different class of CMOS technology. The new trends in CMOS structures include BiCMOS (Bipolar and CMOS) and SOI (Silicon on Insulator) for some military applications.

1.2.2.1 Single Well CMOS

P-well CMOS, historically evolved from early PMOS processes, had been adopted in the industry as a main stream technology in the first part of the 1980's. From the device performance stand point, p-well structures generally offer better balance between p-ch and n-ch MOSFETs in complementary circuits using design rules greater than 3 μm. In a p-well CMOS technology, PMOS devices are built on an n-type starting substrate, whereas NMOS devices are built in a

compensated p-well formed using a boron implant. Due to the overcompensation in the channel doping for the NMOS device, the electron mobility is lower than that of a non-compensated channel. Figure 1-5 shows the effective carrier mobility for electrons and holes for MOS devices built in p-well and n-well with respect to channel length. It is observed that for gate lengths longer than 3.0 μm, the hole mobility in a p-well structure is larger, thus reducing the differences in current drive capability of n-ch over p-ch devices. The speed advantage is further enhanced by lower parasitic capacitances in p-well structures since with adequate p-well surface concentration, an n-channel 'channel stop' implant is not needed.

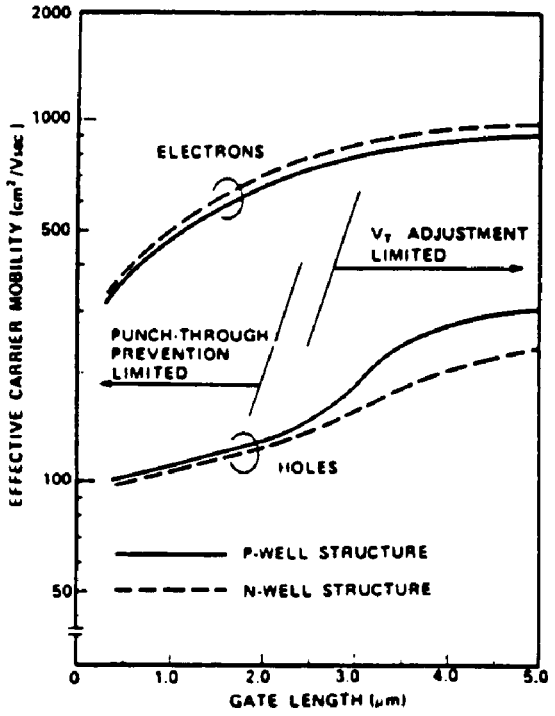


Figure 1-5. Electrons and Holes Mobility on n- and p-well CMOS devices.[41]

N-well technology, on the other hand, emerged from existing and established NMOS processes, where the devices are built on p-type substrates. However, a self-aligned n-well process, (with respect to p-well on p-substrate), is also very attractive in incorporating an npn bipolar transistor on a CMOS chip (BiCMOS). In this case, the npn transistor is isolated with p-type material. For applications that require high sensitivity, such as differential sense amplifiers and analog operational amplifiers, the BiCMOS approach is very attractive. Shown in Figure 1-6

is a cross section of an n-well BiCMOS structure which employs process steps in CMOS such as n+ for emitter and collector, threshold adjust implant for base and p+ for base contact. However, a patterned buried layer is needed to reduce collector series resistance if high performance npn transistor is desired.

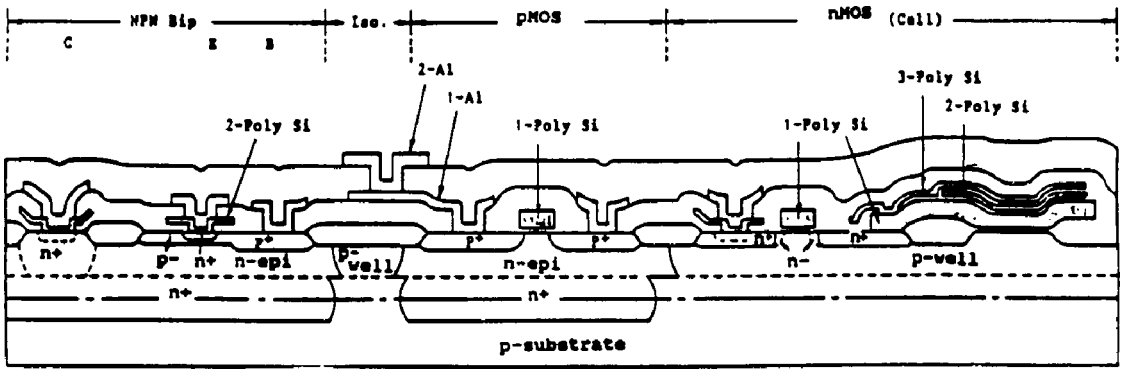


Figure 1-6. n-well BiCMOS structure for SRAM application.[42]

As the progress of CMOS processes moves toward more advanced design rules, the arguments of process compatibility and historical reasons are no longer valid. Since both single well approaches suffer from overcompensating the substrate background doping, the intrinsic mobility of the device placed in the well is severely degraded. Furthermore, the choice of substrate type, n- or p-type, depends on circuit applications, e.g. n-well for EPROM related products and p-well for SRAM and microprocessors making neither technology suitable for all products. An optimal approach is the twin-tub CMOS process which will be described in the following section.

1.2.2.2 Twin-Tub CMOS Technology

The single-mask approach to form twin-tub (or twin-well), which was developed at AT&T Bell Laboratories,[40] allows the independent tailoring of the individual well profiles without having either device type suffer from excessive counter-doping effects. Since a lightly doped n- or p-type substrate can be used interchangeably, the twin-tub approach also allows flexibility in choosing the optimum type of starting material for a particular circuit application. A cross section of twin-tub CMOS structures is shown in Figure 1-7.

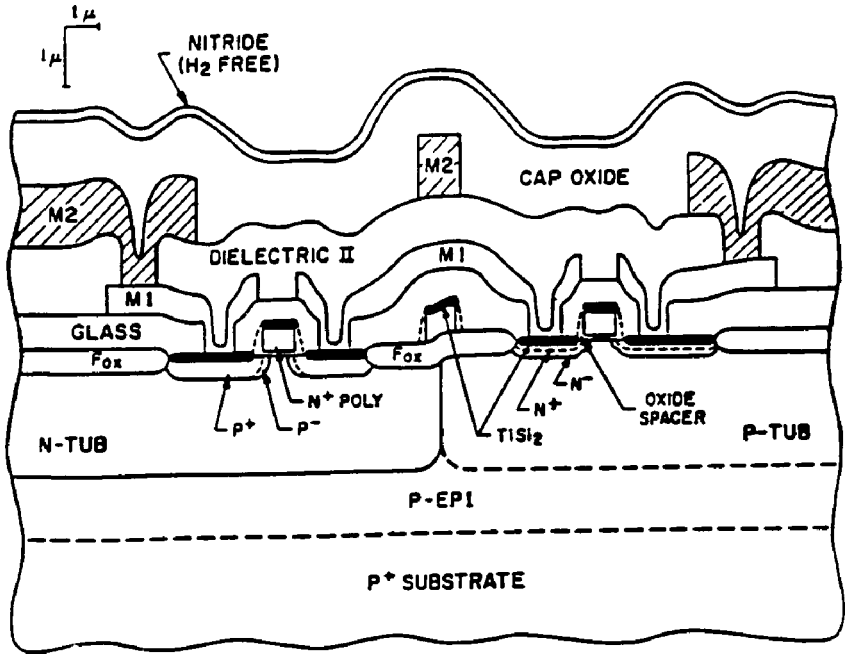


Figure 1-7. Cross section structure of a twin-tub CMOS structure.[43]

The detailed fabrication sequence of this structure will be described in chapter 4. In this technology a surface n-channel and a buried p-ch MOSFET are integrated with n+ doped polycide ($TiSi_2$/polysilicon) gate. The formation of twin-tub is done by a single mask level utilizing the LOCOS (Local Oxidation) process.

The process and device considerations for micron and submicron CMOS technology are

numerous. The factors that affect circuit packing density, besides the usual scaling of the channel length and width, are the device isolation (within tub and across tub boundary), the source/drain junction depth, the local interconnection schemes, and multi-level metal technology. The isolation of the same device type (or intra-tub), i.e. NMOS to NMOS or PMOS to PMOS, is determined by the subthreshold leakage of the parasitic field transistors. In the conventional LOCOS isolation, the minimum spacing between 2 adjacent diffusion regions is controlled by the surface concentration and field oxide thickness in the isolation area. However, as the spacing is reduced, the high-field-induced barrier lowering at the drain of the field device becomes severe, thus limiting the allowable minimum spacing. Alternative approaches include trench isolation,[44] field shield isolation,[45] and selective epitaxial growth of active device areas.[46] Some of these isolation techniques also aim at latch-up prevention.

Some of the earlier considered disadvantages of CMOS technology, such as higher mask count, lower packing density as compared with NMOS, have become accepted as de facto of this technology. A CMOS process, with double level metal, requiring 11 masking steps is considered normal. Some specific circuit applications such as SRAM and DRAM need as many as 23 masks. Low packing density due to large isolation between n and p-ch devices has been accepted and design cleverness can usually overcome these drawbacks. The latch-up susceptibility caused by parasitic npnp structures between the tub boundary are rectified by relaxing layout rules for high current handling circuits, retrograde n-tub, BiCMOS structures with an n+ buried layer, and in cases trench isolation or selective epi.

1.3 DEVICE SCALING ASPECTS

As indicated in the previous section on the packing density of integrated circuits, the active device scaling constitutes only one aspect of the overall scaling. However, it is the physics and technology associated with the fabrication of submicron devices that represents the most challenging barrier to be overcome. In this section we will review the traditional device scaling theory and approaches, and the practical shortcomings thereof which lead to the motivation for this work.

1.3.1 Constant Field Device Scaling

The conventional MOS transistor scaling laws were developed by Dennard et al [47] In this approach, to maintain the constant electric field in the device, the voltage and current are scaled accordingly to the device geometry. Assuming the scaling factor S (where S>1, e.g. for 0.9μm scaled to 0.6μm S=1.5), the device parameters should follow the scaling laws listed in Table 1-1.

TABLE 1-1. Conventional Scaling Law

Parameters	Scale (S>1)
L, W, t_{ox}, X_j	1/S
Doping Concentration	S
VOLTAGE	**1/S**
CURRENT	1/S
Capacitance	1/S
Delay Time	1/S
Power Dissipation	$1/S^2$
Power Delay Product	$1/S^3$
Power Density	1

Note that the voltage and current are scaled by a factor 1/S (<1). This was applicable in the early 1970's when supply voltages of MOS technologies were reduced from 16V, to 12V, to 8V and finally 5V. The 5V supply voltage for MOS ICs has been used for more than a decade for the reason of TTL logic compatibility. Therefore with the design rules of 5μm in the late 1970's to the present 0.7-0.8μm geometry, the voltage has not been scaled. In this case the current drive does not scale. In fact, the current increases with device scaling. The subthreshold current, due the exponential relationship with the gate voltage, does not scale. The transistor width used in circuit is not scaled accordingly, largely because the parasitic capacitances are not scaled very well. This creates reliability problems in contacts and metal conductors caused by electromigration.

1.3.2 Practical Scaling of CMOS Devices

The device parameters of 5 generations of CMOS technology are listed in Table 1-2. The constant supply voltage and the increase in saturation current drive are highlighted. As noted in Table 1-2 both lateral (L_{eff}) and vertical (t_{ox} and X_j) dimensions are decreased while the voltage remained the same. The Gauss' electric field law, governed by the relationship $E=V/d$, indicates that the field must increase both horizontally and vertically.

TABLE 1-2. Practical Scaling of CMOS Devices

CMOS Technologies	2.5μm [48]	1.75μm [49]	1.25μm [50]	0.9μm [43]	0.6μm [51]
L_{eff} (μm)	1.5	1.3	1.0	.75	0.5
t_{ox} (Å)	350	250	200	150	125
$X_{jn,p}$ (μm)	0.8	0.55	0.3,0.5	0.3,0.4	0.2,0.25
N_A (cm^{-3})	3E16	4E16	5E16	1E17	1E17
V_{tn} (V)	0.7	0.7	0.65	0.65	0.60
V_{DD} (V)	**5**	**5**	**5**	**5**	**3.3→4.0**
I_{on} (mA/μm)	**.180**	**.230**	**.280**	**.380**	**.340(@3.3V)**
Drain Design	Conv.	n-DDD	n-LDD	n&p-LDD	n-DDD

(From References[48] [49] [50] [43] [51]).

1.3.3 High Field Effects

High field occurs in small geometry devices in several ways. As the channel length is reduced, short channel effects such as the subthreshold conduction and punch-through current are the concerns for device design. This issue can be addressed by increasing the channel doping concentration. The high channel doping concentration compounded with the thinner gate oxide degrades carrier mobility in the channel as shown in Fig. 1-8. It is noted that the electron mobility is degraded by 50% when the gate oxide thickness is scaled from 350 to 100. This is the first vertical high field effect.

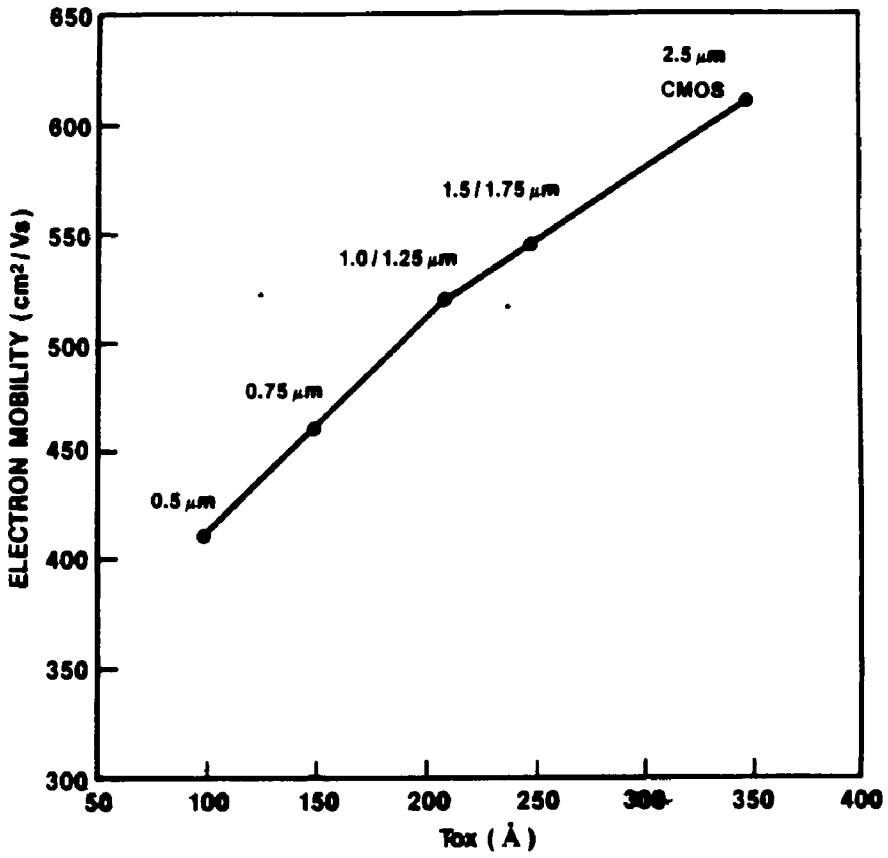


Figure 1-8. Electron mobility versus gate oxide thickness for different CMOS Technologies with appropriate channel doping, (measured at peak transconductance of long channel devices, with V_{DS}=0.1V).

The shallower junction depth together with higher bulk and field isolation doping lower the reverse junction breakdown voltage. If the doping concentration is not chosen properly, junction avalanche breakdown can occur at low voltage. With thinner gate oxide thickness, the issue of time dependent dielectric breakdown (TDDB) is eminent.[52]

The oxide conduction due to high vertical field is limited by the Fowler-Nordheim conduction process, which will take place at a field of 4-5 MV/cm. This means an oxide of 100Å can not be used as an insulator with the supply voltage at 5V. The corner effect of the source/drain junction and gate oxide (gated diode), combined with the spacer oxide pose major reliability problems. However, the most severe high field effects for the constant voltage scaling is the generation and injection of hot carriers due to the high electric field in the channel. This effect has caused degradation in circuit performance due to a reduction in transconductance and a shift in the threshold voltage. An additional reliability consequence of this effect is a weakened

gate oxide that may be broken down prematurely. The excessive substrate current, generated by impact ionization of hot carriers in a high electric field, causes drain-source breakdown, overloads on-chip substrate bias generator, and induces latch-up. The drain to source breakdown caused by the hot carrier generation is one of the factors affecting the maximum operating voltage.

Modified drain-structures such as Double Diffused Drain (DDD) and Lightly Doped Drain (LDD) have been implemented to alleviate this problem. However, with the device dimension in the range of 0.5μm, the modified drain structure proved not to be effective. A proposal to reduce the power supply voltage to 3.3V, has been accepted by some companies, and resisted by others. The mechanisms and characterization of hot carrier effects have been studied by several authors who mostly dealt with DC stressing of single devices. While this method provides some basic understanding of the physical processes involved (oftenly controversial), it is not adequate. Some ac aging results on CMOS inverters have been reported. Process and material issues relating to scaling have also been published. However, a global view in device scaling that encompasses circuit definition, device structures, fabrication processes and material issues has yet to be developed.

The issue of hot carrier effects is not only limited by the drain structure and supply voltage, but also related to processing conditions and the circuit configurations. How it affects the circuit and what the circuit can tolerate in performance are the goals. An integrated approach to deal with high field problems in scaling CMOS devices is the prime motivation for this dissertation.

1.4 SCOPE AND ORGANIZATION OF THE DISSERTATION

The focus of this dissertation is on the device scaling issues due to the high electric field effects resulting from a constant supply voltage and scaled physical dimensions. Hot carrier effects are emphasized based not only on physical understanding but also on what are the realistic limitations of hot carrier effects on actual circuit performance. In order to achieve this new level of understanding, the overall process integration considerations, starting from circuit design requirements, device structures, to process and materials are presented systematically based on theory and the experimental data.

1.4.1 Organization

In this chapter, we have reviewed the historical events that led to the thrust for device scaling. The evolution of CMOS technologies have been described. The scope, organization and contribution of this work are presented in this section.

Chapter 2 deals with the theory of MOS devices in various operation modes, i.e. subthreshold, linear, saturation, and breakdown regimes. The theory is developed to take into account submicron device behavior. Charge sharing and interface state density are introduced in a straightforward analytical form so that the physical insights can be grasped.

Chapter 3 covers the device design and process integration issues. Circuit design expectation is used as a starting point for device parameters. Technology assessments are discussed to help integrating individual processing steps (or modules) to fabricate the desired device structures. CMOS transistor structures, e.g. gate material, channel profile, and drain-engineering, are analyzed. The use of numerical process and device simulators is demonstrated to design and analyze alternative device structures. Device simulator is found very useful for gaining the insights of device behavior and for predicting device performance.

Chapter 4 describes the process sequence to fabricate completed CMOS devices and interconnects. The difficulties and tradeoffs for certain approaches are investigated. Where possible, alternatives or novel approaches are proposed.

Chapter 5 presents the experimental methods for technology characterization and the resulting process and device parameters extracted from these techniques. The device characteristics include subthreshold, linear, and saturation modes of operation. The off current, the threshold voltage, and the body effect are among the topics to be discussed. New small device geometry phenomena, i.e. reverse subthreshold swing and reverse short channel effects are reported. Simple correlation with theory will be made where appropriate.

Chapter 6 treats the generation and injection processes of the hot carriers in NMOS and PMOS devices. Models for substrate and gate currents are developed to explain the experimental results. A new model that accurately relates the peak substrate current and the drain voltage is

derived. The effects of drain structures on substrate current and drain-source breakdown voltage will be reported. The use of thin p-type epi on p+ substrate to improve the drain-source breakdown voltage is presented.

Chapter 7 deals with the experiments and analysis of devices degraded by hot carriers at high drain bias. Experimental results verify the proposed models of stressed devices. The subthreshold curves are used to extract the interface states generation during the accelerated stress. The criteria for predicting device lifetime at operating voltage will be proposed. It is shown that the DC aging on individual devices is useful method to compare device structures and monitor problems in processing. The ultimate degradation criterion is from circuit performance degradation tolerance, and the way the devices are used in circuits. The substrate current during device switching in an inverter is used to predict the device lifetime in this circuit.

Chapter 8 covers more advanced device structures for future generations. Topics in isolation, the emerging BiCMOS technology, new gate dielectric materials, novel local interconnect and drain-engineering structures are discussed. The future trends in CMOS technologies can be predicted based on the present understanding of device physics and technology know-how.

Chapter 9 concludes the dissertation with a summary of the contributions and new understanding and phenomena in submicron CMOS devices. Subjects for future research in university-industry cooperation environment are proposed.

1.4.2 Contributions of this Work toward the Art

During the course of this work and prior publications, we have contributed to the art and technology of scaled CMOS device design and integration in the following areas:

- A methodology for designing CMOS structures based on the understanding of device physics and its limitations, and on circuit application, utilizing existing tools and process capacity available to build and verify optimum device structures.
- We have extended device reliability due to high field effects by optimizing LDD structures, and compared with other device structures by stressing the devices at high field.

- We introduce a technique to extract the spatially distributed damaged regions of hot carrier aged devices by using subthreshold conduction measurements of well designed devices. The measurement set-up and the layout of transistors are well suited for aging multiple devices at the same time.
- We use the substrate current not only to predict the device lifetime, but also to detect process variations in the LDD spacer formation. This is one of the important parameters to monitor in manufacture.
- Reported new phenomena in submicron devices such as reverse subthreshold swing with high drain bias or short channel lengths, and reverse short channel effects in surface NMOS devices.
- Fabricated device structures to investigate the advantages and limitations of advanced techniques for isolation and alternative technologies, e.g. BiCMOS. These topics are important for decision making for future technologies.
- We have used the process and device simulators to analyze and optimize device behavior of different structures before fabrication, thus reducing the processing time.
- We have employed an integrated approach toward the device scaling: the interrelationship of process technology, device physics, circuit design and system applications. This is an important concept for realizing ULSI circuits and systems.

This is a blank page

Chapter 2

SUBMICRON CMOS DEVICE THEORY

2.1 Introduction And Convention

In this chapter, the basic MOS device theory is developed taking into account such effects as mobility degradation due to high vertical field, short channel effects, in particular subthreshold conduction. The conventional scaling theory is reviewed and a practical approach will be proposed based on circuit application requirements, process capability and device reliability. We will derive basic equations that govern the device operations for a surface channel NMOS device. The equations should be applicable to surface PMOS with appropriate sign changes.

The basic structure of n and p-channel MOSFETs are shown in cross-section in Fig. 2-1, where the gate insulator is thermally grown SiO_2, the gate is polycide (silicided polysilicon), and the source and drain doped with the same type as the conduction channel, is self-aligned with the patterned gate. The structure and operation of a modern submicron MOSFET is highly 3-dimensional in nature. Short channel effects, drain induced barrier lowering (DIBL), bulk conduction, and drain/source breakdown voltage are generally considered two-dimensional effects. The third dimension effects include narrow width effect, edge conduction along the bird's beak, and corner effect in the generation of hot carriers.

For an enhancement mode device, the channel is normally off when there is no applied potential to the gate. When sufficient voltage is applied to the gate, an inversion layer is formed on the surface of the channel (hence the name surface device). To simplify the physical understanding, we shall first examine the device behavior with 2-dimensional the device equations. The 3-dimensional effects are discussed where necessary. Considering an NMOS device with non-uniform doping channel, graded source/drain junctions, the device operations are analyzed in 3 regimes: subthreshold, strong inversion in linear mode, and saturation. The x direction is pointing toward the bulk, with its origin at the surface at the source junction. The y direction is along the horizontal surface, pointing from source to drain (see Fig. 2-1). All the terminal voltages reference to the source, which is grounded, unless otherwise noted. A drain

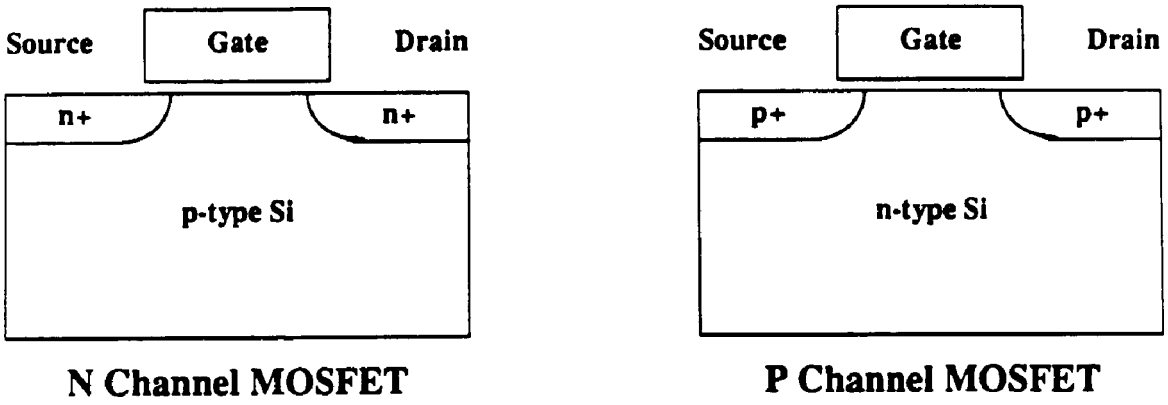


Figure 2-1. Cross-Sections of n and p-channel MOS Transistors

current I_D is a total current measured at the drain, whereas I_{DS} indicates a drain-to-source (or channel) current. For most treatments in this chapter, $I_{DS} = I_{DS} = I_S$, assuming no substrate, gate, or drain-to-bulk leakage currents exist. A band bending downward assumes a positive sign in potential, and negative for an upward band bending.

2.2 Subthreshold Conduction in a Long Channel MOSFET

We first construct the energy band diagram of an MOS capacitor with an n+ doped polysilicon gate. In the band diagram of Fig. 2-2, the Fermi level of an n+ doped gate is assumed constant, and should be very close to the conduction band (ϕ_{FnG} ~0.45 to 0.55eV above mid-gap). At zero gate bias, the Fermi levels in the silicon bulk, E_{Fp} and in the gate electrode, E_{FnG} line up (Fig. 2-2). The work function difference between the gate and the bulk is

$$\Phi_{GB} = \phi_{FnG} - \phi_{Fp} = -\frac{kT}{q}\ln(\frac{N_{DG}}{n_i}) - \frac{kT}{q}\ln(\frac{N_A}{n_i}) \qquad (2.1)$$

where N_B is an assumed constant bulk doping density over a depletion width of the device. For an n-ch device $N_B = N_A$, and $N_B = N_D$ for a p-ch device. Due to the work function difference of the gate material, the surface potential can be bent at zero gate bias and the majority carriers at the surface may be depleted. If this band bending is greater than

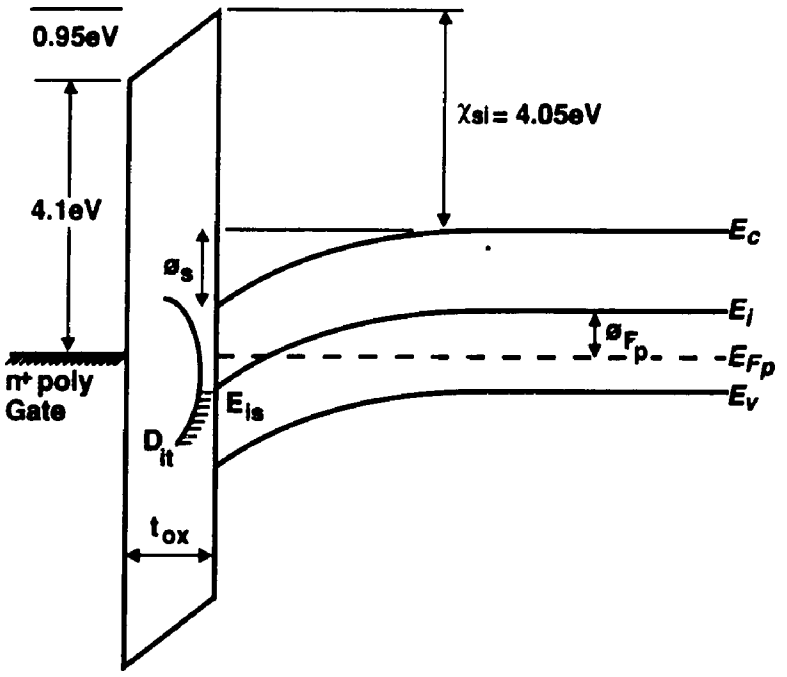


Figure 2-2. Energy Band Diagram of Poly Gate MOS structure.

$$\phi_F = \frac{kT}{q}\ln(\frac{N_B}{n_i}) \tag{2.2}$$

at a given gate voltage V_{GS}, then the channel is in 'weak inversion', and the conduction takes place for a finite drain voltage applied with respect to the source. This conduction current is referred to as the 'subthreshold' current. In a more specific case, off-current, I_{off}, is defined as the drain current with V_{GS} = 0V, and V_{DS} = maximum operating voltage (i.e., 5.5V for 5.0V and 3.6V for 3.3V supply, respectively). In order to address the first high field effect in short channel devices, in this case, punch-through current, we shall develop the off-current relationship with respect to device structure. Once the I_{off} is controlled, punch-through problems are eliminated. The subthreshold current is also important in such circuits as high density DRAM and SRAM, and in low-power, battery-back up circuits.

The surface potential, which is defined as the band bending at the surface compared with the neutral bulk, using either the conduction band or the intrinsic level at mid-gap as the reference is expressed as

$$\phi_s = \frac{E_i - E_{is}}{q} = \psi_s + V_{SB} \tag{2.3}$$

and the surface potential at the onset of weak inversion becomes

$$\phi_s(weak\ inv.) = \phi_F + V_{SB} \tag{2.4}$$

In weak inversion, the channel conduction is diffusion dominated, similar to the n-p-n bipolar collector current, with the channel equivalent to the base. In Eq. (2.3), ψ_s is the surface potential with respect to the source. With the back gate bias, V_{SB}, applied to the bulk, we use the notation $\phi_s = \psi_s + V_{SB}$ for a more generalized case. The channel current, with a bias V_{DS} at the drain, is the sum of the drift and diffusion currents and is given by

$$I_{DS} = q\mu_n nE + qAD_n \frac{dn}{dy} \tag{2.5}$$

The lateral electric field, E, is defined as the change in surface potential with respect to distance y from the source, i.e.

$$E = -\frac{\partial \phi_s}{\partial y} \tag{2.6}$$

where A is the cross section of the current flow, can be a spread in the electron flux depending on the doping profile of the channel region. D_n is the electron diffusion constant and related to the electron mobility, μ_n, by the Einstein relation

$$D_n = \mu_n \frac{kT}{q} \tag{2.7}$$

If a short channel device is designed so that its behavior is in the same manner as the long channel device, then the subthreshold current is independent of the channel length and the drain voltage. In practice, the submicron device is designed so that the I_{off} current is lower than a certain limit at a given minimum channel length. The subthreshold current of a submicron device is, in general, a function of both channel length and drain voltage.

From the drain current equation (2.5) we will develop a first order expression for the subthreshold current as a function of the channel length and drain voltage. In Eq. (2.5), the electron density n at point (x,y) is expressed, for a general 2-dimensional case, in terms of the electrostatic potential, $\phi(x,y)$, as

$$n(x,y) = \frac{n_i^2}{N_A}\exp\left[\left[\phi(x,y)-(V(y)+2\phi_F + V_{SB})\right]/V_T\right] \tag{2.8}$$

where $V(y)$ is the local potential at point y in the channel as a function of distance from the source to the drain, i.e. at the source V(0)~=~0 and at the drain $V(L) = V_{DS}$. To simplify the analysis, we rewrite Eq. (2.8) in terms of surface concentration, n_s, and potential, ϕ_s, and treat the problem in 1-dimension along the y direction. At the surface, i.e. x=0, $\phi(0) = \phi_s$, the surface concentrations of the electrons and holes are derived as:

$$n_s(y) = \frac{n_i^2}{N_A}\exp[\phi_s-(V(y)+2\phi_F + V_{SB})]/V_T \tag{2.9}$$

and

$$p_s(y) = N_A e^{-\phi_s/V_T} \tag{2.10}$$

Taking the derivative of n_s with respect to y we have,

$$\frac{\partial n_s}{\partial y} = \frac{n_s}{V_T}\left[\frac{\partial \phi_s}{\partial y} - \frac{\partial V(y)}{\partial y}\right] \tag{2.11}$$

Substitute equations (2.6) and (2.11) into equation (2.5), the channel current density is written as

$$J_n(y) = q\mu_n n_s\left[-\frac{\partial \phi_s}{\partial y}\right] + qD_n\frac{n_s}{V_T}\left[\frac{\partial \phi_s}{\partial y} - \frac{\partial V(y)}{\partial y}\right] \tag{2.12}$$

From Eq. (2.7), substitute $D_n = \mu_n V_T$ into Eq. (2.12) the channel current equation reduces to

$$J_n(y) = -q\mu_n n_s\frac{\partial V(y)}{\partial y} \tag{2.13}$$

From the band diagram of Fig. 2-2, $V(y)$ is the difference in the band bending along the channel length. We integrate Eq. (2.13) in the x direction from the surface to the depletion width X_d and multiply the current density by the transistor width W to obtain the total drain current,

$$I_{DS} = -W\int_0^{X_d} J_n(y)dx \tag{2.14}$$

For electrons moving in an inversion layer along the surface of an MOS transistor, the general mobility term μ_n, is replaced with the effective electron mobility, μ_{eff}. Equation (2.14) is written as,

$$I_{DS} = W\mu_{eff}\frac{\partial V(y)}{\partial y}\int_0^{sX_{ss}} qn_s dx \tag{2.15}$$

The integral of the electron density over the depletion width is the inversion charge in the channel, Q_n. Then Eq. (2.15) is reduced to

$$I_{DS} = W\mu_{eff}Q_n\frac{\partial V(y)}{\partial y} \tag{2.16}$$

The effective electron mobility μ_{eff} is the combination of surface and bulk mobility over the depletion width, affected by both vertical and lateral electric field, and can be expressed in terms of the intrinsic electron mobility, μ_0, the inversion and bulk charges, Q_n and Q_B, and the effective channel length, L_e, as[53] (refer to Fig. 2-3 for L and L_e):

$$\mu_{eff} = \frac{\mu_0}{1 + \frac{\theta_s}{C_{ox}}\left[Q_n + 2Q_B\right] + \theta_c L_e\frac{\partial V(y)}{\partial y}} \tag{2.17}$$

where

$$L_e = L - w_s - w_d \tag{2.18}$$

$$= L - \alpha\sqrt{V_{SB} + \phi} - \alpha\sqrt{V_{SB} + V_{DS} + \phi} \tag{2.19}$$

and

$$\alpha = \sqrt{\frac{2\varepsilon_s}{qN_A}} \tag{2.20}$$

is the depletion width coefficient. For the long channel case, we can assume that the depletion widths associated with the source and drain, w_s and $w_d \ll L$, then $L_e \simeq L$. (A treatment of the short channel case is found in Section 2.3).

We substitute Eq. (2.17) into (2.16), and the drain current becomes,

$$I_{DS} = \frac{\mu_0 W Q_n\frac{\partial V(y)}{\partial y}}{1 + \frac{\theta_s}{C_{ox}}\left[Q_n + 2Q_B\right] + \theta_c L_e\frac{\partial V(y)}{\partial y}} \tag{2.21}$$

In Eq. (2.21) Q_n is a function of $V(y)$, V_{SB} and V_{GS}, and

$$\theta_s = \frac{\kappa_{ox}}{2\kappa_s t_{ox} E_{cr\perp}} \tag{2.22}$$

is the 'normal field' surface roughness parameter, with t_{ox} the gate oxide thickness and $E_{cr\perp}$ the critical normal field for mobility degradation.

$$\theta_c = \frac{1}{L_e E_{cr\parallel}} \tag{2.23}$$

is the lateral field degradation parameter associated with the saturation velocity

$$v_{sat} = \mu_0 E_{cr\parallel} \tag{2.24}$$

where $E_{cr\parallel}$ is the critical parallel (lateral) field for velocity saturation. We assume that the channel current is constant over the channel length, Eq. (2.21) can be integrated to yield:

$$I_{DS} = \frac{\mu_0 W \int_0^{V_{DS}} Q_n dV}{L_e \left[1 + \frac{\theta_s}{C_{ox} L_e} \int_0^{L_e} (Q_n + 2Q_B) dy + \theta_c V_{DS}\right]} \tag{2.25}$$

For a long channel device we have the gradual channel approximation and the voltage drops linearly along the channel, i.e.

$$V = \frac{y}{L_e} V_{DS} \tag{2.26}$$

We integrate the second term of the denominator of Eq. (2.25) using Eq. (2.26),

$$\frac{\theta_s}{C_{ox} L_e} \int_0^{L_e} (Q_n + 2Q_B) dy = \frac{\theta_s}{V_{DS} C_{ox}} \int_0^{V_{DS}} (Q_n + 2Q_B) dV \tag{2.27}$$

Substitute Eq. (2.27) into Eq. (2.25), the channel current equation becomes,

$$I_{DS} = \frac{\mu_0 W \int_0^{V_{DS}} Q_n dV}{L_e \left[1 + \frac{\theta_s}{C_{ox} V_{DS}} \int_0^{V_{DS}} (Q_n + 2Q_B) dV + \theta_c V_{DS}\right]} \tag{2.28}$$

We now derive the channel charge Q_n as a function of the gate voltage which controls the

surface potential ϕ_s at the silicon surface. In the subthreshold regime, the MOSFET behaves similar to the bipolar transistor, whereby the source acts as an "emitter", the channel can be considered as an "indirect base", and the drain is equivalent to the "collector". The surface potential in the channel, which is equivalent to V_{BE} in the bipolar bias configuration, is the resultant of the capacitive coupling from the gate voltage V_{GS}, the body bias V_{SB}, and the drain bias V_{DS}.

The one-dimensional Poisson's equation in the channel position y, may be written as

$$\frac{\partial^2\phi}{\partial x^2} = -\frac{\rho}{\varepsilon_s} = \frac{q}{\varepsilon_s}[N_A - N_D + n - p]$$

$$= \frac{qN_A}{\varepsilon_s}[1 + \frac{n}{N_A} - \frac{p}{N_A}] \tag{2.29}$$

Using Eqs. (2.9) and (2.10) for n and p, Eq. (2.29) becomes:

$$= \frac{qN_A}{\varepsilon_s}\left[1 + \exp\left[(\phi - 2\phi_F - V_{SB} - V(y))/V_T\right] - \exp(-\phi/V_T)\right] \tag{2.30}$$

In Eq. (2.29) ϕ is a function of x at position y, $\phi = \phi(x)\,|_y$. We solve Eq. (2.29) with the following boundary conditions:

$$\phi(x=0) = \phi_s \tag{2.31}$$

$$\frac{\partial\phi}{\partial x}(x=0) = -E_s \tag{2.32}$$

$$\phi(x=\infty) = 0 \tag{2.33}$$

Integrating both sides of Eq. (2.30) with respect to ϕ, and using conditions Eqs. (2.31) and (2.32), and using $V=V(y)$, the surface charge as function of electric field E_s is expressed by Gauss's law as

$$Q_s = -\varepsilon_s E_s \tag{2.34}$$

$$Q_s = -\sqrt{2\varepsilon_s qN_A}$$
$$\cdot\left[\phi_s + V_T\left[-1 + \exp[(\phi_s - 2\phi_F - V_{SB} - V)/V_T] + \exp(-\phi_s/V_T) + \exp[(-2\phi_s - V_{SB} - V)/V_T]\right]\right]^{1/2} \tag{2.35}$$

In weak inversion conditions, i.e. $\phi_F \le \phi_s \le 2\phi_F$, Eq. (2.35) can be reduced to

$$Q_s \simeq -\sqrt{2\varepsilon_s q N_A}\left[\phi_s + V_T\left[-1+e^{(\phi_s-2\phi_F-V_{SB}-V)/V_T}\right]\right] \tag{2.36}$$

The surface charge can be expressed in terms of the bulk charge as a function of surface potential and the mid-point of weak inversion, referred to as "mid-inversion", when $\phi_s = 1.5\phi_F$, i.e.

$$Q_s = Q_B(\phi_s) = Q_B(1.5\phi_F + V_{SB}) + (\phi_s - 1.5\phi_F - V_{SB})\frac{\partial Q_B}{\partial \phi_s}\Big|_{(1.5\phi_F+V_{SB})} \tag{2.37}$$

$$= Q_B(1.5\phi_F+V_{SB}) - C_D\cdot(\phi_s - 1.5\phi_F - V_{SB}) \tag{2.38}$$

We define the average bulk charge at mid-inversion is $\overline{Q_B} = Q_B(1.5\phi_F+V_{SB})$, Eq. (2.38) can be rewritten as

$$Q_B = \overline{Q_B} - \overline{C_D}\cdot[\phi_s - (1.5\phi_F+V+V_{SB})] \tag{2.39}$$

The interface trapped charge Q_{it} is considered responding to the surface band bending about mid-inversion surface potential, $1.5\phi_F$. Similarly, we define $\overline{Q_{it}} = Q_{it}(1.5\phi_F+V+V_{SB})$, then the interface trap charge as a function of surface potential can be written as

$$Q_{it} = \overline{Q_{it}} - q\overline{D_{it}}\cdot[\phi_s - (1.5\phi_F+V+V_{SB})] \tag{2.40}$$

where

$$q\overline{D_{it}} = \overline{C_{it}} \tag{2.41}$$

is the measured interface trap capacitance. We now relate the applied gate voltage to the surface potential and the charges at interface an in the bulk,

$$V_{GB} = V_{GS} + V_{SB} = V_{FB} + \phi_s - \frac{(Q_B + Q_{it})}{C_{ox}} \tag{2.43}$$

We use Eqs. (2.39) and (2.40) for Q_B and Q_{it}, V_{GB} is expressed as

$$V_{GB} = V_{FB} + \phi_s\left[1 + \frac{(\overline{C_D} + \overline{C_{it}})}{C_{ox}}\right] - \frac{(\overline{C_D} + \overline{C_{it}})}{C_{ox}}\cdot(1.5\phi_F+V+V_{SB}) - \frac{(\overline{Q_B} + \overline{Q_{it}})}{C_{ox}} \tag{2.44}$$

Deriving an expression for V_{GS} from Eq. (2.44), we obtain

$$V_{GS} = V_{FB} + 1.5\phi_F - \frac{(\overline{Q_B} + \overline{Q_{it}})}{C_{ox}} + \phi_s \left[1 + \frac{(\overline{C_D} + \overline{C_{it}})}{C_{ox}}\right]$$
$$- V_{SB} \left[1 + \frac{(\overline{C_D} + \overline{C_{it}})}{C_{ox}}\right] - V\frac{(\overline{C_D} + \overline{C_{it}})}{C_{ox}} \tag{2.45}$$

We define the 'mid-inversion threshold voltage', V_{in}^* as

$$V_{in}^* = V_{FB} + 1.5\phi_F - \frac{(\overline{Q_B} + \overline{Q_{it}})}{C_{ox}} \tag{2.46}$$

and we further define

$$n = 1 + \frac{\overline{C_D} + \overline{C_{it}}}{C_{ox}} \tag{2.47}$$

and

$$m = 1 + \frac{\overline{C_D}}{C_{ox}} \tag{2.48}$$

Equation (2.45) reduces to:

$$V_{GS} = V_{in}^* - 1.5\phi_F n + \phi_s n - V_{SB} n - V(n - m) \tag{2.49}$$

Solve Eq. (2.49) for ϕ_s:

$$\phi_s = \frac{V_{GS} - V_{in}^*}{n} + 1.5\phi_F + V_{SB} + V(1 - \frac{m}{n}) \tag{2.50}$$

For a small drain voltage, the potential along the channel V(y) is small, therefore the last term of Eq. (2.49) can be neglected. From Eqs. (2.47) and (2.48) the interface trap capacitance, $\overline{C_{it}}$, and then the interface trap density, $\overline{D_{it}}$, can be calculated. (The techniques to measure and extract n and m will be described in details in chapter 6, to obtain interface state density.) The space charge equation (2.38) can be rewritten, using the body effect term:

$$\lambda = \frac{\sqrt{2q\varepsilon_s N_B}}{C_{ox}} \tag{2.51}$$

then

$$Q_s = Q_n + Q_B = \lambda C_{ox}\left[\phi_s + V_T[-1 + \exp(\phi_s - 2\phi_F - V_{SB} - V(y)/V_T)]\right]^{1/2} \tag{2.52}$$

At the onset of weak inversion, $\phi_s = \phi_F + V_{SB} + V(y)$, $Q_s \sim Q_B$ at any point along the channel. Rearrange Eq. (2.52), using $V = V(y)$, we obtain,

$$Q_s = \lambda C_{ox}(\phi_s - V_T)^{1/2}\left[1 + \frac{V_T}{\phi_s - V_T}\exp((\phi_s - 2\phi_F - V_{SB} - V)/V_T)\right]^{1/2} \tag{2.53}$$

$$= Q_B\left[1 + \frac{V_T}{2(\phi_s - V_T)}\exp((\phi_s - 2\phi_F - V_{SB} - V)/V_T)\right]^{1/2} \tag{2.54}$$

$$= Q_n + Q_B \tag{2.55}$$

where

$$Q_B = \lambda C_{ox}\sqrt{\phi_s + V_{SB} + V} \tag{2.56}$$

Solve for Q_n:

$$Q_n = \frac{1}{2}\frac{Q_B V_T}{(\phi_s - V_T)}\exp[(\phi_s - 2\phi_F - V_{SB} - V)/V_T] \tag{2.57}$$

from Eq. (2.57), the inversion charge is only significant to the bulk charge Q_B when the surface band-bending ϕ_s is greater than $2\phi_F$. We substitute Eq. (2.57) into the drain current equation (2.28) and integrate to obtain the subthreshold conduction current for the n-ch device,

$$I_{DS} = \frac{W}{L_e}\mu_{n,eff}C_D(\phi_s)\frac{n}{m}V_T^2\exp\left[(V_{GS} - V_{in}^*)/nV_T\right]$$
$$\cdot\exp(-\phi_F/2V_T)\cdot\left[1 - \exp(-mV_{DS}/nV_T)\right] \tag{2.58}$$

In Eq (2.58) n and m are defined as in Eqs. (2.47) and (2.48), i.e. n and m are evaluated at the mid-inversion, where the surface potential $\phi_s = 1.5\phi_F + V_{SB}$.

For PMOS devices, we follow a similar analysis where the surface potential, bulk Fermi level, and applied voltages are negative with respect to ground. We can derive a similar expression for the subthreshold current in a p-ch device,

$$I_{DS} = \frac{W}{L_e}\mu_{p,eff}C_D(\phi_s)\frac{n}{m}V_T^2\exp\left[-(V_{GS} - V_{tp}^*)/nV_T\right]$$
$$\cdot\exp(\phi_{Fp}/2V_T)\left[1 - \exp(mV_{DS}/nV_T)\right] \qquad (2.59)$$

2.3 Charge Sharing and Subthreshold Conduction in a Short Channel MOS Device

In the previous section, the subthreshold conduction equations were derived for the long channel case, where the underlying fundamental physics can be examined with a simple 1-D analysis. In this section we shall formulate a similar analysis for the short channel case. The definition of short channel behavior depends on the transistor design, but in our context of good device structure, we will informally define a short channel device as when the sum of the source and drain junction depth approaches the channel length, i.e. $2X_j \sim L$. In this case the charge sharing from the depletion layers on both source and drain ends becomes significant in comparison with the bulk charge under the gate.

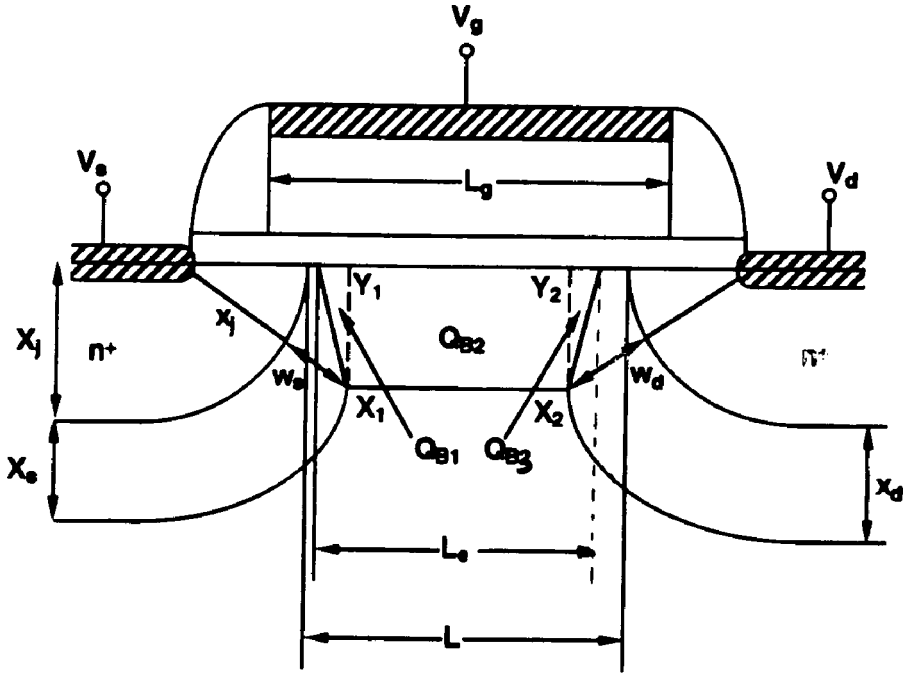


Figure 2-3. Cross Section of a short channel MOS device with charge sharing boundary

Consider a short channel NMOS transistor with source/drain junction depth of X_j and metallurgical channel length L and the device cross section as shown in Fig. 2-3. Although the

LDD drain structure is used extensively for hot carrier suppression, we can assume that the doping concentration of the lightly doped region is much higher than the channel doping (by a factor of ~25 times at the surface for practical devices), and, therefore, a one sided junction approximation can be used for the purpose of this analysis.

In this section, the subthreshold conduction in a short channel device is analyzed with a physically based model of charge sharing. A two-dimensional device simulator is used to determine the field line boundary in a short channel using a practical doping profile of an NMOS transistor in a submicron twin-tub CMOS technology. Figure 2-3a shows a family of equi-potential contours of a 0.5μm gate length NMOS device with V_{DS} = 0.01V, $V_S = V_B$ = 0V. The oxide thickness of the device is 125Å, with the channel doping of ~ 8.0E16 cm^{-3} extended over 0.4μm into the bulk (The simulated doping profile is shown in Fig. 3-4b). In Fig. 2-3, we have defined L_g as the gate electrode length, L as the physical channel length between the 2 metallurgical junctions at the source and drain, and L_e as the effective (or electrical) channel defined between the edge of the 2 depletion boundaries at the source and drain ends. As such, L_e is a function of drain bias and channel doping concentration. Let us first analyze the case with small drain voltage and note the charge sharing boundaries, as shown by the 2-D numerical simulation, are not straight lines. For a circular junction, with a junction depth X_j, the charge sharing boundaries satisfy the loci $y(x)$ with the condition

$$\left[\sqrt{(X_j+y)^2+x^2}-X_j^2\right]^2=\alpha^2\phi_s\left[1-\frac{x}{w_{d,s}}\right]^2 \tag{2.60}$$

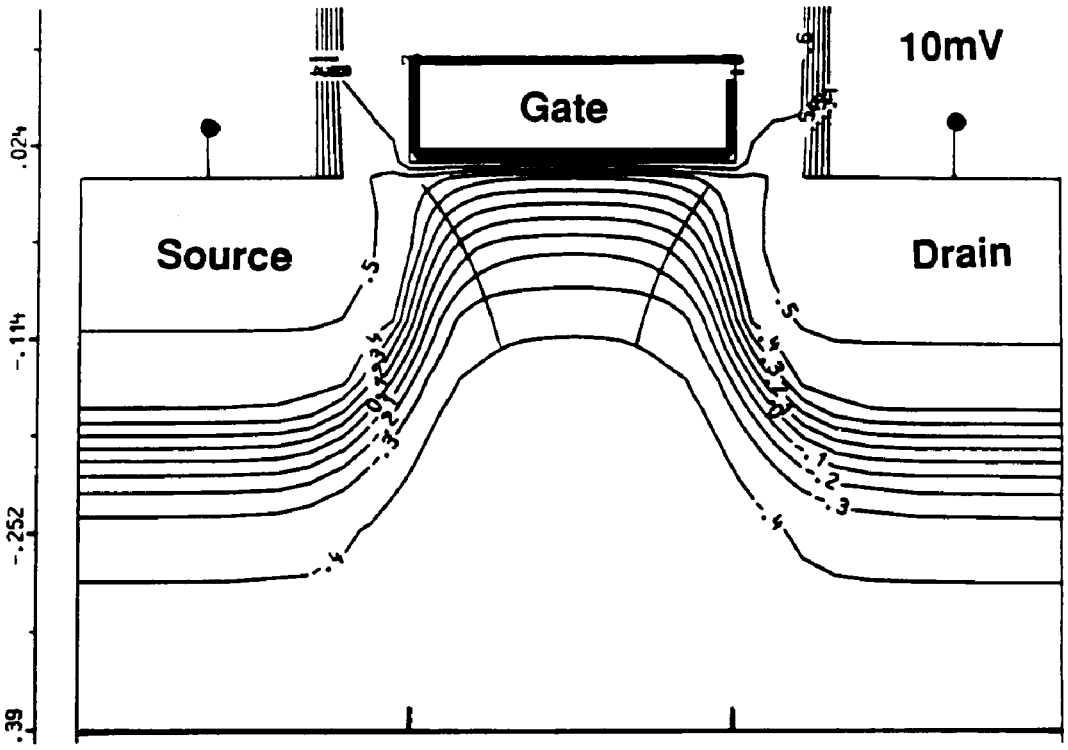
where w_s and w_d are bias dependent depletion widths at source and drain junctions,

$$w_s=\alpha\sqrt{\phi(x,y)} \tag{2.61}$$

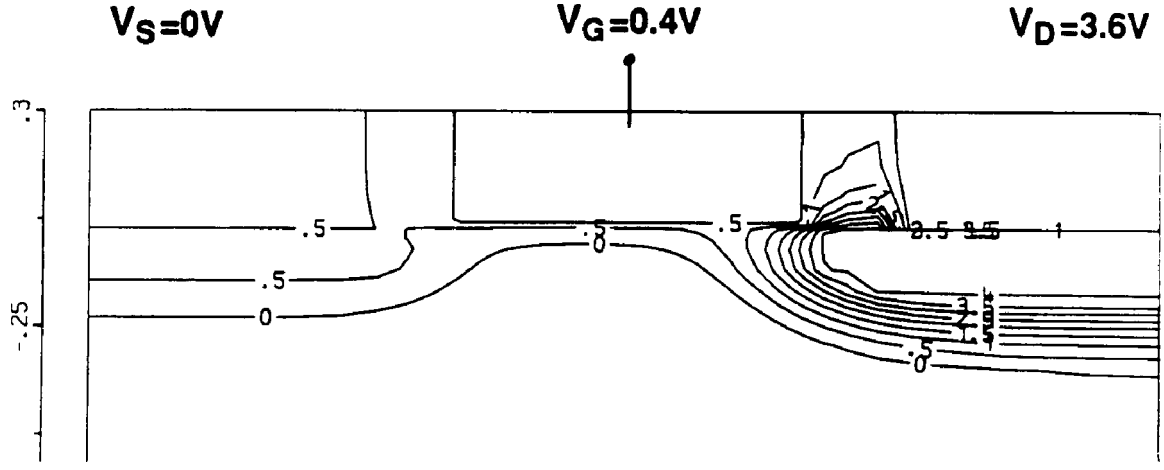
and

$$w_d=\alpha\sqrt{\phi(x,y)+V_{DS}} \tag{2.62}$$

where α is given by Eq. (2.20), and $\phi(x,y)$ is the two dimensional potential depending on the location of the charge sharing boundary along the contour of the junction (see Fig. 2-3). At the surface $\phi(x,y)$ is equal to the surface potential, ϕ_s. In the bulk, i.e. under the source/drain junctions outside the channel region, $\phi(x,y)$ is replaced with the junction built-in potential ϕ_{bi},



(a)



(b)

Figure 2-4. Simulated equi-potential contours of a 0.5μm channel length, V_{GS}= 0.3V, and V_S=V_{SB}=0V, (a)V_{DS}=10mV (b)V_{DS}=3.6V.

and is given by,

$$\phi_{bi} = \frac{kT}{q} \ln\left[\frac{N_A N_D}{n_i^2}\right] \tag{2.63}$$

The locus of Eq. (2.60) can be approximated as a straight line, therefore a straight forward geometrical analysis of the structure can be carried out. We will derive the expression for lateral position y_1 and y_2, so that the gradual channel approximation can be applied within these boundaries. Gradual channel approximation means:

$$\frac{\partial E_y}{\partial y} \ll \frac{\partial E_x}{\partial x} \tag{2.64}$$

(i.e. normal field variation is much larger than the lateral field variation, such that E_y varies gradually with distance). Refer to Fig. 2-3, at triangle 1, we can write:

$$(X_j + y_1)^2 + x_1{}^2 = (X_j + w_s)^2 \tag{2.65}$$

Solve for y_1 we obtain

$$y_1 = \sqrt{(X_j + w_s)^2 - x_1{}^2} - X_j \tag{2.66}$$

Similarly, at the drain end, we can write:

$$(X_j + L_2)^2 + x_2{}^2 = (X_j + w_d)^2 \tag{2.67}$$

Solve for L_2 measured from the drain end

$$L_2 = \sqrt{(X_j + w_d)^2 - x_2{}^2} - X_j \tag{2.68}$$

y_2 with respect to the origin of the coordinates at the source end is simply:

$$y_2 = L_e - L_2 \tag{2.69}$$

$$= L_e - \left[\sqrt{(X_j + w_d)^2 - x_2{}^2} - X_j\right] \tag{2.70}$$

The effective bulk charge in the trapezoid area is the sum of 3 components:

$$0 \le y \le y_1 : Q_{B1} = \lambda C_{ox}(V_{SB} + \phi_s + V_1)^{1/4} \frac{y}{y_1} \tag{2.71}$$

$$y_1 \le y \le y_2 : Q_{B2} = \lambda C_{ox}(V_{SB} + \phi_s + V(y))^{1/4} \tag{2.72}$$

$$y_2 \le y \le L_e : Q_{B3} = \lambda C_{ox}(V_{SB} + \phi_s + V(y))^{1/2} \frac{L_e - y}{L - y_2} \tag{2.73}$$

where λ is the body effect coefficient, and is given by Eq. (2.51).

Integrate the charge over the channel length using gradual channel length approximation, noted that this approximation is only valid in the range of L_e, and for small V_{DS}:

$$y = \frac{V(y)L}{V_{DS}} \tag{2.74}$$

Integrate Q_{B1} over length y_1 and using Eq. (2.74)

$$\int_0^{y_1} \frac{Q_{B1}}{C_{ox}} dy = \frac{\lambda}{y_1}(V_{SB} + \phi_s + V_1)^{1/2} \int_0^{y_1} y dy \tag{2.75}$$

$$= \frac{\lambda}{2}(V_{SB} + \phi_2 + V_1)^{1/2} \frac{V_1 L_e}{V_{DS}} \tag{2.76}$$

Similarly, for Q_{B2}, we evaluate:

$$\int_{y_1}^{y_2} \frac{Q_{B2}}{C_{ox}} dy = \int_{V_1}^{V_2} \frac{Q_{B2}}{C_{ox}} dV \tag{2.77}$$

$$= \frac{\lambda L}{V_{DS}} \int_{V_1}^{V_2} (V_{SB} + \phi_s + V)^{1/2} dV \tag{2.78}$$

$$= \frac{2\lambda L}{3V_{DS}} \left[(V_{SB} + \phi_s + V_2)^{3/2} - (V_{SB} + \phi_s + V_1)^{3/2} \right] \tag{2.79}$$

And for Q_{B3} at the drain end between y_2 and L_e,

$$\int_{y_2}^{L_e} \frac{Q_{B3}}{C_{ox}} dy = \frac{\lambda(V_{SB} + \phi_s + V_2)^{1/2}}{L_e - y_2} \int_{y_2}^{L_e} (L_e - y) dy \tag{2.80}$$

$$= \frac{\lambda(V_{SB} + \phi_s + V_2)^{1/2}}{2}(L_e - y_2) \tag{2.81}$$

$$= \frac{\lambda L_e}{2V_{DS}}(V_{SB} + \phi_s + V_2)^{1/2}(V_{DS} - V_2) \tag{2.82}$$

From Eqs. (2.76), (2.79) and (2.82), we can solve for V_1 and V_2 by using the gradual channel length approximation, Eq. (2.74):

$$V_1 = \frac{V_{DS}}{L_e}\left(X_j + \frac{\varepsilon_s V_{DS}}{qN_A L_e}\right)\left[\left[1 + \frac{2X_j[\frac{2\varepsilon_s}{qN_A}(V_{SB}+\phi_s)]^{1/4}}{(X_j + \frac{\varepsilon_s V_{DS}}{qN_A L_e})^2}\right]^{1/4} - 1\right] \tag{2.83}$$

Similarly, for V_2, we have,

$$y_2 = \frac{V_2 L_2}{V_{DS}} \tag{2.84}$$

$$= L_e + X_j - \left[(X_j + w_d)^2 - {X_2}^2\right]^{1/2} \tag{2.85}$$

Solve for V_2, we obtain,

$$\frac{V_2}{V_{DS}} = 1 + \frac{X_j}{L_e}\left(1 - \frac{\varepsilon_s V_{DS}}{qN_A L_e}\right) - \left[\left(\frac{X_j}{L_e} - \frac{\varepsilon_s V_{DS}}{qN_A L_e^2}\right)^2 + \frac{2X_j}{L_e^2}[\frac{2\varepsilon_s}{qN_A}(V_{SB} + \phi_{bi} + V_{DS})]^{1/2}\right]^{1/2} \tag{2.86}$$

Under very small V_{DS} condition, we let $V_{DS} \rightarrow 0$, Eqs. (2.83) and (2.86) become:

$$V_1 \simeq \frac{V_{DS}X_j}{L_e}\left[(1 + \frac{2w_s}{X_j})^{1/2} - 1\right] \tag{2.87}$$

$$V_1 = V_{DS} \cdot f_s \tag{2.88}$$

where

$$f_s \triangleq \frac{X_j}{L_e}\left[(1 + \frac{2w_s}{X_j})^{1/2} - 1\right] \tag{2.89}$$

and

$$\frac{V_2}{V_{DS}} \simeq 1 - \frac{X_j}{L_e}\left[\left[1 + \frac{2w_d}{X_j}\right]^{1/2} - 1\right] \tag{2.90}$$

$$V_2 = V_{DS} \cdot (1 - f_d) \tag{2.91}$$

where

$$f_d \triangleq \frac{X_j}{L_e}\left[\left[1 + \frac{2w_d}{X_j}\right]^{1/2} - 1\right] \tag{2.92}$$

f_s and f_d are referred to as short channel charge sharing factors at source and drain ends, respectively. To simplify the analysis, one can assume the average charge sharing at both ends by

the factor, $f = \frac{1}{2}(f_s + f_d)$. The total charge in the channel is simply expressed as:

$$Q_B = f \cdot Q_{B0} \tag{2.93}$$

where Q_{B0} is the long channel bulk charge. We will now use the charge-sharing concepts to derive the subthreshold current formulation for a short channel device.

$$I_{DS} = \frac{W\mu_{eff}C_{ox}}{L_e} \cdot \int_0^{V_{DS}} \partial V C_D V_T e^{[\phi_s - (2\phi_F + V_{SB} + V)]/V_T} \tag{2.94}$$

Using surface scattering coefficient, θ_s, Eq. (2.94) is evaluated as:

$$I_{DS} = \frac{\beta_0 C_D V_T}{1 + 2\lambda\theta_s\sqrt{1.5\phi_F + V_{SB}}} \cdot e^{(V_{GS} - V_m^*)/nV_T} \cdot e^{-\frac{\phi_F}{2nV_T}} \cdot \int_0^{V_{DS}} \partial V e^{-\frac{mV}{nV_T}} \tag{2.95}$$

$$= \frac{\beta_0 C_D V_T^2}{1 + 2\lambda\theta_s\sqrt{1.5\phi_F + V_{SB}}} \cdot \frac{n}{m} e^{(V_{GS} - V_m^*)/nV_T} \cdot e^{-\frac{\phi_F}{2nV_T}} \left[1 - e^{-\frac{mV_{DS}}{nV_T}}\right] \tag{2.96}$$

In Eq. (2.96), the channel length shortening is included in β_0 by the expression:

$$\beta_0 = \frac{\mu_0 C_{ox} W}{L - w_s - w_d} \tag{2.97}$$

where m and n are defined with charge sharing factor as

$$n = 1 + \frac{(1-f)\cdot\overline{C_D} + \overline{C_{it}}}{C_{ox}} \tag{2.98}$$

and

$$m = 1 + \frac{(1-f)\cdot\overline{C_D}}{C_{ox}} \tag{2.99}$$

where

$$f = \frac{(f_s + f_d)}{2} = \frac{1}{2}\frac{X_j}{L_e}\left[(1 + \frac{2w_s}{X_j})^{1/2} + (1 + \frac{2w_d}{X_j})^{1/2} - 2\right] \tag{2.100}$$

The drain voltage effect on short channel device is contained in the expressions of f_d, w_d. Therefore n and m are also affected by drain voltage. If V_{DS} is increased, then f_d increases according to Eq. (2.92), n is decreased (Eq. 2.98), i.e. the subthreshold swing is smaller.

For an optimized device, V_{in}^{*} is about midway between $V_{GS}=0$ and the strong inversion threshold voltage V_{in}, I_{off} can be derived , by substituting $V_{GS}=0$ into Eq. (2.96).

$$I_{off}=\frac{\beta_0 C_D {V_T}^2}{1+2\lambda\theta_s\sqrt{1.5\phi_F+V_{SB}}}\cdot\frac{n}{m}e^{(-V_m^*)/nV_T}\cdot e^{-\frac{\phi_F}{2nV_T}}\left[1-e^{-\frac{mV_{DS}}{nV_T}}\right] \qquad (2.101)$$

At high drain bias, the effective channel L_e is dependent on V_{DS}. The experimental results of these new short channel effects, referred to as reverse subthreshold swing, will be reported in chapter 5. If the gate voltage is further increased then the device will enter the threshold of strong inversion, which will be described in the next section.

2.4 Threshold Voltage

The threshold voltage is one of the most important parameters in the design of MOS transistor. As discussed in the previous section, the choice of threshold voltage for a given subthreshold slope affects the off current, and therefore the punch-through prevention in a short channel device. On the other hand, the device performance in the active region depends on the difference between the gate voltage and the threshold voltage, $V_{GS}-V_t$ for a drive current. When the gate voltage is larger than threshold voltage, the device is in strong inversion, the vertical field from the gate to channel affects the mobility, and therefore the conduction current of the device. In this section we will develop a theory, similar to that developed by Krutsick, White, Wong and Booth.[54]

The classical threshold voltage is defined when the surface potential, ϕ_s, is bent $2\phi_F$, based on the band bending at the surface and with the consideration of interface state traps. The gate voltage as a function of surface potential can be derived from the charge conservation in an MOS structure as:

$$Q_G+Q_{ox}+Q_{inv}+Q_B=0 \qquad (2.102)$$

Where Q_G is the charge on the gate, Q_{ox} is the total charge in the bulk oxide and at the interface, Q_{inv} is the inversion charge, and Q_B is the bulk charge in the depletion layer. Equation 2.102 can be expressed in terms of the voltage by dividing the charge by associated capacitance, we obtain

$$V_{GS} = V_{FB} + \phi_s - \frac{Q_B(\phi_s)}{C_{ox}} - \frac{Q_{it}(\phi_s)}{C_{ox}} \tag{2.103}$$

where V_{FB} is the flat-band voltage, expressed as

$$V_{FB} = \Phi_{ms} - \frac{Q_f}{C_{ox}} - \frac{Q_{it}(0)}{C_{ox}} \tag{2.104}$$

where Φ_{ms} is the work function difference between the gate material and semiconductor intrinsic Fermi level. For an n+ doped gate, with a acceptor bulk concentration of $N_A = 8 \cdot 10^{16} cm^{-3}$, Φ_{ms} is in the range of -0.85 to -0.95 eV, depending on the gate dopants (arsenic or phosphorus implant) and the anneal temperature,[55] and Q_f is the positive fixed charge. By the classical definition, the threshold voltage is the gate voltage that induces the the strong inversion on the channel region, at that point, the surface potential $\phi_s = 2\phi_F$. Equation (2.103) becomes:

$$V_{tn} = V_{GS}(\phi_s = 2\phi_F) = V_{FB} + 2\phi_F + \frac{1}{C_{ox}}\sqrt{2\varepsilon_s q N_A (2\phi_F + V_{SB} + V(y))} + \frac{q\overline{D_{it}} \cdot 2\phi_F}{C_{ox}} \tag{2.105}$$

The interface state density above surface potential of $2\phi_F$ is typically very difficult to determine, and normally assumed to have an effective value, $\overline{D_{it}}$. In the above equation we assume that the surface potential is bent $2\phi_F$ at strong inversion. However, Krutsick et al.[54] have introduced an improved model in which the band bending in strong inversion can be over $2\phi_F$, by the factor γ, where

$$\gamma = V_T \ln\left\{\frac{1}{\lambda^2 V_T}[(V_{GS} - V_{tn})(V_{GS} - V_{tn} + 2\lambda\sqrt{\phi_s - V_T})]\right\} \tag{2.106}$$

where $V_T = \frac{kT}{q}$ is a Boltzmann Thermal voltage.

The threshold voltage extrapolated from the maximum slope of a linear g curve, is in fact above the strong inversion threshold voltage,[54] and expressed by

$$V_{GS} = V_{tn} + \gamma + V_{DS}(1 + \delta)/2 \tag{2.107}$$

where

$$\delta = \frac{\lambda}{2\sqrt{\phi_s - V_T}} \tag{2.108}$$

and λ is the body effect coefficient as given by Eq. (2.51)

2.5 Linear Drain Current

The drain current model in strong inversion at low drain bias is expressed as the integral of inversion charge multiplied by the effective mobility under the influence of the drain bias.

$$I_{DS} = \frac{W}{L_e}\int_0^{V_{DS}} \mu_{eff} Q_n \partial V \tag{2.109}$$

The effective mobility, μ_{eff}, is derived by Mathiessen's rule, with the inclusion of surface roughness term, θ_s, is expressed as [54]

$$\mu_{eff} = \frac{\mu_o}{1 + \frac{\theta_s}{C_{ox}}(Q_n + 2Q_B)} \tag{2.110}$$

where θ_s is the mobility degradation term due to surface roughness and the vertical critical field, $E_{crit\perp}$.

$$\theta_s = \frac{C_{ox}}{2\varepsilon_s E_{crit\perp}} \tag{2.111}$$

The critical field in an n-ch device is typically in the range of 1-2×10^5 MV/cm. In the linear regime, the drain bias is assumed to be small, so that the channel shortening effect is negligible. Therefore, the second order drain voltage effects can be ignored in a well designed short channel transistor. Evaluating Eq. (2.109) using the effective mobility term of Eq. (2.110), and taking the source-drain series resistance into account (assuming $R_S = R_D$), we obtain,

$$I_{DS} = \frac{\beta_o[V_{GS} - V_{in} - \frac{V_{DS}}{2}(1+\delta)]V_{DS}}{1 + (\theta_s + 2\beta_o R_s)[V_{GS} - V_{in} + 2\lambda\sqrt{\phi_s - V_T} - V_{DS}(1+\delta)/2]} \tag{2.112}$$

If the drain voltage is maintained small, i.e $V_{DS} < V_T$ (25.8 mV at room temperature), then Eq. (2.112) can be further simplified as

$$I_{DS} = \frac{\beta_o[V_{GS} - V_{in}^*]V_{DS}}{1 + (\theta_s + 2\beta_o R_s)[V_{GS} - V_{in} + 2\lambda\sqrt{\phi_s}\,]} \tag{2.113}$$

It is clear that in short channel devices, the effect of series source-drain resistance is very important for linear transconductance gain, and therefore can not be neglected. In the above

equation, ϕ_s is the surface band bending, for small V_{DS}, over $2\phi_F$ by the term γ defined by Eq. (2.108).

$$\phi_s = 2\phi_F + \gamma + V_{SB} \tag{2.114}$$

To simplify the equation, we use μ_{eff} as an effective electron mobility. For the long channel case, the series resistance is assumed to be small compared with the channel resistance, we obtain the simplified expression for linear current:

$$I_{DS} = \mu_{eff}\frac{W}{L}C_{ox}(V_{GS} - V_{tn} - \frac{V_{DS}}{2}(1+\delta))V_{DS} \tag{2.115}$$

The accurate modeling of linear region for long channel device physical parameters such as channel doping, vertical field induced mobility degradation can be obtained.

2.6 Saturation Regime

When the drain voltage is increased greater than the gate voltage minus threshold voltage, $V_{GS} - V_t$ at strong inversion, the device is biased in saturation regime. Figure 2-5 depicts the device cross section of electron density of an n-ch device operates in saturation mode. It is shown that in the saturation mode, the electrons are steered away from the surface due to the field reversal in the pinch-off region. In the long channel length case it is well known that the drain saturation voltage is

$$V_{Dsat} = \frac{V_{GS} - V_{tn}}{1+\delta} \tag{2.116}$$

Substitute Eq. (2.116) into Eq. (2.115), we can obtain the conventional long channel saturation current:

$$I_{Dsat} = \frac{\mu_{eff}C_{ox}}{2(1+\delta)}\frac{W}{L_e}(V_{GS} - V_{tn})^2 \tag{2.117}$$

When the drain voltage is increased over V_{Dsat}, the channel length modulation occurs. As shown in Fig. (2-5), the pinch-off region, ΔL, increases with increasing drain voltage. ΔL is expressed as:

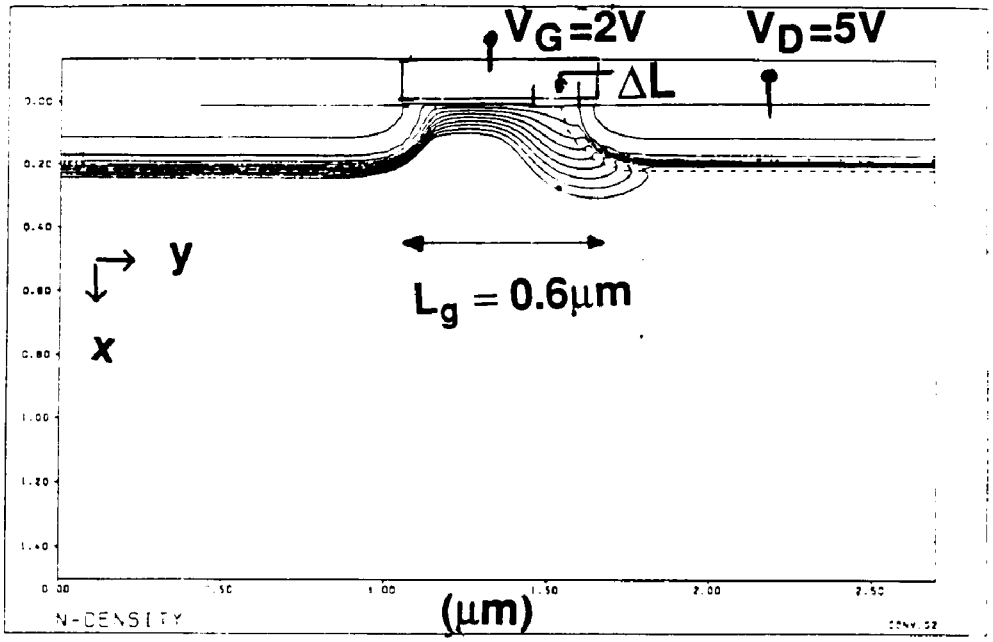


Figure 2-5. MOS device operates in saturation mode.

$$\Delta L = \sqrt{\frac{2\varepsilon_s}{qN_A}(V_{DS} - V_{Dsat})} \tag{2.118}$$

In a long channel device ΔL is negligible compared with L, therefore the drain current remains constant after saturation, i.e., $I_{DS} = I_{Dsat}$. However, in a short channel case, the drain current is expressed as:

$$I_{DS}(V_{DS}>V_{Dsat}) = I_{Dsat}\left[\frac{L}{L-\Delta L}\right] \tag{2.119}$$

V_{Dsat} for short channel device is also differed from the long channel, and can be derived assuming that the linear drain current Eq. (2.115) is valid when when $V_{DS} = V_{Dsat}$ is the onset of saturation and the channel electric field, $E_{sat} = E_{cr\|}$ is the critical parallel field along the channel length of the device, i.e.

$$V_{Dsat} = L_e E_{sat} \tag{2.120}$$

Substitute Eq. (2.120) for V_{DS} in Eq. (2.115), we can equate:

$$I_{Dsat} = \beta[V_{GS} - V_{tn} - \frac{V_{Dsat}}{2}]V_{Dsat} \tag{2.121}$$

$$= \beta[V_{GS} - V_{tn} - V_{Dsat}]L_e E_{sat} \tag{2.122}$$

Neglect the term $\frac{V_{Dsat}^2}{2}$ and solve for V_{Dsat}, we obtain:

$$V_{Dsat} = \frac{(V_{GS}-V_{tn})L_e E_{sat}}{(V_{GS}-V_{tn})+L_e E_{sat}} \tag{2.123}$$

where E_{sat} is the saturation field, experimentally determined in the range of 4-5×10^4 V/cm. L_e is a function of drain voltage as described by Eqs. (2.18) and (2.19). This result is consistent with the empirical results obtained by Sodini, et al.[56]

In obtaining Eq. (2.123), we assume that the term $\frac{V_{Dsat}^2}{2}$ is negligible. A more rigorous treatment, taking into account the drain-induced barrier lowering effect δ, leads to the expression:[57]

$$V_{Dsat} = L_e E_{sat}\left[\sqrt{1 + \frac{2(V_{GS} - V_{tn})}{(1+\delta)L_e E_{sat}}} - 1\right] \tag{2.124}$$

where δ is given by Eq. (1.108).

When the drain voltage is increased several volts higher that V_{Dsat}, the channel electric field increasing due to $(V_{DS} - V_{Dsat})/\Delta L$ term, (since ΔL is pinched), transistor breakdown can occur. The break-down regime in an MOS device is attributed to the generated hot carriers in the bulk when the device is biased into the high saturation regime. The channel to source is forward biased and the device is turned into a bipolar transistor. The drain-source breakdown voltage due to substrate current generation is one of the critical parameters affecting device scaling and the maximum power supply for sub-half-micron (<0.5μm in L_e) CMOS devices. A full treatment of this subject is found in chapter 6.

Chapter 3

DEVICE DESIGN and PROCESS INTEGRATION

3.1 INTRODUCTION

With the background in MOSFET theory as described in the previous chapter, the transistor is designed based on several input factors. The device design is a part of process integration. Process integration is a multi-discipline field in which individual processes are integrated to fabricate desirable device structures and interconnects and to meet certain circuit performance requirements. Transistor design, process flow and fabrication, technology characterization, and design rules verification are all parts of the integration process. In this chapter, we will concentrate on the transistor design aspects of the integration process.

The process of designing an integrated circuit is rather complex. A device designer is a device physicist who understands the device structures, their operations and limitations. The device designer should also be knowledgeable of processing steps, such as oxidation, lithography, etching etc., and how processing conditions can affect the device performance. An understanding in circuit design is useful to determine how the devices should be used. A good material science background is very helpful in all phases of device and process design.

3.2 APPLICATIONS DRIVEN DEVICE DESIGN

In designing submicron device structures, several factors must be taken into consideration. We will examine the design issues from the circuit applications point of view and working toward the central point of the device design process.

3.2.1 Define the Circuit/System Applications of The Technology

There are generally two approaches to the device design. The first approach is a general-purpose technology where the device performance is adequate for most applications. One example of this category is technology developed for general purpose use, e.g. Application Specific IC's (ASIC) and logic oriented products. This approach is appropriate for the diversified products in low to moderate volume. The second approach is to develop a technology for a

particular product where the device design and technology features are tailored to a certain circuit applications. The notorious example of this category is dynamic RAM, which requires a different device structure for the memory cells as compared with other circuit applications. The transistor for DRAM applications often requires low junction and subthreshold leakage, and low body effect for the access device. Similar leakage considerations are used for designing high-performance low stand-by current Static RAM circuits.

3.2.2 Subthreshold Leakage Requirements

Once the circuit requirement is determined, the threshold voltage of a given device can be determined by the I_{off} limitation and current drive expectation, since the subthreshold slope of the device is not scalable as described in section 2.1. Therefore, the off current and saturation current always work against each other, because to lower the off current, the channel doping has to be higher for a given channel length. This high channel doping leads to lower the mobility due to surface scattering and increase the threshold voltage. The higher threshold voltage leads to lower active current drive, thus, degrading circuit performance. However, the analysis of the off current for the device has another advantage in that punch-through prevention in short channel devices can be accomplished by controlling I_{off}, since punch-through is simply an excessive conduction in the off state.

The off current depends strongly on circuit applications. For example, in a DRAM circuit, the access transistor leakage determines the charge hold time in the cell, and as a consequence, the refreshing time of the whole circuit. In SRAM design, the subthreshold leakage of pull-down and access devices affect the standby current in the chip with high density, e.g. greater than 1M bits. However, in logic chips such as microprocessors, or Digital Signal Processors (DSP's), where the active cycles are dominant, the I_{off} requirement is less stringent, leading to higher current drive. In some applications, multiple threshold voltages are preferred to ease in the circuit design. The use of the devices in circuits, whether the product is a memory (Static or Dynamic), a complex microprocessor chip, a gate array, or an ASIC chip, dictates the device design. Moreover, the way the chip is operated also affects the device design constraints. For example, the standby current of a static memory operated by battery highly influences the off current and

junction leakage of the device. At the other extreme, processor chips that constantly operate at high frequency clock rate are more tolerable to the leakage current, and hence current drive performance is more important than the off state limits. A typical I_{off} specification is 1 nA/μm width of transistor, at 125°C junction temperature. However, the high operating temperature due to high clock rate of the processor chips often affect circuit performance and reliability. Finally, the cost/performance expectations from the technology necessitate the decision whether circuit design using the existing technology is sufficient or a new, higher performance technology has to be developed.

3.3 PROCESSING TECHNOLOGY ASSESSMENT

Once aware of the design objective, the next crucial step is to assess the processing capabilities. To realize the submicron MOS device structures, the most important processing steps are the pattern definition (lithography), pattern transfer (etching), source-drain junction formation, low temperature and RTP (Rapid Thermal Processing), gate oxide growth, and interconnect materials. Except for the interconnect which is highly structural and material oriented, we will discuss the factors that directly influence the submicron transistor design.

3.3.1 Lithography Capability

The pattern definition or lithography is the process in which the circuit layout features are printed on the wafers in the form of photoresist. Photoresist is a photo-sensitive organic material, which is polymerized when exposed to the short wave length light source. The exposed area is then developed by a solvent leaving an opening for pattern etching and/or implantation. The minimum resolution, or the smallest reproducible, of the defined feature is dependent on both photoresist property and the wave length of the exposure source. At the present time, the optical lithography stepper using g-line (λ=456nm) and i-line (λ=365nm) light source and lenses, are capable of defining 0.8 and 0.5μm features, respectively, in the manufacture of the devices. More advanced systems such as deep UV (using excimer laser), direct electron beam writing, and X-ray are popular in research laboratories. The deep UV source with the wavelength of 248 nm is capable of producing feature sizes with resolution in the 0.3-0.4μm range and is expected to be

used for the next generation of technologies. I-line optical lithography is expected to reach a 0.35μm resolution with a more advanced resist.

3.3.2 Pattern Transfer

The pattern transfer or etching processes essentially create the permanent features on the wafers from the defined photoresist patterns. In a typical CMOS process (described in more details in Chapter 4), several etching steps are required. Among them, the critical pattern transfer in a submicron device is in etching the transistor gate which the MOS transistor characteristics strongly depend upon. Other etching steps which define the active areas, contact windows, and metal conductors, are no less important. Dry etch processes, using reactive ion plasma techniques, are used almost exclusively to define features at half a micrometer reliably and controllably in production.

In the literature, MOSFETS with lengths as short as 0.1μm have been fabricated using direct E-beam writing on photoresist.[58] However, these fabricated devices are discrete in nature and not integrated into a full functional circuit form. The high performance circuit and systems require small dimension in every design rule: from transistor gate length to contact window size, from isolation of active device areas to metal line/spacing, etc. The final pitch of the transistor, i.e. how close the two transistors can be placed together, is often used as a figure of merit for a particular set of design rules.

3.3.3 Isolation processes

Device isolation affects packing density of regular arrays like memory cell or cell array circuits. As discussed earlier, the most common scheme for intra-tub isolation is the conventional LOCOS (Local Oxidation), where a sandwich layer of pad oxide and nitride serves as an oxidation mask during field oxide growth. As the device dimension is reduced to submicron, the lateral encroachment of the "bird's beak", the gradual transition region from active area to field oxide, is comparable to the the active device geometry. Therefore, more advanced isolation schemes, include shallow trench for intra-tub isolation and variations of poly buffered LOCOS, have been proposed for smaller design rule technologies. More details of isolation techniques will be covered in chapter 8. For the conventional LOCOS and Poly-Buffer LOCOS, the

availability of high pressure oxidation furnace would help in alleviating the segregation effect of the boron doping in p-well during the field oxidation. The lateral diffusion of boron and phosphorus across the well-boundary during the long atmospheric field oxidation is also greatly reduced with the use of high pressure oxidation, hence the inter-tub isolation design rules can be optimized.

3.3.4 Other Processing Capabilities

Other processing capabilities should be considered such as refractive metal deposition (e.g. Titanium, Tantalum, Cobalt) for self-aligned silicided source- drain-gate, rapid thermal processing (RTP) for dopant activation, high energy implants (>200KeV) for well formation or field implant, and selective epitaxial deposition for isolated well formation and/or raised source-drain MOS structure. Some of these processing steps can introduce unwanted defects. Faceting and dislocation in selective epi, lattice slipping in high temperature RTA, and bulk defects by MeV implants are a few examples. Therefore, careful considerations must be made before a process step is committed to the manufacturing of integrated circuits.

3.3.5 Local and Global Interconnect

Moderate resistivity material such as silicide, Titanium Nitride (TiN), or heavily doped polysilicon can be used as local interconnect. A typical scheme for local interconnect is used to strap adjacent source/drain regions of the same or different type by a conductive material. The contact window to the metal conductor can be placed on the local interconnect layer over the field oxide, hence further reducing source/drain area of the active device. A local interconnect layer is very effective in reducing SRAM cell layout size.

The commonly used global interconnect material is aluminum, with a very small percentage of Cu and Si for electromigration considerations. A multilevel (up to 4 layers) metal system is the norm in present-day CMOS technologies. In some CMOS technologies below 0.8μm, with smaller contact window sizes, the sputtered aluminum step coverage on the window wall is poor, therefore, tungsten or polysilicon plugs can be used to fill the contact windows and improve the integrity of the metal system. Tungsten, with higher resistivity compared with aluminum, is being considered for use as the first level interconnect material.[59] ,[60] CVD

(Chemical Vapor Deposition) aluminum[61] and copper are considered as local and global interconnect materials for future technologies.

3.4 DEVICE STRUCTURE DETERMINATION

When all the available processing tools and materials are assessed, the device specifications are established with strong emphasis on reliability. The following is the sequence of considerations to be made when designing devices:

3.4.1 Gate Dielectric Material

The next important physical device parameter is the gate oxide thickness. In practice, the gate oxide can be grown controllably to a thickness as thin as 40 Å. However, the oxide quality, high field breakdown, and charge trapping are the major concerns. The gate oxide thickness, together with the channel length and and threshold voltage constitute the current drive per unit width of the device as described in Eq. (2.117).

The time-dependent dielectric breakdown (TDDB) is mainly governed by the processing conditions of not only the gate oxide growth but also the prior and subsequent processes. The pre-gate oxidation processes include the surface cleanliness and sacrificial oxide growth to remove defects introduced by the field oxidation. During the gate oxidation, the temperature, oxidation ambient and chemicals affect the oxide quality and the charge trapping characteristics. The post-gate processing effects include ion implant charging, corner effects of the patterned gates, the mechanical stress due to material mismatch, and radiation from the plasma etching. All of these processes and effects take part in degrading gate oxide quality. The gate oxide attributes can be examined by several ways. Dielectric breakdown at low field (infant mortality), defect density in a TDDB (Time Dependent Temperature Biased) stress, interface state density, oxide fixed and bulk charges etc. are some measures of gate oxide quality. The problems of hot carrier generation and degradation depend on the drain structure and the gate dielectric quality. For a device with 0.4-0.6μm design rules, the gate oxide thickness is typically in the range of 100-120 Å. Gate oxide quality is a key issue to ensure extended reliability of the devices in operation. Several works[62] ,[63] have been done to improve gate dielectrics quality, in conjunction with improved drain structures, and these subjects will be discussed in chapter 8.

3.4.2 Gate Electrode Material Considerations

n–type-doped polysilicon, so far, is the most commonly used gate material due to its ease of processing and good natural interface with the silicon dioxide. However, the gate to semiconductor work function difference of the n+ doped poly gate requires a high acceptor doping concentration in the channel of NMOS transistor, thereby, lowering the effective electron channel mobility. The availability of tungsten (W) or molybdenum (Mo) would be an improved alternative for a desirable mid-gap, (i.e. Φ_m of the gate is close to E_i/q on Si), work function gate material. However, process instabilities and poor gate-to-oxide interface property of these materials limit their widespread use in practice. For half-micron geometry, p+ doped polysilicon is required as gate material for PMOS devices. We will limit our discussion to the n^+ and p^+ poly gate in Section 3.2.4.1.

3.4.3 S/D Junction Formation

The ability to form shallow source/drain junctions is crucial in device scaling. It is the junction depth that affect the bulk conduction during the off state, channel doping concentration, the overlap capacitance between the drain and the gate that can slow down switching speed due to the Miller effect. For the leakage and punch-through control, shallow junctions are more desirable. However, the shallow junctions increase the channel electric field and enhance the hot carrier generation in the bulk and injection into the gate oxide. Therefore drain engineering to suppress the hot carrier degradation is an important part of device design.

3.5 PRACTICAL DEVICE DESIGN CONSIDERATIONS

We will now concentrate qualitatively on the trade-offs of different device structures, assuming the twin-well foundation is generally accepted as an optimum choice for submicron CMOS technology.

The fundamental question to be answered before designing the deep submicron (less than 0.6μm) device is the maximum operating voltage the MOS devices will withstand. Most of the circuits still operate at 5V power supply down to the channel length of 0.6 μm. In some cases, on-chip voltage regulator has been used to reduce supplying voltage for internal devices of the

circuits. When the operating voltage is determined, (or to be determined when a flexible voltage regulator is used), the device parameters can be specified and predicted, before the fabrication begins. Process and device simulators have been found to be useful in aiding the device design and integration process.

The followings are device parameters chosen for an illustrational 0.5μm CMOS technology using a supply voltage of 4V.

3.5.1 Polysilicon Gate Doping

For the device dimension of ≥0.8μm in channel, the use of N+ doped polysilicon gate for both N and PMOS is the most popular. Figure 3-1 shows the band diagram of n+ polysilicon gate on p and n type substrates.

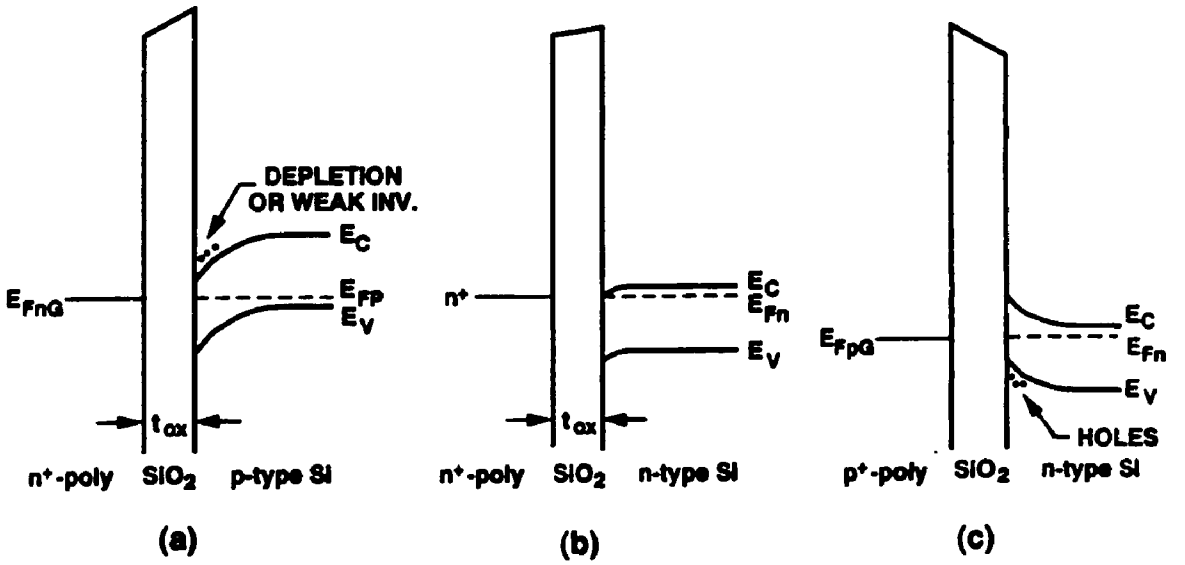


Figure 3-1. Band bending of n+ Polysilicon gate on p- and n-type Si.

Due to the Fermi level of a heavily doped n-type polysilicon being close to the conduction band edge, the work function difference to the p-type substrate (NMOS device) is approximately close

to $E_g/2 + \phi_{Fp}$ or ~-0.9 to -0.95 eV. The band bending when there is no applied gate voltage to line up the Fermi level at the gate and E_{Fp} causes the p-type silicon surface to be depleted (Fig. 3-1a). For an NMOS device an adequate surface concentration is required to obtain reasonable threshold voltage and off current. On the other hand, the PMOS device surface is normally in accumulation at zero gate voltage (Fig. 3-1b), more applied gate voltage is needed to bring the device into inversion, i.e. undesirably high threshold voltage. To rectify this problem, a surface counter doping is needed to compensate for the work function difference, compounded to the positive fixed charge Q_f at $Si-SiO_2$ interface which also accumulates electrons in n-tub, in order to bring the p-ch threshold voltage to a reasonable value. A single threshold adjust implant, using boron or BF_2, can be used to increase V_{tn} and decreases V_{tp}. This implant causes the donor concentration at the surface of the p-ch device to be depleted or even type-converted into p-type. The PMOS device, for that reason, is called a buried channel device. The disadvantage of a buried channel device is excessive I_{off} at short channel length, therefore the threshold voltage has to be adjusted to allow for the worst case short channel. This effect will be presented in Section 5.3.4.

For that reason devices with channel lengths under 0.6μm design rules often use N+/P+ poly as gate electrodes for NMOS and PMOS devices respectively. With the P+ poly gate used for p-channel device, the work function difference is shifted by almost +1V, (~ E_g/q). V_{tp} is reduced accordingly, in this case by Φ_{ms} rather than the use of a BF_2 implant, which cause a buried channel behavior. The donor concentration at the surface of n-tub tends to pile up (see Fig. 3-4a), and as indicated above, the positive fixed charge accumulates electrons. With an appropriate concentration of the n-tub, the PMOS transistor can be an enhancement device with a surface channel, without requiring a threshold adjust implant of either type. The use of N+/P+ doped polysilicon gates requires, however, at least an extra masking step to selectively implant the respective gate areas. The implanted gate MOS structures, however, suffer from reliability issues. It is well known that the phosphorus and/or arsenic doped poly gate is a good gettering source for sodium. With boron doped gate, sodium instability in the gate oxide has been observed.[64] The diffusion of boron through thin gate oxide at high temperatures is another problem. Composite gate dielectric material such as Oxide-Nitride-Oxide (ONO) can be used as

an effective boron diffusion barrier.[62],[65] BF_2 implant has been proposed to reduce the boron penetration range in the poly and subsequently diffused to the poly-oxide by an anneal step. However, the presence of fluorine in the device causing the threshold voltage instabilities as reported by several authors.[66] ,[67] ,[68] Therefore low energy (20-25 KeV) boron implant is preferred for p+ gate doping. The gate oxide quality suffers from the charging effects of implantation, and results will be presented in chapter 8.

3.5.2 Minimum and Nominal Gate Length

Assuming the minimum resolution lithography capability is 0.40μm, and nominal gate length can be patterned at 0.5μm. The line width control, or variation, is typically 20% of nominal feature, i.e. 0.10μm (at 3σ), then in the worst case, the minimum channel length to be designed with is 0.4μm. This is the device geometry that we have to insure proper performance and reliability.

3.5.3 Gate Oxide Thickness

For proper control of subthreshold characteristics and proper current drive, the gate oxide must be scaled. Based on published literature of previous CMOS generations, gate oxide thickness of 100Å is chosen. However, this oxide thickness is constrained by reliability issues such gate oxide breakdown, time dependent dielectric breakdown, and hot carrier effects, and can be relaxed or modified to meet the reliability qualification criteria. This parameter is yield and reliability dominated.

3.5.4 Off Current

From Eq. (2.101), the off current, $I_{off} = I_D \mid_{V_G=0}$ is a function of temperature, channel length, and drain voltage. The I_{off} at high operating temperature, 125°C, is typically specified at 1nA/μm width of the device. From the theory developed in chapter 2, I_{off} is an exponential function of small V_D (~ < 4kT), but this dependence diminishes rapidly as V_D gets larger. For a well designed device, I_{off} is a weak function of drain voltage, as it will be shown experimentally in the characterization section of Chapter 5, and that reduction in operating voltage does not alleviate the I_{off} constraint significantly.

3.5.5 Threshold Voltages

Once the I_{off} is specified at a high temperature and translated to the room temperature value, the subthreshold swing, S, can be determined using the oxide thickness and the known interface state density, N_{it}. Typically, S is in the range of 80-90 mV/decade for a well designed device. Using Eq. (2.101), the threshold voltage at $\psi_s=2\phi_F$ can be calculated. When the operating voltage is reduced, a smallest possible V_t is desired, since the current drive is proportional to $V_{GS} - V_{tn,p}$.

In the 0.5μm geometry CMOS devices, the p-channel MOSFET is a surface device in order to control the off current more effectively. Therefore, it is also beneficial from the circuit design standpoints to have $V_{tn} = V_{tp}$ for a symmetrical CMOS design. The inverter transfer characteristics is more centered at $V_{DD}/2$, and the overlapping current, when both n and p devices conduct, is reduced. A value of $V_{tn} = |V_{tp}| = 0.5$V is chosen.

3.5.6 Channel Doping and Body Effect Considerations

Some circuits applications require low body effect coefficient, λ as defined by Eq. (2.51), for use as access gate in DRAM and SRAM cells, or in a dynamic shift registers. However, in order to control I_{off} and punch-through, the channel doping concentration must be sufficiently high. Unfortunately, high doping concentration leads to higher body effects (refer to Eq. (2.51). There are some techniques to lower total channel doping concentration, and yet maintain a good subthreshold characteristic. One of them is called "halo" drain structure, where a higher dose of channel dopant (boron for n-ch, and phosphorus for p-ch) is implanted at higher energy over the source/drain regions after the gate electrode has been defined. The lateral channel electric field is controlled by the halo tip of the deep implant and yet leaving the channel region at a moderate doping (refer to Fig. 3-7).

3.5.7 Source & Drain Structure

As mentioned before, the junction depth is critical in controlling the off current in short channel devices, and yet it also creates high field effects at the drain region. With the inclusion of salicide (self-aligned silicide) process, fabrication of shallow junctions with low leakage is more

crucial. From the scaling theory derived in chapter 2, for a device with minimum gate length of 0.4μm, a junction depth of 0.15-0.2μm is adequate. Figure 3-2 depicts the proposed device geometry for a 0.4 μm MOS device.

For the proposed structure in this work, we choose a junction depth of 0.2μm from the silicon-gate oxide interface, and the optimum lightly doped drain structures for both n&p channels are used. The oxide sidewall spacers are used to form LDD drain and to avoid the undesirable short effective channel length. Given the minimum gate length of 0.4μm, the lateral diffusion distance of the LDD source-drain is less than 0.05μm per side. Therefore, the effective channel length from metallurgical junctions for a minimum device is only 0.3μm for our example. Shown in Fig. 3-2 is the summary of device dimensions of a minimum 0.4μm MOS device.

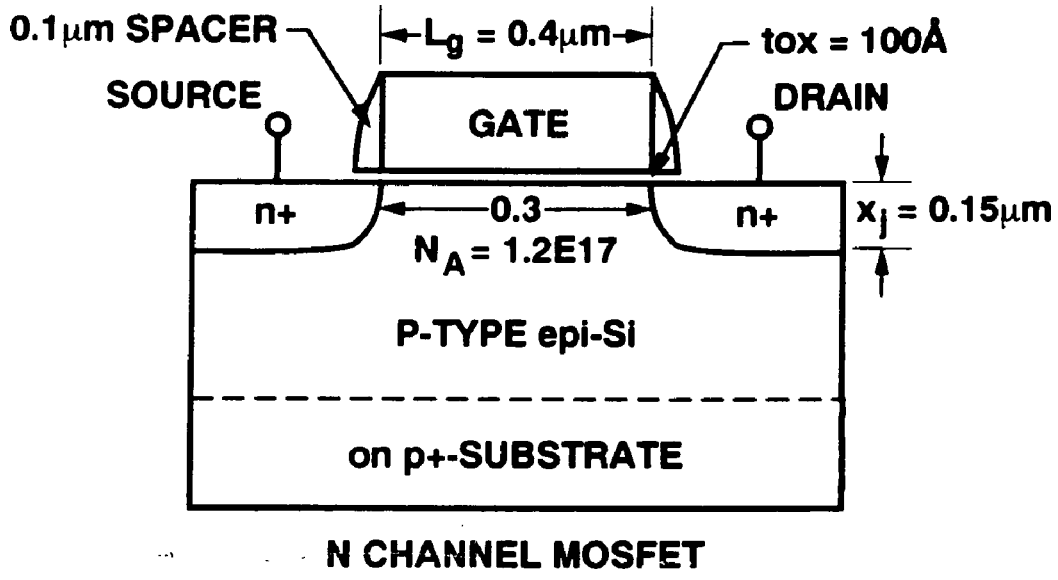


Figure 3-2. Dimensions of a 0.4μm MOS device.

The choice of junction depth and profile must take into considerations of the hot carrier effects. This leads to the subject of drain engineering, which is one of the crucial part of device design

3.6 DRAIN ENGINEERING

In this section we will analyze qualitatively the characteristics of different types of drain structures, with emphasis on n-channel for hot carrier effects. These structures include Conventional, Double Diffused (DDD), Lightly Doped (LDD), Double Implant (or halo), and are shown in Figure 3-3.

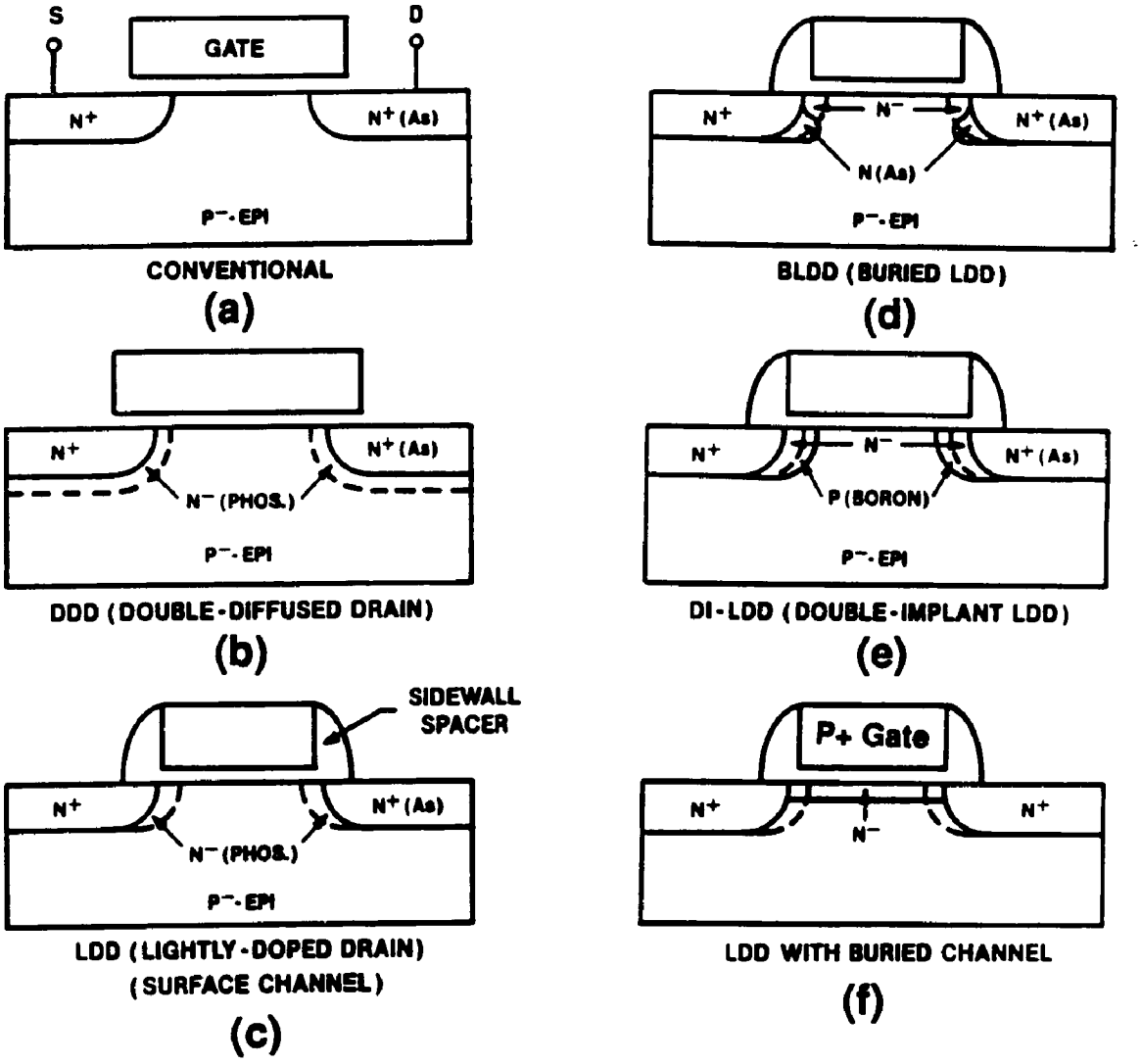


Figure 3-3. Drain Structures for Hot Carriers Analysis.

3.6.1 Conventional Drain

The conventional shallow drain is formed by implanting arsenic with high dose (7E15-1E16 cm^{-3}) for NMOS and boron di-fluoride (BF_2) for PMOS, after the polysilicon gate is etched. Due to the low diffusion coefficient of arsenic, NMOS arsenic source-drain junction profile is typically very steep and shallow, thus creating very high electric field at the drain when the device is operated in saturation mode. The conventional drain was used for devices with channel lengths of 2.0μm or longer. Although, performance degradation in NMOS DRAM circuits was observed at channel lengths >2.0μm using conventional drain structures.[69]

3.6.2 Double Diffused Drain (DDD)

When the channel length was reduced below 2.0μm, the conventional drain structure could no longer tolerate hot carrier drift. A moderate dose (~ 1E14 cm^{-3}) of phosphorus is implanted in addition to arsenic to obtain a more gradual junction profile. Since the phosphorus is diffused much faster than arsenic, the channel region is adjoined with the drain by a more gentle profile of the phosphorus. Although DDD device is simple to fabricate, the more lateral diffusion of phosphorus severes the short channel effect, forcing the poly gate length to be increased. The gate-to-drain overlap capacitance is large, thus degrading ac performance due to Miller coupling effect. The phosphorus diffusion under the field oxide also compromises the isolation distance between adjacent active regions, therefore packing density is compromised. The hot carrier suppression of DDD devices is generally effective for channel lengths in the range of 1.2 to 2.0 μm, as will be shown in chapter 7. However, for submicron channel length device, the LDD device is proved to be more robust.

3.6.3 Lightly Doped Drain (LDD)

LDD CMOS devices was first proposed by Ogura [70] and their derivatives, namely MDD,[71] BLDD,[72] are the subjects for more extensive research and engineering for submicron geometries.

The basic structure of an LDD NMOS is shown in Fig. 3-3c. The lightly doped portion of the drain is formed by a phosphorus implant after gate etch. A low temperature oxide (TEOS)

layer is then deposited and anisotropically etched back to form the spacer of the thickness from 0.2 to 0.35μm. The N+ implant with arsenic is subsequently followed.

The choice of the phosphorus dose, and the spacer thickness affect the device in both active performance and high field effects. We have determined that[50] for the channel length of 0.7 to 1.0μm the phosphorus dose of 4E13cm^{-2}, resulting in a surface concentration of 2E18cm^{-3}, and the spacer width of 0.3μm are optimum. The substrate and gate currents of LDD and DDD devices will be presented in chapter 6.

In the active mode performance, the lightly doped source-drain added the series resistance components to the channel current. The on current is reduced by 15% with optimized $n-$ dose, compared with conventional device with the same effective channel length. However, the reduce gate-drain overlap and smaller coded channel length result in better overall ac performance of the circuit and improve layout density.

3.6.4 Variations of LDD

Other variations of the LDD include BLDD[72] (Fig. 3-3e), where a arsenic is implanted to form a buried layer underneath of phosphorus. However, we have experimented that insignificant improvement in device aging was gained with the BLDD structure. Another form of LDD is an n-ch buried device with p+ poly gate (Fig. 3-3f). The principle advantage of this structure is the conducting channel is below the surface, the generation center is therefore located deeper in the bulk. The hot electron injection mean free path is longer, thus, less damaging to the interface due to hot carriers. This structure is not scaled well as far as subthreshold conduction control, and the problems associated with the p+ gate as discussed earlier.

In addition to the process complexity to form the spacer for an LDD device, during the RSE etch of the deposited TEOS there is no etch stop between the oxide spacer and the field oxide. 50% overetch is usually used to ensure source-drain areas are cleared of oxdie. This TEOS overetch will thin the open field oxide (not protected by gate) and thus creating problems in device isolation and dielectric voidings. Modifications of the oxide spacer has been proposed using a disposable polysilicon spacer,[73] nitride spacer,[74] and L-shape nitride as etch stop for oxide spacer etc.. All of these alternatives have advantages and disadvantages in device

fabrication and performance.

In the next section, we will use process and device simulators to evaluate different device structures.

3.7 PROCESS AND DEVICE SIMULATION

For a complex set of variables and constraints for device design as outlined above, analytical tools are not sufficient to design an optimal structure. The problems of submicron semiconductor devices are highly 2-dimensional, and even 3-dimensional. Therefore, numerical solutions of the device structures in 2D (or 3D, if available) are necessary. With the advents in supercomputer, numerical methods, and knowledge in process and device modeling, the use of numerical simulators is indispensable in the process design and integration. The process modeling is by far the most important and most difficult. The device modeling's accuracy depends on the input of doping profiles and geometry from the output of the process simulator. The physics of semiconductor device is well understood for reliable device modeling. However, in the area of hot carrier generation and injection into gate oxide requires more understanding in energy transport. [75]

We shall investigate the use of process and device simulation tools to design the submicron device geometry based on the conventional knowledge developed in the previous chapter. The active device operations will be investigated in this chapter.

The process simulator BICEPS (Bell IC Engineering Process Simulator) developed by Penumelli.[76] is primarily used for process modeling in this dissertation. Other process simulators include SUPREM III and IV,[77] PREDICT[78] are commonly used. Medusa, a device/circuit simulator developed at Univ. of Aachen, is used for basic device parameters, i.e. V_t, I_{on}, I_{off}. The device simulator Galene 2,[79] which has energy transport equations,[80] is used for hot carrier analysis and substrate current prediction. PISCES IIB[81] is a device simulator using triangular grid structures, capable of handling non-planar structures in bipolar and/or CMOS, can be used for bipolar and latch-up analysis.

3.7.1 BICEPS Process Simulator

BICEPS, originally developed by Penumelli[76] is a two dimensional process simulator, using finite difference discretization method for the grid structures. The program capabilities include:

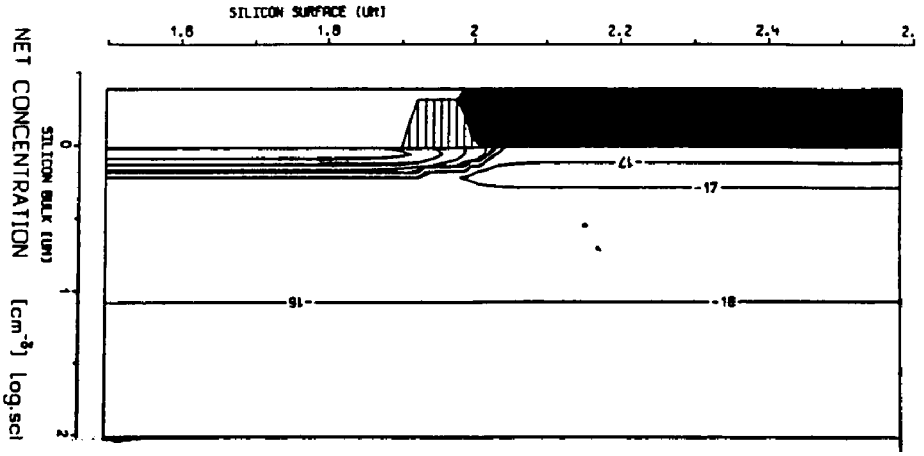
1. Diffusion and Oxidation with dry, wet, HCl oxidant ambient, and partial, atmospheric, and high pressure.
2. Implantation: with arsenic, phosphorus, boron, BF_2, and antimony species.
3. Deposition and Etching of various materials: Oxide, nitride, poly, photoresist.
4. Epitaxial Growth of silicon layers.

All of the above features can be used to simulate 2-dimensional structures of CMOS and bipolar devices. The doping profile of the process simulator can be used for device simulation, or just simply verify some processing steps.

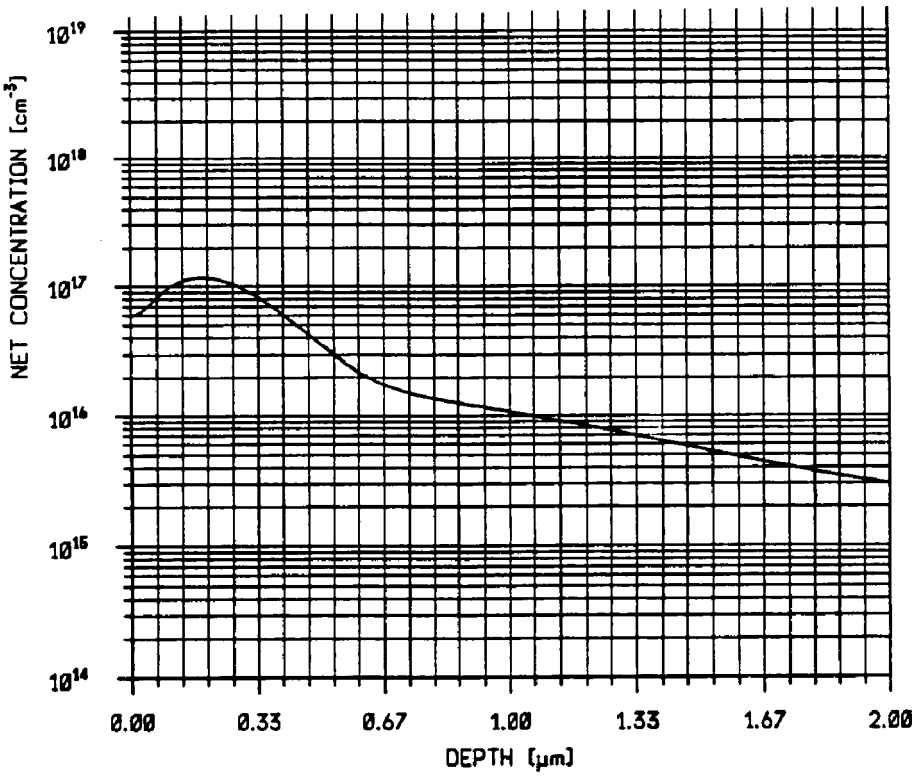
CMOS devices, isolated by a fairly large distance (~ 4μm) across the tub boundaries can be simulated in smaller sections, and then merged together by a device simulator. The n- and p-ch devices are simulated in 2-D and 1-D doping profiles at the middle of the channel and source-drain junction are plotted, Figs. 3-4 and 3-5 for n- and p-ch, respectively. The 2-D tub profiles across tub boundaries is shown in Fig. 3-6, indicating that no lateral diffusion of either dopants into the adjacent tub.

3.7.2 Device Simulation

To predict device characteristics from the process integration stand point, the use of numerical simulator is unavoidable. Although analytical expressions are important in the understanding of device physics, and first order prediction of simple parameters such as threshold voltage and transconductance of large geometry devices. For smaller devices, which are highly 2- and 3- dimensional in nature during processing and operation, a detailed knowledge of doping profiles, lateral diffusion, non-planar structures etc. is needed to provide as a input for a numerical solution to predict device performance.

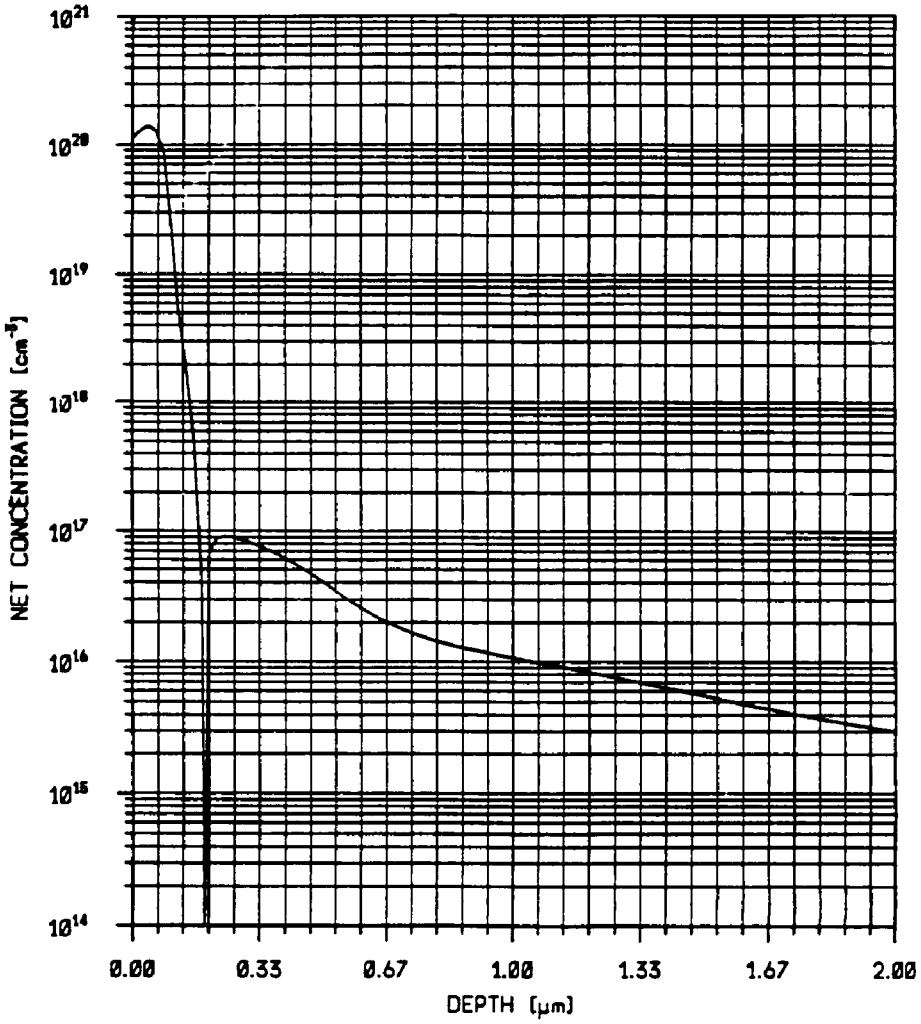


(a)



(b)

Figure 3-4. NMOS doping profiles: (a) 2-D profile of s/d and channel; (b) 1-D profile in the channel; (c) 1-D s/d junction profile.



(c)

Figure 3-4c. n-channel 1-D source/drain junction profile.

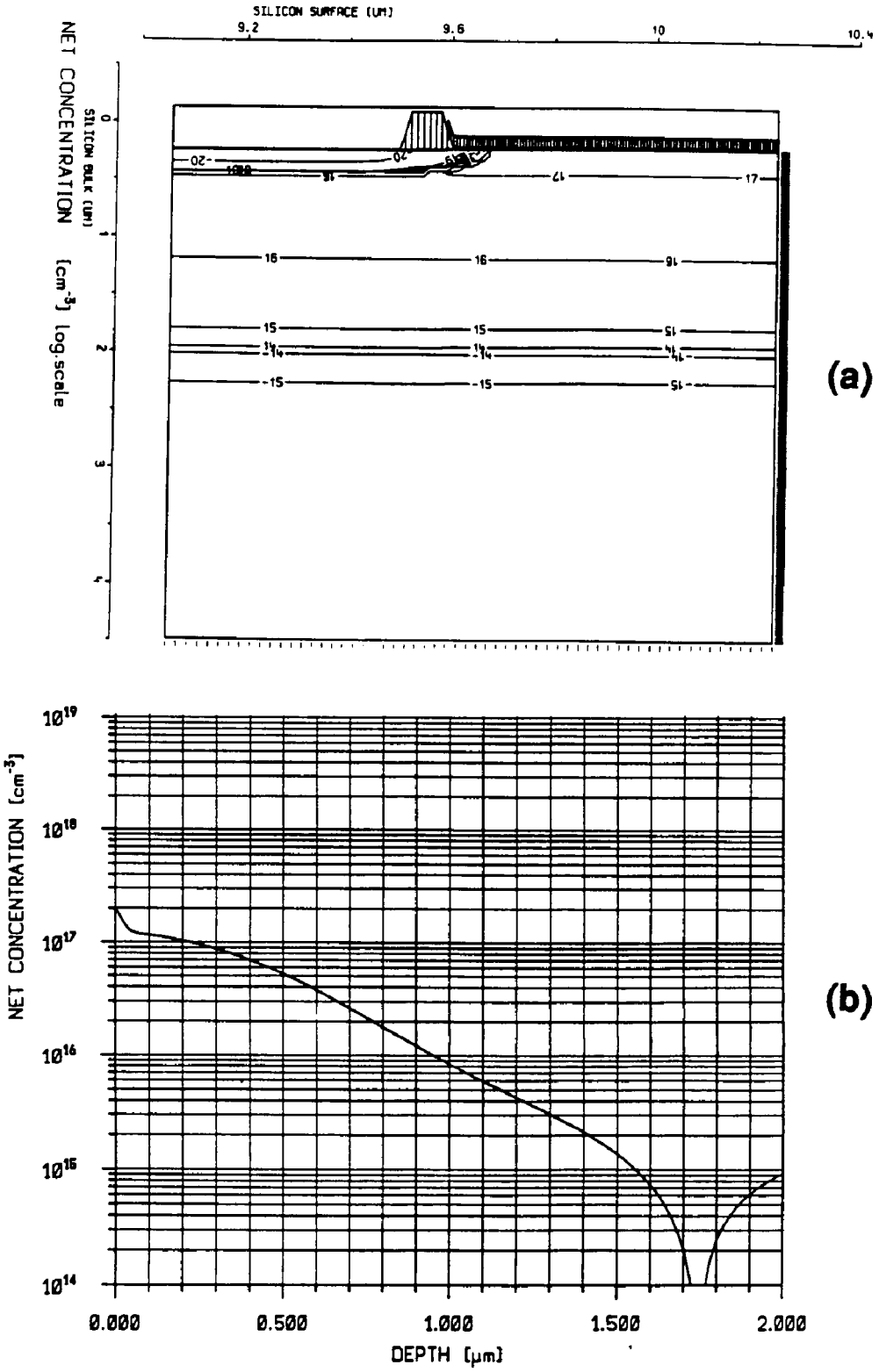
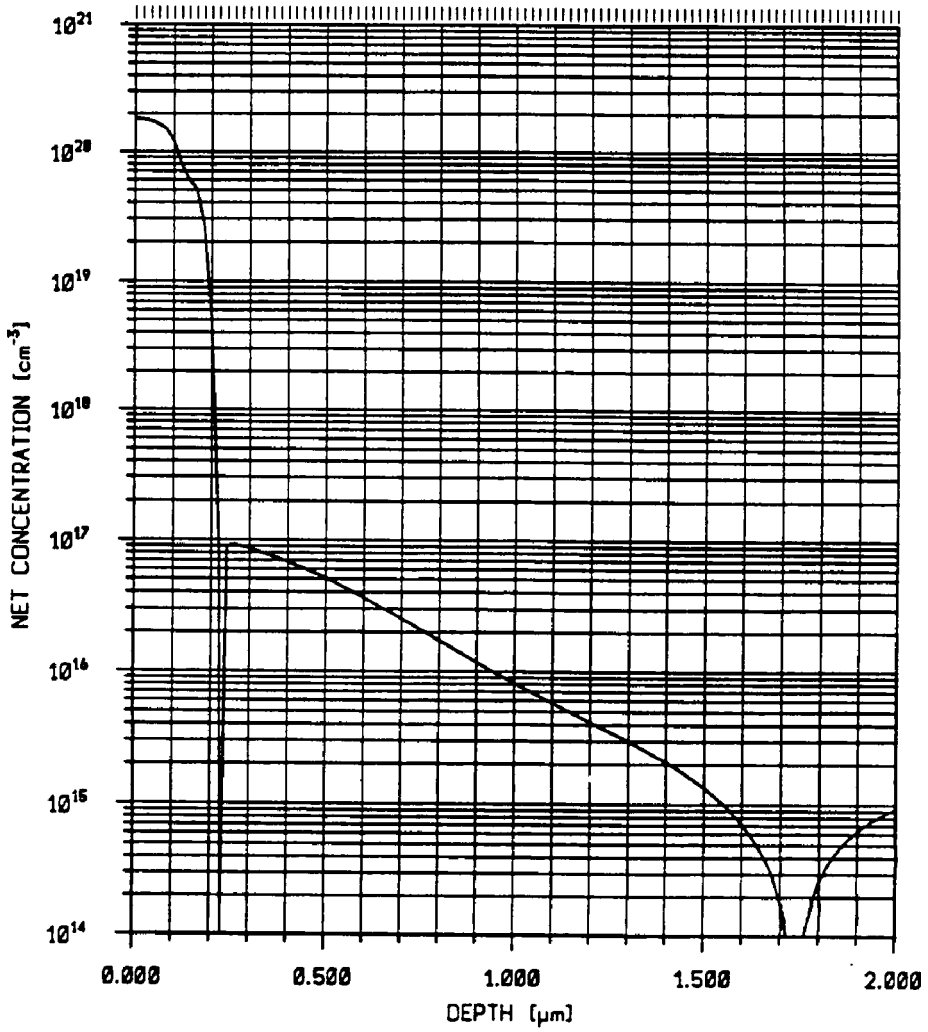


Figure 3-5. PMOS doping profiles: (a) 2-D profile of s/d and channel; (b) 1-D profile in the channel; (c) 1-D s/d junction profile.



(c)

Figure 3-5c. p-channel 1-D source/drain junction profile.

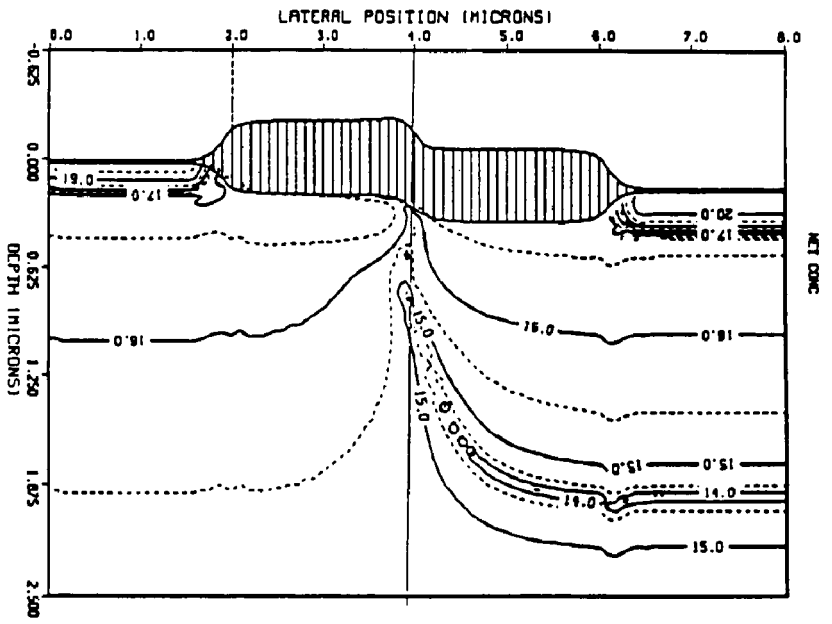


Figure 3-6. 2-D Profile of dopant concentrations at the tub boundary for an advanced twin-tub CMOS using high pressure oxidation for field isolation.

The basic principle of device simulator is to accept input doping profile and geometrical information from process simulator such as BICEPS. The device structure is then discretized into grid points of localized potential, impurity concentration. The boundary conditions are the external applied potential, the known charge distribution, i.e. fixed charge or interface states.

The device simulators solve for the solutions of Poisson and current continuity equations of the forms:

$$\nabla \cdot E = -\nabla^2 \phi = \frac{\rho}{\varepsilon_s} \tag{3.1}$$

$$J_n = q\mu_n nE + qD_n \nabla n \tag{3.2}$$

$$J_p = q\mu_p pE - qD_p \nabla p \tag{3.3}$$

where ρ the total space charge, assuming total ionization of dopant impurities, given by

$$\rho = q(n - p + N_D^+ - N_A^-) \tag{3.4}$$

and μ_n and μ_p are the electron and hole mobilities and D_n and D_p are the corresponding diffusion coefficients. It is further noted that both mobilities and diffusion coefficients depend on temperature, doping level, and electric field. The use of notation ϕ is to be consistent with our definition of the potential with respect to the bulk, as defined in Eq. (2.3). In most of device simulators, ψ is often used for electrostatic potential. The electron and hole concentrations, n and p, in Eqs. (3.2) and (3.3) can be written in terms of the electrostatic potential, ϕ, and the quasi-Fermi levels, ϕ_n and ϕ_p, using Boltzmann statistics, as:

$$n = n_{ie}\exp\left[\frac{q(\phi - \phi_n)}{kT}\right] \tag{3.5}$$

$$p = n_{ie}\exp\left[\frac{q(\phi_p - \phi)}{kT}\right] \tag{3.6}$$

n_{ie} is the effective intrinsic carrier concentration. n_{ie} is a function of bandgap narrowing in a heavily doped semiconductor, and is expressed empirically as:

$$n_{ie} = n_{io}\exp(\frac{\Delta E}{2kT}) \tag{3.7}$$

For low to moderate impurity doping level, n_{ie} approaches the intrinsic carrier concentration, n_i = 1.45E10cm^{-3} at room temperature.

Using the Einstein relation for diffusion coefficients, $D = \mu\frac{kT}{q}$, we can rewrite the continuity equations, (3.2) and (3.3), using Eqs. (3.5) and (3.6), as

$$J_n = -q\mu_n n\nabla\phi_n \tag{3.8}$$

$$J_p = -q\mu_p p\nabla\phi_p \tag{3.9}$$

The continuity equations for electrons and holes can be expressed as:

$$\nabla\cdot J_n = -q(G - R - \frac{\partial n}{\partial t}) \tag{3.10}$$

$$\nabla\cdot J_p = q(G - R - \frac{\partial p}{\partial t}) \tag{3.11}$$

where G and R are the generation and recombination terms respectively.

The total continuity current equation couples the change in electric field vs time to the

current densities of electrons and holes according to the equation:

$$\nabla \cdot J_T = \nabla \cdot (\varepsilon \frac{\partial E}{\partial t} + J_n + J_p) = 0 \tag{3.12}$$

The total current J_T is the sum of conduction currents J_n and J_p and the displacement current $\varepsilon \partial E / \partial t$.

Solving Eq. (3.12) analytically for time and 2- or 3-dimensional space is almost impossible. The use of supercomputers for numerical solution is becoming more practical for more complicated device structures such as bipolar, CMOS, and even latch-up (equivalent to a thyristor) structure. The details of software implementation and numerical techniques are beyond the scope of this dissertation. However, the use of process and device modeling to design different device structures will be described in the following sections. One caveat: the numerical simulation can be used to analyze the trends of changes in process and device parameters, and therefore should be used as a guide during technology development. Characterization of fabricated devices should be used to calibrate the simulators, since the models used in these programs may not reflect the existing process conditions.

3.8 DEVICE DESIGN USING PROCESS AND DEVICE SIMULATORS

In this section, we will report the results of different proposed submicron CMOS device structures and their trade-offs using numerical process and device simulators. Hot carrier effects will be considered for n-channel devices using energy transport in Galene 2 simulator.[80]

3.8.1 Low Body Effect NMOS Using Halo Drain

In short channel device, controlling punch-through can be accomplished by either using higher doping concentration in the channel or a deep implant of the same type as the bulk dopant (boron for n-ch) at the source and drain, also called halo drain. The high channel doping leads to higher body effect coefficient, λ, described in Eq. (2.51), which is undesirable in circuit applications such as access transistor in DRAM and SRAM cell or transfer gate. The halo drain with deep boron implant after the gate etch allows the channel doping to be low in the bulk and therefore low body effect. A cross section of the halo device structure is shown in Fig. 3-7 and the simulated channel doping profile (Fig. 3-8) shows a steep decrease from a surface

concentration of ~ 8E16cm^{-3} to the background concentration (2E15 cm^{-3} in the bulk. The front end fabrication sequence of the device include a background well concentration with the doping level of 1-2E15cm^{-3}. The field isolation channel stop implant is needed using an additional masking step, since the p-well concentration is not sufficient for field device isolation purpose. After the gate oxide is grown and gate electrode is defined, the boron is implanted with 70KeV in energy and moderate dose of 7.0E12cm^{-2}. The shallower implant for the N-LDD is done at this step followed by the spacer formation and the N+ s/d implant.

From the device performance, the n-ch halo device has the following advantages:

- Low body effect coefficient: as seen evident from channel doping profile in Fig. 3-8, where the background concentration of 2E15 is used at depth of ~ 0.3μm. The change in V_{th} when back gate bias varies from 0 to -4.5V is 0.29V (0.47V to 0.76V), while the I_{off} leakage is 7.21E-10A/μm width at 125°C. The experimental results show in chapter 5 are consistent with the predicted results.
- Coded gate length can be shorter, if lithography allows.
- The effective source/drain junction capacitance (integrated over the operating range) is lower, since the background p-tub concentration is significantly low.

and the disadvantages are:

- Higher channel electric field at the drain, due to the halo implant.
- Higher substrate current and lower drain-source breakdown voltage.
- Lower s/d to tub junction breakdown, due to higher acceptor concentration on the p-tub.
- The control of the device due to the halo-implant shadowing effect.

The hot carrier effects was analyzed using the Galene 2 simulator, which employs the energy transport equations in its implementation. The results show that the hot carrier generation center is pushed close to the surface. Figure 3-9 shows this center is 462Å to the surface, as compared with 486Å in a conventional LDD structure. The peak substrate current is ~ 2 times higher than the conventional LDD structure for the same channel length. This higher substrate current translates into a shorter transconductance degradation lifetime by about an order of magnitude

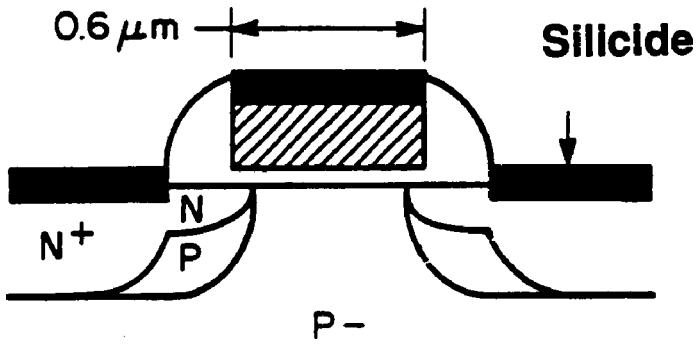


Figure 3-7. Cross Section of a halo-drain NMOS device

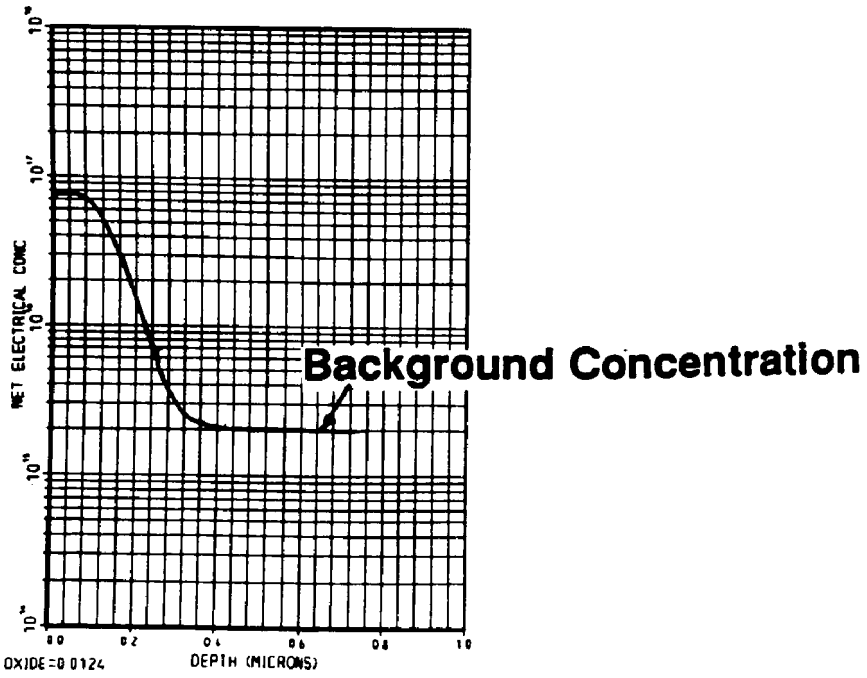


Figure 3-8. 1-D Doping Profile in the channel

($\tau \propto 2^{-3}$). The drain-source breakdown voltage is experimentally verified to be 6.9V for the halo and 8.9V for the convention device, with lower current drive, as shown in Fig. 3-10. Therefore, the scalability of this structure is limited. From the simulated and experimental results shown, the n-ch halo device is not recommended for half-micron NMOS devices operated at 5V supply voltage, when hot carrier degradation is a constraint.

3.8.2 P-Channel Halo-Drain

As discussed earlier, the use of N+/P+ poly gate material for device sizes of 0.5μm or below results in a more scalable structure in terms of I_{off} control. However, the use of single N+ gate for both device types is still preferred for devices in the range of 0.6 to 0.7μm, for process simplicity and to avoid problems associated with P+ gate. We have proposed a modified P-channel with halo drain that uses N+ gate and appropriate off current control at short channel length down to 0.5μm. This structure is attractive for supply voltage at 5.0V since the current drive capability is comparable to that of the surface p-ch device in spite of higher threshold voltage.

The fabrication process requires a protecting layer of nitride or oxide on top of the polysilicon gate. After the nitride/poly sandwich is etched, a deep phosphorus implant at 180KeV is carried out to form a punch-through protection at the source-and-drain. P-ch LDD is implanted with moderate dose of BF_2. Then, the P+ source-drain is formed either by the conventional structure or with salicide. The I_{off} and V_{tp} vs L_{eff} plots are shown in Figs. 3-11 and 3-12. It is noted that the off current is suppressed 4 orders of magnitude compared with conventional structure at the same coded channel length (i.e. 0.7μm), therefore allowing the coded length to reduce to 0.5μm and satisfy the leakage specification of 1nA/μm. The predicted I_{on} is plotted vs coded gate length as shown in Fig. 3-13, indicating that for a nominal gate length device of 0.6μm, I_{on} is 290μA/μm width. The coded length of the p-channel halo is the same as that of the n-ch device, easing the circuit layout and improving circuit speed.

The p-ch halo drain structure is an attractive alternative to maintain the n+ doped gate, and allowing the same n and p-ch gate lengths to be used. The threshold voltage, V_{tp}, however, is rather high, ~ -1.0V, therefore a existing 5V supply voltage is desired for not suffering the current

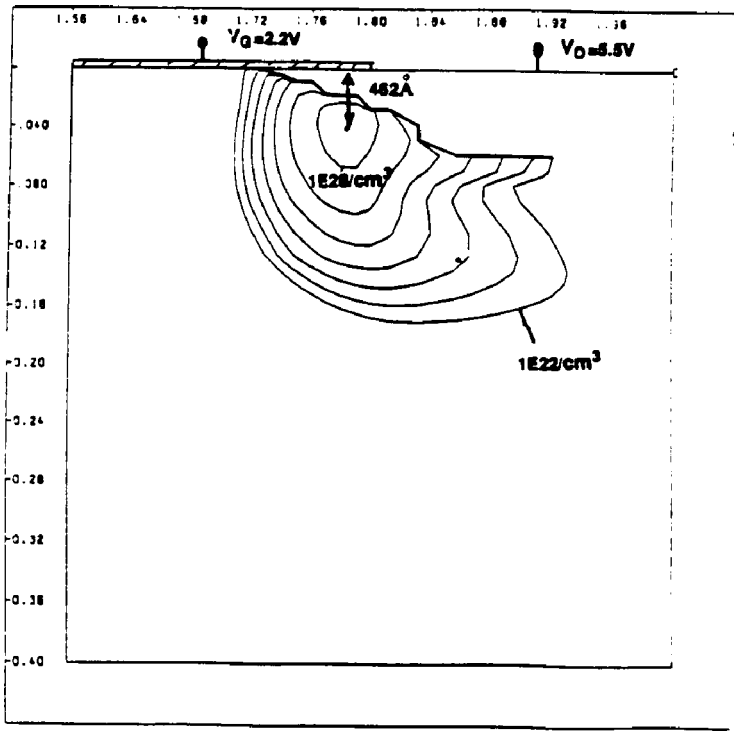


Figure 3-9. Hot carrier generation center in a n-ch halo device.

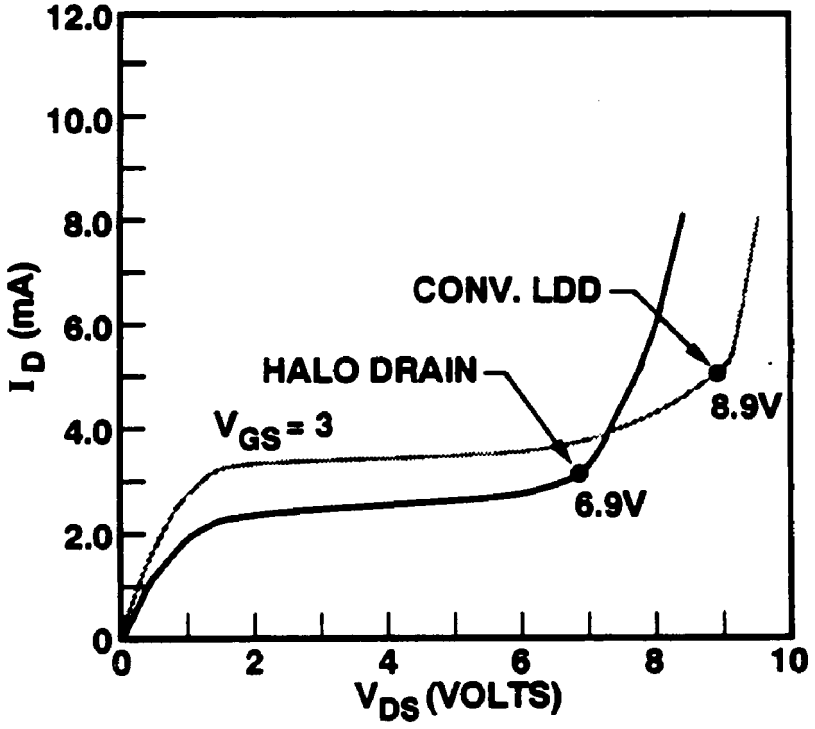


Figure 3-10. Measured Source-Drain Breakdown of conventional LDD and Halo-LDD NMOS Devices with effective channel lengths of 0.8μm.

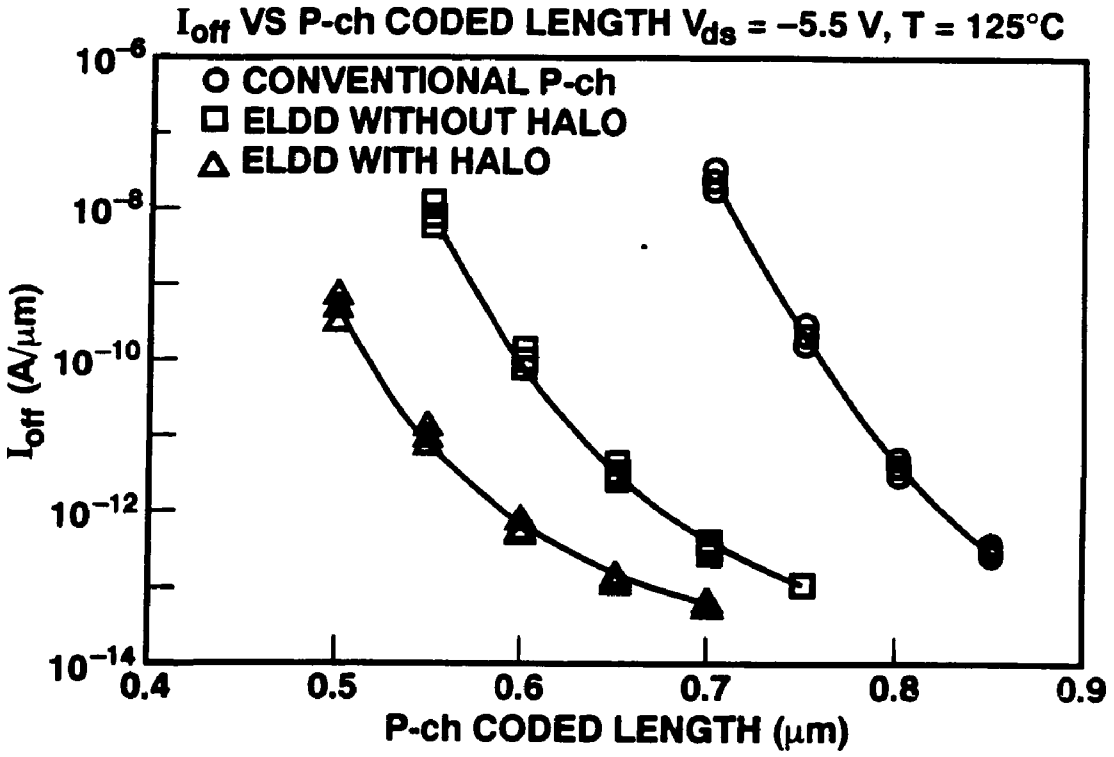


Figure 3-11. I_{off} vs L_{eff} of p-ch halo and conventional devices.

drive. In conclusion, the p-ch halo drain structure is a temporary approach for device geometry > 0.6μm. Further scaling requires the n+/p+ poly or a closer-to-midgap material to be used as gate, for reduced supply voltage technology.

3.9 SUMMARY

In this chapter, we have considered practical approaches to define CMOS device structures to meet desirable circuit and device specifications, and suitable for a given set of processing constraints. The use of process and device simulators have been demonstrated to evaluate alternative structures for n- and p-ch halo drain devices. The performance advantages and drawbacks of each approach can be simulated and analyzed. Proper decision can be made in a timely manner without embarking on expensive and lengthy fabrication processes. We have recommended that the gate material for subhalf micron devices is n+/p+ doped polysilicon for the near future until an established midgap work-function-material is developed.

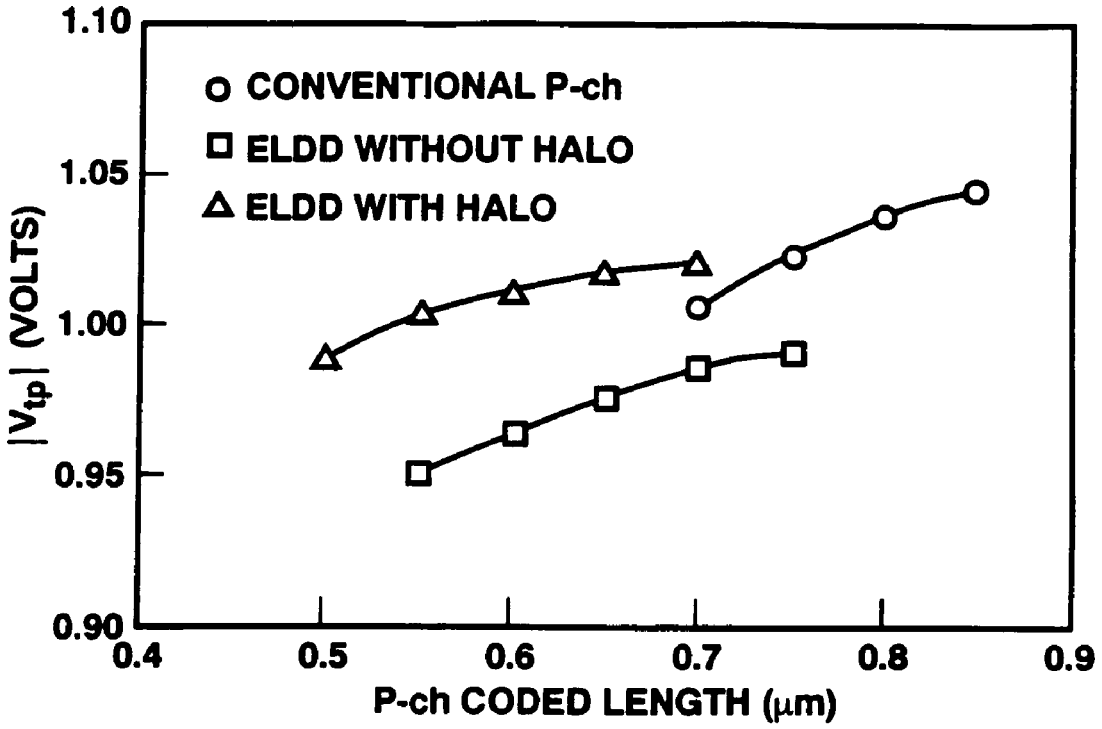


Figure 3-12. V_{tp} vs L_{eff} of p-ch halo and conventional devices.

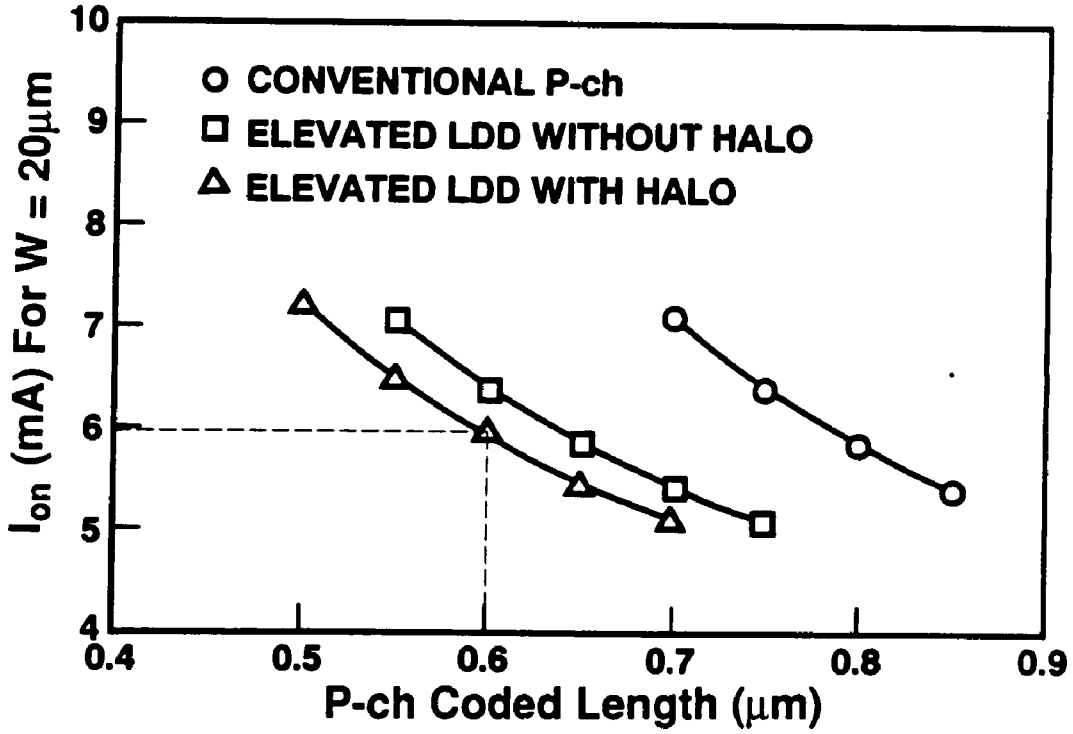


Figure 3-13. I_{on} vs L_{eff} of p-ch halo and conventional devices.

Chapter 4

CMOS FABRICATION PROCESS

4.1 INTRODUCTION

As discussed in section 1.2, the CMOS (Complementary Metal Oxide Semiconductor) technology has evolved from the early 1970's when it served as a low-power, low-speed alternative for bipolar TTL SSI logic applications to the modern high-density, high-speed technology found in dynamic and static RAM's(Random Access Memory), microprocessors and application specific integrated circuits (ASIC). To accomplish this, the process technology evolved from a p-well process, which was derived from NMOS technology for the reasons of process compatibility, to n-well for improved device balance, and finally emerged as the optimized twin-tub structure.[82]

The driving force for the popularity of CMOS technology stemmed from the major drawback of NMOS technology, high power consumption. The high power consumption limited circuit speed and taxed the heat dissipation ability of the operating environment. Moreover, circuits designed in CMOS are less sensitive to process variations and more tolerant to junction leakage than those designed in dynamic NMOS. The objective of this chapter is to describe the twin-tub CMOS process sequence used to fabricate the devices reported in this dissertation.

4.2 TWIN-TUB CMOS FABRICATION SEQUENCE

The self-aligned twin-tub approach, developed at AT&T Bell Laboratories,[82] allows independent tailoring of the individual tub profiles without having either device type suffer from excessive counter-doping effects. The twin-tub approach also allows flexibility in choosing the optimum type of starting material for a particular circuit application because either a lightly doped n- or p-type substrate can be used.

A cross section of twin-tub CMOS structures is shown in Figure 1-8. In this technology a surface n-channel and a buried p-ch MOSFET are integrated with n+ doped polycide ($TiSi_2$ on polysilicon) gate.

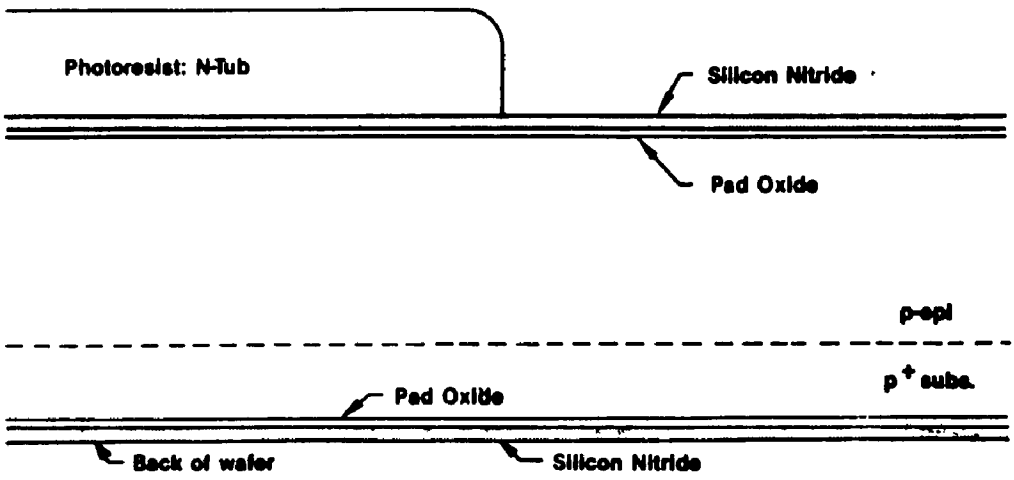


Figure 4-1. N-tub Photo Resist Step.

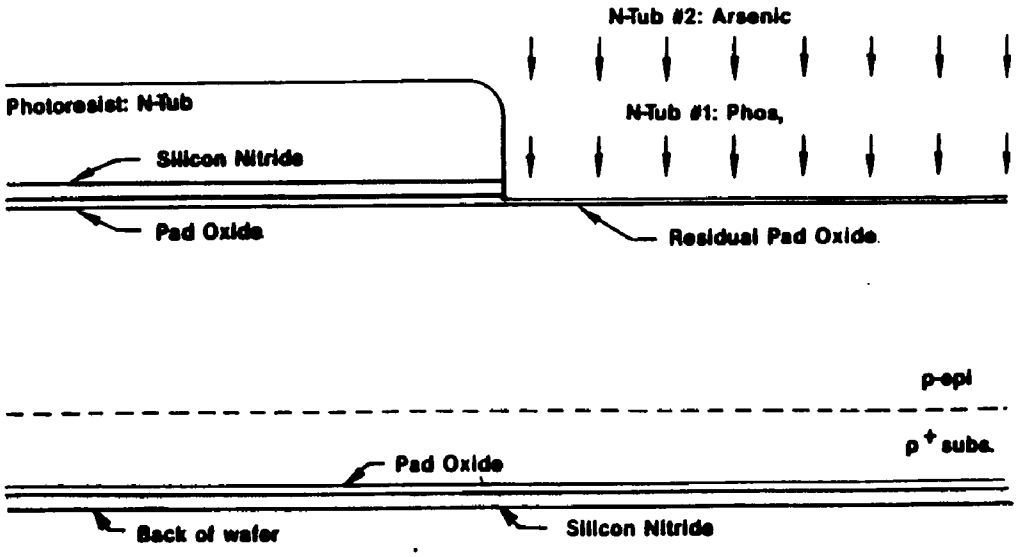


Figure 4-2. N-tub Implant.

The process and device considerations for submicron CMOS technology are numerous.[83] The factors that affect circuit packing density for a given gate length are the device isolation, contact and metalization, and latch-up prevention measures. The isolation of the same device type (intra-tub), i.e. NMOS to NMOS or PMOS to PMOS, is determined by the subthreshold leakage of the parasitic field transistors. In the conventional LOCOS isolation, the minimum spacing between 2 adjacent diffusion regions is controlled by the surface concentration and field oxide thickness in the chanstop (channel stop: the area underneath field oxide and between channel regions of active device) area. However, as the spacing is reduced the barrier lowering at the drain of the field device becomes severe, thus limiting the allowable minimum spacing. Alternative approaches include trench isolation,[84] buried oxide (BOX) isolation,[85] and selective epitaxial growth of active device areas.[86] These techniques also aim at latch-up prevention, but are more prone to the material defects and will not be covered in in the context of this work.

4.3 FABRICATION PROCESS

The process steps to fabricate twin-tub CMOS devices were originally described in ref.[82]. Through evolution and improvements, the modern twin-tub process is slightly different from the original scheme but the basic philosophy remains the same. The process that is described in this section, resembles state-of-the art processing technology and materials characteristic of 0.9 and 0.6μm CMOS technologies.[43] [51] Where applicable, we will attempt to justify the choice of certain processing steps and suggest alternatives that may be a common practice in other technologies.

4.3.1 Starting Material and Tub Formation

The starting material is of lightly doped p-type layer, with the resistivity in the range of 10-20 Ω-cm, epitaxially grown on a p+ substrate wafer. The choice of p+ substrate is to reduce the substrate resistance and thus increasing the source-drain breakdown voltage caused by hot carrier generation, as will be shown in Chapter 6. For DRAM applications, the use of low resistivity p+ layer is known to be effective in collecting α-particles generated from cosmic rays.[87] However, in applications such as SRAMs, an n-type substrate is more effective in reducing single event upsets due to radiation.

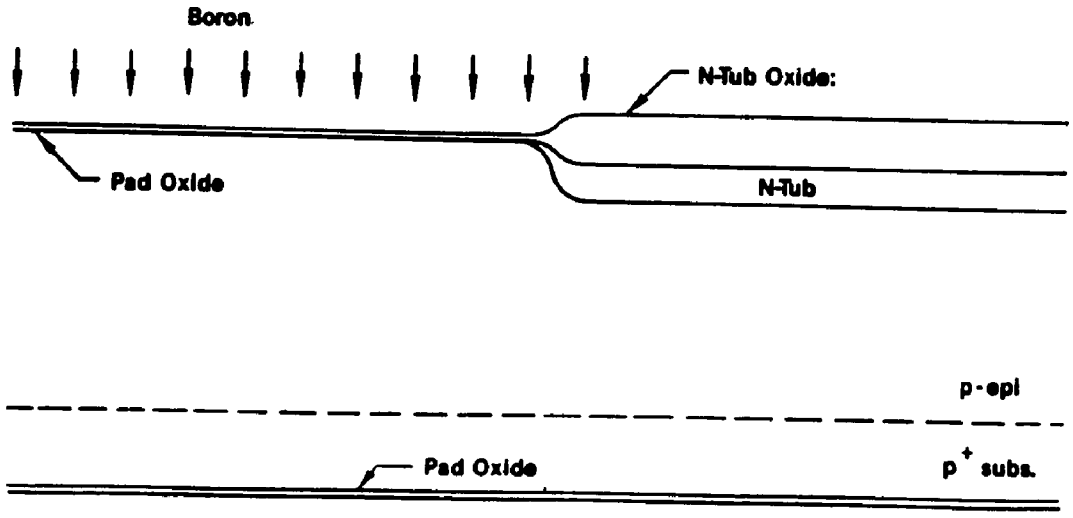


Figure 4-3. N-tub Oxidation and self-aligned P-tub implant.

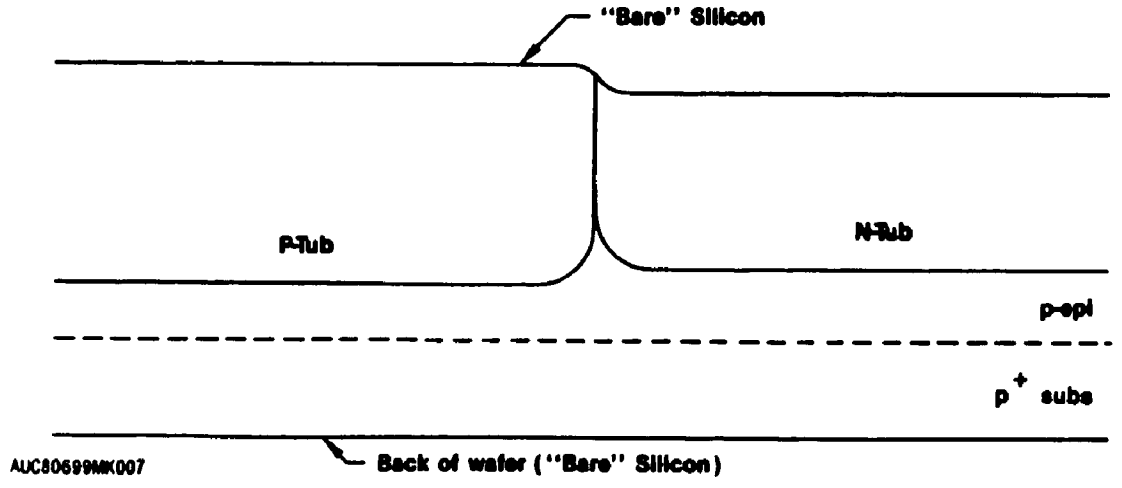


Figure 4-4. Wafer topography after tub formation.

The wafer is first cleaned and etched in a diluted hydrofluoric (HF) acid solution to remove any native oxide on the surface. The initial tub pad oxide is thermally grown to a thickness of 350 Å followed by a chemically vapor deposited (LPCVD) nitride layer of 1200 Å thickness. The nitride layer is used as an oxidation mask for the n-tub oxide, whereas the fairly thick pad oxide is used for stress relief during the local oxidation over n-tub areas. The n-tub regions are defined by a photolithography step, using an optical source for for photoresist exposure in a step and repeat machine. The photoresist is developed leaving the opening in photoresist as shown in Fig. 4-1. The nitride layer is then etched using the pad oxide as an etch stop with some overetch to insure the complete removal of nitride. While the photoresist is still on, Phosphorus and arsenic are implanted into the opening areas (Fig. 4-2).

The photoresist is then removed by an oxygen plasma, followed by wet chemical cleaning steps. 6500 Å of silicon dioxide is grown in a pyrogenic (H_2 and O_2) ambient in n-tub regions which are unprotected by the nitride layer. The main purpose of the n-tub oxide is to be used as blocking oxide for the subsequent self-aligned p-tub implant. With the thickness of 6500Å, this oxide is sufficient to prevent Boron implant energy up to 130 KeV (with range and 3 time standard deviation (σ)) from penetrating to the n-tub silicon surface. This step is also to activate and partially diffuse-in the implanted phosphorus and arsenic impurities (Fig. 4-3). The remaining nitride is removed and the p-tub is implanted with boron whose dose is adjusted to serve as a background concentration to the n-channel device. With the tub oxide still on, both tubs are driven by a high temperature (normally between 1050 to 1150°C) drive-in step resulted in a tub depth of ~2 μm. The final 2-D tub doping profile is shown in Fig. 3-6 from the BICEPS process simulation.[88] An optional second p-tub boron implant is employed to increase the surface and bulk concentration of the active and field isolation of the n-channel device. This step may not be needed if the threshold adjust implant is sufficient to adjust the threshold voltage, and a separate chanstop (or field) implant is used. All of the oxide is removed from the wafer yielding a silicon surface with patterned n and p-tubs separated by a gradual step at the tub boundary (Fig. 4-4). This step is essential for subsequent lithography alignment.

4.3.2 Active Area Definition

The transistors are built on active areas (also known as diffusion, island or thin-ox areas) that are isolated from each other by the field regions. The conventional method to define the active area is by Local Oxidation (LOCOS). It is very similar to that of N-tub formation sequence described in section 4.3.1. A 200 Å pad oxide is grown, followed by an LPCVD nitride film of 1200 Å. A lithography step creates patterns whereby the intended active areas are covered by islands of photoresist (Fig. 4-5). The nitride layer is reactive-ion-etched in NF_3 gas to expose the field oxide-to-be regions. The pad oxide is again used as etch stop layer. It is important to protect the silicon surface from being damaged by the reactive-ion etch. The remaining photoresist is now removed. For the conventional atmospheric field oxidation approach an extra mask would be required to selectively implant the n-chanstop region.

The field oxide is grown by pyrogenic oxidation in high pressure ambient. The advantage of high pressure oxidation is that there is less segregation of the boron into the field oxide and less redistribution of the dopants than during an atmospheric oxidation resulting in the same oxide thickness. Using high pressure oxidation (HiPox) a chanstop masking step can be saved as indicated above. The cross section of the wafer is shown in Fig. 4-6.

4.3.3 Gate Oxide Process

The silicon nitride islands remaining in active area are now removed in hot phosphoric acid. It is well known that the LOCOS process produces a compound of oxynitride of unknown proportion (Si_xN_y)[89] at the edges of the active areas (bird's beak) which will result in a region of thinner oxide in a subsequent oxidation. A sacrificial oxide is grown in wet atmosphere, after the pad oxide is stripped, to remove the Si_xN_y composite. The thickness of this oxide needs to be between 500-800 Å to be effective. The sacrificial oxide is then etched away and a threshold adjust screening oxide is grown to a thickness of 150 in dry ambient. The thickness control of this oxide is critical for the threshold voltage consistency.

A blanket boron implanted is done to adjust the threshold voltage of both n and p devices. The p-channel device is therefore of the buried channel device, as discussed in chapter 3. For a device with channel length less than 0.6μm, the buried device is not effective in controlling

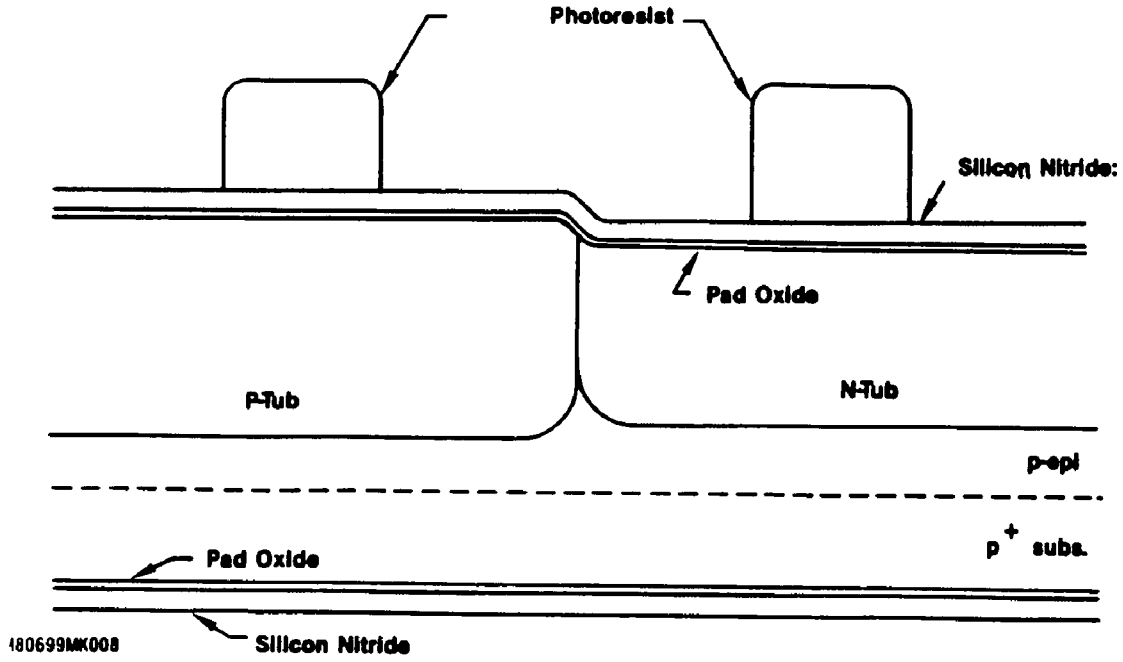


Figure 4-5. Photo Resist Active Area.

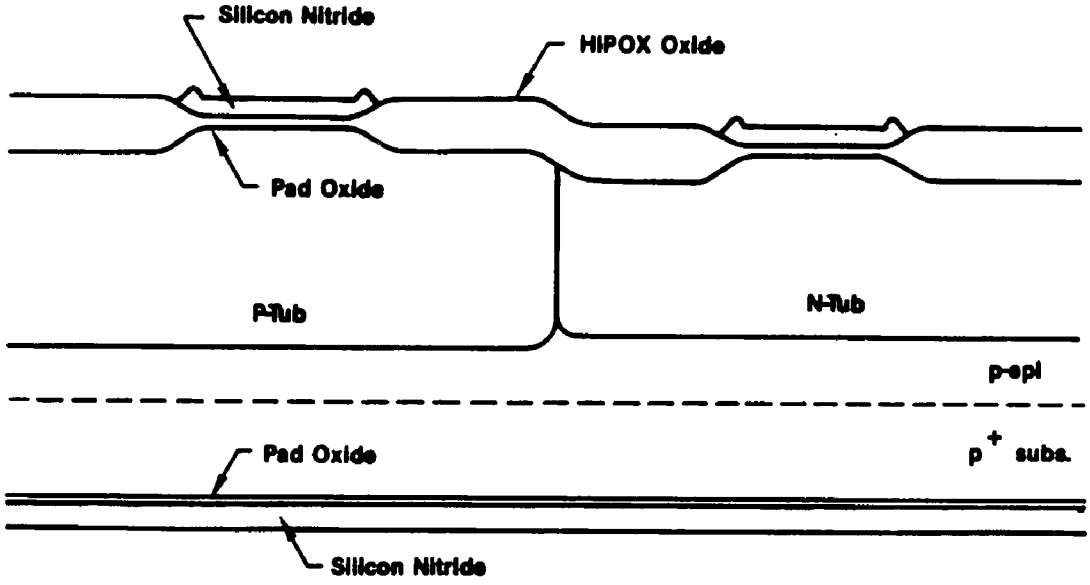


Figure 4-6. Device Cross-Section After Field Oxidation.

subthreshold leakage. In this case a surface p-channel device with p-type doped gate is more appropriate (Section 3.3). In this case the threshold adjust step is optional.

The screen oxide is removed by a diluted hydrofluoric acid solution. The gate oxide is grown in a dry oxygen ambient. Chlorine or Fluorine may be added to the oxygen to improve the quality of the oxide.

4.3.4 Gate Patterning

A 3500 Å polysilicon film is immediately deposited by LPCVD (Low Pressure Chemical Vapor Deposition) at 600°C by pyrolysis of silane (SiH_4). The polysilicon is subsequently doped with phosphorus using PBr_3 in a furnace. This approach will produce an n-type gate for both n and p-ch devices. For subhalf-micron structures with n+/p+ gates, the polysilicon need not be doped before etching since it can be doped by the source/drain implants.

There are two approaches that are commonly used to reduce the sheet resistivity of the polysilicon runners (used for interconnect). The conventional approach is to deposit a layer of refractive material such as tantalum, molybdenum, or tungsten. The gate is then defined by etching the sandwich layer to form the gate. This approach has a major disadvantage in channel length control due to the difficulty in etching the dual layers.

A more elegant technique is the self-aligned silicide (salicide) process in which the polysilicon gate is etched (Fig. 4-7), the lightly doped source/drain implants are done, spacer are formed on the edges of the polysilicon, and a selective silicide is formed on the gate and source/drain.

4.3.5 Source and Drain Formation

There are several approaches to form source/drain depending on the desired drain structures. We will describe the two sequences to fabricate LDD and DDD respectively for n and p MOS (Fig. 4-8). After the polysilicon gate is etched and PR is removed:

1. For LDD: a light dose of phophorus (~ 4E13 cm^{-2}) is implanted nonselectively. An short oxidation step is done to drive in phosphorus into the channel to insure the channel continuity under the gate edges. This consideration is due to the implant shadowing and

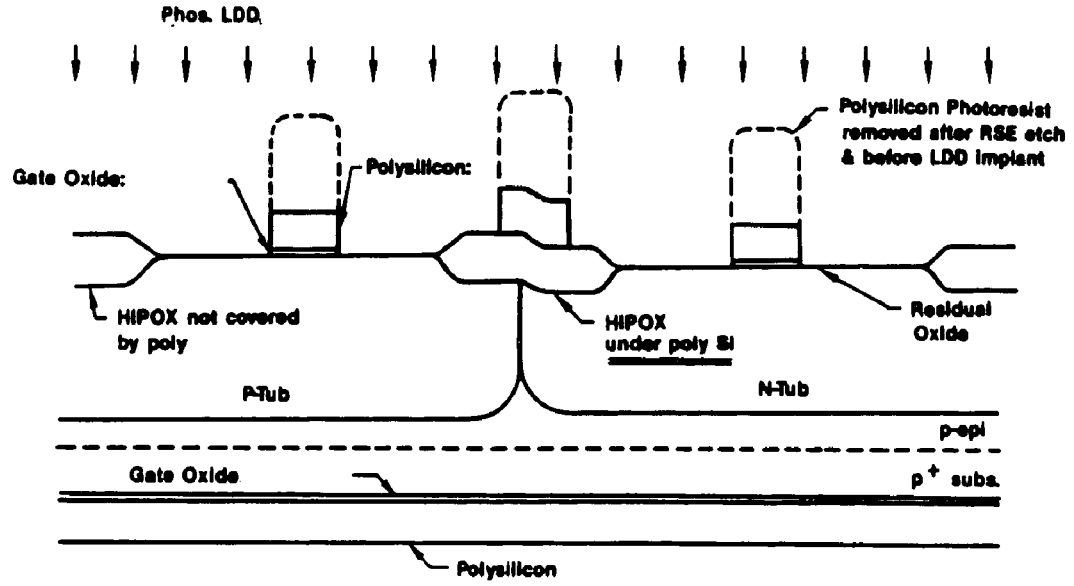


Figure 4-7. Polysilicon Gate Photoresist and Etching.

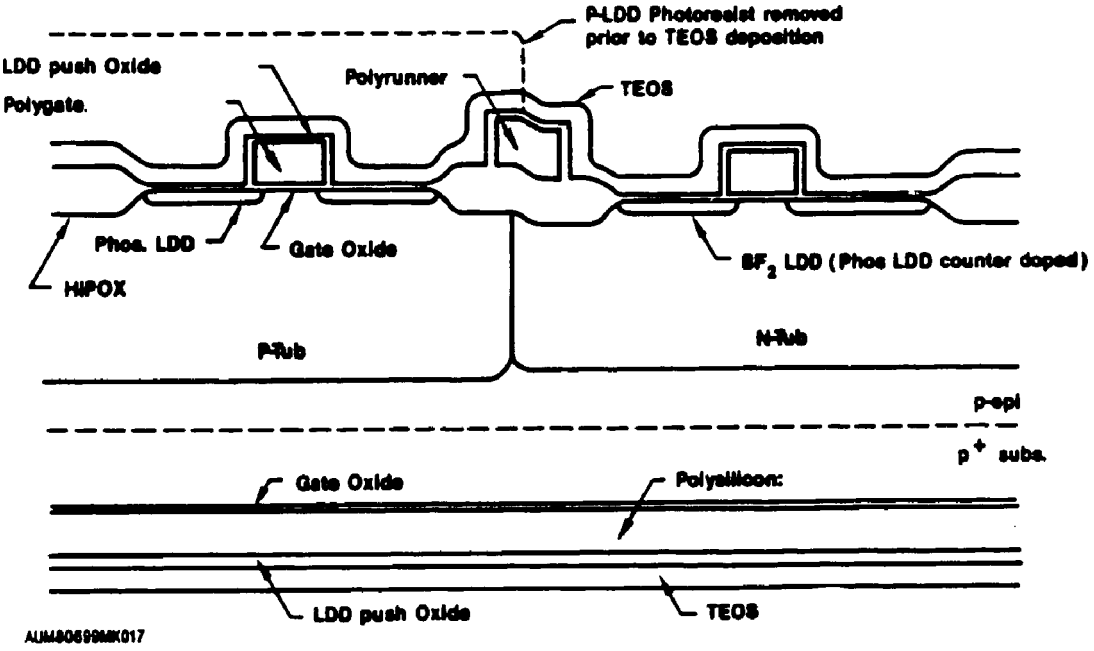


Figure 4-8. Sidewall Spacer Formation.

undercut in gate profile. The oxidation step also grows a sidewall thermal oxide on the sides of the poly gate. An implant mask is used to expose the p-ch area, where a moderate dose (1E14) of BF_2 species is used for PLDD implant. A low temperature oxide is deposited to the thickness of 3500 Å. An anisotropic RSE etch is performed to form the sidewall oxide spacer. The final width at the foot of this spacer is 2500-3000Å for a 5V technology. N+ S/D implant using arsenic is done with a lithography step. One more masking step is needed for P+S/D using BF_2 implant.

2. For DDD: A sidewall spacer is first formed by thermal oxidation to a thickness of 200Å, followed by an RSE etch back of 1000Å deposited TEOS, resulting in a spacer width of ~800Å. A lithography step is used to open the n-channel device active areas. A moderate dose of phosphorus (1E15) and arsenic are implanted into n-channel device to form DDD structure. Another masking step for the P+s/d implant, where BF_2 is implanted to p-ch devices.

The width of the spacer, particular in the case of LDD devices, is critical in optimizing the hot carrier generation as described in ref.[50] and in detail in chapter 6. During the spacer etch, the source and drain silicon surface is exposed. The n-channel source/drain is implanted with a high arsenic dose of 5.0×10^{15} at low energy (80 KeV) using a selective mask. P source/drain is implanted using BF_2 at high dosage (3.0E15) and low energy (50 KeV) to form shallow junctions. The ability to form shallow junction is one of the most critical processing step to realize scaled devices as was discussed in chapter 3. The source/drain junctions are driven-in by an anneal step using nitrogen as the major carrier gas. The wafer is cleaned in a diluted HF solution to remove any grown oxide before the refractive metal deposition (Fig. 4-9).

4.3.6 Self-Align Silicide (Salicide)

The choice of refractive metal depends on many factors: sheet resistance, reactions to certain chemicals, interaction with dopants, and most importantly the adhesion to silicon and polysilicon. For example, titanium is known to react with hydrofluoric acid, while cobalt is not. Cobalt has another advantage that it can be used as a source for out-diffusion (similar to that of polysilicon) to form a very shallow junction. However, cobalt is a magnetic material, and

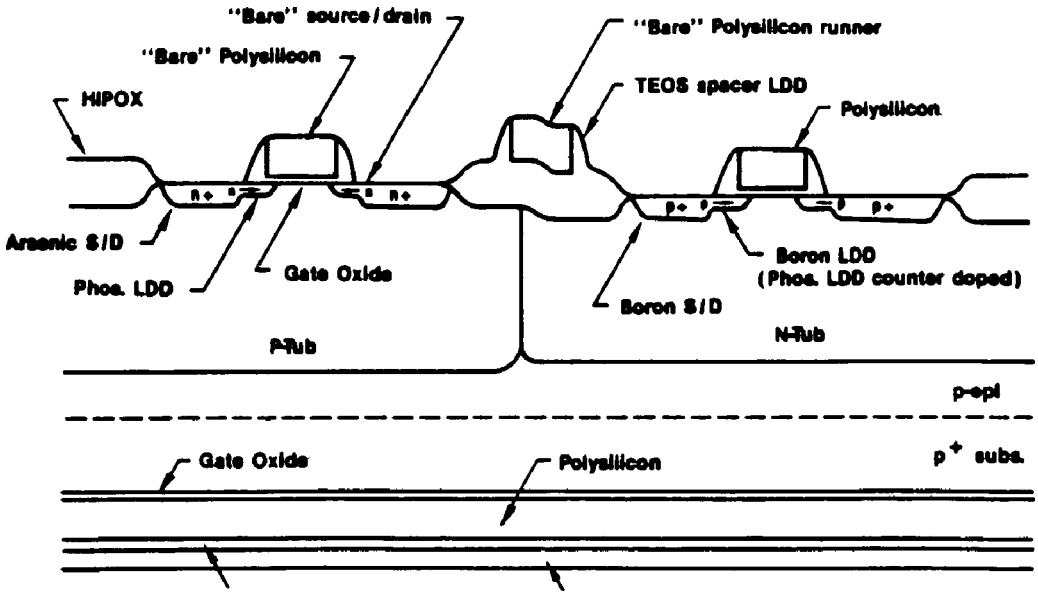


Figure 4-9. Device Structure Before Ti Deposition.

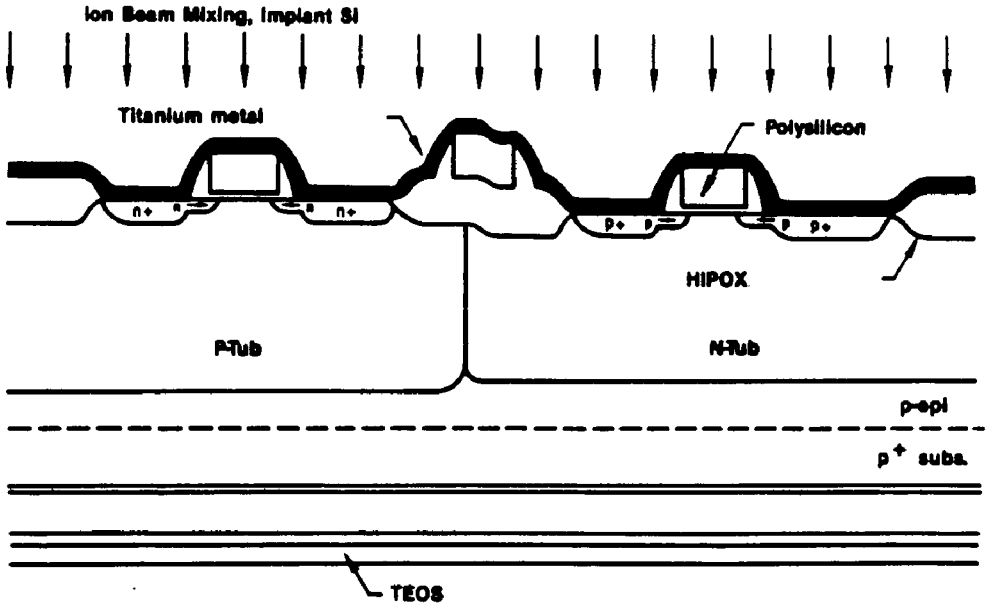


Figure 4-10. Ti Deposition and Ion Mixing Implant.

equipment contamination is the serious drawback for its use in practice. In this work, we will describe a salicide process using titanium to form $TiSi_2$. It is observed from Fig. 4-9 that the source/drain silicon and gate polysilicon surface are cleaned to remove any native oxide. A layer of Titanium is sputtered to the thickness of ~800 Å. Silicon atoms are implanted for ion mixing of Ti and Si. This step is to facilitate the silicidation process (Fig. 4-10).

A low temperature (600-700°C) Rapid Thermal Anneal (RTA) is done to react Ti and Si to form Ti-mono-silicide. This reaction converts Ti into TiSi only in areas where Ti is in contact with Si or polysilicon. Titanium on field oxide and the sidewall spacer slopes is not reacted. A wet chemical solution containing HF is used to selectively etch the unreacted Ti. Another RTA step is done at higher temperature (900°C) for a short time to convert TiSi (Ti mono-silicide) into $TiSi_2$ (Ti disilicide) which is a more stable compound (Fig. 4-10). At this point the basic transistor is formed.

4.3.7 First Level Dielectric and Metal

The first dielectric layer is to separate the active device terminals (drain gate source) to the first level of metal. A layer of 2000Å undoped TEOS is deposited, followed by 4000Å BPSG (Boron Phospho-Silicate Glass) or BPTEOS deposition (Fig. 4-11). The incorporation of Boron and phosphosilicate is to reduce the flow temperature to smooth the glass topography. This is crucial for a shallow junction, since the thermal cycles for shallow junction formation are strictly limited. With the proper glass concentration, the flow temperature can be reduced to 800°C.[90] A more advance approach is to planarize the intermetal dielectric by photoresist etch back or mechanical grinding.[91] The planarized dielectric surface would greatly improve the metal coverage over the underlying topography and avoid notching over dielectric steps therefore improve the electromigration. The difficulties associated with a planarized dielectric is the different depths of contact windows opened to gate and source/drain (~ 3500Å in height difference). This poses a challenge for window etching since there is no good etch stop on polycide gate. Moreover, the window need to be filled (plugged) with material such as tungsten or polysilicon, since step coverage of Al in high aspect ratio windows is very poor. For the purpose of obtaining active devices for study, we adopt a simple approach to make contact with the

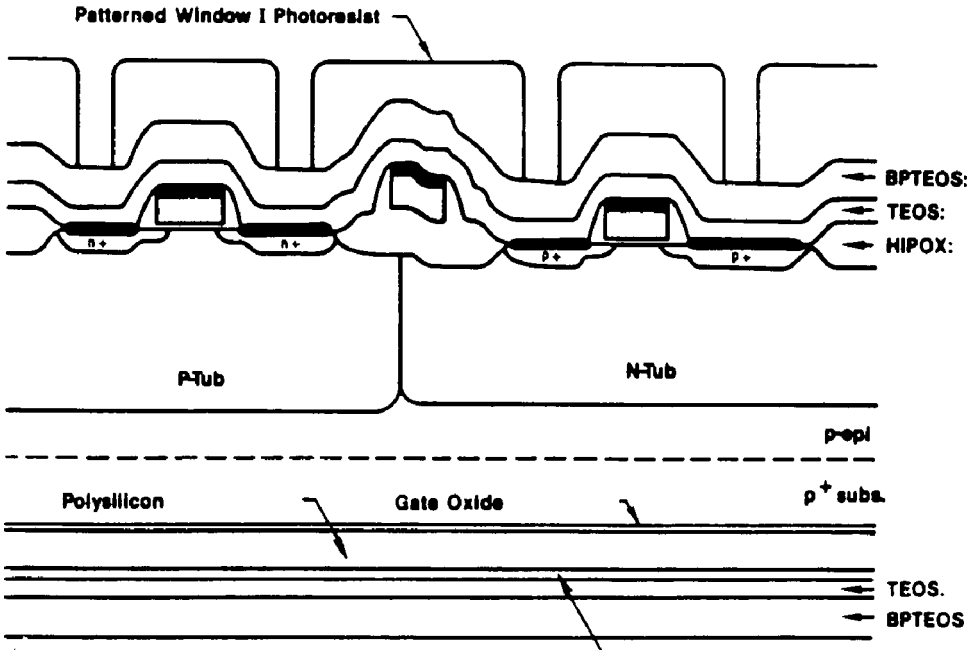


Figure 4-11. Selective Silicide Reaction, Dielectric 1 and Window 1 PR.

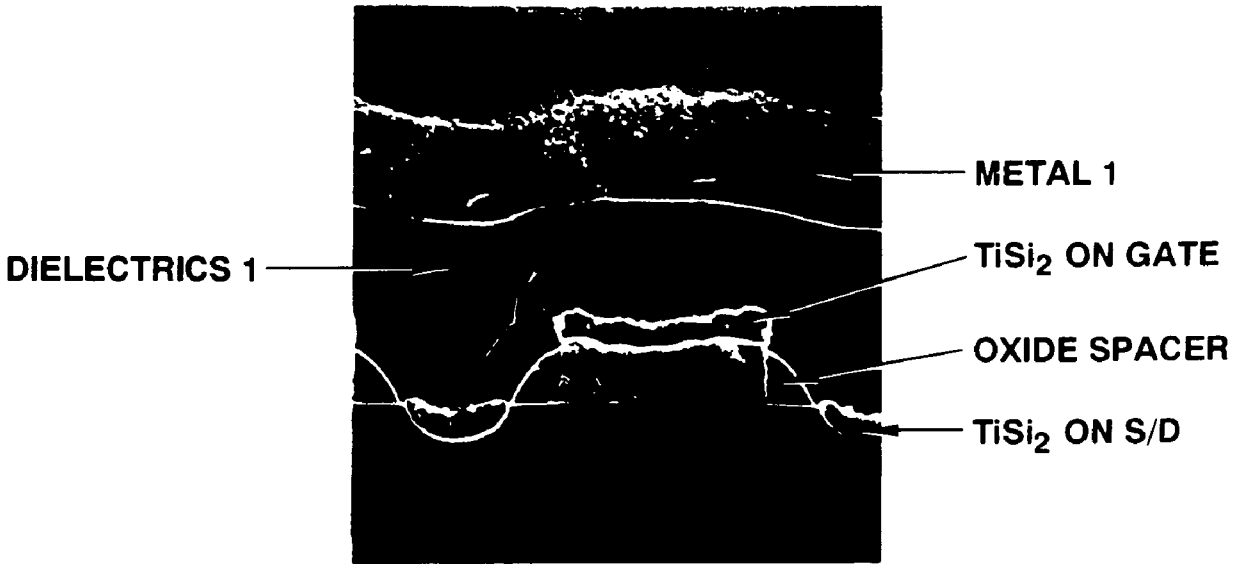


Figure 4-12. SEM micrograph of a finished MOS device with Ti silicide and first level metal dielectric.

device. Contact windows to active device are etched after a lithography step. The window size and its alignment to the gate and the source/drain periphery constitute a major penalty for an MOSFET size with respect to the gate size. Typically the extension of drain (or source) beyond the gate edge is more than 2 times larger than the minimum gate size. The large source and drain area not only increases the pitch and thus reducing the packing density in the layout of circuits but also increases the junction capacitance of the drain (source), degrading the circuit speed. More advanced structures employing elevated Source/Drain and shared contacts to reduce device size will be discussed in chapter 8.

Aluminum with a small percentage (0.5-1%) of Cu and (~ 1%) silicon is deposited and patterned to define interconnect wiring. The multi-level (up to 3) metal interconnects are the standard of modern CMOS integrated circuits technology for improve packing density and ease in circuit layout. The process technology to form multi-level metal systems include depositing a second layer of dielectric (D2), window 2 lithography and etch, window plug processes, metal 2 deposition, lithography and etch. The details of multi-level metal processes are beyond the scope of this dissertation.

The final capping material, such as plasma enhanced silicon nitride (SiN), or a sandwich cap (Oxide/Nitride), is deposited at low temperature and pressure. The pad openings are etched for wire-bonding. A low temperature (350-375°C) sinter step is done in H_2 ambient. As with most of the plasma enhanced CVD process, hydrogen is incorporated into the film and trapped there. The hydrogen contents will have a direct impact on the hot carrier aging in MOS devices as will be reported in chapter 7. The final device cross section obtained by SEM is shown in Fig. 4-12. The next chapter covers the analysis and characterization of the fabricated devices.

Chapter 5

TECHNOLOGY CHARACTERIZATION

5.1 INTRODUCTION

After the desired device structures and circuits have been fabricated by the integrated processes, the technology is characterized by different methods. The ultimate purpose of integrated circuit processing is to built functional circuits with high yield. For the device physicist, achieving the good and reliable device characteristics is the goal. The device and process monitor structures, for most practical purposes, will detect any potential problems which occurred during device fabrication, due to the processed material, and the device characteristics themselves. Following is a list of structures that are essential in process and device characterization:

1. Contact resistance: to gate, source/drain, intermetal contacts, local interconnect to active areas.
2. Sheet resistivity: for n-tub (or buried layer in the case of BiCMOS), polysilicon gate, active n+ and p+ source-drain diffusion areas, and the interconnect levels.
3. Capacitances: Junctions (n+/p-tub, p+/n-tub, n-tub to p-tub), MOS (gate, field oxide), dielectric layers and their composites.
4. Transistors: n and p-ch MOS devices, parasitic field devices, active and parasitic bipolars. The device layouts should include different lengths, widths and electrical design rule variations.
5. Isolation: active areas within the same (intra) tub and between (inter) tub.
6. Design Rule checking structures: variations in design size and placement to verify the misalignments and size variations duing photolithography steps.
7. Reliability Features: gate oxide, electromigration testers, latch-up, ESD (Electrostatic Discharge).

8. Physical Analysis: Special Structures for SEM and TEM cross-sections; combinations of dopings with areas sufficient for SIMS (Secondary Ion Mass Spectroscopy) analysis.

To cover all aspects of the above structures is beyond the scope of this dissertation. We will focus mainly on electrical characteristics of active devices in order to gain insight understanding the physics of short channel devices. The physical analyses of processing materials will be described where appropriate to explain the device behavior.

Although it is difficult to separate the process and device characterization, we can loosely consider process parameters are those related to physical quantities affected by the many processing conditions. The device parameters on the other hand can affect the circuit performance due to the active transistor behavior. Contact and sheet resistance, and capacitance components of various structures can be used as process monitor tools. Variations in parasitic resistance and capacitance can also affect device and circuit performance.

5.2 PROCESS CHARACTERIZATION

There are many techniques to characterize process related parameters such as doping profiles, oxide charges and quality, and resistance of contacts and various doped layers. These methods can use physical, chemical, mechanical, and electrical principles. We will discuss some practical techniques used in VLSI process analysis.

5.2.1 Doping Profiles

In CMOS device structures, the doping profiles of the tub (well) concentration, the channel and the source/drain doping profile are critical for active and parasitic device operations. From the MOS transistor theory developed in section 2.1, the channel doping profile affects the threshold voltage, the body effect, and the subthreshold conduction current. The source-drain junction profile, including depth, gradient and peak concentration, is the dominant factor in the short channel device behavior and hot carrier generation. The source/drain junction doping profiles determine the junction depth and the sheet resistivity of the diffusion areas. The drain-to-tub junction concentrations will affect the parasitic capacitance and therefore speed of the circuits.

For a CMOS technology built on a p-type substrate, the n-tub profiles in the vertical direction not only affect the active devices, but also the vertical punch through of a parasitic pnp device (formed by p+ s/d, n-tub and p-substrate). If the n-tub depth is too shallow, the vertical pnp bipolar gain is high, and collector-emitter punch-through may occur. Under the field oxide, the tub-concentrations and the field oxide thickness determine the field threshold voltage and the drain induced barrier lowering between 2 source-drain regions on the two sides of the field. The interaction of the n- and p-tub at the tub boundary is critical for inter-tub isolation. The inter-diffusion of boron and phosphorus from n&p-tub implants and channel-stop implants during subsequent heat treatments is a severe case for n+ to p+ isolation in a twin-tub process using atmospheric wet field oxidation (finger effect).[92] In this case, a 2-dimensional process simulator can be used to analyze the 2 dimensional characteristics of the lateral doping profile. For the above cited reasons, the knowledge of doping profiles during or after processing is critical in characterizing the processes and devices.

SIMS (Secondary-Ion Mass Spectrometry) is by far the most popular technique in VLSI material analysis. Its particular usefulness is in determining of dopant distribution in silicon of a variety of impurities. Boron, phosphorus, arsenic, antimony are among the species that SIMS can detect. The SIMS technique employs energetic ion beams to bombard the material. As the primary ion beam energy is increased, the ion penetration depth increases till the onset of sputtering. In this process, the secondary ions are emitted. The detection of these emitted ions is used to analyze the material distribution vs depth. the channel doping profile. Both C-V and spreading resistance methods profile the electrically active species, while the SIMS technique analyzes all the chemical species in the sample.

Spreading resistance measurement is a mechanical and electrical technique to extract the profile of electrically active dopants. This method requires an angled lapping of the sample. Two point probes are used to measure sheet resistivity of a particular position on the angled surface. The accuracy of this technique is limited by the spatial resolution of the mechanical stepping distance of the probes and the angle of the lapping, and therefore usually is useful for deeper junctions.

Figures 5-1a shows a shallow n+ source/drain junction profile formed by an arsenic implant with a dose of $3e15cm^{-2}$, and an energy of 70KeV. Figure 5-1b shows profiles of a deep (3μm) antimony buried layer and an n+ arsenic doped emitter of an npn bipolar transistor. SIMS is a very useful technique for dopant profile in VLSI processing, in particular for concentration greater than $5\times10^{15}cm^{-3}$.

5.2.2 Analytical Tools for Material and Structural Analysis

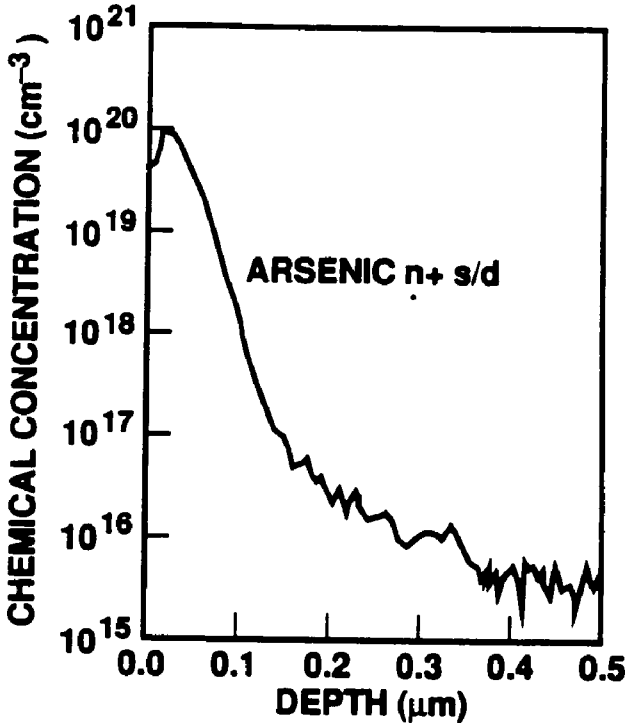
SEM and TEM (Scanning and Transmission Electron Microscopy) are used extensively to analysis the microscopic structures of the devices, and to detect the defects introduced by processes. For the case of submicron device geometry, SEM and TEM have been used as a metrology for line width control of patterned photoresist lines and etched features of submicron structures. TEM and SEM results are shown along with the discussions of device structures in this dissertation.

Other physical and chemical methods useful for analyzing contamination impurities such as carbon, iron and other metallic contaminants, include Auger spectroscopy, X-ray, Rutherford Back-scattering (RBS), Infared Spectroscopy etc..

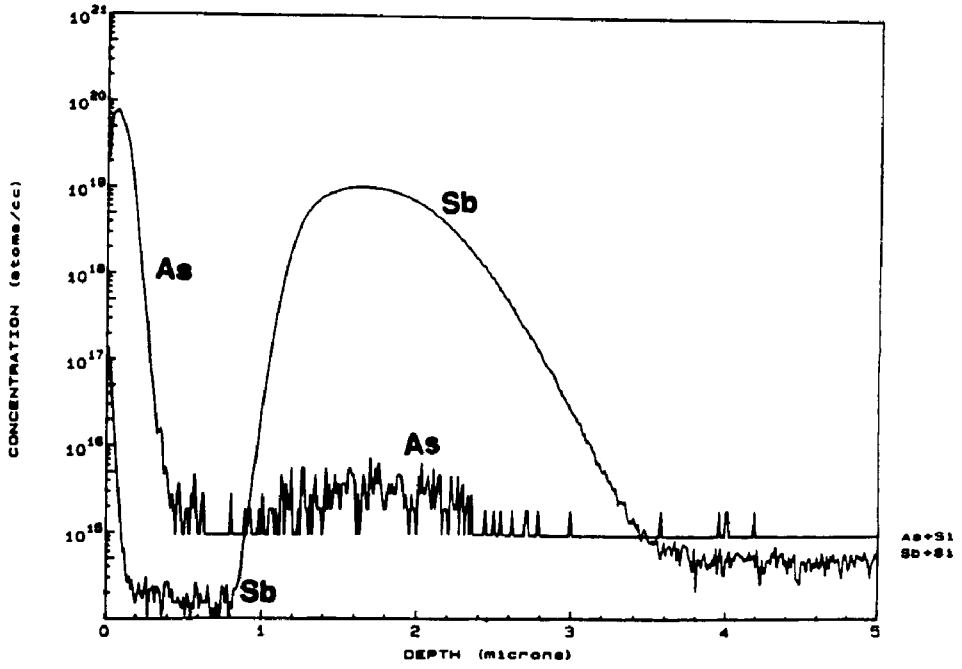
We will concentrate on electrical measurement techniques to analyze devices and to detect any processing problems. The following parameters, although affected by processing, have profound effects on device performance as discussed in chapter 2 and in the remaining pf this chapter.

5.2.3 C-V Techniques for MOS Structure Characterization

The gate oxide quality is the most fundamental requirement for the reliable operation of MOSFETs. The gate oxide quality attributes include the control of charges in the oxide and interfaces, the integrity of the oxide with respect to breakdowns and defects, and the control of thickness uniformity. The control of charges at the Si-SiO_2 interface has been a subject of extensive research since the conception of MOS transistor until the present time. The three charge components in the SiO_2 are the fixed charge, Q_f, the interface trap charge, Q_{it}, and the bulk oxide charge, Q_{ox}. Capacitance-Voltage (C-V) measurements on an MOS capacitors,



(a)



(b)

Figure 5-1. SIMS profiles of an n+ source/drain junction (a), and an npn buried layer and emitter (b)(SIMS work done by F. Stevie at AT&T Bell Labs).

including high-low frequency and quasi-static methods, have been used extensively, not only to study charges at the Si-SiO_2 interface, but also to investigate the semiconductor bulk properties such as impurity doping profiles, minority carrier recombination lifetimes, and electrically active impurity trap densities and energy levels within the silicon forbidden bandgap.[93] The quasi-static method measures the displacement current in an MOS capacitor when a linear voltage ramp is applied. The displacement current is proportional to the low frequency capacitance of the device. The key element of either C-V technique is that the ramp rate has to be sufficiently slow so that the device is maintained in thermal equilibrium. The slow ramp rate also allows the interface state, also referred to as slow states, to respond to the changes in surface potential changes. For the modern silicon material quality needed for advanced CMOS devices, the minority carrier lifetime is excessively long on the orders of minutes. Therefore the voltage ramp rate is kept slow to maintain the device in equilibrium when an MOS capacitor is biased in depletion and inversion regimes. However, if the linear voltage is applied at a higher rate, deviations from thermal equilibrium can allow bulk information to be extracted.

From the CV curve of an MOS capacitor on p-substrate shown in Fig. 5-2, the oxide thickness is calculated from the accumulation capacitance, $C_{acc} = C_{ox}$, by:

$$t_{ox} = \frac{\varepsilon_{ox}}{C_{ox}} \tag{5.1}$$

The fixed charge, Q_f, can be calculated from the flat-band voltage by the relation:

$$Q_f = C_{ox}(V_{FB} - \Phi_{ms}) \tag{5.2}$$

From the minimum high frequency capacitance, $C_{\min}$, measured in inversion as shown in Fig. 5-2, an effective doping concentration, N_A, in the silicon bulk can be obtained by equation:

$$\frac{C_{\min}}{C_{ox}} = \frac{1}{1 + \frac{2C_{ox}}{q\varepsilon_s}\sqrt{\frac{\varepsilon_s kT \ln(N_A/n_i)}{N_A}}} \tag{5.3}$$

The interface state density, D_{it}, can be determined from the difference in the high frequency and quasi-static CV curves as the depletion region is formed. The quasi-static method was developed by Castagne and Vapaille,[94] and has been modified and improved by several

researchers. A good treatment on this subject can be found in the book, MOS Physics and Technology, by Nicollian and Brews.[95] The interface state density as a function of energy extracted from the MOS capacitor is shown in Fig. 5-3, the midgap density is about 2E10 $cm^{-2}eV^{-1}$.

In addition to the oxide parameters extraction described above, the quasi-static C-V technique is also very useful in detecting the activation of the implanted dopants in a polysilicon gate. This method is particularly important in sub-half micron CMOS devices where polysilicon gates are implanted with arsenic/phosphorus or boron impurities separately. The C-V curves of a PBr_3 doped gate, which is fully activated, MOS capacitor is shown in Fig. 5-2. Figure 5-4 shows similar C-V curves for an arsenic implanted gate . In the strong inversion region, which corresponds to the positive gate voltage, the inversion capacitance of the fullly activated gate device shown in Fig. 5-2 is approaching $C_{max} = C_{ox}$. On the other hand, the inversion capacitance of the partially activated arsenic gate device shown in Fig. 5-4 is about 85% of the maximum capacitance. The equivalent gate capacitance is now comprised of a depleted poly gate capacitance in series with the gate oxide capacitance. Therefore the total capacitance is less than the oxide capacitance.

In half-micron and below CMOS structures, the CMOS gates are implanted with arsenic and boron for n+ and p+ poly respectively. Low temperature heat treatments are used in the later stages of processing to preserve the shallow junction depth. RTA (Rapid Thermal Annealing) is often used to activate implanted dopants. Therefore, the quasi-static measurement, in addition to the conventional interface state density extraction capacity, is a very powerful method to ensure the dopants are fully activated. If the dopants at the polysilicon- oxide interface are not fully activated or lightly doped, the threshold voltage of the device is shifted due to the change in fermi level in the poly. This results in a lower transconductance, i.e. proportional to the linear gain, because the gate oxide capacitance is lower (refer to Eq. (2.86)).

Other applications of the CV techniques include the bias temperature stress to monitor the sodium drift in oxide. In an extension of this method, called TVS (Triangle Voltage Sweep),[96] a large magnitude peak-to-peak ramp is applied to an MOM (Metal-Oxide-Metal) structure heated

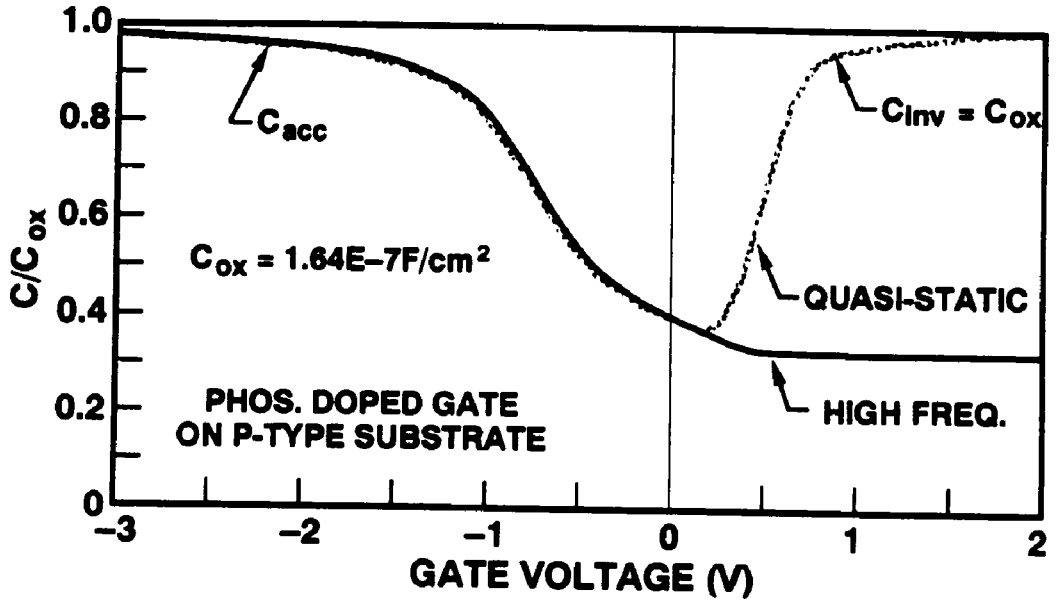


Figure 5-2. High frequency and quasi-static C-V curves of an MOS capacitors with a phosphorus diffused gate, with an area of 0.56E-3cm^2, t_{ox} 208Å.

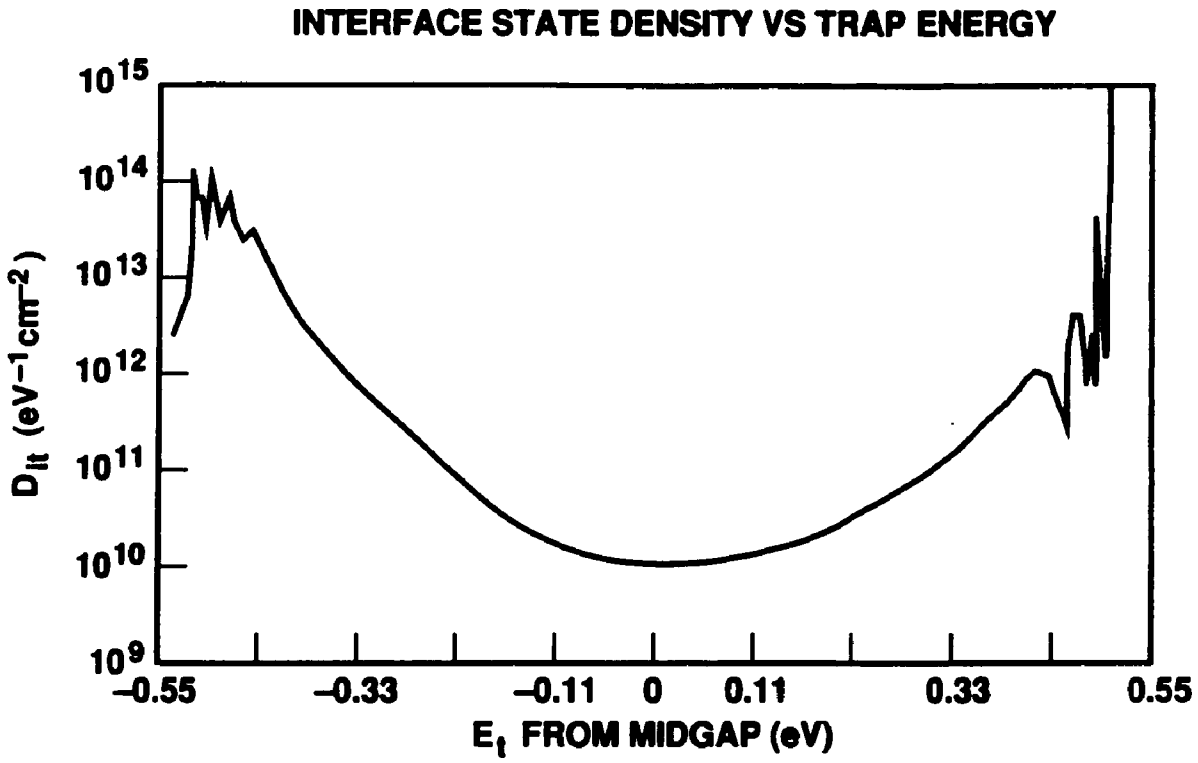


Figure 5-3. Extracted interface density, D_{it}, vs energy.

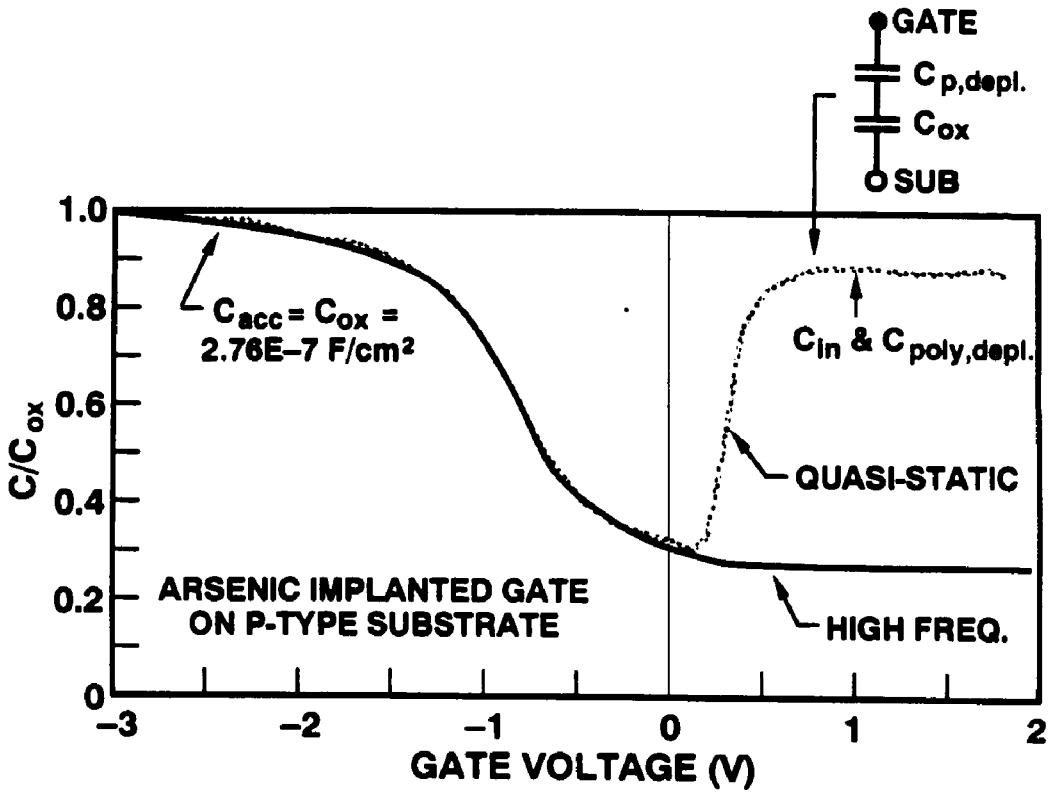


Figure 5-4. Quasi-Static C-V curves an arsenic implanted, not fully activated poly gate capacitor.

to temperature of about 300°C. Any sodium presence will result in a peak of displacement current of the I-V curves. This is a very simple and powerful method for detecting sodium incorporated into insulator, e.g. photoresist contamination, during process development.

5.2.4 Resistivity

Sheet and Contact resistance are important for submicron device performance. For an MOS transistor the series resistances at source and drain caused by the contact to the source/drain and the diffusion sheet resistance degrade device performance, both in linear and saturation modes. Therefore efforts have been made to reduce the parasitic resistances. The silicide or salicide processes have been used to alleviate this problem. However, with the silicide layer on top of the source-drain-gate areas, silicide-to-silicon interfacial resistance can be inadvertently introduced, if the silicidation process is not done properly. MOS transistor structures have been proposed to characterize this interface by Lynch and Ng.[97]

For metal-to-silicon or gate contacts, the Kelvin contact structure has been used extensively to monitor the contact resistance. The sheet resistivity is normally measured from a Van der Pauw structure. The following table lists typical values of contact and sheet resistances of various structures.

TABLE 5-1. Typical resistance values for different types of material.

Structures	No Silicide	With Silicide
n+ s/d sheet (Ω/□)	25-30	1-3
p+ s/d sheet (Ω/□)	100-200	1-3
n+ poly gate (Ω/□)	40-80	2-5
metal to n+s/d (Ω)	5-10	< 1
metal to p+s/d (Ω)	5-100	< 1
metal to poly (Ω)	1-10	< 1

It is clear from the Table 5-1 that with the addition of silicide layer cladding over silicon, the sheet and contact resistance reduce significantly. However, the added source-drain series resistance due to silicide-silicon interface, and the Lightly doped drain structure would degrade the device performance as will be addressed in section 5.3.

5.3 DEVICE INTEGRITY IN SUBTHRESHOLD

Transistor characterization deals with device parameters which ultimately effect device and circuit performance. These parameters are directly related to device structure, processing conditions and process variations. Assuming all the process related structures have been characterized, i.e. good oxide properties, correct gate oxide thickness, good contact resistance, low junction leakage, good silicide formation etc., the next step is to analyze the device characteristics.

5.3.1 Subthreshold Conduction: The *g* Curves

A set of drain current versus gate voltage (I_D vs V_G), referred to as *g* curves with 2 different drain voltages, V_D = 0.1 and $V_{D,max}$ = 3.6 V for a 3.3 V technology, are measured using a low

leakage, low noise set-up. Shown in Figures 5-5 and 5-6 are the g curves for N and PMOS transistors with effective channel lengths of 0.45μm and 0.43μm, respectively.

Device integrity and many important parameters can be obtained from this simple measurement:

- The subthreshold swing, S_n, of the device, is the inverse slope of $\log I_D$ vs V_G, from which the effective interface trap density can be derived. From Eq. (2.96), S_n is defined as:

$$S_n = 2.3n\frac{kT}{q} \tag{5.4}$$

  where a calculation of interface density can be carried out by measuring n and using Eqs. (2.98) and (2.99). (A more detail extraction of interface states is treated in chapter 7 for device aging experiments). Typical subthreshold swing is in the range of 80-90 mV/dec. An unusually high subthreshold swing indicates a high density of states, or other processing related problems.

- The I_{off} current measured at V_D = 3.6V and V_G= 0V can also be readily obtained, yielding an I_{off} of −24 pA at room temperature for both $n-$ and $p-$ ch devices. The punch-through is not detected in these devices at V_D = 3.6 V. However, in a 0.35μm channel length device fabricated by the same technology, the subthreshold conduction is excessive as can be seen from Fig. 5-7. In these curves, the subthreshold swing gradually increases with increasing V_{DS}, showing the device structure is not designed properly for this geometry.

- Junction Integrity: with a low noise, low leakage measurement set-up, the drain junction leakage can be detected by measuring the drain current at high drain voltage ($\geq V_{D,max}$). A significant higher drain current, which is gate voltage independent, compared with the current measured at the source indicates a leaky drain-to-substrate (or tub) junction. (Refer to Fig. 7-9 for a leaky drain junction after hot carrier stressing).

- Drain Induced Barrier Lowering: A shift in 2 g curves at $V_D = 0.1V$ and $V_{D,max}$ at a constant drain current is a measure of drain-induced barrier lowering, ΔV_G. ΔV_G is very small for long channel devices, and becomes larger at shorter channel devices. At the current level of 10 nA, ΔV_G is measured to be ~ 70mV for both devices (refer to Figs. 5-5

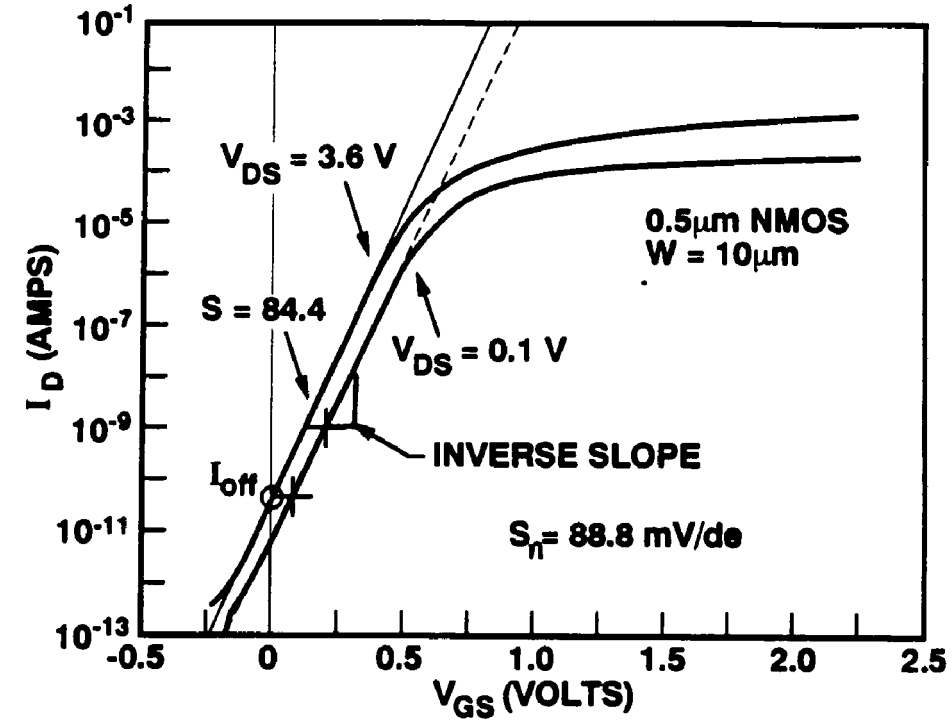


Figure 5-5. Log(I_D) vs V_G curves of an NMOSFET with L_{eff}=0.45μm and at 2 different drain bias.

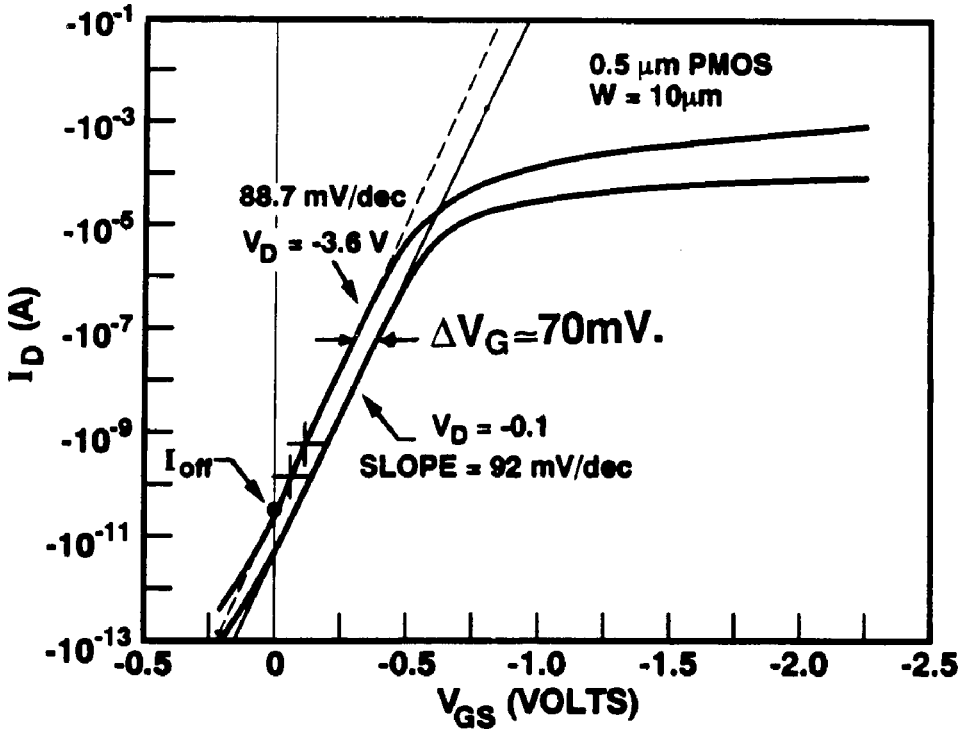


Figure 5-6. Log(I_D) vs V_G curves of a PMOSFET with L_{eff}=0.43μm and at 2 different drain bias.

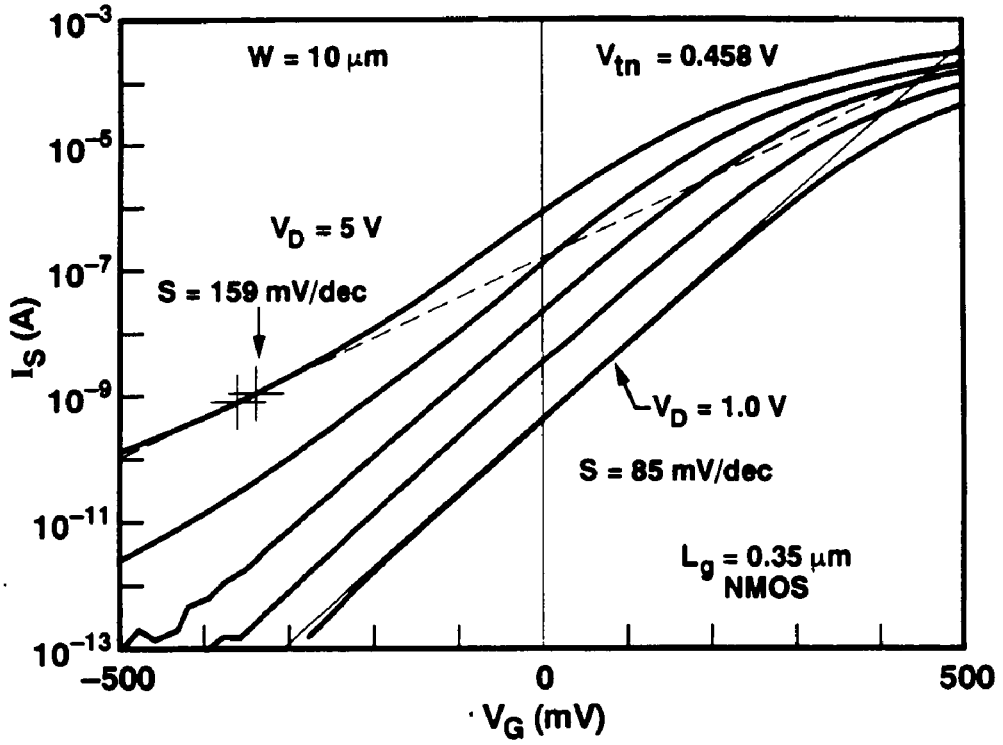


Figure 5-7. Subthreshold g-curves of a 0.3μm channel length NMOSFET operates near punch-through.

and 5-6), which is quite acceptable.

- Surface potential band bending, ψ_s, can be extracted as a function of gate voltage, from the point of weak inversion to strong inversion. Therefore, the charges associated with an MOS system can be studied.

Using the g-curves approach, we have observed that the subthreshold swing in a well designed short channel device is indeed decreased with increasing drain voltage within a reasonable range, as will be analyzed in the next section.

5.3.2 Reverse Subthreshold Swing

As we have observed from Figs. 5-5 and 5-6, the subthreshold swings of the short channel CMOS devices at $V_D = 3.6V$ is smaller than that of the same device measured at $V_D = 0.1V$, (84.4 vs 88.8 and 88.7 vs 92 mV/dec for N and PMOS, respectively). This phenomena is referred to as reverse subthreshold swing, and is first reported here in this work. We have measured subV$_t$

swing S of a series of short channel devices as a function of drain bias. The results plotted in Fig. 5-8, show a rather interesting behavior of well designed short channel devices. The drain voltage is increased over almost 3 decades, from 10mV to 7V. S initially decreases with increasing V_D until the device enters into the punch-through regime, where S is rapidly increases. The punch-through effect is most apparent for the 0.3μm device in which S increases abruptly at V_D= 3V. An analytical model was derived to explain the phenomena in Chapter 2 (Eq. 2.98). A similar device behavior has been reported in our previous work,[50] where S decreases with shortening channel length for surface NMOS devices before punch-through, whereas S monotonically increases as the channel length of buried-channel PMOS devices decreases (Fig. 5-9).

5.3.3 I_{off} Optimization

The object of most MOS technologies is to obtain the highest performance devices without suffering excessive short channel length effects. The fall-off of either linear or saturation threshold voltage as the effective channel decreasing has often been used to characterize short channel effects (see Fig. 5-16 in the next section). However, this does not describe the increase in leakage at short channel lengths. we have dealt with the short channel effects based on the off current (I_{off}) as the criteria, normally specified at 10^{-10} A/μm width at 90°C, or 1nA/μm width at 125°C, as described in ref[50].

Circuit considerations define a maximum I_{off} per unit μm of channel width, where I_{off} is defined as I_D at $V_D = V_{D,max}$, and $V_{GS} = 0V$. The device can be designed such that for the shortest allowable channel length and the lowest possible threshold voltage, the off current still conforms to the leakage specifications. Shown in Fig. 5-10 is a plot of I_{off} vs L_{eff} curves, measured at 125°C for the worst case drain voltage of 6V. The linear threshold voltages, V_{tn} and V_{tp} are measured at room temperature from the nominal channel length devices. As can be seen in Fig. 5-10, the subthreshold behavior is well preserved, with I_{off} << 1nA/μm at 125°C, for channel length of 0.6μm for this 1.0μm CMOS technology.[50] Figure 5-11 depicts the experimental procedure to obtain a minimum L_{eff} for a range of $V_{tn,p}$'s to meet an $I_{off} \leq 0.1$ nA/μm width at different temperatures. It is noted that the NMOS devices are more sensitive to temperature variation as compared with PMOS devices. From this plot, a pair of V_t and L_{eff} can be obtained

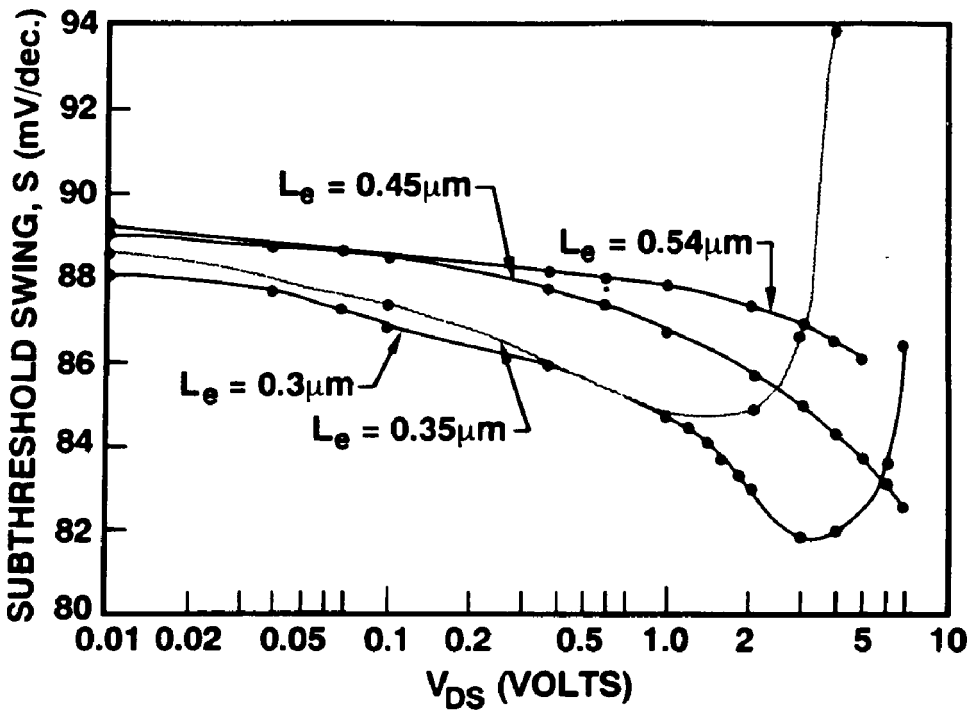


Figure 5-8. Reverse SubV_t Swing vs Drain Bias for sub-half micron NMOS devices.

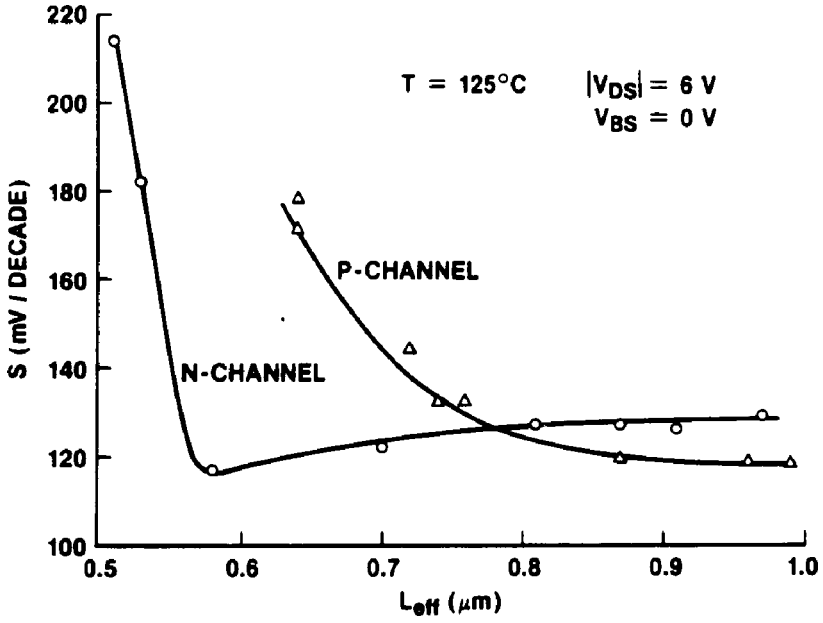


Figure 5-9. Reverse SubV_t Swing on NMOS with respect to channel length, while PMOS buried channel devices shows monotonically increasing S.[50]

for an I_{off} specification.

5.4 ACTIVE DEVICE PARAMETERS

5.4.1 Threshold Voltage

The g-curves of Figs. 5-5 and 5-6 are re-plotted on a linear scale with their derivatives (which is the transconductance $g_m = \partial I_D / \partial V_G$) for the drain bias of 0.1V, in Figs. 5-12 and 5-13. In linear mode, the channel is strongly inverted when V_G is higher than threshold voltage. The linear threshold voltage can be derived from this plot by extrapolating the slope of the linear I_D–V_{GS} curve at peak transconductance, $g_{m,max}$, and extrapolate to the V_G axis. The linear threshold voltage, is calculated by subtracting $V_D/2$ from the intercept point. This method is reasonably accurate for most devices. However, with a device damaged by hot carriers or radiation, the band bending is a complex function of gate voltage. The point of inflection method is prone to errors. Wong et al.[95] have developed a method, in which the peak of the derivative of the transconductance is used to determine the $2\phi_F$ point. This extraction will be described in chapter 7.

The saturation threshold voltage is a measure of short channel effects at maximum operating voltage on the drain. Shown in Figs. 5-14 and 5-15 are $\sqrt{I_D}$ vs V_G plots, where the intersect of the maximum slope of $\sqrt{I_D}$ versus V_G onto the V_G axis is the saturation threshold voltage. Several devices with different channel lengths are measured. The conventional threshold voltages versus the effective channel can be plotted as shown in Figs. 5-16[50] for both types of devices. The usual V_t falloff at short L_{eff} can be observed. As we indicated earlier, the I_{off} approach to characterizing short channel effects is more robust for verifying the device integrity, and optimizing the active mode performance.

5.4.2 Reverse Short Channel Effects in Threshold Voltage

Figure 5-17 shows the threshold voltage of an NMOS device as the function of effective channel lengths. The channel doping profile of the 0.5μm device is shown in Fig. 3-4, where a boron concentration at the surface is lower than in the bulk. This depletion of dopants at the surface is due to the boron segregation into the silicon dioxide during the sacrificial and threshold

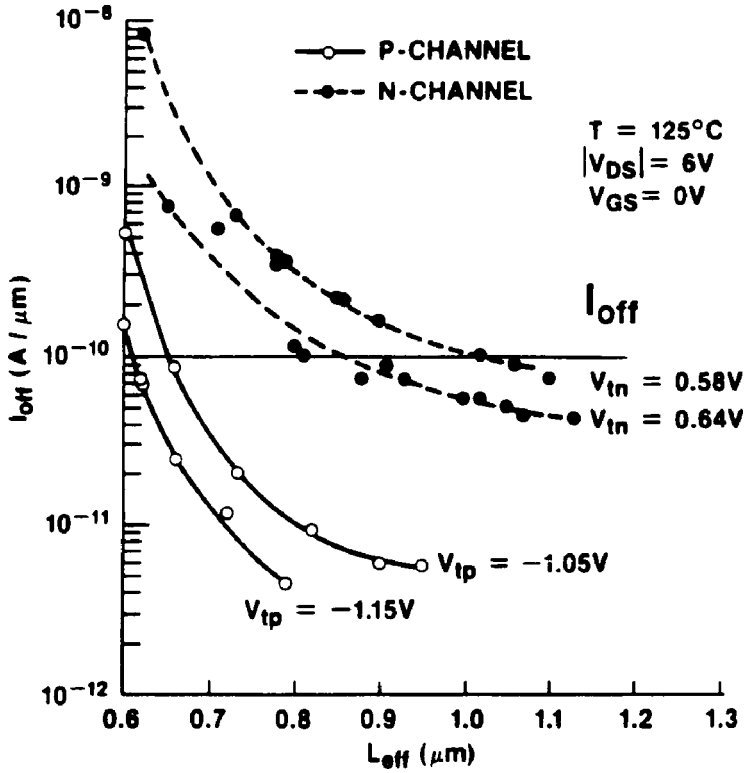


Figure 5-10. I_{off} vs L_{eff} for N and PMOS devices at 125°C and at different V_{tn}'s and V_{tp}'s.

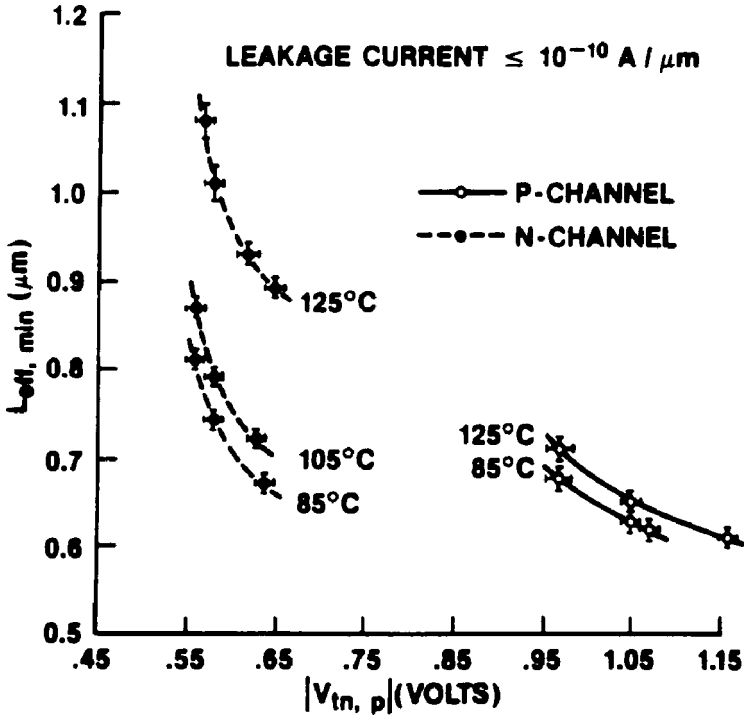


Figure 5-11. Optimization of threshold voltage with respect to minimum channel length and I_{off} requirements at different temperatures.

adjust oxide growths. The lack of a threshold adjust implant in a symmetrical CMOS technology is to avoid over-compensating the n-tub, on which the surface PMOS device is desired. From Fig. 5-17, it is noted that the threshold voltage increases ~ 68mV higher than the longer channel device. This phenomena is called the reverse short channel effects. The reverse short channel effect was attributed by several factors. The enhanced lateral diffusion in a short channel from an OED (oxidation enhanced Diffusion) during the source-drain reoxidation, or an LDD push oxidation, resulting in dopants (boron in the NMOS case) piling up in the middle of the channel, increasing threshold voltage.[96] $TiSi_2$ encroachment has also been attributed to the reverse short channel effects.[97] We have tried a mild threshold implant to fill up the surface depletion, and hence reducing the lateral gradient for diffusion. The effect is reduced, although the difference in threshold voltage is still ~ 40mV. The reverse short channel effect may not be a severe problem, however controlling the threshold difference is an added burden for V_t control and circuit design modeling in submicron CMOS.

5.4.3 Transconductance and Linear Gain

Transconductance is an important parameter to determine the performance of an MOS device. It is defined as the rate of change in channel current with respect to the change in the gate voltage. the transconductance is derived by taking the derivative of Eq. (2.113) to obtain:

$$g_m(V_{GS}) = \frac{\partial I_D}{\partial V_{GS}} = \frac{\beta_0 V_{DS}[1 + 2\lambda\theta_{sr}\sqrt{V_{SB} + 2\phi_F}]}{[1 + \theta_{sr}(V_{GS} - V_{tn} + 2\lambda\sqrt{V_{SB} + 2\phi_F})]^2} \tag{5.5}$$

where θ_{sr} includes the series resistance and mobility reduction due to interface states:

$$\theta_{sr} = \theta_s(0) + 2R_s\beta_0 \tag{5.6}$$

and

$$\beta_0 = \frac{\mu_{I,L}(\frac{W}{L_e})C_{ox}}{(1 + \alpha_{it}\overline{D_{it}})} \tag{5.7}$$

The maximum transconductance, $g_{m,max}$ is determined by taking the second derivative of Eq. (5.5), and find a V_{GS} at which the second derivative is equal to 0. We obtain:

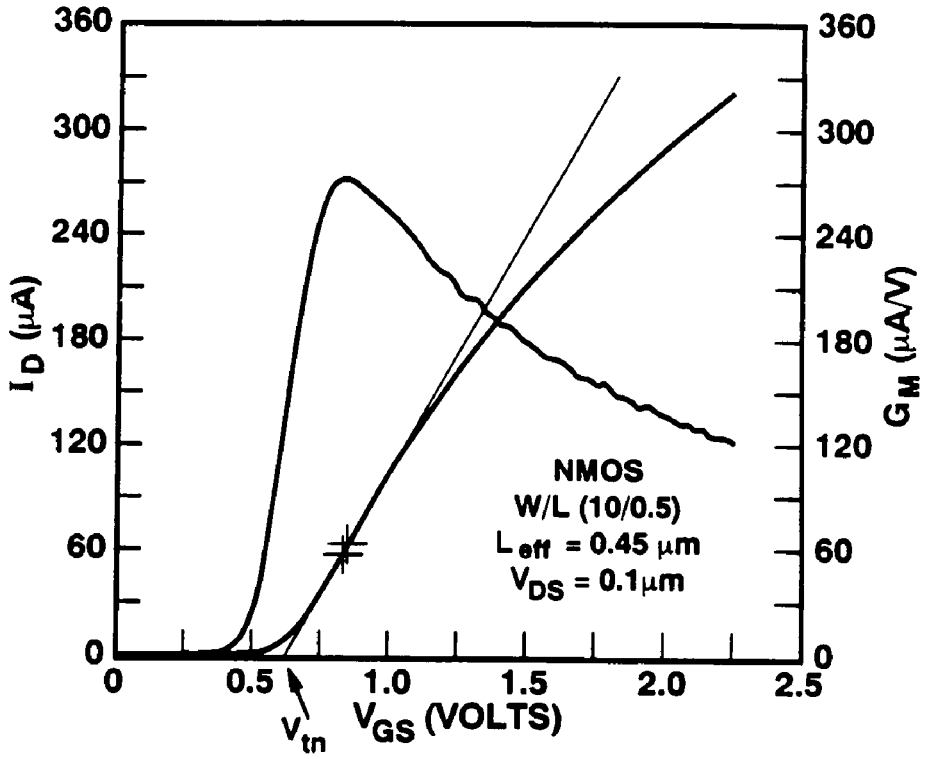


Figure 5-12. Linear drain current and its transconductance as a function of gate voltage for a 0.45μm NMOS device.

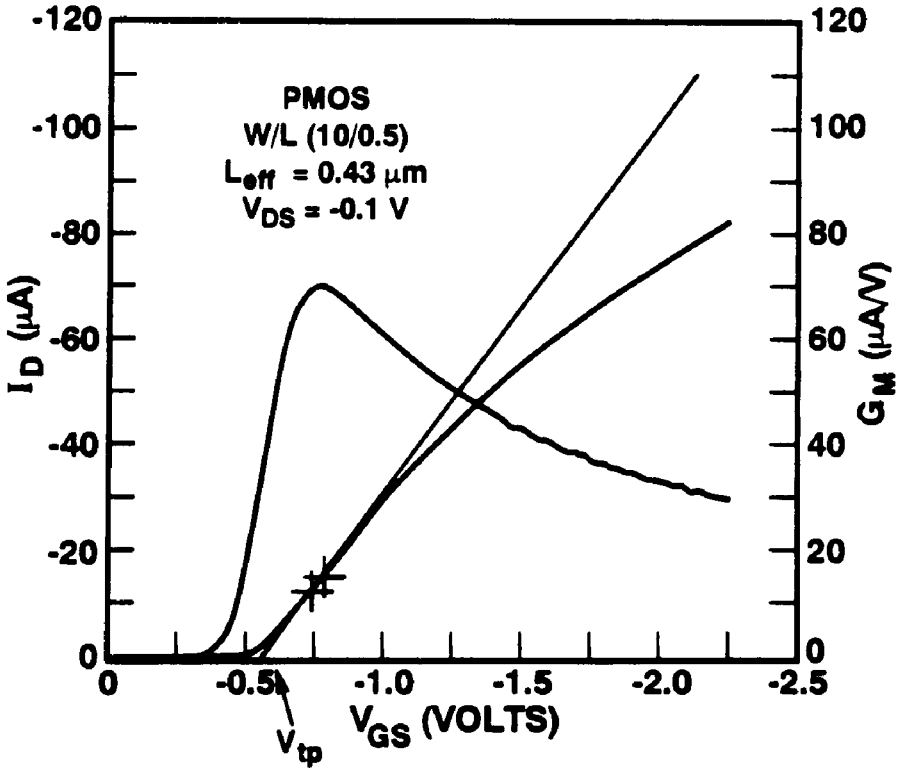


Figure 5-13. Linear drain current and its transconductance as a function of gate voltage for a 0.43μm PMOS device.

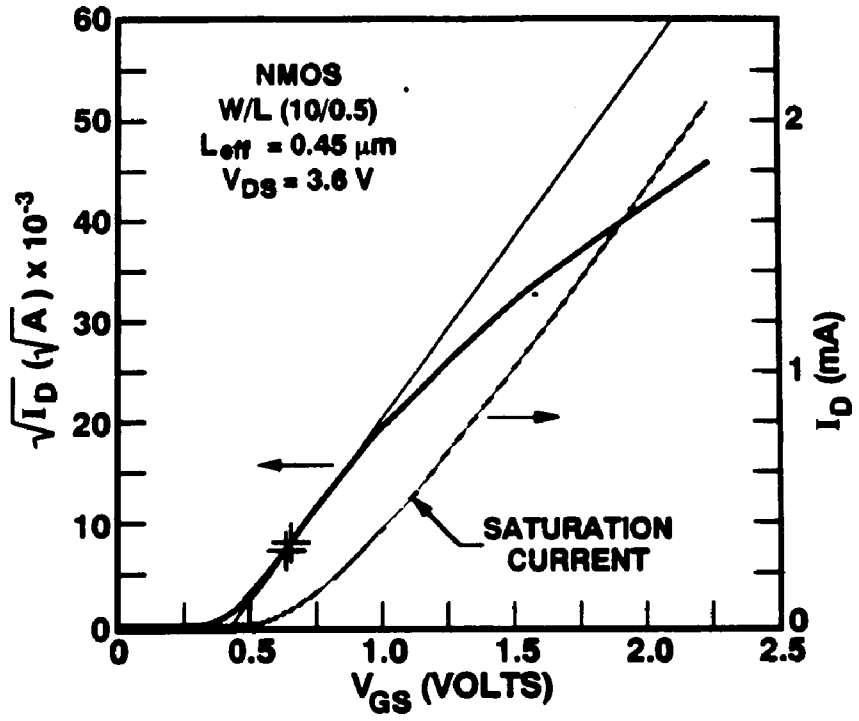


Figure 5-14. Plot of I_D and $\sqrt{I_D}$ for an NMOS device, with the drain biased at V_D =3.6V.

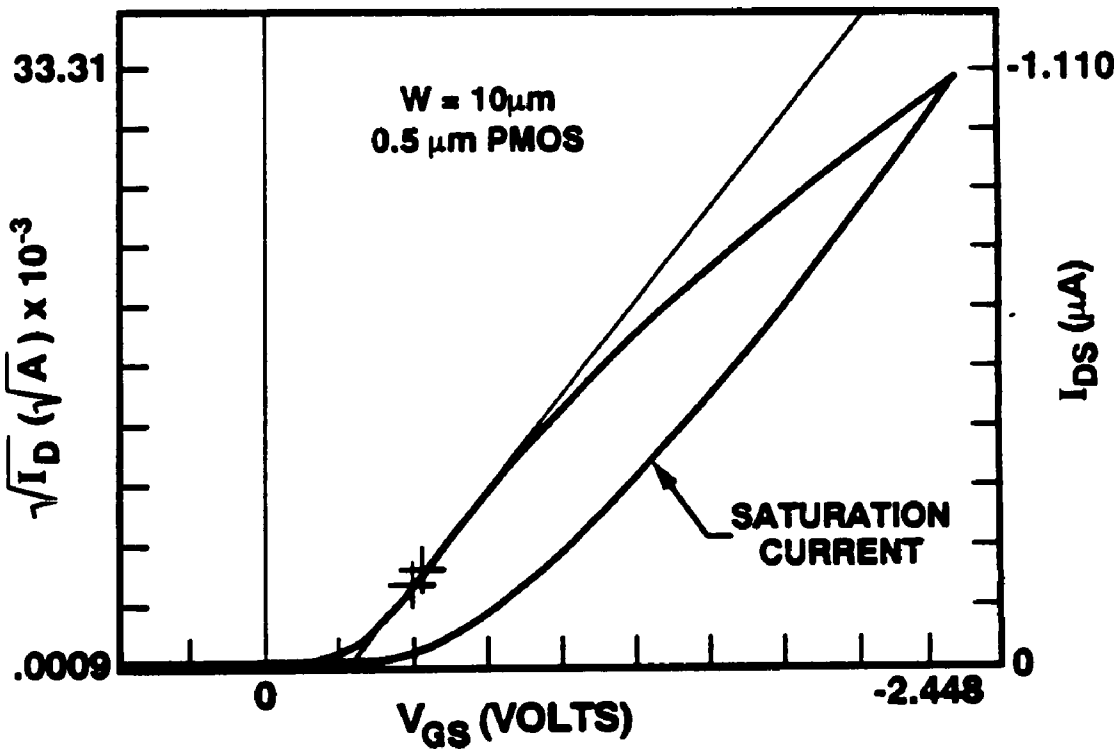


Figure 5-15. Plot of I_D and $\sqrt{I_D}$ for an PMOS device, with the drain biased at V_D =-3.6V.

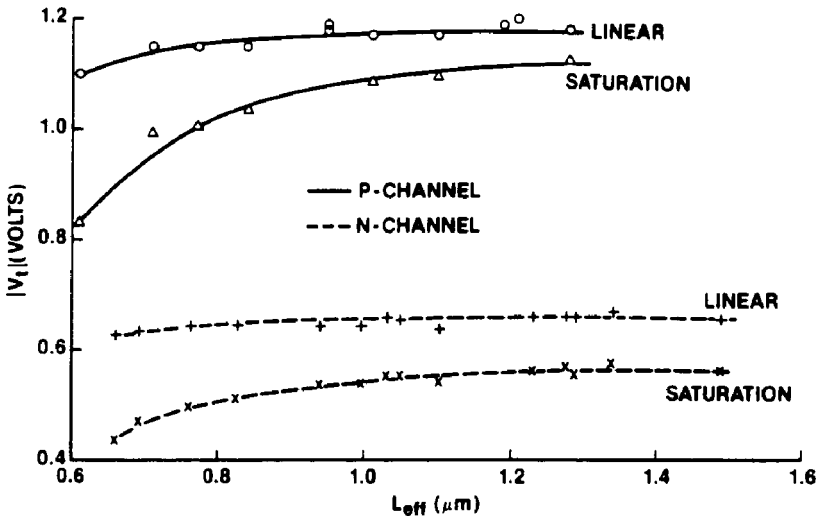


Figure 5-16. Linear and saturation threshold voltage vs. L_{eff} for n and p-channel transistors.[50]

$$g_{m,max} = \frac{\beta_0 V_{DS}}{1 + 2\lambda\theta_{sr}\sqrt{V_{SB} + 2\phi_F}} \tag{5.8}$$

$g_{m,max}/V_{DS}$ is referred to as linear gain and is an important parameter for the device operates in linear region. From Figs. 5-10 and 5-11, the transconductance as a function of gate voltage is calculated from the measure g-curves of n- and p-ch devices. In short channel length devices, a transconductance overshoot with back gate bias is observed. This is attributed to the charge sharing in short channel devices which causes the effective channel length to be shorter, increasing β_0 as governed by Eq. (5.7). This effect is more dominant than the mobility reduction due to high vertical field (V_{GB} is higher for negative back gate bias V_{BS}). The transconductance overshoot observed in this work should be modeled correctly for circuit analysis.

5.4.4 Effective Channel Length And Source-Drain Series Resistance

The difference between the coded channel and the final electrical channel length measured from the fabricated devices can be extracted using the transconductance method described below. The difference between the coded and effective lengths is a combination of the lateral diffusion of

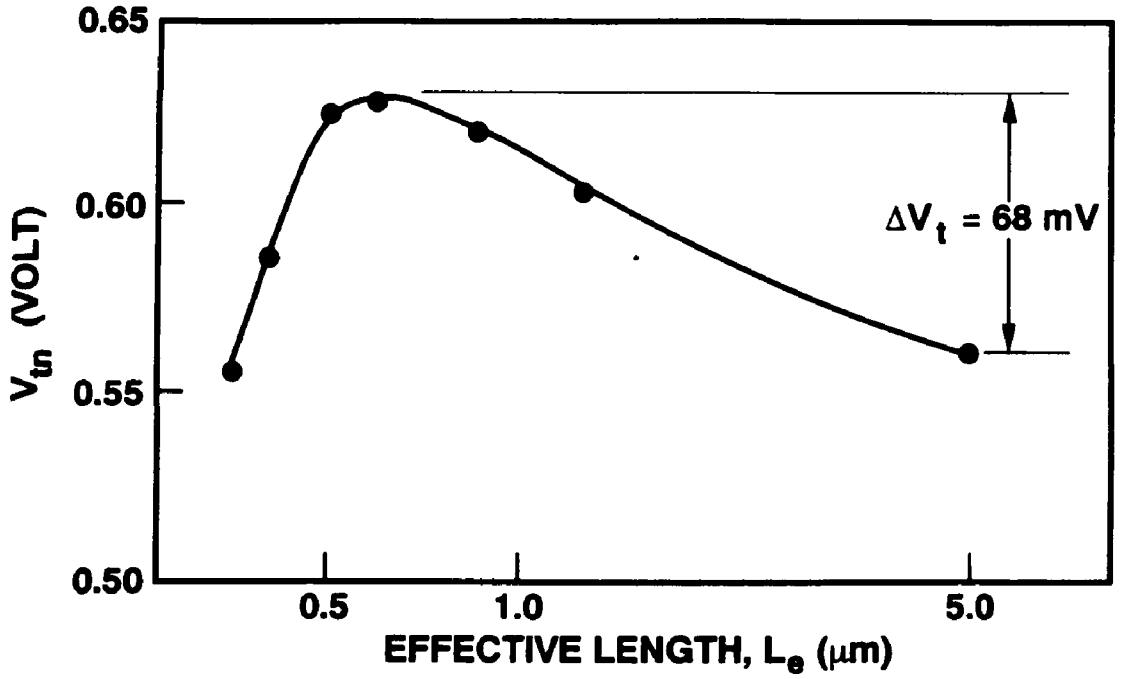


Figure 5-17. Reverse short channel effect in NMOS Devices with lower surface concentration.

the source-drain junction, and the line width variation during the gate lithography and etching,

The maximum transconductance is directly related to the linear gain of the device by Eq. (5.8), and can be used to calculate the effective channel length of a device if a set of different coded lengths devices are available. In strong inversion, the maximum transconductance is inversely proportional to channel resistance.

In short channel device, the channel resistance is becoming smaller, and therefore the series source and drain resistance can not be ignored. A series resistance model can be developed using the linear current equation with θ_s and series resistance. We rewrite Eq. (2.113) with the δL term accounting for the change in coded mask length L_m due to lateral diffusion and variations in processing, assuming the series resistance and ΔL are the same for adjacent devices with different channel length:

$$I_D = \frac{\dfrac{\beta_o}{(1-\dfrac{\delta L}{L_m})}[V_{GS} - V_{tn}]V_{DS}}{1 + (\theta_s + \dfrac{2\beta_0 R_s}{(1-\dfrac{\delta L}{L_m})})\left[V_{GS} - V_{tn} + 2\lambda\sqrt{|V_{SB}| + 2\phi_F}\right]} \tag{5.9}$$

We define

$$V_0 = V_{GS} - V_{tn} + 2\lambda\sqrt{|V_{SB}| + 2\phi_F} \tag{5.10}$$

and

$$K_0 = L_m\beta_0 \tag{5.11}$$

Equation (5.9) can be reduced to:

$$K_0(V_{GS} - V_{tn})\frac{V_{DS}}{I_{DS}} = L_m(1 + \theta_s V_0) - \delta L(1 + \theta_s V_0) + 2R_s K_0 V_0 \tag{5.12}$$

Plotting $K_0(V_{GS} - V_{tn})\dfrac{V_{DS}}{I_{DS}}$ vs different L_m's, the intersect, x_0, on the x-axis is:

$$x_0 = \delta L - \frac{2R_s K_0 V_0}{(1 + \theta_s V_0)} \tag{5.13}$$

or

$$\delta L = x_0 + \frac{2R_s K_0 V_0}{(1 + \theta_s V_0)} \tag{5.14}$$

The slope of this curve is $(1 + \theta_s V_0)$, from which the surface scattering term θ_s can be extracted.[54] Equation (5.14) shows that the series resistance and the surface scattering term θ_s affect the δL determination.

5.4.5 Saturation Mode, and I_{on}

The theory of devices operate in saturation mode was developed in section 2.6. The current drive of the device, I_{on}, is defined as the drain current at normal power supply voltage applied to the gate and the drain. For instance, $V_G = V_D$ = 5V, for a 5V voltage supply. Therefore at this bias, the device is in saturation mode. The I_{on} is found to be a strong function of effective channel length and gate oxide thickness. From Eq. (2.117) in chapter 2, I_{on} is expressed as:

$$I_{on} = \frac{\mu_{eff} C_{ox} W}{2(L_e - \Delta L)} (V_{DD} - V_{t,sat})^n \tag{5.15}$$

where n can take a value of 1.8-2.5, V_{DD} is the operating voltage, and $V_{t,sat}$ is the saturation threshold voltage as was defined using plots of Figs. 5-14 and 5-15. A plot of I_{on} vs L_{eff} is shown in Fig. 5-18 for n and p-ch devices for a 5V power supply technology.

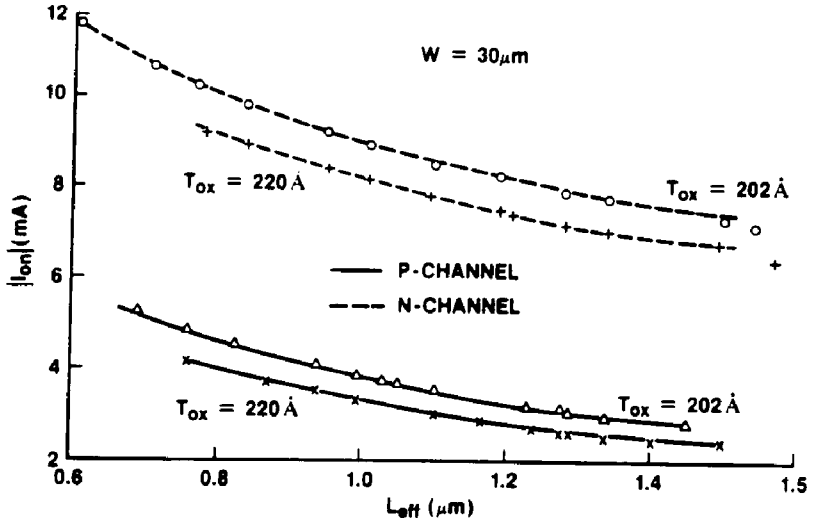


Figure 5-18. I_{on} vs L_{eff} of n- and buried p-ch MOSFETs.

From this plot the current drive variation with the 2 processing related parameters, t_{ox} and L_e, can be determined. Statistical data accumulated in a mature technology is used to refined the variation limits used in the circuit design process.

5.4.6 Body Effects and Channel Doping Profile

When the back gate bias is applied with respect to the source, V_{BS}, the threshold voltage changes according to the equation (2.94). Assuming the flat band voltage and the interface state

density do not change with back gate bias, the change in threshold voltage ΔV_t with respect to $V_{tn}(V_{BS}=0)$ is then derived directly from Eq. (2.105):

$$\Delta V_{tn} = V_{tn}(V_{SB}) - V_{tn}(V_{SB}=0) = \lambda\left[\sqrt{V_{SB}+2\phi_F} - \sqrt{2\phi_F}\right] \tag{5.16}$$

where λ is the body effect coefficient, For a channel doping concentration, N_B, the body effect coefficient is expressed as (Eq. 2.97):

$$\lambda = \frac{\sqrt{2\varepsilon_s q N_B(x)}}{C_{ox}} \tag{5.17}$$

λ is the slope of the curve ΔV_{tn} vs $(\sqrt{V_{SB}+2\phi_F} - \sqrt{2\phi_F})$ as shown plotted in Fig. 5-19 for a square ($20\times20\mu m^2$) NMOS device with gate oxide thickness of 150Å. In most practical devices, the channel doping varies with depth toward the bulk. Therefore, λ is not a constant when the space charge region is extended by increasing the back gate voltage. The bulk charge is expressed in term of depletion width, X_d, as:

$$Q_B = -\int_0^{X_d} qN_B(x)dx \tag{5.18}$$

where X_d is a function of bulk voltage, V_{SB}, as:

$$X_d = \alpha\sqrt{V_{SB}+2\phi_F} \tag{5.19}$$

α is defined by Eq. (2.20).

The channel doping profile can be extracted from this method, assuming within the incremental depletion layer for each back gate bias, the concentration $N_B(X_d)$ is constant. A large geometry device is used to obtain N_B from λ as shown in Fig. 5-19. As seen from the curve, λ is fairly constant, ~ $0.759V^{1/2}$ for the depletion width extended to 0.4μm from the surface. For shorter channel devices, the body effect is lower due to the charge sharing effects as discussed in chapter 2, where the effect of bulk bias is lessened due to the 2-dimensional field lines from the source and drain.

One circuit which is sensitive to body effects is the source follower circuit. This circuit is in fact the same as the access transistor in DRAM and SRAM memory cells, or the transfer gate in dynamic shift register circuit. The halo drain structure, as discussed in section 2.3.4, is to

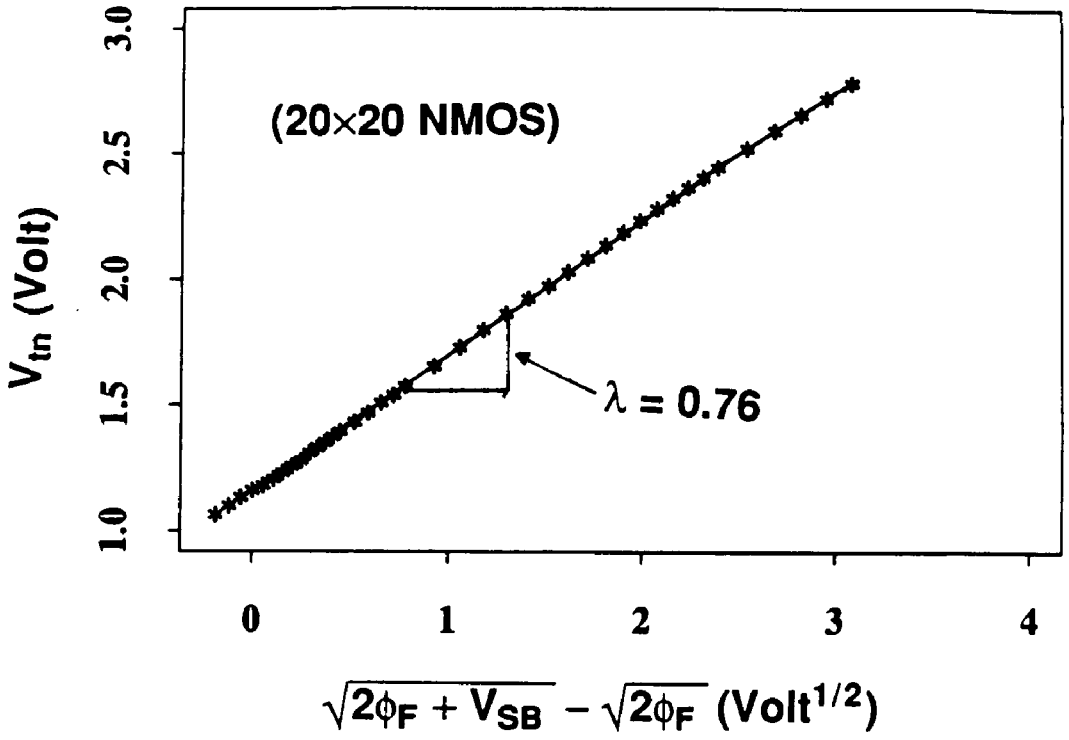


Figure 5-19. Change in threshold voltage vs $\sqrt{V_{SB} + 2\phi_F} - \sqrt{2\phi_F}$.

reduce the body effect by engineering the channel profile low enough for the λ term to be small, and yet the bulk punch-through is controlled by the boron implant at the source-drain junction. Figures 5-20 and 5-21 show the shift in linear voltage with back gate bias for the conventional device structure and the halo drain. The conventional device with a $V_{tn}(0)$ =0.65V shows a much larger shift in V_t than the halo device, $V_t(0)$=0.8V when V_{BS} varies from 0 to -4V. One measure of the body effect in terms of circuit design is how much the source voltage of a device drops if a full supply voltage is applied to the gate and drain. Shown in Fig. 5-22 are the source voltages of the conventional and halo devices connected in the source follower configuration. As indicated, the halo structure has 0.5 V higher V_S at V_G=5V than the conventional device, although the $V_t(0)$ is higher for the halo structure compared with the conventional (0.8V vs 0.65V). The body effect is important when the power supply voltage is scaled, thus affecting the storage high node in DRAM and SRAM memory cells.

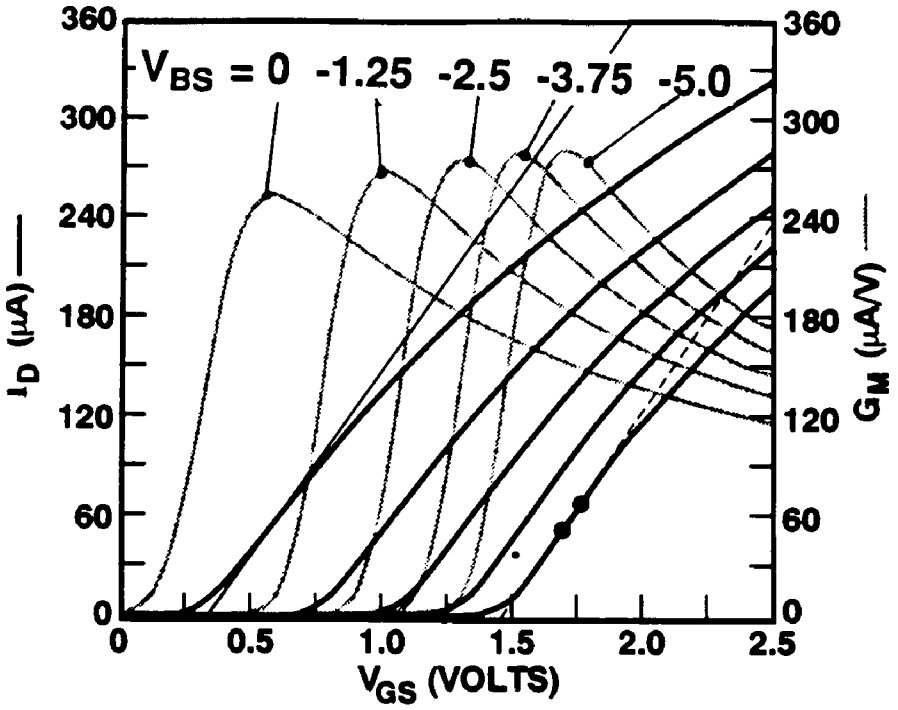


Figure 5-20. Shift in threshold voltage for conventional as a function of back gate voltages.

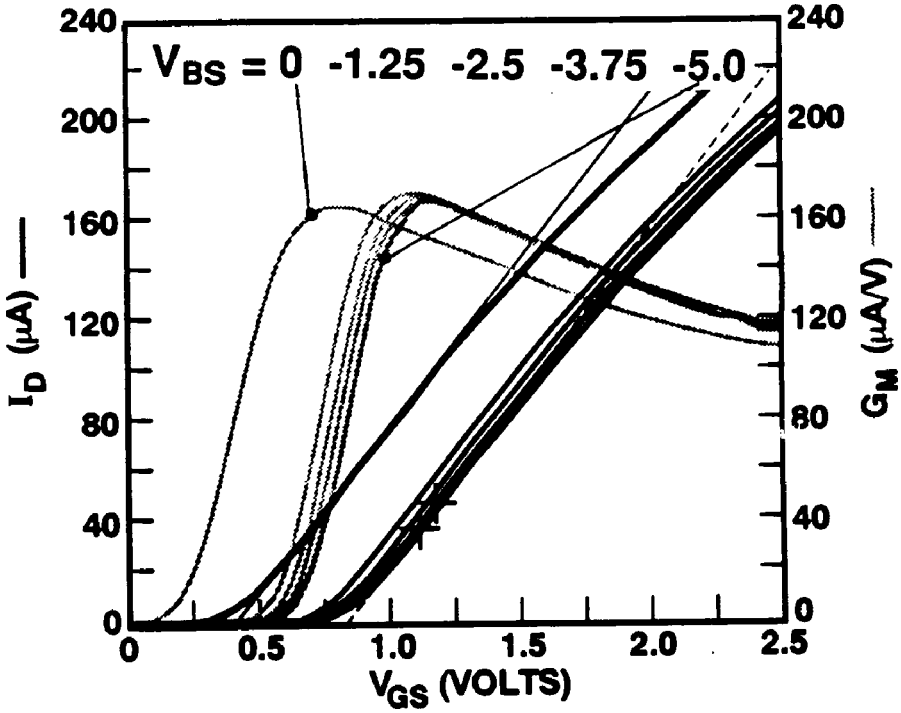


Figure 5-21. Shift in threshold voltage for halo drain device with $V_{in}(0)$=0.8V, as a function of back gate voltages.

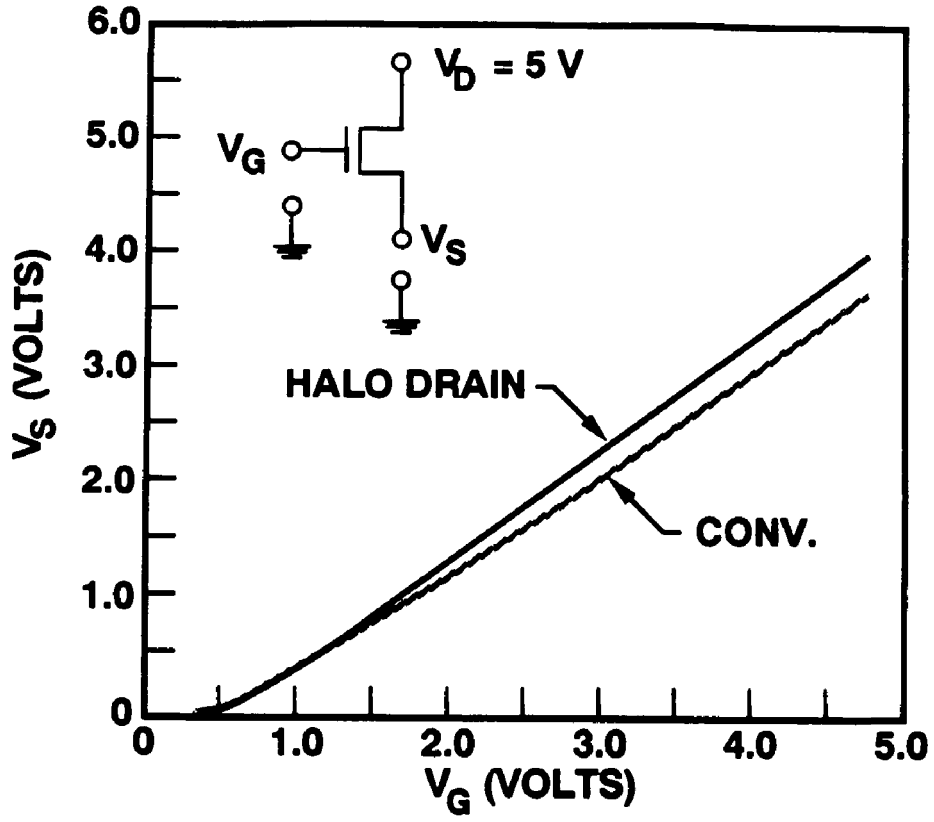


Figure 5-22. Source Voltage of a source follower transistor (or transfer gate),shows the different in V_s for the conventional and halo-drain device.

5.5 PARASITIC CHARACTERIZATION

5.5.1 Junction Leakage

Diode leakage is a severe problem in the formation of shallow source-drain junctions. This is particularly true, with silicided source-drains and metalization processes, such as tungsten plug and the use of barrier layers. Characterizing reverse-bias junction leakage is, therefore, an integral part of process development, but this will not be discussed in detailed in this dissertation.

Within the scope of this dissertation is the other cause for junction leakage from the device stand point, hot carrier induced drain junction leakage which will be reported in Chapter 7.

5.5.2 Junction Capacitance

In addition to the active device driving capability, circuit speed is limited by the presence of parasitic resistances and capacitances. Junction capacitance is one of the components slowing

down circuit speed, since the drain (or source) of a device is tagged with a capacitor that is to be charged or discharged every time the device in a circuit switches. A drain-to-tub junction capacitance is comprised of 2 components: the area, C_{JA} and perimeter, C_{JP} terms. The area term is determined by the vertical direction of the junction to the substrate (or tub), the perimeter term is the capacitance per unit length of the junction and the channel stop field isolation. A test structure with 3 different combined areas and perimeters is used to extract C_{JA} and C_{JP}. Shown in Fig. 5-22 and 5-23 are the junction capacitances of 3 different pair of diodes (n+/p and p+/n). The parameters are extracted by solving a set of 3 linear equations.

The junction capacitance is a function of both doping concentration (of substrate and channel stop) and the reverse applied voltage, V, and is expressed by (assuming one sided junction and nonuniform substrate doping $N_B(x)$):

$$C_J(V) = \frac{\varepsilon_s}{W_d} = \sqrt{\frac{q\varepsilon_s N_B}{2}}(V_{bi}+V)^{-n} \tag{5.20}$$

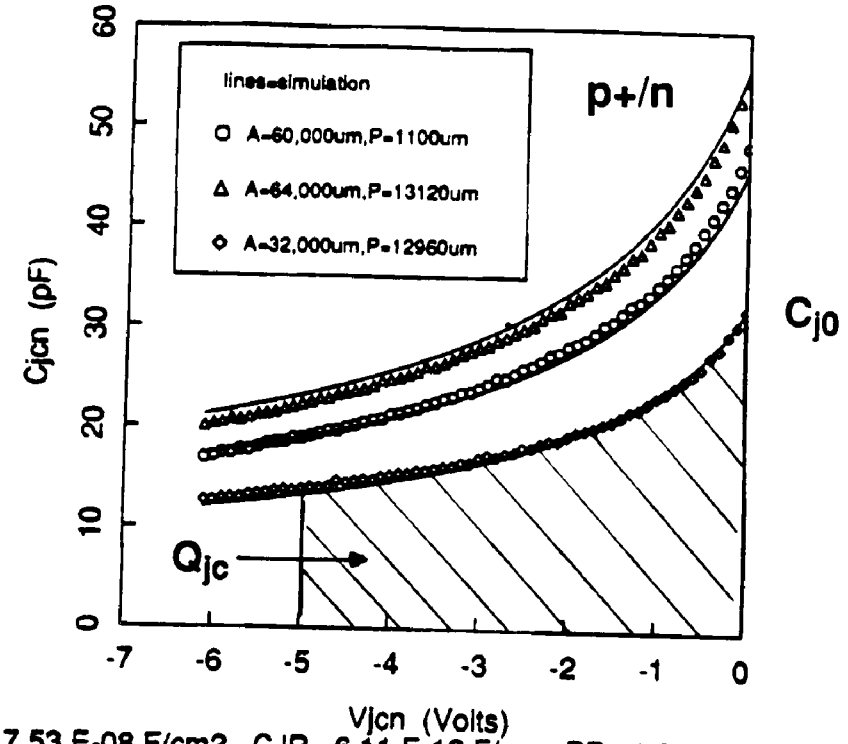
where V_{bi} is the junction built-in voltage, and n is the factor depending on the doping profile, ranging from 0.3 to 0.7. We define the figure of merit of a junction capacitance, the charge Q_{jc}, by integrating Eq. (5.20) over the operating voltage range:

$$Q_{jc} = \int_0^{V_{DD}} C_J(V)dV \tag{5.21}$$

The charge Q_{jc} is essentially the shaded area of Figs. 5-23 and 5-24. For a graded substrate profile, such as in the halo transistor, the capacitance C_{J0} at V=0 may be larger, but when the depletion width is extended into the lower doping substrate, C_J decreases rapidly, resulting in a lower integrated charge. Therefore, the total charge that a transistor has to move at a given junction node affects circuit speed more than the junction capacitance itself. This is one important aspect of device design with circuit performance considerations.

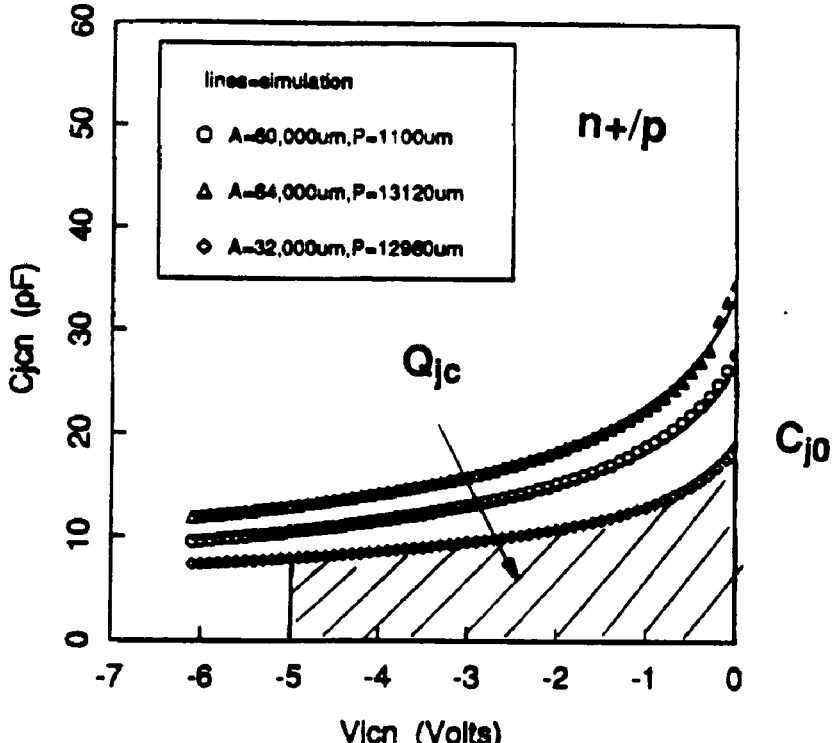
5.5.3 Isolation Leakage

The isolation between devices is a critical issue in CMOS structure integration. The intra-tub isolation is defined as the distance between adjacent devices within one tub separated by the field oxide. The isolation is characterized by a parasitic transistor with field oxide as the



CJA= 7.53 E-08 F/cm2, CJP= 6.11 E-12 F/cm, PB= 1.382 V, MB= 0.576 V

Figure 5-23. Junction Capacitances of 3 n+/p diodes with different areas and perimeters.



CJA= 4.4 E-08 F/cm2, CJP= 4.09 E-12 F/cm, PB= 0.782 V, MB= 0.467 V

Figure 5-24. Junction Capacitances of 3 p+/n diodes with different areas and perimeters.

dielectric. The conduction through the field device is caused by a channel formation by a poly line acting as the gate over field oxide, or by a drain-induced-barrier lowering if the two source/drain junctions are close together. The field transistor threshold voltage, V_{tF}, is designed to be higher than the operating voltage. It is typically chosen at least 2 times V_{DD}. Excessive channel-stop doping, to obtain high V_{tF}, can cause high junction capacitance, lower source-drain reverse breakdown voltage, and higher junction reverse leakage. These parameters should be optimized in the integration process.

In a CMOS technology using polysilicon gate and an LDD source/drain, the isolation width between the 2 active areas is affected by the poly gate etch and the oxide spacer etch. These etches can cause field oxide thinning in the open field areas not protected by the poly gate (or runners). This field oxide thinning will bring isolated source-drain features closer together, and drain-induced barrier lowering can cause leakage without a poly gate to help turn-on the device.

The inter-tub isolation is the distance across the n+ s/d and p+ s/d between the tub boundary. For leakage considerations, the inter-tub isolation distances can be considered two field devices, an n device formed between n+, p-tub and n-tub and a p device formed between p+ in n-tub and p-tub. A more severe effect of inter-tub isolation is latch-up, caused by transient current during circuit switching, or by inadvertently forward biased junction in either tub. Circuit layouts with more spacing between n+ and p+ active areas, and proper guard rings to collect minority carriers are effective ways to alleviate this problem.

5.6 SUMMARY

We have reported results of electrical device characteristics ranging from the subthreshold conduction regime, using g curves, to active mode characteristics in linear and under saturation bias conditions. The technology characterization verifies if the devices behave properly as predicted during the design phase, and the process parameters are well controlled. Analytical methods to detect process problems have also been introduced. Doping profiles in the MOS channel region and under source/drain junction are characterized, using back gate bias methods and Q_c figure of merit. Technology characterization is very important, not only to understand device physics and operations, but also to relate the interactions between processing conditions

and device performance.

In the next 2 chapters, we will focus on the high field effects that generate the substrate and gate currents which in turn may cause the drain-source breakdown and degradation in MOS devices.

Chapter 6

HOT CARRIER GENERATION

IN SHORT CHANNEL MOS TRANSISTORS

6.1 INTRODUCTION

As we have discussed at several occasions in this dissertation that hot carrier effects in the submicron CMOS devices are the most severe high field phenomena affecting the device scaling. The long term reliability and circuit performance can be degraded due to charge injection into the gate oxide and the spacer oxide over the drain region.

In this chapter, we will formally treat the various aspects of hot carrier generation process by means of characterizing substrate and gate currents. The roles of drain structures such as LDD and DDD in submicron devices are analyzed in terms of channel electric field reduction and device performance tradeoffs. The substrate current is a function of the device geometry, i.e. channel length, gate oxide thickness, drain structures; and bias conditions. The drain-source breakdown voltage, due to source injection, is a function of substrate current, effective bulk resistance and the placement of the tub tie contact. This parameter is preferred to be as high as possible to provide safe operating voltage margin and high reliability burn-in voltages.

We have found that the relationship between the peak substrate current and the applied drain voltage is exponentially proportional to the inverse of the square root of V_{DS}. This relationship is important in order to predict the lifetime of the device operated at nominal drain voltages from acceleration experiment. Comparison with the Double Diffused Drain (DDD) devices showed an order of magnitude improvement in device lifetime in favor of the LDD structure at shorter channel length.

6.2 HOT CARRIER GENERATION AND INJECTION MECHANISMS

It has been established that hot carrier effects in MOS transistors can lead to degradation in performance and long term reliability, especially when device geometry is scaled down to submicron dimension in channel lengths as in the 1.0 μm and below CMOS technologies. The

nominal gate oxide thickness used in these technologies are typically in the range of 125 to 210 Å, and the corresponding minimum effective channel length in NMOSFETs is in the range of 0.40 to 1.0 μm for technologies with design rules of 0.5μm to 1.0μm respectively. In order to alleviate hot carrier aging problems in the n-channel device we have incorporated an LDD (Lightly Doped Drain) structure to reduce the high channel electric field that causes carrier injection. The experimental optimization of the LDD structure for 1.0μm NMOS transistor has been described.[50] The n- phosphorus dose is chosen to be 4.0E13 ions/cm^2 (2.0E18 cm^{-3} on the surface) and the spacer width is 0.3 μm. The experimental results will be reported as an empirical formulation for predicting the device lifetime at different operating voltages. Comparisons among drain structures in different technologies will be made.

The hot carrier effects in short channel NMOSFETs have been well documented in the literature. Although hot carrier effects in submicron PMOSFET has been gaining much attention, the dominant effects are in the n-channel. We use NMOS transistor for the purpose of analysis of device degradation. The high channel electric field is the result of scaling both vertical and lateral device dimensions. This high channel electric field is in turn causing impact ionization of carriers and lattice near the drain region of the transistor. In this process hot electrons and holes are generated. The hot holes are collected by lower potential at the substrate contact or tub-tie constituting substrate current (Fig. 6-1). Hot electrons are collected at the drain causing the drain current to increase from its saturation value. If excessive substrate current is generated and not being collected effectively, the local potential in the channel close to the source junction can become positive with respect to the source potential. When this positive bias is greater than 0.7 V, the channel to source junction becomes forward bias and more electrons are generated into the channel. The MOSFET is now acting as a bipolar transistor where the channel bulk acts as the base, the source as emitter and drain as collector. Most of the electrons are collected at the drain because the lower electron potential. However, very small fraction of hot electrons gains high enough energy to surpass the silicon dioxide barrier causing conduction through gate oxide. This is the gate current.

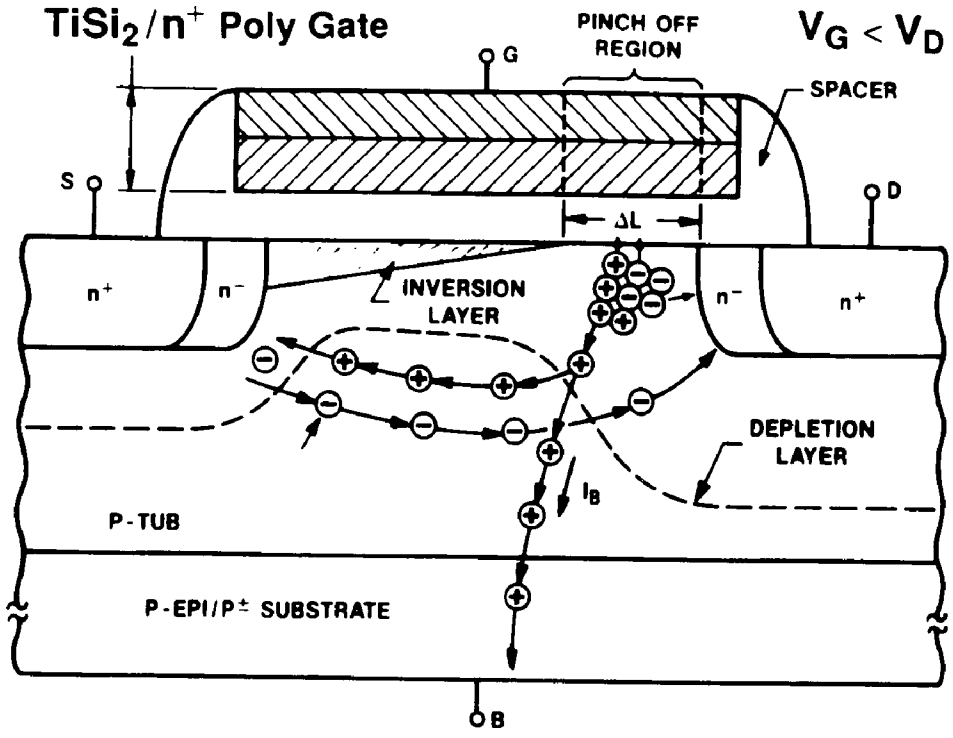


Figure 6-1. The processes of hot carrier generation, injection and drain-source breakdown conditions.

We will now formulate the analytical equations to model the channel electric field, the substrate current, gate current and drain-source breakdown due to hot carrier generation.

6.2.1 Channel Electric Field

When the channel electric field exceeds the critical field for surface electron mobility to saturate, the hot carrier effects become significant. This critical field is in the range of 4-5×10^4 V/cm. Maximum channel electric field is given by:[101]

$$E_m = \left[\frac{(V_d - V_{dsat})^2}{\Delta L^2} + E_{sat}^2 \right]^{1/2} \tag{6.1}$$

where

$$\Delta L = \sqrt{\frac{2\varepsilon_s}{qN_A}(V_{DS} - V_{Dsat})} \tag{6.2}$$

and

$$V_{dsat} = \frac{(V_g - V_t)L_e E_{sat}}{(V_g - V_t) + L_e E_{sat}} \tag{6.3}$$

E_{sat} is the critical field, 4-5×10^{4}V/cm, and L_{eff} is the effective channel length. Along the channel from the source to drain, the channel electric field is increased exponentially toward the drain-end and reaching E_m as shown in Fig. 6-2, from device simulation of a 0.5μm NMOS device.

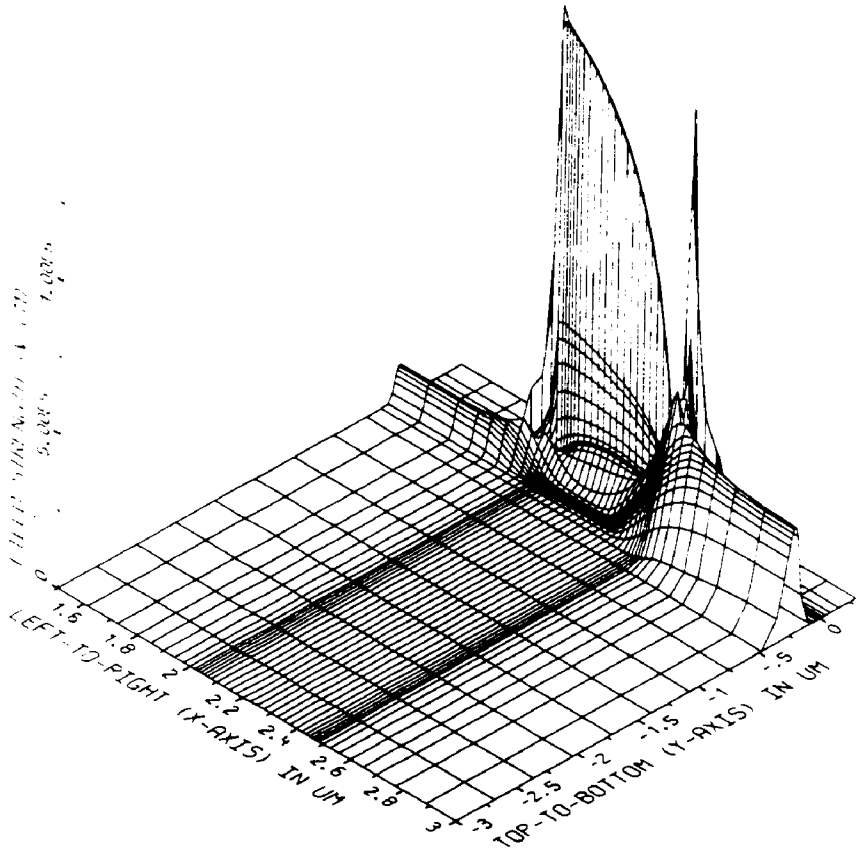


Figure 6-2. 3-D channel electric field for a DDD drain structure with V_D=5V and V_G=2V, E_m is calculated to be 2E6V/cm.

6.2.2 Substrate Current

The substrate current generated by impact ionization process when the device is biased in saturation regime can be derived as the following. The impact ionization coefficient, i.e. the events produced by one carrier per unit length, is expressed as a function of electric field as:[102]

$$M = A_i \exp^{-B_i/E} \tag{6.4}$$

Substrate current, for a given saturated channel current I_S (measured at the source end), is the sum of dI_{sub} at a interval dy in the pinch-off region (Fig. 6-1):

$$I_{sub} = \int_{L-\Delta L}^{L} I_S A_i e^{-B_i/E(y)} dy \tag{6.5}$$

Substitute dy with $(dy/dE)dE$ and using Eq. (6.1), we can derive

$$I_{sub} = \frac{A_i}{B_i}(V_D - V_{Dsat}) I_D \cdot \exp\left[\frac{-\Delta L B_i}{V_D - V_{Dsat}}\right] \tag{6.6}$$

where A_i/B_i = 1.2 V^{-1} and B_i = 1.7 - 2.0 $\times 10^6$ V/cm.

Another form to express the substrate current empirically is in terms of the pinch-off region ΔL and the impact ionization coefficient M:

$$I_B = I_S \cdot M \cdot \Delta L \tag{6.7}$$

where I_S is the channel current measured at the source, M is impact ionization coefficient, and ΔL Length of pinch-off region.

The multiplication factor M=I_B/I_S is often used as a figure of merit in terms of substrate current generation. The source, or channel, current is used since in the heavy saturation regime, the drain and source currents are not the same, as can be verified experimentally, the drain current is the sum of channel, substrate, and gate currents.

$$I_D = I_S + I_B + I_G \tag{6.8}$$

Typically, the gate current at high field is measurable, but insignificant compared with the channel current and can be ignored.

Figures 6-3 and 6-4 show measured substrate and gate currents vs V_G for different drain biases of a 0.5μm NMOS and PMOS devices. These devices use a spacer Double Diffused Drain.

It is observed that at the drain bias of 6V, the peak substrate current of the p-ch device is about 2 orders of magnitude lower than the n-ch device's $I_{B,max}$. However, the p-ch gate current not only occurs at lower gate voltage, but also is slightly higher in magnitude compared with its n-ch counterpart.

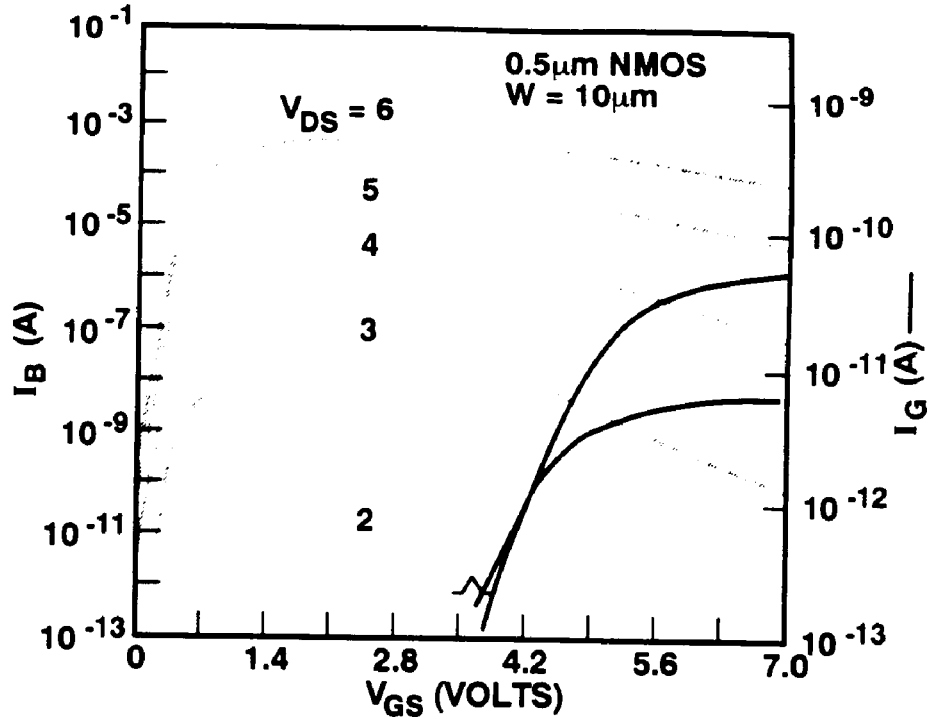


Figure 6-3. Substrate current and gate currents as a function of V_{GS} at different V_{DS}'s for a 0.45 μm NMOSFET.

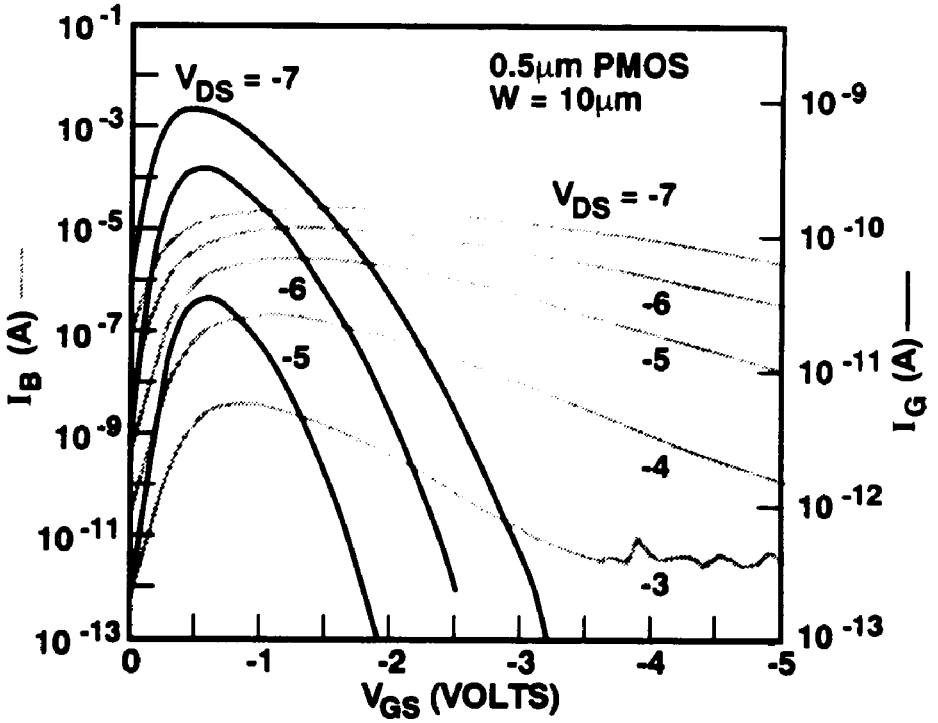


Figure 6-4. Substrate current and gate currents as a function of V_{GS} at different V_{DS}'s for a 0.43 μm PMOSFET.

It is well known that the devices are aged most severely if the drain and gate are biased in a condition that generates maximum substrate current. The maximum, or peak, substrate current in a short channel device extends over a wide range of gate voltage at high drain bias. The relationship of $I_{B,max}$ and V_{DS} is important in analyzing the aging results. The substrate current peaks at different gate voltages for different drain biases. We have found experimentally that for the range of applied voltage of interest (i.e. 3-6V or 3-8V for a 3.3V and 5V devices respectively), the gate voltage at which the substrate current peaks in a n-ch device is found to be:

$$V_{GS} = 0.4 \cdot V_{DS} \tag{6.9}$$

This result is fairly consistent for devices fabricated in different technologies and drain structures. We will use this result to determine our maximum aging conditions presented in chapter 7.

The relationship between the maximum substrate current to the applied drain voltage is important to extrapolate the aging lifetime of a stressed device to the predicted lifetime at the operating voltage. Using Eq. (6.2) for the pinch-off length ΔL in the substrate current Eq. (6.6) at maximum field, we have found that the maximum substrate current, $I_{B,max}$, is exponentially proportional to the inverse square root of the V_{DS} for an n-ch device, i.e.

$$I_{B,max} = A \cdot \exp(-\beta/\sqrt{V_{DS}}) \tag{6.10}$$

In arriving at Eq. (6.10), we have assumed that the saturation voltage V_{Dsat} is small as compared with the drain voltage at maximum field. Equation (6.10) can be experimentally verified, with a plot of log $I_{B,max}$ vs $1/\sqrt{V_{DS}}$ for 3 different channel lengths showing the straight lines fit (Fig. 6-5).[103] The same fit can be obtained for the 0.45μm NMOS device, with the substrate current as a function of gate and drain voltages as shown in Fig. 6-3. This device is fabricated with a different technology. This new relationship is different from that reported by Takeda and Suzuki.[104] In their paper, they reported a linear relation between log I_B and $1/V_{DS}$ which is not the case for the devices under consideration, as shown evidently in Fig. 6-6.

Other authors report a $1/(V_D - V_{Dsat})$ relationship, which shows a better fit with experimental data.[105] With their approach, V_{Dsat} has to be extracted for every drain and gate

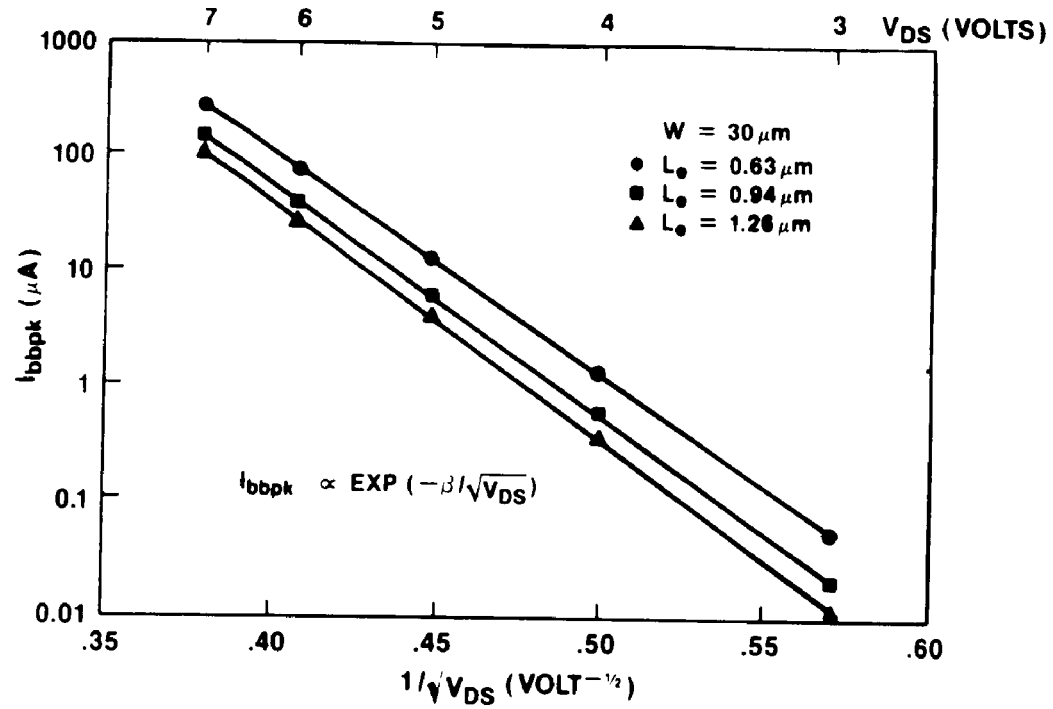


Figure 6-5. The maximum substrate current vs $1/\sqrt{V_{DS}}$ plotted on semi-log axes.

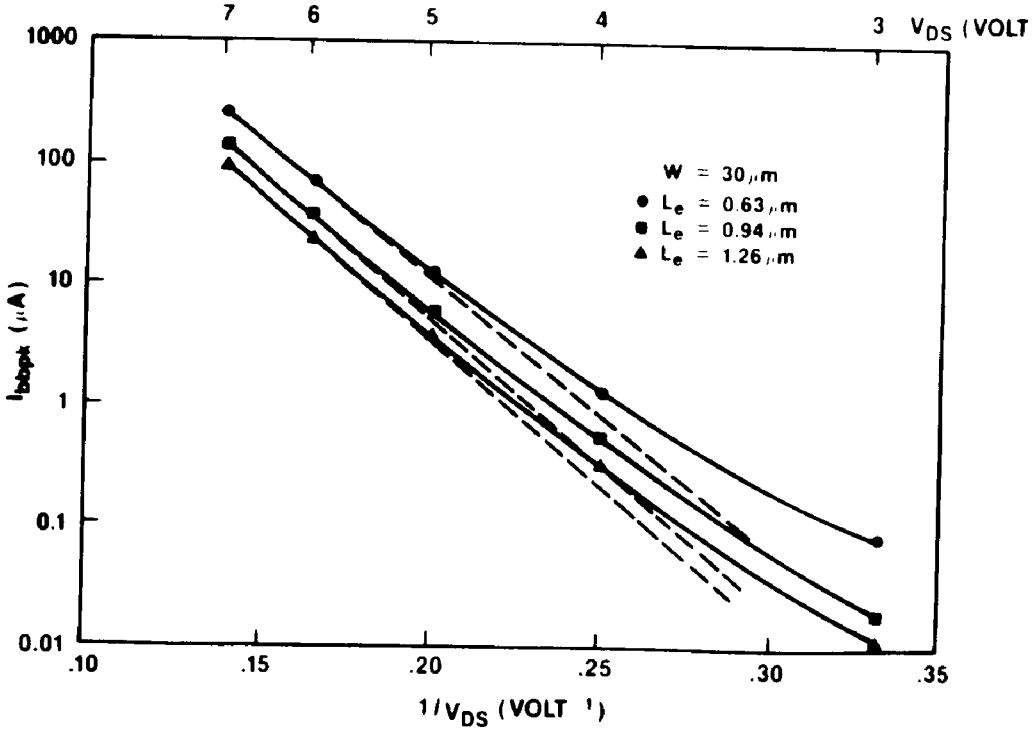


Figure 6-6. A conventional plot of $I_{B,max}$ vs $1/V_{DS}$ [104], indicating a deviation from the straight lines at low V_{DS}'s.

bias.

6.2.3 Analytical Model for Gate Current

In this section, the analytical expression of gate current is derived using lucky-electron model approach, similar to that of ref.[106] The lucky-electron concept was actually conceived by Shockley[107] to analyze the p-n junctions.

As we stated earlier in this chapter, the gate current is a small faction of substrate current, and starts conducting when the gate voltage is increased to high drain voltage as shown in Fig. 6-3. In the case of NMOS, the gate current is comprised of electrons, whereas for PMOS, holes are injected through the gate dielectrics. We will concentrate our analysis on the NMOS electron gate current.

Examining Figure 6-3, we observe that the measurable gate current appears at the gate voltage slightly lower than drain voltage, i.e. against the vertical field. Therefore, these hot-electrons must gain sufficient kinetic energy from channel carrier impact ionization in order to surpass the SiO_2 barrier. The process can be explained by the simulated contours of an NMOS device operated at a drain bias of 5.5V and a gate voltage of 2.2V, as shown in Fig. 6-7. From the generation center, the hot electrons acquire enough momentum by the favorable gate potential to reach the gate. The distance from the generation center to the Si-SiO_2 interface, d, so the carrier is accelerated by a vertical electric field E_x and surmount the in interfacial barrier Φ_b. This distance d is equal to Φ_b/E_x. Within this distance the probability of a traveling electron without any lattice collision is $e^{-d/\lambda}$, where a hot electron can surmount the barrier height without loosing energy and momentum. Therefore, the probability for a redirected electron to reach to Φ_b at the surface is:

$$P = e^{-(\Phi_b/E_x\lambda)} \tag{6.11}$$

The probability of an electron possessing energy $\Phi=\Phi_b + \Delta\Phi$ to surmount the interfacial barrier is approximated by:[108]

$$\frac{1}{2}\left[1-\sqrt{\frac{\Phi_b}{\Phi_b+\Delta\Phi}}\right] \approx \frac{\Delta\Phi}{4\Phi_b} \tag{6.12}$$

The approximation is valid for $\Delta\Phi \ll \Phi_b$. For a small increment of $d(\Delta\Phi)$, the probability of an

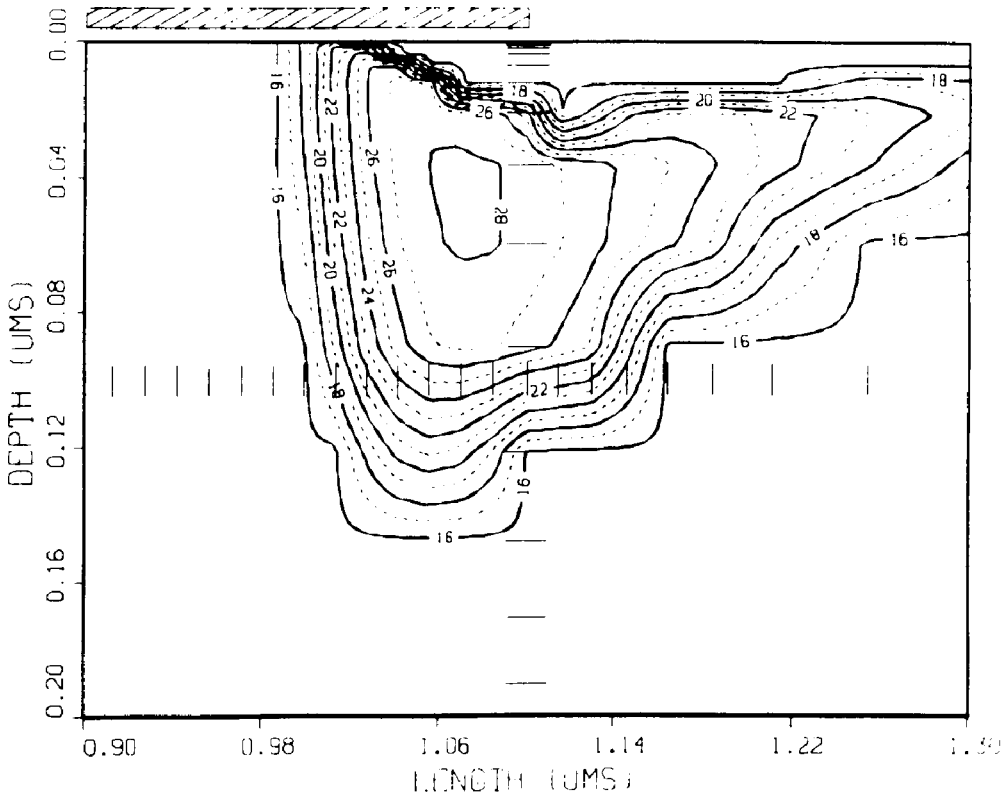


Figure 6-7. Generation center of Hot Carriers and Hot Electron Injected Gate Current.

electron to have this kinetic infinitesimal energy above $\Phi_b + \Delta\Phi$ is derived from Eq. (6.10)

$$\frac{d(1-e^{-(\Phi_b+\Delta\Phi)/E_x\lambda})}{d(\Delta\Phi)}d(\Delta\Phi) = \frac{e^{-[(\Phi_b+\Delta\Phi)/E_x\lambda]}}{E_x\lambda}d(\Delta\Phi) \tag{6.13}$$

Integrate the product of Eqs. (6.11) and (6.12) over all solid angle $\Delta\Phi$ of redirected electrons toward the interface.

$$P_{\Phi_b} = \int_{\Delta\Phi=0}^{\Delta\Phi=\infty} \frac{\Delta\Phi}{4\Phi_b}\frac{e^{-[(\Phi_b+\Delta\Phi)/E_x\lambda]}}{E_x\lambda}d(\Delta\Phi)$$

$$= 0.25\cdot\frac{E_x\lambda}{\Phi_b}\cdot e^{-(\Phi_b/E_x\lambda)}. \tag{6.14}$$

After being redirected, the probability P_1 for a hot electron to reach the surface without a collision different depths from the interface. At a y position along the channel close to the drain

region, the electron concentration at distance x from the surface is $n(x)$, then P_1 can be expressed as:

$$P_1 = \frac{\int_{x=0}^{x=\infty} n(x)e^{-(x/\lambda)}dx}{\int_{x=0}^{x=\infty} n(x)dx} \tag{6.15}$$

Solving the 2-D Poisson equation, we can obtain the electrostatic potential $\psi(x)$ at a given position y:

$$\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} = -\frac{q}{\varepsilon_s}(N_A + n) \tag{6.16}$$

By assuming the channel is inverted (i.e. $V_G > V_D$ for gate current) and the gradual channel approximation, i.e. $(\partial E_x/\partial x) \gg (\partial E_y/\partial y)$, Eq. (6.16) can be solved for $n(x)$ and substitute into Eq. (6.14) to obtain:

$$P_1 = 1 - \alpha e^{\alpha} E_1(\alpha) \tag{6.17}$$

where

$$\alpha = \frac{6kTt_{ox}}{q\lambda(V_G - V_D)} \tag{6.18}$$

Once the hot electron is injected over the oxide barrier, it has to survive the scattering in the oxide image-potential well before reaching the gate. Here, the probability for the carrier to reach the gate without being trapped or scattered, P_2, can be written as:[108]

$$P_2 = e^{-(x_0/\lambda_{ox})} \tag{6.19}$$

where

$$x_0 = \sqrt{\frac{q}{16\pi(V_G - V_D)C_{ox}}} \tag{6.20}$$

and λ_{ox} is the scatter-free mean free path in the bulk of oxide, and was found to be 32 Å by Young.[109]

Substitute all the constants, we obtain a simple expression:

$$P_2 = e^{(-300/\sqrt{(V_G - V_D)/t_{ox}})} \tag{6.21}$$

$$= e^{(-300/\sqrt{E_{ox}})} \tag{6.22}$$

where E_{ox} is the oxide electric field between gate and drain.

The product of P_1 and P_2 is the probability of a hot electron to surpass the barrier peak and arrive at the gate without collision:

$$P(E_{ox}) = P_1 \cdot P_2 \tag{6.23}$$

$$= [1 - \alpha e^{\alpha} E_1(\alpha)] \cdot e^{-(x_0/\lambda_{ox})} \tag{6.24}$$

The gate current is then expressed in terms of the probability factors P_{Φ_b} and $P(E_{ox})$ integrated over the channel length:

$$I_G = I_S \int_0^L P_{\Phi_b} P(E_{ox}) \frac{dy}{\lambda_r} \tag{6.25}$$

Assuming the peak field occurs at the drain end in the pinch-off region ΔL, Eq. (6.22) is approximated as:

$$I_G = I_S \frac{\Delta L}{\lambda_r} [P_{\Phi_b} P(E_{ox})]_{\max} \tag{6.26}$$

Evaluate Eq. (6.23) for the case $V_G \geq V_D$, and normalized to channel current, using substrate current Eq. (6.6), the relationship between gate, substrate and channel currents can be obtained:[101]

$$\frac{I_G}{I_S} = K \cdot t_{ox}^{1-\Phi_b/B,\lambda} \cdot P(E_{ox}) \left(\frac{I_{sub}}{I_S} \right)^{(\Phi_b/B,\lambda)} \tag{6.27}$$

The above derivations serves as a first order understanding of substrate and gate currents generation process. Inaccuracy caused by the many assumptions often leads to poor correlation with experimental results. Another approach to calculate the substrate and gate currents is to use numerical analysis, as described below.

6.2.4 Numerical Simulation of Substrate and Gate Currents

The modeling of the substrate and gate current as a function of gate voltage has been a great challenge for the device simulator developers. The one approach has shown good correlation

with measure data using the energy transport and lucky electron concept by Meinerzhagen.[80] The simulation using GALENE 2 device simulator,[79] with the energy transport, is compared with measured result for an LDD NMOS transistor with channel length of 0.88 μm, using threshold energy for impact ionization of 1.7V, and the distance from the center of impact ionization to the Si-SiO_2 was found in the range of 90-100 A, as shown in Fig. 6-7. Although numerical simulation is useful to compare different drain structures, as reported in Chapter 3, a close agreement with experimental results is still desirable. In general, the simulated substrate current is within a factor of 2-3 as compared with measured data. We, therefore, use the device simulator for comparison and device analysis purposes only, but not in an absolute treatment of hot carrier problem, particularly in the areas of breakdown and device aging. We will resolve to experimental measurements to characterize the substrate current and breakdown, before subjecting the devices to hot carrier aging.

6.3 SUBSTRATE AND GATE CURRENTS OF LDD AND DDD NMOS TRANSISTORS

In this section we will present the substrate and gate current of LDD devices fabricated with different spacer widths and LDD phosphorus doses. The characterization of the fabricated devices as shown in Fig. 6-1 has been published in ref.[50]. The optimization of LDD doping and spacer width led to an improvement over the early concept of LDD devices, as published by other authors.[70],[110] The substrate and gate currents of an NMOS device with 2 different LDD implants are shown in Fig. 6-8. The normalized peak substrate current with respect to source current is shown in Fig. 6-9. It is observed that the higher NLDD dose (2E18cm^3) device produces less substrate current and that the multiplication factor is also lower as compared with device with lower LDD dose (1E18cm^3). The double humps in substrate current associated with lighter LDD dose, reported by Hui et al.,[110] was not observed here (Fig. 6-8). A moderate NLDD dose has an additional advantage on the active current drive, since the source-drain series resistance is lower. Further development of the moderate LDD and its variations have been published. [72]

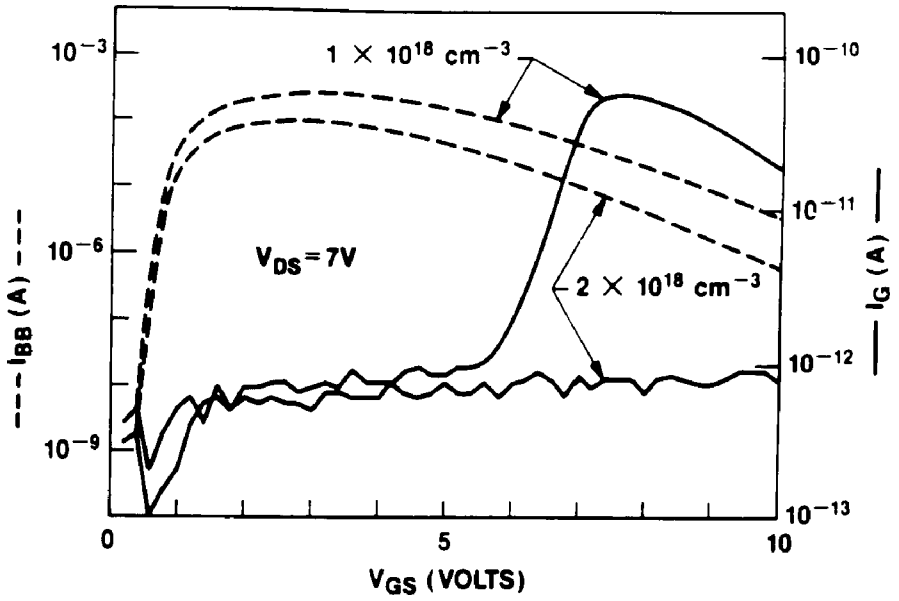


Figure 6-8. Substrate and gate current for the .98μm NMOS transistors with spacer width of 0.24μm and 2 different n- implants.[50]

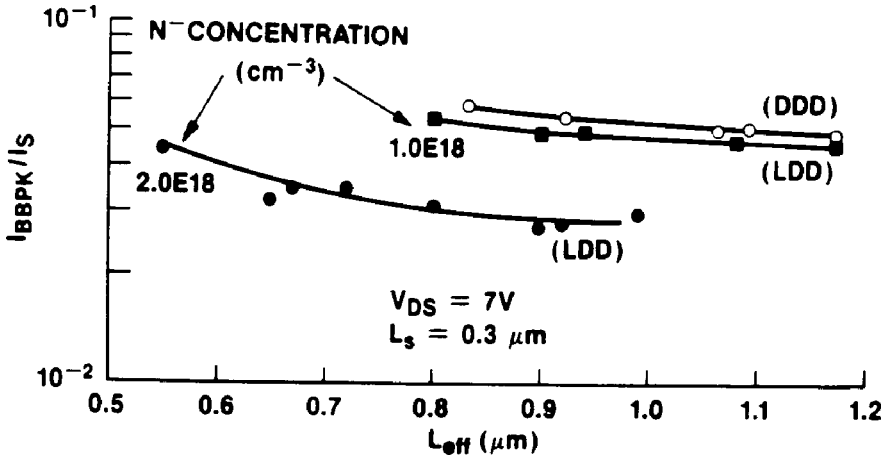


Figure 6-9. Normalized peak substrate current ($I_{sub,max}/I_S$) vs L_{eff} for LDD and DDD n-ch devices.[50]

6.4 DRAIN-SOURCE BREAKDOWN VOLTAGE

As we mentioned earlier, the electron-hole pairs generated by impact ionization are collected at the drain (electrons) and the substrate (holes). As the holes move toward the negative potential of the substrate (or tub tie), they can create a local positive potential near the source of the device. When this potential is greater than ~ 0.6 V, the source-substrate junction is forward biased. The device now behaves as a bipolar transistor being turned on. The electrons emitted from the source (emitter) into the base (channel bulk) get collected at the drain (collector) (refer to Fig. 6-1). When this situation occurs, the drain current increases abruptly as shown in Fig. 6-10 and the voltage at the knee of I-V curve is defined as drain-source breakdown voltage, V_{DSB}. The breakdown voltage occurs at different gate voltages, as shown plotted in Fig. 6-11. This contour is in agreement with the bell shape curves of substrate current shown in Fig. 6-3, i.e. V_{DSB} is at minimum when I_B is at its peak.

In Figure 6-12, the V_{DSB} measured at V_G = 3 V is plotted against the effective channel length for 1.0 μm and 1.5 μm CMOS technologies. The V_{DSB} as a function of L_{eff} is expressed as:

$$V_{DSB} \propto (L_{eff})^m \tag{6.28}$$

where m is ~ 0.27.

The effect of the bulk resistance, by varying the p-type epitaxial layer thickness on a p+ substrate, on the drain-source breakdown voltage is shown in Fig. 6-13. As the epi thickness is reduced the effective bulk resistance from the channel region to the p+ substrate decreases. Therefore a higher drain voltage is needed to produce the substrate current necessary to forward bias the bulk-to-source junction. As seen in Fig. 6-13, at L_{eff} = 1 μm, an improvement of ~ 3 V and 1 V in V_{DSB} is obtained for 7 μm and 10 μm epi devices respectively as compared with the control devices (16 μm epi). The differences in V_{DSB} for other epi thicknesses (13 and 22 μm) are not obvious.

From the operating voltage margin stand point, one would like to have as high a V_{DSB} as possible. This can be achieved by decreasing I_B, using an alternative drain structure such as the

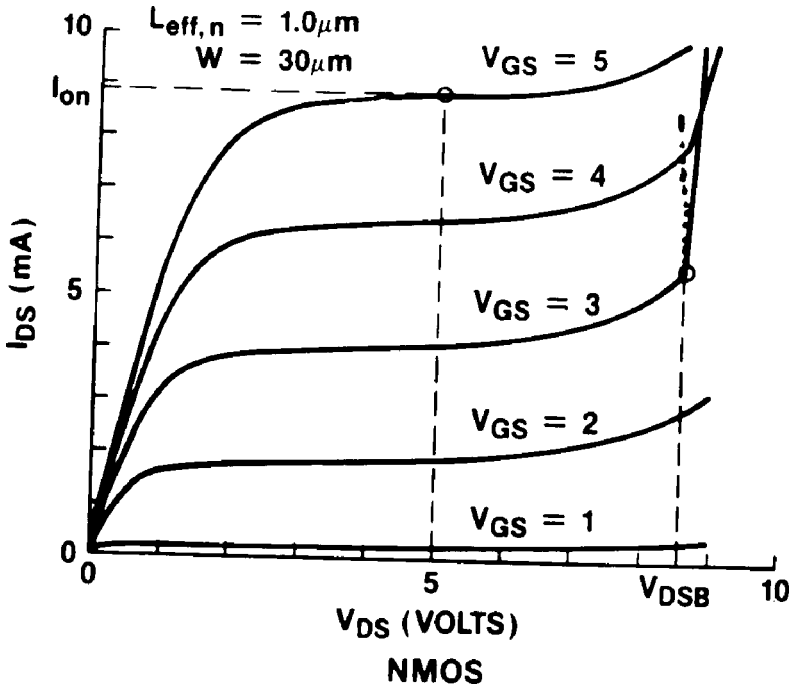


Figure 6-10. A typical I_{DS} vs V_{DS} illustrates the measured breakdown voltage, V_{DSB}.

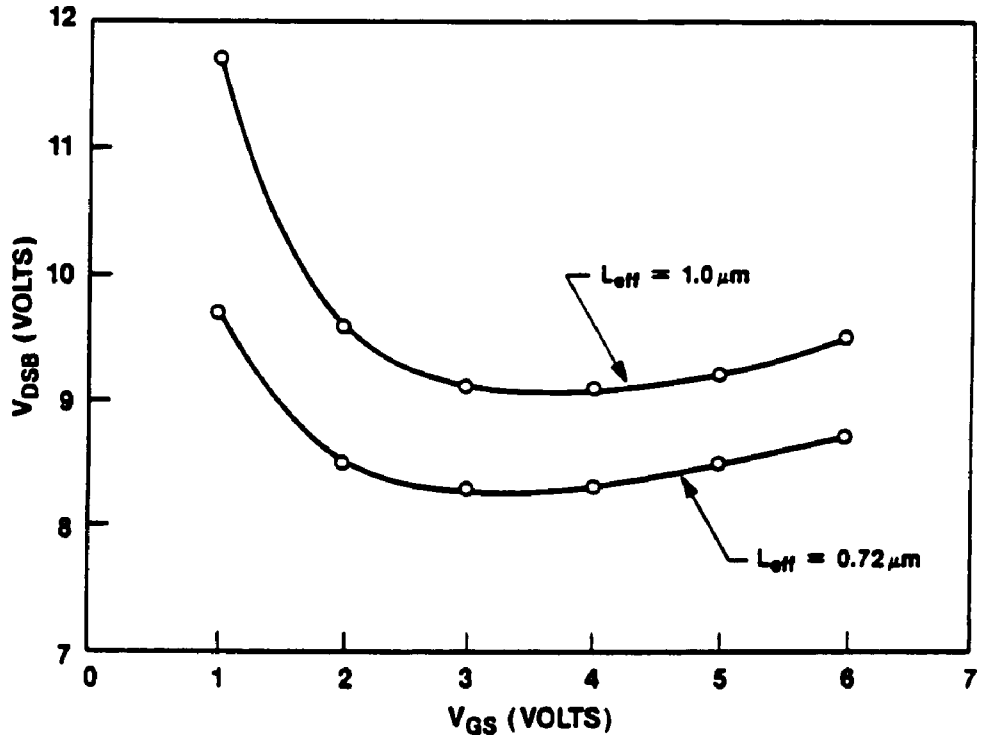


Figure 6-11. Breakdown voltage as a function of gate voltage for an NMOS device.

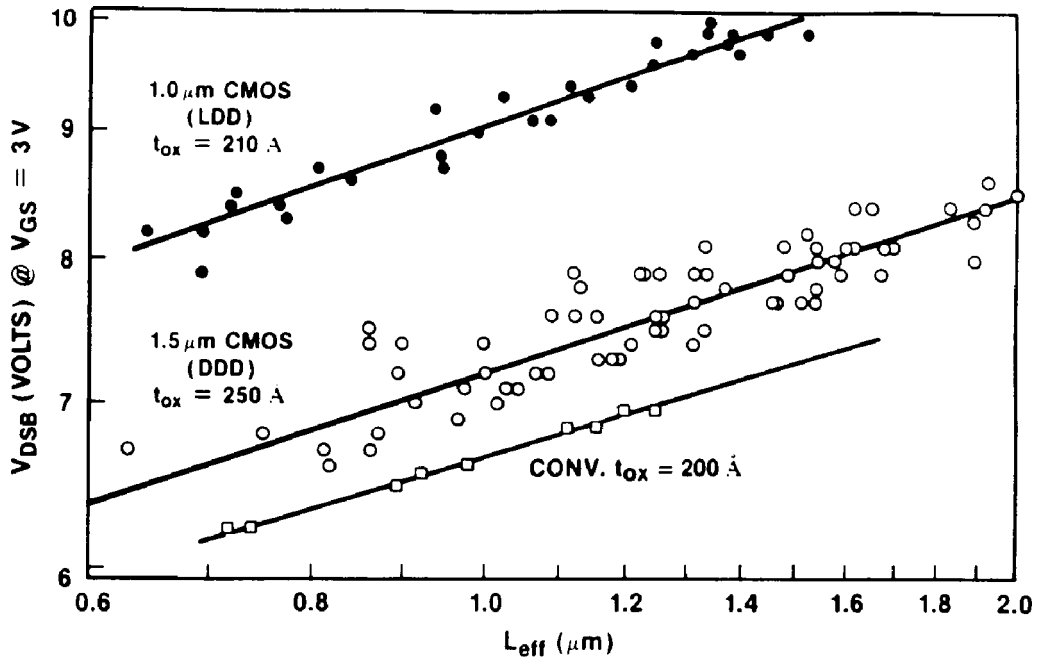


Figure 6-12. V_{DSB} at V_G=3V as a function of effective channel length for different NMOS device structures, 1.0 μm (LDD); 1.5 μm (DDD) devices.

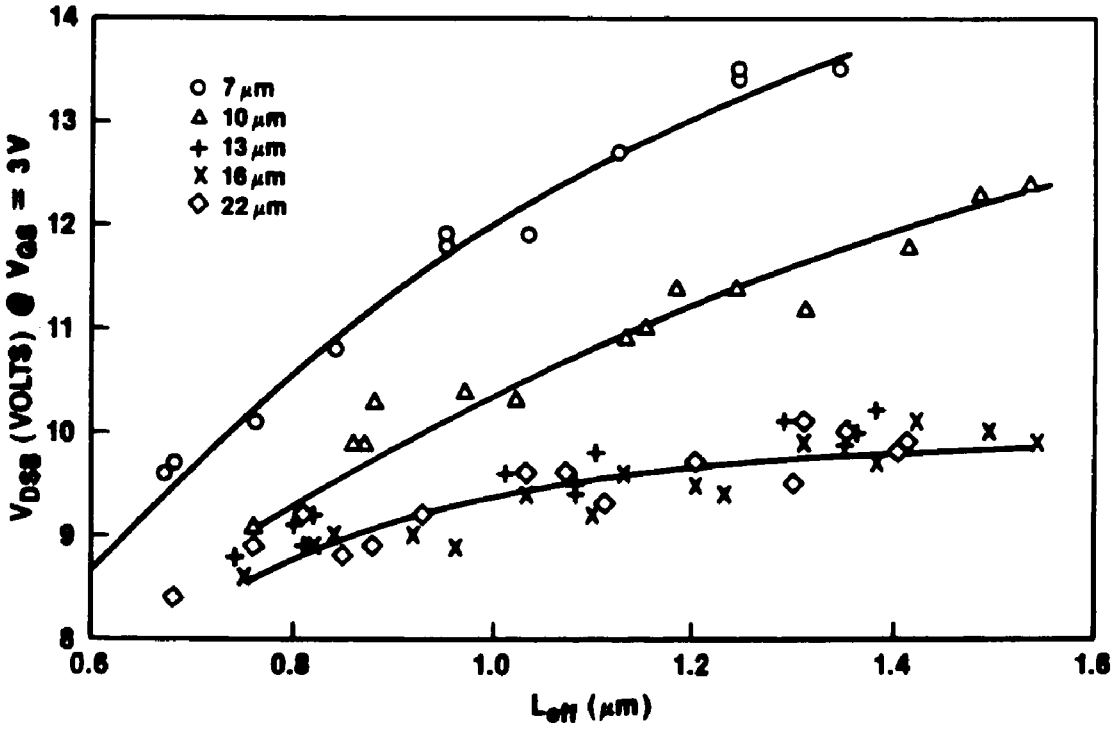


Figure 6-13. Effects of epi thickness (effective bulk resistance) on breakdown voltages.

LDD; or by reducing the effective bulk resistance by using a thinner epi thickness (Fig. 6-13). The choice of starting substrate material and epi thickness has an important impact on device integration. p-epi on p+ substrate is preferred for high break down voltage in an NMOS device. For a given drain structure, the V_{DSB} and maximum substrate current $I_{sub,max}$ are inversely related as shown in the experimental data in Fig. 6-14.

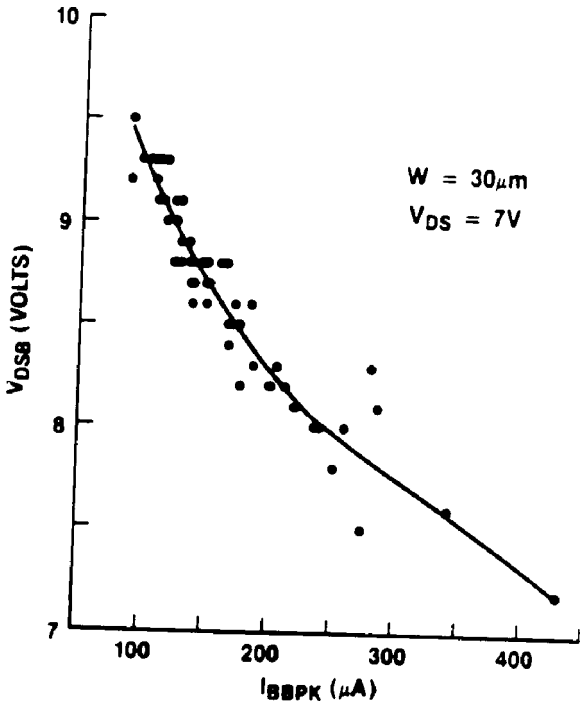


Figure 6-14. Correlation between drain-source breakdown voltage and peak substrate current, $I_{B,max}$. Substrate current measured at V_{DS} = 7 V. Channel lengths are varied.

6.5 SUMMARY

In this chapter, we have determined that the hot carrier generated substrate currents lead to charge injection through the gate oxide at high gate bias. A more severe affect on the maximum operating voltage is the source-drain breakdown due to the bipolar action in the channel. This effect has been characterized and the use of different drain structures to reduce the substrate current, and the thinner epi thickness to reduce the effective bulk (or equivalent to a base resistance) resistance can gain higher breakdown voltages. Once the maximum operating voltage limitation is overcome, the hot carrier aging affects on the devices operated at lower than $V_{D,max}$ can cause severe device degradation. This subject will be treated in the next chapter.

Chapter 7

CMOS DEVICE AGING BY HOT CARRIER INJECTION

7.1 INTRODUCTION

In this chapter, the aging experiments and analysis of the results will be described in details. The hot carrier aging measurements at accelerated biases under dc stress of LDD n-channel MOSFETs with channel lengths from 0.4 to 0.8 μm will be reported and analyzed using the subthreshold conduction of MOS device to characterize the interface trap distribution and relate to a degraded transconductance $\Delta g_m / g_{m0}$ model.

In the aging stress experiments, it was observed that the transconductance, g_m, degradation is the most sensitive parameter when the device was aged initially. After that the interface state is built up at the drain-end of the device and spread out toward the source end. The fixed negative charge is also increases resulting in a positive threshold shift when $\Delta g_m / g_{m0}$ changed by more than 10 %. Post-stress output conductance increases in forward mode indicating channel shortening effect. However, in the reverse mode the reduction in saturation was observed. The lifetime of the device and the substrate current during aging are found to be governed by an inverse power law relationship. The lifetime prediction at normal operating voltage will be presented.

7.2 DEVICE AGING EXPERIMENTS AND OBSERVATIONS

The channel lengths of the devices aged are in the range of 0.4-1.0μm, with oxide thickness in the range of 125 to 210 Å. Most devices used had passivation layer of SiN or Nitride/Oxide caps, and aged at room temperature unless otherwise noted. The transistor widths varied from 10, 15, 20 and 30 μm. The substrate current is normalized to the transistor width to ease in the analysis and comparison.

7.2.1 Experimental Set-up

An array of 3 transistors of the same coded length with separate drains and gates, and common source were aged at different drain and gate voltages. This technique allows transistors

of about the same channel lengths to be stressed at different fields in order to extrapolate the lifetime at operating voltage more rapidly. The layout of the transistors are shown in Fig. 7-1 for the coded channel lengths of 0.5μm and 0.6μm for both n and p-ch devices.

The measurement set-up is shown in Fig. 7-2, where a computer controlled Semiconductor Analyzer, HP 4145A, was used to analyze the initial device characteristics; to apply the bias during the timed stress, and to measure the device characteristics at time intervals. The terminal connections are shown in Fig. 7-2, where the 3 drains are connected to SMU1, 2, and 3. The substrate, or well, is connected to SMU4 for substrate current measurements. The gate of transistor 1 is connected to source voltage Vs1, and the gates of transistors 2 and 3 are connected to Vs2. For the more sensitive current measurements used for detail analysis, one transistor is aged at a time, and shielded coaxial probes were used for low current measurements.

The devices to be stressed were screened for any defects, such as junction leakage or abnormal characteristics. Since the 3 devices of the same coded length are in the vicinity of each other, the effective channel lengths were in general matched, thus allowing the similar field distribution for the same device geometry.

Initial device parameters such as threshold voltage, V_t, transconductance, g_m, current drive I_{on}, I_{off}, and the substrate current, I_B, at operating voltage are first measured. Then the voltages are applied to device terminals for stressing. During aging, the terminal currents are measured at time intervals, typically on logarithmic scale. The subthreshold slopes at different drain biases at constant subthreshold current are measured for the spatial interface trap density calculation, as will be discussed in the next section. The all results were recorded and subsequently analyzed using graphics software tools for data analysis.

7.2.2 Observations

The recorded terminal currents show that the gate current is not measurable for drain and gate bias at peak substrate current condition, i.e. $V_G = 0.4V_D$. Therefore we can neglect the charge conduction through oxide that could fill the bulk trap in the oxide. The substrate currents measured at stress and operating voltages changed very little and they will not be analyzed. We will concentrate on linear drain current for transconductance calculation. We have observed that

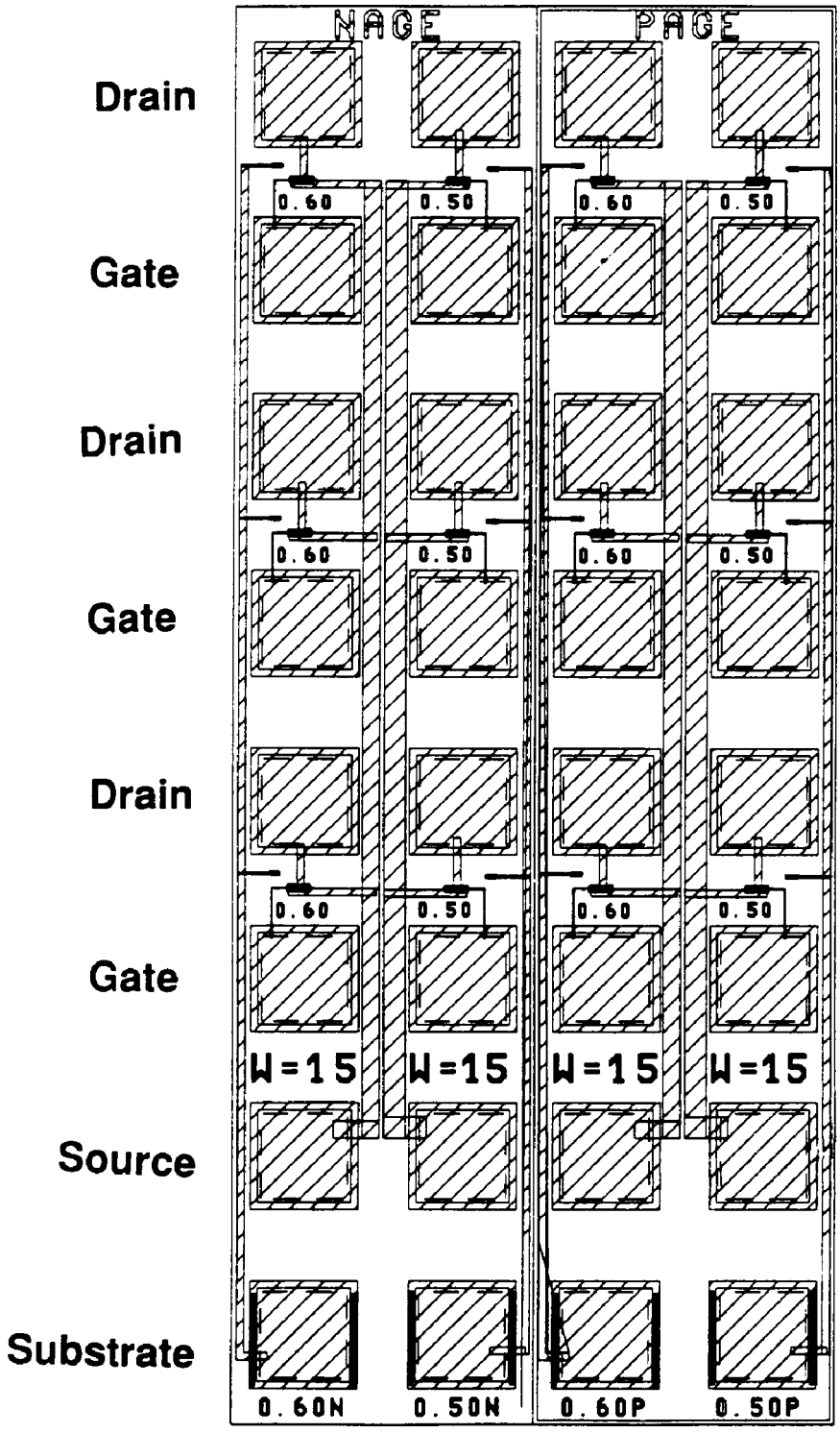


Figure 7-1. Layout of 0.5 and 0.6μm CMOS devices for DC aging experiments.

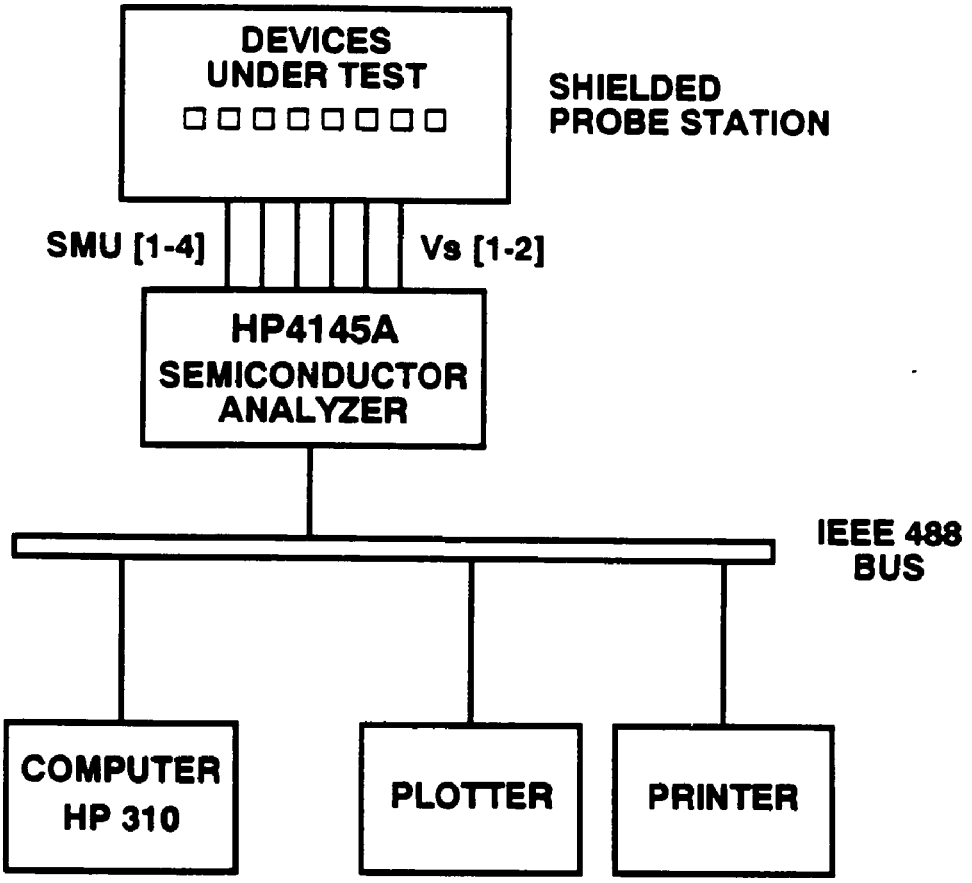


Figure 7-2. Aging Stress Measurement System with a shielded probe station.

the maximum transconductance, g_m, degrades most noticeably under stress whereas the threshold voltage shift is small initially and then becomes more pronounced as g_m degradation surpasses 10%.[100] The criteria of 10 % change in transconductance has been chosen for device lifetime analysis. This choice will be proved conservative later. The I-V characteristics of a 0.71 μm effective channel length device before and after 1310 minute stress at V_D = 8 V and V_G = 3 V are shown in Figs. 7-3 and 7-4 and will be explained in the following sections.

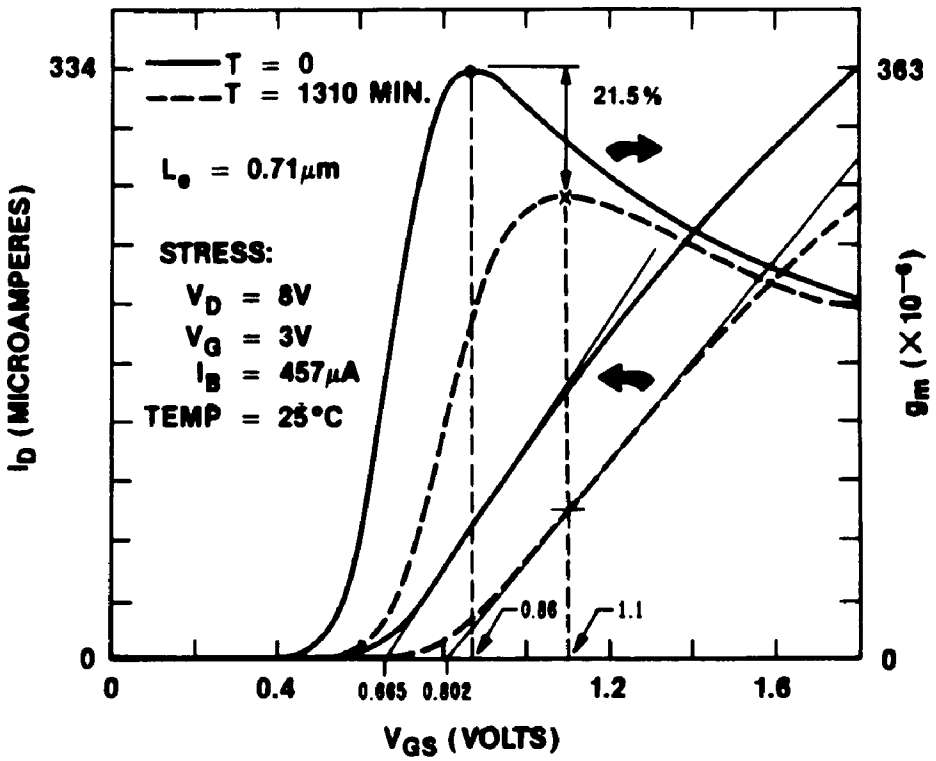


Figure 7-3. Typical transfer characteristics in linear mode before and after dc bias stress.[100]

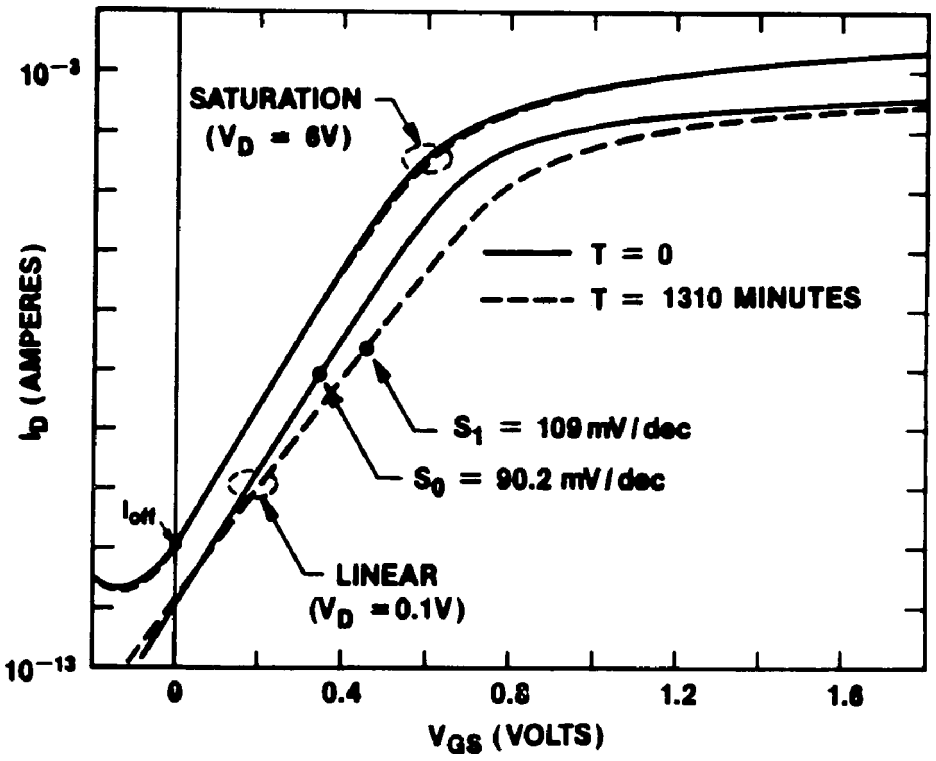


Figure 7-4. Subthreshold curves in linear and saturation mode to monitor subthreshold swing S and I_{off}.[100]

7.2.3 g_m and V_t Changes

The g_m degradation with respect to stress time is governed by an empirical expression:

$$\frac{\Delta g_m}{g_{m0}} = A \cdot t^n \tag{7.1}$$

where A is the proportional constant depending on the technology, t is the stress time, and n is in the range of 0.5 to 0.7. Equation (7.1) is generally valid for $\Delta g_m/g_{m0}$ of less than 15 %. Figure 7-6 shows the percentage of change in g_m and V_{tn} vs stress time. For longer stress times, the degradation tends to saturate. The threshold voltage, on the other hand, shifts in the negative direction initially ($\simeq$ -2mV) then changes to a positive shift as $\Delta g_m/g_{m0}$ degrades more than 10 % as seen in Fig. 7-6. This positive shift in V_t can be explained by the trapping of negative interfacial charge at the Si-SiO_2 interface.

7.2.4 Changes in Saturation Characteristics

In addition to drifts in g_m and V_{tn}, the saturation current drive I_{on} also decreases slightly in the forward mode (The drain and source are in the same order as during stress). The output conductance ($\delta I_{DS}/\delta V_{DS}$), i. e. the slope of the saturation I-V curves, increases after stress (Fig. 7-6). This is an indication of short channel effects due to channel shortening by a net negative trapped charges at the drain, serving as a drain extension. This degradation mechanism should be modeled properly for circuit applications such as access transistor or bilateral switch.

The reverse characteristics (i.e. the drain-source interchanged), on the other hand, show significant reduction in saturation current especially for low gate voltages (Fig. 7-7). The triode region of forward and reverse curves after stress are essentially the same indicating at low field the mobility reduction is independent of the mode of the drain and source are interchanged. However, at high drain bias, the reverse mode produces less channel current. The parallel slopes of the reverse I-V curves in saturation regime (dotted lines in Fig. 7-6) indicate the output conductance is unchanged. Figure 7-7 shows the channel and substrate currents at V_{GS} = 3 V after stressing in forward and reverse modes. The substrate current in the reverse mode is lower due to a reduction in channel current due to increasing threshold voltage. The drain-to-source breakdown voltage is therefore higher in reverse mode as shown in Fig. 7-7, although the

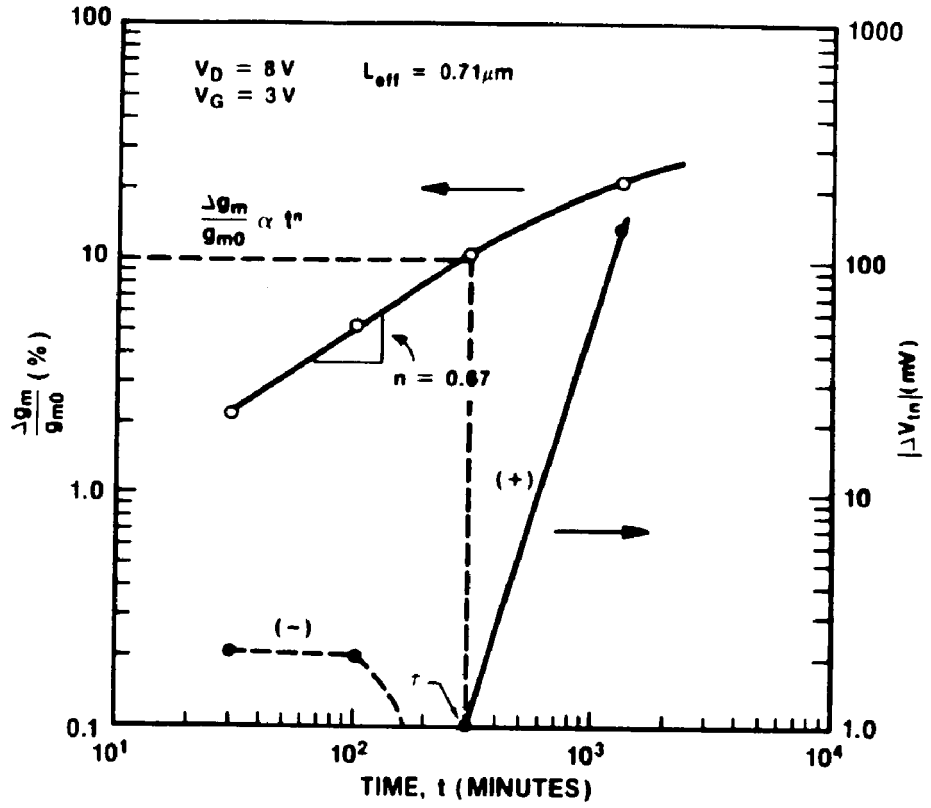


Figure 7-5. Changes in transconductance, g_m, and threshold voltage, V_{th}, vs stress time.

multiplication factor M = I_B/I_S at breakdown is about the same in both modes.

Figure 7-8 is an example of substrate current vs V_{GS} curves for a 0.73 μm L_{eff} device after 1200 minutes stress at V_D=7.5 V and V_G=3 V. The I_B curves are higher in forward mode and lower in reverse mode as compared with the initial curves. This is a direct consequence of saturation channel current being different in the two modes of operation as discussed above.

7.2.5 Hot Carrier Induced Drain Junction Leakage

For the devices with very short channel length, i.e. smaller than the minimum allowable L_{min} for a given technology, subjected to aging at accelerated voltage close to breakdown, drain junction leakage after stress was observed. Figure 7-9 shows the subthreshold I-V curves of a 0.61μm device (for a nominal 1.0μm channel technology) stressed at V_D = 8V and V_G = 3V for 170 minutes. The I_{off} current is measured at V_D = 6V and V_G = 0V indicating a substantial increase, in the μA range. However, this current is not observed at the source end (dotted line), indicating that the leakage is from the damaged drain junction. This phenomena was also

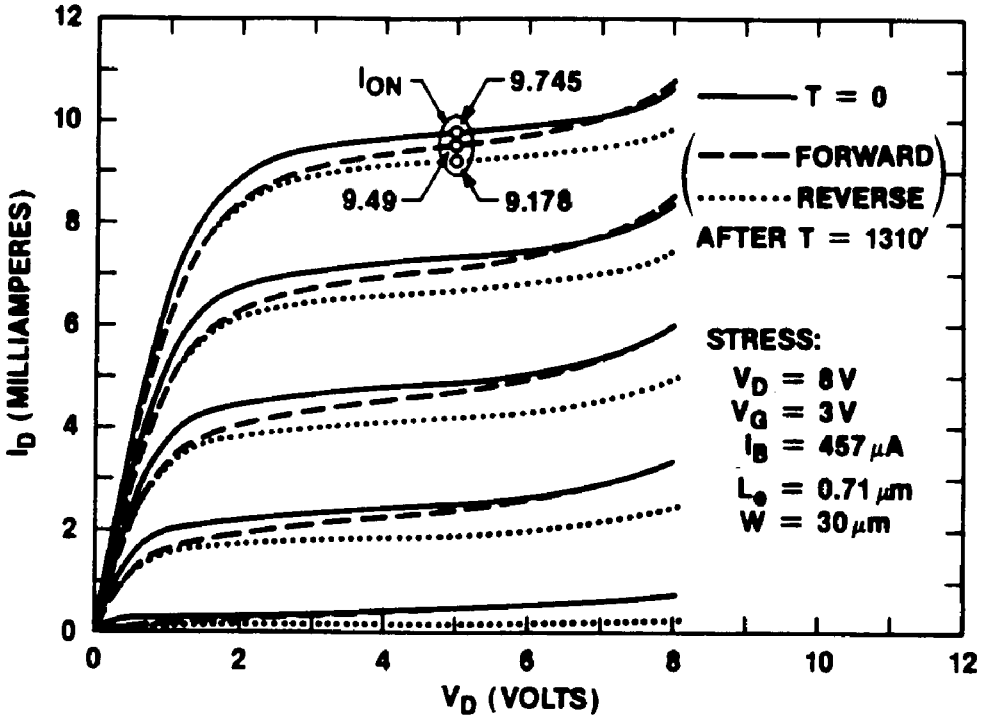


Figure 7-6. Forward (dashed) and reverse (dotted curves) I_D-V_D characteristics before and after stress.

confirmed by other authors.[108] This drain leakage mechanism, although less emphasis in the literature, is a very severe reliability problem for DRAM and high value resistor load SRAMs. Since it would degrade the charge retention in a DRAM and affect the cell balance in an SRAM circuit.

From the observations from our published results, we will concentrate on the analysis of interface state built-up that causes device degradation.

7.3 THE ANALYSIS AND MODELING OF HOT CARRIER AGING MECHANISMS

In this section, we will describe in details the subthreshold conduction method, used to extract the spatial distribution of generated interface states as a result of hot carrier stressing. Attempts to determine spatial distribution of interface states had been made by using the conductance method,[109] and the charge pumping technique.[110] In the conductance method, the authors had attempted to evaluate the length of the damaged region, ΔL, without further

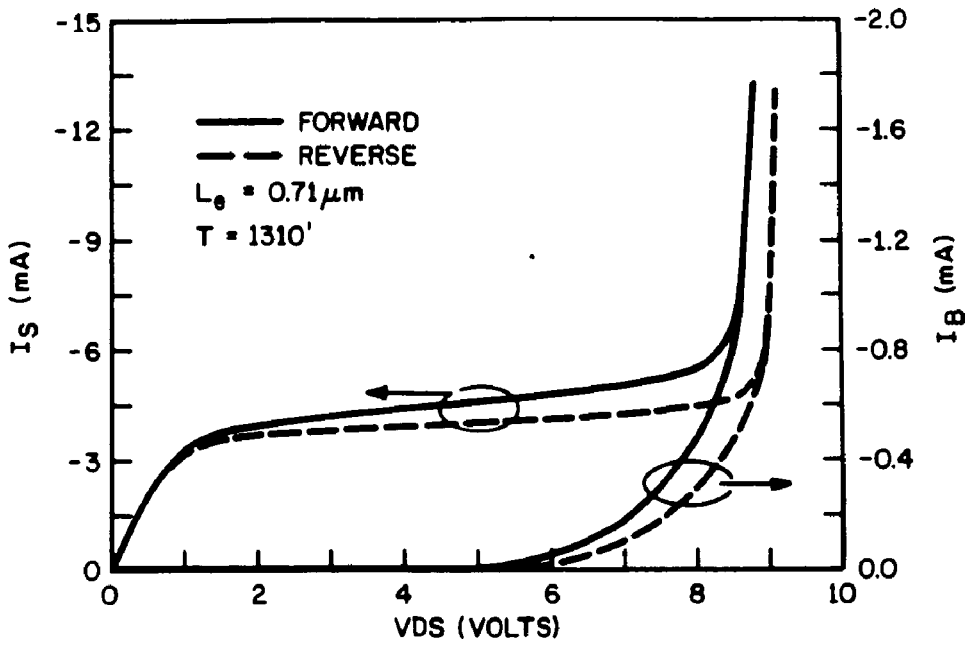


Figure 7-7. Channel and substrate currents after stress in forward and reverse modes indicate an increase in breakdown voltage in reverse mode.

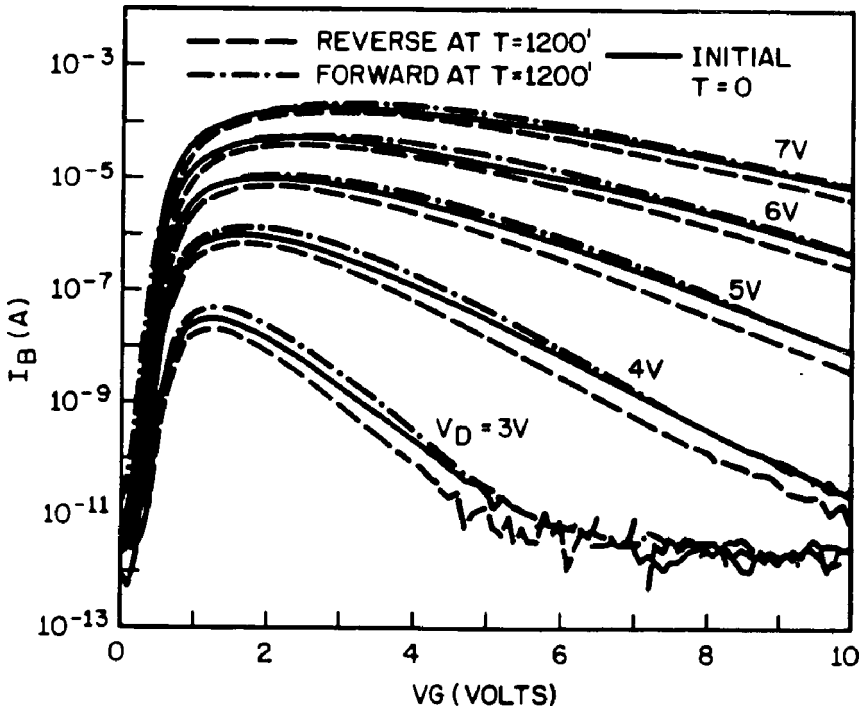


Figure 7-8. Substrate currents before stress (solid curves) and after stress in forward and reverse modes .

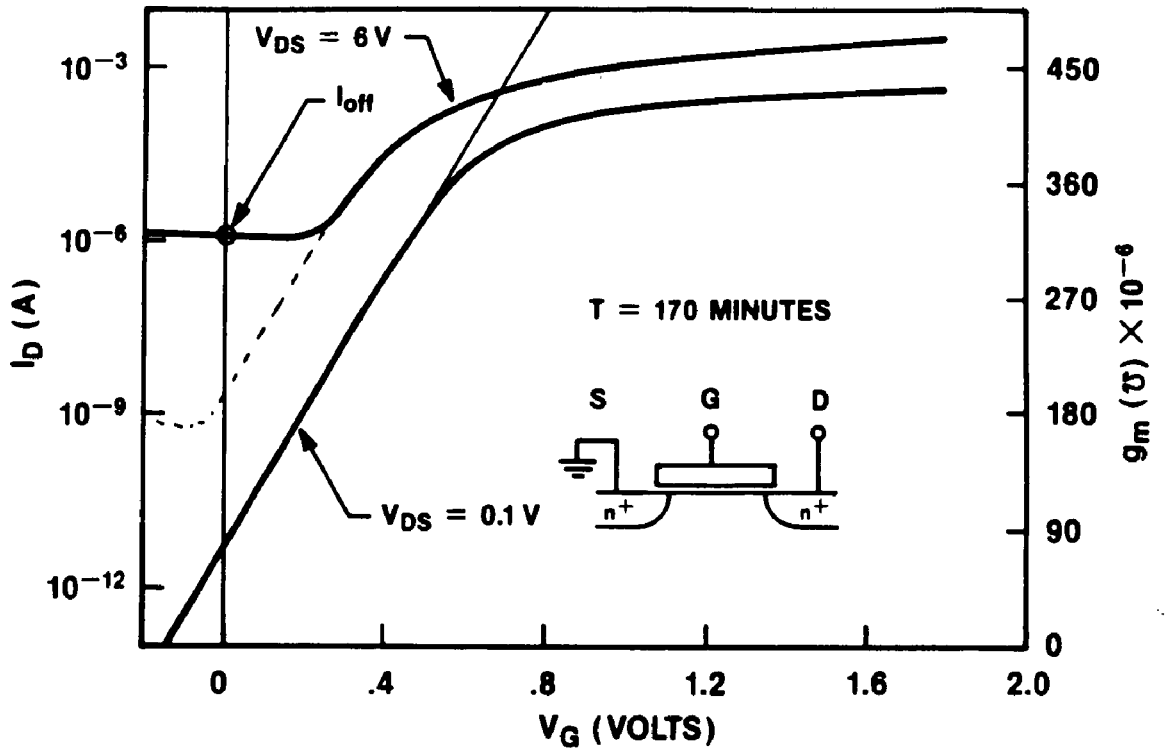


Figure 7-9. Drain junction leakage after 170 minutes stress on a 0.61 μm transistor. Dotted lines show reverse mode channel current measured at the source.

information about the spatial distribution. The charge pumping technique, in principal, allows the average interface state distribution over the channel length and trap energy at midgap to be obtained, by using ac pulsed gate and measuring generated substrate current.[111] In Ref.[110] the authors extracted the spatial distribution by varying the drain voltages. The disadvantage of this method is in the charge pumping set-up, where different systems are used for dc stressing and charge-pumping (ac) measurements. The data extraction and analysis are also complicated.

Meyer and Fair[112] first observed the spatial dependency of interface states in longer channel devices (2.8μm), using subthreshold I-V curves, but did not analyze the spatial distribution.

In this work, the subthreshold I-V technique is used to characterize the device behavior, the degradation mechanisms, and the spatial distribution of the interface states. This method offers the advantage of simplicity in measurement set-up, and ease of data analysis. Justifications to use

the subthreshold conduction method are several-folds:

- As discussed in Chapter 5, the device integrity, particularly at short channel length, can be analyzed using the subthreshold characteristics.
- The subthreshold conduction is important in device applications such as DRAM, SRAM and transfer gates in dynamic circuits, The use of transistor to obtain information related to the way the device is used in circuits.
- The interface trap density, oxide quality, punch-through, drain-induced-barrier-lowering, and source/drain junction leakage can be readily observed and analyzed with the subthreshold current curves.
- In addition, the distribution of interface state density can be extracted using drain (or bulk) bias variation.

We will now present the theoretical background based on the device theory developed in chapter 2, and introduce a methodology to extract interface and fixed charge density.

7.3.1 Subthreshold Current and Interface Trap Density

From chapter 2, the subthreshold current of an MOS device is expressed in terms of the gate and drain voltages as:

$$I_{DS} = \frac{W}{L_e}\mu_n C_D(\phi_s)\frac{n}{m}V_T^2 \exp\left[(V_{GS} - V_{in}^*)/nV_T\right] \cdot \exp(-1.5\phi_F/nV_T)\left[1 - \exp(-mV_{DS}/nV_T)\right] \tag{7.2}$$

At a constant drain bias, the drain current, I_{DS}, is an exponential function of the gate voltage, V_{GS}, as formulated by Eq. (7-2). The values of m and n are evaluated at a surface band bending of $1.5\phi_F$, i.e. mid-point in the weak and strong inversion range, as is referred to as "mid-inversion". The inverse slope of the $\log(I_D)$ vs V_{GS}, or subthreshold swing, is derived by taking partial differentiation of $\log(I_D)$ vs V_{GS},

$$S_n = \ln 10\left[\frac{\partial(\log I_{DS})}{\partial V_{GS}}\right]^{-1} = 2.3 \cdot nV_T \tag{7.3}$$

where the thermal voltage $V_T = kT/q$ is ~25.8mV at room temperature. In the ideal case, i.e. no interface state, subthreshold swing S_1 is ~ 60mV, as can be obtained from the Gummel plot of a bipolar device (Fig. 8-9, Chapter 8). In an MOS device, S_n typically ranges from 80-90 mV/decade. n is a function of surface potential, as defined in Eq. (2.98)

$$n(\phi_s) = 1 + \frac{(1-f)C_D(\phi_s) + C_{it}(\phi_s)}{C_{ox}} \tag{7.4}$$

However, when ϕ_s is in the range of ϕ_F to $1.5\phi_F$, n is found to be almost constant (Fig. 7-12).

The ratio $\frac{m}{n}$ can be determined by differentiating Eq. (7.2) with respect to the drain voltage at fixed gate voltage, i.e. same surface potential. Equation (7.2) can be simplified as

$$I_{DS} = I_{dss}\left[1 - e^{\frac{-mV_{DS}}{nV_T}}\right] \tag{7.5}$$

where

$$I_{dss} = \frac{W}{L_e} C_D \frac{n}{m} V_T{}^2 e^{(V_{GS}-V_m)/nV_T} \tag{7.6}$$

From Eq. (7.5), when V_{DS} is greater than $8V_T$, I_{DS} is approaching I_{dss} within 0.04%. For a given V_{GS}, I_{dss} can be measured (Fig. 7-11). Divide I_{DS} by I_{dss} and change side, Eq. (7.5) becomes:

$$1 - \frac{I_D}{I_{dss}} = e^{\frac{-mV_{DS}}{nV_T}} \tag{7.7}$$

Take inverse and then natural log both sides of Eq. (7.7), we obtain:

$$\ln\left[1 - \frac{I_D}{I_{dss}}\right]^{-1} = \frac{mV_{DS}}{nV_T} = R_{mn} \tag{7.8}$$

where R_{mn} is defined as normalized channel current to aid in calculating the parameter m. From Eq. (7.8), a plot of $\log_{10}\left[1 - \frac{I_{DS}}{I_{dss}}\right]^{-1}$ versus small change in V_{DS} can yield m/n (Fig. 7-11). For convenience, we define the inverse slope, S_m, as:

$$S_m = \frac{2.3n \cdot V_T}{m} \tag{7.9}$$

S_m is measured at V_{DS} at which the drain current is 63.2% of the subthreshold saturation current I_{dss} at V_D=200 mV (~ $8V_T$).

We make the following assumptions for a well designed device:

- In a small V_{DS} range, 10 mV to 200 mV, the short channel effect is insignificant. Therefore, the channel shortening due to increasing V_{DS} in L_e term in Eq. (7.2) can be ignored.
- At V_{DS}=200 mV, i.e. ~ $8 \cdot V_T$ at room temperature, $0.5 \le m/n \le 1$, the term $e^{-(mV_{DS}/nV_T)}$ is much less than 1. Therefore, the drain current is at a fixed V_{GS} approaches I_{dss}.

7.3.2 Determine Parameters n and m

Parameter n is obtained from I_D vs V_{GS} transfer curve at mid-inversion point. First, the threshold voltage at $\phi_s = 2\phi_F$ point is determined using the transconductance change (TC) method.[113] In this method, the derivative of transconductance with respect to gate voltage, $\frac{\partial g_m}{\partial V_{GS}}$, is calculated. The peak of this curve corresponds to the threshold voltage at $2\phi_F$. From this point on the I-V curve, any other point in the range $\phi_F \le \phi_s \le 2\phi_F$, can be evaluated by the drain current expression:

$$I_{DS}(\phi_s) = I_{DS}(2\phi_F)e^{-(2\phi_F - \phi_s)/nV_T} \tag{7.10}$$

At mid-inversion, the drain current is:

$$I_{DS}(1.5\phi_F) = I_{DS}(2\phi_F)e^{-(0.5\phi_F)/nV_T} \tag{7.11}$$

The drain current at mid-inversion is recorded for later reference and the gate voltage at mid-inversion can be determined from the I–V curve. The parameter n is calculated from the subthreshold swing, S_n:

$$n = \frac{S_n}{\ln 10 \cdot V_T} = \frac{S_n}{59.4} \text{ (at room temperature)} \tag{7.12}$$

Fig. 7-10 shows the I_{DS} versus V_{GS} of a 0.5μm device, and its first and second derivatives with respect to V_{GS}. The threshold voltage measured by the transconductance change method was

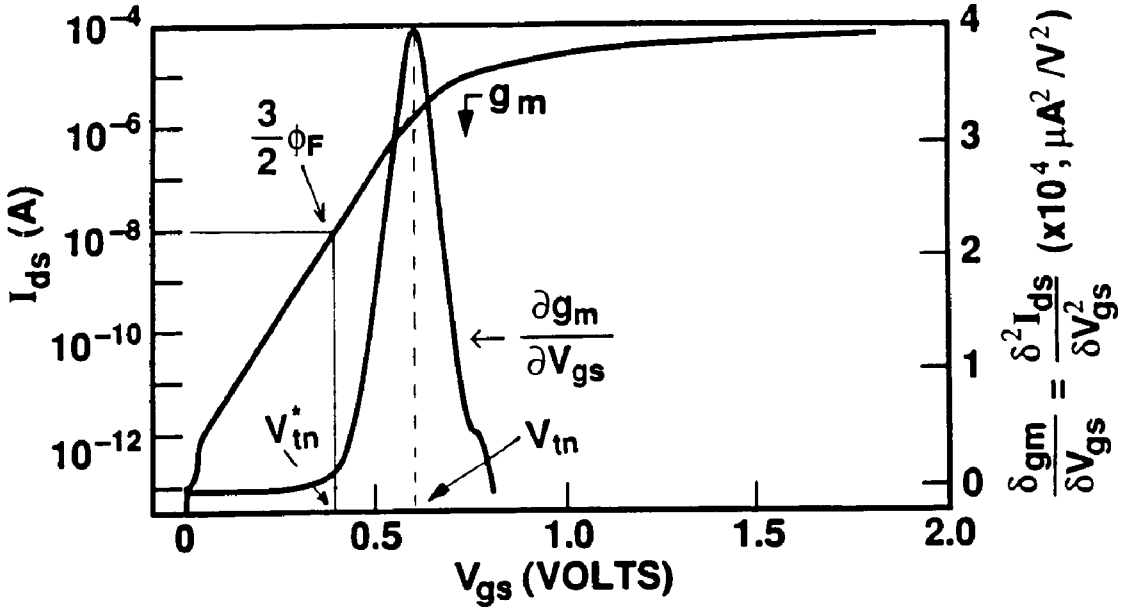


Figure 7-10. I_D vs V_G transfer curve and its first and second derivatives.

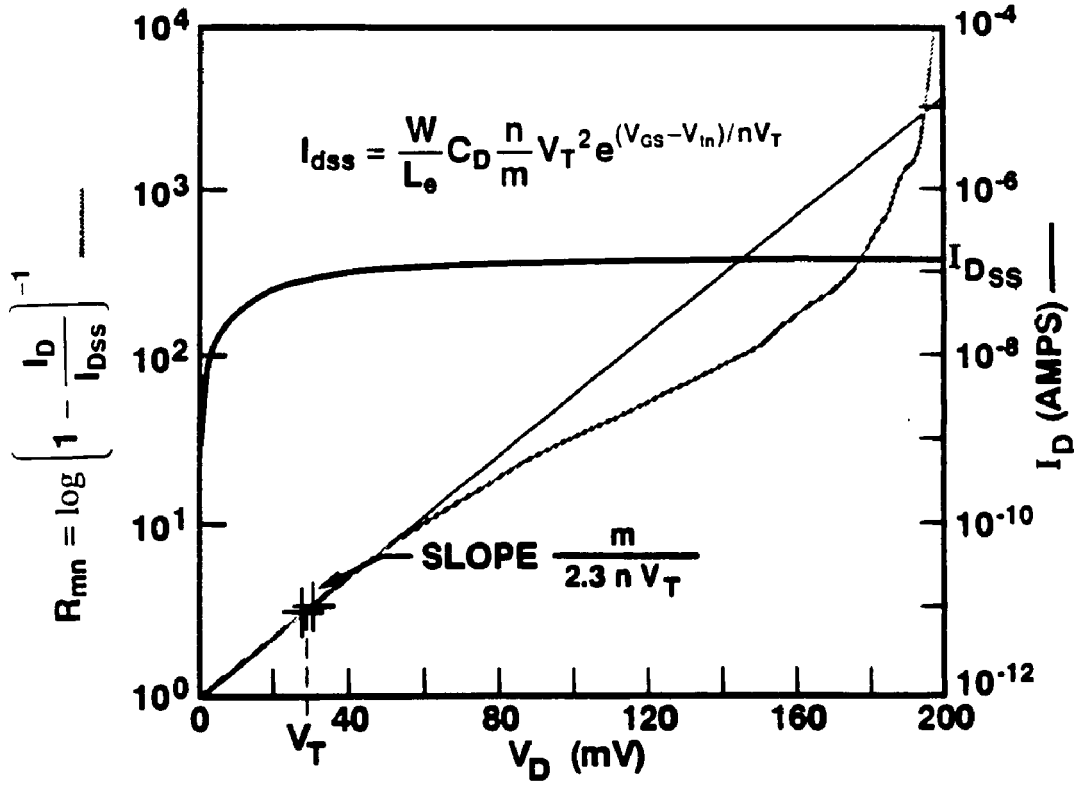


Figure 7-11. Normalized R_{mn} and I_{DS} plots vs V_{DS} for m/n evaluation.

0.605 V corresponding to a current of 2.009μA. The calculated mid-inversion channel current is 10.3nA, using Eq. (7.11). The gate voltage at $1.5\phi_F$ is determined to be 0.384V. The calculation uses a ϕ_F value of 0.40V, from the effective channel doping of N_A of 8E16/cm^3. The subthreshold swing measured around $V_{GS}(1.5\phi_F)$ in this example was 87.0 mV/decade, i.e. n=1.46, and is fairly constant over the weak inversion range.

To extract parameter m, the channel current is measured at a fixed $V_{GS} = V_{GS}(1.5\phi F)$ when the drain voltage is varied to 200mV (~ $8V_T$). This channel current is defined as I_{dss} and is substituted into Eq. (7.8) to plot log(I_{mn}) versus V_{DS}. For a very small change of V_{DS}, the inverse slope of this curve at small V_{DS}, i.e. 25mV, is $S_m = n \cdot V_T / m$. We have extracted n=1.46 from the subthreshold swing (Fig. 7-10), therefore m can be calculated. Figure 7-11 shows the typical plot of I_{mn} vs V_{DS}. m is calculated to be 1.33. Once m and n are determined, using Eq. (2.47) and (2.48), the average interface state density is extracted as:

$$\overline{D_{it}} = \frac{\overline{C_{it}}}{q} = \frac{C_{ox}}{q}(n-m) \tag{7.13}$$

For the unstressed devices, the average interface state density at ϕ_s=1.5 ϕ_F was found to be $1.98\times10^{11} eV^{-1}cm^{-2}$ for the devices with t_{ox}=210Å and $2.24\times 10^{11} eV^{-1}cm^{-2}$ for devices with t_{ox}=125Å. One should be noted that the inverse slope S_m, is very sensitive to the drain voltage variation, as shown in Fig. 7-11 for short channel device, therefore, verification with depletion capacitance extraction from the long channel device is needed for consistency checking.

7.3.3 Spatial Distribution of Interface State Density By Hot Carrier Stress

Follow the same procedure, the interface state density of stressed devices is extracted from measured data at different time intervals during stress. Several authors have used the charge pumping technique[111] with varying drain voltage to profile the lateral distribution of the damage region. As discussed earlier, in this dissertation, we use the subthreshold conduction method to extract the spatially distributed trap density. Shown in Fig. 7-12 is the g curves of an 0.54μm NMOS device, before and after 1000 minutes of stress, with drain bias maintained at 10mV. As seen from Fig. 7-12, S_n increases from 88 mV/dec initially to 114 mV/dec after the stress. The parameter n is plotted using Eq. (7.12), indicating an increase from 1.47 initially to 1.98 after stress.

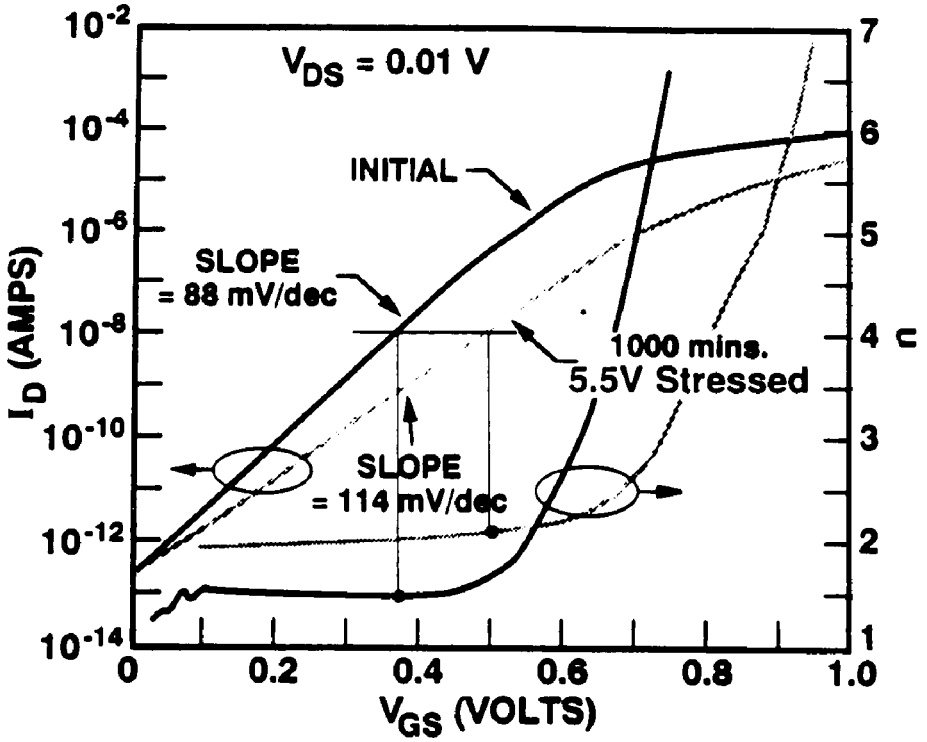


Figure 7-12. Subthreshold *g* curves before and after 1000 mins. stress for a 0.54μm NMOS devices. The parameter *n* is extracted at a constant drain current of 10nA.

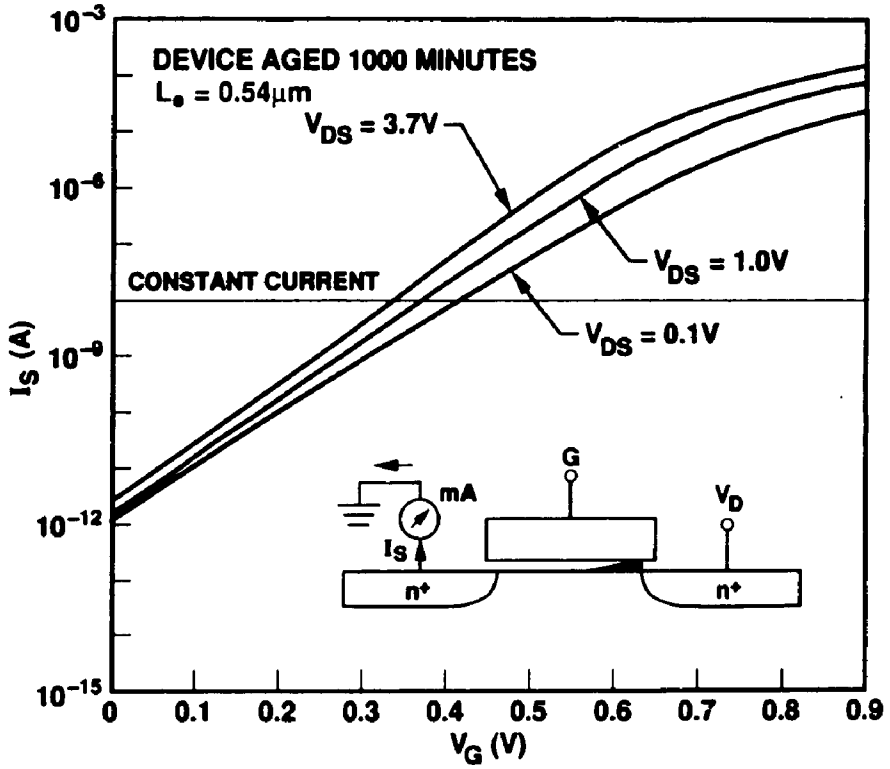


Figure 7-13. The *g* curves of an aged device at different drain bias, *n* is extracted at constant channel (source) current of 10nA.

To obtain the spatial distribution along the channel length, there are two approaches to probe the localized interface trap distribution extended from the drain.

- The drain voltage can be increased so that the depletion width extended from the drain covers the damaged area. This is a more straight forward approach, assuming the device does not have short channel effects, at the channel length under consideration. However, in shorter channel length, this method will increase the channel modulation, and then the subthreshold slope can be misleading due to the short channel effects.
- In a equivalent way, the back gate bias can be used to extend the depletion region beyond the drain and the source, and yet still allowing the drain-to-source potential to be at a small value, i.e. < V_T. To accomplish this, we can either maintain the source voltage, V_S, at ground, and vary the back gate bias negatively with respect to the source for an NMOS, and vice versa for PMOS. Another way to achieve the same effect is to increase both source and drain voltages by the same offset, and keeping the bulk bias at ground. We choose the grounded source approach, since all the terminal voltages can be easily referenced to the source.

One restriction of this method is the source and drain junction leakage must be low enough so that the leakage current does not interfere with the channel current during measurements.

7.3.3.1 Varying Back Gate Voltage

Table 7-1 shows the subthreshold swing, S_n, for an aged 0.5μm NMOS devices, with back gate bias at -0.015, -0.09, -0.39, -0.89, -1.59, -2.59, and -3.59. The drain voltage was biased at 0.010 V for all cases. We observed the slope S_n changes according the following table:

TABLE 7-1. Changes in Subthreshold Swing vs. Back Gate Bias.

V_{BS}(V)	0	-0.015	-0.09	-0.39	-0.89	-1.59	-2.49	-3.59
S_n (mV/dec)	119	116	114	107	95.7	91.9	86.4	84.2

The above results were extracted using the source currents (= channel current + small source-bulk leakage current). The analysis of the interface state density is very complicated since the depletion width extended from the source and drain along the channel length and the channel

charge-sharing effect must be taken into account.

7.3.3.2 Varying Drain-to-Bulk Bias

For a well-designed device, we prefer the use the drain voltage variation method to extract the interface state density. From the channel concentration $N_A = 8\times10^{16} cm^{-3}$, and the depletion width approximation, we can calculate the depletion width extended from the drain intrinsic depletion edge, due to the drain to channel bias.

$$\Delta L_d = \sqrt{\frac{2\varepsilon_s}{qN_A}}\left[\sqrt{\phi_{bi} + V_{DB}} - \sqrt{\phi_{bi}}\right] \tag{7.14}$$

where

$$\phi_{bi} = \frac{kT}{q}\ln(\frac{N_A N_D}{n_i^2}) \tag{7.15}$$

is the built-in potential between the drain junction and the channel surface. For the NLDD concentration of 2×10^{18}, and the channel surface concentration of 8×10^{16}, the built-in potential is calculated to be 0.87 eV. The depletion distance from the drain for different drain to substrate bias is then calculated. The numerical calculation of the surface electron density of the same device structure with gate length of 0.5μm, with the drain bias varied from 10mV to 3.6V is shown in Fig. 7-14.

The g curves of the aged device are measured at different drain biases. Figure 7-13 shows the subthreshold swing is indeed decreasing with increasing drain bias. The parameter n is extracted at a constant current. The average interface state density $\overline{D_{it}}$ and the distance from the drain are calculated using Eqs. (7.13) and (7.14). The results are tabulated in Table 7-2.

TABLE 7-2. Interface State Density as a function of distance ΔL_d from the drain junction.

V_D	0	-0.025	-0.10	-0.40	-0.90	-1.60	-2.50	-3.60
ΔL_d (nm)	0	1.95	7.66	28.54	58.46	93.97	132.86	173.87
$\overline{D_{it}}$ ($x10^{11}$)	12.3	11.42	10.84	8.81	5.54	4.43	2.8	2.20

The initial average interface state density of the device was found to be $2.24\times10^{11} eV^{-1}cm^{-2}$.

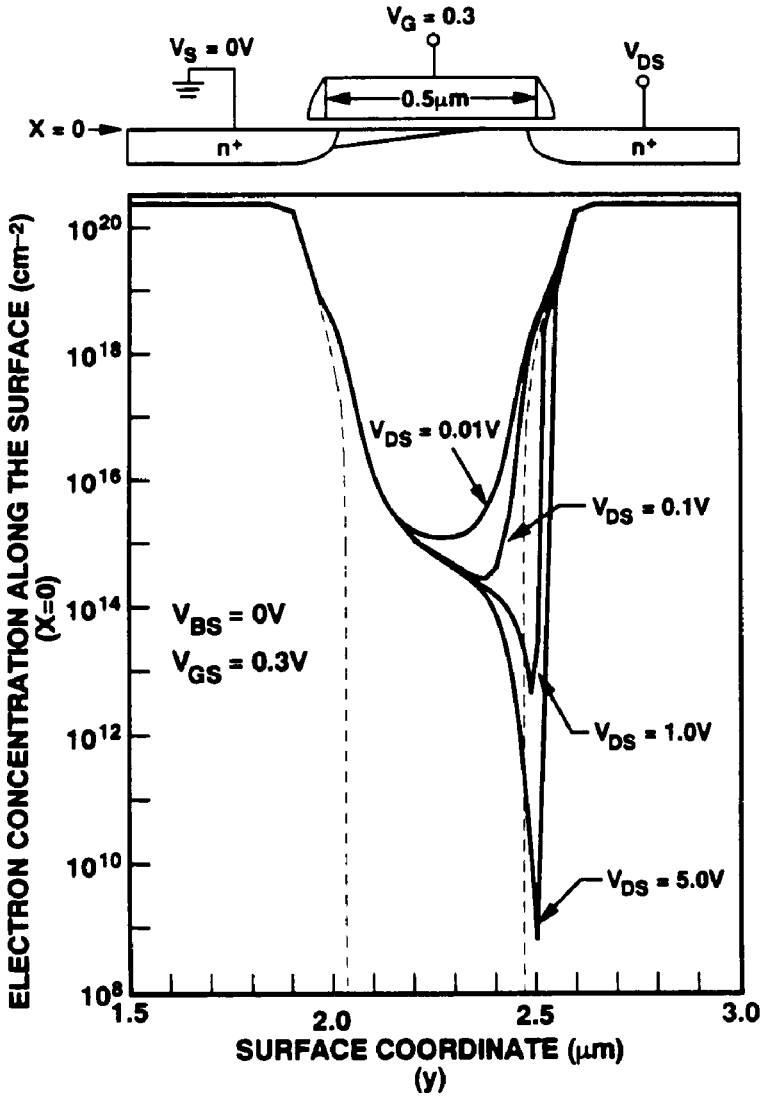


Figure 7-14. Surface electron density showing a depletion of electrons at the drain when the drain bias is increased.

7.3.4 Formulation of the Lateral Interface Trap Density

Assuming the interface trap is distributed along the channel length, i.e. in the y direction. And the measured $\overline{D_{it}}$ is the sum of the interface states in the undamaged region L_0, and that of the damage region ΔL_d. The measured interface trap density is an integrated of the density spatially distributed along the channel, from the 2 depletion edges of the source and drain.

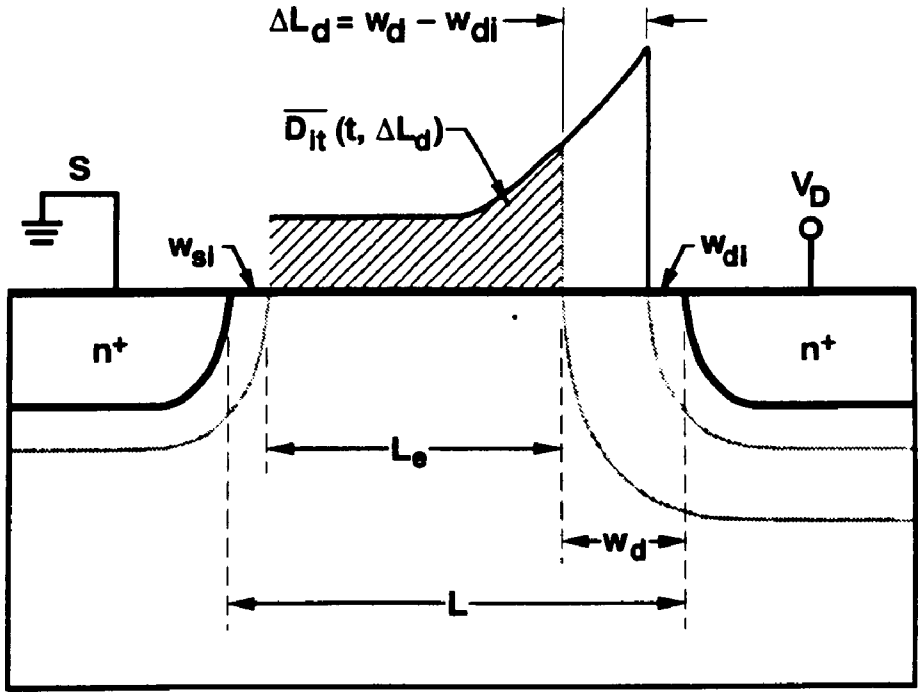


Figure 7-15. Schematic illustrates the distribution of interface states along the channel length.

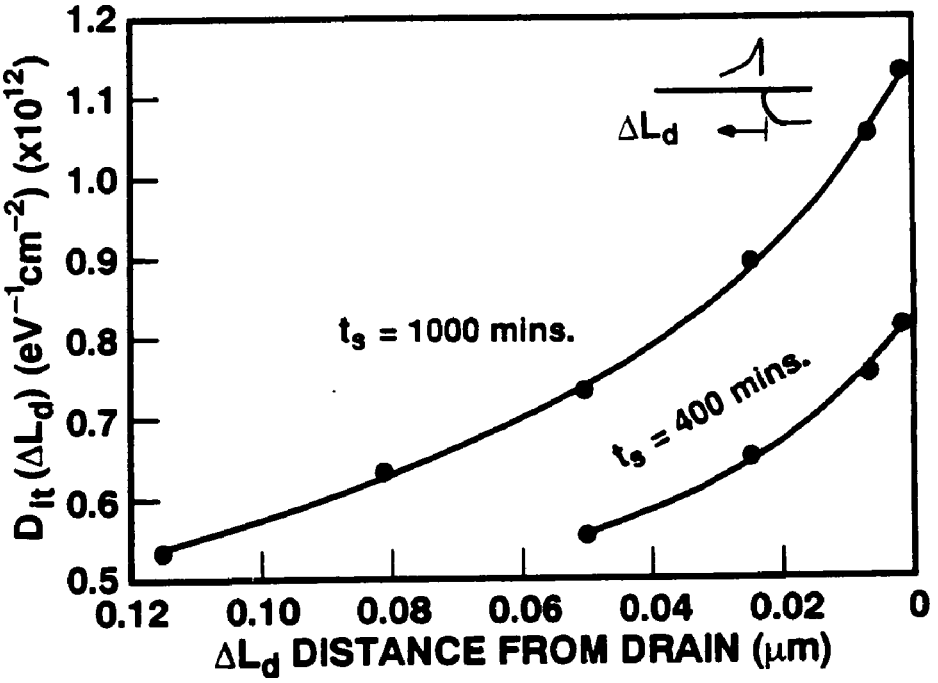


Figure 7-16. Experimental spatial distribution of interface state density at the drain of an NMOSFET with L_{eff}= 0.54 μm aged at 5.5V, for 400 and 1000 mins.

$$\overline{D_{it}} = \frac{1}{L_e}\int_{w_s}^{L-w_d} D_{it}(y)dy \tag{7.16}$$

At stress time t, $\overline{D_{it}}$ is a function of both time and location, i.e. $D_{it}(t,y)$:

$$\overline{D_{it}}(t,y) = \frac{1}{L-w_s-w_d}\int_{w_s}^{L-w_d} D_{it}(t,y)dy \tag{7.17}$$

where $\Delta L_d = w_d - w_{di}$, is the extension of drain depletion edge beyond the intrinsic (V_D=0) built-in junction depletion, w_{di}, which is a constant. Therefore,

$$\frac{\partial \overline{D_{it}}}{\partial \Delta L_d} = \frac{\partial \overline{D_{it}}}{\partial w_d} \tag{7.18}$$

Differentiate Eq. (7.17) with respect to ΔL_d, we obtain

$$\frac{\partial \overline{D_{it}}}{\partial \Delta L_d} = \frac{1}{(L-w_s-w_d)^2}\int_{w_s}^{L-w_d} \overline{D_{it}}(t,y)dy - \frac{D_{it}(t,\Delta L_d)}{L_e} \tag{7.19}$$

In our spatial interface extraction method, the drain (or bulk) is biased to extend the depletion layer into the channel. Shown in Fig. 7-14 is the surface electron concentration as a function of the drain bias, obtained from the device simulation results. Any damage to the surface has occurred in the region above the drain is shielded by the drain electric field, and therefore the traps lie in this distance do not respond to the the change in surface potential sweeping during the subthreshold measurement. At zero drain bias, the depletion with from the drain is extended by a built-in space charge layer from the drain to channel junction. The extension from drain ΔL_d into the channel can be calculated using the depletion width approximation (Eq. 7.14),

$$\Delta L_d = w_d - w_{di} \tag{7.20}$$

For a given drain bias, the measured average $\overline{D_{it}}$ is the integrated density of $D_{it}(t,y)$ from w_s to $L-w_d$, i.e. to the edge of ΔL_d if measured from drain (refer to Fig. 7-15). Solving Eq. (7.19) for $D_{it}(t,\Delta L_d)$, and neglecting the effect of w_s, we obtain,

$$D_{it}(t,\Delta L_d) = \overline{D_{it}(t,\Delta L_d)} - (L-\Delta L_d)\frac{\partial \overline{D_{it}}}{\partial \Delta L_d} \tag{7.21}$$

The interface state density vs the distance from the drain, $D_{it}(t,\Delta L_d)$, can be evaluated from the

measured data, is shown plotted in Fig. 7-16 for the 2 time intervals during stress. The spatial distribution of the interface state density extended from the zero-bias drain depletion edge, is plotted in Fig. 7-17 for 2 positions of ΔL_d's. This is a very interesting result indicating that the interface state density are built-up very slowly initially (toward the channel), but after 100 minutes, it increases with time by a power law relationship.

$$D_{it}(t, \Delta L_d) = K \cdot (t)^{n'} \tag{7.22}$$

where n' is calculated to be 0.387 for the channel position ΔL_d= 17 Å away from the intrinsic drain depletion edge. This result is consistent with the D_{it} extracted using charge pumping method.[114] For the corresponding g_m degradation of this device is found plotted in Fig. 7-22, where the lifetime of the device is ~ 43 minutes. During this time the change in the channel interface states is insignificant, as seen from the top curve in Fig. 7-17. The charges injection during this period take place directly above the drain, which causes the drain series resistance to increase, and the g_m to decrease. This built-up of interface charges outside of the channel region can not be measured by the changes in the slopes of the g curves. The built-up of interface trap density is plotted as a function of distance from the drain at different stress time and is shown in Fig. 7-18. We have shown that the interface state density built-up by hot carrier injection is more severe than the transconductance degradation.

7.4 THE VALIDITY OF TWO-TRANSISTOR EQUIVALENT MODEL FOR A DAMAGED DEVICE

For the hot carrier degraded MOS device, the damage region is localized at the drain end and extending toward the source as the aging time prolongs. In order to model the degraded device, 2-series-transistor transistor models have been proposed by several authors.[115] [116] However, these models assume the fixed charge and constant distribution of interface state, i.e. the transition between the undamage and the damaged regions is assumed to be abrupt.

A 2-transistor model that includes the laterally distributed interface states as the function of channel length is shown schematically in Fig. 7-19. The channel length portion of the undegraded device is labeled L_o, and the damaged length at the drain side is denoted as ΔL_d. Therefore

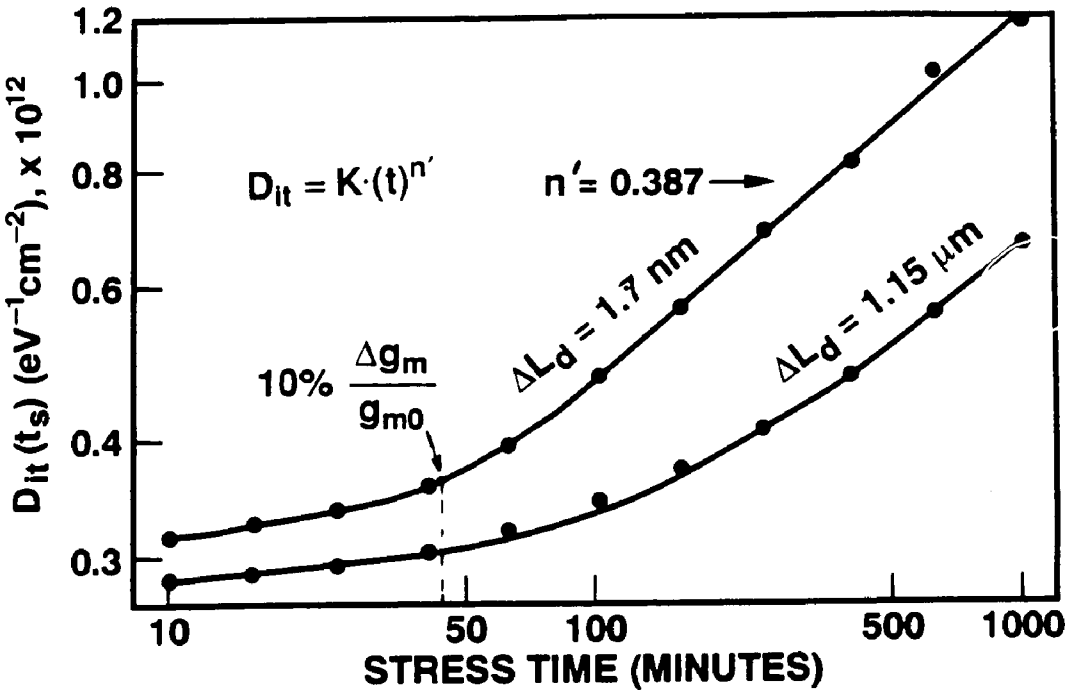


Figure 7-17. Interface trap built-up with time at the 2 edges of ΔL_d.

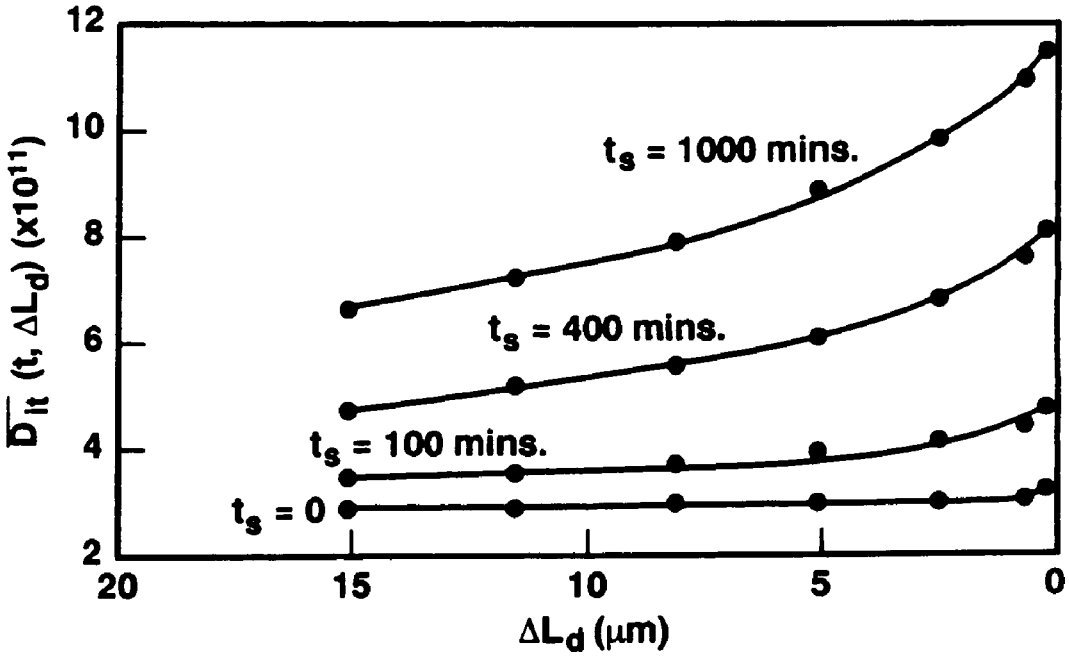


Figure 7-18. $\overline{D_{it}}$ built-up along the drain at various stress times.

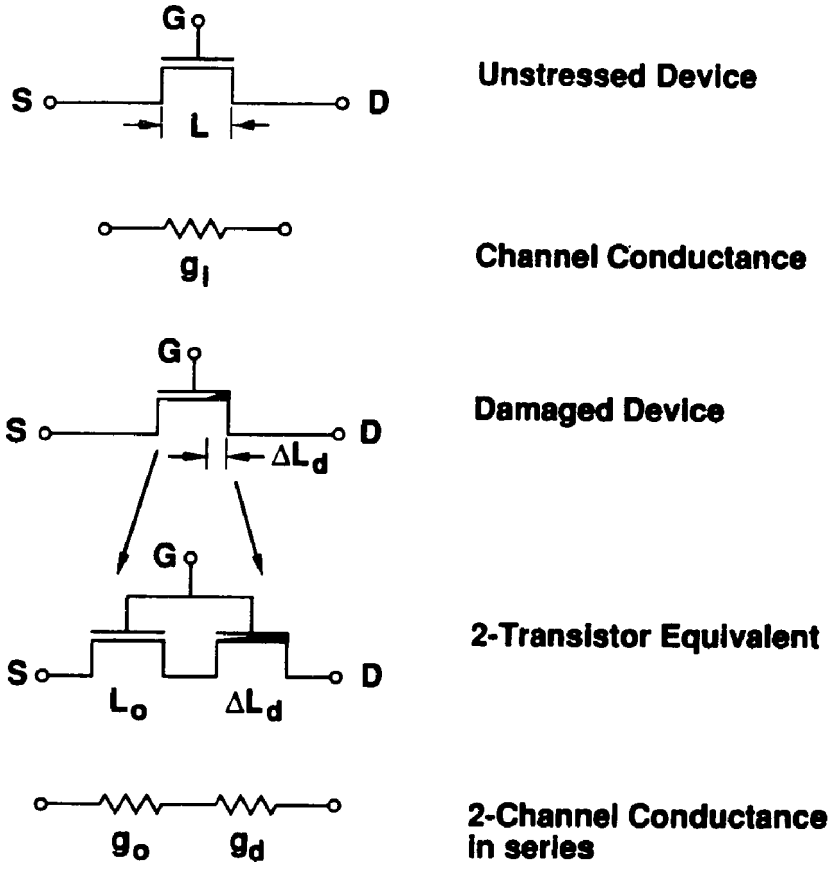


Figure 7-19. Two-Transistor Model with non-uniformed defect region.

$$L = L_o + \Delta L_d \tag{7.23}$$

In previous section, we have determined that the interface state is locally distributed at the drain end of the device. The damaged region ΔL_d can be extracted. The distribution of D_{it} vs ΔL_d is approximated by a $1/\Delta L_d$ relationship. It is difficult to determine the break point between L_0 and ΔL_d for a short channel device, as the distribution of interface states will extend toward the source end, i.e. the entire channel is damaged as shown in Fig. 7-18. Therefore, the 2-transistor equivalent circuit is valid only when the device is moderately aged. We have examined this situation by interchanging the source and drain terminal of an aged device. The subthreshold

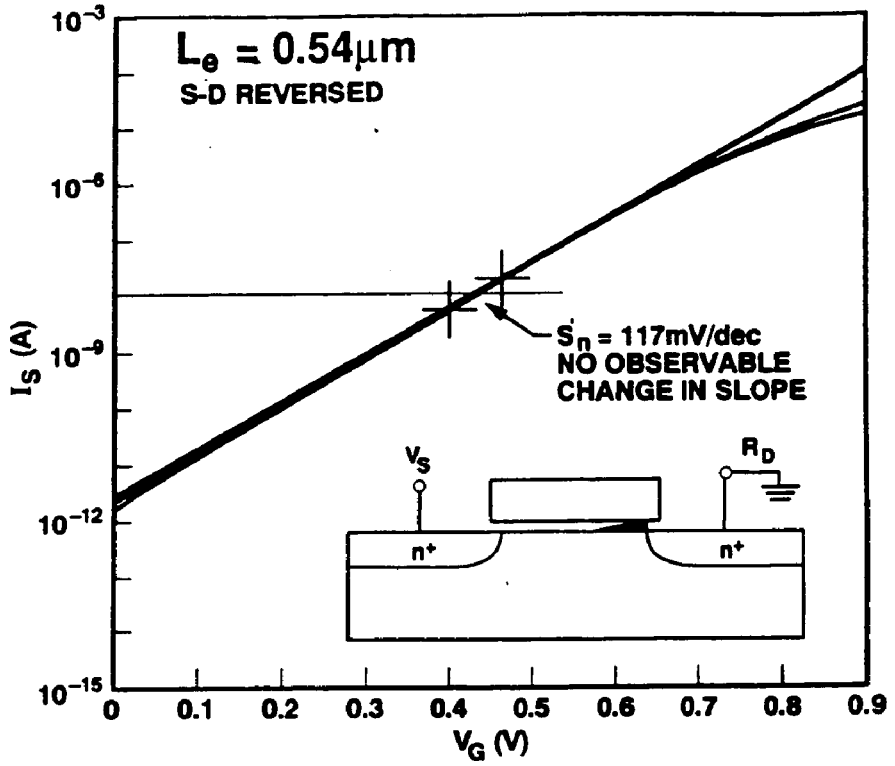


Figure 7-20. Subthreshold g curves for an aged device measured in the reverse-mode, the sub V_t swings do not change as the reversed source voltage varies.

curves measured in this reverse mode show no changes in subthreshold slopes when the source (formerly drain during stress) voltage varies. The reverse mode g curves are shown in Fig. 7-20.

7.4.1 Transconductance Degradation as a Function of Channel Length

We have stressed devices with different short channel lengths ranging from 0.73μm to 1.22 μm in L_{eff}. The devices were aged at the same time with V_D = 8V and V_G =3.0V. As shown in Fig. 7-21, the g_m degradation for the longer device (1.22μm) reach 10% mark slower than the shorter devices, as expected. Also, the $\Delta g_m/g_{m0}$ saturates with time for the 2 shorter channel devices at different levels (18% and 12%). In order to understand this phenomena, we will derive the maximum transconductance of the damaged device including the interface state density built-up during stressing. From the linear current Eq. (2.113) in chapter 2, we rewrite a simplified drain current as:

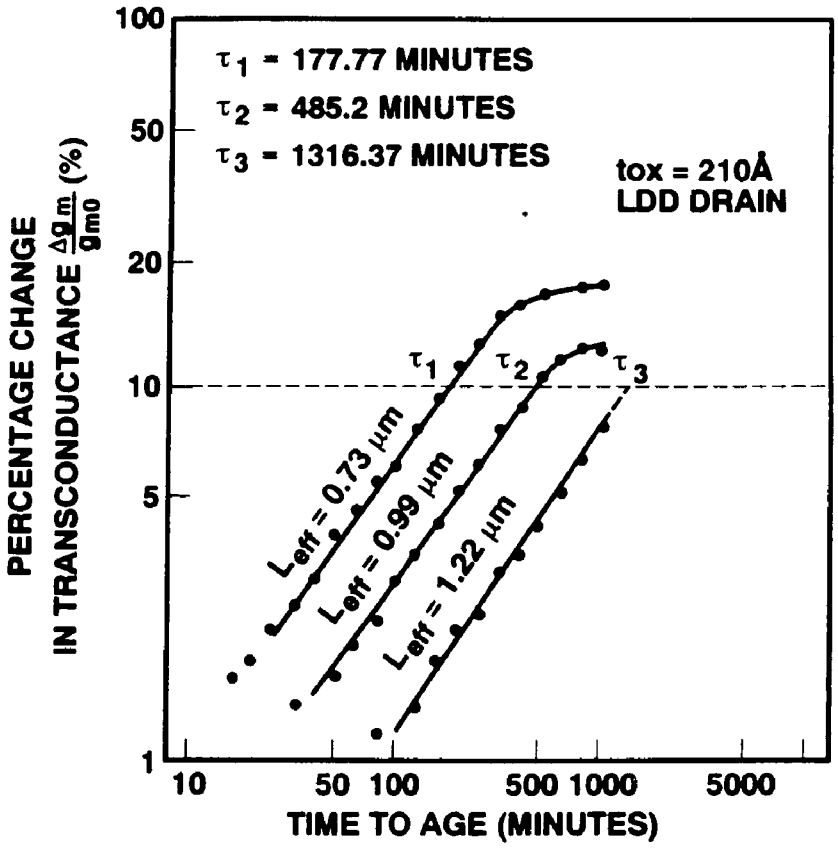


Figure 7-21. g_m degradation for 3 LDD NMOS devices with different channel lengths, t_{ox}=210Å.

$$I_D = \frac{\beta_0[V_{GS} - V_{in} - \frac{V_{DS}}{2}]V_{DS}}{1 + (\theta_s + 2\beta_0 R_s)[V_{GS} - V_{in} + 2\lambda\sqrt{|V_{SB}| + \phi_s^*} - V_{DS}/2]} \tag{7.24}$$

in which the effect of series resistance R_s is included, and

$$\theta_s = \theta_s(0) + 2\beta_0 R_s \tag{7.25}$$

and

$$\theta_s(0) = \frac{C_{ox}}{2\varepsilon_s E_{cr\perp}} \tag{7.26}$$

Take the derivative of I_{DS} in Eq. (7.24) with respect to V_{GS}

$$g_m = \frac{\partial I_D}{\partial V_G} = \frac{\beta_0 V_D[1 + 2\lambda\theta_s\sqrt{V_{SB} + 2\phi_F}\,]}{1 + (\theta_s + 2\beta_o R_s)[V_{GS} - V_{tn} + 2\lambda\sqrt{|\,V_{SB}\,| + \phi_s^*} - V_{DS}/2]} \tag{7.27}$$

β_0 is expressed in terms of low field mobility and device geometry as

$$\beta_0 = \mu_0(\frac{W}{L})C_{ox} \tag{7.28}$$

and

$$\mu_0 = \frac{\mu_{I,L}}{1 + \alpha_{it}\overline{D_{it}}(\frac{\Delta L_d}{L})} \tag{7.29}$$

where $\mu_{I,L}$ is intrinsic low field mobility, and α_{it} can be theoretically calculated or empirically measured.[117] The maximum transconductance $g_{m,max}$ is derived from Eq. (7.27) as:

$$g_{m,max} = \frac{\mu_{I,L}(\frac{W}{L})C_{ox}V_D}{(1 + \alpha_{it}\overline{D_{it}}(\frac{\Delta L_d}{L}))\left[1 + 2\lambda\sqrt{2\phi_F + V_{SB}}\left[\theta_s(0) + \frac{2R_s\beta_{00}}{1 + \alpha_{it}\overline{D_{it}}(\frac{\Delta L_d}{L})}\right]\right]} \tag{7.30}$$

Let's define:

$$A \triangleq 1 + 2\lambda\sqrt{2\phi_F + V_{SB}}\left[\theta_s(0) + \frac{2R_s\beta_{00}}{1 + \alpha_{it}\overline{D_{it}}(\frac{\Delta L_d}{L})}\right] \tag{7.31}$$

and

$$B \triangleq 2\lambda\alpha_{it}\theta_s(0)\sqrt{2\phi_F + V_{SB}} \tag{7.32}$$

Then the maximum transconductance, $g_{m,max}$, can be expressed in terms of A, B and the ideal transconductance as:

$$g_{m,max} = \frac{g_{m,max}^{ideal}}{A + BD_{it}(\frac{\Delta L_d}{L})} \tag{7.33}$$

When the device is aged, the interface state density is built up to an average density $D_{it,f}$, corresponding to the transconductance g_{mf}. The percentage of change in the transconductance is then expressed as,

$$\frac{\Delta g_m}{g_{m0}} = \frac{g_{mf} - g_{m0}}{g_{m0}} \tag{7.34}$$

Using Eqs. (7.30)-(7.32), we can simplify the transconductance degradation equation as,

$$\frac{\Delta g_m}{g_{m0}} = \frac{-B[D_{itf} - D_{it0}](\frac{\Delta L_d}{L})}{A + BD_{itf}(\frac{\Delta L_d}{L})} \tag{7.35}$$

From Eq. (7.35) we can explain the $\Delta g_m / g_{m0}$ saturation for different channel devices as shown in Fig. 7-21. This result indicates that using the 10% g_m degradation as the criteria for device lifetime can be misleading, since for the longer than nominal device, the change in $g_{m,max}$ may never reach the 10% point. The lifetime extrapolation from 3 devices with different lengths also results in erroneous prediction as will be discussed in the next section.

7.5 THE LIFETIME ANALYSIS

In this section Figure 7-22 shows the transconductance degradation vs time for 3 devices with the same coded channel length, 0.5μm, using the test structure shown in Fig. 7-1. The devices are fabricated with spacer DDD drain as discussed in chapters 3 and 4, and intended for 3.3V supply voltage. The devices are aged at V_D of 5.5V, 5V, and 4.5V and $V_G = 0.4V_D$. The transconductance degradation curves crossing the 10% line is defined as the lifetime of the device for that particular drain voltage and the corresponding substrate current during aging. The lifetimes for 3 different stress bias are then plotted versus the substrate current during aging as shown in Fig. 7-23. τ is found to be strongly dependent on the substrate current at stress for different drain biases and channel lengths. As seen from Fig. 7-23, τ is an inverse power law function of I_B and can be expressed as:

$$\tau = C \cdot (I_B)^S \tag{7.36}$$

where s is calculated in this example to be -2.79, which is in the typical range of -2.5 to -3.0. The lifetime expectancy at operating voltage (3.3V) is extrapolated to the maximum substrate currents measured at 3.3V and 3.6V (0.15 and 0.36μA/μm respectively) before stressing. The lifetimes at 3.3V and 3.6V are extrapolated to be 45.86 years and 3.83 years, respectively. These results indicates that the device lifetime is very sensitive to operating voltage, and therefore substrate

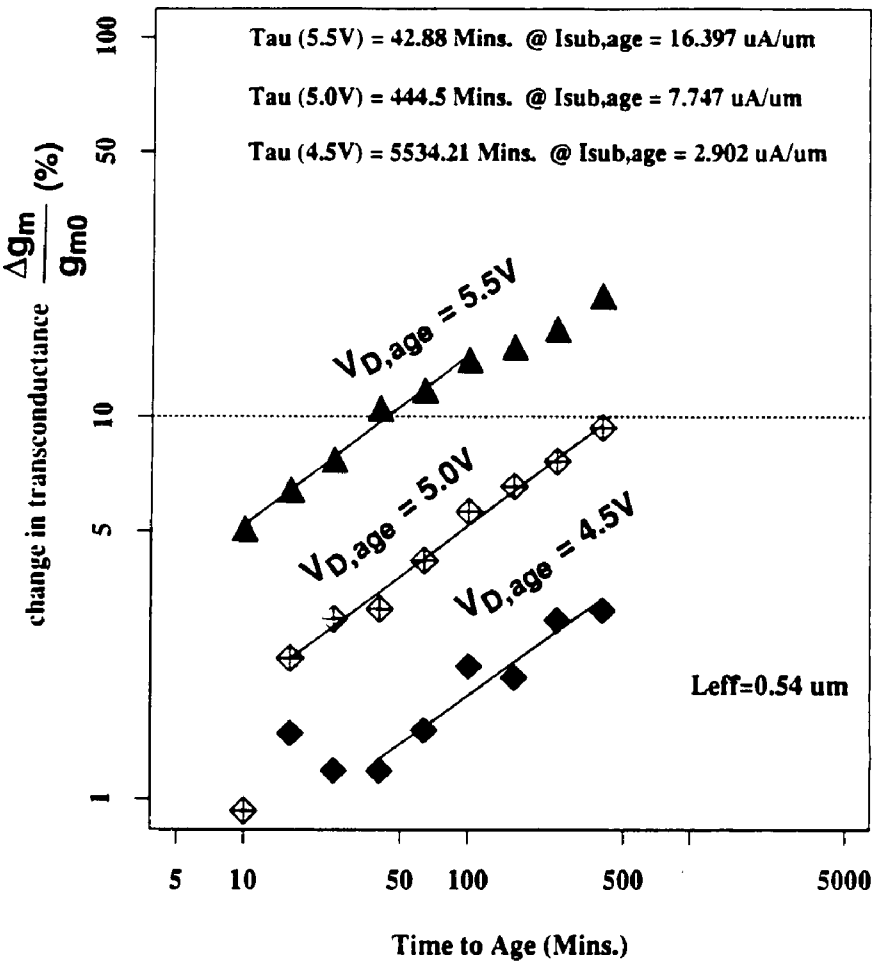


Figure 7-22. Typical g_m degradation curves for 3 devices with the same effective lengths of L_e=0.54μm.

current.

In the lifetime extrapolation using substrate current as the reference point, one should be careful about the channel length dependent of the aging. As we discussed in the previous section, the channel dependent of hot carrier aging can lead to errors in lifetime extrapolation, when a long channel device is used in the data set. If we use the data shown in Fig. 7-21 for 3 different channel length devices, designed for 5V operation, and extrapolate to the substrate current measured at 5.5V, the lifetime is predicted to be 81 years at 5.5V. The same set of 3 devices with L_{eff}=0.99μm (same coded lengths of 1.25μm) on the same wafer aged at different voltages, 7, 7.5 and 8V. The extrapolation is shown on the same plot in Fig. 7-24, the extrapolation predicts only

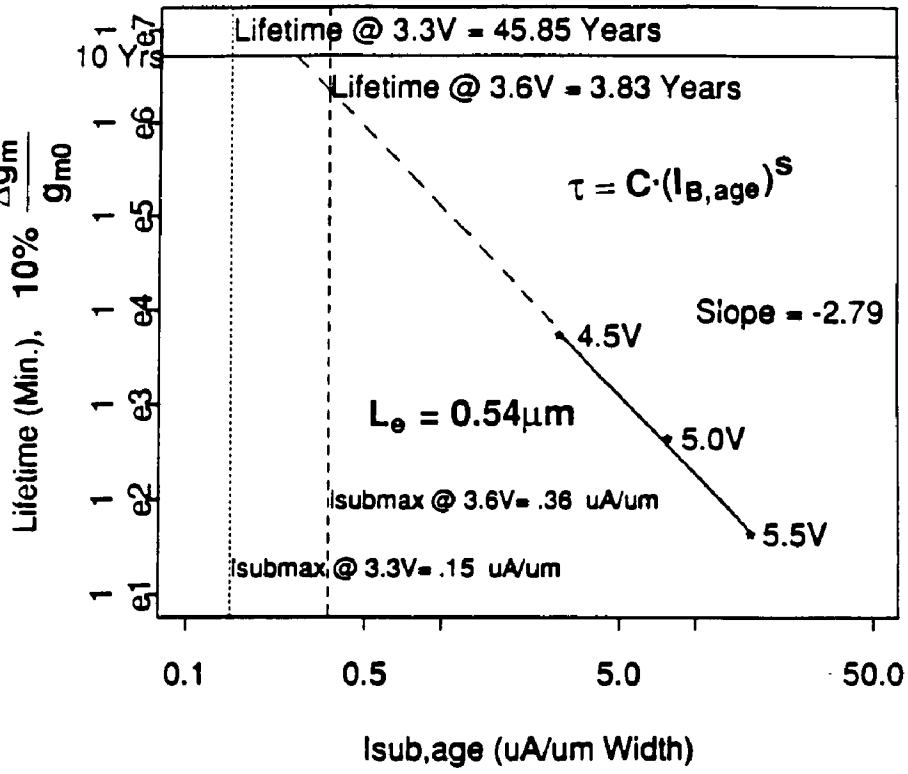


Figure 7-23. Lifetime vs substrate current during aging for the devices shown in Fig7-22.

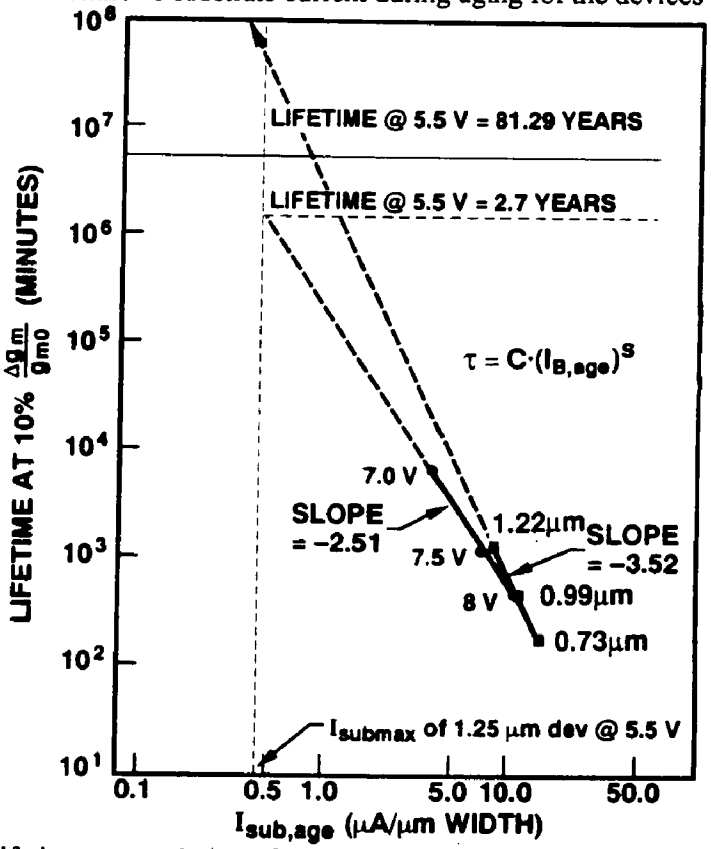


Figure 7-24. Lifetime extrapolation of devices aged with different channel lengths and voltages.

2.7 years at 5.5V. This discrepancy can be explained in light of the analysis in the previous section. For longer channel length device the g_m degradation reaches to the 10% with different acceleration factor. Therefore, the acceleration factor l is calculated to be 3.52 as compared with the value of 2.51, if the voltage variation is used. From this result, we should use the same channel length and varying the drain/gate voltage to achieve the lifetime vs substrate current acceleration factor. The channel field vs drain voltage relationship is consistent for the same L_e, i.e. the same V_{Dsat} (Eq. 6.3). Based on this result, the multiple devices of the same channel length, placed in the vicinity of each other (Fig. 7-1) is very test structure for device aging experiments.

Another method to extrapolate the device lifetime to the operating voltage is to use Eq. (6.10), i.e. $I_{B,max}$ is exponentially proportional to $1/\sqrt{V_{DS}}$, and is rewritten here for convenience.

$$I_{B,max} = A \cdot \exp(-\beta/\sqrt{V_{DS}}) \tag{7.37}$$

Using Eq. (7.37) and substitute into (7.36) to obtain the lifetime as a function of $\sqrt{V_{DS}}$ as:

$$\tau = K \cdot \exp(-s\beta/\sqrt{V_{DS}}) \tag{7.38}$$

Therefore, if equation (7.38) is valid for all V_{DS}'s then, one ought to plot τ in log scale vs $1/\sqrt{V_{DS}}$ as shown in Fig. 7-24, to maintain the consistency with equation (7.38). This gives a more conservative estimate of device lifetime as compared with the $1/V_{DS}$ extrapolation reported by Takeda and Suzuki[101] and other. The two extrapolation methods, namely $I_{sub,age}$ and $1/\sqrt{V_{DS}}$, are equivalent, as dictated by Eq. (7.37), for devices of the same channel lengths. As an example, using the data from wafer 8408-44 in Fig. 7-25, the lifetime extrapolated from $1/V_{DS}$ method would be 95 years, whereas using $1/\sqrt{V_{DS}}$ method the predicted lifetime is only 55 years at 5 V operating voltage.

In order to compare the effectiveness of the LDD devices in suppressing hot carrier aging, we plot the device lifetime for the 2 technologies using the same aging conditions, V_D=7V and V_G=3V (Fig. 7-26). It is apparent that at the minimum allowable effective channel length for each respective technology, i.e. 0.75μm vs 1.0μm, the LDD device has an order of magnitude longer lifetime that the DDD device used in an earlier generation. For the NMOS devices with

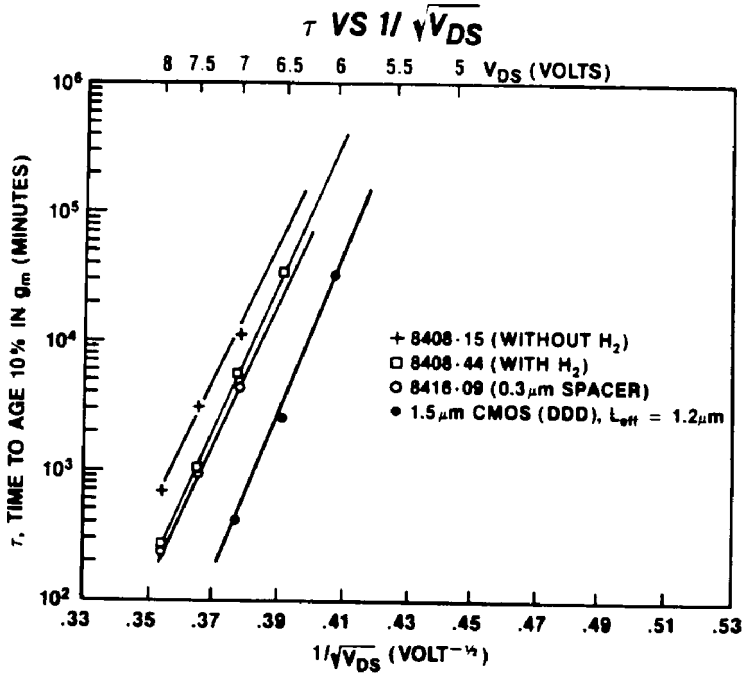


Figure 7-25. A more accurate plot of lifetime vs $1/\sqrt{V_{DS}}$, to extrapolate to 5 V and below.

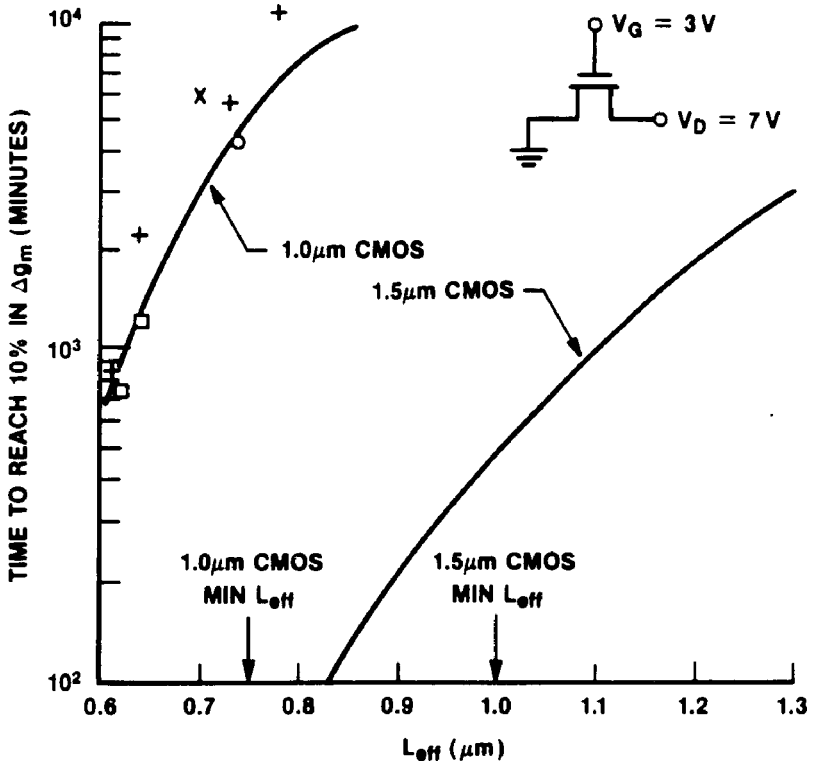


Figure 7-26. Lifetime vs L_{eff} for the 1.0μm CMOS devices, using NLDD with t_{ox}=210Å and 1.5μm NMOS with DDD drain and t_{ox}=250A.[50]

thinner gate oxide (150Å) using similar LDD structure,[43] the device lifetime is also one order of magnitude shorter than our 1.0μm NMOS 's lifetime.

In addition to the use of substrate current to extrapolate the lifetime, we also obtain the $I_{B,max}$ distribution as a function of the channel lengths. A plot of $I_{B,mas}$ vs $1/L_{eff}$ is constructed from data obtained from 'good' wafers (Fig. 7-27). Any deviation from the spacer etch that will result in high substrate current is detected. The method is a very useful tool to electrically detect the consistency of an etching process.

7.6 SUBSTRATE CURRENT IN A CMOS INVERTER AND CIRCUIT AGING

In order to investigate the substrate current in a CMOS inverter during a transient switching, and inverter with gate length of 0.5μm and W_n = 10μm and W_p = 20μm. The input-output transfer characteristics were measured with the channel current, I_{SS}, and substrate current generated during the switch. As seen from Fig. 7-28, the n-ch substrate current occurs during the output node is being pulled down of the n-ch device, and the p-ch bulk current (generated in n-tub), are generated when the out voltage is low. We have made a distinction between substrate for n-channel and bulk current for p-channel, since the bulk current generated by the p-channel is collected at the n-tub contact, but not at the substrate. The peak p-ch bulk current is about 2 orders of magnitude lower than that of the n-channel. Two distinct loops of bulk currents are observed for the p-ch device. We will concentrate on the substrate current of the NMOS device. The average substrate current, $\overline{I_{B,n}}$, during switching is integrated over the transition period, V_{in}, as is shown in Fig. 7-29. Note that the n-ch substrate current increases during the n-ch device turns on, but disappears rapidly as the $V_{DS,n}$ decreases during switching. I_{SS}, measured at V_{SS} node is the channel current when both devices are conducting.

$$\overline{I_{B,n}} = \frac{1}{V_{DD}} \int_{V_{in}=0}^{V_{in}=V_{DD}} I_B(V_{in}) dV_{in} \tag{7.39}$$

With the average substrate current at the operating condition, we can use the results presented in the previous section to predict lifetime of circuit. From the data shown in Fig. 7-29, the average substrate current during switching at a power supply of 4V was integrated to be 0.3μA for a 10μm device, or equivalent to an $I_{sub,age}$ of 0.03μA/μm. From the dc aging data shown in Fig.

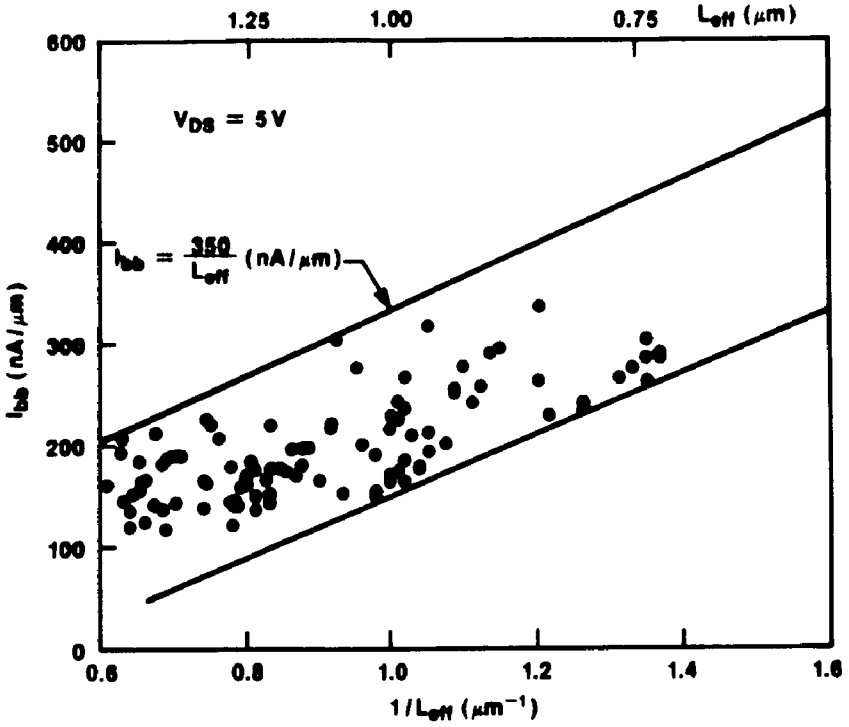


Figure 7-27. Maximum substrate current as a function of $1/L_{eff}$ for devices with L_{eff} from 0.75μm to 1.3μm.

7-22, the device lifetime is extrapolated to be almost 4000 years!. This simple assumption is not valid in actual circuit since the time the transistor takes to pull down the output node due to capacitive load is another factor that affect the ac circuit aging. It should be noted that in the ac mode, the hot carriers are generated and injected in a packet of charge, then followed by a relaxation time. This relaxation time is another advantage of using circuit for aging.

We have fabricated a 32-bit microprocessor using the optimized LDD.[50] This chip was previously manufactured with a previous technology using DDD for n-ch. The chip supply voltage was at 7V and the drift in the input low voltage level V_{IL} is monitored. Shown in Fig. 7-30 are the results of the circuit aging of the devices made with 2 different transistors structures with the DDD devices having thicker gate oxide than the LDD devices (250Å vs 210Å). After 1024 hours, the drift was very insignificant in for the LDD devices (5-10mV), whereas the DDD devices show a drift of 20-70mV. This result is consistent with the dc device aging shown in Fig. 7-18.

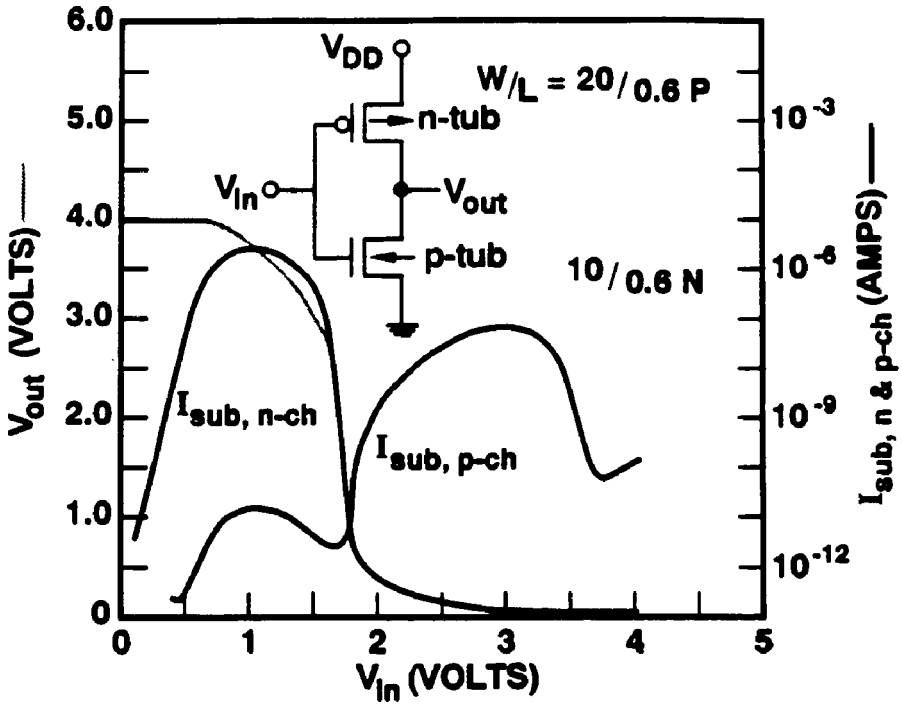


Figure 7-28. n & p-ch substrate (bulk) currents during switching of a CMOS inverter.

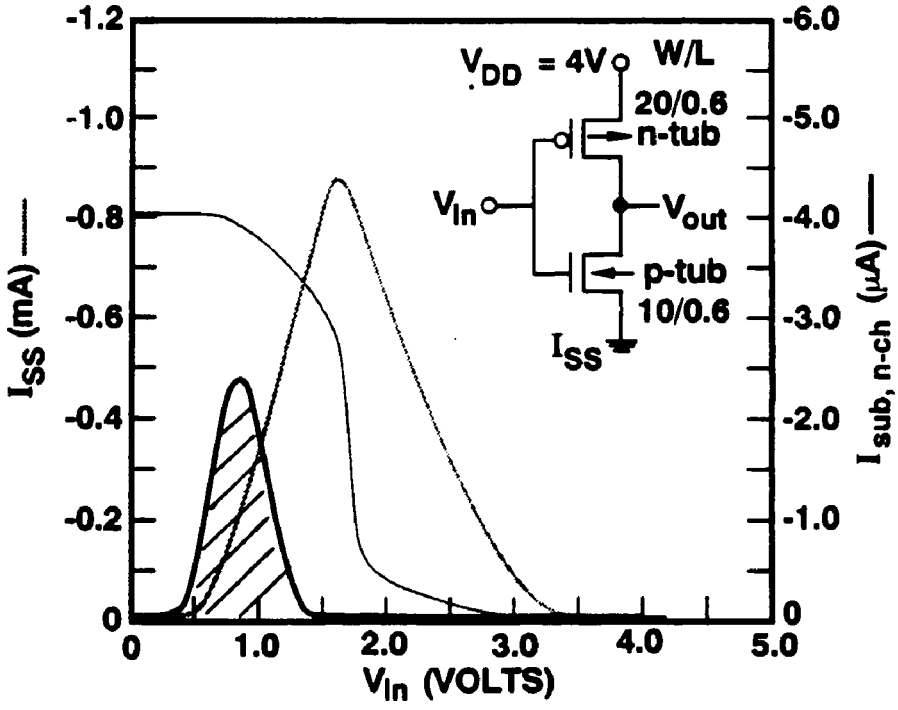


Figure 7-29. Averaged substrate current during inverter switching.

7.7 DISCUSSIONS

In this chapter, we have proved that the 10% g_m degradation is a very conservative approach to determine the device lifetime. With so many stringent requirements for submicron CMOS devices, this criteria may be extremely to meet even though power supply is reduced. The change in interface state density is a more severe reliability concern. We have designed boostrap circuits in NMOS technology, using 2 transistors connected in series where the high voltage node is a concern (Fig. 7-31). The operating principle of this additional device (M1) is rather simple. The source follower of transistor M1 limits the drain voltage V_{D2} to a threshold drop below V_{REF} that is tied to the gate of transistor M1. Therefore the drain-to-source voltage of transistor M2 will never exceed $V_{REF} - V_{tn2}$. This circuit technique was used in the NMOS DSP chip using conventional drain structure.[118] and in a CMOS 1Mb DRAM memory circuit.[119] The use of longer channel length device is also helpful to improve the lifetime in particular, the input/output buffers.

With the data obtained, we can explain the hot carrier aging process following:

1. Hot carrier injection at the pinch-off region near the drain creates damage in the oxide that leads to the generation of interface states. The damages region is initially created directly above the drain region in a form of an effect negative charge that would deplete the drain region. This in effect increases the series resistance of the aged device. During this phase, the interface states built-up is negligible as measured from the channel region using subthreshold conduction method, as seen in Fig. 7-17. During this phase, the g_m degradation is observed.

2. When the interface states, assumed to be the acceptor type, extended into the channel region, as can be measured by the subV_t method, the density increases more rapidly with a power law relationship with time. The built-up starts at the drain end, extends over the channel, and could reach the source if the stress time is sufficiently long, or the channel length is short. During this phase, the mobility is reduced and the threshold voltage is shifted toward the positive direction.

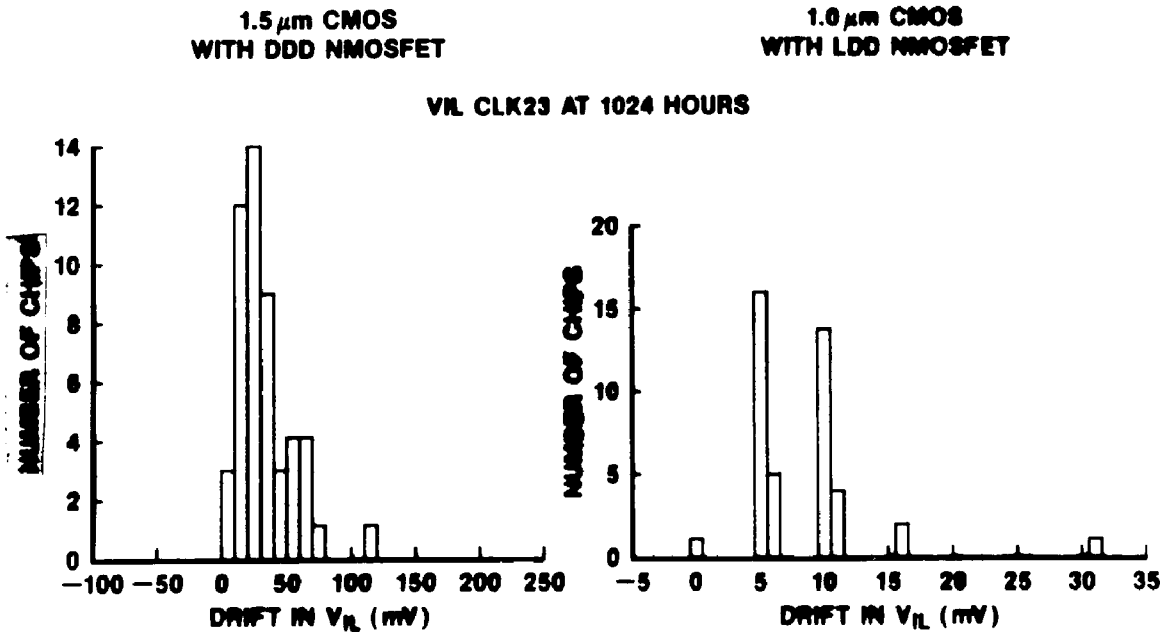


Figure 7-30. Circuit drift from hot carrier stressing. (The circuit aging was performed by P. Kempsey).

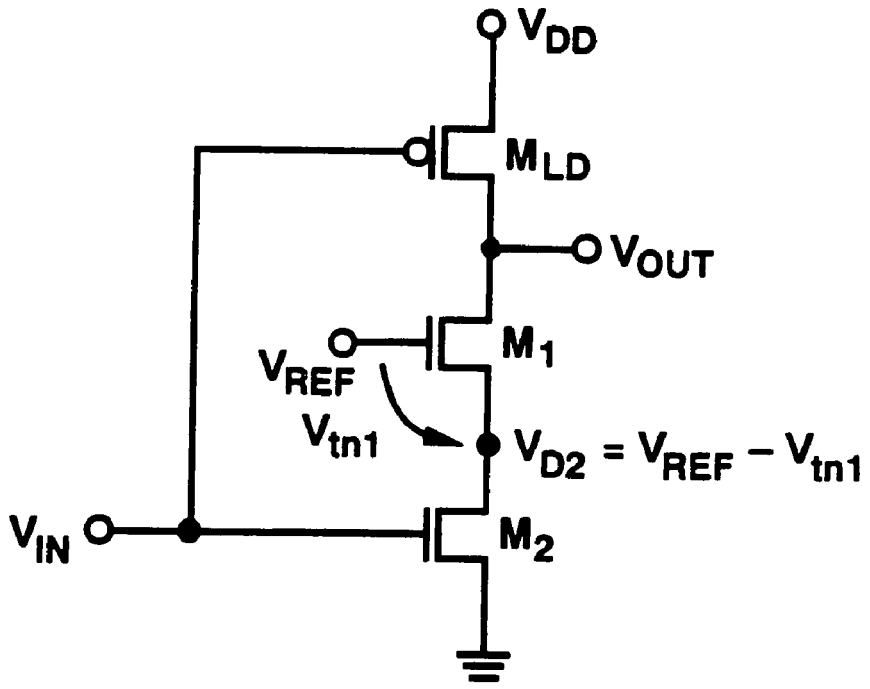


Figure 7-31. Series transistors connected in a cascode configuration to reduce hot carrier aging.

3. When the device is measured in the saturation mode after stressing, the reverse mode produces less channel current due to two reasons: the threshold shift due to negative charge, now at the source, and the reduction of the linear gain in the triode region due to the interface state built-up.

4. The impact ionization coefficient is the same for both forward and reverse modes at breakdown, although the substrate currents generated by both modes are different due to changes in the channel current.

Hot carrier effect is "a fact of life" in small geometry devices. And we, as device physicists, technologists and circuit designers must learn how to design and use these devices intelligently and to lessen the unnecessary burdens for a particular field to resolve the overall problem.

7.8 SUMMARY

We have presented and analyzed the experimental results of hot carrier effects on the n-channel LDD MOSFETs used in the 1.0 μm CMOS technology and spacer DDD devices with 0.5μm in channel length for 3.3V power supply. An interface density probing technique has been introduced that characterizes the built-up of states during hot carrier stress. Channel length dependence of lifetime extrapolation has been discussed. A more accurate expression of peak substrate current and applied drain voltage was obtained to accurately extrapolate the device lifetime to lower operating voltages.

Chapter 8

ADVANCED TOPICS and FUTURE WORKS

In this chapter, we will discuss the advanced topics related to device scaling and technology integration. Proposals for future works needed to realize sub half-micron technology will be described.

The key processing steps to realize the small geometry device is the lithography and pattern transfer. Assuming the sub-half micron lithography tools are available, then the issues to be resolved by device technologists are numerous. These include:

1. Device isolation to take full advantage of the pattern transfer capability.
2. Latchup prevention: The use of BiCMOS technology with the heavily doped N+ buried layer will improve the latch-up immunity. Deep trench isolation and selective epi are other alternatives.
3. Local Interconnects to improve packing density and reducing parasitic junction capacitance.

8.1 ISOLATION

With the scaling of CMOS technology down to a submicron geometry, the isolation between active devices has become more important to reduce the overall dimension of integrated circuits.

8.1.1 Drawbacks of Conventional LOCOS

The conventional Local Oxidation (LOCOS or SORT) pioneered by Kooi and Appels[27] using pad oxide and nitride has been used extensively in the industry as a standard technique for device isolation (refer to section 4.3.1 and 4.3.2 for LOCOS process sequence). The main drawback of this technique is the bird's beak encroachment during the field oxidation. This length of the bird's beak is in the order of 0.3 to 0.4μm per side (see Fig. 8-1).

For an active area feature of 0.7μm wide, a total compensation of ~+0.7μm is needed on both sides of the active area. Therefore, for submicron features, LOCOS isolation has its limits.

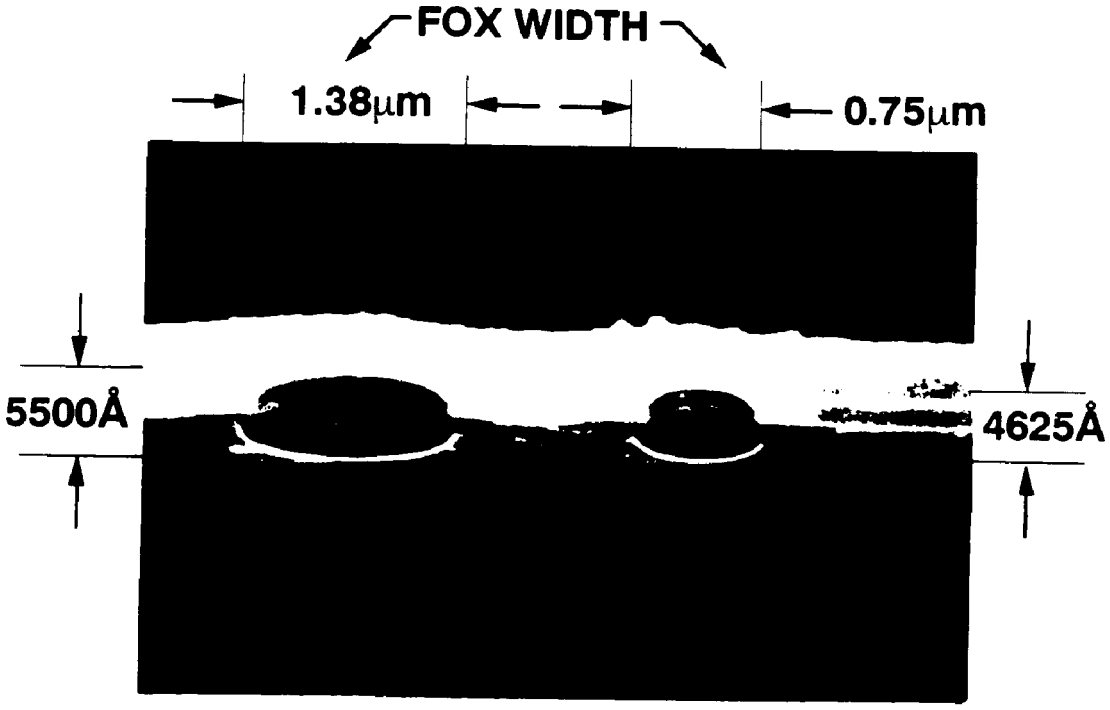
Figure 8-1. Conventional LOCOS Isolation.

Several techniques have been attempted to either reduce the bird's beak or to etch a vertical trench into the silicon to form recessed oxide (ROX) structure. All of these alternatives involve either process complexity or stress-induced defects.

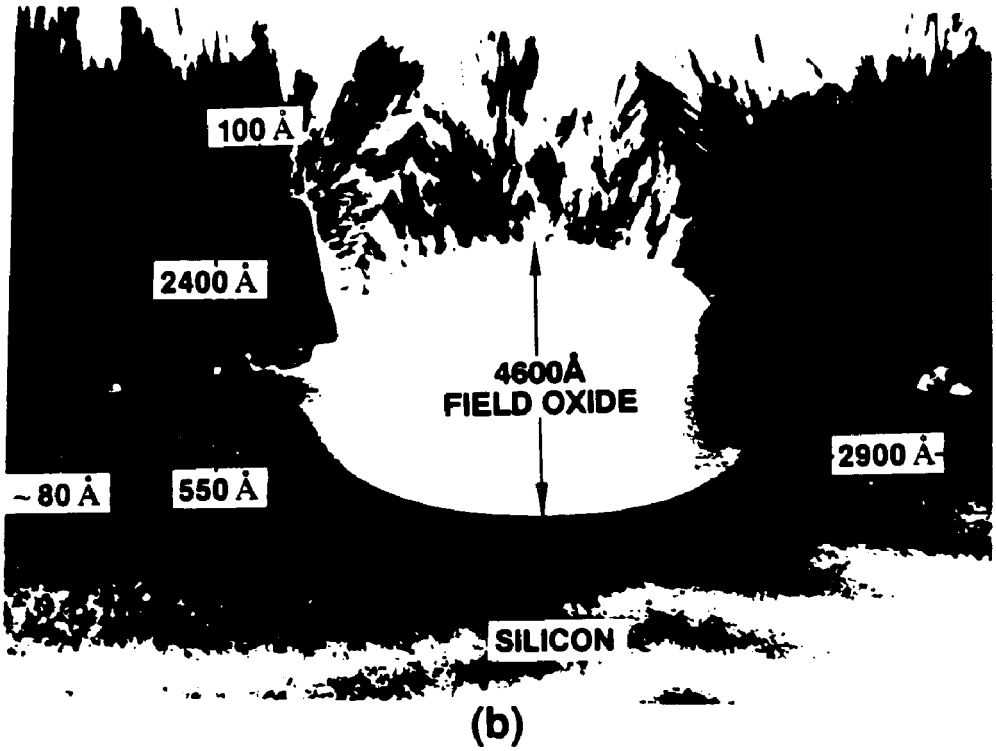
A straight forward modification of LOCOS employs a polysilicon layer as the buffer layer for stress relief. This technique is called Poly Buffered LOCOS (PBL) or Modified Local Oxidation (MLOCOS)

8.1.2 PBL Structure

It is known that the bird's beak in a LOCOS process can be reduced by thinning the pad oxide and thickening the nitride layer. However, this approach leads to the stress induced defects at the edges of the active area regions. These defects cause excessive junction leakage and transistor conduction along the width edge. In the poly buffer LOCOS process, a thin pad oxide is grown to the thickness of 100 to 130 Å. An undoped polysilicon of 500-700Å is then



(a)



(b)

Figure 8-2. Field oxide thinning in narrow space, (a) for .75μm spacing, oxide is thinned by 875Å compared with 1.38μm space. (b) TEM cross section shows the stress region on the poly layer.

chemically deposited in a low pressure chamber at 600°C. The thicker nitride in the range of 2000-2400Å is deposited. The nitride/poly stack is defined by a photo resist step and reactive ion etch to stop on the thin pad oxide. This approach produces more consistent field oxide control than the scheme proposed by ref[120].

The Nitride/Poly stack is etched in a reactive ion etcher using a combination of freon and oxygen plasma. The etch selectivity with respect to oxide is important to ensure that breakthrough of the thin pad oxide does not occur. The wafer is then cleaned and the field oxide is grown to the thickness of 6500Å.

8.1.2.1 Field Oxide Thinning in Narrow Space

One problem associated with the poly buffered LOCOS process is the field oxide thinning in narrow spaces. Shown in Fig. 8-2a is the SEM photograph of the field oxide at 0.75μm space between the nitride/poly edges. The measured thickness using SEM in this space is 4625Å. For a space of 1.38μm the field oxide (FOX) is grown to 5500Å, whereas the same field oxide in the wide open field measured 6300Å. A reduction of 27% and 13% is observed. Using TEM cross section of the structure, no stress at the bird's beak on the silicon substrate is observed. However, the stress is evident from the bright grain spots at the tips of the bird's beak in the poly layer. The final bird's beak after actual gate oxide and polysilicon gate process is less than 0.1μm as shown in Fig. 8-3. As mentioned earlier, the main advantage of PBL is in the device isolation. For the lithography and etching capability, the active device spacing of 0.8μm can be realized as seen in Fig. 8-2. The nitride/poly stack on the active region is removed by a series of oxynitride etch, nitride removal in wet chemical and poly removal by a dry etch. The pad oxide was then removed, the wafer was cleaned and the sacrificial oxide was grown to remove the Si_xN_y compound at the bird's beak. In the PBL process, care must be taken in etching back the sacrificial oxide, because the bird's beak is short, oxide over etch can cause a notch at the device edge to the field oxide.

Due to the field oxide thinning in narrow field oxide, this isolation scheme is an interim solution for a half-micron design rule. A more scalable isolation technology has to be developed. Shallow trench for fully recessed isolation is an attractive structure, although silicon etching

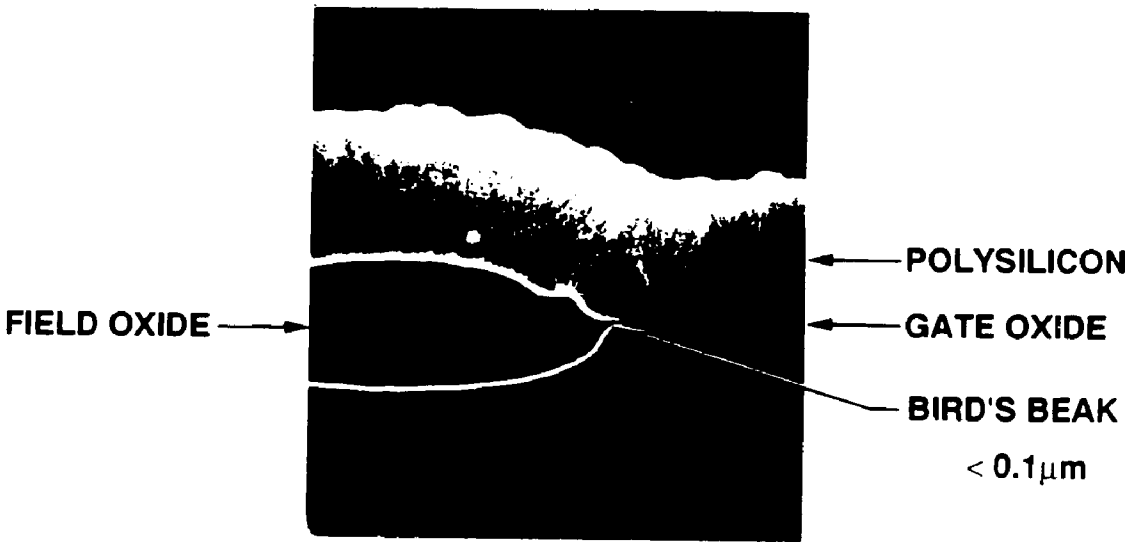


Figure 8-3. The final Bird's Beak at the edge of field oxide and gate oxide.

causing damage to the active areas sidewall and corners are serious device reliability issues.

8.2 BiCMOS

Although CMOS is the workhorse technology for VLSI and ULSI applications, its limitations have been discussed throughout this dissertation. The current drivability of CMOS is compromised when high capacitance load is encountered. The Electro-Static Discharge (ESD) protection at the input/output buffers poses a major challenge as the gate oxide is scaled down. This effect is compounded with the use of silicide to reduce the sheet resistance of the active area. In circuit applications such as sense amplifiers in memory, the small signal handling and transistor matching is crucial. The MOS device is less sensitive in this area. Designing input/output drivers compatible with ECL logic for fast operation is another drawback for CMOS, since the CMOS circuits used to perform the level conversion is very complicated, slow and inefficient. The bipolar transistor in a dominantly CMOS IC is an ideal complement for these shortcomings.

The bipolar transistor is known to be relatively insensitive to large capacitance load. The transconductance to load capacitance ratio, $\frac{g_m}{C_{load}}$, remains fairly constant for the bipolar device. The small signal sensing and transistor matching is excellent for bipolar transistor. ECL compatible, analog, I/O buffers and voltage regulators (for step-down power supply for submicron CMOS transistors) are naturally well suited with the use of bipolar devices.

The incorporation of bipolar devices into CMOS technology is called BiCMOS. The BiCMOS technology combines the best of bipolar and CMOS features, i.e. high packing density of CMOS and higher speed in performance. The power consumption for BiCMOS is comparable to that of CMOS if I/O circuits use in a totem pole configuration. A typical improvement in performance is about a factor of two for the same set of design rules. Therefore, with BiCMOS technology, capital equipment such as lithography tools and etching are fully utilized. The product offerings in BiCMOS technologies are more versatile, and range from SRAMs, microprocessors, to gate arrays. Additional complexity is the price to be paid for BiCMOS process. The processing steps to integrate an n-p-n transistor usually range from a simple 1 masking step for a low performance transistor, to 5 masks to achieve high performance devices with features such as poly emitter, buried collector, self-aligned emitter-base etc. In this section, we will describe a typical process sequence to realize a high performance bipolar device, with cut-off frequency, f_t, in the order of 10 GHz.

8.2.1 High Performance BiCMOS Fabrication Process

The most fundamental process for high performance bipolar transistor is to form the highly-doped buried collector. The collector series resistance is important to the bipolar frequency response as is the source-drain series resistance to MOS performance. The buried layer is also use as the retrograde n-tub and is very effective in latch-up suppression.

8.2.1.1 Buried Layer

The pad oxide and nitride sandwich layer is first formed in similar fashion as the LOCOS process. The buried layer is defined by a lithography and nitride etching step as shown in Fig. 8-4a. With the photoresist mask on the wafer, antimony (or arsenic) impurity is implanted into the

opening regions. The photo resist is then stripped, the wafer is cleaned and a buried layer protection oxide is grown at high temperature (1200-1250°C for Sb, 1000-1050°C for As). This oxidation step is to activate and to diffuse the implanted ions into the silicon substrate.

The n-p-n transistors are to be isolated from each other and from the NMOS and PMOS device. A buried p-type isolation is needed for this purpose. The nitride layer in between buried layers is removed. With the pad oxide serving as a screen, boron is implanted with the buried layer protection oxide serves as an implant mask(Fig. 8-4b). An anneal step is carried out to activate the boron species. The dosage of the boron implant is determined by the tradeoff between the sidewall capacitance of the buried layer and the isolation distance between two adjacent buried layers. The oxide is etched from the wafer to create a phobic surface. A lightly doped epitaxial layer of 1.3 to 1.5 μm thick is grown on top of the silicon substrate. The epitaxial layer is grown in an LPCVD reactor at temperature of 1000-1100°C.

8.2.1.2 Tub Formation

The twin-tub formation is similar in approach to that of standard CMOS process described in section 4.3.1, with the exception that the n-tub lithography is now aligned to the buried layer. The tub drive-in temperature is reduced so that a fairly shallow tub depth for the collector resistance consideration, and to prevent the buried layer antimony impurity from up-diffusing to the n-tub's PMOS devices.

8.2.1.3 Deep Collector Implant

The active area is formed in an identical procedure as described in section 4.3.2. After the field oxide is grown and the nitride on the active-device-to- be is removed, a photoresist step is done to define the implant mask opening to the collector contact area. Phosphorus is implanted at high energy (180 KeV) using the pad oxide as an implant screen (Fig. 8-5). This step is intentionally performed immediately after the active device formation to take advantage of subsequent thermal heat treatments to diffuse the phosphorus atoms to adjoin the buried layer. The purpose of the deep collector implant is to reduce the collector resistance from the contact (on the surface) to the buried layer.

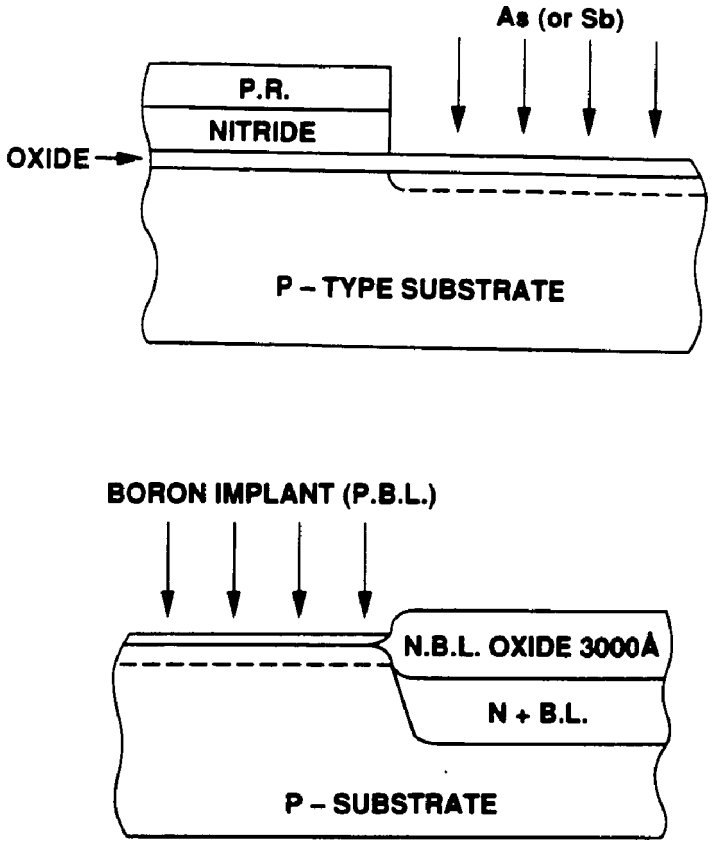


Figure 8-4. Buried layer protection oxide and the self-aligned p isolation implant.

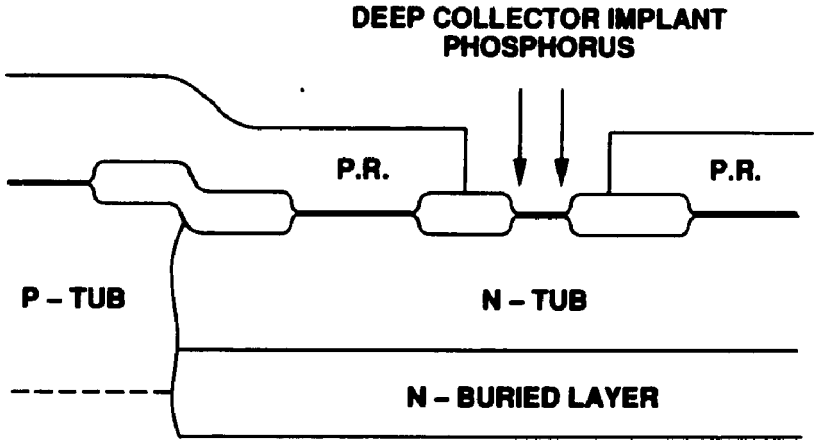


Figure 8-5. Deep-collector implant.

8.2.1.4 CMOS Gate Oxide Process

The CMOS sacrificial oxide is processed in similar manner as prescribed in section 4.3.3. The dosage of the threshold adjust implant is different from the CMOS since the resulting tub concentrations are different in the BiCMOS structure with a thinner epi-layer and due to bipolar collector considerations. The incorporation of the bipolar base is deferred as late in the process as possible. There are processes that the base implant is done before the gate oxidation.[121] In this work, we will describe a novel method to form the base after the gate oxide growth, referred to as Reverse Emitter Window (REW). [122]

8.2.1.5 Base Definition

After the gate oxide is grown, a thin undoped polysilicon layer of 600Å-800Å thick is deposited to protect the MOS gate oxide integrity. A thin separation oxide is grown to a thickness of 50Å on the polysilicon surface. A 2000Å nitride layer is deposited. The top view layout of the transistor and cross sections of processing steps are shown in Fig. 8-6. The nitride/oxide/poly stack is etched, the photo resist is removed, similar to the poly buffer LOCOS process described in section 8.1. The remaining gate oxide over the bipolar device active area is wet-etched and a base oxide is grown to a thickness of 800Å (Fig. 8-6d). A self-aligned base-emitter can be realized with a removable TEOS or nitride spacer. A boron base link-up is implant to reduce the base series resistance, and yet separated to the emitter window by the spacer (Fig. 8-6e). The spacer and the oxidation mask nitride are removed in wet chemicals, leaving oxide/poly structure on the wafer except where the base is to be contacted (Fig. 8-7). A photo resist mask is used to selectively implant the base area directly underneath the emitter window. An optional deep phosphorus implant with moderate dose can be used to increase the doping concentration of the collector beneath the base, while at the same time pushing up the boron base impurity profile, and reducing the base width. Figure 8-8 shows the doping profile of the emitter-base-collector with a local collector implant.

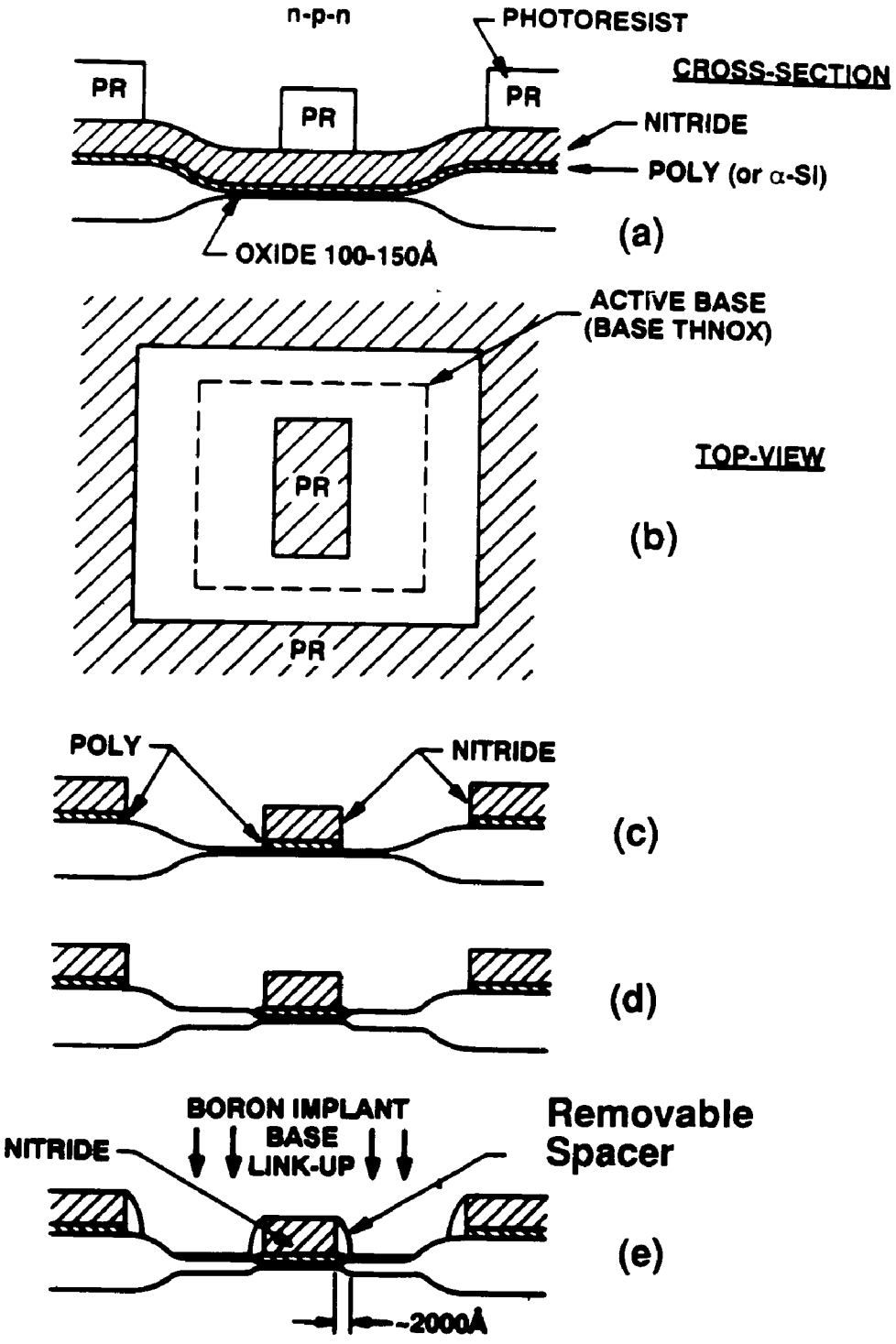


Figure 8-6. Reverse Emitter Window process to form emitter and collector implant.

8.2.2 Emitter Formation

With the base photo resist still on, the oxide/poly in the exposed area is removed by a dry etch. The photoresist is then stripped. A wet oxide etch is used to etch the gate oxide over emitter area and the poly-oxide. The bare silicon emitter window is now exposed. An undoped polysilicon layer with thickness in the range of (2000-2500Å) is deposited. This layer of polysilicon serves both as the poly emitter and to thicken the MOS gate to the desired thickness. The MOS gate and bipolar poly emitter are patterned by a single masking step. Although the poly thickness is different by ~ 600Å between MOS gate and emitter, the 500 Å final base oxide is used as etch stop for the polysilicon etch. Figure 8-7c shows the poly-emitter after gate etch. The LDD implants for CMOS devices are carried out as described in section 4.3.5. The spacer is formed on the MOSFET gates and emitter sidewalls. The P+ source/drain implant also dopes the p+ base contact area. The remaining of the fabrication sequence is compatible with the CMOS process as described in Chapter 4.

8.2.3 Bipolar Transistor Characteristics

We have fabricated BiCMOS devices, using 0.9μm design rules. The data presented here is for the purpose of explaining the difference between the bipolar transfer curves and MOS subthreshold characteristics described in Chapters 2 and 5. Shown in Fig. 8-9 is the Gummel plot of the collector and base currents, I_C and I_B, versus the base-emitter voltage V_{BE}. The slope of the I_C vs V_{BE} curve is obtained from the bipolar current equation:

$$I_C = I_{ss} e^{(V_{BE}/V_T)} \tag{8.1}$$

where the inverse slope is derived by taking logarithm of both sides from Eq. (8.1) as:

$$S_{bip} = \ln(10) V_T = 2.3 V_T \tag{8.2}$$

As seen from Fig. 8-9, the inverse slope was measured to be 60mV/decade at 300°K, indicating that the junction is ideal. If we referred to the MOS subthreshold slopes shown in Fig. 5-5 and 5-6, the n factor in this case is 1.

The gain, β or h_{FE}, is defined as

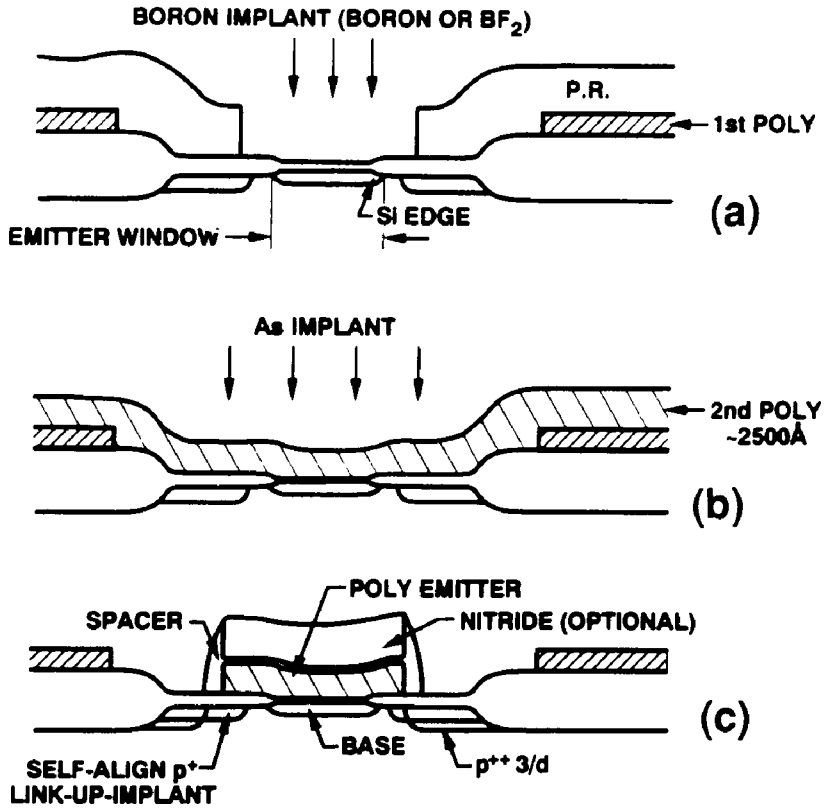


Figure 8-7. Emitter window and base and optional local intrinsic collector implants.

$$\beta = h_{FE} = \frac{I_C}{I_B} \tag{8.3}$$

and is plotted as a function of collector current I_C (Fig. 8-10). The gain is fairly constant over 7 decades of collector current indicating low recombination current from the base, and the base current tail which indicates the "Sah-Noyce-Shockley" recombination current, is not observed here (Fig. 8-9). For completeness, the I_C versus V_{CE} curves are shown plotted in Fig. 8-11. The collector-emitter breakdown voltage is higher than 7.5V, ensure a safe 5V operation. The Early voltage, equivalent to the saturation short channel effect in MOS device, was measured ~ -60V.

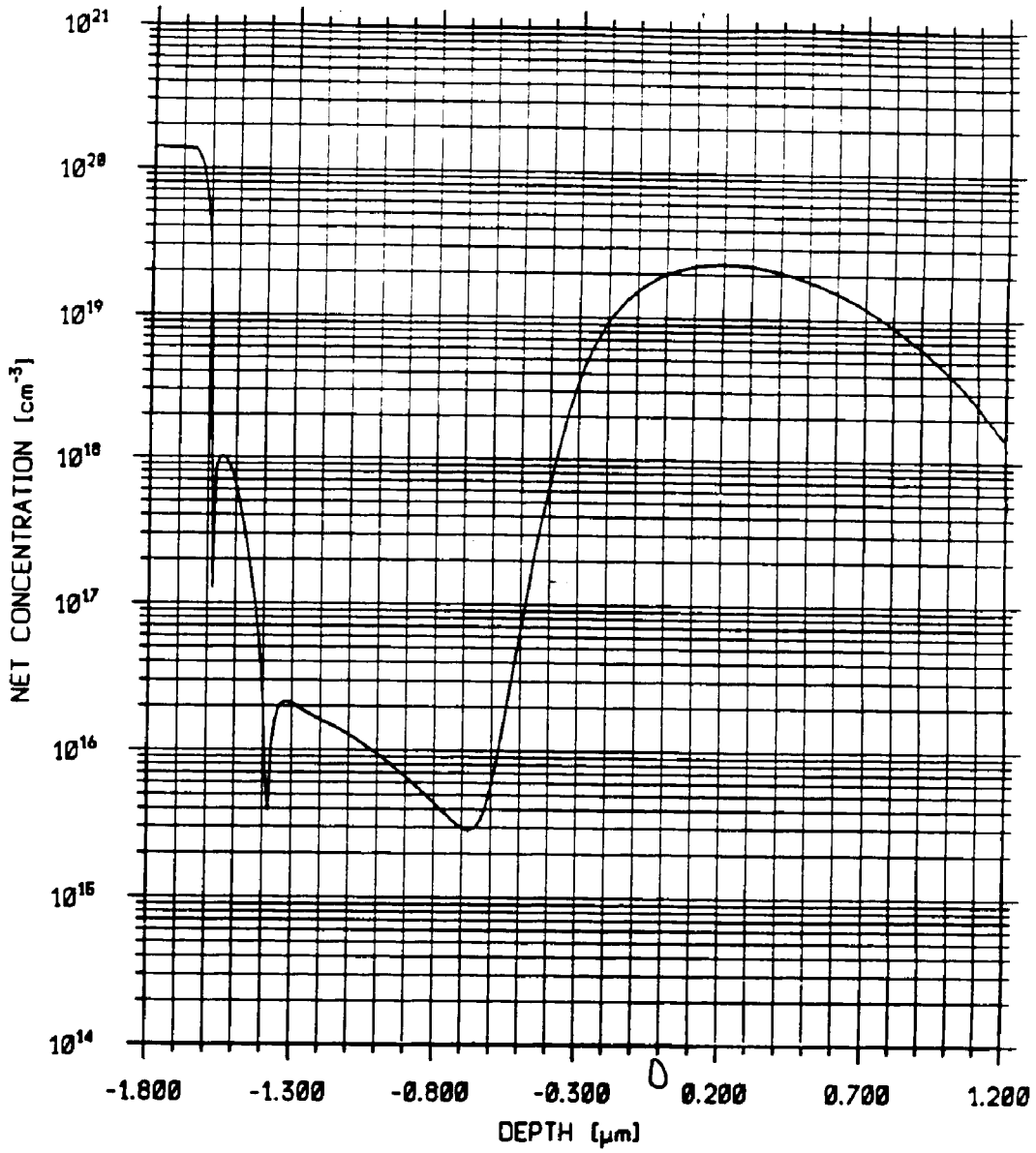


Figure 8-8. Doping Profile of Emitter-Base-Collector with a local collector implant.

For the submicron BiCMOS technologies, integrating high performance npn bipolar devices into a CMOS process is quite challenging. The high field and hot carrier effects observed in CMOS devices are also observed in bipolar structures. If a base-emitter junction is biased close to reverse junction breakdown, charges can be injected into the base oxide due to soft avalanche injection. The trapped charged increases the base surface recombination current, and therefore reduce the bipolar gain. This aging phenomena is equivalent to the transconductance

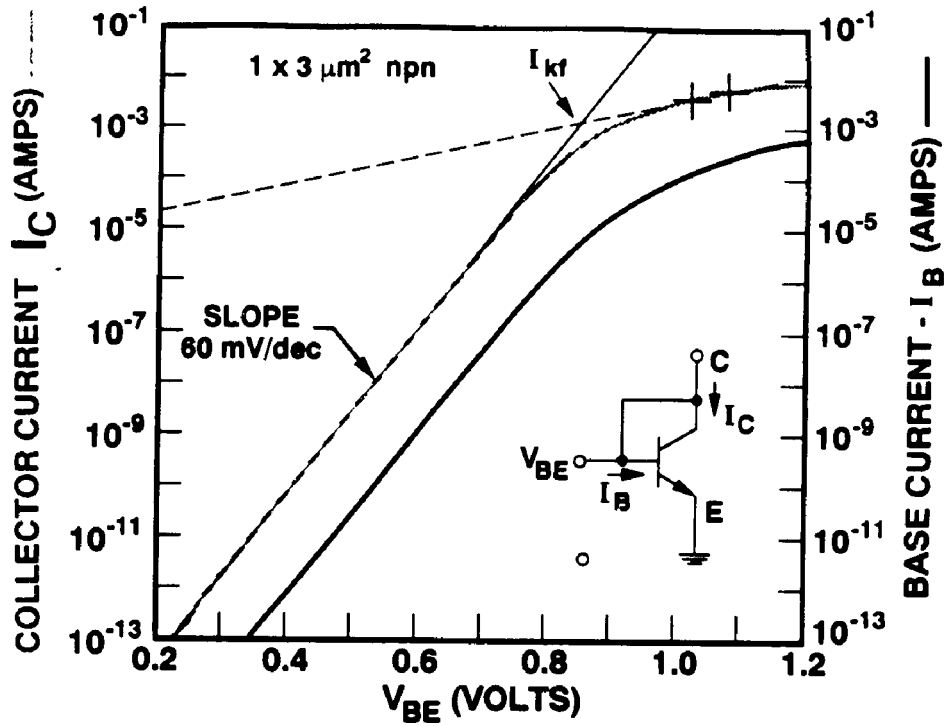


Figure 8-9. Gummel Plot of an npn bipolar transistor with 1×3 μm^2 Emitter window.

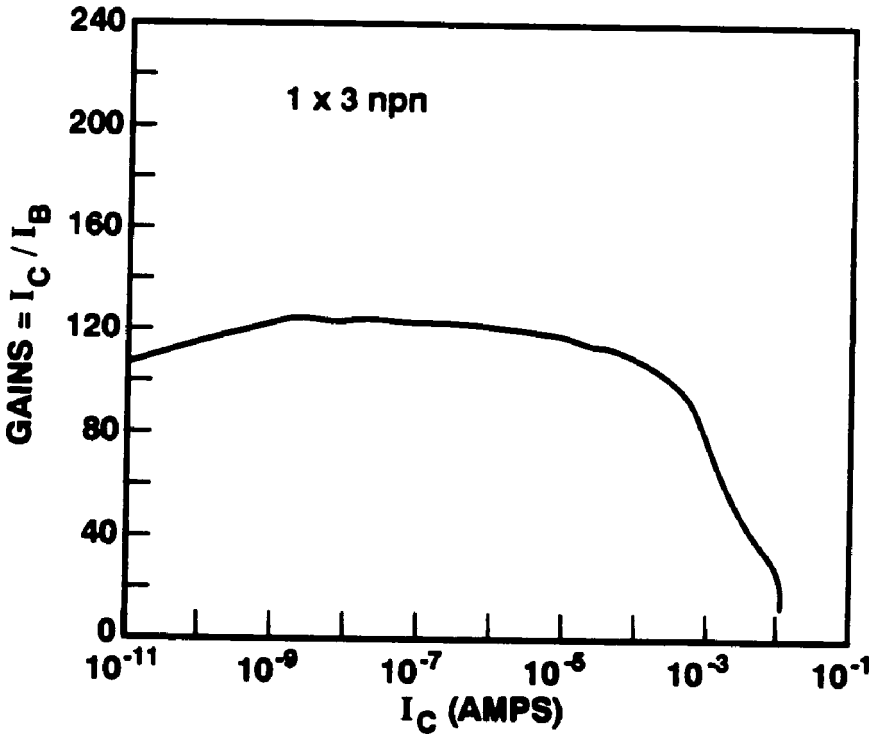


Figure 8-10. Bipolar gain vs collector current.

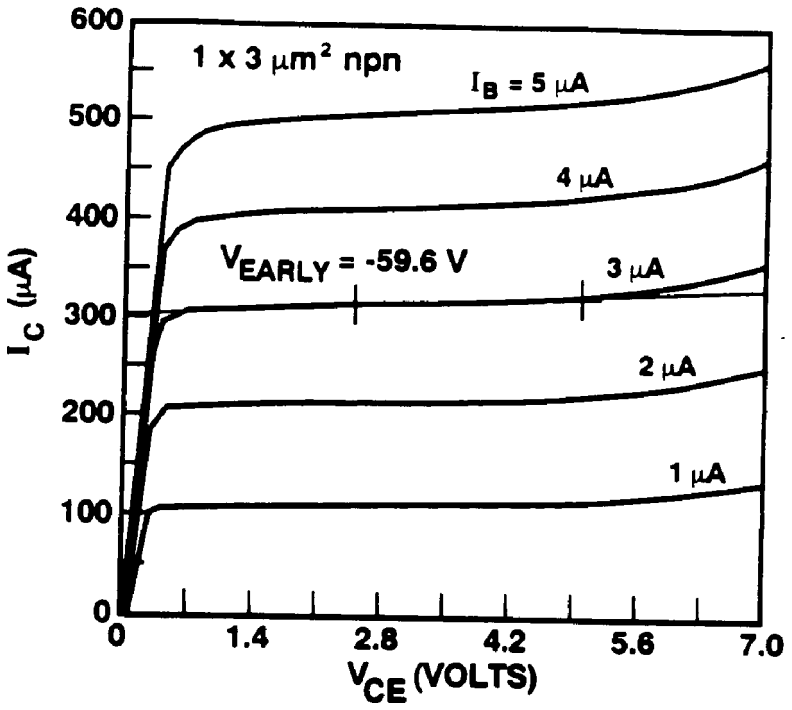


Figure 8-11. I_C vs V_{CE} characteristics.

degradation in an MOS device.

8.3 GATE DIELECTRIC MATERIAL

We have stated in chapter 3 that the gate oxide quality is crucial for the scaling of MOS devices into submicron (or sub-half-micron) geometry. As the required oxide thickness becomes less than 100Å, the defects introduced during pre- and post-gate oxidation processes are detrimental to the device reliability. Several alternative gate dielectric materials have been proposed. Oxide Nitride Oxide (ONO) sandwich,[65] ,[123] fluorinated oxide,[124] ,[125] and thermal oxide-TEOS-thermal oxide[63] have all been proposed to improve the oxide quality of less than 100Å gate oxides. All of the proposed schemes aim at improving oxide quality for under 100Å thick oxide.

The use of implanted gate with arsenic (or phosphorus) and boron to form n+/p+ gate is required for sub-half-micron devices. The effect of implanted poly gate on gate oxide quality is a

serious concern. We have designed an experiment that uses the large area MOS capacitors as a tool to characterize oxide quality.[126] A 125Å gate oxide is grown in the active areas after field oxidation. The polysilicon gate is deposited, and subsequently doped by phosphorus diffusion or arsenic/phosphorus implants. The dielectric breakdown of the oxide is measured by a voltage ramp technique, when the irreversible oxide current at breakdown is monitored. The oxide breakdown field statistics are collected for each experimental split. The defect density, D_0, is defined by a yield model that related to the area of the capacitors. The results show that the implanted gate MOS capacitors have higher defect densities (D_0=4-6 defects/cm^2) than D_0 of the oxide with the phosphorus diffused gates (D_0 <1). For the p+ (boron doped) gate MOS capacitors, a small trace of sodium is detected using the TVS (Triangle Voltage Sweep) as described in Chapter 5.

From the experimental results described above, gate oxide quality, in addition to the hot carrier aging, is prime concern for submicron devices. Potential dielectric material should possess the following properties:

- Ability to withstand the charging effect of ion implantation, plasma processing that can cause the microdefects during fabrication.
- Good barrier to sodium diffusion.
- High dielectric breakdown.
- And most importantly, low defect density.

Research in thin dielectric material continues to be an important area for submicron device fabrication.

8.4 LOCAL INTERCONNECTS & RAISED SOURCE/DRAIN

The local interconnect is becoming more critical in dense circuits like SRAMs. The local interconnect layer ties the adjacent source-drain of devices of the same type and thus eliminates the contact windows to metal. An additional advantage of the local interconnect is that the drain junction area can be minimized resulting in a reduction in capacitance and an improvement in circuit speed.

Polysilicon has been proposed as the raised source-drain alternative to silicided source-drain. The source/drain dopants are out-diffused into silicon from the polysilicon source, similar to the poly emitter formation in a bipolar structure. Selective silicon epi-layer has also been proposed for the elevated source-drain and to reduce hot carrier effects.[127]

8.5 OTHER RELIABILITY ISSUES

The active device reliability problems have been addressed and analyzed extensively throughout the course of this dissertation. Other reliability issues related to parasitics and materials are discussed briefly here for completeness.

Latch-up: The parasitic pnpn thyristor in CMOS structure can be turned on and latched into a high current state that can destruct the device. This phenomenon is called latch-up in CMOS a structure. Preventive measures include relaxed layout rules for circuits prone to latch-up, trench isolation between tubs to cut down the parasitic bipolar gain, and the use of retrograde tubs or n+ buried layer to reduce the gain of the pnp vertical bipolar device.

ESD: Electro-Static Discharge is due to the external charge built-up at Input/Output devices of MOS circuits. Some popular circuit prevention schemes for ESD include junction diode or MOS device to limit the voltage surge at the I/O terminal. From the device stand point, the gate oxide should be able to withstand high field without breakdown before the protection circuit activates. One solution to this problem is the use of BiCMOS technology, where bipolar transistors are used in the input/output circuits.

Electromigration due to high current density in a small geometry conductor (e.g. aluminum) is a serious material and structural reliability problems. Electromigration can result in open lines in a circuit or a short in a junction due to metal spiking. The overall circuit scaling is affected by the contact size and metal conductor design rules. Aluminum with small percentage of Cu-Si to help reducing the electromigration rate is commonly used for global interconnect in an integrated circuits. For submicron technology, layers of different conductors (Ti/Al) can alleviate the electromigration problem. Material research in this area is crucial for future technologies.

8.6 SUMMARY

In this Chapter, we have described some advanced topics needed to realized future device structures. The BiCMOS technology is interesting not only from the device physics and process integration stand points but also from the economic side of capital equipment investment.

The poly buffer LOCOS isolation is an alternative for device isolation in the upper-half micron technologies, but is limited by the field oxide thinning in narrow spaces, therefore is not suitable for sub-half-micron isolation. The shallow trench or selective epitaxial schemes have material defect issues to be resolved.

The local interconnect using doped polysilicon, if used in conjunction with raised source-drain, promises novel device structures. Local interconnect using TiN or TiW as conductors is also popular in SRAM applications.

Chapter 9

CONCLUSIONS And RECOMMENDATIONS

In this dissertation work, we have established a methodology for good submicron device design, based on device physics, circuit applications, and processing capabilities. In this process, we have contributed to an optimum LDD NMOS structure for a CMOS process, that extends the device reliability as compared with other CMOS technologies. During the course of the research, we have focused on the understanding of hot carriers degradation mechanisms in submicron CMOS devices. We have developed a simple characterization technique, based on the subthreshold conduction in an MOS device, to determine the spatial distribution of interface states in the damaged area close to the drain of the transistor during accelerated aging stress. The analytical model of hot carrier generated interface density was developed and using the extracted data to relate to the transconductance degradation. The technique can be used to routinely monitor the device aging to detect any potential problem in the fabrication process.

We further conclude that the DC aging method where the device is biased at the worst case peak substrate current is useful to compare the figure of merit of a given technology. The choice of 10% g_m degradation as a criterion for device lifetime is arbitrary, as is often found in the literature. This method does not reflects the actual circuit degradation due to hot carrier effects. Long channel devices are found to saturate in $\Delta g_m / g_{m0}$ at a lower level than the short channel devices for the same technology. We determine that the combination of an LDD structure and a thin p-type epitaxial layer on a heavily doped p+ substrate improves the drain-source breakdown voltage, which is one of the limiting factors in device scaling. We have instituted the use of substrate current as a means to monitor and detect any variations in the drain processes, in particular the spacer etch, in the routine manufacture environment. We have presented an integrated approach to tackle the hot carrier problems. Expertise in different areas, include device structures, e.g. drain-engineering; processing conditions, and circuit design techniques, all contribute to the solution of the hot carrier effects. We also employed a circuit design technique in a conventional NMOS process to reduce the hot carrier degradation.

We use the numerical process and device simulators to evaluate the n- and p-ch "halo" drain structures and for device design and analysis. We have also studied the new small device behaviors, namely the reverse subthreshold swing, reverse short channel effects on threshold voltage. We have examined the BiCMOS technology as an alternative for device scaling. The combined use of bipolar and CMOS devices leads to the next generation technology's performance while utilizing the existing processing equipment.

The external power supply voltage is still at 5 volts for CMOS circuits designed at 0.6μm channel length, although an "on-chip" voltage regulator has been used to reduce the internal voltage to 3.5-4.0V range. The next voltage standard will be 3.3V, and currently is intended to be used for the 64M DRAM (or 16M SRAM) generation of memory products. Therefore, drain-engineering is still an important area even with a reduced supply voltage, since ones cannot possibly change the supply voltage for every generation of technology.

For future research directions, the process of realizing submicron device structures is a multidisciplined field. Research on device structures should concentrate on drain structures and gate oxide quality. Correlation between dc aging on discrete devices with ac aging in an actual circuit will lead to an understanding of relaxation in hot carrier injection. New gate dielectric material is very important for future MOS devices. Fluorinated oxide and composite dielectrics (i.e. oxide/nitride/oxide, oxide/TEOS etc.) are promising candidates. In the area of device physics, velocity overshoot, ballistic transport, and energy transport are important areas to understand and to model devices smaller than 0.25μm. In the areas of materials research, new gate electrode materials (i.e. tungsten or molybdenum) are desirable for mid-gap work function. Insulators and conductors for metallization system are crucial for device interconnects.

The scope of microelectronic research is enormous. In the future, no single corporation can afford to devote resources to study and investigate all the areas for sub-quarter-micron device sizes. Cooperation among corporations, universities, and governments is the key to the success of the microelectronic industry.

REFERENCES

1. J. E. Lilienfeld, U.S. Patent 1,877,140 (filed in 1928, issued in 1932), and U.S. Patent 1,900,018 (filed in 1928, issued in 1933).
2. D. Kahng and M. M. Atalla, "Silicon-silicon dioxide field induced surface devices," in *IRE-AIEE Solid-Sate Device Res. Conf.*, (Carnegie Inst. of Technol., Pittsburgh, PA), 1960.
3. W. Brattain's notebook entry on 16 December, 1947, at Bell Laboratories.
4. W. Shockley, "The path to the conception of the junction transistor," *IEEE Trans. Elec. Dev.*, vol. ED-23, No. 7, pp. 597-620, 1976. Or *IEEE Trans. Elec. Dev.*, vol. ED-31, No.11, pp. 1523-1546, 1984.
5. W. Shockley, U.S Patent no. 2,569,347, filed 26 June 1948, issued 25 September 1951.
6. W. Shockley, "The theory of p-n junctions in semiconductors and p-n junction transistors," *Bell Syst. Tech. Jour.*, vol. 28, pp. 435-489, 1949.
7. G. C. Dacey, I.M. Ross, "A unipolar filed-effect transistor," *Proc. IRE., B50,* 1953.
8. W. L. Brown, "*N*-type surface conductivity on *p*-type germanium," *Phys. Rev.*, vol. 91, pp. 518-527, 1953.
9. I. M. Ross, U.S. Patent 2,791,760 (filed 1955, issued 1957).
10. M. M. Atalla, U.S. Patent 3,206,670, filed in 1960, issued in 1965.
11. D. Kahng, U.S. Patent 3,102,230, filed in 1960, issued in 1963.
12. M. M. Atalla, E. Tannenbaum, E. J. Scheibner, *Bell Syst. Tech. J.*, vol. 38, p. 749, 1959.
13. J. S. Kilby, "The Invention of the Integrated Circuit," IEEE. Trans. on ED, pp. 648-654, 1976.
14. R. Noyce, "Planar silicon BJT," 1959.
15. G.K Teal, M. Sparks, and E. Buehler, "Growth of germanium single crystal containing p-n junctions," *Phys. Rev.*, vol. 81, p. 637, Feb. 1951.
16. G. K. Teal and E. Buehler, "Growth of silicon single crystals and of single crystal p-n junctions," *Phys. Rev.*, vol. 87, p. 190, 1952.
17. W. G. Pfann, "Techniques of zone melting and crystal growing," *Solid State Physics,*" vol. 1, pp. 423-521, NY, 1957.
18. W. Shockley, U.S. Patent 2,787,564, filed October 28, 1954, issued April 2, 1957.
19. C. J. Frosch and L. Derrick, "Surface protection and selective masking during diffusion in silicon," *J. Electrochemical Soc.*, **104,** pp. 547, 1957.
20. H. C. Theuerer, H. H. Loar, J. J. Kleimack, H. Christensen, "Epitaxial diffused Transistors," *Proc. IRE.*, vol. 48, pp. 1642-1643, 1960.
21. B. T. Murphy, "Collector diffusion islolated integrated circuits," *Proc. IEEE.*, pp. 1523-1527, 1969.
22. L. M. Terman, "An investigation of surface states at a silicon/silicon oxide interface employing metal-oxide-silicon diodes," *Solid State Electronics,* vol.5, pp. 285-299, 1962.
23. J. L. Moll, IRE Wescon Convention Record Part 3, p.32, 1959.
24. B. E. Deal, A. S. Grove, "General relationship for the thermal oxidation of silicon," *J. Appl. Phys.*, vo. 36, p. 3770, 1965.
25. E. H. Snow, A. S. Grove, B. E. Deal, and C. T. Sah, "Ion transport phenomena in insulating films," *J. Appl. Phys.*, vol.36, no.5, pp. 1664-1673, 1964.
26. D. R. Kerr, J. S. Logan, P. J. Burkhardt, W. A. Pliskin, "Stabilization of SiO_2 passivation layers with P_2O_5," *IBM J. of Research and Development,* vol. 8, pp. 376-384, 1964.

27. J. A. Appels, E. Kooi, M. M. Paffen, J. J. H. Schatorje, and W. H. C. G. Verkuylen, "Local Oxidation of silicon and its application in semiconductor device technology," *Philips Research Report*, vol. 25, pp. 118, 1970.

28. J. C. Sarace, R. E. Kerwin, D. L. Klein, and R. Edwards, "Metal-nitride-oxide- silicon field-effect transistors, with self-aligned gates," *Solid State Electronics*, vol.11, pp. 653-660, 1968. Paper was presented at the Metallurgical Society (AIME) Meeting, New York, August 1967.

29. F. M. Wanlass and C. T. Sah, "Nanowatt logic using field-effect metal-oxide semiconductor triodes," *ISSCC Technical Digest.*, pp. 32-34, 1963.

30. M. H. White and J. R. Cricchi, "Complementary MOS transistors," *Solid-Sate Electronics*, vol. 9, pp. 991-1008, 1966.

31. S. M. Sze, "Historical developments and future performances of MOSFET," *VLSI Tech. Symp. Digest*, pp. 232-236, Taiwan, 1987.

32. R. H. Dennard, "Field effect transistor memory," U.S. Patent 3,387,286, June 8, 1968.

33. Digest of Technical Papers, International Solid-State Circuits Conference 1963-1968.

34. ISSCC Digest of Technical Papers, 1988, 1989 and 1990.

35. Y. Kakagome et al., "A 1.5V Circuit Technology for 64Mb DRAMs, *Symposium on VLSI Circuits*, pp. 17-18, 1990.

36. L. Kohn and S-W Wu, "A 1,000,000 transistor microprocessor," *ISSCC Tech. Digest*, pp. 54-55, 1989.

37. G. E. Moore, "Progress in digital integrated circuits," in *IEDM Tech. Dig.*, pp.11-13, 1975.

38. W. P. Hays et al, 32bit Floating-Point DSP, ISSCC Best Paper Award, 1986.

39. J. W. Beyers et al, "A 32b VLSI CPU chip," *ISSCC Digest of Tech. Papers*, pp.104-105, 1981.

40. L. C. Parrillo et al., "Twin-Tub CMOS, A Technology for VLSI Circuits," *IEDM Technical Digest*, pp. 752, 1980.

41. S. Kohyama, J. Matsunaga and K. Hashimoto, "Directions in CMOS Technology," *IEDM Technical Digest, 1983* pp. 151-154.

42. Y. Maki et al., "A 6.5ns 1Mb BiCMOS ECL SRAM," *ISSCC Tech. Dig.*, pp. 136-137, 1990.

43. M-L Chen et al., "A High Performance Submicron CMOS Process With Self-Aligned Chan-Stop and Punch-Through Implants (Twin-Tub V), *IEDM Tech. Dig.*, pp. 256-259, 1986.

44. H. Mikoshiba, T. Homma, K. Hamano, "A new trench isolation technology as a replacement of LOCOS," *IEDM Tech. Dig.*, pp. 578-581, 1984.

45. L. Manchanda et al., "Radiation-hard and enhanced isolation field oxide technology for GHz Si-CMOS VLSI," *IEEE Trans. Elec. Dev.*, pp. 651-658, 1990.

46. S. Nagao et al., "Application of Selective Silicon Epitaxial Growth for CMOS Technology," *IEDM Tech. Dig.*, pp. 593-596, 1984.

47. R. H. Dennard, et all, "Design of ion implanted MOS transistor," *IEEE J. of Solid State Circuits*. vol. SC-9, p. 256, 1974.

48. L. C. Parrillo et al., "Twin-tub CMOS II - An advanced VLSI technology," *IEDM Tech. Dig.*, pp. 706-709, IEDM 82.

49. Agraz-Guerena et al., "Twin tub III-A third generation CMOS technology," *IEDM Tech. Dig.*, pp. 63-66, 1984.

50. L. Tran, R. Ashton, B. Jones, C. Lawrence, D. McGillis, "Device Characterization of a 1.0μm CMOS Technology for ASIC," *CICC Tech. Dig.*, pp. 46-50, 1986.

51. S. J. Hillenius et al., "A symmetric Submicron CMOS Technology," *IEDM Tech. Dig.*, pp. 252-255, 1986.

52. K. Boyko, "Time Dependent Dielectric Breakdown of 210Å Oxides," *Proc. IRPS*, pp. 1-8, 1989.

53. M. H. White et al., "High-accuracy MOS models for Computer-Aided-Design," *IEEE Trans. Elec. Dev.*, vol. ED-27, pp. 899-906, 1980.

54. T. J. Krutsick, M. H. White, H.-S. Wong and R. V. Booth, "An improved method of MOSFET modeling and parameter extraction", *IEEE Trans. Elec. Dev.*, Vol. ED-34, pp. 1676-1680, 1987.

55. N. Liftshitz, "Dependence of the Work-Function Difference Between the Polysilicon Gate and Silicon Substrate on the Doping Level in Polysilicon," *IEEE Trans. Elec. Dev.*, vol. ED-32, no. 3, 1985.

56. C. Sodini et al., "Effect of High Fields on MOS Device and Circuit Performance," *IEEE Trans. Elec. Dev.*, vol. ED-31. pp. 1386-1393, 1984.

57. R. Booth and M. White, "An experimental method for the determination of the saturation point of a MOSFET," *IEEE Trans. Elec. Dev.*, pp. 247-251, 1984.

58. W. Fitchner et al., "0.15μm Channel-length MOSFETs fabricated using E-beam lithography," *IEDM Tech. Dig.*, pp. 722-725, 1982.

59. C. Kaanta et al., "Submicron wiring technology with tungsten and planarization," *IEDM Tech. Dig.*, pp. 209-212, 1987.

60. R. A. Chapman et al., "0.5 micron CMOS for high performance at 3.3V," *IEDM Tech. Dig.*, pp. 52-55, 1988.

61. T. Amazawa, H. Nakamura and Y. Arita, "Selective growth of aluminum using a novel CVD system," *IEDM Tech. Dig.*, pp. 442-445, 1988.

62. K. Kobayashi et al., "Dielectric breakdown and current conduction of Oxide/Nitride/Oxide multi-layer structures," *Symp. on VLSI Tech. Dig.*, pp. 119-120, 1990.

63. P. K. Roy et al., "Synthesis and characterization of high quality ultrathin gate oxides for VLSI/ULSI circuits," *IEDM Tech. Dig.*, pp. 714-717, 1988.

64. Y. Hiruta et al., "+BT Instability in P+ Poly Gate MOS Structure," *IEDM Tech. Digest.*, pp. 578-581, 1987.

65. A. Roy, F. R. Libsch, M. H. White, "Investigations on ultra-thin nitride scaled SONOS/MONOS memory transistors," *Proc. of the Symp. on Silicon Nitride and Silicon Dioxide Insulating Films*, 170th ECS Meeting, San Diego, 1986.

66. F. Baker et al., "The influence of Fluorine on Threshold Voltage Instabilities in P+ Polysilicon Gated P-channel MOSFETs," *IEDM Tech. Dig.*, pp. 443-446, 1989.

67. J. Sung et al., "Fluorine effect on Boron Diffusion of P+ Gate Devices," *IEDM Tech. Dig.*, pp.447-450, 1989.

68. C. Y. Wong et al., "Mobile Ion Gettering in Passivated P+ Polysilicon Gates," *Symp. on VLSI Technology Tech. Dig.*, pp. 123-124, 1990.

69. A. Sabnis and J. Nelson, "Characterization of Si/SiO_2 interface degradation due to hot-carrier injection," *IEDM Tech. Dig.*, pp.52-55, 1985.

70. S. Ogura, P. J. Tsang, "Design and Characteristics of the Lightly Doped Drain-Source (LDD) IGFET," *IEEE Trans. Elec. Dev.*, vol. ED-27, no.8, pp. 1359-1367, 1980.

71. M. Kinugawa et al., *Symp. on VLSI Tech.*, p. 116, 1985.

72. H. R. Grinolds, M. Kinugawa, M. Kakumu, "Reliability and Performance of Submicron LDD NMOSFET's," *IEDM Tech. Dig.*, pp. 246-249, 1985.

73. L. Parrillo et al., "A versatile, high-performance, double-level-poly double-level-metal, 1.2-micron CMOS technology," *IEDM Tech. Dig.*, pp. 244-247, 1986.

74. T. Mizuno et al., "High dielectric LDD spacer technology for high performance MOSFET using gate-fringing field effects," *IEDM Tech. Dig.*, pp. 613-616, 1989.

75. B. Meinerzhagen, private communication.

76. B. R. Penumelli, "A Comprehensive two-dimensional Process Simulation Program, BICEPS," *IEEE Trans. Elec. Dev.*, pp. 212-213, vol. ED-30, September 1983.

77. M. Law, C. Rafferty, and R. Dutton, "SUPREM IV Users Manual," Stanford University, 1986.

78. R. Fair, "PREDICT: Process Estimator for the Design of Integrated Circuit Technologies," MCNC, 1985.

79. GALENE II User's Guide, University of Aachen, 1990.

80. B. Meinerzhagen, "Consistent gate and substrate current modeling based on energy transport and the lucky electron concept," *IEDM Tech. Digest*, pp. 504-507, 1988.

81. M. Pinto, R. Dutton, PISCES IIB User's Manual, Stanford Univ., 1985.

82. L. C. Parrillo et al., "Twin-Tub CMOS, A Technology for VLSI Circuits," *IEDM Technical Digest*, pp. 752-755, 1980.

83. L. C. Parrillo, "Process and Device Considerations for Micron and Submicron CMOS Technology," *IEDM Technical Digest*, pp. 398-402, 1985.

84. R. D. Rung, "Trench Isolation Prospects for Application in CMOS VLSI," *IEDM Technical Digest*, pp. 574-577, 1984.

85. T. Shibata, R. Nakayama, K. Kurosawa, S. Onga, M. Konaka, and H. Iizuka, "A Simplified BOX (Buried-Oxide) Isolation Technology for Megabit Dynamic Memories," *IEDM Technical Digest*, pp. 27-30, 1983.

86. N. Endo, N. Kasai, A. Ishitani and Y. Kurogi, "CMOS Technology using SEG Isolation Technique," *IEDM Technical Digest*, pp. 31-34, 1983.

87. D. S. Yaney, J. Nelson, "The use of epitaxial layer on p+substrate for DRAM", IRPS, 1980.

88. B. R. Penumelli,"A Comprehensive two-dimensional Process Simulation Program, BICEPS," *IEEE Trans. Elec. Dev.*, pp. 212-213, vol. ED-30, September 1983.

89. R. A. Ashton unpublished work, and also Agraz-Guerena et al., "Twin tub III-A third generation CMOS technology," *IEDM Tech. Dig.*, pp. 63-66, 1984.

90. L. Tran, "Platinum Doped MOS Devices," *M.A.Sc Thesis*, University of Toronto, 1980.

91. R. Castagne and A. Vapaille, *Surface Sci.*, vol. 28, p. 557, 1971.

92. E. H. Nicollian and J. R. Brews, "MOS Physics and Technology," published by John Wiley & Sons, Inc., 1982.

93. C. W. Lawrence, M. J. Thoma, and A. Harrus, "Sodium drift in interlevel- dielectric of a two-level metal process," unpublished work.

94. W. T. Lynch and K. K. Ng, "A Tester for the contact resistivity of self-aligned silicides," *IEDM Tech. Dig.*,, pp. 352-355, 1988.

95. Wong, White, Krutsick, and Booth, "Modeling of transconductance degradation and extraction of threshold voltage in thin oxide MOSFET's," *Solid State Electronics*, vol. 30, pp. 953-968, 1987.

96. M. Orlowski, C. Mazure, and F. Lau, "Submicron short channel effects due to reoxidation induced lateral interstitial diffusion," *IEDM Tech. Dig.*, pp. 632-635, 1987.

97. C-Y Lu and J. Sung, private communication.

98. P. K. Ko et al., "A unified model for hot-electron currents in MOSFET's," *IEDM Tech. Dig.*, p.600, 1980.

99. A. G. Chynoweth, "Ionization rates for electrons and holes in silicon," *Phys. Rev.*, vol. 109, no.5, pp. 1537-1540, 1958.

100. L. Tran, "Hot Carrier Aging of 1.0 μm CMOS NMOSFET Devices," *Proceedings of the 1987 International Symposium on VLSI Technology, Systems and Applications*, p. 36, 1987.

101. E. Takeda and N. Suzuki, "An Empirical Model for Device Degradation Due to Hot-Carrier Injection," *IEEE-ED Letters*, pp.111-133, 1983.

102. C. Hu et al, "Substrate Current Model for Circuit Model," Berkeley Report.

103. S. Tam, P-K. Ko, and C. Hu, "Lucky-Electron Model of Channel Hot-Electron Injection in MOSFETs," *IEEE Trans. Elec. Dev.*, vol. ED-31, pp. 1116- 1124, 1984.

104. W. Shockley, "Problems related to p-n junctions in silicon," *Solid-State Electronics*, vol. 2, pp.35-67, 1961.

105. C. N. Berglund and R. J. Powell, "Photoinjection into SiO_2. Electron scattering in the image force potential well," *J. Appl. Phys.*, vol. 42, pp. 573-579, 1971.

106. D. R. Young, "Electron current injected into SiO_2 from p-type Si depletion regions," *J. Appl. Phys.*, vol. 47, pp. 2098-2102, 1976.

107. J. Hui, F. C. Hsu, and J. Moll, "A new substrate and gate current phenomenon in short-channel LDD and minimum overlap devices," *IEEE Elec. Dev. Let.*, vol. EDL-6, no. 3, pp. 135-138, 1985.

108. H. Sasaki, M. Saitoh, and K. Hashimoto, "Hot-carrier induced drain leakage current in n-channel MOSFET," *IEDM Tech. Dig.*, pp. 726-729, 1987.

109. H. S. Haddara and S. Cristoloveanu, "Parameter extraction method for inhomogeneous MOSFETs locally damaged by hot carrier injection," *Solid-State Electronics*, Vol. 31, No. 11, pp. 1573-1581, 1988.

110. J-J. Shaw and K. Wu, "Determination of spatial distribution of interface states on submicron, lightly doped drain transistors by charge pumping measurement," *IEDM Tech. Digest*, pp. 83-86, 1989.

111. G. Groeseneken et al., "A reliable approach to charge-pumping measurements in MOS transistors," *IEEE Trans. Elec. Dev.*, Vol. ED-31, No. 1, pp. 42-53, 1984.

112. W. G. Meyer and R. B. Fair, "Dynamic behavior of the buildup of fixed charge and interface states during hot-carrier injection in encapsulated MOSFET's," *IEEE Trans. Elec. Dev.*, Vol. ED-30, No. 2, pp. 96-103, 1983.

113. H-S. Wong, M. H. White, T. J. Krutsick, and R. V. Booth, "Modeling of transconductance degradation and extraction of threshold voltage in thin oxide MOSFET'S," *Solid-State Electronics*, Vol. 30, No. 9, pp. 953-968, 1987.

114. R. Bellens et al., "On the channel-length dependence of the hot-carrier degradation of n-channel MOSFET's," *IEEE Elec. Dev. Lett.*, pp. 553-555, 1989.

115. R. Booth, "Simulation and Measurement of Hot-Carrier Injection and Degradation in Short Channel MOS Transistors," Ph.D. Dissertation, Lehigh University, 1989.

116. H. Haddara and S. Cristoloveanu, "Two-Dimensional Modeling of Locally Damaged Short-Channel MOSFET's Operating in the Linear Region," *IEEE Trans. Elec. Dev.*, vol. ED-34, No. 2, pp. 378-385, 1987.

117. S. C. Sun and J. D. Plummer, "Electron mobility in inversion and accumulation layers on thermally oxidized silicon surface," *IEEE Trans. Elec. Dev.*, vol. ED-27, p. 1497, 1980.

118. L. Tran, "A 32K Bit High Speed On-Chip DRAM for Digital Signal Processor," *Proceedings of the IEEE Custom Integrated Circuits Conference*, pp. 166-169, 1985.

119. H. C. Kirsch et al, "1Mb CMOS DRAM," *ISSCC Tech. Dig.*, pp..., 1985.

120. R. A. Chapman et al., "An 0.8μm CMOS Technology for high performance logic applications," *IEDM Tech. Dig.*, pp.362-365, 1987.

121. R. H. Havemann, et al., "An 0.8μm 256K BiCMOS SRAM Technology," *IEDM Tech. Digest*, pp. 841-843, 1987.

122. C. L. Tran, patent applied.

123. T. Hori and H. Iwasaki, "Ultra-thin re-oxidized nitrided-oxides prepared by rapid thermal processing," *IEDM Tech. Dig.*, pp. 570-573, 1987.

124. R. Jaccodine, Elec. Chem. Soc., 1990.

125. T. P. Ma, "Fluorinated oxide," *Semi. Interf. Spec. Conf.*, San Diego, 1986.

126. K. Boyko, L. Tran, and D. Gerlach, "Oxide breakdown with different gate doping techniques," unpublished work.

127. A. Tasch et al., "A new structural approach for reducing hot carrier generation in deep submicron MOSFETs," *VSLI Symp. on VLSI Tech.*, pp. 43-44, 1990.

VITA

Conn-Luan Tran was born in Phong Dinh, Vietnam, on December 10, 1953 to Mr. Tran Van Trien and Mrs. Vo Thi Kiem. He received the B.S. degree in Electronics Engineering from National Institute of Technology, Saigon, Vietnam in 1974, and an M.A.Sc. degree in Electrical Engineering from University of Toronto, Canada in 1980. In the spring of 1984, he started his graduate studies for the Ph.D. degree in electrical engineering at Lehigh University, Bethlehem, PA, under Professor Marvin H. White. The financial assistance of this study was funded by AT&T Bell Laboratories.

Since 1982 he has been employed at AT&T Bell Laboratories, Allentown, PA, working in the areas of CMOS and NMOS technology development, device physics and DRAM design. From 1980 to 1982 he was with Philips Research Labs in Sunnyvale, CA, involved in work on EEPROM and polysilicon anti-fuse devices. His current research interests include device physics and process integration of submicron CMOS and BiCMOS devices, especially hot carrier effects and small geometry phenomena due to device scaling. He is also active in the areas of technology development and circuit applications of the devices.

Mr. Luan Tran is the co-recipient of the ISSCC Best Paper Award in 1986 for his work on the design of a 32-bit floating point digital signal processor. He is a member of IEEE and Sigma Xi Society.

PUBLICATIONS

1. L. Tran, "Hot Carrier Aging of 1.0 μm CMOS NMOSFET Devices," *Proceedings of the 1987 International Symposium on VLSI Technology, Systems and Applications*, p. 36, 1987.

2. L. Tran, R. A. Ashton, B. R. Jones, C. W. Lawrence, D. A. McGillis, "Device Characterization of A 1.0 μm CMOS Technology For Logic and Custom VLSI Applications," *Proceedings of the IEEE Custom Integrated Circuits Conference*, pp. 46-50, 1986.

3. K. H. Lee, B. R. Jones, C. Burke, L. V. Tran, J. A. Shimer and M. L. Chen, "Lightly Doped Drain Structure for advanced CMOS (Twin-tub IV)," *IEDM Technical Digest*, pp. 242-245, 1985.

4. With W. P. Hays et al., "A 32-bit VLSI Digital Signal Processor," *Journal of Solid State Circuits*, pp. 998-1004, October, 1985.

5. L. Tran, "A 32K Bit High Speed On-Chip DRAM for Digital Signal Processor," *Proceedings of the IEEE Custom Integrated Circuits Conference*, pp. 166-169, 1985.

6. With R. N. Kershaw et al., "A Programmable Digital Signal Processor with 32b Floating Point Arithmetic," *ISSCC Technical Digest*, pp. 92-93, 1985. *(1985 ISSCC Best Paper Award.)*

7. H. Schauer, L. Tran, L. Smith, "A High-Density, High-Performance EEPROM Cell," *IEEE Trans. Electron Devices*, Vol. ED-29, pp. 1178-1182, 1982.

8. D. W. Greve and L. Tran, "Polysilicon n^+-p-n^+ Structures for Memory Redundancy," *IEEE Trans. Electron Devices*, Vol. ED-29, pp. 1313-1318, 1982.