CMOS-Compatible High-Voltage Transistors

Thesis submitted by
Martin Russell Duncan
for the degree of
Doctor of Philosophy

Edinburgh Microfabrication Facility,
Department of Electrical Engineering,
University of Edinburgh
Scotland.

1992

ABSTRACT

Bipolar transistors are known to be the most suitable for high-voltage and power applications, due to their inherently greater current handling capability. In contrast, MOS technology is preferable for logic applications, due to its superior packing density. Therefore the 'ideal' solution to the smart power problem of integrating control elements on the same die as power switches is a marriage of the two different technologies. This results in a complex process that can only be cost effective in high volume applications. For ASIC applications and low volume product runs a less expensive compromise solution is needed.

By analyzing both bipolar and MOS, low and high voltage devices, it was found that if more than one power transistor is needed on the circuit, and a single technology is to be used, then MOS power transistors are inherently easier to integrate into a low voltage process. In particular the lateral double-diffused transistor (LDMOS) with all terminal contacts on the surface is to be preferred. Analyzing a CMOS process, common processing steps were found for both the low and high-voltage devices, leading to a smart power solution that does not need many masking levels.

By making small changes to an established n-well CMOS process, and developing a novel power transistor structure with a field oxidation separating the channel and drain, a 120 Volt n-channel power transistor could be realised within a conventional process with no additional processing steps. By adding one further masking layer, a complementary p-channel power transistor that supported -55 Volts was fabricated.

If these transistors were fabricated on a p- epitaxial layer on an n- substrate then by changing the p-channel power device structure, a breakdown voltage of -95 Volts could be achieved using only nine masking layers. The significant development of this work is that complementary LDMOS power transistors can be fabricated without any additional masks to the standard low voltage CMOS process. The only additional processing step is the use of an epitaxial layer used for isolation of the power transistors.

Simulation results indicate that if the epitaxial layer thickness is optimised, the RESURF effect, where the electric field at the surface of the transistors is reduced, can be applied. This results in n-channel power transistors that can withstand 200 Volts and p-channel transistors that support -180 Volts. These transistors are fully compatible with a low-voltage CMOS process.

Acknowledgements

I would like to thank all of my friends and colleagues who stood by me during the painfully long write-up period. A special apology is due to my wife Vittoria, who has had to put up with me while I thumped the PC keyboard night after night. I hope that now the work is finished we will be able to have normal conversations in the evening.

Particular thanks are due to my two supervisors Prof.J.M.Robertson and Dr.R.J.Holwill who gave me support when I most needed it.

An acknowledgement is necessary for all the EMF technicians, staff and PhD students who helped out with cleanroom processing, testing of finished wafers and drinking copious quantities of Belhaven 80' on various nights out. A special mention goes out to Robbie Hannah who helped out with the dirty work, i.e. binding and handing in. The fact that I bribed him with Barolo did nothing to influence his generosity in performing these tasks.

Finally, a thank you to Richard Rodrigues of Seagate, who helped me both morally and financially during this project.

Contents

List of figures

List of Tables

List of Symbols used

a	impurity gradient
a_{11}	constant in bipolar current equations
a_{12}	constant in bipolar current equations
a_{21}	constant in bipolar current equations
a_{22}	constant in bipolar current equations
A	cross-section area
BV_{DSS}	drain breakdown voltage
C_D	depletion capacitance
C_{it}	interface trap capacitance
C_o	gate oxide capacitance
D_n	electron diffusion coefficients
E	electric field
E_{cr}	critical electric field
E_{eff}	effective electric field
E_g	band gap energy
E_M	maximum electric field
E_{pt}	punchthrough electric field
E_s	surface electric field
E_{si}	peak field at the silicon surface
F_{max}	maximum operating bandwidth
g_m	transconductance
h	Planck's constant
I_B	base current
I_C	collector current

I_D	drain current
I_{Dsat}	drain saturation current
I_E	emitter current
I_G	gate current
I_o	diode reverse current
J	current density
k	Boltzmann's constant
K	process gain factor
K'	process gain factor at 300K
L	MOS transistor length
L'	effective channel length
L_a	accumulation layer length
L_D	drift region length
L_{GNS}	gate to drift region (nwell) separation
m	electron mass
M	multiplication factor
n_i	intrinsic charge density
n(0)	electron density at source
n(L)	electron density at drain
n_{po}	equilibrium electron density on p side of pn junction
N_A	acceptor impurity density
N_B	background dopant density
N_D	donor impurity density
q	electron charge
Q_B	Gummel number (bipolar)
Q_B'	effective Gummel number (bipolar)

Q_B	bulk charge (MOS)
Q_C	inversion region charge
Q_D	depletion region charge
Q_G	gate charge
Q_{inv}	charge induced by inversion layer mobile electrons
Q_S	surface charge
Q_{SS}	fixed oxide charge
R_{on}	on resistance
R_1	resistance
R_2	resistance
R_3	resistance
R_4	resistance
R_D	depletion mode device on-resistance
R_E	enhancement mode device on-resistance
R_{jfet}	JFET resistance
S	gate voltage swing
t_{on}	on time
T	period
T	temperature
V_A	accumulation layer voltage
V_B	junction breakdown voltage
V_B'	bulk voltage (MOS)
V_{br}	breakdown voltage of the base-collector diode
V_{br}'	reachthrough voltage of the base-collector diode
V_C	channel voltage
V_{CB}	collector base voltage

V_{CE}	collector emitter voltage
V_{cy}	cylindrical junction breakdown voltage
V_D	drain voltage
V_{Dsat}	drain voltage at saturation
V_{EB}	emitter base voltage
V_G	gate voltage
V_{in}	input voltage
V_{out}	output voltage
V_{pt}	punchthrough voltage
V_{sp}	spherical junction breakdown voltage
V_T	threshold voltage
V_{Tdep}	threshold voltage of depletion mode transistor
V_{Tenh}	threshold voltage of enhancement mode transistor
V_T^*	effective threshold voltage
W	MOS transistor width
W_B	base width
W_B'	effective base width
W_C	collector width
x_D	depletion region width
x_{dp}	depletion region width in the base
x_j	junction depth
α	common base current gain (bipolar)
α_T	base transport factor
β	common emitter current gain (bipolar)
β'	transport gain factor (MOS)
γ	emitter effeciency (bipolar)

γ'	back bias constant (MOS)
Ψ_B	fermi level difference
Ψ_S	surface potential
η	radius of junction curvature
μ_E	electron mobility
μ_{eff}	effective mobility
μ_N	drift mobility of channel electrons
μ_{max}	maximum mobility for a given doping level
Φ_{MS}	work function
ϵ_0	vacuum permittivity
ϵ_S	semiconductor relative permittivity
ϵ_{Si}	silicon relative permittivity

CHAPTER ONE : INTRODUCTION

The worldwide electronics industry has grown rapidly over the years. Total sales of electronics now rival that of traditional heavy industries such as automobiles, chemical and steel. Electronics equipment sales have risen from $114 billion in 1981 to $265 billion in 1989. They rose further to $300 billion by 1991 and are predicted to reach $500 billion the turn of the century, see figure 1.1 [1,2]. Similarly total solid-state shipments have risen from $45 billion in 1988 to $60.6 billion in 1991. It can also be seen from figure 1.2 that ICs have continued to rise at a faster rate than discrete device sales.

As the total sales increased, production and capital spending has had to increase to keep pace with requirements [3]. From figure 1.3, it is seen that total spending is predicted to increase worldwide from $78 billion in 1990 to $159 billion in 1994. This represents a compound annual growth of 15%. This is despite the fact that the industry went though a small recession in 1990. An upward trend is predicted for 1992 onwards.

There has been a larger than average increase of spending in Europe. Several companies have either opened or commissioned plants in time for 1992 and the single market. The European Community has ruled that the first diffusion process defines a chip's origin. Therefore to maximise profit margins, American and Asian companies have been obliged to set up fabrication facilities within the Community, as extra-communitary chips will be subject to higher taxation.

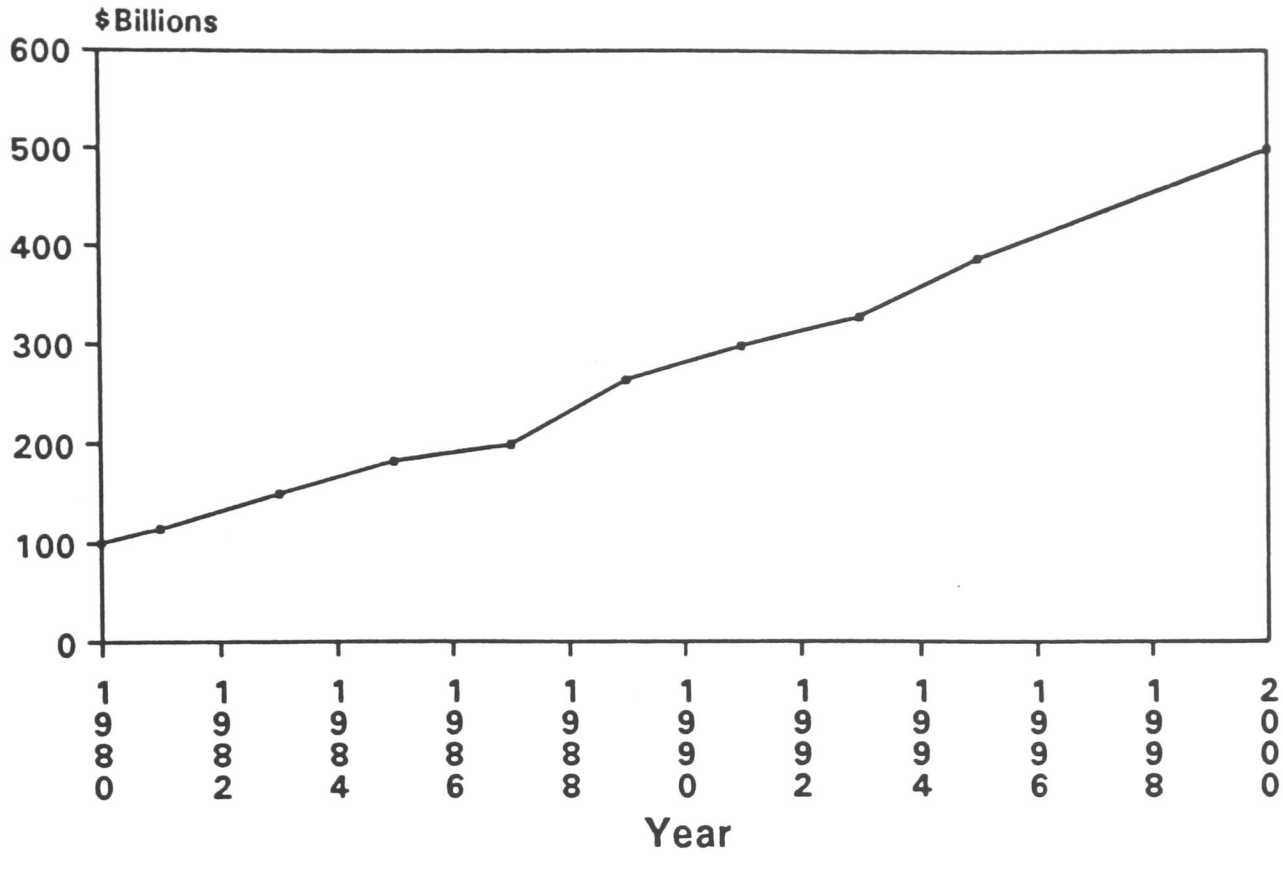


Figure 1.1 The increase of Worldwide Electronics Sales from 1980 projected to 2000

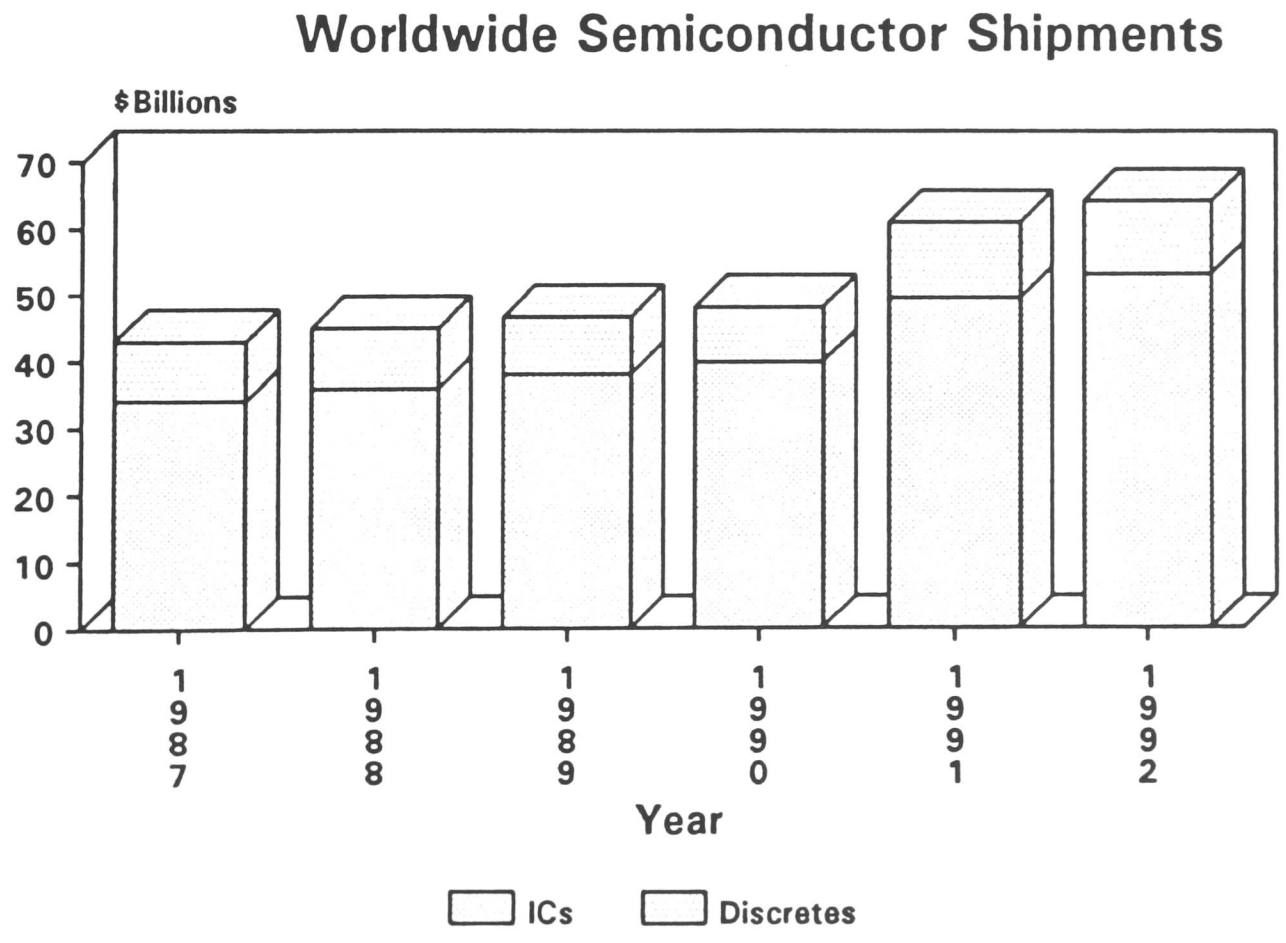


Figure 1.2 The increase of Worldwide Semiconductor Shipments from 1987 to 1992 (for both integrated and discrete devices).

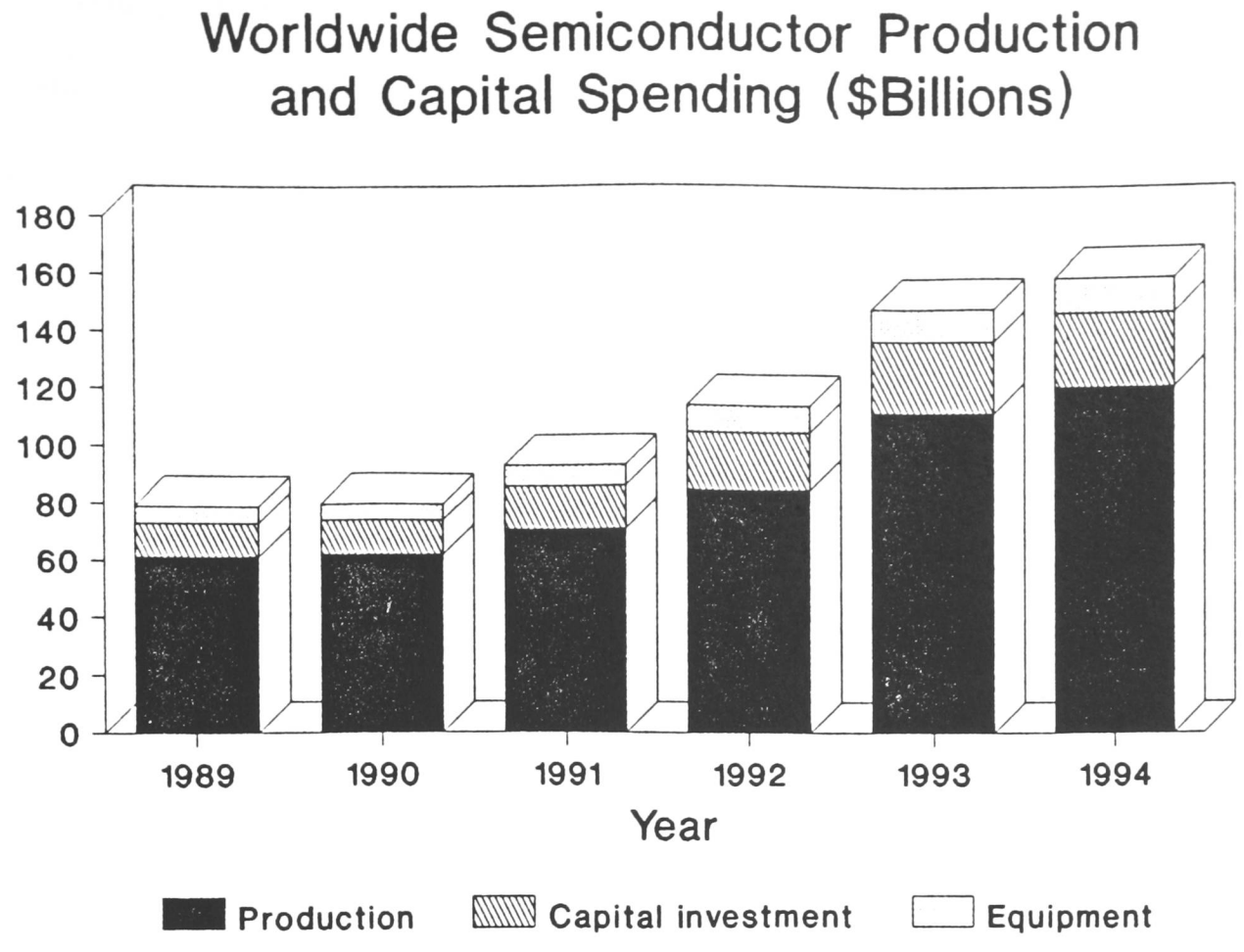


Figure 1.3 The predicted Semiconductor Production and Capital Spending from 1989 to 1994.

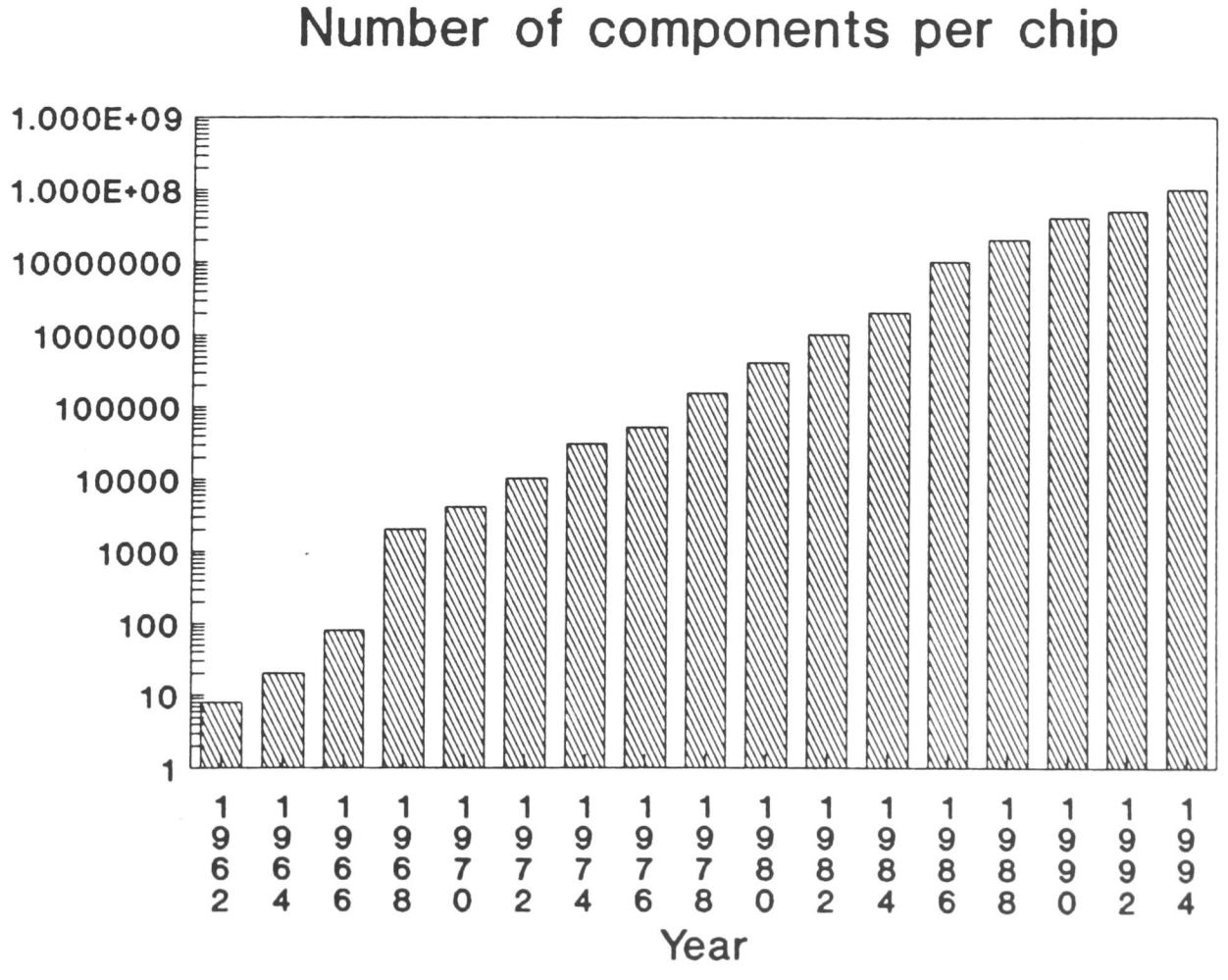


Figure 1.4 The increase of the number of components per chip from 1962 to 1994.

There has been a change in demand for different types of semiconductors over the years. In 1947 the semiconductor industry was born with the invention of the bipolar transistor by Shockley et al [4]. The first integrated circuits were first fabricated in 1958 by Noyce et al at Fairchild [5], and by the mid-sixties sales overtook those of discrete devices. Since 1975, however, digital MOS ICs have dominated the market place. This is due to their advantages in device miniaturisation, low power dissipation, high yield and inherent simplicity. However as MOS structures become more complex, the operating characteristics and parasitic effects mean that both bipolar and unipolar phenomena have to be understood. It is not enough for circuit designers to be competent in only one of the fields. With the development of mixed bipolar and MOS technologies this is becoming more and more true.

The number of components on a chip has grown exponentially over the years, as shown by figure 1.4. IC complexity has advanced from small scale integration (SSI) through medium scale (MSI) and large scale (LSI) to very large scale integration (VLSI). As an example, 16M DRAMs are now being produced with approximately 4×10^7 components per chip. The next generation of 64M DRAMs is planned to be introduced in 1994 and will have 10^8 components. Asian, American and European suppliers are all trying to rewrite Moore's law (which predicts an exponential decrease in feature size with time) by accelerating the change from one generation of products to the next. This development drives prices down, and it is predicted that bit-for-bit pricing will crossover for the 4M and 16M DRAM in the second half of 1992.

To stay competitive die sizes and therefore minimum feature sizes have had to shrink accordingly. The minimum feature size for memory chips has changed from 1.5 μm for 256K, through 0.8 μm for 4M, to 0.5 μm for 16M (see figure 1.5). This has meant changes in photolithography from g-line to i-line steppers and it is predicted that future generations will need excimer laser or even x-ray sources.

Whilst memories (DRAMs, SRAMs and EPROMs) remain the driving technologies for miniaturisation, other semiconductor product sectors have either improved vastly or have been newly created. Power integrated circuits (PICs), long a sleeper in the electronics industry are stirring to life after a decade of incubation in the laboratory. PICs are being investigated for use in power conversion and control of high voltages and high currents in automobiles, television and audio equipment, home appliances, robotics, aircraft, motors, switching devices, flat-panel, electroluminescent displays and electrostatic printing.

When a power device is integrated with a control circuit on a single chip there are substantial improvements in both performance and cost. The savings in cost come from eliminating the many packages to house individual chips, abolishing the interface circuit between power transistors and control ICs and shrinking the overall system size. Having an integrated solution can also lead to faster system speed and greater reliability. Furthermore, the integration of power devices and control ICs brings about a range of functions such as temperature control, over-voltage and under-current protections that have been unavailable with discrete power devices.

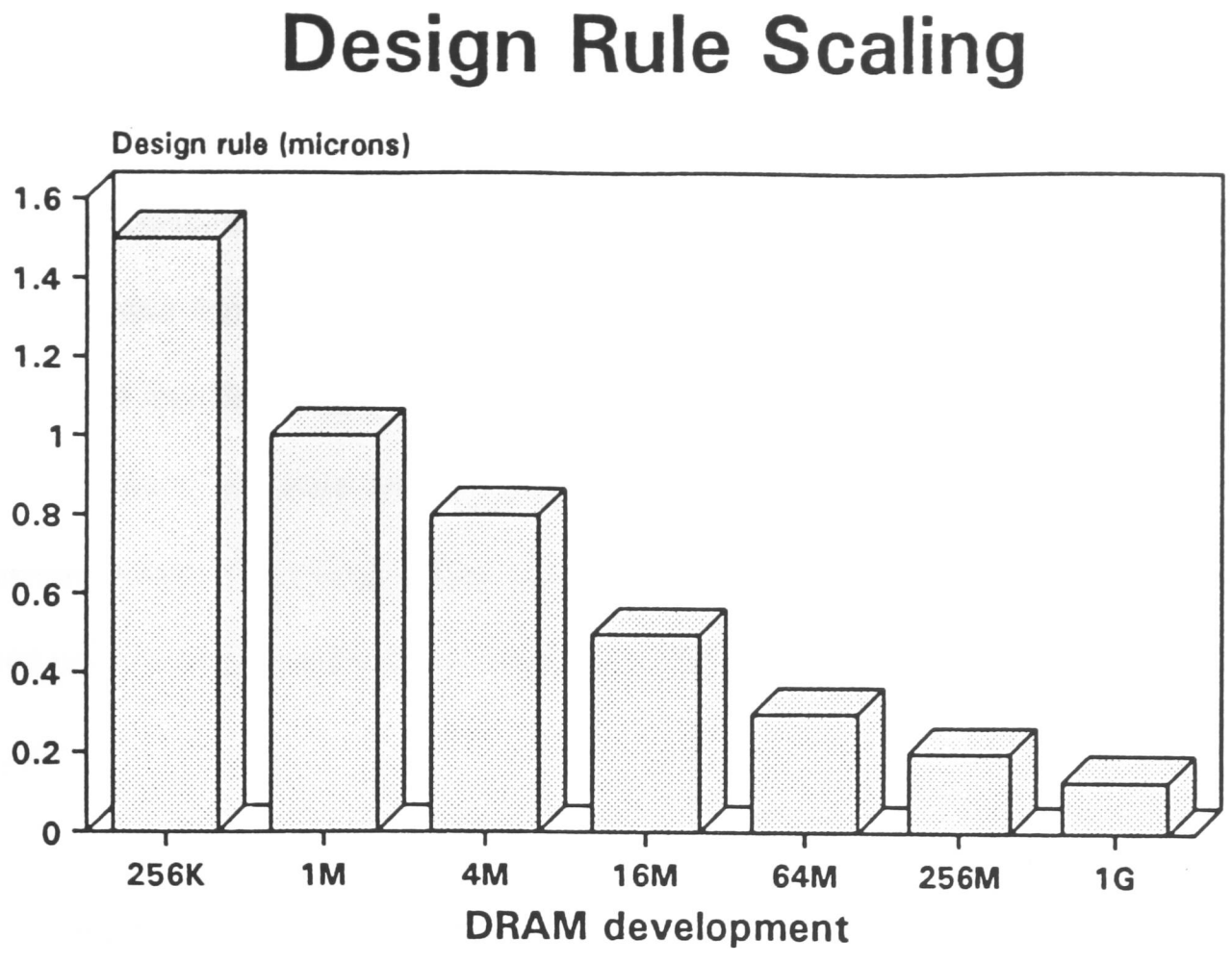


Figure 1.5 The decrease of design rules for various generations of DRAMs.

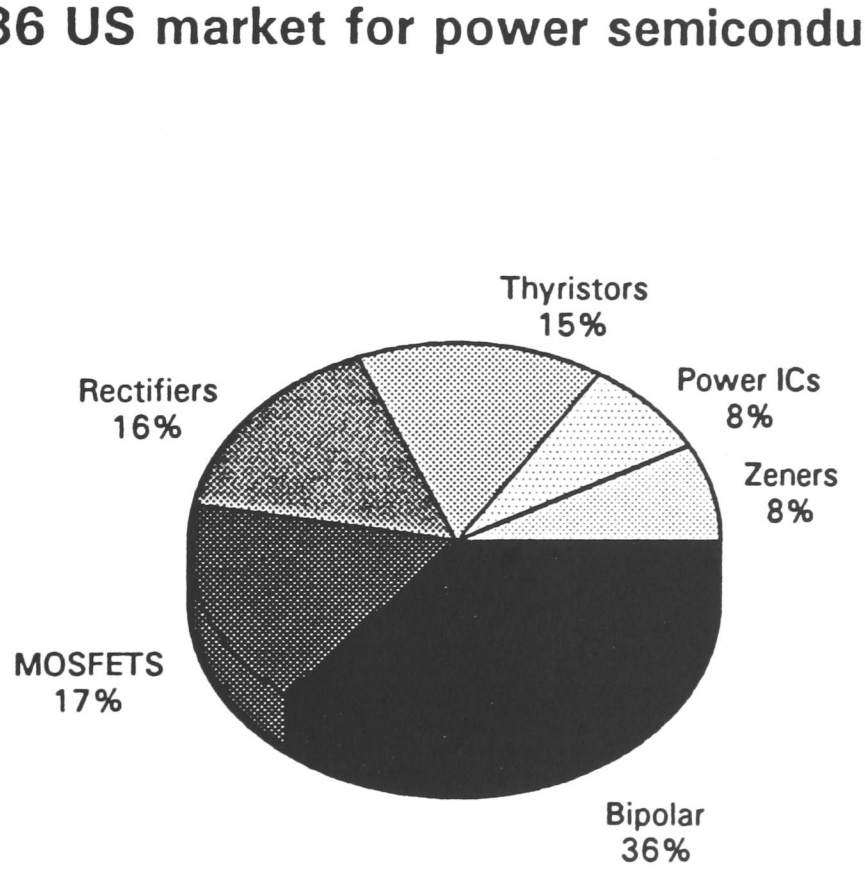


Figure 1.6 The split of the 1986 US Market for power semiconductors by product group.

It is important to distinguish PICs from high-voltage integrated circuits (HVICs) which are designed to deliver relatively low output current. HVICs generally consist of a few high voltage switches combined on one substrate to realise simple functions. Further, PICs can be split into two categories ; true smart power and intelligent switches. An intelligent switch consists of a single power device integrated with logic. This type of technology can be considered a type of smart power ; but with a low I.Q. True smart power consists of multiple power transistors and complex logic , and it is this technology that is developed in this project.

As in the area of digital and small-signal analogue electronics, the adjacent domain of power electronics can be characterised by the issue of FET versus bipolar transistor technology. Although the trend is away from bipolar towards DMOS-based MOSFETs for discrete devices, there is still debate on the most suitable solution for smart power ICs. Consequently both technologies will be discussed in later chapters.

Up to 60V, MOSFETS have a lower on-resistance than the same sized bipolar transistor. Above this value a penalty has to be paid in terms of larger die area for lower on-resistance. For applications at greater than 500V breakdown voltage, it is generally acknowledged that bipolar transistors remain the more attractive in terms of low on-resistance. However, bipolar transistors exclude themselves from many applications, or put constraints upon designers because of their large rise, storage and fall times, which can be two orders of magnitude slower than unipolar devices.

From a 1986 report of the US situation for power semiconductors by Frost & Sullivan [6], it can be see that power ICs took 8% of a $1 billion dollar market. Further, figure 1.7 shows that bipolar technology dominated the world power IC market with a 52% share compared to a 19% share for MOS devices. Figure 1.8 shows the predicted growth of the world power device market. It can be seen that there is a substantial growth in the smart power sector as well as a 20% per annum growth in total sales. By 1993 it is predicted that monolithic power solutions will take 60% of a $ 1.8 billion dollar market.

A report from BIS-Mackintosh [7] predicts that the automotive industry will be the biggest user of smart power with 40% of the market, by 1995. Therefore it seemed appropriate to develop processes that address the requirements of this type of application.

There are two types of solution available, lateral and vertical. It is acknowledged [8] that lateral smart power chips are more suitable to applications such as computer peripherals and automotive products. Vertical technologies are better suited to devices that are simpler but more powerful. They can be used in applications such as solenoid driving, power supplies and electronic ignition. The power transistors used in both of these technologies will be described later in this work.

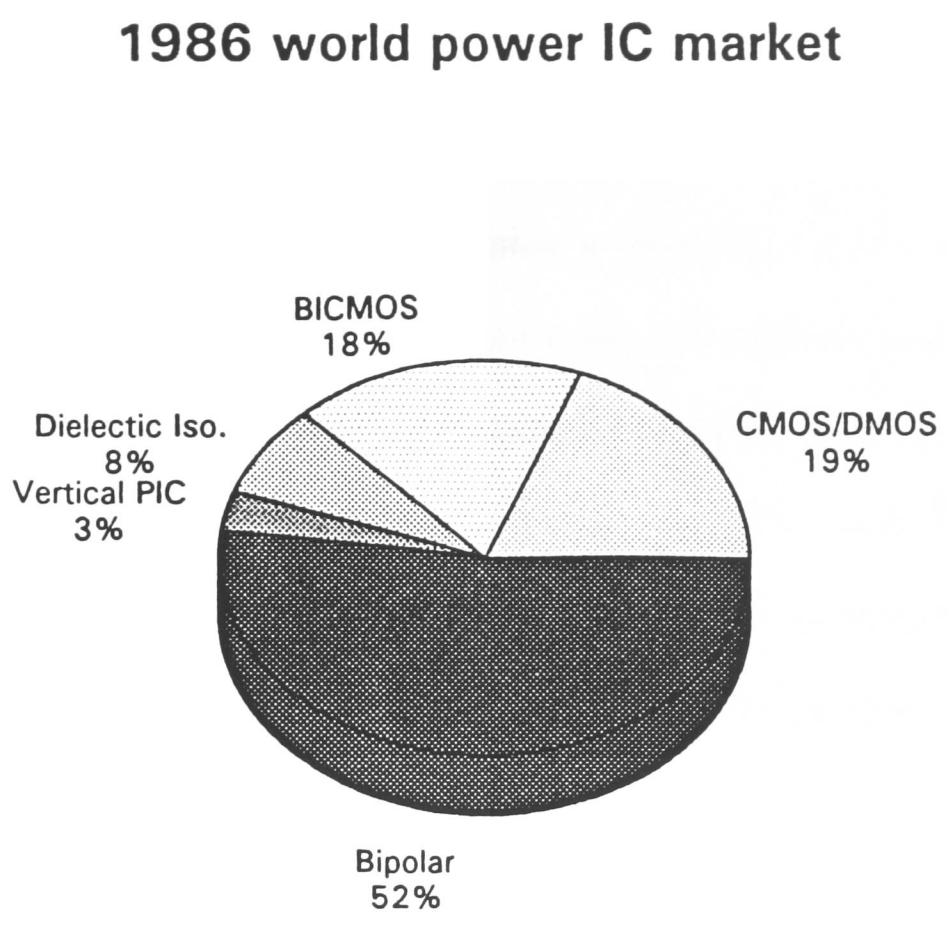


Figure 1.7 The divisions of the 1986 World Power IC Market by technology type.

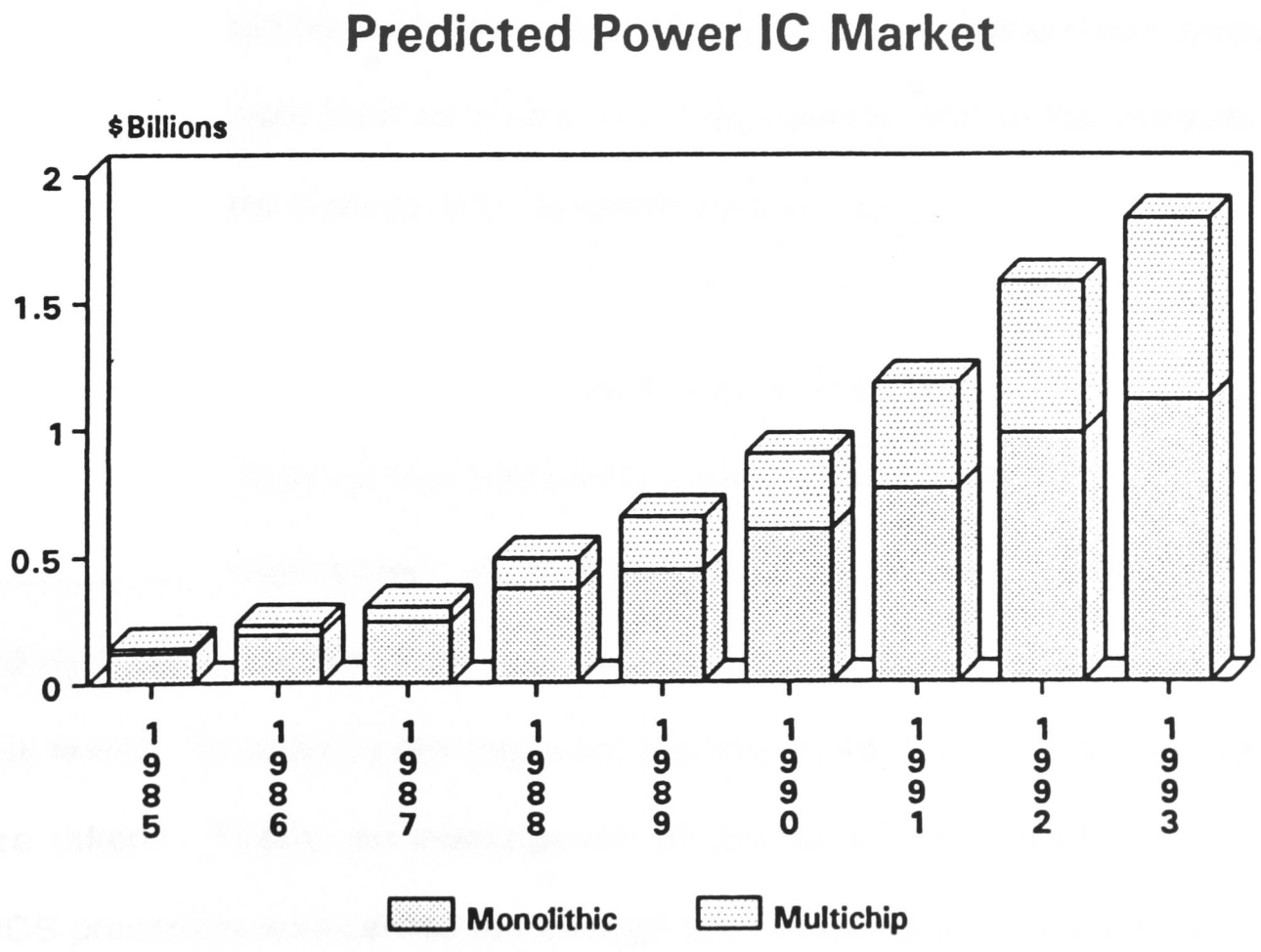


Figure 1.8 The Predicted increase of the World Power IC market from 1985 to 1993.

The type of smart power solution depends heavily on the system complexity and budget. It is not very simple to integrate power and logic transistors on the same chip. Power MOS transistors have a different device structure to logic transistors, and therefore require different mask and process steps. The simplest solution to the smart power problem is to take a logic process and a power process and to add them to create a smart power process. The result is a costly process with many mask levels. For military applications where performance is paramount and all transistors need to be optimised, this type of solution is possible. For commercial applications such as automotive, it is often necessary to slightly compromise performance in order to meet a realistic market price. To do this some parts of the high and low voltage elements are formed with the same process steps. However, current solutions are still too expensive to be used for small volume runs, where process complexity and design time need to be minimised. If a process could be realised that had power transistors created with a minimum of additional processing steps to a standard logic process, then a niche in the market could be exploited. By realising a simple building block, design route it would be possible both to create new applications, and to tap markets usually served by discrete devices with separate control logic.

The approach taken in this work was to start with a logic process and to investigate how high voltage transistors could be incorporated. To do this, it was necessary to design power transistors that could use the maximum number of existing CMOS process steps, and that needed the minimum number of additional mask levels. In order to develop such a smart power solution the following steps were taken. Firstly, an investigation of the fundamental limits of low voltage CMOS processes was carried out through both measurement and simulation. Then

new power transistor architectures were developed and integrated into the existing process without compromising the logic devices.

This approach yielded three different , but closely related smart power solutions. The first of these, based on a CMOS process yielded 40 volt complementary transistors. These high voltage transistors were obtained by modifying source/drain and channel doping concentrations. The basic transistor structure was left intact. In order to further increase the safe operating voltage, it was necessary to modify the transistor structure. This was done by separating the high voltage region from the gate controlled channel. This can be thought of simply, as the introduction of a resistor in series with the low voltage transistor. In the second smart power solution, it was found that using the same mask levels as the low voltage CMOS process, with the appropriate process modifications, that 120 Volt n-channel power transistors could be realised. In the third variant, by adding an epitaxial layer, it was found that both 120 Volt n-channel and -100 Volt p-channel transistors could be realised with the same number of masks as the original low voltage process.

These developments lead to a smart power solution that could be used in various scenarios, but the major applications of these power ranges are automotive, telecommunications and computer peripherals. In addition to the these smart power applications, these results show that it is possible to adapt existing low voltage circuits for more harsh environments. For example, by adding active protection using the power transistors developed here, it is possible to adapt existing memory products for automotive applications.

References

[1] Electronic Industries Associates, Electronic Market Data Book 1987, 1987, Washington D.C.

[2] "World Market Forecast for 1991", Electronics, vol 64, 1991.

[3] "Ten year Worldwide Forecast for Electronic Equipment and Components", Electronic Business, p.92, February 1981.

[4] W.Shockley, "The Path to the Conception of the Junction Transistor", IEEE Transaction on Electron Devices, vol.ED-23, pp.597-620, 1976.

[5] C.-T.Sah, R.N.Noyce and W.Shockley, "Carrier Generation and Recombination in P-N Junctions and P-N Junction Characteristics", Proceedings of the IRE, vol.45, pp.1228-1253, 1957.

[6] Frost and Sullivan, 1986 US Market for Power Semiconductors.

[7] BIS-Mackintosh, "Power Devices - Product Focus", Electronic Engineering, vol.July, pp.48-68, 1987.

[8] P.Pistorio, "Interview", Electronic Engineering, vol.Sept., pp.32-43, 1989.

CHAPTER TWO : BIPOLAR DEVICES

2.1 Introduction

In this work on smart power, it was initially not clear which is the best technology to use ; bipolar, MOS or a hybrid of the two. In order to make the choice it is necessary to consider the merits of both technologies. On a more basic level, as both technologies have pn junctions incorporated in their structures, it is necessary to understand the limitations of this simple semiconductor device before passing on to transistors.

In order to choose between bipolar and MOS transistors for smart power applications, it is necessary to understand the strengths and weaknesses of both types. Another reason for studying both technologies is the fact that sometimes parasitics can dominate circuit performance. One example of this is the bipolar parasitics involved in latchup of a CMOS inverter. This thesis follows the historical evolution, therefore as bipolar transistors were realised in silicon first, they shall be considered first in this chapter before passing to MOS transistors in chapter three. Then the choice of technology used in this work will be discussed.

The first transfer resistor (transistor)was developed by Shockley et al., at Bell Labs, in 1947 [1]. These fragile bipolar transistors first described by Shockley in the classic paper of 1949, gave birth to the semiconductor industry [2]. Various developments, including diffusion [3], epitaxial-growth [4] and ion-implantation [5,6], have lead to the advanced structures of today.

The alloyed construction technique may have been superceded by diffused, implanted and epitaxial structures, but the original operational theory as described in the next section is still valid. This section also documents the extensions to the basic theory that model the secondary effects not explained by the original equations.

As this work is interested in high voltage limitations, there follows a description of the mechanisms of voltage breakdown. Firstly pn diodes are discussed before progressing to bipolar transistors. Barrier tunnelling, impact ionisation and thermal destruction are treated. It is noted that avalanche breakdown is the dominant mechanism in the over-voltage breakdown of bipolar transistors. The effects of corner rounding, punchthrough and reach-through are discussed. These effects will be seen later to be relevant also for MOS transistor structures. The section concludes by describing the phenomenon of secondary breakdown, an important factor in power bipolar transistors.

Then the various developments that have resulted in power bipolar transistors, both discrete and integrated are treated. The various strengths and weaknesses of different bipolar structures are then listed.

2.2 Low Voltage Bipolars

In this section we shall consider the basic dc characteristics of npn and pnp bipolar transistors. The two different types of transistor are shown in figure 2.1. The arrows indicate the direction of (conventional) current flow, with the emitter junction forward biased and the collector junction reverse biased, i.e. normal operating conditions.

If it is assumed that the current-voltage relationships of the emitter and collector junction are given by the ideal diode equations [2],then it can be shown that the dc emitter and collector currents can be given by equations 2.1 and 2.2 , where the constants, a_{11} a_{12} a_{21} a_{22} , depend on junction dopings and physical dimensions [7].

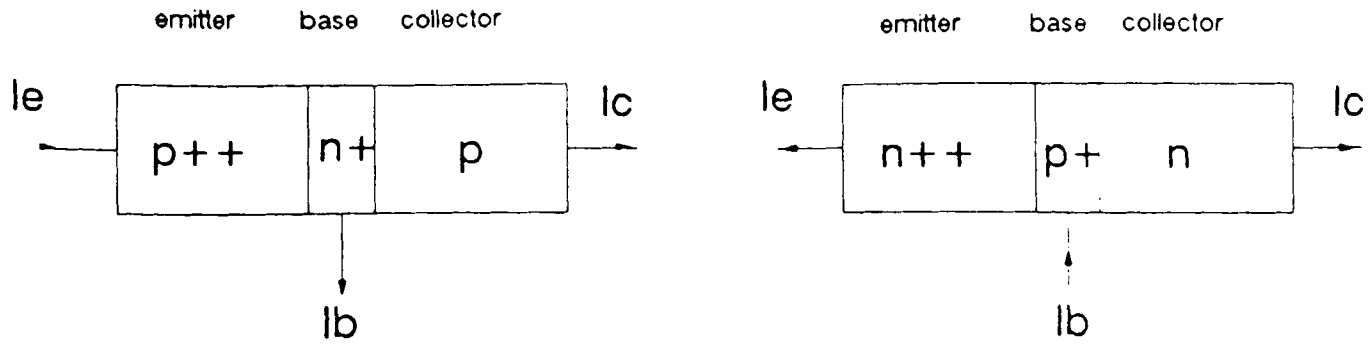
$$I_E = a_{11}[exp\left(\frac{qV_{EB}}{kT}\right)-1]+a_{12}[\exp\left(\frac{qV_{CB}}{kT}\right)-1)] \qquad 2.1$$

$$I_C = a_{21}[exp\left(\frac{qV_{EB}}{kT}\right)-1]+a_{22}\exp\left[\left(\frac{qV_{CB}}{kT}\right)-1)\right] \qquad 2.2$$

The difference between these two currents is small and appears as the base current.

$$I_B = I_E - I_C \qquad 2.3$$

The number of impurities per unit area in the base, the Gummel number is given by equation 2.4, and is typically 10^{12} to 10^{13} cm^{-2}.



(a) pnp (b) npn

Figure 2.1 Schematic designs showing conventional current flow for (a)pnp and (b)npn bipolar transistors.

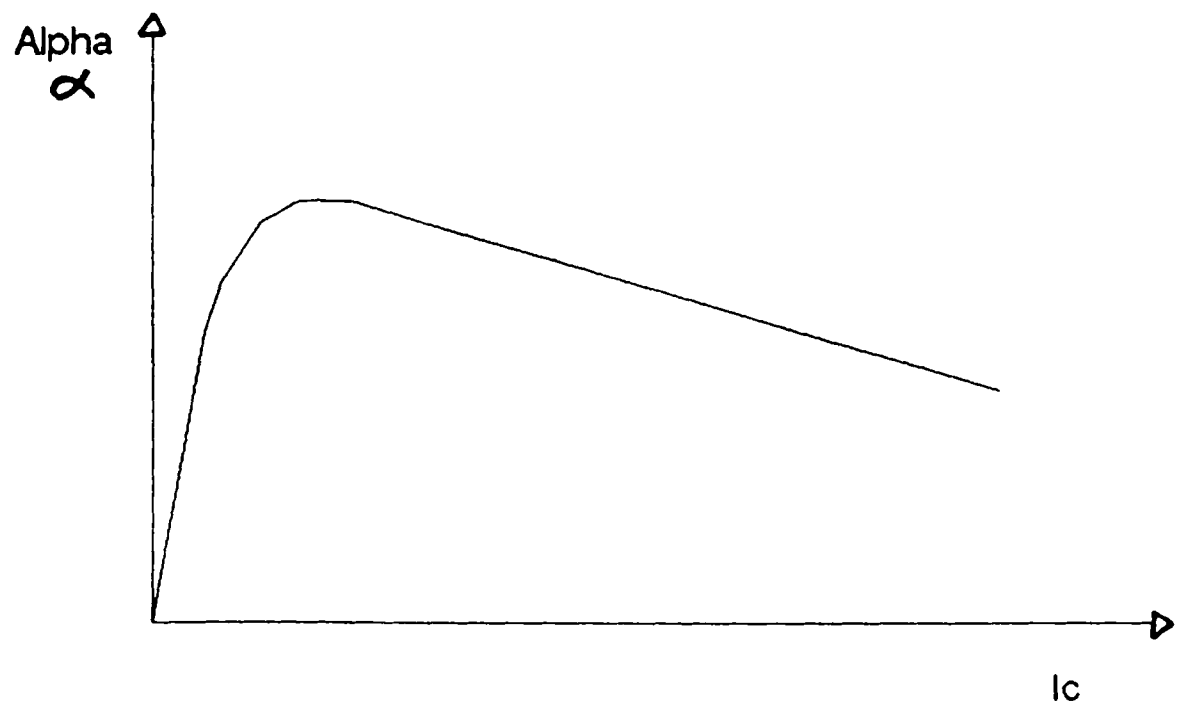


Figure 2.2 The variation of common-base current gain with collector current.

$$Q_B = \int N(x)dx \qquad 2.4$$

When a pnp transistor is biased into its active region the emitter current consists of two parts ; the hole component injected into the base, and the electron component injected from the base into the emitter regions. Similarly , the collector current also consists of two components. The static common-base current gain is defined as the rate of change of collector current with emitter current [8].

$$\alpha = \frac{\partial I_C}{\partial I_E} = \gamma \alpha_T M \qquad 2.5$$

Where γ is the emitter efficiency, α_T is the base transport factor and M is the collector multiplication factor.

The static common-emitter current gain is given by equation 2.6.

$$\beta = \frac{\partial I_C}{\partial I_B} \qquad 2.6$$

These are related by equation 2.7.

$$\beta = \frac{\alpha}{1-\alpha} \qquad 2.7$$

The value of α in well-designed transistors is close to unity, therefore ß is generally much larger than 1. For example, if $\alpha = 0.99$ then ß = 99.

The common-emitter current gain generally varies with collector current as in figure 2.2. At low collector currents there is a drop in the value of current gain. This is caused by the recombination of carriers at the surface, the recombination of carriers in the emitter-base space-charge layer [9], and the formation of emitter-base surface channels. These three effects add to the base current and reduce the gain. At high injection levels, the injection of minority carriers into the base region is significant with respect to the majority carrier concentration. Since the space-charge neutrality is maintained in the base, the total majority carrier concentration is increased by the same amount as the total minority carrier concentration. This effect of the excess majority carriers has been calculated by Webster [10], who showed that at high levels the collector current asymptotes to the value given in equation 2.8.

$$I_{B\,(high\,-level)} \propto \exp\left(\frac{qV_{BE}}{2kT}\right) \qquad 2.8$$

The current gain is also heavily dependent on the emitter doping concentration N_E. To improve gain, the emitter should be much more heavily doped than the base. However, when the emitter doping becomes very high, bandgap narrowing and Auger recombination cause reductions in the gain.

Auger recombination is the direct recombination between an electron and a hole , accompanied by the transfer of energy [11] to another free hole. This process involving two holes and one electron occurs when electrons are injected into a heavily doped p+ region, as in the emitter of a p+np transistor. Auger recombination is the inverse process of avalanche multiplication.

In a bipolar transistor with a lightly doped epitaxial collector, there is a high-current fall-off in the gain of the transistor as a direct result of a spreading of the neutral base layer into the collector region of the device. This high-field redistribution phenomenon is referred to as the Kirk effect [12] , which increases the effective Gummel number causing a reduction in ß.

At high V_{CB} values, the extension of the depletion region into the base must be considered. Depleted-base charge and thickness must be subtracted from the low-voltage values in order to obtain the effective base charge Q_B and base width W_B, (both are functions of V_{CE}). It has been shown [12] that these effective values are given by equations 2.9 and 2.10.

$$Q_B'(V_{CE}) = Q_B - \left(\frac{2\epsilon N_D V_{CE}}{q}\right)^{0.5} \qquad 2.9$$

$$W_B'(V_{CE}) = W_B - x_{dp} \qquad 2.10$$

If these effective values are used then the expressions used for current gain are still valid. This so-called Early effect is observed as an increase of ß with voltage, which is more consistent when the total base charge Q_B is small. It should be noted that when punch-through occurs Q_B becomes zero, current gain becomes infinite and there is an unlimited current flow.

The two most used circuit configurations of bipolar transistors are the common-base and the common-emitter. If the transistor is connected in the common-base configuration as in figure 2.3 then the output characteristics are as in figure 2.4. The collector current is practically equal to the emitter current and virtually independent of V_{CB}. The collector current remains practically constant, even down to zero voltage where the excess holes are still extracted by the collector. To reduce the collector current to zero, a small forward voltage must be applied to the collector, which sufficiently increases the hole density at the base-collector edge to make it equal to that of the emitter-base edge. As V_{CB} increases to the value BV_{CBO}, the collector current starts to increase rapidly. This is usually due to the avalanche breakdown of the collector-base junction, but if the base is relatively lightly doped or very narrow, the breakdown is caused by the punch-through effect.

In the common-emitter configuration of figure 2.5, the output characteristics are as in figure 2.6. The collector-emitter voltage is developed across both junctions, and is given by equation 2.11.

$$V_{CE} = V_{CB} + V_{BE} \qquad 2.11$$

In the active region of operation, where V_{CE} exceeds about 0.2 V, the emitter junction has a small forward bias, which is almost constant to maintain the base current constant, and the collector has a much larger reverse bias. In this region I_C rises slowly with V_{CE}, owing to the fall in base width as V_{CE} rises and the corresponding fall in α. This effect is much more noticeable than in the common-base configuration, due to the1-α factor. The increase in I_C due to avalanche breakdown becomes apparent at a lower collector voltage as the change in α is magnified in its effect on I_C.

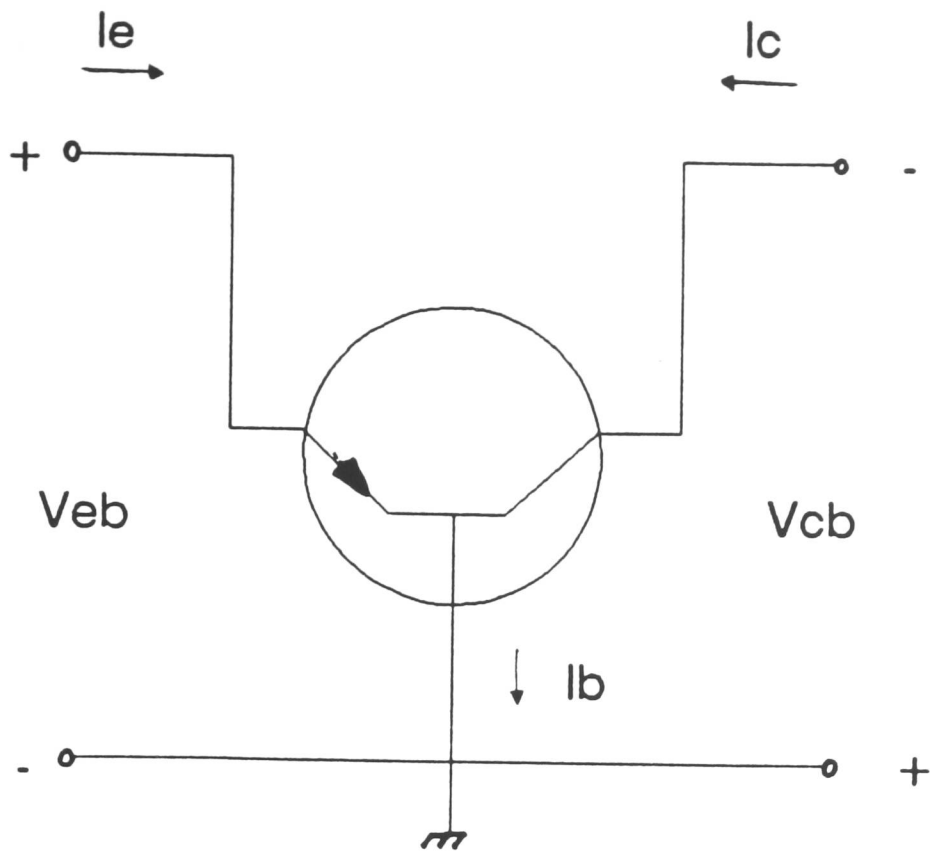


Figure 2.3 The terminal connections of a pnp transistor in the common-base configuration.

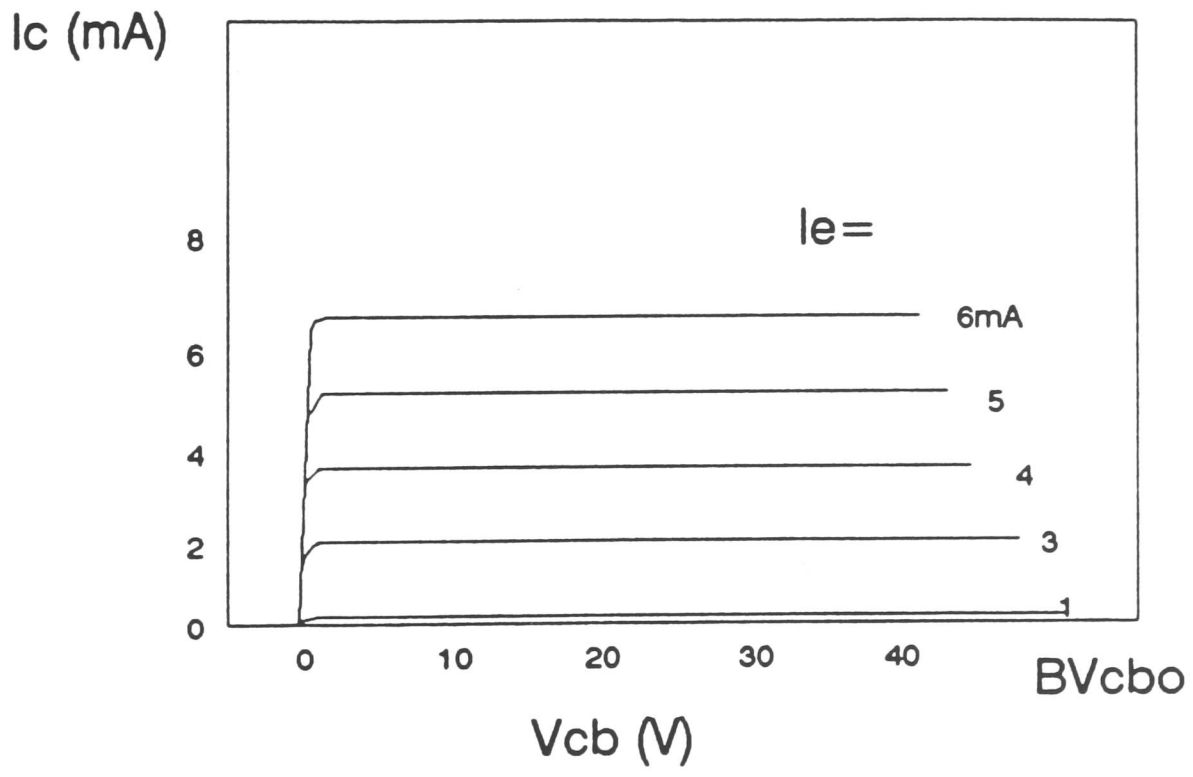


Figure 2.4 The output characteristics of a pnp transistor in the common-base configuration for various values of emitter current.

For values of V_{CE} below about 0.2 V, I_C falls very rapidly as V_{CE} is reduced, and the transistor is working well within the saturation region. As V_{CE} is reduced, the collector-base diode becomes forward biased and current flows in the reverse direction to normal operation resulting in a rapidly decreasing collector current.

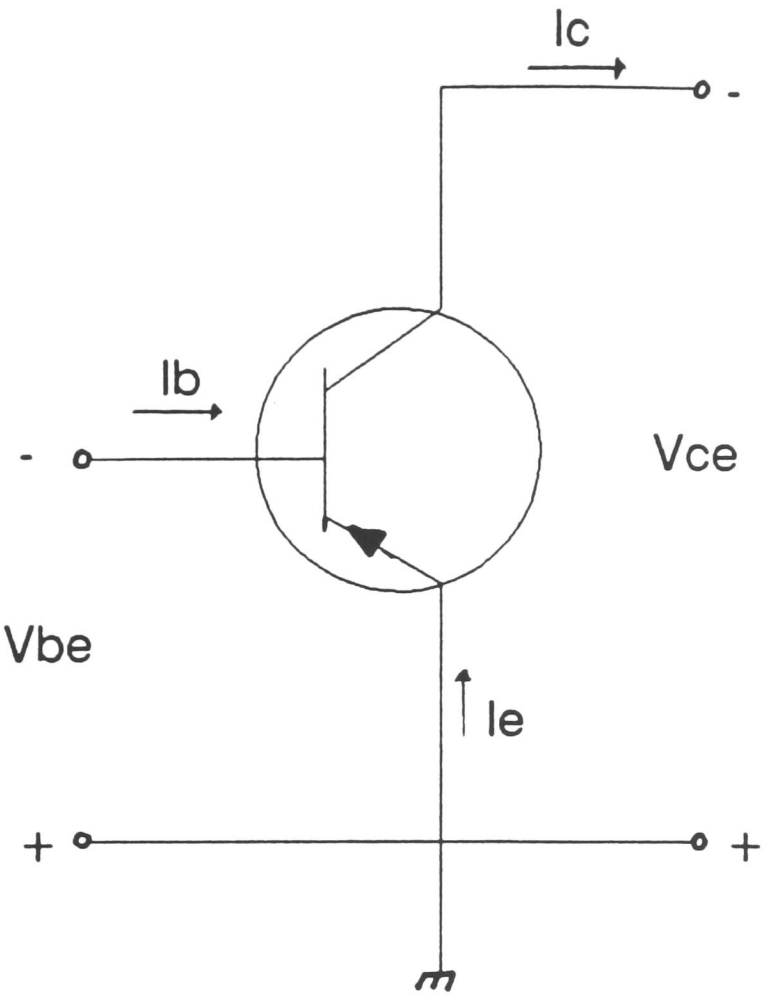


Figure 2.5 The terminal connections of a pnp transistor in the common-emitter configuration.

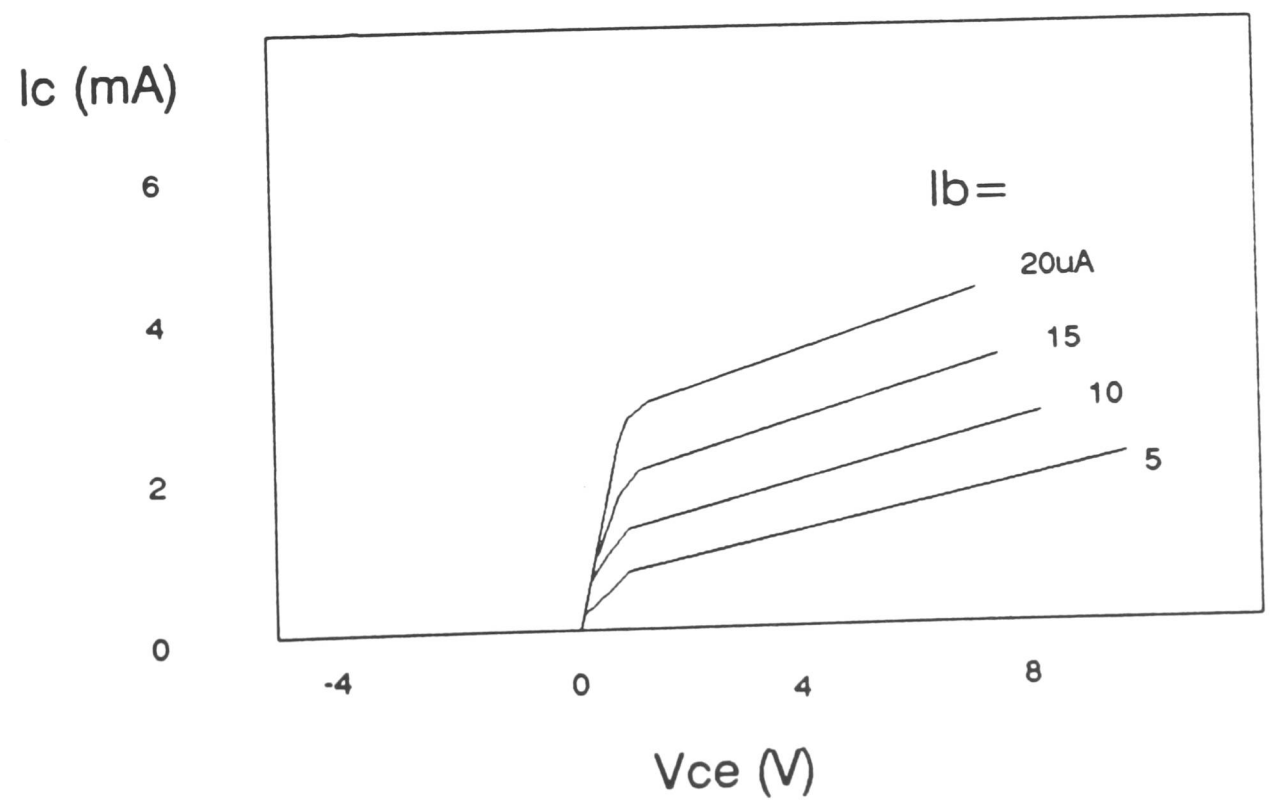


Figure 2.6 The output characteristics of a pnp transistor in the common-emitter configuration for various values of base current.

2.3 Breakdown Mechanisms

Bipolar power transistors must be designed to withstand high-voltage, current and power. Wide operating areas in both the forward and reverse base-driving conditions and good switching speed are also often required. One important phenomenon that must be considered is avalanche breakdown. Before passing onto transistor structures, consider first the simple pn junction.

In practice the reverse bias current remains constant in a pn junction only until a voltage V_b is reached, the reverse breakdown voltage. At this voltage the high electric field causes the junction to conduct a very large current [13]. Breakdown can occur due to one of three principal mechanisms, but combinations of two or more of these can occur. These mechanisms are as follows.

1. Barrier Tunnelling

2. Impact Ionisation

3. Thermal Breakdown

The first mechanism is that of barrier tunnelling. It can be shown [14] that the tunnelling current density for a pn junction is given by equation 2.12.

$$J = (2mq^3)^{1/2} \mathrm{E} \frac{V}{4\pi h^2 Eg^{1/2}} \exp(-4(2m)^{1/2} \frac{Eg^{3/2}}{3q\,\mathrm{E}h}) \qquad 2.12$$

where m is the effective mass

E is the electric field at the junction

V is the applied voltage

Eg is the bandgap.

When the electric field approaches 10 Volt per cm, significant current begins to flow by means of the band-to-band tunnelling process. The mechanism of breakdown for Si and Ge n + p + junctions with breakdown voltages less than about 4Eg/q is found to be due to the tunnelling effect. Figure 2.7 shows an example of this effect. The breakdown voltage has a negative temperature coefficient: that is voltage decreases with increasing temperature.

In practice it is found that the most important mechanism in junction breakdown is avalanche multiplication [15,16,17] (or impact ionisation). For one-sided abrupt junctions it can be shown [18] that the breakdown voltage is related to field and doping according to equation 2.13. Similarly equation 2.14 applies to linearly graded junctions.

$$V_b = E_m \frac{W}{2} = \epsilon_s \frac{E_m^{1/2}}{2qN_B} \qquad 2.13$$

$$V_b = 4E^{3/2}/3(2\frac{\epsilon_s}{aq})^{1/2} \qquad 2.14$$

where N_B is the background impurity concentration of the lightly doped side

a is the impurity gradient

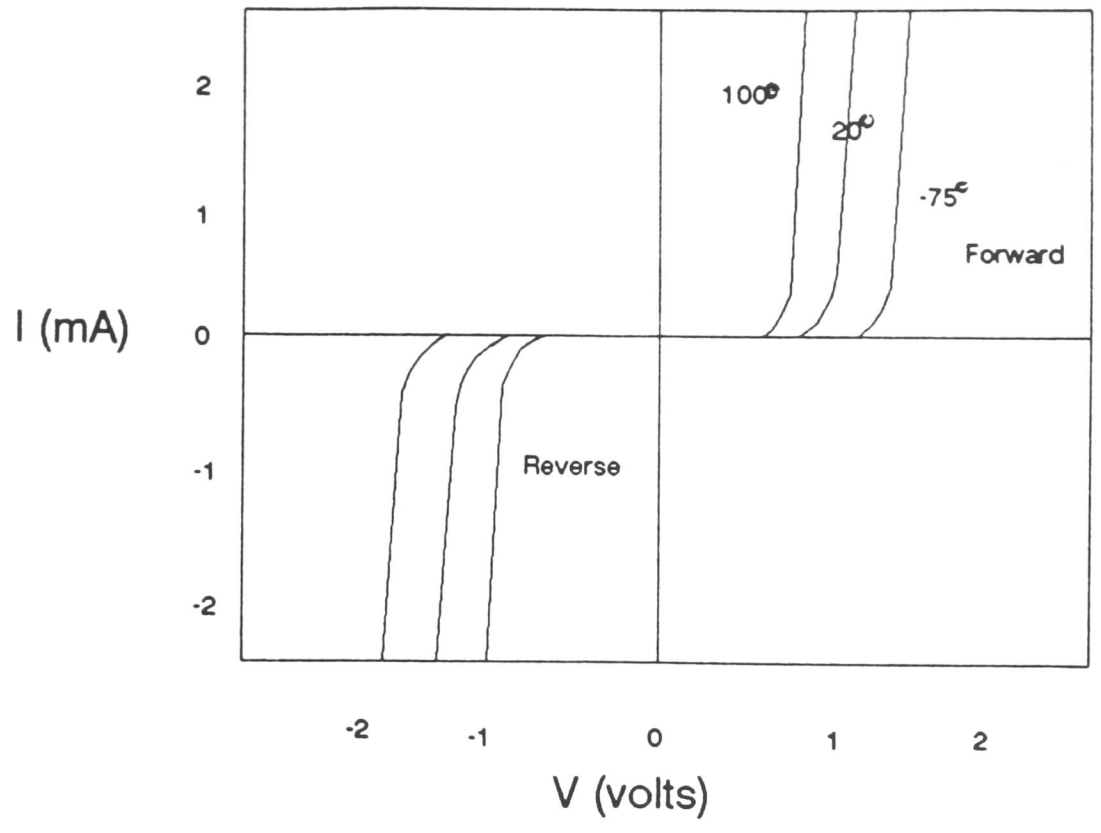


Figure 2.7 The variation of junction breakdown voltage with temperature.

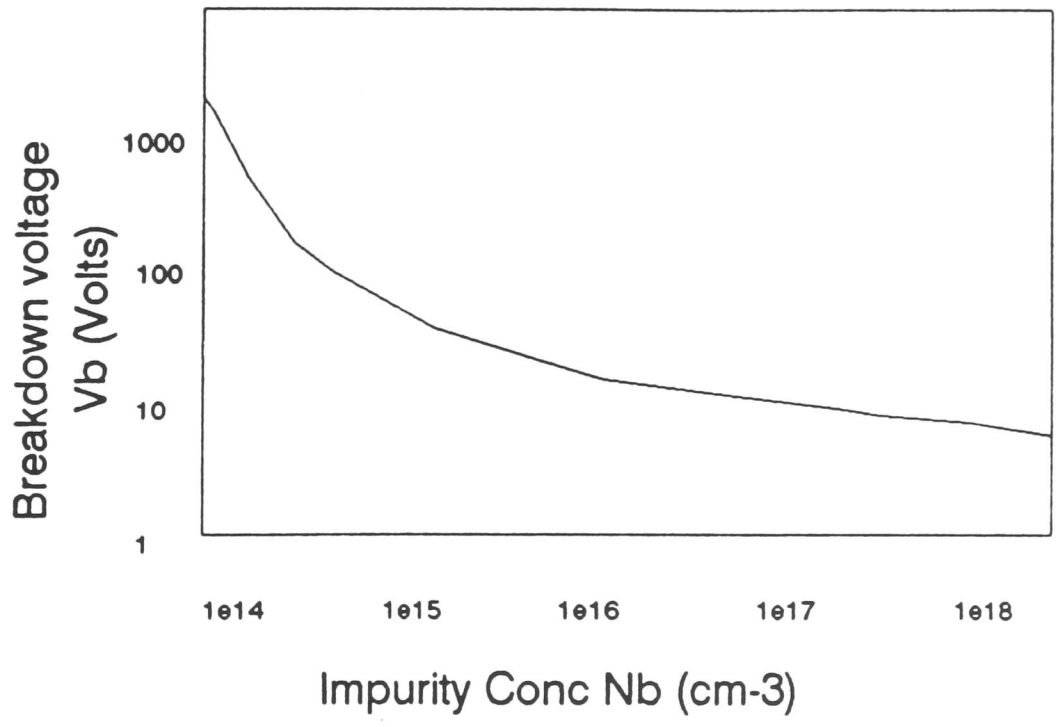


Figure 2.8 The variation of breakdown voltage with background doping concentration.

The variation of calculated breakdown voltage as a function of N_B is shown in figure 2.8, and as a function of a in figure 2.9. This mechanism is prevalent in junctions with breakdown voltages in excess of 6Eg/q.

The calculated values of the maximum field and the depletion-layer width at breakdown are shown in figures 2.10 and 2.11. It can be shown that the maximum field for Si junctions can be expressed as in equation 2.15 [19].

$$E_m = \frac{4x10^5}{1-1/3\log(N_B/10^{16})} V/cm \qquad 2.15$$

where N_B is in cm^{-3}

If the semiconductor layer is not thick enough to support the depletion-layer width at breakdown, the device will be punched through prior to breakdown. The breakdown voltage V_{PT} for the punched through device is given by equation 2.16. This breakdown mode proves to be the limiting factor in low voltage metal-oxide-semiconductor field effect transistors, MOSFETs.

$$\frac{V_{PT}}{V_B} = \frac{W}{W_m(2-W/W_m)} \qquad 2.16$$

where W is the physical width and W_m is the depletion layer width

The variation of avalanche breakdown voltage with temperature and doping concentration is shown in figure 2.12.

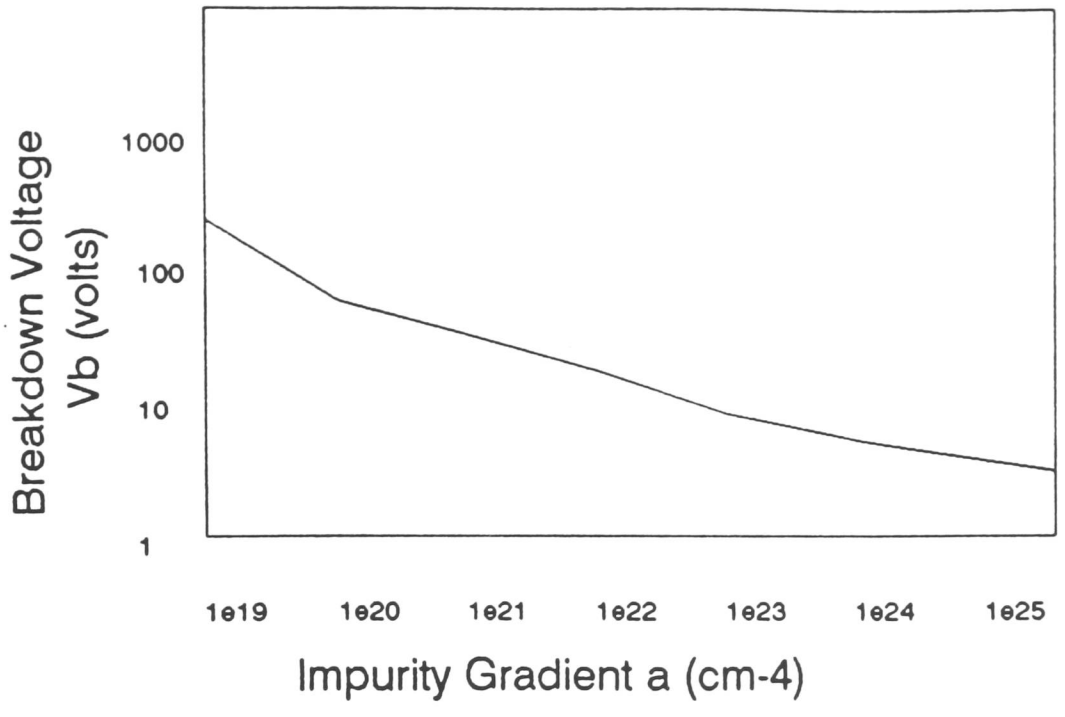


Figure 2.9 The variation of breakdown voltage with background impurity gradient.

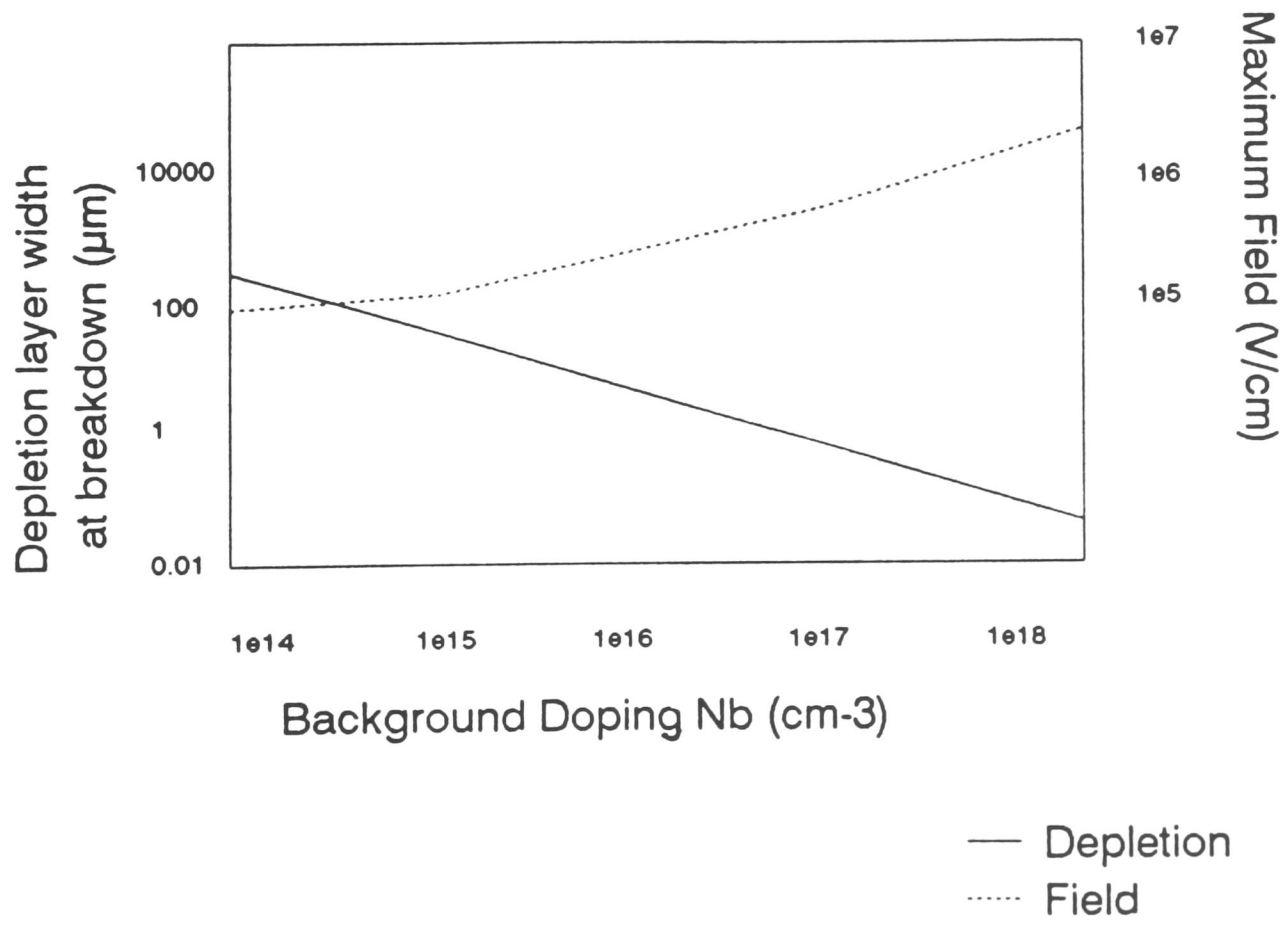


Figure 2.10 The variation of depletion layer width at breakdown and the maximum electrical field with background doping concentration.

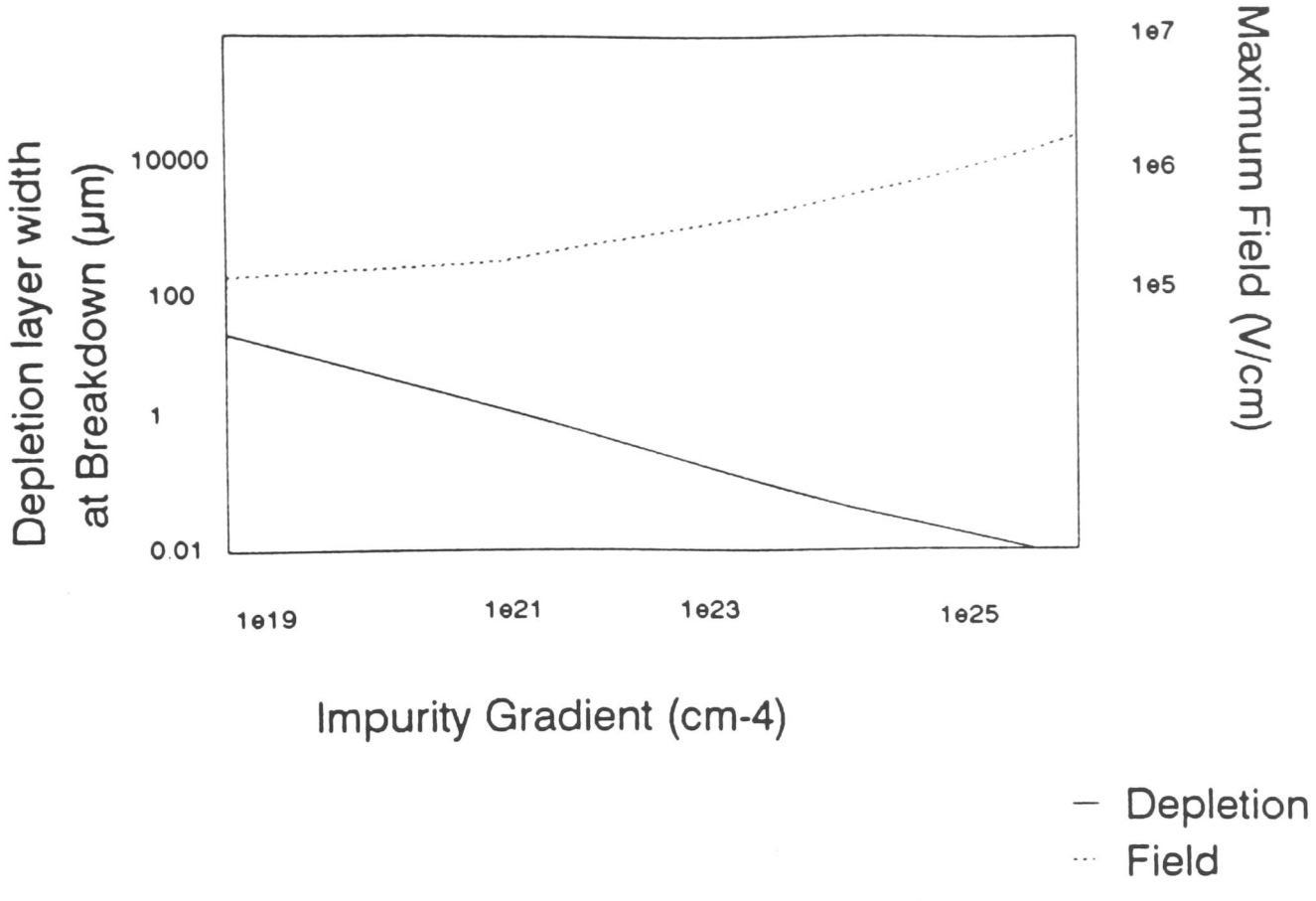


Figure 2.11 The variation of depletion layer width at breakdown and the maximum electrical field with background impurity gradient.

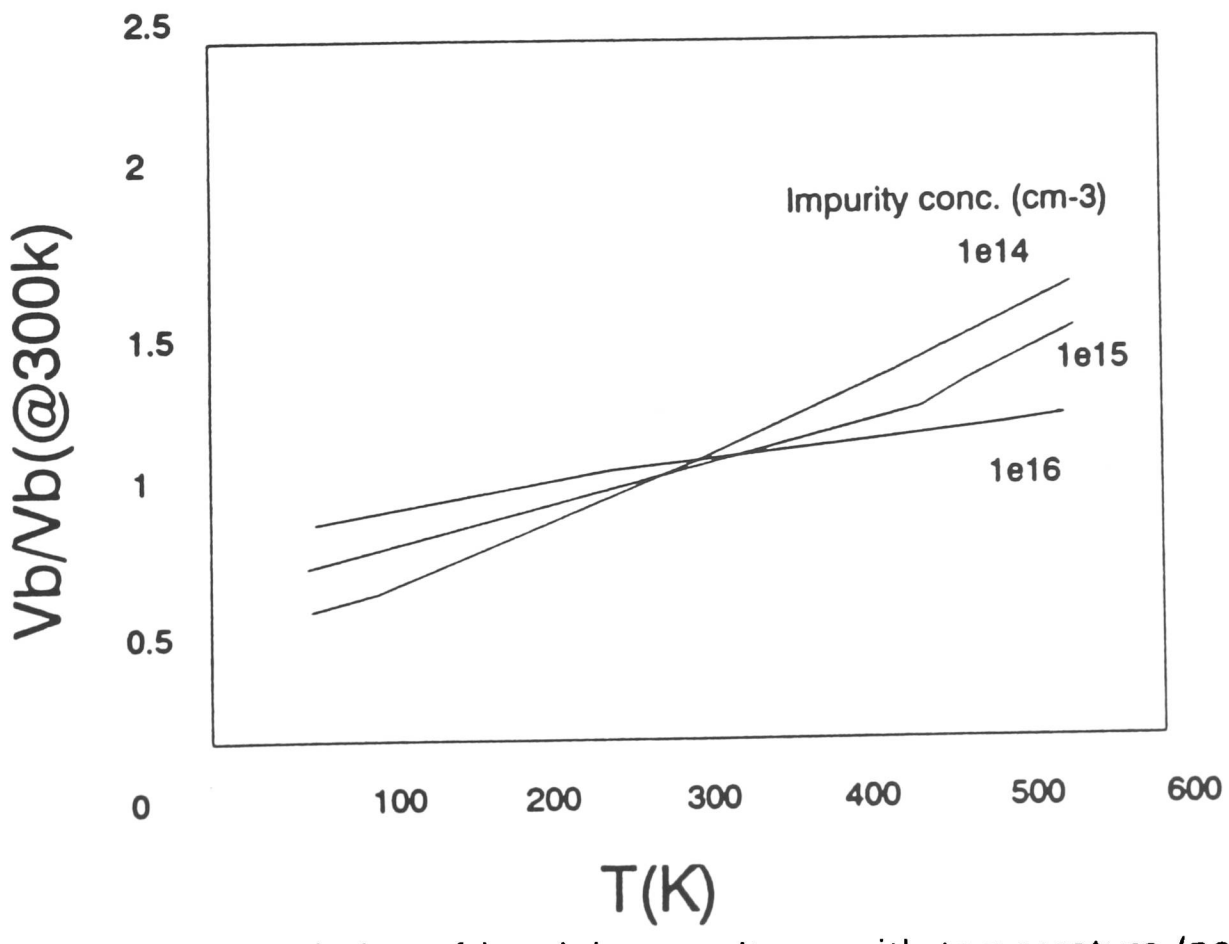


Figure 2.12 The variation of breakdown voltage with temperature (normalised to 300K).

The effect of junction curvature on breakdown voltage is given in figure 2.13. These curved areas have a higher field intensity and the avalanche breakdown voltage is determined by these regions. It can be shown that the cylindrical, V_{cy}, and spherical, V_{sp}, breakdown voltages can be given by equations 2.17 and 2.18 respectively. These "corrected" breakdown voltages are more realistic estimations of breakdown in actual devices. Typically it is junction edges that determine voltage limits not the bulk breakdown value [20].

$$V_{cy} = V_B [0.5(\eta^2 + \eta^{6/7}) ln\,(1 + 2\eta^{-8/7}) - \eta^{6/7}] \qquad 2.17$$

$$V_{sp} = V_B [\eta^2 + 2.14\eta^{6/7} - (\eta^3 + 3\eta^{13/7})^{2/3}] \qquad 2.18$$

where V_B is the bulk breakdown value

and η is the radius of curvature divided by the depletion layer width.

It has been shown [21] that there are two different kinds of avalanche breakdown. The first is characterised by a very small localised area of breakdown voltage, the so-called microplasma. Microplasmas have been shown to be caused by imperfections in the crystal lattice. The second type of breakdown, which occurs in the absence of imperfections is distributed over the whole area.

Shockley found that the voltage dependence of the multiplication at voltages approaching the breakdown voltage V_b was given by equation 2.19 [2].

$$\frac{1}{M} = (V_B - V)\frac{n}{V_B} \qquad 2.19$$

where n is a constant.

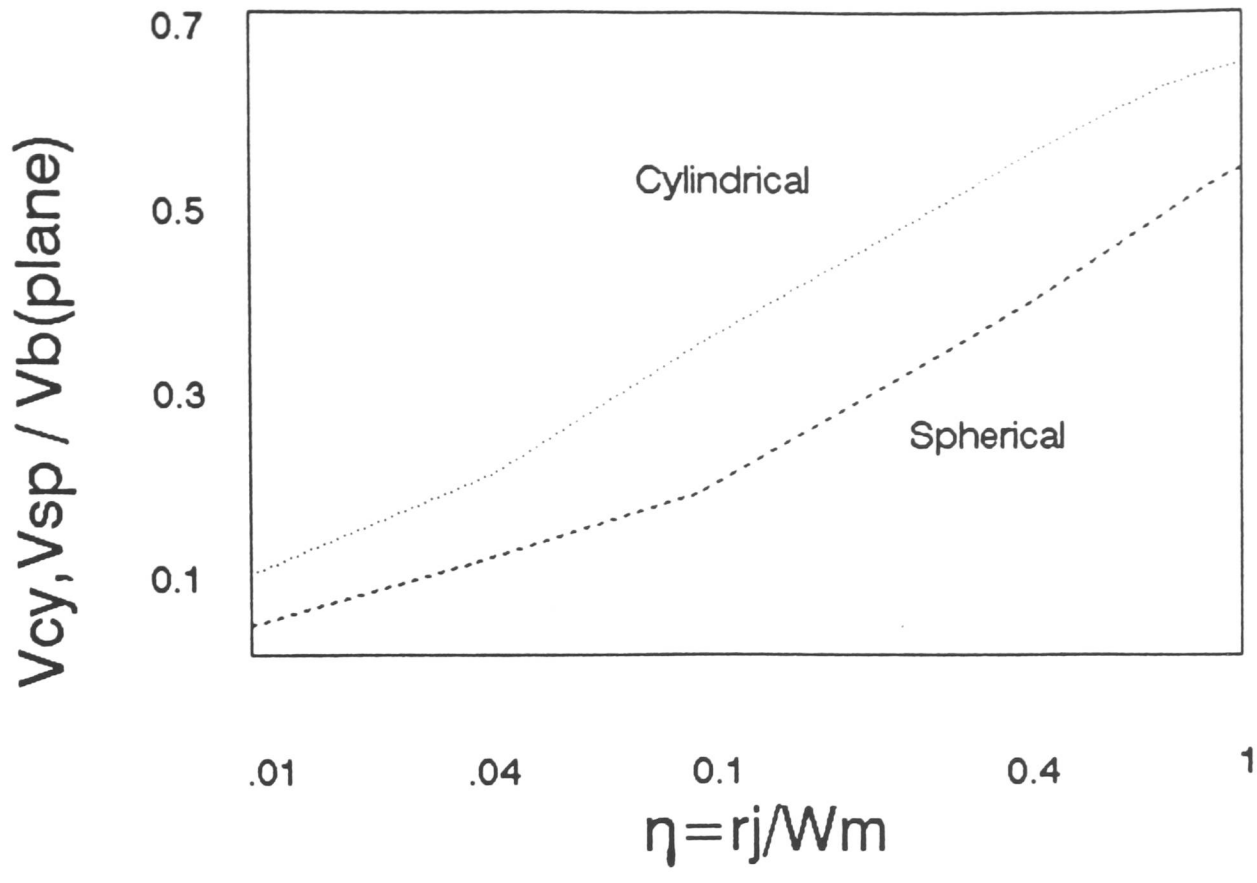


Figure 2.13 The effect of curvature on breakdown voltage for cylindrical and spherical junctions.

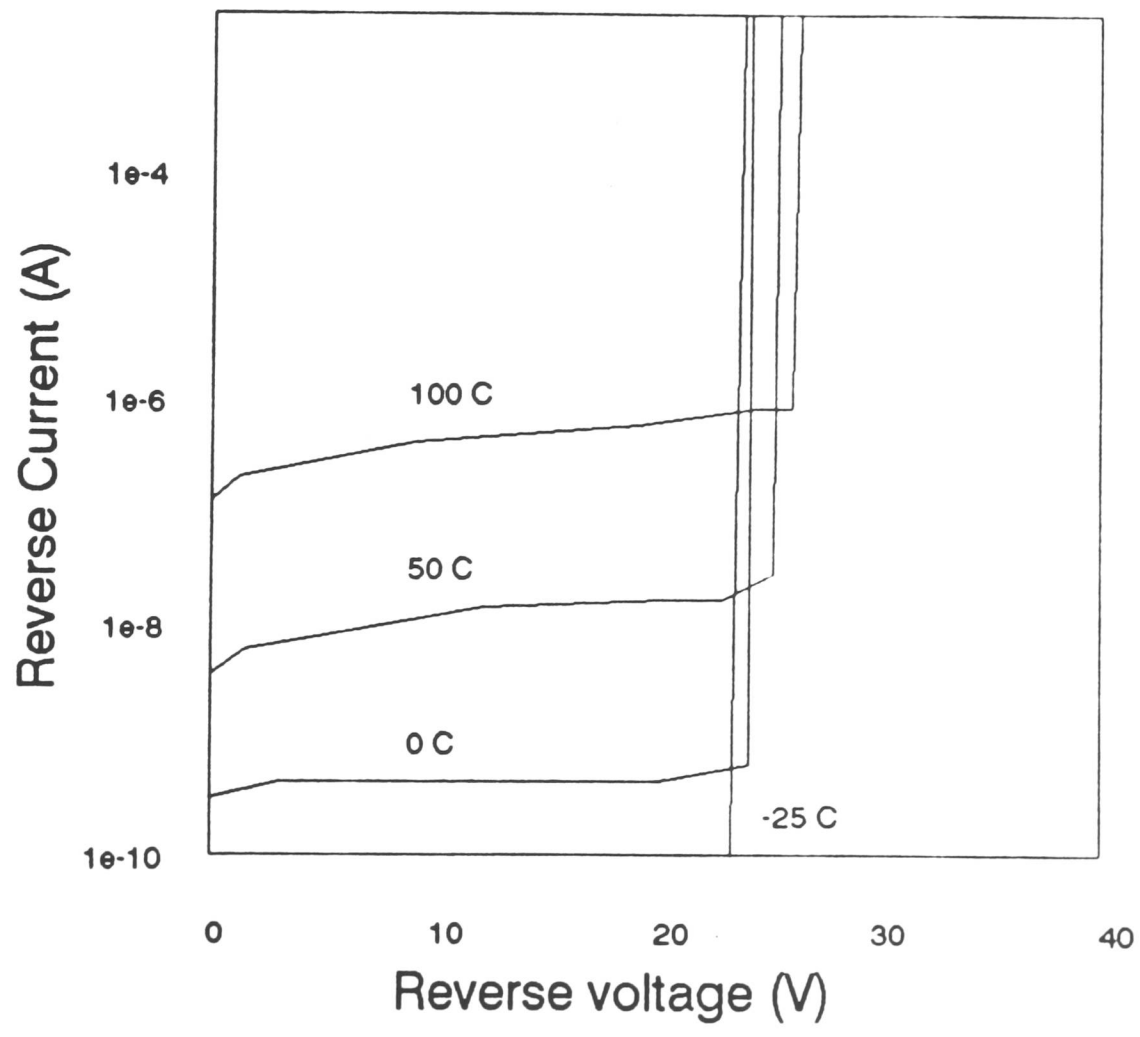


Figure 2.14 The effect of temperature on reverse current and breakdown voltage for a pn junction.

This expression is in very good agreement with experimental data provided the area in which multiplication occurs has a uniform breakdown voltage. If there is an area of significantly lower breakdown voltage i.e. a microplasma, the multiplication law is altered. There is a deviation from the linear 1/M law at high multiplication due to the space-charge effect of the multiplied carriers, and eventually a saturation of the multiplication factor occurs.

Breakdown due to thermal instability is a major effect in semiconductors with relatively small bandgaps (eg Ge), although it often occurs in conjunction with the other breakdown mechanisms. Because of the heat dissipation caused by the reverse current at high reverse voltage, the junction temperature increases. This temperature increase, in turn increases the reverse current in comparison to its value at lower voltages. This effect is shown in Figure 2.14. The reverse current is found to vary as equation 2.20 [22].

$$I_0 \alpha T^{3+\tau/2} \exp \frac{-E_g}{kT} \qquad 2.20$$

where τ is a constant

The heat dissipation at high reverse voltage results in a negative differential resistance and the junction will be destroyed by 'thermal runaway'. This mechanism is important for Ge junctions and even at room temperature devices can be thermally unstable [16,23].

The dominant mechanism of breakdown in bipolar transistors is avalanche breakdown. The first junction we shall consider is the collector to base junction. In a plane unlimited junction the junction breakdown voltage is given by equation 2.14. However in actual cases the dimensions of the p and n-type regions are limited and two phenomena may occur; reach-through and punchthrough.

Reach-through can occur in vertical structures, eg. the npn in figure 2.15. The collector-base junction can be considered one sided (P+ N), as the base doping is much greater than the collector. The collector width, W_C, is the vertical distance between the junction and the N+ buried layer. The base-collector structure is well-modelled as a P+NN+ abrupt diode with the collector width much greater than the depletion width on the collector side of the junction. The electric field at breakdown of the diode modelling the base-collector junction is shown in figure 2.16.

When the width of the n- region of the collector is much greater than the depletion width, E_{cr} is found not to vary appreciably from the unlimited junction value [16]. If the n-region is less than the depletion width, then the breakdown voltage is reduced accordingly.

$$V_{br}' = V_{br}\left[1-\left(\frac{W_C}{x_{dN}}\right)^2\right] \qquad 2.21$$

Punchthrough is a phenomenon that can occur in lateral bipolar transistors such as figure 2.17. It occurs when the depletion region in the base touches the emitter at some value of junction field E_{pt}. An emitter-collector short circuit via the depleted base occurs, allowing an unlimited current flow between collector and emitter. The electric field cannot be raised above E_{pt} and breakdown occurs at a voltage V_{pt} which is lower than V_{br}. In this case the collector-base junction breaks down via the emitter-base junction. This effect depends mainly on the base width and doping.

$$V_{pt} = \frac{\epsilon E_{pt}^2}{2qN_D} \qquad 2.22$$

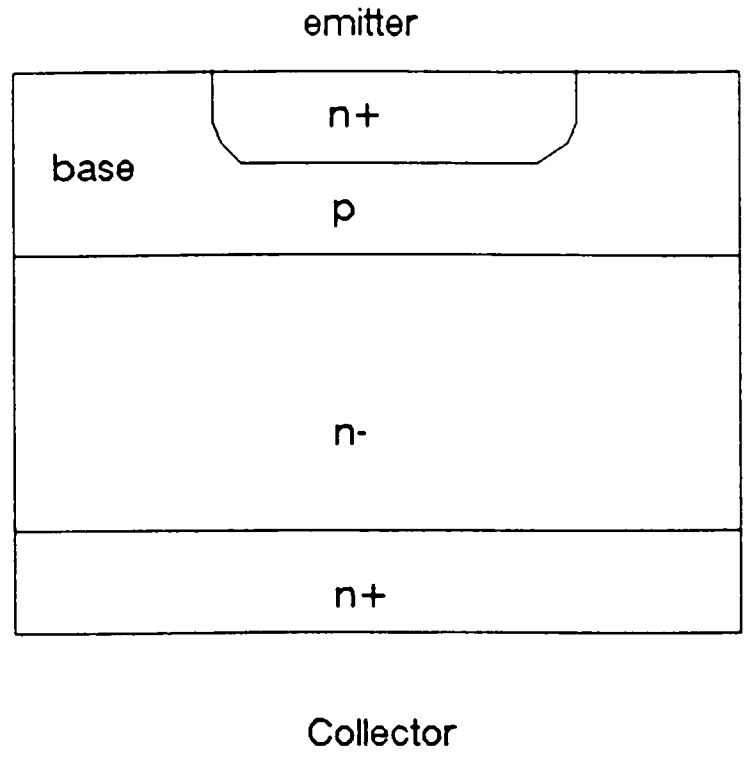


Figure 2.15 A vertical npn transistor structure.

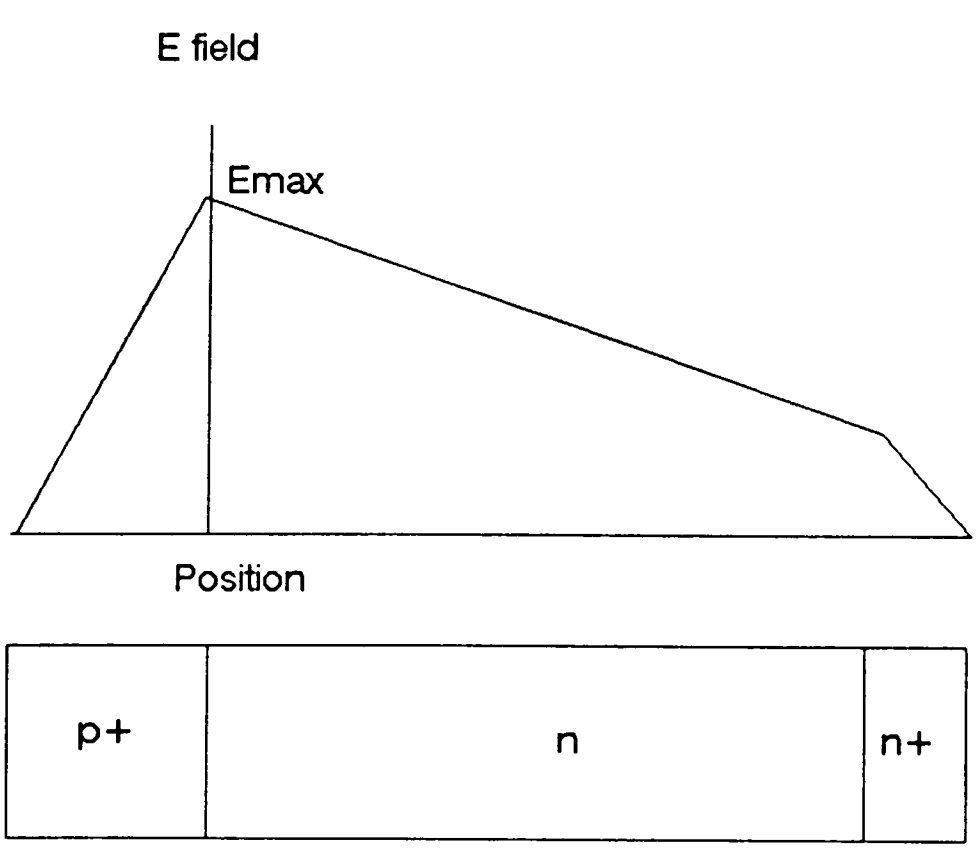


Figure 2.16 A diagram of the electrical field at breakdown for the p+nn+ diode used to model the base-collector junction.

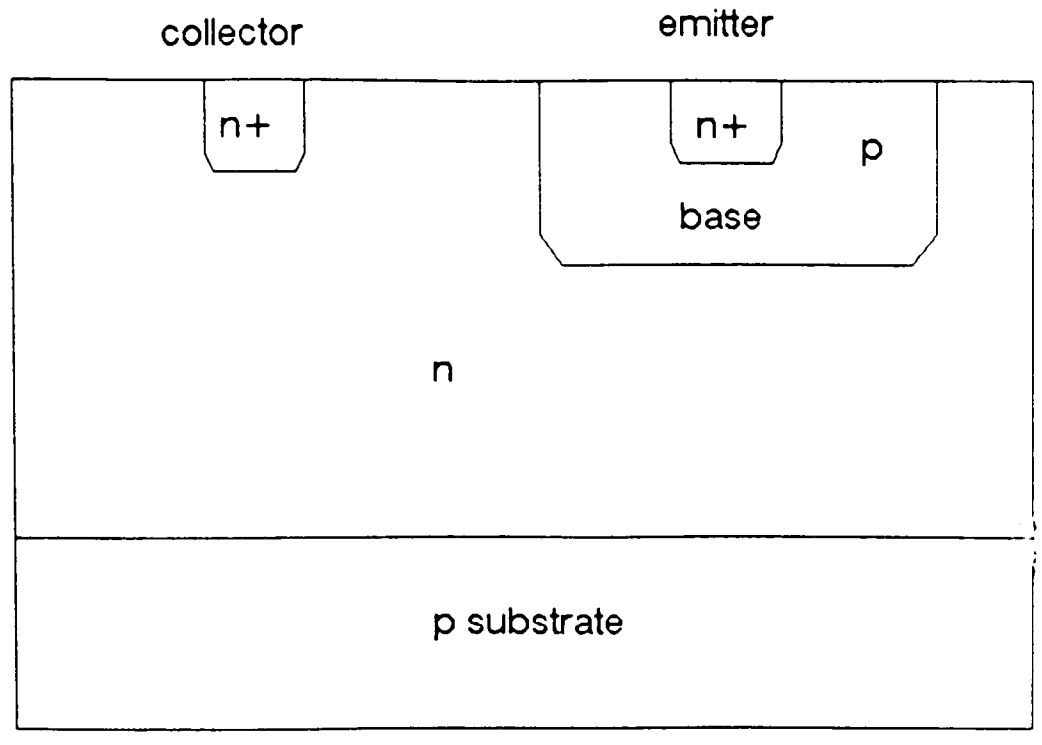


Figure 2.17 A lateral npn transistor structure.

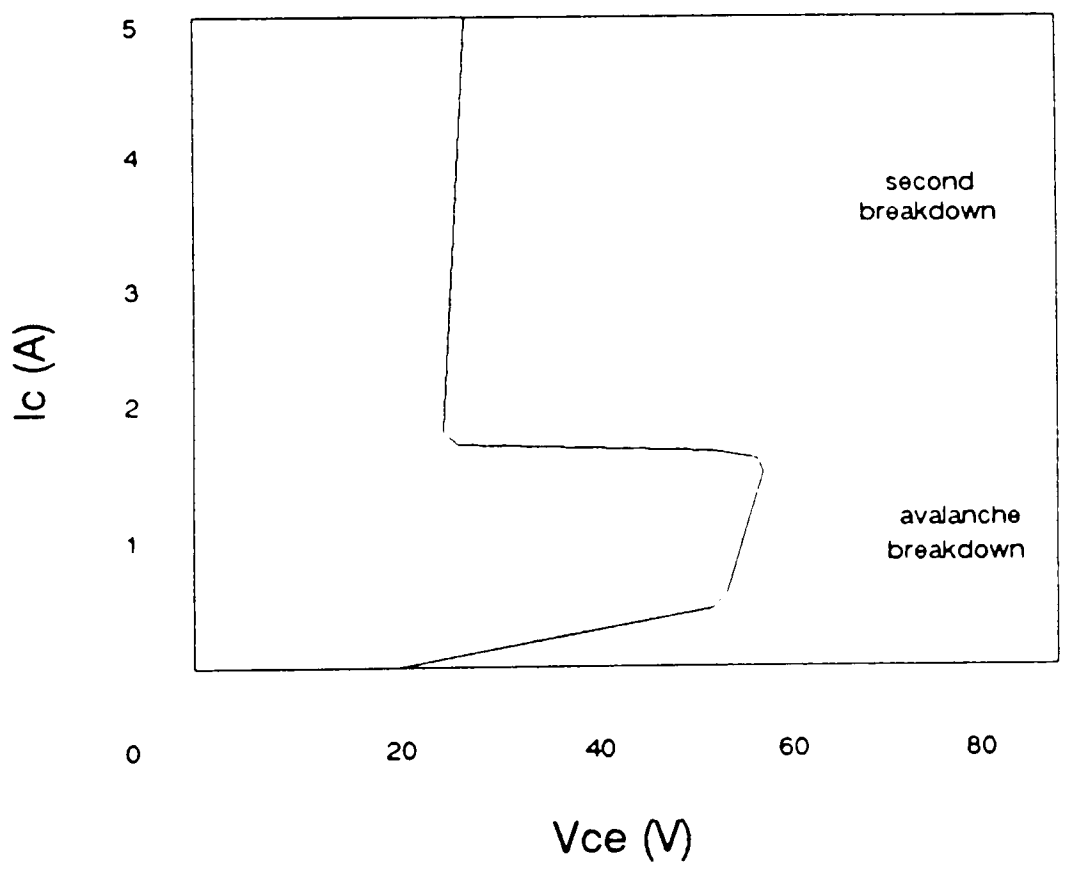


Figure 2.18 The current voltage characteristics of a transistor under second breakdown.

The breakdown of the collector-base junction can be either simple planar breakdown or the reach-through phenomenon.

Maximum voltage, current and power do not completely define the safe operating area (SOA) of bipolar transistors. Second breakdown (SB) is known to occur well within the absolute maximum ratings and the SOA must be reduced accordingly. Figure 2.18 shows the general features of the Ic vs Vce characteristics of a transistor under second breakdown conditions. The avalanche (first) breakdown occurs when the applied emitter-collector voltage reaches BV_{CEO}. As the voltage increases further, second breakdown occurs. There are four stages : the first stage leads to instability at the breakdown voltage; the second is the switching from the high to low voltage region; the third phase is the move to the low-voltage high-current range; the fourth stage leads to destruction.

The initiation of instability is mainly caused by the temperature effect. When a pulse with a given power $P = I_C\ BV_{CEO}$ is applied to a transistor, a time delay follows. The device is then triggered into the second breakdown condition. After instability the voltage collapses across the junction. During this second stage of the breakdown, the resistance of the breakdown spot becomes drastically reduced. In the third low-voltage stage, the semiconductor is at a high temperature and degenerates in the vicinity of the breakdown spot. As the current continues to increase, the breakdown spot melts, resulting in the fourth stage of destruction. To safeguard a transistor from permanent damage a reduced safe operating area must be specified.

2.4 High-Voltage Bipolar Transistors

Power bipolar transistors are very similar to low voltage transistors in terms of device structure. As this work is concerned with smart power applications, this section will describe power transistors that can be easily integrated with low voltage transistors.

An important parameter in power transistor design is the maximum voltage that can be imposed on the collector with respect to the emitter. The resulting common-emitter breakdown is the lowest expected breakdown configuration. Breakdown results from a current gain multiplication of the collector-base junction leakage. This collector avalanche voltage can be limited by several effects including junction curvature, as described in the previous section.

There are several phenomena that should be considered when describing power transistors. Although the simple theory suggests that common-base current gain, α, is independent of I_E and I_C, α varies with I_C as shown in figure 2.19. The initial low value of α is due to the loss of injected minority carriers at the surface of the base outside the edge of the emitter-base junction. At high currents the density of injected carriers is large and requires a correspondingly large base current to replenish the majority carriers which have recombined in the base. This recombination consists of both the Shockley-Hall-Read and Auger mechanisms. The base conductivity is reduced and the emitter efficiency also falls [24].

Also the base area between emitter and collector becomes negatively biased due to the accumulation of space charge, so that only the edge of the emitter can inject carriers which will reach the collector. This reduces the base transport factor and both these effects will cause the high current gain fall-off. This can be counter-acted by making the edge of the emitter as long as possible where the base

and emitter connections are in the form of stripes as in figure 2.20. In power transistors the number of stripes is increased to handle larger currents and to distribute the current more uniformly [25]. Current gain is dependent on the emitter doping concentration. To improve gain, the emitter should be much more heavily doped than the base. However, as the emitter doping becomes very high, a reduction in gain due to the Auger effect and bandgap narrowing occurs.

Another technique widely used in practice is the use of a graded base doping. The collector-base junction is graded but the emitter-base junction is effectively abrupt due to the heavy emitter doping. This helps to accelerate electrons from emitter to collector, and as the emitter is heavily doped the potential falls rapidly across the emitter-base junction and prevents any significant retarding field returning electrons to the emitter.

In high-power transistors localised hot-spots can lead to catastrophic failures. There are typically hundreds of emitters connected in parallel and the entire device will fail if any single emitter fails. For example, a localised defect in the emitter may start drawing excess current, causing additional Joule heating at the location. This increase in temperature then induces a still higher current. The regenerative mechanism will eventually destroy the device.

This problem can be reduced by introducing a distributed series emitter resistance [26], using a technique called emitter ballasting. Any increase in the current through a particular emitter will be limited by the resistor, so that every emitter will tend to draw approximately the same amount of current.

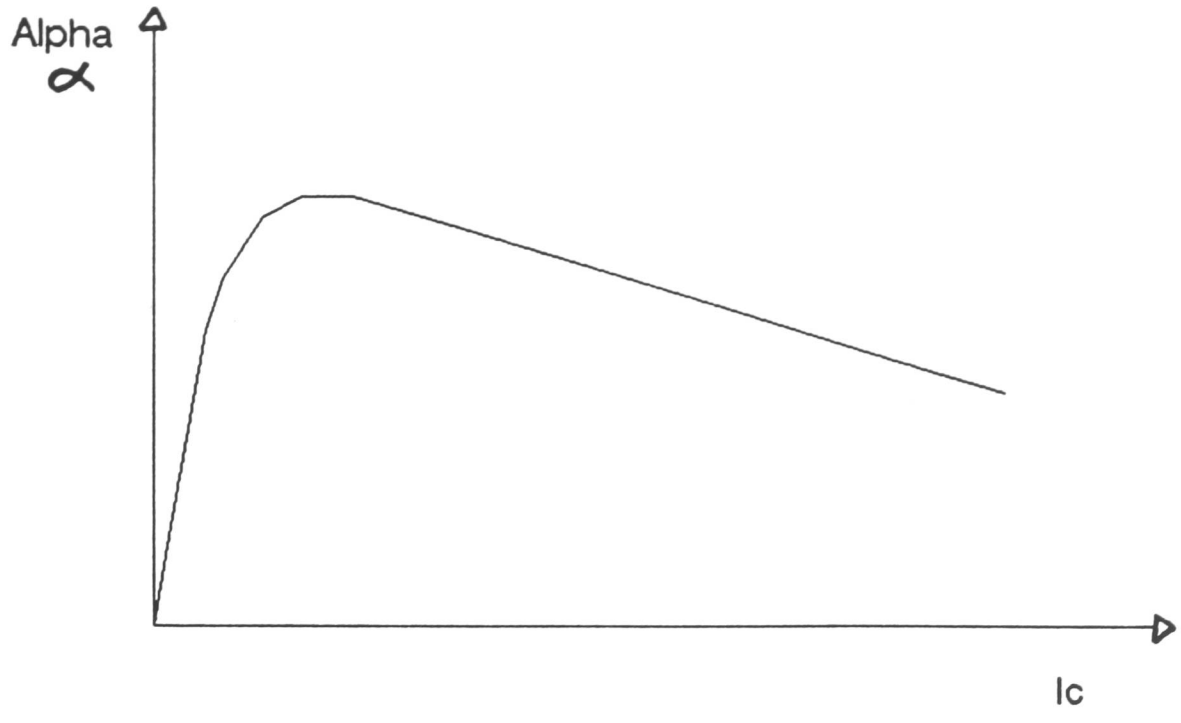


Figure 2.19 The effect of collector current on common-base current gain.

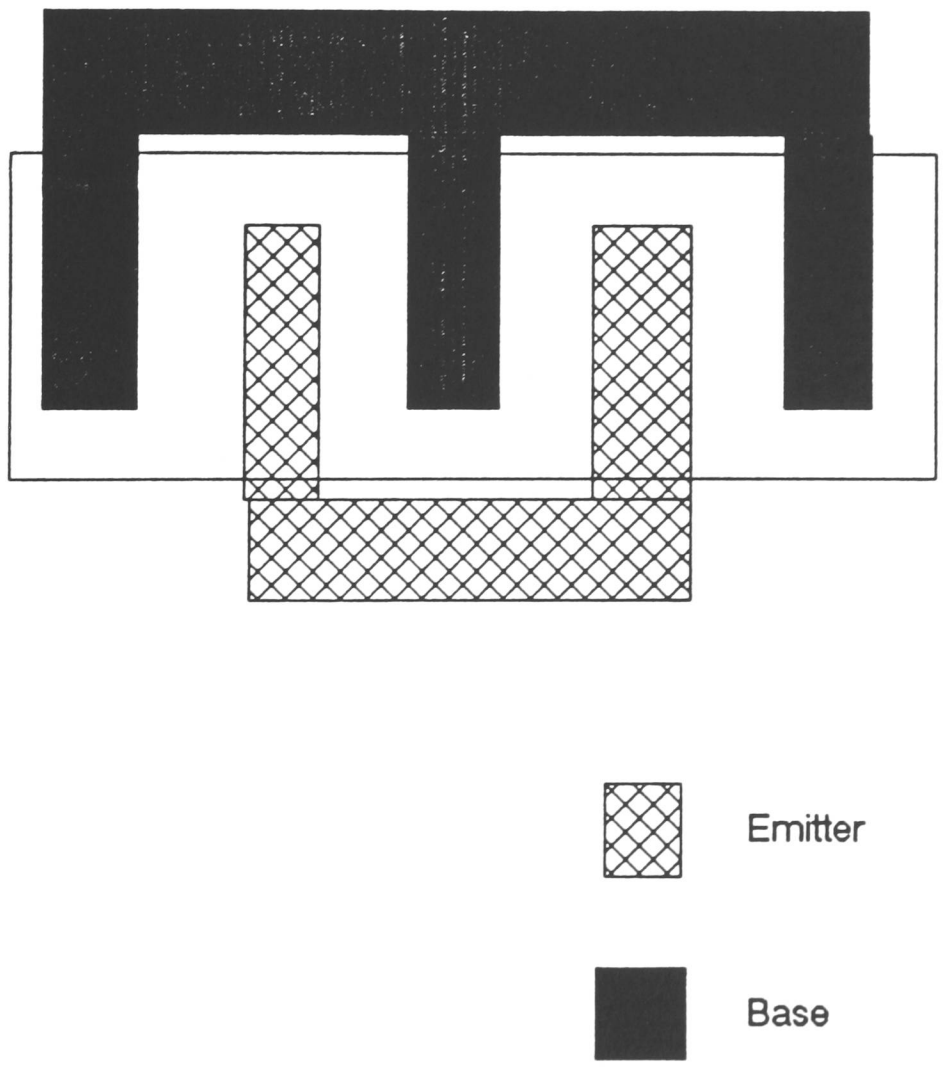


Figure 2.20 A plan view of an interdigitated transistor structure.

In order to integrate npn transistors the collector contact has to be on the surface of the chip rather than the reverse, as shown in figures 2.21 and 2.22. This results in a much larger value of collector resistance. The dominant component of collector series resistance is the path linking the N+ buried layer with the collector, since the epitaxial layer has a comparatively high resistance. This resistance R_1 can be limited by using a sinker diffusion as in figure 2.23. In addition, to minimise the lateral resistance R_2 , an interdigitated structure with a continuous buried layer is usually used.

Integrated pnp transistors do not have as good performance as their npn counterparts. There are two basic types that can be realised ; the vertical pnp and the lateral pnp transistor [27].

In the vertical pnp structure of figure 2.24 current flows vertically from the emitter, which is realised using the npn base diffusion, towards the collector on the bottom of the device. The base is formed using the n- epitaxial layer, and to reduce resistance the contacts are made with n+ areas formed using the npn emitter diffusion. The most severe limitation of the device is that the collector is at ground potential so it can only be used as an emitter follower. Isolation is not possible, therefore these structures are rarely used in practice.

The lateral pnp of figure 2.25 is realised using the npn base diffusion for the emitter and collector regions and the npn emitter diffusion to form the base contact. In this case the buried layer is used to reduce the efficiency of the parasitic vertical pnp transistor. The lateral pnp is punch-through limited and operation takes place on the emitter periphery facing the collector region. Current injected vertically is lost to the substrate. In order to optimise the perimeter to area the circular structure of figure 2.26 is often used. To improve emitter efficiency

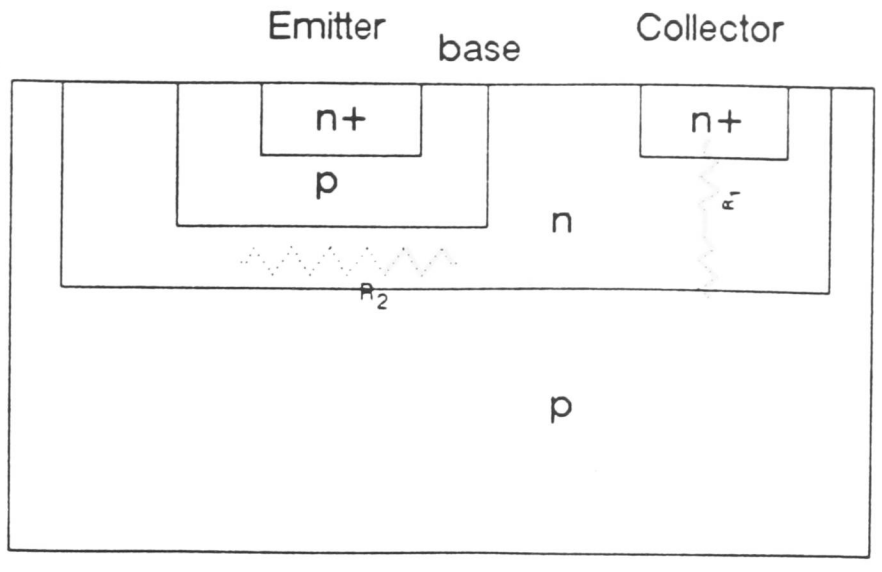


Figure 2.21 A simple lateral npn structure that can be used in integrated applications.

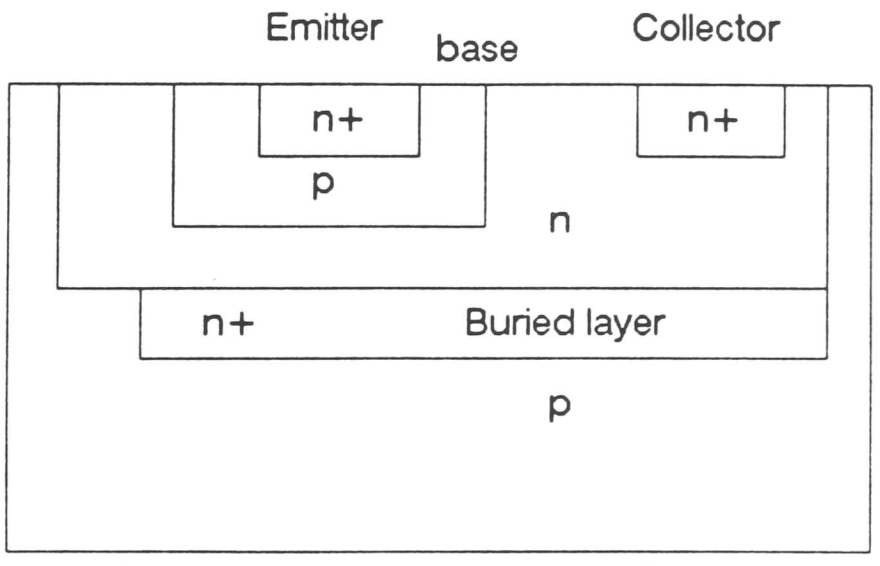


Figure 2.22 A npn transistor structure with a n+ buried layer, that can be used in integrated applications.

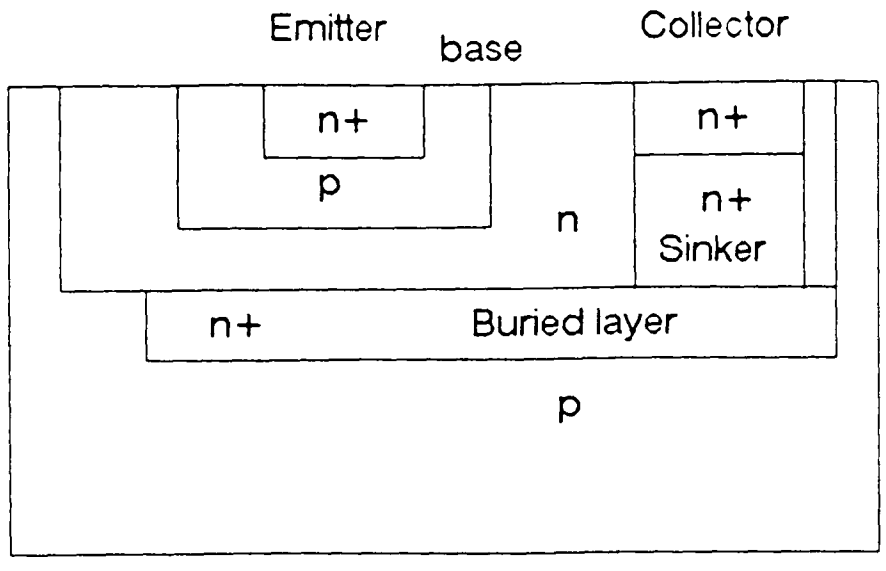


Figure 2.23 A npn transistor structure with a n+ buried layer and a n+ sinker diffusion, that can be used in integrated applications.

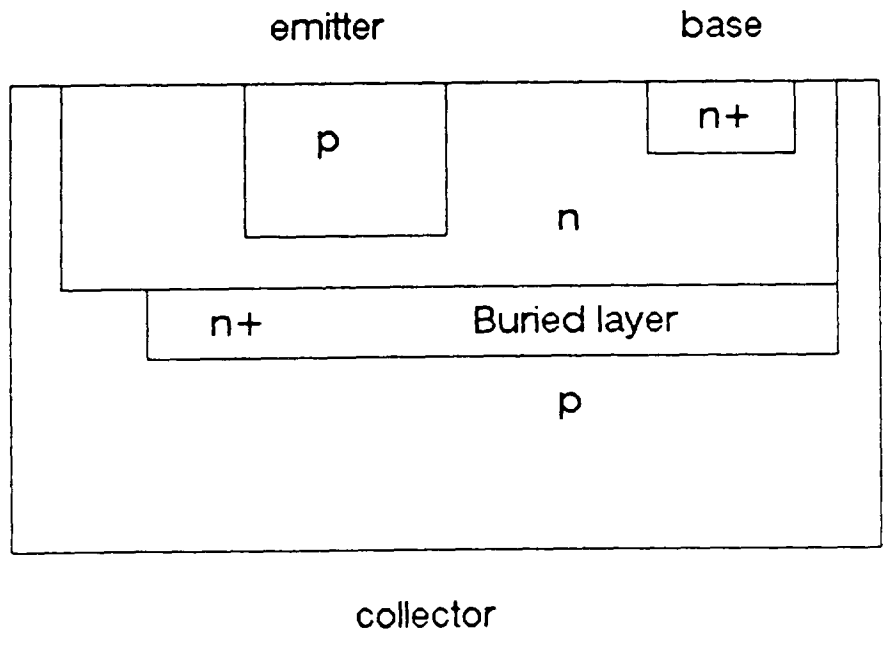


Figure 2.24 A simple vertical pnp transistor structure.

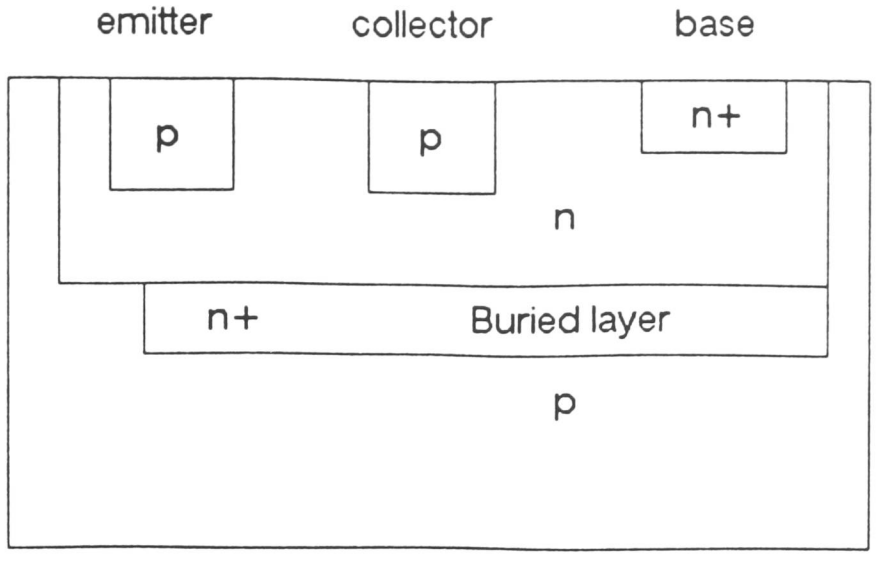
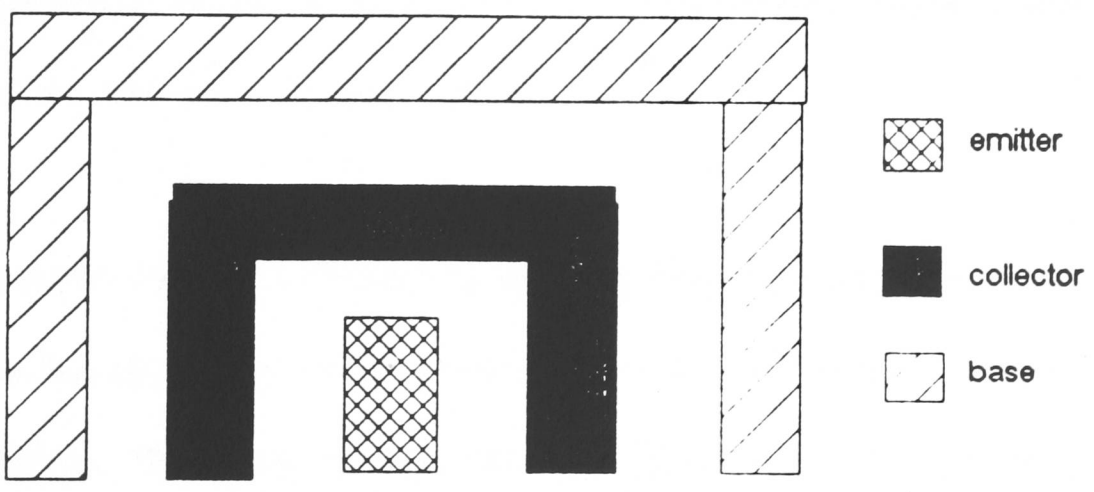


Figure 2.25 A lateral pnp transistor structure that can be used in integrated applications.



Plan View

Figure 2.26 Schematic plan view of a layout used to optimise the perimeter-to-area ratio.

and the perimeter-to-area ratio, one additional masking step and a P+ diffusion can be added to the standard pnp of figure 2.25 to give the structure shown in figure 2.27. This structure will have a higher peak current gain and current handling capability [32].

These lateral pnp transistors have the advantage in that they require no masking and diffusion steps beyond the standard npn bipolar process, but gain, maximum collector current and cut-off frequency are degraded by the device layout.

In order to improve these performance limitations, the epitaxial-base vertical pnp transistor [33,34] was developed. It is based on the standard npn process with the addition of a p-type buried collector diffusion in the n+ buried layer. A p-buried region can be formed above the n- buried layer as boron will out-diffuse faster than either arsenic or antimony into the epitaxial layer. The surface contact for the collector is made using the p+ isolation diffusion, thus forming a completely isolated pnp transistor, as shown in figure 2.28.

These pnp transistors have low frequency response due to large base width, high saturation voltage due to collector series resistance and low current capability due to high base resistivity. The triple-diffused pnp structure of figure 2.29 has a smaller base width [35,36] with lower resistivity. A lightly doped n- layer overdoped with a p+ layer is formed first in the substrate. Then the epitaxial n-layer is grown. A p- region is driven down from the surface to merge with the p+ buried layer. The p+ isolation implant is used to form collector contacts. Then an n-type base is diffused from the surface followed by a highly-doped emitter. This fabrication process requires many more steps than the previous techniques but will result in better performance, as the devices have a smaller, lower resistance base

and a reduced collector resistance.

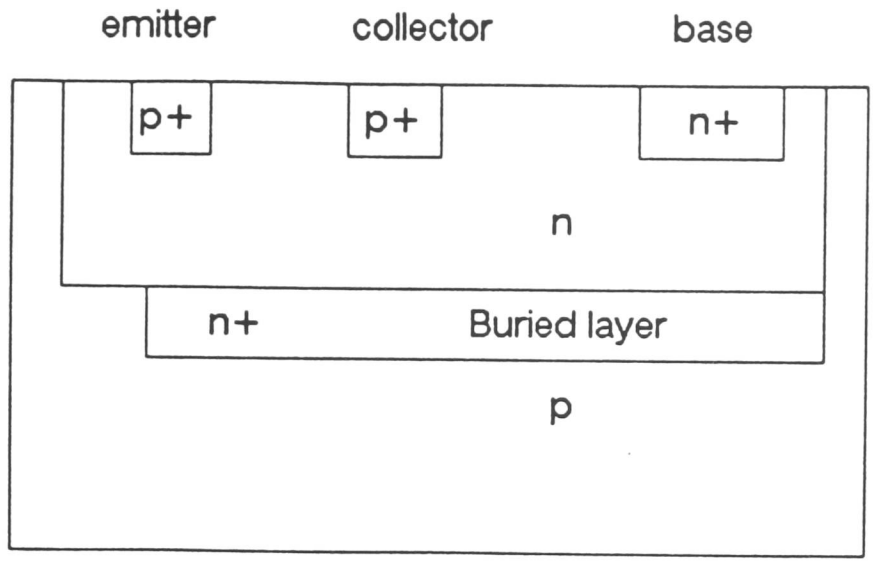


Figure 2.27 A lateral pnp transistor structure that differs from the transistor in figure 2.25 in that it uses an additional p+ implant for the emitter and collector.

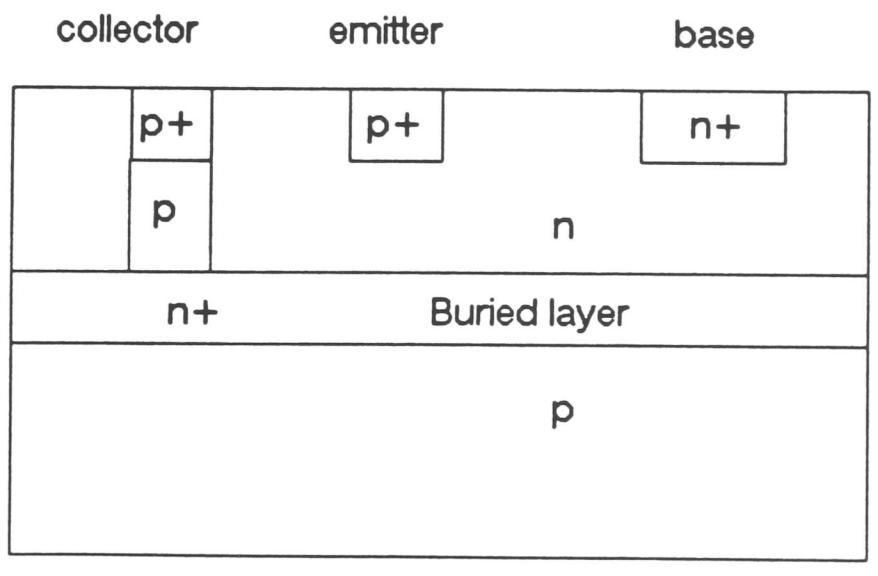


Figure 2.28 An epitaxial base pnp transistor using a p sinker diffusion.

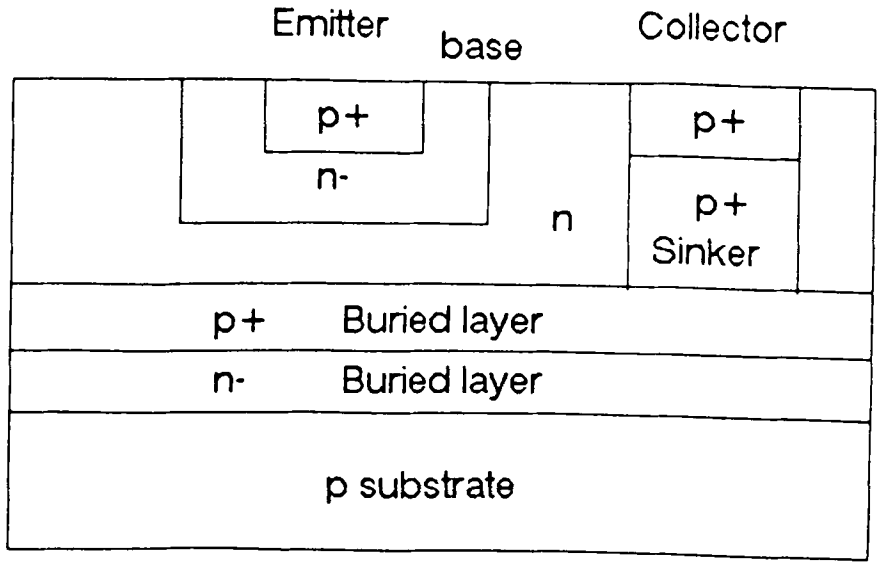


Figure 2.29 A triple diffused pnp transistor using a p sinker diffusion and a p+ buried layer.

2.5 Conclusions

Even though the techniques for producing devices have changed dramatically, it is a testimony to the pioneers of the field that the original operational theories are still valid. There have been modifications to the original device equations throughout the years, but these have concentrated on secondary effects that were generally a consequence of device scaling. These updates model effects such as high electric field relocations and base punchthrough and are necessary for considering the advanced structures of today.

As a bipolar transistor can be considered as two rectifying junctions back to back, it is necessary to consider the high electric field limitations of pn junctions before studying full device structures.

Firstly ideal conditions, and then terms to model more realistic conditions such as corner rounding and reach-through were introduced. Also discussed were the three modes of breakdown; barrier tunnelling,impact ionisation and thermal failure. The emphasis was on impact ionisation, or avalanche multiplication, which is in practice the dominant factor. It can be seen that in infinite structures the breakdown is determined by the doping of the lightly-doped side of junctions. However in practice the peak values are reduced due to the physical size of the junction or its shape. These factors will be seen to apply when more complex structures, i.e. transistors, are discussed.

It is found that in addition to simple avalanche breakdown, other effects such as device punchthrough, junction reachthrough and secondary breakdown can be limiting factors in transistors. Again in practice it is important to consider all the possible failure modes of a transistor. It is often the case that the less obvious failure modes are the limiting factors. All the single failure modes must be

evaluated are to determine the safe operating area for a transistor. The pn junction equations here applied to bipolar devices can equally be applied to other semiconductor devices such as MOSFETs, thyristors or triacs. In fact in the following chapter these techniques are used to determine limiting factors in MOSFET device operation. Further, it is important to remember that both diodes and bipolar transistors can appear as parasitics in integrated MOS structures. Often these parasitics are the main limitation in terms of performance.

The penultimate section in this chapter dealt with the developments that have lead to bipolar power transistors. Firstly discrete structures were introduced, then as this thesis is concerned with smart power, integrated devices were considered. There are two basic structures, lateral and vertical. It is found that for power density vertical is superior and for integration, lateral. To improve on the simple structures epitaxial layers and more masking stages are needed. This results in optimum devices but at a cost. For military or high volume runs this is acceptable, but for ASICs it is not. Therefore the designer must always carefully balance the two opposing goals of price and performance.

In the next chapter unipolar devices are discussed in order to make clear which technology is best suited to smart power, and in which applications. At first these structures seem remote from bipolar, but deeper investigation reveals they are not. For example, punchthrough voltage limitation or parasitic bipolars that cause latch-up in CMOS are both bipolar subjects. Further, the MOS power structures are in fact very similar to their bipolar counterparts , and merged MOS-bipolar devices such as COMFETs (or IGTs), look to be an excellent solution for medium power discrete applications.

References

[1] W.Shockley, "The Path to the Conception of the Junction Transistor", IEEE Trans. Electron devices, vol.ED-23, pp.597-620, 1976.

[2] W.Shockley, "The Theory of p-n Junctions in Semiconductors and p-n Junction Transistors", Bell System Technical Journal, vol.28, pp.435-489, 1949.

[3] M.Tanenbaum and D.E.Thomas, "Diffused Emitter and Base Silicon Transistors", Bell System Technical Journal, vol.35, pp.1-22,1956.

[4] H.C.Theurer, J.J.Kliemack, H.H.Loar and H.Christensen, "Epitaxial Diffused Transistors", Proc.IRE, vol.48, pp.1642-1643, 1960.

[5] M.P.Lepselter, "Beam-Lead Technology", Bell System Technical Journal, vol.45, pp.233-253, 1966.

[6] R.S.Payne, R.J.Scavuzzo, K.H.Olson, J.M.Nacci and R.A.Moline, "Fully Ion-Implanted Bipolar Transistors", IEEE Trans. Electron Devices, vol ED-21, pp.273-278, 1974.

[7] J.J.Ebers and J.L.Moll, "Large-Signal Behaviour of Junction Transistors", Proc.IRE, vol.42, pp.1761-1772, 1954.

[8] J.L.Moll and I.M.Ross, "The Dependence of Transistor Parameters on the Distribution of Base Layer Resistivity, Proc.IRE, vol.44, pp.72-78, 1956.

[9] C.-T.Sah, R.N.Noyce and W.Shockley, "Carrier Generation and Recombination in P-N Junctions and P-N Junction Characteristics", Proc.IRE, vol.45, pp.1228-1253, 1957.

[10] W.M.Webster, "On the Variation of Junction-Transistor Current-Amplification Factor with Emitter Current", Proc.IRE, vol.42, pp.914-920, 1954.

[11] S.M.Sze, "Physics of Semiconductor Devices", Second edition, p.446, Wiley-Interscience, New York, 1981.

[12] C.T.Kirk, "A Theory of Transistor Cutoff Frequency Falloff at High Current densities", IRE Trans.Electron Devices, vol.ED-9, pp.164-174, 1962.

[13] J.M.Early, "Effects of Space-Charge layer Widening in Junction Transistors", Proc.IRE, vol.40, pp.1401-1406, 1952.

[14] C.Y.Chang, S.S.Chai and L.P.Hsu, "Dependence of breakdown Voltage in silicon Abrupt P-N Junctions", IEEE Trans. Electron devices, vol.ED-18, pp.391-393, 1971.

[15] B.Senitzky and J.L.Moll, "Breakdown in Silicon", Physical Review, vol.110, pp.621-620, 1958.

[16] S.M.Sze and G.Gibbons, "Avalanche Breakdown Voltages of Abrupt and Linearly graded p-n Junctions in Ge, Si, GaAs. GaP", Applied Physics Letters, vol.8, p.111, 1966.

[17] D.P.Kennedy and R.R.O'Brien, "Avalanche Breakdown Characteristics of a Diffused P-N Junction", vol.ED-9, pp.478-483, 1962.

[18] K.Shenai, "Analytical solutions for Avalanche-Breakdown of Single-Diffused Gaussian junctions", Solid-State Electronics, vol.26, pp.211-216, 1983.

[19] S.M.Sze, "Physics of Semiconductor Devices", Second edition, p.102, Wiley-Interscience, New York, 1981.

[20] J.Ueda and N.Totsuka, "Breakdown Voltage Analysis of Planar p-n Junctions taking into Account the Radius of Curvature of the Corners of the Patterning Mask", Solid-State Electronics, vol.28, pp.1245-1249, 1985.

[21] R.H.Haitz, A.Getzberger, F.M.Scarlett and W.M.Shockley, "Avalanche effects in silicon p-n junctions 1. Localised Photomultiplication Studies on Microplasmas", Journal of Applied Physics, vol.34, pp.1581-1590, 1963.

[22] W.J.Stepowicz, "D.C. Characteristics of Silicon p-n Junctions at Avalanche Breakdown Including Self-heating", Solid-State Electronics, vol.22, pp.7-13, 1979.

[23] R.S.Cobbold, "Theory and Applications of Field Effect Transistors, pp.265-275, J.Wiley and Sons, New York, 1970.

[24] E.J.McGrath and D.H.Navon, "Factors Limiting Current Gain in Power Transistors", IEEE Trans. Electron Devices, vol.ED-24, pp.1255-1259, 1977.

[25] J.Seymour, "Electronic Devices and Components", pp.136-138, Pitman, London, 1981.

[26] R.P.Arnold and D.S.Zoroglu, "A Quantative Study of Emitter Ballasting", IEEE Trans. Electron devices, vol.ED-21, pp.385-391, 1974.

[27] F.Bertotti and P.Antognetti, " Power Integrated Circuits", Chapter 4, pp.1-4, McGraw-Hill, New York, 1986.

[28] B.Murari, "Power Integrated Circuits : Problems, Tradeoffs and Solutions", IEEE Journal of Solid-State Circuits, vol.SC-13, pp.307-319, 1978.

[29] S.C.Su and J.D.Meindl, "A New Complementary Monolithic Transistor structure", IEEE Journal of Solid-State Circuits, vol.SC-7, pp.170-174, 1972.

[30] S.C.Su and J.D.Meindl, "A New Complementary Bipolar Transistor structure", IEEE Journal of Solid-State Circuits, vol.SC-7, pp.351-357, 1972.

[31] D.W.Aull, D.A.Spires, P.C,Davis and S.F. Moyer, "A High-Voltage IC for a Transformerless Trunk and subscriber Line Interface", IEEE Journal of Solid-State Circuits, vol.SC-16, pp.261-266, 1981.

[32] J.H.Green, S.F.Moyer and J.J.Nahas, "A Channel Unit Signal Controller for Shared Codec D-Type Channel Banks", IEEE Journal of Solid-State Circuits, vol.SC-16, pp.341-348, 1981.

CHAPTER THREE : MOS DEVICES

3.1 Introduction

The MOSFET or metal-oxide-semiconductor field-effect transistor is one of a family of transistors that have a high input impedance and voltage control of output current. This contrasts with bipolar transistors which have low input impedance, and base current controlling output current.

The controlling voltage is applied to the gate of the device. The resulting electric field modulates the current flowing between the drain and source of the device, i.e. the output current. Therefore the operation can be expressed as the change in output current as a function of input voltage change, in units of AV^{-1}. Typically, MOS transistors have transconductances of about 0.25 mAV^{-1} and input capacitances of about 0.02 pF. In order to compare bipolar and MOS performance, it is necessary to examine the current controlled, as the gain of a bipolar and the transconductance of a MOS cannot be directly compared. Using this criteria, it is normally the bipolar transistor which has the higher current handling capability and the lower on-resistance. The redeeming factor for MOS transistors is that they are several times smaller than their bipolar counterparts. This small size is of paramount importance as circuit complexity is increased, resulting in the dominance of MOS over bipolar for high density LSI and VLSI applications.

Another important property of MOS transistors is the fact that the input capacitance leads to the possibility of storing signals at the input. This results in a device that can be used in memory applications.

Figure 3.1 shows a cross-section of an n-channel MOS transistor. The structure is formed by a conductor-insulator-semiconductor layer structure with two electrical contacts to the semiconductor located at either edge of the lower plate of the capacitor-like structure. The top plate of the capacitor is made from polycrystalline silicon (or polysilicon) and is called the gate. Earlier transistors used metal, usually aluminium, as the gate. However, polysilicon predominates in modern applications. The insulator is an intrinsic dielectric formed on the semiconductor substrate. The dielectric most commonly used is very pure silicon dioxide (SiO_2). Because the original transistors were made of a metal oxide semiconductor sandwich , the name MOS transistor was given. This name has remained despite the change from metal to polysilicon gates. The device is also referred to as an insulated gate field-effect transistor, or IGFET, because of the near perfect insulating property of silicon dioxide.

The output terminals of the device are called the drain and source, and because of the device symmetry, these are interchangeable. Circuit convention means that the source is the terminal which is common to the input and output circuits, and the drain is the terminal that is connected to the output load. Because of this, the output current is known as the drain current I_D , whose value is a function of the drain voltage V_D , and the voltage applied to the gate electrode, V_G.

Ohmic contacts are made to both the source and diffused areas. The drain current flows near the surface between the drain and source diffusions. This area is called the channel. In an ideal MOS transistor all the current flowing from the drain will reach the source. This is because there is a high impedance insulator

(gate oxide) above the channel, and an isolating depletion region is formed beneath the channel and around the source and drain junction, as in figure 3.2. This conservation of current contrasts with the bipolar transistor where substrate current flows in the base lead. Another difference from bipolar devices is that current flow is due to majority carriers only. Consequently the MOSFET is known as a unipolar device as opposed to a bipolar transistor, where both majority and minority carriers contribute to current flow.

The bulk semiconductor can be either p-type or n-type, corresponding to n-channel or p-channel transistors. The primary advantage of n-channel transistors is that the mobility of electrons, the majority carrier, is about three times that of holes. This results in a corresponding increase, in the operating speed and gain, so n-channel devices are generally preferred. As a consequence, for the rest of this chapter all equations will refer to nMOS transistors unless otherwise stated. The source and drain regions which make the pn junctions are formed by diffusing, a n+ dopant into these areas.

The transistor performance is directly related to the length of the channel, and therefore depends on gate length L, and width W. In current production processes the minimum gate lengths are in the region 0.5 - 3 microns [1]. The actual transistor length is larger than this due to the source and drain lengths. However, the basic MOS structure is several times smaller than the equivalent bipolar one. Transistors with channel lengths down to 0.1 microns have been produced in laboratories [2] , but this size reduction has not been achieved on a production scale, yet.

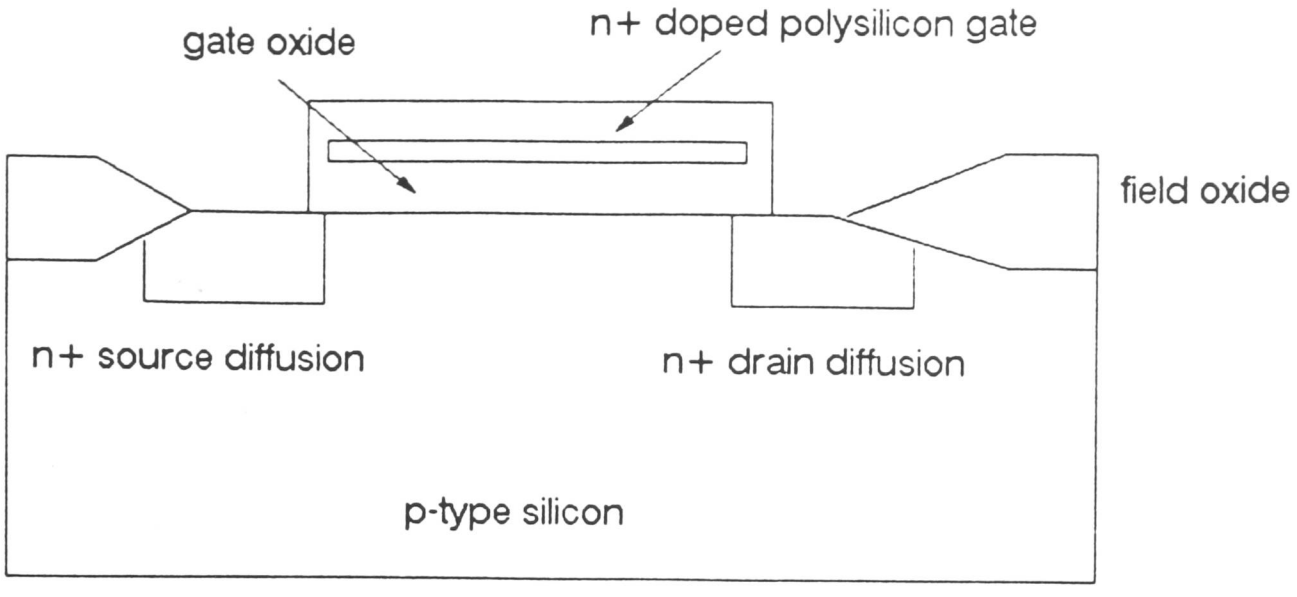


Figure 3.1 Cross section of a NMOS transistor showing drain, source and gate.

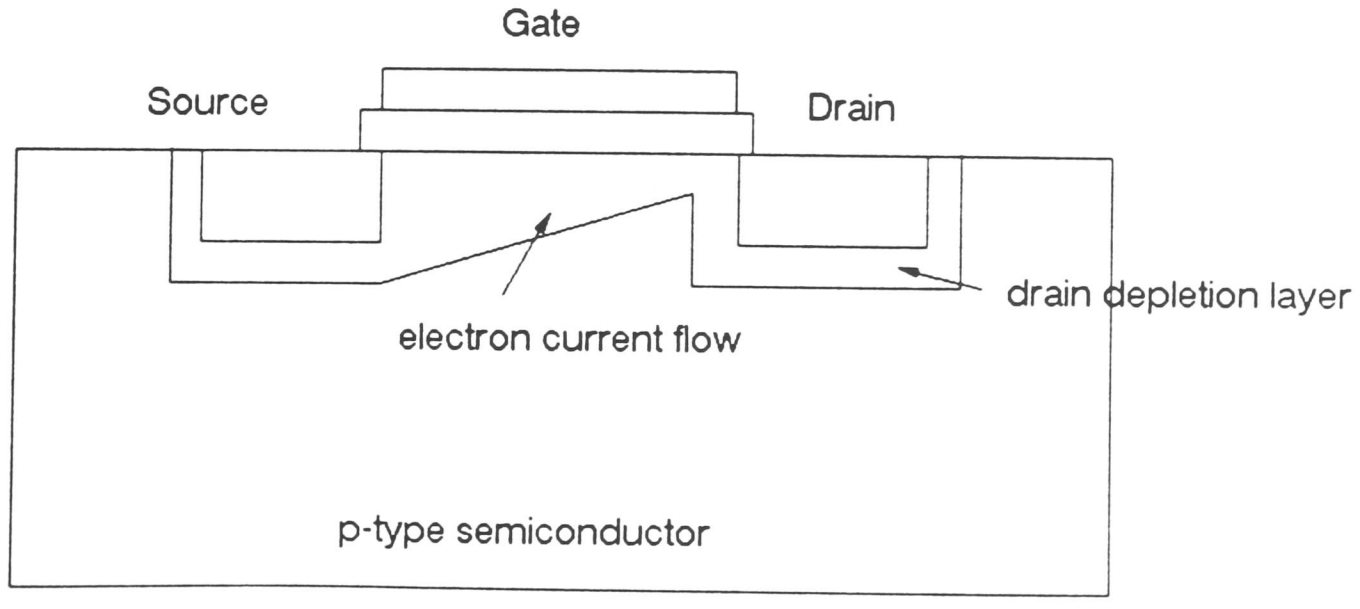


Figure 3.2 Cross section of a NMOS transistor when current is flowing in the channel

3.2 Device Characteristics

3.2.1 Threshold Voltage

In a MOS transistor the gate voltage at which the current can flow between the drain and source is referred to as the threshold voltage, V_T. When $V_{GS} > V_T$ for an n-channel device the device conducts and is said to be on. The value of threshold voltage can be found by considering the MOS capacitor structure of figure 3.3. This simplification of structure is justified as it is the field across the gate insulator which sets the carrier density in the channel.

Consider a sectional volume of width δy, taken vertically through the middle of the channel. The capacitance has two constituent parts. The first is C_O, due to the gate oxide, and the second C_D is due to the depletion region. The energy band diagrams of figure 3.4 show the effect of varying the gate voltage in a MOS capacitor.

The Fermi level is defined as the energy at which the probability of finding an electron within the energy gap is 0.5. The position of this Fermi level is a function of doping and type of semiconductor. In this case, the p-type substrate, the Fermi level will lie below the intrinsic Fermi level E_i. This difference is often defined using a parameter Ψ_B which is defined as :

$$\psi_B = \frac{E_i - E_F}{q} \qquad 3.1$$

$$\psi_B = \frac{kT}{q} ln\left(\frac{N_A}{n_i}\right) \qquad 3.2$$

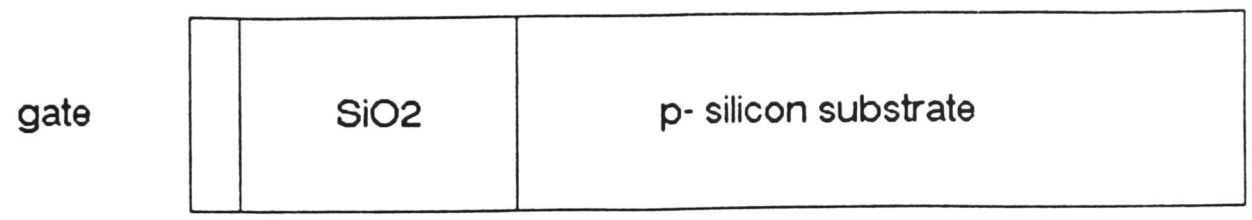


Figure 3.3 Cross section of a MOS capacitor

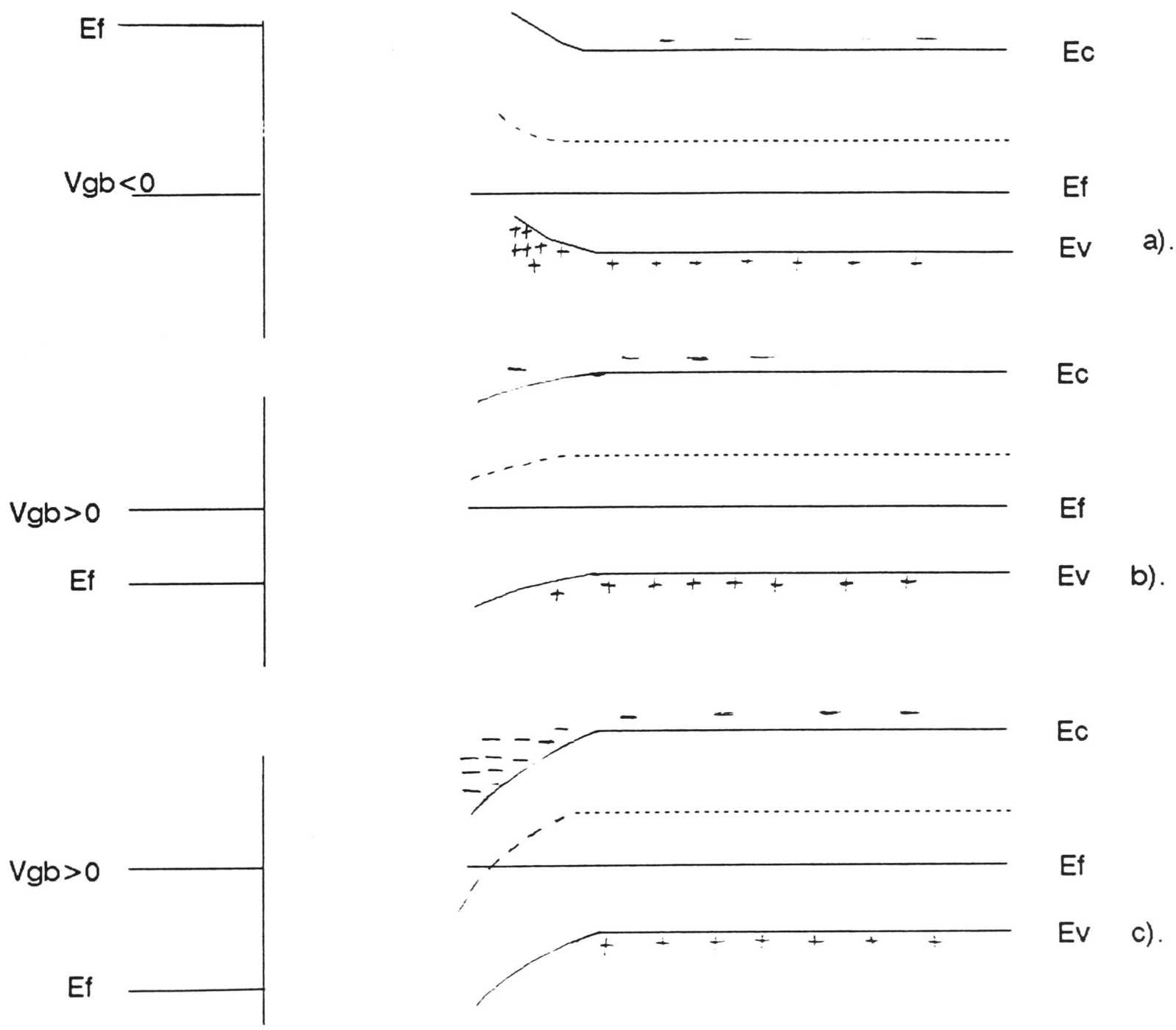


Figure 3.4 The energy band diagrams of a MOS capacitor in (a)accumulation, (b)depletion and (c)inversion.

When this MOS capacitor is biased with positive or negative voltages, three cases may occur at the surface. If a negative bias is applied, $V_{GB} < 0$ as in figure 3.4(a), the top of the valence bands bend upward and are close to the Fermi level. In the ideal case, no current flows in the structure, so the Fermi level remains constant in the semiconductor. Since the carrier density depends exponentially on the energy difference, holes from the substrate accumulate at the surface. If a positive bias is applied, $V_{GB} > 0$, as in figure 3.4(b), holes from the substrate are repelled from the surface and the energy bands bend down towards the surface. This effect is called depletion, and the resulting change in concentration of minority carriers means that Ψ_B is varied near the oxide-semiconductor boundary. The surface potential Ψ_S is therefore a measure of the band bending at this boundary. When the bias is further increased, as in figure 3.4(c), the surface is further depleted and inversion takes place when $\Psi_S >= \Psi_B$. At this point the number of electrons (minority carriers) at the surface is larger the the number of holes, and the surface is thus inverted. Heavy inversion is defined when $\Psi_S = 2\,\Psi_B$ and at this point the minority carrier electron density equals the magnitude of the majority carrier hole density in the bulk.

It is also assumed that if the bias is further increased, the depletion region will remain approximately constant, and the extra voltage will appear across the oxide capacitance. This means that the total induced surface charge Q_S can be separated into two components. The first of these is Q_D, the charge in the depletion region. The other is Q_C, the excess electrons in the inversion layer which cause conduction in the n-channel device. Due to charge conservation the total induced surface charge is equal in magnitude, but opposite in sign to the gate charge Q_G.

$$Q_G + Q_C + Q_D = 0 \qquad 3.3$$

If we regard the depletion region as a rectangular charge distribution, and the surface channel charge a δ function, then the conditions are the same as an abrupt pn junction with a potential $2\,\Psi_B$ across it. The voltage across the gate insulator is V_G - $2\,\Psi_B$ therefore the depletion charge and excess electron charges are given by :

$$Q_D = -\sqrt{(2\epsilon_0\epsilon_S qN_A)}\sqrt{(2\Psi_B)} \qquad 3.4$$

$$Q_C = -V_G C_0 + 2\Psi_B C_0 + \sqrt{(2\epsilon_0\epsilon_S qN_A)}\sqrt{(2\Psi_B)} \qquad 3.5$$

It was assumed earlier that the voltage across the depletion region remains constant at $2\,\Psi_B$,at and beyond inversion, and that no channel current will flow in a MOS transistor until $V_G > V_T$. Then it is seen that V_T is given by :

$$V_T = 2\Psi_B + \frac{\sqrt{(2\epsilon_0\epsilon_S qN_A)}\sqrt{(2\Psi_B)}}{C_0} \qquad 3.6$$

The simplistic expression does not take into account the effect of surface states that exhibit a net positive charge layer Q_{ss} located within the silicon dioxide. Measurements have shown that Q_{ss} is dependent on crystallographic orientation, but is independent of doping type. This fixed oxide charge is located just within the silicon dioxide and typical values lie between 10^{15} - 10^{16} charges per square

metre. Another effect that must be considered is the energy band bending due to the unequal work functions of the gate and the silicon. This work function difference Φ_{MS} can vary from around +1.0 volt for p+ poly on n substrate to -1.0 volt for n+ poly on p substrate. Therefore equation 3.6 becomes :

$$V_T = \Phi_{MS} - \frac{Q_{SS}}{C_0} + 2\Psi_B + \frac{\sqrt{(2\epsilon_0\epsilon_S qN_A)}\sqrt{(2\Psi_B)}}{C_0} \qquad 3.7$$

This can be written as :

$$V_T = V_{FB} + 2\Psi_B + \gamma\sqrt{(2\Psi_B)} \qquad 3.8$$

where :

$$\gamma = \frac{\sqrt{(2\epsilon_0\epsilon_S qN_A)}}{C_0} \qquad 3.9$$

$$V_{FB} = \Phi_{MS} - \frac{Q_{SS}}{C_0} \qquad 3.10$$

Up until now we have discussed the MOS transistor without considering the influence of the source and drain terminals. When a voltage is applied across the source-drain contacts the MOS structure enters a non-equilibrium state. The quasi-Fermi levels or imrefs are defined as the potentials that apply under bias when the product of the minority carrier densities no longer equals n_i^2. These imrefs are determined using the following equations, for electrons and holes respectively.

$$\phi_n = \Psi - kT/q \ln(n/n_i) \qquad 3.11$$

$$\phi_p = \Psi + kT/q \ln(p/n_i) \qquad 3.12$$

Under bias, the imref of the minority carriers is lowered from the equilibrium Fermi level, In figure 3.4(c) the equilibrium condition under the gate bias that causes surface inversion was shown. If now a positive bias is applied to the drain, then the electron imref is lowered towards the drain contact, whilst the hole imref remains at the bulk Fermi level. Due to the fact that an inversion layer can only be formed when the potential at the surface crosses over the imref of the minority carrier, the gate voltage required for inversion at the drain is larger than the equilibrium case.

There exist two types of n-channel MOSFETs. If at zero bias the channel conductance is very low in an nMOS transistor, then a positive gate bias must be applied to the gate to form the channel. In this case, the transistor is called a (normally-off) enhancement MOSFET. If an n-channel exists at zero bias, then a negative bias needs to be applied to deplete the carriers in the channel to reduce

channel conductance. This type of transistor is called the (normally-on) depletion MOSFET.

For the enhancement device, a positive bias greater than the threshold voltage must be applied before substantial current flows. For the depletion MOSFET, a large current can flow when $V_G = 0$, and this current can be modulated by varying the gate voltage.

3.2.2 Drain Current Equation

The MOS transistor is in fact a four terminal device, there being a connection to the substrate on the back of the device. This contact can act as a second gate, and if a negative bias V_B is applied at this terminal, this second gate will try to turn off the transistor. This so-called body effect has to be taken into account as it will increase the threshold voltage. The charge in the depletion region has to be modified to include this effect, and at a point y from the drain :

$$Q_D(y) = -\gamma[V(y) + 2\Psi_B - V_B]^{\frac{1}{2}} \qquad 3.13$$

It can be shown that the mobile inversion charge is given by :

$$Q_C(y) = [V_G - \Phi_{MS} + \frac{Q_{SS}}{C_0} - 2\Psi_B - V(y)]C_0 + \gamma[V(y) + 2\Psi_B - V_B]^{\frac{1}{2}} \qquad 3.14$$

If it is assumed that the mobility of electrons μ_N is constant throughout the channel length L, then the voltage drop in an elemental section along the channel length (see fig 3.5) is given by :

$$dV(y) = I_D\, dR = \frac{I_D\, dy}{W \mu_N |Q_C(y)|} \qquad 3.15$$

Now if this expression is integrated along the channel length from y=0 to y=L with V(0) = 0 and V(L) = V_D then :

$$I_D = \frac{W}{L}\mu_N C_0[(V_G - \Phi_{MS} + \frac{Q_{SS}}{C_0} - 2\Psi_B)V_D - \frac{V_D^2}{2} - \frac{2}{3}\gamma'((V_D + 2\Psi_B - V_B)^{\frac{3}{2}} - (2\Psi_B - V_B')^{\frac{3}{2}})] \qquad 3.16$$

In simple circuit calculations, for small V_D and no V_D, a simpler expression is often used :

$$I_D = \frac{W}{L}\mu_N C_0[(V_G - V_T)V_D - \frac{V_D^2}{2}] \qquad 3.17$$

This reduced form does not include the effect of the substrate-source potential V_B'. It can be shown that the change in threshold voltage due to the body effect is given by [3] :

$$V_T^* = V_T + \gamma'(-V_B')^{\frac{1}{2}} \qquad 3.18$$

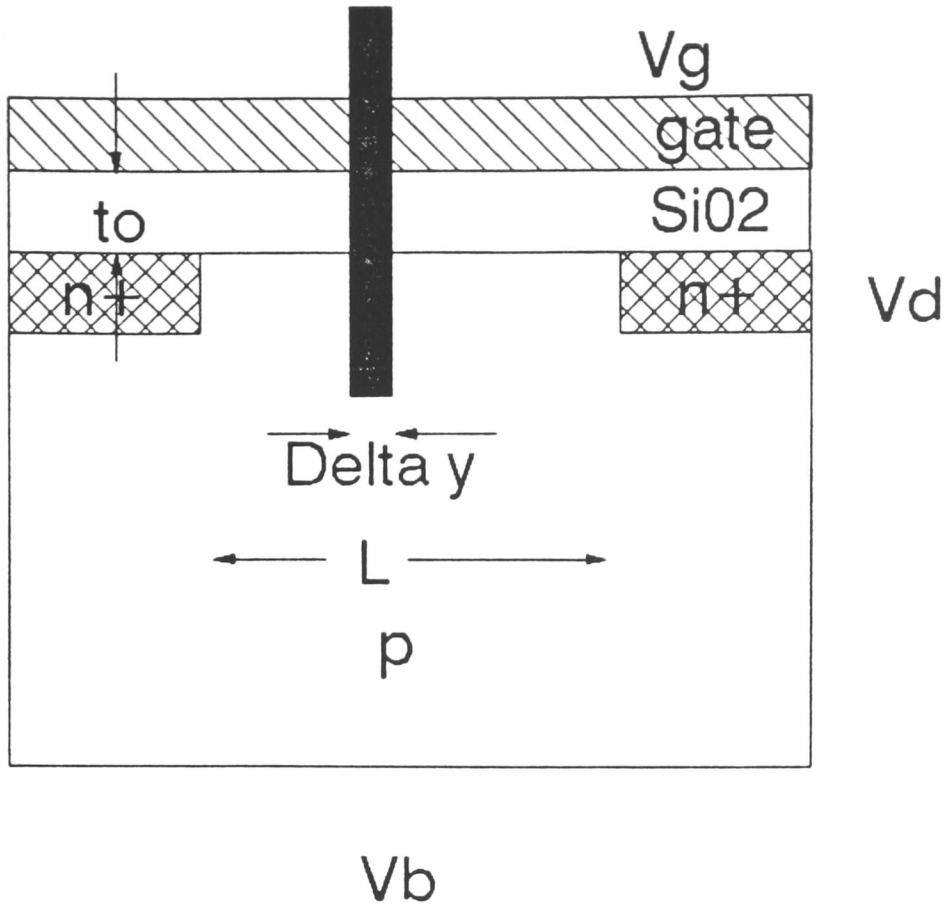


Figure 3.5 Cross section of a MOS transistor showing sectional volume delta y (δy).

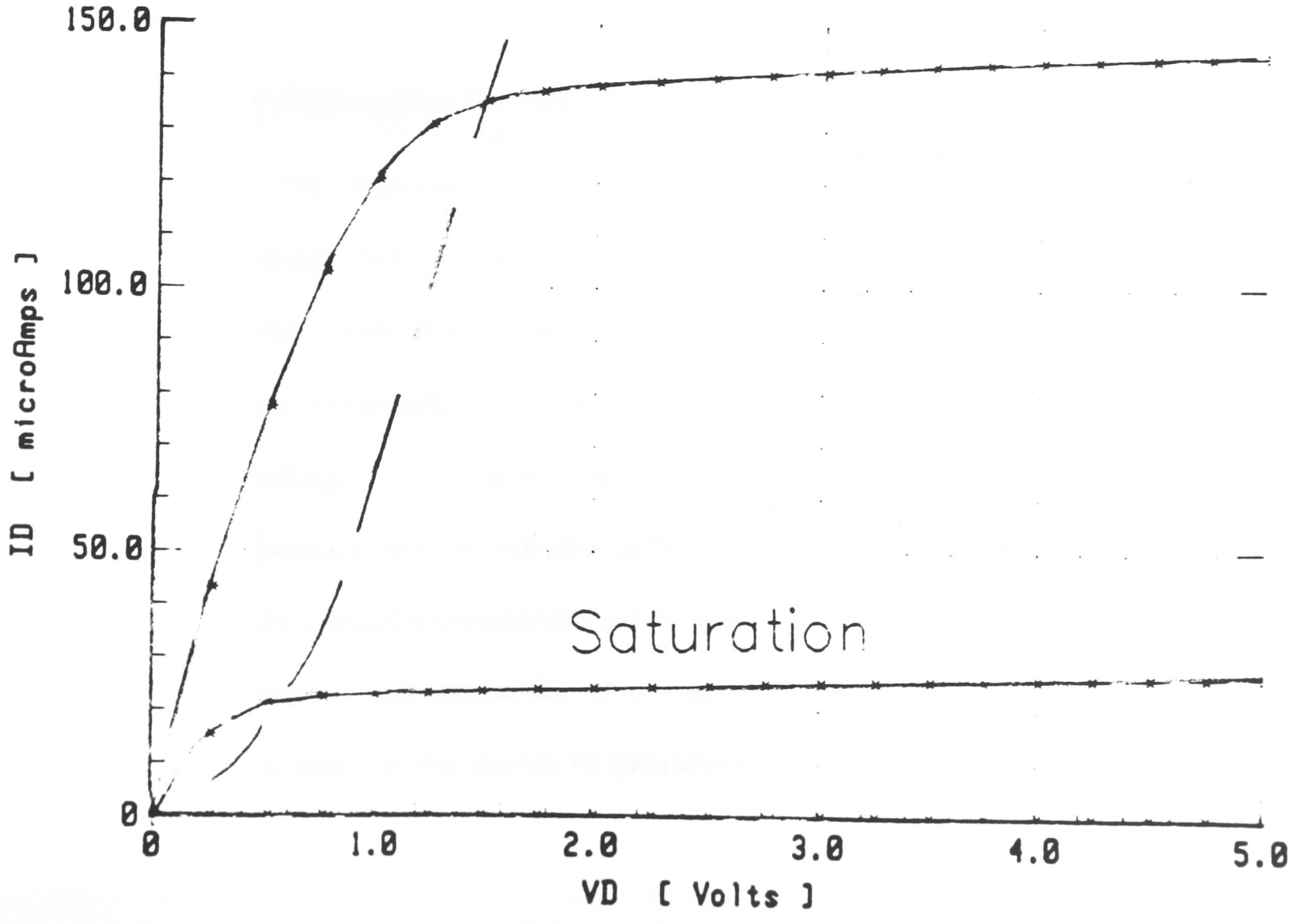


Figure 3.6 Drain current characteristics for a NMOS transistor.

The general effect of this is to make an enhancement MOS have a larger threshold and a depletion MOS transistor to be less depleted. It can be seen from equation 3.17 that the drain current is directly proportional to the ratio of gate width to gate length. This aspect ratio means that the design engineer can set the drain current, gain and switching speed by altering the gate length and width during layout. In order to simplify the equation further, the term $\mu_N C_o$ is often referred to as the process gain K'. The subsequent product K'W / L is called the transistor gain factor ß ,and using these terms, equation 3.17 reduces to that of equation 3.19 :

$$I_D = \beta[(V_G - V_T)V_D - \frac{V_D^2}{2}] \qquad 3.19$$

3.2.3 Saturation Conditions

The gradual channel approximation, where the contribution of the longitudinal field gradient to the charge in the surface space charge region is neglected, was used in the previous section. For a fixed gate voltage, the field gradient, particularly in the vicinity of the drain junction, increases with increasing drain voltage. For a device operating in the saturation region, this approximation becomes less valid as one proceeds along the channel towards the drain. Near the drain the potential distribution in the space charge region becomes two dimensional as a result of the interaction of the gate and drain electrodes. To properly model the operation of the device in saturation this effect should be considered.

From the expressions for drain current, it can be seen that I_D will tend to a maximum value at a particular value of drain voltage, V_{Dsat}, for a constant gate voltage, V_G. This current limiting or saturation can be seen in figure 3.6. Above the saturation point the current curves level off.

Consider what happens as V_D is increased to V_{Dsat} as in figure 3.7. At this point the voltage across the gate insulator at the drain end of the channel is exactly V_T. Therefore inversion is supported and because of current conservation the carrier velocity under the lateral drain field is sufficient to maintain current flow. If V_D is now raised above this value, as in figure 3.8, then the drain current will remain practically constant. The excess drain voltage is dropped across the drain depletion region. As the drain potential increases the depletion region extends both into the bulk and towards the surface. When the potential voltage is further increased, the depletion region expands causing an effective electrical shortening of the channel length. Above a certain voltage the drain depletion can reach the source (punchthrough) causing breakdown with usually destructive consequences. The distance X_D of the length of the drain depletion region at the surface, assuming simple pn junction theory, is given by :

$$X_D = [\frac{2\epsilon_0\epsilon_s}{qN_A})\sqrt{(V_D - V_{Dsat}}]^{\frac{1}{2}} \qquad 3.20$$

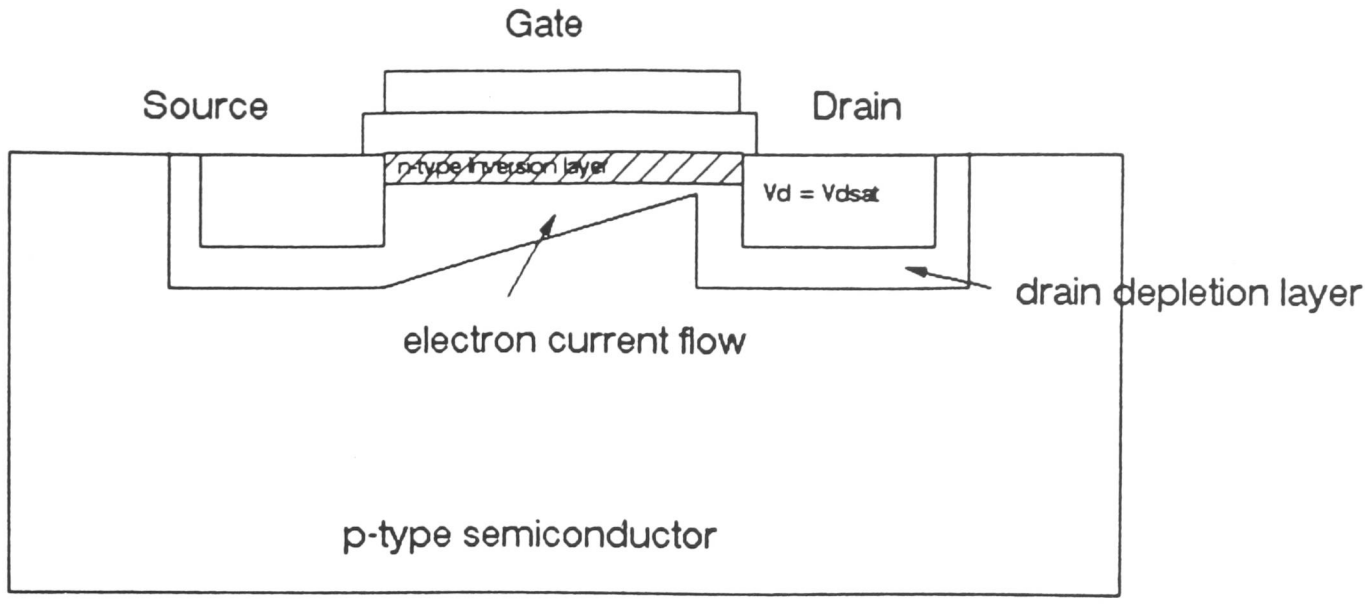


Figure 3.7 Cross section of a NMOS transistor at the saturation point.

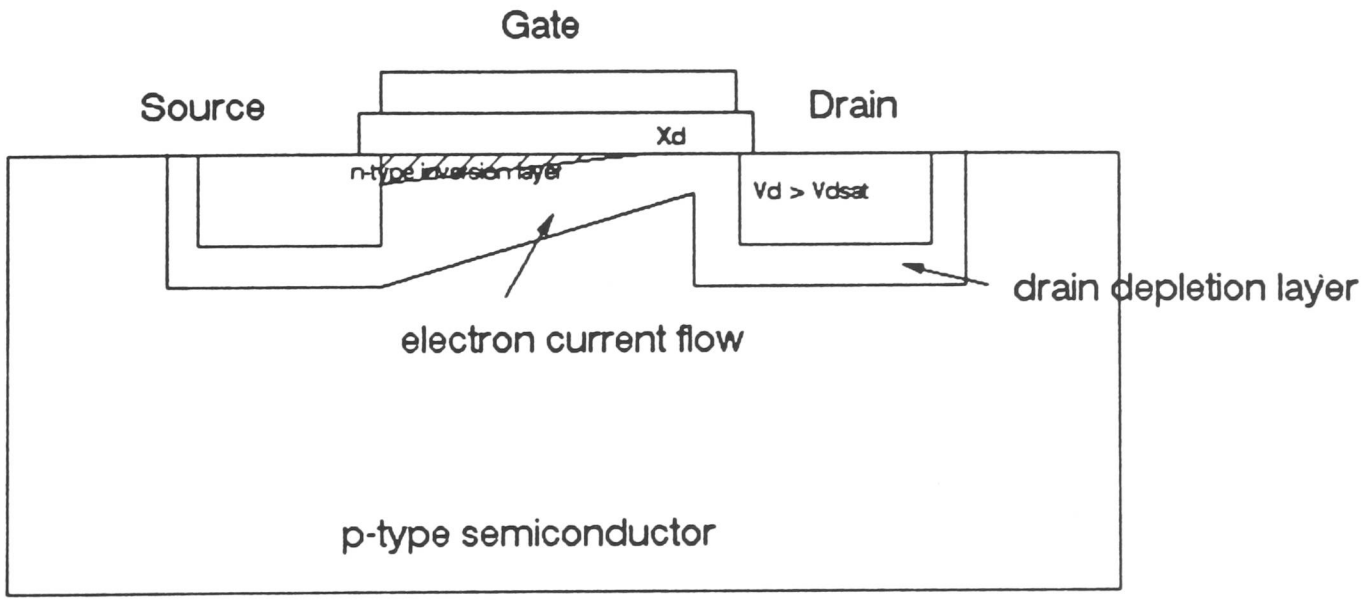


Figure 3.8 Cross section of a NMOS transistor above the saturation point.

The drain current in saturation is found by first determining V_{Dsat} by differentiating the drain current equation and setting the result to zero. This yields the following expression for the simplified drain current expression :

$$V_{Dsat} = V_G - V_T \tag{3.21}$$

Substituting into equation 3.19 gives :

$$I_{Dsat} = \beta \frac{(V_G - V_T)^2}{2} = \beta \frac{(V_{Dsat}^2)}{2} \tag{3.22}$$

3.2.4 Subthreshold Region

The subthreshold region is defined as the working area where the gate voltage is below the threshold voltage and the semiconductor surface is in weak inversion [4,5]. The subthreshold current is dominated by diffusion terms, and is calculated in similar fashion to the collector current in a bipolar transistor, and is given by :

$$I_D = -qAD_n \frac{dn}{dy} = qAD_n \frac{n(0) - n(L)}{L} \tag{3.23}$$

Where A is the cross section of current flow and n(0) and n(L) are the electron densities at the source and drain respectively. These electron densities are given by :

$$n(0)=n_{p_0}\exp(\beta'\Psi_s) \qquad 3.24$$

$$n(L)=n_{p_0}\exp(\beta'\Psi_s-\beta V_D) \qquad 3.25$$

Where Ψ_s is the surface potential at the source. The weak inversion surface field is given by :

$$E_S=-\frac{Q_B}{\epsilon_S}=\sqrt{(\frac{2qN_A\Psi_S}{\epsilon_S})} \qquad 3.26$$

Therefore I_D is now given by [6,7] :

$$I_D=\mu_N\frac{W}{L}\frac{aC_i}{2\beta'^2}(\frac{n_i}{N_A})^2(1-\exp(-\beta'V_D)\exp(\beta'\Psi_S)(\beta'\Psi_S)^{\frac{-1}{2}} \qquad 3.27$$

Where a is given by :

$$a=\frac{\sqrt{2}(\epsilon_s/L_D)}{C_i} \qquad 3.28$$

The gate voltage swing S needed to reduce the current by one decade is given by :

$$S = \ln 10 \frac{dV_G}{d(lnI_D)} = \frac{kT}{q} \ln 10 [1 + C_D(\Psi_S)/C_i] / 1 - (\frac{2}{a^2})[C_D \frac{\Psi_S}{C_i}]^2 \qquad 3.29$$

For a >> C_D / C_i this reduces to :

$$S = \frac{kT}{q} \ln 10 (1 + \frac{C_D}{C_i}) \qquad 3.30$$

If there are interface traps then there is a capacitance C_{it} associated with these traps, and C_D has to be replaced by (C_D + C_{it}) in equation 3.30. If a substrate bias is applied, the depletion layer capacitance is reduced and therefore subthreshold swing is reduced accordingly.

3.2.5 Mobility

Up until now we have considered the mobility of carriers in the channel to be constant. However the mobility and drift velocity are dependent on the inversion layer thickness. From Gauss' Law, the peak field at the silicon surface is given by [8] :

$$E_{si} = \frac{1}{\epsilon_{si}}[Q_{inv} + Q_B] \qquad 3.31$$

Where Q_{inv} is the contribution from the induced mobile electrons in the inversion layer. This peak field is not what most of the electrons will encounter. It can be shown [9,10,11] that the carriers are distributed in a Gaussian profile with the peak inside the silicon surface, and that the electric field in the inversion layer is given by :

$$E_{eff} = \frac{1}{\epsilon_{si}}[\frac{1}{2}Q_{inv} + Q_B] \qquad 3.32$$

The dependence of mobility on the gate field can be described by an empirical approximation [12] :

$$\mu_{eff} = \mu_{max}(\frac{E_C}{E_{eff}})^{C1} \qquad 3.33$$

Where μ_{max} is the maximum value for a given doping level, C1 is an empirical constant of the order 0.22 - 0.26 , and E_C is the critical electric field value.

When a bias is applied to the substrate the bulk depletion charge Q_B is altered. The threshold voltage is increased and the new value for effective field becomes :

$$E_{eff}(V_{BS}) = \frac{C_0}{\epsilon_{si}}\left[\frac{1}{2}[V_G - V_T(V_{BS})] + \frac{Q_B(V_{BS})}{C_0}\right] \qquad 3.34$$

3.2.6 Temperature effects

The operation of any semiconductor device is dependent on temperature. Parameters such as mobility effect the process gain factor. The empirical relationship of equation 3.35 can be used to model this.

$$K' = K_0'\left(\frac{T}{T_0}\right)^{\frac{3}{2}} \qquad 3.35$$

Where K'_0 is the value of K' at temperature T_0, and T is the operating temperature of the transistor. The threshold voltage V_T is a strong function of temperature as is Ψ_B. The change in threshold voltage is given by [13,14] :

$$\frac{dV_T}{dT} = \frac{d\Psi_B}{dT}\left(2 + \frac{1}{C_i}\sqrt{\left(\frac{\epsilon_S q N_A}{\Psi_B}\right)}\right) \qquad 3.36$$

Where :

$$\frac{d\Psi_B}{dT} = \pm\frac{1}{T}\left[\frac{E_G(T=0)}{2q} - |\Psi_B(T)|\right] \qquad 3.37$$

The general effect is that the threshold voltage increases with decreasing temperature at a rate of approximately -2 mVK^{-1}. The other changes are a reduction in subthreshold swing, higher mobility, lower power consumption, lower junction leakage current and lower metal-line resistance. In order to profit from these effects however, there is a major disadvantage in that the transistor has to be forcibly cooled resulting in additional equipment being needed.

3.2.7 Small-Signal Parameters

The previous sections gave the d.c. characteristics, but dynamic a.c. parameters are needed for both deriving analogue performance and determining frequency limits for digital switching. The transconductance g_M can be found by differentiating equation 3.19 to give :

$$g_m = \beta' V_D \qquad 3.38$$

In saturation this becomes :

$$g_m = \beta' V_{Dsat} \qquad 3.39$$

The on-resistance R_{on} depends on both geometry and applied potential, and in practice this should be minimised for good circuit performance. This can be realised by designing the transistor with a large aspect ratio and driving it with a high gate voltage.

The maximum operating bandwidth F_{max} of a small-signal amplifier, where the frequency at which the modulus of current gain is unity, is given by :

$$F_{max} = \frac{\mu_N}{2\pi L^2} V_{Dsat} \qquad 3.40$$

In practice however, capacitive loading on the output will reduce the practical operating bandwidth.

3.3 Power MOSFETs

3.3.1 Introduction

There are some severe limitations in using conventional MOS structures in power applications. The main problem is that the reverse-biased body-to-drain junction depletes further into the body region than the drain region because the drain is more heavily doped than the body (see figure 4.11 for an example of this effect). If high voltages are to be obtained, the depletion of this junction into the body both laterally and vertically means that the channel length has to be increased accordingly.

The first solution to compensate for this problem [15,16,17,18] allowed some of the drain depletion region to spread into the N- drain area, to increase the high-voltage capability. However, as can be seen from figure 3.9, the channel length is still very long.

In the early seventies a MOS structure was proposed which eliminates many of the voltage and on-resistance limitations of conventional MOS transistors. It was called the double-diffused or DMOS transistor [19].

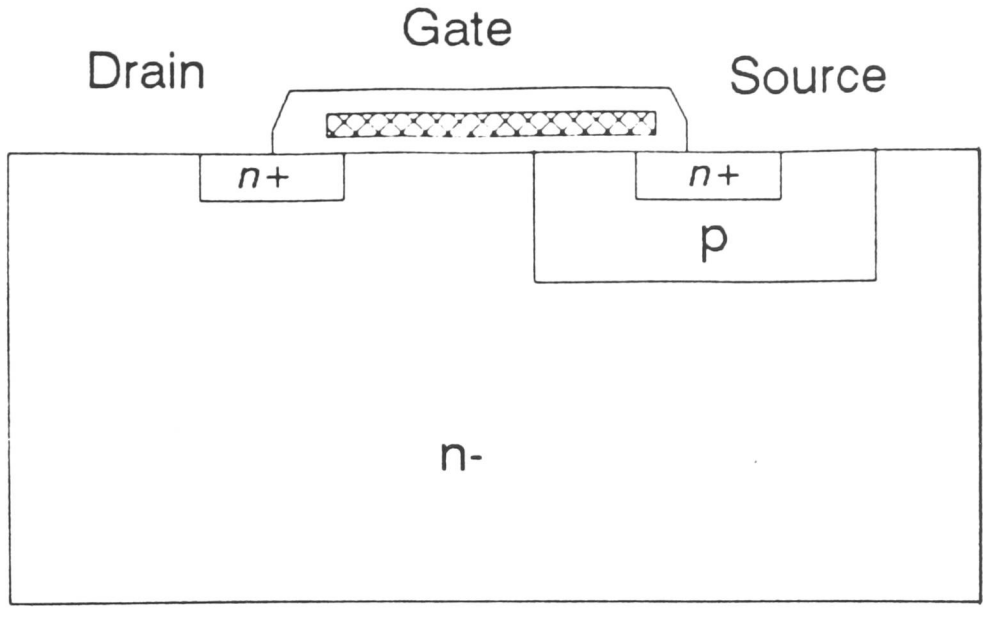


Figure 3.9 An early n-channel double diffused DMOS transistor.

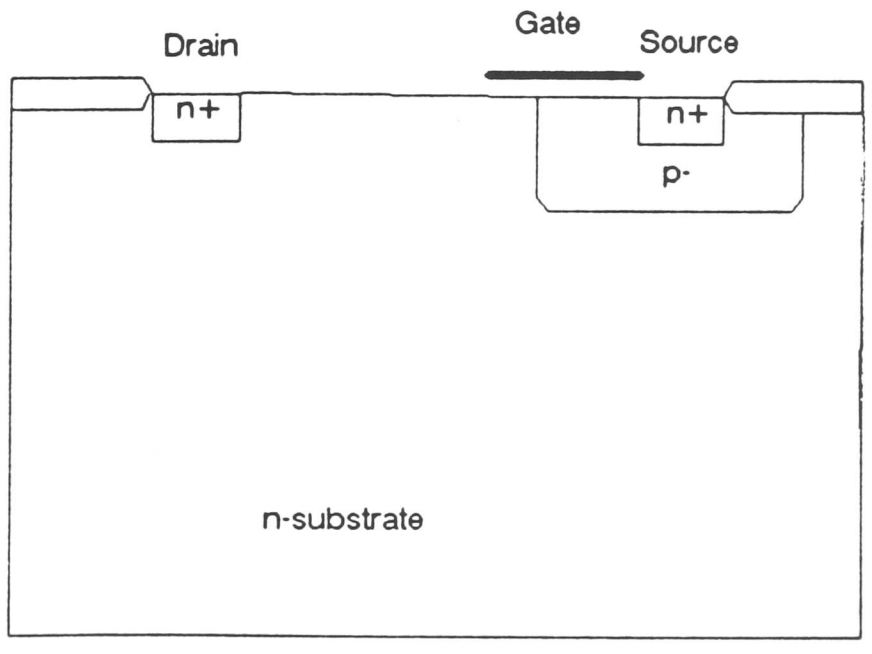


Figure 3.10 n-channel lateral double diffused LDMOS transistor.

3.3.2 LDMOS : Lateral Double-diffused MOS

In order to increase power handling a lightly doped region needs to be introduced into the structure. This can be done in a vertical or a lateral structure. The first type considered is the lateral double-diffused MOS or LDMOS.

Double-diffused MOS structures use the difference in diffusion lengths of sequentially introduced impurities through a common masking level to define the channel. The channel length is defined in a similar way to the base width of a double-diffused bipolar transistor. The channel length and peak dopant are controlled by the dopant introduced at each step and the sequential diffusions. A typical device structure is shown in figure 3.10.

The two main differences between this device and a conventional MOS transistor are that the channel length is defined by two sequentially diffusions in the same direction and that the body region (channel in this case) is more heavily doped than the lightly doped drain region (substrate).

This heavily doped body means that the drain-to-body junction depletes further into the drain region than into the body region. This difference means that significantly higher voltages can be applied across the body-to-drain junction without affecting the electrical channel length of the transistor. This structure where the active channel region is separate from the region that sustains the drain-to-source voltage is similar to that used in double-diffused bipolar transistors.

Since their introduction a few years ago [20], LDMOS transistors have been used in several types of applications including microwave devices [19], high-voltage

switches [21], and as high-speed logic [22,23]

The device structure consists basically of an enhancement-mode transistor in series with a depletion-mode transistor (surface accumulation region), itself in parallel with a bulk resistance, in series with a second bulk resistor R_2 [24]. The middle two components model the gate controlled accumulation in parallel with a bulk resistor; the current through the device will divide between these two in a manner determined by the gate voltage (i.e. the conductance of the accumulation layer). The third series component R_2 models the region of the device between the gate electrode and the drain contact. The equivalent circuit is shown in figure 3.11.

In high voltage applications the contributions of the depletion-mode transistor and the bulk resistor R_1 to the total device resistance are generally small [24]. Therefore the depletion mode device can be neglected and R_1 can be included in the calculation of R_2.

The enhancement mode channel resistance is given by

$$R_E = \frac{W}{L_{EFF}} C_0 \mu_E (V_G)(V_G - V_{TE}) \qquad 3.41$$

The bulk resistor R_2 can be modelled by considering the current flowing from a line source with radius r_1, at the end of the channel to a line sink with radius r_2 at the n+ drain contact [25].

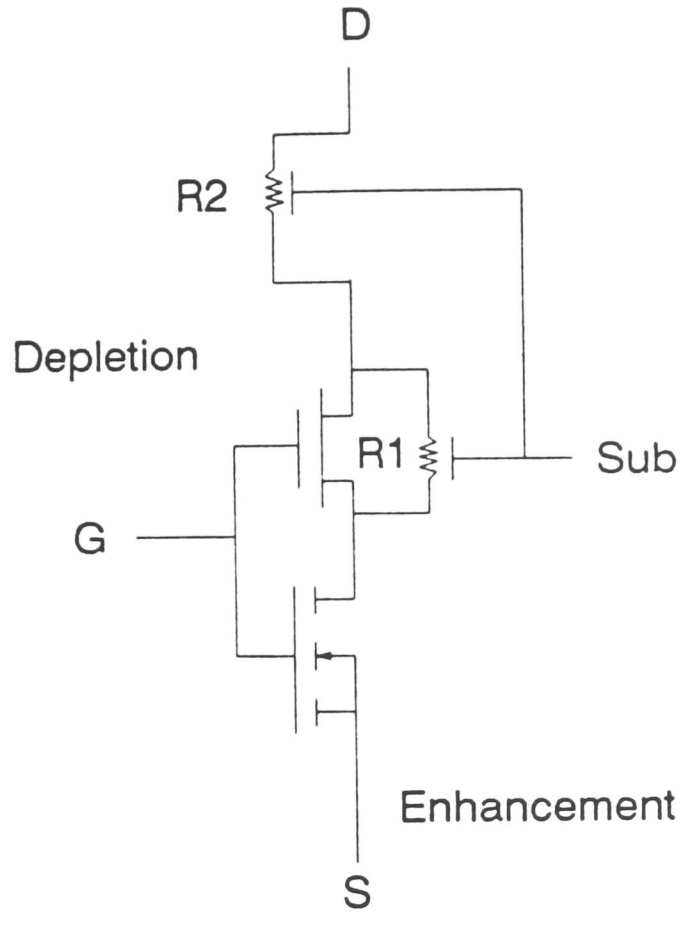


Figure 3.11 The equivalent circuit of the LDMOS transistor.

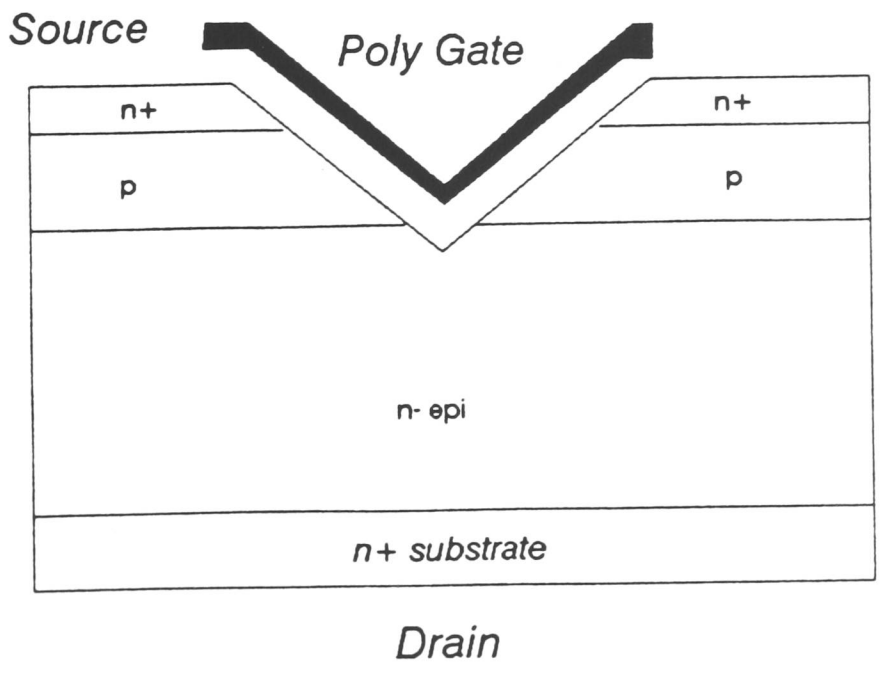


Figure 3.12 n-channel vertical V-groove VMOS transistor.

$$R_2 = \frac{\rho}{W\pi}\left[\ln\left(\frac{L' - r_1}{r_1}\right) + \ln\left(\frac{L' - r_2}{r_2}\right)\right] \qquad 3.42$$

In its simplest form then, the equivalent circuit of the LDMOS device is that of an enhancement-mode transistor in series with a bulk resistor.

It can be shown [26] , that the drain current for the LDMOS for $V_A > V_G$ is given by:

$$I_D = \frac{3\mu_E C_0 W}{2L}\left[(V_G - \frac{V_A}{2})V_A - (V_G - \frac{V_C}{2})V_C\right] \qquad 3.43$$

Where V_C and V_A are the potentials at the end of the channel and at the end of the accumulation layers respectively. When V_A reaches the saturation voltage, which is close to the gate potential value, the length of the accumulation region starts decreasing. The saturation value for the drain current is found by substituting $V_A = V_G$ into equation 3.43.

$$I_D = \frac{3\mu_E C_0 W}{4L_s}\left[V_G^2 - (2V_G - V_C)V_C\right] \qquad 3.44$$

Where L_s is the accumulation layer length up to the saturation voltage.

For the n-LDMOS of figure 3.10, in high-voltage operation, breakdown can be due to either the avalanching of the p-channel to the n+ drain diode or punchthrough of the p-n- depletion layer through the p-channel diffusion to the n+ source contact. If the gate material extends over the n+ drain region then the breakdown voltage is reduced. Similarly if the gate voltage is increased then the breakdown voltage is reduced accordingly [27].

$$BV_{DS} = BV_{DSS} + mV_G \qquad 3.45$$

where m is approximately 1

This effect is similar to pn junctions that have been covered by a field plate. However in the LDMOS device, the range of the gate voltage is such that partially covering the pn junction with the gate metal does not lower the surface breakdown of the junction, but helps to increase it by reducing the surface electric field in this region. It has been found [28] , that the origin of breakdown is localised in the n-n+ graded region, so that the gate-depletion region overlap has to be carefully controlled to optimise the breakdown value.

One of the limitations of the conventional LDMOS is the high on-resistance. If a semi-insulating polysilicon layer is deposited over the drift region oxide then this SIPOS layer acts to reduce the surface fields in the device. If the oxide thickness is sufficiently thin (< 1000 A) then a strong accumulation region can form at the surface of the drift region. This accumulation layer provides an

additional current path in parallel with the conduction in the drift-region bulk. This effect can reduce the on-resistance by a factor of five. However, there is a penalty in that the increased capacitance will reduce the operating speed of the device.

3.3.3 VMOS : V-groove MOS

The other possibility is to introduce the lightly doped region in a vertical structure. The vertical MOS or V-groove MOS is one possible solution. Lateral power transistors have several disadvantages in discrete device applications. The first of these is that the drain contact is on the top surface of the device. If the drain contact is made on the bottom as in discrete bipolar structures then the device will be more compact. The fact that the depletion region of the body-to-drain diode spreads laterally means that the surface area is used to support the applied voltage. If the depletion region spreads vertically, then less surface area will be required. Lastly many of the field shaping techniques used to allow higher voltages to be reached can only be applied to vertical structures. There are two basic vertical power transistors that can be used, the VMOS and the VDMOS structures. This subsection describes the VMOS.

The VMOS or V-groove MOSFET is a non-planar device. The fabrication of non-planar FETs involves either an anisotropic or an isotropic etching process to generate grooves in the silicon substrate. Anisotropic etching is used to etch V-grooves in (100) orientation silicon. If the etching of these grooves is accurately

aligned with the [110] direction on the surface of the slice then the etching process is self-stopping. The resulting walls of the groove slope down from the horizontal at 54.7° . This etch is normally done with wet chemistry [29].

The channel region of the VMOS transistor is first formed by the sequential diffusion of the body and source regions from the surface of the wafer. The channel region is next exposed with a preferential etch sufficiently deep to reach through the body of the device to the drain. A gate oxidation followed by gate deposition results in the VMOS structure of figure 3.12.

The current flow in the VMOS is due to carriers moving from the source to the drain through a channel at the surface of the V-groove, induced by a voltage on the gate. The channel length of a VMOS made with the same diffusion schedule as a LDMOS will be approximately one and a half times longer due to the 54.7° angle. The mobility in the inversion layer in a (111) surface is less than that along a (100) surface. There is also a three-fold increase in Q_{ss} (fixed oxide charge) due to the (111) orientation of the channel. This will increase the threshold voltage of the device.

The equivalent circuit for the VMOS device is given in figure 3.13. It is identical to that of the LDMOS in figure 3.10, except that there is no influence on R_1 and R_2 due to the substrate as the VMOS is an isolated discrete device. The thin gate oxide extends down to the bottom of the groove, resulting in a surface accumulation layer under the gate which minimises current-crowding effects at the end of the enhancement channel and reduces overall on-resistance. This region should be modelled as a distributed network of depletion-mode transistors in series

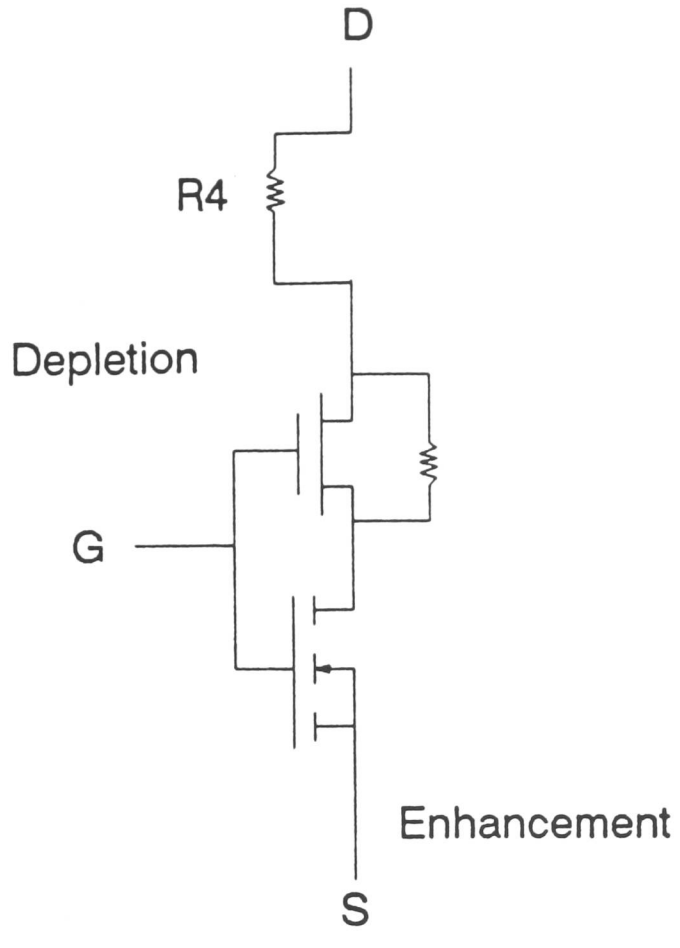


Figure 3.13 The equivalent circuit of the VMOS transistor.

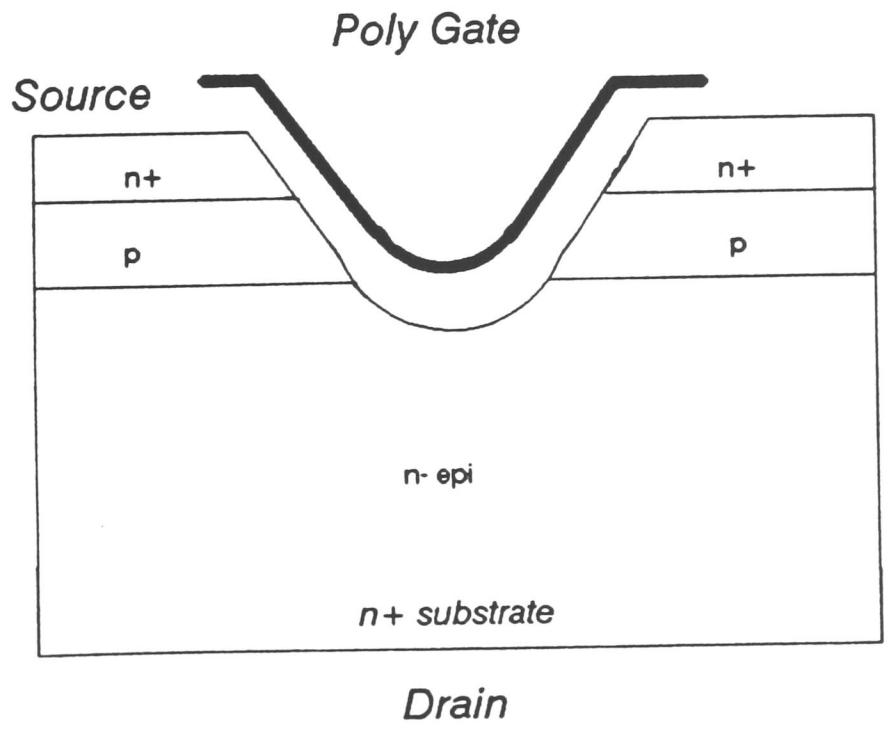


Figure 3.14 n-channel vertical U-groove UMOS transistor.

with bulk resistors, but for simplicity is usually modelled as a single transistor and a series resistor. The total on-resistance of the VMOS structure is given by:

$$R_{ONvmos} = R_E + R_D + R_3 + R_4 \qquad 3.46$$

R_E is the enhancement-mode device on-resistance and is given by the same expression as used for the LDMOS device.

The depletion-mode device on-resistance is given by:

$$R_D = \frac{1}{3}\left[\frac{W}{L'_{EFF}} C_0 \mu_D (V_G)(V_G - V_{TD})\right]^{-1} \qquad 3.47$$

The factor of a third appears due to the two-dimensional nature of the current flow from the accumulation layer into the bulk n-region and has been verified experimentally [30,31].

The resistance R_3 can be calculated by assuming that the geometry is a sector of a circle with an included angle of 54.7° . The current is assumed to uniformly enter this region, and integration yields :

$$R_3 = 0.477 \frac{\rho}{W} \qquad 3.48$$

The resistance R_4 of the bulk resistor can be shown to be

$$R_4 = \frac{\rho}{W}\frac{1}{\tan\alpha}\ln(1+2\frac{h}{a}\tan\alpha) \qquad 3.49$$

Where $\alpha = 28° - \frac{h}{a}$ if $h \geq a$

and $\alpha = 28° - \frac{a}{h}$ if $h < a$

Where a is the distance between adjacent n+ source regions and h is the distance into the bulk from the bottom of the groove (as in figure 3.11).

There can be an additional on-resistance component due to the constriction in carrier flow at the notch of the groove. This can be avoided by using the truncated V-groove or U-groove structure as in figure 3.14. This requires careful control of etching to create the correct profile.

3.3.4 VDMOS

The major disadvantage of the VMOS is the non-planar structure. An alternative vertical structure , the VDMOS or Vertical Double-diffused MOS transistor has been more readily adopted commercially. Like the VMOS the current flows from a drain contact on the back plane of the device to the source at the surface under the voltage control of the gate electrode.

The structure is shown in figure 3.15, where the carriers flow from the source through the channel at the surface of the body region, to the lightly-doped drain region under the influence of voltages on the gate and drain. However, when the electrons reach the lightly doped drain region below the gate, they flow away from the surface of the silicon towards the more heavily doped drain region.

The equivalent circuit for the VDMOS device is shown in figure 3.16. This equivalent circuit is similar to that of the LDMOS, but differs in two ways. Firstly, the depletion-mode transistor and its parallel bulk resistor must be regarded as distributed devices as the current flow changes in direction from horizontal to vertical along the length of the accumulated surface between the source regions. Secondly, a JFET is present in the equivalent circuit. This arises physically because of the pinching of the current between the adjacent p bulk diffusions. The spacing between these diffusions (channel width) can be of great importance in determining the fraction of the devices total on-resistance contributed by the series JFET.

If contact resistances are neglected then the VDMOS on-resistance is given by

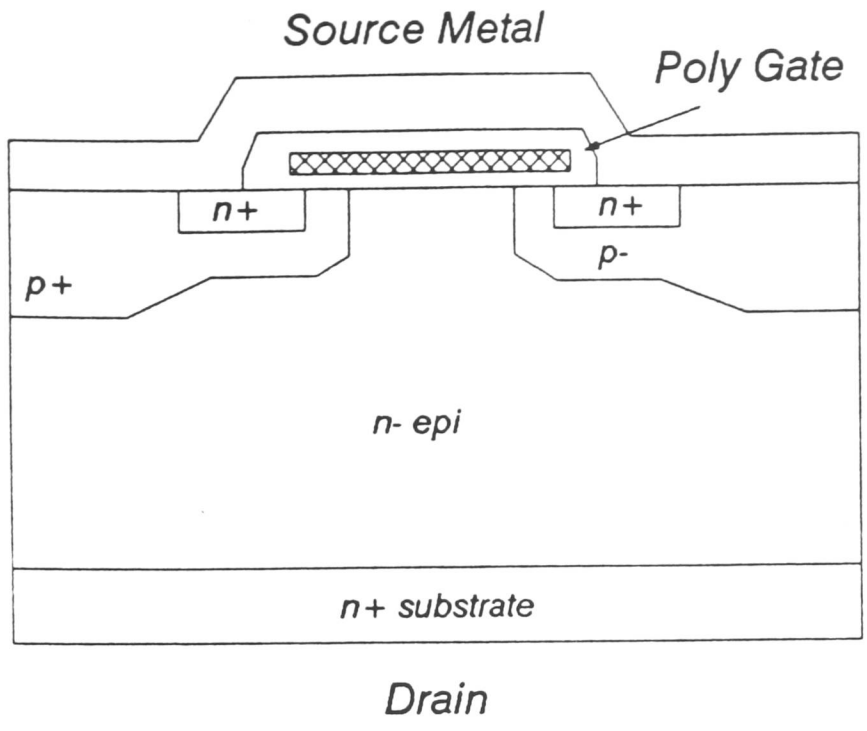


Figure 3.15 n-channel vertical double diffused VDMOS transistor.

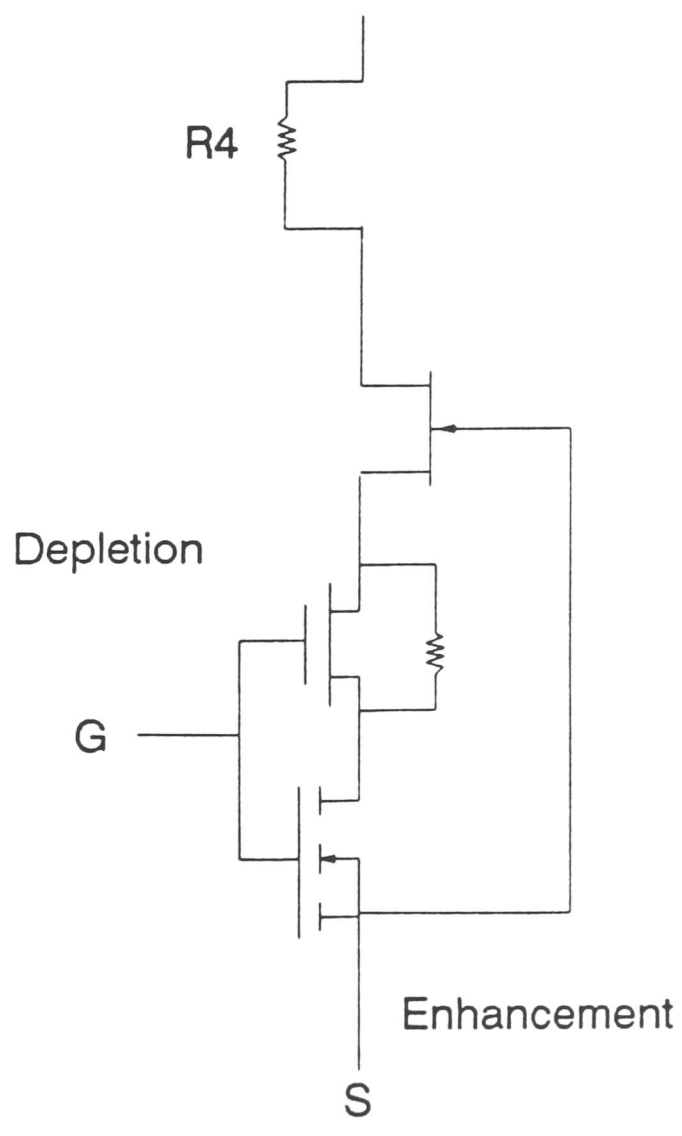


Figure 3.16 The equivalent circuit of the VDMOS transistor.

$$R_{ONvdmos} = R_E + R_D + R_{JFET} + R_4 \qquad 3.50$$

Where R_E is given by the same expression as in the modelling of the LDMOS device. R_D and R_4 are both given by the same expressions as in the VMOS case.

The JFET component of the VDMOS equivalent circuit is modelled as a portion of a circle with the origin offset from the original mask edge by 0.15 x_j to take into account the difference in vertical and lateral diffusion. The current is assumed to originate uniformly from the surface accumulation layer and it has been shown that the resistance is given by [24] :

$$R_{JFET} = 2\frac{\rho}{W}\left[\frac{1}{\sqrt{(1-(2X_j/L)^2)}}\tan^{-1}\sqrt{\frac{2(L+2X_j)}{(L-2X_j)}} - \frac{\pi}{8}\right] \qquad 3.51$$

The VDMOS structure has been found to be more area-efficient for discrete devices in terms of R_{on} per unit area for applications above 100 Volts than the LDMOS structure. This is due to the drain contact being on the bottom surface. However, when integrated devices are needed then this is not always an advantage. Additional improvements in on-resistance have been reported by using technologies normally used in advanced low voltage processes such as LPCVD tungsten to lower contact resistivity and therefore increase current handling [32].

3.4 Other Power Mosfets

3.4.1 Introduction

Improvements in these three basic structures have been made, mainly in the areas of on-resistance per unit area and high frequency performance. They are versions of the original lateral and vertical power mosfets, but in this review they will be referred to by their separate names.

3.4.2 RMOS : Rectangular-groove MOS

The on-resistance of vertical power MOSFETs with high source-to-drain breakdown voltage is predominantly determined by the resistance of the lightly doped epitaxial drain region. By having rectangular grooves the RMOS device reduces the resistance of the channel region. The channels of the RMOS transistor are formed along the vertical walls of rectangular grooves that penetrate into the n- buffer region through n+ source and p-body regions as in figure 3.17. The channel length is shorter and packing density is higher than that of conventional VMOS and DMOS structures. As a result the channel resistance per unit area is lower than in previous devices. A reactive ion-beam etching (RIBE) technique is employed to form the rectangular grooves, since it permits anisotropic etching [33].

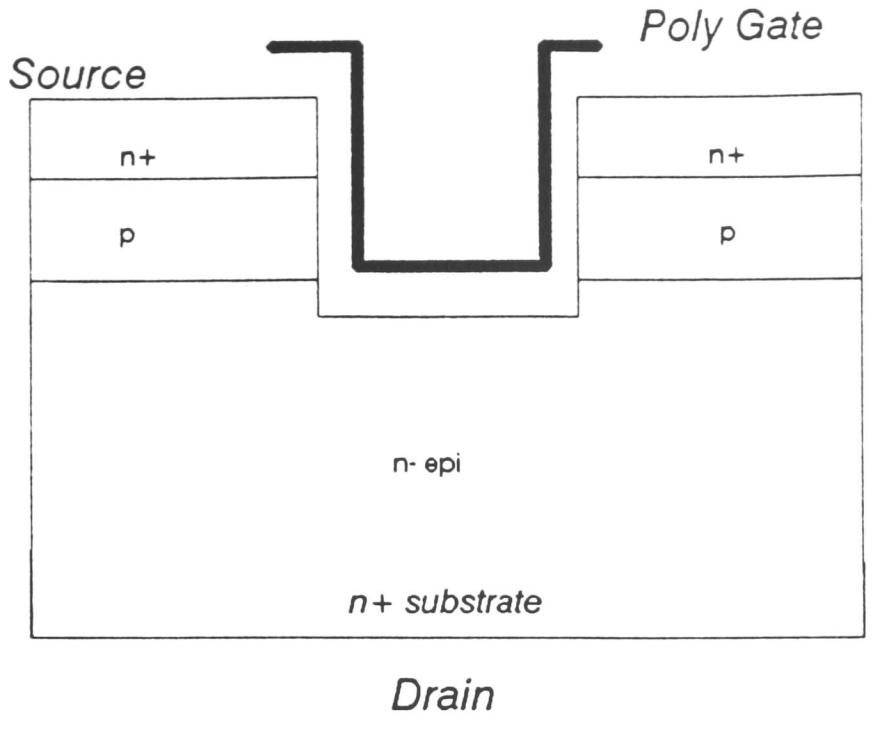


Figure 3.17 n-channel vertical R-groove RMOS transistor.

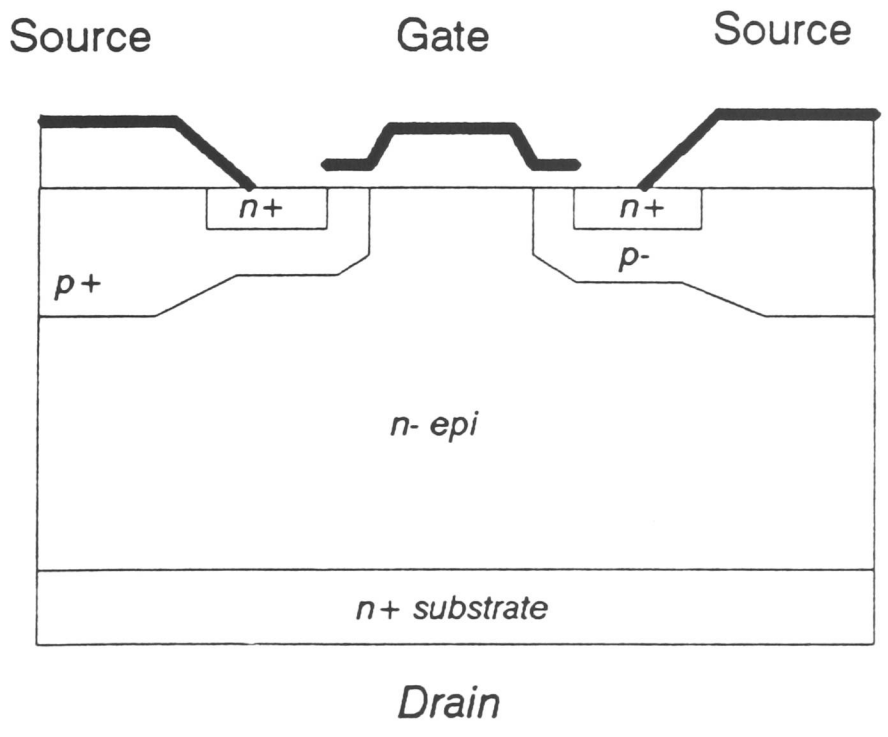


Figure 3.18 n-channel self-aligned terraced gate STGMOS transistor.

3.4.3 STGMOS : Self-aligned Terraced Gate MOS

The high-frequency performance of conventional vertical double-diffused power MOSFETs are poor, with typical cut-off frequencies of several megahertz. This low cut-off frequency is primarily caused by the large Miller capacitance resulting from the thin gate oxide and the relatively high gate resistance of the self-aligned polysilicon process.

An alternative approach is with an aluminium gate, instead of polysilicon, that permits reduction of Miller capacitance by forming a thick oxide on the gate-drain overlapping region. The new MOSFET called STGMOS offers a cut-off frequency of ten times higher than a conventional power MOSFET, but requires a more complex fabrication process (see figure 3.18).

In the STGMOSFET the oxide film under the gate electrode has a different thickness than over the source, body and drain regions. A self-aligned terraced-shaped thick oxide layer over the drain region reduces the gate-to-drain Miller capacitance. The oxide film over the source region is three times thicker than the oxide over the channel region. This reduces the device's input capacitance. The unique gate structure permits significant reductions in parasitic capacitances resulting in a cut-off frequency as high as 100 MHz for some devices [34].

3.4.4 DSCMOS : Drain Separated from Channel MOS

Another lateral device is the grooved gate MOSFET of figure 3.19, with its drain separated from the channel-implanted regions (DSC structure). This structure obtains higher breakdown voltages: drain sustaining voltage and highest applicable voltage . Non-implanted regions between channel implanted and source/drain regions are a unique feature of the device structure. The self-aligned non-implanted region in the channel is obtained by using silicon dioxide and resist overhangs. These overhangs are fabricated by grooving the silicon substrate. The non-implanted region introduces a resistance in series with the channel and drain that allows higher drain terminal voltages by reducing the electric field near the junction. However, the relatively large capacitance this introduces will limit device switching speeds in use [35].

3.5 Merged MOS-Bipolar structures

There is a new class of power semiconductor devices that combine the advantages of MOS gate drive simplicity with the current handling capability of bipolar devices. Bipolar operation is achieved by the injection of minority carriers into the bulk of the device through a forward biased p-n junction. If the device is sufficiently thin, compared to the minority carrier diffusion length, the device will be filled by an electron-hole plasma resulting in an increase of its conductivity with a strong reduction of its on-state voltage drop.

This conductivity modulation principle has been applied to MOS gated devices by adding a p-n junction at the drain side of the device [36,37] , and to JFET devices by using the gate electrode to inject minority carriers [38,39].

The IGT (Insulated Gate Transistor) or COMFET (Conductivity Modulated FET) cross-section of figure 3.20 shows that the structure is very similar to that of a VDMOS power transistor. The difference is that the substrate is p+ instead of n+, adding an extra p-n junction to the structure. When the device is turned on, electrons pass into the n- region lowering its potential and forward biasing the p+ / n- junction. This causes holes to be injected from the p+ substrate into the n- epi layer. The excess electrons and holes modulate the conductivity of the high-resistivity n-region, which dramatically reduces the on-resistance of the device.

The main disadvantage of the IGT is a larger fall time than power MOS devices, but this time is still acceptable for a large number of applications. However it is possible to improve the fall-time in IGTs by using techniques such as electron-beam irradiation [37] and doping with life-time killers [40].

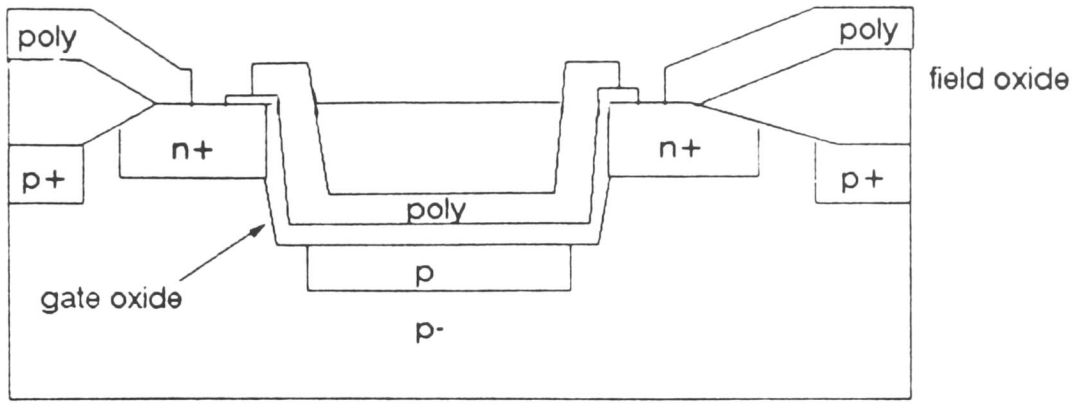


Figure 3.19 n-channel drain separated from channel DSCMOS transistor.

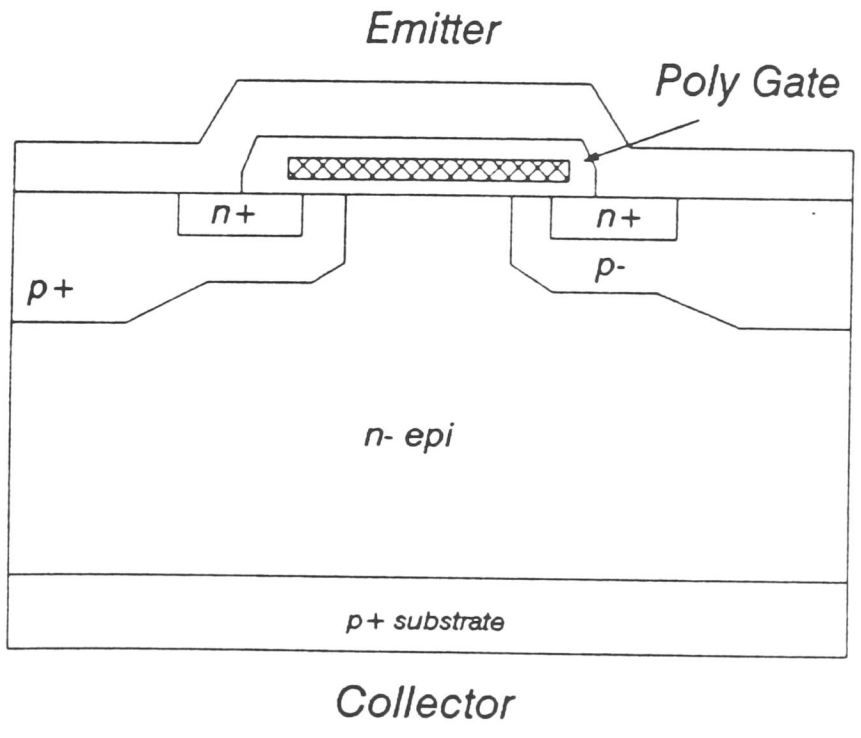


Figure 3.20 Conductivity modulated COMFET structure.

3.6 Conclusions

In this chapter the fundamental MOS principles of operation and the resulting equations have been discussed. The two main differences between MOS and bipolar transistors are that MOS devices involve majority carriers for conduction and use input voltage to control output current, whilst bipolar devices involve majority and minority carriers for conduction and use input current to control output current.

Although proposed first, the MOS transistor was not realised in silicon until much later than its bipolar equivalent. Despite this its use has grown and since 1975, MOS sales have overtaken bipolar sales. The main reasons for this are the ease of integration, lower power consumption and smaller dimensions. Although MOS transistors can be used in the linear region the majority of applications are as a switching element. Its inherent capacitance is also exploited in memory circuits.

By introducing a high resistance region in series with the simple MOS transistor a power device can be formed. There are many solutions merging these two areas either laterally or vertically. The best structure depends on the power range required. For voltages up to 60V lateral devices are more efficient, above 200V vertical devices dominate. In the middle range the choice is not straight forward.

A late comer to the scene is a merged-MOS-bipolar power switch the COMFET or IGT. This device is particularly suited for low switching speed, discrete $>$ 200V applications. It has two major drawbacks for use in smart power applications. The first is the large fall time, limiting operating speed. The second

is the difficulty in integrating the structure, in order to have all terminal connections on the surface. This second point is a pre-requisite for real smart power applications, where only one power transistor and a few logic transistors are needed.

In comparison with bipolar devices it can be noted that for low voltage applications the MOS solution results in a much more compact transistor. This leads to the dominance of MOS over bipolar in solutions where a large number of transistors are required. For high voltage solutions it can be seen that the two solutions are similar in terms of area used. However bipolar devices have an intrinsically lower on-resistance. For this reason it is bipolars that dominate single transistor and simple power integrated circuit applications.

For smart power solutions where both low voltage logic and high voltage power switches are needed, the ideal solution is MOS logic and bipolar power transistors. This results in a complex and costly process which may be suitable for large volume runs or clients with big budgets, which are not too common in this type of application. There exists a large niche market for suppliers who can put logic and power on the same chip in the same technology, bipolar or MOS, at a competitive price. The choice between bipolar and MOS here depends on how much logic and how much power is required. It was seen in chapter two that both low and high voltage bipolar transistors have similar structures. However, as they operate in different regimes the doping concentrations and junction depths must be different for both types. This means that in order to create an entirely bipolar smart power process a large number of processing steps are needed.

Contrastingly, it has been seen in this chapter that for MOS technology the low and high voltage transistors are different in structure. It is known that for a full smart power process where more than one power transistor is needed, all terminal connections must be in the top surface of the device. The power transistor structure that most easily satisfies this condition is the LDMOS. It can be seen that if the source, the highly doped part of drain and the gate can be formed with the same processing steps as the low voltage transistors, then a smart power process with a minimum number of mask levels can be realised. Further, as there is a large market for existing logic parts with "bolted on" power transistors for applications such as automotive, then it was decided to develop a smart power process that was compatible with an existing low voltage CMOS process.

References

[1] B.Santo, "Technology '89 : Solid State", IEEE Spectrum, vol.26, pp.47-49, Jan.1989.

[2] M.Miyake, T.Kobayashi and Y.Okazaki, "Subquarter-Micrometer Gate-Length p-Channel and n-Channel MOSFETs with Extremely Shallow Source-Drain Junctions", IEEE Trans. Electron Devices, vol.ED-36, pp.392-398, 1989.

[3] J.Mavor, M.A.Jack and P.B.Denyer, "Introduction to MOS LSI Design", pp.18-39, Addison-Wesley, London, 1983.

[4] M.B.Barron, "Low Level Currents in Insulated Gate Field Effect Transistors", Solid-State Electronics, vol.15, pp.293-302, 1974.

[5] W.M.Gosney, "Subthreshold Drain Leakage Currents in MOS Field-Effect Transistors", IEEE Trans. Electron Devices, Vol.ED-19, p.213, 1972.

[6] J.R.Brews, "A Charge-Sheet Model of the MOSFET", Solid-State Electronics, vol.21, pp.345-355, 1978.

[7] W.Fichtner and H.W.Potzl, "MOS Modelling by Analytical Approximations. I. Subthreshold Current and Threshold Voltage", International Journal of Electronics, vol.46, p.33, 1979.

[8] S.C.Sun and J.D.Plummer, "Electron Mobility and Accumulation layers on Thermally Oxidized Silicon Surfaces", IEEE Trans. Electron Devices, vol.ED-27, pp.1497-1508, 1980.

[9] H.C.Lin, "Integrated Electronics, p.110, Holden-Day, San Fransisco, 1967.

[10] V.G.C.Reddi, "Majority carrier surface mobilities in thermally oxidized silicon", IEEE Trans. Electron Devices, vol.ED-15, pp.151-160, 1968.

[11] R.F.David, "Computerized thermal analysis of hybrid circuits", ECC, pp.324-332, 1977.

[12] D.Sharma, J.Gautier and G.Merckel, "Negative dynamic resistance in MOS devices", IEEE J. Solid-State Circuits, vol.SC-13, pp.378-380, 1978.

[13] L.Vadasz and A.S.Grove, "Temperature Dependence on MOS Transistor Characteristics Below Saturation", IEEE Trans. Electron Devices, vol.ED-13, pp.863-866, 1966.

[14] R.Wang, J.Dunkley, T.A.DeMassa and L.F.Jelsma, "Threshold Voltage Variation with Temperature in MOS Transistors", IEEE Trans. Electron Devices, vol.ED-18, p.386, 1971.

[15] R.W.Bower, H.G.Dill, K.E.Aubuchon and S.A.Thompson, "MOS Field Effect Transistors Formed by Gate Masked Ion Implantation", IEEE Trans. Electron Devices, vol.ED-15, pp.757-761, 1968.

[16] H.G.Dill, "A New Insulated gate tetrode with High Drain Breakdown Potential and Low Miller Feedback Capacitance", IEEE Trans. Electron Devices, vol.ED-15, pp.717-728, 1968.

[17] R.A.Blanchard, P.A.Gargini, G.A.May and R.D.Melen, "High-Voltage Simultaneous Diffusion Silicon-Gate CMOS", IEEE J. Solid-State Circuits, vol.SC-9, pp.103-110. 1975.

[18] I.Yoshida, M.Kubo and S.Ochi, "A High Power MOSFET with Vertical Drain Electrode and a Meshed Gate Structure", IEEE J. Solid-State Circuits, vol.SC-11, pp.472-477, 1976.

[19] H.J.Sigg, G.D.Vendelin, T.P.Cauge and J.Kocsis, "D-MOS Transistor for Microwave Applications", IEEE Trans. Electron Devices, vol.ED-19, pp.45-53, 1972.

[20] Y.Tarui, Y.Hayashi and T.Sekigawa, "Diffusion Self-Aligned Enhance-Depletion MOS IC", Supplement J.Jpn. Soc. Appl. Physics, vol.40, pp.193-198, 1971.

[21] J.D.Plummer, J.D.Meindl and M.G.Maginness, "An ultrasonic imaging system for real-time cardiac imaging", Int. Solid-State Circuits Conf., pp.162-163, Feb. 1974.

[22] Y.Tarui, Y.Hayashi and T.Sekigawa, "Diffusion self-aligned MOST : A new approach for high speed device", Proc. 1st. Conf. on Solid State Devices, pp.105-110, Tokyo, 1969.

[23] M.D.Pocha, A.G.Gonzalez and R.W.Dutton, "Threshold Voltage Controllability in Double-Diffused MOS Transistors", IEEE Trans. Electron Devices, vol.ED-21, pp.778-784, 1974.

[24] S.C.Sun and J.D.Plummer, "Modelling of the On-Resistance of LDMOS, VDMOS and VMOS Power Transistors", IEEE Trans. Electron Devices, vol. ED-27, pp.356-366, 1980.

[25] M.D.Pocha and R.W.Dutton, "A Computer-Aided design Model for High-Voltage Double Diffused MOS (DMOS) Transistors", IEEE J. Solid-State Circuits, vol.SC-11, pp.718-726, 1976.

[26] S.Colak, "Effects of Drift Region Parameters on the Static Properties of Power LDMOST", IEEE Trans. Electron Devices, vol.ED-28, pp.1455-1466, 1981.

[27] M.J.DeClerq and J.D.Plummer,"Avalanche breakdown in High-Voltage DMOS Devices", IEEE Trans. Electron Devices, vol. ED-23, pp.1-4, 1976.

[28] A.S.Grove, O.Leistiko Jr. and W.W.Hooper, "Effects of Surface Fields on the Breakdown of Planar Silicon P-N Junctions", IEEE Trans. Electron Devices, vol. ED-14, p.157, 1967.

[29] J.B.Price, "Anisotropic etching of Silicon with KOH-H_2O-isopropyl alcohol", Semiconductor Silicon, pp.339-353, The Electrochemical Society, 1973.

[30] A.B.Phillips, Transistor Engineering, p.216, McGraw-Hill, New York, 1962.

[31] S.E.-D.Habib, "The ALDMOST : A New Power MOS Transistor", IEEE Electron Device Letters, vol.EDL-8, pp.257-259, 1987.

[32] K.A.Shenai, P.A.Piaceto, R.Saia and B.J.Baliga, "Blanket LPCVD Tungsten Silicide Technology for Smart Power Applications", IEEE Electron Device Letters, vol.EDL-10, p.270, 1989.

[33] D.Ueda, H.Takagi and G.Kano, "A New Vertical Power MOSFET structure with Extremely Reduced On-Resistance", IEEE Trans. Electron Devices, vol, ED-32, pp.2-6, 1985.

[34] D.Ueda, H.Takagi and G.Kano, "A New Vertical Double-Diffused MOSFET : The Self-Aligned Terraced-Gate MOSFET", IEEE Trans. Electron Devices, vol, ED-31, pp.416-420, 1984.

[35] E.Takeda, K.Hume and S.Assai, "New Grooved-Gate MOSFET with Drain Separated from Channel Implanted Region", IEEE Trans. Electron Devices, vol. ED-30, pp.681-686, 1983.

[36] Y.Tarui, Y.Hayashi, T.Sekigawa, J.P.Russell, A.M.Goodman, L.A.Goodman and J.M.Neilson, "The COMFET - A New High Conductance MOS-Gated Device", IEEE Electron Device Letters, vol.EDL-4, pp.63-65, 1983.

[37] B.J.Baliga, "Fast-Switching Insulated Gate Transistors", IEEE Electron Device Letters, vol.EDL-4, pp.452-454, 1983.

[38] A.Caruso, P.Spirito and G.Vitale, "New Semiconductor Active Device : The Conductivity-Controlled Transistor", Electronics Letters, vol.15, pp.267-268, 1979.

[39] S.Bellone, A.Caruso, P.Spirito, G.Vitale, G.Busatto, G.Cocorullo, G.Ferla and S.Musumeci, "High-Voltage Bipolar Mode JFET With Normally Off Characteristics", IEEE Trans. Electron Device Letters, vol.EDL-6, pp.522-524, 1985.

[40] M.Melito, "Power Switches For Automotive Ignition", IEE Colloquium : Power Integrated Circuits, p.7/1, March 1989.

CHAPTER FOUR : SIMULATION

4.1 Introduction

With the advent of VLSI, computer simulation has become a necessity. No longer is it possible for the designer of today to rely on human experience as the numbers of transistors per device increase by one order of magnitude every four years. He has to rely on computer tools to solve some of the problems for him. Computers can be used for process, device and circuit simulation, for measurement control, for statistical analysis, circuit layout and computer-aided-manufacturing. In this chapter we consider their role in simulation.

With the reduction in device geometries it is no longer intuitively obvious how transistors will perform. The cost and time of processing test chips to verify developments means that this route can only be used when absolutely necessary. It is imperative that new chips work on first pass, or at worst on the second attempt to keep the commercial edge on the competition. Therefore the simulator's role has changed from being a research tool to being a useful production and design tool.

In this section we shall consider process, device and circuit simulators and the interaction between each, as none of these areas can be considered independently any more. The packages discussed here are the ones used in later chapters of this project. Newer developments such as the drive to faster simulations through parallel processing [1], and the integration of process simulation with computer aided manufacturing have not been used in this project and will therefore not be considered in this review.

4.2 Process Simulation

Process modelling is now a mature simulation tool . Initially it was considered to be an academic exercise conducted mainly within the American academic community, particularly Stanford University, but now is used in virtually every commercial organisation [2].

Process modelling has become important due to the increasing complexities of device fabrication. With devices consisting of more than fifty processing steps, and the subsequent interactions between steps, it is important to be able to predict doping profiles and layer thicknesses before any transistors are fabricated. This is due to the fact that each iteration in silicon can take eight weeks and cost £100k or more. Therefore if the initial predictions can be made more accurately, the development cost is cut and the product can be introduced to the market much more rapidly.

This commercial need for increasingly more accurate and complex process simulation has resulted in several computer programs addressing the problem. It has also lead to more sophisticated process models within the programs. In this section the most well known and used simulators are reviewed, from the pioneering SUPREM to its state-of-the-art two-dimensional relative SUPREM 4-B. Also included is a description of one of the lithography simulators available, DEPICT.

4.2.1 SUPREM II

SUPREM [3], the Stanford University Process Engineering Models is the best known computer program capable of simulating IC fabrication steps. The output of the program consists of one-dimensional profiles of all the dopants present in the silicon and silicon-dioxide. These profiles can be output either numerically or graphically.

The models implemented in SUPREM are :

a) Ion Implantation

b) Chemical Pre-deposition through the surface

c) Oxidation / Drive-in

d) Epitaxial growth

e) Etching

f) Oxide deposition

In SUPREM, different impurity species can be modelled as a discrete profile with a maximum number of four hundred grid points. Each concentration value corresponds to a point in the discrete space defined along a vertical axis, with its origin at the surface of the material; Si or SiO_2.

The program uses a finite difference numerical solution to the problem and splits the grid points into three uniformly spaced regions. The grid spacing is usually different within each of these regions. The first region represents the oxide and is allocated between ten and fifty points automatically by the program. The number of points in this region is zero if no oxide is present.

The silicon distribution is contained within the next two regions. These

regions differ only in their grid spacings, the second region having a much finer mesh than the third. This second region usually has a higher resolution grid as this is where the steepest concentration gradients are encountered. The third region models the bulk silicon and therefore its modelling is not so critical. Often during a processing step, the physical dimensions of the simulated discrete space may change , as happens during oxidation, etching, deposition and epitaxy, and the grid has to be automatically altered using a cubic spline interpolation routine.

It has been found that several of the models present in SUPREM are inadequate for quantified analysis of real device structures [4]. By comparing simulated profiles with measurements obtained by spreading resistance and SIMS techniques, some significant differences were noted [4]. Therefore, some modifications have been made which, for example, allow better modelling of single boron implants. In this modified version, the implant profile is calculated from stored tables containing LSS data (after Lindhard, Scharff and Schott) and the profile tail is simulated by an exponential weighting technique dependent upon ion energy and channelling options. The inert annealing of high temperature boron uses a modified diffusion model which accommodates for clustering effects, a phenomenon which results in a static electrically inactive near-surface peak, and is only evident for high concentrations ($> 2.0\ 10^{15}\ cm^{-2}$). As an example of SUPREM II output, figure 4.1 shows the net doping concentration after a boron implant into a phosphorous substrate. In this case it can be noted that the silicon-oxide interface has moved into the silicon (and is now at 0.2μm depth). This is due to the consumption of silicon during the oxidation steps. The junction is at a depth of 0.45μm, but the graphics of this program version make it difficult to read of this value.

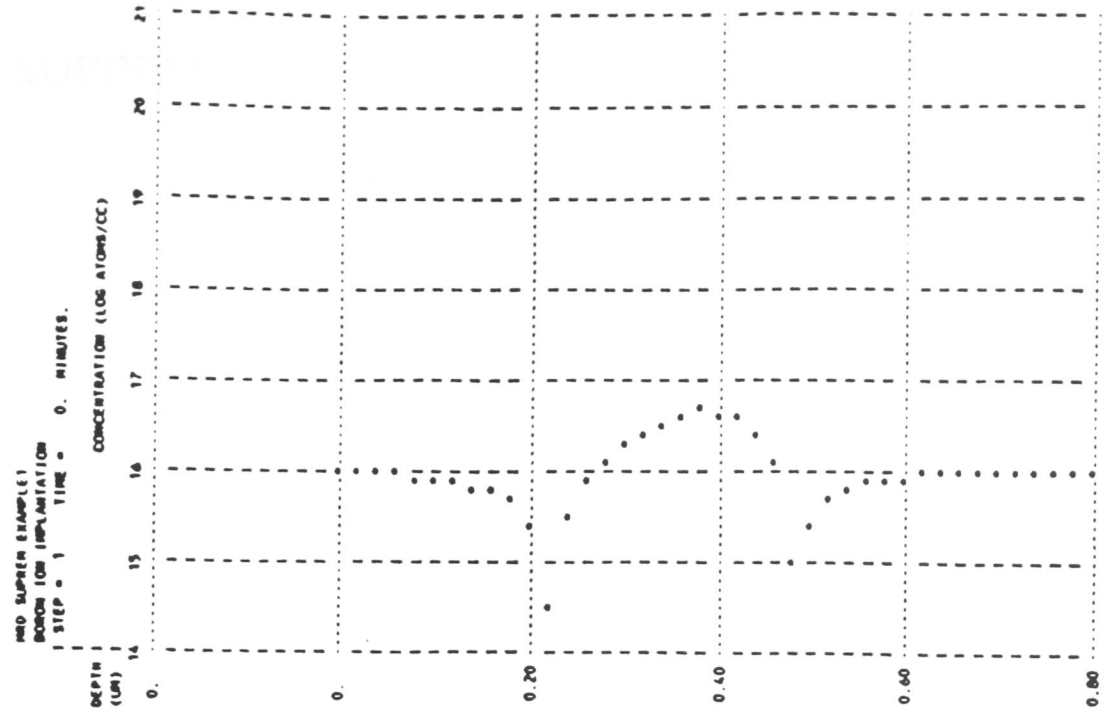


Figure 4.1 A simulation of the net doping concentration of a boron implant into a phosphorous substrate using SUPREM II.

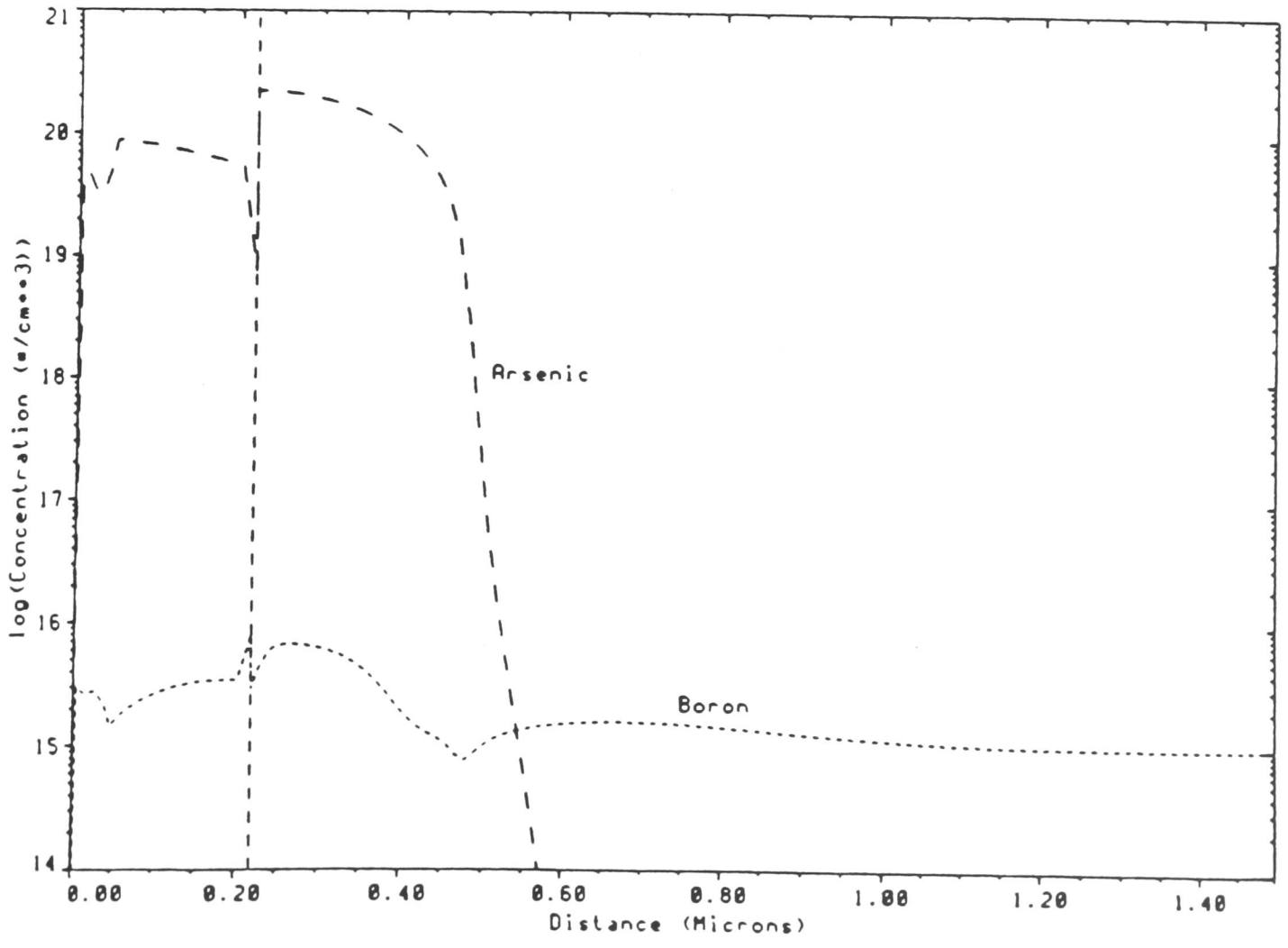


Figure 4.2 A SUPREM III simulation of the Arsenic Source/Drain implant for a nmos transistor produced using the EMF's 5μm CMOS process.

4.2.2 SUPREM III

The third generation of process simulators from Stanford, SUPREM III substantially upgrades the modelling capabilities. It also allows up to ten material layers, instead of the two allowed in the previous version. It can now model Si_3N_4 and polysilicon as well as Si and SiO_2. Also available are five user-defined materials. Six impurities are allowed ; antimony, arsenic, boron, phosphorus and two user-defined materials. Significantly, newer and more accurate models for oxidation, diffusion, epitaxy and ion-implantation are included [5].

The grid in SUPREM III is more flexible than earlier versions. There are now five hundred nodes and there can be any number of grid sections in each material layer leading to a highly non-uniform grid which can handle moving material boundaries.

Unfortunately the models are based on a small subset of the data available for each processing step. These models have been developed in isolation from each other, and interactive effects between subsequent processing steps are unaccounted for [6]. This means that a good understanding of the physics involved is needed to obtain valid simulations. With the models given it is possible to alter coefficients to 'fit' a process, but this method can only be a short term approach. More accurate models still are needed before accurate process modelling becomes easily accessible.

Figure 4.2 shows the source doping for an nmos transistor produced with the EMF's 5μm process discussed in the next chapter. The region nearest the surface is oxide, the rest is silicon. There has been an arsenic source doping and

a boron threshold adjust implant. The junction can be seen at 0.54μm. There is a significant improvement in the graphics output of this program over SUPREM II. Figure 4.3 shows an extrapolation to predict the threshold voltage an nmos device without the threshold adjust implant. If the implant is added then the predicted threshold voltage is 1.4 Volts. The calculation of threshold voltage is based on the channel dopant and does not take into account short-channel effects.

This simulator is probably the most widely used in the industry. This is due to various factors. It is relatively user friendly, it can be used to 'fit' individual processes and there is an easy interface between it and device simulators. For these reasons it was the process simulator most widely used during this work.

4.2.3 ICECREM

ICECREM is similar to SUPREM in that it is a one-dimensional process simulator. It is capable of modelling implantation, wet and dry oxidation, diffusion and etching. The models used are similar to that of SUPREM II with the exception of the ion-implantation model. This is modelled using a Pearson distribution instead of look-up tables.

The main difference between SUPREM II and ICECREM is that ICECREM can be run interactively allowing the user to see the effect of each processing step as it happens. Its main limitations are that it cannot model silicon nitride or polysilicon and that some model coefficients have to be altered in order to fit real

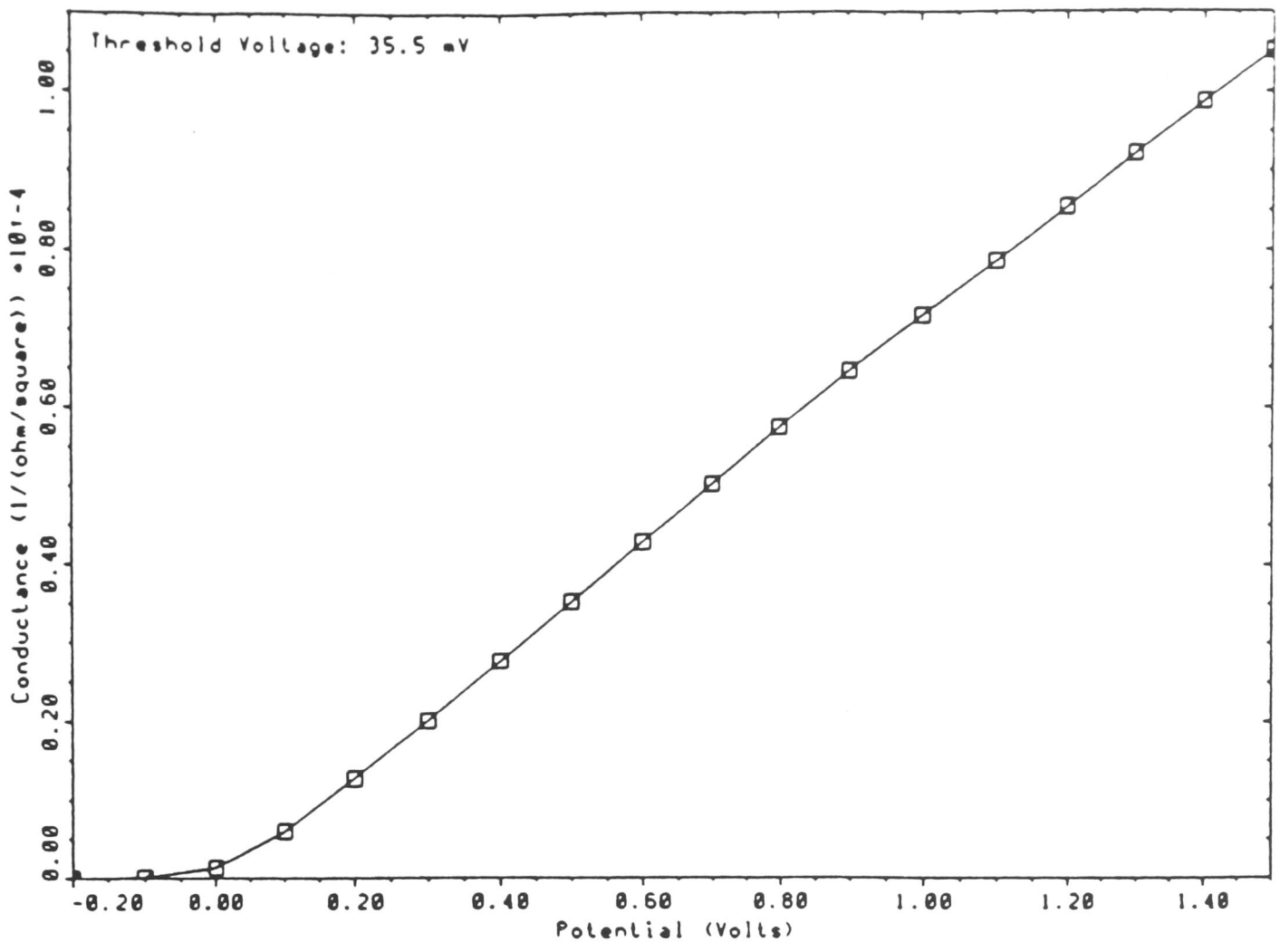


Figure 4.3 A SUPREM III extrapolation of the threshold voltage of a nmos transistor without the threshold adjust implant.

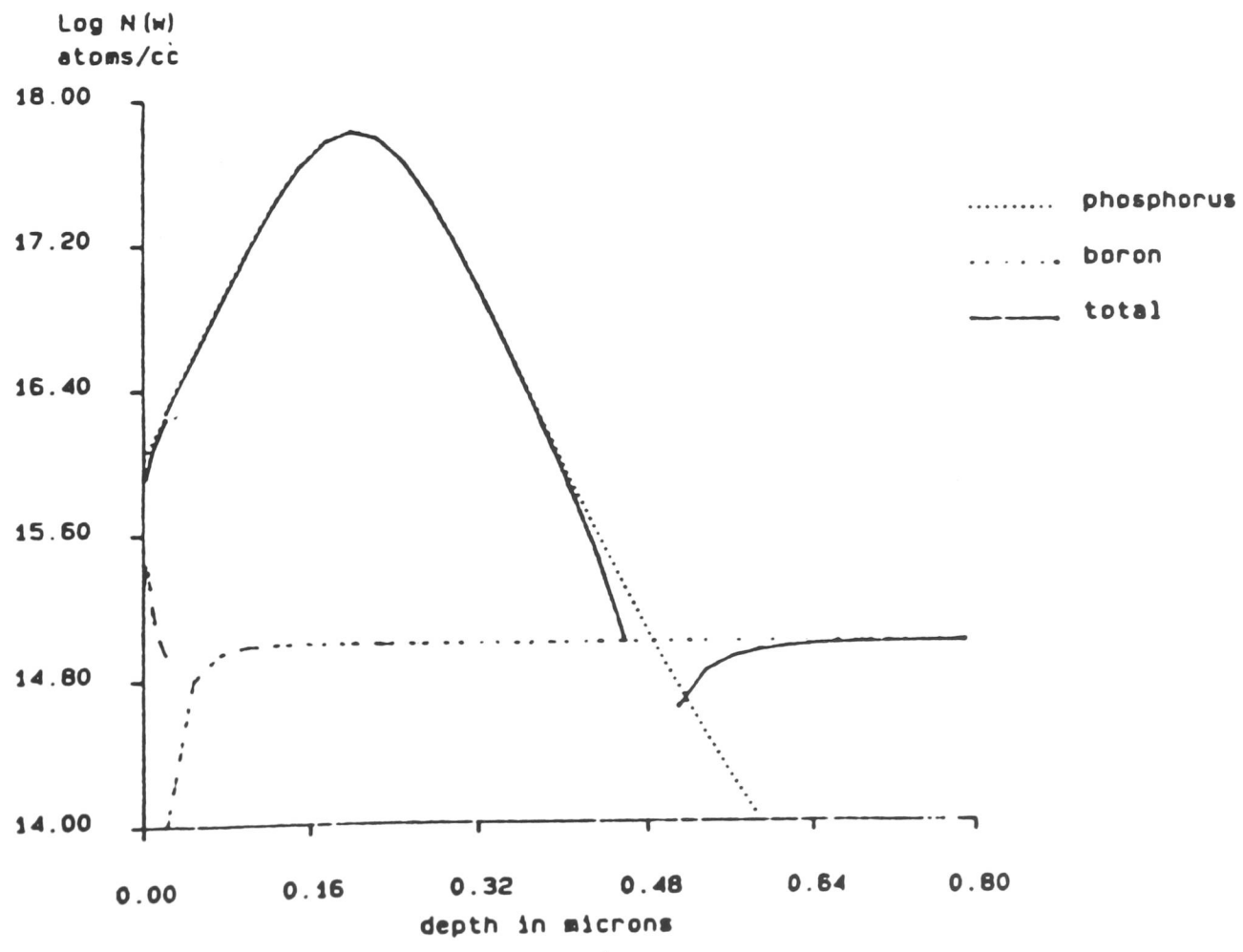


Figure 4.4 An ICECREM simulation of a phosphorus implant into a boron substrate used as the Source/Drain for the EMF's 6μm NMOS process.

data [7]. These features mean that it can be made to model well the processes of the EMF for example, but the main limitation is the fact that there does not exist any easy file transfer between it and device simulators. This limitation meant that this simulator was not extensively used during this work.

Figure 4.4. shows the doping concentrations for a phosphorus implant into a boron substrate for an nmos transistor produced with the EMF's 6μm NMOS process described in chapter five. The junction depth in this case is 0.46μm.

4.2.4 SUPRA

Traditionally devices have been modelled as one-dimensional structures, for which one-dimensional process simulators such as SUPREM are adequate. With the rapidly diminishing device geometries now used, the device widths are becoming comparable to their depths. Therefore these structures can no longer be treated one-dimensionally and true two-dimensional simulators are needed [6].

One such two-dimensional simulator is SUPRA. This program can handle local oxidation, implantation through arbitrary mask edges, non-planar surfaces and high concentration diffusions. It has been developed at Stanford University and therefore process models and input commands are similar to the industry standard one-dimensional SUPREM program [8].

The grid is non-uniform rectangular with a maximum of a thousand nodes in any direction, and again a numerical finite difference method is used to calculate impurity distributions. Due to the similarity with SUPREM, profiles can be passed between the two process simulators. It is also possible to pass profiles to the device and lithography simulators in the TMA suite [9].

The main drawbacks with SUPRA are the limited number of nodes and the absence of some of the models that appear in SUPREM III. It has also been reported that there is a discrepancy between the model coefficients in the two programs [10].

It is best used to investigate two dimensional effects such as LOCOS bird's beak, but as input to device simulators the accuracy of doping profiles is severely limited. For this reason it was used to determine design rules such as minimum physical distances between two oxidations, but not as direct input to device simulators.

Figure 4.5 shows half an nmos transistor. The source is seen to be doped with arsenic and the contact is on the right. There is a deep boron implant under the field oxide on the left to prevent leakage from one transistor to another. The field oxide was formed using LOCOS, and the characteristic bird's beak can be noted.

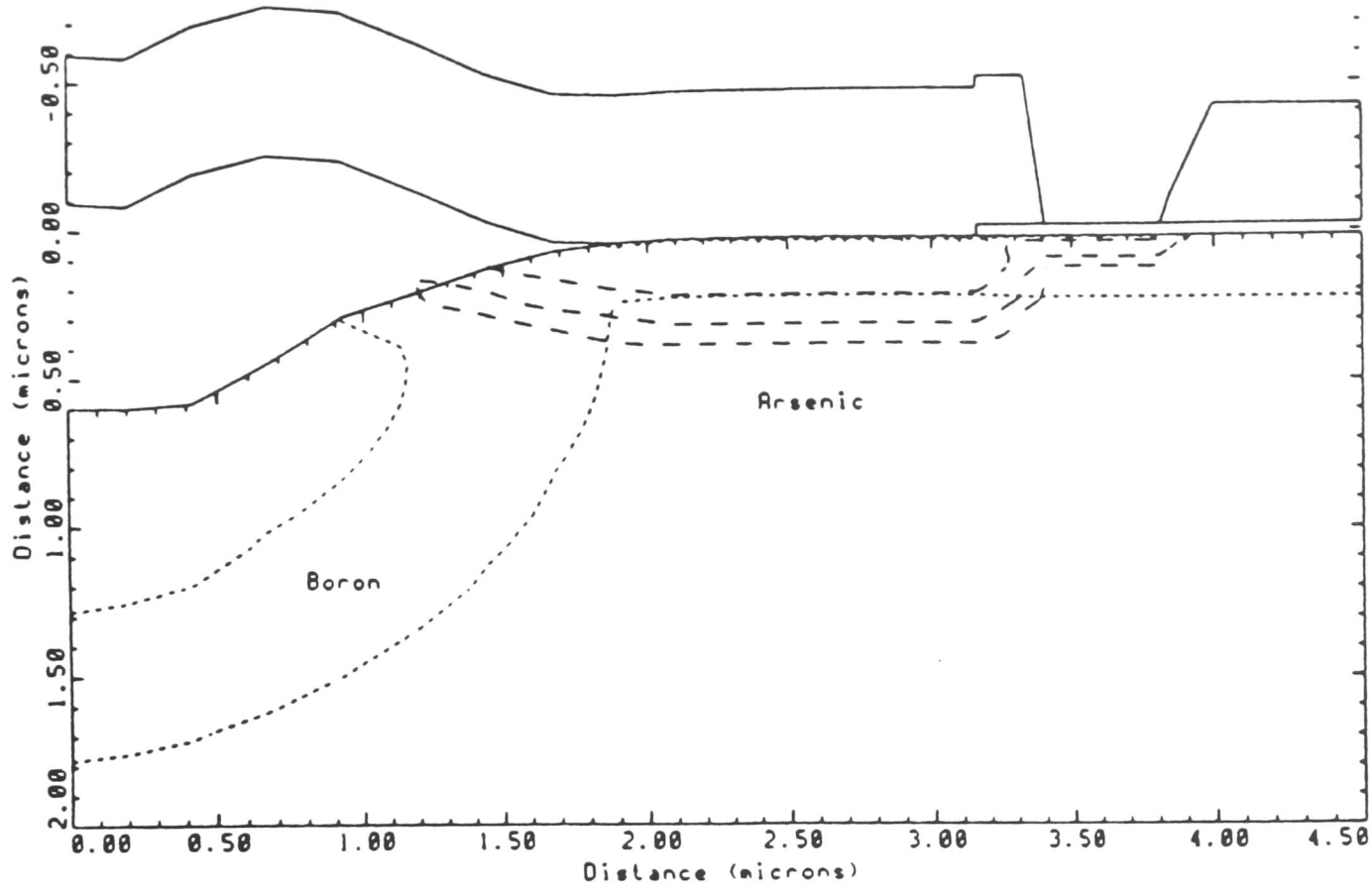


Figure 4.5 A SUPRA simulation of the source region of a transistor, highlighting the bird's beak phenomenon of the LOCOS process.

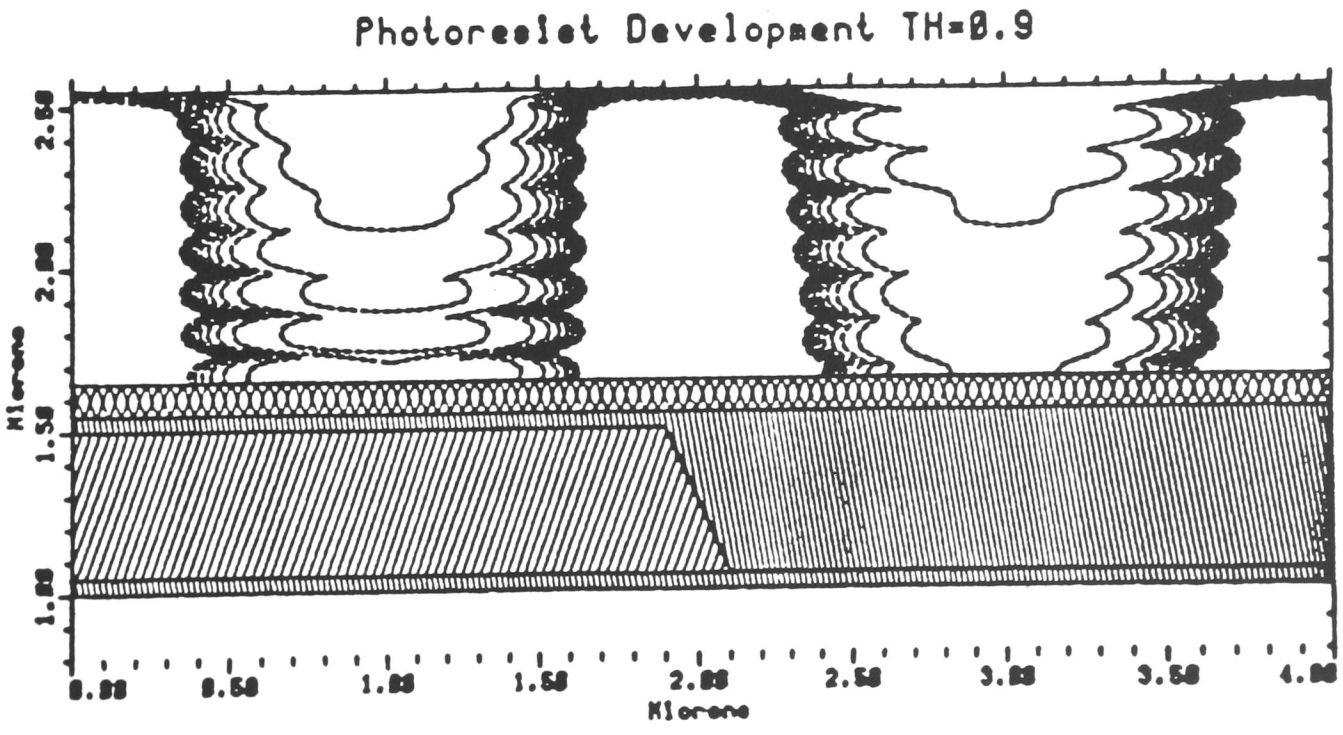


Figure 4.6 A DEPICT simulation of the development of resist over two different materials.

4.2.5 SUPREM IV

A further advance on SUPRA is the new process simulator from TMA, TSUPREM-4 [11]. It is a de-bugged version of Stanford University's SUPREM IV B. This program is now compatible with SUPREM III and features all the same modelling capabilities. In addition it will now simulate the diffusion of point defects simultaneously with impurity diffusion [12], enabling an accurate two-dimensional simulation of oxidation-enhanced diffusion. It also includes a complete silicon oxidation model, simulating the diffusion of oxidising species, the growth of oxide at the interface and the viscous oxide flow. The oxidation simulation includes a calculation of the oxidation-induced stress, an important cause of silicon dislocation, which can lead to junction leakage and stress.

The grid is triangular and has a maximum of three thousand nodes like SUPRA, but because of the more sophisticated models, it is a lot more CPU intensive and solutions are not so stable. Therefore SUPRA can be used as a first-pass approximation, before TSUPREM IV is used to obtain the more accurate results. One limitation of this package is that the temperatures cannot be ramped in the diffusion stages.

The main limitation however is the time needed to run a simulation, an order of magnitude more than SUPREM III. When an accurate simulation of a few process steps is required this is an ideal tool, but it is not adapted to a simulation of a full process flow.

4.2.6 EQUIPS

EQUIPS, the Edinburgh and Queens Universities Interactive Process Simulator, is a menu driven two-dimensional process simulator that has been developed in two British universities [13,14]. As the specification of the mesh in process simulation has been found to be so critical, EQUIPS allows the user to prepare and adjust the mesh at different stages of the simulation. This is done graphically using either a joystick or a mouse and is rapid enough to be used interactively.

The program can also be run in batch mode, which is recommended for meshes of around one thousand nodes. It has been tested and compared with SUPREM II, SUPREM III and ICECREM and has been found to give comparable results [15]. The limitation for this work is that other simulators are more compatible in passing data from one program to another.

4.2.7 DEPICT

As device sizes shrink and the critical dimensions approach the optical limits, it is important to be able to model the processing steps that affect the topography of a device. DEPICT [16], is a two-dimensional process simulation program that can model the photolithography, deposition and etching stages in a process. It can simulate devices with up to ten different layers consisting of forty different materials. The structure can be passed to the process simulators SUPREM and SUPRA.

DEPICT models processing by advancing discrete nodes which describe the upper surface of the device structure as it evolves in a deposition, etching or development process. In the deposition and development sections this motion is described using a string algorithm where the boundary between processed and unprocessed regions is approximated by a series of points joined by straight line segments [17].

In the etching section a threaded-tree structure is used. Each layer interface is represented by a series of nodes comprising a branch of this tree, where each node contains the coordinate value of the interface point it represents. Therefore, a deposition process is represented by adding a new branch on the tree and an etching or development process alters existing branches of the tree.

In the deposition stage new material can be added to the existing structure in one of five ways :

(a) Isotropically

(b) Uni-directionally

(c) Dual-directionally

(d) Hemispherically

(e) Planetary

Machine models exist and additional ones can be created to give a match between simulations and SEM data. The etching models can represent dry and wet etching by using an isotropic model, directional model, angle-dependent model or

combination of the above.

The Expose statement in DEPICT is used to simulate the imaging and exposure aspects of projection optical lithography. Particular machines and masks can be called up from the existing library or created by the user. It is also possible to simulate defocus of the imaging system with respect to the plane of perfect focus. There is also a module that can simulate different development conditions.

With these models the user can analyze and evaluate different imaging systems in order to optimise the photolithography process. It is a valuable tool for analysis of process latitude and sensitivity, and can be used to detect potential problems with metallisation step coverage. Figures 4.6 and 4.7 show the development of photoresist for a low voltage MOS transistor. It can be seen that during development the resist above the oxide develops slower than the resist over poly. This is due to the fact that during exposure poly reflects more UV than the oxide causing a higher exposure level. The effect of the polysilicon angle on the resist profile should also be noted in figure 4.7. It is possible to simulate the effects of topography on imaging. For example, this tool can be used to examine the effects of focus offset on contacts to both poly and active area, in order to determine the best focus for the two situations.

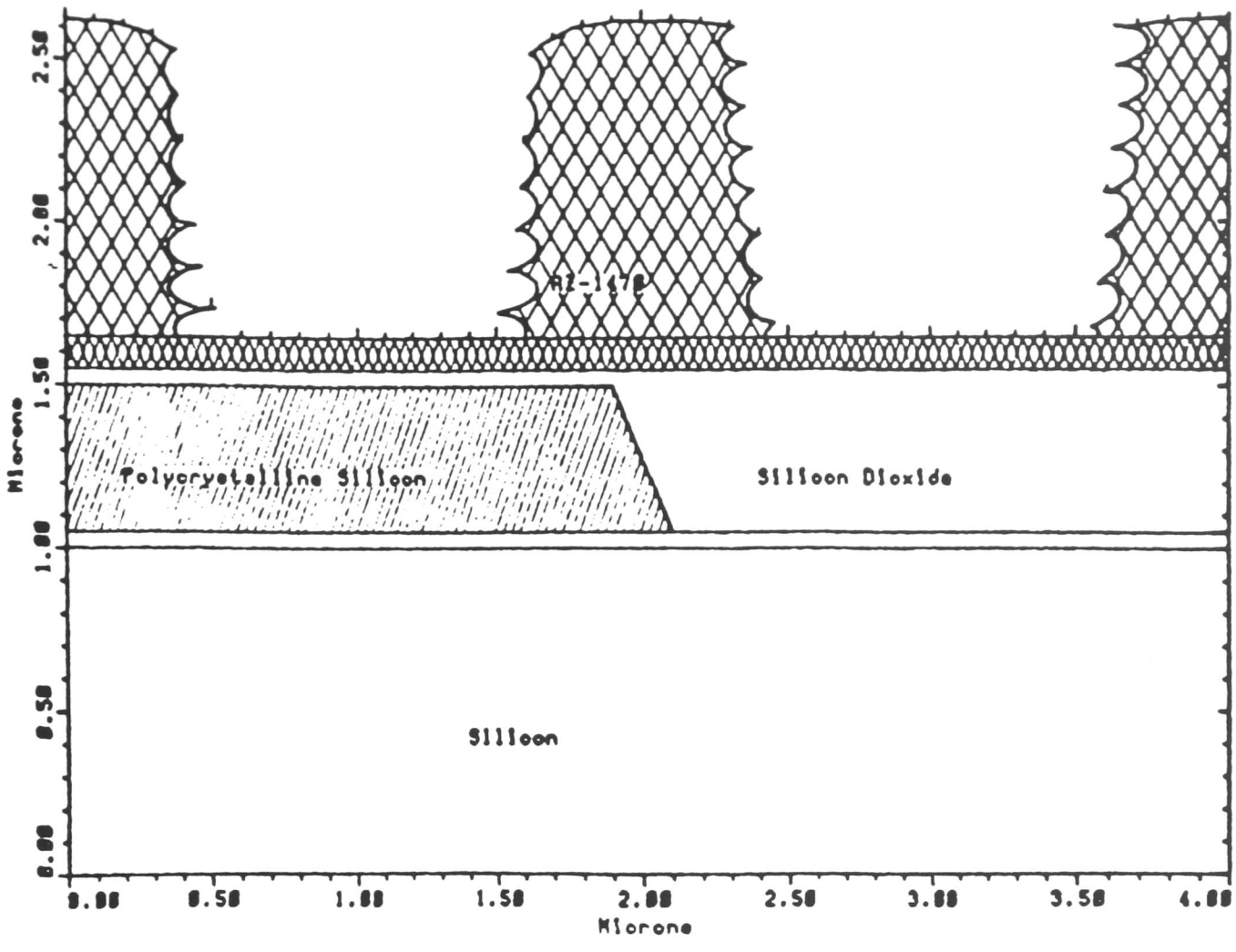


Figure 4.7 A DEPICT simulation after development of resist over two different materials, polysilicon and oxide.

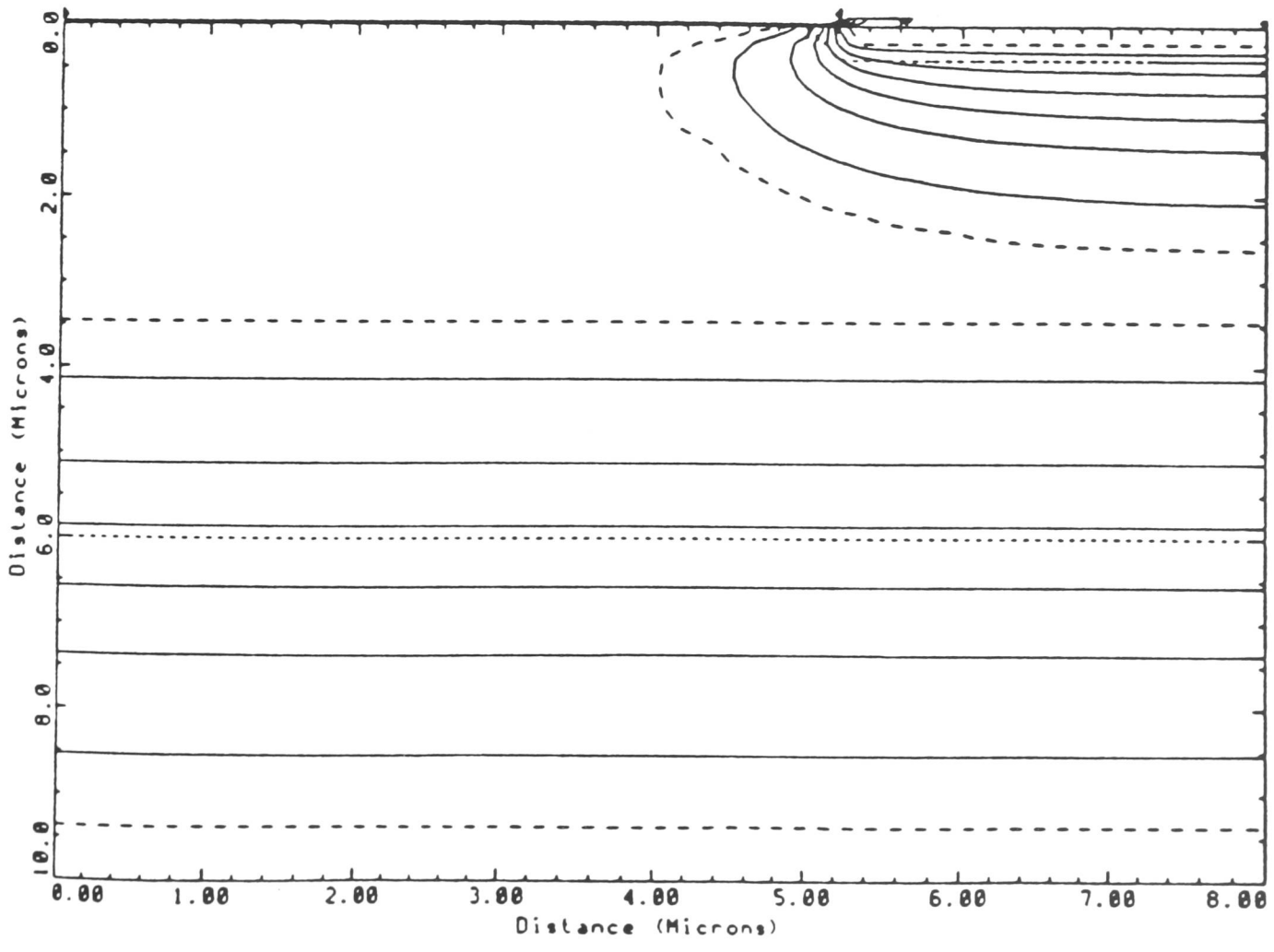


Figure 4.8 A CANDE simulation showing the equipotentials and depletion region around the drain under a bias of 5V on both the drain and the substrate.

4.3 Device Simulation

Device simulation is important for two user groups. The device engineer wants to understand how a device operates and hence is primarily interested in the internal device mechanisms, whilst the circuit designer seeks a quantative description of the terminal behaviour only, which should be as accurate as necessary and as simple as possible. Most approaches emphasise the device designer's view on modelling and hence stress numerical models. This review follows this trend [18,19].

The device designer needs to preserve the links between device physics and simulation in order to understand better how changes affect performance. Therefore it is important that he can view the electric-field strength and electrostatic potential as well as the electric current density in the device. Device simulation cannot really be considered in isolation from process simulation, but for simplicity only device modelling will be discussed in this section. Process simulation has been considered in an earlier section. An important point to remember is that accurate simulation of devices depends on good input from process simulators. It is therefore important to have a good file transfer between process and device simulation packages.

In the remainder of this section, a representative selection of simulators are considered. The four chosen are probably the most well-known and used of the large number available.

4.3.1 MINIMOS

MINIMOS, is perhaps the best known device simulator. As its name suggests it simulates two-dimensional planar MOS transistors. Its popularity stems from the fact that it is available from its authors for the handling costs only and that it is not CPU intensive. In fact it is now available in a PC version.

The device geometry is limited, so that only fairly standard MOS structures can be simulated. It is possible to input doping profiles from SUPREM using the interface file SUPMIN [20], or point by point. This last method, although complicated, facilitates the simulation of DMOS and LDDMOS devices. These types of devices cannot be simulated directly in MINIMOS, but by inputting doping profiles generated by a process simulator such as SUPREM this type of simulation becomes possible.

The mesh in MINIMOS is non-uniform and is generated automatically to account for bias values and the doping profile. The maximum mesh size is sixty by sixty nodes.

Validation of MINIMOS has shown some anomalies. The major cause for concern was the threshold model. A new model has been developed [20] that extrapolates back from the drain current - gate voltage slope to the intercept on the gate voltage axis. This new version agrees better with observed characteristics. Other changes to the MINIMOS code under this work [21], were to ensure that the SUPREM doping profiles were correctly reproduced in MINIMOS, and to correct the causes of some instabilities in the original code.

The main benefit of MINIMOS over other simulators is its price. It costs two orders of magnitude less than some of its contemporaries. However, when CANDE and PISCES are available, their flexibility in type of device used and their ease of file transfer mean that MINIMOS was not used to a great extent in this thesis.

4.3.2 BAMBI

The program BAMBI (Basic Analyzer of MOS and Bipolar devices) is a general purpose device simulator [22]. The simulation structure is described by a closed polygon which is divided into a main section (semiconductor) and optionally one to ten adjoining dielectric sections. It can therefore, in principle, simulate any semiconductor device [23].

The simulator has a non-uniform automatically generated grid similar to that of MINIMOS, which is refined as the program progresses. It runs in three passes and is found in practice to be slow to run, and has difficulty in converging when trying to simulate near the avalanche region of operation [24,25]. For this motive it was tried, but not used in this project.

4.3.3 CANDE

CANDE is a two-dimensional MOS device simulator that models potential distributions and carrier concentrations in order to predict a device's electrical characteristics under terminal biases [9,26]. It will give a one-carrier (i.e. electron or hole not both) solution to current flow in the device. It is more versatile than MINIMOS in that it can model virtually any device structure composed of a semiconductor substrate covered by an insulating layer, but because it does not consider minority carriers it is not suitable for bipolar or COMFET devices.

The grid has a total of three hundred nodes in either direction, with a total maximum of ten thousand nodes and solves Poisson's equation using a finite difference algorithm. The grid is user-defined and is constant throughout the whole simulation. It can also be defined automatically but this is only suitable for the most simple of device structures, and should not be used if accurate simulations are required.

The impurity profiles can be defined using built-in analytic functions, SUPREM III or SUPRA profiles. If accuracy is required it is best to use several SUPREM III profiles rather than SUPRA as the former can have a much finer grid and therefore better represents the profile.

CANDE can generate complete I-V characteristics for MOS devices, calculate the electron and hole ionisation integrals and derive capacitance - bias characteristics. It can do this for several MOS structures including LDDMOS, LDMOS and VDMOS devices.

Figure 4.8 shows an nmos half transistor structure under bias, with 5 volts on both the substrate and the drain, with no bias on the gate. Both the equi-potentials (solid lines) and depletion regions (dotted lines) are shown. It can be seen that the well-substrate depletion is separate from the drain-well depletion, but if the drain voltage is increased, as in figure 4.9, then these regions merge.

The fact that CANDE is flexible in the type of MOS transistor modelled, and that it has an easy file transfer to other simulators, lead to its extensive usage in the design and modelling of both low voltage and power transistors during this project. In the following chapters, a comparison between measured and simulated results shows this choice to be valid.

4.3.4 PISCES

PISCES is a more general purpose two-dimensional device simulator than CANDE. It solves Poisson's equation and both the electron and hole current-continuity equations simultaneously. This means that it is suitable for both bipolar and unipolar devices. It can calculate terminal characteristics, punchthrough characteristics, electric fields, capacitances, resistances of cross-sections and CMOS latchup conditions [27].

The grid is non-uniform triangular and can be generated manually or automatically, and has a maximum of three thousand nodes. The grid can be refined automatically as the simulation progresses. These regrids can be based on the midgap potential, electric field, quasi-Fermi potentials, carrier concentration or

the doping concentration. Like CANDE, the doping profiles can be input from SUPREM III or SUPRA. SUPREM III profiles will lead to more accurate results.

In addition to the standard version there are three option modules that are available from TMA for PISCES 2-B. The first of these models impact ionisation and calculates substrate and gate current. This is useful for the designer interested in hot electron effects. The second module calculates the ionisation integrals to analyze the avalanche breakdown of reverse biased junctions. The third option allows different semiconductor material parameters to be specified for different regions of the device structure. This capability provides an approximate method of treating structures that contain more than one semiconductor material or that have semiconductor properties that vary with location.

Figure 4.10 shows the grid after two successive automatic doping regrids for a vertical npn bipolar transistor. It can be seen that the nodes are concentrated around the junctions. This type of grid is suitable for low voltage simulations, but under high bias the regions of interest can be far from the junctions, and a regrid based on potentials should be used. Figure 4.11 shows the current vectors flowing under bias in this npn bipolar device when Vbe = 0.90 Volts and Vce = 3.0 Volts.

PISCES was more difficult than CANDE to use and so the latter was used wherever possible. However there were cases where PISCES is the only simulator capable of modelling the given structure. This program has the benefit of the ease of data transfer common to all programs in the TMA suite of software.

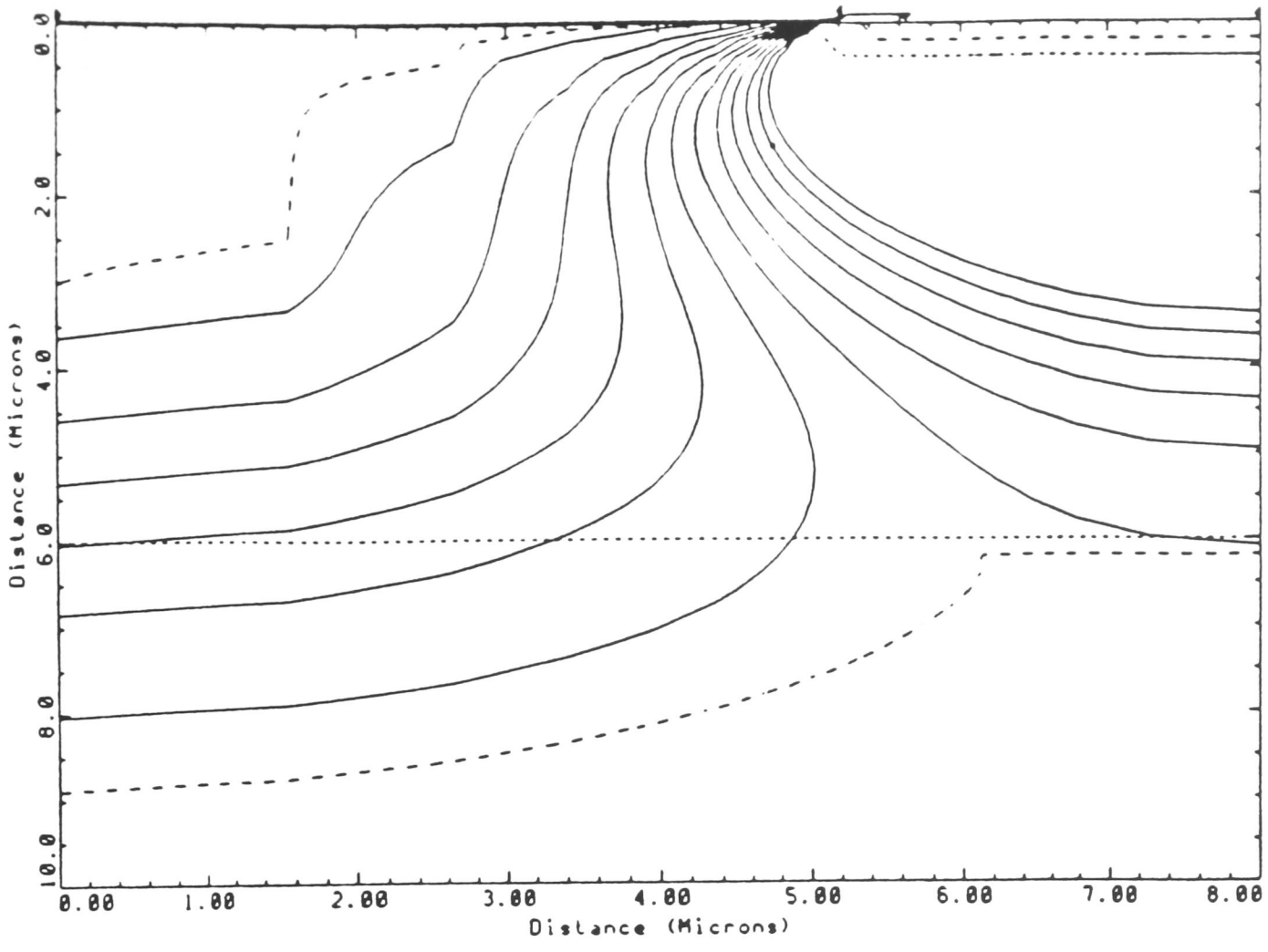


Figure 4.9 A CANDE simulation showing the equipotentials and depletion region around the drain under a bias of 10V on the drain and 5V on the substrate.

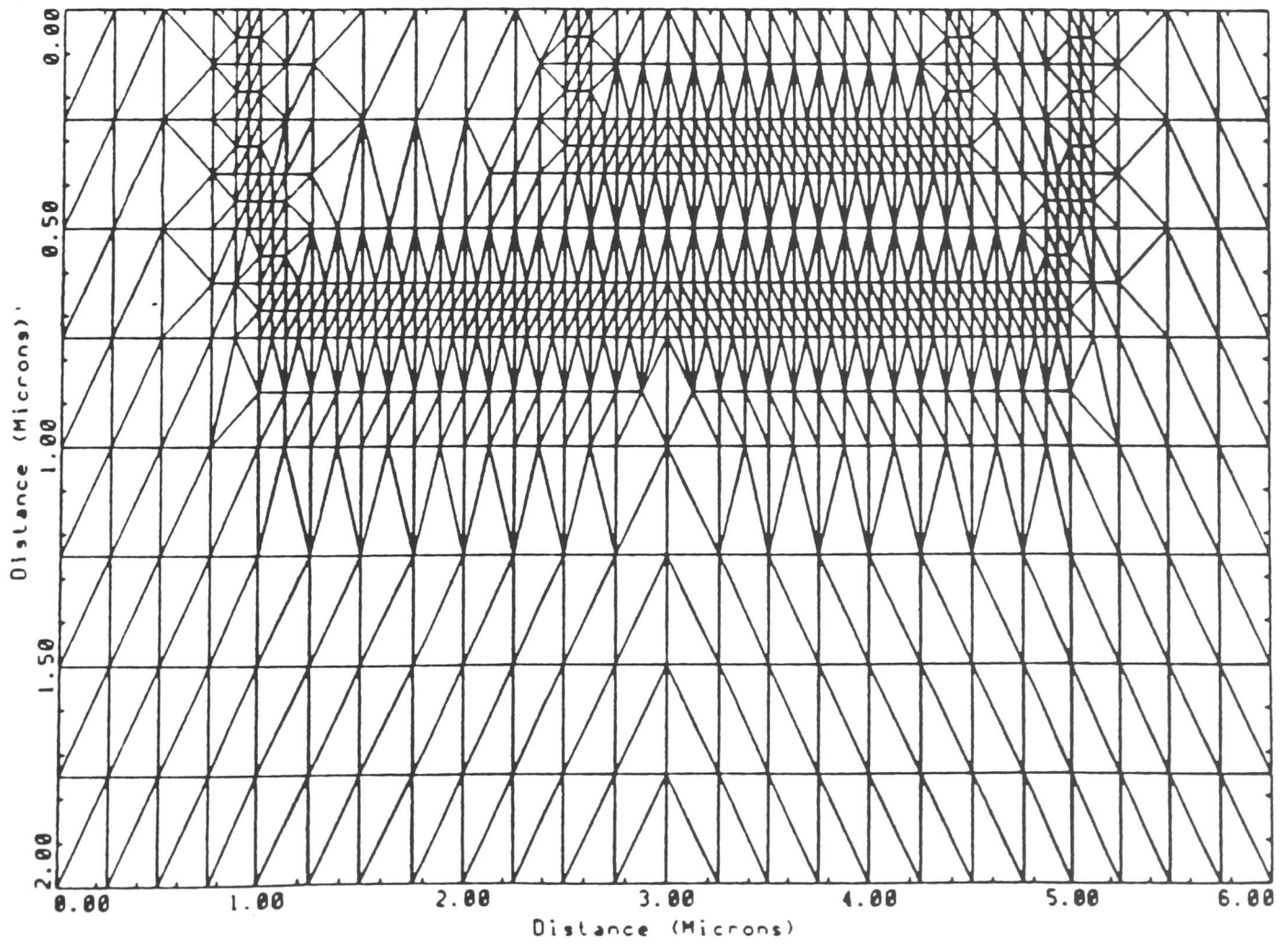


Figure 4.10 The Grid in PISCES of a bipolar transistor after two automatic regriding operations based on the doping concentrations.

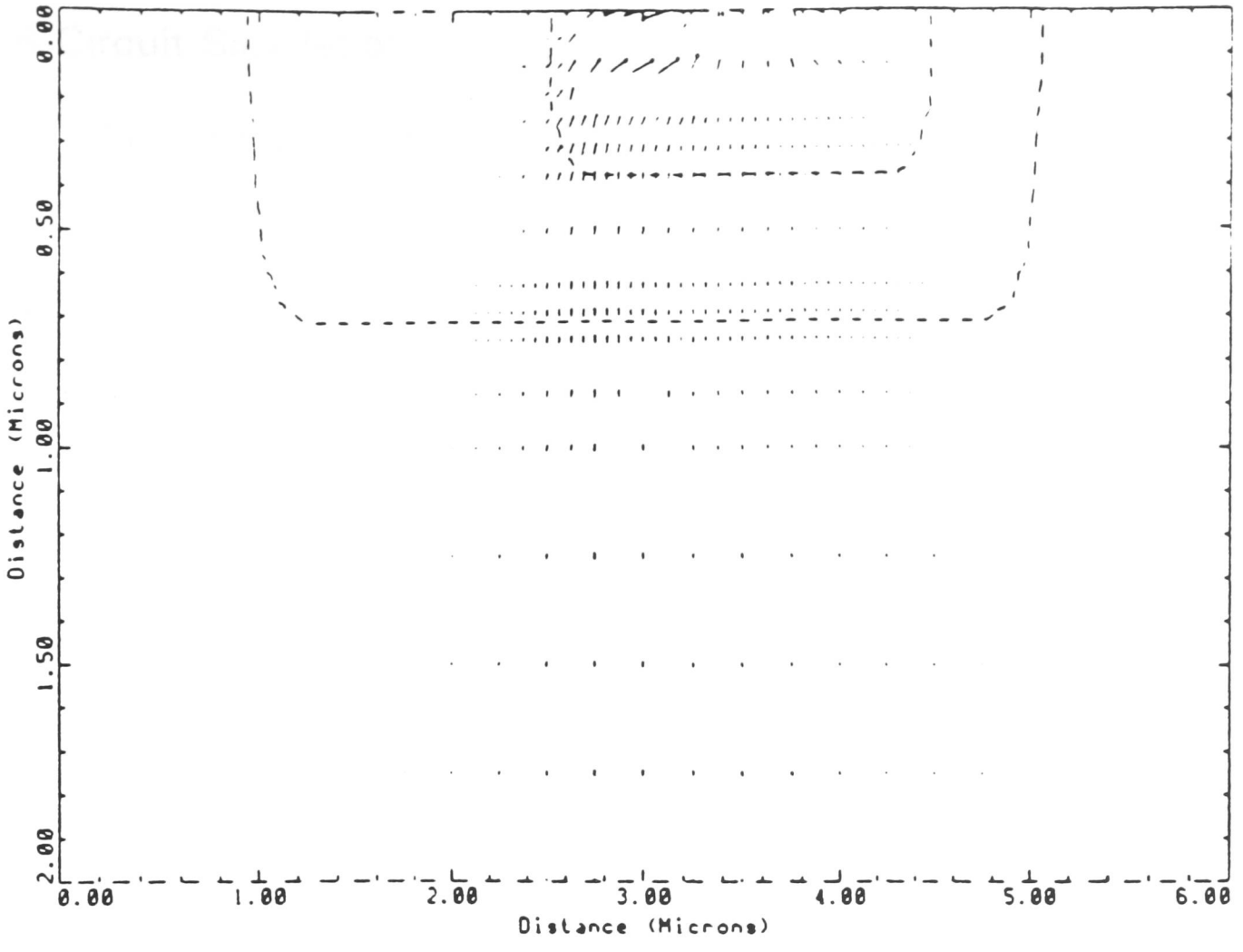


Figure 4.11 A simulation of the current flowing in an npn transistor under bias using PISCES.

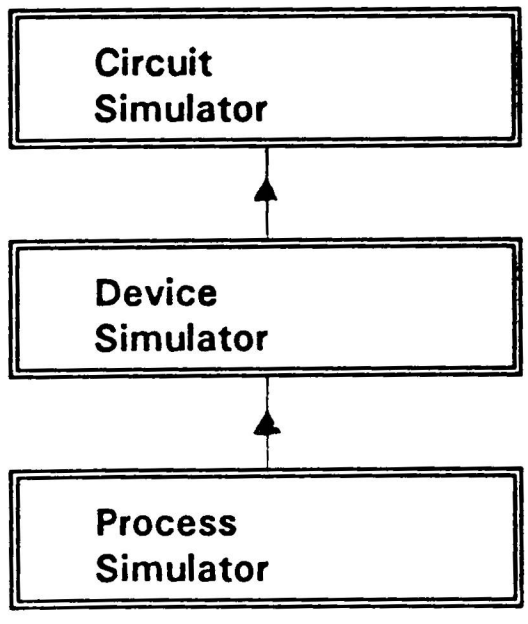


Figure 4.12 Diagram of the hierarchy of process, device and circuit simulators.

4.4 Circuit Simulation

Circuit simulators are hierarchically the next step up from process and device simulators (see figure 4.12). They generally require accurate input from the device simulators to output valid circuit parameters. There are three ways in which a circuit simulator can operate [28], and these are shown in figure 4.13.

The first of these is to run using measured or simulated data for each of the components involved. This however means that a vast amount of data needs to be stored to cover all possible operational points. The second method is to use a reduced data set with only some of the characteristics stored, and interpolation is used between measured curves [29]. The third option is the most common and data-efficient method uses parameter extraction techniques.

In parameter extraction, each device is represented by an analytic, empirical or semi-empirical model. By using a model instead of I-V points, there can be a reduction of up to three orders of magnitude in the data needed. The models are usually based on basic semiconductor equations, for example the Gummel-Poon model for bipolars with some added correction terms. Sometimes, however,the new simulated data points do not match the measured ones and so-called optimisation techniques are needed. It is important to note that optimisation of parameters does not always lead to a physically valid parameter set.

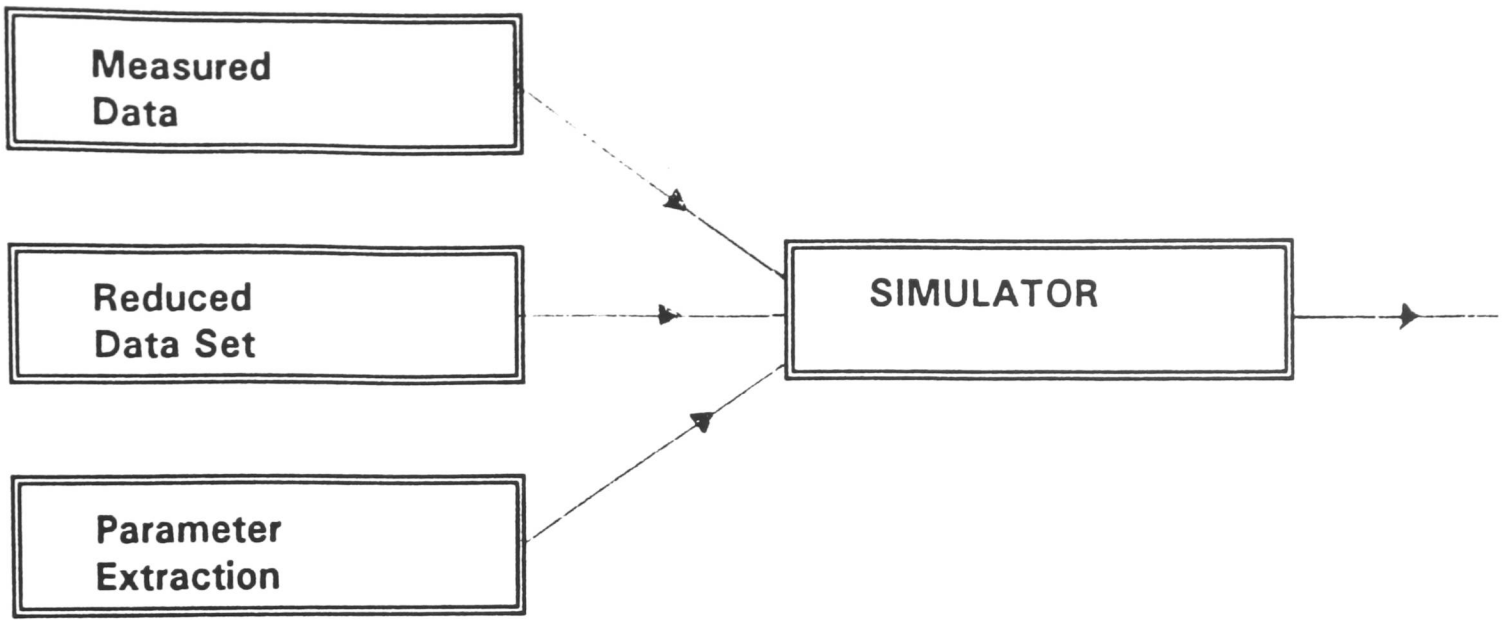


Figure 4.13 Schematic of the three possible inputs to a circuit simulator.

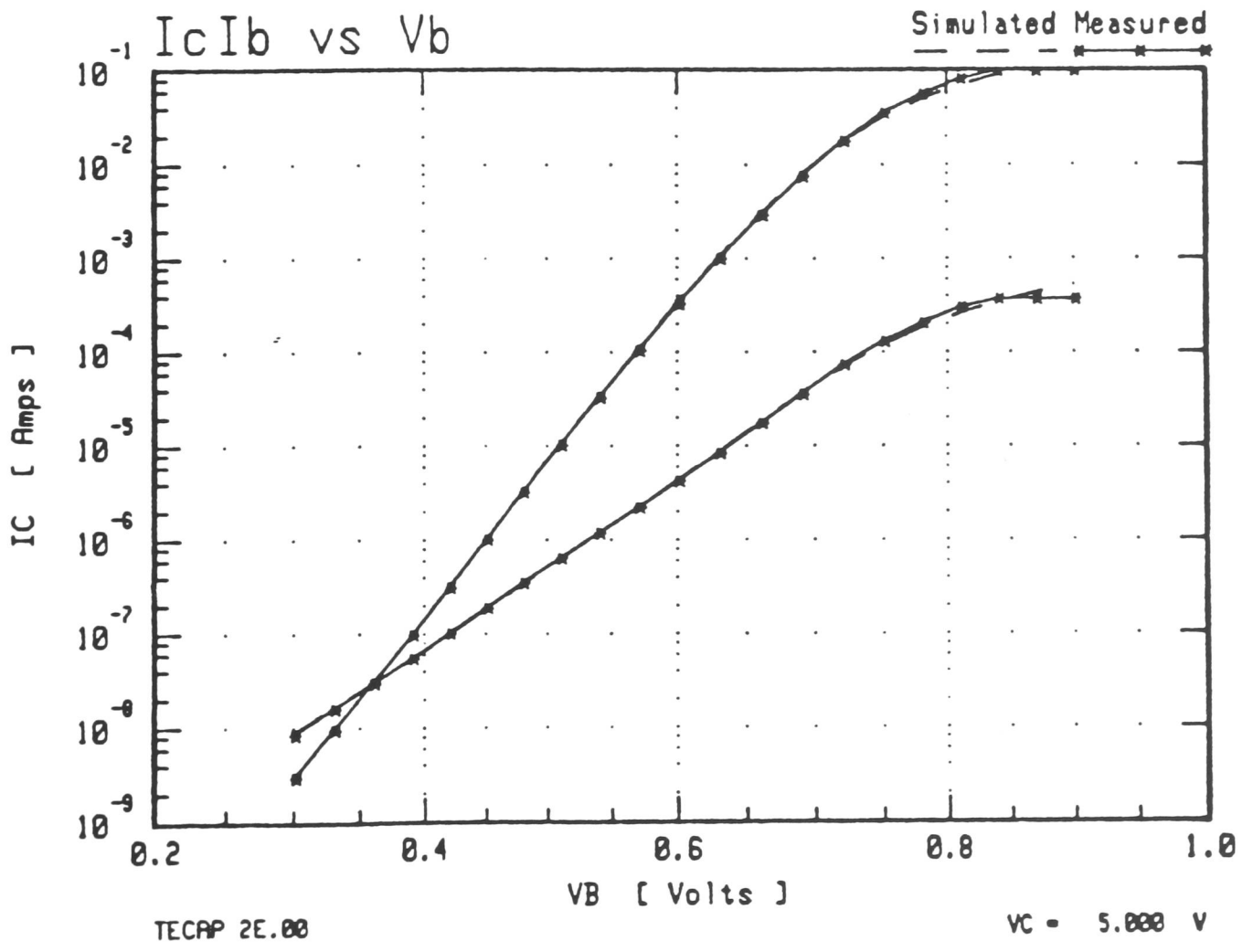


Figure 4.14 The measured and simulated Gummel curves of an npn transistor after parameter extraction using TECAP.

4.4.1 SPICE

The most widely used circuit simulator is SPICE which was developed at the University of California, Berkeley. The SPICE circuit simulation program can represent passive components, active components, ideal functions and sub-circuits as detailed below [30].

Passive Components

Resistors

Capacitors

Non-linear Capacitors

Inductors

Non-linear Inductors

Mutual Inductors

Loss-less Transmission Lines

Active Components

Junction Diodes

Bipolar Transistors

Junction Field Effect Transistors

Insulated Gate MOS Transistors

Ideal Functions

Linear Voltage Controlled Current Sources

Non-linear Voltage Controlled Current Sources

Linear Voltage Controlled Voltage Sources

Non-linear Voltage Controlled Voltage Sources

Linear Current Controlled Current Sources

Non-linear Current Controlled Current Sources

Linear Current Controlled Voltage Sources

Non-linear Current Controlled Voltage Sources

Sub-Circuits

In SPICE 2.F there are three different MOS models that are available to the user. The Level 1 model is the simple Scichman-Hodges model, the Level 2 model is an analytical one-dimensional model which incorporates some second-order effects of small geometry devices and the Level 3 model is a semi-empirical model. This last model is described by a set of curve fitting parameters rather than physical ones [31,32].

Improvements and fixes for these models have been reported in the literature including the BSIM short channel IGFET model [33 - 39].

The bipolar model used in SPICE is an adapted Gummel-Poon model. It differs from the original model in some of the input parameters and the lack of any base push-out modelling [40].

There are several parameter extraction programmes available which produce the input necessary for SPICE circuit simulation. The next sections will discuss TECAP, TOPEX, and PARAMEX, the programs that were available during this project ; and IPSEX, a bipolar parameter extraction program developed during this project.

4.4.2 TECAP

One parameter extraction program is Hewlett-Packard's TECAP [41]. This program includes the Berkeley models for MOS, bipolar and gallium-arsenide devices. The software system controls a DC source/monitor unit, a capacitance meter and a network analyzer. It then extracts parameters from measured data. At this stage the parameters can be used to simulate the measured curves. If the fit is not good enough then optimisation may be used.

The optimisation algorithm used is a modified Levenberg-Marquadt one which combines the steepest descent and the Gauss-Newton method. Care has to be taken here to limit the optimisation to the region of interest and to set parameter tolerances to physically valid limits. If this is not done then the algorithm will invariably converge to physically invalid numbers.

Figures 4.14 and 4.15 show the measured and re-simulated extracted values for a bipolar transistor. The first diagram shows the Gummel curves, the second shows the gain characteristics. It can be seen that the fit in both cases is reasonably good. There is some discrepancy at high current due to the inadequacy of the Gummel-Poon model used in Spice.

The software routines form the basis for the EXACT parameter extraction routines [42]. In the EXACT version all decisions on where the device should be biased during measurements, whether data is valid and how to extract the parameters are made by the software system, removing operator control. In this version no optimisation is used. This means that the final parameters should be more physically valid. However this software is not commercially available at

present and was therefore not used.

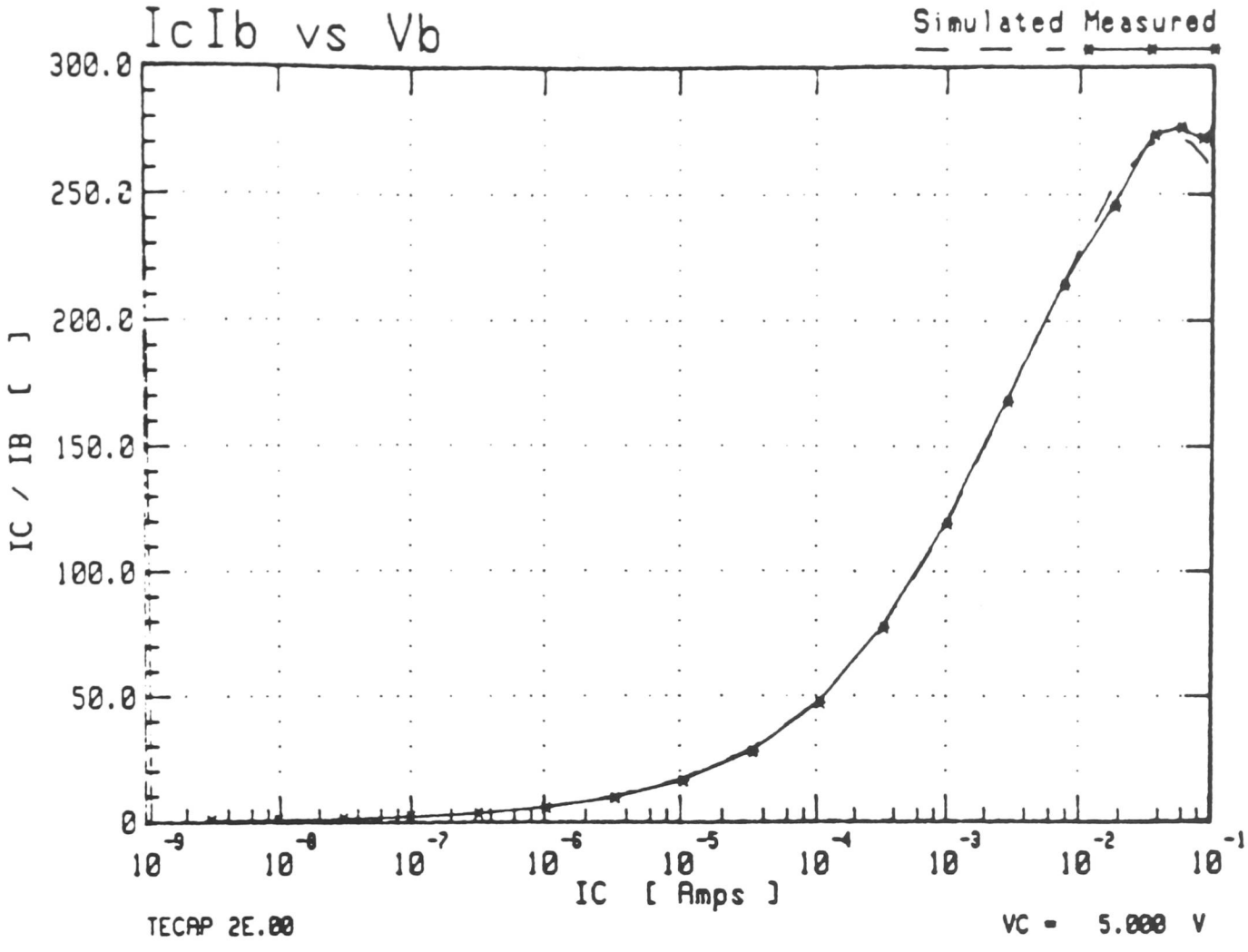


Figure 4.15 The measured and simulated gain curves of an npn transistor after parameter extraction using TECAP.

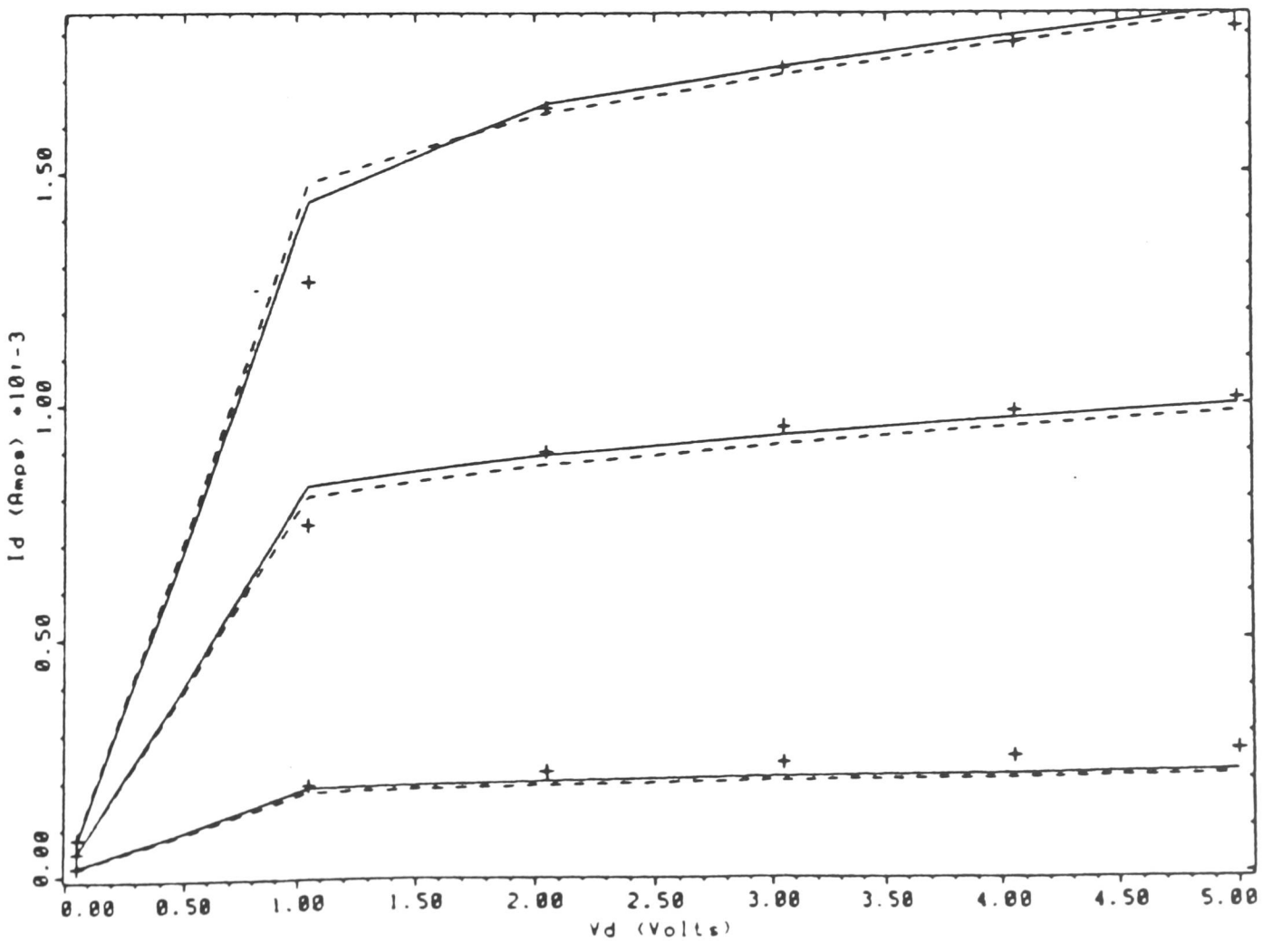


Figure 4.16 The measured and simulated drain curves of an nmos transistor after parameter extraction using TOPEX.

4.4.3 TOPEX

In order to simulate the whole semiconductor process it must be possible to take output from device simulators as input to circuit simulators. TOPEX [43] is a general-purpose parameter extraction program that can operate on either simulated or measured data. This data only has to be provided in the same form as that produced by PISCES and CANDE.

The main use of TOPEX is to extract SPICE parameters, and it is equipped with MOSFET, bipolar, JFET and JCAP models. It also has the potential to use user-defined models and extraction routines, and be used as a model development tool. However most users will be content with the given models.

TOPEX is fitted with the same optimisation algorithm as TECAP, which is a sophisticated least squares fit. Parameters can be varied over user-defined ranges and the output will give sensitivities of fit to each parameter, and estimates of how the values of the parameters depend on each other.

Figure 4.16 shows simulated drain characteristics and re-simulated data using extracted parameters. The fit between the curves is reasonable. Similarly figure 4.17 shows the threshold characteristics for two values of substrate bias. The extracted threshold for a nMOS transistor (without a threshold adjust implant) obtained for no substrate bias is 0.20 Volts. When the implant is added the TOPEX simulated value is 1.2 volts compared with a measured value of 1.4 and a SUPREM III value of 1.4 Volt. This poor fit is probably due to the mismatching of grids when information is passed from SUPREM III through CANDE to TOPEX, as the same SUPREM III simulations were used as the source of the

doping profiles. This highlights the fact that extreme care must be taken when passing information from one simulator to another.

4.4.4 PARAMEX

An alternative to using parameter extraction with numerical optimisation is to use physical parameter extraction. In physical extraction each parameter is extracted individually in a specific sequence rather than in parallel as in TECAP and TOPEX.

A program called PARAMEX [32] adopts this strategy for extracting the d.c. SPICE 2 Level 3 MOSFET parameters. It concentrates on accurate extraction of meaningful parameters without resorting to optimisation techniques. This difference means that the user does not need to know how parameters interact, and means that the parameters derived can be used for process control. Like TECAP it uses a Hewlett-Packard controller to drive a HP 4145 semiconductor parameter analyzer. The resulting fits have been shown to be comparable with TECAP [44].

Figure 4.18 shows both the measured and simulated data for a NMOS transistor produced using the EMF's 6μm process. The fit is reasonably good in the linear region, but there are some discrepancies in the saturation region.

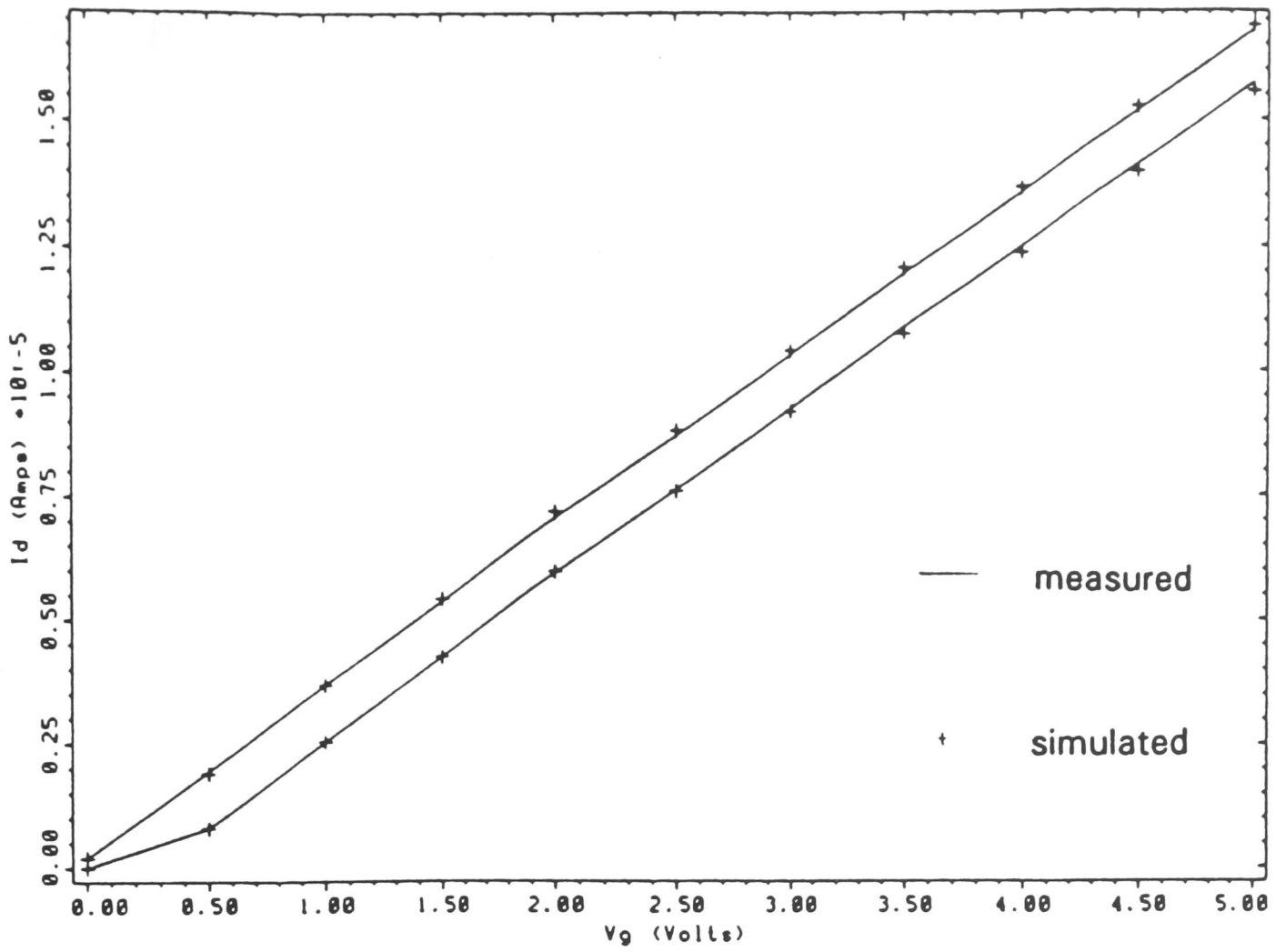


Figure 4.17 The measured and simulated Id-Vg curves of an nmos transistor after parameter extraction using TOPEX.

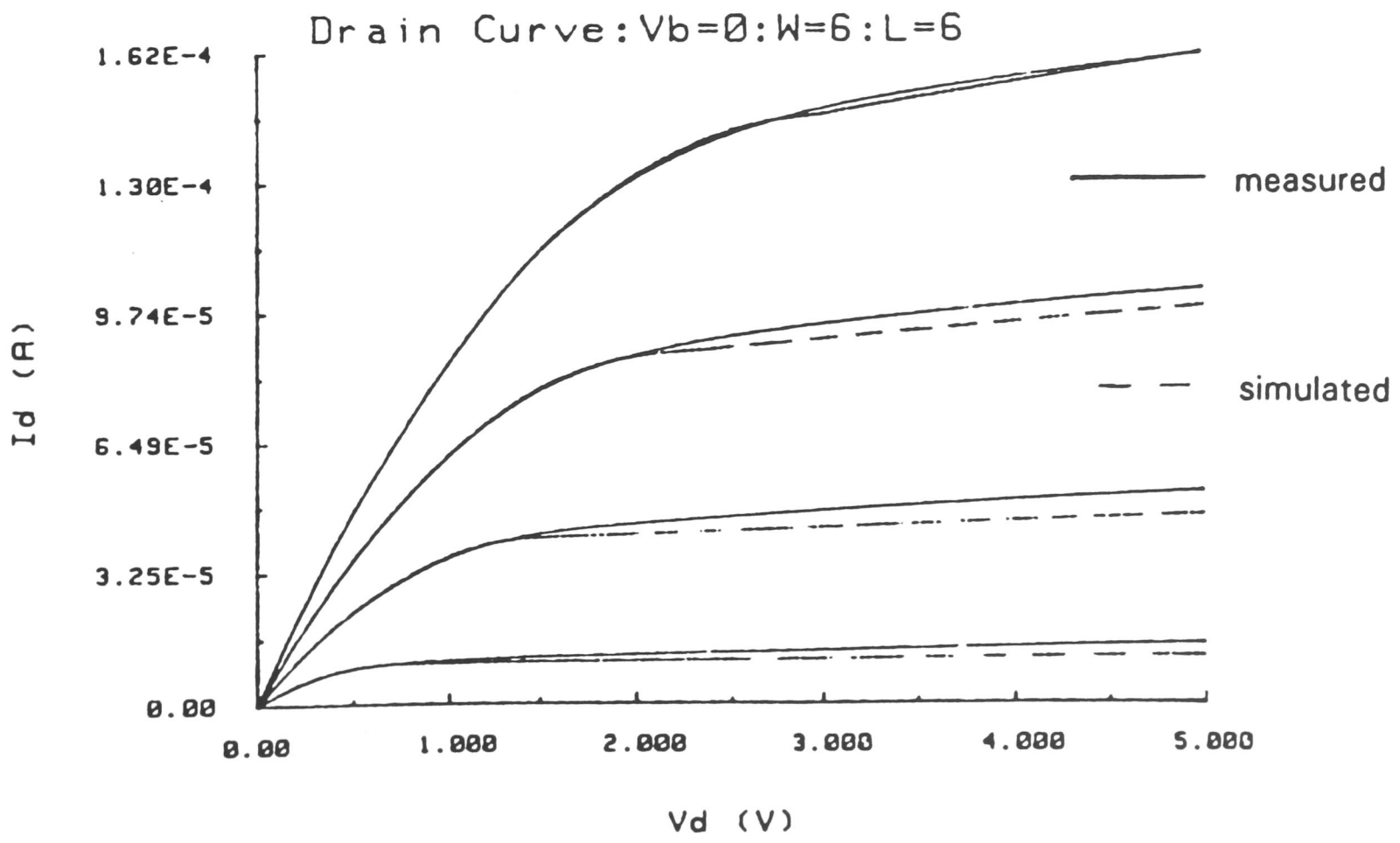


Figure 4.18 The measured and simulated drain curves of an nmos transistor after parameter extraction using PARAMEX.

4.4.5 IPSEX

IPSEX is a dc bipolar parameter extraction routine that was developed during this project [45]. The model used for bipolar transistors is based on the work of Gummel and Poon as described by Getreu [40].

The IPSEX extractor attempts to obtain fifteen dc parameters by driving a semiconductor parameter analyzer and fitting parameters to measured curves. The parameters extracted are :

ISS the transistor saturation current

BF the forward common-emitter, large signal gain

BR the reverse common-emitter, large signal gain

RE the emitter resistance

RC the collector resistance

VA the Early voltage

VB the inverse Early voltage

IKF the knee current

IKR the reverse knee current

Theta, C2 and NEL that model forward gain

Theatar, C4 and NCL that model reverse gain

These are measured using the techniques suggested by Getreu. The exceptions being that the inverse Early voltage is measured using the same technique as the forward Early voltage to improve accuracy. There is also an alternative method for extracting the large signal gains . Using this program it is possible to change the region in which the parameters are extracted to reflect the

way the transistor is driven in a circuit. In this way it is possible to have an accurate fit in the high-current region despite the inadequacies of the model used in Spice.

The program can be run in either interactive mode or the whole measurement and extraction procedure can be run automatically. The main display screen of the program shows the parameters that can be extracted and their current values. There are three modes of operation namely measurement, extraction and simulation. Each of these have plotting routines associated with them. The program is menu-driven in interactive mode, and there are eight measurement routines available :

Gummel

Reverse Gummel

Early voltage

Reverse Early voltage

Emitter resistance

Collector resistance

Knee current

Reverse Knee current

There are also nine extraction routines that can be called up. The main table is updated after each extraction. In order to fit lines a double-differential technique is employed. A high second-derivative value indicates a high rate of change of slope and is used to detect where the curve departs from a straight line. Once a line has been found it is possible to extrapolate in both directions to obtain

intercept values.

Figures 4.19 and 4.20 show the measured and simulated values for the Gummel curves and gain characteristics of an npn transistor. in this case the simulator was set up in order to model accurately the mid-range of currents. As the model used in Spice doe not compensate for high-current gain fall off, the fit is poor in the high current region. However, it is possible with this program, to generate a separate set of parameters that model well the high current region, if this is region of interest to the designer.

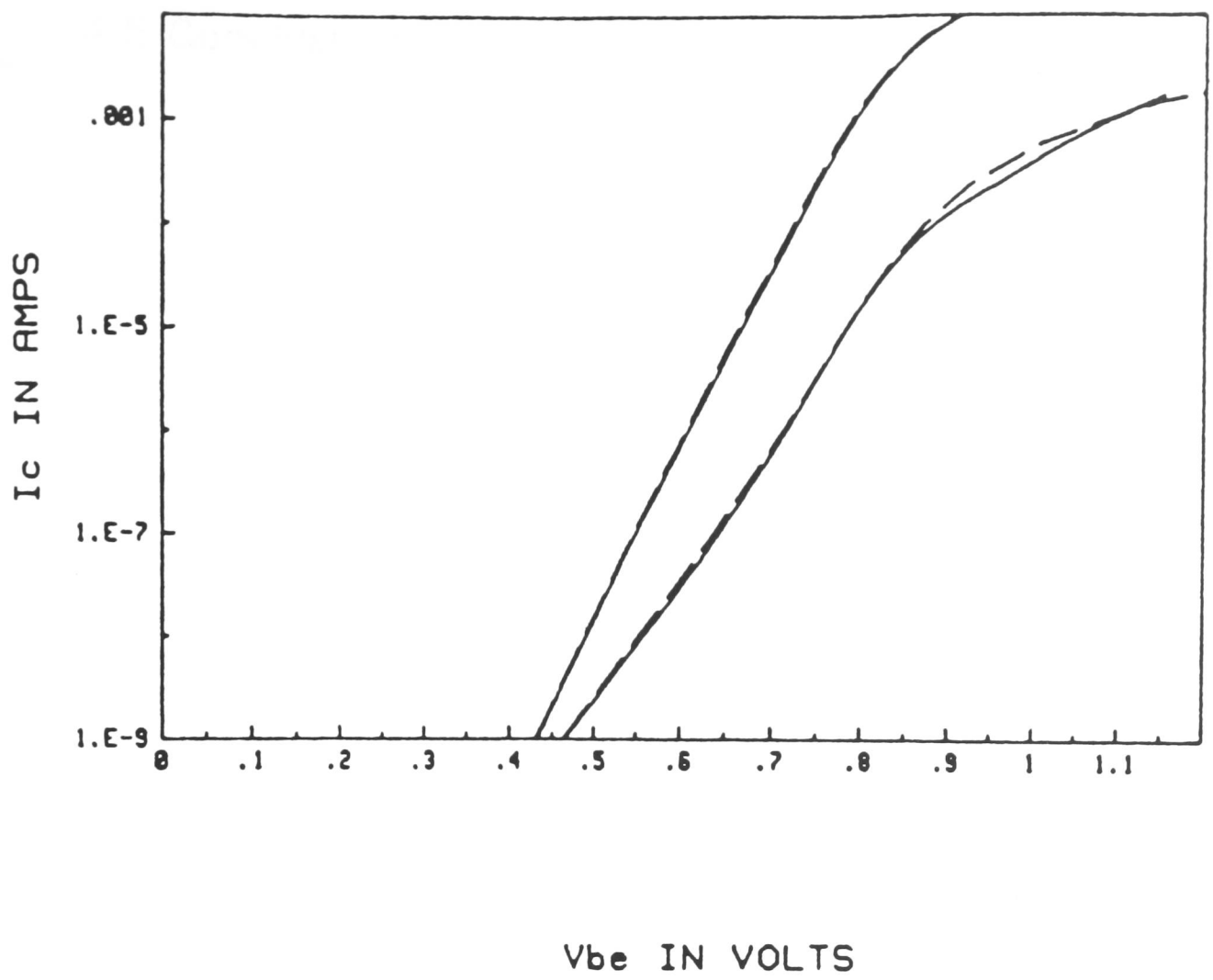


Figure 4.19 The measured and simulated Gummel curves of an npn transistor after parameter extraction using IPSEX.

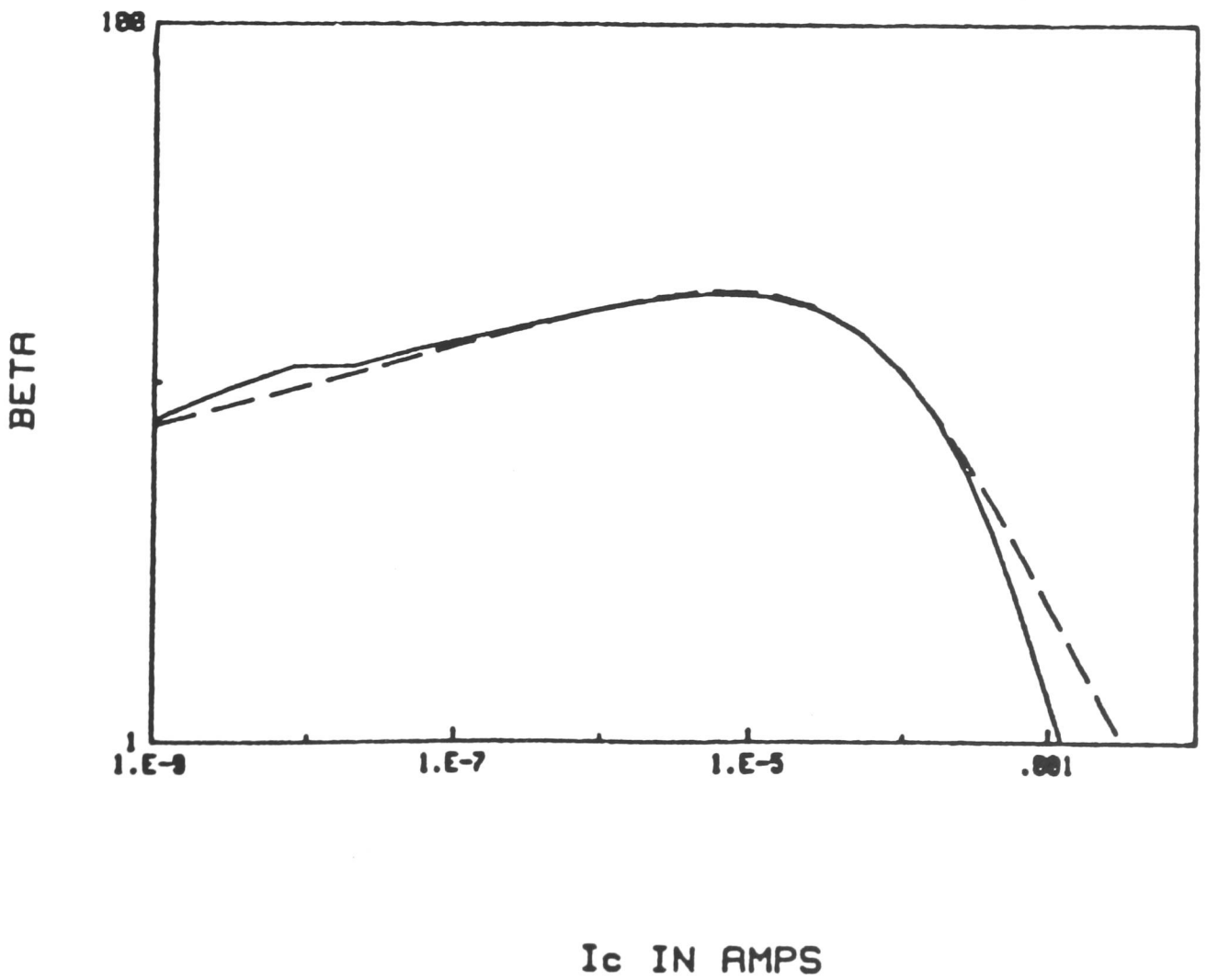


Figure 4.20 The measured and simulated gain curves of an npn transistor after parameter extraction using IPSEX.

4.5 Conclusions

The advanced processes and device structures used in today's circuits are too complex to be designed by humans alone. The circuits themselves have become so large that designers need computerised help to predict their performance. Simulation has become indispensable in the design of new processes and even in process modifications. When, for example, an ion implant energy is changed it is also necessary to simulate the subsequent thermal treatments to be sure of the effect. With today's twenty mask processes a rough guess is not enough, predictions of device performance and circuit yield are also needed. It is necessary to combine the three disciplines of process, device and circuit simulation in order to satisfy these demands.

With this approach it is possible to develop a new circuit or to determine process windows almost entirely at a terminal. Practical experiments cannot be replaced since, by definition models always follow measurements, but the number of physical runs can be reduced, thus saving both time and money.

The approach chosen in using simulation can be summarised as follows; first calibrate the simulation against measured devices, then model new structures. When these structures are eventually fabricated, compare the measurements with the simulated values. If the differences are significant, then correct the variable coefficients, then re-simulate before using the package on other processes or structures. This methodology of feedback between measurement and simulation is the best way of ensuring a reasonable chance of a good fit when developing new processes and device structures using computer packages.

In this project simulation packages were used extensively in order to reduce the development time of new transistor designs and process flows. As a first pass it is necessary to compare simulation results against known processes and device structures. The simulators described previously were used to characterise two known and extensively measured processes; the EMF 6 μm NMOS and 5 μm CMOS processes. It was seen that some of these simulators could be made to fit adequately the measured values. The process simulators which best fitted were SUPREM III and ICECREM. However the latter program does not have the capability to model silicon nitride or polysilsicon. Further, ICECREM does not have the same ease of file transfer as SUPREM III. For these reasons it was the TMA package that was used predominantly in this work. Similarly for the device simulators, it was CANDE and PISCES that were extensively used. The reasons for this were again the ease of file transfer and the fact that simulated results agreed well with measured values for transistors produced in the EMF.

References

[1] W.J.C.Alexander, "The Implementation of Process Simulation on Transputers For the Production of Application Specific ICs", Microelectronics'89, Brisbane, 12-14th July, 1989.

[2] B.R.Penumalli, "Physical Models and Numerical Methods for VLSI Process Simulation", Process and Device Modelling, pp.1-29, Elsevier Science Publishers B.V, Amsterdam, 1986.

[3] D.A.Antoniadas, S.E.Hansen and R.W.Dutton, "SUPREM II - A Program for IC Process Modelling and Simulation", Stanford Electronics Laboratories Technical Report No.5019-2, 1978.

[4] A.L.Butler and S.Duncan, "Use Guide and Assessment of SUPREM II.5", Plessey Technical Report A11/285/85, pp.1-27, Towcester, Northants, 1985.

[5] Technology Modeling Associates TMA, "Suprem-3", One-Dimensional Process Analysis Program, 1986.

[6] C.P.Ho, J.D.Plummer, S.E.Hansen and R.W.Dutton, "VLSI Process Modelling - SUPREM II", IEEE Trans. Electron Devices, vol.ED-30, pp.1438-1453, 1983.

[7] A.J.MacDonald and A.J.Walton, "Simulation of the EMF 6 micron Silicon Gate NMOS Process Using ICECREM", Edinburgh Microfabrication Facility Internal Report, April 1987.

[8] Technology Modeling Associates TMA, "Supra", Two-Dimensional Process Analysis Program, 1987.

[9] Technology Modeling Associates TMA, 1987 European Product Seminar agenda, 1987.

[10] A.J.MacDonald and A.J.Walton, "Simulation of the EMF 6 micron Silicon Gate NMOS Process Using SUPREM III", Edinburgh Microfabrication Facility Internal Report, April 1987.

[11] Technology Modeling Associates TMA, "Tsuprem-4", Two-Dimensional Process Analysis Program, 1988.

[12] Newsletter TMA, 31st May 1988.

[13] R.Ferguson, A.Walton and M.Fallon, "Interactive Process Simulator (EQUIPS)", Process and Device Simulation, Alvey Club Meeting, pp.65-67, 1986.

[14] R.Ferguson, A.Walton and M.Fallon, "An Interactive Process Modelling Simulator : its Capability and Performance", Process and Device Simulation, Alvey Club Meeting, pp.54-72, 1988.

[15] R.S.Ferguson and J.G.Doherty, "An Interactive Two-Dimensional Finite Element Process Modelling Package for a Single User Mini-Computer", Solid-State Electronics, vol.27, pp.1043-1054, 1984.

[16] Technology Modeling Associates TMA, "Depict-2", Two-Dimensional Process Simulation Program for Photolithography and Deposition, 1988.

[17] W.L.Engl, H.K.Dirks and B.Meinerzhagen, "Device Modelling", Proc.IEEE, vol.71, pp.10-33, 1983.

[18] J.A.Greenfield and R.W.Dutton, "Nonplanar VLSI Device Analysis Using the Solution of Poisson's Equation", IEEE Trans. Electron Devices, vol.ED-27, pp.1250-1531, 1980.

[19] S.Selberherr, A.Schutz and H.W.Potzl, "MINIMOS - A Two Dimensional MOS Transistor Analyzer", IEEE Trans. Electron Devices, vol.ED-27, pp.1540-1549, 1980.

[20] T.C.Denton and G.A.Armstrong, "An Improved Threshold Model Using Minimos 2", Process and Device Simulation, Alvey Club Meeting, p.7, 1987.

[21] T.C.Denton, "Validation of MINIMOS 2-2", Process and Device Simulation, Alvey Club Meeting, p.77, 1986.

[22] A.F.Franz and G.A.Franz, "BAMBI Users Guide", Technische Universitat Wien, Vienna, 1989.

[23] A.F.Franz, G.A.Franz and S.Selberherr, "BAMBI- A Design Model for POWER MOSFETS", Proc.ICCAD, pp.179-181, Santa Clara, 1984.

[24] D.Byrne, A.McCowen and K.Board, "Simulation of Bipolar Transistors Using Bambi", Process and Device Simulation, Alvey Club Meeting, p.64, 1986.

[25] D.Byrne, A.McCowen and K.Board, "Use of Bambi and Pisces for Bipolar Device Simulation", Process and Device Simulation, Alvey Club Meeting, pp.81-82, 1988.

[26] Technology Modeling Associates TMA, "Cande", Two-Dimensional Device Analysis Program, 1986.

[27] Technology Modeling Associates TMA, "Pisces-2B", Two-Dimensional Device Analysis Program, 1988.

[28] T.Shima, H.Yamada and R.L.Dang, "Table Look-Up MOSFET Modeling System Using a 2-D Device Simulator and Monotonic Piecewise Cubic Interpolation", IEEE Trans. Computer-Aided design of ICAS, vol.CAD-2, pp.121-126, 1983.

[29] T.Shima, T.Sugawara, S.Moriyama and H.Yamada, "Three Dimensional Table Look-Up MOSFET Model for Precise Circuit Simulation", IEEE Journal of Solid-State Circuits, vol.SC-17, pp.449,454, 1982.

[30] A.Vladeimirescu and D.Liu, "The Simulation of MOS Integrated Circuits Using SPICE2", University of California, Berkeley Memorandum No.UCB/ERL MBO/7, 1980.

[31] S.Liu and L.W.Nagel, "Small-Signal MOSFET Models for Analog Circuit Design", IEEE Journal of Solid-State Circuits, vol.SC-17, pp.983-998, 1982.

[32] A.Gribben, "MOSFET Characterisation and its Application to Process Control and VLSI Circuit Design", Ph.D Thesis, University of Edinburgh, 1988.

[33] M.G.Beuhler and Y.S.Lin, "Comparison of Results from Simple Expressions for MOSFET Parameter Extraction, IEEE Proceedings on Microelectronic Test Structures, vol.1, pp.91-96, Long Beach, Feb. 1988.

[34] M.H.White, F.Van de Wiele and J.-P.Lambot, "High-Accuracy MOS Models for Computer-Aided Design", IEEE Trans. Electron Devices, vol.ED-27, pp.899-906, 1980.

[35] B.A.Molin, "On MOS transistor model accuracy", Microelectronics Journal, vol.18, pp.5-11, 1986.

[36] G.T.Wright, "Simple and continuous models for the computer-aided design of VLSI", IEE Proceedings, vol.132 Pt.1,pp.187-194, 1985.

[37] G.T.Wright, "Physical and CAD Models for the Implanted-Channel VLSI MOSFET", IEEE Trans. Electron Devices, vol.ED-34, pp.823-833, 1987.

[38] P.Yang and P.K.Chaterjee, "SPICE modelling for Small Geometry MOSFET Circuits", IEEE Trans. Computer-Aided Design of ICAS, vol.CAD-1, pp.169-182, 1982.

[39] T.J.Jrutsick, M.H.White, H.-S.Wong and R.V.H.Booth, "An Improved Method of MOSFET Modelling and Parameter Extraction" IEEE Trans. Electron Devices, vol.ED-34, pp.1676-1680, 1987.

[40] I.E.Getreu, Modelling the Bipolar Transistor.

[41] A.Oakley, "Parameter Extraction and Optimisation Using TECAP Software", IEE Colloquium on Large Signal Device models and Parameter Extractions for Circuit Simulation", vol.1988/56, London, 19 April 1988.

[42] A.D.Hart and I.H.Leventhal, "Numerical Techniques Lead to Exact Solution of Berkeley Bipolar Model", Proceedings of the International Electron Device Meeting, 1986.

[43] Technology Modeling Associates TMA, "TOPEX", Parameter extraction and Optimization Program, 1987.

[44] A.Gribben, A.J.Walton and J.M.Robertson, "Accurate Physical Parameter Extraction for Small Geometry devices", Semiconductor International Conference Proceedings, pp.186-202, Birmingham, 1986.

[45] M.Duncan, "IPSEX Users Guide", Edinburgh Microfabrication Facility Internal Report, 1988.

CHAPTER FIVE : EXTENDING LOW VOLTAGE PROCESSES FOR 50 VOLT APPLICATIONS

5.1 Introduction

Traditionally power control has been realised using separate power transistors and logic circuits. The drive of this work was to integrate both of these functions to produce so-called smart power circuits [1]. These circuits can be used for automotive,display driver, telecommunications and domestic appliances. The reduction in chip count should lead to an increase in system reliability and a decrease in cost.

There are various strategies for realising this class of IC. One is to optimise all process steps for both power and logic. Unfortunately this solution is only cost effective when manufactured on a large scale, as this type of approach results in a complex process with many mask levels [2]. In order to address the ASIC market it is necessary to trade-off performance with cost. As there are large cell libraries and well characterised low voltage processes for logic in many semiconductor companies , it seemed logical to try to develop a power transistor that was compatible with a low voltage process with a minimum of additional masking steps. Since there are two such processes in the EMF, they formed a starting point. The first is a 6μm nMOS process. In order to understand how to develop a power process it was necessary to study the limitations of this low voltage process first.

The first experimental lot produced was used to determine the limitations of the EMF's standard 6μm NMOS process. This mature process was varied to determine the effect of different junction profiles and transistor channel dopant concentrations. By analyzing both transistors and junctions it was possible to determine the dominant mechanisms for breakdown in the transistors, and to postulate on how these limitations could be overcome.

The second stage was to try to modify both the process and the basic transistor structure in order to cope with higher voltages. The second experimental lot was based on the EMF's 5μm CMOS process since the objective of this work was to realise both high voltage n and p-channel transistors. The 5μm CMOS process was developed originally using contact printing lithography. With the introduction of steppers, it has been shown that this process can be scaled down to 1.5μm without any changes to the basic process. Obviously not all rules scale, for example the retention of the 5μm deep well means that the distance between p- and n- active areas remains constant. Power transistors are generally much larger than logic transistors, therefore the high voltage parts should be designed with one set of design rules and control logic another much tighter set. A matrix of process and device variants was devised to try to optimize the performance of the high voltage transistors.

5.2 Exploring the Limitations of the 6μm NMOS Process

In order to develop high voltage transistors an understanding of the breakdown mechanisms is needed. The theory for avalanche breakdown and the effects of limited junctions and corner rounding, were discussed in chapter two, and this section describes work done to try and correlate between theoretical and measured values.

This work was based on the EMF's 6 micron NMOS process. The wafers used were 3" <100> 14-20 Ohm.cm p-type, and the basic process is summarised as follows.

After an initial clean, a 500Å thermal oxide is grown as a buffer layer. Then 500Å of silicon nitride is deposited followed by a thin surface masking oxide. The buffer zone is needed to relieve the stress caused by the different thermal expansion coefficients of silicon and silicon nitride. Then photoresist is deposited and exposed to define the active areas. The nitride is etched in the exposed areas and the region, which will become field, is implanted with Boron (2×10^{13} atoms/cm^2 at 130 keV). Then a thick field oxide (1.3μm) is grown. This oxide combined with the implant helps prevent the formation of parasitic conduction paths. Now the remainder of the nitride is etched from the active regions.

The second photolithography stage opens up windows to form the resistor in the depletion transistors. There is now an Arsenic depletion implant (1.5×10^{12} atoms/cm^2 at 90 keV). After the resist strip an 800 Å gate oxide is grown in all active areas. This is followed by a threshold adjust implant of Boron (4×10^{11} atoms/cm^2 at 40 keV). A short anneal is used to activate and drive-in this channel

implant.

Buried contacts are defined with the third photomask layer. The gate oxide is etched away where polysilicon is required to contact directly to the diffused layer. Next polysilicon is deposited (5000 Å) followed by an oxidation to form a thin polyoxide layer. This polyoxide allows good adhesion with the photoresist.

The fourth photo layer defines the polysilicon gates and interconnections. This is followed by a blanket phosphorous deposition which diffuses into the polysilicon (and increases its conductivity) and into the exposed areas of silicon to define the source and drain regions. The process is self-aligning and therefore the Miller-Capacitance is reduced to near zero. This is the major advantage of using a self-aligned polysilicon process instead of metal for the gates.

A 2000 Å polyoxide is grown over the entire wafer to provide a base for the pyrolytic oxide deposition. It also helps reduce metal-polysilicon and metal-diffusion shorts caused by pinholes in the pyrolytic oxide. Heavily Phosphorous doped silicon dioxide (7500 Å) is deposited followed by a high temperature anneal which softens this phospho-silicate glass (PSG) to provide smooth step coverage. Then a short wet oxidation is performed to leach out some of the phosphorous from the surface of the pyrolytic oxide and to help adhesion of photoresist for the next mask level.

The fifth photo layer defines the contacts. Oxide is removed to form contact windows and another anneal is performed to round off these windows to

provide better metal step coverage. Aluminium is now sputtered to form a (1.2μm) thick layer and a sixth photo mask defines where the metal interconnects should be. The remaining metal is removed and a low temperature anneal (sinter) is performed to provide good aluminium-silicon contacts.

The seventh masking layer is used to define pads in a second, lightly-doped pyrolytic oxide. This oxide is used to protect circuits from ionic contamination and mechanical damage, but for test chips this stage is often omitted.

From simulations, it was predicted that the dominant mechanism in transistor failure would be punchthrough. Therefore it was necessary to determine the minimum channel length where this effect was not present. Simulation also predicted that for diode structures the breakdown voltage would be higher than the transistors, but in order to confirm this and to investigate effects such as junction curvature, various diode and gated-diode structures were designed. It was clear from simulations that simply enlarging the channel length could increase the peak operating voltage, but this may not be the best solution. By investigating discrete diodes, transistors and proximity effect structures , some insight in how to create a new transistor structure that has two regions, one to support voltage and one to control current, could be found.

For these reasons, in addition to the basic process, some variants were tried to investigate their effects on breakdown performance. These variations are summarised in table 5.1. The field implant was varied from the standard 2×10^{13} atoms/cm^2 , a lower 5×10^{12} atoms/cm^2 and finally omitting the implant altogether. The idea was to investigate if this implant was necessary under the normal design

rules, and at what distance of field isolation spacing it becomes critical.

Wafer #	Field Implant atoms/cm^2	Threshold Implant	Source Drain Doping	N-region around drain	N-well used as n-region around drain
1	0	Yes	Arsenic	No	No
2	5e12	Yes	Arsenic	No	No
3	2e13	Yes	Arsenic	No	No
4	0	No	Arsenic	No	No
5	0	Yes	Phos.	Yes	No
6	0	Yes	Phos.	No	No
7	2e13	Yes	Phos.	No	Yes
8	2e13	Yes	Arsenic	No	Yes

Table 5.1 A summary of the process variations used. (the dose of the field implant, the use of the threshold adjust implant, n- around the drain, n-well as a n- region around the drain and the dopant used for the source & drain implant)

Another process split was used to see the effect of omitting the threshold adjust implant. From simulation it was predicted that the natural threshold voltage was high enough to be immune to noise, and it was shown that this implant reduces the transistor breakdown voltage due to the increased doping concentration in the channel region. The effect of using an arsenic implant instead of a phosphorus diffusion for source and drains was investigated, as was the effect of using an lightly doped n- region surrounding the source/drain regions. This latter variant is shown in figure 5.1. In this structure the lightly doped n- region should help to support high electric fields, but at the price of an increase in transistor on-resistance. Finally an investigation of deep n-well breakdown was attempted by depositing phosphorus in the active areas followed by a long (17 hours, 1050°C) drive-in, as used in the 5μm CMOS process. Then the normal n+ source/drain diffusion is used to form a lower resistance contact to the metal layer. This structure is shown in figure 5.2. The purpose of this was to see if this n-well diode could be used to form part of the hybrid transistor-resistor structure discussed previously.

In order to investigate both process and design variations a test chip was designed using the 'Caesar' design package. The resulting 'cif' files were sent to the Rutherford laboratories for mask making. The test chip contained an nMOS test transistor matrix with various aspect ratios. These transistors varied from 2μm : 1μm through to 30μm : 30μm. The matrix is such that all transistors are labelled with their respective aspect ratios using the metal layer. This makes identification of devices much easier for testing purposes.

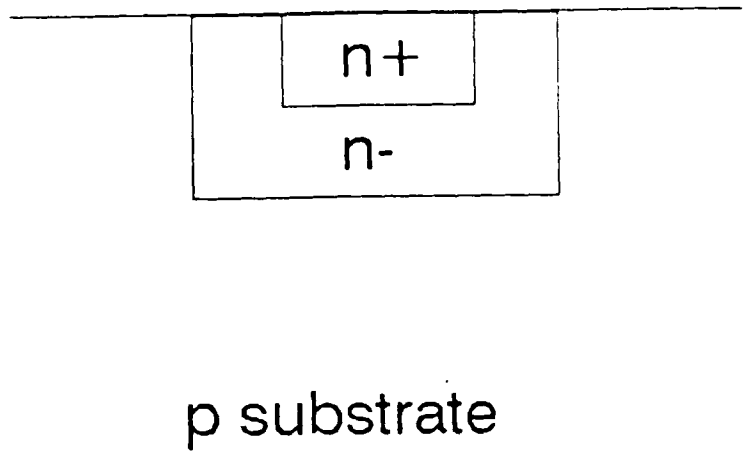


Figure 5.1 n+n-p diode formed using a double diffusion.

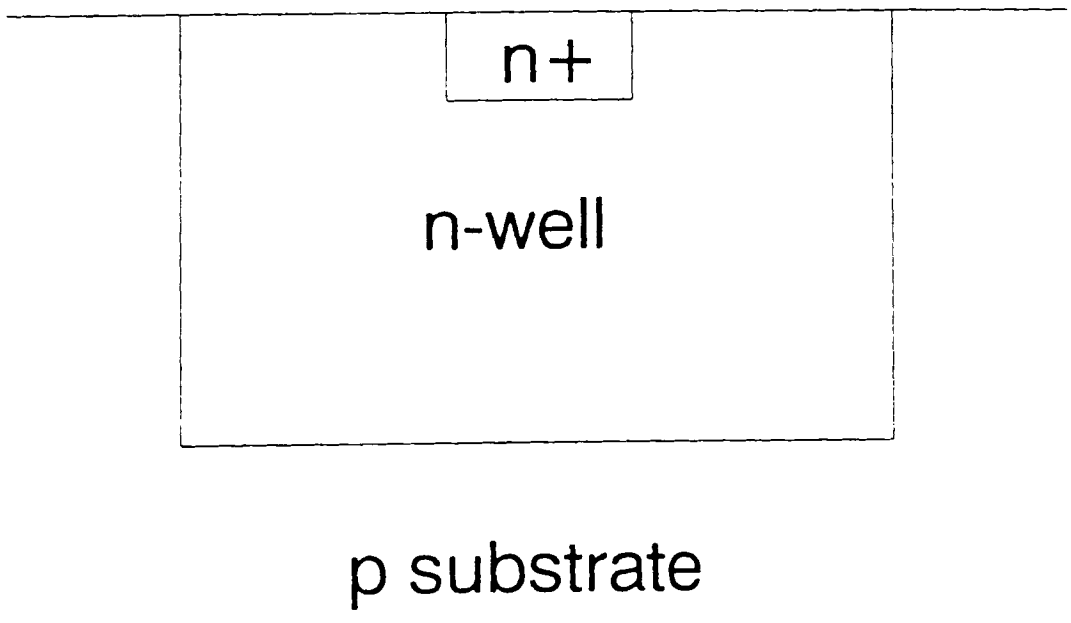


Figure 5.2 n+n-p diode formed using the n-well diffusion.

To investigate possible field distortions within transistors due to transmitted signals above active areas, a group of transistors with voltage controllable metal patterns were laid out.

Various diffusion structures were laid out to investigate junction breakdown as compared with the gate-assisted junction breakdown of the MOS transistors. These structures included square diffusion patterns, sharp angled diffusion patterns and structures to look at proximity effects of closely placed, independent junctions.

This test chip was fabricated in the EMF between 12/2/87 and 13/3/87. The batch number was 713 and the mask set number was EU643.

5.3 Results of Diode Structures and Transistor Variants

The drain breakdown voltage of the transistors was measured using an HP 4145 semiconductor parameter analyzer connected to a Wentworth manual probe station. For the test the gate and substrate were grounded, whilst the drain voltage was biased. When the drain-substrate current exceeded 1μA then breakdown was said to be occurring. It was found that in some of the transistors that this breakdown was in two stages, as in figure 5.3. The lower voltage is breakdown due to either local lattice defects or corner effects. The bulk breakdown was typically around 2 Volts higher. The results of drain breakdown

voltage are summarised in table 5.2. By observing the structures at breakdown using a microscope in the dark, the critical points, for example angles of structures, could be observed in order to determine which part of the structure failed first.

Wafer #	**BV transistor gate < 3μm**	**BV transistor gate > 3μm**	**Threshold Voltage**
1	31V	21V	2.0V
2	32V	21V	2.0V
3	28V	21V	2.0V
4	34V	35V	1.1V
5	33V	29V	0.1V
6	33V	22V	2.0V
7	n/a	n/a	n/a
8	n/a	n/a	n/a

Table 5.2 Breakdown and threshold voltages for transistors with gate lengths greater and less than 3μm.

There were two distinct values found for breakdown voltage, a high value for those transistors with gate length (on mask) of 3μm and less and a lower one for those transistors with gate lengths above 3μm. From SUPREM3 simulations

it was found that this distance is approximately twice the sideways diffusion of the source/drain areas, therefore it seems that the source and drains are shorted as in figure 5.4.

In the transistors that did not have the threshold adjust implant, the breakdown voltage was found to be higher than those transistors with the implant. The reduction in BV_{DSS} in the transistors with the implant is due to the fact that the drain breaks down into this channel implant area with its higher doping as in figure 5.5, rather than into the bulk as shown in figure 5.6.

The large geometry (> 3μm) transistors had the same breakdown value with both the phosphorus and the arsenic source/drain doping. The breakdown here is limited by the threshold adjust implant. For the merged source/drain transistors there was a difference in breakdown voltage due to the different impurity profiles.

It was seen that for these transistors the field implant did not have an effect on the breakdown voltage. This can be explained by the fact that the transistors are separated by at least 240μm, therefore parasitic device effects will be negligible. However this implant is necessary in circuits where the packing density is typically fifty times higher. Simulations have shown that if this separation is less than 15μm there is the chance of leakage between one transistor and another, and that at 10μm there is the possibility of turn on of a parasitic MOS transistor if a metal line runs over the field oxide at this point. For the double-diffused transistors, the breakdown voltage was found to be higher. This can be explained

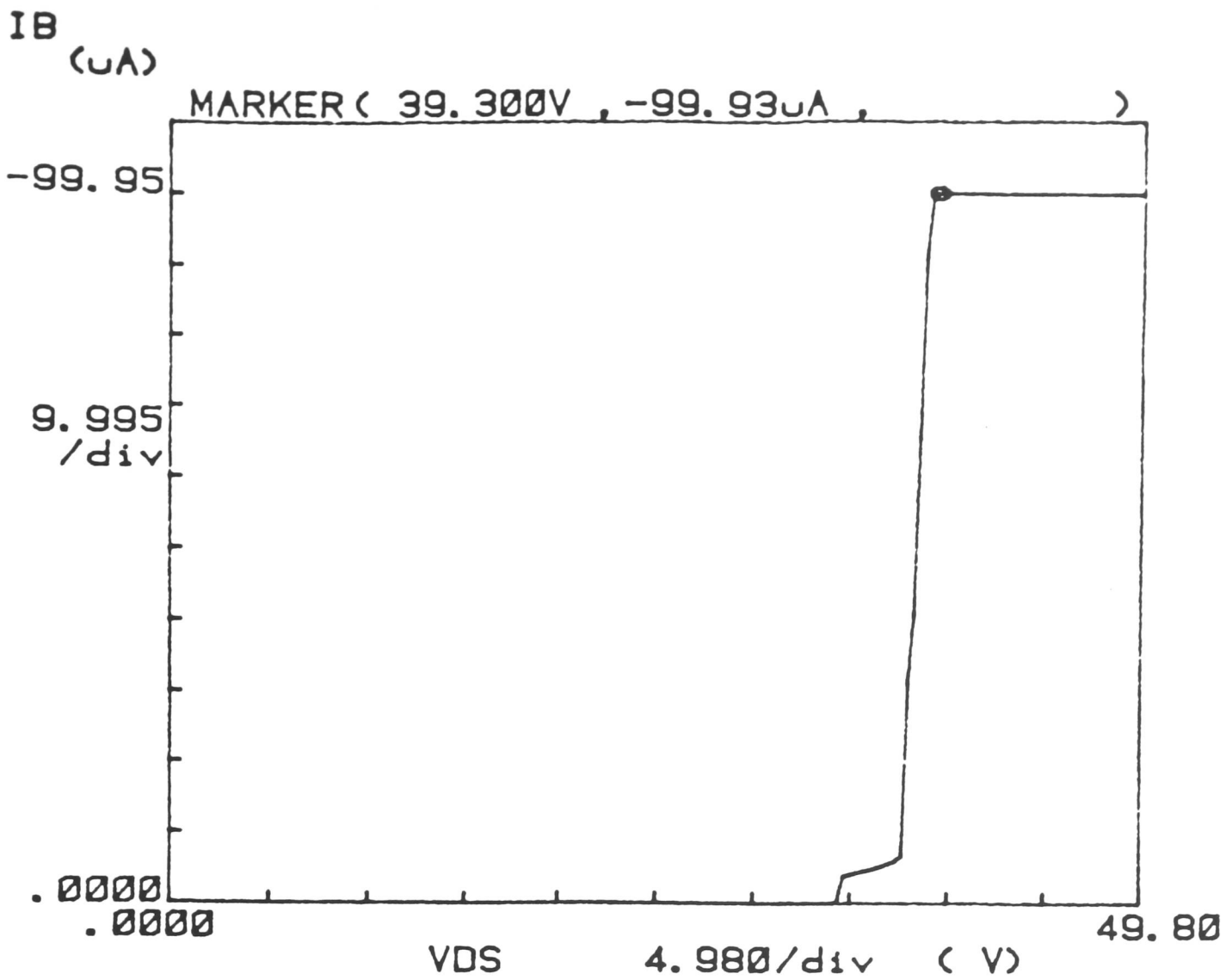


Figure 5.3 I-V breakdown characteristics showing the two stages of breakdown.

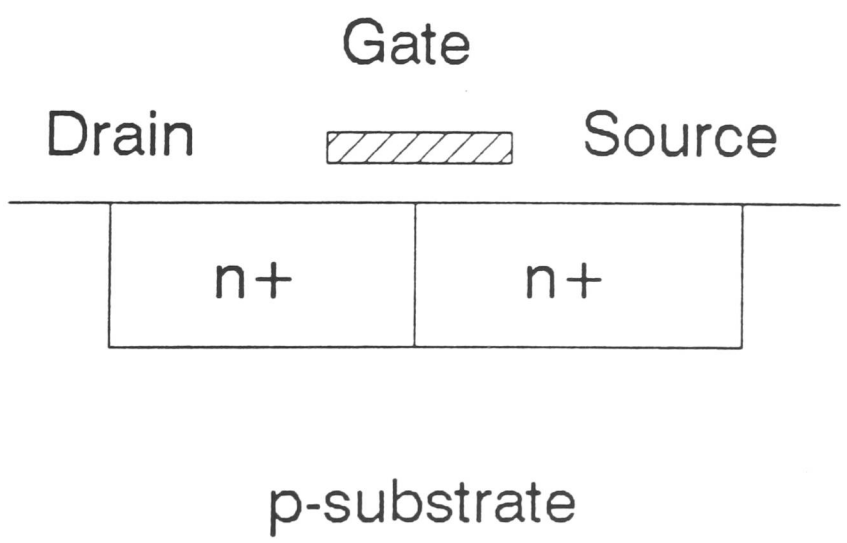


Figure 5.4 Shorted out transistor that results when the gate length is less than 3μm.

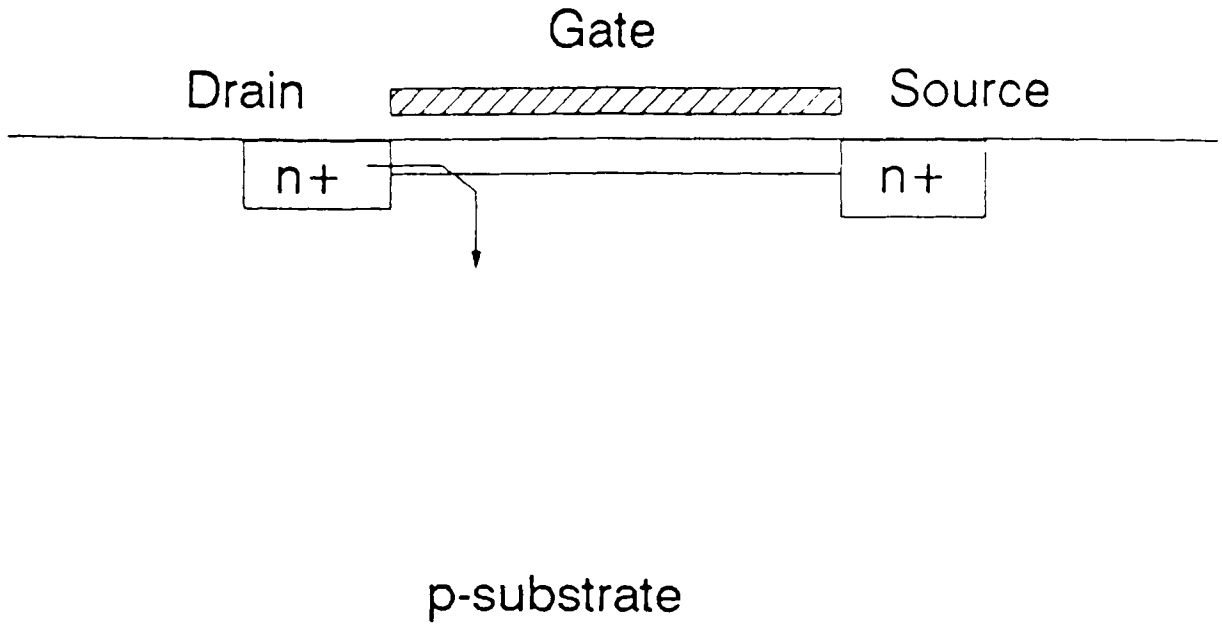


Figure 5.5 Breakdown through the channel in a transistor with the threshold adjust implant.

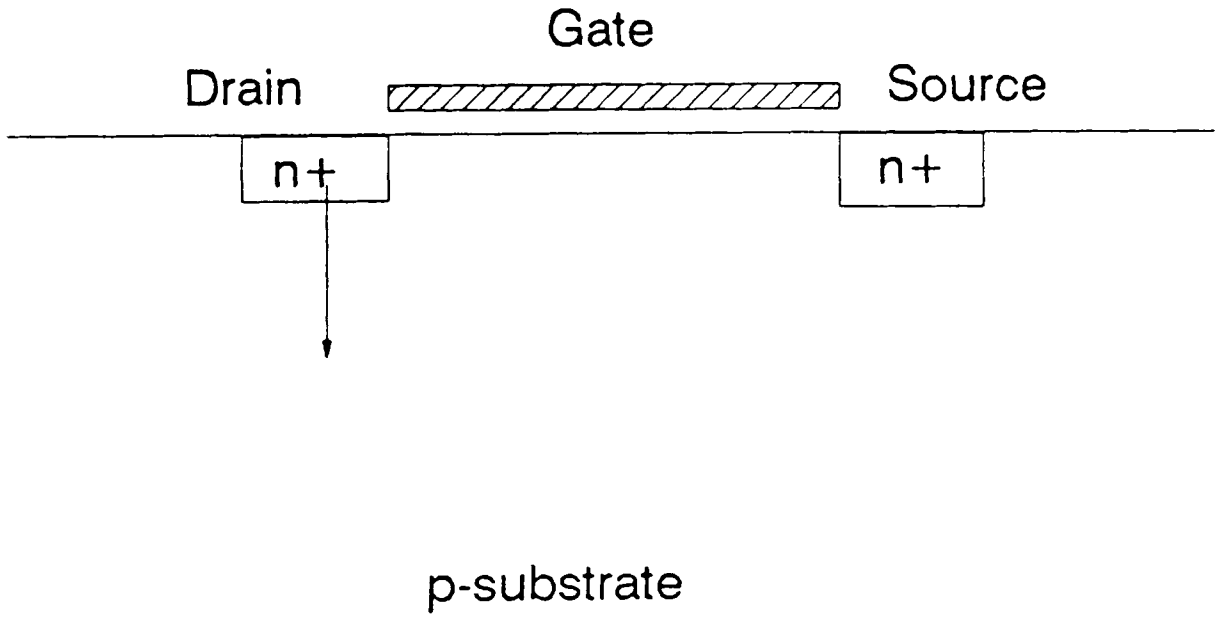


Figure 5.6 Breakdown to the substrate in a transistor without the threshold adjust implant.

by the fact that the extra temperature step after the threshold adjust implant drives the channel dopant deeper than before, therefore lowering its peak concentration and so increasing the maximum safe operating voltage.

The results of measurements on threshold voltage are shown in table 5.2. From this it can be seen that although the double-diffused transistors have a higher breakdown, they are not practical enhancement-mode transistors as the threshold voltage is too low. This is due to two effects. Firstly, the n- region reduces the electrical channel length. Secondly, the drive-in of the n-region moved the threshold adjust implant away from the surface. The Arsenic source/drain devices without threshold adjust implant have the highest breakdown and have a practical value of threshold voltage. This confirms the simulated results, therefore it is proposed that this implant is omitted in future for all runs using the standard nMOS process.

Finally, the gate-insulator dielectric breakdown strength was measured and found to be 80 Volts. This is substantially higher than the drain breakdown and therefore does not prove to be a limitation at this voltage range. However, it could not be used in 100 Volt transistors without modifying the transistor structure.

Within the diode structure matrix the breakdown voltage varied considerably. Structures with acute angles and thin conductors tended to breakdown at lower values than the merged short-channel transistors (typically 25-30 Volts). For example, the cross structure has two conductors of width 12μm at right angles to one another. The breakdown of this structure depends on the field implant. When this implant is increased, the breakdown decreases. A similar effect is seen

with the 90 degree angle structure, which consists of two conductors of width 12μm with a right-angle bend. It is also seen by comparing these results with the 45 degree angle structure, which has two join between the two conductors at 135 degrees, that the angle is also an important factor. This in confirmed, in that when a large square (25μm x 25μm) is added to the centre of the cross structure the breakdown voltage increases. It was found that the simple diffusions had higher breakdown voltages (typically 40-50 Volts) for the 25x25μm and 75x75μm diodes. Therefore as predicted from the junction theory discussed in chapter 2, acute angles and thin tracks should be avoided wherever possible. Table 5.3 shows a summary of all structures and all wafers measured.

Wafers 7 and 8 used the n-well as part of the diode, therefore only the large structures could be measured. The X-mas structure consisted of an equilateral triangle with sides of 25μm and a 'trunk' of 12μm width. The concertina consists of eight 25μm equilateral triangles in a line point to base. However, these triangles are overlapping such that the point is 10μm inside the base of the successive triangle. Therefore the thinnest part of the structure is 10μm.

The wafer that had the n-well doping used as a diode showed breakdown voltages of the range 95-105 Volts. This value is significantly higher than the shallower junctions and indicates that deeper junctions with lower peak doping concentrations will be required in order to support the high voltages necessary for power devices. Further, it was found using SUPREM III and PISCES simulations, that this breakdown voltage was limited by the overlap used for the n-well around the drain diffusion. By increasing this overlap it is predicted that a diode formed by the n-well could support nearly 200 Volts.

Structure \ Wafer	#1	#2	#3	#4	#5	#6	#7	#8
X-mas tree	47	37	27	39	39	35		
Concertina	47	42	30	36	36	36		
Contact 5 x 5μm	57	47	32	35	35	36	66	97
33x12μm rectangle	58	47	32	35	35	37	72	110
75x75μm square	58	47	32	30	39	37	72	110
12x14μm rectangle	58	47	32	34	47	36	74	93
Cross with centre	50	39	29	34	34	34	70	101
Cross	30	25	23	33	33	33		
45 degree angle	46	37	27	34	34	34		
90 degree angle	30	25	23	34	34	34		
2 x 90 degree angle	52	41	29	34	34	36		
4 x 90 degree angle	53	41	29	34	34	35		

Table 5.3 Breakdown voltages for various diode structures (given in volts) for the eight wafers described in table 5.1.

5.4 Extending the 5μm CMOS Process

After exploring the limitations of the nMOS process, the next step was to investigate CMOS devices. The reason for this is that the objective of this work was to produce both high voltage nMOS and high voltage pMOS transistors on the same die. The other goal was that if possible, the high voltage devices should remain compatible with existing low voltage transistors so that existing cell libraries can be exploited.

It seemed prudent to attempt to increase the voltage in stages and this first iteration set out to raise the maximum voltage from 15 Volts to 50 Volts. Transistors with this level of performance can be used in motor drives, telecommunications and automotive applications. This voltage level is seen by industry as a natural first level in smart power product ranges [2].

In order to achieve this performance level, modifications were needed in both the process and in the device structure. These changes will be detailed later, but first consider the standard EMF 5μm CMOS process.

To produce complementary transistors a well must be introduced. Either n- or p-wells may be selected, but as most modern single channel processes are nMOS, the EMF 5μm process, which uses n-wells with the p-channel transistors formed inside tubs, was selected. Like the nMOS process the 5μm nMOS process uses 3" <100> 14-20 Ohm.cm p-type substrates.

The first masking stage in this process defines the n-wells. There is a Phosphorus implant (2 x 10^{12} atoms/cm^2 at 100 keV) followed by a well drive-in.

This consists of an initial sacrificial oxide growth followed by a 17 hour diffusion in nitrogen at 1150°C. This drive-in creates 5μm deep n-wells.

An oxide-nitride-oxide sandwich is then deposited over the wafer, as for the nMOS process. The second photolithography stage now defines the active areas, and the nitride is removed from the other areas. Then a third masking layer is used to cover all the n-well areas during the field implant. This field implant of Boron (7×10^{13} atoms/cm^2 at 25 keV) helps prevent parasitic transistors being formed. The resist is then stripped and a thick field oxide of 1.3μm is grown over the wafer. Since this oxide grows at a much slower rate on top of the silicon nitride, removal of the sandwich will leave bare silicon active areas separated by thick local oxide (LOCOS) that further prevents parasitic devices.

Next an 800 Å gate oxide is grown in the active areas. This is followed by a Boron threshold adjust implant (5×10^{11} atoms/cm^2 at 40 keV). A short anneal is used to activate and drive-in this channel implant. A 500 Å polysilicon layer is then deposited over the wafer and is doped n+ with a blanket phosphorous deposition.

The fourth photo layer defines the polysilicon gates and interconnects. The excess polysilicon is etched away to leave bare silicon in the active areas. The fifth photo layer defines the areas for the n+ implant. It covers all PMOS source/drains in n-wells and all p+ substrate contacts, whilst trying to maximise the area of polysilicon interconnects subjected to the n+ implant. The implant is Phosphorous with a dose of 7×10^{15} atoms/cm^2 at 90 keV.

The sixth mask for p+ implant is a field reversal of the n+ mask. Windows are opened around all p+ source/drains in n-wells and around all p+ substrate contacts. The implant is Boron with dose of 2×10^{15} atoms/cm^2 at 35 keV.

A wet oxide is grown over the entire wafer to provide a base for pyrolytic oxide. Then heavily phosphorus doped silicon dioxide (PSG) is deposited followed by a wet oxidation to leach out some of the phosphorus and to help adhesion of photoresist.

The seventh photo layer defines the metal contacts. Oxide is removed to form contact windows and another anneal is performed to round off the windows and to further drive in the n+ and p+ implants. Aluminium is now sputtered to form a 1.2μm thick layer, and the eighth photo mask defines where the metal interconnects will be. The remainder of the process consists of sintering, overlay pyro deposition and a ninth photo layer to define holes in the overlay glass to allow probing of pads.

The test chip contains various groups of transistors. Each transistor is labelled with the name of the group and the dimensions. Each of these groups will be referred to in the remainder of this chapter by the label used in the design. For example, the first group of transistors to be described will be referred to by its label name, NN. The NN transistors are simple NMOS devices. They have channel dimensions of 5μm x 5μm, 10μm x 10μm, 15μm x 15μm and 20μm x 20μm (see fig 5.7). The different sizes may show slightly different effects, but should have similar current handling performance as the aspect ratio is kept constant. This group is used as a control, and to help calibrate the simulators so that simulation

can be used in future to develop new structures.

Figure 5.7 illustrates the PP devices, which are standard PMOS transistors as fabricated normally in the EMF 5μm CMOS process. These devices do not, however, have any well contact. In order to investigate the effect of a butting n+ contact between the source and well, PB devices have been fabricated. They have a n+ region butting against the p+ source to give a contact to the n-well.

It is possible to achieve higher breakdown at the expense of higher on-resistance by lightly doping the source and drain. However, to achieve a good contact a high doping is necessary. One solution is the double-diffused (DDMOS) technique. By using the n-well mask to create a lightly doped source and drain, then implanting with the n+ mask into the same region, the WN, DDMOS transistors are formed (as shown in figure 5.8).

The power-handling capability of MOS devices can be increased by altering the basic MOSFET structure. This is achieved with the incorporation of a lightly-doped drift region. This region largely supports the applied drain potential because its doping level is chosen to be much smaller than the channel region. These new structures effectively separate the active portion of the device (channel) which determines device gain, from the region which supports the applied voltage (drift region). This separation is analogous to modern bipolar transistors in which a lightly doped collector region largely supports the applied potential and a narrow, more heavily doped base region largely determines device gain.

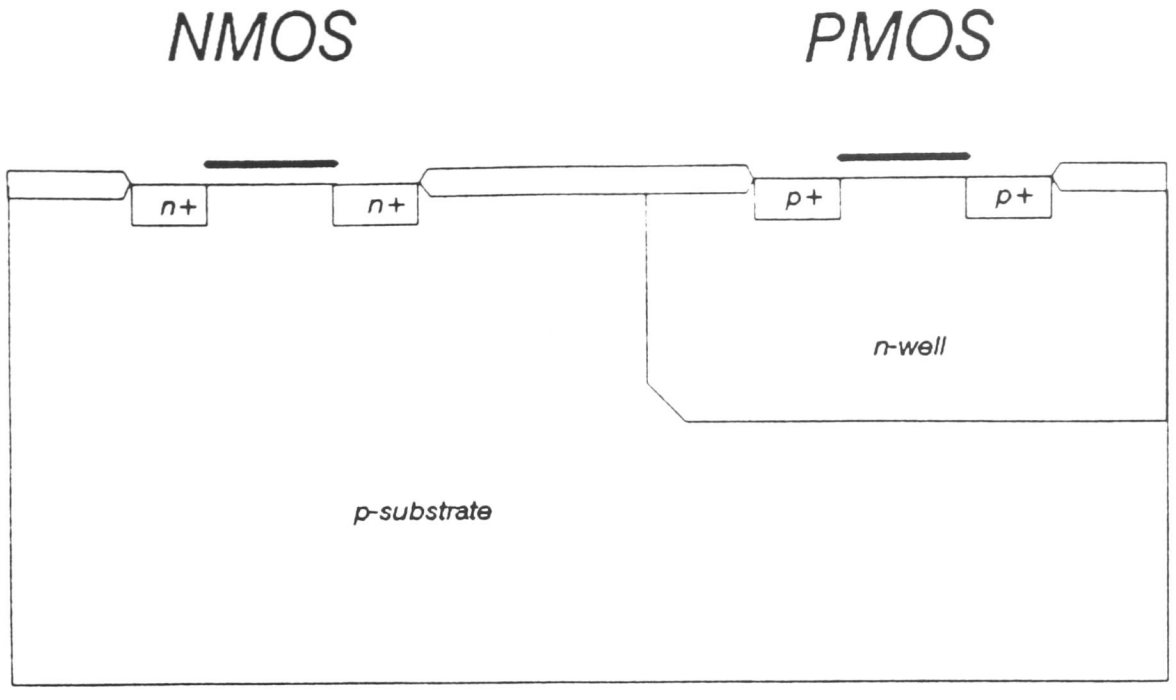


Figure 5.7 The simple transistors labelled "NN" for the nmos and "PP" for the pmos.

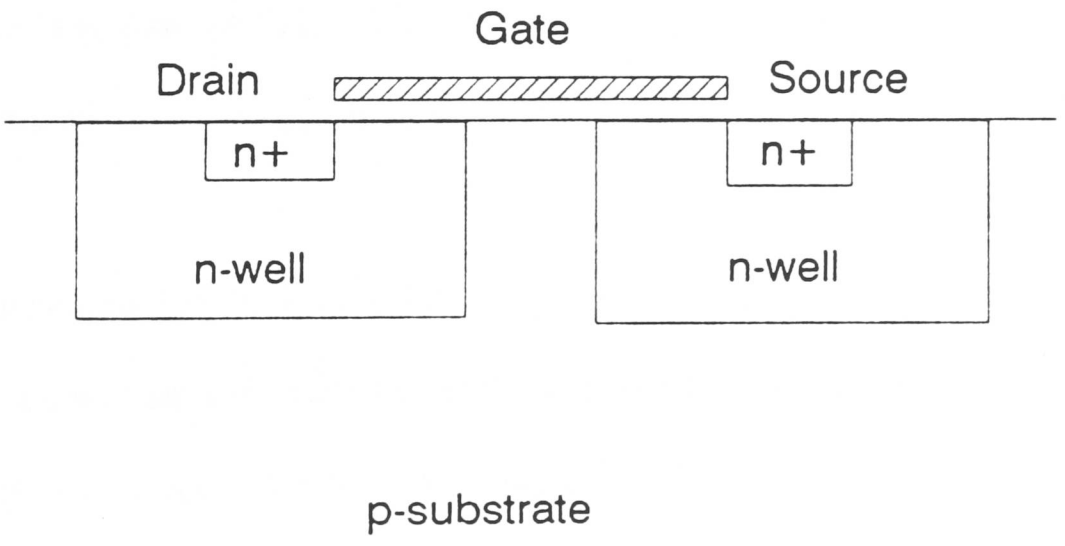


Figure 5.8 The double-diffused DDMOS transistors formed using the n-well as the lightly doped region.

Two different implementations are possible, either a lateral or a vertical structure as discussed in chapter 3. The two structures most suitable for integration were previously found to be the LDMOS for lateral solutions [3] and the VDMOS for vertical applications [4] (see figures 5.9 and 5.10). The first of these is the lateral double-diffused transistor (LDMOS). The fabrication process is as for standard MOS transistors, but with the following changes. Consider first the p-channel transistor. Instead of the n-well containing the whole transistor, it only overlaps the source. The result of this is that the well acts as the channel of the device and the substrate acts as the drift region. In this case the effect of scaling the device is to change the drift region, the part that supports the high voltage. In order to look at the effect of changing the channel length both LP (figure 5.11) and LPB (figure 5.12) variants have been designed. The LP devices have a drawn channel length of 5μm greater than the LPB devices. By reducing the channel length, there is a reduction in on-resistance but a greater susceptibility to punch-through. A third variant tried was the PL devices. These use an n+ implant instead of the n-well to produce the channel. The structure is similar to those discussed above, but the channel region has a higher doping than before. Therefore it is expected that these devices will have a lower on-resistance, but also a lower breakdown voltage than the other two types.

The n-channel devices are constructed in a similar manner, but this time the substrate is used as the channel, and the n-well is used as the drift region. They are named LN (figure 5.13), LNB (figure 5.14) and NL respectively, and the variations in structure are analogous to the p-channel devices.

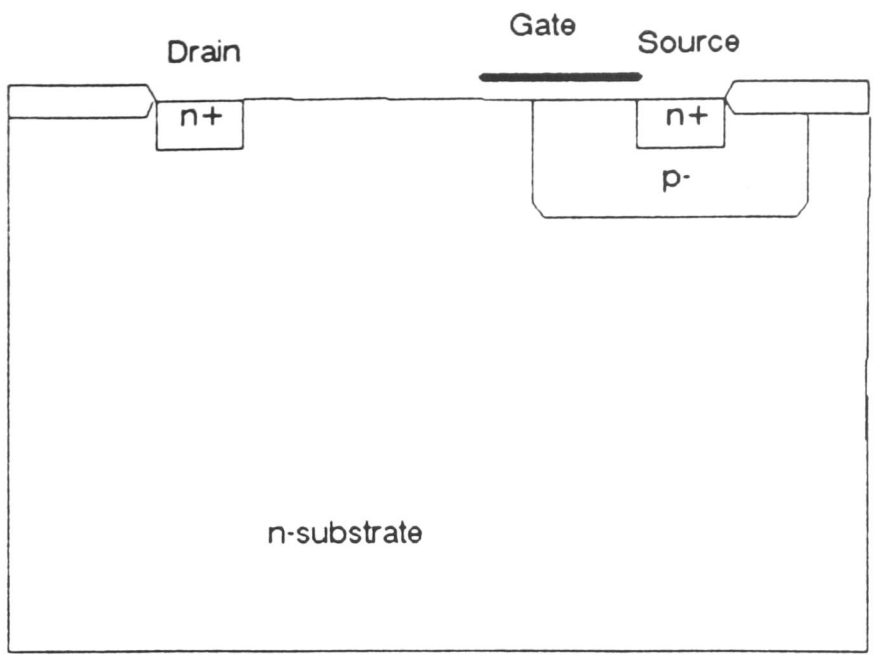


Figure 5.9 A lateral double-diffused LDMOS transistor.

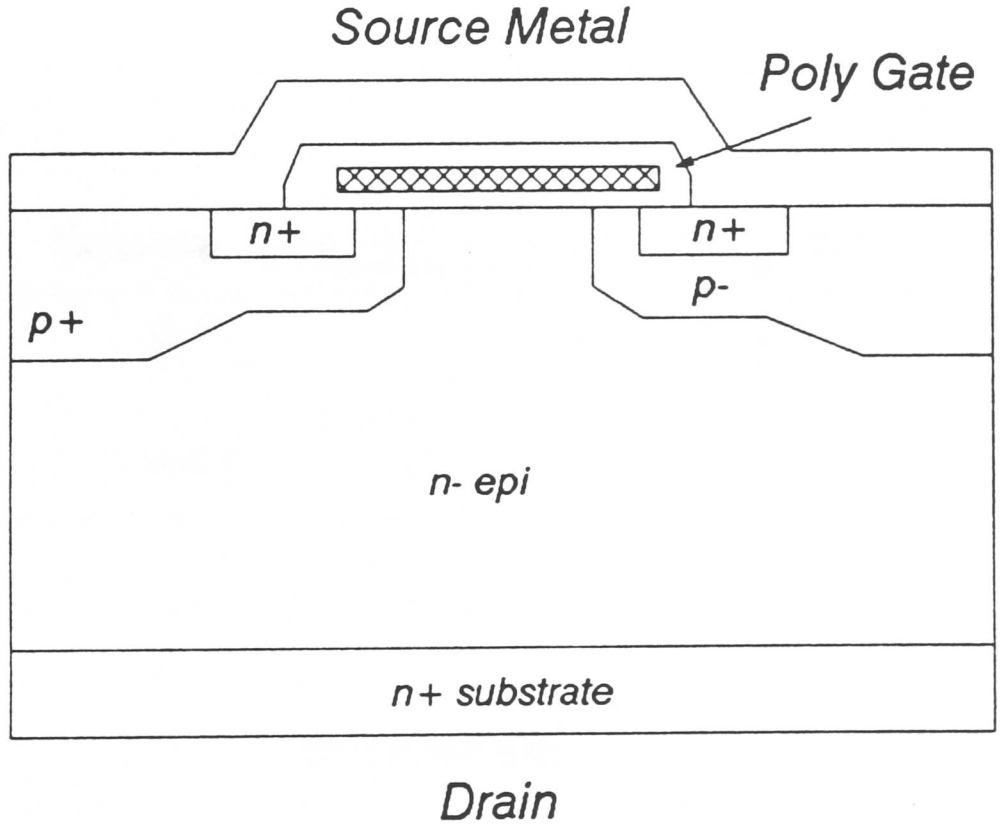


Figure 5.10 A vertical double-diffused VDMOS transistor.

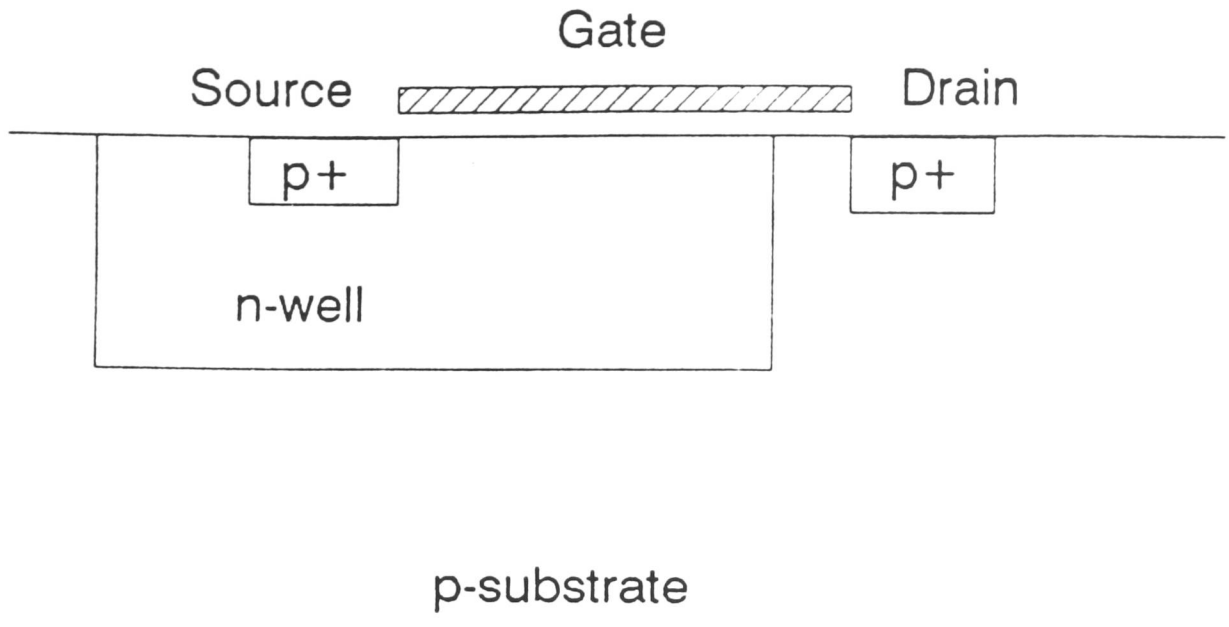


Figure 5.11 The "LP" p-channel LDMOS transistor using the substrate as the lightly doped region. The n-well forms the channel and extends far under the gate.

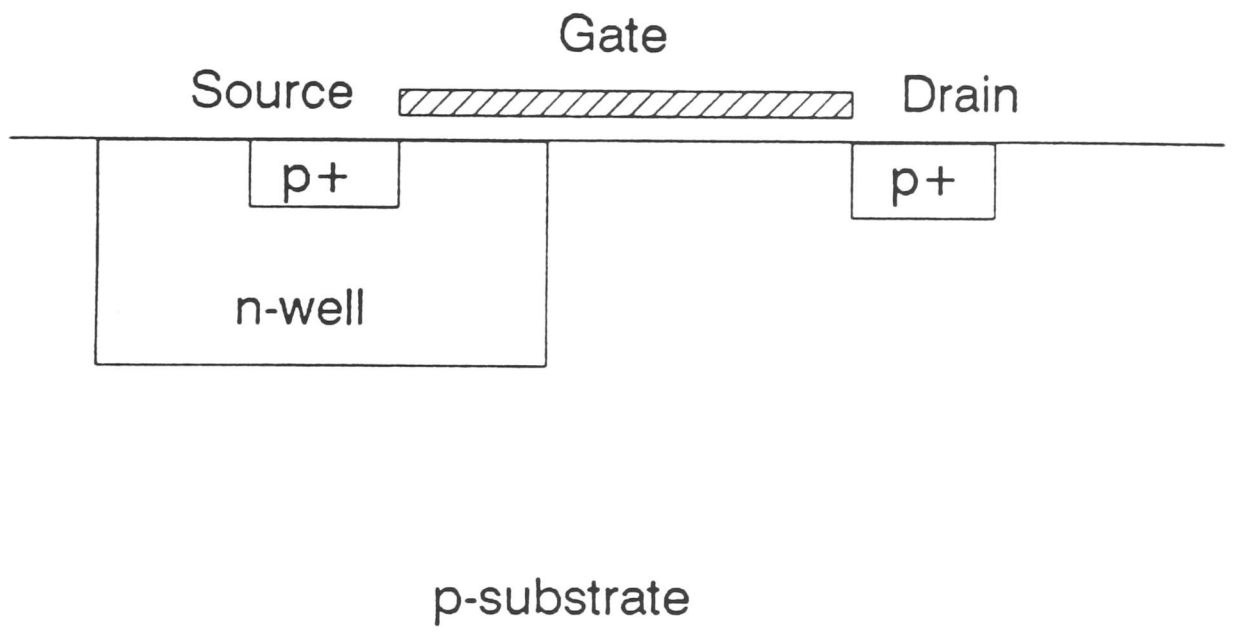


Figure 5.12 The "LPB" p-channel LDMOS transistor using the substrate as the lightly doped region. The n-well forms the channel and extends a short distance under the gate.

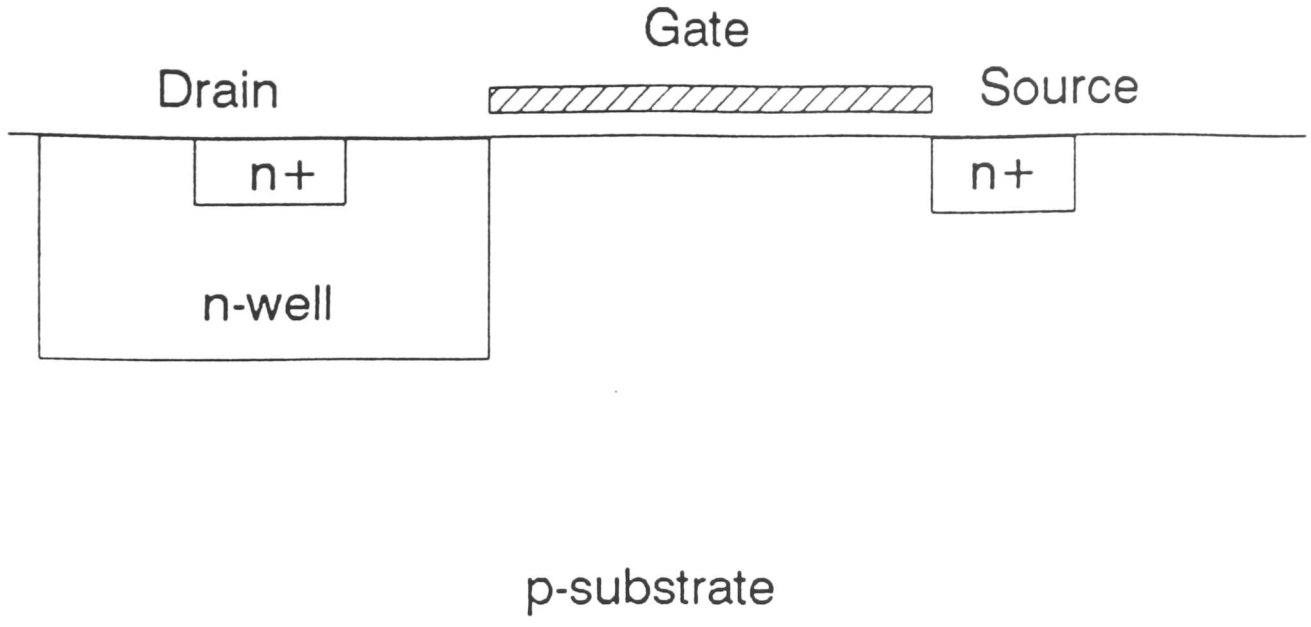


Figure 5.13 The "LN" n-channel LDMOS transistor using the n-well as the lightly doped region. The n-well is defined so as not to extend under the gate.

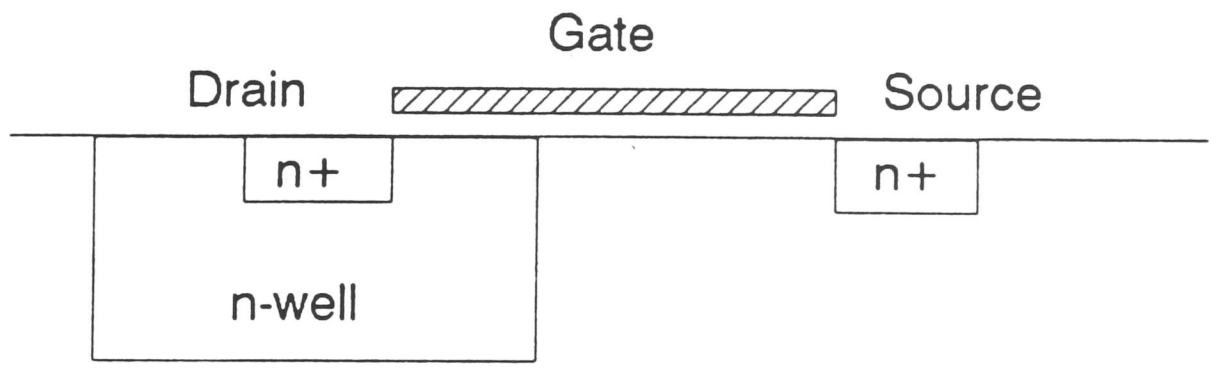


Figure 5.14 The "LNB" n-channel LDMOS transistor using n-well as the lightly doped region. The n-well is defined so as to extend a short distance under the gate.

The other major type of power transistor investigated here was the vertical double-diffused VDMOS. With this structure there are two source contacts on the surface. These are located in the same place as the source and drain of a conventional MOS device. The sources are implanted with the opposite type of impurity to the substrate. This is then driven in and an implant of the same type as the substrate is performed. The drain contact is made to the backside of the wafer. The wafer is typically heavily doped to form a good contact at the drain, with a lightly doped epitaxially-grown layer used to form the drift region of the transistor.

The limiting factor of the research was to develop a technique of fabricating power transistors alongside logic transistors with a minimum of extra process steps. This meant that the vertical devices fabricated could not be truly vertical, as up-down sinkers, epitaxial layers and buried layers were not adopted. The current flow is initially down into the substrate and the current is collected at the surface using a lateral drain. There will be current loss into the substrate and leakage to other devices, but the primary consideration was to determine if this type of approach was feasible or not (see figure 5.15).

Consider first the n-channel transistors. The VN devices are formed using the same process as the PP device but with the following variations. There is a fourth terminal which is an n-type connection to the well. This terminal acts as the drain. There are two source terminals, the original source and drain of the PP devices. The channel is formed by first implanting p+ into the active areas. This is followed by a short drive-in and a subsequent n+ implant to form the sources.

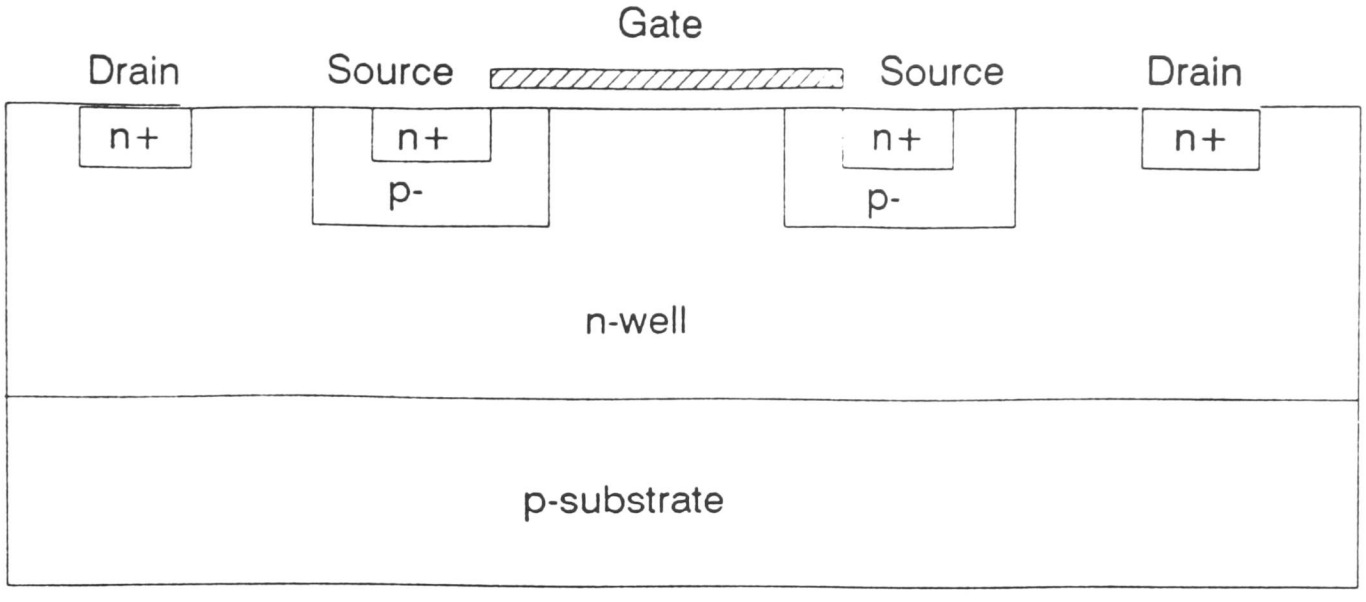


Figure 5.15 A vertical n-channel VDMOS transistor formed with all contacts on the surface.

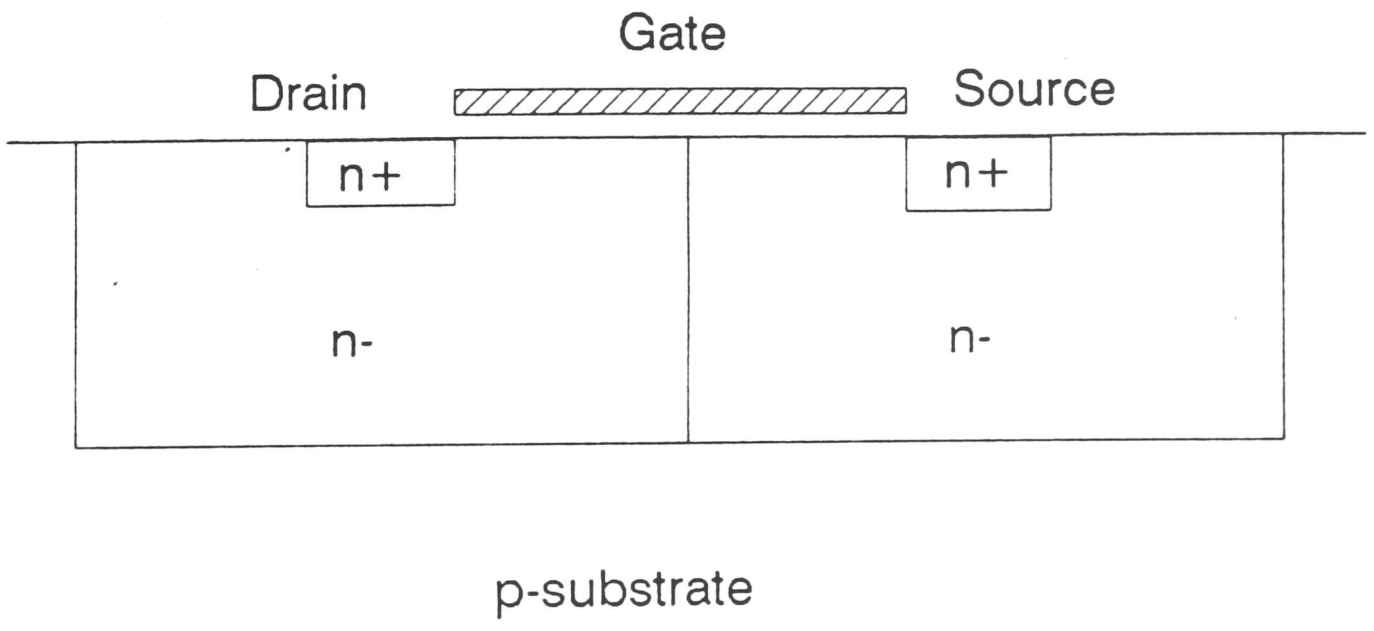


Figure 5.16 A double-diffused DDMOS transistor with a 5μm drawn channel length. The drain and source are shorted out due to the lateral diffusion of the n- dopings under the gate.

The NV devices are similar to the VN devices but have larger drains in order to collect more of the current from the sources (see figure 5.15). The VP and PV devices are the p-channel equivalents of the VN and NV devices respectively.

In addition to the geometrical variations, some process variations were also made. The steps that have been modified are the n-well drive-in and the n+ and p+ source/drain implants.

By using a process simulator (ICECREM) it was found that the normal well depth is around 5μm. In some of the devices, the well is being used as a drain and the standard well may be too deep. An alternative well was designed that has similar concentration, but a depth of 2μm. It may sound strange that a 2μm well can be used in a 5μm CMOS process, but the EMF CMOS process is in fact scalable to 1.5μm. The reason for this is that the process was developed several years ago for use with contact printing. However, with the use of steppers and the subsequent improvement in resolition and alignment, it is possible to scale some of the design rules whilst maintaining the same process steps.

For both the n+ and p+ implant three different doses have been tried. The object of this is to try and find the optimum trade-off between breakdown voltage and on-resistance. Table 5.4 shows which process variations were used on each of the wafers.

Wafer #	N-well junction depth (μm)	N+ dose atoms/cm^2	P+ dose atoms/cm^2
1	5	7e15	2e15
2	5	1e14	1e14
3	5	1e13	1e13
4	2	7e15	2e15
5	2	1e14	1e14
6	2	1e13	1e13
7	5	*	2e15
8	5	7e15	#

where * signifies B11 2e15 @ 35keV + 30 min drive in

+ P31 1e13

and # signifies P31 7e15 @ 90keV + 30 min drive in

+ B11 1e13

Table 5.4 Summary of the process variations (well depth and source/drain dopings) used for the CMOS split lot.

In addition to the variations described previously, two others were used to produce the VDMOS devices. The first of these was to implant p+ through the n+ mask, drive in the p+ for half an hour and then implant n+. The second was to implant n+ through the p+ mask, drive in for half an hour and then implant p+. In total there were eight different wafers processed.

5.5 Results of the Transistor Variants

The measured and simulated results of the process variations on the simple nMOS and pMOS transistors are summarised in table 5.5. It can be seen from this table of results that wafers 1&4, 2&5 and 3&6 have similar values for breakdown voltage. This shows that there is no effect on breakdown by decreasing the n-well depth if the surface concentration is kept the same. It was also found that there was no effect on threshold voltage either. Therefore this shallow depth could be adopted for the standard process. Since the vertical depth is reduced, the lateral diffusion is less and therefore packing density could be increased. Table 5.5 also shows the results of simulation using SUPREM III and CANDE to determine threshold and breakdown voltage. It can be seen that the fit is reasonably good. This increases our confidence when using these simulators in a predictive mode to develop new structures in later work [5].

For the standard nMOS devices (NN), the breakdown voltage is found to be dependent on n+ doping concentration. It varies from 18 volts to 34 volts as the source/drain implant dose is reduced. Similarly for the standard pMOS devices the breakdown voltage increases from 18 volts to 42 volts with decreasing source/drain doping density. However, by reducing the source/drain implants there is an increase in on-resistance and therefore a decrease in current density. For the pMOS transistors it was found that butting n+ source contacts did not effect either the threshold or breakdown voltage.

Wafer #	#1	#2	#3	#4	#5	#6
Measured BV NMOS	18	24	33	18	24	34
Measured Vt "	1.4	1.4	1.4	1.4	1.4	1.4
Simulated BV "	20	26	34	20	26	34
Simulated Vt "	1.4	1.4	1.4	1.4	1.4	1.4
Measured BV PMOS	-20	-29	-42	-19	-28	-42
Measured Vt "	-1.5	-1.5	-1.5	-1.5	-1.5	-1.5
Simulated BV "	-20	-30	-43	-20	-30	-43
Simulated Vt "	-1.4	-1.4	-1.4	-1.4	-1.4	-1.4

Table 5.5 Simulated and measured breadown and threshold voltages for the simple CMOS transistors (all values are in volts).

For all of the above devices there was no variation of breakdown for the four different channel lengths tried i.e. 5μm, 10μm, 15μm and 20μm. This shows that at the minimum distance used, punchthrough is not the limiting breakdown mode. Simulation confirmed that the breakdown is into the channel and then the substrate. This was not the case for the double-diffused and the lateral-diffused transistors. The p-channel LDMOS transistors were non-functional. The reason for this can be seen from figure 5.12. The n-region quickly punches through and the device acts as a gate-modulated resistor. The NL LDMOS devices with the N+ drift region are also poor transistors as the drift region was not driven in deep

enough. Therefore there is a current path from drain to source and the device acts as a resistor.

The DDMOS devices (WN) were found to exhibit similar breakdown voltages to the standard nmos (NN) transistors when the drain was highly and medium doped. This was the case for channel lengths of 10μm and above. The 5μm channel length transistors were shorted. This was due to the merging of the source and drain diffusions under the gate as is shown in figure 5.16. For the lightly-doped drains the breakdown voltage was around 40 Volts, about 20% higher than the NN case (see table 5.5). Additionally, this type of structure offers no advantage over the LDMOS as only one lightly-doped region is needed at the drain of the power transistor. For these reasons the DDMOS was rejected as a candidate for smart power solutions.

For the lateral double-diffused devices,labelled LN and LNB, there was an increase in breakdown voltage for both the medium and lightly doped case. The latter produced a BV_{DSS} of around 47 volts. The LN devices were functional for source/drain separations of 10μm. This increased to 15μm for the LNB devices due to the extra 5μm drift region that encroaches under the gate. This demonstrated that the drift region only helps increase breakdown voltage when it is separated from the influence of the gate. The drift region under the gate only reduces the electrical channel length of the transistor.

The fact that the breakdown still depends on the doping density of drain could be resolved simply by making the n- drift region larger. However with the simple transistor structure used here this in not possible without encroaching under

the gate because the source/drain mask is self-aligned to the polysilicon gate mask. It was seen above that this does not help increase the safe operating voltage. It is necessary to re-design the LDMOS transistor in order to separate the drift region from the channel region. Only by doing this will there be the twin benefits of high breakdown voltage and low drain contact resistance. The next chapter will detail how these modifications can be made.

It was seen that for the p-channel LDMOS devices, LP and LPB that the breakdown voltage is similar to that of the simple pmos with the lightly-doped drain, ie -43V. There is however the advantage of the reduction in drain contact resistance in using the lateral double-diffused structure. As a conclusion, it has been shown that by using simple nMOS and pMOS transistors for logic and n-LDMOS and p-LDMOS for the power transistors that a 40V full CMOS smart power process can be realised without any increase in process complexity from the CMOS logic process.

It can be seen that for both the DDMOS and LDMOS transistors that simulation using SUPREM III and CANDE gives a reasonable fit for both breakdown voltage and threshold voltage (see tables 5.6 and 5.7). It has been seen in this chapter that simulators can be used to accurately predict threshold and breakdown voltages of a variety of transistor structures. If simulation is to be used as a predictive tool for future modifications, we can be reasonably confident that the results are indicative. This is important in that there is at least two orders of magnitude of difference in turn round time of development by computer compared with development in silicon. Therefore many more variations can be tried out at the terminal than could be fabricated in the clean room. This should increase the

probability of realising high voltage LDMOS transistors that are compatible with the original low voltage CMOS process.

Wafer #	**#1**	**#2**	**#3**	**#4**	**#5**	**#6**
Measured BV	18	25	41	19	25	40
Measured Vt	1.6	1.6	1.6	1.6	1.6	1.6
Simulated BV	18	25	41	18	24	41
Simulated Vt	1.5	1.5	1.5	1.5	1.5	1.5

Table 5.6 Simulated and measured breadown and threshold voltages for the DDMOS transistors (all values are in volts).

Wafer #	**#1**	**#2**	**#3**	**#4**	**#5**	**#6**
Measured BV	19	31	47	19	30	42
Measured Vt	1.6	1.7	1.7	1.6	1.7	1.7
Simulated BV	22	24	50	18	22	41
Simulated Vt	1.5	1.6	1.6	1.5	1.6	1.6

Table 5.7 Simulated and measured breadown and threshold voltages for the LDMOS transistors (all values are in volts).

It was found that all the vertical devices were non functional. This can be explained by considering figure 5.15. The n+ source region of the VN device

overdopes the p-channel and shorts to the n-well body of the device. This results in the structure behaving as a resistor, not a transfer resistor. From this we can conclude that it is difficult to realise a vertical power transistor using the same mask set as a CMOS logic process. It could be possible to implement a vertical transistor by using additional masks and process steps, but as the results for lateral power transistors are more promising, it was decided to follow that route.

5.6 Conclusions

In order to develop smart power devices an understanding of the limitations of both logic and high voltage parts is needed. The first section of this chapter investigated the limitations of a well proven nMOS process. In this section an attempt was made to isolate and comprehend dominant mechanisms by using process, device and structure variants. This work was also used to compare results obtained with simulators and measured values. The fact that the simulators chosen gave a reasonable fit meant that in future work they could be used in a predictive mode to help develop both new transistor structures, and to evaluate the effect of process modifications.

It was found that the main limitation on breakdown voltage for the nMOS transistors was the threshold adjust implant. In transistors with this implant source/drain doping did not effect breakdown voltage. In contrast, the transistors without the implant showed an increase in breakdown with decreasing source/drain doping. Also, the threshold voltage of these devices was found to be greater than 1 Volt, meaning that it was strictly necessary to use this implant for enhancement mode transistors. The omission of the threshold voltage implant and the reduction of source/drain implant dose was shown to yield transistors with 40 volt performance.

This was not to be expected. The nMOS process has been used for several years without being modified. The implant shifts the threshold voltage by 0.3 Volts as predicted by simulations. However the absolute value is 0.2 Volts higher experimentally than the simulated value, using SUPREM II and ICECREM. If however, SUPREM III is used in conjunction with CANDE then the values obtained are correct. This confirms that this combination of simulators is the best choice for modelling the process. The simulations of the effect of the implant on breakdown voltages give reasonable results (within 5%) and this means that a 40V N-channel HVIC can be realised within the 6μm process by simply omitting one implant. No further process or device modifications are necessary.

Improvements were found using the double-diffused DDMOS device, but at the expense of increased device area. It was found that the gate-oxide breakdown of 80 volts would not be a limitation at these voltage levels. By analyzing the pn diode structure matrix it was found that the breakdown of a transistor is typically 20-30% less than that of a simple junction. This is due to

several phenomena. The first can be explained by comparing the different diode structures used. It can be seen that thin conductors and sharp angles reduce the peak breakdown voltage, as discussed in chapter 2. The second factor can be the effect of punchthrough between the source and drain junctions. The third factor is the influence of the gate electrode. In order to further increase voltage the gate must be separated from the drain region, the junction supporting the high voltage must be lightly or doubly-doped, and thin conductors and sharp angles must be avoided. In addition, it was found that the n-well of a CMOS process used as diode could withstand 200 Volts. It is necessary however, to find a way of incorporating that type of performance in a transistor structure.

The second half of the chapter investigated and extended a standard CMOS process. For conventional pMOS transistors formed in an n-well it was found that reducing the well depth from 5μm to 2μm had no effect on BV_{DSS} if the surface doping concentration was kept constant. This reduction in well depth could improve the packing density of the logic circuitry. The well depth was 5μm for historical reasons, not for physical ones. By using process simulation of CMOS invertors, it is predicted that if the field oxide isolation is greater than 3μm then there are no detrimental effects on latch-up performance. This distance is less than the given design rules for the process, therefore no problems should be encountered. If these two results are used in a more aggressive design rule set, then a 10% shrink in circuit die size could be realised. This type of shrink in a production environment can mean the difference between profit and loss.

The breakdown of the nMOS and pMOS transistors was found to be dependent on source/drain doping, and could be increased to 34 Volts and 42 Volts

respectively. The down side of this is the increase in device on-resistance and decrease in operating speeds. The result obtained is similar to that of the 6μm process when the threshold adjust implant is removed. This suggests that the breakdown mechanism is into the substrate rather than reach-through. Further, it was seen that the lateral double-diffused MOS transistors were found to have a higher breakdown voltage than conventional nMOS devices. This meant that by small process and device changes the breakdown voltages can be more than doubled : 18 Volts to 47 Volts for nMOS ; 18 Volts to 42 Volts for pMOS transistors.

The increase in breakdown voltage due to the change in source/drain doping contrasts with that of the nMOS process with the threshold adjust implant where no change was seen. The thermal budget and dopant density are comparable in the two processes, therefore there must be a difference in channel doping. This is in fact true, the channel doping in the CMOS n-channel is lower than in the nMOS case. Therefore the depletion region extends further under the channel for a given voltage. Thus in the CMOS process, the punchthrough voltage is more dependent on drain doping than before.

This effect was also seen in the LDMOS transistors where the lightly doped region further reduces drain doping and therefore reduce the depletion length under the channel. For the pMOS transistors there was no further increase in breakdown voltage due to the lateral region. This was probably due to a limiting value being reached determined by the minimum doping level. It was seen from both measurement and simulation that the limitation of these LDMOS devices is that they are self-aligned. The gate overlaps the lightly-doped drift region. To further

increase the breakdown voltage this drift region, which supports high voltage, will have to be separated from the channel, which controls current flow. This requires a different structure.

It can be seen that from comparisons of measured and simulated values of several types of transistor, that the simulators chosen model well the parameters of breakdown and threshold voltage. This means that they can be applied to speed up the development of a new transistor structure which separates the channel and drift regions [5].

References

[1] V.Rumennik, "Power Devices are in the chips", IEEE Spectrum, pp.42-48, July 1985.

[2] C.Contiero, A.Andreini and P.Galbiati, "Multipower BCD - A Versatile technology from 60V to 250V that integrates Bipolar, CMOS and DMOS Devices", IEE Colloquium on Integrated Intelligent Power Devices, March 1989.

[3] Y.Tarui, Y.Hayashi and T.Sekigawa, "Diffusion Self-Aligned Enhance-Depletion MOS-IC", Suppl.J.Jpn.Soc.Appl.Phys., vol.40, pp.193-198, 1971.

[4] K.A.Shenai, P.A.Piaceto, R.Saia and B.J.Baliga, "Blanket LPCVD Tungsten Silicide Technology for Smart Power Applications", IEEE Electron Device Letters, vol. EDL-10, p.270, 1989.

[5] M.R.Duncan, J.M.Robertson, R.J.Holwill and R.Rodrigues, " Design of CMOS-Compatible High Voltage Devices", ICSSICT, 2nd International Conference on Solid-State and Integrated Circuit Technology, October 22nd-28th 1989, Beijing, China.

CHAPTER SIX : DEVELOPING A SMART POWER PROCESS FOR 120 VOLT APPLICATIONS

6.1 Introduction

The strategy adopted here for developing smart power processes has been to increase the voltage capability in stages. In the last chapter the limitations of a low voltage process were first investigated. Then process modifications were made which resulted in a doubling of safe operating voltage. The devices produced were the first step in the evolution of a high voltage smart power process.

The second stage of this evolutionary process was to further modify the device structure and process flow. It was seen that the basic self-aligned MOSFET structure needed to be changed. In order to support high voltages, a lightly-doped drift region has to be inserted in series with the channel. In addition, to realise complementary high-voltage transistors an extra masking layer is needed to form a drift region for the high-voltage p-channel devices.

These process and device changes were made without any degradation of the low voltage CMOS elements. This allows the use of existing designs and cell libraries.

The last stage of the process development was to produce a working circuit that utilised both low-voltage and high-voltage devices. To this end a chopper circuit was designed, fabricated and tested.

6.2 Choice of Device Structure

In order to increase the safe operating voltage of MOS transistors it is necessary to change the device structure. A high resistance region has to be inserted between the source and drain in order to sustain the increase in terminal potentials. There is obviously a trade-off between device on-resistance and breakdown voltage. Therefore careful design has to be made for each application. For smart power circuits both low voltage and high voltage transistors are needed and the numbers of each will influence this decision.

There are two feasible solutions for increasing breakdown voltage. One is to use a lateral MOS transistor with a lightly-doped region at the drain end, i.e. LDMOS. The other is to change from lateral current flow to vertical VDMOS or VMOS. The VMOS, or V-groove MOSFET of figure 6.1 is not suitable due to its non-planar structure. To realise a VMOS transistor it is necessary to add the following process modifications. Firstly, an etch is needed to define the V-groove. Secondly, a second thicker gate oxide is needed for the power transistors and lastly, a thicker polysilicon gate is needed to ensure electrode integrity in the groove. Further, as the drain contact is on the rear of the device, only one power transistor per circuit can be realised. This is not a problem for intelligent switches, but as the aim of this project is smart power the VMOS transistor was discarded as a possible solution. This leaves two possibilities.

(i) In the lateral double-diffused LDMOS of figure 6.2 there are two distinct regions. There is the relatively highly-doped channel region and separated from this there is a lightly-doped region around the drain. This high resistivity area sustains the high voltage. The simplest model of this type of device is a resistor in series with the

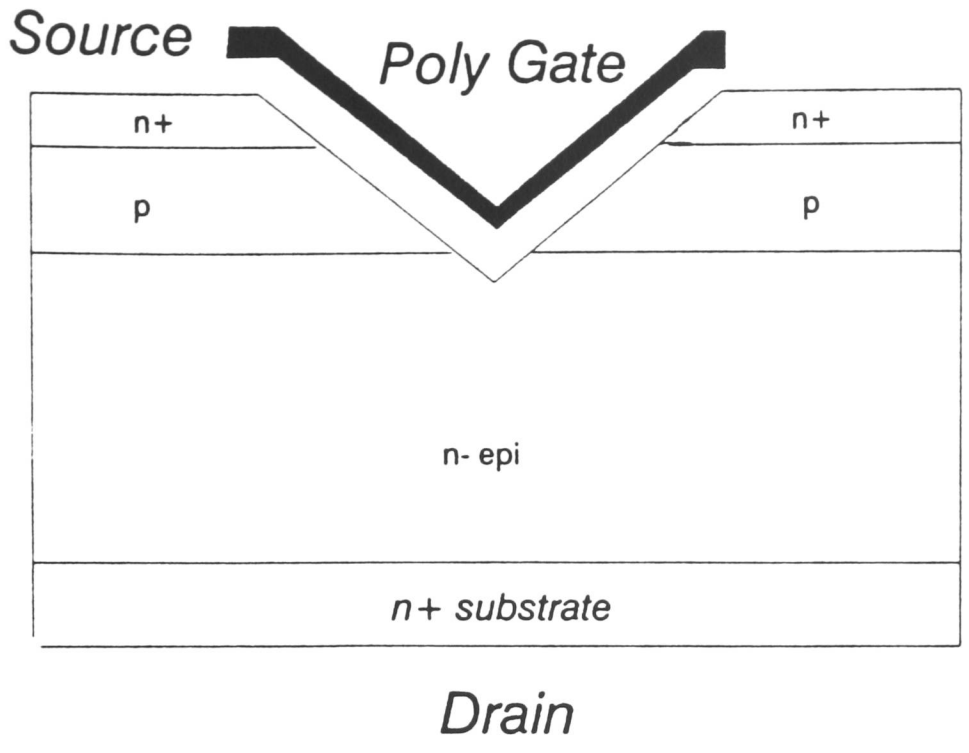


Figure 6.1 A vertical V-groove transistor.

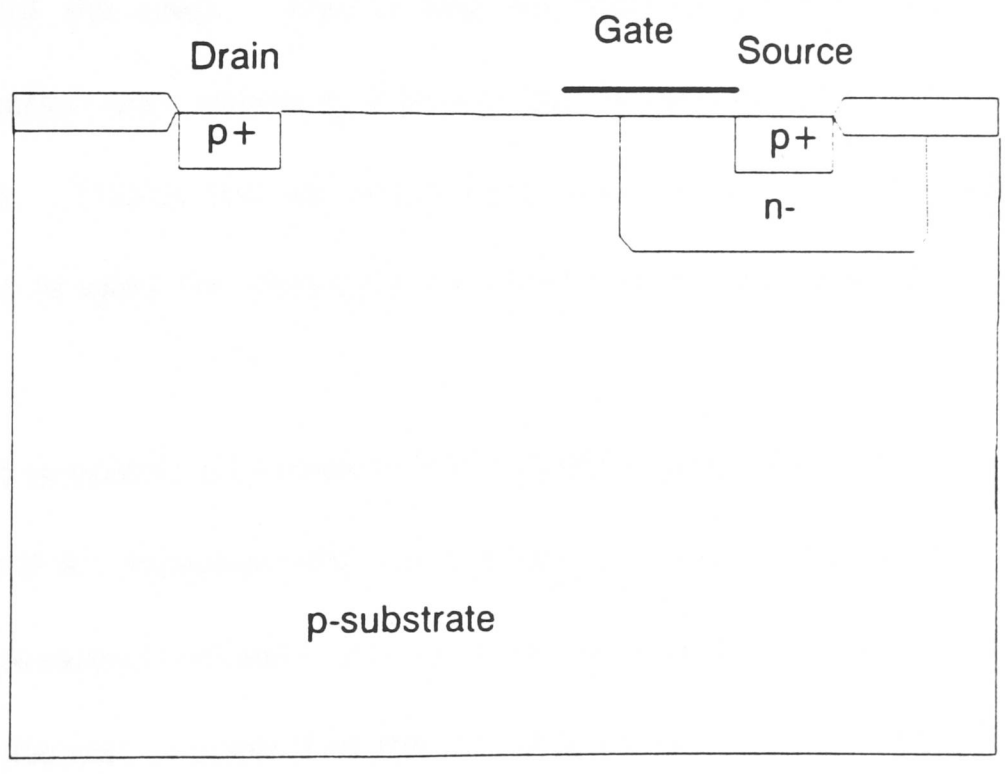


Figure 6.2 A lateral double-diffused LDMOS transistor.

MOS transistor. This type of transistor was discussed in more detail in section 3.2.

(ii) The alternative is the vertical double diffused MOS, VDMOS, of figure 6.3. In this structure the current flow is no longer lateral, but flows vertically from the drain on the back surface of the wafer through the semiconductor bulk to the source on the top surface. Current flows under the influence of the gate electrode which is on the top surface. The voltage is supported by the bulk region, n- as indicated in figure 6.3. A low resistivity region is diffused into the back surface to provide a better contact between the drain and the metallisation layer. The channel region is diffused in first, followed by the counter doped source diffusion. This is the double-diffusion process that gives the structure its name.

The main drawback of this device is that the drain electrode is on the rear surface of the chip. This is fine for discrete devices but in smart power this configuration only allows one power device per die, or a circuit where all drains are common. This is too serious a limitation for true smart power, but this type of structure is used for Vertical Intelligent Power ,VIP, applications.

It is possible to modify the VDMOS structure to have a surface contact as in figure 6.4. However the extra diffusions and contacts add to the complexity of the process and increase the ratio of cell size to on resistance. Further, it was seen in the last chapter that the lateral solution can be integrated easily with either no or little modification to the logic process. For this reason the lateral solution was adopted in this work.

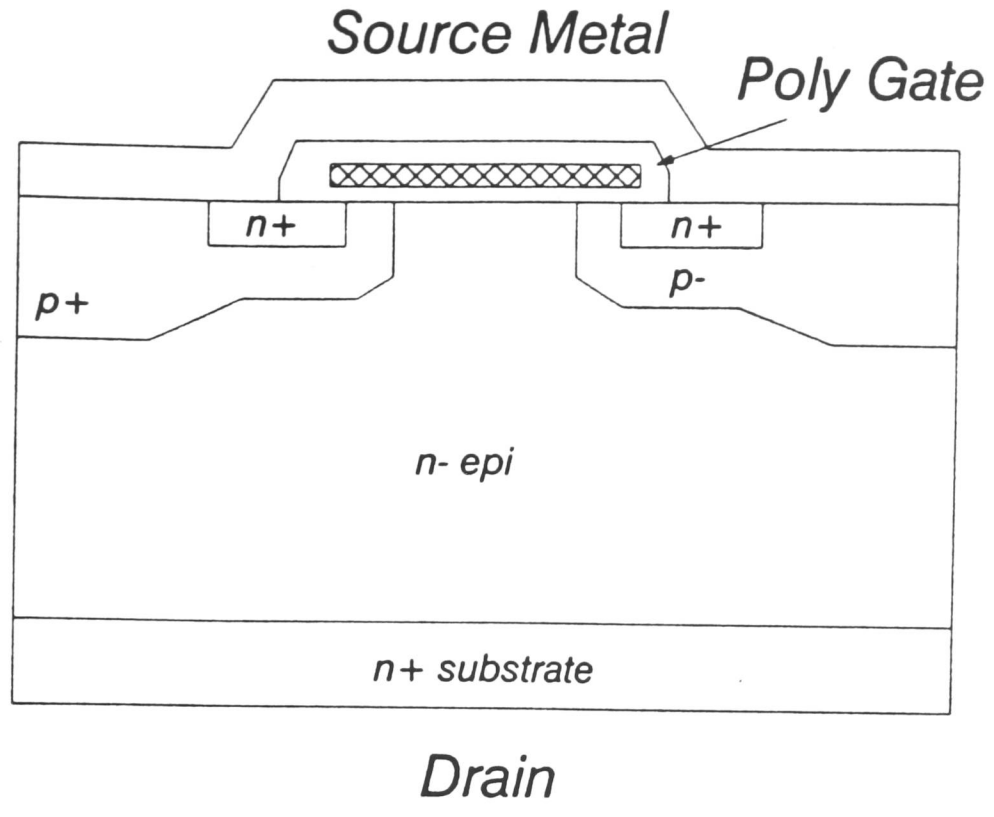


Figure 6.3 A vertical double-diffused VDMOS transistor.

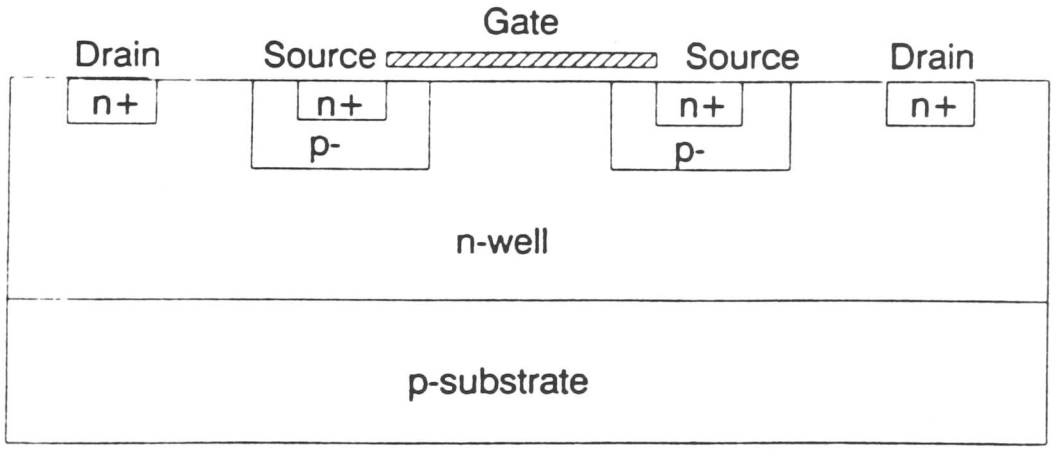


Figure 6.4 A vertical n-channel VDMOS transistor formed with all contacts on the surface.

An evaluation of the on-resistance of both lateral and vertical solutions was made. Usually lateral power devices are used for discrete transistors that have a breakdown voltage of less than 50 V and vertical power transistors are used for voltages above 100 V. In between these ranges, both device types are used. With the modified vertical structure these rules do not apply. Using 5 μm design rules some comparisons were made. The structures of figures 6.2 to 6.4 were simulated using SUPREM3 and CANDE for different target breakdown voltages. By comparing the current density divided by the area as in table 6.1, it can be seen that lateral devices are preferable for breakdown voltages of 120 V, the target of this project. In addition to the better current density, the LDMOS uses a simpler, more economical process. Therefore lateral structures were used in this work.

In order to evaluate the choice between a single LDMOS transistor and a VDMOS transistor using a surface contact, various simulations were made. These simulations were run based on the EMF 5 μm CMOS design rules, and used SUPREM III for doping profiles and CANDE for simulation of device breakdown. Table 6.1 summarises the results obtained. The device lengths given are for a complete LDMOS and a half-device for the VDMOS, as it is a symmetrical structure (as can be seen from figure 6.3). The results obtained from the simulations are for the current that flows per micron of device width. Using these numbers the current density was obtained for various target breakdown voltages.

It can be seen that in all cases from 50V to 200V using the 5 μm CMOS process that the LDMOS is the better choice. Above 300V it was found that vertical devices reign supreme in terms of performance, but as the goal of this project is 120V for automotive and telecommunication applications, the LDMOS

was the natural choice.

BV required (Volts)		50 V	100V	150V	200V
LDMOS	Device length μm	35	40	50	60
LDMOS	Current/Width mA/μm	2	1	0.8	0.1
LDMOS	Current density $\mu A/\mu m^2$	57	25	10	1.7
VDMOS	Device length μm	25	30	35	40
VDMOS	Current/Width mA/μm	0.5	0.1	0.08	0.01
VDMOS	Current density $\mu A/\mu m^2$	20	3.1	2.3	1.7

Table 6.1 A comparison of simulated breakdown voltages for LDMOS and VDMOS transistors.

6.3 Details of the Process Used

In order to minimise the mask count it was decided to base the smart power objective on an existing CMOS process. This has the added benefit of maintaining compatibility with existing logic designs and cell libraries. The process chosen was the EMF 5 μm n-well CMOS one. All existing masking stages of this process were included. A description of this CMOS process was given in chapter 5.

As will be discussed in the next section, the power transistor chosen was the lateral double-diffused MOS or LDMOS. In these transistors the drain is separated from the channel by a drift region. In the previous chapter it was seen from diode structures that the n-well had a much higher breakdown voltage that all the other diffusions used. For this reason, the n-well could be used as the lightly doped drift region for the n-channel LDMOS, as shown in figure 6.5. However for the p-channel transistors an extra masking stage is needed to create a p- lightly doped drift region. This involved a triple doping to form the drain region, i.e. n-well into p-substrate, p- into n-well followed by p+ on top to form the drain contact.

This p-channel LDMOS structure is shown in figure 6.6 for a device using the standard n-well. The triple doped profile obtained is not tolerant to the processing fluctuations that could be expected in a production environment. Table 6.2 shows that the well diode structure would break down at a low voltage due to punchthrough and that this value is heavily dependent on small changes in junction depth. Therefore it was decided to alter the n-well doping profile in order to realise a more stable power p-channel transistor.

If possible, it is preferable to keep the threshold voltages of the low voltage pMOS transistors constant in order to continue using existing cell libraries. To realise this the doping density at the surface of the device has to be the same for both the conventional and the modified well. In order to support the triple diffusion the well must be driven in to around 8 μm depth, as is shown in table 6.2. These results were obtained using CANDE simulations. A SUPREM3 simulation of the new well profile is shown in figure 6.7. The simulated value for V_t using SUPREM3 for both wells is shown in table 6.3. Here, it seems that there is no

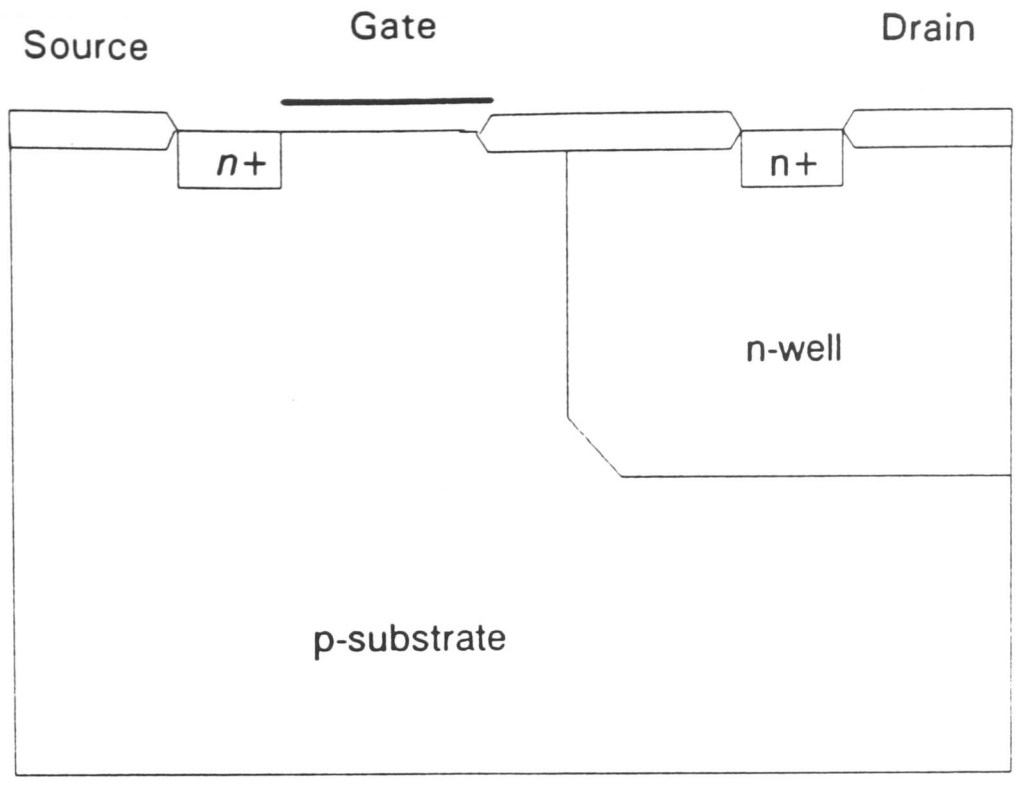


Figure 6.5 The n-channel LDMOS transistor, using the n-well as the lightly-doped drift region.

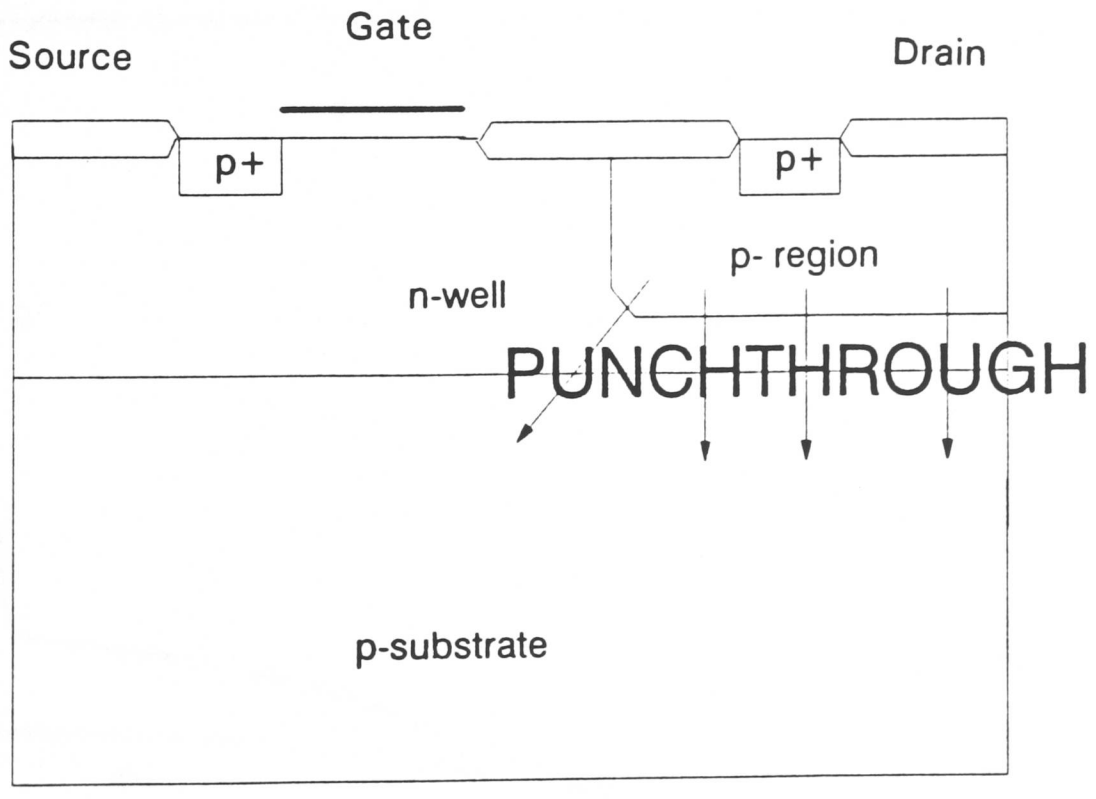


Figure 6.6 The n-channel LDMOS transistor, showing the hazard of vertical punchthrough.

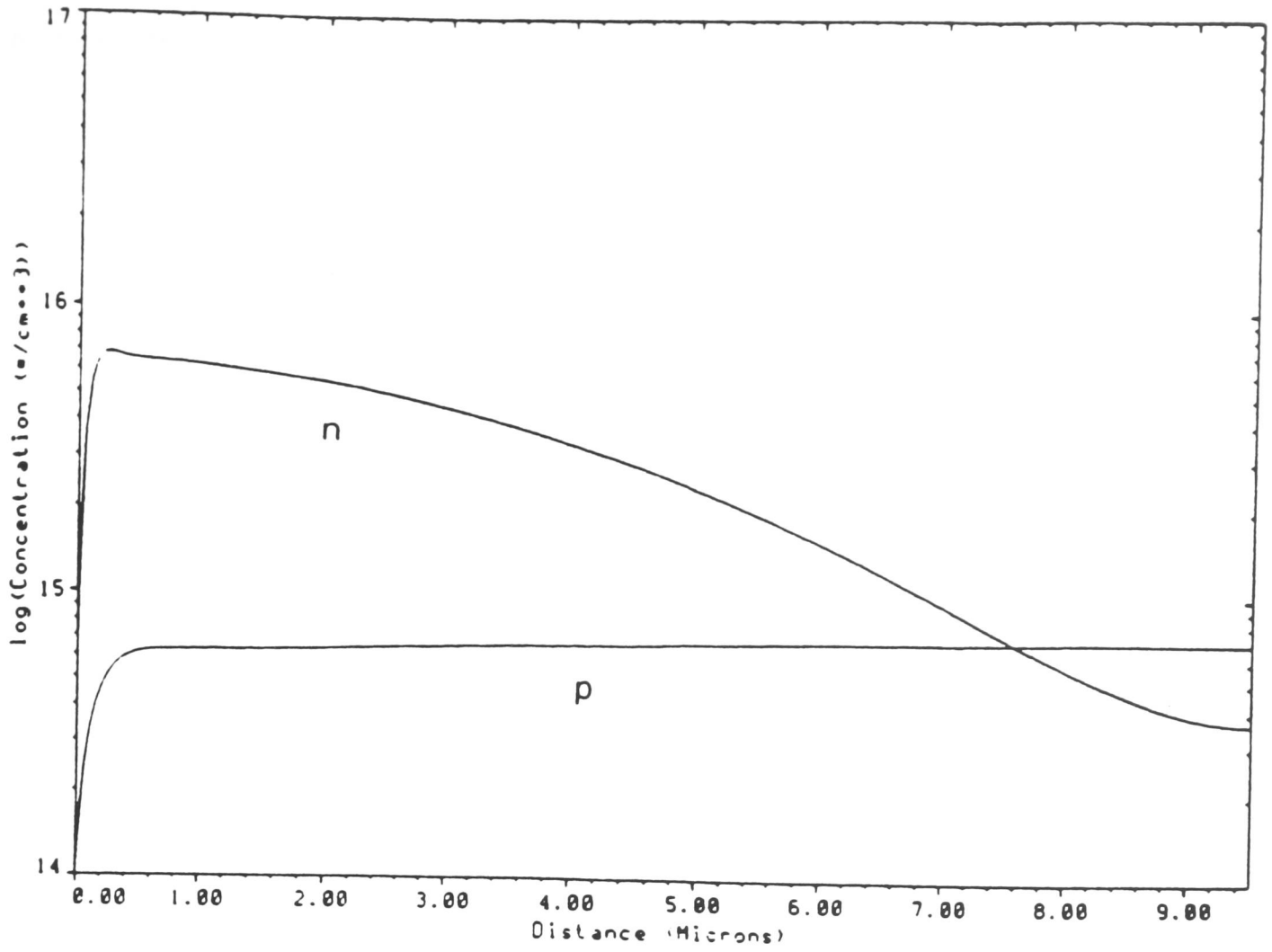


Figure 6.7 A SUPREM III simulation of the p and n-type doping densities using the modified n-well. The junction is seen at 7.6μm.

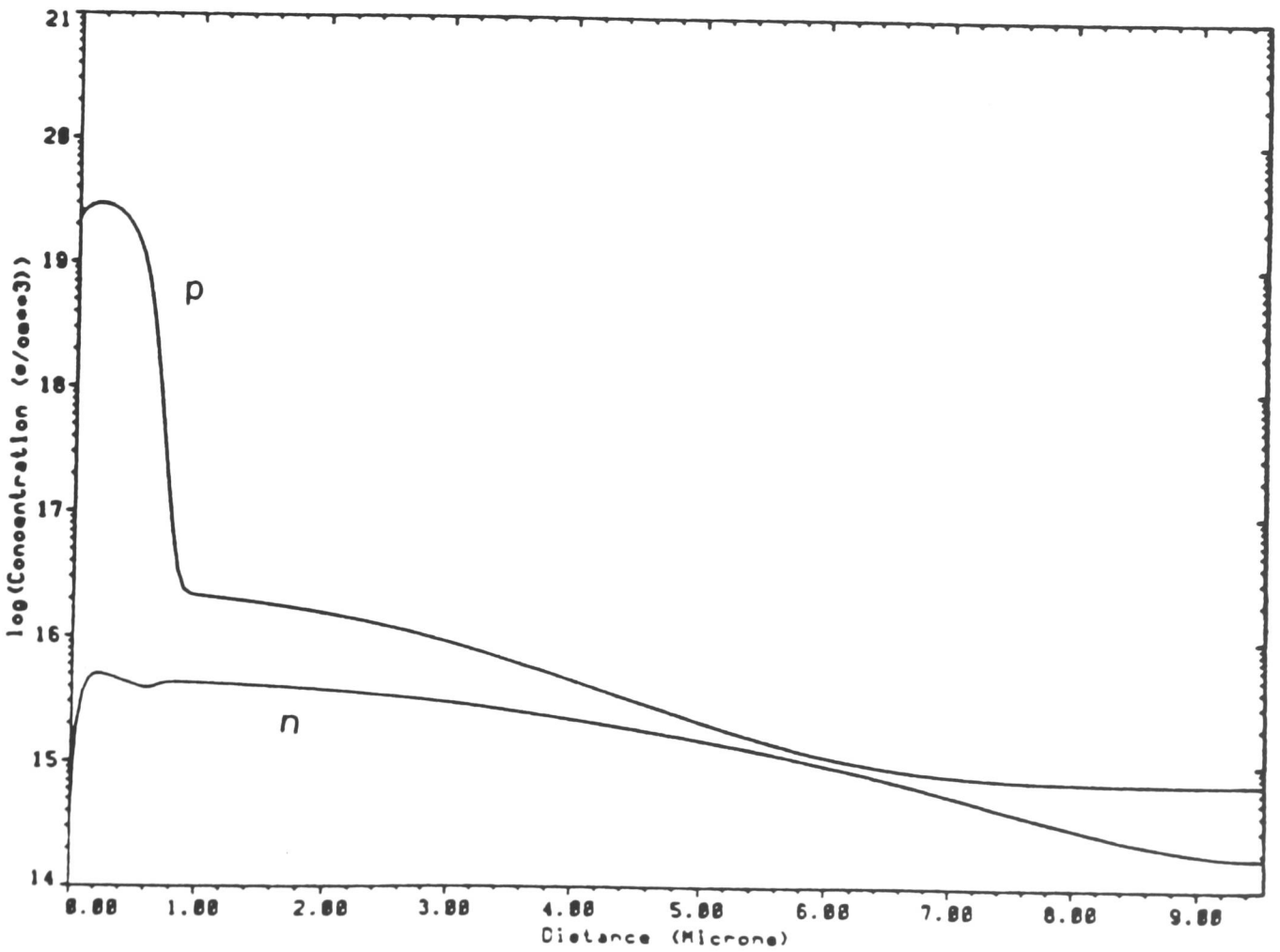


Figure 6.8 A SUPREM III simulation of the p and n-type doping densities in a cross-section of the p-LDMOS drain.

significant change in V_t due to the modified well for the low voltage transistors.

Similarly the simulated drain breakdown voltages of the low voltage pMOS transistors are shown in table 6.3. No change in BV_{DSS} was seen with various well depths. In order to define the high voltage p-channel transistor drift region an extra masking step has to be introduced into the process. This p- region mask is added after n-well but before the definition of the active area. The optimum depth of this drift region is dependent on device characteristics and was found to be around 4 μm depth. Figure 6.8 shows a cross section through an n-well that has the p- drift region. The junctions can clearly be seen at 4 μm and 8 μm .

N-well x_j	P-x_j 3μm	P-x_j 4μm	P-x_j 5μm	P-x_j 6μm	N-well BV
5μm	-20V	-	-	-	175V
6μm	-26V	-18V	-	-	180V
7μm	-60V	-40V	-19V	-	188V
8μm	-98V	-102V	-80V	-45V	207V
9μm	-98V	-102V	-96V	-72V	220V

Table 6.2 Simulated breakdown voltages for the p-channel LDMOS with different n-well and p-region junction depths.

Parameter	N-well 5μm	N-well 8μm
nmos BV (V)	18	18
nmos Vt (V)	1.4	1.4
pmos BV (V)	-20	-20
pmos Vt (V)	-1.4	-1.4

Table 6.3 Simulation results for low voltage nmos and pmos transistors with different n-well depths.

Since neither of these process changes effect the nMOS low voltage transistors and there is no measured difference in pMOS low-voltage characteristics, the existing logic designs and cell libraries can be used with this smart power process. The extra well depth will encroach further laterally, but as the design rules for the n-well are pessimistic, no problems would result if the old designs respected the rules given.

6.4 Simulation and Results for the Power Transistors

To achieve high-voltage operation and to maintain compatibility with low voltage devices, LDMOS structures were used as the power elements. The LDMOS transistor supports higher voltages by having a lightly-doped drift region in series with the channel. For the n-channel power transistors this drift region is the n-well as shown in figure 6.9.

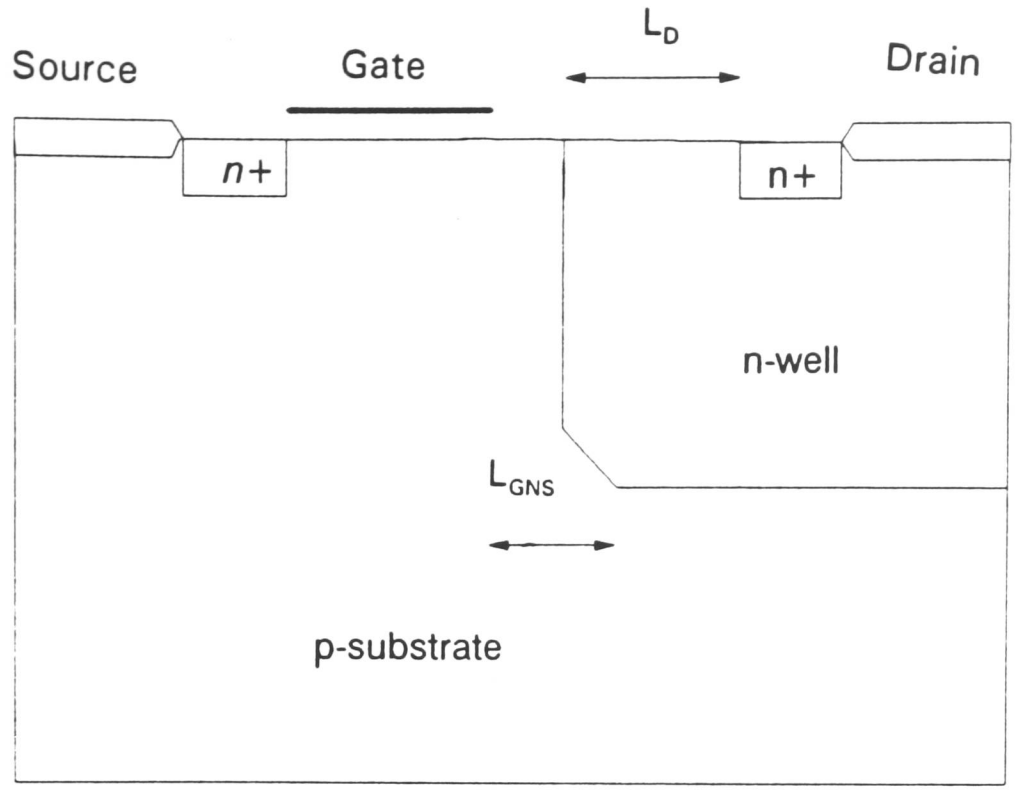


Figure 6.9 The n-channel LDMOS without the field oxide inserted in the channel.

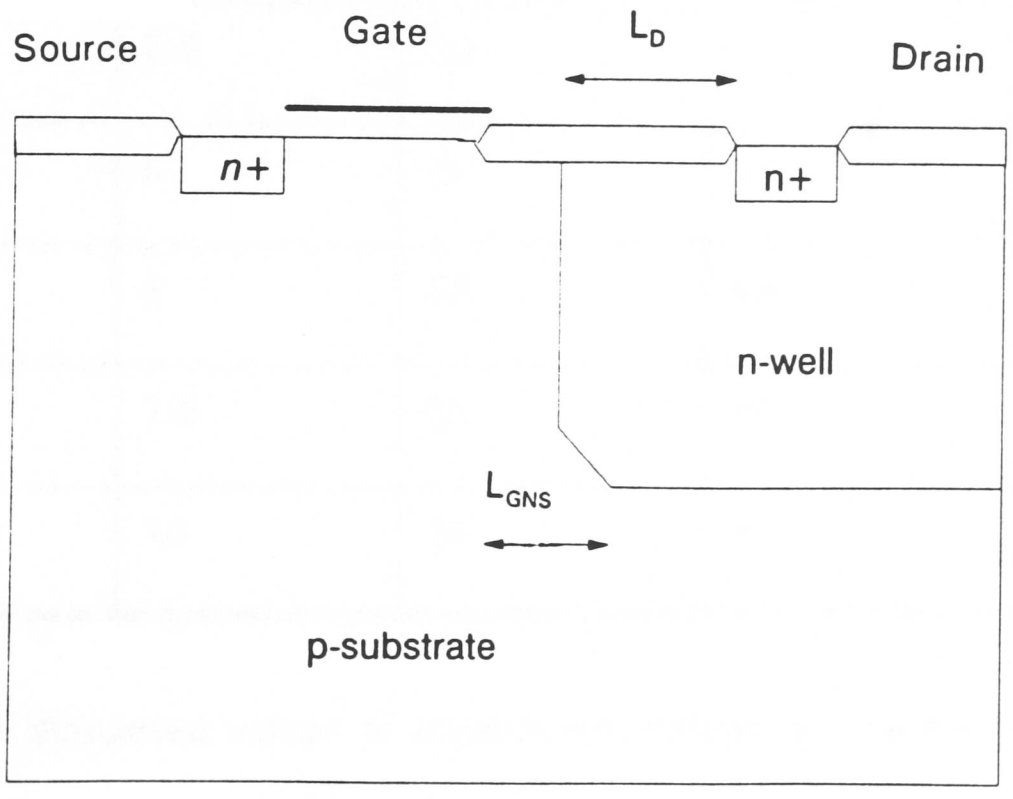


Figure 6.10 The n-channel LDMOS with the field oxide inserted in the channel.

In the test transistor matrix, various values of drift region length, L_D , were used. Increasing L_D will increase BV_{DSS} up to a certain point, after which the breakdown voltage saturates and the only effect will be to increase the on-resistance of the device. Optimum values of L_D were also calculated using SUPREM3 and CANDE simulations, and are shown in table 6.4. The separation between the end of the gate and the start of the n-well was also varied. In practice, the optimum separation is 0 in the silicon, but because of side-diffusion this will correspond to 0.8 x_j (n-well) as drawn on the mask. All values of drift region length and separation will be referred to by their drawn values.

Drift length L_D (μm)	**Gate-Space L_{GNS} (μm)**	**BV_{DSS} no field + no Vt implant (Volts)**	**BV_{DSS} no field oxide (Volts)**	**BV_{DSS} with field oxide (Volts)**
5	2.5	25	29	29
7.5	5	30	38	108
10	5	33	44	142
12.5	7.5	35	50	142
15	10	35	50	142

Table 6.4 Simulated values of breakdown voltage for the n-LDMOS transistors without the and without the field oxidation inserted in the device.

The third parameter varied was the inclusion and exclusion of LOCOS field isolation between the source and drain. Compare figure 6.9 without the field oxide and figure 6.10 with the oxide isolation. Simulations showed this oxide to be necessary to maximise breakdown voltage, but this needed to be verified. Without this isolation the transistor will be punchthrough-limited. The motive for this is that there is a blanket threshold adjust implant performed on the active area. This increases the surface doping concentration and therefore reduces the maximum safe voltage at the device surface. However, if a field oxide region is added then this isolation and the lack of active area threshold adjust implant mean that the effect of the punchthrough-limited breakdown at the surface can be reduced. Simulations showed that the oxide was the main reason for the increase in performance. This is due to the fact that when grown, the oxide consumes silicon and therefore physically separates the channel and drift region even further.

Another process variation tried was the exclusion of the channel stop implant. This implant is normally used to prevent current leakage between adjacent transistors. The value used in this process was found not to increase the immunity to latch-up of low voltage CMOS, and did not effect the breakdown voltage of the transistors. Due to the cost of an implant step this process could safely be removed from the flow giving a reduction in cycle time and fabrication cost.

For the p-channel power transistors, the whole device is formed inside the n-well. The drift region is obtained using an additional p-region. This triple-diffusion process was defined with the aid of simulations, as described in section 6.3. As demonstrated previously in table 6.2, the n-well has to be deeper

than normal to allow this. To confirm this, wafers were processed having both the conventional n-well and the deeper well. According to simulations, the wafers with the conventional well would have only n-channel power transistors functional, whilst the wafers with the deeper well should have dual channel power transistors. The p-channel structure is shown in figure 6.11. As for the n-channel device the distances L_D and L_{GNS} were varied to determined optimum values. Similarly, devices were fabricated with and without the LOCOS field isolation region, and the channel stop implant. Results of simulations of various p-LDMOS transistors are shown in table 6.5. From these simulations, we can expect to realise 140V n-channel and -60V p-channel power transistors using a 5μm CMOS process with a deep n-well and an additional p- mask level.

Drift length L_D (μm)	**Gate-Space L_{GNS} (μm)**	**BV_{DSS} without field oxide (Volts)**	**BV_{DSS} with field oxide (Volts)**
5	2.5	-24	-20
7.5	5	-31	-44
10	5	-31	-58
12.5	7.5	-31	-58
15	10	-31	-58

Table 6.5 Simulated values of breakdown voltage for the p-LDMOS transistors with and without the field oxide inserted in the transistor.

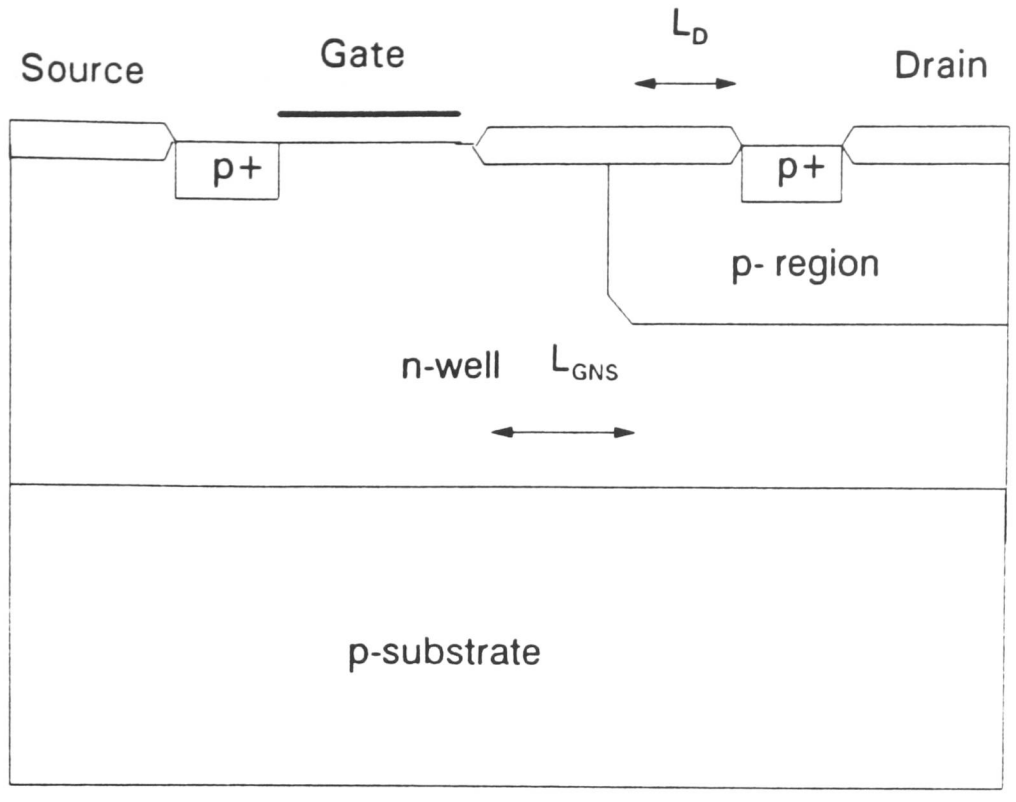


Figure 6.11 The p-channel LDMOS with the field oxide inserted in the channel.

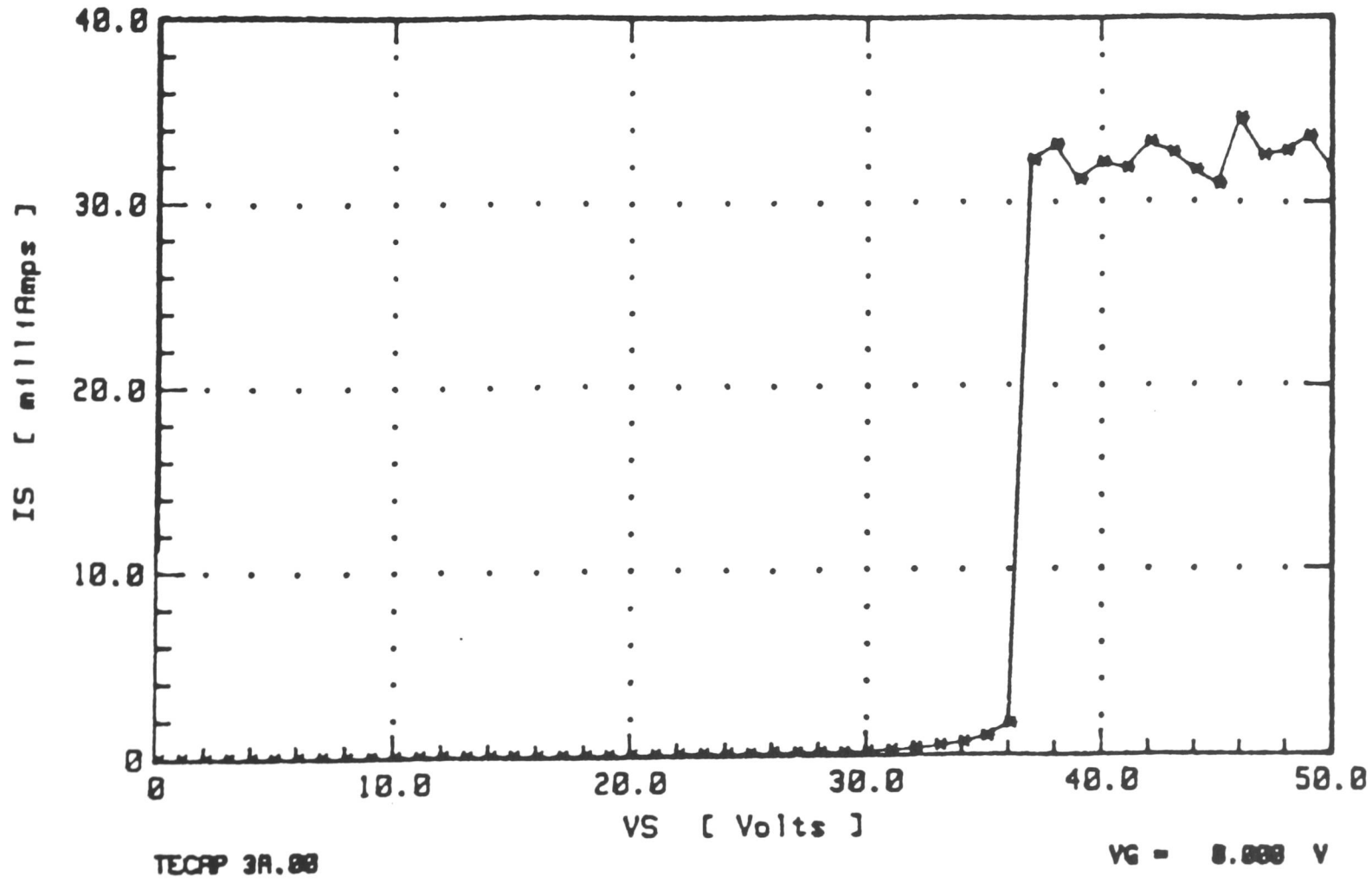


Figure 6.12 Drain breakdown characteristics for the n-LDMOS without the field oxide.

The results of the n-channel transistors without the LOCOS isolation in terms of measured breakdown and threshold voltages are given in table 6.6. The values obtained are slightly higher than the results of the low voltage transistors, but are certainly not high enough to be useful as high-voltage transistors. These low breakdown voltages are caused by two mechanisms : the lack of the isolation oxide between drift region and channel, and the active area threshold adjust implant. This implant into the drift region means that the device will breakdown earlier as shown in figure 6.12. This confirms the results obtained from simulation, previously shown in table 6.4.

Drift length L_D (μm)	**Gate-Space** L_{GNS} (μm)	**BV_{DSS}** (Volts)	**Threshold** (Volts)
5	2.5	25	1.4
7.5	5	30	1.6
10	5	30	1.6
12.5	7.5	30	1.6
15	10	30	1.6

Table 6.6 Measured values of breakdown and threshold voltage for the n-LDMOS transistors without the field oxidation.

Similarly for the p-channel transistor without the LOCOS isolation the measured breakdown voltages are given in table 6.7. These values are similar to

the low-voltage case and therefore do not offer a significant improvement, so can be discounted as power transistor structures,(see figure 6.13). The simulated 3D doping profiles for these devices are shown in figures 6.14 and 6.15. The threshold implant is clearly seen in the channel near the device surface. It is this implant that causes the low breakdown voltage.

Drift length L_D (μm)	**Gate-Space** L_{GNS} (μm)	**BV_{DSS}** (Volts)	**Threshold** (Volts)
5	2.5	-19	-1.4
7.5	5	-29	-1.6
10	5	-29	-1.6
12.5	7.5	-29	-1.6
15	10	-29	-1.6

Table 6.7 Measured values of breakdown and threshold voltage for the p-LDMOS transistors without the field oxidation.

However, there is a significant improvement when the LOCOS isolation is inserted between drain and channel regions. From tables 6.8 and 6.9, it can be seen that for the n-channel power transistors above the critical values of L_D and L_{GNS} the breakdown voltage saturates and is 120 V. The n-LDMOS device structure is shown in figure 6.16. The values of L_D and L_{GNS} quoted are the drawn mask dimensions, not the actual dimension. When the dopings are such that the n-well does not overlap the drain, and the drift region is 10 μm (on the mask, for the 8

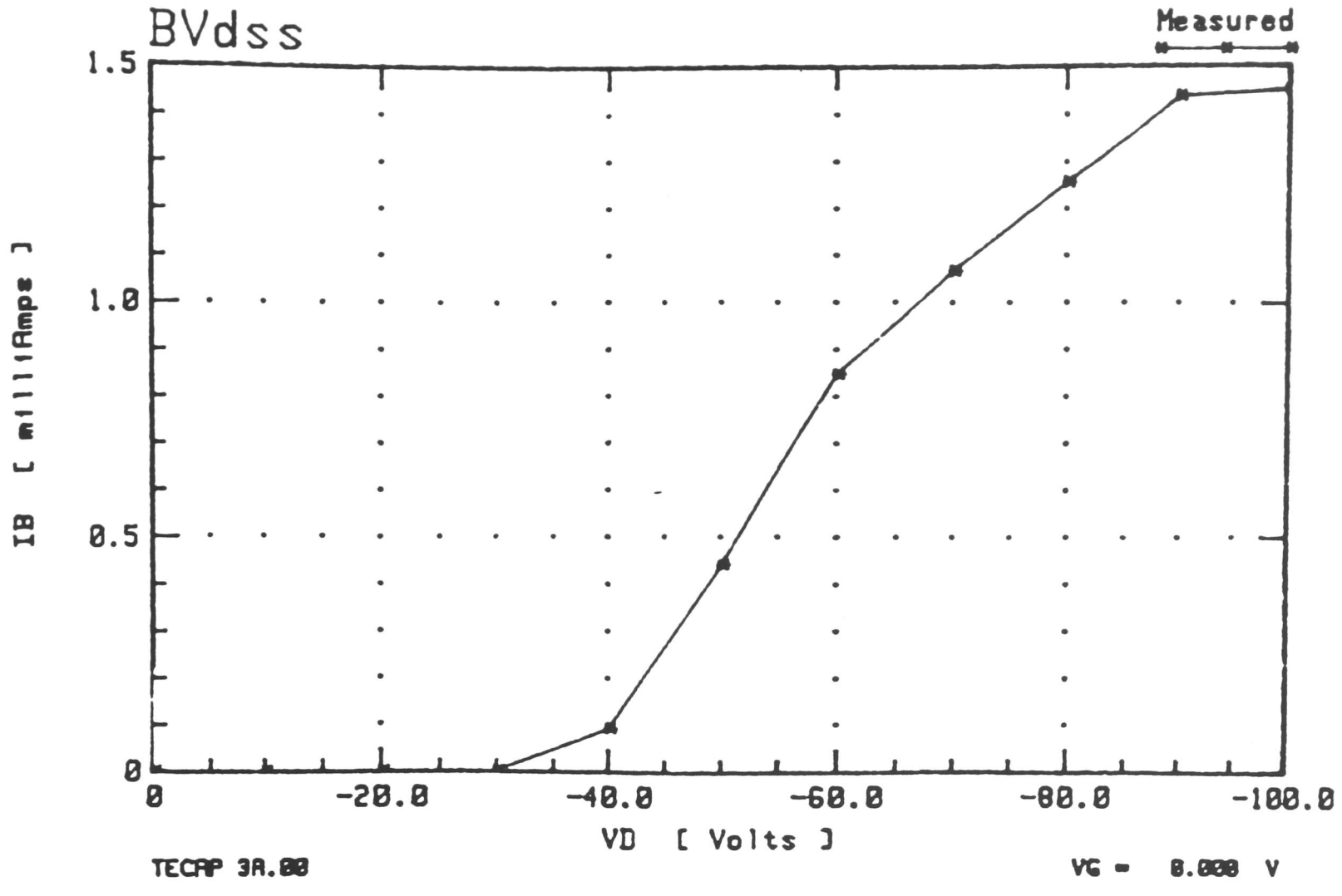


Figure 6.13 Drain breakdown characteristics for the p-LDMOS without the field oxide.

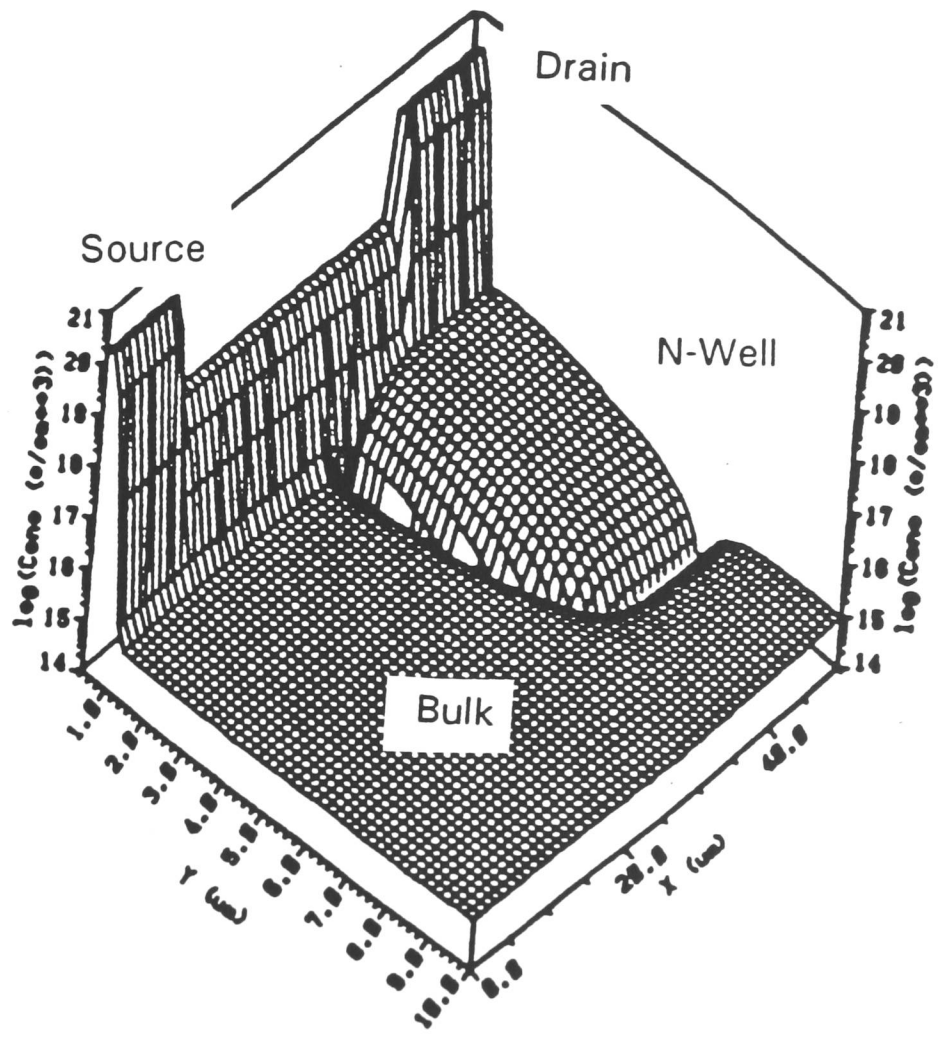


Figure 6.14 A SUPRA simulation showing the 3D net doping profiles of the n-LDMOS without the field oxide.

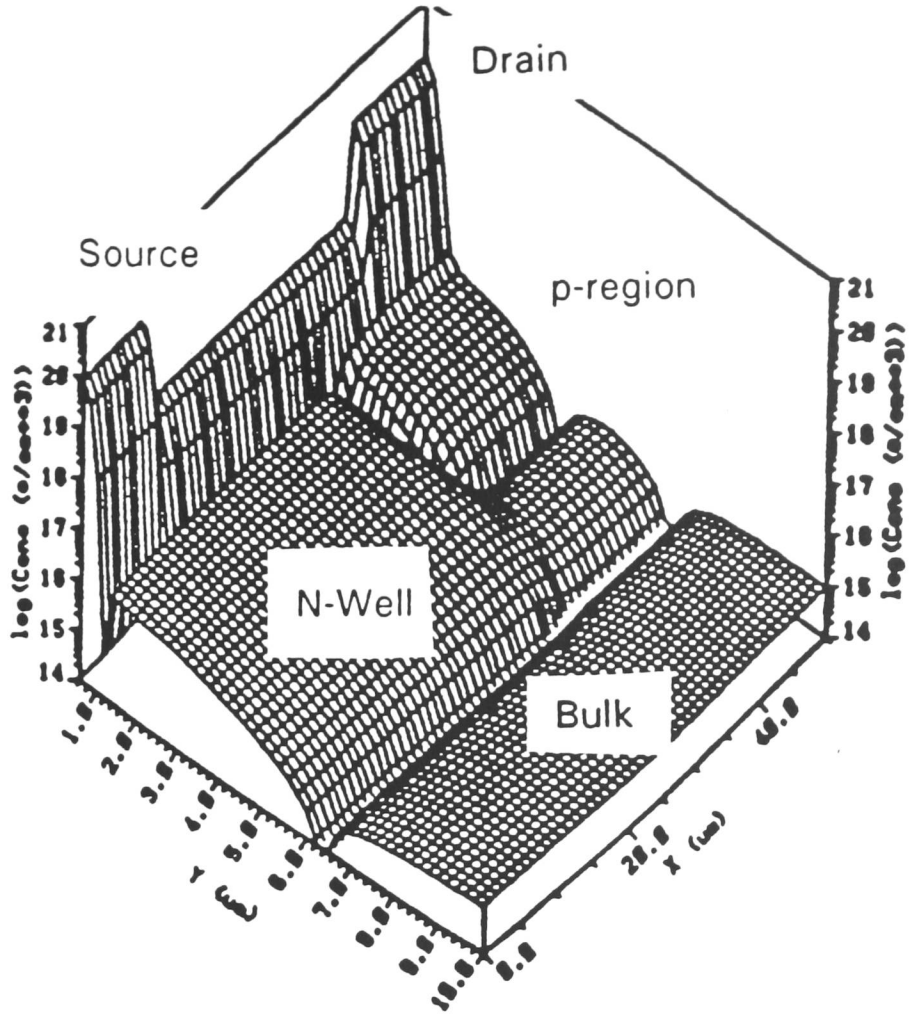


Figure 6.15 A SUPRA simulation showing the 3D net doping profiles of the p-LDMOS without the field oxide.

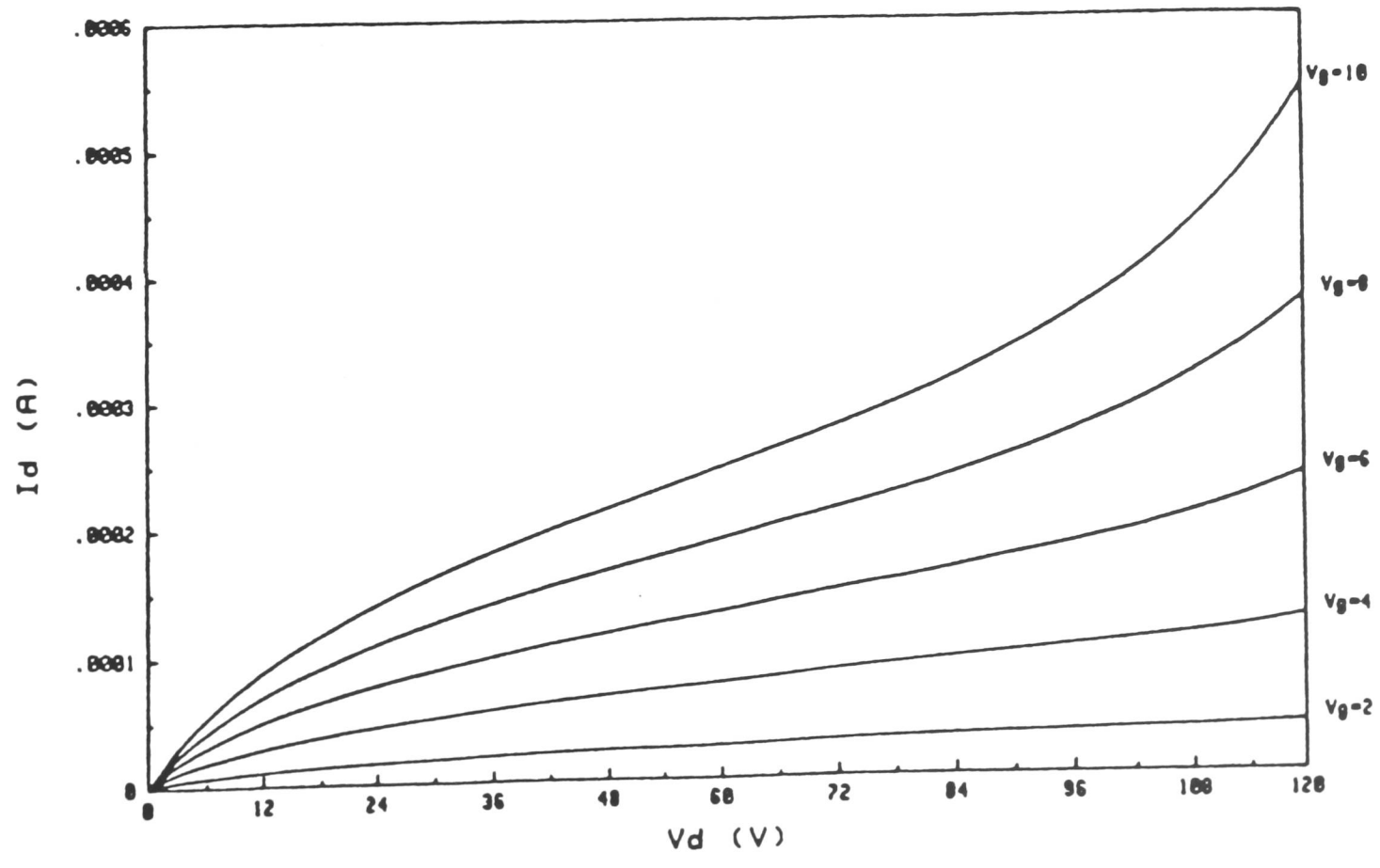


Figure 6.16 Drain current characteristics for the n-LDMOS with the field oxide.

μm well case) then the maximum BV_{DSS} is obtained. Further increases in the drift region length result in an increase of on-resistance only.

Drift length L_D (μm)	Gate-Space L_{GNS} (μm)	BV_{DSS} (Volts)	Threshold (Volts)
5	2.5	-	-
7.5	5	89	1.9
10	5	119	2.0
12.5	7.5	123	2.0
15	10	121	2.0

Table 6.8 Measured values of breakdown and threshold voltage for the n-LDMOS transistors with the field oxidation and 8μm n-well.

Drift length L_D (μm)	Gate-Space L_{GNS} (μm)	BV_{DSS} (Volts)	Threshold (Volts)
5	2.5	20	1.6
7.5	5	120	2.0
10	5	122	2.0
12.5	7.5	122	2.0
15	10	122	2.0

Table 6.9 Measured values of breakdown and threshold voltage for the n-LDMOS transistors with the field oxidation and 5μm n-well.

Similarly for the p-channel transistors shown in figure 6.17, the maximum breakdown voltage occurs when the p-drift region does not overlap the gate. The results are shown in table 6.10. It was predicted from simulations that the saturation of breakdown voltage should occur at a shorter drift region length than the n-channel LDMOS , but the step size used on the test chip was too coarse to detect this. The p-channel transistors formed using the standard 5 μm well had the drain shorted out to the substrate, as predicted by the simulations. Therefore, in order to realise power p-LDMOS transistors it is obligatory to use the deeper n-well.

Drift length L_D (μm)	**Gate-Space** L_{GNS} (μm)	**BV_{DSS}** (Volts)	**Threshold** (Volts)
5	2.5	-21	-1.4
7.5	5	-39	-1.6
10	5	-55	-1.7
12.5	7.5	-55	-1.7
15	10	-55	-1.7

Table 6.10 Measured values of breakdown and threshold voltage for the p-LDMOS transistors with the field oxidation and 8μm n-well.

For the n-channel transistors fabricated using the standard CMOS process with the shallower n-well, the breakdown voltages are similar to those with the

deeper junction. This demonstrates that the extra well depth does not positively affect performance, and should only be used if both n- and p-channel power transistors are needed in the design. The 3D doping profiles for both of these transistors are shown in figures 6.18 and 6.19. In comparison with figures 6.14 and 6.15, the peak doping level in the drift region is much lower.

Figures 6.16 and 6.17 show typical drain current voltage characteristics for these devices. It can be seen from the n-channel devices that the drift region resistance influences the trace giving the characteristic slope. This effect is not so evident in the p-channel transistors due to the drift region's lower resistance value.

It was described in chapter 3 how the LDMOS structure could be modelled simply as an enhancement mode transistor with a resistance in series. This resistance is the effect of the lightly-doped drift region. In the devices shown here the two zones, channel and drift-region, were physically separated by the field oxidation. It can be clearly seen from figure 6.16 that this resistance is present, and is modulated by the gate voltage. Therefore, it is not electrically separated from the channel region. In order to model this effect, a more accurate model such as that of section 3.3.2 is necessary. However, as a first approximation the simple model of a resistor in series with the transistor, is adequate in circuit simulation programs such as SPICE.

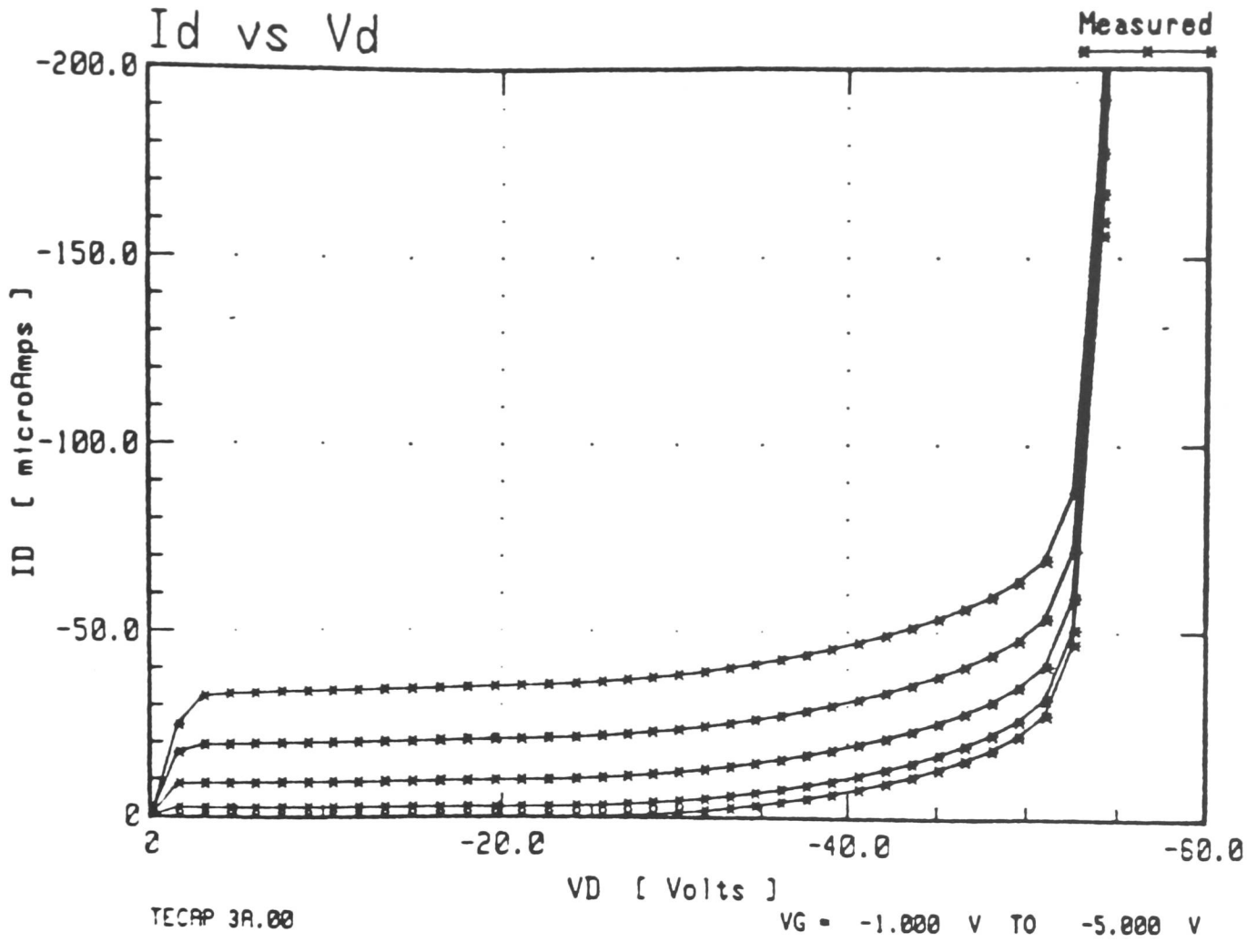


Figure 6.17 Drain current characteristics for the p-LDMOS with the field oxide.

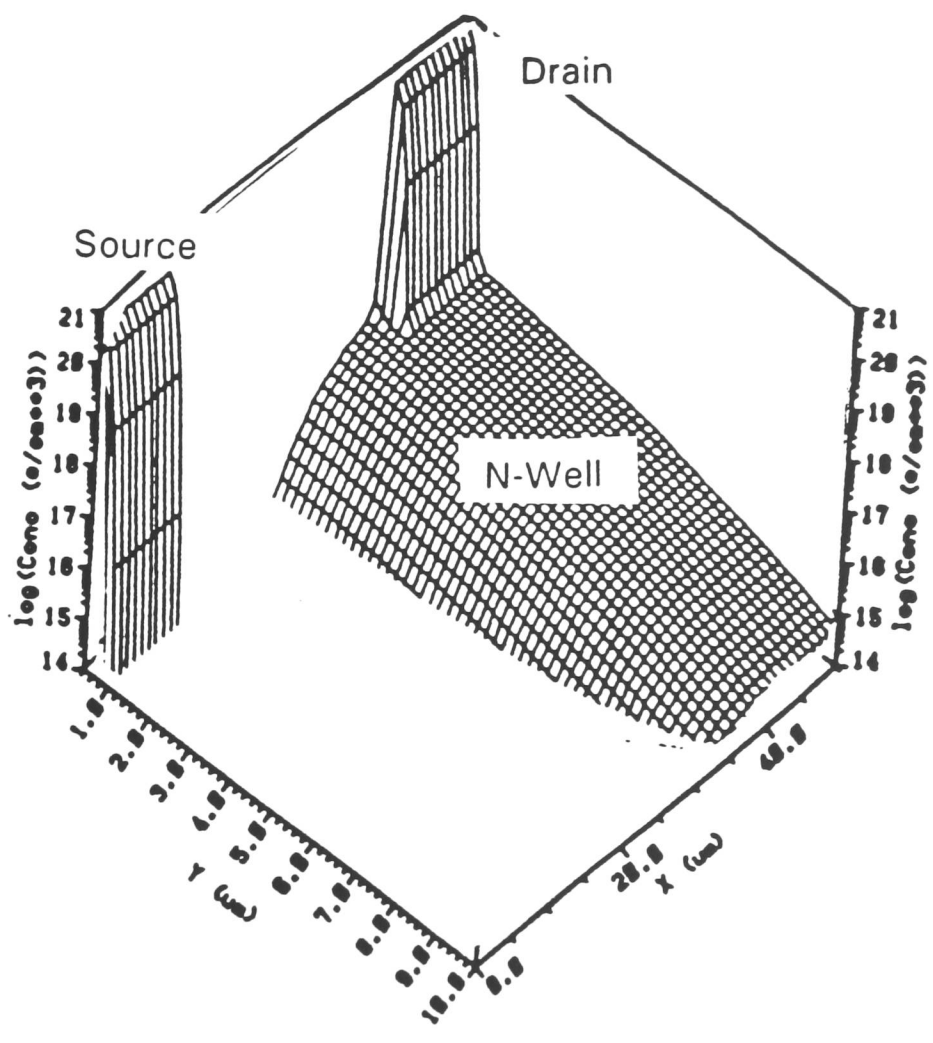


Figure 6.18 A SUPRA simulation showing the 3D n-type doping profiles of the n-LDMOS with the field oxide.

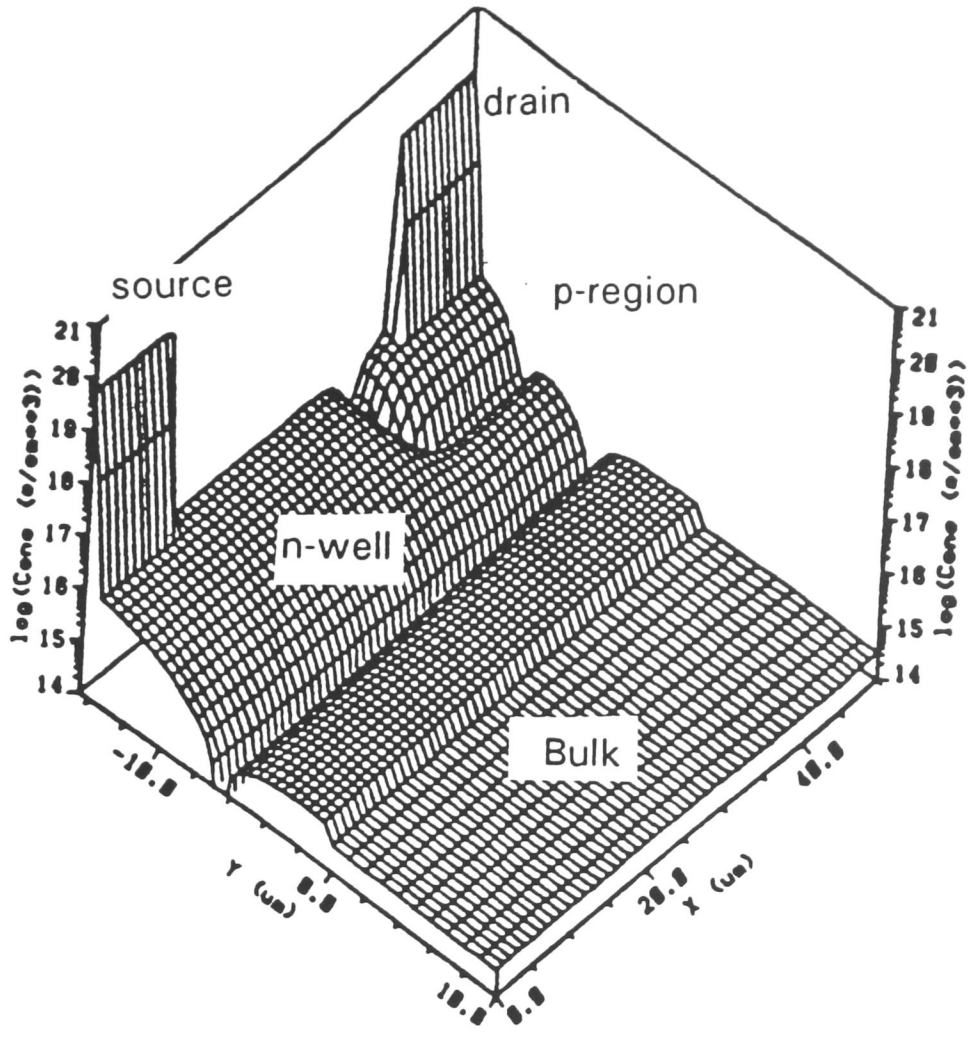


Figure 6.19 A SUPRA simulation showing the 3D net doping profiles of the p-LDMOS with the field oxide.

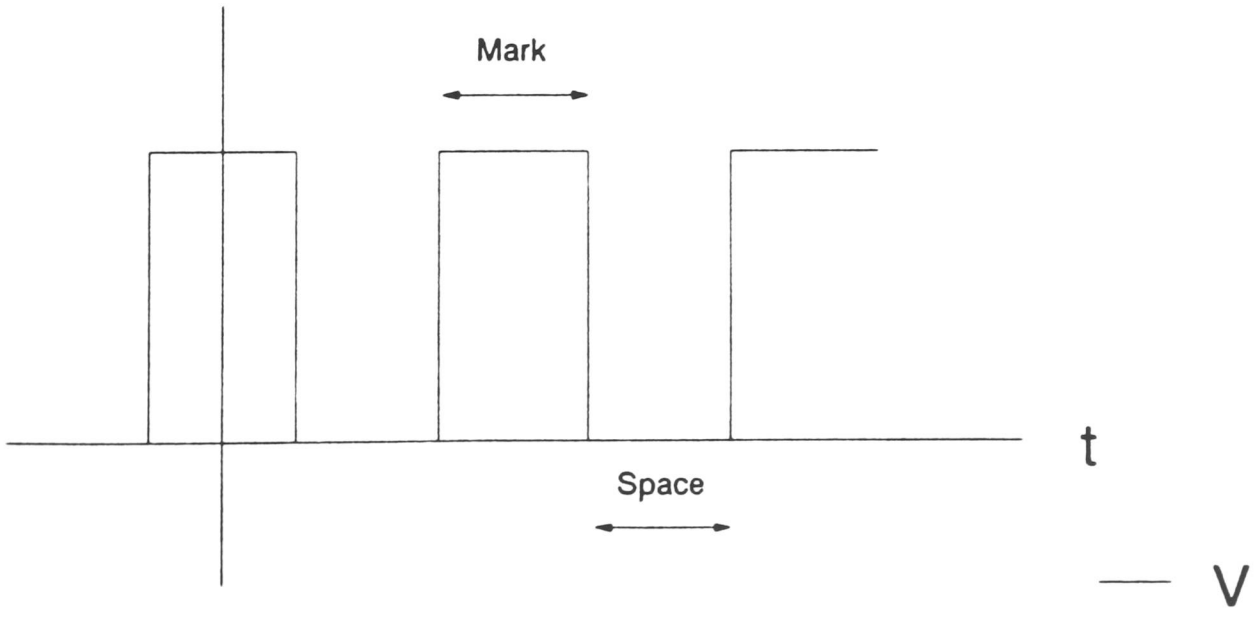


Figure 6.20 The input and output waveforms of a d.c. chopper circuit.

Simulation was used extensively in this work to speed up design time. After fabricating test chips, it is interesting to compare the predicted and measured results. We saw in table 6.1 that the maximum predicted breakdown values for the n- and p- junctions were 207V and -102V respectively. However, for the transistors the values reduce to 142V for the n-LDMOS and -58V for the p-LDMOS transistor. The measured values were somewhat lower than this, 122V and -55V respectively. This was the first major discrepancy noted between simulated and measured values, and signifies that at this voltage the process and device models that were adequate for the low voltage transistors are in need of some modifications.

The breakdown voltage in CANDE is obtained when the ionisation integrals exceed unity. In other terms, when the number of electrons and holes generated by impact ionisation exceeds the number consumed, then avalanche breakdown is occurring. It was found empirically that by substituting unity with 0.8 a good fit to the measured values was obtained, 125V for the n-LDMOS and -54V for the p-LDMOS transistors. It should be noted that simulations predicted the effect of the field oxide inserted in the device structure. Using CANDE, it was possible to see that it was the surface fields that the limited the breakdown of lateral DMOS transistors.

This section demonstrates that high-voltage transistors can be fabricated using a CMOS process without compromising the logic devices. The modifications result in a process with a mask count of nine if only n-channel power transistors are requested, or ten if dual-channel power transistors are necessary.

6.5 Chopper Circuit Design

So far only single test transistors have been fabricated. The next step was to design a circuit using the smart power process. As the circuit was only intended to demonstrate the functionality of the process, a simple design was chosen that would be easily tested. The circuit was a dc chopper.

The function of a chopper circuit is to take a dc voltage as input and to give a lower dc voltage as output. It does this by switching power on and off at high frequency. By varying the mark-space ratio, as shown in figure 6.20, the output changes as follows.

$$V_{out} = V_{in}.t_{on} \ / \ T \qquad 6.1$$

This type of circuit is obviously crude as an electronics power supply but is widely used in motor speed control applications. The function performed is similar to that of a resistor chain, as in figure 6.21, but is a much more efficient way of converting power. This is due to the fact that in the chain circuit all the power dissipated across the top resistor is lost.

If the load is resistive then the output current simply follows the voltage shape as in figure 6.22. However if the load is inductive, the current becomes the integral of the square wave, see figure 6.23 for an ideal inductor. In motors, the current is a combination of the two ideal waveforms, as motors have both inductive and resistive components.

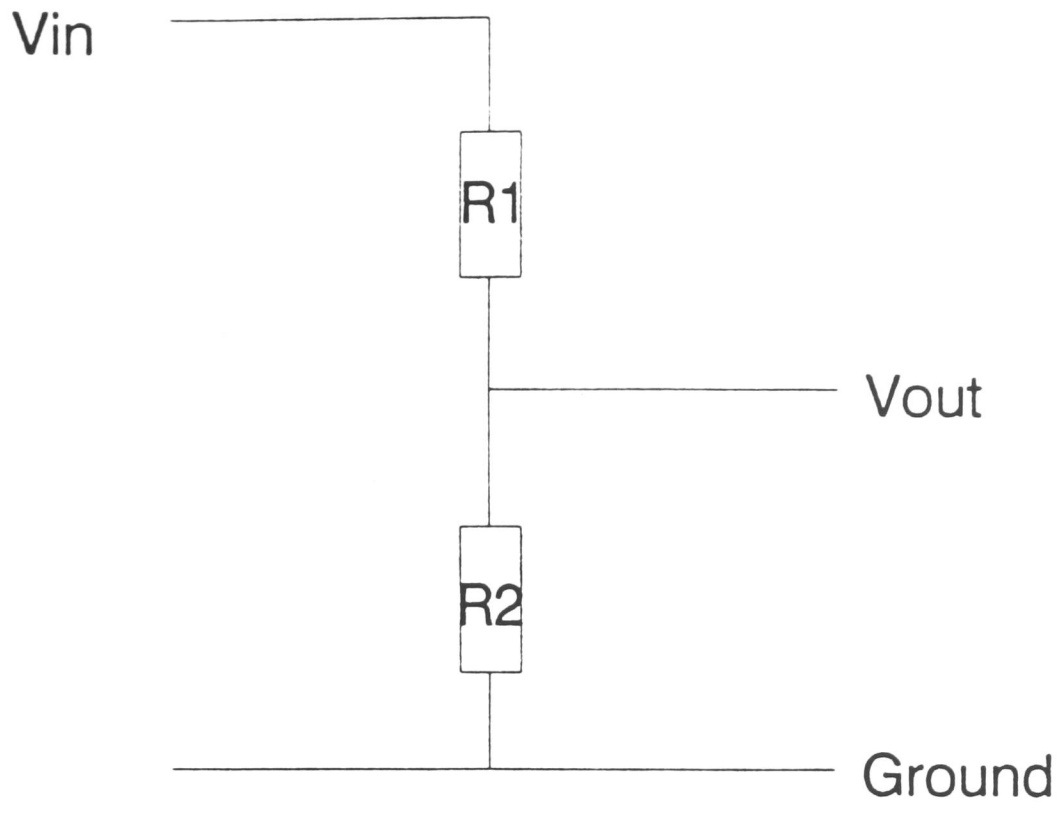


Figure 6.21 The use of a resistor chain to reduce output voltage.

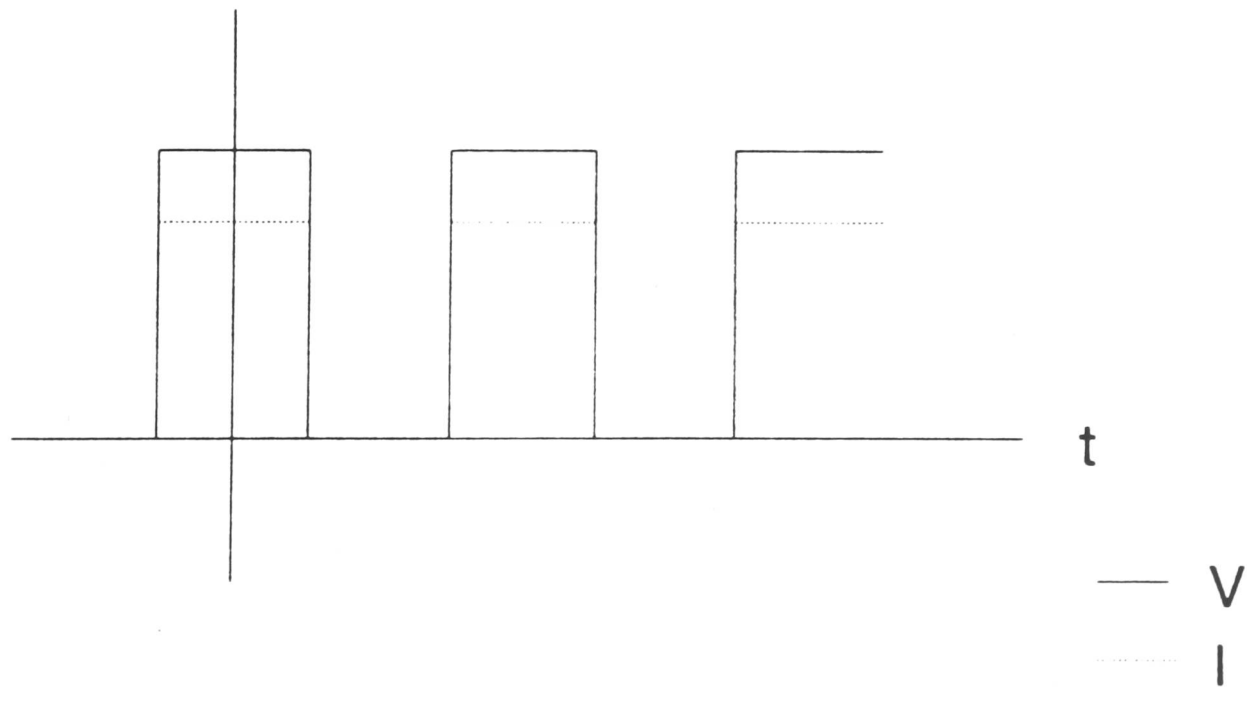


Figure 6.22 The ideal output waveforms with a resistive load.

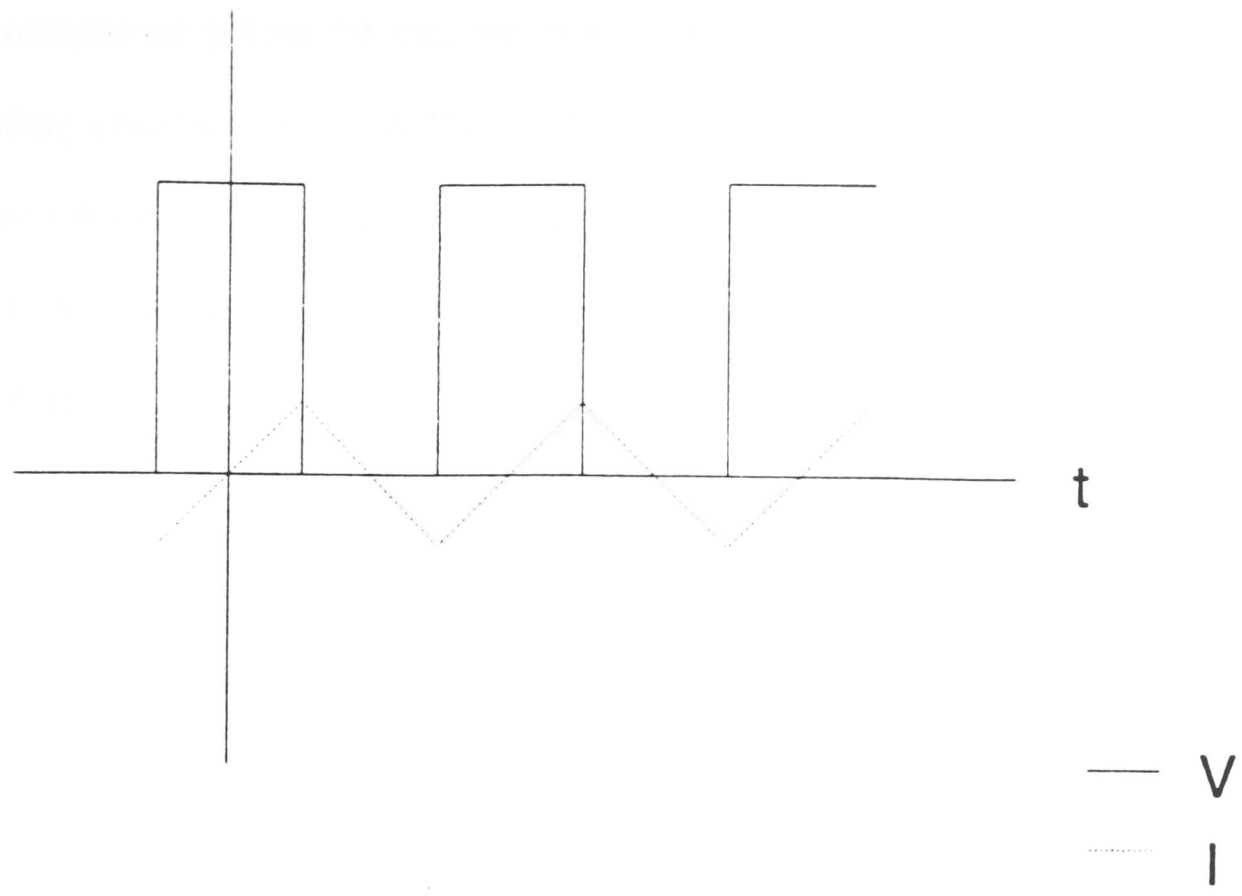


Figure 6.23 The ideal output waveforms with an inductive load.

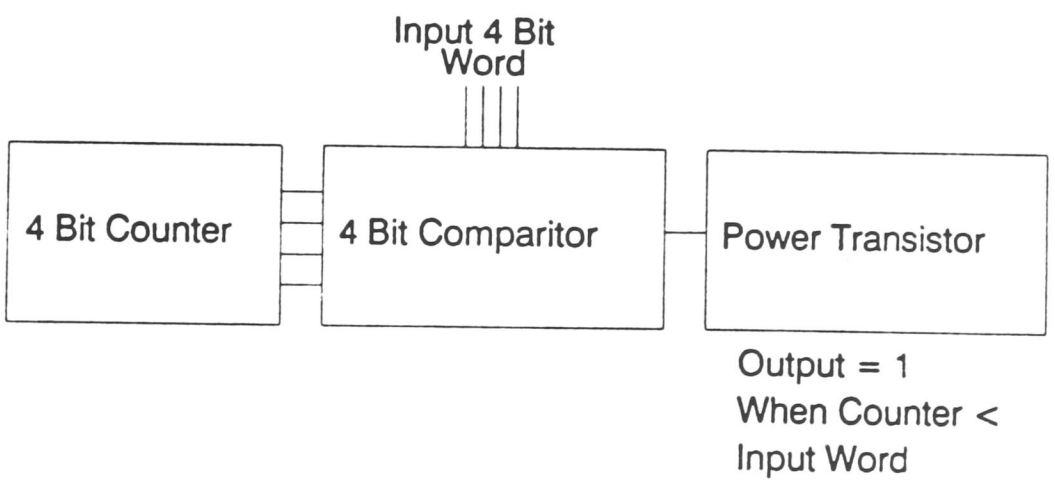


Figure 6.24 A block diagram of the d.c. chopper circuit.

The control logic block diagram of the circuit used is shown in figure 6.24. It consists of a four bit counter and a four bit comparator. The binary counter cyclicly counts from 0 to 15. Input to the comparator is a four bit word. If the output from the counter is less than the word, the power transistor is turned on by applying five volts to the gate. This means that the mark space ratio can vary from 1:14 to 14:1.

The frequency is set by an external clock. However if the CMOS process used could generate analogue elements then this clock could be generated on-chip.

The logic circuitry was constructed using ES2's cell library, SOLO. It is based on a 2 μm n-well CMOS process. In order to make the design compatible with the smart power process it was scaled up by a factor of three to 6 μm design rules.

In addition an extra mask level had to be added. This is the field implant mask which in the EMF's 5 μm CMOS process is just an oversize of the well mask. Due to the complexity of the circuit, interconnections were made using a double level metallisation. This is not a standard feature of the EMF's process but is an option that requires two extra masking levels, vias and metal 2. It would have been possible to use only single level metal and design the logic circuitry from scratch, but due to the time limitation of a PhD, the former approach was adopted.

The power transistor was similar to the optimised discrete power transistors described earlier in this chapter. The difference was that the transistor consists of striped drains and sources, and is 1mm x 1mm in size. The design consists of one

power transistor driven by the control circuitry and one separate power transistor to test performance under variable conditions.

6.6 Chopper Circuit Results

The logic design was done using the SOLO suite on a SUN workstation. The resulting CIF files were transferred to SERC's PRIME F computer. There they were converted to GAELIC. The power transistor was designed in CIF and converted to GAELIC on the PRIME. Then the two parts of the design were merged to form the circuit. This completed the smart power chip design. In addition to this a discrete power transistor was put on the mask set. The reason for this was to be able to test at least the power part if the logic circuit failed.

Then the files were converted to David Mann format on the EMF's VAX 11/750 in preparation for mask making. The masks were then fabricated in-house using an emulsion process. The design was then fabricated in the EMF using a double level metal CMOS process.

When the circuit was tested, it was found that the power transistor in the circuit did not work. The reason for this was an error on the active area mask. It was later found that during conversion from one file format to another a rounding error occurred on mask level two, active area. This probably happened during the transfer from CIF to GAELIC and was found to be fatal. Due to the removal of the

PRIME computer from service it was not possible to construct another active area mask using the existing design, and unfortunately the circuit could not be used as a demonstrator for the smart power process. However, the discrete power transistor was found to have a breakdown voltage of 125 V. Also, it should be remembered that in section 6.4 both low voltage and high voltage transistors were fabricated using the same process as this circuit. This demonstrated the possibility of a nine mask CMOS-compatible smart power process. The only process modification used here for the chopper circuit was the second level of metal used for the logic.

6.7 Conclusions

The goal of this work was to realise a working 120 Volt smart power process. Additionally, if possible, it would be an advantage if the low voltage CMOS transistors were not affected by process changes so existing cell libraries could be incorporated into smart power designs.

There existed two possible solutions using either vertical or lateral power transistors. The power devices used were of the lateral double-diffused MOS (LDMOS) type as this was found to be the most area-efficient and the easiest to integrate at the required power levels (from simulations). A matrix of different mask dimensions and process variations was performed in order to determine the optimum cell size and the simplest process flow. From measurements, it was found that for the power transistor the addition of a field oxide region between the drain and source could dramatically increase the immunity to necessary to punchthrough. This structure has not been seen previously in the literature, and is novel to this work. Its effect is crucial in separating the drift region and the channel of the device.

It was found that the channel stop implant was not necessary to prevent latch-up of the CMOS logic transistors in this process, and could therefore be eliminated simplifying the process flow and reducing the fabrication costs. The minimum dimensions were found for gate length, drift region length and gap between gate and drift region, to define a design rule set for the power transistors. In order to have a good design these distances could be guardbanded by adding tolerances to compensate for mask misalignment and process variations. It was also shown that the n-channel power transistors could support 120V and the

p-channel transistors -55V. This opens up applications as display drivers, and for automotive and telecommunication applications.

Therefore in this chapter, a process which realises 120V high voltage power transistors with a nine mask process that is compatible with an existing low-voltage CMOS process was demonstrated. The novelty of using a field isolation to separate the drift region and channel is the main reason that this kind of performance can be realised using such a simple process flow. Additionally, a one mask option can add -55V PMOS power transistors to the designers possibilities.

Again, this p-LDMOS was a structure previously unseen. The performance of this transistor is much superior to simple pMOS transistors using a lower drain doping as proposed in chapter 5, and this leads to a relatively low cost solution to the problem of realising complementary high-voltage MOS transistors for a smart power process.

References

[1] V.Rumennik, "Power Devices are in the chips", IEEE Spectrum, pp.42-48, July 1985.

[2] C.Petersen and A.R.Barlows, "High Voltage Circuits in standard CMOS processes", IEDM Tech. Digest, pp.77-80, 1982.

[3] C.Contiero, A.Andreini and P.Galbiati, "Multipower BCD - A Versatile technology from 60V to 250V that integrates Bipolar, CMOS and DMOS Devices", IEE Colloquium on Integrated Intelligent Power Devices, March 1989.

[4] Z.Parpia, C.A.T.Salama and R.Hadaway, "Modelling and Characterization of CMOS-compatible High-Voltage Device Structures", IEEE Trans. Electron Devices, vol.ED-34, pp2335-2343, 1987.

[5] Y.Tarui, Y.Hayashi and T.Sekigawa, "Diffusion Self-Aligned Enhance-Depletion MOS-IC", Suppl.J.Jpn.Soc.Appl.Phys., vol.40, pp.193-198, 1971.

[6] K.A.Shenai, P.A.Piaceto, R.Saia and B.J.Baliga, "Blanket LPCVD Tungsten Silicide Technology for Smart Power Applications", IEEE Electron Device Letters, vol. EDL-10, p.270, 1989.

[7] M.R.Duncan, J.M.Robertson, R.J.Holwill and R.Rodrigues, " Design of CMOS-Compatible High Voltage Devices", ICSSICT, 2nd International Conference on Solid-State and Integrated Circuit Technology, October 22nd-28th 1989, Beijing, China.

[8] M.R.Duncan, J.M.Robertson, R.J.Holwill and R.Rodrigues, "CMOS-Compatible High Voltage Complementary LDMOS Devices", ESSDERC, 19th European Solid State Device Research Conference, September 11th-14th 1989, Berlin, Germany.

CHAPTER SEVEN : IMPROVED SMART POWER PROCESS

7.1 Introduction

In the last chapter a process was developed that realised 120 BV_{DSS} nMOS and -55V BV_{DSS} pMOS power transistors. It used ten mask levels which was one more than the standard low voltage CMOS process on which it was based. In order to realise this performance, a new transistor structure was developed, where the channel region was separated from the lightly-doped drift region by a field oxidation layer. These structures were similar to the discrete lateral double-diffused MOS power transistors described in chapter 3, and could be fully integrated into a CMOS logic process. There were however two notable drawbacks in this solution.

Firstly, it was found that the n-channel LDMOS could be fabricated with a 5 μm or a 8 μm well, but in order to support complementary power transistors the deeper well was necessary. This deeper well did not improve the nLDMOS device performance but was detrimental to cell size. It also meant a 12 hour increase in processing time. For the proposed smart power process this represents 2% of cycle time and should be eliminated if possible.

Secondly, it was found that the processing of the triple-diffused structure used for the high-voltage pMOS devices was particularly critical. From an analysis of the difference in device performance compared with small fluctuations in the process, it was seen that a variation of 0.5 μm in the junction depth of the p-region corresponds to an increase of only 20°C during the p- drive-in, and this kind of processing tolerance is more than possible in a manufacturing environment. This type of variation of the junction depth will reduce the breakdown value of the pLDMOS from -55V to -40V. From this it can be seen that the process would be difficult to control in production.

For the two reasons cited above, it seemed appropriate to try to improve on the high-voltage pMOS device structure and to use the shallowest well possible. As explained earlier, for reasons of packing density and ease of integration, lateral power structures were to be preferred.

7.2 Device Structures

Since the lateral NMOS power transistors was shown to have an excellent performance with the 5μm well used as the lightly-doped region, then it was a reasonable objective to keep it in the process flow if possible. An alternative structure for the p-channel power transistor was therefore needed.

A modification in device layout for both n and p-channel devices was made at this stage. By forming the power transistors of concentric rings the structure is self-guarding and so eliminates the need for extra guard rings. This change in structure will lead to an increase in packing density : see figure 7.1 for a plan view of an n-channel power transistor.

The proposed modification of the pLDMOS is as follows. If instead of a p-substrate, the substrate was n- with a p- epitaxial layer then the lateral DMOS of figure 7.2 could be realised. Without this epitaxial layer the device would be like that of figure 7.3, and the substrate would act as the channel and would short out. This type of structure would be suitable for a discrete device where the substrate could be left floating, but it could not be used in an integrated circuit as the substrate

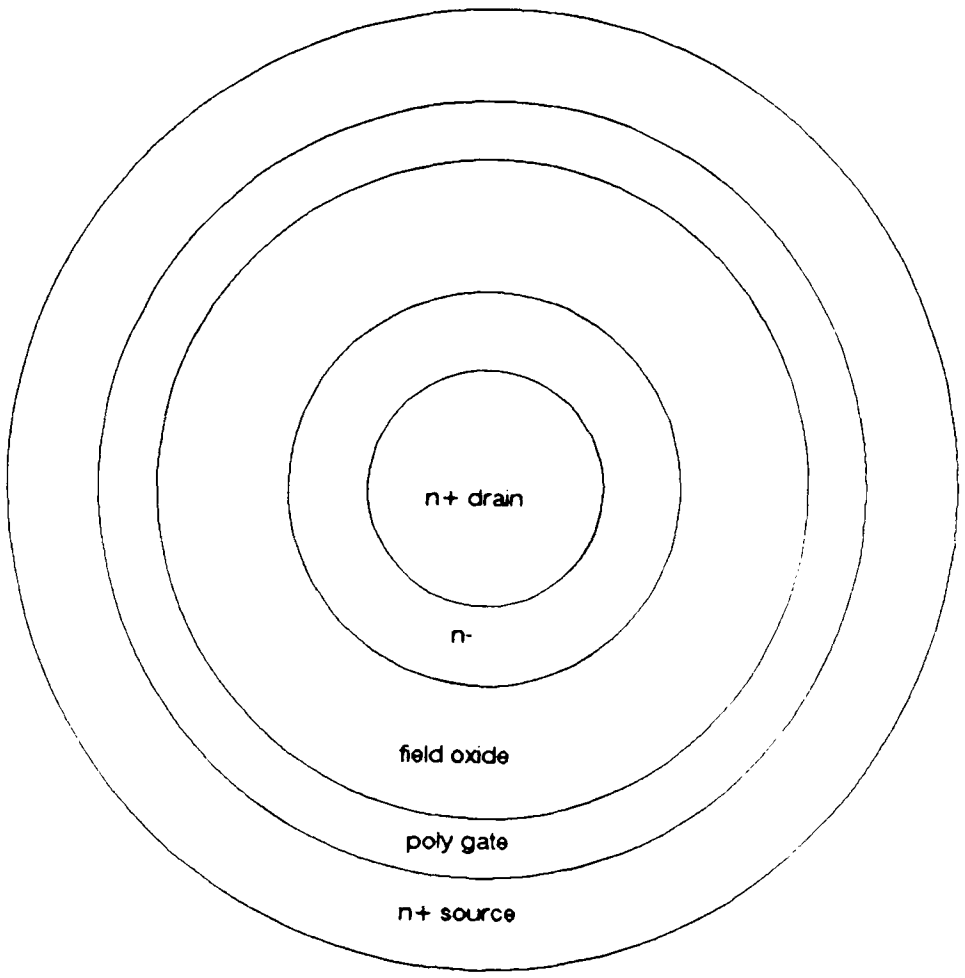


Figure 7.1 A plan view of a self-guarding n-LDMOS power transistor.

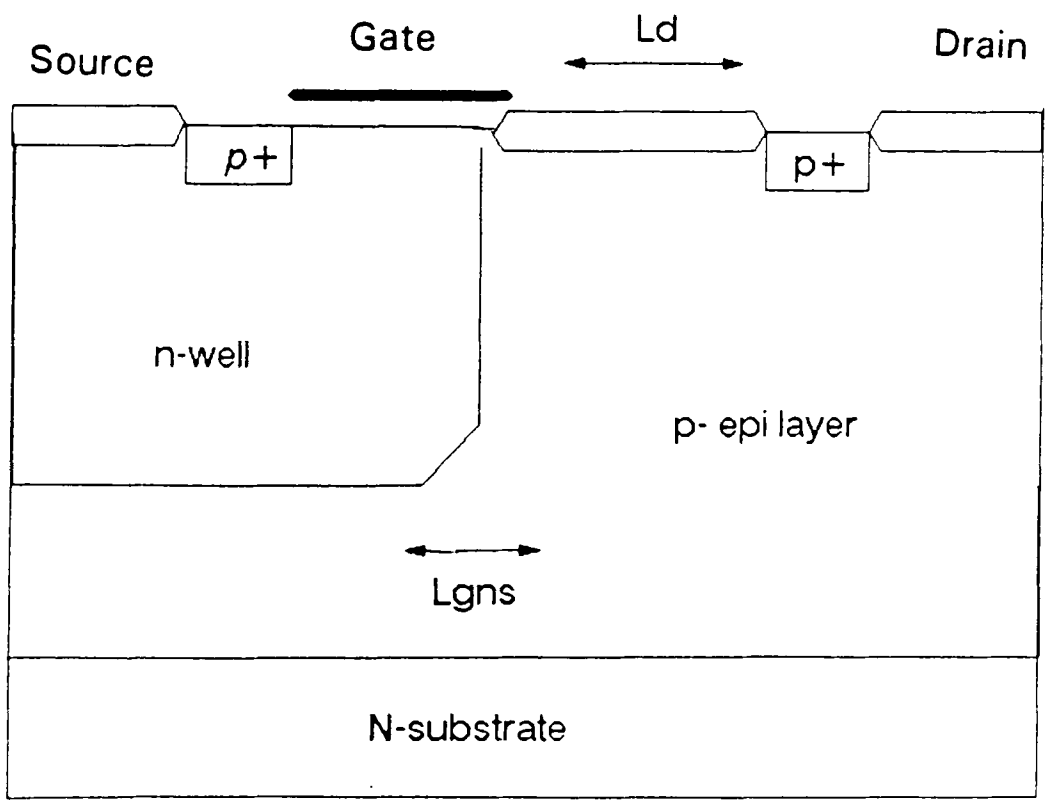


Figure 7.2 The p-channel LDMOS with the field oxide inserted in the channel, fabricated on an epitaxial layer.

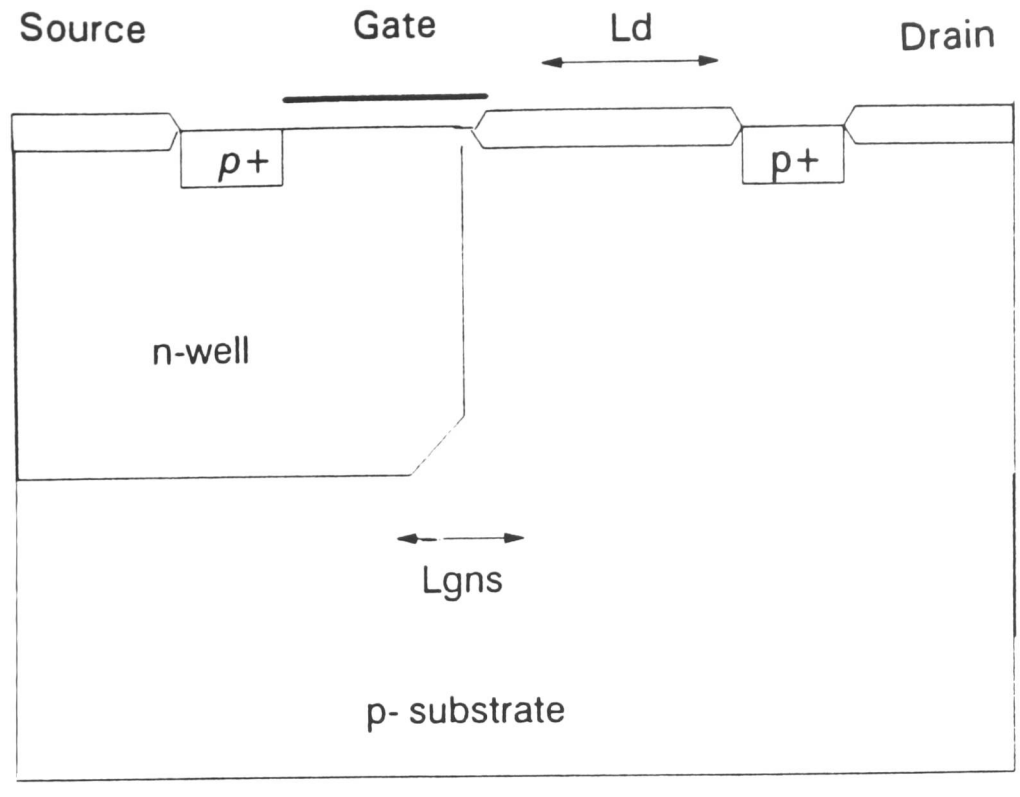


Figure 7.3 The p-channel LDMOS with the field oxide inserted in the channel, using the p-substrate as the lightly-doped drift region.

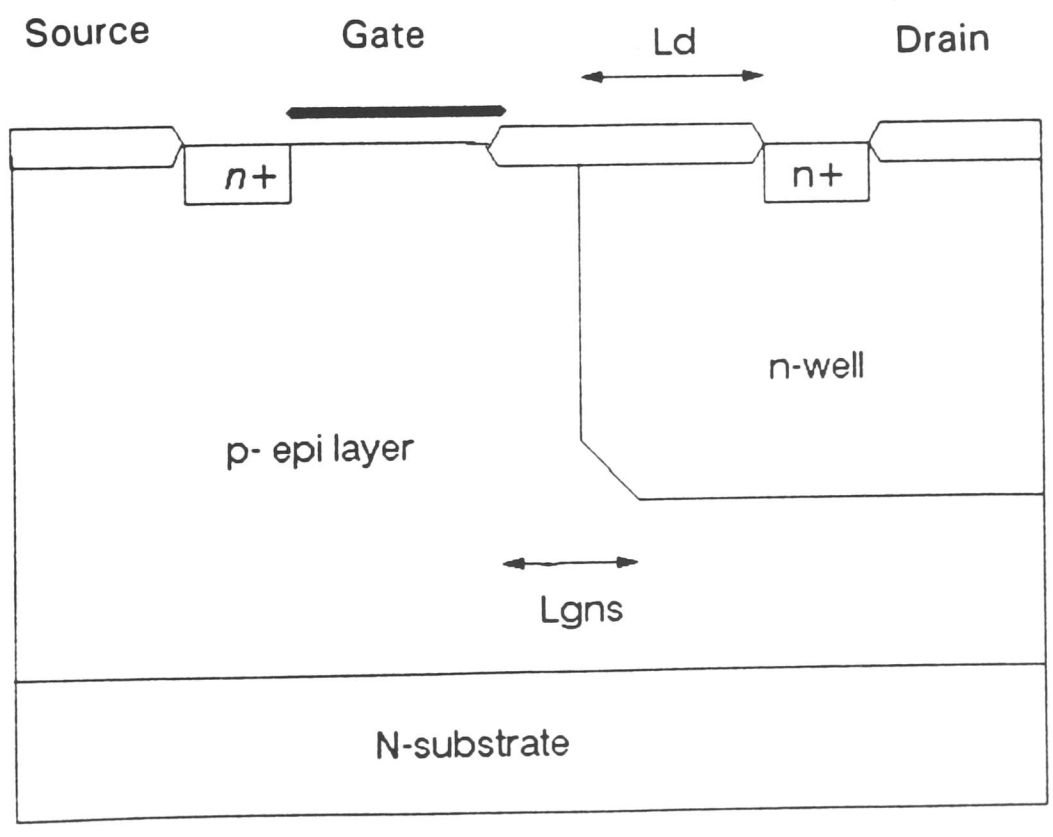


Figure 7.4 The n-channel LDMOS with the field oxide inserted in the channel, fabricated on an epitaxial layer.

in generally connected to ground. The epitaxial layer changes the n-LDMOS to that of figure 7.4. Despite this change, from simulations, there should be no change in either the performance of the power transistor or the low voltage device behaviour, if the epitaxial layer is sufficiently thick and has the same resistance as the original substrate.

The p-channel device, however, should show a substantial improvement in operational performance. As it is formed with deeper junctions and with lower doping concentrations than the previous structure, increased safe operating voltages were expected. By using SUPREM3 and CANDE simulations, predicted avalanche breakdown voltages for the p-channel LDMOS were found to be -110V. It can be seen in figure 7.5 that the ionisation integrals exceed unity at -110V. This signifies that the number of electrons and holes generated by avalanche breakdown exceed the number consumed, and the device is in breakdown. This simulation was performed cosidering a thick epitaxial layer of 40μm. However, increased voltages can be expected if the RESURF technique is employed.

The effect of the RESURF method can be explained using the diode structure of figure 7.6. When two of these diodes were formed identically except for epitaxial layer thickness, it was found that the diode with the thin layer (15μm) exhibited a significantly higher breakdown voltage than the diode with the thick layer (50μm) . The difference was found to be 1150V compared with 370V (using SUPREM3 and PISCES2B simulations).

The horizontal length L, of the n- region between the n+ contact and the p+ junction, is greater than the epitaxial layer thickness. Both the horizontal and vertical junctions can be approximated by the plane junction case. The depletion at the vertical junction is one-sided, in contrast with the horizontal junction where

depletion is two-sided. This means that the critical electrical field will be reached first at the surface.

Therefore if the surface field could be reduced, it would be possible to apply a much greater potential until the critical field at the horizontal junction was reached. This is the case when the epitaxial layer is made thinner.

The influence of the depletion from the bottom side of the vertical junction becomes considerable. Due to the interaction between the vertical and the horizontal junctions the depletion layer stretches along the surface over a much longer distance than in the plane junction case. This causes a large reduction of the surface electric field. Further increase of the reverse junction voltage will lead to total depletion of the layer and subsequent failure.

In conclusion, by reducing the epitaxial layer thickness the peak electric field at the surface is reduced and higher breakdown voltages can be achieved. The n-channel power transistors of the previous chapter did not use an epitaxial layer whereas the structures proposed here do. By varying the epitaxial layer thickness then it is possible that by taking advantage of the RESURF effect, greater safe operating voltages can be obtained. This proposed process has the added advantage over the process of the previous chapter, in that it only uses nine masking layers, bringing a reduction in both costs and production cycle time.

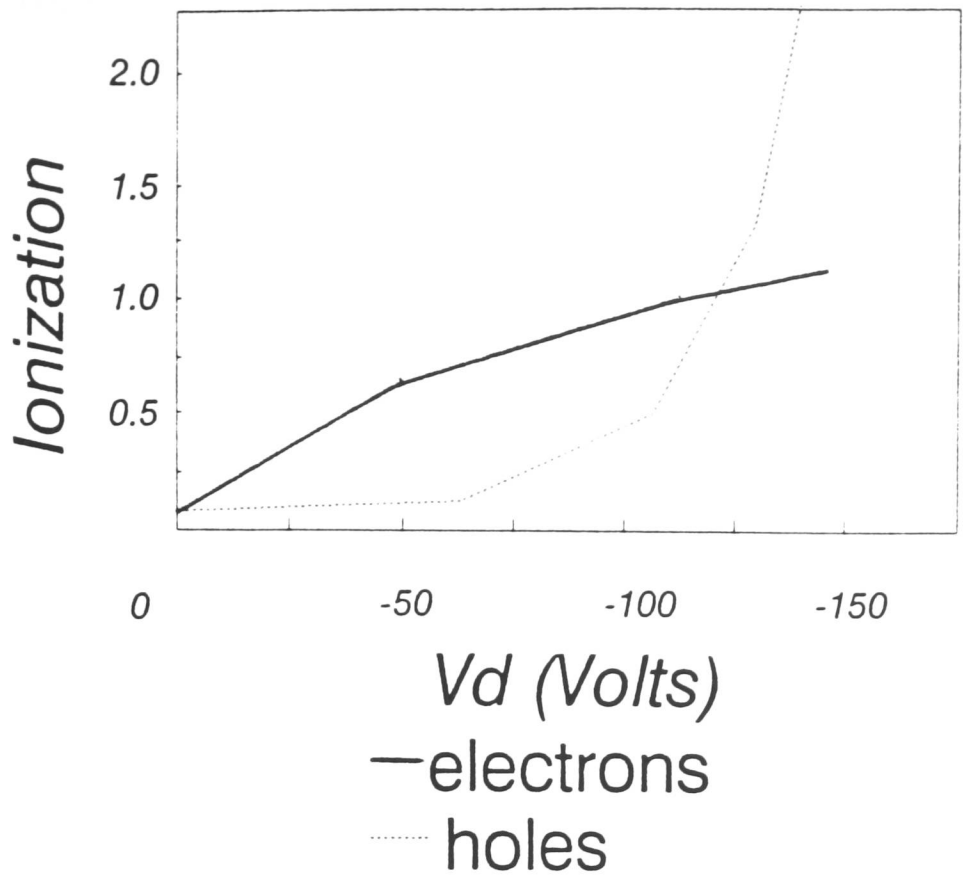


Figure 7.5 A CANDE simulation showing the ionization integrals (to determine BV_{DSS}) for the p-LDMOS transistor.

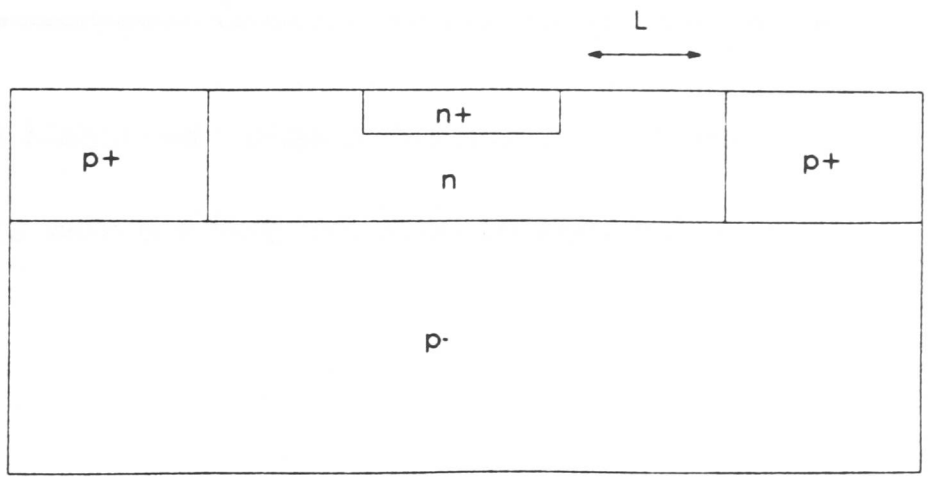


Figure 7.6 The RESURF Diode structure.

7.3 Results

Due to the fact that an epitaxial layer could not be deposited in house, only one thickness, 25μm was used. In order to optimise transistor dimensions, devices were fabricated with different drain and channel lengths. There was also a series of devices to determine the best drift region to channel separation L_{gns} . The basic structure for the p-LDMOS is shown in figure 7.2

Drift length L_D (μm)	**Gate-Space** L_{GNS} (μm)	**BV_{DSS}** (Volts)	**Threshold** (Volts)
5	2.5	18	1.6
7.5	5	118	2.0
10	5	125	2.0
12.5	7.5	125	2.0
15	10	124	2.0

Table 7.1 Measured values of breakdown and threshold voltage for the n-LDMOS transistors with the field oxidation on epitaxial wafers.

The results for the n-LDMOS are summarised in table 7.1. It can be seen that the optimum gate length is 5μm . Similarly the drain length and separation are given by 10μm and 5μm respectively. The highest breakdown voltage achieved was 120 Volts. This is exactly the same as for the devices in the last chapter without the epitaxial layer. There is a slight improvement in on-resistance, but this is probably due to repeatability of sheet resistances from

batch to batch, as the on-resistance is dominated by the n-well drift zone. Therefore for n-channel power transistors there is no conclusive evidence to show an improvement in performance due to the epitaxial layer. The Drain current characteristics are shown in figure 7.7.

The highest breakdown voltage obtained for the triple-doped p-channel power transistors of chapter six was -55V. It can be seen from table 7.2 that the epitaxial p-LDMOS devices have a vastly superior breakdown voltage. It is nearly -100V, as can be seen from the current characteristics of figure 7.8. An extrapolation of the threshold voltage is shown in figure 7.9.

Drift length L_D (μm)	Gate-Space L_{GNS} (μm)	BV_{DSS} (Volts)	Threshold (Volts)
5	2.5	-21	-2.0
7.5	5	-91	-2.5
10	5	-95	-2.5
12.5	7.5	-96	-2.5
15	10	-96	-2.5

Table 7.2 Measured values of breakdown and threshold voltage for the p-LDMOS transistors with the field oxidation on epitaxial wafers.

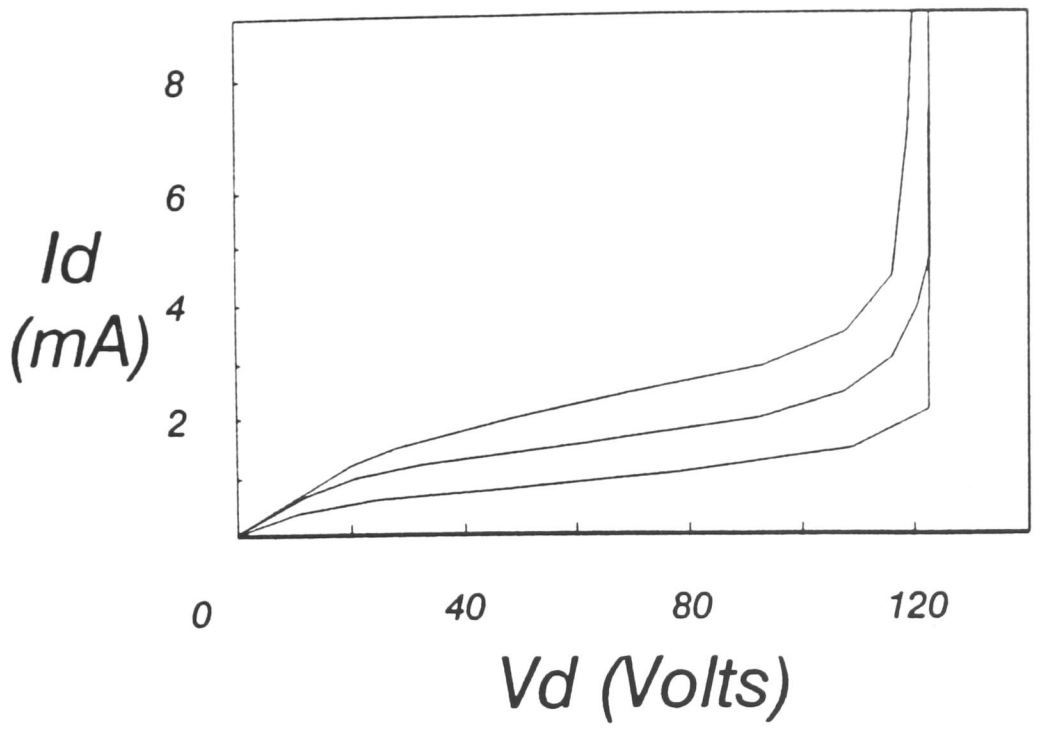


Figure 7.7 The drain current characteristics of the n-LDMOS transistor, fabricated on an epitaxial layer.

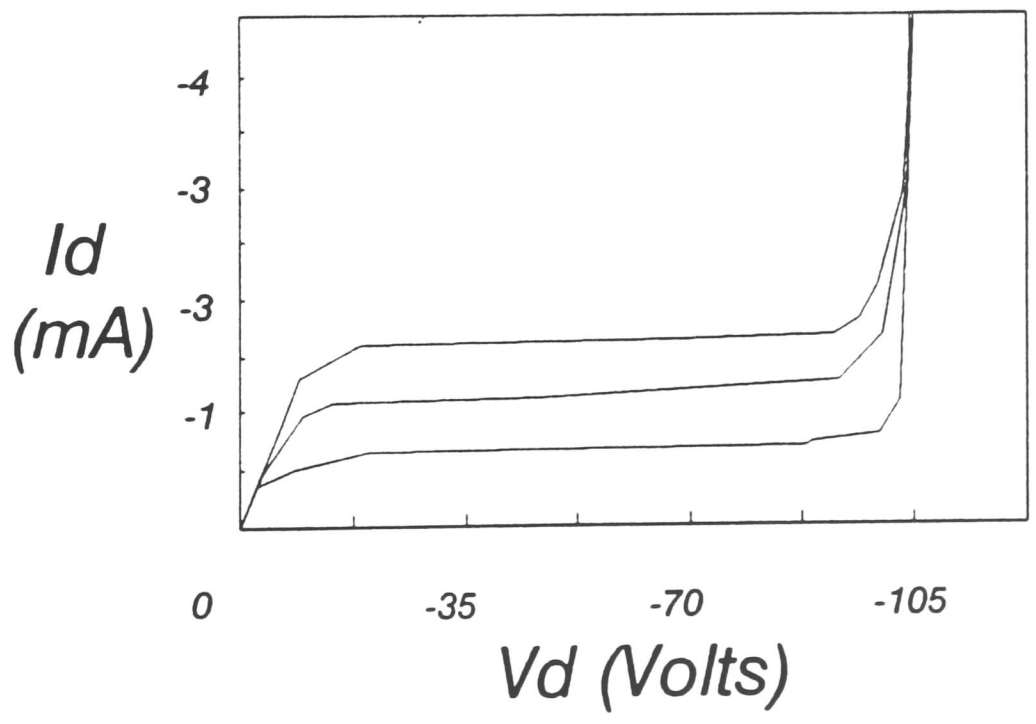


Figure 7.8 The drain current characteristics of the p-LDMOS transistor, fabricated on an epitaxial layer.

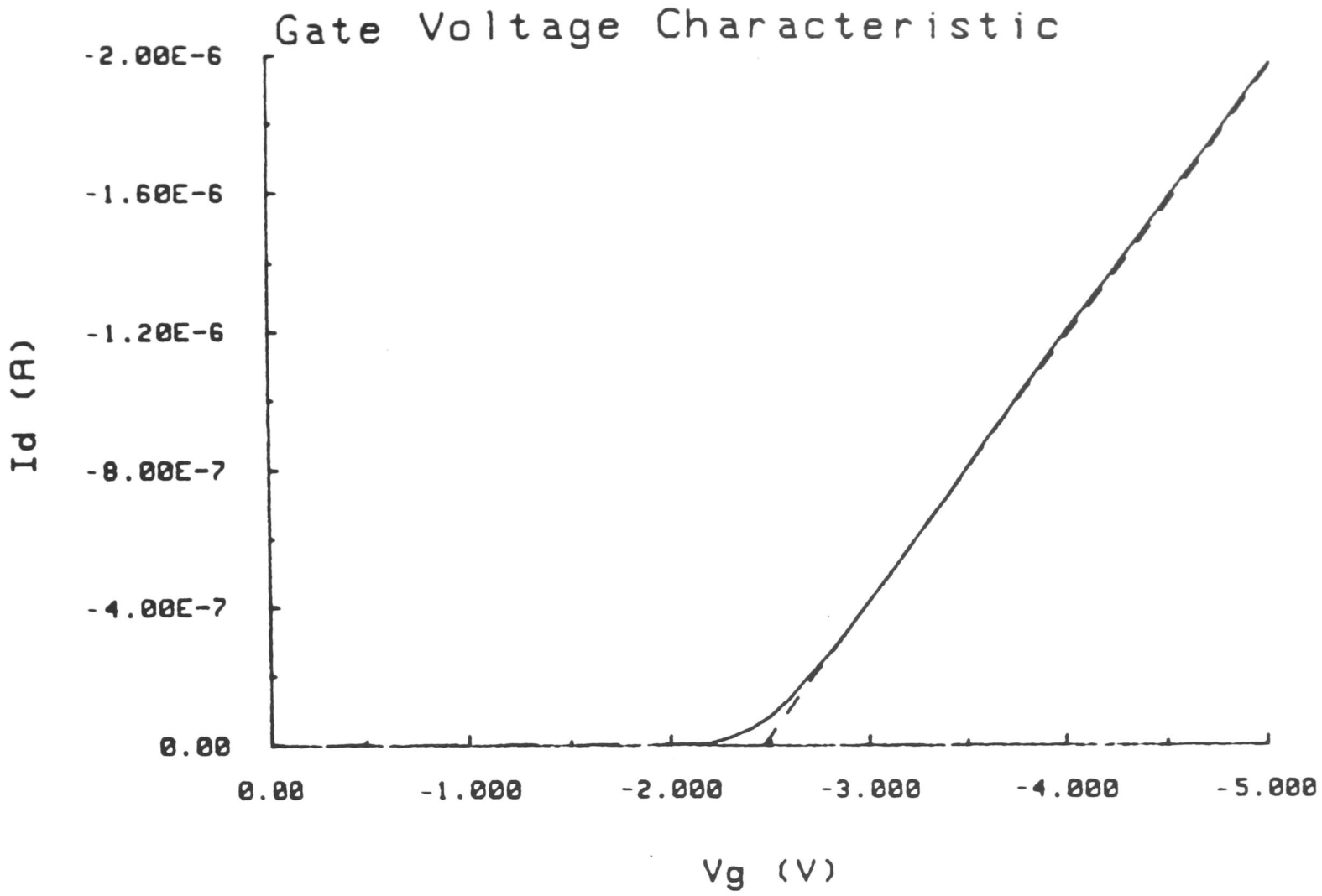


Figure 7.9 An extrapolation using PARAMEX to determine the threshold voltage of the p-LDMOS transistor.

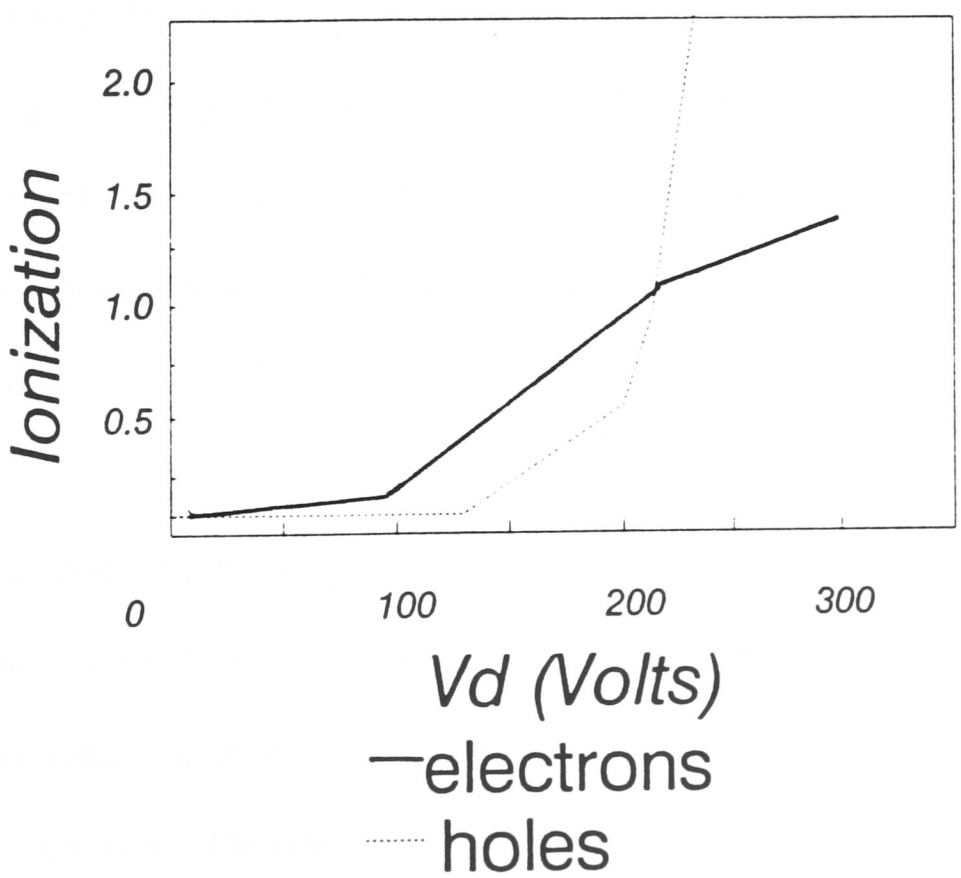


Figure 7.10 A CANDE simulation showing the ionization integrals (to determine BV_{DSS}) for a n-LDMOS transistor fabricated on a 20μm epitaxial layer.

The improvement in safe operating voltage is due to two factors. Firstly, the doping level of the drift region is lower than before. Secondly, the previous structure was vertically punchthrough-limited due to the triple doping. In contrast, by using PISCES simulations and looking at current flow it was found that as for the n-LDMOS devices, this structure is limited by the lateral junctions. In this case, it is reach-through from the drift region to the channel that causes failure. The onset of this failure is contained by the field oxidation inside the device structure. This field oxidation helps to reduce the peak electric fields near the surface. Without this oxidation, breakdown occurs sooner as the drift region is more strongly influenced by the potential of the gate electrode. Table 7.2 shows that the best distances for gate length, drain length and separation are found to be 5μm , 10μm and 5μm respectively. Larger values increase on-resistance without increasing breakdown voltage.

It was found by comparing simulated and measured values for both the n- and p-channel power transistors, that simulations gave higher estimates of the breakdown voltage. This phenomenon was noted in the previous chapter. If the criteria for judging the onset of breakdown is changed in the simulations to when the ionisation integrals exceed 0.8 then a reasonable fit to measured values is obtained (see table 7.3).

It is important to optimise the epitaxial layer thickness for performance. Unfortunately due to the lack of an epitaxial reactor at the EMF, only one thickness of epitaxial layer was available for this experiment. This layer was produced at Southampton University. This meant that investigation into the best thickness due to the RESURF effect could not be carried out in the clean room. The epitaxial layer used was 25μm.

Drift length L_D (μm)	Gate-Space L_{GNS} (μm)	BV_{DSS} when ioniz. integral = 1.0 (Volts)	BV_{DSS} when ioniz.integral = 0.8 (Volts)
5	2.5	-24	-21
7.5	5	-110	-98
10	5	-112	-100
12.5	7.5	-112	-100
15	10	-112	-100

Table 7.3 Simulated values of breakdown voltage for the p-LDMOS transistors with different coefficients of the ionization integral.

In order to investigate the RESURF phenomenon, some simulations of power transistor breakdown as a function of epitaxial layer thickness were carried out. The results are summarised in table 7.4. The empirically obtained value of 0.8 for the ionisation integrals was used. It can be seen that above 25μm, the RESURF effect does not occur and the breakdown voltage is constant. This is confirmed by the measured results obtained for the n-LDMOS power transistors where the breakdown voltage for the 25μm epitaxial layer is the same as for the transistors realised on p- substrate, without an epitaxial layer.

Epitaxial layer thickness	Simulated BV nLDMOS	Simulated BV pLDMOS
10μm	74 V	-61 V
15μm	144 V	-107 V
20μm	195 V	-182 V
25μm	125 V	-100 V
30μm	125 V	-100 V
100μm	125 V	-100 V

Table 7.4 Simulated breakdown voltages for different thicknesses of epitaxial layer. Calculated when the ionization integrals exceed 0.8.

However, if the epitaxial layer is reduced to 20μm then a dramatic increase in breakdown voltage is seen. Figures 7.10 and 7.11 show the ionization integrals for the n- and p-channel LDMOS respectively with a 20μm epitaxial layer. By further reducing the thickness to 15μm then the positive effect starts to reduce, and at 10μm the devices punch-through vertically into the substrate. Therefore we can conclude that if an epitaxial layer of 20μm is used with the device structures of figures 7.2 and 7.4, then an n-channel power transistor with a breakdown of nearly 200V, and a p-channel power transistor with a breakdown of over -180V can be realised.

In figure 7.12 a contour plot of electric potentials is shown for the n-LDMOS (on 20μm epi) with 200V on the drain and 0V on the gate. This is at the onset of breakdown. It can be seen that the device is breaking down vertically in this case.

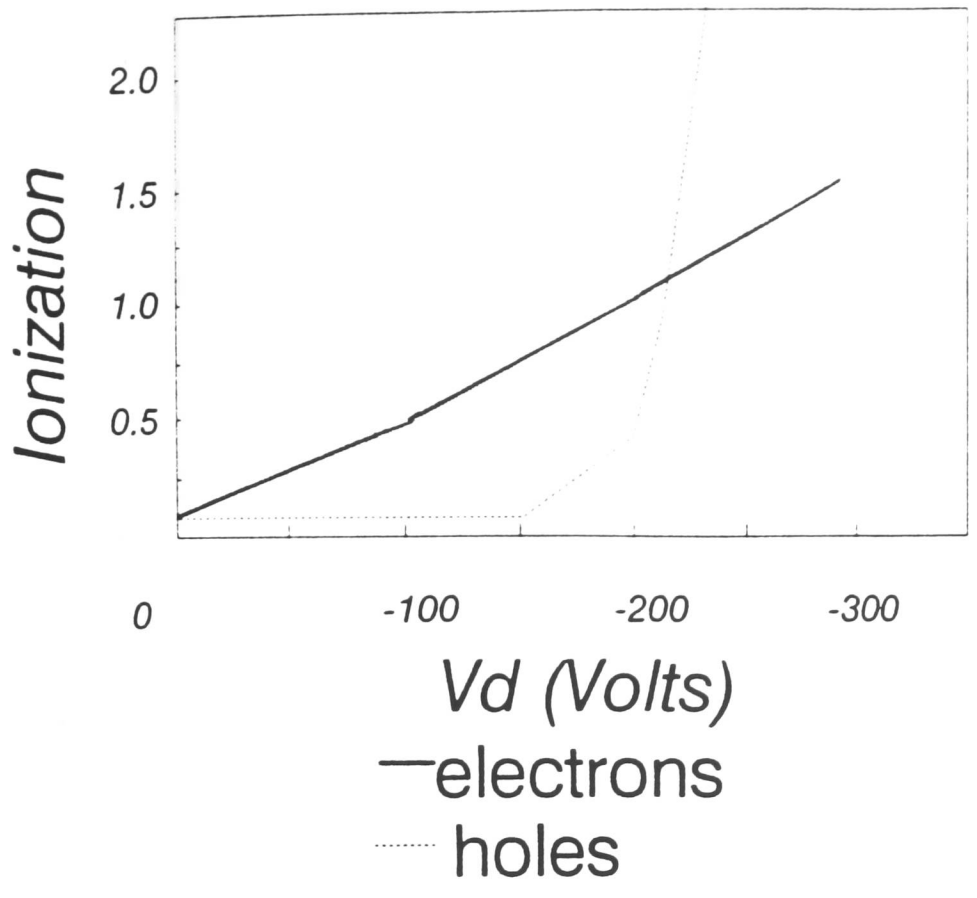


Figure 7.11 A CANDE simulation showing the ionization integrals (to determine BV_{DSS}) for a p-LDMOS transistor fabricated on a 20μm epitaxial layer.

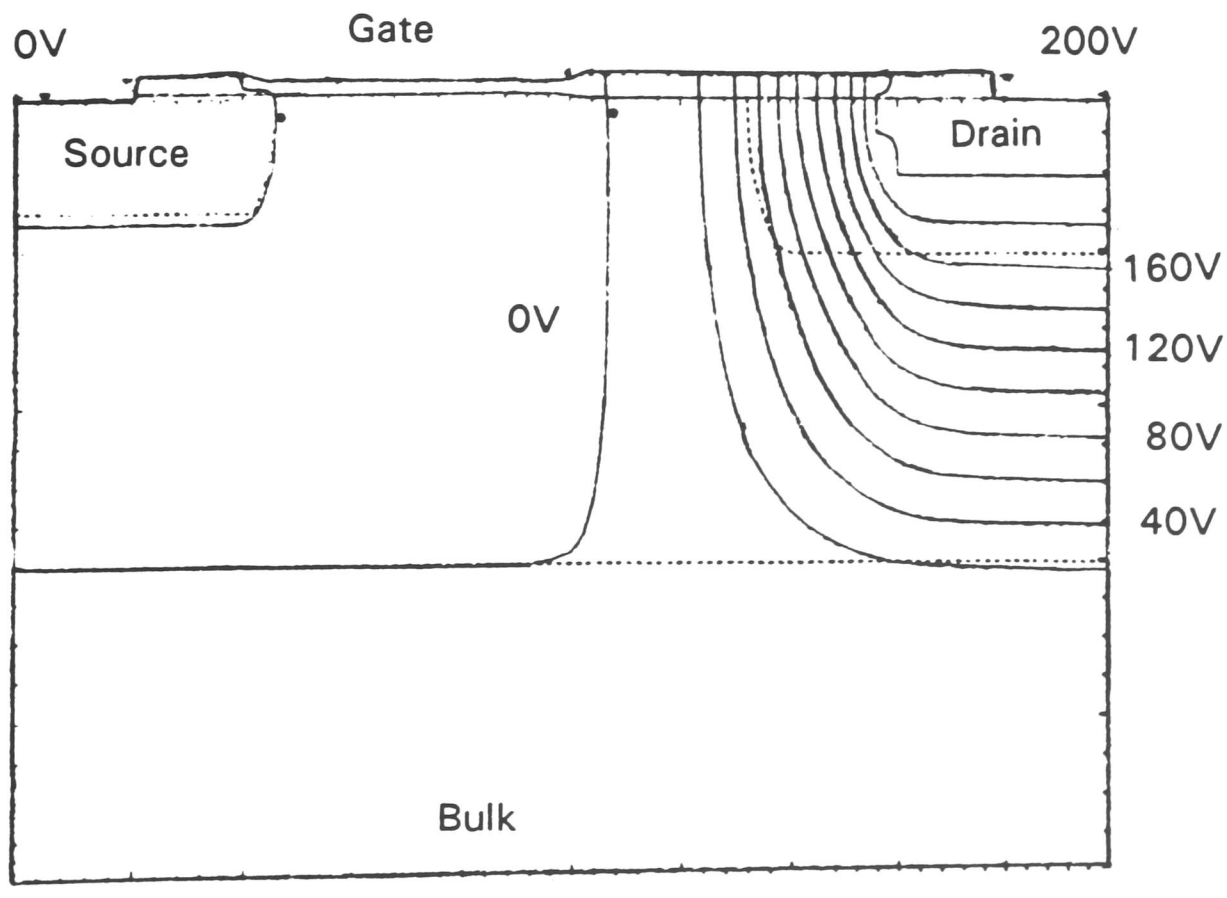


Figure 7.12 A CANDE simulation showing the voltage contours for the n-LDMOS transistor on 20μm epi, with 200V on the drain terminal and with all other terminals at ground. The potential lines (solid) are in 10V steps.

It was found that if the well depth was increased to 8μm, then the breakdown voltage did not change, but the critical point moved to the surface and the device becomes limited by lateral punchthrough breakdown. In order to further improve on these values, the electric field at the surface under the field oxidation would have to be reduced. This cannot be achieved using the RESURF effect alone.

7.4 Conclusions

The main objective of this chapter was to improve on the performance of the high voltage p-channel transistor. The structure proposed in the previous chapter was shown to withstand -55V but would be difficult to produce in volume due to its tight process margins. The low voltage transistors and the high voltage n-channel transistor structures used here had been fabricated previously without the epitaxial layer. It was found that there were no differences in performance with or without a 25μm epitaxial layer.

The new structure proposed for high-voltage p-channel devices achieved a higher breakdown value, -95V, with a similar on-resistance value to the previous structure. In addition to this, one masking layer was saved. The modified process flow removes one PR stage, one implant and shortens the well drive-in by 12 hours. The only additional process requirement is for the initial epitaxial layer. Therefore there is a net shortening of cycle time, and a corresponding saving in chemicals and cost. These two results of better performance and lower cost demonstrate the effectiveness of the new device structure.

Further improvements in the performance can be expected if the epitaxial layer thickness is optimised. By finding the ideal epitaxial thickness and exploiting the RESURF effect, safe operating voltage can be significantly increased. The fact that the equipment necessary to characterise this effect was not available, meant that only one thickness could be tried experimentally. However from simulations, it was found that the breakdown voltage could be significantly improved for both the n- and p-channel power transistors. The results are very impressive, 200V for the n-channel and -180V for the p-channel devices. It is important to remember that this is obtained using a nine mask process that is fully compatible with a low voltage CMOS process. The breakdown voltages for the high-voltage transistors can be contrasted with the logic transistors which formed the starting point of the work. The safe operating voltages of the low voltage parts are an order of magnitude lower.

CHAPTER EIGHT : CONCLUSIONS

The objective of this project was to develop a working smart power process. This process had to be compatible with either TTL or CMOS logic and had to sustain higher power or voltage levels than conventional transistors.

It is known that for discrete power devices, bipolar transistors are preferable to MOS as they have a greater current capacity for the same die size. In contrast, MOS technology is preferred to bipolar for logic circuits due to its increased packing density. Therefore, the ideal smart power device would have bipolar power parts and MOS logic. The drawback of this is the complexity of process needed. A compromise needs to be made. By comparing bipolar power and logic transistors, it can be seen that the devices are superficially similar in structure. However, all junction depths and doping levels are different. Therefore, separate masking levels and process steps would be required to realise both low voltage and high voltage transistors on the same die.

A second drawback for bipolar technology is that current flow in power transistors is vertical. If more than one power transistor is required, as in a smart power circuit, then up-down sinker diffusions and a buried layer are needed in order to have all terminal connections on the top surface. These diffusions add to the process complexity and decrease packing density. It was clear that using bipolar technology for smart power, rather than as intelligent switches, a complex process is required.

In contrast to bipolar transistors, it can be seen that power MOSFETs are structurally different to logic transistors in that they have two distinct regions. The first region is lightly doped and supports the high tension. The second region is the channel area. Various configurations are possible, both lateral and vertical. Vertical structures however, suffer from the same processing and integration problems as bipolar power transistors, and are therefore more suited as intelligent switches, where only one power transistor is required per die. It was found using simulations that for 300 V and above vertical structures are dominant, but for 100 V applications lateral devices are the better choice. They have higher current density, greater packing density, are easier to integrate and can be realised with a simpler process.

It was seen that it may be possible to use many of the same masking levels of a low voltage process in order to form the channel part of a lateral power device structure. It is clear that the lightly-doped drift region of the power LDMOS will need to be formed in another manner.

It can be noted that LDMOS transistors are inherently easier to integrate into a smart power process than bipolar transistors. For this reason, and the fact that smart power has the emphasis on the clever part(i.e. logic), if a low cost solution is needed then, MOS technology is the natural choice.

In order to develop a smart power process it is necessary to understand the limitations of low voltage processes. An initial investigation was made into the voltage limitations of an nMOS process. It was seen from diode structures that breakdown voltage is strongly effected by geometry. Sharp angles and thin conductors reduce the breakdown voltage and should be avoided in power structures. This is due to localised high electric field causing point breakdown, before the bulk breakdown level. It was also seen that the breakdown of diodes was dependent on doping level and junction depth. By using a 5 μm n-well doping as a diode a breakdown voltage of 200V could be obtained. This compares with 50 V for a diode with a 1 μm n+ drain diffusion. Therefore, for power transistors it is clear that the drains need to be formed using deeper junctions with lower doping concentrations than were used for low voltage transistors.

It was found that transistors broke down at lower voltages than planar pn junctions. This was mainly due to the influence of the gate electrode. Further, if the drain and source are sufficiently close, their depletion regions will merge. This causes punchthrough breakdown, rather than breakdown to substrate as in the planar junction case, lowering the maximum safe operating voltage. Also, it was found that the threshold adjust implant which increases peak doping concentration in the channel greatly reduces the transistor breakdown voltage for this process. This led to the conclusion that for low voltage transistors drain junction breakdown was gate assisted and limited by the channel doping. Further, it was found that this implant was not necessary for the low voltage devices and could be safely eliminated from future process runs.

As the goal of this work was complementary high-voltage transistors, the next step was to try and increase the safe operating voltage of both n- and p-channel transistors. This work was based on a CMOS process, the objective being to increase BV_{DSS} of both n and p-channel devices. To increase the breakdown voltage both channel and source/drain dopings were varied. There was obviously a direct trade-off with on-resistance, but breakdown voltages of around 40V for nMOS and -50V for pMOS transistors could be obtained without drastically altering the basic device structure.

Some variations in transistor structure were tried. It was seen that by using a lightly-doped region around the drain of the low voltage transistor, a similar breakdown voltage to that of a transistor with very lowly doped source and drain could be realised, without the drawback of increasing the transistor on-resistance. This simplistic LDMOS looked promising as a power transistor structure. The double-diffused DDMOS, where both the source and drain have the lightly doped region , had a similar type of performance, but had a larger device size. For this reason it was discarded as a possible solution.

In order to further increase breakdown voltage, the basic transistor structure needs to be changed. The region supporting the high potential must be separated from the region controlling current flow.

To speed up development time, various process and device simulators were calibrated with the results obtained on low voltage transistors and diodes. It was found that SUPREM III fitted both nMOS and CMOS processes well. For MOS transistors CANDE was the most suited, and for diodes and more complex structures such as a CMOS inverter, PISCES was chosen.

The smart power concept was then taken a step further. It was known that the power transistors would need to be structurally different from the logic case. As simulations were shown to model the low voltage process well, they were used to determine that the most suitable transistor for 100V smart power applications was the lateral double-diffused MOS, LDMOS. In contrast, above 300V VDMOS transistors were found to have a higher current density and therefore should be chosen in that range.

In order to physically separate the drift region from the channel region, a thick oxide was grown between the drain and the gate region. The thick oxide used was the field oxidation, therefore no additional process steps were needed. This oxide limits the influence of the gate electrode on the drift region by reducing the electric field potentials in the area where the gate edge meets the drift region. The lightly-doped region of the n-LDMOS was formed using the n-well. Again no additional steps are needed. It was possible to produce this type of transistor using the existing masking stages of a low voltage CMOS process. Further, if more than one power transistor were needed in the design, then as the design is of a self-isolating type, it is by far the most area efficient at this power level.

To implement complementary power transistors one more masking level was added in order to produce a lightly-doped p- region around the drain of the p-LDMOS. It was also necessary to alter the doping of the n-well. The junction had to be made deeper than the standard in order to avoid vertical punchthrough of the triple-doped drain of the p-channel transistor. This new well needed to have the same surface concentration as the original 5μm well to avoid degradation of the low voltage p-channel transistors, and to allow continued use of existing low voltage logic design libraries. A well of 8 μm was found to satisfy both the requirements of the p-LDMOS and the logic transistors.

These modifications gave rise to n-channel power transistors with BV_{DSS} of greater than 120 V and p-channel power transistors with BV_{DSS} in excess of -55 V. Further, it can be seen that if only n-channel power transistors are needed then the process reduces to only nine masking levels, exactly the same as the low voltage CMOS process.

The crucial factor in increasing the performance was the introduction of the field oxide between the drain and channel regions. This oxide helps to reduce the influence of the gate on the drift region, and results in a doubling of the safe operating voltage of the n-channel power transistors.

In order to improve on the p-LDMOS transistor performance, a change in device structure was necessary. Instead of the triple-diffused structure a double-diffused structure, as for the n-LDMOS was implemented. By using a p- epitaxial layer on n- substrate, it was possible to bias the p- region. The n- well was used this time as the channel region and the p- epitaxial layer was used as the

lightly-doped drift region. This device could be realised using the same nine masking layers as a CMOS process. Using a 25 μm epitaxial layer the p-LDMOS transistors could support 100 V. This performance was achieved using one less masking layer than the previous process and therefore results in a reduction in manufacturing costs, as well as an increase in breakdown voltage.

Additionally, by exploiting the RESURF technique, simulations showed that it was possible to increase the safe operating voltages even further, to 200 V for the n-LDMOS and -180 V for the p-LDMOS, by simply reducing the epitaxial layer to 20 μm. This represents an improvement of a factor of ten over the original low voltage transistor performances. This notable result was achieved by developing new transistor structures which are similar to discrete lateral double-diffused MOS transistors. The novelty lies in :

1) The fact that these new transistors use a field oxidation to physically and electrically separate the two regions of the power transistor.

2) The power transistors are easily integrated in a low voltage process.

3) The only additional processing step is the use of an epitaxial layer to realise the complementary power transistors. If only n-channel power transistors are required, no additional steps are needed.

This means that performance normally associated with much more complex processes [1,2,3] can be realised with only nine masking levels [4,5]. Therefore the benefits of smart power over separate power transistors and logic circuits, can be reaped for niche markets not usually exploited, such as ASICs and low volume runs. It also opens up new possibilities for products, such as memories, that with the addition of a few high voltage transistors for charge pumps and active protection could be used in the harsh environment of automotive applications. To summarise, with this type of process the market share of smart power could grow even faster than predicted.

References

[1] C.Petersen and A.R.Barlows, "High Voltage Circuits in standard CMOS processes", IEDM Tech. Digest, pp.77-80, 1982.

[2] C.Contiero, A.Andreini and P.Galbiati, "Multipower BCD - A Versatile technology from 60V to 250V that integrates Bipolar, CMOS and DMOS Devices", IEE Colloquium on Integrated Intelligent Power Devices, March 1989.

[3] Z.Parpia, C.A.T.Salama and R.Hadaway, "Modelling and Characterization of CMOS-compatible High-Voltage Device Structures", IEEE Trans. Electron Devices, vol.ED-34, pp2335-2343, 1987.

[4] M.R.Duncan, J.M.Robertson, R.J.Holwill and R.Rodrigues, " Design of CMOS-Compatible High Voltage Devices", ICSSICT, 2nd International Conference on Solid-State and Integrated Circuit Technology, October 22nd-28th 1989, Beijing, China.

[5] M.R.Duncan, J.M.Robertson, R.J.Holwill and R.Rodrigues, "CMOS-Compatible High Voltage Complementary LDMOS Devices", ESSDERC, 19th European Solid State Device Research Conference, September 11th-14th 1989, Berlin, Germany.