Edge Effects in Silicon IGFETs.

Thesis submitted by

James Arthur Serack

for the degree of

Doctor of Philosophy.

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

1988

Abstract.

With the trend for increased miniaturisation in integrated circuits the effect of edges on the d.c. electrical characteristics of insulated gate field effect transistors have become more important. Investigation of these edge effects has been impeded by the lack of suitable test structures. Construction of edge effect test structures requires a microfabrication technique with greater control of relative edge positions than that available with standard microfabrication process.

A general technique for that purpose, called the progressional offset technique, was developed during this research. It was applied to the construction of an array of incompletely gated field effect transistors using a custom 1 μm non-self-aligned metal gate enhancement NMOS process. The terms source gap transistors (SGTs), and drain gap transistors (DGTs), were used for the resulting transistors with gaps in gate-to-channel coverage on the source side, and drain side, of the channel.

The electrical characteristics of SGTs differ from the normally gated transistors (NGTs) with their increased threshold voltage, increased series resistance, reduced subthreshold performance, and flatter saturation current behaviour. DGTs have a drain-voltage-dependent subthreshold current, and an increased threshold voltage. DGTs with large gaps do not exhibit drain current saturation but have an extended linear region of operation. DGTs are also more sensitive to hot electron degradation. The drain voltage dependent subthreshold swing of DGTs was used to study drain depletion boundary motion and to extract the surface impurity doping concentration.

The thesis also contains a comprehensive review of silicon microfabrication techniques, suggestions for other uses of the progressional offset technique, and possible applications for incompletely gated transistors.

Declaration.

I declare that this thesis has been composed by myself and that all the work included in this thesis is entirely my own except where otherwise indicated.

James Arthur Serack

Acknowledgements.

I would like to thank my supervisors Prof. J.M. Robertson and Dr. A.J. Walton for their guidance throughout this Ph.D. program. I would also like to thank Mr. A. Gundlach and Dr. J.T. Stevenson for the many useful suggestions that they made. Thanks also to the staff of the Edinburgh Microfabrication Facility for their technical support.

The personal financial support by the Natural Science and Engineering Research Council of Canada and Bell Northern Research, and the indirect support by the Committee of Vice-Chancellors and Principals of the Universities of the United Kingdom, were greatly appreciated.

To my wife.

Marie Therese Northcote

Contents.

Chapter 1. Introduction.

1.1. Historical View.

We can not help but be aware of improvements in electronic technology and so should not require any justification for change. However, a closer look at the way the technology has changed will explain the increased importance of edge effects in IGFETs.

1.1.1. Motivation for improvement.

The motivation for improvement in electronics was at first driven by the desire to reduce the space required for a function to be performed. Success at that soon lead to motivation to increase the amount of functionality in a given area. As the area required for a function decreased, the speed of operation increased, and eventually speed of operation also gained importance as a motivator. In microelectronics the cost of a function is proportional to the area it takes, so the cost of a function fell dramatically as the size was decreased. Tables 1.1, 1.2, 1.3, and 1.4, summarise these trends.

Nand Gate Equivalent Circuits.		
Epoch	Circuit	Area (mm^2)
mid 1950's	Vacuum tube and resistors	2620.0
early 1960's	Discrete transistors and resistors	491.0
1970	Silicon Gate PMOS MSI	0.02
1978	LSI integrated circuit	0.0016
1986	VLSI integrated circuit	0.00088

Table 1.1, Area required for a standard function.[1]

In short, the overall market effect over time has been to provide, slightly larger integrated circuits, at slowly increasing costs, with immensely increased functionality. The net effect being a dramatic decrease in the cost of a function over time.

Transistor Area		
Year	Transistor	Area (μm^2)
1959	Discrete Planar Bi-polar	590000
1964	Bi-polar DTL IC	23000
1970	Bi-polar RAM IC	6600
1981	VLSI MOS IC	88
1986	VLSI MOS IC	42

Table 1.2, Transistor area.

Circuit Speed.		
Year	Process	Shortest Gate Delay (nS)
1972	Enhancement NMOS	14
1976	Depletion NMOS	4
1977	HMOS	1
1980	HMOS II	0.5
1986	CMOS	0.2

Table 1.3, Circuit Speed.[2]

Cost per bit of memory.		
Year	Memory Size	Cost (¢ US)
1973	256 Bit.	0.6
1975	1K Bit	0.25
1977	4K Bit	0.12
1979	16K Bit	0.06
1981	65K Bit	0.04
1984	256K Bit	0.004
1987	A character in a text book	0.003

Table 1.4, The cost of storing data.

1.1.2. Important Process Advances.

The dramatic improvements in the effectiveness of electronics owe a lot of credit to clever circuit improvements, however the underlying improvements in microfabrication technology still account for the majority of increased functionality per unit area. Table 1.5 summarises some important events in processing. The advances presented have been selected for impact on the basic transistor, the front end of the process, other equally impressive innovations have occurred in the interconnection technologies and lately in improved isolation techniques, both have impacted packing densities.

Processing Advances	
Year	Invention or Advance
1926	Surface Field Effect Transistor Proposed (Lilienfeld).
1948	Field Effect Demonstrated (Shockley + Pearson).
1952	Single Crystal Silicon Produced.
1953	Parasitic Surface FET channel identified in Bipolar Transistor.
1955	Insulated Gate FET structure proposed (Ross).
1959	Thermal SiO_2 for gate insulator proposed (Atalla).
1960	First Modern Surface FET (Atalla + Kahng).
1965	Sodium Contamination Problems Conquered.
1966	Self Aligned Silicon Gate Technology Demonstrated.
1968	Ion Implanted Self Aligned Metal Gate Attempted.
1970	LOCOS isolation applied.
1973	Ring Field Projection Lithography Invented (Markie).
1974	Ion Implanted Threshold Adjusted MOSFET's.
1979	10:1 Reduction Direct Step Lithography Applied.
1979	First Dry Processed Integrated Circuit.
1980	Reactive Ion Etching.
1980	Silicide Gates used.
1981	Rapid Thermal Annealing used.
1982	Gate Sidewall Spacer Technology used.

Table 1.5, A summary of important processing advances.[3-7]

1.1.3. Relative Dimensions and Scaling Failures.

It should not be too surprising that, with the motivation for decreasing the area required for a function, and the steady processing improvements previously listed, that the basic IGFET dimensions have decreased over time. Table 1.6 lists four basic dimensions; the gate length (L), the gate overlap of source or drain (L_{over}), the source and drain junction depth (X_j), and the gate insulator thickness (T_{ox}), for a number of processes.

Relative Dimensions.						
Year	L (μm)	L_{over} (μm)	L_{over} %	X_j (μm)	T_{ox} (Å)	Process
1969	20	4	40	2.5	1500	Al Gate PMOS
1970	10	1.8	36	2.5	1200	Si Gate PMOS
1972	6	1.4	47	2.0	1200	Enhancement NMOS
1976	6	1.4	47	2.0	1200	Depletion NMOS
1977	6	0.81	27	1.20	1000	NMOS
1977	3.5	0.6	34	0.8	700	HMOS
1980	2	0.4	40	0.8	400	CMOS
1982	2.5	0.25	20	0.4	450	CMOS
1983	1.3	0.25	38	0.4	300	CMOS
1986	1.1	0.05	9	0.35	250	NMOS (Sidewall Spacers)

Table 1.6, Relative Transistor Dimensions.

Similar decreases in the width of transistors, as shown by table 1.7 have been supported by the same technology.

LOCOS Isolation.		
Epoch	Minimum Width (μm)	Transition Region Proportion (%)
late 1960s	20	8
early 1970s	10	15
mid 1970s	6	25
late 1970s	3	50
early 1980s	2	75†
mid 1980s	1.5	100†

Table 1.7; Increase in width edge effects.

The reduction in length, gate overlap, and transistor widths have been following simple scaling laws developed to try to minimise the departure of transistor's electrical characteristics from that of the larger transistors. Various criteria exist for how the scaling should be done. One of the most common, summarised in table 1.8, is based on keeping the internal electric field magnitudes constant. Table 1.9 shows that only some aspects of the scaling laws have been adhered to. The penalty for the scaling exceptions have been a departure from large transistor behaviour. The systems constrained power supply of 5 volts results in high electric magnitude field problems, such as punch though and hot electron effects‡.

There are however other features which do not scale. They are almost always associated with edge effects. The built in potential of the source and drain, fringing capacitance off of the gate structure, and subthreshold current dependence on drain bias are some features that do not scale‡.

---

†Variations on the LOCal Oxidation of Silicon (LOCOS) isolation process such as Sealed Interface LOcos (SILO), and SideWAll Masked Isolation (SWAMI) address this problem. Completely different techniques such as Buried OXide (BOX) isolation, and dielectric refilled trenchs etched into the silicon (TRENCH) isolation allow for denser circuitry and no birds beak (transition zone). They have been successfully used for dense DRAM memory production, but are not without their own edge effects.

‡ See chapter 3 for definitions.

Scaling Laws.	
Transistor Dimension	Scaling Factor
Width, Length, Insulator Thickness	$\frac{1}{K}$
Doping Concentration	K
Voltage	$\frac{1}{K}$
Power	$\frac{1}{K^2}$
Power delay product	$\frac{1}{K^3}$

Table 1.8, Scaling Laws for Constant Electric Fields.

Scaling Practise.			
Feature	1970	1983	
	Real	Scaled	Real
Length (μm)	10	1.3	1.3
Tox (Å)	1200	156	300
Junction Depth (μm)	2.5	0.33	0.4
Overlap (μm)	1.8	0.23	0.25
Drain Voltage (V)	12	1.56	5
Speed Power Product (pJ)	20	0.04	0.25

Table 1.9, An example of the scaling practise.

As the main dimensions of the transistors decrease, the "fixed" contributions of edge effects become more predominate. They can no longer be considered negligible as they once were in larger transistors. It is therefore important to study edge effects to appraise the magnitude of the effects and develop techniques for characterising them.

1.2. Thesis topic definition.

The topics of most Ph.D. thesis lie at the edge of active research in their areas. This thesis is no exception, it is literally about the edge of active areas in the silicon insulated gate field effect transistor (IGFET).

Before proceeding to discuss edge effects in detail, it is best to properly define the topic. What is considered an edge?, what are important edges in an IGFET?, and why bother studying them?, are all valid questions.

1.2.1. What is considered an edge?

There are many dictionary definitions of edge. [8] Some of the applicable ones are; the meeting-line of two surfaces of a solid; the narrow surface of a **thin object;** and the boundary line or surface of a **region.** Modern IGFETs are composed of **thin films** of conducting materials formed over the surface of a single crystal of silicon containing **regions** of impurities. It is appropriate then, to apply the last two of these definitions to the IGFET.

1.2.2. Some important region edges.

A closer look at the actual structure of an IGFET will reveal how these definitions apply. Figure 1.1 shows an IGFET cross section, both in length, revealing the source, channel, and drain, and in width, showing the drain and isolating oxide region.

Channel to Source/Drain Boundaries.

The interface between the source and channel, and the channel and the drain are very important to the operation of an IGFET, since that is the direction of current flow. The edges involved between these regions are between areas of differing charge carrier concentration. The differences in concentration come about from variations in both impurity atom concentration in the crystal, and internal electric field magnitudes. The electric fields result from the relative voltages applied to the crystal bulk, the source, the drain and the gate. The degree of the gate edge overlap to the channel edges can effect the internal electric fields and hence the transistor operation. These

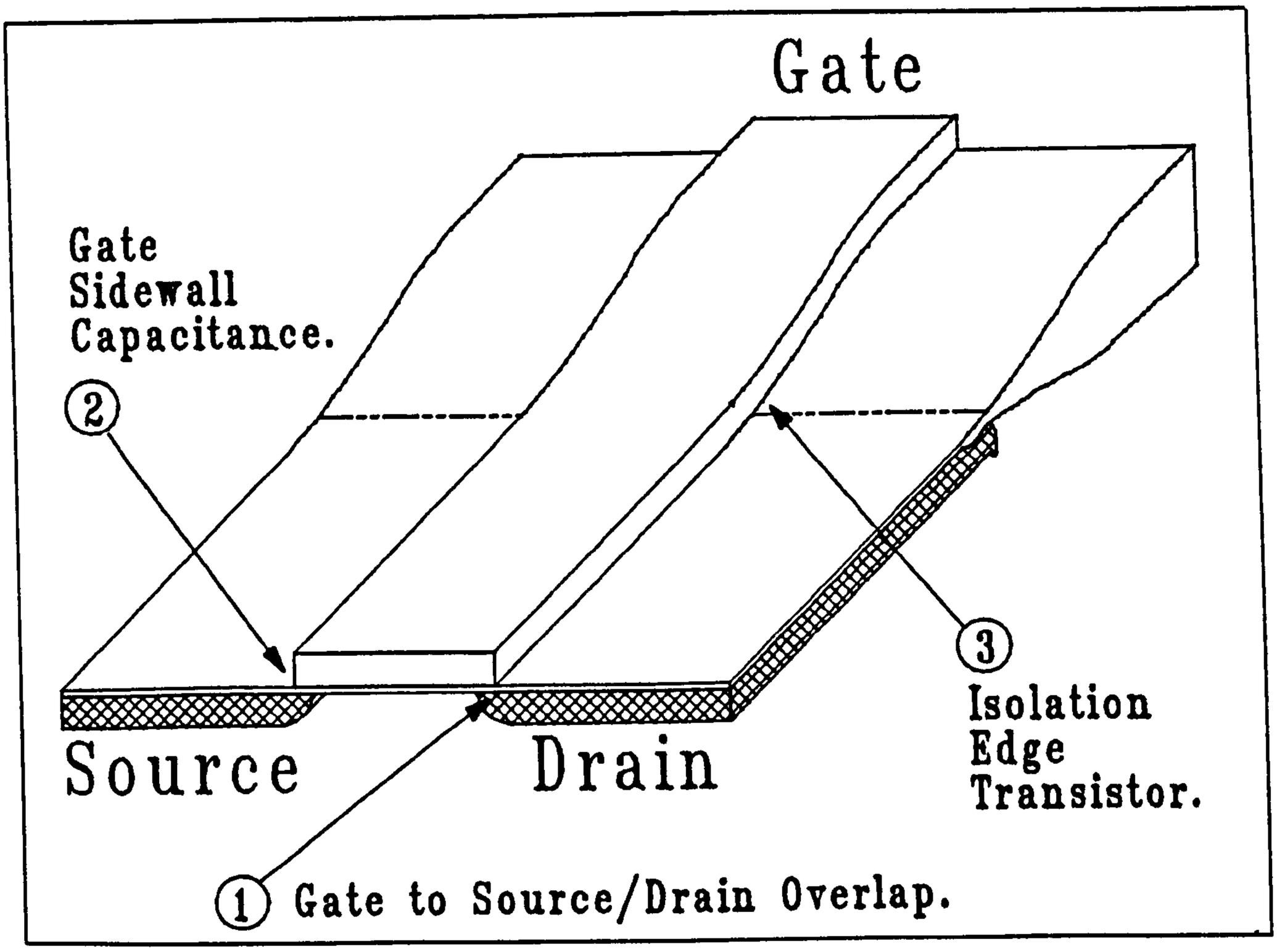


Figure 1.1, Important Edges in an IGFET.

edge effects are important to operation of transistors of any size.

Gate Sidewall.

The edge of the gate electrode, or gate sidewall, becomes important when the capacitance of the gate, and threshold voltage of the transistor are considered. The fringing electric field from the edges of the gate results in increased gate capacitance which can lead to poorer transistor switching speeds. Gate sidewall capacitance is of greater importance in transistors with shorter channel lengths.

Channel to Isolation Region Boundary.

The channel to isolation region boundary edge is important since the width of the transistor is one of the variables which influences the magnitude of the current flow. Any uncertainty in the position of the edge of the channel, and hence the transistor width, will obviously effect the predicted magnitude of current. The isolation region

edges of the transistor also suffer from variations in vertical electric field magnitude due to the transition to a thicker insulator. Operation at the edges is therefore different to the rest of the channel. Width effects are of greater importance in narrower channel transistors.

1.2.3. Importance for Study.

The question of "why bother studying edge effects?", has two answers. Firstly, information about edge effects can be important to process designers when considering trade-offs between choices in methods of fabrication. Secondly, an understanding of edge effects could also lead to novel device design.

When a process designer is considering how to achieve a certain structure in a new process, he or she must usually decide between a number of techniques, each of which has advantages and disadvantages. Often the designer will identify a difference between two techniques but have no information on how side effects of the technique will influence transistor operation. Recent advances in computer simulations of structures can give useful answers to the designers questions, but direct experimental research often causes questions to be asked which would not arise during simulations. The answers to these questions can have a profound influence on the choices the designer must make.

Experimental research can also lead to the invention of novel devices. The bipolar transistor control mechanism of minority carrier injection was discovered from unaccountable errors in a field effect experiment. It is unlikely that minority carriers would have even been included in a simulation program, and an important opportunity missed.

So at least two good reasons for bothering with direct experimental work exist. The experiments deal with the complete set of physical phenomenon and allow for insights which might lead to novel structures.

1.3. Rational for Thesis Content.

Now that the research topic of edge effects has been defined, it is worthwhile taking a short overview of the thesis content to see how this subject has been addressed. The thesis has been divided into seven chapters excluding this introduction and the appendix.

Processing.

Microfabrication process improvements have been the key to continued advances in miniaturisation. They are also the key to the internal structure of insulated gate field effect transistors (IGFETs). Since the internal structure of an IGFET is important to its electrical characteristics, it follows that a detailed knowledge of processing is essential for understanding the resulting electrical edge effects. The chapter on processing gives a comprehensive review of silicon processing techniques.

IGFET Physics.

A review of the physics behind transistor operation is also important to the understanding of edge effects. The chapter on IGFET physics is a review of the important principles of operation and applicable theory for edge effects.

Progressional Offset Technique.

Creating test structures for the edges of minimum dimension transistors, seems to necessitate fabricating structures smaller than the minimum possible dimensions. However, there is a way of trading off maximum die yield for the maximum potential for knowledge. Chapter four describes a progressional offset technique, developed during this research, which allows suitable structures to be built.

Misaligned Gate Experiment.

The misaligned gate experiment applies the progressional offset technique to the study of gate-to-channel overlap. Chapter five describes the design of the experiment. Included are; the design of a non-self-aligned small geometry metal gate process, the design of a test chip using the progressional offset technique, and development of a test

program for the resulting transistors.

Asymmetrical Transistor Characteristics.

Analysis of the resulting asymmetrical transistor characteristics is the topic of the sixth chapter. It also contains computer simulations of the asymmetric transistor operation to aid in understanding internal operation. The simulated and experimental DC characteristics are presented along with experimental evidence for heightened hot electron degradation in asymmetric transistors.

Drain Depletion Motion.

Further analysis of the asymmetric transistors as test structures for studying drain depletion region motion is presented in chapter seven. In that chapter the drain voltage threshold effect of Drain Gapped Transistors (DGTs) is used to evaluate drain depletion motion and surface impurity concentrations. Simulations are again used along with experimental data to highlight internal operation.

Conclusions and Further Research.

The concluding chapter also suggests where effort could be beneficially applied to continued research. It contains proposals for other applications of the progressional alignment technique, and ideas for how asymmetric transistors might be used for environmental sensors, analogue circuits, and as a technology for faster digital switches.

Finally, a word about changes during the course of this study. The metal gate process described in chapter five was designed using some computer process simulations. During the analysis of the resulting asymmetric transistors, (about 18 months later), a much improved package of software became available. The advanced software was then employed on the simulation of the internal operation of those transistors and for the process simulation of the example process in chapter two.

1.4. References

1. Forester, T., "The Microelectronics Industry.," in *The Microelectronics Revolution.*, ed. Forester, T., Basil Blackwell, Oxford, 1980.

2. Millman, J., *Microelectronics, Digital and Analog Circuits and Systems.*, McGraw-Hill Book Company., New York, 1979.

3. Seidel, T.E., "Ion Implantation.," in *VLSI Technology.*, ed. Sze, S.M., McGraw-Hill Book Company., New York, 1983.

4. Colclaser, R. A., *Microelectronics, Processing and Device Design.*, John Wiley & Sons., New York, 1980.

5. Parrillo, L.C., "VLSI Process Integration.," in *VLSI Technology.*, ed. Sze, S.M., McGraw-Hill Book Company., New York, 1983.

6. Glaser, A.B. and Subak-Sharpe, G.E., *Integrated Circuit Engineering.*, Addison-Wesley Publishing Company., London, 1977.

7. Kahng, D., "A Historical Perspective on the Development of MOS Transistors and Related Devices.," *IEEE Transactions on Electron Devices*, vol. Ed 23., no. 7, pp. 655-657, IEEE, July 1976.

8. *The Concise Oxford Dictionary*, Oxford University Press, Oxford, 1979. Ed. 6

Chapter 2. Silicon Processing.

2.1. Introduction.

Like all other man-made objects, the suitability, performance, and reliability of each transistor in an integrated circuit depends both on the materials from which it is made and the manner in which it is fabricated. It is therefore essential to understand silicon processing in order to fully appreciate transistor operation.

Literally hundreds of integrated circuit processing functions have been developed over the last three decades, which have been combined to produce hundreds of complete microfabrication processes. The interaction between the functions results in each process's characteristics topography. That high degree of interaction creates a difficulty in describing silicon processing. Although it is the whole of each process that makes a unique product, simply describing a single process does not adequately cover the details of the component funtions and considering only the detail of the functions neglects their interaction in a process.

Some texts on the subject describe an entire process as an introduction to semiconductor processing and others present great depth on a single technique but neither imparts an understanding of the variety of processing methods available today. Although there was a text which treated the subject with both generality and some depth,[1] the rapid development of new processing functions has made it incomplete. The approach taken here is to present a reasonably comprehensive collection of production processing techniques and then to illustrate their interaction through an example process.

Processing functions can be broadly grouped into seven classes, six of which directly affect the device structure. The next sections will cover the production of starting materials, addition of secondary layers, pattern transfer, selective removal of material, change of layer composition, and change of layer shape. The last section before the process example will describe the final class which provides support functions such as measurement, cleaning and gettering.

2.2. Production of Starting Material.

Integrated circuit microfabrication requires single crystal wafers of high purity and uniform doping as a starting material. Nearly all source wafers are made using the method described in this subsection.

The raw material for silicon wafers is the common sand quartzite (SiO_2). Although the source material is abundantly available it requires much refinement before it is suitable for integrated circuit fabrication.

In the first step of refinement the Quartzite is mixed with various forms of carbon, such as coal, coke, and wood chips, in a submerged-electrode arc furnace. The carbon combines with the oxygen to form volatile gasses in the 2000°C melt leaving molten silicon behind. The process requires about 13 kWh of energy per kg of silicon produced.[2]

The solid silicon produced by that reaction is called metallurgical grade silicon (MGS) and is about 98% pure. MGS is used for aluminium alloys, silicone chemicals, transformer steel, and a semiconductor source material. Further refinement is necessary for electronics and the MGS is pulverised and treated with hydrochloric acid to form the room temperature liquid trichlorosilane. It is then fractionally distilled before being subjected to a hydrogen reduction reaction which produces solid electronics grade silicon (EGS) around a resistance-heated silicon nucleation rod. It takes several hundred hours to deposit 200 mm diameter rods.[3] The impurity level in EGS is measured in parts per billion. In 1982 over three million kilogrammes of EGS was produced in the world using this method.

The EGS must be converted into single crystal wafers before it is useful for semiconductor fabrication. There are three possible methods of growing single crystal ingots. They are Czochralski crystal growth, float zone growth, and the Bridgeman Technique.

The crystal growth method, shown in figure 2.1, was named after its 1917 inventor Czochralski and is responsible for 80 to 90% of the silicon crystals prepared for electronics.[1,2]

The crystal growth is initiated by dipping the edge of a 5 to 8 mm diameter seed crystal of either <100> or <111> orientation, depending on the desired ingot

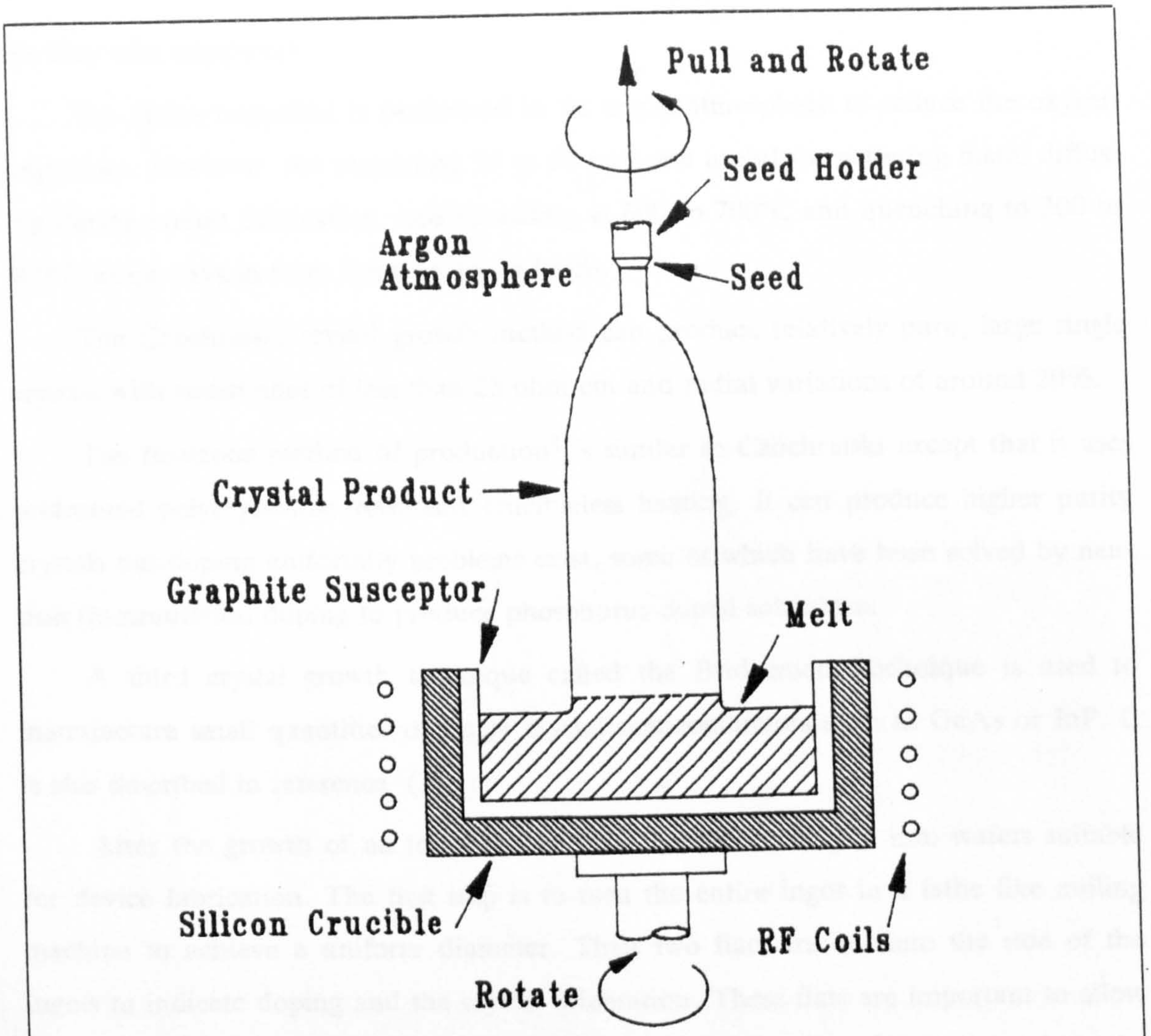


Figure 2.1. Czochralski Crystal Growth. (After2).

orientation, into a radio frequency (RF) heated crucible containing a melt of the EGS and any intentional impurities at just over 1412°C. As seed crystal and melt are rotated in opposite directions, in order to mix the melt and reduce temperature non uniformities, the seed crystal is slowly pulled from the melt allowing growth from the seed into a ingot crystal.

Today, microcomputers control the temperature, rotation, and pulling rates so that uniform crystals of 10 to 15 cm diameter and typically 1 metre long can be produced.[4] Microcomputer control also allows programmable pull and rotation rates so that an initial necking or thinning of the ingot can be preformed to reduce stress and allow defect free crystals to be grown, which was discovered by Dash in 1958 (without

the help of a computer).

The entire operation is performed in an argon atmosphere to reduce the oxygen impurities. However, the remaining 10 to 50 ppm are useful for gettering metal diffusing during circuit fabrication, and annealing at 600 to 700°C and quenching to 300 to 400°C keeps oxygen from forming donor levels.

The Czochralski crystal growth method can produce relatively pure, large single crystals with resistivities of less than 25 ohm-cm and radial variations of around 20%.

The floatzone method of production[3] is similar to Czochralski except that it uses preformed polycrystalline rods and crucibleless heating. It can produce higher purity crystals but doping uniformity problems exist, some of which have been solved by neutron transmutation doping to produce phosphorus doped substrates.

A third crystal growth technique called the Bridgeman Technique is used to manufacture small quantities of single crystal semiconductors such as GaAs or InP. It is also described in reference (3).

After the growth of an ingot is complete it must be shaped into wafers suitable for device fabrication. The first step is to turn the entire ingot in a lathe like milling machine to achieve a uniform diameter. Then two flats are cut into the side of the ingots to indicate doping and the crystal orientation. These flats are important to allow later alignment of the product die to the cleavage plane. Figure 2.2 shows how the wafers can be identified by their primary and secondary flats.

After the ingot shaping, wafers are sliced from it using a diamond saw. The slicing determines the surface orientation, thickness, taper, and bow of the wafers. Both sides of the wafers then have their edges rounded and the surfaces are lapped using a mixture of Al_2O_3 and glycerine to produce a typical flatness of around $\pm 2\mu m$. The damage and contamination caused by lapping is removed by etching 50 to 80 μm off the surfaces in an isotropic etchant. (HF, HNO_3, CH_3COOH).

The front surface must be polished further to achieve the optical flatness required for lithography. That is accomplished by mounting the wafers on carriers and buffing them against first a coarse and then fine mat while using a colloidal suspension of fine SiO_2 in an aqueous alkaline solution as a polishing compound.

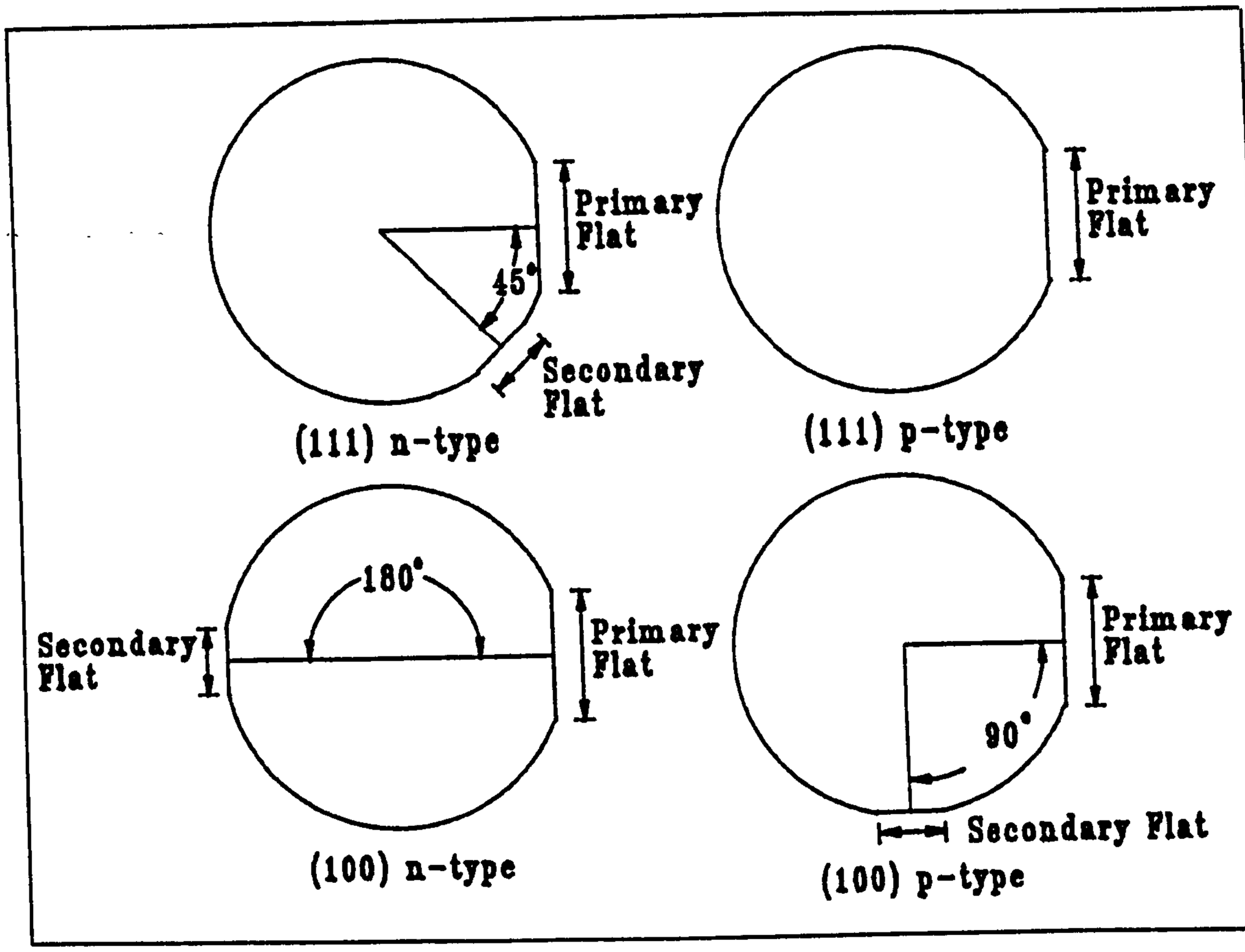


Figure 2.2. Wafer identification.

Finally the wafers are cleaned using solvents, mechanical scrubbing with detergents, multiple de-ionised (DI) water rinses, and drying in warm nitrogen.

2.2.1. Section Summary.

The production of wafers for integrated circuit fabrication, from the refining of raw material to the shaping and cleaning of wafers, has been described in this section. The following section describes how these wafers can be used as substrates to support the thin films that makeup integrated circuits.

2.3. Addition of Layers.

Secondary layers can be formed on the bulk semiconductor to act as insulators, conductors or merely as a masking material for some other processing step. It is also possible to provide secondary semi-conducting layers.

There are two distinct classes of films formed on the surface. They can either be grown from the surface or deposited onto the surface. Grown films include epitaxial growth of semi-conducting layers and oxidation of the silicon to produce silicon-dioxide insulators. Other insulators, metals, and amorphous semi-conducting layers are deposited on the surface either through a chemical reaction or condensation at the surface or by precipitation. These techniques are covered in the following sections.

2.3.1. Epitaxy.

The growth of a thin layer of semiconducting crystal on a monocrystal substrate by epitaxy,[1,2,5-8] literally meaning "arranged upon", allows the construction of crystal substrates with layers having drastic variations in composition.

The ability to produce these buried layers within a crystal structure is essential to the manufacture of bipolar (for which it was developed), optical, and III-IV compound semiconductor processes. Epitaxial processing, is also applied to planar MOS integrated circuits, although it is not necessarily required. It can be used to form abrupt wells, merged bipolar-CMOS processes, and when grown on a non-silicon insulating substrate a more versatile CMOS process.

There are three types of epitaxy, chemical vapour epitaxy (which is the most common), liquid phase epitaxy, and molecular beam epitaxy.

Chemical Vapour Epitaxy.

Chemical Vapour Epitaxy (CVD) is carried out in a closed vessel called a reactor, into which reactant gasses of carefully controlled ratios are introduced. There are three common reactors defined by either horizontal, pancake, or barrel shaped wafer susceptors which are described in the literature.[2,4] All use electrically conducting graphite susceptors coated with either boron-nitride, silicon-carbide, or quartz and heated by RF induction to temperatures about 50 to 70% of the melt growth temperature of the crystal. Since residual oxides are a primary cause of epitaxial film defects, in situ cleaning[5] by $HCl - H_2$ gas etching at 1200°C is used prior to the actual epitaxy.

Three liquids, silicon tetrachloride ($SiCl_4$), dichlorosilane (SiH_2Cl_2) and trichlorosilane ($SiHCL_3$) and the gas silane (SiH_4) are common silicon sources for

epitaxy. The liquid sources are transported into the reactor by passing hydrogen gas through or over a bath of the liquid and then into the reactor. Silicon tetrachloride is the most common source, since it has been the best studied. Equations 2.1, 2.2, and 2.3 describe the hydrogen reduction and pyrolysis used to deposit silicon on the wafer surface.

$$SiCl_{4\,(gas)} + 2H_{2\,(gas)} \xleftarrow{1250\,^{\circ}C} Si_{(solid)} + 4HCl_{(gas)} \tag{2.1}$$

$$SiH_2Cl_{2\,(gas)} \xleftarrow{1150\,^{\circ}C} Si_{(solid)} + 2HCl_{(gas)} \tag{2.2}$$

$$SiH_{4\,(gas)} \xrightarrow{1000\,^{\circ}C} Si_{(solid)} + 2H_{2\,(gas)} \tag{2.3}$$

Dopant gasses such as diborane (B_2H_4), phosphane (PH_3) or arsine (AsH_3) can be incorporated into the gas flows to produce doped epitaxial films. Sometimes hydrogen gas is added to the reactions in equations 2.1 and 2.2 to inhibit the reverse silicon etching reaction.

Since the deposition rate is a complex interaction of surface chemistry and mass transport, chemical vapour epitaxy is difficult to model but still must be carefully controlled. As the lattice association of epitaxial silicon atoms takes a finite amount of time, deposition at rates higher than the lattice association rate can result in polycrystalline films or film defects.[9] Typical industrial growth rates are about 1 μm per minute.

Liquid Phase Epitaxy.

Liquid phase epitaxy is epitaxial growth directly from contact with a crystal melt. It is used for thin layers formation in III-V compound growth,[2] which find application mainly in optical device production.

Molecular Beam Epitaxy.

Molecular-Beam Epitaxy (MBE) is an epitaxial process involving the reaction of one or more thermal beams of atoms or molecules with a crystalline surface under ultrahigh vacuum conditions ($approx\,10^{-10}Torr$).[2] The distinguishing features of molecular beam epitaxy are; great control of the doping profile, a wide choice of dopants, ultrahigh purity, low temperature processing, in situ cleaning, and sophisticated in situ

analysis.

The equipment of an MBE system[9] consists of an ultra high vacuum chamber, a substrate holder-heater, Knudsen Cell effusion ovens (where dopants are heated to a sufficient temperature that simple kinetics cause them to be emitted from the cell), electron beam evaporators to cope with higher melting point materials such as silicon, inert gas sputter cleaning sources, low voltage (1 to 2KeV) ion implanters and finally sophisticated analytical techniques to allow in situ analysis.

Since MBE is performed under ultrahigh vacuum, the mean free path of epitaxial molecules is very long and deposition rates can be controlled by source parameters. MBE has none of the lattice association time constant or mass transport problems of chemical vapour epitaxy thereby allowing very complex processes to be run with predictability. Growth rates vary from 60 to 600 $\mathring{A}\ minute^{-1}$ [8] and although throughput is 10 to 15 times slower than CV epitaxy the doping control is extremely precise.

There are two doping methods possible with MBE, one is to provide co-flux of dopant and substrate atoms, and the other is to use in situ low energy ion beam implantation concurrent with epitaxy. In situ cleaning is performed by inert ion sputtering followed by a short anneal to restore surface crystal order.

In situ analysis can be performed by Auger spectrometry, secondary mass spectrometry or electron reflection-diffraction and the results can be used to control the process.

2.3.2. Oxidation.

Silicon dioxide is an insulator with excellent electrical and mechanical properties.[7] It makes the metal-oxide-semiconductor (MOS) processes possible. Silicon dioxide (SiO_2) is used as; a high quality gate insulator, the field insulator, a diffusion mask, a passivation layer, a stress relief layer, and as an adhesion promoting layer. Several techniques exist to produce oxides; anodisation, chemical vapour deposition, and rapid thermal oxidation; but it is thermal oxidation in conventional furnace tubes that produce the highest quality films.

Thermal Oxidation.

Thermal oxidation of silicon is the most widespread of all the integrated circuit fabrication techniques and has been extensively studied.[10] Conventional furnace tubes, as shown in figure 2.3, which are usually constructed from quartz and radiantly heated to between 800°C and 1200 °C, are used for oxidation at pressures around 1 atmosphere. Wafers are held vertically, (and parallel to the gas flow), in a "boat". The "boat", usually made of quartz, is slowly inserted into the center region of the tube, either resting on the tube itself or supported by a canti-levered paddle in automated systems, where the temperature is controlled to within ± 1°C. Mass flow controllers precisely control the ratios of gas flows (total velocity is 1 to 2 $cm\ sec^{-1}$). Microcomputers connected to the mass flow controllers, temperature controllers and auto-loading systems, manage the necessary interactions to meet the process recipe.

Two of the main oxidation reactions are;

$$Si_{(solid)} + O_{2\,(gas)} \rightarrow SiO_{2\,(solid)} \qquad (2.4)$$

and,

$$Si_{(solid)} + 2H_2O_{(vapour)} \rightarrow SiO_{2\,(solid)} + 2H_{2\,(gas)} \qquad (2.5)$$

Equation 2.4 describes the oxidation during processing called dry oxidation, since there is no water vapour present. Note that silicon from the wafer surface is consumed in the reaction so the silicon interface moves into the wafer at 0.44 times the oxide thickness. Equation 2.5 describes the reaction during wet oxidation. The water vapour is usually added to the furnace tube by pyrolysis of oxygen and hydrogen at the mouth of the tube, or through a gas bubbler.[11] The wet oxidation rate is considerably higher than for similar conditions of dry oxidation. The resulting oxide is also less dense because as the H_2 diffuses away from the interface and to the surface it can form a hydroxyl group (OH) in place of an oxygen atom in the oxide.[2] The resulting oxide is weaker, more porous and as a result less dense.

Other additives such as HCl provide important improvements in oxide quality. The halogen improves the oxide-silicon interface quality while the chlorine getters impurities from the silicon by converting them to volatile chlorides. The combination results in reduced mobile oxide charge, and improved carrier lifetime and breakdown

voltage. Some chlorine also remains in the oxide, which can react with mobile sodium to form immobile compounds, to protect oxide quality after processing. HCl also increases the oxidation rate.

The thermal oxidation rate is effected by temperature, ambient gasses, crystal orientation, and impurity concentrations. The main thermal oxidation model is the Deal and Grove model.[3,8]

$$\frac{Zo^2}{B} + \frac{Zo}{B/A} = t + \tau \qquad (2.6)$$

Where;

Zo is the oxide thickness,

t is the oxidation time,

B/A is the linear rate constant,

B is the parabolic rate constant, and

τ is the initial oxide thickness as an equivalent time.

It was developed by considering the gas-phase mass-transfer of the oxidant from the ambient into the oxide. Then the flux of the oxidising species through the oxide was considered to occur by diffusion and finally the oxidation rate was assumed proportional to the oxidant concentration at the interface. The result (equation 2.6) is the well known linear-parabolic relationship[10] of oxide growth verses time. It has provided accurate predictions for both wet and dry ambients within ranges of; temperature between 700 to 1300 °C, pressure from 0.2 to 1.0 atmospheres, and oxide thickness between 250 and 20000Å. Oxides thinner than 250Å, which are becoming popular, cannot be modelled with equation 2.6 because the linear rate constant (B/A) only becomes established after that thickness. Current theories suggest that charged oxidants and internal electric fields in the oxide are important in oxidant transport in oxides thinner than 150Å, explaining the increased initial oxidation rate and the failure of the Deal-Grove (diffusion transport) model. Modern models also consider the effect of point defects and defect migration on oxidation rate.

The oxidation of polysilicon is difficult to model since the oxide growth depends not only on the growth method and temperature but also on the doping concentration, grain size and morphology.

Another consideration during oxidation is the segregation of dopants during the high temperature processing.[6] As the silicon is consumed, an electro-chemical potential balance across the $Si-SiO_2$ boundary, is maintained by the displacement of dopants. The diffusivity and activity coefficients of the dopants and the rate of the $Si-SiO_2$ boundary advance effect the result. Boron has a segregation coefficient [10] less than one resulting in a high concentration on the oxide side of the interface and depleted surface concentration in the silicon. Phosphorus, however, has a segregation coefficient greater than one resulting in an increased phosphorus concentration on the silicon side of the interface.

Anodisation.

The oxidation of silicon through anodisation depends on electrical transport [7] of oxygen through the silicon dioxide to the silicon surface whereas thermal oxidation depends on diffusion transport of the oxidant. Two types of anodisation exist; wet, and plasma.

Wet anodisation, where the wafer is submerged in an electrolytic bath and connected to a constant current source, suffers from poor uniformity and limited maximum thickness. It is mostly of historic interest.

Plasma anodisation[10] however has been shown to produce reasonably thick films ($\approx 1\mu m$) at rate of up to 1μm an hour. Interest in plasma anodisation stems from attempts to reduce high temperature processing in order to limit impurity diffusion and through the advantage it has in selective oxidation without the bird's beak formation problems inherent to LOCOS (local oxidation of silicon) processing.

Plasma anodisation occurs when wafers are exposed to an oxygen plasma discharge and biased positively relative to the plasma so that the active charged oxygen species collect at the silicon surface. Plasma density, the substrate temperature and substrate doping all effect the anodisation rate. The mechanism is not well understood but theories[10] suggest that oxygen moves through the oxide towards the interface and oxygen vacancies move from the interface to the oxide surface. However a recent experiment [12] using O^{18} and nuclear microanalysis has shown that the motion of the oxygen species is the major factor. Other work[13] has shown that the addition of

chlorine has a similar effect as in thermal oxidation and that the resulting films have good physical properties and similar electrical properties to thermal films. However there are still problems to be overcome before plasma anodisation becomes a production tool.

Rapid Thermal Processing.

Modern thin gate-insulator (<300Å) processes use low temperature processing in order to increase the controllability of oxide thickness by extending the processing time. However, good electrical qualities such as low oxide charge and low interface state densities require higher temperature processes. Rapid Thermal Processing (RTP), where single wafers are exposed to a bank of tungsten-halogen lamps for a few tens of seconds[14] producing surface temperatures between 1100 to 1200°C, has been used to produce controlled film thicknesses with good physical qualities and suggestions of equivalent electrical qualities to conventional thermal oxidation.

2.3.3. Chemical Vapour Deposition.

Oxides, as well as other thin films, can be deposited on the surface of wafers through a chemical reaction at the surface of the wafer. This technique is called Chemical Vapour Deposition or CVD for short.

There are a number of variations on this technique such as; Low Pressure Chemical Vapour Deposition (LPCVD), Plasma Enhanced Chemical Vapour Deposition (PECVD),[15] and Metal Organics Chemical Vapour Deposition (MOCVD). With the exception of PECVD most chemical vapour deposition techniques make use of a standard diffusion furnace tube, shown in figure 2.3, as the basis for the equipment. Some lower temperature processes, such as passivation, make use of a horizontal reactor where the wafers are processed on a continuous feed hot-plate type reactor where the reaction occurs above the wafer surface.[16] The LPCVD equipment, also shown in figure 2.3, is a CVD tube fitted with a vacuum pump and load door to allow lower pressure processing. MOCVD uses either CVD or LPCVD equipment. PECVD is performed in a parallel plate plasma reactor at relatively low temperatures and pressures.

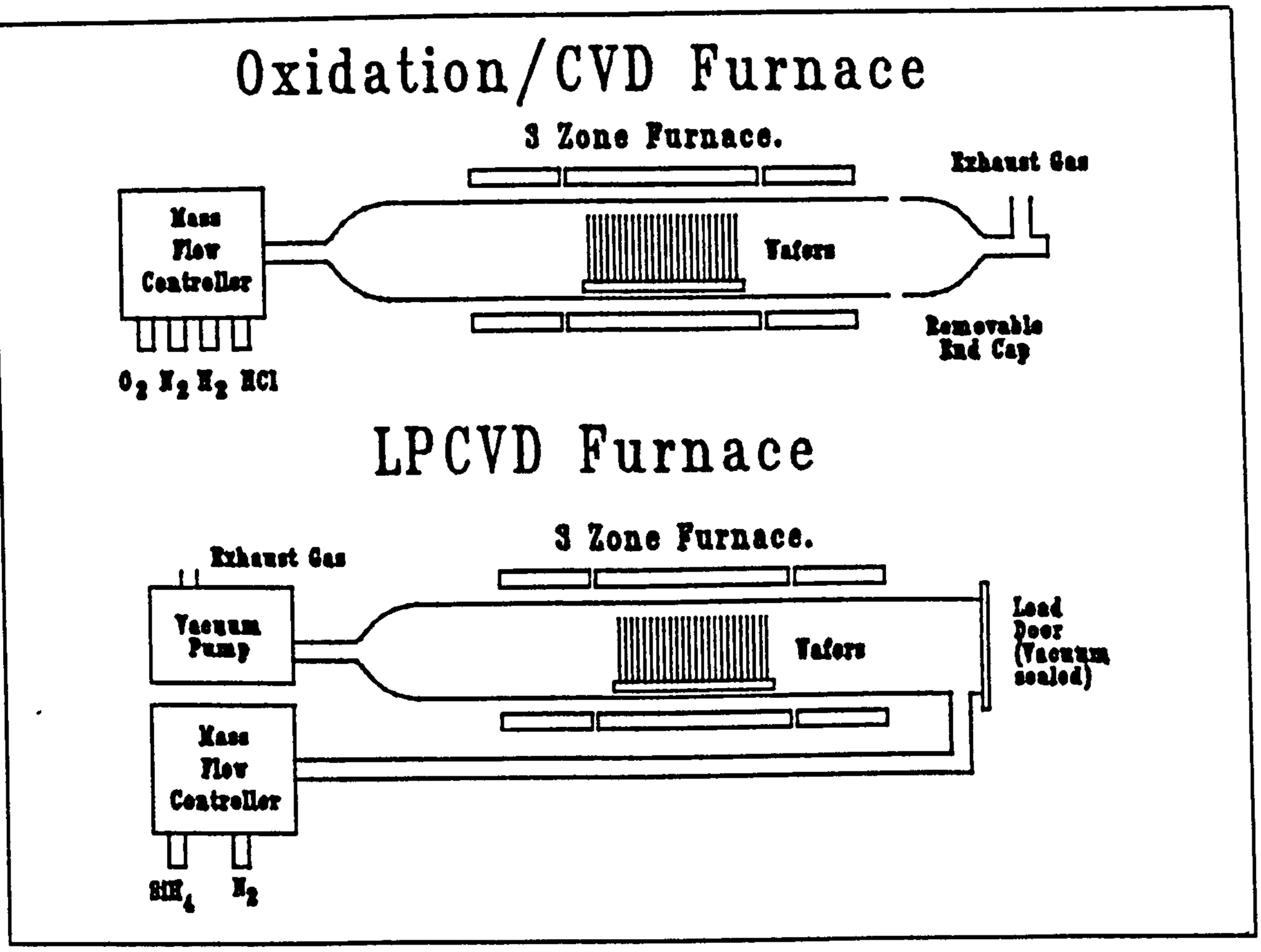


Figure 2.3. Standard Furnace Tubes.

CVD, LPCVD, and MOCVD work through endothermic reactions on the surface of the wafers driven by the thermal energy of the wafers. The gasses in the ambient react on the hot wafer surfaces (also on the hot tube walls) forming a film that bonds to the wafer. Other products of the reaction remain volatile and are removed from the ambient in the exhaust gas flow. In PECVD most of the energy to drive the reaction is provided by the plasma,[15] so substrate temperatures can be lowered to between 100 and 400 °C[17] resulting in less dopant diffusion and allowing the films to be deposited over low melting temperature materials such as aluminium. PECVD nitride films for passivating integrated circuits to be bonded in plastic packages has given them nearly as good of reliability as those in hermetically sealed packages, and at a fraction of the cost.

Although most films that are deposited by LPCVD could also have been deposited by CVD, LPCVD enjoys a number of advantages although the equipment is more sophisticated.[3] In comparison to CVD LPCVD has; increased uniformity of deposition,

increased through-put, significantly lower carrier gas volumes, and slightly lower processing temperatures, thus giving LPCVD a large economic and slight capability advantage over CVD.

A number of important dielectric films are deposited by CVD or LPCVD, MOCVD has seen some use in depositing conductors, and PECVD is used for passivation films. Some of the materials deposited are polycrystalline silicon (polysilicon), silicon-nitride, and both phosphorus doped and undoped silicon-dioxide. Metals such as aluminium, and tungsten can also be deposited. The chemistry for each material differs and the more popular reactions for each material are presented in the next subsections.

Polysilicon.

Polysilicon is a very common material for the gate conductor of insulated gate field effect transistors (IGFETs) as well as being used as a signal carrying conductor. It is most commonly produced in a LPCVD reactor with good uniformity in batches of up to 150 wafers.[11] Pressures between 0.2 and 1.0 Torr, and temperatures between 600 and 650°C are used to produce deposition rates of up to 10 *nm* min^{-1}. The resulting crystals are predominately <110> orientation and sizes of 0.03 to 0.3 μm before and 1.0 μm after heat treatment. Higher temperature processes can produce epitaxial films if a seed is present. Most processes uses the thermal decomposition of silane;

$$SiH_{4\,(gas)} \xrightarrow{600\,^{\circ}C} Si_{(solid)} + 2H_{2\,(gas)} \qquad (2.7)$$

Other processes dilute 20% silane in nitrogen at similar total pressure to reduce the deposition rate allowing increased control and uniformity.

Silicon Nitride.

Silicon nitride is used as an oxidation mask in the common LOCOS process as well as a passivation layer to sodium and water contamination. There are a number of ways to achieve a silicon nitride film, but LPCVD is the most popular. The reaction produces stoichiometric silicon nitride at reduced pressures in the LPCVD reactor shown in figure 2.3. It is the reaction of dichlorosilane with ammonia.

$$3SiCl_2H_{2\,(gas)} + 4NH_{3\,(gas)} \xrightarrow{700-800\,^{\circ}C} Si_3N_{4\,(solid)} + 6HCl_{(gas)} + 6H_{2\,(gas)} \qquad (2.8)$$

Non-stoichiometric silicon nitride films containing hydrogen can be produced at low temperatures, 250 to 350°C, in PECVD reactors. Although the reactions inside a PECVD reactor depend on numerous variables and reactions with complex interaction the net reactions are usually assumed to be quite simple. One reaction, such as that in equation 2.9, uses an argon plasma to drive the reaction of silane and ammonia. Another uses a nitrogen discharge to drive a silane nitrogen-reaction (equation 2.10). The properties of the films produced in this manner can vary greatly depending on the processing conditions,[16] however, as mentioned earlier the passivating qualities of these films are excellent.

$$SiH_{4\,(gas)}+NH_{3\,(gas)} \rightarrow SiNH_{(solid)}+3H_{2\,(gas)} \qquad (2.9)$$

$$2SiH_{4\,(gas)}+N_{2\,(gas)} \rightarrow 2SiNH_{(solid)}+3H_{2\,(gas)} \qquad (2.10)$$

Silicon nitride can also be deposited in a CVD reactor by reacting silane and ammonia;

$$3SiH_4+4NH_3 \xrightarrow{700-900\,^{\circ}C} Si_3N_4+12H_2 \qquad (2.11)$$

Both of these reactions have similar deposition rates to polysilicon.

Although it is possible to grow silicon nitride at temperatures around 1100°C the films are thin and have poor compositional consistency.

Silicon Dioxide.

CVD silicon dioxide is used as an interconductor insulator and as a planarising material when it is phosphorus doped and as a passivation layer without doping. The phosphorus doped silicon dioxide, commonly called P-glass, is important for re-flow smoothing of contact steps and gettering of sodium contamination (both processes are covered in later sections). The method of deposition and reactants used depends on the use of the oxide. Oxide as an insulator over polysilicon is deposited in a CVD furnace tube by the reaction of dichlorosilane with nitrous oxide.

$$SiCl_2H_{2\,(gas)}+2N_2O \xrightarrow{900\,^{\circ}C} SiO_{2\,(solid)}+2HCl_{(gas)}+2N_{2\,(gas)} \qquad (2.12)$$

It gives good uniformity but chlorine incorporation into the film can attack the polysilicon causing cracking and long term reliability problems. Another possible reaction is

the decomposition of tetraethoxysilane (TEOS) in a LPCVD reactor. It produces by-products of organics and organosilicates, but has the advantage of excellent uniformity and step coverage.

$$Si(OC_2H_5)_{4\,(vapour)} \xrightarrow{650\text{–}750\,^{\circ}C} SiO_{2\,(solid)} + by\text{-}products \qquad (2.13)$$

For applications requiring lower deposition temperatures and re-flow glass the most widely used process is the reaction of silane with oxygen and dopants with oxygen. For P-Glass, reactions in equations 2.14 and 2.15 are run concurrently with the product being 6 to 8% phosphorus oxide. The common reactor configuration for this process is the horizontal hot-plate type.

$$SiH_{4\,(gas)} + O_{2\,(gas)} \xrightarrow{500\,^{\circ}C} SiO_{2\,(solid)} + 2H_{2\,(gas)} \qquad (2.14)$$

$$PH_{3\,(gas)} + 5O_{2\,(gas)} \xrightarrow{500\,^{\circ}C} 2P_2O_{5\,(solid)} + 6H_{2\,(gas)} \qquad (2.15)$$

Silicon dioxide can also be produced in a PECVD reactor driven by an argon plasma. The product film is usually used as part of a passivation scheme. The reaction is described in equation 2.16.

$$SiH_{4\,(gas)} + 4N_2O_{(gas)} \rightarrow SiO_{2\,(solid)} + 4N_{2\,(gas)} + 2H_2O_{(vapour)} \qquad (2.16)$$

Tungsten.

Tungsten is becoming a more common gate material[18] when combined with silicon to form silicides and as a barrier region in contact formation. The decomposition of tungsten hexafloride, equation 2.17, in a LPCVD reactor can selectively deposit tungsten on to silicon allowing self aligned contacts, and self-aligning tungsten to polysilicon caps to make polycides. It also has the advantage of conformal step coverage,[19] however the reaction's nasty exhaust products can cause some difficult problems.

$$WF_{6\,(gas)} + 3H_{2\,(gas)} \rightarrow W_{(solid)} + 6HF_{(gas)} \qquad (2.17)$$

Other Metals.

It is possible to deposit aluminium as well as metals such as titanium (Ti), tantalum (Ta), and molybdenum (Mo) with LPCVD techniques.[17] The only advantage this deposition has is that it is conformal over steps since it is a surface reaction product. Other deposition techniques such as those described in the next section suffer from step coverage problems. The MOCVD of aluminium has been demonstrated[20] through decomposition of tri-isobutyl-aluminium, equation 2.18, but production cost effectiveness might not be possible.

$$2[(CH_3)_2CHCH_2]_3Al_{(vapour)} \rightarrow 2Al_{(solid)} + 3H_{2\,(vapour)} + by\text{-}products_{(vapour)} \quad (2.18)$$

Many of the other metals are deposited through the hydrogen reduction of metal chlorides, as shown in equation 2.19 (where the metal of interest is substituted for M) in a LPCVD reactor. Most of these processes are still at an experimental stage.

$$2MCl_{5\,(gas)} + 5H_{2\,(gas)} \xrightarrow{600-800\,^{\circ}C} 2M_{(solid)} + 10HCl_{(gas)} \quad (2.19)$$

2.3.4. Physical Vapour Deposition.

Physical Vapour Deposition (PVD) is routinely used to deposit metallic conducting films. It differs from CVD in that no surface reaction is involved, the physical vapour just condenses on the wafer surface forming a thin film. PVD systems can be divided into two classes depending on how the vapour is produced. There is the original system of evaporators and the newer standard of sputtering machines.

All systems require the source and substrates to be contained in a vacuum to promote vapour transport and to reduce incorporation of impurities into the film. One common problem is that the film thickness on a surface is a function of the angle the surface makes with the incident vapour flux. Small angles, such as sidewalls of steps, reduce the film thickness and result in step coverage problems. Most systems usually employ planetary substrate fixtures, not only to allow batch processing, but to improve step coverage by continually changing the orientation of the wafer surface to the vapour flux. Reference (6) provides excellent detail both on types of equipment and the physics behind their operation.

Evaporators.

Evaporation systems were originally the main source of metallic films but have to a large extent been replaced by sputtering systems. Vacuums of between 10^{-6} and 10^{-7} Torr are required to transfer material from the evaporation source to the target. A number of evaporation sources, such as resistance and inductively heated sources, have been used but the most popular was electron beam evaporators. With e-beam evaporators deposition rates of up to $0.5\mu m\ min^{-1}$ can be achieved.

The drawback to evaporators came when it became necessary to use alloyed films to overcome reliability problems.[19] Alloyed sources cannot be used in evaporators because each constituent of the alloy would evaporate at a different rate. Multiple sources can be used to solve the problem but such a system is complex and difficult to control.

Sputtering Systems.

A sputter deposition source depends on momentum transfer from an incident inert ion beam to the atoms of a composite target. The energy transferred during the impact gives surface atoms on the target sufficient energy to become volatile and travel to the deposition substrate. The advantage over evaporators is that, as long as bonding energies are exceeded, all constituents of an alloy sputter equally, resulting in films with similar compositions to the source. Secondary advantages are; that sputtering systems use lower acceleration voltages therefore subjecting the wafers to less radiation damage, and it is possible to re-direct the sputtering beam to the wafers thereby providing in situ cleaning for low resistance contacts.

The most common alloys are *Al-Si* and *Al-Si-Cu*, but other metals; *Ti*, *Pt*, *Au*, *Mo*, *W*, *Ni* and *Cu* as well as dielectrics; Al_2O_3 and SiO_2 are often deposited with sputtering equipment.

There are a number of sputtering source designs available;[6] diode, triode, reactive and RF excited sputtering sources, but by far the most popular source for metals is the magnetron source.[19] The RF excitation source is used for dielectrics.

Magnetron sources confine the bombarding ions, usually argon, at pressures around 10^{-3}Torr, in a magnetic field around the target. The magnetic field causes

increased electron densities by effectively lowering the target impedance. This results in an increased deposition rate. Deposition rates can be as high as 1.0μm for Al and its alloys.

2.3.5. Silicides

Silicides are a combination of silicon and refractory metals. They are used to reduce the electrical resistance of transistor gates without introducing the instabilities of pure refractory metals. There are a lot of problems with stability of these films both mechanically and chemically. Polycide, a sandwich of polysilicon and silicide, deposited by LPCVD solves many of the problems and is used in production.[21]

Research into pure silicides continues with a number of deposition approaches. Co-evaporation,[19] co-sputtering,[22] LPCVD,[18] ion beam mixing (metal is deposited over silicon and ion bombardment is used to mix the metal and the silicon)[23] and sintering (metal is deposited over silicon and the combination is heated in a furnace or RTP system to achieve interdiffusion mixing) are all currently being evaluated.

2.3.6. Section Summary.

In this section a number of methods for adding a thin film to the surface of a wafer were discussed. From epitaxial and oxide growth, through surface chemical reactions, to straight physical deposition, thin films of a variety of materials can be formed on the surface of a wafer. With the exception of diffusion barrier masked oxide growth and selective tungsten deposition all these techniques cover the surface as a whole without regard for the pattern required to make useful integrated circuits. The next two sections, on pattern transfer and selective layer removal, will describe how unwanted areas of these films are removed to leave the desired circuit patterns.

2.4. Pattern Transfer.

Photolithography is the key process in microfabrication which makes the production of semiconductor integrated circuits possible. It involves the transfer of the design conception of a pattern into the reality of a layers morphology. Each integrated circuit is composed of a number of layers of insulating, semiconducting and conducting

materials. Various shapes on each layer interact, (conductor through contact hole in insulator to conductor), in order to produce the final product.

The information regarding how these layers must interact in order to produce a required function is created by a circuit design engineer on a computer aided design (CAD) system. The information is then separated into the content required for each layer. A pattern generator is used to transfer the design information onto a mask for each layer (see figure 2.4). An exposure tool is used to transfer the design information contained on the mask to a resist material on the surface of the wafer. The resist material then prevents certain areas of the wafer from being processed (etched, or ion implanted) thereby translating the design information into a layer shape on the wafer.

Since most of the transfers are usually achieved by photography techniques the process is called photolithography, using photomasks and photoresists. Here the processes are broken into three manageable sections, mask making, exposure tools, and photoresists, but in actuality they are highly interdependent.

Photolithography received a great deal of attention in the late 1970's and early 1980's in order to break through packing density barriers, which resulted in a wealth of papers on the subject. [24-31] Reference (29) provides a useful overview of all the exposure methods. Reference (31) gives a assessment of current production lithography including resist schemes, and reference (25) discusses the problems of the developing future lithography techniques and attempts to predict the outcome. The best introductory reference for optical lithography considerations is reference (32) and reference (25) provides the best economic viewpoint on lithography.

There are a number of important parameters in lithography systems and they have attracted numerical descriptions for the purposes of comparison and calculation. Some of them are:

Magnification:	the relative size of the mask image to the wafer image.
Usable Resolution:	the minimum feature size which can be reliably reproduced by the system.
Image Field:	the maximum size of area which can be exposed at one time.
Depth of focus:	the range of variation in depth which will still result in a focused image.

Resolution: the minimum feature size which a resist can reproduce.

Throughput: the number of wafers per hour that the system can expose.

Overlay Accuracy: the tolerance to which two masking levels can be repeatably aligned to each other.

Distortion: the magnitude of line width variations due to the exposure tool.

Resist Contrast: the log ratio of maximum to minimum exposure energy a resist is sensitive to.

Resist Adhesion: the ability of the resist to stick to the substrate.

Resist Temperature: the maximum temperature the resist can tolerate without losing shape.

Resist Hardness: the degree to which the resist resists attack during subsequent processing.

Resist Sensitivity: the minimum energy required to expose a resist.

Resist Thickness: the minimum resist thickness required to avoid deleterious effects.

Exposure Wavelength:the wavelength of light used for the exposure. It effects some of the previous parameters.

There are other measures of optics systems such as numerical aperture, modulation transfer function, and aerial images, all of which are described in reference (32, 33) and others.

2.4.1. Mask Making.

Some exposure systems use a mask containing tens to hundreds of actual size images of an integrated circuit, laid out in the pattern that would cover a whole wafer. They are made from an enlarged version, usually five or ten times the size of a single die (or integrated circuit), called a reticle. Some exposure systems use the reticle directly and others can dispense with masks entirely.

Possible Integrated Circuit layout Information Flow.

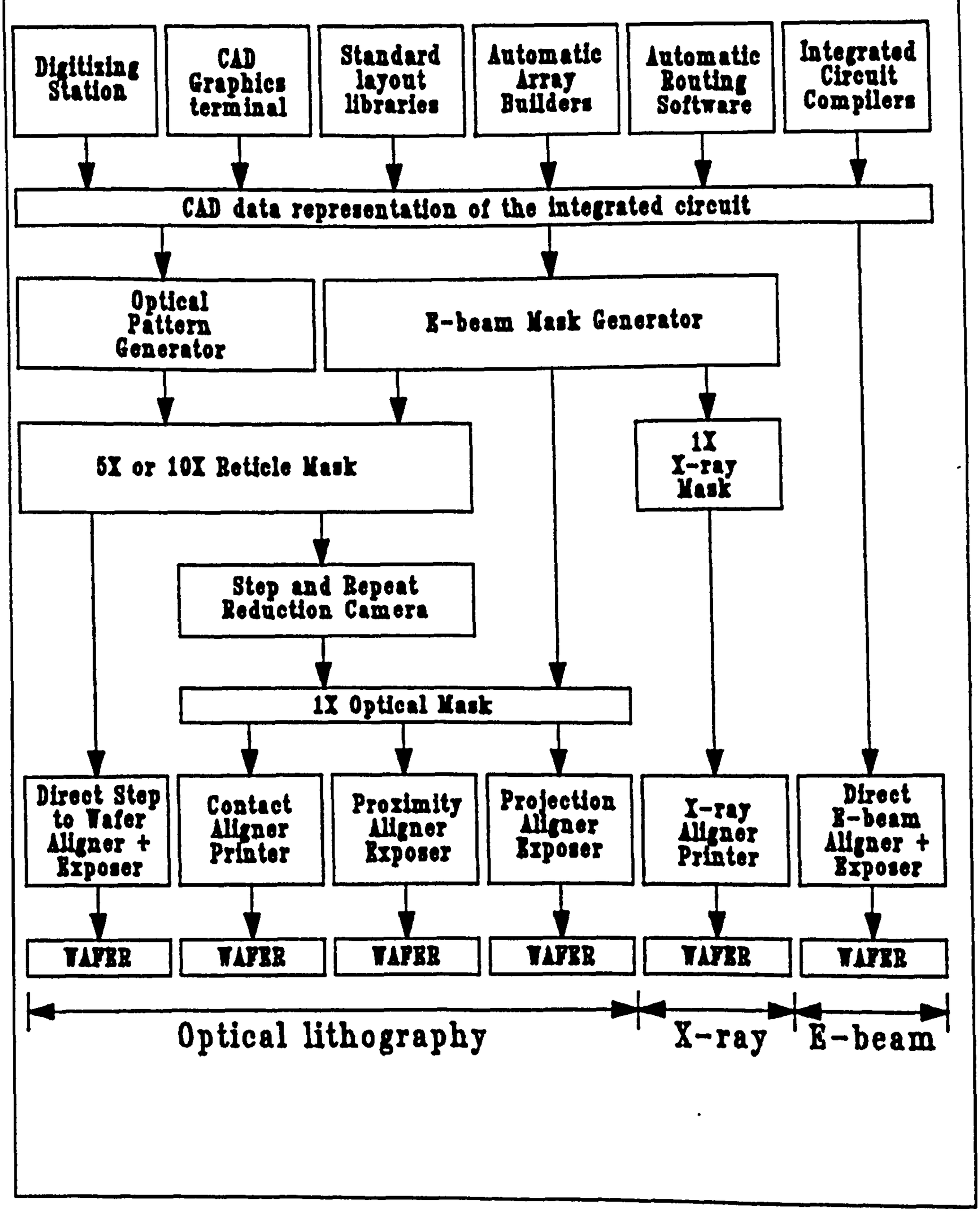


Figure 2.4, Routes through lithography.

Figure 2.4 shows a number of routes for layout information from design to the final wafer. Mask or reticle requirements depend on the exposure system for which they are produced. Both optical and electron beam pattern generation is briefly discussed here.

Optical reticle and masks are made from similar materials. They are usually an optically flat glass plate (5×5 inches × 2mm), which is coated with chromium and then a photon or electron sensitive resist. In some cases low cost emulsion-only reticles are used.

For exposure in an optical pattern generator, the reticle is then clamped in an accurately controlled X-Y stage. A pulsed light source shaped by a variable dimension slit is focused onto the resist. A computer controls the shape of the slit and position of the X-Y table from the CAD data describing the integrated circuit pattern. When they are the correct size and in position the light source is "flashed" producing a tiny portion of the final image. This operation may take several hours and up to several hundred thousand flashes before the required shapes are built up on the reticle.

The enlarged image of the chip is then developed by chemical developers before the optional chromium layer is etched. The resulting reticle can then be used directly or it can be placed into a step and repeat camera and copied across another plate in rows and columns of actual size die. After being developed and etched it is then ready for use as a mask in 1:1 exposure tools.

An alternate way of producing masks or reticles is to place an electron beam resist coated plate onto an X-Y stage in the e-beam pattern generators vacuum chamber. The plate is moved to the start of the first integrated circuit site and the electron beam is modulated and directed by computer control to expose the image of the die. Two beam scanning schemes exist; raster scan, like a television where the beam is pulsed on and off, and vector scan, like an X-Y plotter where the shape is drawn by the beam. After a die is completed the stage moves to the start of the next die and the process repeats until a mask is completed. Five or ten times reticles are also made using e-beam pattern generators either for direct use by exposure tools or for the step and repeat camera.

X-ray masks are more complicated and composed of a 10μm thick sandwich of gold, polyimide and boron nitride. The patterning of masks is currently achieved by e-

beam technology. The boron-nitride and polyimide are substrate materials and the gold is a masking X-ray absorber.[2] Historically, photolithography reticles were produced by 10-25:1 reduction cameras from hand drawn artwork cut into sheets of rubylith film. This procedure, used as little as 10 to 20 years ago, would have taken several days. Table 2.1, compares the speed of current techniques.

Machine ☞ / Features	Optical P.G.	E-beam P.G.
Minimum feature size	2.5 μm	0.25 μm +
Positional accuracy	0.5 μm	0.2 μm
10cm×10cm Exposure time (100000 features)	10 hours	0.5 to 2.5 hours

+ can be smaller but increases writing time.

Note: both machines can produce 10:1 reticles.

Table 2.1, Mask Making.

2.4.2. Exposure Tools.

There are a number of exposure systems available for lithography both in experimental and production systems. Contact, proximity, projection and direct-step-on-wafer optical systems, direct write electron beam, and X-ray systems will be presented here. With the exception of X-ray lithography, all of these systems are production machines. Other experimental techniques are discussed in reference (29).

Table 2.2 contains a summary of the features of these systems, figure 2.5 shows schematic representations of them, and the following sub-sections contain a brief description of their operation. All exposure tools are required to perform two tasks.

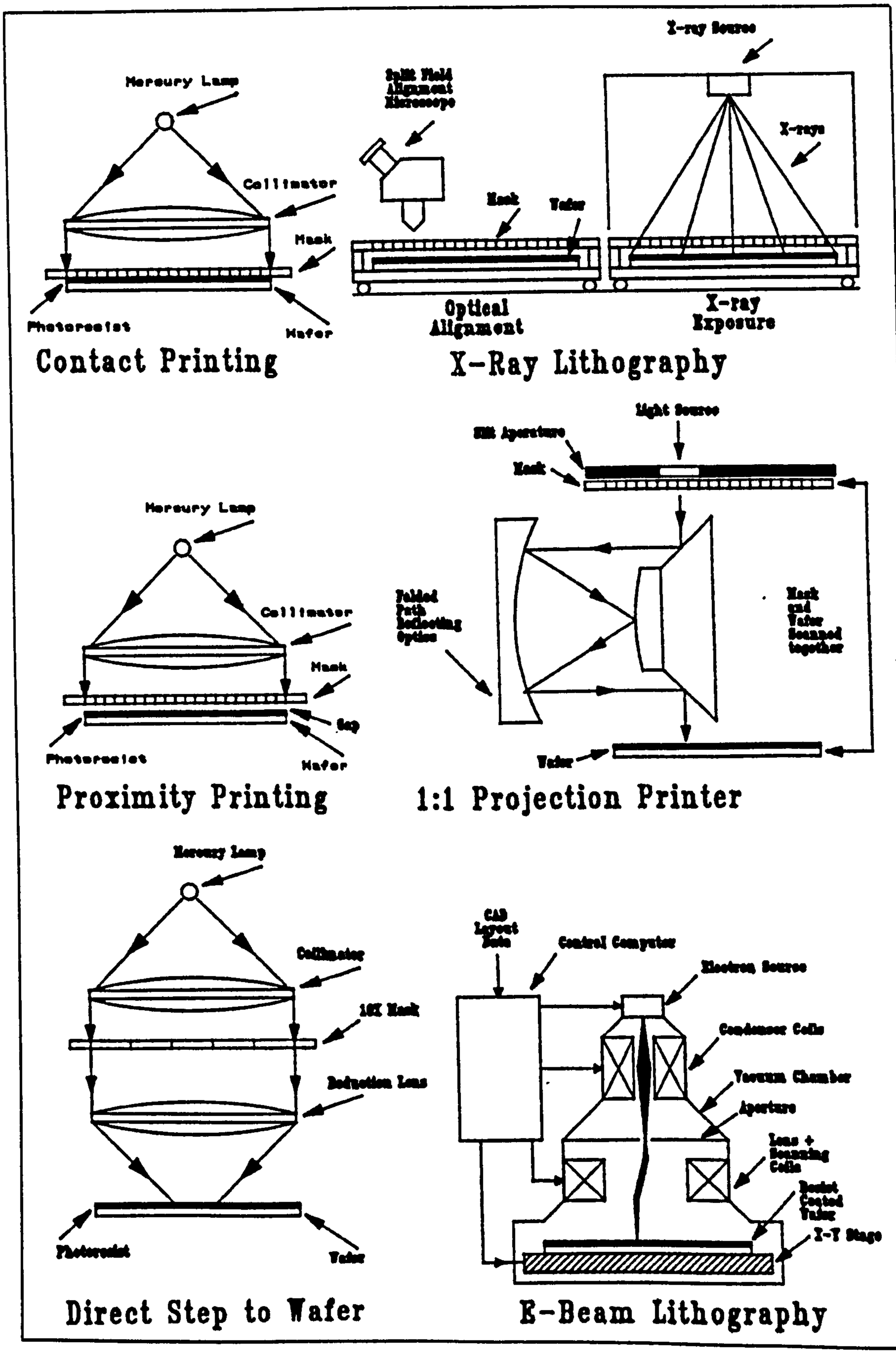


Figure 2.5, Lithography Exposure Tools.

1) Align the wafer, with possible previous layers, to the current mask.
2) Expose each die on the wafer to create a copy of the mask pattern in the photoresist.

Contact Printer.

A contact printer is the simplest of all exposure systems. The operation is as follows:

1) A mask is held slightly (50μm) above the wafer surface.
2) An operator rotates and translates the wafer until alignment marks on opposite sides of the wafer, as seen through a split field microscope, align with the mask.
3) The wafer is then pressed tightly against the mask and the image is transferred by even illumination from a mercury UV lamp.

Proximity Printing.

A proximity printer is very similar to a contact printer except that the mask never comes into contact with the wafer but is held in close proximity (about 10 μm) during exposure. The advantage to this is that the wafer yield and mask lifetime are increased but the resolution is poorer than for contact printing. Most proximity printers use only part of the mercury lamp spectrum.

Projection Printing.

The advantage of non-contact printing without the diffraction limitations of a proximity printer can be had from a "ring" field scanning 1:1 projection system. This exposure tool uses circular reflecting optics to project a crescent shaped portion of the mask onto the wafer. The wafer and the mask are jointly scanned across the apertures in order to expose the entire wafer surface. The crescent shape is about 1mm in width and 80mm in length. Alignment is achieved with a similar method to contact and proximity printing. Projection printing was the main exposure tool before step and repeat systems were introduced and they still accounts for a large portion of volume I.C. production.

Direct Step-to-Wafer.

Direct Step-to-Wafer exposure machines use 5:1 or 10:1 reducing optics, (both reflective and refracting systems exist), to project the reticle image directly onto the wafer. Each die site is exposed separately as it is moved into position by an accurate X-Y stage. Alignment of the wafer to the stage can be performed automatically from alignment marks on die on opposite sides of the wafer. Then exposure positioning for each die can be achieved by automated die by die alignment or just by relative stage movement for less critical alignment. Die by die alignment can compensate for wafer warpage. Reduction systems in general have better resolution because the larger reticle features are not effected by diffraction. Exposure defects can be reduced by fitting pellicles, (a thin film at a distance from the mask surface), to the mask to keep dust out of the focal plane. The common refractive systems use a single mercury spectrum "line" to avoid chromatic aberration.

Electron Beam Direct Write.

Electron beam direct write machines expose one die at a time on the wafer under computer control with the CAD data as an input instead of a mask. The resist coated wafer is clamped to a stage in the vacuum chamber of the machine. Electromagnetic deflection is used to draw the pattern. The stage is moved to select another die for patterning. Alignment is achieved by measuring the change in back scattered electron current caused by a reduced current beam striking suitable alignment structures. The minimum spot size and hence feature size is not limited so much by the e-beam optics but by the time required to write a pattern. Although features as small as 10 nm have been written, the time to expose a wafer would be impossibly long. As a result e-beam machines tend to use spot sizes one quarter of the minimum dimension required.

Direct write E-beam machines are not used for volume production because they have low throughput. However, they have a unique niche in fast turn around design-to-product cycles, since there is no time spent on mask making and checking.

X-ray Lithography

X-ray lithography is an area of very active research. X-ray's smaller wavelength and resulting greater depth of field and resolution, coupled with the high throughput of 1:1 masks makes it a likely choice for future lithography systems. The biggest problem to overcome is that of a reliable and small X-ray source. Current work on synchrotron sources will likely produce a multiple exposure station in the 1990's. There are also difficulties in making masks. It is expected that only high volume high density memory products would be a cost effective use of X-ray lithography.

The likely configuration of a system would be to use wafer-mask carriers which allow optical alignment of the wafer to the mask followed by transport into the X-ray exposure system. Reference (2) discusses all of these systems and provides more detail on the mechanisms of their limitations.

System ☞ Feature	Contact Printing	Proximity Printing	Projection Printing	Direct Step on wafer	Direct E-Beam Exposure	X-Ray Exposure
Mask:Wafer Ratio	1:1	1:1	1:1	5:1 10:1	no mask	1:1
Exposure area	Wafer	Wafer	Strip of Wafer	Die	Feature	Wafer/ or die
Useful Resolution	1.0 μm	3-5μm	2-3μm	1μm	0.25μm	0.35μm
Alignment Variation	1.0μm	1.0μm	0.7μm	0.15μm	0.1μm	0.1μm
Mask Life	Short	Long (cleaning)	Very Long (pellicles)	Very Long (pellicles)	no mask	long (cleaning)
Exposure Induced Defects	High	Medium	Low	low	Very Low (Vacuum)	Medium
Overlay Error	1.5μm	1.5μm	1.2μm	0.3μm	0.1μm	0.2μm
Throughput (Wafers/hr)	100	100	60	10-40+	1-2	15-60
Price (1984)	$55K	$55K	$340K	$730K	$1.0M	$68M† (8 stations)
Depth of Focus	-	-	8μm	4μm	50μm	sync source very good
Capital Cost per Wafer *	$0.25	$0.25	$2.60	$13.30	$228.10	$404.30
Projected Area of future use.	1	2	3	4	5	6
	1: Large discrete components. 2: old 5μm processes 3: >1.5μm processes (high volume) 4: 1 to 2μm processes (logic and analogue) 5: High speed turn around and specialist processes. 6: <1μm high volume processes. (64Mbit Chips).					

Table 2.2, Exposure Systems Features.[34]

+ dependent on # of die/wafer.
* 5 year depreciation, 50% use, 10 layer processes.
† Estimates vary from $16M in 1984 to $100M in 1988.

2.4.3. Resist Systems.

There are a number of requirements of a good resist. Some of the important ones are:

1) To be sensitive to the exposure illumination, which could be photons, electrons, or X-rays, such that a chemical change occurs.
2) To respond to a development treatment which leaves, or removes, the exposed areas.
3) To adhere strongly to the wafer surface and to be easily removed under suitable conditions.
4) To be easily applied to the wafer surface.
5) To be resistant to damage during subsequent processing and to act as a mask of the wafer surface during that processing.

Exposure and development.

The basic exposure mechanism is the same for all resists whether they are positive or negative and whether they are sensitive to photons, electrons, or X-rays.

The exposing photon or electron changes the chemical structure of the resist so that it becomes either more resistant (negative), or less resistant (positive), to removal in a development solution.

There are many types of resists with different sensitivities, contrasts, and resistance to etching. [2,31,32,35]

Most integrated circuits are made by photolithography with positive photoresists. Positive photoresists are currently more popular than negative resists because positive resists can support resolutions approximately equal to their thickness. Negative resists can only support resolutions of 2 to 3 times their thickness. Attempts at reducing negative resist thicknesses resulted in pin holes forming. However, negative resists have contrasts 3 to 4 times better than positive resists which gives greater process latitude. Only positive resists will be discussed here.

The most common positive photoresists are made by adding photo sensitisers to resins. Novolac (phenol-formaldehyde) resins become photo sensitive when

diazonapthoquinone is added. Other additives are used to modify the physical characteristics of the films. The sensitiser in its unexposed state reacts with the resin in an alkaline developer (such as sodium or potassium hydroxide) to form a cross linked polymer which has reduced solubility in the developer. The exposed area of the resist is removed since the resin is naturally soluble in the developer and the sensitiser has been photolytically decomposed into indene carboxylic acid which does not link with the resin. Thus exposed areas of the resist are removed in the developer.[36]

Other positive photoresists exist which use deep UV exposure to cause main chain polymer fissions in the exposed area of the resist. These weakened areas have greater solubility during development resulting in exposed areas being removed. One such material is PMMA (Poly(methyl methacrylate)) which is also used as an electron beam resist. The developer is isobutyl ketone in isopropyl alcohol.

Resist Application and Hardening.

In order for there to be good adhesion between the wafer and the photoresist the wafers must be prepared before coating.

The wafers are prepared by washing and spinning dry in warm nitrogen. Then the wafers are subjected to hexamethydisilizane (HMDS) vapour which forms an oxide layer after reacting with water vapour on the wafer surface. Other primers such as trichlorophenysilane (TCPS) or bistrimethylsilyacetamide (BAS) provide a similar function. The wafers then are immediately coated with resist.

Resists are usually supplied as viscous liquids containing the photoresist in an organic solvent. A uniform resist thickness can be achieved across the whole wafer by spin-on equipment.

A spin-on-coater attaches a horizontal wafer to a vacuum spindle. An amount of resist is dispensed onto the center of the wafer which is then spun at a predetermined speed between 3000 and 6000 rpm. After spinning for a fixed time between 15 to 60 seconds a uniform layer of resist will cover the wafer.

The thickness of the film depends on the spin-speed, viscosity of the resist, and initially on the spin time. Resist thicknesses are usually between one and several microns thick. Care must be taken in both choosing and maintaining a constant resist

thickness. Reflecting materials, such as metal, produce optical standing waves in the resist which can lead to notching and sidewall terracing. Dyed resists have been produced to minimise the problem but using minimum exposure energy can result in fillets near steps in underlying topography due to local thickness variations (see appended paper).

After coating the wafers are heated either in batches in an oven or individually on a hot plate in order to drive off the solvent and make the resist hard enough to withstand handling and development. The "soft" bake temperature must be high enough to drive off the solvents but lower than the thermal exposure temperature. This is usually just over 100°C.

After development the wafers are usually baked again at a higher temperature to drive off the remaining solvents both promoting adhesion and making the resist more resistant to attack in subsequent processing. The temperature of "hard" bake is usually between 120 and 180 °C. The temperature must not be too high since the resist will start to flow and lose shape, or it will become completely dehydrated and gasses adsorbed on the surface will release and cause resist lift off.

Advanced Photoresist Systems.

Modern experimental variations on photoresists are aimed at improving; resolution, sidewall slope, resist strength, and coping with the smaller depth of field of projection exposure tools. Some variations are, image reversal, multilayer resists, and single layer silyated resists with dry development.

Image reversal is one technique to give higher contrast, steeper sidewalls, and increased line width control effectively increasing the depth of field.[31] Conventionally exposed NOVOLAC resists are exposed to amine containing gasses which combines with the photoproducts in exposed areas to form insoluble compounds. The wafer is then flood exposed and the previous unexposed areas are removed in an alkaline developer.

Multilayer resists[37] attempt to improve the same factors as image reversal with the addition of better line width control over steps. The main feature behind the system is that only a very thin upper layer of the composite resist needs to be exposed. The

image is then developed and used as a mask to expose the lower level resist with deep UV radiation or as an etch mask for R.I.E. patterning of the resist in a tri-level scheme. The lower level of the resist has two purposes, one is to planarise the wafer for the photolith resist layer to be deposited on and the second is to provide a thick mask for subsequent processing.

A resist called DESIRE[38] (Diffusion Enhanced Silyating Resist) gives improved performance in almost all the categories and is still a single spin-on resist. Following normal coating, pre-bake and exposure the resist is treated with a silyating gas such as HMDS. It diffuses preferentially into the exposed regions of the resist and produces silicon bonds with the resin. The high silicon content in the exposed areas protect them from oxygen plasma etching of the resist. Unexposed areas are thereby removed. The two main advantages of the system are low exposure energy, since only the top 200 to 300 nm must be exposed, and dry etching to give steep sidewalls. That allows for the use of thick anti-reflective planarising resists with obvious advantages.

2.4.4. Section Summary.

In this section the transfer of a pattern from the creation on a CAD tool to the shape of a resist layer were described. Some of the various routes through mask making, and pattern exposure were discussed along with the structure and mechanisms of photoresists, which act as masks in the selective processing which produces integrated circuits.

2.5. Selective Removal of Material

As discussed in a preceding section, micro-electronics takes a blanket approach to the deposition of new layers. It follows then, that selective removal of material is necessary to produce layer patterns which will function as components of an integrated circuit. It is also important to realise that the topography of lower layers must not be significantly disturbed during the patterning exercise.

Historically, wet chemical etches were used to remove areas of a layer that were not protected by a photoresist. Lately, dry plasma or gas etching methods have nearly replaced chemical etching in order to meet the more stringent requirements necessary

for modern levels of miniaturisation. The following sub-sections discuss the reasons behind the change in techniques as well as the mechanisms behind them.

2.5.1. Wet Chemical Etching.

Wet or chemical etching is exclusively used in wafer production and has been used in device fabrication when feature sizes were over 3.0μm. The basic scheme of all wet etches is to submerge the photoresist-masked wafers in a liquid containing etches which are selective to the material to be removed. Some systems provide agitation, heating, and particle control.

The essential mechanism of all wet chemical etches is;

1) Transport of etchant species to the wafer surface.

2) Reaction of the etchant with the material to be removed.

3) Transport of the reaction products away from the surface.

The amount of material etched is controlled by the chemical composition, solution temperature, amount of agitation, and the duration of the etch.

Considerations for chemical choices are, selectivity to the material to be etched, etch rate, resist adhesion, and etch uniformity. Reference (2) lists common etchants and the etch rate of various materials. Reference (3) gives a practical review.

Wet etches have the advantage of being highly selective to the material to be removed. However, their disadvantage lies in the isotropic nature of the etch (the film edge is laterally undercut at the mask edge for a distance equal to the film thickness etched), and that they are sometimes sources of particulate contamination.

There are some non-isotropic chemical etches in the form of orientation dependent etches. Some etchants, potassium hydroxide (KOH), will etch a given silicon crystal plane faster than another. Although not a main branch of integrated circuit fabrication, it is possible to precisely etch "U" and "V" groves in silicon[2] to make high voltage (punch through limit) field effect transistors. The majority of integrated circuit thin films can be etched isotropically by the chemistry described in the following subsections.

Silicon.

Silicon is chemically etched in a mixture of nitric acid (HNO_3) and hydrofluoric acid (HF) in either water or acetic acid (CH_3COOH). The overall reaction, given in equation 2.20, is the result of two sub reactions.

$$Si_{(solid)} + HNO_{3\,(liquid)} + 6HF_{(liquid)} \rightarrow H_2SiF_{6\,(solute)} + HNO_{2\,(liquid)} + H_2O_{\,(liquid)} + H_{2\,(gas)} \qquad (2.20)$$

The nitric acid oxidises the silicon through an auto-catalytic hole producing reaction with the HNO_2. The resulting silicon dioxide is then etched by the hydrofluoric acid to produce the solute product in equation 2.20 which it transported away by the solution. Depending on concentrations etch rates of up to several hundred microns per minute can be achieved. An oxide mask, which is etched slower than the silicon,[39] is needed because photoresist adhesion is not good.

Polysilicon.

Polysilicon is etched with similar etches to silicon. Since photoresist is attacked by nitric acid, an oxide mask,[39] patterned by the photoresist layer, is used to protect the areas to remain. The polycrystalline film etches slightly faster than the monocrystalline silicon.

Silicon Dioxide.

Silicon dioxide can be etched with hydrofluoric acid at room temperature, but the etch rate is too high and also attacks the resist-oxide interface. Diluted HF in water can be used but the solution rapidly loses its acidity making control of the etch difficult. The popular solution, called buffered HF, contains the buffering agent ammonium fluoride (NH_4F) which helps to maintain a constant acidity of the solution.

Thermal oxide can be etched at about 1500Å $minute^{-1}$ at 30°C in a 4:1 $NH_4F:HF$ solution. Less dense, and therefore faster etching, pyrolytic oxides are often etched in a 25:1 solution.

Silicon Nitride.

Silicon nitride can be etched with buffered HF but the etch rate is slow and obviously not selective to SiO_2. Orthophosphoric acid (H_3PO_4) will etch silicon nitride a few hundred Angstroms per minute at 165°C. Unfortunately an oxide mask must be made first as resist will not tolerate the high temperature.

Aluminium.

Aluminium can be etched in a number of etchants[3] however a mixture of orthophosphoric acid (H_3PO_4), nitric acid (HNO_3) and acetic acid (CH_3COOH) is the most common. Some dilutions can etch aluminium at rates up to 1μm $minute^{-1}$. The production of hydrogen gas bubbles by the reaction necessitates constant agitation to achieve uniform results.

2.5.2. Dry Etching.

Dry etching uses gasses and plasmas to etch the various layers present during integrated circuit fabrication where wet etching used liquids. Dry etching, as evidenced by the number of papers readily available,[40-52] has now become an indispensable technique in VLSI microfabrication. There were seven reasons to change from wet to dry etching. They are, roughly in order of importance;

1) Excellent fine line definition.
2) High directional etching (Anisotropic).
3) Good selectivity.
4) Smaller chemical volumes, giving lower chemical costs, and
5) Less environmental impact.
6) It is easier to automate than wet etching.
7) There is greater repeatability between batches.

The drawbacks are; higher equipment costs, possibility of radiation damage to wafers, and contamination from apparatus self-sputtering. Reference (44) gives an excellent review of present methods and reference (53) gives good depth on widespread techniques.

Etching Mechanisms.

Although there are at least a dozen dry etching configurations, they can be grouped into four classes based on the etching mechanism. The etching mechanisms are; Physical etching, chemical etching, chemical-physical etching, and photo-chemical etching. The selectivity and degree of anisotropy are the primary considerations when comparing etches. Both factors are of increased importance in small geometry processes. Anisotropic etches are important in patterning short transistor gates which would otherwise be completely undercut in an isotropic etch. Selectivity is important to avoid damage to thin underlying and masking layers.

All etching mechanisms, with the exception of photo-chemical,[44] and some chemical[50] mechanisms, use a plasma as the etchant. Most of the plasmas are created by subjecting the etchant gas to R.F. excitation of a few hundred watts at 13.56 Mhz, with pressures between 10^{-4} and 1.0 Torr. The resulting plasma contains positive ions, electrons, negative ions, and neutral gas atoms. It may also contain neutral free radicals (dissociated variants of the feed gas) which are extremely reactive. Etching equipment is designed to favour a particular etching mechanism which in turn determines the plasma constituent which does the etching. The etching mechanisms are described in the following sub-sections.

The physical etching mechanism is quite simply the sputtering of the wafer surface by an ion formed in a chemically inert gas plasma. The process is highly anisotropic, but since everything is sputtered equally non-selective. This mechanism is not intentionally used in microfabrication.

The chemical etching mechanism occurs when a reaction between the layer to be etched and some component of the plasma, or a spontaneous etchant,[50] produces a volatile product. The process is isotropic and highly selective making it ideal for some processing steps. A number of successful chemical etchers exist, including the original plasma asher for removing photoresists.

Chemical-physical etching is a synergistical combination of the two separate mechanisms. It has increased selectivity over physical etching and increased anisotropy over straight chemical etching. It is the basis of the most popular reactive ion etcher (R.I.E.). Essentially this mechanism uses plasma generated line of sight energetic ions

to bombard the surface along with chemically active etching species. The synergy between the two mechanisms can take four forms.

1) Chemically enhanced sputtering. Chemically weakened bonds facilitate physical sputtering.
2) Physically enhanced reactivity. Physical damage by bombarding ions increases the chemical reactivity of the surface material.
3) Chemical Sputtering. Energy from bombarding ions drives the chemical reaction.
4) Auto-resist film formation and sputtering. The chemical reaction produces a reaction stopping film on the surface of the material. Line of sight energetic ions sputter away the film on horizontal surfaces thereby causing anisotropic etching.

Depending on the etcher configuration and conditions any of these mechanisms might concurrently be active.

Photo-chemical etching uses line of sight photons to drive surface chemical reactions which produce volatile products from the material to be etched. As a result it is a highly anisotropic etch and can have good selectivity. The photons may assist the chemical reaction by excitation of the; gas phase reactants, the adsorbed species, or the solid. This mechanism is a promising area of research but is not yet commercially applied.

Etching equipment.

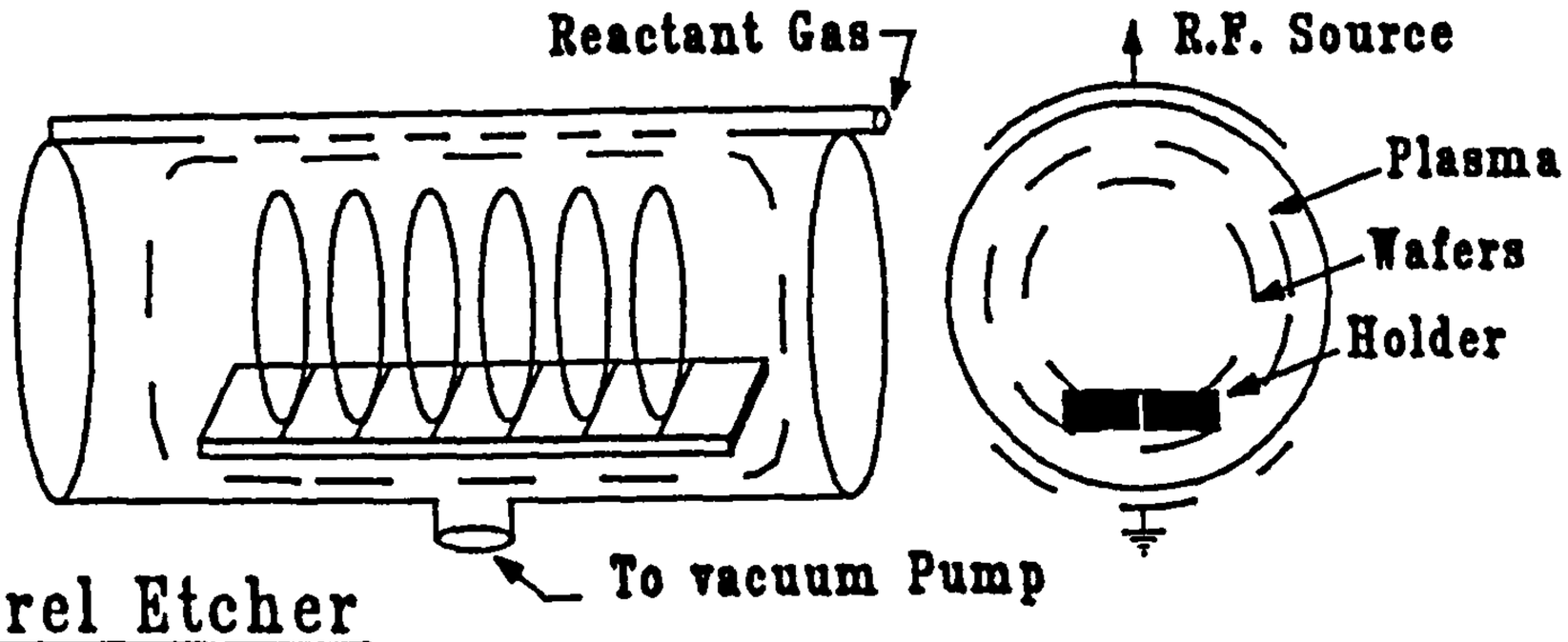
There are a number of etcher configurations and confusing nomenclature for them,[44] only three of the most common, barrel etchers, plasma etchers, and reactive ion etchers (R.I.E.) will be discussed here.

A barrel etcher, see figure 2.6, processes batches of wafers positioned vertically and electrically floating in the center of etcher. To avoid radiation damage, the wafers are usually shielded by some sort of screen or tunnel from all but neutral radicals in the plasma. The plasma is created by R.F. energy which is capacitively coupled to electrodes usually outside the container. Gas inlet and vacuum exhaust ports are provided in the container. Etching pressures vary from 300 mTorr to 5 Torr.[39] The etching mechanism is chemical so the resulting profile is isotropic. Drawbacks of the barrel etcher are; isotopic etching, flow dependent uniformity, and it cannot etch materials

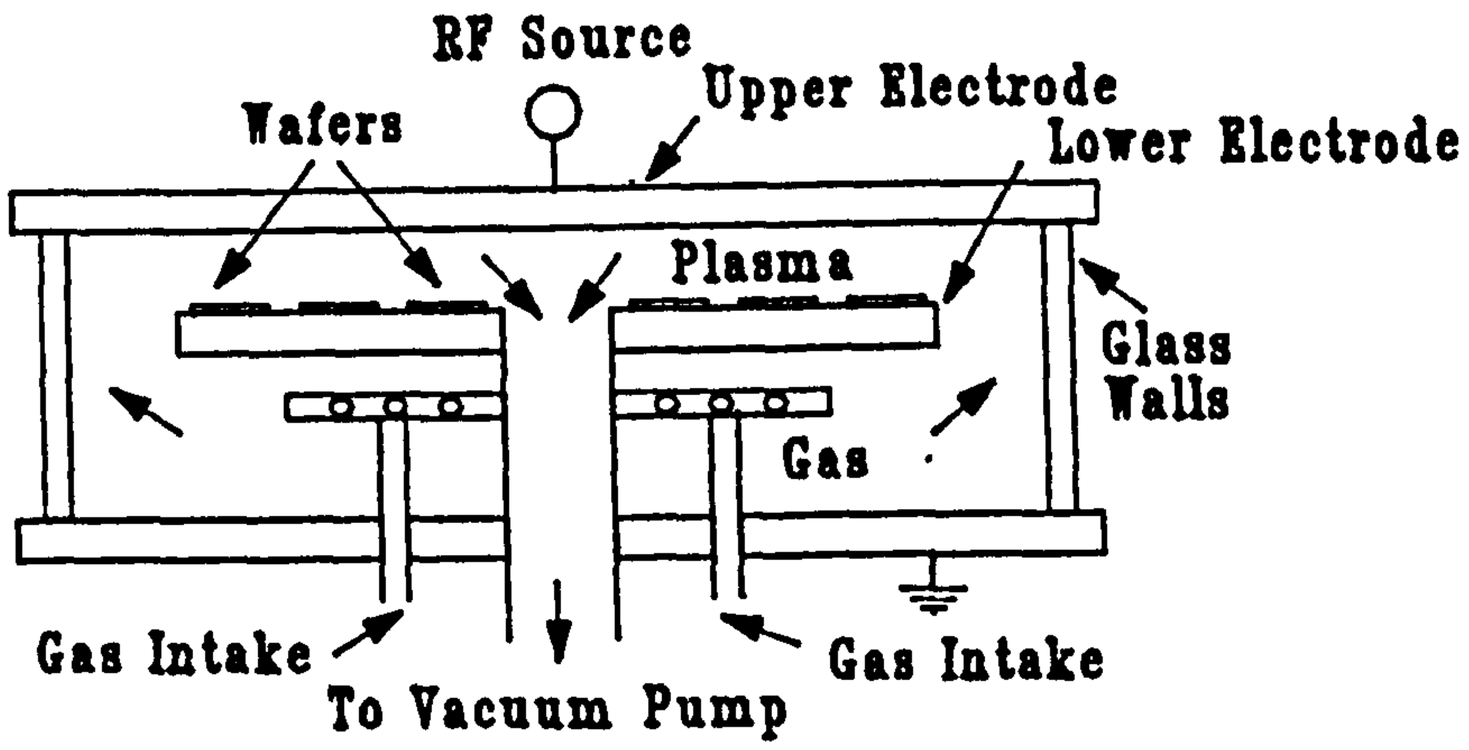
like SiO_2 on silicon or Al-Si alloys. Although it is possible to etch polysilicon and silicon nitride its main use now is isotropic stripping of resist. In that application it has the advantage of high selectivity, and minimum chemical waste problems. Etch rates of up to 6 $\mu m\ min^{-1}$ have been reported.[44]

Planar etchers, also shown in figure 2.6, (which are sometimes called parallel plate or plasma etchers) differ from barrel etchers in a number of ways. The wafers sit flat on their backs on the grounded electrode directly in the plasma. A parallel electrode above the wafers is driven by the R.F generator through an impedance matching network. Processing pressures are controlled somewhere between 100 and 300 mTorr. Processing gasses are introduced at the circumference of the lower plate and drawn off through a central exhaust vacuum port to provide a radial gas flow. The etching mechanism is mostly chemical with some physical-chemical action depending on operating conditions.

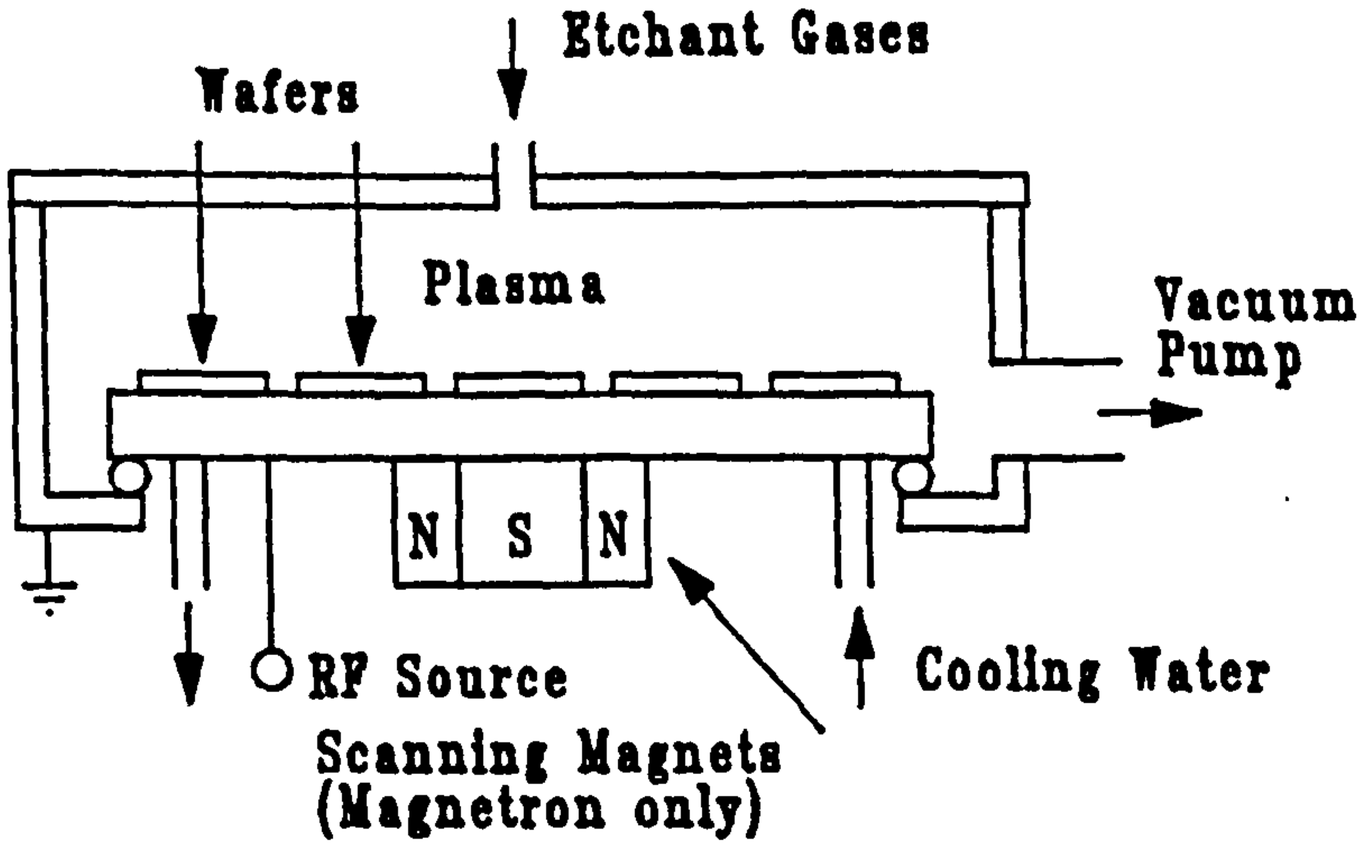
It has the advantage over barrel etchers of increased etch rates due to ionised species and it has some directionality as a result of sputtering by negative ions or electrons. The maximum sputtering energy is the plasma potential multiplied by the ion or electron charge. Reference (53) shows a plasma potential distribution. Disadvantages are that a high plasma potential is required to achieve directional etching which can lead to loss in selectivity, radiation damage to oxides, and contamination from sputtering of the container walls. The main uses are to etch oxides with high selectivity (100:1)[5] over silicon with high etch rates ($10^3 \mathring{A}\ minute^{-1}$).



Barrel Etcher



Planar Plasma Etcher



R.I.E (magnetron) Etcher

Figure 2.6. Common Dry Etcher Configurations.

Reactive ion etching uses a strong physical component of a chemical-physical etching mechanism. Wafers are supported on the driven electrode, in a planar reactor or specialist R.I.E. etcher configuration (figure 2.6). The wafers along with the driven platter achieve a negative potential of about half the peak R.F. potential. This can be several hundred volts negative relative to the plasma resulting in positive ion bombardment allowing high degrees of anisotropic etching. The grounded electrode is usually made much larger to avoid negative ion and electron sputtering. The only other difference between planar and R.I.E. etchers is that the pressure range tends to be lower operating at a few tens of mTorr. R.I.E has the advantage of controlled anisotropy[47,51] over planar etching. Its disadvantages are reduced selectivity, lower etch rates ($\approx 500 Å\ minute^{-1}$), and possible radiation damage.

R.I.E was not considered a viable process as recently as 1983[53] but its importance now[45,46,48,49] probably got its boost when commercial processing went smaller than 3 μm. They are now probably the mainstay of commercial microfabrication etching everything from silicon, polysilicon, and nitride to a variety of metals.

A recent advance is the magnetron reactive ion etcher,[42,54] shown in figure 2.6. It uses magnetic fields to increase electron paths in the plasma to increase the number of free radicals generated and boost the etch rate. This has made single wafer etchers, which have increased uniformity, possible. It also reduces the power required for a given etch rate thereby reducing the probability of radiation damage.

Etching chemistry.

The result of plasma etching a layer depends of course on the material etched but also the conditions and etchant gasses used. The amount of damage to substrate layers also varies with chemistry and conditions. The effect of various chemistries and conditions are briefly described below.

Table 2.3, after reference (43), summarises the effect of various parameters on the etch profile. Reference (41) have used these variables to achieve controlled variations in edge shape.

Etcher Parameter	Magnitude Favouring	
	Anisotropic Etching	Isotropic Etching
Reactant Concentration	Low	High
Gas Reactivity	Low	High
Unsaturate Concentration	High	Low
Product Volatility	Low	High
Gas pressure	Low	High
R.F. Power	High	Low
R.F. Voltage	High	Low

Table 2.3, Effect of etcher parameters on profile.

The actual chemical reactions occurring during plasma etching are not well understood and are certainly beyond the scope of this review. However, a simple summary is possible. Four main reactants are used for etching almost all materials; oxygen is used to etch organic polymers, silicon dioxide and silicon nitride use carbon-fluorine compound radicals as the etchant, and other materials are etched with either fluorine or chlorine based reactants. Common feed gasses are;

Group 1: Chlorine containing gasses to produce Cl radicals are; Boron-trichloride (BCl_3), Carbon-tetrachloride (CCl_4), Chlorine (Cl_2), and Silicon-tetrachloride ($SiCl_4$).

Group 2: Fluorine containing gasses to produce F radicals are; Carbon-tetrafluoride (CF_4), Silicon-tetrafluoride SiF_4, and Silicon-hexafluoride (SiF_6

Group 3: Fluorocarbon containing gasses to produce CF radicals are; CF_4, CHF_3, C_2F_6, and C_3F_8.

Group 4: Pure oxygen plasma is used to produce energetic oxygen to etch hydrogen and carbon containing films.

Other gasses, such as oxygen and hydrogen, are often added to gas mixtures to increase radical formation, alter polymer formation, or improve selectivity. Table 2.4

summarises etching of important materials.

Material	Feed	Reactant	Product (state)	Etch	Important Selectivities	Note
Polymers (Photoresist)	4	O	$H_2O_{(vapour)}$ $CO_{(gas)}$ $CO_{2\,(gas)}$	I	metals (high) silicon (high) oxide (high)	
Silicon	2	F	$SiF_{4\,(gas)}$	I	SiO_2 (10-40:1) Si_3N_4 (5-10:1)	1
	1	Cl, Cl_2	$SiCl_{4\,(volatile)}$	A	SiO_2 (10-50:1)	2
Silicon Dioxide	3	CF_x	$SiF_{4\,(gas)}$ $CO_{(gas)}$ $CO_{2\,(gas)}$	A	Si (10:1)	3
Silicon Nitride	3	CF_x	$SiF_{4\,(gas)}$ $N_{2\,(gas)}$	A	Si (15:1)	3
Tungsten	2	F	$WF_{6\,(volatile)}$	A	SiO_2 (high)	4
Copper	1	Cl, Cl_2	$CuCl_{3\,(non-volatile)}$	-	-	5
Aluminium	1	Cl, Cl_2	$AlCl_3$ (quasi-volatile)	A	SiO_2 (25:1) Si (1:2)	4,6, 7

A: Anisotropic Etching

I: Isotropic Etching

Table 2.4, Dry Etching Chemistry

Notes:

1: Oxygen is a common additive gas to stop unsaturates, C_xF_{2x}, and S_xF_{2x} from forming.

2: The anisotropic etch is due to physically enhanced reactivity.

3: Hydrogen is a common additive gas to remove fluorine radicals and increase

selectivity to silicon.

4: Etch initialisation is complex due to native oxides. Sputtering is usually employed to break through the oxide. Water vapour compounds the problem.

5: Copper-chloride deposits are a problem in Al-Cu alloy films. The sputtering component must be high to avoid residue.

6: Immediate post etch cleaning is important to stop HCl formation and etching.

7: Anisotropy is likely achieved by sidewall polymerisation (CCl_2).

2.5.3. Dry Plasmaless Etching.

There have been reports[50] of dry plasmaless etching of silicon using fluorine containing spontaneous etchants. Etchant gasses such as XeF_2, ClF_3, BrF_3, BrF_5 and IF_5 were all shown to spontaneously etch silicon with high selectivities over silicon dioxide. Although there are drawbacks such as isotropic etching, and highly expensive gasses, research interest is supported by the high selectivities achieved without any radiation damage.

2.5.4. End Point Detection.

Because of variations in etch rates both between wafers and runs in wet and dry etching it is essential to have some method of end-point detection. It allows some control over etch uniformity, line width control by avoiding overetching, and in poor selective etches avoidance of substrate layer damage.

In wet etching an operator looks for de-wetting to indicate that the etch has completed (liquid adheres to SiO_2 and Si_3N_4 but not bare silicon).

Dry etching poses the difficulty that the wafers are contained within a reactor. A small view-port is usually fitted to allow an operator to observe the colour changes associated with varying absorption of the thinning films. When the rainbow effects stop the layer is etched through. Although reasonably successful the technique is neither automatable nor highly consistent.

A number of automatable techniques have been tried but as yet there is no clear winner. The techniques are to monitor; the gas species or reaction product either through optical emission or mass spectroscopy, the film thickness on the wafer by laser

reflectometry, some parameter of the glow discharge, or by detecting a change in plasma impedance.

2.5.5. Section Summary.

Methods to selectively remove regions of a layer by either wet chemical or dry processing techniques has been described in this section. These techniques are necessary to pattern the layers to achieve some integrated circuit design. As will be shown in the next section, patterning alone is not sufficient but selective changes in the layers composition is also required to produce functional circuits.

2.6. Changing Layer Composition

The intentional introduction of electrically active impurities into a semiconducting crystal, or doping, is an essential step to all forms of integrated circuit fabrication.

Changes to the impurity concentration of a layer can be achieved by introducing dopants into a layer or the bulk crystal by either diffusion or implantation. Redistribution of the dopants in a layer can be achieved by thermal diffusion.

Other thermal techniques such as annealing, which can be used to re-order the crystal lattice and electrically activate the dopants, and sintering which is used to improve the interface between layers, also effect layer composition.

These techniques are considered in turn in each of the following sub-sections.

2.6.1. Diffusion.

Most books on integrated circuit processing have a section on diffusion.[2,4-6,8,55-57] A good introduction to diffusion and models is reference (5). References (55) and (56) contain the greatest detail.

Diffusion is the mechanism by which impurity or dopant atoms migrate through the semiconducting crystal. The diffusion action might be intentional, with a source of dopants supplied to the crystal surface in order to change the electrical characteristics of an area of the crystal, or, diffusion may be an unintentional side-effect of some other process where the total concentration of dopant atoms in the crystal remains constant but are distributed in a larger volume of the crystal.

Diffusion Mechanisms.

There are presently four recognised mechanisms for diffusion; vacancy substitution, interstitial motion, interstitialcy mechanism, and the crowdion mechanism.

Vacancy substitution requires unfilled crystal lattice sites, called point defects, in order to occur. A diffusing atom can take up position at a point defect, it then waits for another point defect to form in the direction of diffusion before moving into it and vacating the previous one for a following atom. The rate of diffusion depends on the point defect generation rate, which in turn depends on the temperature. Most dopants, usually group 3 and 5 elements, in silicon diffuse by this mechanism. For common dopants the temperature must be over 1000°C to give the lattice enough energy to generate sufficient point defects for appreciable levels of diffusion.

Interstitial motion refers to diffusion of impurity atoms through the space between the lattice sites. Since interstitial sites are normally vacant, diffusion is faster by this method as the diffusing atom need not wait for a place to go. As a result the activation energies and therefore temperatures are much lower. Impurities, especially group 1 and 8 elements which have small atomic radii, diffuse by this method.

The interstitialcy mechanism is essentially transport by interstitial motion but then the diffusing atom exchanges places with lattice atom which then becomes interstitial.

The crowdion mechanism of diffusion is based on crystal structure defects. A diffusing atom sits between two lattice sites but not interstitially, a lattice atom is displaced by its presense to a similar position to the opposite side,[55] resulting in displacement of the defect and the diffusing atom.

Modelling Diffusion.

If the doping concentrations are very high, or dislocations are predominate, or other diffusing impurities are present, then complicated mechanisms are required to model the diffusion.[55] For the simpler case Fick's law can be used to describe the diffusion process.

Fick's laws which were developed in 1855, to deal with constant isotropic diffusion, can be applied to impurity diffusion in silicon crystals since the lattice is cubic and therefore isotropic. Fick's first law states that the flux density of diffusing atoms is

proportional to the concentration gradient of the impurity or;

$$J = -D \nabla N \tag{2.21}$$

Where:

J is the flux density of diffusing atoms.

D is the diffusivity coefficient.

N is the concentration of the diffusing atoms.

For one dimensional diffusion into the surface from a source it can be simplified to;

$$J = -D \frac{\partial N}{\partial x} \tag{2.22}$$

Fick's second law, equation 2.23, can be derived from this using continuity and allowing the concentration to be a function of time.

$$\frac{\partial N}{\partial t} = D \frac{\partial^2 N}{\partial x^2} \tag{2.23}$$

It states that the rate of change of impurity concentration with time is proportional to the second spatial derivative.

Solutions for various boundary conditions, such as constant surface concentration (diffusion doping from gas, solid, or deposited surface dopant sources), or constant total concentration (redistribution after ion implantation) can be determined in closed form as a function of position and time if the diffusion coefficients are considered constant. Solutions for temperature and concentration dependent diffusivities can be determined with numerical methods on a computer. The effect of internal electric fields can be modelled similarly. References (5) and (4) give diffusion constants as a function of reciprocal temperature for common dopants.

The maximum concentration of dopant which can be dissolved in a solid at a given temperature is known as the solid solubility limit. It is an important factor to consider during diffusion as it limits both the maximum doping concentration‡ and the maximum surface concentration. For common dopants such as arsenic, phosphorus, or boron the solid solubility is around 10^{21} *atoms* cm^{-3}. References (4,5), and (8) give detailed graphs of the solid solubility limit as a function of temperature. The solid

‡ It is possible to exceed these limits using ion implantation and RTA.

solubility limit often makes modelling of diffusion doping simpler since as long as the ambient has abundant quantities of the dopant the surface concentration of the crystal can be assumed to be at the solid solubility limit. Therefore no complicated gas flows or dopant material diffusions need to be considered.

Some CAD modelling programs use a more sophisticated model which also considers other diffusion mechanisms and diffusion of impurities out of the wafer during heat treatment.

Diffusion Technology.

Dopant addition can be accomplished by diffusion from ambient gasses, solid sources, or spin-on sources, during heating in a standard furnace tube.

Gas source diffusion is usually performed in two stages. First the wafers are subjected to a lower temperature ambient of the doping gas to achieve a certain doping concentration. Then a high temperature neutral ambient is used to drive in the dopants to the required depth. The temperature ranges are 800 to 1200 °C. Table 2.5, contains a list of common dopant gasses, and liquids. Dopants from liquid sources are transported to the wafers by bubbling a carrier gas through a heated liquid bath. In all cases doping occurs from an oxide that forms on the wafer surface.

Dopant	Source	State
Boron	B_2H_6	gas
	BBr_3	liquid
Phosphorus	PH_3	gas
	$POCl_3$	liquid
Arsenic	AsH_3	gas
	$AsCl_3$	liquid

Table 2.5, Gas diffusion doping sources.

Although gas doping appears straight forward some dopant gasses present extreme safety problems due to their toxicity. One solution to that problem is to use solid-

source wafers which are non-toxic at room temperature.

Solid-source dopant sources have the same dimensions as wafers. They are sometimes referred to as planar diffusion sources.[58] In practice they are placed between the active surfaces of two product wafers. At high temperatures the sources are reduced and dopant atoms diffuse across the small gap to the surface of the product wafers. The sources usually stay in place for the entire doping step. Again surface oxides are formed during the process and must be removed before other processing steps. Table 2.6, lists some solid-source dopant materials.

Dopant	Source
Boron	BN
Arsenic	As_2O_3, $AlAsO_3$[58]
Phosphorus	P_2O_5

Table 2.6, Solid-Source Dopants.

Spin-on dopants as the name implies are applied directly to the surface of wafer by a machine similar to that used for photoresists. They are dopant incorporated glasses which are suspended in an organic solvent. The solvents will evaporate in about an hour at 200°C after which the film is ready for use. The wafers are then placed in a standard furnace with a nitrogen, and 1% oxygen getter, ambient at 800 to 1200 °C for around 1 hour to allow diffusion into the wafer surface.[59] The remaining glass is then removed by an oxide etch. Borosilicate glasses, and borazole (boron-nitrogen) polymers are common dopants. Arsenic and phosphorus sources also exist. Problems with doping profile consistency with spin-on dopants require careful control of deposition conditions such as humidity. Batch to batch variations can also be caused by the materials poor shelf life.

Another method of providing a source film for diffusion is to deposit SiO_2 with incorporated dopants using CVD techniques. After the source film has been deposited by either of the two methods a two stage process is used to achieve the required doping. First a short doping thermal process is used to dope the surface of the wafer, then the wafers are cleaned and transferred to a clean diffusion tube, where a long high

temperature drive-in stage is used. That method avoids the diffusion of contaminates into the wafer.

It is possible to define the area of a wafer that is to be doped by covering non-doped areas with an oxide mask. The oxide acts as a diffusion barrier to the dopant. The thickness of the oxide must be tailored to the diffusivity of the dopant and to the time and temperature of the diffusion step.

2.6.2. Implantation.

Ion implantation is a widely used alternative to diffusion for introducing dopants into specific areas of integrated circuits. Ion implantation involves production of an ion beam of dopants, selection of specific dopant species from that beam, and acceleration of the selected ions to sufficient energy to penetrate a specific distance into a silicon target wafer. Energy ranges of 3 KeV to 2 MeV have been used to give concentrations between 10^{14} and 10^{21} *atoms* cm^{-3} and depths between 100Å and approximately 1 μm. Ion implantation has advantages over diffusion which have allowed the present state of integration. However implantation and diffusion should be thought of as complementary technologies not competing ones. Table 2.7, shows some comparisons between the two technologies.

There are a number of reference sources on ion implantation,[2,4-7,60-66] including dedicated books.[67,68] Reference (61) gives an excellent review and both references (67) and (68) give good depth on some of the problems.

Comparison	Ion Implanter	Diffusion Sources
Range of Doping Concentrations (*Atoms cm*$^{-3}$)	10^{14} to 10^{18}	10^{17} to 10^{23}
Depth of Profile	<1μm ≈2μm with 1MeV	up to 25μm (shallow control poor)
Dopant Profiles	Engineerable well controlled	Limited to gaussian distributions and error functions
Masking Materials	oxides polysilicon nitride photoresist (physical blocking)	oxides polysilicon nitride (diffusion blocking)
Processing Temperature	Low (room temp.)	high (800-1200 °C)
Time Step	10 minutes (medium dose)	Several Hours
Lateral Spread	Low (scattering)	High (diffusion)
Abrupt Junctions	Possible	Very difficult
Damage Caused	Crystal Lattice Disorder Possible Radiation Damage	Wafer Warpage Surface Contamination
Control of Dose	easy	temperature and time dependent
Dopant Purity	High (mass selected)	Dependent on source
Concentration Uniformity	excellent	good - for high doses poor - for low doses
Equipment Cost	High	Medium
Equipment Complexity	High	Low
Throughput	good	high

Table 2.7, A comparison of diffusion and implantation doping.

Profile Control

Impurity profile control can be broken into two aspects, total dose and position. The total dose is easily controlled since it is basically the product of the ion current and the implant time. The position of the implanted ions is slightly more complicated.

The mean range of the implanted ions is the position where the ions have, on average, lost all their incident energy through collisions with lattice atoms. The range is therefore proportional to the implant energy for a given dopant species and target. There are two energy loss mechanisms in operation. Elastic (ion-nuclei) interactions which cause both energy reduction and incident ion trajectory changes, and inelastic (electronic interactions) which cause only energy loss. Nuclear collisions are important for lower energy higher atomic mass implants, and electronic stopping is more important for the higher energy lower mass implants.

The spread of the range is statistical in nature with some ions suffering fewer collisions and penetrating deeper than others which stop quicker after more frequent collisions. The spread of the range is also incident energy dependent.

Vertical Distribution.

On a first order approximation the distribution of ions in the crystal can be considered a gaussian profile because of the statistical nature of ion stopping. There is however a distinct skewness to the distribution giving a tail deeper into the substate.

Although Lindhard, Schoff, and Schiott have developed a widely used theory for the range of implanted atoms in an amorphous material called LSS,[67,68] it is usually an empirical model called the Pearson-IV solution which is the basis of CAD programs for ion implantation. It considers; projected range, projected range straggle, skewness of straggle, and kurtosis of the profile. It ignores channeling of ions, and two dimensional effects. Modelling of implantation is further complicated by implanting though an oxide, either the 30 angstrom native oxide, or an intentional thicker oxide designed to prevent surface contamination and to direct more of the implant to the surface of the semiconductor.

Silicon crystals have a structure which if aligned correctly to the impinging ion beam presents channels through which the ions may pass with relatively fewer collisions

than they would have in an amorphous solid. This behaviour is called channeling, and presents a problem in controlling ion range since the channeled range may be 2 to 50 times larger than the amorphous target range.[64] Once channeling starts collisions steer the ion along the channel deeper into the substrate which accounts for the long tails on the impurity profiles.

The standard solution to this problem has been to tilt the plane of the incident ions relative to the wafer so that the wafer appears amorphous. However some ions can still channel after nuclear collisions steer them into channels. The standard anti-channeling angle of 7 degrees, which was taken as half the angle between the major <100> channeling axes of <001> and <133>, has recently been shown[64] to have been just a lucky choice (since it is very close to a minor channeling axis) by a map of the entire channeling characteristics which suggest that just under 4 degrees would have been a better choice.

Another solution to the problem is to pre-implant silicon or argon into the target wafers to amorphise the surface.[66] Dopants can then be implanted normal to the surface which has a number of advantages in certain process designs.

Both the anti-channeling implant angle and the lateral straggle of ions due to direction changing collisions cause appreciable concentrations of dopants under the implant mask edges. Since the channel length of MOS transistors depends on this masking, lateral straggle is of increased importance in small geometry processes. The amount of lateral straggle is roughly proportional to the depth of the implant. It is about 50% of the boron, 25% of the arsenic, and 40% of the phosphorus doping profile depth. These values are usually increased by unintentional diffusion during subsequent thermal processing.

Similar statistics and modelling to the vertical profile can be applied to achieve computer models of the lateral profile. There are some computer processing simulators which can perform two dimensional modelling of implant profiles.

Because of nuclear collisions during implantation the crystal structure of the semiconductors become disordered.[2,67] In some cases of higher doses the surface can become amorphous and annealing of this damage is required after implantation.

Implanter Apparatus.

There are three different types of implanter design which can be classified by the beam current being low (100μA), medium (1mA) or high (10mA). Low and medium current machines are similar. High current machines are the most popular. Reference (62, 67) compares target scanning mechanisms for uniformity and complexity. A schematic high current implanter, shown in figure 2.7, will be briefly described here.

The main components of implanters are Ion Sources, Mass Analyser, Accelerator Tube, Beam Line, Implant Chamber, Vacuum Systems.

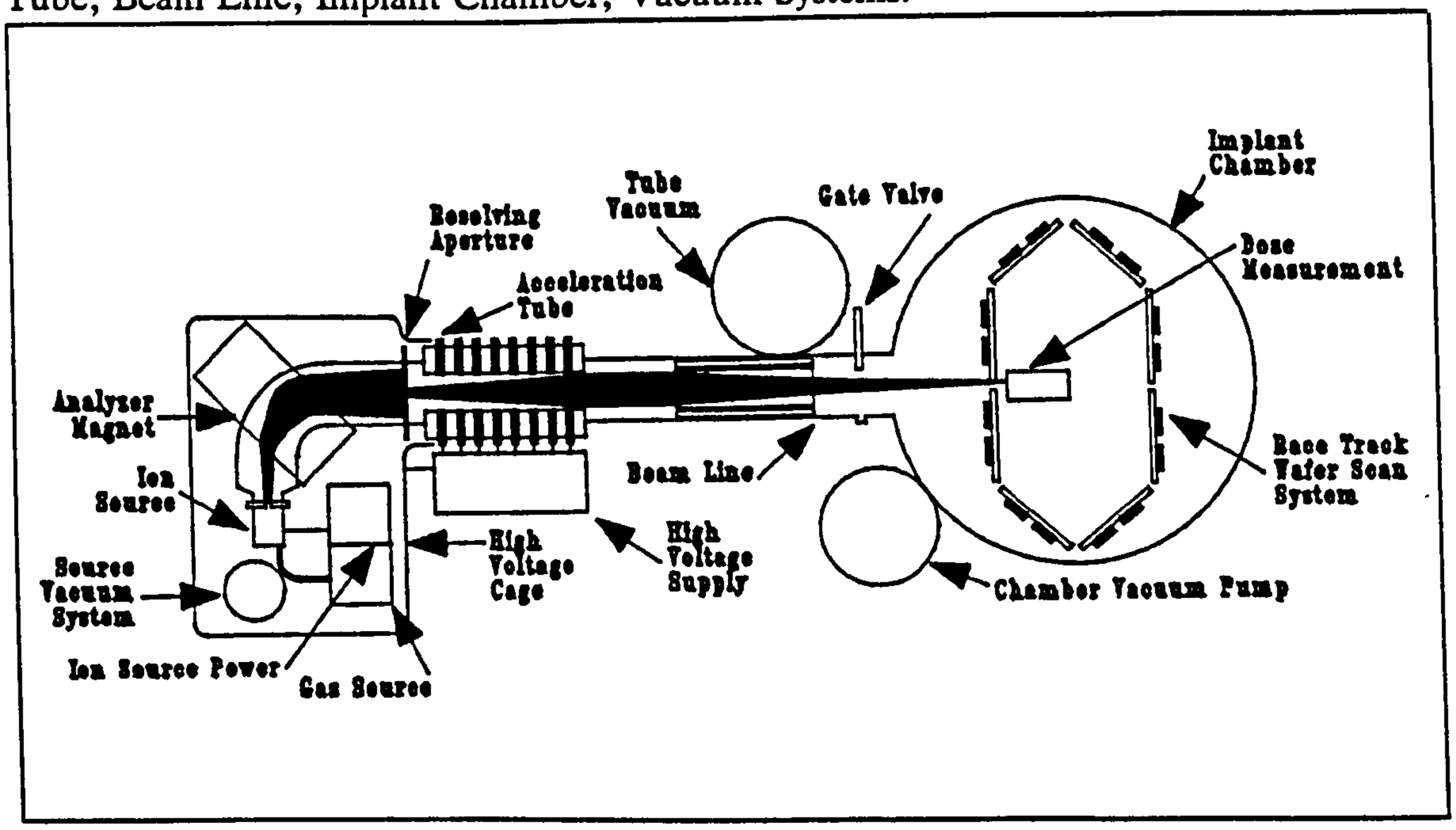


Figure 2.7, A schematic of a high current implanter.

Most ion sources produce ion beams from an ionised plasma of a dopant feed gas created between an anode and cathode. High current machines use helical hot filaments or hot tungsten rods as cathodes. Often magnetic fields are used to increase ionisation efficiency. An ion beam is extracted through a slit aperture and across a pre-analysis acceleration gap (usually between 25 and 50 KeV). Most feed sources are gasses to allow quick change of dopants, however solid sources are also used to avoid toxic gas problems and to allow special dopants. Solid sources are vapourised in ovens to produce ionised species. Table 2.8 lists some of the possible dopant sources.

Dopant	Source	State
Boron	BF_3	gas
	BCl_3	gas
Phosphorus	P	solid
	PCl_3	gas
	PH_3	gas
	PF_3	gas
Arsenic	AsH_3	gas
	AsF_3	gas
	As	solid
	$GaAs$	solid
Silicon	SiF_4	gas
Antimony	Sb	solid
	Sb_2O_3	solid

Table 2.8, Common Ion Source Feed Material.

Mass selection is required to separate the compound ion beam into the separate components so that the impurity profile can be controlled. Mass selection is invariantly done by magnetically bending the beam through 60 or 90 degrees (see figure 2.7). The magnetic field is adjusted so that only the desired dopant is introduced into the acceleration tube through a selection slit. For example BF_3 ionises into (in order of frequence); B^{11+}, BF_2^+, F^+, BF^+, B^{11++} and isotopes. Obviously B^{11+} is normally selected, but B^{11++} is sometimes used to get twice the energy out of the accelerator, and BF_2^+ is used to get minimum energy boron for shallow implants.[65] Non-selected ions simply fail to make it through the slit and after sputtering a disposable selector liner get taken away by the vacuum system.

Since modern implanters use post analysis acceleration to improve beam current control and reduce the size of the analysis magnets, the ion sources and analysis system operate at elevated potentials and for safety are remotely controlled by fibre-optic

controls.

The acceleration tube consists of glass rings with aluminium electrodes between them. Resistors divide the acceleration potential evenly along the tube to give a constant acceleration. The final stage is usually a 2KeV decelerator to suppress electron bombardment and X-ray generation. Although the tube is designed to keep the beam together it is divergent as it leaves the tube. Modern very high potential systems use twin accelerators.[63]

The beam line contains the final focusing elements for the beam. Electrostatic or quadrapole lenses are used. A neutral trap is also provided for machines which use electrostatic scanning.

Since the beam is not large enough or does not have enough uniformity to implant an entire wafer at a time it must be scanned over the wafer. Low and medium current machines are single wafer oriented and tend to have electrostatic scanning.[67] High current machines are batch oriented and characterised by their large implant chamber housing some type of mechanical scanning system.[62] The target chamber also usually contains the mechanisms for monitoring the dose. The current of a Faraday Cup[67] is often used to do this.

There are a number of vacuum systems required in modern implanters. They are usually either oil diffusion pumps, with liquid nitrogen traps to reduce contamination of wafer surfaces with pump oil vapour, or cryogenic pumps, which are popular in newer systems as they completely avoid oil vapour problems. Table 2.9 lists some of the requirements of an implanter vacuum system. The lower pressure in the flight tube is required to avoid beam blowup and self neutralisation.

Area	Pressure	Special Requirements
Ion Sources	10^{-3} Torr	Elevated Voltages
Flight Tube	$<10^{-5}$ Torr	Low pressure
Implant Chamber	10^{-5} Torr	Fast Pumping

Table 2.9. Implanter Vacuum Requirements.

2.6.3. Annealing.

The lattice structure of the semiconductor crystal becomes disordered or damaged during ion implantation, and because crystal order is necessary for good electrical characteristics, the order of the crystal must be restored. The thermal treatment required to achieve this is called annealing.

A separate problem which exists after ion implantation is that most dopant atoms do not occupy a lattice site, and as such are not electrically active. Diffusion during annealing allows them to take up lattice occupancy and become electrically active, which is why annealing is sometimes called activation.

Most books which have a section on ion implantation also mention annealing.[5,6,61] References (60, 67), and (2) however have good detail on the subject. Since annealing is a thermal process it also gives rise to unwanted re-distribution of dopant concentrations. Rapid Thermal Annealing (RTA) has recently been developed to avoid that problem, and judging by the number of references available,[65,66,69-72] it is quite a hot research topic.

Furnace Annealing.

A standard furnace tube, figure 2.3, with a nitrogen ambient can be used to anneal silicon wafers. Temperatures in the range of 600 to 1000 °C are used for periods around 30 minutes. Lattice damage is either healed by thermal vibration in low doping cases, or by solid-phase epitaxy in heavily damage amorphous regions. Almost all dopant atoms will become electrically active during this process. Boron implantation rarely produces an amorphous layer and therefore requires a higher annealing temperature of around 900°C, Phosphorus and arsenic implants usually produce amorphous layers with can be annealed by solid-phase epitaxy at temperatures around 600°C. Solid-phase epitaxy rates of up to 500 Å per minute are possible making anneal times as short as 10 minutes possible which reduce the amount of dopant redistribution. Obviously annealing by solid-phase epitaxy has some important advantages.

Electrical activation of non-amorphous layers, such as boron implanted silicon, has three possible mechanisms[60] depending on the furnace annealing temperature, with the maximum dopant activation requiring 30 minutes at 900°C. Solid-phase epitaxy

produces total dopant activation during the short anneal.

Rapid Thermal Annealing.

Rapid thermal annealing uses high power sources to heat the circuit side of a wafer quickly to high enough temperatures for solid-phase epitaxy to restore crystal order and activate dopant atoms. A good review of available equipment can be found in reference (69). Table 2.10, compares RTA technology to furnace annealing for boron doping. In all systems the wafer is thermally isolated so radiation dominates the heating and cooling processes. Ultra short annealing is possible with laser beams,[60] ion beams, and electron beams.[2] These methods heat the wafer surface above its melting temperature to allow liquid phase epitaxial re-ordering and activation. They are mostly experimental as there are uniformity and stress problems caused by their small spot sizes and therefore surface scanning requirement.

Equipment					Example (Boron Implant Anneal)			
Energy Source	Throughput	Wafers/hr	Orientation	Anneal Mech	Wafer Surface Temp. (°C)	Anneal Time (sec)	Diffusion Distance (Å)	Notes
Furnace Tube	360	B	Diff	950	1800	1000-2000	1
Planar Graphite Heater	200	W	SPE	1100	10	800	2
Multiple Tungsten-Halogen Lamps	60	W	SPE	1100	30	200	3
Single Water Wall Arc Lamp	60	W	SPE	1100	25	Unavailable	4

SPE: Solid Phase Epitaxy.

Diff: Diffusion.

Notes:

1: Possible Contamination. Surface requires post anneal cleaning.

2: Pulse heating Mode Only.

3: Uniformity requires careful consideration. (Example is BF_2 implant).

4: Needs a lot of support equipment. Thermal ramping is possible.

Table 2.10, Comparison of Conventional and Rapid Thermal Annealing.

2.6.4. Sintering.

Sintering is a technique used to make good contact between dissimilar materials. There are two applications in microfabrication, aluminium-silicon contacts, and silicides.[18,20,73] There are two requirements of a sintering process; it must promote the breakdown of native insulting films, such as SiO_2, between the materials and provide interdiffusion mixing of the two materials.[73]

Drastic improvements in conductivity of aluminium-silicon contacts can be made by sintering. Aluminium is capable of reducing the native SiO_2 barrier thereby reducing the resistance.[19] However, great care must be taken in choosing and maintaining a sintering temperature to avoid aluminium spiking into the silicon. The chosen temperature must be such that the solubility of silicon in aluminium at that temperature is equal to the alloyed concentration of silicon in the aluminium film. For the common 0.5% silicon in aluminium alloy that temperature is 435°C.

Silicides can be produced, as noted earlier, by a number of methods. One common method is to selectively deposit a refractory metal over polysilicon. Sintering is then used to mix the two layers at their interface.

Sintering can be performed in a standard furnace tube, figure 2.3, with a nitrogen ambient. Between 400 and 500 °C for 30 minutes is used for aluminium-silicon contacts. Rapid thermal sintering is also a possibility.

2.6.5. Section Summary.

The change of layer composition through dopant addition using either diffusion sources or ion implantation, the re-ordering of crystal structure and electrical activation of dopants using thermal annealing, and improvements in layer interfaces by sintering were discussed in this section. These techniques play important roles in the electrical characteristics of integrated circuits whereas the reliability and yield of integrated circuits are the main considerations in the next section.

2.7. Changing Layer Topography

The need for topography smoothing techniques is largely a recent one caused by denser packing in VLSI processes. The first problem encountered was achieving adequate step coverage at edges of contact holes. The second problem, a more recent one, is the need to use short depths of field with projection lithography tools (see the later section on lithography). Both of these problems are solved by some topography modifying technique.

2.7.1. Step Slope Modification.

With the advent of dry anisotropic etching came the problem of depositing metal on the sides of vertical contact holes. Stresses at the corners can be high enough to cause cracking and conductor failures. A popular solution has been to use dielectrics which can be reshaped by thermal reflow leaving sloped sidewalls and rounded corners on contact holes.

Heating of a layer to lower its viscosity and to allow it to reflow over sharp steps is a popular method of solving step topography problems. Phosphorosilicate glass (phosphorus doped silicon dioxide) or P-Glass is a popular reflow glass in many technologies.[2] P-Glass will reflow at 1050°C in about 20 minutes to a fixed angle dependent on the phosphorus concentration. Angles range from as deposited to 45 degrees with concentrations from zero to 6%. Concentrations of higher than 8% are avoided because of aluminium corrosion by the phosphorus. The range of 6 to 8% is typical.

Borophosphosilicate glass (BPSG)[70,74] reflows to a greater extent at lower temperatures than P-Glass. Rapid thermal processing [74] has also been applied to glass reflow. Angles of 30 degrees have been achieved in 10 seconds at 1100°C for P-Glass (8%) and in 20 seconds at 1030°C for BPSG (4%) without significant dopant redistribution. RTP reflow looks good for future processes.

2.7.2. Planarisation

The purpose of planarisation is two fold. One is to reduce depth of field so that fine lines can be reliably patterned onto higher levels. The second is that lower selective dry etches require more uniform films to avoid damaging underlying layers.

Second level metal layers must be patterned on a planarised surface or otherwise thickness variations to 2 and 3 times occur near steps in underlying layers.

There are a number of techniques to produce planarisation levels, two of which are covered in the next section. A wholly different approach to the problem is to design processes which are inherently planar instead of the popular LOCOS[3] process. Such schemes exist and a popular one for dense memory structures is trench isolation.[2]

Etch Back Planarisation.

Etch back planarisation is a straight forward approach.[75] A dielectric layer of greater than required thickness is deposited. Photoresist is spun on over the dielectric and then hard baked above its reflow temperature (≈200°C). The wafers are etched in a dry anisotropic etch with 1:1 selectivity of resist to dielectric until all the resist is removed. The result is a planar dielectric surface. The drawbacks are that there must be tight controls over film thicknesses and etching conditions.

A variation is bias sputtering,[75] which uses physical etching to sputter away the high points of the dielectric layer and to allow recombination to deposit dielectric at the low points. It provides for better step coverage but is not a true planarising processes and suffers the disadvantage of incorporating impurities in the recombinate dielectric.

Spin Coated Films.

Low viscosity liquid source films can be used to planarise substrates and to act as dielectrics. The method is similar to etch back; coat the wafer with the film, bake the wafers to drive off volatile solvents and to provide some reflow, and the surface is ready. The drawbacks are that surfaces are never perfectly planar, etch back gives better results, and the purity and shelf life of the films are a problem. Some spin coated films are; polyimides (organic polymers), epoxies[76] and spin-on-glass (SiO_2 particles in a volatile organic solvent).

2.7.3. Section Summary.

The techniques of topography modification, allowing increased yield and reliability, were the topic of this section. The next section on support functions describes other techniques which, although important to producing good products, are also not always apparent.

2.8. Support Functions

There are many supporting functions which are essential to microfabrication. They range from the obvious such as measurement techniques, to those that are sometimes taken for granted such as cleaning and gettering. The following sub-sections consider each in turn.

2.8.1. Measurement tools.

Measurements are required to control microfabrication processes. Feedback is necessary to get the desired film thicknesses, compositions, patterns and impurity profiles. Table 2.11 contains a list of measurement techniques and the following subsections describe their principles of operation. Only techniques commonly used for microfabrication are discussed here, other research techniques are covered in references (77) and (78).

Measurement Techniques.					
Technique	Measures	Uses	Destruct -tive.	Problems	Ref.
Spreading Resistance	Impurity Profile	Current	Yes	can damage surface, destructive for profiling.	6, 55, 67, 79-81
4 point Probe	sheet resistance and chemical composition	Voltage	No	can damage surface	3, 6, 55, 67, 79, 81
CV	Impurity Profile	Capacitive	Dedicated	requires expert data interpretation	55, 67
Electro-Chemical	Impurity Profile	"	yes	Very slow but enhanced resolution	3, 67, 80
SIMS	Chemical Composition	Secondary Ions	yes	profile can be distorted by knock-on effects	67, 77, 80, 82
Auger Spectroscopy Scanning	Chemical Profile	Auger Secondary Electrons	yes	Trade off between spot size and sensitivity	8, 78, 83
	Surface Map		no		

Measurement Techniques.					
Technique	Measures	Uses	Destruct -tive.	Problems	Ref.
Optical Microscope	Morphology	light	no	limited to technologies > 1μm.	8,77
SEM	Morphology Topography	Reflected Electrons	no yes	E-beam induced damage and surface contamination	8,77,78,84
TEM	Topography	Transmitted Electrons	yes	requires special preparation.	8,77,78
Interference Thickness Measurement	Film Thickness	Reflected Intensity	no	need to know refractive index.	6
Ellipsometry	Thickness	Polarisation	no	resolution poor in places due to graphical computation	6,56
Electro-mechanical Transducers	Topography	Stylus Displacement	no	can damage surface.	6,8

+ Achieved by angling sample and measuring laterally.

* Nomarski interference method.

Species dependent.

" As Spreading Resistance or CV.

Table 2.11, Measurement Techniques.

Measurement Techniques.						
Technique	Range (field)			Resolution Limit		
	Depth	Laterial	Chemical	Depth Å	later -ial	Chem -ical %
Spreading Resistance	1 → 20 μm	Wafer	10^{14}→$10^{20} cm^{-3}$	2000	20μm	10
4 point Probe	>1μm	Wafer	10^{1}→$10^{6}\Omega/\square$	0.2μm	1mm	0.5
CV	up to 100μm	Wafer	10^{13}→$10^{18} cm^{-3}$	1000	1mm	5
Electro-Chemical	"	"	"	"	200	"
SIMS	1 → 2μm	sample size	10^{15}→$10^{23} cm^{-3}$ #	25 → 120	0.5 → 5	10ppm
Auger Spectroscopy	as above	as above	10^{19}→$10^{23} cm^{-3}$	as above	100μm	100ppm
Scanning	-	500→3000Å	10^{19}→$10^{23} cm^{-3}$	-	300Å	0.01
Optical Microscope	$1\times res.$	40μm → 10mm	-	200 *	2000Å	-
SEM Morpho. SEM Topo.	$300\times res.$ $300\times res.$	1μm → 10 mm	- -	100Å + 100Å	100Å 100Å	- -
TEM	-	100 → $500\times res.$	-	-	10Å	-
Interference Thickness Measurement	>150Å	Wafer	-	14Å	10 μm	-
Ellipsometry	1 → 2400Å	Wafer	-	14Å	1 → 3 mm	-

Measurement Techniques.						
Technique	Range (field)			Resolution Limit		
	Depth	Laterial	Chemical	Depth Å	later -ial	Chem -ical %
Electro-mechanical measurement	25Å → 100μm	10μm → 2.5 mm	-	25Å	2.5μm	-

+ Achieved by angling sample and measuring laterally.

* Nomarski interference method.

Species dependent.

" As Spreading Resistance or CV.

Table 2.12, Measurement Techniques (cont.).

Spreading Resistance.

Spreading resistance is used to determine the resistance of doped silicon. Two tungsten probes, approximately 4μm in diameter and separated by 10 to 20 μm, are pressed in contact with the wafer surface. A small voltage is applied and the resulting current is measured. Then the measured resistance, which can be considered localised at the probe tips, can be used to determine the resistivity of the material which is also a measure of the impurity doping concentration.

Bevels are sometimes cut into samples to allow impurity profile measurements. The technique suffers from frequent need of calibration and is mostly used for comparative measurements.

Four Point Probes.

A linear array of four equally spaced tungsten probes can be used to measure the sheet resistance of a thin film. Two arrangements are used; current is either forced between the outer probes, and voltage measured on the inner ones, or current is forced between probes 1 and 3 and voltage measured between 2 and 4, in order to determine the local sheet resistivity. This procedure can be repeated over the surface of the wafer to determine the uniformity of doping.

Capacitance Voltage Measurements (CV).

Dedicated measurement structures with one electrode of a capacitor being the semiconducting crystal can be used to determine impurity doping concentrations. Suitable structures are reverse biased pn junctions and MOS capacitors (which can also be used to analyse $Si-SiO_2$ interface conditions).

The technique works by measuring the depletion layer capacitance for various bias voltages. The depletion layer capacitance depends on the ratio of electrically active impurities on each side of the depletion layer. The depletion layer depth increases with increasing bias. An impurity profile by depth can be determined by plotting a function of the depletion capacitance versus a function of the bias voltage.

Electro-chemical techniques.

Spreading resistance and CV techniques for measuring impurity profiles can be improved by slowly removing the wafer surface whilst analysing the doping. The electro-chemical technique uses an electro-chemical reaction to remove a small controlled thickness from the sample and an electrical technique to determine the doping concentration. Controlled surface removal can be achieved by a constant current anodising arrangement followed by an oxide etch. Measurements can then be made with a spreading resistance probes. Another method is to use a non-anodising electrically driven etching reaction with con-current AC capacitance measurements, (using the electrolyte as one electrode), to determine the doping profile.

Secondary Ion Mass Spectrometry (SIMS)

Secondary ion mass spectrometry is a technique for analysis of the chemical composition of the substrate as a function of depth. The technique involves; placing the sample in a high vacuum, sputtering the surface with a beam of 5 to 15 KeV ions, collecting the secondary ions sputtered from the surface, mass analysing the secondary ions to determine composition of the surface, and building up a surface profile with depth as the surface is slowly sputtered away.

The technique has the advantage and drawback in that the chemical composition and not the electrically active dopant concentration is what is measured. The technique also suffers from the knock-on effect whereby the sputtering ions drive dopant ions farther into the sample.

Auger Electron Spectroscopy

Auger electron spectroscopy provides a method of surface chemical analysis. Auger electrons can be emitted from a sample if an incident electron or photon succeeds in knocking an inner shell electron from its orbit and starts a chain reaction. The reaction continues with a higher energy orbital electron dropping into the vacancy transmitting energy to an outer valency (Auger) electron which is emitted. The energy of the emitted electron is characteristic of the atom.

Auger electron spectroscopes use an electron beam to initiate that reaction and then capture and measure the Auger electrons in order to determine the surface composition of the sample. Since Auger electrons only have an escape depth of about 50Å, the spectroscope is sometimes used in conjunction with a ion sputtering system to give depth profiles.

Optical Microscope.

Optical microscopes provide morphological information for control of pattern transfer. In almost all cases the microscope images light reflected from the sample.

One variation is the Nomarski Interference Microscope. It breaks the incident light beam into two polarised beams and introduces a path extension into one beam. The reflected light, which is the reconstituted beam, is viewed. The result is enhanced

contrast between edges of different layers through either constructive or destructive interference which provides some depth resolution to the microscope.

Another variation, which has been automated using computers for image recognition, is the image shearing microscope which is used for measuring line widths. The reflected light from the sample is broken into two paths, which are separately coloured in the manual models, and then recombined introducing a known shift in the path of one beam. The result is an overlapping binocular view of the sample. By adjusting and measuring the shift required to misalign the image to each side of the line to be measured the line width can be determined.

Scanning Electron Microscope (SEM).

A scanning electron microscope can be used to examine the surface morphology of a patterned integrated circuit. If the sample is cleaved or cut and polished it is also possible to obtain a cross-section of the topography.

A scanning electron microscope consists of;

1) Electron Gun.
2) Focus and scanning electron optics.
3) An electron reflecting sample target.
4) A scintillating crystal.
5) A photomultiplier.
6) Vacuum system for the electron path.

The SEM works by scanning an electron beam of 1 to 40KeV over the sample. The back-scattered and secondary electrons from the sample are detected by the scintillator and photomultiplier whose current can then be used to modulate the intensity of a linked scanning display.

Possible problems with this system are; electron beam induced damage, and surface charging of insulators. Destructive gold plating can reduce the charging problems. Modern systems with low intensity single scan beams and digital picture storage and enhancement can also get around these problems.

Analysis of x-rays produced as the beam hits the sample can give surface chemical indications. Using selective etches on cross-sections before examination can delineate

metallurgical junctions.

Transmission Electron Microscope (TEM).

Transmission electron microscopes are similar to electron microscopes except that higher energies, (60 to 350KeV), are used to completely penetrate a thin film sample of the material to be studied. The transmitted electrons are collected, as with the SEM. The relative intensity variations of transmitted electrons as the beam is scanned over the sample can be used to determine crystal and even atomic structure. The major drawback of this technique is that samples are difficult to prepare.

Interference Film Thickness Measurements.

Non-opaque dielectric film thicknesses can be measured by using the interference of the multiple internal reflections to alter the reflected intensity of a particular wavelength of the illuminating light. Measurements are usually made by focusing an optical system on the sample, and scanning a optical spectroscope through the range of reflected light noting the spectral frequency of the successive maxima and minima, from which the film thickness can be determined.

Ellipsometry

Dielectric film thicknesses can be determined by ellipsometry along with the dielectric constant if certain optical properties of the substrate are known. Ellipsometry is based on the relative change in phase angle of a circularly polarised wavefront reflecting from a thin film. The procedure is to illuminate the sample with an elliptically polarised light beam from a NeHe laser, then pass the reflected beam through a rotatable polarising filter on its way to an optical intensity meter, and finally rotate the polarising filter to determine minimum intensity. The angle of the polarising filter can then be used to determine the film thickness and dielectric constant.

Electromechanical Transducers.

Electromechanical transducers can be used to determine surface topography by dragging a stylus across the area of interest. An electromechanical transducer generates a signal from the stylus motion which after calibrated amplification can be used to determine the surface variation with distance. The method is best used for determining thicknesses of layers with edges and on the edge shape.

2.8.2. Cleaning

Since MOS transistors are surface devices the cleanliness of the wafer surfaces are essential to achieve reasonable yield and large scale integration. Attempts are made to keep the environment of the wafers clean by using "clean" air laminar flow units over work surfaces and by having personnel dress in particle barrier "clean room suits". However the wafers still require cleaning to remove particles after certain processes.

Most steps only require a thorough wash in de-ionised water (D.I. water) and a spin dry in warm nitrogen to remove physical particles. D.I. water is used to avoid contamination, especially sodium which can adversely affect surface physics.

Some processing steps such as thermal oxidation of silicon, and especially before the critical gate oxide step, require a chemical clean to prepare for high quality film growth. A popular four step process[3,10,11] called the 'RCA Clean' is as follows.

1) Chemically oxidise organic contaminates.

2) Etch away native oxide and oxides formed by step one.

3) Wash in D.I. water.

4) Spin Dry in warm nitrogen.

Step one is achieved by a mixture of de-ionised water, hydrogen peroxide and ammonium hydroxide. The hydrogen peroxide oxidises organics and provides some mechanical agitation by hydrogen gas bubble formation. The ammonium hydroxide removes some organics by forming a solution with them. The combination can also complex some group 1 and 2 metals such as copper, silver, nickel, cobalt, and cadmium. A final cleaning with hydrochloric acid in hydrogen peroxide can remove heavy metal impurities completes step 1. A dip in dilute hydrofluoric acid in D.I. water

removes the oxides in step 2. After a final multiple spray rinse in D.I. water the wafers are dried and ready for oxidation.

Other solutions are possible, and oxygen plasmas have been used in place of the oxidising chemicals to reduce chemical costs and environmental impact.

2.8.3. Gettering.

Gettering of contaminates is an important part of semiconductor processing. It drastically improves the performance of the silicon-silicon dioxide interface by trapping impurities away from the surface where devices are built.

Gettering can be classified into four groups;[85]

1) Pre-Oxidation Gettering to the Other side. (POGO).
2) Internal Gettering (IG).
3) Silicon Gettering by Segregation (SGS).
4) External Gettering. (EG).

Gettering action requires three physical effects; the release of impurities or decomposition of extended defects, diffusion of impurities or defect constituents to a capture zone, and the capture of the impurities to some sink.

The best references for gettering are (1, 60) and (86) which lists a table of gettering mechanisms.

P.O.G.O

Pre-oxidation gettering to the other side, or damage gettering, traps heavy metal impurities at extended defects on the backside of the wafers. Defects are impurity sinks because they locally increase the solid solubility of the impurities in their region. Impurities diffuse quickly[1] to these sites during thermal treatments.

Backside damage can be made by; sand blasting, laser melting, ion implantation, silicon nitride deposition (thermal stress), backside polysilicon (dislocations at grain boundaries which do not anneal out), and by high concentrations of diffusing phosphorus.[87]

Internal Gettering.

Internal gettering, often called intrinsic gettering, uses the wafer bulk as a capture zone. A pre-processing wafer preparation step denudes the surface of oxygen, incorporated during crystal growth, by high temperature evaporation. The wafer is then annealed around 800°C to supersaturate oxygen and SiO_x precipitates at the denuded boundary. Those precipitates locally increase the solid solubility of heavy metal ions resulting in an effective capture zone.

S.G.S.

Silicon Gettering by Segregation, also called ion pairing, is a reaction which occurs between phosphorus and metal impurities. Two mechanisms are possible. Phosphorus diffusion in the bulk segregates impurities by driving them ahead of the phosphorus concentration (n^+ wafers only). Phosphorus donates a large number of electrons to the metals (Cu, Au, Fe, etc.) which then attach themselves to the phosphorus atom by coulomb bonding and are dragged along with it as it diffuses at temperatures above 900°C. Ion pairing with phosphorus at lower temperatures in P-Glass dielectrics is important for gettering metals, such as Na, at the latter stages for processing.

External Gettering.

External gettering, which is also called ambient gettering, removes heavy metal ions to the ambient gasses during oxidation with some chlorine present. Dissociated chlorine from HCl gas included in the gas flow during oxidation can form heavy metal chlorides[88,89] from some impurities producing volatile chlorides which are removed through the ambient gasses. Other chlorine sources are being suggested[88] since HCl may attack its gas bottle and actually supply impurities.

2.8.4. Section Summary.

Support functions like measurement, cleaning, and gettering techniques were described in this section. Although they are often not explicitly noted in a process description, those functions are essential to high yielding reliable processes.

2.9. Example Process (E.M.F. 6 μm NMOS).

This section contains an example process of the Edinburgh Microfabrication Facility's 6 micron NMOS process, in order to show the way that processing steps are combined in order to achieve integrated circuits. With the help of cross-sections and impurity distribution profiles generated by computer simulation each step in the process will be briefly discussed.

2.9.1. Computer Simulation.

The CAD tools SUPRA, a two dimensional device simulator, and DEPICT, a topography modeling tool, (both products of Technology Modelling Associates) and some custom fortran programs, to model re-flow steps, were applied to model an enhancement n channel MOSFET. The minimum design rule layout was taken as the simulation input.

2.9.2. Process Steps.

The E.M.F. 6μm NMOS process is a LOCOS isolation, polysilicon gate process with one metal layer. It is a seven mask process with eight photolithography steps (the contact mask is used twice). The process flow is shown diagrammatically in figures 2.8 and 2.9 with the lithography subprocess shown in figure 2.10. The following paragraphs will describe some of the importance of each of the 61 steps.

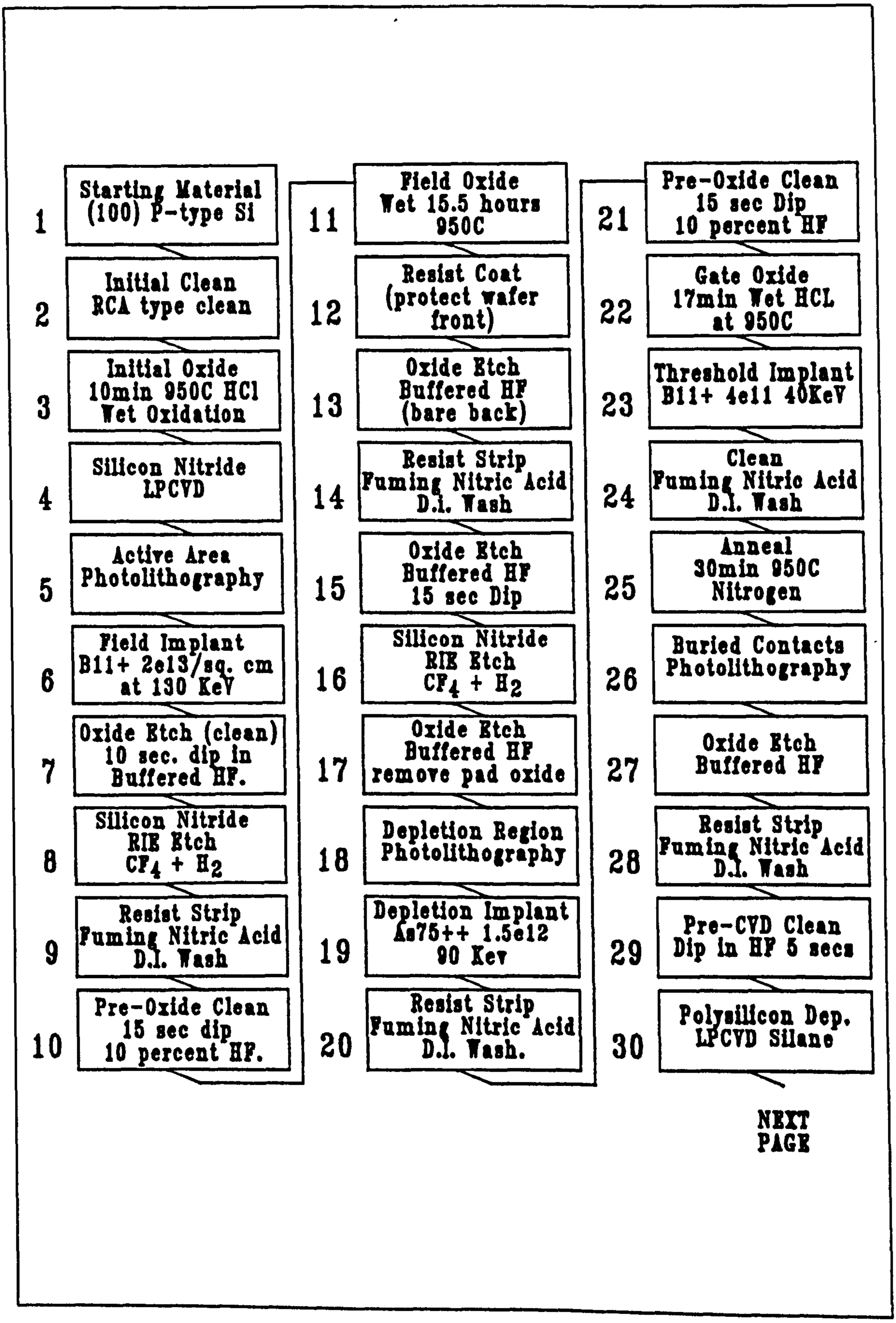


Figure 2.8, E.M.F. 6μm NMOS Process Flow.

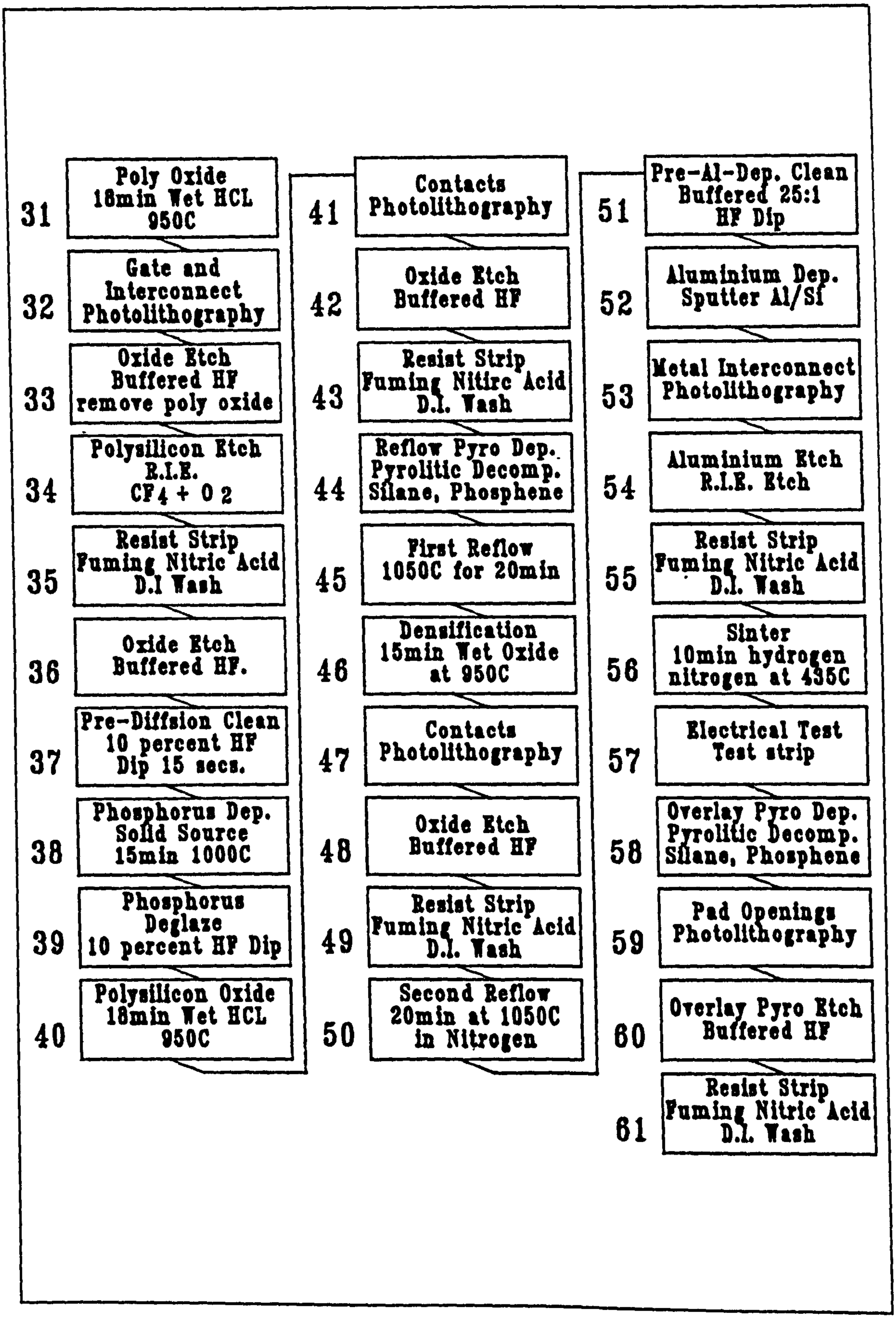


Figure 2.9, E.M.F. 6μm NMOS Process Flow (continued).

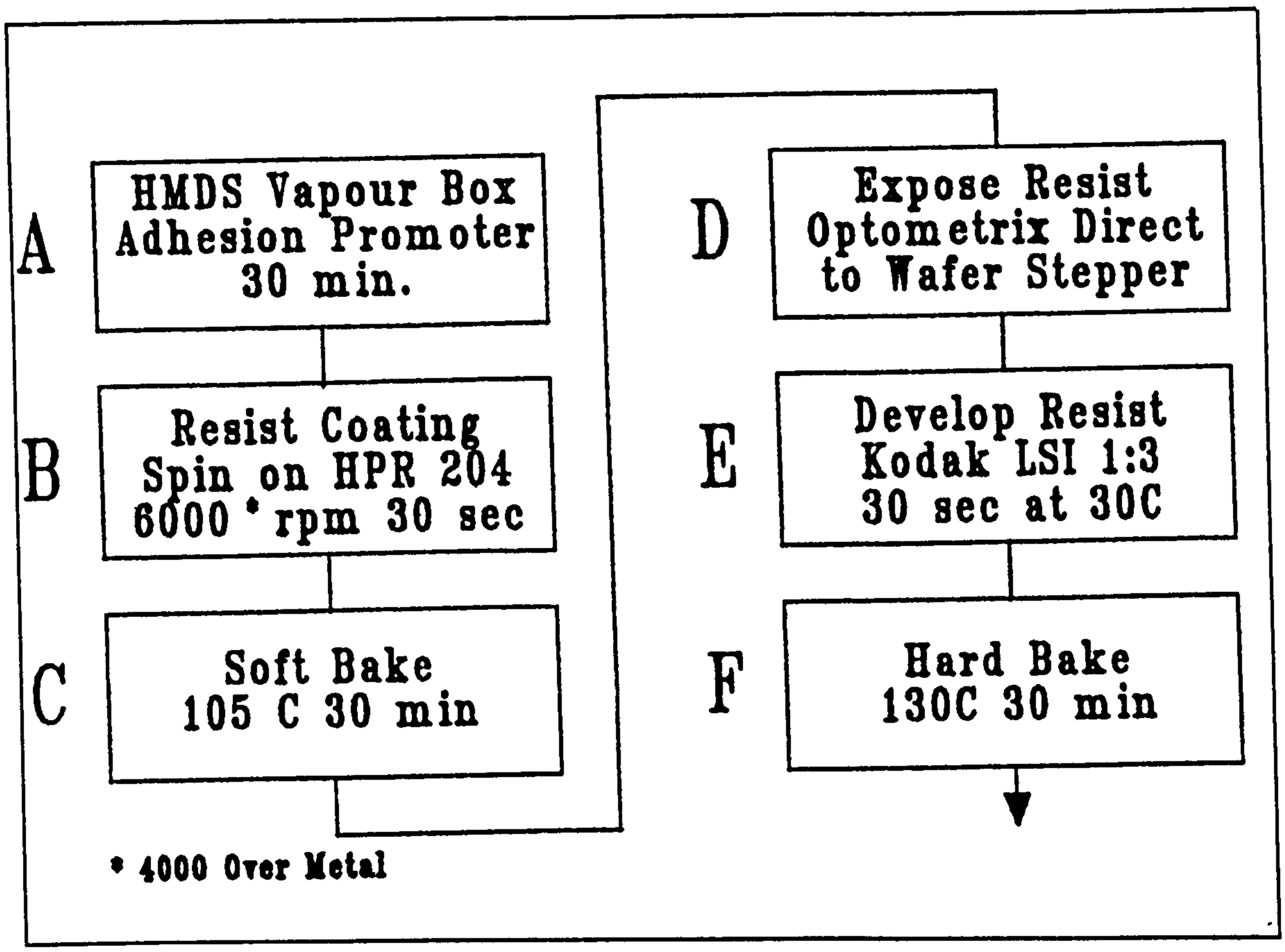


Figure 2.10, E.M.F. 6μm NMOS Lithography Process Flow.

Step 1: The start of wafer processing is the straight forward unpacking of stock wafers.

Step 2: An initial clean is performed, however lately wafers are packed cleaner than the result of this step. It involves the standard R.C.A. clean of sulphuric acid in hydrogen preoxide to remove and organic contamination followed by a dip in hydrofluoric acid and a D.I. wash and dry.

Step 3: Next an initial oxide is thermally grown on the wafer surface. It is sometimes called the pad oxide since its purpose is to provide a pad to reduce stress during the cooling of deposited nitride. It is measured with an interference type thickness measurement machine.

Step 4: Silicon nitride is deposited by LPCVD to act as an oxygen barrier mask for the field oxidation step. Its thickness is also measured by the interference technique.

Step 5: The first photolithography stage, figure 2.10, is to define the active areas of the wafer. The wafer is treated by subjecting it to HMDS vapour to promote adhesion by producing a dry skin of SiO_2 on the surface. Without the oxide positive photoresist adheres poorly to silicon nitride. Next HPR-204, a positive photoresist, is spun onto the wafer to give approximately 1μm film. The wafers are baked to remove the bulk of the solvents, increase photosensitivity, and further improve adhesion. The pattern is then exposed and developed. Following a microscopic inspection the wafers are placed in a hard bake oven to drive off the remaining casting solvent to further increase adhesion but also to stop the remaining photosensitivity.

Step 6: The active area mask is first used to block boron implantation into all but the field regions. The boron implant there increases the field inversion voltage to prevent parasitic transistors. Figure 2.11, shows a likely cross-section after this step.

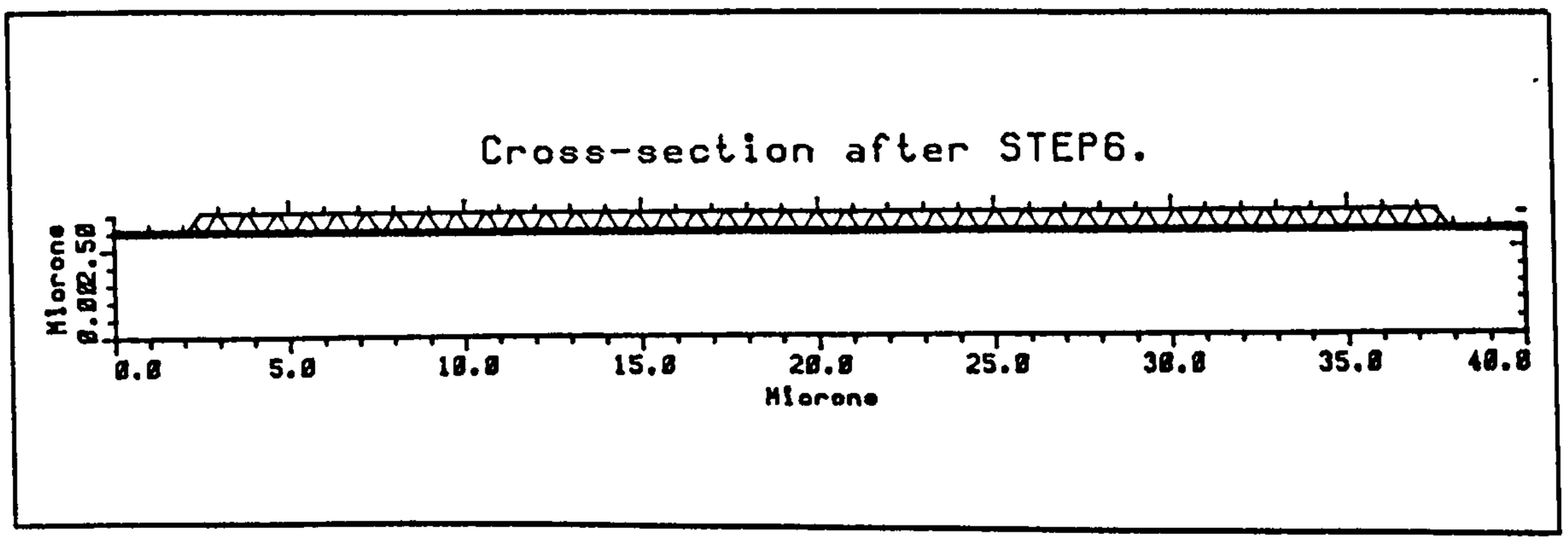


Figure 2.11, Cross-section after nitride mask definition.

Step 7: The brief hydrofluoric acid dip in this step is used to remove any natural oxides from the top of the nitride to prepare it for etching.

Step 8: Both sides of the wafer are etched in a reactive ion etcher. The active surface is etched to expose the field areas and all of the backside is etched to allow oxide to grow on the backside of the wafer and balance the surface side strain.

Step 9: After the patterning of the nitride the photoresist is no longer required and is removed in a fuming nitric acid etch.

Step 10: A pre-oxidation clean of the wafer by dipping in hydrofluoric acid removes the pad oxide in the field regions and cleans the surface ready for oxidation.

Step 11: The field oxidation is performed in five stages. First the wafers are slowly inserted into the furnace with oxygen only flowing. This allows the wafers to warm up gradually and is a safety measure to avoid gassing clean room staff. Next there is a 5 minute dry oxidation with HCl to clean the wafer surface to promote even oxidation. Then a half hour wet oxidation with HCL is performed to getter sodium and promote oxide growth. The main 15.5 hours of wet oxidation forms just over 1 micron of field oxide. A final five minutes of dry oxidation facilitates removal of the wafers. The field oxide thickness is then measured by the interference technique.

Step 12: The entire front of the wafer is coated with photoresist which is immediately hard baked.

Step 13: The back side of the wafers can then be etched back to bare silicon in buffered *HF* .

Step 14: The front side resist can be removed.

Step 15: Any oxide which had formed on the nitride mask is removed to facilitate the nitride removal.

Step 16: The need for the masking silicon nitride is over and is removed in a RIE etcher.

Step 17: The pad oxide is removed from the active regions in a short *HF* dip which also removes some of the field oxide but leaves a profile, as shown in figure 2.12, that is ready for the fabrication of the transistors. The transistors will be built on the silicon surface, and the field region will allow room for interconnect levels and contacts between them.

Step 18: The second photolithography step covers the enhancement active regions with photoresist to block the depletion implant.

Step 19: The active areas where depletion transistors are to be built have the threshold voltage adjusted by an arsenic implant making a natural n channel. It is implanted before gate oxidation unlike boron since it does not get taken

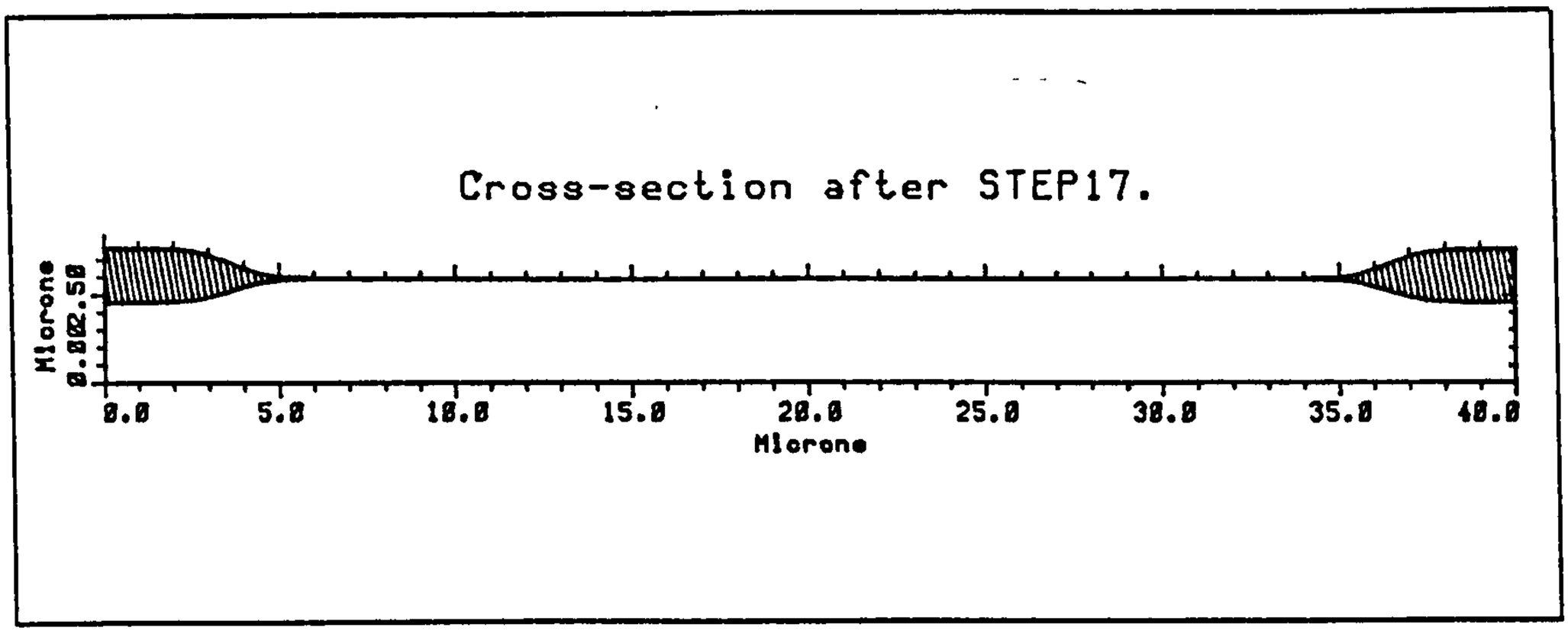


Figure 2.12, The cross-section after LOCOS isolation completion.

up into the oxide the same way during oxidation.

Step 20: The photoresist implantation mask is removed with fuming nitric acid.

Step 21: Next the all important pre-oxidation clean is performed for the gate oxide. It is important in cleaning the surface and removing native oxides and those formed by the weakly oxidising resist strip to present bare silicon for oxidation.

Step 22: The gate oxide is thermally grown using a wet oxidation with HCl in four parts as with the field oxide. At this stage the gettering effect is very important to the performance of the transistors. The 17 minute oxidation produces a gate oxide around 1000 $\mathring{A}$ thick, which is usually measured by the interference technique.

Step 23: Next the threshold adjusting boron implants are made through the gate oxide for two reasons. One is that the oxide slows the ions and results in a higher concentration at the silicon surface. The second reason is that the oxide protects the interface from impurities picked up during implantation.

Step 24: This resist strip is really a cleaning of implanter contamination from the surface without damaging the gate oxide with the usual HF dip.

Step 25: The anneal re-orders the silicon crystal and activates the implants in a controlled manner. Subsequent thermal processing adds to this effect.

Step 26: The third photolith stage prepares for contacts between the silicon and polysilicon. All areas but where a contact hole should be are masked.

Step 27: A buffered *HF* dip is used to cut the contacts.

Step 28: The contact masking photoresist is removed yet again in fuming nitric acid.

Step 29: Another brief *HF* dip prepares the surface for polysilicon deposition and to remove any films which got into the contact windows during the resist strip.

Step 30: About 0.75μm of polysilicon is deposited by chemical vapour deposition.

Step 31: An 18 minute wet oxidation of the polysilicon provides an oxide layer. It acts as a etching mask for the polysilicon in the later RIE etch. The oxide produces a better etch mask than photoresist which is attacked by the oxygen in the plasma.

Step 32: The fourth photolith stage masks areas of polysilicon to be conductors and gates. At the same time a half a test wafer, which has been through the other steps but without fine patterning, is exposed and developed to aid in endpoint detection.

Step 33: The polysilicon oxide is etched in buffered *HF* in areas not masked by the photoresist.

Step 34: An RIE etch in carbon tetrafluoride and oxygen is used to remove the exposed polysilicon in areas not covered with oxide and photoresist. The etch is isotropic and will partially undercut the etch mask. Variation in etch rate across the wafer will result in some etching of the underlying oxide before the polysilicon has completely etched. Figure 2.13, shows how DEP-ICT has represented the cross-section after the completion of the etch.

Step 35: The masking photoresist is removed from the polysilicon oxide using fuming nitric acid.

Step 36: Next the polysilicon oxide and the gate oxide over the source and drain regions are removed using a buffered *HF* etch. The test wafer exposed in step 32 is important in determining the etch endpoint and avoiding undercutting the actual gate oxide. The oxide must be completely removed from these areas to allow diffusion doping of the source and drain and the

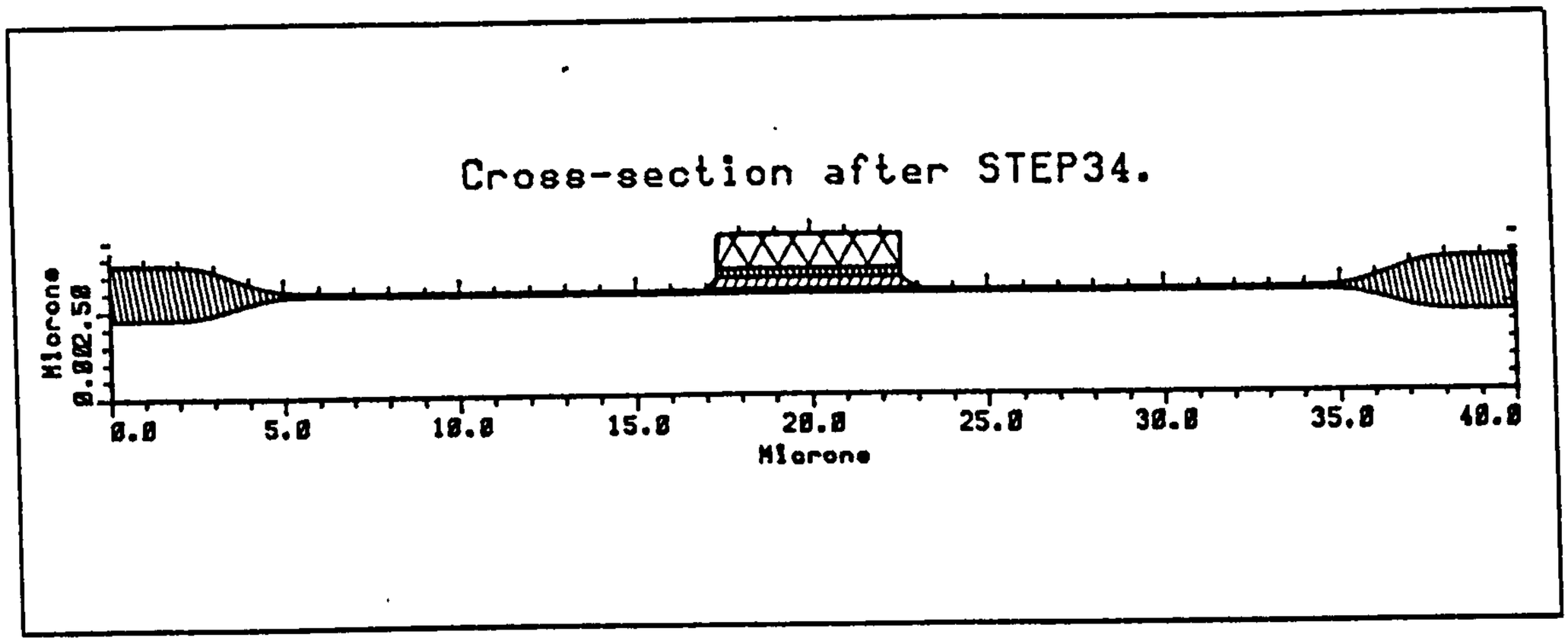


Figure 2.13, Cross-Section after polysilicon definition.

polysilicon gate.

Step 37: A pre-diffusion clean of a *HF* dip is used to remove any native oxides immediately prior to the doping diffusion. Incidentally just previous to this etch is an ideal wait stage if it is necessary to halt a batch at some point since almost all the exposed surface is some form of silicon which can be cleaned by this step.

Step 38: In this step the source and drain and polysilicon gate regions are phosphorus doped from solid source dopant wafers in a furnace tube. The crystalline silicon doping profile is shown in the 3D computer simulation output in figure 2.17. The vertical axis is the concentration of the dopant. The Y axis represents the distance into the silicon and the X axis represents distance along a cross section running from LOCOS isolation through the source, across the channel, through the drain, and back to the LOCOS isolation. The increased depth of diffusion at the isolation side of the source and drain is caused by the isolation technology. There is an electric field set up between the source-drain dopants and the field inversion dopants which affects the source-drain profile. Notice the sharp distinction between the source and channel regions before the following thermal steps. After this step a test wafer which has been following through with the batch can be used to determine the sheet resistance of the N^+ (source and drain) and polysilicon Regions. A four point probe is used

for this.

Step 39: The next step is a phosphorus deglaze that removes the phosphorus rich oxide which forms on the source, drain and gate regions. It also removes the top of the isolation oxide which has become phosphorus doped.

Step 40: A clean oxide is grown over the gate and the source and drain regions next by thermal wet oxidation. Its purpose is to insulate those regions but also to provide a dense seed oxide for pyrox deposition.

Step 41: The fifth photolithography step defines contact windows to the source , drain or gate.

Step 42: A buffered *HF* oxide etch is used to etch the contact holes. This etch may appear premature since the main dielectric isolation has not yet been deposited, however, differing etch rates between thermal oxide and pyro oxide cause significant sideways etching of the pyrox before the thermal oxide is etched through if both layers are etched at one time.

Step 43: Fuming nitride is once again used to remove the organic photoresist.

Step 44: About 7500 $\mathring{A}$ of pyrolytic silicon dioxide is deposited over the wafer. Figure 2.14, shows a representation of the surface after the pyrox deposition. Note that the surface is quite irregular as underlying topography affected the deposition. The oxide however is phosphorus doped which will allow smoothing by thermal reflow. The test wafer is used again here. It has its oxide removed and then four point probes are used to establish the sheet resistance of the polysilicon. It is discarded after this step.

Step 45: The P-Glass deposition is immediately followed by a thermal reflow in an oxygen ambient. This is an important step since it complexes the P_2O_5 into phosphoric glass and prevents the hyrolysis into phosphoric acid.

Step 46: The first reflow is followed by a wet oxidation. The oxidation densifies the pyrolytic oxide and will also leach phosphorus out of the top of the reflow glass which improves photoresist adhesion.

Step 47: The sixth photolith is just the contacts revisited.

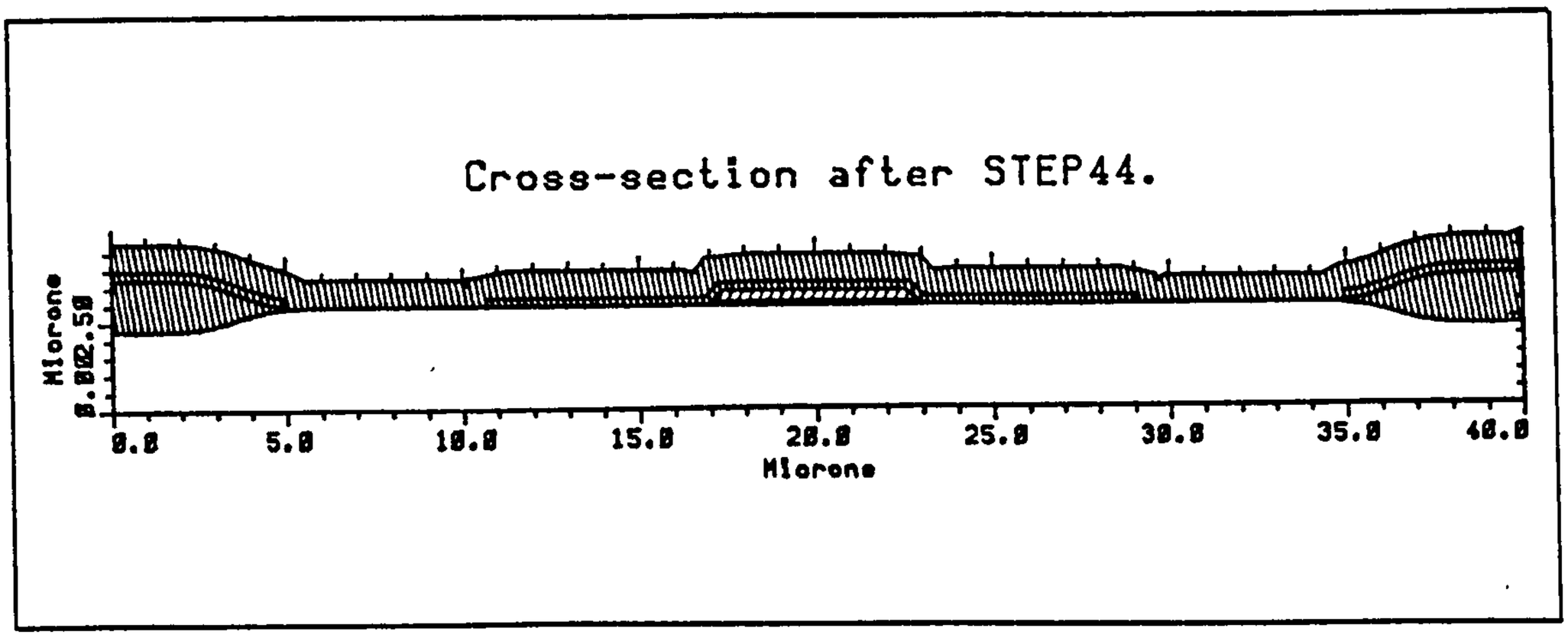


Figure 2.14, Cross-section after pyrolytic oxide deposition.

Step 48: A buffered *HF* solution is used to etch the contact windows through the pyrox to the silicon or polysilicon. The profile at this point will have vertical walled contact holes of up to a micron deep.

Step 49: Fuming nitric acid is used to remove the masking photoresist before the next high temperature step.

Step 50: The second thermal reflow of the P-glass is designed to reduce the contact hole sidewall slope to improve metallisation step coverage. It also has the effect of further smoothing the rest of the cross-section. Figure 2.15, shows a cross-section of the transistor at this point.

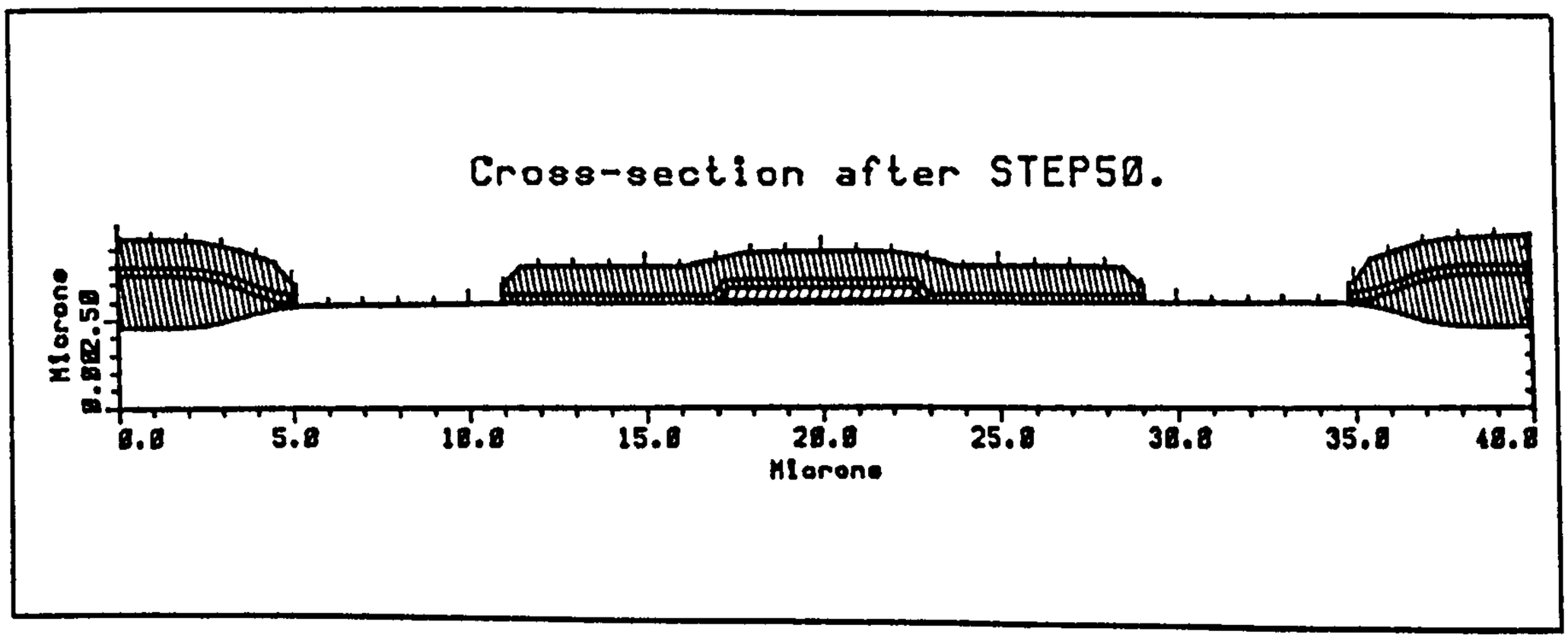


Figure 2.15, Cross-section after contact window rounding.

Step 51: A dilute buffered *HF* solution is used next to clean the natural oxide from the contact windows immediately prior to the metallisation step.

Step 52: Metallisation is performed by deposition of sputtered aluminium/silicon. Just over a micron of aluminium is deposited.

Step 53: The seventh photolithography stage is the metal definition step. The photoresist covers metal that is to remain. A slower spin speed is used in the resist coating operation to achieve a thicker resist. The thicker resist is needed to cover the undulating topography and to guard against complete erosion during the metal etch.

Step 54: The aluminium etch is performed by carbon-tetrachloride in a RIE reactor. The process is anisotropic and produces vertical sidewalls.

Step 55: After etching the wafers are immediately washed to remove any possibility of HCl etching, and then the resist is stripped in fuming nitric acid. After drying a mechanical stylus measurement can be used to determine aluminium thickness.

Step 56: The next step in the process is a thermal sinter. Its purpose is to promote good contact between the aluminium and silicon, without causing local melting or junction spiking.

Step 57: The next step is an electrical test to measure test inserts. The results give information on the processing and the success of the run.

Step 58: The final layer of overlay pyrolytic oxide is deposited to at as a passivation layer, a mechanical surface protector, and as an electrical insulator.

Step 59: The eighth and final photolithography stage provides a resist mask to allow etching of the pad opening windows in the overlay oxide so that electrical connections can be made.

Step 60: Buffered *HF* is used to etch openings where the pyrox is not covered by photoresist.

Step 61: The final step in the process is a fuming nitric removal of the photoresist, a wash and spin dry. The wafers are then complete. Figure 2.16 shows the likely cross-section of the enhancement transistor after completion. Figure

2.18 shows the doping concentration in the substrate after processing. Note how much less abrupt the junctions are in comparison to when they were formed. This unintentional diffusion is a difficult problem for smaller geometry processes.

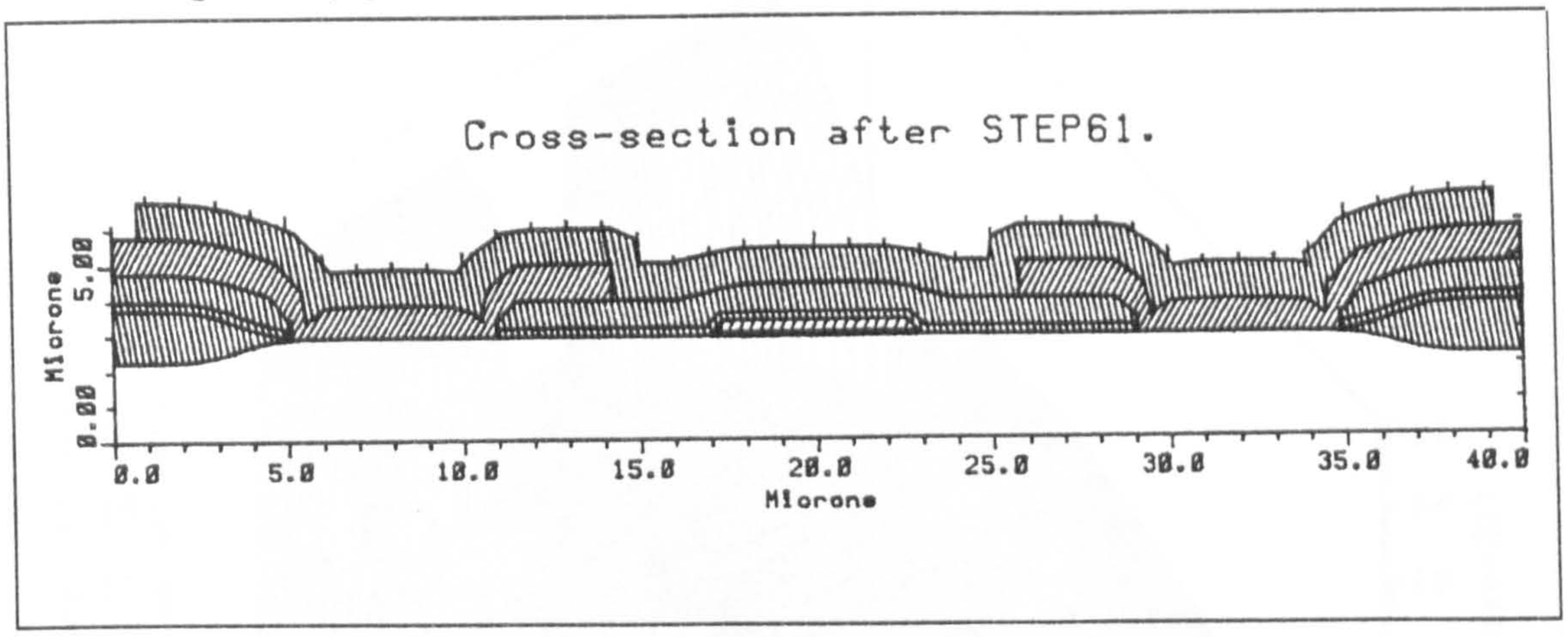


Figure 2.16, Final Cross-section.

2.9.3. Section Summary.

The E.M.F. NMOS process was the feature of this section. Computer simulations of the process along with descriptions of the essential features of each processing step were used to show how the many processes of silicon microfabrication come together to produce integrated circuits.

2.10. Chapter Summary.

In this chapter details on microfabrication processes to produce starting materials, add secondary layers, transfer the design pattern, selectively remove material, modify layer composition and topography, measure aspects, and control impurities were presented. How the techniques can be combined into a microfabrication process was shown through analysis of an example process. The next chapter will review the operation of Insulated Gate Field Effect Transistors (IGFETs) fabricated using some of these techniques.

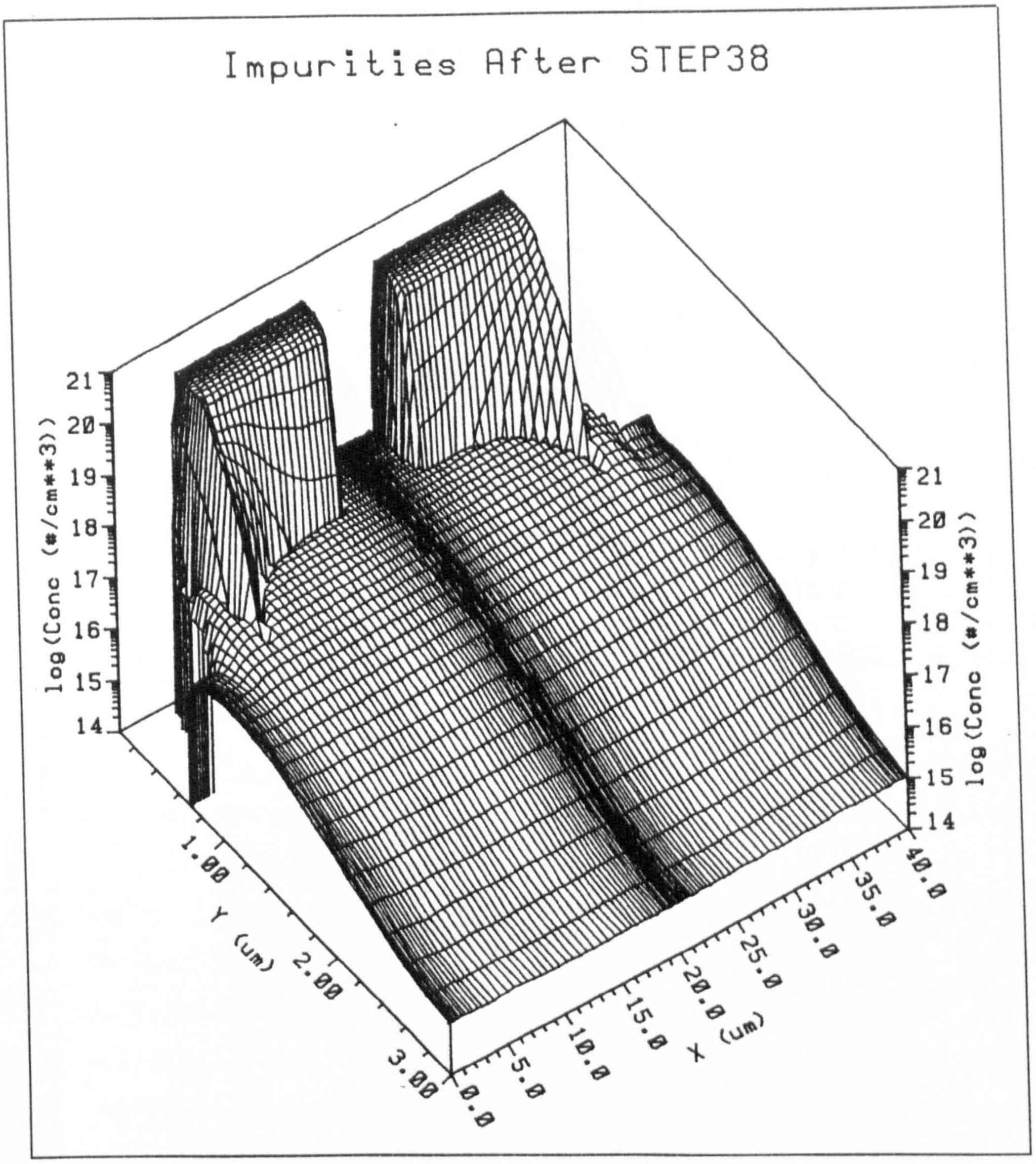


Figure 2.17, Impurity Concentration after Doping.

2.11. References.

1. Pearce, C.W., "Crystal Growth and Wafer Production," in *VLSI Technology*, ed. Sze, S.M., McGraw-Hill, 1983.

2. Sze, S.M., *Semiconductor Devices, Physics and Technology.*, John Wiley & Sons, Inc., New York, 1985.

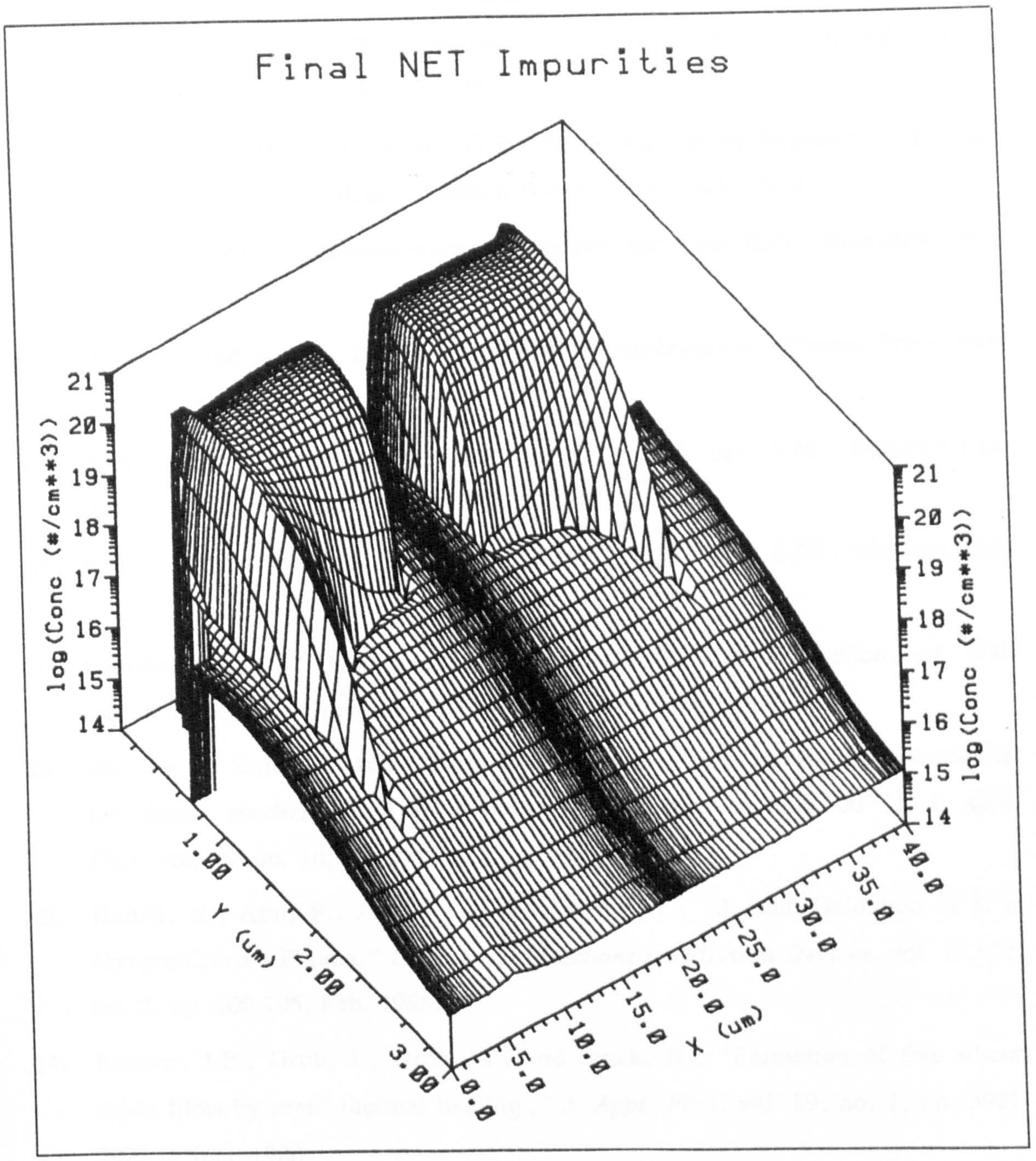


Figure 2.18, Final Impurity Concentration.

3. Colelaser, R.A., *Micro Electronics, Processing and device Design.*, John Wiley & Sons, Inc., New York, 1980.

4. Muller, R.S. and Kamins, T.I., *Device Electronics for Integrated Circuits (2nd Ed.)*, John Wiley & Sons, Inc., New York, 1977.

5. Till, W.C. and Luxon, J.T., *Integrated Circuits: Materials, Devices, and Fabrication.*, Prentice-Hall, New York, 1982.

6. Glaser, A.B. and Subak-Sharpe, G.E., *Integrated Circuit Engineering. Design, Fabrication and Applications.*, Addison-Wesley, New York, 1977.

7. Fogiel, M., *Modern Microelectronics*, Research and Education Association, New York, 1972.

8. Brodie, I. and Muray, J., *The Physics of Microfabrication.*, Plenum Press, New York, 1982.

9. Pearce, C.W., "Epitaxy.," in *VLSI Technology*, ed. Sze, S.M., McGraw-Hill, 1983.

10. Katz, L.E., "Oxidation.," in *VLSI Technology*, ed. Sze, S.M., McGraw-Hill, 1983.

11. Gundlach, A.M., "Oxidation.," in *SERC School on Microfabrication.*, ed. Walton, A.J., E.M.F. University of Edinburgh, Edinburgh, June 1987.

12. Perriere, J., Siejka, J., and Chang, R., "Study of oxygen transport processes during plasma anodization of Si between room temperature and 600°C," *J. Appl. Phys*, vol. 56, no. 10, pp. 2716-2725, 15 Nov. 1984.

13. Haneji, N., Arai, F., Asada, K., and Sugano, T., "Anodic Oxidation of Si in Oxygen/Chlrine Plasma," *I.E.E.E. Transactions on Electron Devices*, vol. ED-32, no. 2, pp. 100-105, Feb. 1985.

14. Ponpon, J.P., Grob, J., Grob, A., and Stuck, R., "Formation of thin silicon oxide films by rapid thermal heating.," *J. Appl. Phys*, vol. 59, no. 1, pp. 3921-3923, 1 June 1986.

15. Veprek, S., "Plasma Induced and Plasma Assisted Chemical and Physical Vapour Deposition," in *Summer Course on Device Impact of New Microfabrication Technologies.*, ed. Claeys, C.L., vol. 3, I.M.E.C., Leuven, Belgium, 1986, June 9-13.

16. Adams, A.C., "Dielectric and Polysilicon Film Deposition.," in *VLSI Technology*, ed. Sze, S.M., McGraw-Hill, 1983.

17. Haworth, L., "CVD," in *SERC School on Microfabrication.*, ed. Walton, A.J., E.M.F. University of Edinburgh, Edinburgh, June 1987.

18. Pauleau, Y., "Interconnect Materials for VLSI Circuits.," *Solid State Technology*, pp. 61-67, Feb 1987.

19. Fraser, D.B., "Metallization.," in *VLSI Technology*, ed. Sze, S.M., McGraw-Hill, 1983.

20. Pauleau, Y., "Interconnect Materials for VLSI Circuits. Part III.," *Solid State Technology*, pp. 101-105, June 1987.

21. Gargini, P.A., "Interconnect and Contact Technologies for VLSI.," in *Summer Course on Device Impact of New Microfabrication Technologies.*, ed. Claeys, C.L., vol. 4, I.M.E.C., Leuven, Belgium, 1986, June 9-13.

22. Murarka, S.P. and Vaidya S., "Cosputtered Cobalt Silicides on silicon, polycrystalline silicon, and silicon dioxide.," *J. Appl. Phys*, vol. 56, no. 12, pp. 3404-3412, 15 Dec. 1984.

23. Fathy, D., Holland, O., and Narayan, J., "Formation of Ion beam mixed silicides on Si(100) at elevated substrate temperatures.," *J. Appl. Phys*, vol. 58, no. 1, pp. 297-301, 1 July 1985.

24. Hassen, J. and Sarkary, H., "Lithography for VLSI: An overview.," *Solid State Technology*, pp. 49-54, June, 1982.

25. Oberai, A., "Lithography - Challenges of the Future.," *Solid State Technology*, pp. 123-128, Sept., 1987.

26. Neukermans, A.P., "Current Status of X-Ray Lithography Part I.," *Solid State Technology*, pp. 185-188, Sept., 1984.

27. Stengl, G. et. al., "Ion Projection Microlithography," *Solid State Technology*, pp. 104-109, Aug., 1982.

28. King, H.N.G., "Electron Lithography," *Solid State Technology*, pp. 102-105, Feb., 1982.

29. Broers, A.N., "The Submicron Lithography Labyrinth," *Solid State Technology*, pp. 119-126, June, 1985.

30. Piwczyk, B. and Williams, A., "Electron Beam Lithography for the 80's.," *Solid State Technology*, pp. 74-82, June, 1982.

31. Fuller, G.E., "Optical Lithography Status," *Solid State Technology*, pp. 113-118, Sept., 1987.

32. King, M.C., "Principles of Optical Lithography.," in *VLSI Electronics Microstructure Science*, ed. Einspruch, N.G., vol. 1, Academic Press, New York, 1981.

33. Lepselter, M.P. and Lynch, W.T., "Resolution Limitations for Submicron Lithography," in *VLSI Electronics Microstructure Science*, ed. Einspruch, N.G., vol. 1, Academic Press, New York, 1981.

34. Stevenson, J.T.M., "Lithography.," in *SERC School on Microfabrication.*, ed. Walton, A.J., E.M.F. University of Edinburgh, Edinburgh, June 1987.

35. McCreary, M.D., "Photoresist Schemes for Submicron Optical Lithography," in *Summer Course on Device Impact of New Microfabrication Technologies.*, ed. Claeys, C.L., vol. 3, I.M.E.C., Leuven, Belgium, 1986, June 9-13.

36. Stevenson, J.T.M., "Resists.," in *SERC School on Microfabrication.*, ed. Walton, A.J., E.M.F. University of Edinburgh, Edinburgh, June 1987.

37. McGillis, D.A., "Lithography.," in *VLSI Technology*, ed. Sze, S.M., McGraw-Hill, 1983.

38. Coopmans, F., "Advanced Dry Developed Resist Systems," in *Summer Course on Device Impact of New Microfabrication Technologies.*, ed. Claeys, C.L., vol. 3, I.M.E.C., Leuven, Belgium, 1986, June 9-13.

39. Holwill, R., "Etching.," in *SERC School on Microfabrication.*, ed. Walton, A.J., E.M.F. University of Edinburgh, Edinburgh, June 1987.

40. Ephrath, L.M., "Etching Needs for VLSI," *Solid State Technology*, pp. 87-92, July, 1982.

41. Bergeron, S. and Duncan, B., "Controlled Anisotropic Etching," *Solid State Technology*, pp. 98-103, Aug, 1982.

42. Okano, H., Yamazaki, T., and Horiike, Y., "High Rate Reactive Ion Etching Using a Magnetron Discharge.," *Solid State Technology*, pp. 166-170, April,1982.

43. Mucha, J.A., "The Gases of Plasma Etching: Silicon Based Technology," *Solid State Technology*, pp. 123-127, March, 1985.

44. Fonash, S.J., "Advances in Dry Etching Processes - A review.," *Solid State Technology*, pp. 150-158, Jan, 1985.

45. Choe, D., Knapp, C., and Jacob, A., "Production RIE - 1. Selective Dielectric Etching.," *Solid State Technology*, pp. 177-183, April 1984.

46. Chang, J.S., "Selective Reactive Ion Etching of Silicon Dioxide.," *Solid State Technology*, pp. 214-217, April 1984.

47. Castellano, R.N., "Profile Control in Plasma Etching of SiO_2," *Solid State Technology*, pp. 203-206, May 1984.

48. Bolling, S., Lida, S., and Matsumoto, O., "Reactive Ion Etching: Its Basis and Future Part I.," *Solid State Technology*, pp. 111-117, May, 1984.

49. Bolling, S., Lida, S., and Matsumoto, O., "Reactive Ion Etching: Its Basis and Future Part II.," *Solid State Technology*, pp. 167-173, June, 1984.

50. Ibbotson, D., Mucha, J., Flamm, D., and Cook, J., "Plasmaless dry etching of silicon with fluorine.," *J. Appl. Phys*, vol. 56, no. 10, pp. 2939-2943, 15 Nov 1984.

51. Nagy, A.G., "Sidewall Tapering in Reactive Ion Etching.," *J. Electrochem. Soc.*, vol. 132, no. 2, pp. 689-693, March 1985.

52. Light, R.W., "Reactive Ion Etching of Aluminum/Silicon," *J. Electrochem. Soc.*, vol. 130, no. 11, pp. 2225-2230, Nov. 1983.

53. Mogab, C.J., "Dry Etching.," in *VLSI Technology*, ed. Sze, S.M., McGraw-Hill, 1983.

54. Coopmans, F. and Brasseur, G., "Magnetron Ion Etching," in *Summer Course on Device Impact of New Microfabrication Technologies.*, ed. Claeys, C.L., vol. 3, I.M.E.C., Leuven, Belgium, 1986, June 9-13.

55. Tsai, J.C.C, "Diffusion.," in *VLSI Technology*, ed. Sze, S.M., McGraw-Hill, 1983.

56. Burger, R.M. and Donovan, R.P., *Fundamentals of Silicon Integrated Device Technology.*, 1, Prentice-Hall, New Jersey, 1967.

57. Rodrigues, R., "Diffusion.," in *SERC School on Microfabrication.*, ed. Walton, A.J., E.M.F. University of Edinburgh, Edinburgh, June 1987.

58. Tressler, R. et. al., "Present Status of Arsenic Planar Diffusion Sources.," *Solid State Technology*, pp. 165-177, Oct 1984.

59. Justice, B.H., Wouster, G.S., Aycock, R.F., and Saunders, D.R., "A Novel Boron Spin-On Dopant," *Solid State Technology*, pp. 153-158, October 1984.

60. Seidel, T.E, "Ion Implantation.," in *VLSI Technology*, ed. Sze, S.M., McGraw-Hill, 1983.

61. Haworth, L., "Ion Implantation.," in *SERC School on Microfabrication.*, ed. Walton, A.J., E.M.F. University of Edinburgh, Edinburgh, June 1987.

62. Wittkower, A.B, "The Effect of Ion Implanter Design Upon Implant Uniformity.," *Solid State Technology*, pp. 77-81, Sept 1982.

63. Pramanik, D. and Current, M., "MeV Implantation for Silicon Device Fabrication.," *Solid State Technology*, pp. 211-216, May 1984.

64. Ziegler, J.F. and Lever, R.F., "Channeling of ions near the silicon <001> axis.," *App. Phys. Lett.*, vol. 46, no. 4, pp. 359-361, 15 Feb. 1985.

65. Lunnon, M.E. and Chen, J.T., "Structural and Electrical Properties of BF_2, implanted, rapid annealed silicon.," *App. Phys. Lett.*, vol. 45, no. 10, pp. 1056-1059, 15 Nov. 1984.

66. Seidel, T.E., Knoell, R., Poli, G., and Schwartz, B., "Rapid Thermal annealing of dopants implanted into preamorphized silicon.," *J. Appl. Phys*, vol. 58, no. 2, pp. 683-687, 15 July 1985.

67. Ziegler, J.F., *Ion Implantation, Science and Technology*, Academic Press Inc., New York, 1984.

68. Carter, G. and Grant, W.A., *Ion Implantation of Semiconductors.*, Edward Arnold, London, 1976.

69. Harrison, H. B., "Rapid Thermal Processing as an alternative to Conventional Furnaces for VLSI technology.," in *Summer Course on Device Impact of New Microfabrication Technologies.*, ed. Claeys, C.L., vol. 3, I.M.E.C., Leuven, Belgium, 1986, June 9-13.

70. Wilson, S., Paulson, W., and Gregory, R., "Rapid Annealing Technology for Future VLSI," *Solid State Technology*, pp. 185-190, June, 1985.

71. Downey, D.F., Russo, C.J., and White, J.T., "Activation and Process Characteristics of Infrared Rapid Isothermal and Furance Annealing Techniques.," *Solid State Technology*, pp. 87-93, Sept., 1982.

72. Narayan, J. and Holland, O., "Rapid Thermal annealing of ion-implanted semiconductors.," *J. Appl. Phys*, vol. 56, no. 10, pp. 2913-2921, 15 Nov. 1984.

73. Pauleau, Y., "Interconnect Materials for VLSI Circuits. Part II.," *Solid State Technology*, pp. 155-162, April 1987.

74. Mercier, J.S., "Rapid Reflow of Doped Glasses for VLSI Fabrication," *Solid State Technology*, pp. 85-91, July, 1987.

75. Shah, P.L. and Havemann, R.H., "Scaled MOS and Bipolar Technologies for VLSI.," in *VLSI Electronics Microstructure Science*, ed. Einspruch, N.G., vol. 7, Academic Press, New York, 1981.

76. Stillwagon, L.E., "Planarization of Substrate Topography by Spin - Coated films. A Review.," *Solid State Technology*, pp. 67-71, June, 1981.

77. Marcus, R.B., "Diagnostic Techniques.," in *VLSI Technology*, ed. Sze, S.M., McGraw-Hill, 1983.

78. Larrabee, G.B., "Materials Characterization for VLSI.," in *VLSI Electronics Microstructure Science*, ed. Einspruch, N.G., vol. 2, Academic Press, New York, 1981.

79. Brennan, R. and Dickey, D., "Determination of Diffusion Characteristics Using Two-and-Four-Point Probe Measurements.," *Solid State Technology*, pp. 125-132, Dec., 1984.

80. Ehrstein, J.R., *Emerging Semiconductor Technology.*, American Society for Testing and Materials, New York, 1987.

81. Yeh, T.H., *Atomic Diffusion in Semiconductors.*, Plenum Press, London, 1973.

82. Vandervorst, W., Maes, H., and De Keersmaecker, R., "Secondary Ion Mass Spectrometry: Depth profiling of shallow As implants in silicon and silicon dioxide.," *J. Appl. Phys*, vol. 56, no. 5, pp. 1425-1433, 1 September 1984.

83. Pantel, R., "Auger voltage contrast depth profiling of shallow p-n junctions.," *App. Phys. Lett.*, vol. 43, no. 7, pp. 650-653, 1 Oct., 1983.

84. Kaitna, R. and Wernisch, J., "SEM-Analysis of the Electrically Active Subsurface P-N Junctions in MOS Configuration.," *Solid State Technology*, pp. 98-101, March 1982.

85. Cerofolini, G. and Polignano, M., "A comparison of gettering techniques for very large scale integration.," *J. Appl. Phys*, vol. 55, no. 2, pp. 579-585, 15 Jan. 1984.

86. Voltmer, F.W., "Manufacturing Process Technology for MOS VLSI.," in *VLSI Electronics Microstructure Science*, ed. Einspruch, N.G., vol. 1, Academic Press, New York, 1981.

87. Beale, J.R.A., Emms, E.T., and Hilbourne, R.A., *Microelectronics*, Taylor and Francis Ltd., London, 1971.

88. Hattori, T., "Contamination Control and Gettering for VLSI processing.," in *Summer Course on Device Impact of New Microfabrication Technologies.*, ed. Claeys, C.L., vol. 2, I.M.E.C., Leuven, Belgium, 1986, June 9-13.

89. Howes, M.J. and Morgan D.V., *Large Scale Integration.*, John Wiley & Sons, Ltd., London, 1981.

Chapter 3. IGFET Physics.

3.1. Introduction.

In the last chapter it was said that an understanding of processing techniques is essential to appreciate the details of transistor operation. It should be obvious that an understanding of semiconductor physics is just as essential. In this chapter the basic core of semiconductor physics which applies to the IGFET will be presented.

The review will start with the basic methods which are used to describe charge carriers and their energies in a crystal. The two sub-structures of the IGFET, the pn junction diode and MOS Capacitor, will be summarised. The basic structure, principal operating regions, and formulae of the IGFET will then be presented. The effect of ion implantation and small geometries on IGFET characteristics along with unusual operation modes will be covered before the final section summarising CAD circuit and device modelling.

Unlike the material presented in the last chapter, there are a number of complete texts on semiconductor physics for IGFETs.[1,2] Therefore only the basic device equations and those necessary for the research presented in later chapters are given here.

3.2. Background Physics.

Semiconductor devices rely on the way electrons can move through a monocrystalline structure as if they were in a modified form of free space. The analytic analysis of their behaviour also relies on the periodic structure of the crystal. That simplifies the description of electron position, which is done in a statistical manner.[3] A brief review of this underlying model is useful in the understanding of device models.

3.2.1. Band Model of Solids.

Band Diagrams.

From quantum mechanics[3] we known that electrons of an isolated atom can have only discrete energy levels. However, as the nuclei of two atoms are brought closer together, they perturb the allowed energy states so that each discrete state can have two

values. As more and more atoms are brought into close proximity, the influences increase until a band of allowed energies exist about the previous discrete levels. This is what occurs in a crystal lattice. When the atoms are positioned at their lattice sites the discrete levels have merged into two bands of allowed energies, as shown in figure 3.1.

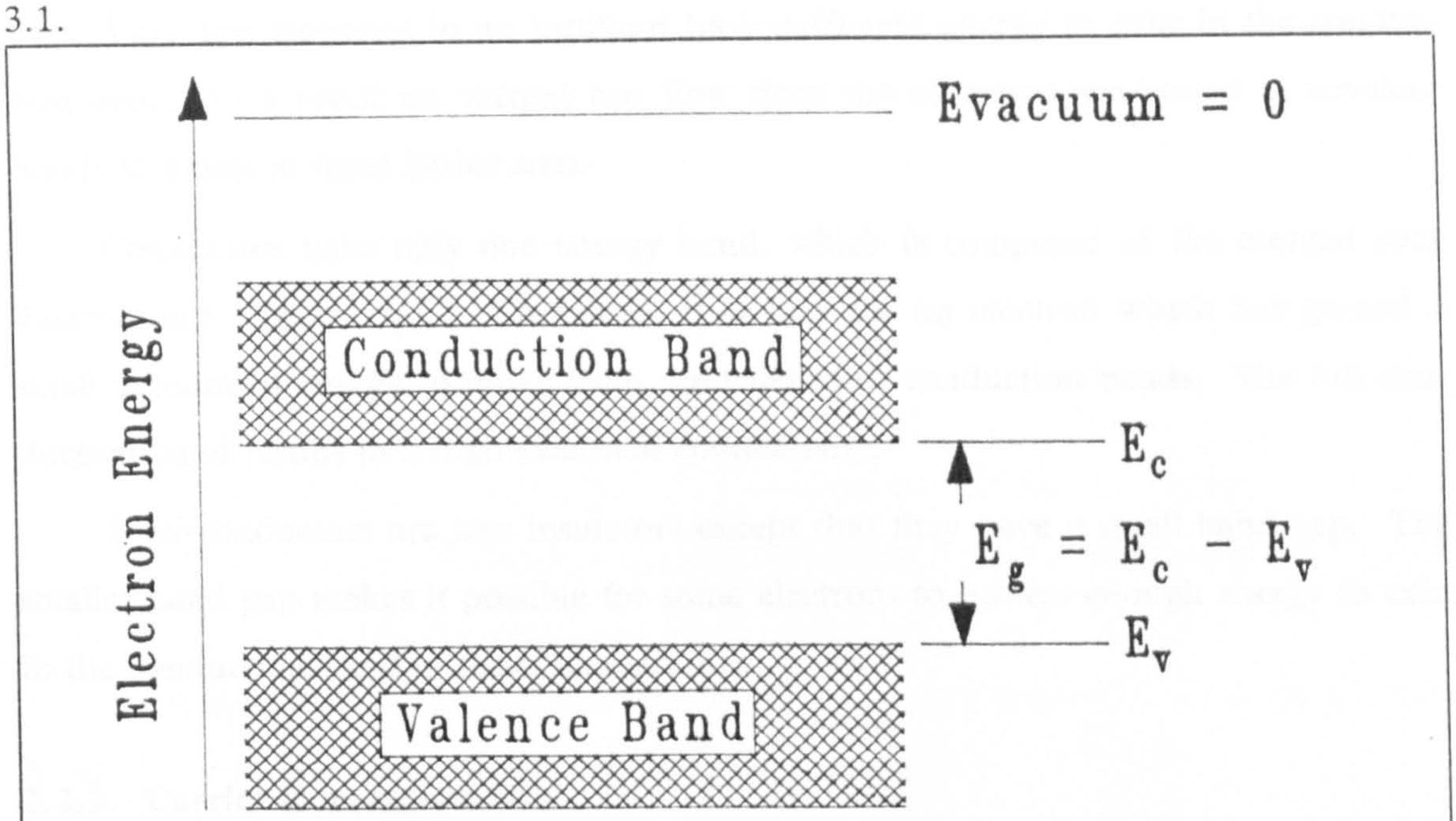


Figure 3.1, An Electron Energy Band Diagram.

An energy diagram expresses the energy of the electrons on the vertical axis, which increases with the zero reference energy taken as the energy of an electron in a complete vacuum. The horizontal axis can be used to express a physical position or another factor about the electrons such as their momentum.

Electrons in the upper band, in figure 3.1, called the conduction band have higher energy than those in the lower valence band. The gap in allowed energies between the two bands is called the band gap. Electrons with energies in the valence band are bound to remain in orbit between the atoms which share them. Electrons which receive sufficient energy to exceed the band gap can join the conduction band and are free to through the lattice.1. The size of the band gap can be thought of as the amount of energy required to break a covalent bond.

By convention, the energy at the bottom of the conduction band is called E_c, as in figure 3.1, and the top of the valence band is called E_v, the difference between them is E_g.

Insulators have very large band gaps compared to the thermal energy of an electron. Very few electrons in an insulator have sufficient energy to exist in the conduction band. As a result no current can flow since the electrons are bound in covalent bonds to atoms at fixed lattice sites.

Conductors have only one energy band, which is composed of the merged conduction and valence bands. Therefore, it is easy for an electron which has gained a small amount of energy to move from the valence to conduction bands. The full conduction band results in a high electrical conductivity.

Semi-conductors are like insulators except that they have a small band gap. The smaller band gap makes it possible for some electrons to possess enough energy to exist in the conduction band at room temperature.

3.2.2. Carrier Concentrations.

It is obviously important to be able to describe the number of charge carriers available in a material if the electric current through the material is to be predicted.

Electrons and Holes.

In a semiconductor, when an electron moves into the conduction band it leaves behind a vacant electron orbit, which is called a hole. Because an electron in an adjacent atom's orbit can move into the vacancy without having conduction band energy, the hole appears able to move. A useful analogy is that of a bubble in a liquid. We commonly think of the bubble as rising whereas it is mostly liquid that is falling.

Because a hole is "not an electron" it has positive charge. Although hole motion is really the motion of a number of electrons, the physics is made simpler by considering it as a particle itself, albeit a fictitious one.

Put simply, electrons conduct charge in the conduction band, holes conduct charge in the valence band.

Further simplifications towards classical physics can be made if the electrons and holes are assigned equivalent masses m_n and m_p respectively.

Fermi Statistics.

It is possible to predict the number of electrons in each energy state in a crystal, which in turn allows the number of charge carriers available to be predicted.

The probability of finding an electron with a given energy E is given by the Fermi-Dirac distribution function

$$f_D(E)=\frac{1}{1+e^{\frac{E-E_F}{kT}}} \tag{3.1}$$

where k is Boltzmann's Constant, and T is the absolute temperature. Since a reference energy is necessary, it is defined as the mean electron energy, and is called the Fermi energy E_F.

Since an allowed state is either occupied, or not, the probability of finding a hole is equal to one minus the probability of finding an electron.

It is possible to find the number of electron energy states allowed in a given physical volume of a crystal by considering the number of allowed wave functions in a confined space, and through consideration of momentum.[4] The result is called the density of states, which is;

$$N(E)=4\pi\left[2\frac{m_n}{h^2}\right]^{\frac{3}{2}}E^{\frac{1}{2}} \tag{3.2}$$

per unit volume. Where h is Planck's Constant.

It is then possible to determine the number of electrons in the conduction band by integrating the product, of equation 3.1 (the probability of a state being full) by equation 3.2 (the density of states), from the bottom to the top of the band energies. For electrons in the conduction band the result is,

$$n=N_c e^{-\frac{E_c-E_F}{kT}} \tag{3.3}$$

and for holes in the valence band,

$$p = N_v e^{-\frac{E_F - E_v}{kT}} \tag{3.4}$$

where N_c and N_v are the effective density of states in the conduction and valence bands.

$$N_c \equiv 2\left[2\pi m_n \frac{kT}{h^2}\right]^{\frac{3}{2}} \tag{3.5}$$

$$N_v \equiv 2\left[2\pi m_p \frac{kT}{h^2}\right]^{\frac{3}{2}} \tag{3.6}$$

For an intrinsic semiconductor, one with no significant impurities, the free electrons and holes are created in pairs by thermal generation. Therefore, $n = p \equiv n_i$ where n_i is the intrinsic carrier density. The intrinsic carrier density can be calculated by taking the product of the hole and electron concentrations which is called the mass action law.

$$n_i^2 = n \times p \tag{3.7}$$

This allows simplification of equations 3.3 and 3.4 to,

$$n_i^2 = N_c N_v e^{\left(\frac{-E_g}{kT}\right)} \tag{3.8}$$

so,

$$n_i = \sqrt{N_c N_v}\, e^{\left(\frac{-E_g}{2kT}\right)} \tag{3.9}$$

The mass action law is valid for both intrinsic and extrinsic semiconductors, since exceeding the intrinsic carrier density allows recombination to occur which reduces the number of excess carriers. At room temperature, the intrinsic carrier concentration for silicon is $1.45 \times 10^{10} cm^{-3}$.

3.2.3. Doping.

Control of the electron and hole concentrations in selected areas of a semiconducting crystal is the key to designing electronic devices. The term applied to adjusting carrier concentrations by intentional introduction of impurities is called doping.

Donors and Acceptors.

When controlled amounts of suitable impurity atoms are introduced into a crystal lattice, and its energy levels and carrier densities are altered, the semiconductor is said to be extrinsic. The effect of the alteration on the carrier density depends on the type and quantity of impurity.

Silicon is a column four element in the periodic table. Column three atoms have one less electron in their valency and as impurities "accept" another electron leaving a deficit in the lattice, which becomes a hole available for conduction. This effect is shown in figure 3.2. Such dopants are called acceptors.

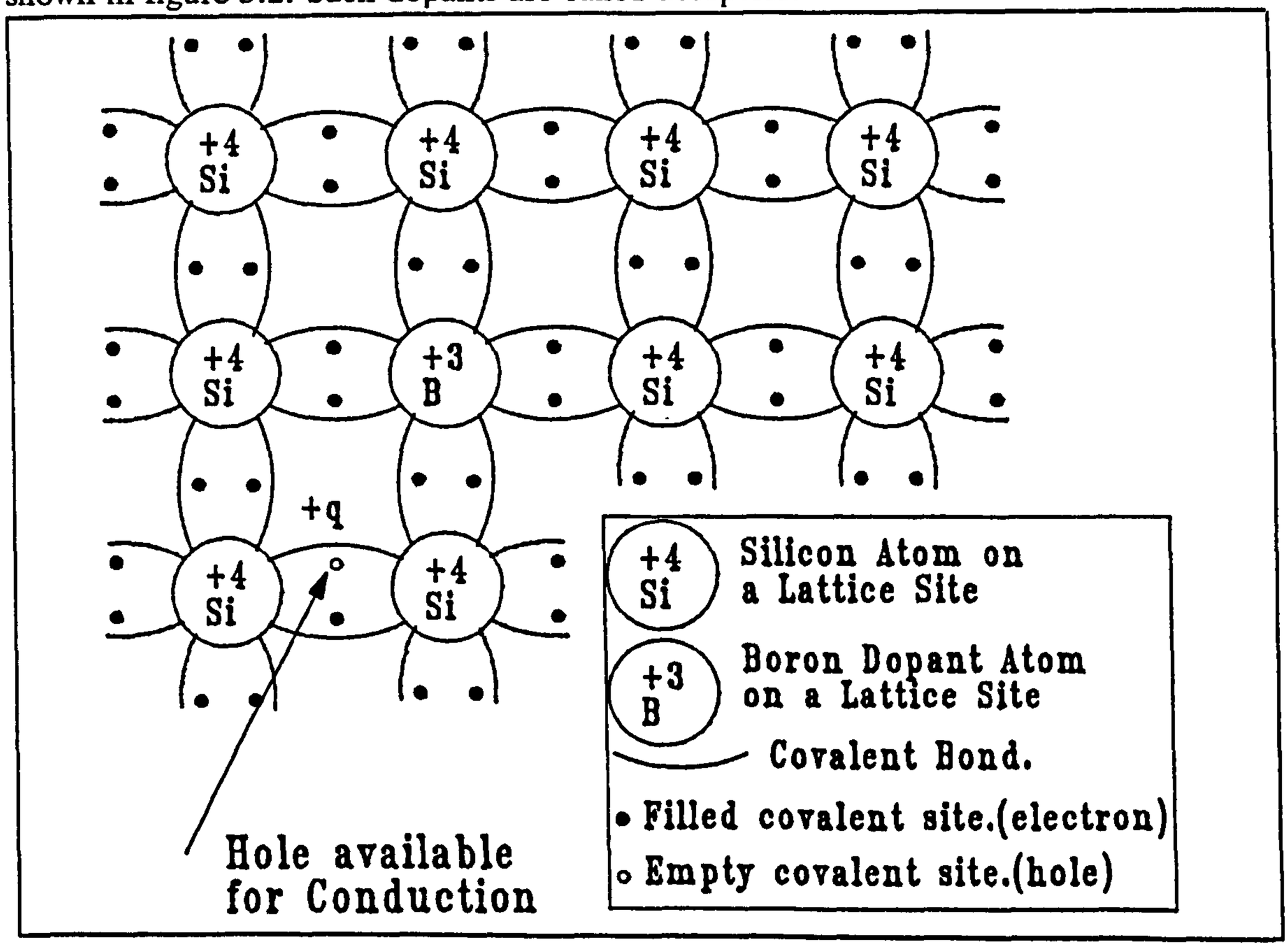


Figure 3.2, Acceptor Doping.

If dopants from column five are used, as in figure 3.3, the electron concentration is increased because the dopant atom has an extra electron in its valence shell allowing it to donate an electron to the conduction band. Column five dopants are called donors. Not all column three and five atoms are suitable dopants for electronics. Table 3.1, lists the commonly used dopants in MOS processes.

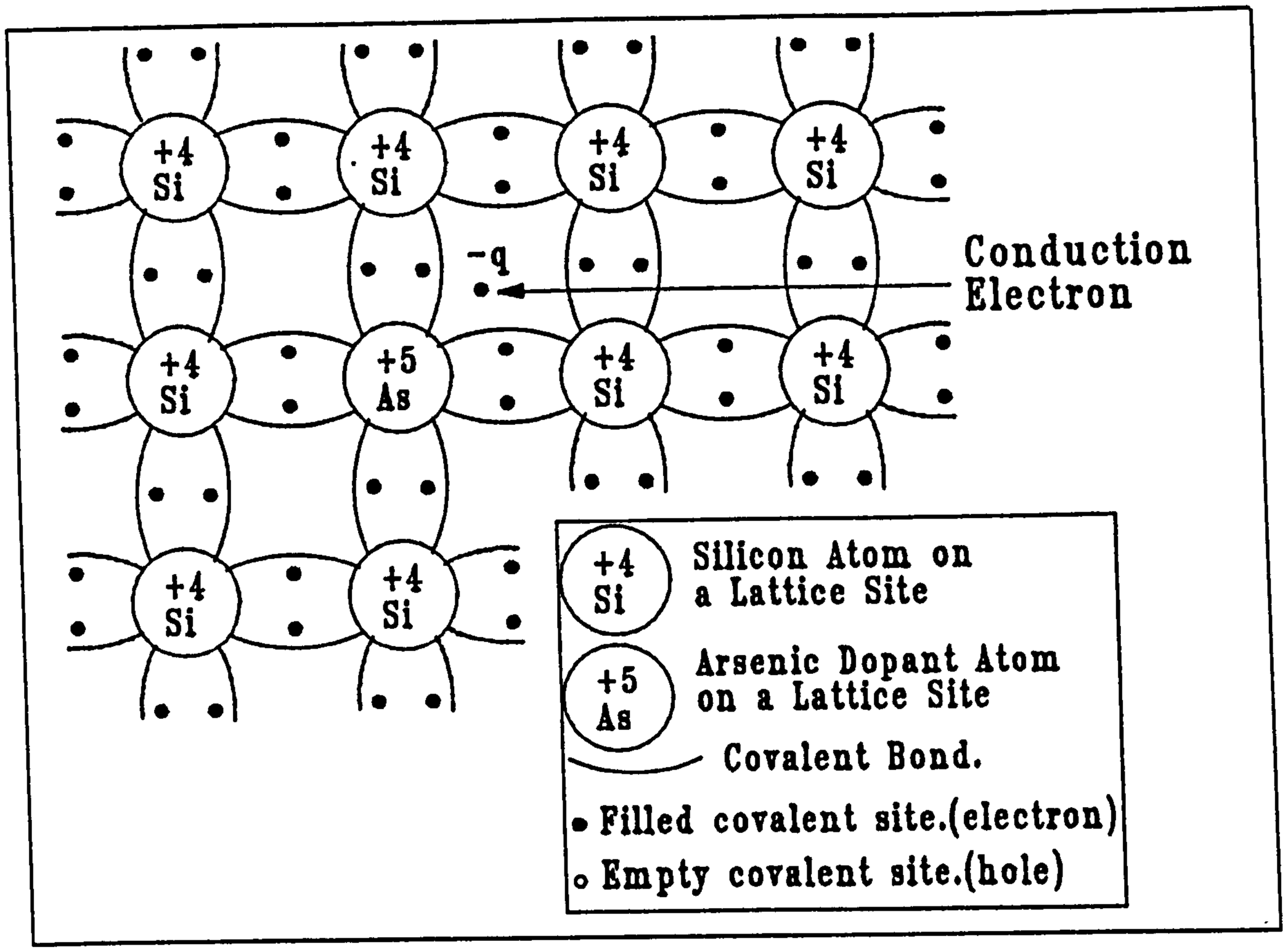


Figure 3.3, Donor Doping.

Donors	Acceptors
Phosphorus Arsenic	Boron

Table 3.1, Common Silicon Dopants.

Surface Typing.

When a region of the crystal has more than the intrinsic density of electrons, ie. is doped with a donor, it is called n-type silicon. When the area is doped with acceptors, the hole concentration exceeds the intrinsic density and the silicon is called p-type. If a region is heavily n-type it is called n^+, and a heavily p-type region is referred to as p^+.

Another definition of importance is the minority carrier. An n-type semiconductor is rich in free electrons, so naturally the majority of the charge will be carried by electrons, the majority charge carrier. The minority charge carrier is then obviously the hole. In p-type silicon the holes become the majority carrier and the electrons the minority carrier.

If the dopant atoms density exceeds the intrinsic carrier density, then the number of carriers is dominated by the dopant induced carriers. In practice the number of dopants are sufficiently high that the carrier concentration is equal to the doping concentration.† ($n = N_D$, or $p = N_A$ where N_D and N_A are the donor and acceptor doping densities respectively).

Obviously the introduction of dopants will also perturb the electron energy states in the lattice. If $n = N_D$ can be assumed then equation 3.3 can be re-arranged to give

$$E_C - E_F = kT \times \ln\left[\frac{N_C}{N_D}\right] \tag{3.10}$$

and similarly

$$E_F - E_V = kT \times \ln\left[\frac{N_V}{N_A}\right] \tag{3.11}$$

So as more donors are added, the Fermi level shifts towards the conduction band, and as more acceptors are added, the Fermi level shifts towards the valence band.

Compensation.

When a crystal lattice has both acceptor and donor impurities present, which is often the case, the predominate dopants will either donate or accept electrons until the effect of the minor dopant is compensated for. The process of compensation results in effective doping densities;

$$N_{D\ effective} = N_D - N_A \quad when \quad N_D > N_A \tag{3.12}$$

$$N_{A\ effective} = N_A - N_D \quad when \quad N_A > N_D \tag{3.13}$$

---

† Providing the thermal energy exceeds the dopant ionisation energies.

Although intrinsic concentration levels can be achieved in a fully compensated lattice the "local" effects on carrier mobility and lifetime are very important.

3.2.4. Free Carriers.

Conduction electrons, which will just be referred to as electrons from now on, and holes can be considered as free particles. Although they are effected by electronic interaction with the lattice sites, that is taken into account in their effective mobility.

Mobility and Drift Velocity.

The electrons and holes in a crystal are in constant motion due to thermal energy.

Although there is carrier motion, the direction is random, so there is no net motion or current transfer.

When an electric field is applied to the crystal the electrons and holes acquire some direction to their random motion and start to drift towards areas of lower energy. The net carrier velocity is called the "drift velocity". It can be calculated[1] by considering an equilibrium between the energy gained from the electric field between collisions and the energy lost to the lattice on each collision. The result is that the drift velocity is proportional to the electric field magnitude, and the "constant" of proportionality is called the mobility.

$$V_d = \mu E \tag{3.14}$$

When energy equilibrium is considered[1,4] it turns out

$$\mu_n = q\frac{\tau_c}{m_n} \tag{3.15}$$

where m_n is the effective electron mass, τ_c is the mean time between collisions, and q is the elemental charge.

The current density flowing as a result of an applied field is then the average velocity times the charge times the number of carriers, (both electrons and holes). So

$$J = nqv_{d_n} + pqv_{d_p} \tag{3.16}$$

or,

$$J = (nq\,\mu_n + pq\,\mu_p)E \tag{3.17}$$

The formulation of the simple mobility equation considers only scattering off thermally agitated lattice sites. Other scattering possibilities exist. Scattering off ionised impurity sites reduces the overall mobility of an electron or hole and is dependent on the sum of all acceptors and donors present. An empirical equation to predict mobility based on doping density is often used.[1]

$$\mu = \mu_{min} + \frac{\mu_{max} - \mu_{min}}{1 + \left(\frac{N}{N_{ref}}\right)^{\alpha}} \tag{3.18}$$

Typical values for the constants are presented in table 3.2, where N is the total dopant concentration. Since mobility is related to scattering from thermally agitated lattice atoms, there is also a dependence of mobility on temperature.[1] In general as temperature increases mobility decreases.

Dopant	μ_{min} $(\frac{cm^2}{Vs})$	μ_{max} $(\frac{cm^2}{Vs})$	N_{ref} (cm^{-3})	α
Arsenic	52.2	1417	9.68×10^{16}	0.680
Phosphorus	68.5	1414	9.20×10^{16}	0.711
Boron	44.9	470.5	2.23×10^{17}	0.719

Table 3.2, Typical mobility constants for equation 3.18

Diffusion Currents.

Drift currents are dominant in metals and important in semiconductors, but there is also another carrier motion mechanism in semiconductors called diffusion current.

It is possible to have regions of varying electron or hole density in a crystal lattice. Due to thermal vibrations carriers will diffuse from regions of higher to regions of lower concentration. This is analogous to dopant diffusion in thermal processing. The rate of diffusion is proportional to the carrier concentration gradient, which means the diffusion current is also proportional to the carrier concentration gradient and can

be described by equation 3.19 for electrons and equation 3.20 for holes, where D_n and D_p are the diffusion constants for one dimension.

$$J_n = qD_n \frac{\partial n}{\partial x} \tag{3.19}$$

$$- J_p = qD_p \frac{\partial p}{\partial x} \tag{3.20}$$

By considering the equal portion of energy[1,4] it is possible to derive the Einstein relationship

$$D_n = \left[\frac{kT}{q}\right]\mu_n \tag{3.21}$$

$$D_p = \left[\frac{kT}{q}\right]\mu_p \tag{3.22}$$

between diffusion and drift current constants.

The total current flowing in a semiconductor is the sum of the diffusion and drift currents of both the electrons and holes. In three dimensions it is

$$\mathbf{J}_n = q\,\mu_n\, n\,\mathbf{E} + qD_n \nabla n \tag{3.23}$$

$$\mathbf{J}_p = q\,\mu_p\, p\,\mathbf{E} - qD_p \nabla p \tag{3.24}$$

$$\mathbf{J}_T = \mathbf{J}_n + \mathbf{J}_p \tag{3.25}$$

Minority Carrier Lifetime.

When minority carriers are injected into a region of majority carriers they temporarily disturb the equilibrium $np = n_i^2$. Although the process of recombination of electrons and holes in a indirect bandgap semiconductor such as silicon[4] is quite complex, the process of recombination can be modelled by a parameter called "the minority carrier lifetime" (τ_n τ_p). The rate of change of hole concentration in a n-type material can then be given by

$$\frac{\partial p}{\partial t} = -\frac{p'}{\tau_p} + G_p - \frac{1}{q}\nabla \cdot \mathbf{J}_p \tag{3.26}$$

where $p' = p - p_0$ is the excess carrier concentration, and G_p is the generation rate of minority carriers. A similar equation for electrons in a p-type semiconductor is

$$\frac{\partial n}{\partial t} = -\frac{n'}{\tau_n} + G_n + \frac{1}{q}\nabla \cdot \mathbf{J}_n \tag{3.27}$$

These two equations are termed the current continuity equations. A useful value from this is the mean diffusion length of a minority carrier.

$$L_t \equiv \sqrt{D_t \tau_t} \tag{3.28}$$

Where t is either n *or* p type carriers.

3.3. pn Junctions.

The pn junction is a key component of almost all integrated circuit devices.[5] It is composed of two regions, one of each type of carrier concentration, in a continuous crystal. It is useful because it blocks current flow in one direction, and allows current to pass in the other. Figure 3.4, shows the physical construction, doping concentrations, and resulting energy band diagrams for the pn junction at equilibrium and in forward and reverse bias.

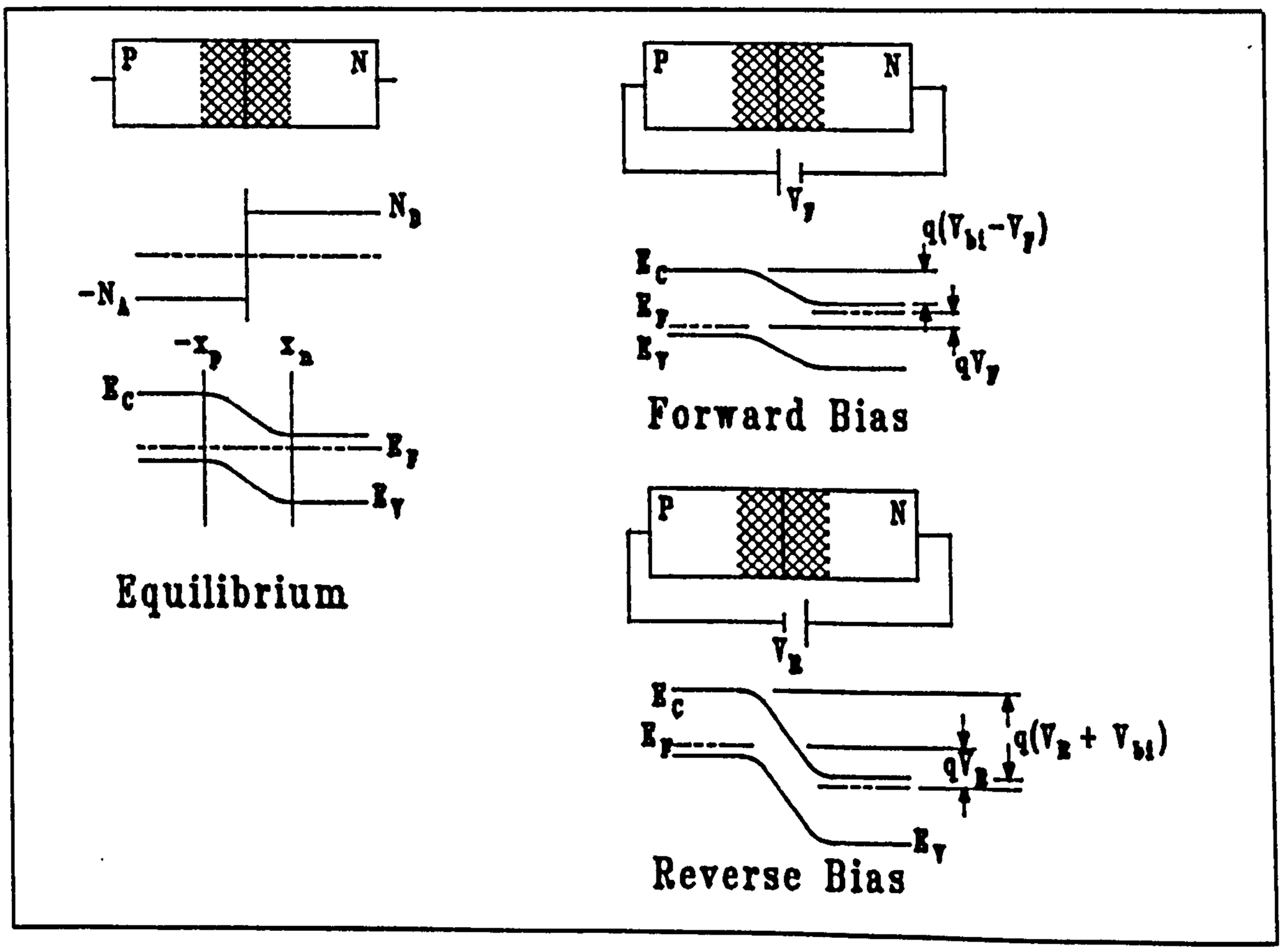


Figure 3.4, Bias conditions, and resulting band diagrams for the pn junction.

3.3.1. At Equilibrium.

Due to the high gradient of carrier concentration, electrons and holes diffuse across the junction of the two regions until an electric field is set up, by their displacement, which is sufficient to balance the diffusion current with a drift current. The displaced charge results in a space charge region, or a depletion region (since the natural charge carriers are depleted), of a given width.

Built-in Potential.

The magnitude of the potential across the junction is an important parameter. It is called the built-in potential (V_{bi}). The Einstein relation and the current continuity equations can be combined and integrated across the junction ($-x_p$ *to* x_n) to give the built-in potential for an abrupt junction.†

$$V_{bi} = \left[\frac{kT}{q}\right] \ln \left[\frac{N_D N_A}{{n_i}^2}\right] \tag{3.29}$$

Which gives the n-type material a positive voltage relative to the p-type. The abrupt junction solution is often used as a first approximation. Real junctions have some form of dopant gradient, and solutions of more complex integrals often require recourse to numerical methods.

3.3.2. Reverse Bias.

When an external bias is applied to the junction with the same polarity as the built-in potential, the p-n diode is said to be in reverse bias. The effect of this is to increase the potential energy required for charge to flow from one conduction band to the other (see figure 3.4). The net result is that no current flows. The increased barrier height also means that the space charge region must increase to maintain equilibrium, so the depletion width must increase.

---

† All of these developments contain approximations. Reference (2) goes into detail.

Depletion Width.

The depletion width can be calculated for a one sided abrupt junction with reverse bias V_R by combining the effective barrier height ($V_C = V_{bi} + V_R$) with equation 3.29, and Gauss's law to give[4]

$$\dot{W} = \left[\frac{2\epsilon_s}{qN_B} (V_R + V_{bi}) \right]^{\frac{1}{2}} \tag{3.30}$$

where N_B is the larger of N_A *or* N_D if one greatly exceeds the other. The abrupt junction approximation gets better with increasing reverse bias.

Depletion Capacitance.

Another characteristic of a reverse biased p-n junction is that of depletion layer capacitance. The capacitance can be derived on the basis of $C = \frac{\partial Q}{\partial V}$ which results in the intuitive capacitance per unit area of $C' = \frac{\epsilon_s}{W}$ where ϵ_s is the permittivity of the semiconductor.

Maximum Field.

A third parameter of use is that of the maximum electric field in the depletion region. The solution for a step junction is simply [1]

$$E_{max} = \frac{2(V_{bi} + V_R)}{W} \tag{3.31}$$

since it has a linear field distribution. The maximum field is important because if a critical value is exceeded junction breakdown occurs and current will flow, possibly destructively.

3.3.3. Forward Bias.

If a potential is applied in an opposite sense to the built-in potential it reduces the barrier height, see figure 3.4, and encourages charge to flow. A pn junction biased in that manner is said to be forward biased. The effect of the bias is to reduce the depletion width. In the forward bias case there is a relationship between current flow and applied voltage. At applied voltages lower than the built-in potential the applied

voltage reduces the electrostatic potential at the junction. The drift current is therefore reduced and diffusion current is allowed to carry the charge across the junction. Electrons drift from the n side into the p and holes the other way. Current flow is therefore by minority carrier injection.

IV Characteristics.

Through consideration of the minority carrier injection and equilibrium carrier concentrations, and assuming no generation and recombination (so that the current continuity equations can be applied), the ideal diode equation can be developed.[4]

$$J = J_s(e^{\frac{qV}{kT}} - 1) \tag{3.32}$$

Where

$$J_s \equiv \frac{qD_p p_{no}}{L_p} + \frac{qD_n n_{po}}{L_n} \tag{3.33}$$

is the saturation current and;

n_{po} is the electron density in the p region at equilibrium.

p_{no} is the hole density in the n region at equilibrium.

Lp,n *is the diffusion lengths of holes and electrons.*

V *is the applied positive bias.*

Other effects such as generation and recombination, series resistance, and high current injection levels, tend to influence the characteristics of real diodes.

Gated Diode.

An electric field at right angles to the current flow in a pn junction, as shown in figure 3.5, can effect the reverse leakage current of the junction. Such a structure exists at the source and drain regions of a surface IGFET. When there is a electric field in such a manner as to deplete the carrier concentration at the surface, the reverse bias leakage current of the diode is affected.[1] It is firstly increased by generation and recombination in the additional volume of depletion region caused by the transverse field. The second increase comes from additional surface generation recombination sites in the gated area. This effect can also occur when the structure is

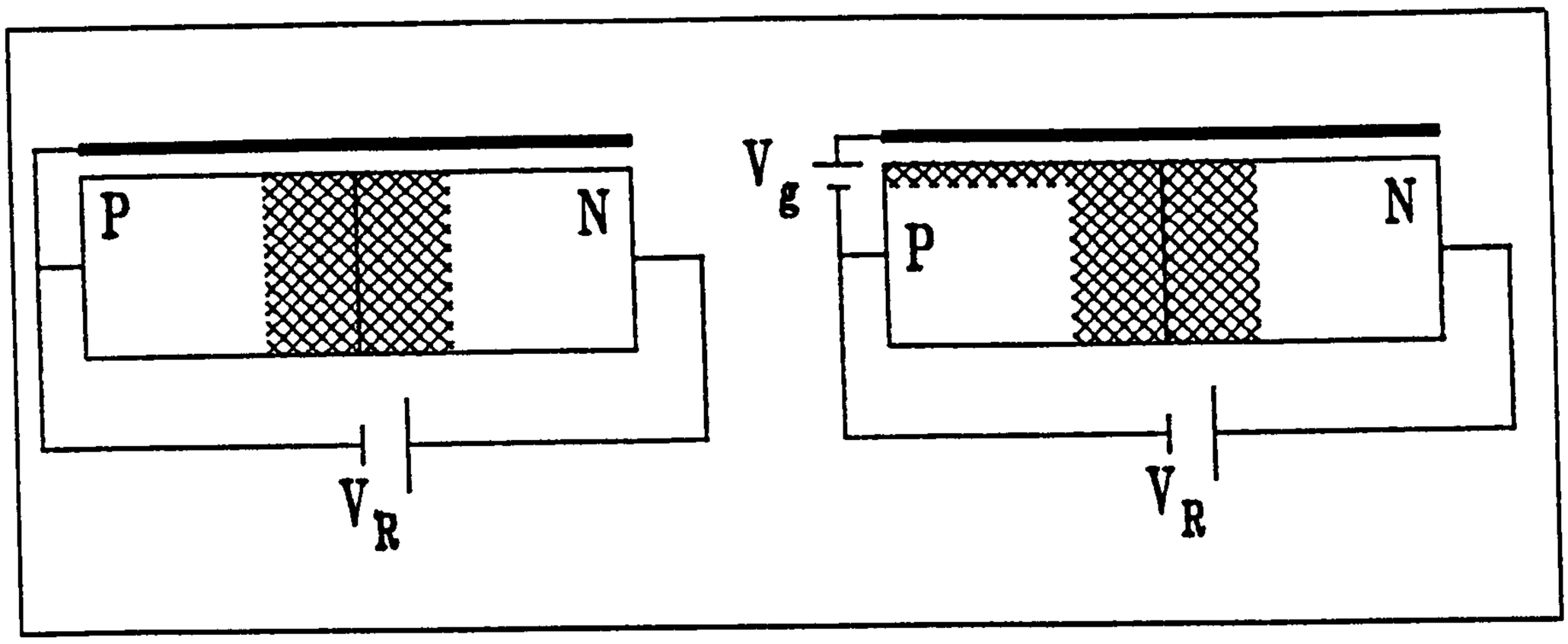


Figure 3.5, A gated pn diode, and the surface depletion region.

forward biased but the effect is small compared to the forward current.

3.3.4. Junction Breakdown.

When a sufficiently high reverse potential is applied to a pn junction, the "barrier" breaks down and allows current to flow. If heating of the junction is limited, by externally limiting the current flow, the phenomenon can be reversed by lowering the potential. There are two mechanisms for breakdown.

Zener.

Zener breakdown, or tunneling, occurs when a the electric field is sufficiently high that a valence electron on one side of the junction can make a transition directly into the conduction band on the opposite side of the junction. This can occur in silicon where the maximum field is higher than $10^6 V\ cm^{-1}$. Usually this will only occur when doping densities are greater than $5\times10^{17} cm^{-3}$.

Avalanche.

Avalanche multiplication breakdown is more likely to occur in junctions with lower doping densities, or one-sided junctions. There an electron gains sufficient energy from the electric field to cause the generation of an electron-hole pair on collision with a lattice site. The electron and hole are swept apart by the electric field. The electron, along with the incident electron, can then reach sufficient energy to repeat

the process once each. The multiplication avalanches until a substantial current flows and external series resistance limits the current or the junction is destroyed. Reference (2) provides details of calculating onset voltages.

3.4. MOS Capacitor.

The Metal Oxide Semiconductor (MOS) Capacitor is the basic element of all MOS devices. Although the IGFET structure is more complicated, the basic theory developed for the MOS Capacitor also applies to an IGFET.

3.4.1. MOS Structure.

A MOS capacitor, sometimes called a MOS diode, consists of a sandwich of a metal conductor on top of a silicon dioxide insulator over a silicon substrate. A schematic cross section in shown in figure 3.6. The term MOS is still used although other materials, such as polysilicon and the conductors mentioned in chapter 2, have been used for the top electrode.

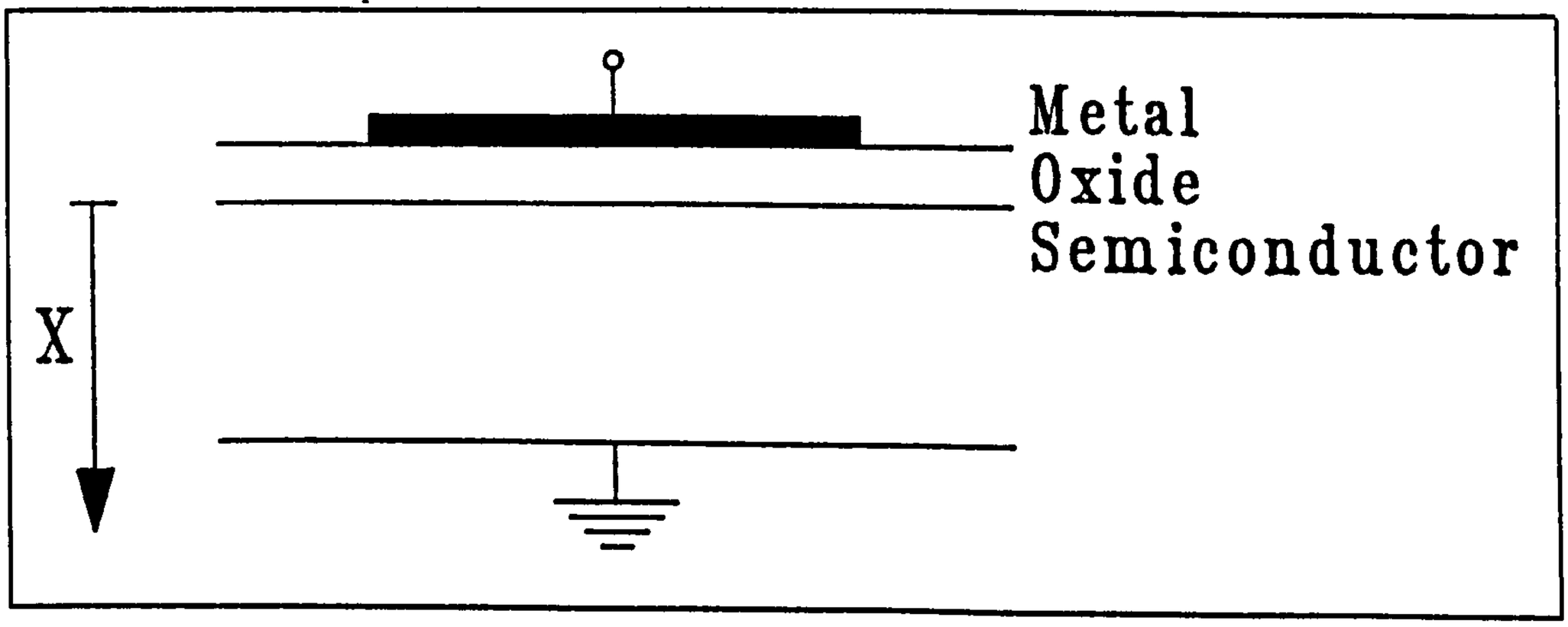


Figure 3.6, The basic structure of an MOS Capacitor.

The silicon is doped either p or n type, in this case p-type will be used as an example. The n-type case can be determined similarly by considering the change in polarity of the carriers. Most doping densities are on the order of 10^{15} *or* $10^{16} Atoms\ cm^{-3}$.

Band Diagrams.

Ignoring a number of practical influences on the electron energy levels, which will be considered later, the band diagrams for an MOS capacitor, and any MOS system, are shown in figure 3.7.

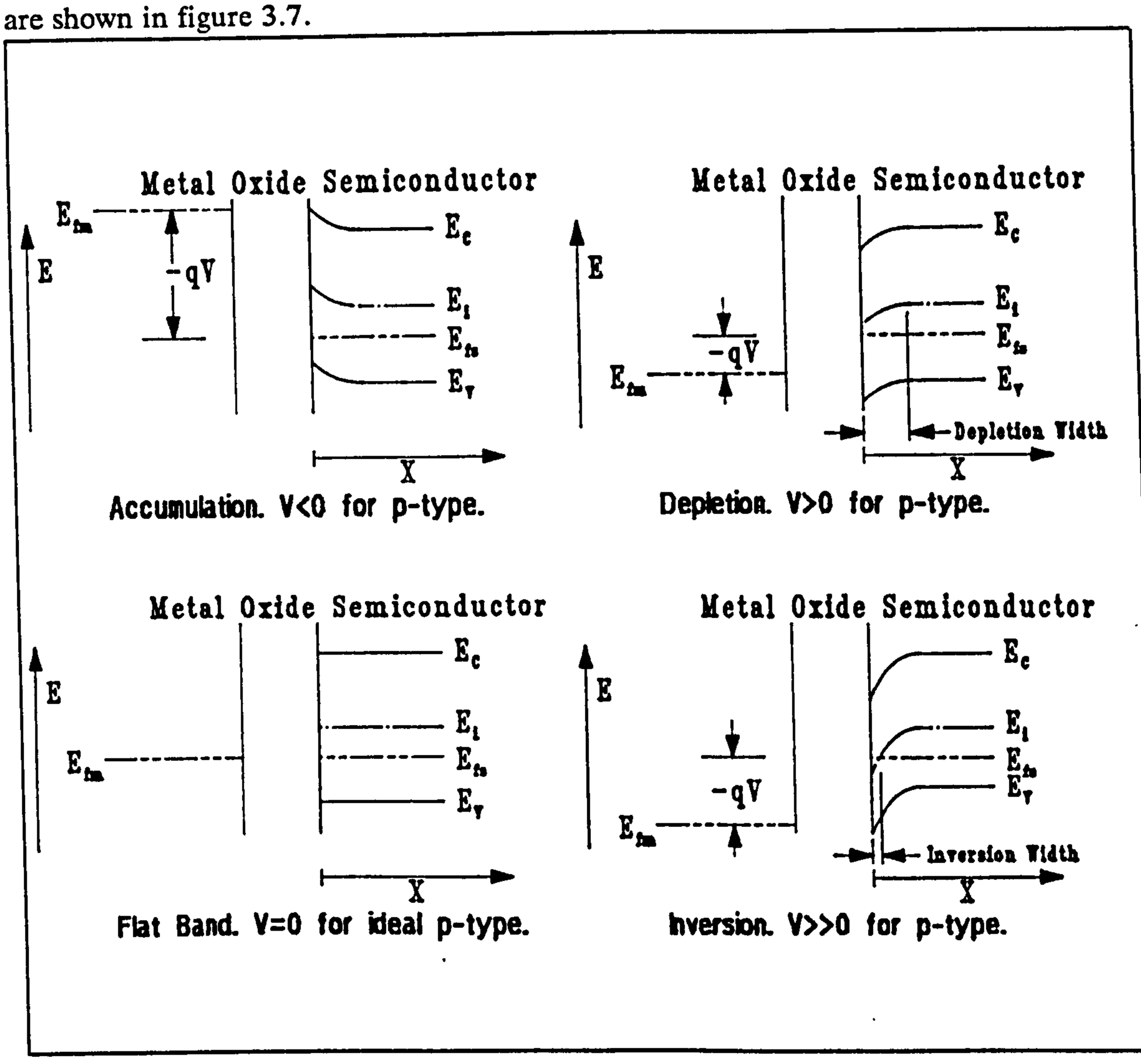


Figure 3.7, Bias conditions and energy diagrams for the MOS Capacitor.

By convention, the x-axis is taken as increasing into the semiconductor bulk from its origin at the silicon surface. Again E_C and E_V are the energies at the edge of the conduction and valence bands. E_F is again the Fermi energy level in the semiconductor and E_i is introduced as the intrinsic Fermi level. The convention in band diagrams is to take the Fermi energy level as a reference. As a result it may appear confusing to some that everything else changes, where in reality it is the centroid of the electron

energy probability which is altered.

3.4.2. Ideal Effect of Bias Voltage.

As the bias, V_G, between the top, gate, electrode and the bulk is changed four distinct regions can be identified from the band diagrams.

Accumulation.

Accumulation occurs, for a p-type semiconductor, when the gate is biased negatively relative to the substrate. In this state, see figure 3.7, holes are attracted to the surface of the semiconductor, and it becomes more p-type as they are accumulated.

Flat Band.

As the bias is reduced, to zero in the ideal case, the MOS structure achieves a state called Flat Band where the carrier densities are influenced by the doping only. The term flat band comes from the fact there is no band bending in this state.

Depletion.

As the bias is increased from flat band, the bands bend downwards at the surface as holes are repelled and electrons attracted. The surface becomes depleted of its natural carriers. The name depletion is given to this state, which has a characteristic depletion depth, also shown in diagram 3.7.

Inversion.

If the bias is further increased the surface will be completely depleted of its intrinsic carrier, holes in this case, and will start to collect electrons thereby inverting the surface polarity. This mode is called inversion and is characterised in band diagrams by the fermi level crossing the intrinsic fermi level band. An important parameter is the inversion layer depth, which is shown in diagram 3.7 as well.

3.4.3. MOS Formulae

Formulae for the MOS diode[1,2,4,7-10] can be developed if we assume that;

1) The work function difference between the metal and the semiconductor is zero.
2) The only charges which exist, are those in the semiconductor and their mirrors at the metal oxide interface.
3) The resistivity of the oxide is infinite.

A useful construct, for analysis of MOS systems, is the electrostatic potential (Ψ), which is defined as being zero in the semiconductor bulk. At the surface, the electrostatic potential is given a subscript (Ψ_S) and is called the surface potential. It is then possible to define the electron and hole concentrations at the surface.

$$n_s = n_i e^{\frac{q(\Psi_S - \Psi_B)}{kT}} \tag{3.34}$$

$$p_s = p_i e^{\frac{q(\Psi_B - \Psi_S)}{kT}} \tag{3.35}$$

The states of the system can then be mathematically described as in table 3.3.[1]

Ideal p-type semiconductor			
Surface Potential	Mode	Surface Carriers Concentration	Gate Voltage
$\Psi_S < 0$	Accumulation	$p_s > N_A$	$V_G < 0$
$\Psi_S = 0$	Flat Band	$p_s = N_A$	$V_G = 0$
$\Psi_B > \Psi_S > 0$	Depletion	$p_s < N_A$	$V_T > V_G > 0$
$\Psi_S = \Psi_B$	Mid Gap	$p_s = n_i$	
$\Psi_S = 2\lvert\Psi_B\rvert$	Inversion	$n_s = N_A$	$V_G = V_T$

Table 3.3, The states for a MOS System.

The potential as a function of distance into the bulk of the silicon can be determined from the one dimensional Poisson equation,

$$\frac{\partial^2 \Psi}{\partial x^2} = \frac{-p_s(x)}{\epsilon_s} \tag{3.36}$$

where $p_s(x)$ is the total space charge density at point x. If the surface is depleted then the charge within the semiconductor can be given as

$$p_s = -qN_A \tag{3.37}$$

It is then possible to solve for the surface potential,[4]

$$\Psi_S = \frac{qN_A W^2}{2\epsilon_s} \tag{3.38}$$

where W is the depletion width. If the criteria for strong inversion is that the electron concentration equals the doping concentration, then equations 3.34 and 3.38 can be used to give

$$\Psi_S(inv) \approx 2\Psi_B = \frac{2kT}{q} \ln\left[\frac{N_A}{n_i}\right] \tag{3.39}$$

which agrees with what is obvious from band bending. The maximum equilibrium depletion width can be given by,

$$W_m = L_D\sqrt{2\beta\Psi_S} \tag{3.40}$$

where

$$L_D = \left[\frac{kT\epsilon_s}{q^2 N_A}\right]^{\frac{1}{2}} \tag{3.41}$$

and

$$\beta = \frac{q}{kT} \tag{3.42}$$

L_D is a characteristic length called the Debye length.

The capacitance of the structure is composed of two terms in series. The oxide capacitance per unit area, which is

$$C_{ox} = \frac{\epsilon_{ox}}{T_{ox}} \tag{3.43}$$

and the depletion layer capacitance

$$C_D = \frac{\epsilon_s}{W} \tag{3.44}$$

so

$$\frac{1}{C}=\frac{1}{C_D}+\frac{1}{C_{ox}} \tag{3.45}$$

in accumulation $C=C_{ox}$ and at flat band $\Psi_S=0$ so $C_D=\frac{\epsilon_S}{L_D}$.

The voltage at the metal surface is

$$V=V_{ox}+\Psi_S \tag{3.46}$$

where V_{ox} is the voltage across the oxide

$$V_{ox}=\frac{Q_s}{C_{ox}} \tag{3.47}$$

where Q_s is the charge at the surface. A useful value, called the threshold voltage (V_T), is the voltage at which the surface will be in strong inversion. From equation 3.39 one gets

$$V_T=V_{ox}+2\Psi_B=\frac{Q_s}{C_{ox}}+2\Psi_B \tag{3.48}$$

Since a depletion approximation is being used the total surface charge per unit area is just the electron density multiplied by the depletion width. The substrate doping can be substituted for n_s by definition of strong inversion. So

$$Q_s=qn_sW=qN_AW \tag{3.49}$$

and

$$V_T=\frac{qN_AW_m}{C_{ox}}+2\Psi_B\approx\frac{\sqrt{2\epsilon_s qN_A(2\Psi_B)}}{C_{ox}}+2\Psi_B \tag{3.50}$$

the corresponding minimum system capacitance at that bias is

$$C_{\min}=\frac{\epsilon_{ox}}{T_{ox}+(\frac{\epsilon_{ox}}{\epsilon_s})W_m} \tag{3.51}$$

3.4.4. Practical Effects.

Like most theoretical analysis, the simplifying assumptions made are often wrong and corrections need to be applied. The corrections necessary are for; charges in the oxide and at the interface, work function differences, and external effects.

Oxide and Interface Charges.

There are four types of charge associated with the oxide and the interface; Q_f the fixed interface charge density, Q_m the mobile oxide charge, Q_{ot} the oxide trapped charge, Q_{it} the interface trapped charge.

The fixed interface charge density is mostly due to incompletely compensated silicon in the surface of the oxide, left over as the oxidation was finished. It is fixed in the sense that it can neither be charged or discharged through a wide range of surface potential.

Mobile oxide charge is more of a problem and is usually caused by mobile ionic impurities, such as sodium, which have been incorporated into the oxide.

Oxide trapped charge is caused by electrons or holes attaching themselves to defects in the oxide layer. These defects can be caused by radiation during processing.

Interface trapped charge is caused by the incomplete bonding of the surface silicon to silicon dioxide. It is surface orientation dependent with <100> having a lower density than <111>. In <100> silicon, with modern processing, interface traps are often considered negligible.

The overall oxide and interface charge density

$$Q_o = Q_f + Q_m + Q_{ot} + Q_{it} \tag{3.52}$$

has an effect on the threshold voltage by introducing an offset

$$\Delta V_T = -\frac{Q_f + Q_m + Q_{ot} + Q_{it}}{C_{ox}} \tag{3.53}$$

Work Function Differences.

The work function for a material is the energy required to raise the energy level of an electron from that of a conduction electron to that of a free vacuum electron. It provides a way of referencing energy levels in different solids. If the work function of the metal is Φ_m and the silicon is Φ_s then the difference $\Phi_{ms} = \Phi_m - \Phi_s$ is the work function difference and has the effect of influencing the threshold voltage. The sum of all the offsets to the threshold voltage gives a voltage

$$V_{FB} = \Phi_{ms} - \frac{Q_f + Q_m + Q_{ot} + Q_{it}}{C_{ox}} \tag{3.54}$$

which is the voltage that would have to be applied between the crystal bulk and the gate to achieve flat band conditions. it is known as the flat band voltage. The threshold voltage then becomes

$$V_T \approx V_{FB} + \frac{\sqrt{2\epsilon_s q N_A (2\Psi_B)}}{C_{ox}} + 2\Psi_B \qquad (3.55)$$

Some work function differences which are doping concentration dependent are given in table 3.4.[4]

Materials	Doping cm^{-3}	
	10^{14} Φ_{ms} (V)	10^{18} Φ_{ms} (V)
Aluminium to n-type silicon	-0.36	-0.15
n^+polysilicon to n-type silicon	-0.52	-0.35
Aluminium to p-type silicon	-0.81	-0.95
n^+polysilicon to p-type silicon	-0.96	-1.20

Table 3.4, Work function difference for common I.C. gate materials.

External Effects.

There is also the possibility of local changes to the flat band voltage after processing. These effects can occur in devices with high fields parallel to the surface of the semiconductor. Electrons in these fields can gain sufficient energy to breach the silicon-silicon dioxide barrier and to become trapped in the oxide. There they increase the fixed charge density which changes the flat band voltage. Such electrons are usually called hot electrons because their energy is many times that of a thermal electron.

3.5. IGFET's.

An IGFET is composed of two back to back pn junctions with an MOS Capacitor linking them. Voltage on the gate electrode of the IGFET controls the current passing from the source to the drain. Since the charge must be transferred through the inversion region created by the gate field, only one type of carrier is predominately

responsible for the current. Figure 3.8, shows the basic IGFET structure and defines the important features.

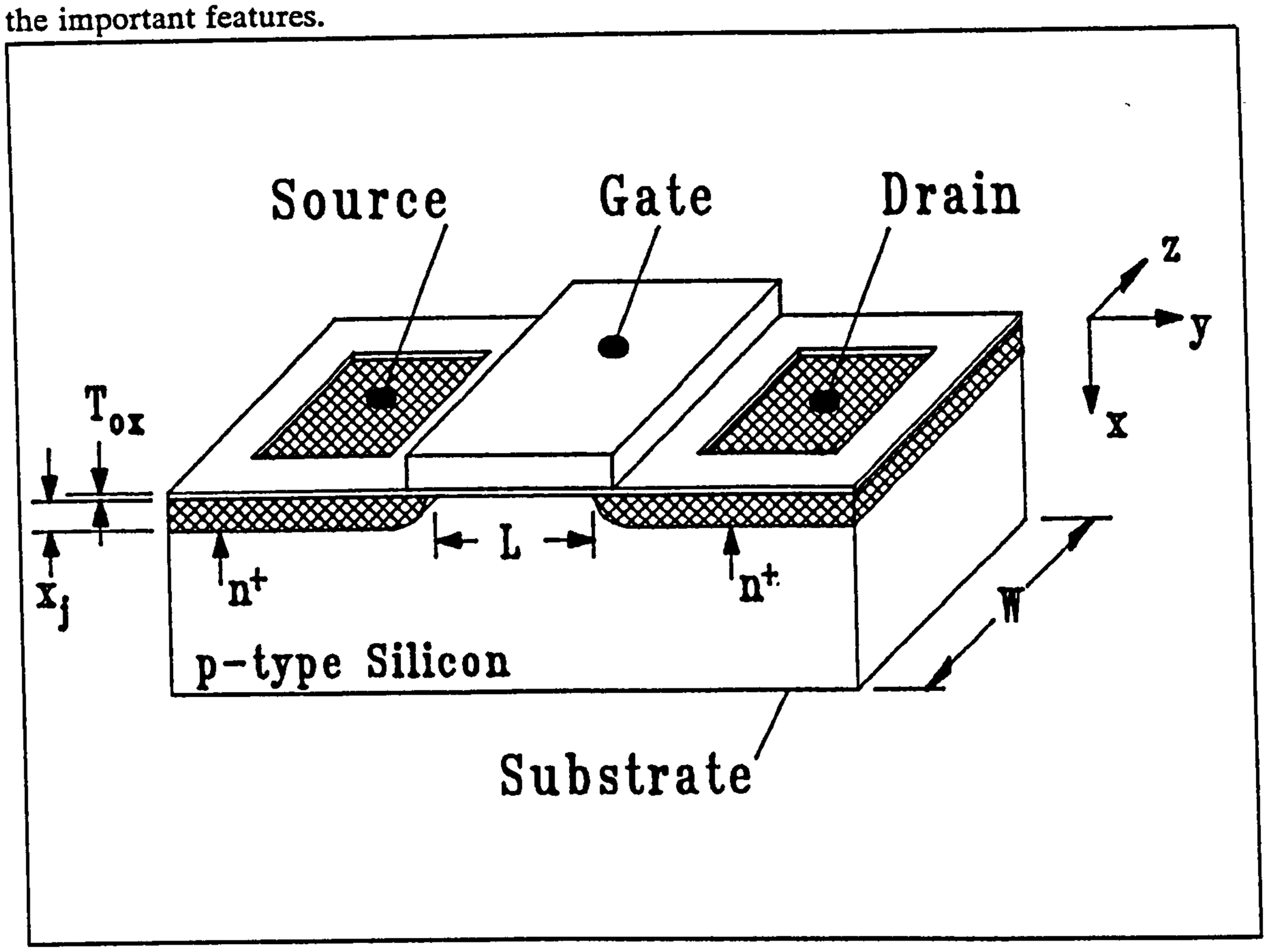


Figure 3.8. Basic IGFET Structure.

3.5.1. Structure and Basic Theory.

There are many models for the source to drain current as a function of the electrode biases. All of them have a restricted range of validity.[11-17]

A simple intuitive, but very restricted model[1,18] can be derived by considering a uniform inversion layer in the channel region (the area below the silicon-silicon dioxide interface and between the source and drain depletions), as in figure 3.9.

The charge in that inversion layer is

$$Q = V_G C_{ox} = V_G LW \frac{\epsilon_{ox}}{T_{ox}} \tag{3.56}$$

for a device with zero flat band voltage ($V_{FB} = 0$). If a very small drain voltage relative

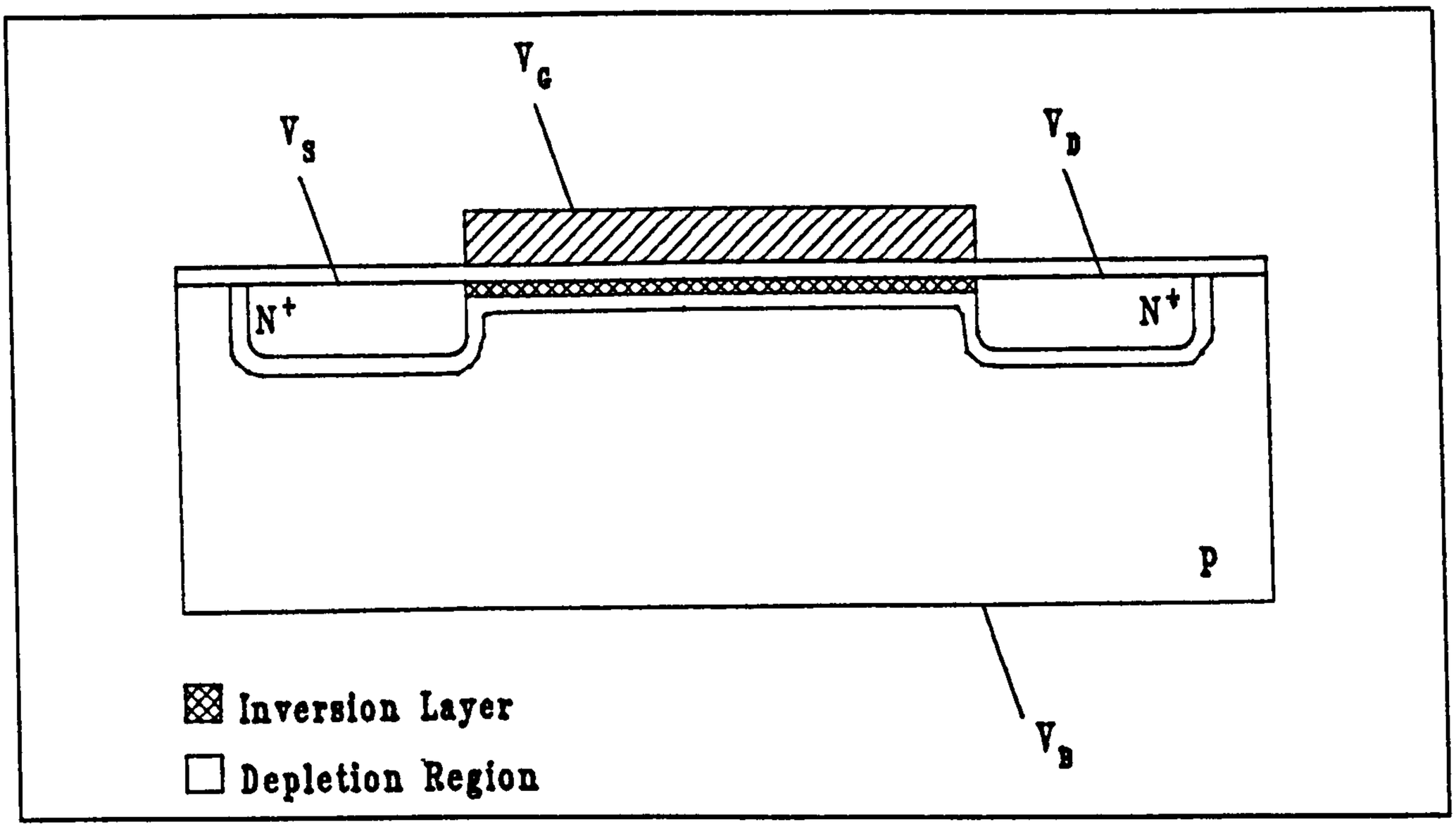


Figure 3.9, IGFET in inversion.

to the source is applied so that the uniformity of the inversion layer is not changed, the current at the drain would just be the charge in the channel divided by the transit time along the channel.

$$I = \frac{Q}{t} \tag{3.57}$$

The transit time is given by the length of the channel divided by the carrier velocity.

$$t = \frac{L}{v} \tag{3.58}$$

$$v = \mu E \tag{3.59}$$

$$E = \frac{V}{L} \tag{3.60}$$

so

$$t = \frac{L^2}{\mu V_D} \tag{3.61}$$

and the current becomes

$$I_D = \left(\frac{\mu V_D}{L^2}\right)\left(\frac{LW\epsilon_{ox}}{T_{ox}}\right)V_G \tag{3.62}$$

which rationalises to

$$I_D = (\frac{\mu \epsilon_{ox}}{T_{ox}})(\frac{W}{L}) V_D V_G \quad for \quad V_G >> V_D \tag{3.63}$$

However, an equation which is valid for a wider range of biases is necessary for transistor models.

Operation Regions.

There are three main modes of transistor operation. The mode of operation depends not only on device construction but also on the relative potential of each of the electrodes.

Two break points define the limits of the regions of operation. The first is the threshold. With a low drain voltage and sufficient voltage on the gate of the transistor, an inversion region can be formed along the entire length of the channel. This is equivalent to the threshold of the MOS capacitor. The second point is the onset of saturation. If the drain voltage is increased eventually the drain depletion region will extend into the channel region and deplete the inversion layer of electrons. Any further increase in drain potential does not result in an increase in current since the inversion region is supplying the maximum amount of charge possible for that gate voltage. The result is a saturated drain current that in the ideal case is not effected by increases in drain bias. The point in the channel at which the drain depletion has reduced the inversion layer width to zero is called the pinch off point.

The three regions of operation are then; the subthreshold region where small amounts of charge are transferred without an strong inversion layer existing, the linear region between the threshold and saturation points where the drain current is linearly proportional to the drain voltage and the constant of proportionality depends on the gate voltage, and the saturation region where the drain current is independent of drain voltage but controlled by the gate voltage.

3.5.2. Conventional Models.

Conventional analysis[1,2,4,12,13,15,18] of the IGFET, which gives a good fit for most long channel devices, considers the charge contributions in the channel from both the inversion layer, caused by the gate electric field, and the depletion layers, caused by the gate and drain electric fields.

As with the previous simple model, some simplifications must be made which also limit the range of validity of the model. In the conventional analysis it is assumed that the electric field perpendicular to the current flow is much larger than the tangential electric field caused by the drain bias. This assumption is made so that one dimensional descriptions of the carrier density can be used so Poisson's equation can be easily solved. This approximation is called the "gradual channel approximation" since the vertical field varies gradually across the channel. The assumptions that charge is carried only by a single carrier type in the channel, and that the channel is long compared to the junction depths, are also made.

Linear region.

The drain current of an IGFET biased below saturation must, by reasons of continuity , flow through any incremental section of the channel. The current through an incremental section, as shown in figure 3.10, can be calculated by considering

$$I_D = \frac{\partial Q}{\partial t} \tag{3.64}$$

where ∂Q is the charge in the incremental section, and ∂t is the time required to transfer that charge across the incremental section. As before the time is given by

$$\partial t = \frac{\partial y}{v_D} = \partial y \times \frac{\partial y}{\mu \partial V_c} \tag{3.65}$$

where ∂V_c is the incremental voltage drop. Therefore

$$I_D = \frac{\partial Q \, \mu \partial V_c}{(\partial y)^2} \tag{3.66}$$

and if $Q_n(V_G)$ is the inversion charge per unit of surface area then

$$\partial Q = Q_n(V_c) W \, \partial y \tag{3.67}$$

which gives

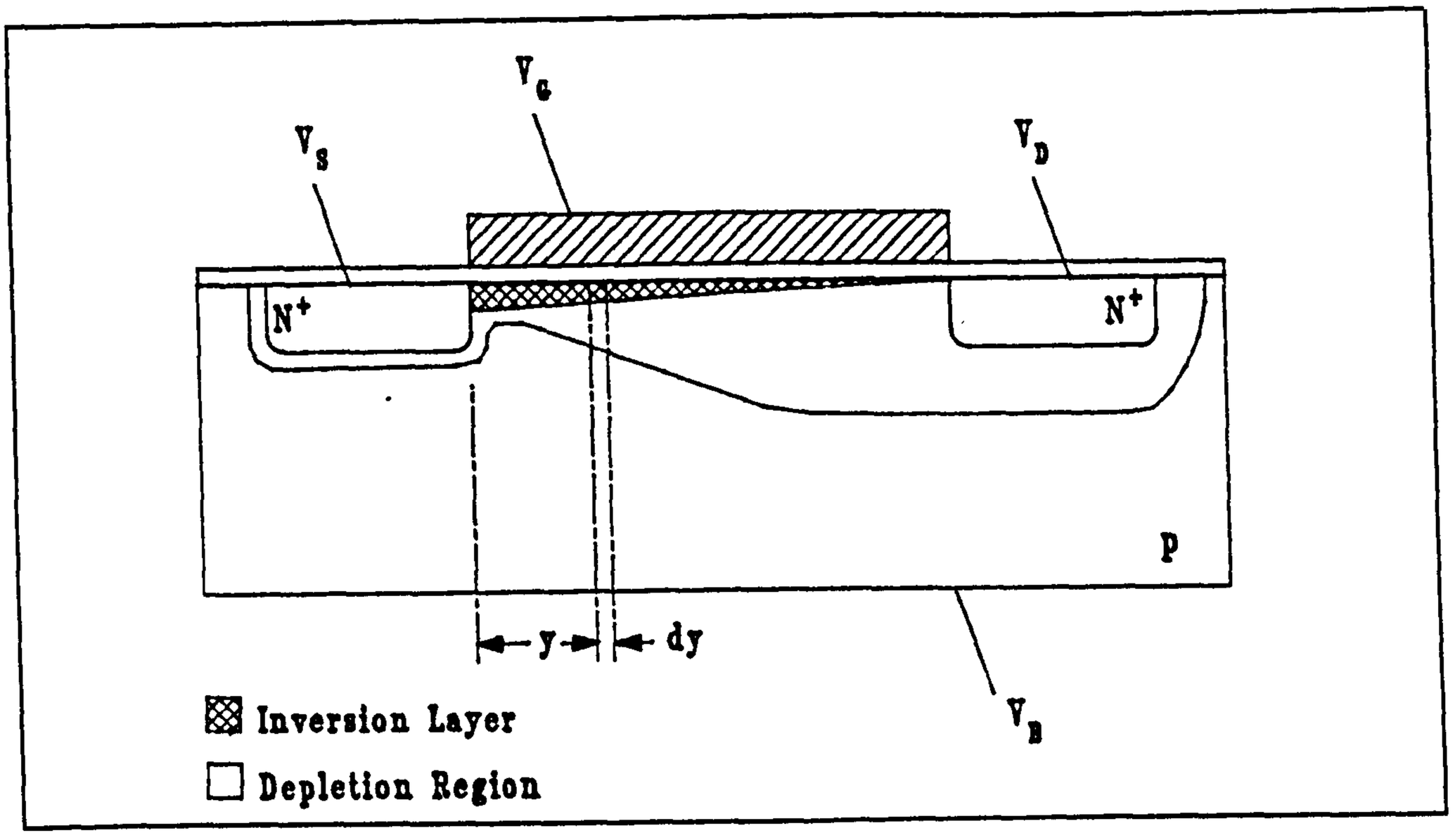


Figure 3.10, IGFET near saturation.

$$I_D = Q_n(V_c) W \mu \frac{\partial V_c}{\partial y} \tag{3.68}$$

Then by re-arranging and integrating equation 3.68 over the channel

$$\int_{y=0}^{L} I_D \, \partial y = W \mu \int_{y=0}^{L} Q_n(V_c) \partial V_c \tag{3.69}$$

the expression

$$I_D = -\mu_n \frac{W}{L} \int_{V_c=V_s}^{V_D} Q_n(V_c) \partial V_c \tag{3.70}$$

for the drain current can be found.

The inversion layer charge in the channel is equal to the total charge minus the depletion charge,

$$Q_n(V_c) = Q_s - Q_D \tag{3.71}$$

or using equations 3.40, 3.41, 3.42, and 3.49 one gets

$$Q_D = \sqrt{2qN_A \epsilon_s \Psi_S} \tag{3.72}$$

The surface potential of the depletion region, Ψ_S which is measured relative to the bulk can be calculated considering the voltage of the surface of the semiconductor and

working down. The surface at inversion is defined as having twice the difference between the intrinsic and doped Fermi levels, or

$$\Psi = 2|\Psi_B| \tag{3.73}$$

relative to the bulk Fermi level. The depletion at point y has a relative potential of $V_c - V_B$. So the surface potential must be

$$\Psi_S = 2|\Psi_B| + V_c - V_B \tag{3.74}$$

therefore

$$Q_D = \sqrt{2q\epsilon_s N_A (2|\Psi_B| + V_c - V_B)} \tag{3.75}$$

The total charge on the gate can be given by

$$Q = C_{ox} V_{ox} \tag{3.76}$$

where V_{ox} is the voltage across the insulator. It is given by

$$V_{ox} = V_G - V_{FB} - 2|\Psi_B| - V_c \tag{3.77}$$

yielding

$$Q = C_{ox}(V_G - V_{FB} - 2|\Psi_B| - V_C) \tag{3.78}$$

but

$$Q_s = -Q \tag{3.79}$$

by balanced charge across the oxide. So finally combining equations 3.71, 3.75, and 3.78 yields

$$Q_n(V_c) = -C_{ox}(V_G - V_{FB} - 2|\Psi_B| - V_c) - \sqrt{2q\epsilon_s N_A (2|\Psi_B| + V_c - V_B)} \tag{3.80}$$

Which put back into equation 3.70 gives the solution for the drain current.

$$I_{DS} = \frac{W}{L}\mu_n C_{ox}\left[(V_G - V_{FB} - 2|\Psi_B| - \frac{V_D}{2} - \frac{V_S}{2})V_{DS}\right.$$

$$\left. - \frac{2}{3}\sqrt{2\epsilon_s q N_A}\left((2|\Psi_B| + V_D - V_B)^{\frac{3}{2}} - (2|\Psi_B| + V_S - V_B)^{\frac{3}{2}}\right)\right] \tag{3.81}$$

Where $V_{DS} = V_D - V_S$.

This formula is valid for as long as the channel has an inversion region from the source to drain. That is for $V_G > V_T$ and $I_D < I_{D_{sat}}$, where $I_{D_{sat}}$ is the saturation current.

A simpler form of equation 3.81 can be derived by neglecting the variation in depletion charge across the channel in equation 3.80 to give

$$Q_n(V_c) = -C_{ox}(V_G - V_{FB} - 2|\Psi_B| - V_c) - \sqrt{2q\epsilon_s N_A 2|\Psi_B|} \tag{3.82}$$

and considering the threshold voltage equation 3.55 one can get

$$Q_n(V_c) = -C_{ox}(V_G - V_T - V_c) \tag{3.83}$$

which on integration yields

$$I_{DS} = \mu_n \frac{W}{L} C_{ox} \left[(V_G - V_T)V_{DS} - \frac{{V_{DS}}^2}{2} \right] \tag{3.84}$$

This solution is called the charge control model because by neglecting the variation of the depletion depth with position in the channel the entire depletion charge must be assumed to be controlled by the gate only.

Saturation region.

When V_D exceeds $V_{D_{sat}}$ the drain current stays at a fixed value and does not appreciably increase when the drain voltage increases. In the charge control model the saturation voltage can be determined as follows; [1]

At the onset of saturation the charge density at the drain side of the channel will be zero and $V_c = V_D = V_{D_{sat}}$ the drain voltage at saturation, or

$$Q_n(L) = 0 = -C_{ox}(V_G - V_T - V_{D_{sat}}) \tag{3.85}$$

so

$$V_{D_{sat}} = V_G - V_T \tag{3.86}$$

and

$$I_{D_{sat}} = \frac{\mu_n W C_{ox}}{2L}(V_G - V_T)^2 \tag{3.87}$$

with the limit $V_G < V_D + V_T$. A similar derivation can be made for the complete solution.

Subthreshold region.

The other important region of IGFET operation is the subthreshold region.[6,15-17,19-21] When the gate voltage is less than the threshold voltage there is still some inversion charge in the channel, which can give rise to a subthreshold current.

The subthreshold region is characterised by low free carrier densities. The carrier densities are so low that drift currents (which are proportional to the carrier density) are less than diffusion currents (which are proportional to the gradient of the carrier density). The subthreshold region of operation is therefore dominated by diffusion currents and has a similar operation to an npn bipolar transistor (for a p-type doped channel). The current can be given by

$$I_D = -qAD_n \frac{\partial n}{\partial y} \tag{3.88}$$

where A is the channel cross section normal to the current flow.

Since the current through the channel must be constant, due to continuity, the free carrier gradient must be constant also. Therefore,

$$\frac{\partial n}{\partial y} = \frac{n(L)-n(0)}{L} \tag{3.89}$$

Then,

$$I_D = qAD\frac{n(0)-n(L)}{L} \tag{3.90}$$

and using

$$n = n_i e^{\beta(\Psi-\Psi_n)} = n_i e^{\beta(\Psi-\xi-\Psi_p)} \tag{3.91}$$

where Ψ_n is the quasi Fermi level, and $\xi = \Psi_n - \Psi_p$ is the difference between quasi Fermi levels. Since there is no hole current $\Psi_p = \Psi_F$. The difference in the quasi Fermi levels at each end of the channel is the applied voltage relative to the bulk, so;

$$n(L) = n_i e^{\beta(\Psi_s - V_{DB} - \Psi_F)} \tag{3.92}$$

and

$$n(0) = n_i e^{\beta(\Psi_s - V_{SB} - \Psi_F)} \tag{3.93}$$

The area of the current flow can be determined by taking the product of the transistor width W and the effective channel thickness. Since there is an exponential dependence it is adequate to lump the current into an effective depth defined by the first order of the exponent of the carrier concentration. In other words, the thickness is the distance by which Ψ decreases by β^{-1}. So from the definition of the electric field one can get

$$t = \frac{V}{E} = \frac{\nabla\Psi}{E_s} = \frac{1}{\beta E_s} \tag{3.94}$$

where E_s is the electric field at the surface. From Gausses law, and neglecting the small inversion charge, the electric field supports the depletion charge. Or,

$$\epsilon_s E_s = -Q_B = \sqrt{2\epsilon_s q N_B \Psi_S} \tag{3.95}$$

Therefore equation 3.90 becomes

$$I_D = \frac{WqDn_i}{L}\left[e^{\beta(\Psi_S - V_{SB} - \Psi_F)} - e^{\beta(\Psi_S - V_{DB} - \Psi_F)}\right]\beta^{-1}\left[\frac{\epsilon_s}{2qN_B\Psi_s}\right]^{\frac{1}{2}} \tag{3.96}$$

Using the definition of the Fermi level

$$N_B = n_i e^{\beta\Psi_F} \tag{3.97}$$

and defining the intrinsic Debye length as

$$L_i = \left[\frac{\epsilon_s}{2qn_i\beta}\right]^{\frac{1}{2}} \tag{3.98}$$

and using $V_{DB} = V_{DS} + V_{SB}$ then equation 3.96 becomes

$$I_D = \frac{WqDn_iL_i}{L}\left[\frac{e^{\beta(\Psi_S - V_{SB} - 1.5\Psi_F)}}{\sqrt{\beta\Psi_S}}(1 - e^{-\beta V_{DS}})\right] \tag{3.99}$$

Using $C_{ox} = \frac{\epsilon_s}{T_{ox}}$ the surface potential can be related to the gate voltage by

$$V_{GB} - V_{FB} = \Psi_S - \frac{Q_B}{C_{ox}} = \Psi_S + \frac{1}{C_{ox}}\sqrt{2\epsilon_s q N_B \Psi_S} \tag{3.100}$$

and if the source is used as the reference voltage by substituting

$$\Psi_S = \Psi_S + V_{SB} \tag{3.101}$$

into the previous equations one gets the solutions

$$I_D = \frac{WL_i qDn_i}{L} \frac{e^{\beta(\Psi_S - 1.5\Psi_F)}}{\sqrt{\beta(\Psi_S + V_{SB})}} \left[1 - e^{-\beta V_{DS}}\right] \tag{3.102}$$

and

$$V_{GS} - V_{FB} = \Psi_S + \frac{1}{C_{ox}} \sqrt{2q\epsilon_s N_B (\Psi_s + V_{SB})} \tag{3.103}$$

to describe the subthreshold region.

An explicit relationship between I_D and V_{GS} can be developed by defining a depletion layer capacitance

$$C_B = \frac{\partial(-Q_B)}{\partial \Psi_s} \tag{3.104}$$

which yields

$$C_B = \left[\frac{\epsilon_s qN_B}{2(\Psi_S + V_{SB})}\right]^{\frac{1}{2}} \tag{3.105}$$

Using this equation along with the Einstein relationship allows equation 3.102 to be rearranged to give

$$I_D = \mu \frac{W}{L} \left(\frac{kT}{q}\right) \times qL_i n_i \frac{e^{\beta(\Psi_S - 1.5\Psi_F)}}{\sqrt{\beta(\Psi_S + V_{SB})}} \left[1 - e^{-\beta V_{DS}}\right] \tag{3.106}$$

but equation 3.105 can also be rearranged by using equation 3.97 to give a similar form

$$C_B = \frac{L_i qn_i \beta e^{\beta \frac{\Psi_F}{2}}}{\sqrt{\beta(\Psi_S + V_{SB})}} \tag{3.107}$$

Substitution yields

$$I_D = C_B \mu \frac{W}{L} \left(\frac{kT}{q}\right)^2 e^{\beta(\Psi_S - 1.5\Psi_F)} \left[1 - e^{-\beta V_{DS}}\right] e^{-\beta \frac{\Psi_F}{2}} \tag{3.108}$$

Now by differentiating equation 3.100 by $\partial \Psi_S$ and rearranging, it is possible to define a simple linear function

$$m(\Psi_S) \equiv 1 + \frac{C_B(\Psi_S)}{C_{ox}} = \frac{\partial V_{GS}}{\partial \Psi_S} \tag{3.109}$$

which allows further simplification of equation 3.108. If two constants are defined about the mid point of the subthreshold region $\Psi_F < \Psi_S < 2\Psi_F$, $m^* = m(1.5\Psi_F)$ *and* $V_{GS}{}^* = V_{GS}(1.5\Psi_F)$, then for small offsets

$$V_{GS} - V_{GS}{}^* = m^* \times (\Psi_S - 1.5\Psi_F) \tag{3.110}$$

which can be rearranged to give

$$\Psi_S - 1.5\Psi_F = \frac{1}{m^*}(V_{GS} - V_{GS}{}^*) \tag{3.111}$$

so equation 3.108 becomes

$$I_D = C_B \mu \frac{W}{L}\left(\frac{kT}{q}\right)^2 \left[e^{\frac{\beta}{m^*}(V_{GS}-V_{GS}{}^*)-\beta\frac{\Psi_F}{2}}\right]\left[1-e^{-\beta V_{DS}}\right] \tag{3.110}$$

A popular parameter for describing the subthreshold behaviour is the voltage swing required to reduce the drain current by one decade, or

$$S \equiv ln(10) \times \frac{\partial V_{GS}}{\partial(ln(I_D))} \tag{3.113}$$

which after taking the natural log of equation 3.112 and doing the derivative yields the expression

$$S = \frac{kT}{q} ln(10) \times \left[1 + \frac{C_B(1.5\Psi_F)}{C_{ox}}\right] \tag{3.114}$$

This model of the subthreshold ignores the inversion charge in the channel. A more complete model[19] considers the bulk charge to include the inversion charge. Equation 3.37 then becomes

$$p(x) = q(N_D - N_A + p_p - n_p) \tag{3.115}$$

which results in a more complex Poisson equation to solve. It is possible to solve that equation and the resulting subthreshold equations are of interest for later chapters.

$$a \equiv \sqrt{2}\left(\frac{\epsilon_s}{\epsilon_{ox}}\right)\left(\frac{T_{ox}}{L_D}\right) \tag{3.116}$$

$$L_D = \sqrt{\frac{kT\,\epsilon_s}{q^2 N_A}} \tag{3.117}$$

$$C_D = \frac{\epsilon_s}{\sqrt{2}L_D} \frac{\left[1 - e^{-\beta\Psi_S} + \left(\frac{n_{po}}{p_{po}}\right) e^{(\beta\Psi_S - 1)}\right]}{F\left(\beta\Psi_S, \frac{n_{po}}{p_{po}}\right)} \tag{3.118}$$

$$F \equiv \left[(e^{-\beta\Psi} + \beta\Psi - 1) + \frac{n_{po}}{p_{po}}(e^{\beta\Psi} - \beta\Psi - 1)\right]^{\frac{1}{2}} \tag{3.119}$$

All of these needs to be evaluated at $\Psi_S = 1.5\Psi_F$ to obtain the swing.

$$S = \frac{kT}{q} ln(10) \times \left[1 + \frac{C_B}{C_{ox}}\right]\left[1 - \left(\frac{2}{a^2}\right)\left[\frac{C_D}{C_{ox}}\right]^2\right]^{-1} \tag{3.120}$$

Carrier Transport.

In previous equations the mobility of the electrons in the channel, μ, has been used. This mobility, the channel mobility, is generally less than the bulk mobility for silicon. Inversion layer mobility studies [22] suggest that surface-related scattering centres reduce the mobility. The nature of the $Si - SiO_2$ interface is such that there is not an abrupt transition but rather a transition zone of 20Å to 30Å thick. The channel mobility is reduced by scattering off this zone as well as off impurity atoms, and crystal defects. In general as the gate electric field increases the mobility decreases, which is probably due to the carrier density increasing at the surface and scattering events becoming more likely.[23-25]

Theoretical expressions for the channel mobility are difficult to develop and implement due to a lack of knowledge about the interface. CAD device models which take mobility variations into account use look up tables and empirical analytic expressions to model the effect of transverse electric field, temperature, and doping densities.

3.5.3. Implantation Effects.

Ion implantation can be used to adjust the doping concentrations of the channel.[26] As mentioned in chapter 2 the distribution of ions in the crystal will have a Gaussian like profile after annealing. This negates the constant doping concentration assumed in the earlier developments. The correct solution of the previous equations

requires integration of the actual doping profile. Although this approach is available in numerical simulators a simplifying approximation is required for analytic equations.[1]

A common simplification which is made is to lump the Gaussian distribution into a "box" distribution so that a constant density N_{AI} can be assumed to some depth x_I. The total implanted dose is then $D = N_{AI} \times x_I$. This simplification can be used to adjust other parameters.

Threshold Voltage.

The calculation of the effect of ion implantation on the threshold voltage is very simple if the depth of implantation is deeper than the depletion depth at inversion. The doping can be considered as $N_A = N_A + N_{AI}$ and the change in threshold voltage that results is

$$\Delta V_T = \frac{qD}{C_{ox}} \tag{3.121}$$

Subthreshold Slope.

Like the threshold voltage calculation, two cases exist for the subthreshold swing calculation. If the implant is deep enough and uniform enough then the effective doping concentration can be used in the swing calculation.

Otherwise a far more complicated equation is required to model the swing.[19] The swing is strongly dependent on the depletion width, and the depletion width is in turn effected by the dose and depth of the implant. As the implant depth is increased the depletion layer width is decreased, so the swing is increased and closely tied to the implant depth. Eventually as the implant depth in increased further the depletion capacitance starts to be reduced and as a result the swing value is also reduced. Further detail about implantation effects are given in reference (19).

3.5.4. Small Geometry Effects.

As IGFET's have become smaller departures from long channel approximations have become important in the modelling of the devices.[27-48]

The subthreshold current, the threshold voltage, and the saturation current and voltages have all been effected. There are also new effects, brought on by higher electric fields due to the failure to scale supply voltages. These new hot carrier effects have presented difficult problems.

Threshold Voltage.

As the channel length of an IGFET decreases the threshold voltage of the transistor has also been seen to decrease. This effect, called the short channel effect, has been extensively studied. A transistor is considered short when the channel length is comparable to the source and drain depletion depths. A review of the models,[28] shows a multitude of solutions to the problem of modelling short channel transistors. The source of the problem stems from the depletion charge in the channel being not only under gate voltage control, but also that of the source and drain.

Most solutions are based on a "charge sharing solution" which divides the charge in the channel into sections allowing the gate control of only a reduced portion. The most common allows the gate to control a trapezoid of charge defined by L at the surface and L' at the bottom of the depletion, where L' is the channel length minus the source and drain depletion widths. The net effect on the threshold voltage formula is to reduce the depletion charge seen by the gate, thereby reducing the voltage required to support it.

Another threshold voltage effect caused by down scaling is the narrow width effect. As the width of LOCOS isolated transistors decreases the transition region becomes important in the device models. A device is considered narrow if the width of the channel is of the same order of magnitude as the depletion depth. The narrow width effect is characterised by an increase in threshold voltage as the width decreases†. It is due to the field inversion preventing implant diffusing laterally into the edges of the channel. The charge stored under the thick edges of the gate oxide increases the gate voltage required to invert the rest of the channel. A number of modelling[28,32,33] and processing[31] solutions exist.

† For standard LOCOS isolation processes.

Subthreshold Conduction.

As the source and drain diffusion spacing is reduced, effects caused by their interaction become important.[22,34-36] Drain induced barrier lowering is one of these. In the subthreshold region of operation the channel acts as a potential barrier resisting the diffusion currents. If the drain depletion is sufficiently close to the source depletion the potential barrier is reduced and more subthreshold current will flow. The main effect of barrier lowering is seen as subthreshold current dependence on the drain voltage beyond the normal thermal exponent contribution. A few models of drain induced barrier lowering exist.[22,35]

2-D Effects above Threshold.

Small geometry effects occur above threshold also.[22,27,34] There are the obvious effects on saturation currents and voltages because of the change in threshold voltage, but there are other effects as well. Channel length modulation, field dependent mobility, and velocity saturation are important effects.

Channel length modulation occurs as a result of effective channel shortening as the drain depletion region is biased into the channel. Although some models have attempted to take this factor into account by simple adjustments to the channel length used in the previous models, other factors have made those models inaccurate. When channel length modulation becomes important, so do two-dimensional electric field components in the Poisson equation. Channel length modulation is thought to account for some dependence of drain saturation current on drain voltage.

The increased tangential electric field magnitude in the channel, due to shorted channel lengths, effects the carrier mobility.[2,22-25,36] As the fields become higher, the mobility of the electrons decrease, in that the drift velocity fails to increase in proportion to the increase in the electric field. There are a number of models which attempt to explain this effect.[23-25]

Eventually the carrier velocity fails to increase at all with increasing electric field, which is called velocity saturation.[1] The phenomenon can be thought of as the collision frequency increasing to the point where there is insufficient time between collisions for a carrier to gain more momentum than that which would be lost in the next collision.

The effect on IGFET characteristics is a lower saturation voltage than predicted by simple theory, and much lower saturation currents, which are almost drain voltage independent. The maximum electron and hole velocities in silicon are $10^7 cm\ s^{-1}$, and $8\times10^6 cm\ s^{-1}$ respectively, and occur for field magnitudes of approximately $7\times10^3 V\ cm^{-1}$.

3.5.5. Unusual Operating Modes.

Two unusual operating models which stem from increased internal electric field in small geometry IGFETs are punch through and hot carrier effects.

Punch Through

If a transistor's channel is sufficiently short, or if the drain bias is sufficiently large, it is possible for the drain depletion region to extend all the way to the source depletion region, thereby allowing current to flow regardless of the gate potential. The onset of this effect can be easily predicted since it is when the sum of the source and drain depletion widths equals their separation. A second deep channel implant, which reduces the drain depletion motions, is the processing solution to control this effect. [1,36] Avalanche breakdown of the channel is also possible if the field is high enough, but that usually occurs only in transistors with very high doping concentrations.

Hot Carriers

In small geometry transistors the electric fields can be large enough to accelerate charge carriers to energies of a few electron volts. The equivalent thermal temperature of such electrons is tens of thousands of degrees K. Hence the name hot carriers. Problems in IGFETs start to occur when the carriers reach energies sufficiently high to allow them to cross the silicon - silicon dioxide interface. That energy is around $3.1eV$ for electrons.

If the carriers reach sufficient energy to be injected into the oxide they may become trapped and influence the characteristics of the transistor. This is probably the most important aspect of hot electrons and certainly the most studied.[37-48]

Other effects such as increased gate current, substrate currents, and biasing of parasitic transistors are also a problem. What makes oxide injection and capture the main problem is the way that it can slowly alter the transistor characteristics over a long period of time.

There are three mechanisms for hot electron generation in IGFETs. They are substrate hot electrons (SHE), channel hot electrons (CHE), and avalanche hot electrons (AHE).

Substrate hot electrons arise from a strong gate electric field. As shown in figure 3.11, thermal electron-hole pairs are split by the field and the electrons are accelerated towards the silicon-silicon dioxide interface. If they gain enough energy they will penetrate the oxide and possibly be trapped. The probability of emission is the same for the entire channel, but fairly low compared to the other mechanisms.

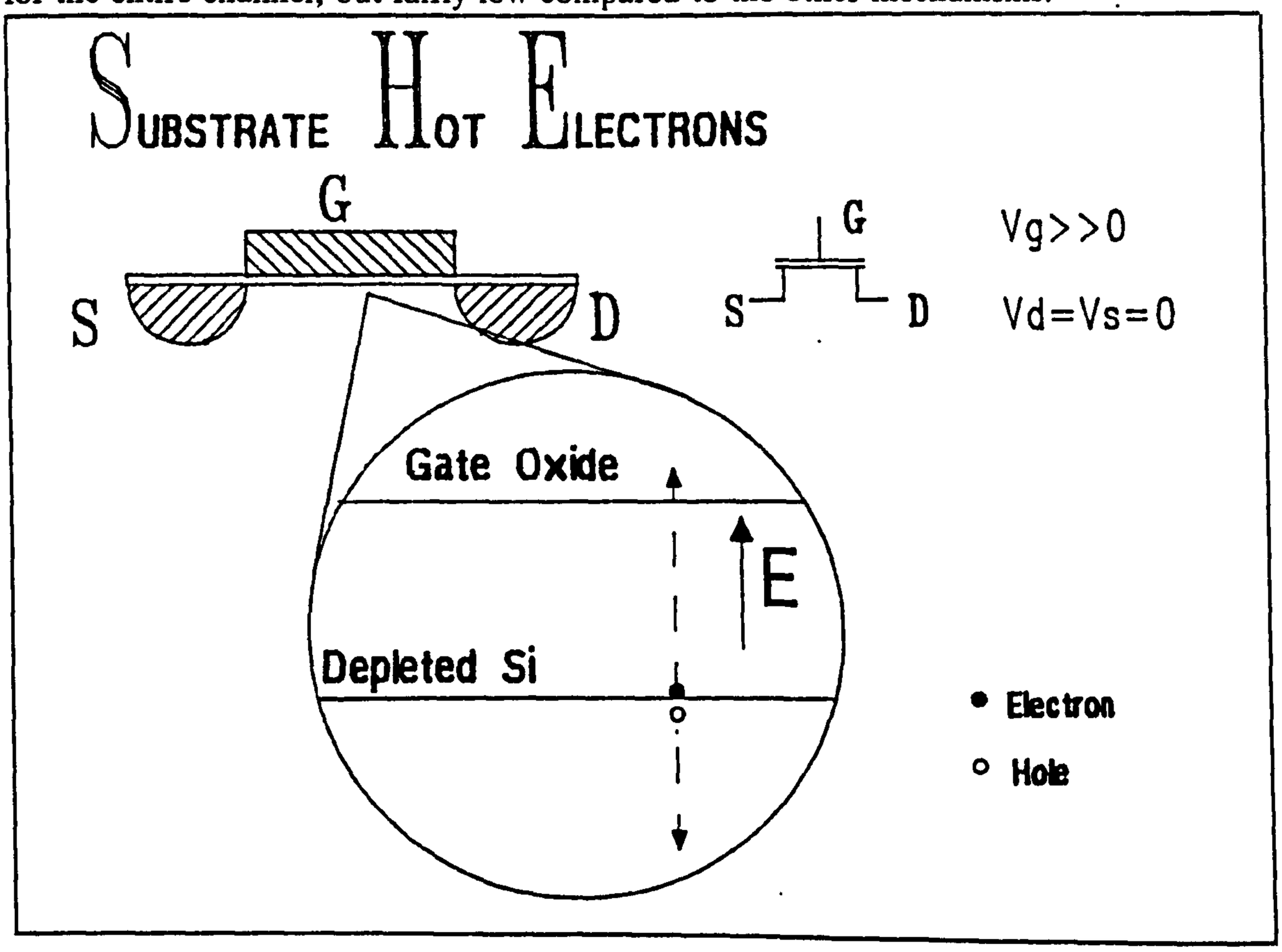


Figure 3.11, Substrate Hot Electron Mechanism.

Channel hot electrons occur under different bias conditions than SHE. Figure 3.12 shows that the channel must be in strong inversion with a high tangential field. A

"lucky electron"[45] gains energy from the tangential field before having an energy conserving but direction changing collision with a lattice atom. The electron is deflected at the interface, and, if it has sufficient energy, into the oxide. CHE's have an increased probability of occurring towards the drain end of the channel. They are more probable than SHE and less than AHE in current small geometry transistors.

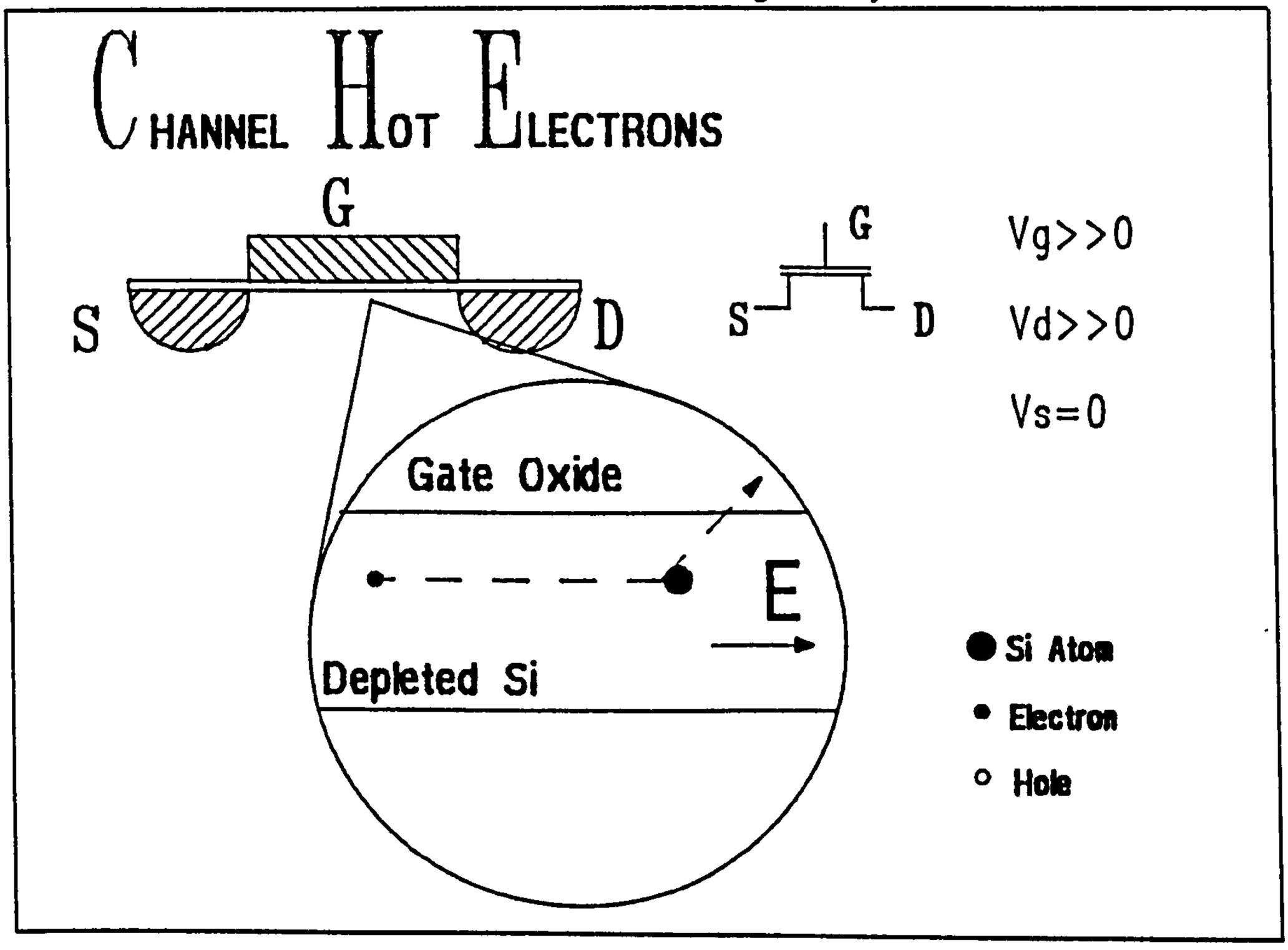


Figure 3.12, Channel Hot Electron Mechanism.

Avalanche hot electrons occur at the drain end of the channel in the drain depletion region, as shown in figure 3.13. The bias conditions best suited to AHE are a high drain bias and a gate bias just above the threshold voltage. The mechanism involves channel charge carriers being injected into the high field region in the drain depletion region. There they gain sufficient energy to create an electron hole pair on collision with a lattice atom. After the collision the electron may still have enough energy to be injected into the oxide. The electron generated in the collision may be accelerated by the field and cause avalanche multiplication, the hole may gain sufficient energy to also be injected into the oxide.

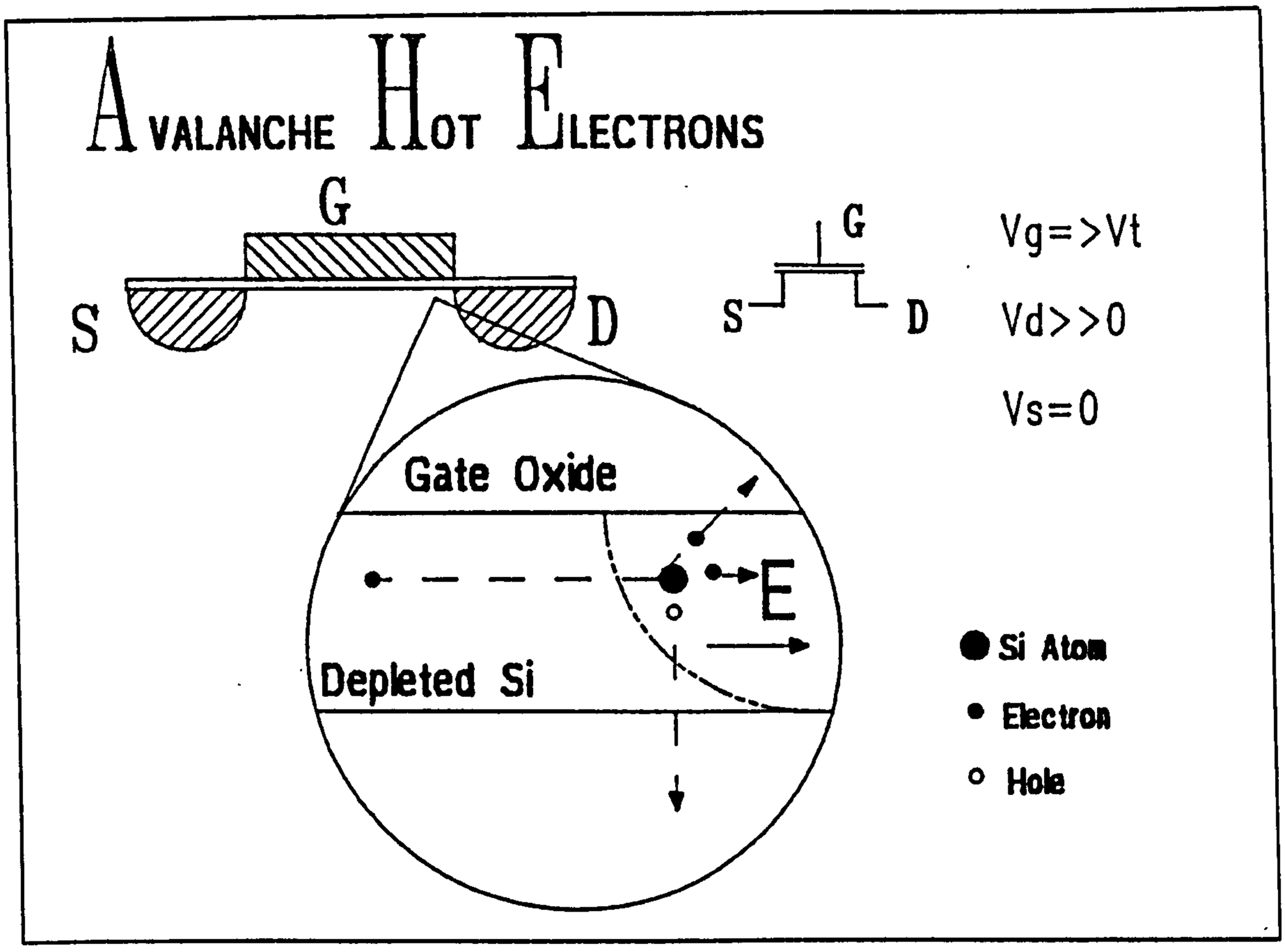


Figure 3.13, Avalanche Hot Electron Mechanism.

What actually happens in the oxide once the electrons and holes are injected has been the subject of many theories.[39,44,46,47] The current theories[38,40] suggest a number of mechanisms are responsible for characteristics degradation, depending on the bias conditions. Degradation of device parameters can occur because of trapped charge in the oxide and though the creation of interface traps. It is thought that hot holes form neutral interface trap centres. Then hot electrons injected under different bias conditions become trapped at these centres. This explains earlier observations that bias cycling increased degradation. The current thought is that degradation in transconductance and subthreshold slope can be caused just by oxide trapped hot electrons, but not just by generated interface traps. Enhanced degradation occurs when electrons are injected into a region previously subjected to hot holes which have generated interface trap centres.

A simple circuit model of an IGFET after hot carrier damage is shown in figure 3.14. It is a model using a normal transistor and a damaged transistor in series with a

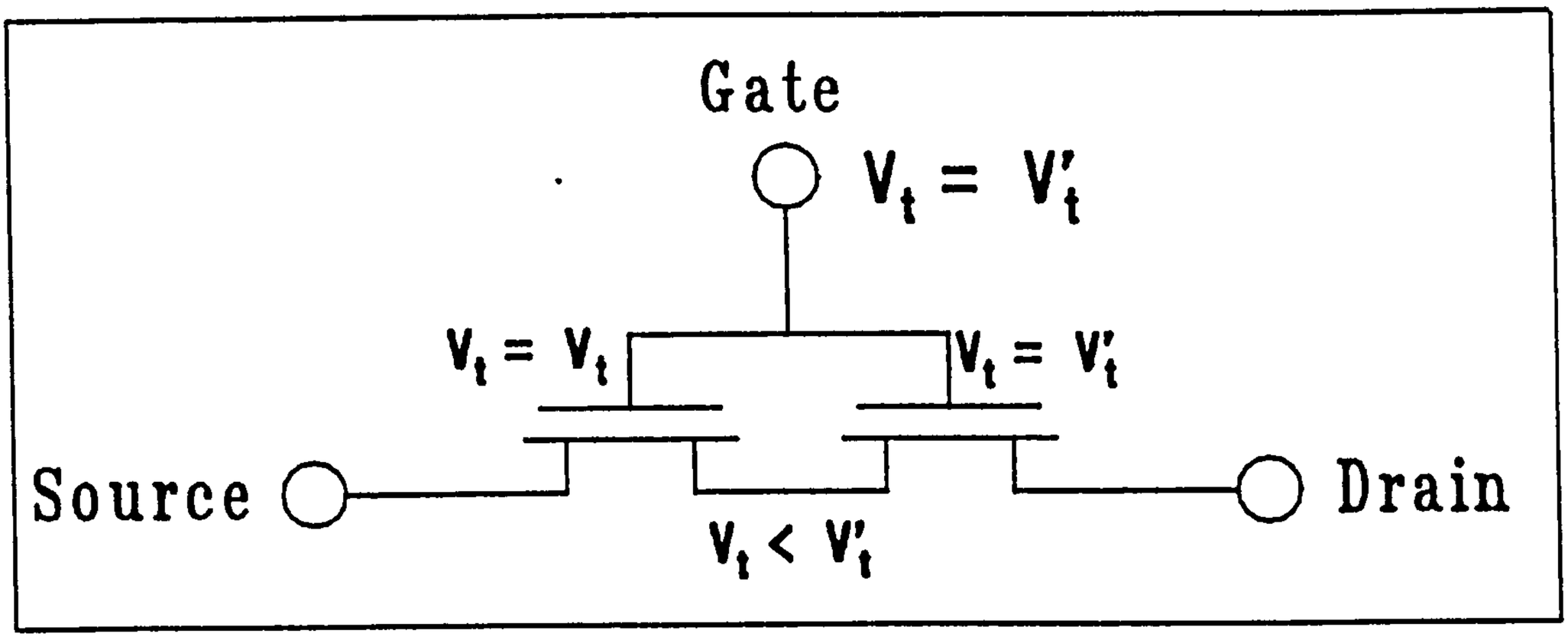


Figure 3.14, An equivalent model for a Hot Electron Damaged IGFET.

common gate bias. The damaged transistor has an increased threshold due to trapped charge. Although it is a simplistic model, it has some physical merits in that AHE damage occurs only at the drain end of the channel.

3.6. Models.

Because of the long times required to fabricate an integrated circuit (6 to 22 weeks), and the high degree of interaction between a complex array of design variables, it is useful and almost essential to perform computer simulations to predict their device and circuit performance before they are fabricated. Simulation is also a useful way to investigate internal device phenomena.

3.6.1. Circuit Models.

Since there are many transistors in a large intergrated circuit, the CAD model of an IGFET for circuit simulators must be simple enough to allow fast computation, but complex enough that the essential characteristics are modelled.

There are a number of such models.[12-14] One of the most widely used models is called SPICE.[14,18,49]

SPICE is based on semi-physical equations, similar to what was described previously, and uses empirically determined parameters. SPICE, level 1, uses the following parameters, the transistors drawn length and width, parameters describing the reduction of those lengths due to processing, the oxide thickness, the threshold voltage and a

parameter relating it to substrate bias, and the mobility and a parameter relating it to gate voltage, in order to predict the drain current for a given set of terminal biases. It is successful for transistors larger than about 3 microns. SPICE, level 3, uses additional empirical parameters to adjust the threshold voltage for small geometry effects, to model velocity saturation, adjust the subthreshold slope, and model change length modulation,[49] in smaller geometry devices.

3.6.2. Device Models.

CAD models for a single transistor tend to be much more complex than for circuits because the designer wants to know more and relatively greater amounts of computer time are feasible for one transistor than for many.

The main use of device simulators is not just to predict the behaviour of the terminal characteristics of the transistor but to show the internal operation (carrier densities, electric potentials and fields) for various operating conditions.

To achieve this aim, most device simulators use a grid of points representing positions in the bulk and the oxide of a transistor in two or three dimensions. Each point can then be assigned material characteristics, doping characteristics, electric field and potential magnitudes and charge carrier densities, as numerical solutions to the differential equations relating these parameters are found. When convergent solutions for each external bias condition are found the values at each grid point for carrier density and velocity, electric potential and field magnitudes as well as the terminal characteristics of the transistor can be plotted in some meaningful manner. The nature of the display depends on what the designer wishes to know.

Some popular CAD software packages for this purpose are MINIMOS,[50] PISCES,[51] and CANDE.[52]

MINIMOS

MINIMOS was probably one of the first successful MOSFET CAD device models. It is restricted to the standard MOSFET structure, but choices of three gate materials, oxide thickness, and vertical doping profiles, are allowed by the user. It uses a rectangular two-dimensional grid and solves the semiconductor equations using

finite-difference solutions between grid nodes. It uses analytic models for mobility and can model carrier generation and recombination for most mechanisms including impact ionisation.

MINIMOS solves on the basic Poisson equation and models both electron and hole currents considering both diffusion and drift mechanisms. It depends on Boltzmann statistics for carrier density, and uses the Einstein relationship between diffusion and mobility constants.

PISCES

PISCES is a general purpose semiconductor device modelling program. It can produce solutions for any device structure and it models both types of carriers concurrently making it useful for bipolar as well as MOS transistors. This is at the expense of large amounts of computing time to reach a solution.

PISCES uses a rectangular mesh grid based on a bisected rectangular starting grid. The solution method is continuous Gummel (three different equations solved sequentially with only one variable varying at one time) and Newton's method (where all variables vary concurrently). The choice of solutions depends on the type of device modelled, and must be specified by the designer. A number of substrate and surface layer materials are available with the geometry freely described by the user, or entered from a process simulator output.

The basic equations solved are the semiconductor Poisson equation, along with carrier continuity and equations to handle a number of generation and recombination mechanisms. The carrier currents again are modelled for both drift and diffusion mechanisms, but Fermi-Dirac statistics are used to model carrier concentration and the diffusion to mobility constant relationship. Multiple mobility models are available, along with boundary conditions for a variety of semiconductor surface contacts, and models for incomplete donor and acceptor ionisation.

In short, PISCES embodies into its models most of the semiconductor physics known today. Unfortunately this requires a lot of effort by the user to describe a device and requires huge amounts of computer time to find the solution.

CANDE

The capabilities of CANDE exists midway between those of MINIMOS and that of PISCES. It allows more flexibility of transistor geometry than MINIMOS and only uses the relevant models from PISCES in order to increase computation speed.

CANDE uses a rectangular grid and accepts two dimensional doping profiles. Poisson's equation is solved using a finite difference approximation over the grid node points. It uses similar statistics and continuity equations to PISCES but uses an analytic mobility model, with a look-up table for doping, that includes both tangential and perpendicular electric field effects. Work function differences between materials are also modelled. To speed solutions, only one carrier type is modelled for any structure, depending on the channel impurity type. That simplification greatly increases computer speed, and only slightly limits the range of problems that CANDE can solve.

3.7. Chapter Summary.

In this chapter the basic physics required to model an IGFET were discussed. Included was, the background physics for describing electron densities in crystals, the effect of introducing dopants, and charge mobility mechanisms. The simple pn junction and MOS capacitor were also presented. The subthreshold, linear and saturation regions of IGFET operation were described along with the effect of ion implantation and small geometry effects. Finally a brief description of CAD tools for modelling IGFET behaviour was given.

3.8. References.

1. Muller, R.S. and Kamins, T.I., *Device Electronics for Integrated Circuits*, John Wiley & Sons, Inc., New York, 1986. Ed. 2

2. Sze, S.M, *Physics of Semiconductor Devices*, John Wiley & Sons, Inc., New York, 1981. Edition 2

3. Kittel, C., *Introduction to Solid State Physics*, John Wiley & Sons, Inc., New York, 1976. Ed. 5.

4. Sze, S.M, *Semiconductor Devices Physics and Technology*, John Wiley & Sons, Inc., New York, 1985.

5. Warner R.M., *Integrated Circuits, Design Principles and Fabrication.*, McGraw Hill Book Company, New York, 1965.

6. Grove, A.S. and Fitzgerald, D.J., "Surface Effects on PN junctions: Characteristics of surface space-charge regions under non-equilibrium conditions.," *Solid State Electronics*, vol. 9, pp. 783-806, Pergamon Press., England, 1966.

7. Colclaser, R.A., *Microelectronics, processing and device design.*, John Wiley & Sons, Inc., New York, 1980.

8. Robertson, J.M., "CV," in *SERC School on Microfabrication*, ed. Walton, A.J., University of Edinburgh, Edinburgh, June 1987.

9. Mavor, J., "FET Devices and Technology," in *Large Scale Integration. Devices, Circuits, Systems.*, ed. Howes, M.J., John Wiley & Sons, Inc., New York, 1981.

10. Glaser, A.B. and Subak-Sharpe, G.E., *Integrated Circuit Engineering. Design, Fabrication, and Applications.*, Addison-Wesley Pub. Co., London, 1977.

11. Sah, C.T., "Characteristics of the Metal Oxide Semiconductor Transistors.," *IEEE Transactions on Electron Devices*, pp. 324-345, IEEE, New York, July 1964.

12. Rogers, D.M., Hayden, J.D., and Rinerson D.D., "Model for the Channel implanted Enhancement Mode IGFET," *IEEE Transactions on Electron Devices*, vol. ED-33, no. 7, pp. 955-964, IEEE, New York, July 1986.

13. Wright, G.T., "A Simple and Continuous MOSFET Model.," *IEEE Transactions on Electron Devices*, vol. ED-32, no. 7, pp. 1259-1263, IEEE, New York, July 1985.

14. Yang, P. and Chatterjee, P.K, "SPICE Modelling for Small Geometry MOSFET Circuits," *IEEE Transactions on Computer Aided Design of Integrated Circuits and Systems.*, vol. CAD-1, no. 4, pp. 169-182, IEEE, New York, Oct. 1982.

15. Brews, J.R., "A Charge-sheet Model of the MOSFET," *Solid State Electronics*, vol. 21, pp. 345-355, Pergamon Press., England, 1978.

16. Fichtner, W. and Potzl, H., "MOS Modelling by Analytical Approximations. I. Subthreshold Current and Threshold Voltage.," *International Journal of Electronics*, vol. 46, no. 1, pp. 33-55, 1979.

17. Lin, P.S. and Wu C.Y., "A New Approach to Analytically Solving the Two Dimensional Poisson's Equation and its application in Short-Channel MOSFET Modelling.," *IEEE Transactions on Electron Devices*, vol. ED-34, no. 9, pp. 1947-1956, IEEE, New York, Sept 1987.

18. Robertson, J.M., "Device Operation," in *SERC School on Microfabrication*, ed. Walton, A.J., University of Edinburgh, Edinburgh, June 1987.

19. Brews, J.R., "Subthreshold Behavior of Uniformly and Nonuniformly Doped Long Channel MOSFET.," *IEEE Transactions on Electron Devices*, vol. ED-26, no. 9, IEEE, New York, 1979.

20. Barron, M.B., "Low Level Currents in Insulated Gate Field Effect Transistors," *Solid State Electronics*, vol. 15, pp. 293-302, Pergamon Press., England, 1972.

21. Gosney, W.M., "Subthreshold Drain leakage Currents in MOS Field Effect Transistors," *IEEE Transactions on Electron Devices*, vol. ED-19, no. 2, IEEE, New York, Feb, 1972.

22. Ferry, D.K., "Physics and Modelling of Submicron Insulated Gate Field Effect Transistors. I.," in *VLSI Electronics Microstructure Science*, ed. Einspruch, N.G., vol. 6, Academic Press, Inc., New York, 1981.

23. Ohno, Y. and Okuto, Y., "Electron Mobility in n-Channel Depletion Type MOS Transistors.," *IEEE Transactions on Electron Devices*, vol. ED-29, no. 2, pp. 190-194, IEEE, New York, Feb. 1982.

24. Krutsick T.J., White, M.H., Wang H.S., and Booth, R.V., "An Improved Method of MOSFET Modeling and Parameter Extraction.," *IEEE Transactions on Electron Devices*, vol. ED-34, no. 8, pp. 1676-1679, IEEE, New York, Aug. 1987.

25. Majkusiak, B. and Jakubowski, A., "The Dependence of MOSFET Surface Carrier Mobility on Gate-Oxide Thickness.," *IEEE Transactions on Electron Devices*, vol. ED-33, no. 11, pp. 1717-1721, IEEE, New York, Nov 1986.

26. Risch, L., Werner, C., Muller, W., and Weider, A., "Deep Implant 1 μm MOSFET structure with Improved Threshold Control for VLSI Circuitry.," *IEEE Transactions on Electron Devices*, vol. ED-29, no. 4, pp. 601-606, IEEE, New

York, April 1982.

27. Wang. P.P., "Device Characteristics of Short-Channel and Narrow-Width MOSFET's.," *IEEE Transactions on Electron Devices*, vol. ED-25, no. 7, pp. 779-786, IEEE, New York, July 1978.

28. Akers, L.A. and Sanchez, J.J., "Threshold Voltage Models of Short, Narrow, and Small Geometry MOSFET's: A review.," *Solid State Electronics*, vol. 25, no. 7, pp. 621-644, Pergamon Press., England, July 1982.

29. Wang, C.T., "A Threshold Voltage Expression for Small-Size MOSFET's Based on an Approximate Three Dimensional Analysis.," *IEEE Transactions on Electron Devices*, vol. ED-33, no. 1, pp. 160-164, IEEE, New York, Jan. 1986.

30. Viswanathan, C.R., Burkey, B.C., Lubberts, G., and Tredwell, T.J., "Threshold Voltage in Short-Channel MOS Devices," *IEEE Transactions on Electron Devices*, vol. ED-32, no. 5, pp. 932-940, IEEE, New York, May 1985.

31. Lai, P.T. and Cheng, Y.C., "Comparison of Threshold Modulation in Narrow MOSFET's with Different Isolation Structures.," *Solid State Electronics*, vol. 28, no. 6, pp. 551-554, Pergamon Press., England, June 1985.

32. Kasai, R., Yokoyama, K., Yoshii, A., and Sudo, T., "Threshold-Voltage Analysis of Short and Narrow Channel MOSFET's by Three Dimensional Computer Simulation.," *IEEE Transactions on Electron Devices*, vol. ED-29, no. 5, pp. 870-876, IEEE, New York, May 1982.

33. Kroell, K.E. and Ackermann, G.K., "Threshold Voltage of Narrow Channel Field Effect Transistors.," *Solid State Electronics*, vol. 19, pp. 77-81, Pergamon Press., England, 1976.

34. Sodini, C.G, Ko, P.K., and Moll, J.L., "The Effect of High Fields on MOS Device and Circuit Performance," *IEEE Transactions on Electron Devices*, vol. ED-31, no. 10, pp. 1386-1393, IEEE, New York, Oct. 1984.

35. Chamberlain, S.G. and Ramanan, S., "Drain Induced Barrier Lowering Analysis in VLSI MOSFET Devices Using Two-Dimensional Numerical Simulations.," *IEEE Transactions on Electron Devices*, vol. ED-33, no. 11, pp. 1745-1753, IEEE, New York, Nov. 1986.

36. Eitan, B. and Frohman-Bentchkowsky, D., "Surface Conduction in Short-Channel MOS Devices as a Limit to VLSI Scaling.," *IEEE Transactions on Electron Devices*, vol. ED-29, no. 2, pp. 254-266, IEEE, New York, Feb. 1982.

37. Shiono, N, and Hashimoto, C., "Threshold-Voltage Instability of n-Channel MOSFET's under Bias-Temperature Aging.," *IEEE Transactions on Electron Devices*, vol. ED-29, no. 3, pp. 361-368, IEEE, New York, Mar. 1982.

38. Haddara, H. and Cristoloveanu, S., "Two-Dimensional Modeling of Locally Damaged Short-Channel MOSFET's Operating in the linear region.," *IEEE Transactions on Electron Devices*, vol. ED-34, no. 2, pp. 378-385, IEEE, New York, Feb. 1987.

39. Tsuchiya, T., Koboyashi, T., and Nakajima, S., "Hot-Carrier-Injected Oxide Region and Hot-Electron Trapping as the Main Cause in Si nMOSFET Degradation.," *IEEE Transactions on Electron Devices*, vol. ED-34, no. 2, pp. 386-391, IEEE, New York, Feb 1987.

40. Tsuchiya, T., "Trapped Electron and Generated Interface Trap Effects in Hot-Electron-Induced MOSFET Degradation.," *IEEE Transactions on Electron Devices*, vol. ED-34, no. 11, pp. 2291-2292, IEEE, New York, Nov. 1987.

41. Ning, T.H., "Hot Electron Emission from Silicon into Silicon Dioxide.," *Solid State Electronics*, vol. 21, pp. 273-282, Pergamon Press., England, 1978.

42. Takeda, E., Kume, H., Toyabe, T., and Asai, S., "Submicrometer MOSFET Structure for Mimimizing Hot Carrier Generation.," *IEEE Transactions on Electron Devices*, vol. ED-29, no. 4, pp. 611-618, IEEE, New York, Apr. 1982.

43. Tam, S. and Hu, C., "Hot Electron-Induced Photon and Photocarrier Generation in Silicon MOSFET's.," *IEEE Transactions on Electron Devices*, vol. ED-31, no. 9, pp. 1264-1273, IEEE, New York, Sept. 1984.

44. Cottrell, P.E., Troutman, R.R., and Ning, T.H., "Hot Electron Emission in N-Channel IGFET's.," *IEEE Transactions on Electron Devices*, vol. ED-26, no. 4, pp. 520-533, IEEE, New York, Apr. 1979.

45. Tam, S., Ko, P.K., and Hu, G., "Lucky-Electron Model of Channel Hot-Electron Injection in MOSFET's.," *IEEE Transactions on Electron Devices*, vol.

ED-31, no. 9, pp. 1116-1125, IEEE, New York, Sept. 1984.

46. Schmitt-Landsiedel, D. and Dorda, G., "Novel Hot-Electron Effects in the Channel of MOSFET's Observed by Capacitance Measurements.," *IEEE Transactions on Electron Devices*, vol. ED-32, no. 7, pp. 1294-1301, IEEE, New York, July 1985.

47. Radojcic, R., "Some Aspects of Hot-Electron Aging in MOSFET's.," *IEEE Transactions on Electron Devices*, vol. ED-31, no. 10, pp. 1381-1386, IEEE, New York, Oct. 1984.

48. Tanaka, S., Saito, S., Atsumi, S., and Yoshikawa, K., "A Self-Consistent Pseudo-Two-Dimensional Model for Hot-Electron Current in MOST's.," *IEEE Transactions on Electron Devices*, vol. ED-33, no. 6, pp. 743-753, IEEE, New York, June 1986.

49. T. Gribben, "Device Modeling.," in *A Ph.D. Thesis*, University of Edinburgh, Edinburgh, July 1988.

50. Hull, T.H., *Using Minimos.*, Department of Electrical and Electronic Engineering, Queens' University Belfast., Belfast.

51. Greenfield, J.A., "PISCES-2B Models.," in *PISCES-2B Users Manual*, Technology Modeling Associates, California, 1987.

52. Greenfield, J.A., "CANDE Models and Solution Methods.," in *CANDE Users Manual*, Technology Modeling Associates, California, 1987.

Chapter 4. Progressional Offset Technique.

4.1. Introduction.

The importance of edge effects to the IGFET, and the need to study them, was outlined in chapter one. In this chapter a novel design technique will be presented which advances positional precision in experimental structures. The improved resolution, from the progressional offset technique, is applied to the study of source and drain edge effects in the following chapters.

Investigation of edge characteristics is of the most importance to minimum geometry transistors. Test structures for studying edge effects should ideally be constructed in the edge region of those transistors. However, by definition, it is impossible to fabricate any structure smaller than the "minimum geometry transistor" using the same processing techniques.

By examining the tolerances which define the minimum geometry for a certain processing technology, it is possible to avoid the need of producing structures smaller than the minimum geometry by the use of a technique for greater control in the positioning of relative edges. With that technique it is possible to build suitable test structures for studying IGFET edge effects.

4.2. Standard Fabrication Positional Tolerance.

The design of standard fabrication processes has the objective of achieving the maximum number of functional integrated circuits per wafer. That objective clearly must influence the specification of the minimum transistor size. The combined tolerances of the pattern generation resolution, mask alignment accuracy, and processing related effects must be considered when determining the minimum allowed dimensions. There are two aspects to the minimum geometry of a feature. They are, the absolute size of the structure (ie the length and width), and the relative position of that feature to other features in the device.† It is the latter which is of interest here. Figure 4.1, shows the definitions which will be used to mathematically describe the relationship

† Incidently, the complete set of minimum sizes and spaces for a given process are called the "design rules", which also include the electrical specifications of the process.

between features on one mask, and between features on two masks.‡

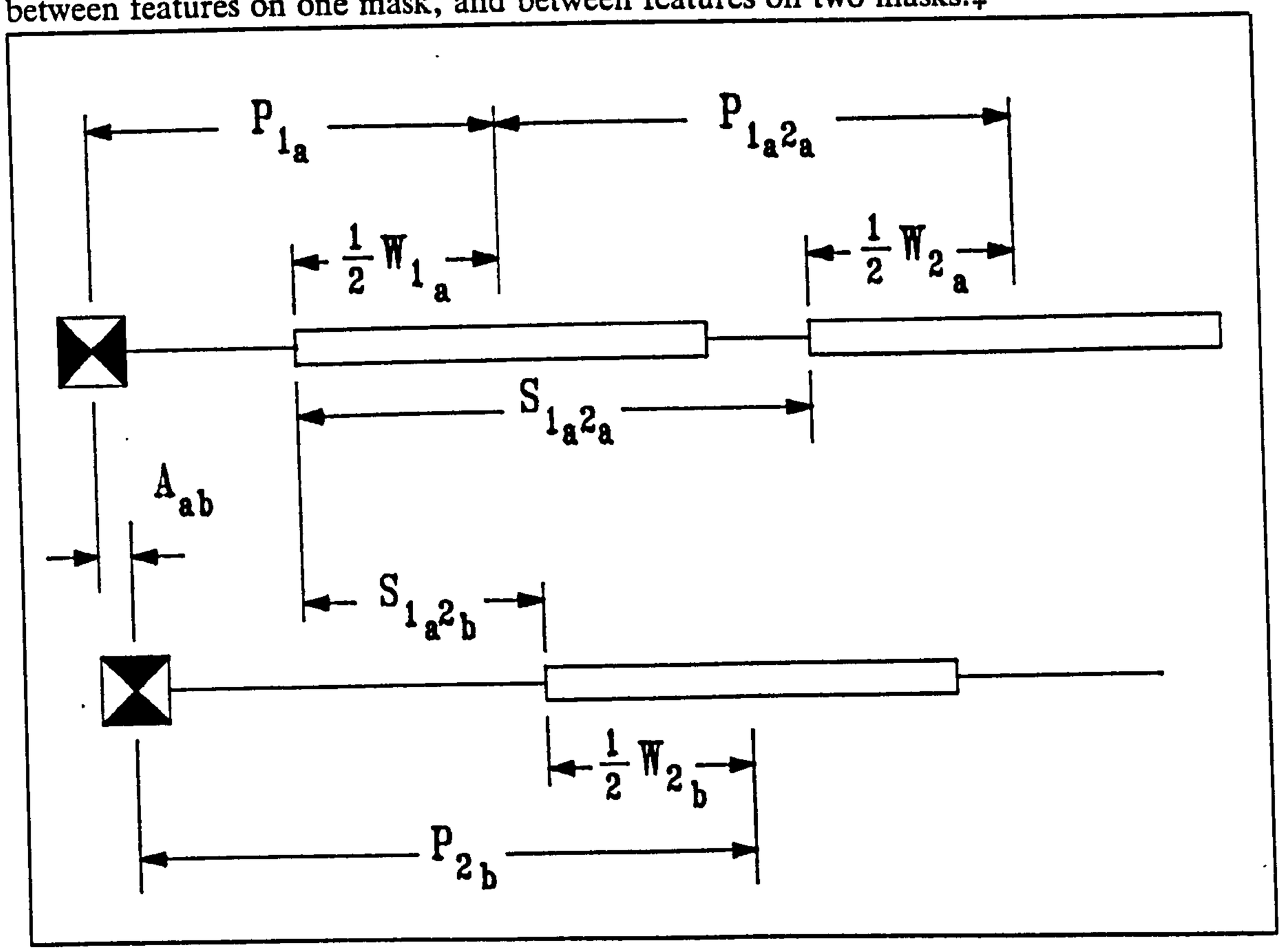


Figure 4.1, The alignment of features on different mask levels.

They are; P_{1_a}, $P_{1_a2_a}$, and $P_{1_a2_b}$, which are positions of the feature relative to another feature on the same mask, W_{1_a}, W_{2_a}, and W_{2_b}, which are the widths of a feature, and A_{ab}, which is the displacement in alignment between masks. The separations between edges can then be calculated with the following formulae.

$$S_{1_a2_a} = \frac{1}{2}W_{1_a} + P_{1_a2_a} - \frac{1}{2}W_{2_a} \tag{4.1}$$

$$S_{1_a2_b} = \frac{1}{2}W_{1_a} - P_{1_a} + A_{ab} + P_{1_a2_b} - \frac{1}{2}W_{2_b} \tag{4.2}$$

The relative error in the separations can be readily determined by differentiating with respect to x,

‡ The assumption of perfect orthogonality enables each dimension to be considered independently. In reality, the errors in orthogonality are usually small enough that they can be neglected.

$$\frac{\partial S_{1_a 2_a}}{\partial x} = \frac{1}{2}\frac{\partial W_{1_a}}{\partial x} + \frac{\partial P_{1_a 2_a}}{\partial x} - \frac{1}{2}\frac{\partial W_{2_a}}{\partial x} \tag{4.3}$$

$$\frac{\partial S_{1_a 2_b}}{\partial x} = \frac{1}{2}\frac{\partial W_{1_a}}{\partial x} - \frac{\partial P_{1_a}}{\partial x} + \frac{\partial A_{ab}}{\partial x} + \frac{\partial P_{1_a 2_b}}{\partial x} - \frac{1}{2}\frac{\partial W_{2_b}}{\partial x} \tag{4.4}$$

and substituting for the partial differential by the uncertainty (Δ) in each variable.

$$\Delta S_{1_a 2_a} = \frac{1}{2}\Delta W_{1_a} + \Delta P_{1_a 2_a} - \frac{1}{2}\Delta W_{2_a} \tag{4.5}$$

$$\Delta S_{1_a 2_b} = \frac{1}{2}\Delta W_{1_a} - \Delta P_{1_a} + \Delta A_{ab} + \Delta P_{1_a 2_b} - \frac{1}{2}\Delta W_{2_b} \tag{4.6}$$

The magnitude of the errors in position, width and alignment can be determined by considering the contribution of each step in the definition of a feature.

Pattern Generation Resolution.

The pattern definition stage of feature definition can effect both the width and positional tolerances of a feature. Whether the pattern is defined by an optical, or electron beam pattern generator the errors can be classified as; the error in defining the shape size ($\Delta W = \pm dSS$), and the error in stage positioning ($\Delta P = \pm dSM$). Typical magnitudes of these errors were given in table 2.1 on page 36

Mask Alignment Accuracy.

When the mask for one layer is aligned to the mark left after pattern definition of a previous layer, there is a possibility of misalignment. The resolution of the alignment mechanism, and alteration by previous processing to the alignment mark, contribute to the misalignment. The tolerance can be expressed as $\pm\Delta A$. The alignment accuracy for a number of photolithography exposure tools was given in table 2.2.

Processing Related Accuracy.

A number of processing steps affect the width and relative position of a feature. Probably the most obvious is that of etching. Over and under etching of a material effects the width of a feature ($\Delta W = \pm\Delta etch$). The width may also vary with depth, depending on the etching conditions. Variations of a few hundred angstroms are considered acceptable.

Oxidation will also effect the width of a feature. The effect may be to increase the feature size, as the bird's beak does to field oxide width, or decrease the feature size by consuming material, as polysilicon oxidation does to the gate width. In either case the effect is only on the feature width ($\Delta W = \pm \Delta ox$). Changes caused by oxidation can range from a few hundreds of angstroms in gate width, to nearly a micron of field oxide encroachment.

Sideways diffusion under a dopant mask, or during a post-implant anneal, can also effect the feature width. Again, depending on whether the feature was the doped area or the undoped area, the tolerance can take either sign ($\pm \Delta diff$). Some values for sideways diffusion are given in table 1.6.

Another effect of high temperature processing, such as diffusion, can be wafer warpage which results in a scaling of the wafer relative to the mask. If the ratio of the mask to wafer dimensions is $(1 - \alpha)/1$ then the relative errors introduced are $\Delta W = \pm \alpha W$ and $\Delta P = \pm \alpha P$. The effect on feature width is small, but the effect on position across a wafer can be on the order of microns.

Minimum Dimensions Using Standard Fabrication.

By considering the maximum error in relative edge positions, the minimum dimensions of a device can be determined. The two cases, relative separation between features defined by a single mask, and relative separation between features defined by two masks, will be considered separately.

First consider the case of two features on one mask. The worst case error would be when all the errors combine to cause the greatest displacement. The error in the width of the first feature would be

$$\Delta W_{1_a} = \Delta SS + \Delta etch + \Delta diff + \Delta ox \tag{4.7}$$

if the wafer warpage is small. The worst case errors in the second width would be for errors in the opposite direction. However, the second feature is processed in exactly the same way as the first, so, with the exception of the pattern generation term ΔSS, the errors must have the same sign.

$$\Delta W_{2_a} = -\Delta SS + \Delta etch + \Delta diff + \Delta ox \tag{4.8}$$

The error in the positional term $P_{1_a 2_a}$ is only from the pattern generator, or

$$\Delta P_{1_a 2_a} = \Delta SM \tag{4.9}$$

The error in the relative position of feature edges on one mask level can be determined by substitution of equations 4.7, 4.8, and 4.9 into equation 4.5. The result after rationalisation is

$$\Delta S_{1_a 2_a} = \Delta SS + \Delta SM \tag{4.10}$$

The second case, relative position between features on different masks, yields a significantly different solution. The error in the width of the first feature takes a similar form, except for the addition of superscripts to differentiate between processing on the two layers.

$$\Delta W_{1_a} = \Delta SS + \Delta etch^A + \Delta diff^A + \Delta ox^A \tag{4.11}$$

The errors in position for worst case are;

$$\Delta P_{1_a} = -\Delta SM \tag{4.12}$$

$$\Delta P_{1_a 2_b} = \Delta SM \tag{4.13}$$

$$\Delta A_{ab} = \Delta A \tag{4.14}$$

This time the error in the width of the second feature is free to take any sign since it is from independent processing actions.* The worst case is

$$\Delta W_{2_b} = -\Delta SS - \Delta etch^B - \Delta diff^B - \Delta ox^B \tag{4.15}$$

Again substitution and rationalisation gives the total error, which is

$$\Delta S_{1_a 2_b} = \Delta SS + 2\Delta SM + \Delta A_{ab} + \frac{1}{2}\left(\Delta diff^A + \Delta ox^A + \Delta etch^A + \Delta diff^B + \Delta ox^B + \Delta etch^B\right) \tag{4.16}$$

Of course all of the etching, diffusion, and oxidation terms cannot exist at once, but the error is still larger than that between features on the same mask.

---

* In a complex process there may be oxidation, and diffusion occurring for both levels at once.

The minimum allowed feature size must be large enough that the calculated variation in feature width would not make a significant difference to device performance. Uncertainty in the edge position also constricts the feature width by requiring a minimum overlap or gap.

A couple of example tolerances might help to put things in perspective. The spacing between two polysilicon bars, the result of etching after exposure using an optically generated mask, would have a tolerance in the range of 0.05 to 0.10 μm. The tolerance between a polysilicon bar and a contact hole, with R.I.E. etching after direct step to wafer exposure, would be at least 1.0 μm.

4.3. Progressional Offset Technique.

Most of the important edge regions described in chapter one exist between features on two levels which are patterned using different techniques for each level. The gate, which can be taken as the reference level, is patterned using R.I.E. etching. The width of the channel is defined predominately by oxidation after a mask is defined by R.I.E. etching. The length of the channel is defined by implantation using the gate as a mask, but then altered by sideways diffusion. If the "edges" are the transition zones defined by the tolerance in length and width, then clearly the two level tolerances of current production technology prevent the construction of test structures small enough to fit in those regions. Another approach must be used.

Yield Objectives.

The production yield objective is usually also appropriate to the construction of test structures, but, it is not essential. The objective of a test structure to study edge effects in IGFET's is to gain knowledge about the transistor operation at the edge regions. It is sufficient to have a few structures to test. As long as a "correct" structure can be guaranteed to occur, and can be identified once it does, it is not necessary to have millions of identical structures on a wafer. This is a key concept in the progressional offset technique.

Progressional Offset Scheme.

It was previously shown that the control over relative edge positions is around an order of magnitude better between features on the same mask than between features on different masks. That property can be employed in a progressional offset scheme to allow manipulation of edge relationships with a fine degree of control. Figure 4.2, illustrates the progressional offset technique.

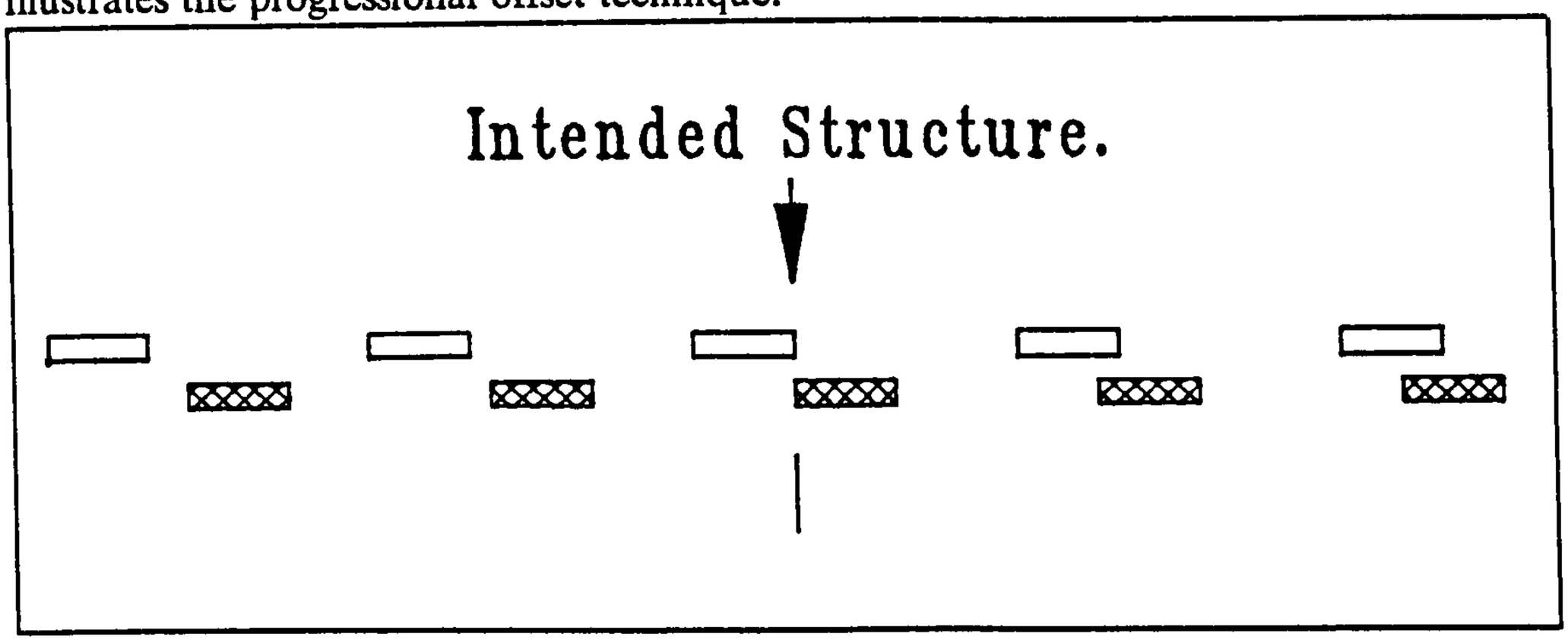


Figure 4.2, The progressional offset technique.

As it can be seen, tolerances are built into the mask set so that any variations in processing can be compensated for by one of the structures in the scheme. The offsets in the structures are implemented in a progressive manner including displacements to each side of the intended structure. In that way, variations in processing only alter the position of the intended structure in the progression. Figure 4.3, shows how this scheme could compensate for processing effects. The progressional offset technique allows the construction of precision structures by coping with processing uncertainties.

Resulting Minimum Resolution.

The resulting minimum resolution of the progressional offset technique is just the uncertainity in position of features on one mask, which is given by equation 4.10. Thus, at least an order of magnitude improvement in control of edge relationships is available for design of test structures. The resolution is defined by the mask making mechanism, and a structure "as drawn" on the mask can be created on a wafer. The cost of the improvement is the loss of knowledge of exactly where that structure is. If

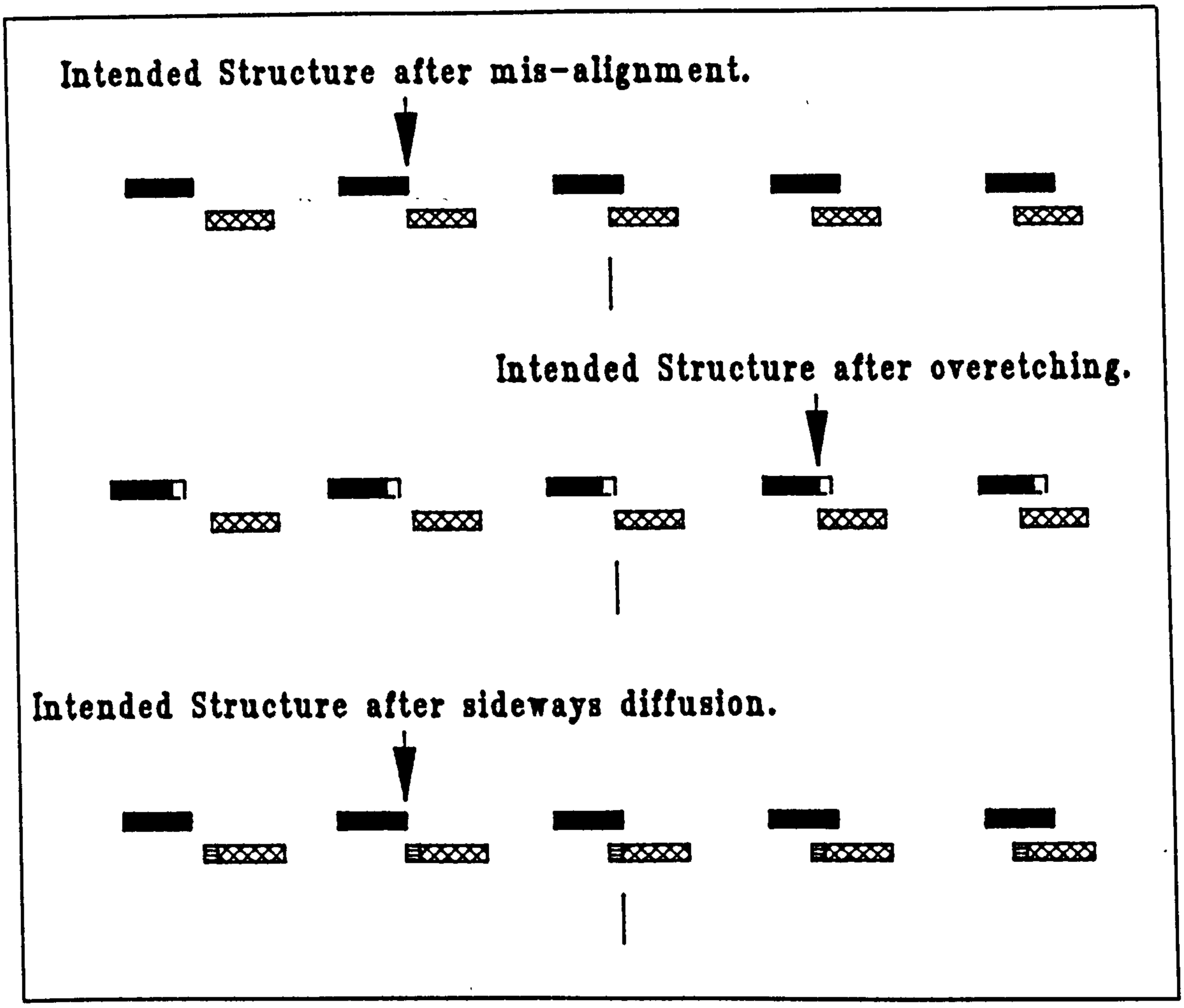


Figure 4.3, Processing variation compensation through the progressional offset technique.

the maximum designed offset is sufficient to accommodate the processing variations, the intended structure is guaranteed to exist, somewhere in the range of the progressional offsets.

4.4. Progressional Offset Alignment Technique.

It is clear that, to be able to perform electrical tests on the precision test structure it must somehow be located. It is obvious that destructive physical methods of location or alignment would not be appropriate. Therefore some simple form of electrical location or alignment is necessary for locating the intended structure.

Alignment Technique Requirements.

The main requirement of an electrical alignment technique is that the measured value should be strongly related to the separation or overlap of the sub-structures to be aligned. It may be current flow between contacting layers, the magnitude of the field effect on a channel, or the capacitance between structures. A secondary requirement, which is not essential but quite useful, is that the technique should be able to be calibrated. It could then be used to measure the magnitude of the misalignment, not only on test structures but also in the analysis of badly fabricated production samples.

Confirmation using a cross-over pattern.

A useful technique in confirming the position of the perfectly aligned or intended structure, is that of designing a "cross-over" pattern into the progressional offset scheme. It can help pin-point the "center" of the pattern.[1] Figure 4.4, shows the example of aligning a conductor to a contact hole.

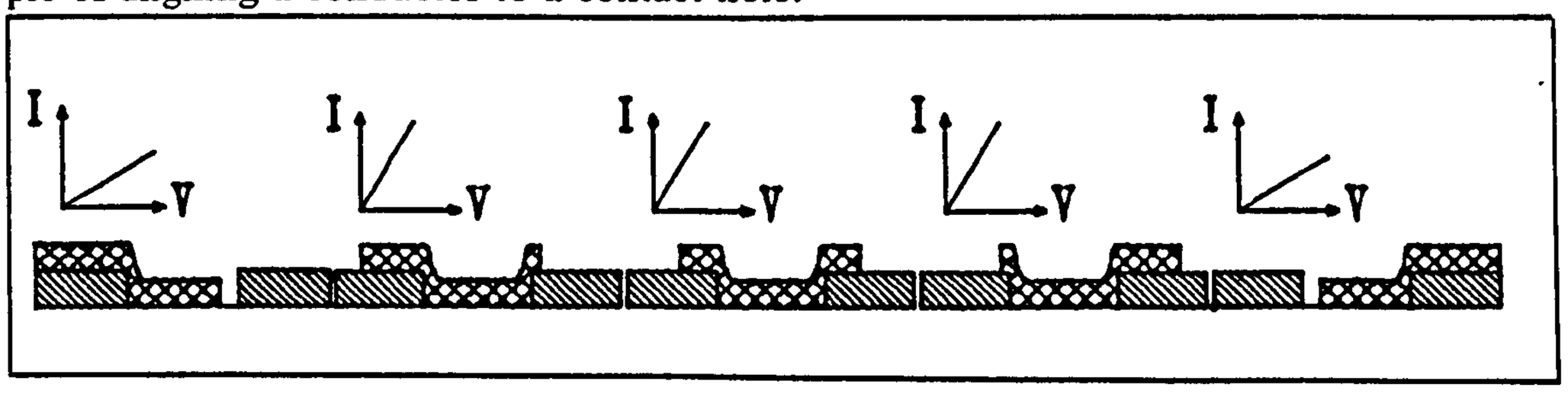


Figure 4.4, Application of cross-over for alignment detection.

It can be seen that the intended device lies between the two misaligned structures. If current through the contact is used as the alignment technique, then, as the devices are measured from left to right, the current will first increase, then stay steady for one or more structures, and eventually fall again. The cross-over is the midpoint between the rise and fall, and also the location of the aligned device.

Physical Verification.

Physical verification of the relative edges positions after tests are complete would be a useful confirmation of the electrical alignment technique. Optical microscopes could look at features visible from the surface, but their resolution (see table 2.12)

would not likely be sufficient. Scanning Electron Microscopes can easily measure displacements between features on the surface, but can not "see" through oxides to features below. A cross section and staining technique is useful for most structures, but relationships to edges of doped areas are dogged by problems in delineating the edge of a junction. It is possible to determine the metallurgical junctions but not the carrier concentrations this way. Although physical verification is always desirable, it is often very difficult and inconclusive.

4.5. Processing factors influencing implementation.

There are a few processing factors which can influence the implementation of the progressional offset technique and limit its resolution. They are variation in etching rates, both locally and across a wafer, and the effect of underlying topography.

Edge definition.

Edge variations caused by variations in etching rates in polycrystalline materials can effect the resolution of the technique. Figure 4.5, shows a schematic plan view of an edge. Small variations in the etch rate can cause inconsistencies in the position of an edge, effecting both the separation between like edges in the progressional offset pattern and the linearity of a single feature. The resolution of the technique is therefore limited to the magnitude of such variations plus the pattern generator uncertainty.

Topography.

Variations in the underlying topography can cause problems for a progressional offset scheme. Figure 4.6, shows a steep step in the underlying topography where perhaps metal may have to be patterned. Since the application of the photoresist provides some measure of planarisation, one structure in the progressional offset will have a thicker resist ($t=T$) to expose, with the same energy density, than another ($t<T$). That will result in offsets in the position of the etching mask edge that are not common to all sites in the progressional offset scheme. Two conclusions are; to avoid progressional offsets across drastic changes in topography, and if it is necessary to construct structures there, ensure that the topography variation is much smaller than the resist thickness.

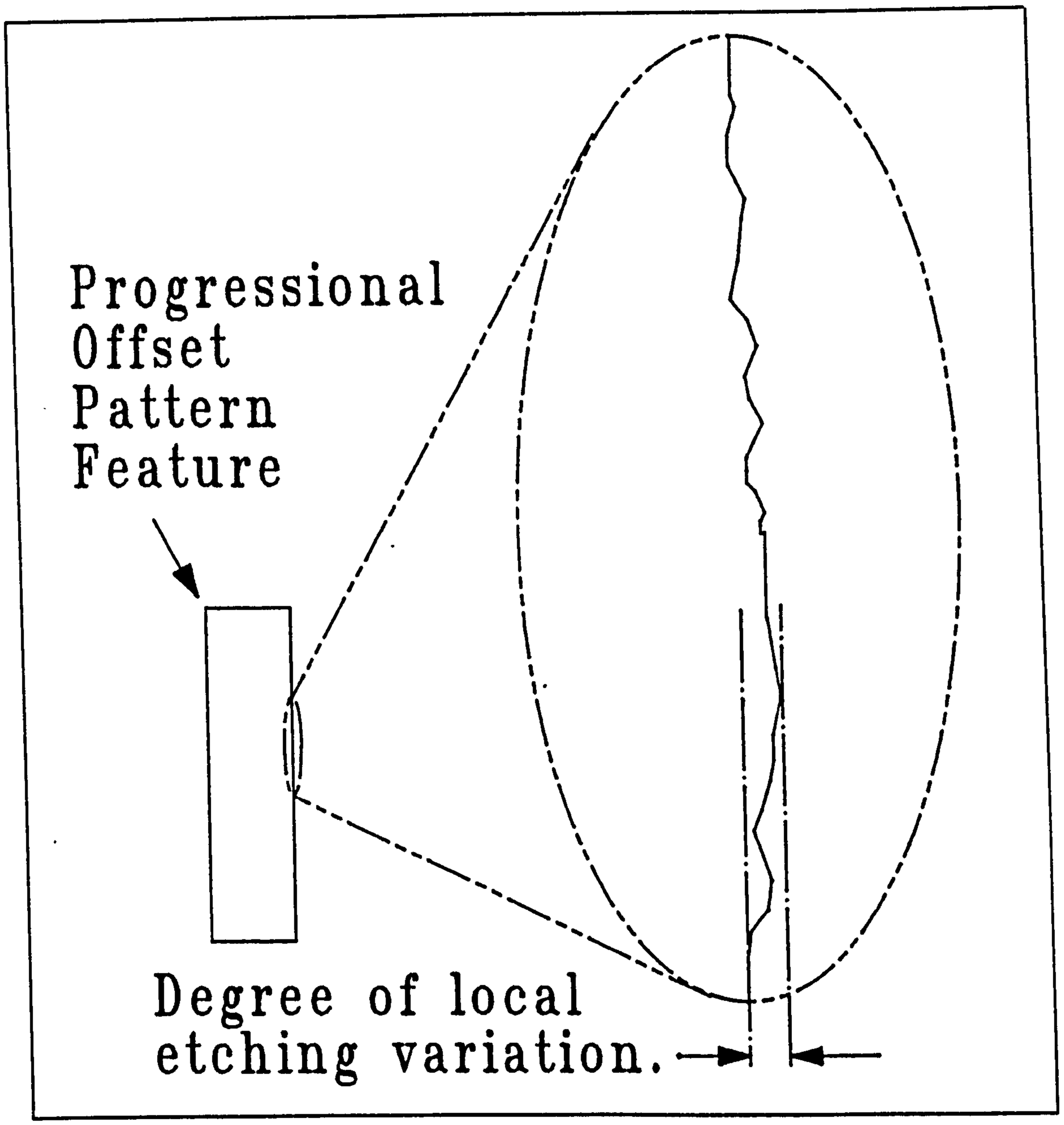


Figure 4.5, Local variations in etch rate.

Variations over distance.

Variations in processing over a wafer could have profound effects on the progressional offset technique. Thermal variations in furnace tubes, implanter dose consistency, photolithography, and etching variations across a wafer, have the potential to violate the assumption of the progressional offset. Therefore it is important to keep the progressional offset pattern compact to avoid aberrations to the pattern.

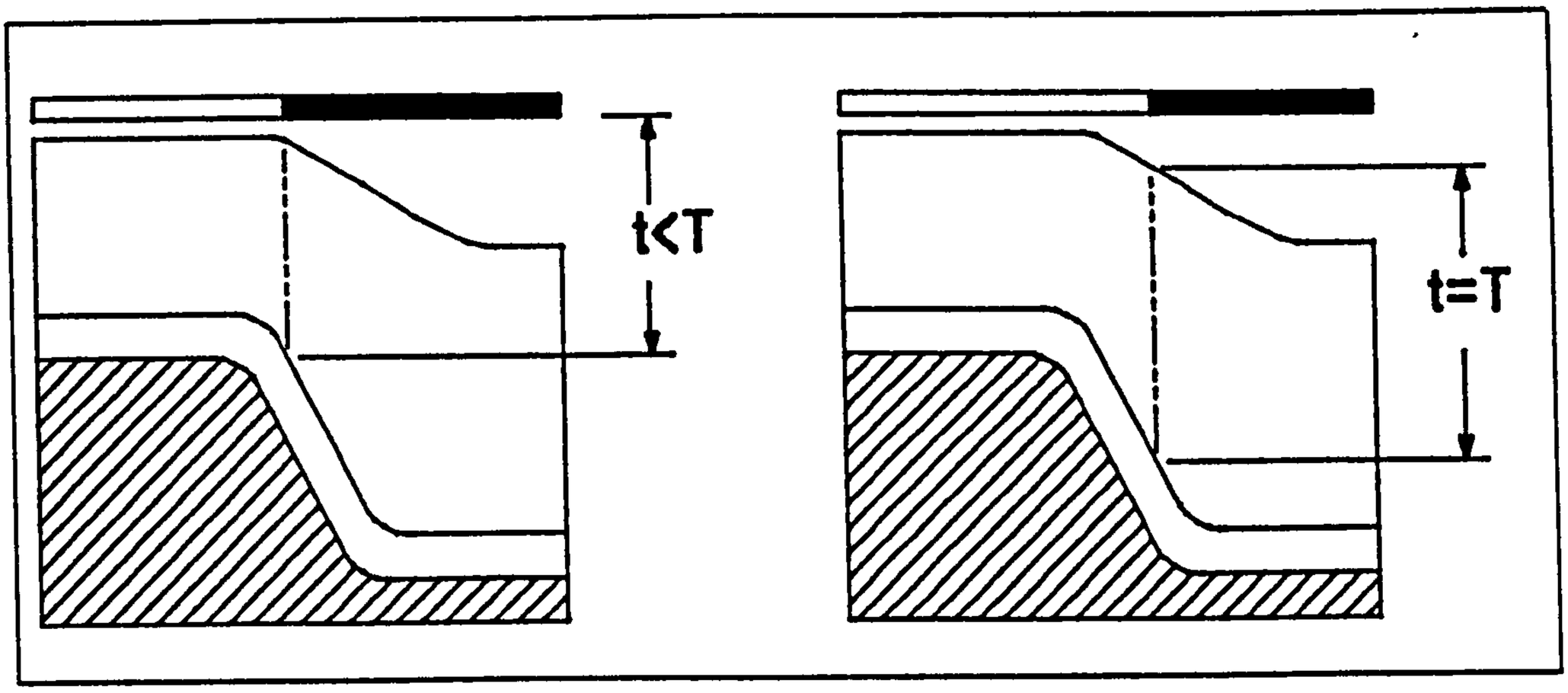


Figure 4.6, Topography variations can effect the scheme.

4.6. Chapter Summary.

A technique to build test structures with greater relative positional control than that available with standard production procedures has been described. The technique was given the name "progressional offset technique". The maximum resolution of the method was theoretically shown to be equal to the pattern generators positional accuracy, even after variations caused by processing. The technique requires an electrical method of determining the position of the intended structure in progressional array and it was suggested that a cross-over pattern designed into the structures would make confirmation easier. Finally, local edge position variation, steep topography, and processing variations across the array were discussed as possible complications.

4.7. References.

1. Henderson, B.M., Gundlach, A.M., and Walton, A.J., "Integrated-Circuit Test Structure which uses a Vernier to Electrically Measure Mask Misalignment," *Electronics Letters*, vol. 19, no. 21, 13 Oct. 1983.

Chapter 5. Mis-aligned Gate Experiment.

5.1. Introduction.

In chapter one, the region of an IGFET where the gate overlaps the source and drain was identified as an important edge region. In this chapter an experiment into the effect of this region on electrical characteristics, using the progressional offset technique to build suitable test structures, is described. The objective of the experiment was to build transistors with varying degrees of gate to drain (and source) overlaps, including gaps, on the same wafer, so that adequately controlled experiments on the effect of the overlap could be performed.

5.2. Background.

During the period of this Ph.D. research there was increased interest in source and drain edge effects. A few papers appeared during 1986,[1-5] more followed in 1987,[6-9] and one is promised in a major conference in 1988.[10]

That work can be classified into two groups. They are; papers by researchers interested in the effect of the gate electric field overlapping the source and drain,[3,4,7,8] and those interested in detecting and characteristing transistors where there is no gate overlap of the source or drain.[1,2,5,6,9,10] Both groups are briefly discussed in the following sections.

Gate to Channel Overlap Effects.

The most important feature of gate overlap of the source and drain is the way it increases gate capacitance (Miller capacitance). The increased gate capacitance affects the transistor's a.c. characteristics to the detriment of circuit speed. This effect is well known and was the motivation for the reduction in gate-to-source and gate-to-drain overlaps brought about by the processing advances mentioned in chapter one. There are, however, other effects.

The source and drain regions of an IGFET also effect the d.c. electrical characteristics since they represent a resistance in series with the channel. Attempts are usually made to keep the contribution to the total resistance small, but the series resistance

is still significant. The resistance of the transition region from the source, or drain, doping to the channel doping can be influenced by the gate electric field. As the channel lengths become smaller the transition region resistance becomes more important. Since the transition region is effected by the gate electric field, and the gate electric field is affected by the degree of overlap of the gate to the source and drain, it follows that this is an important area for edge effect experiments. The effect on LDD† transistors is of greater importance since the lightly doped region can be influenced to a greater extent by the gate electric field. The experimental work in this field has so far been limited to characterisation of gate-voltage-dependent series resistance in standard and LDD transistors.

Asymmetric Transistors.

Figure 5.1 shows how asymmetric transistors can result from the off-axis anti-channeling implant and sidewall-spacer technology.

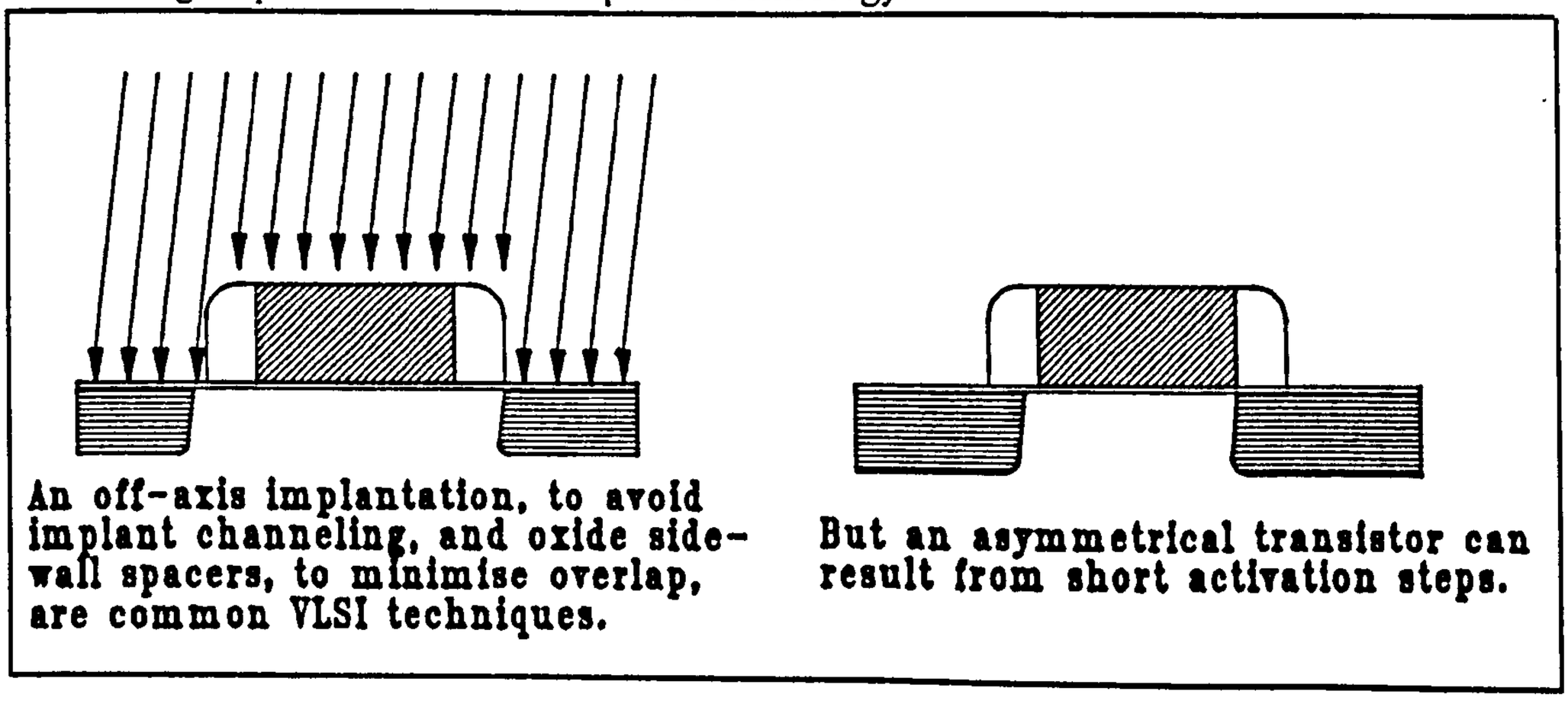


Figure 5.1, "Accidental" Asymmetric Transistor Formation.

Since sidewall spacer technology (Shown in figure 5.2) is routinely used in LDD technologies (to avoid hot electron effects) and minimum geometry technologies (to avoid gate-drain capacitance) the problem could be a widespread one. Research has shown that not only do asymmetric transistors result in changes in predicted operating

† Lightly Doped Drains (LDD).

currents, but cause entirely different operation in some configurations.[1] In some cases LDD transistors actually have worse hot electron performance than standard transistors, and it is now thought that this is due to asymmetric effects.[9]

Proposals for Novel Devices.

There was even a proposal predating the "accidental" asymmetric transistors, to use incompletely gated transistors to overcome some of the "small geometry effects".[11] The proposed structure was to have a gap in gate-to-channel coverage on both the source and drain sides of the channel. Computer simulations were used to compare the proposed structure against the standard MOSFET for such features as threshold voltage reduction and punch through voltage. They were found to be superior to those of the standard transistor.

Worldwide Experimental Work.

The first experimental characteristics of incompletely gated field effect transistors followed from that theoretical study.[12] It appears that the gate and gap lengths were one micron, but no details of the fabrication process, or the gate material, were given. Since the authors were mainly concerned about the punch through and threshold voltages, insufficient details about the other characteristics were given to make comparisons with this Ph.D. research. There has not been any follow-up of this work in any of the major journals.

There was however, one other experimental program concurrent with this Ph.D. research. It was aimed at identifying and characterising accidental asymmetric transistors.[1,2,6] It used sidewall spacer technology, shown in figure 5.2, to build a number of transistors with gaps in gate-to-channel coverage. Those transistors were called "weakly overlapped". The main purpose of the experiment was to allow asymmetric transistors to occur by the off axis anti-channeling implant being shadowed by the gate-sidewall-spacer structure.

In that experiment a number of wafers were prepared with large sidewall spacers using a joint process up to step five in figure 5.2. Then different annealing times were used to achieve a variety of gaps and overlaps in gate-to-source and gate-to-drain

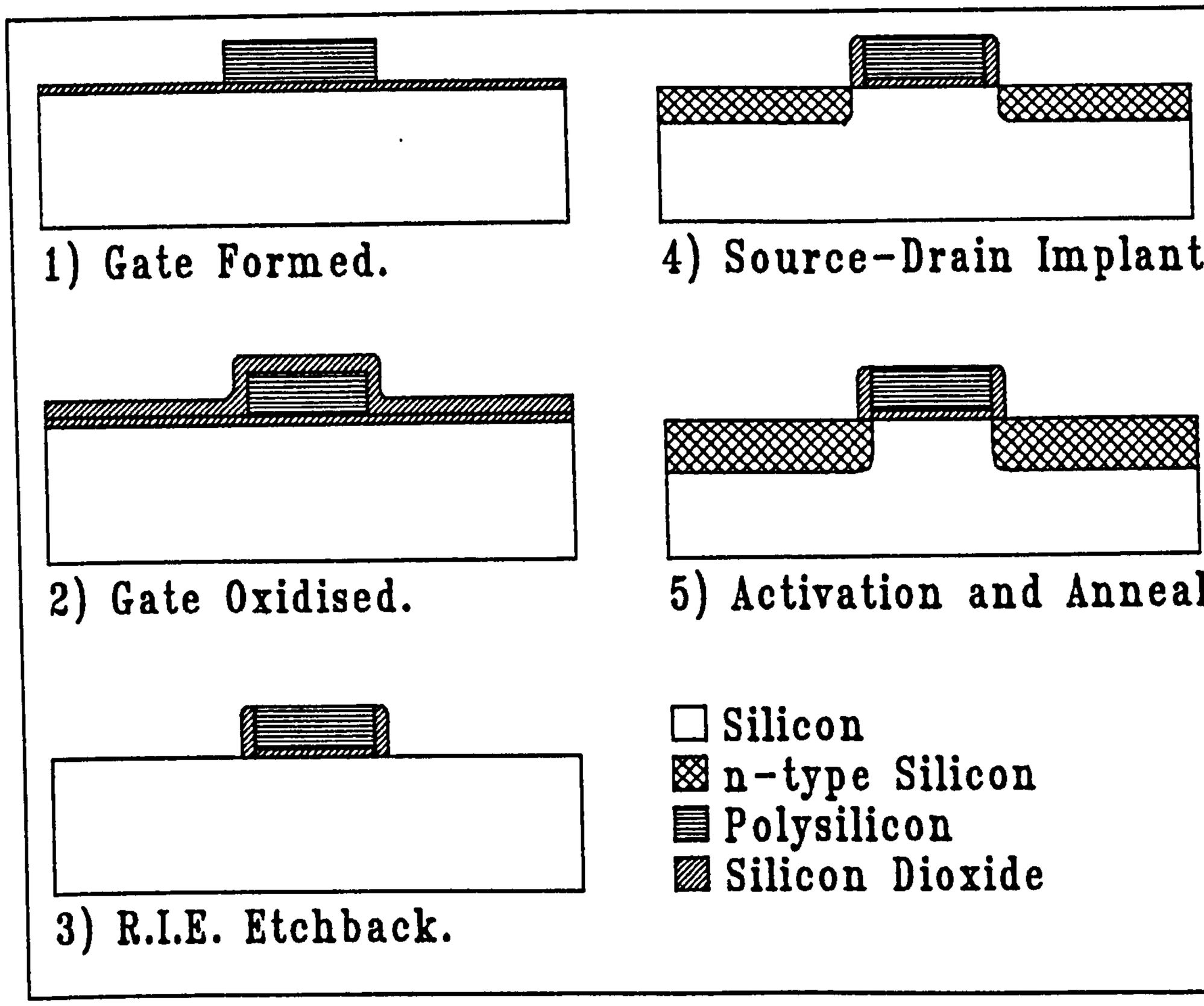


Figure 5.2, Sidewall Spacer Technology.

coverage. That experiment did succeed in demonstrating asymmetric effects, but one must wonder about the other differences between the transistors. The technique required transistors with different gap sizes to be on different wafers because of the approach of varying the heat treatment. Other factors, such as channel doping and oxide thicknesses, can also vary from wafer to wafer and might well have influenced the results. There is also the question of how the variation in sideways diffusion of the source and drain affects the junction behaviour.

Although there were drawbacks in the method of that research, it was important since it made the international community aware of the importance of edge effects in standard production IGFETs.

Furthermore, the term "weakly overlapped" may be appropriate for transistors whose gates slightly overlap, or perhaps even for those that nearly overlap the channel,

but it is completely misleading for the transistors which have a gap in gate-to-channel coverage. The terms used in this thesis are; a source gap transistor (SGT) for a transistor with a gap in gate-to-channel coverage at the source side of the channel, a drain gap transistor (DGT) for a transistor with a gap in gate-to-channel coverage at the drain side of the channel, and a normally gated transistor (NGT) for a transistor with complete gate-to-channel coverage.

5.3. Experiment Design

The design of an experiment for studying the effect of gate-to-channel coverage on the IGFET required not only the design of a test chip but also the design of a non-self-aligned-gate process with which to fabricate it. From the beginning, the conscious decision was taken to keep the design of both the chip and the process as simple as possible. That simplicity was to have allowed the highest chance of first time success along with the fastest design-to-chips turn around time. That decision also meant that in theory any unexpected effects would be easier to identify on the simpler design.

The criteria for process simplicity was met by using a non-self-aligned metal gate enhancement mode MOSFET process. Although commercially obsolete, it was updated to take advantage of the two orders of magnitude resolution improvement through the use of ion-implantation, direct-step-to-wafer photolithography, and reactive ion etching.

5.3.1. Sub-experiment for Edge Variation.

In the previous chapter, it was noted that local etching variations limit the resolution of the progressional offset technique. It was important to gain some insight into that limitation before designing the test chip. The objective of the sub-experiment was to establish the magnitude of the local etching variation and to assess the effect of metal thickness on the variation. The information required for the main experiment was; what was the optimal metal gate thickness, and what was the resulting local edge variation for that thickness.

Experiment Design.

A general purpose test recticle, containing lines, spaces, holes and isolated squares in a number of sizes from 0.9 to 5.0 microns, was used to pattern aluminium using the following procedure.

1) Thermal wet oxidation was used to grow a 550 Å oxide over bare silicon to mimic the "gate oxide" in the main experiment.
2) High purity aluminium was deposited by evaporation to obtain three different thicknesses on separate wafers.
3) An HPR-204 photoresist film was spun on the aluminium, and patterned by the general purpose test recticle using direct-step-to-wafer exposure.
4) The pattern was transferred to the aluminium using a chlorine chemistry R.I.E. system. Three separate etches were required to accommodate the different thicknesses.
5) The patterned wafers were then transferred to a furnace for low temperature sintering of (non-existent) contacts, as they would be in the main experiment.

Measurement Technique.

Three separate sets of measurements were made on the wafers to establish the metal thickness, and the pattern pitch and edge variations.

Four point probe electrical conductivity measurements were made after step 2 as one method of determining the metal thickness. After the wafers were patterned, a mechanical stylus (Sloan Dektak) was used to measure the wafer thickness at each die site on the wafer. Then the wafers were gold coated and examined using an SEM. The SEM measurements were used to determine the edge roughness and pitch for 1.5 and 2.0 micron lines and spaces. The edge roughness was defined as the magnitude of the edge variation along a length of line that was five times its width.

All the measurements were stored using the program "dbase" written for this work and briefly described in the appendix.

Results.

The results of the experiment contained some process specific information. The evaporator deposition rate at the power used was found to be 110 $Å s^{-1}$, with a radial variation of 1.3% and a standard deviation between wafers of 1.6%. It was found that there was good correlation between the electrical and mechanical thickness measurements, and a calibration factor was determined.

The results of importance to the main experiment are shown in the plot in figure 5.3.

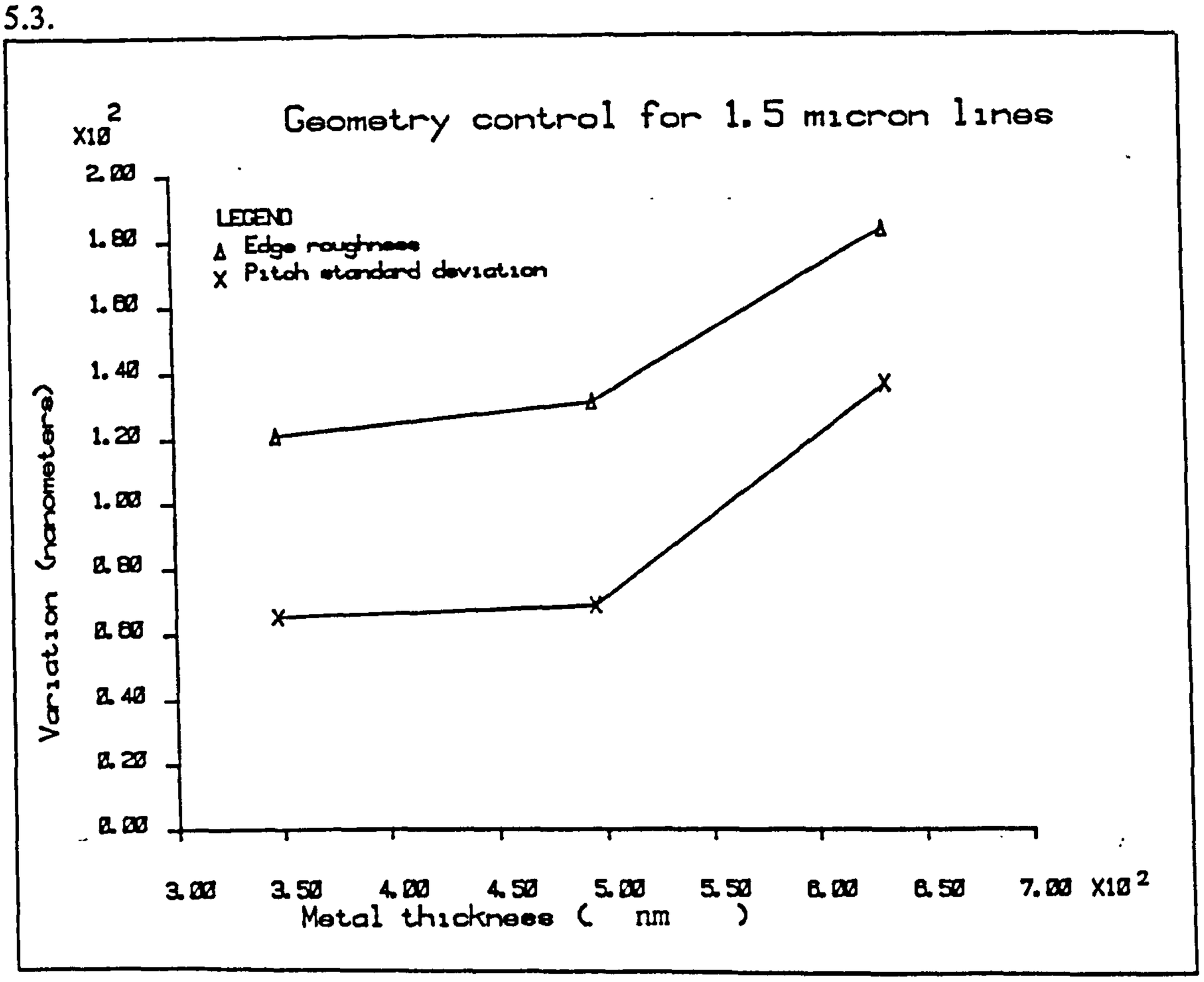


Figure 5.3, Geometry control for 1.5 μm lines.

It can be seen from the plot that the variation in edge roughness and the standard deviation in the pitch are closely related, which was what was predicted in the previous chapter. It can also be seen that the edge roughness decreases as the metal thickness decreases. It is also apparent that if thicker metal is necessary for pads (to prolong pad life under probe contact) the best choice is the 0.5 μm value as it has smaller edge

variations than the thicker metal. A similar plot resulted for the 2.0 μm geometries.

The important conclusion for the main experiment was that 0.5 μm thick aluminium should be used as the gate material with the minimum offset step size being 150 nm.

5.3.2. Chip Structure.

Figure 5.4 shows a three dimensional schematic of a progressional offset array for studying channel edge effects in small geometry IGFETs.

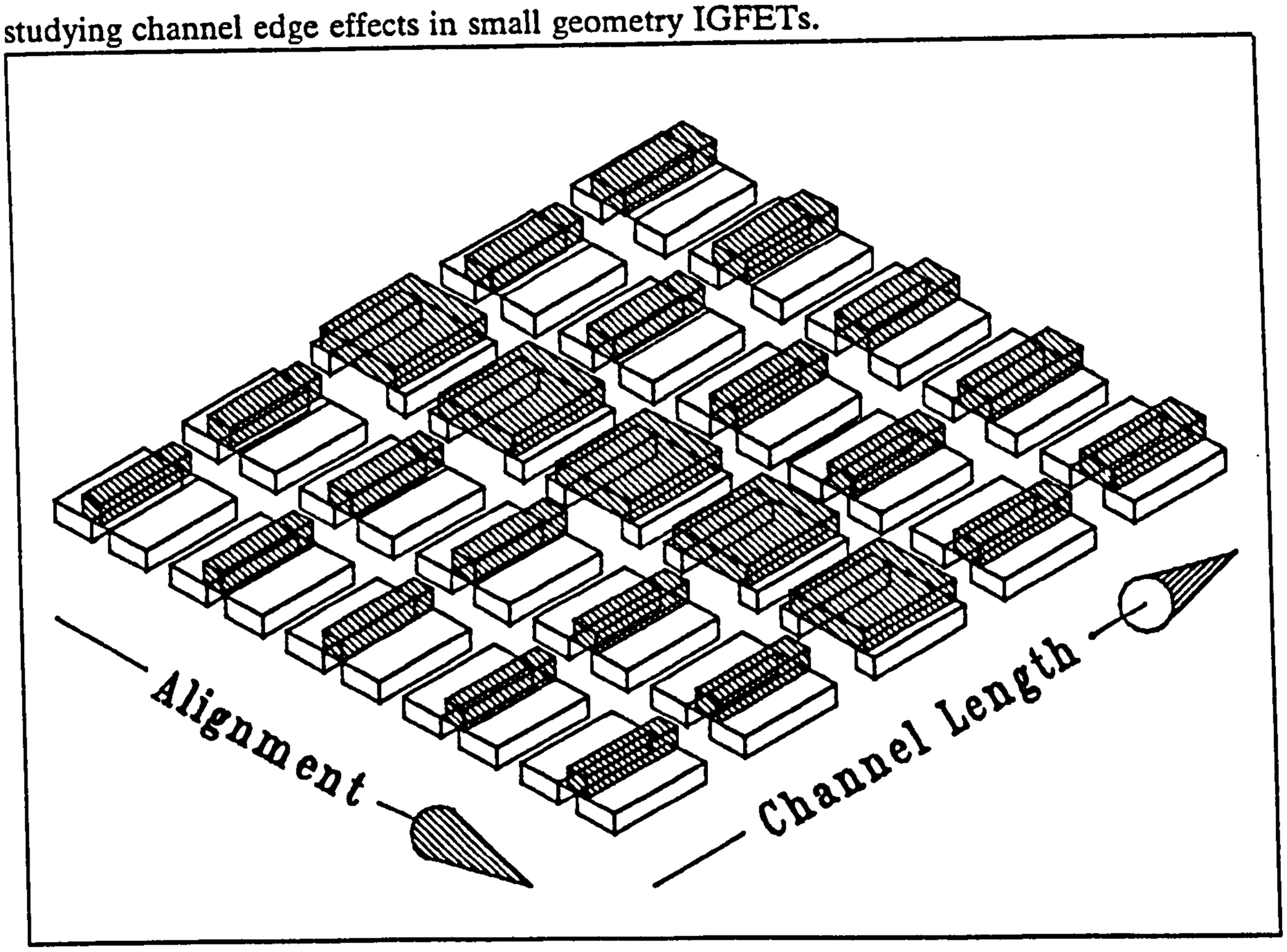


Figure 5.4, 3-D schematic of the progressional offset array chip structure.

Down the columns of the arrays the alignment of the gate to the channel was varied. Across a row the channel and gate lengths were increased. Large completely gated test structures were also included on every third column to act as controls for processing variations. In the actual layout the transistors were centered between three pads so that a "2 by N" probe card could be used to make connection to the source, drain, and gate. The backside of the wafer was used as the substrate connection.

5.3.3. Layout Methodology

The layout methodology was unusual in that a graphical layout editor was not used to create the layout. Instead the data base described in the appendix was used.

First a suitable entity was created to contain the rectangles that would make up the layout. The entity was called "layout" and contained the fields, "object", "layer", "xcenter", "ycenter", "xsize", "ysize", "xsite", "ysite", "type", "xcorner", and "ycorner". The "object" field was a text field containing the name of the layout object described by the rest of the fields which were numerical. The "layer" field contained the mask layer number. The "xcenter" and "ycenter" fields described the location of the center of a rectangle relative to the center of a transistor site, and the fields "xsize" and "ysize" defined the width and and length of it. The "xsite" and "ysite" fields contained the number of the transistor site. The "type" field was a flag for whether or not the transistor was a reference transistor, and the "xcorner" and "ycorner" fields contained the displacement vector from the chips lower left hand corner to the rectangles lower left hand corner. The database used the scale of 100 units to a micron.

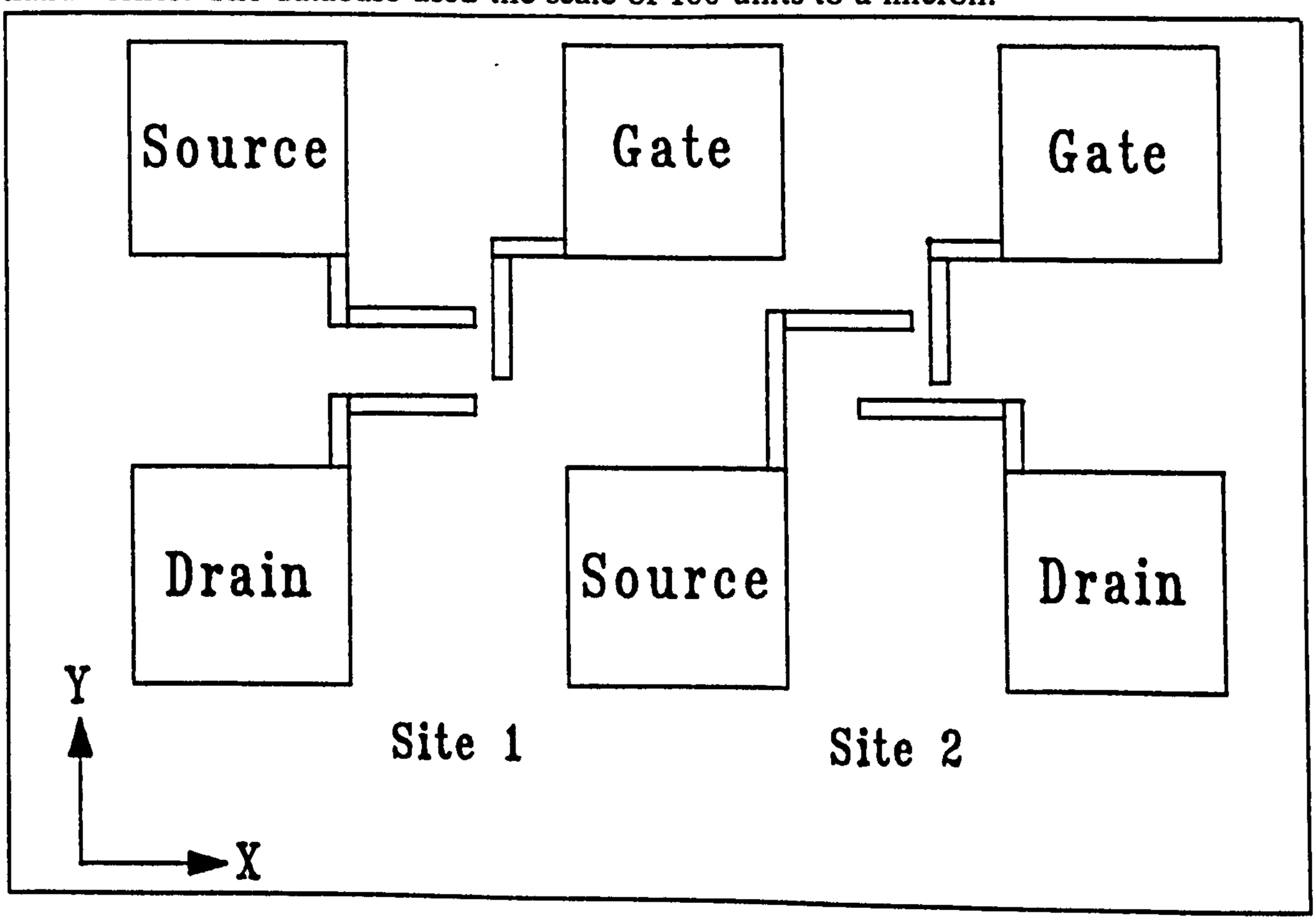


Figure 5.5, The basic six pad "shell".

To achieve the most effective use of the "2 by N" array of pads, two transistors used six pads in the basic "shell" shown in figure 5.5. In that way a square chip 2.8 mm on a side had an array of 18×18 pads which allowed for an array of transistors containing nine rows and twelve columns. Table 5.1, contains the database description of the pad "shell" as it was typed in.

Object	layer	xcenter	ycenter	xsize	ysize	xsite	ysite
gpad	4	12000	12000	12000	12000	2	1
dpad	4	12000	−12000	12000	12000	2	1
spad	4	−12000	−12000	12000	12000	2	1
interconnect	4	3500	6300	5000	600	2	1
interconnect	4	6300	−3500	600	5000	2	1
interconnect	4	−6300	−2500	600	7000	2	1
interconnect	4	1300	2750	600	6500	2	1
interconnect	4	−3000	1300	7200	600	2	1
interconnect	4	2700	−1300	6600	600	2	1
gpad	4	12000	12000	12000	12000	1	1
dpad	4	−12000	−12000	12000	12000	1	1
spad	4	−12000	−12000	12000	12000	1	1
interconnect	4	3500	6300	5000	600	1	1
interconnect	4	−6300	3500	600	5000	1	1
interconnect	4	−6300	−3500	600	5000	1	1
interconnect	4	−2700	1300	6600	600	1	1
interconnect	4	−2700	−1300	6600	600	1	1
interconnect	4	1300	2750	600	6500	1	1

Table 5.1, Database description of the six pad "shell".

The basic "core" transistor was then defined to fit in the center of the pads. The "core" transistors structure is shown in figure 5.6, and was entered into the database using the entries in table 5.2.

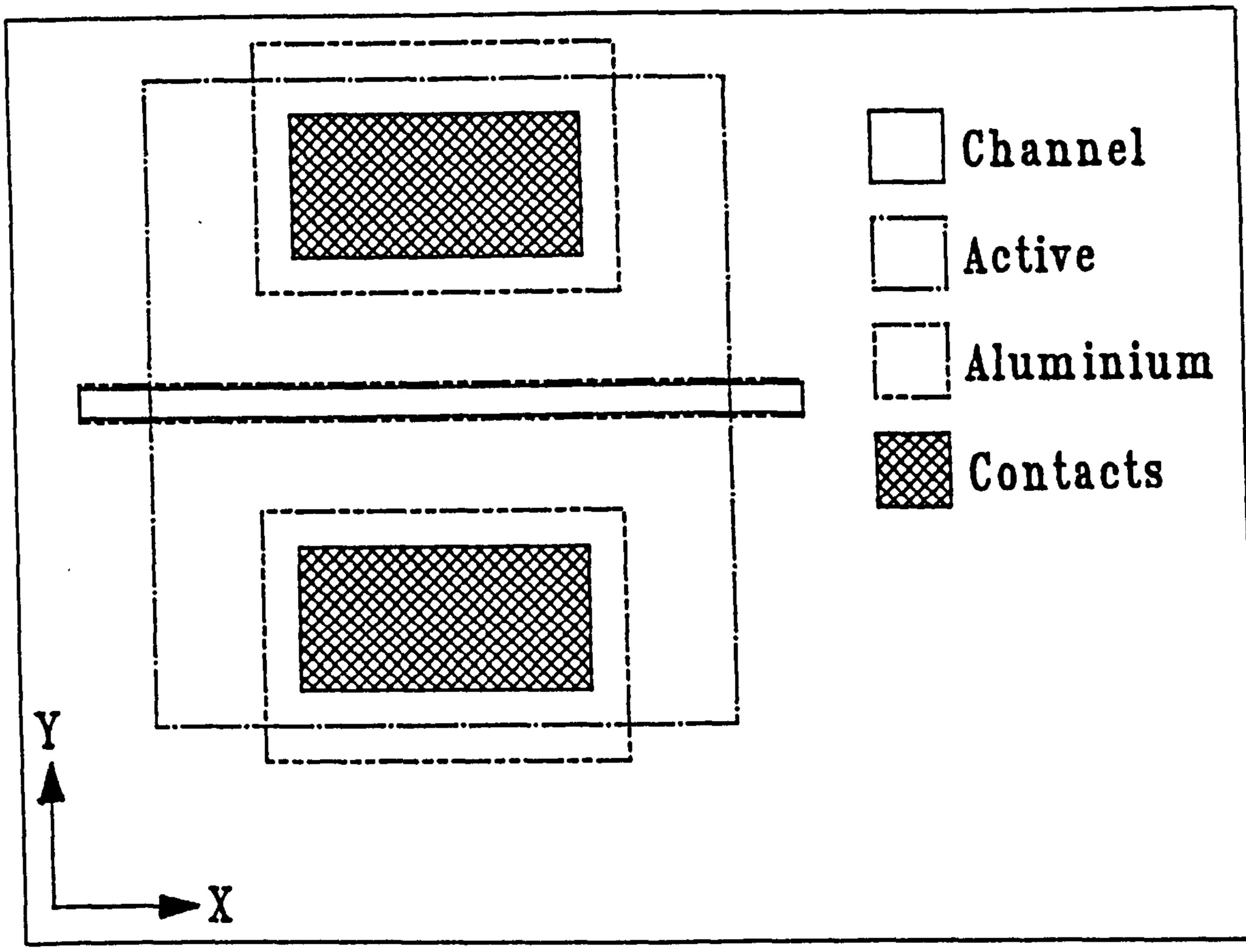


Figure 5.6, The "core" transistor.

Object	layer	xcenter	ycenter	xsize	ysize	xsite	ysite	type
active	1	0	0	1600	2800	1	1	1
scontact	3	0	900	800	400	1	1	1
dcontact	3	0	–900	800	400	1	1	1
interconnect	4	0	800	1000	400	1	1	1
interconnect	4	0	–800	1000	400	1	1	1
channel	2	0	0	2000	100	1	1	1
gate	4	0	0	2000	100	1	1	1

Table 5.2, The database entries for the "core" transistor.

The "core" transistor was copied into the second site with the command.

"find layout where type=1"

"copy 7 with xsite=xsite+1"

The site usage was defined to be of varying channel length in the x direction, and varying alignment in the y direction for test transistors. Table 5.3, lists the xsite usage and table 5.4 the "ysite" usage. Using that scheme every test transistor would have a reference transistor as a near neighbour. The range of offsets in the y direction was designed to exceed the direct-step-to-wafer projection exposure alignment error performance of 0.3 μm.

Xsite	Channel length	Use
1	5.00 μm	Reference.
2	1.00 μm	Test.
3	1.15 μm	Test.
4	5.00 μm	Reference.
5	1.30 μm	Test.
6	1.45 μm	Test.
7	5.00 μm	Reference.
8	1.60 μm	Test.
9	1.75 μm	Test.
10	5.00 μm	Reference.
11	1.90 μm	Test.
12	2.05 μm	Test.

Table 5.3, The "xsite" usage in the progressional offset array.

Ysite	Test transistor gate	Reference transistor gate
1	0.60 μm mis-alignment to the drain	Aligned with 0.30 μm overlap.
2	0.45 μm mis-alignment to the drain	Aligned with 0.30 μm overlap.
3	0.30 μm mis-alignment to the drain	Aligned with 0.30 μm overlap.
4	0.15 μm mis-alignment to the drain	Aligned with 0.30 μm overlap.
5	Aligned with zero overlap	Aligned with 0.30 μm overlap.
6	0.15 μm mis-alignment to the source	Aligned with 0.30 μm overlap.
7	0.30 μm mis-alignment to the source	Aligned with 0.30 μm overlap.
8	0.45 μm mis-alignment to the source	Aligned with 0.30 μm overlap.
9	0.60 μm mis-alignment to the source	Aligned with 0.30 μm overlap.

Table 5.4, Use of the "y" transistor sites.

To achieve that layout, the pad "shell" with its "cores" was copied in the following manner.

"find layout"

"copy with xsite=xsite+2"

"c w xs=xs+4"

"find layout"

"c w xs=xs+6"

This created a row of twelve sites. The row was then copied in a similar manner to form an array of twelve columns of nine sites, each with a unique "(xsite,ysite)" identifier. After those commands each site there would have had a one micron transistor with an aligned gate.

In order to introduce the variation into the array, the reference transistors were first identified and then their "type" changed so that subsequent changes were able to exclude them. Then the channel and gate lengths were varied. First the channel and gate rectangles were selected using the following "dbase" commands.

"find layout where object=channel"

"find layout where object=gate"

"union 1 2"

This resulted in a set containing all the gates and channel rectangles. The length was

then modified by the command.

"change 300 with ysize=70+15*(xsite-int((xsite-1)/3))"

Then the only the reference channels were selected, and their size was set to five microns,

"find 1 where type=1"

"ch 36 w ysize=500"

and then the gates of the reference transistors were selected, and their size made large enough to ensure overlap after processing.

Next the variation in alignment of the gate to the channel was made for the test transistors using the following commands;

"find layout where obj=gate"

"find 1 where type<>1"

This selected the test transistors gates, and was followed by a command to adjust the alignment to get the desired offsets,

"change 300 with yce=90 - 15*ysite"

to complete the array.

Scribe channels were added simply by typing them in. Finally the transistor rectangles "xcorner" and "ycorner" were defined in terms of the chip reference system by the commands,

"find layout where object<>scribe"

"change 2000 w yco=yce - ysize/2 + ysite*48000 - 13000"

"change 2000 w xco=xce - xsize/2 + (3*int((xsite-1)/2)-xsite+2*int(xsite/2))*24000+59000"

The rectangles were dumped into file using the database report commands and some custom software was written to fracture the rectangles into sizes small enough for the pattern generator.

Since the fine detail is lost in a plot large enough show the whole array, a schematic plan view, shown in figure 5.7, was drawn to demonstrate the resulting chip.

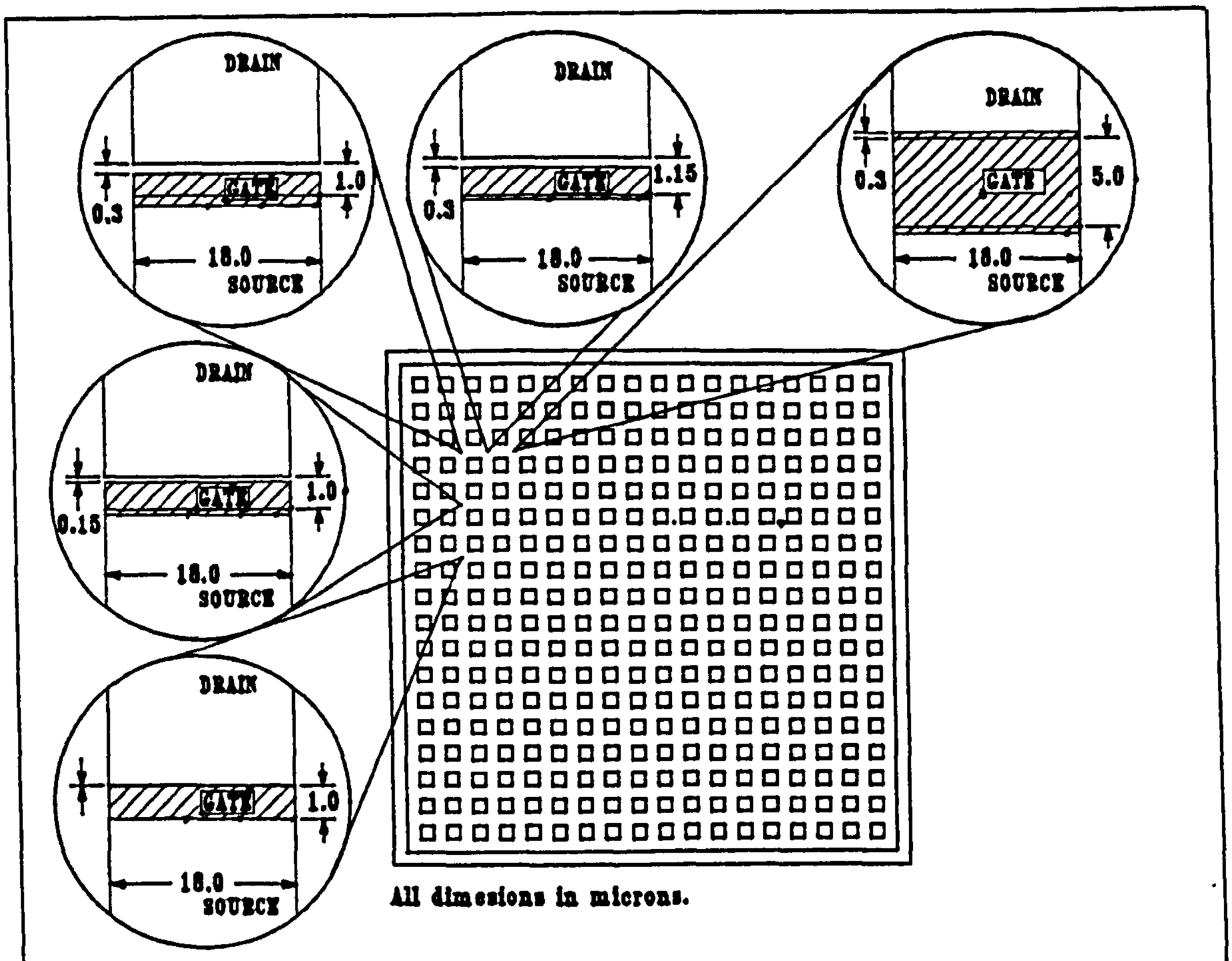


Figure 5.7, A schematic plan view of the test chip.

5.3.4. Process Summary.

A non-self-aligned metal gate process using LOCOS isolation was designed. The objective of the process was to be able to build the chip layout previously described. The twenty-nine step process is described in the following paragraphs and through the likely cross-sections in figures 5.8 to 5.13.

Step 1 Eleven p-type (100) three inches in diameter wafers with resistivities between 14 to 20 ohm-cm were used as the starting material. They were cleaned before starting a standard LOCOS isolation process.

Step 2 The first processing began with the growth of a pad oxide.

Step 3 Silicon-nitride was deposited, using a CVD furnace tube, to act as an oxide mask for the field oxidation.

Step 4 A standard photolithography cell, as was described in chapter two, was used to provide a resist layer masking the active area. The active area was defined by layer 1 in the database description of the layout.

Step 5 A Boron implant of 2×10^{13} *atoms* cm^{-2} at 130 KeV. Its purpose was to increase the inversion voltage of the field regions to prevent parasitic transistors.

Step 6 The surface was cleaned using a buffered HF solution, to remove any contamination from the implanter pumps, before the silicon nitride was etched in a $CF_4 + O_2$ plasma.

Step 7 The resist mask was removed using fuming nitric acid and the surface cleaned using a short HF dip before the next high temperature step.

Step 8 The field oxide was grown using a 9.5 hour wet oxidation in a 950 °C furnace.

Step 9 The front side of the wafer was protected by photoresist while the field oxide was etched off the back of the wafer. Then the silicon nitride oxidation mask was removed using a plasma similar to step 6.

Step 10 The pad oxide was removed using a buffered HF chemical etch which left bare silicon in the active areas.

Step 11 The surface was again cleaned with an HF dip before the growth of a sacrificial gate oxide. The sacrificial gate oxide was used to getter nitrogen from the edges of the field oxide. If left it would have caused a thinning of the gate oxide at the edges of the channel, which is called the "white ribbon" effect.

Step 12 The sacrificial oxide was removed using buffered HF again. At that point the LOCOS isolation was complete, and the surface of the active areas was ready for the transistor fabrication.

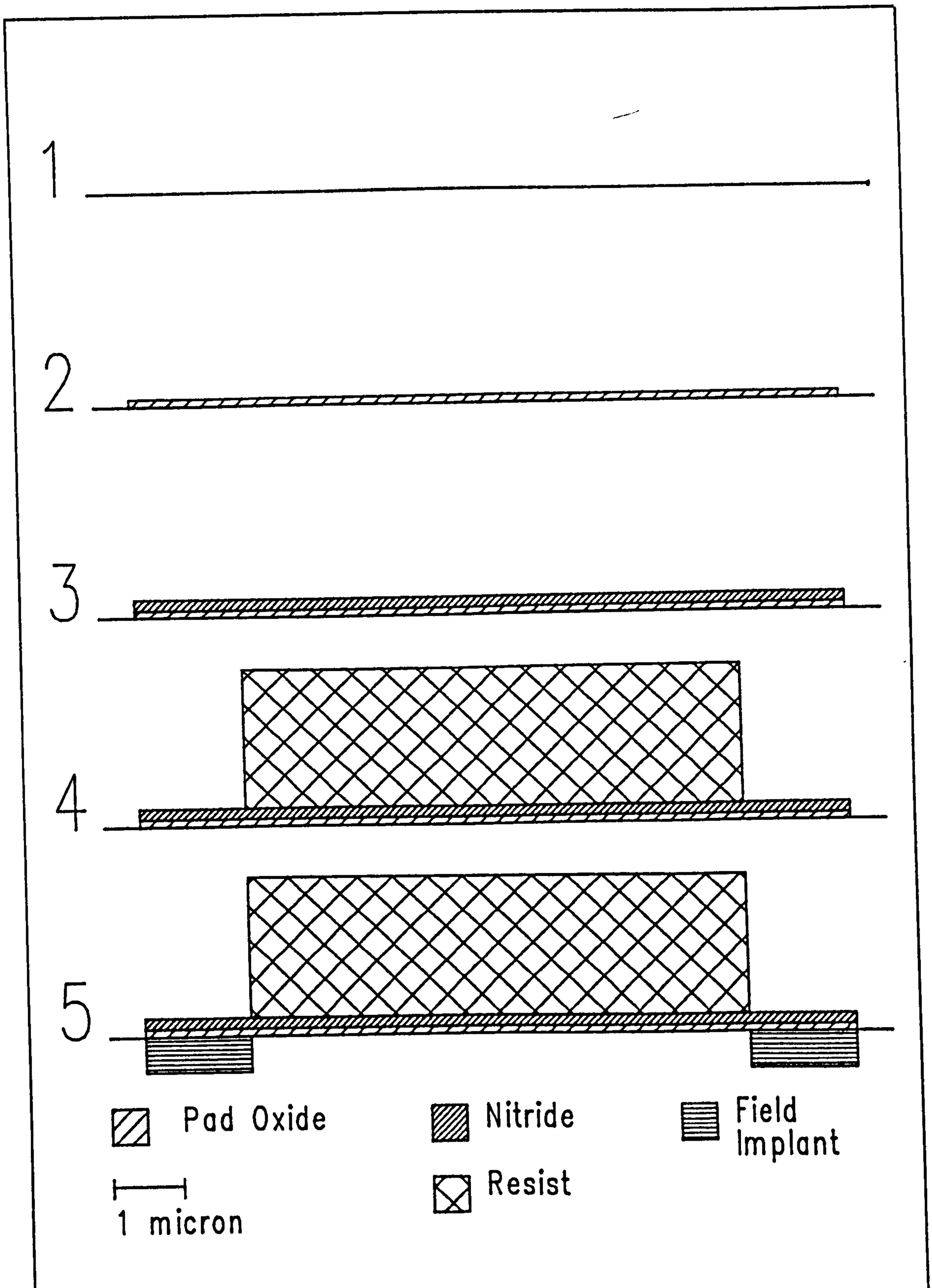


Figure 5.8, Non-self-aligned Metal Gate Process.

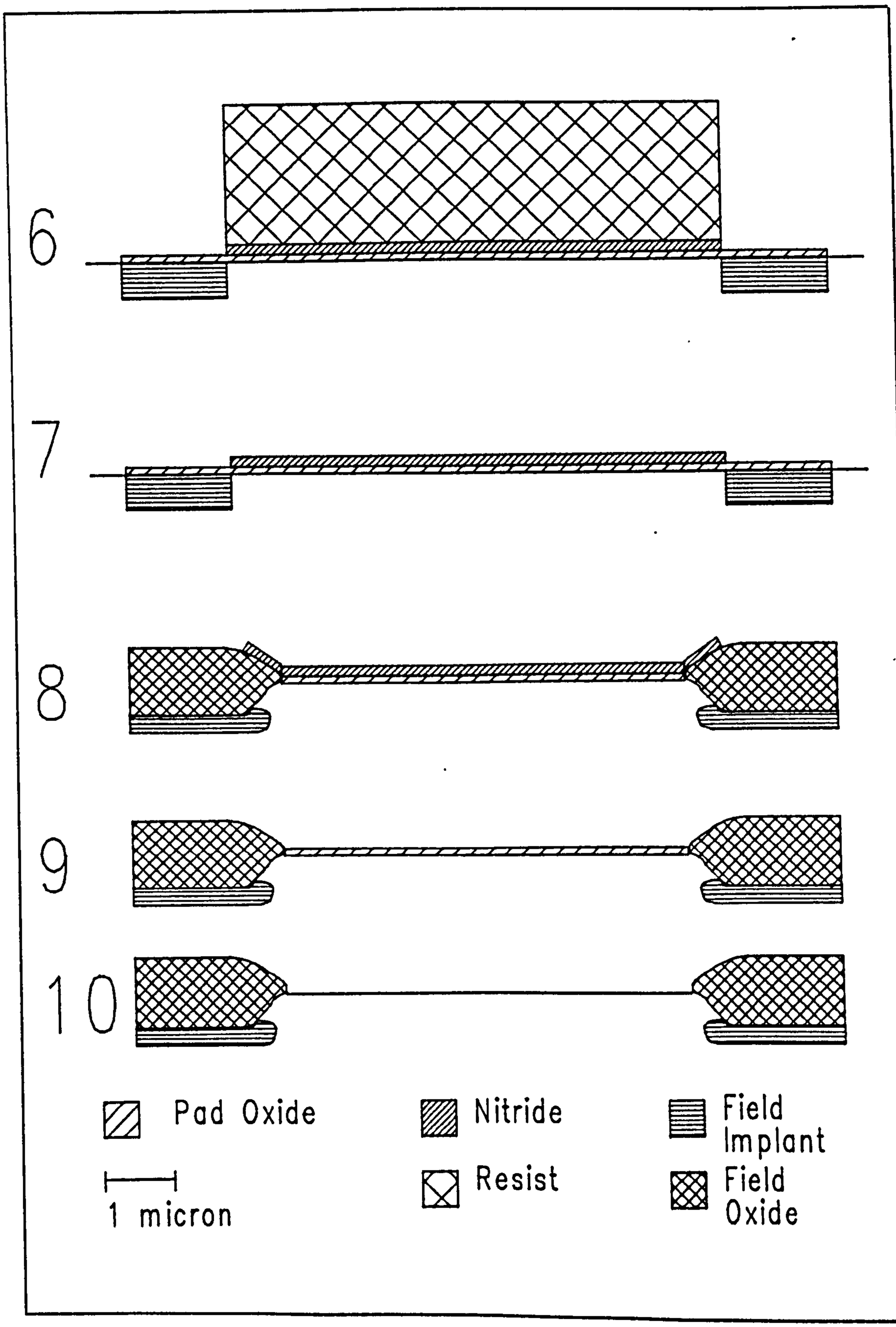


Figure 5.9, Non-self-aligned Metal Gate Process.

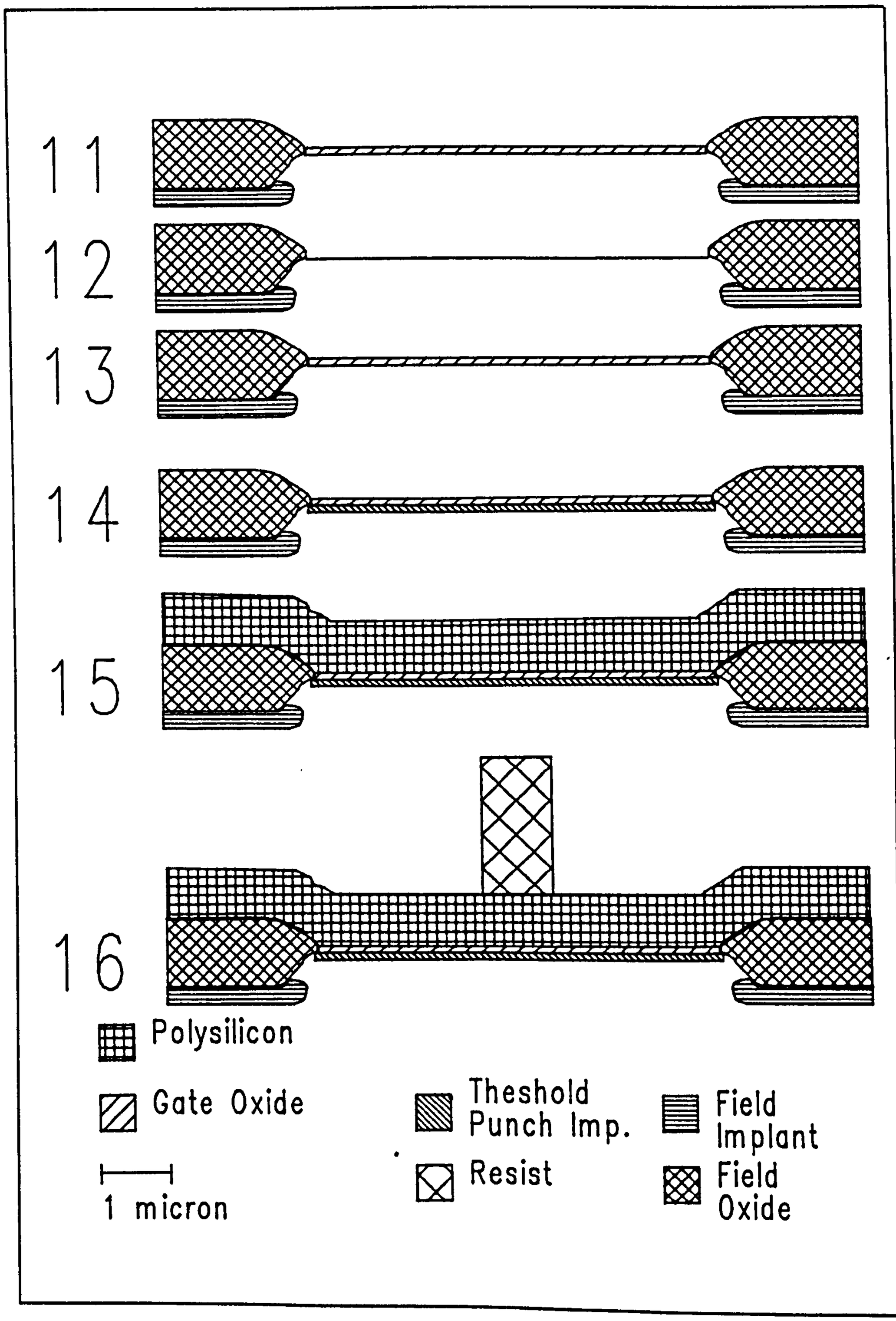


Figure 5.10, Non-self-aligned Metal Gate Process.

Step 13 The surface was again cleaned using an HF dip before the gate oxide growth. The gate oxide was grown by a 5 minute dry oxidation with an HCl getter, followed by 30 minutes of wet oxidation with HCl gettering, and finished by a 5 min dry oxidation. All were performed in a 950 °C furnace.

Step 14 The channel doping was modified by two boron implants through the gate oxide. One implant, 7×10^{11} *atoms* cm^{-2} at 40 KeV, was to increase the threshold voltage. The second, 2×10^{12} *atoms* cm^{-2} at 130 KeV, was to prevent punch through. The surface was then cleaned and annealed at 950 °C for 30 minutes in a nitrogen ambient to provide a controlled activation of the implants.

Step 15 A polysilicon layer was deposited by pyrolytic decomposition of silane gas.

Step 16 The second photolithography cell provided a resist mask defining the channel of the transistors. It was defined by mask layer 2 in the database definition.

Step 17 The channel mask layer was then transferred into the polysilicon layer using a chlorine chemistry reactive ion etcher.

Step 18 The photoresist was removed using a fuming nitric acid etch. The polysilicon was chosen as a source-drain implant mask since it had a fine edge definition and could stand the higher temperatures caused by high energy and heavy dose implants. High temperature photoresist was not available at the time.

Step 19 The source and drains were formed by two arsenic implants through the gate oxide. Two implants were used to tailor the impurity profile. One was 8×10^{13} *atoms* cm^{-2} at 340 KeV and the other was 8×10^{15} *atoms* cm^{-2} at 170 KeV.

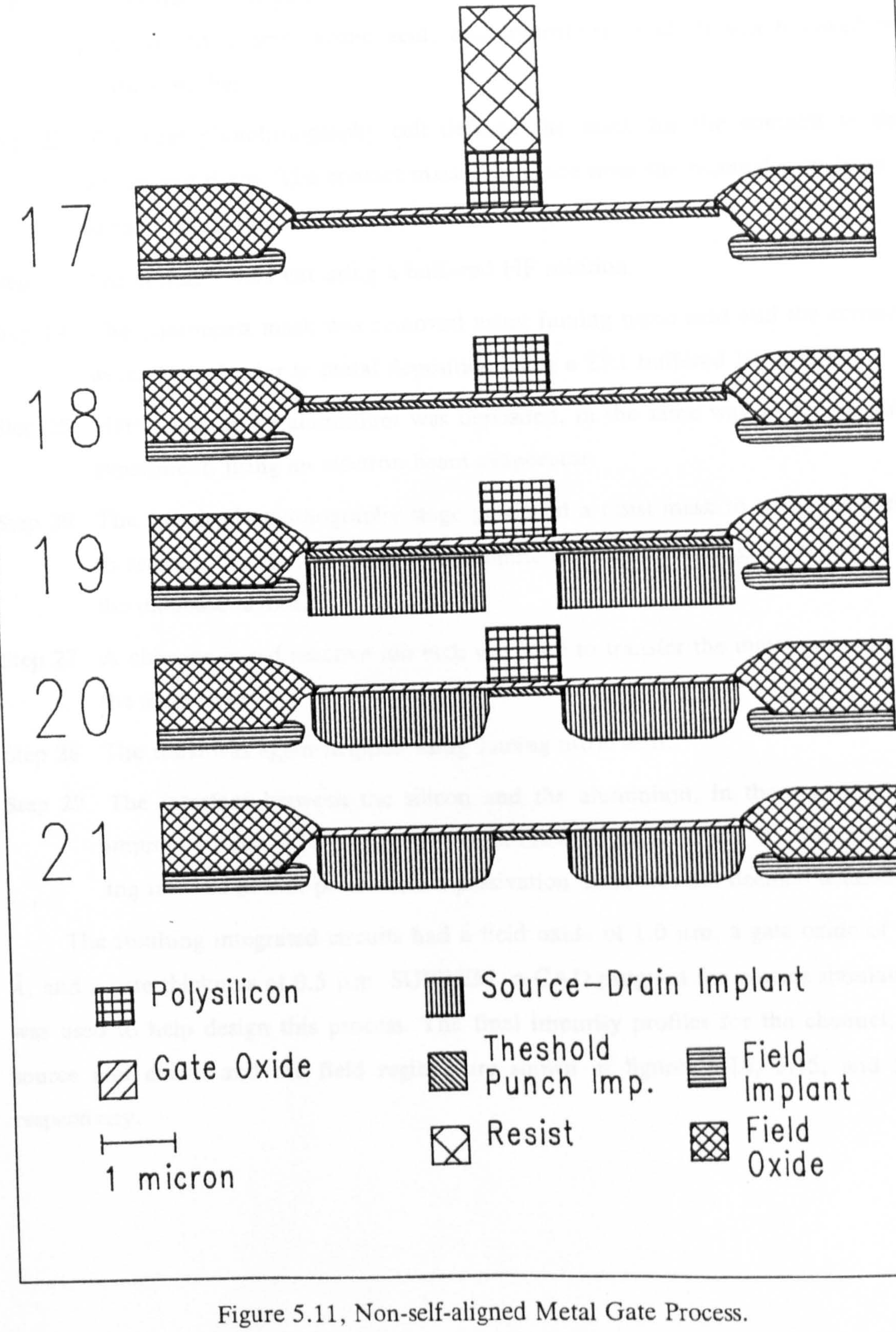


Figure 5.11, Non-self-aligned Metal Gate Process.

Step 20 The implants were annealed in a nitrogen ambient at 950 ° C for 30 minutes.

Step 21 The polysilicon channel mask was then removed using a chemical etch composed of nitric acid, acetic acid, and hydrofluric acid. It was followed by multiple washes.

Step 22 The next photolithography cell defined the mask for the contacts to the source and drain. The contact mask was made from the rectangles on layer 3 of the database definition.

Step 23 The contacts were cut using a buffered HF solution.

Step 24 The photoresist mask was removed using fuming nitric acid and the contacts were cleaned prior to metal deposition using a 25:1 buffered HF solution.

Step 25 Half a micron of aluminium was deposited, in the same way as for the sub-experiment, using an electron beam evaporator.

Step 26 The fourth photolithography stage produced a resist mask to cover the metal to remain. The fourth mask was composed of all the rectangles on layer 4 in the database layout.

Step 27 A chlorine based reactive ion etch was used to transfer the metal pattern into the aluminium.

Step 28 The resist was again stripped using fuming nitric acid.

Step 29 The interface between the silicon and the aluminium, in the contacts, was improved using a 10 minute hydrogen-nitrogen sinter at 435 °C. The processing finished at that point since a passivation mask was not deemed necessary.

The resulting integrated circuits had a field oxide of 1.0 μm, a gate oxide of 500 Å, and a gate thickness of 0.5 μm. SUPREM, a CAD program for process simulation, was used to help design this process. The final impurity profiles for the channel, the source and drain, and the field regions are shown in figures 5.14, 5.15, and 5.16 respectively.

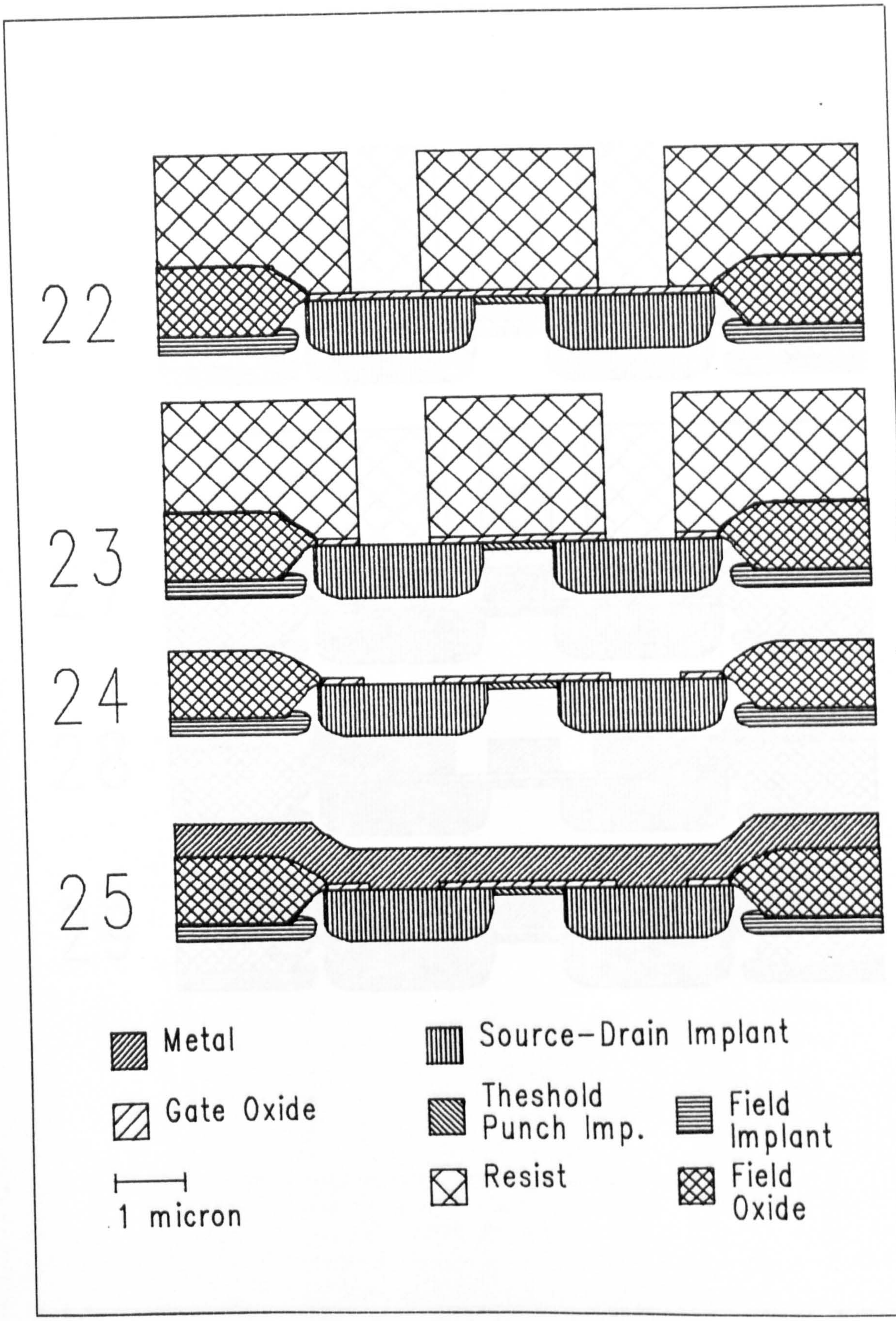


Figure 5.12, Non-self-aligned Metal Gate Process.

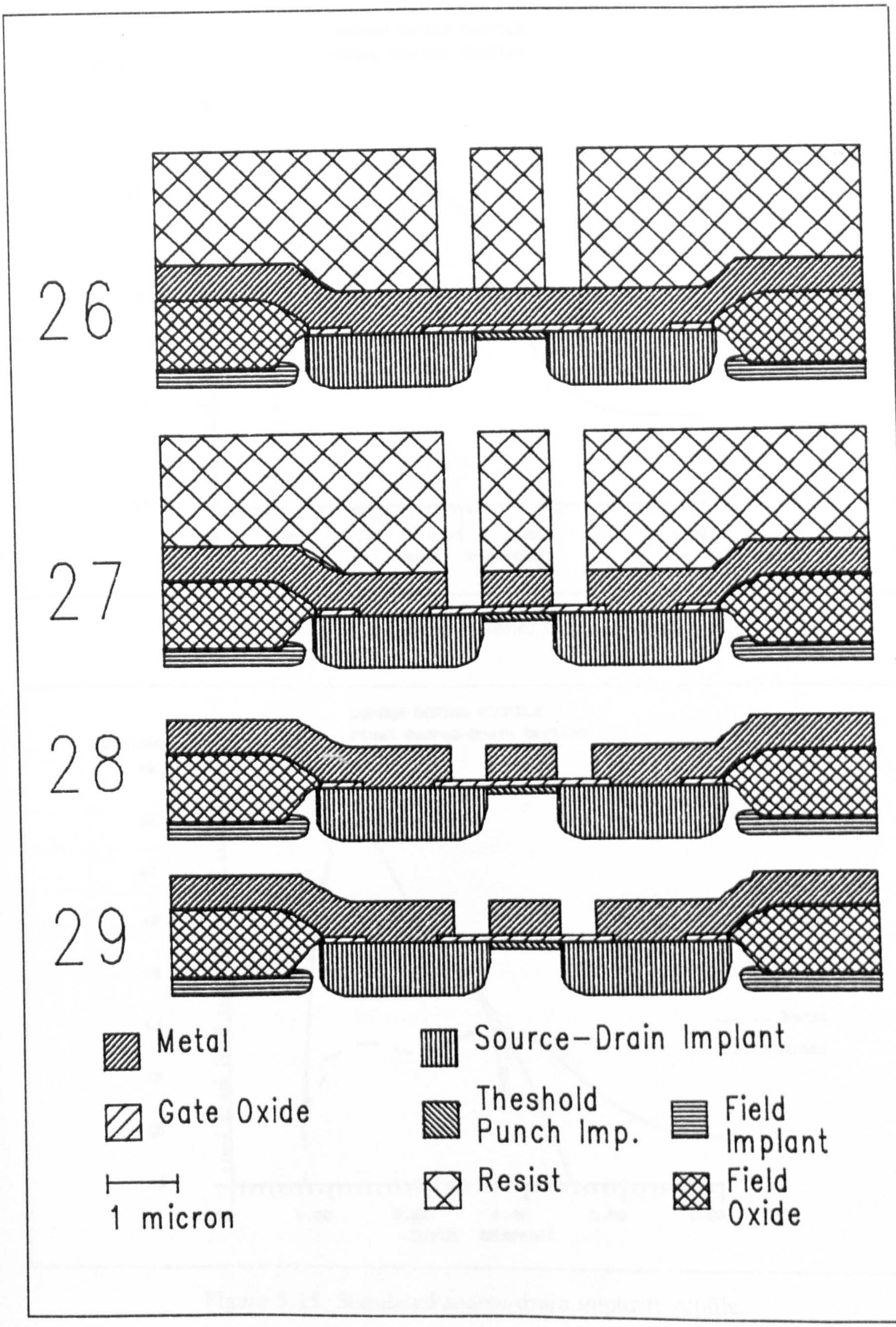


Figure 5.13, Non-self-aligned Metal Gate Process.

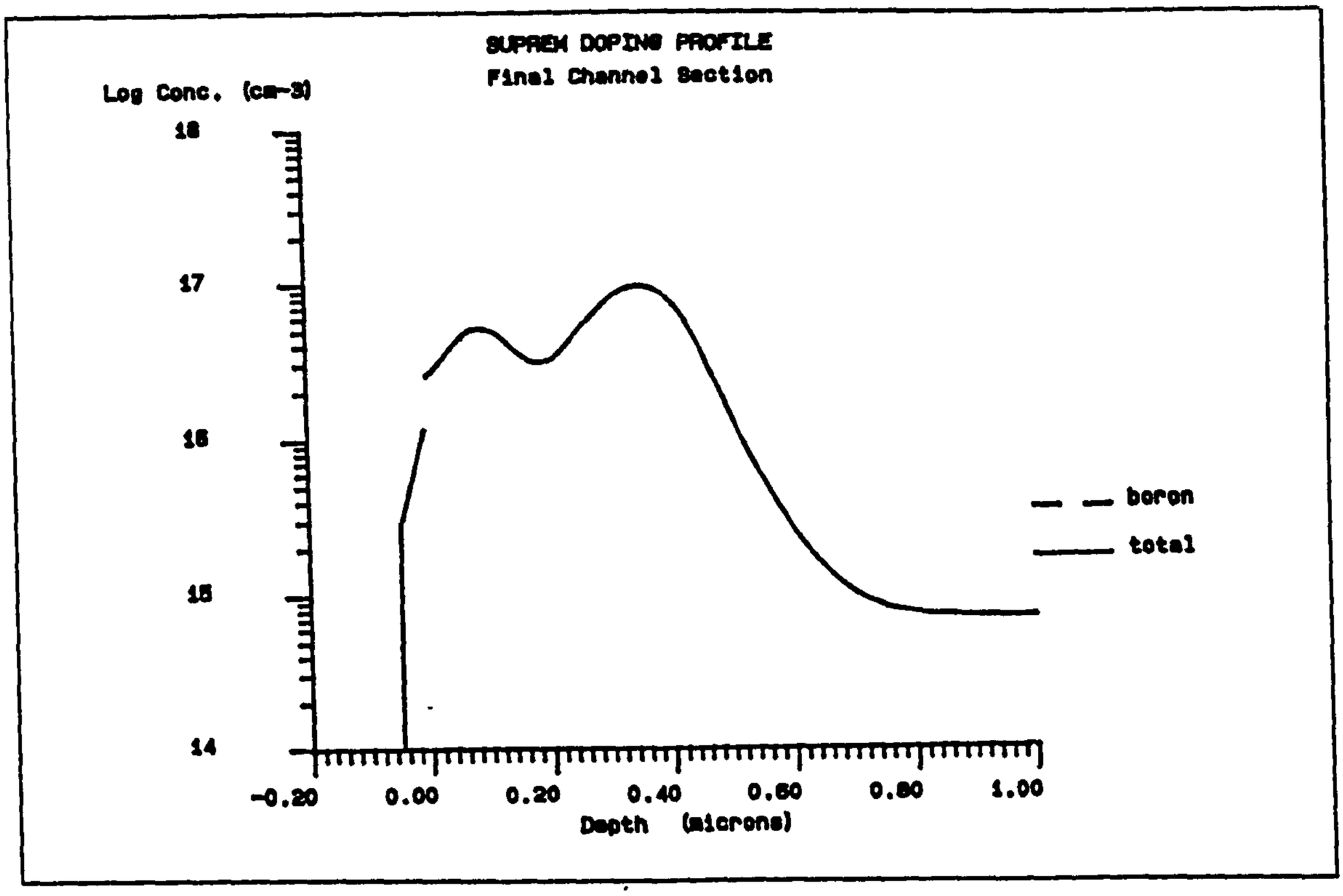


Figure 5.14, Simulated channel impurity profile.

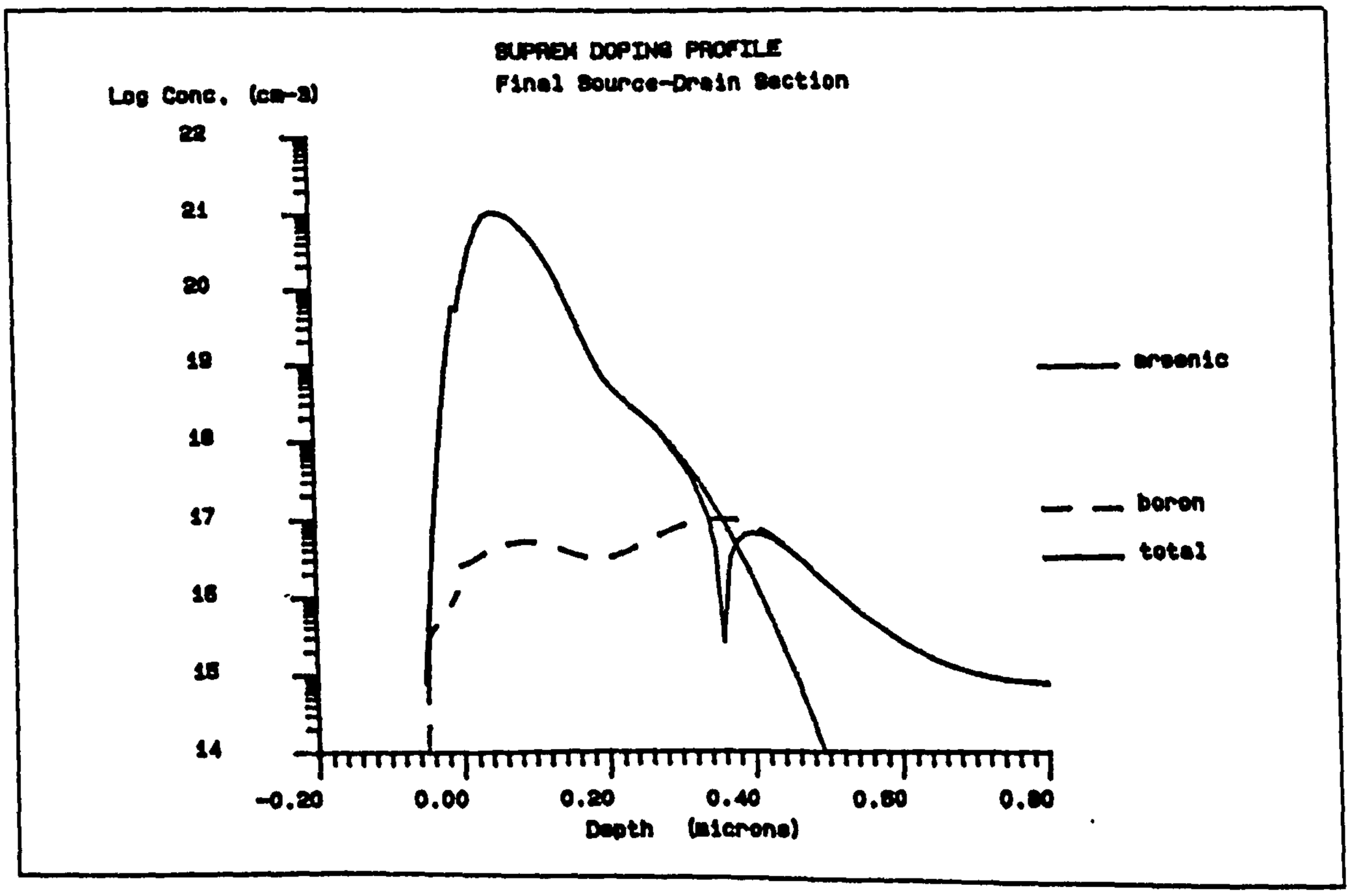


Figure 5.15, Simulated source-drain impurity profile.

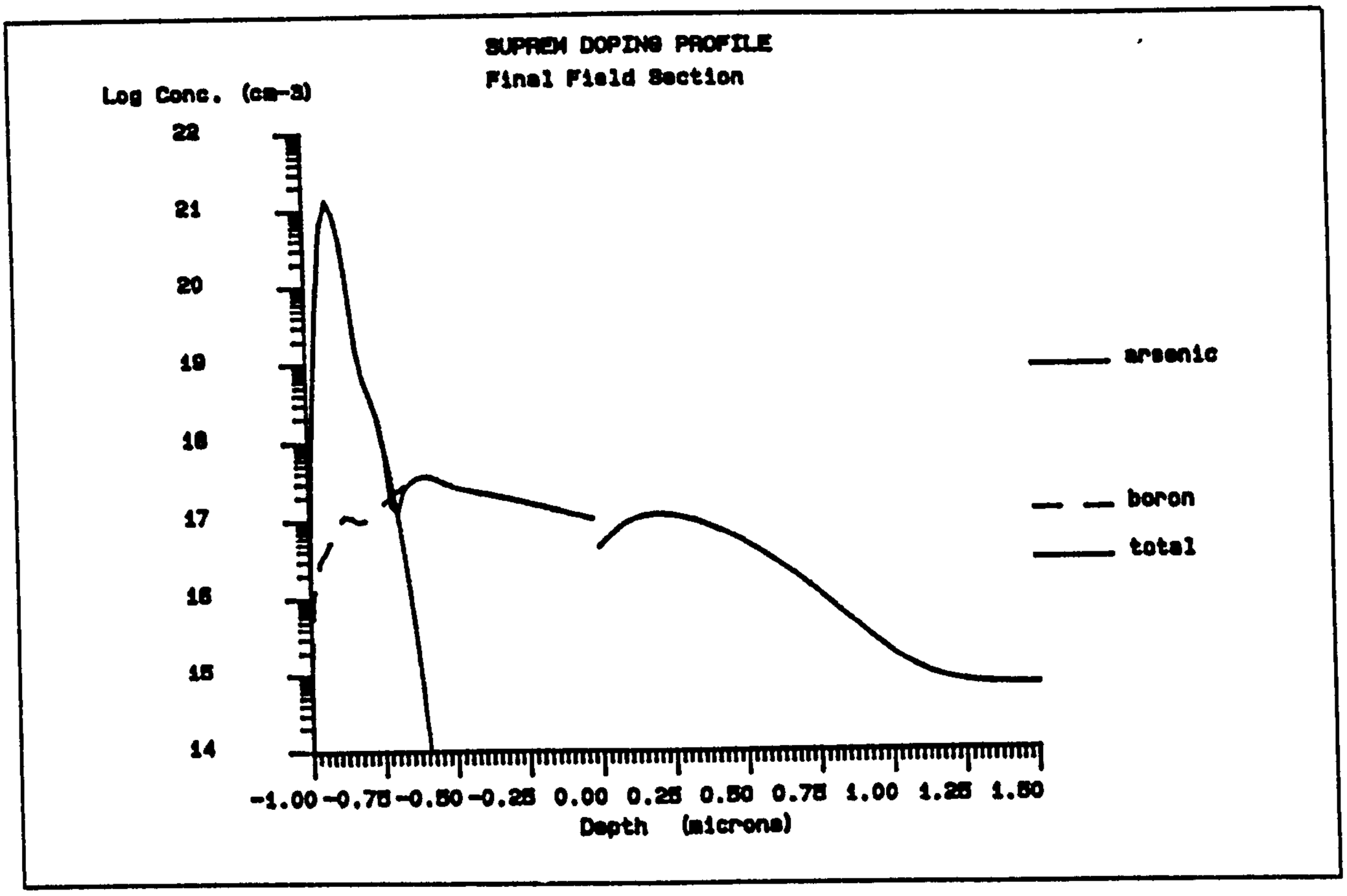


Figure 5.16, Simulated field impurity profile.

5.4. Fabrication Anomaly.

In chapter two the topics of selectivity and endpoint detection for plasma based etching were discussed. The problem of etching the gate oxide below the polysilicon to be removed in step 16, was identified as a possible problem with the process because of those limitations. The selectivity of the etch was low and the endpoint detection was difficult since little polysilicon area was to remain after the etch was complete. However, since high temperature photoresists were not available, there was no other solution. The precaution was taken to hold part of the batch at step 21.

Gate to Drain Shorts.

The first half of the batch was completed and 5 reference transistors were tested. All of them exhibited shorts (47.5Ω) between the gate and either the source or the drain. Obviously, they were not the insulating gate transistors envisaged.

The channel and near-channel oxide of the incompletely processed half batch of wafers was examined by a mechanical stylus and reflectometry to establish the oxide thicknesses. The channel regions consistently had oxide thicknesses of around 580 Å. The overetched near-channel oxide was only 180 Å to 220 Å thick. It was clear that the near-channel oxide had been seriously damaged by the polysilicon etch. Although the near-channel thickness would have been sufficient if it was a good film, it was not, and as a result presented plenty of pin-holes through which the gate to drain shorts formed during the sintering of the first half of the batch.

Thermal Oxide Repair.

One possible solution to the problem would have been to remove the gate oxide and regrow it. The difficulty with that solution was that the new gate oxide would have grown faster over the source and drain than it would over the channel region because of the increased doping. Such a variation would have reduced the usefulness of the array.

However, the original gate oxide provided a lucky opportunity for merely a repair of the damaged regions. It was thought that the gate oxide over the channel was good. It was known that any further oxidation would occur faster over the source and drain, and especially where the oxide was thinner. Therefore a simple "repair" oxidation of; 5 minutes of dry, 5 minutes of wet with an HCl getter, and a further five minutes of dry, oxidation all at 950 °C was performed based on the rational that the oxide over the source and drain would grow quicker and catch-up with the oxide thickness over the channel.

That is in fact what happened. The oxide over the source and drain grew 1.4 times as fast as the channel oxide. The final channel oxide was 675 Å and the near-channel oxide was 520 Å. The "new" oxide was then thick enough to resist pin holes and had a smooth transition from the channel to the source and drain. Processing was then completed on those wafers.

Opportunity for SEM Alignment Check.

The damage to the oxide on the first half of the batch provided a unique opportunity to test the success of the progressional offset scheme. Since the position of the channel was betrayed by the etching of the gate oxide where it was not protected by the channel mask, an SEM could be used to compare the position of the gates to the position of the channel. A series of scanning electron micrographs down a column of 1.0 μm test transistors shows the source and drain contacts and the gate in metal, and the outline of the channel in the gate oxide. Figure 5.17, shows a mis-alignment of 0.45 μm, figure 5.18, the mis-alignment of 0.30 μm, figure 5.19, the 0.15 μm mis-alignment, and figure 5.20, the perfect alignment of the gate to the channel. The feasibility of the progressional offset technique is demonstrated by these SEMs.

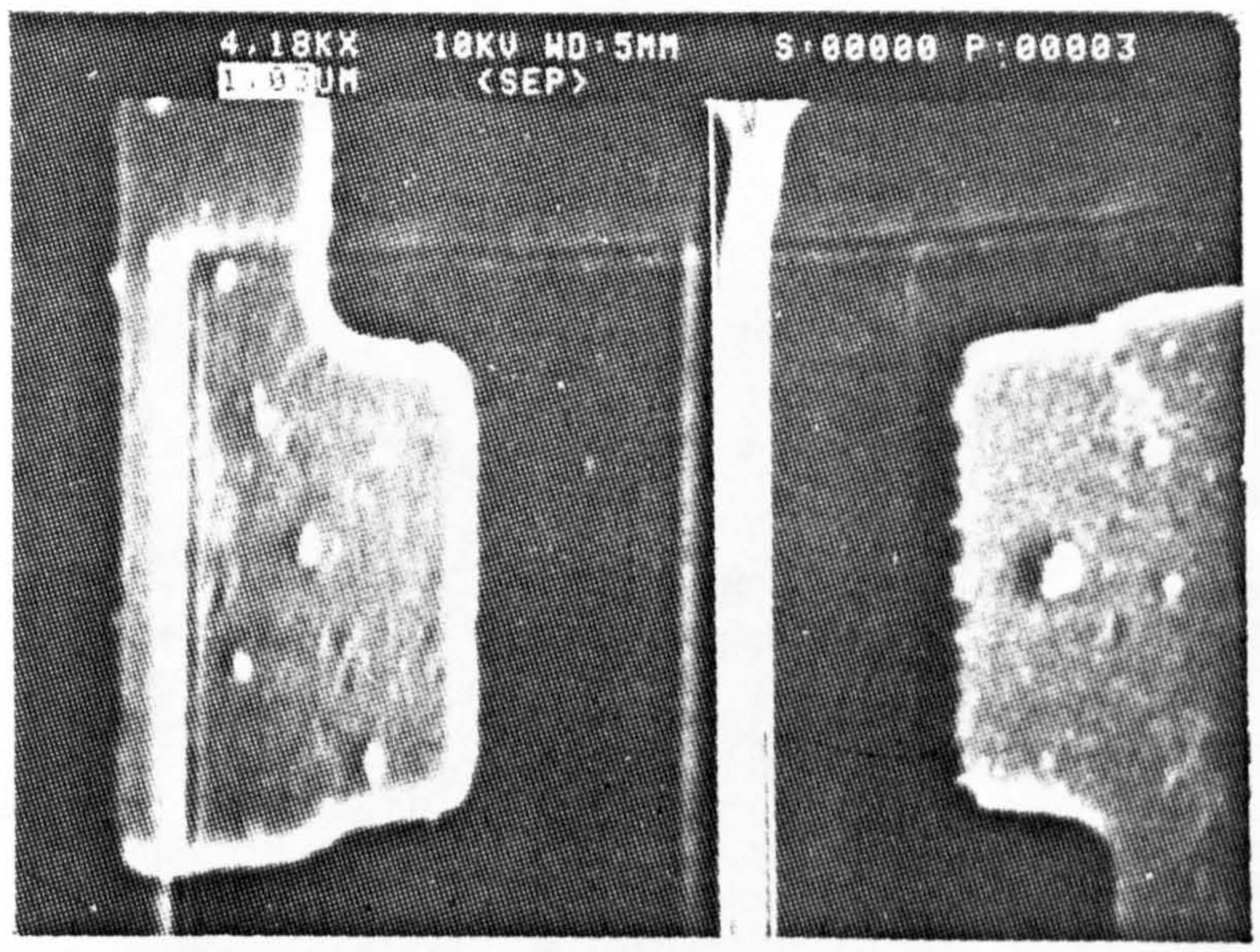


Figure 5.17, Progressional offset of 0.45 μm.

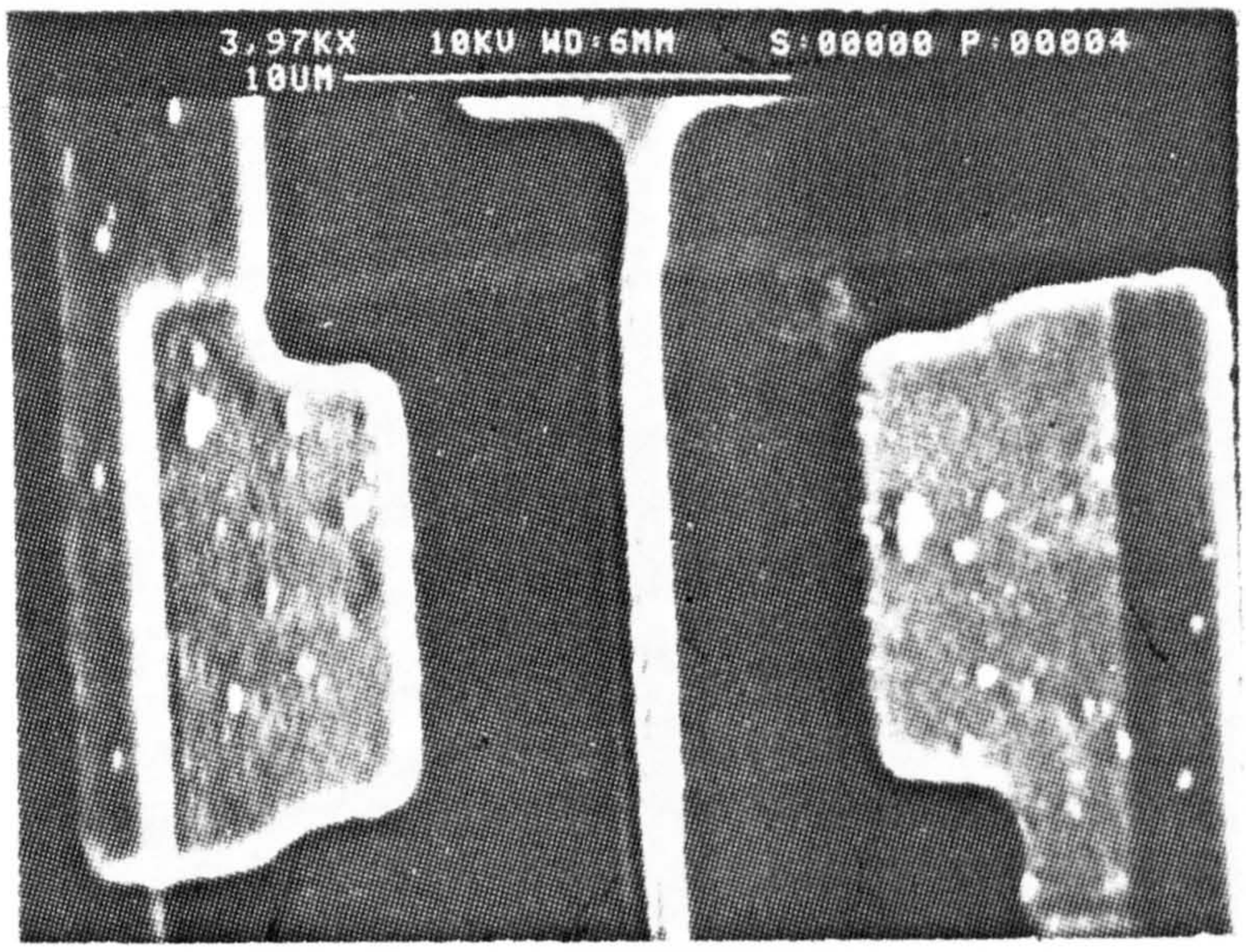


Figure 5.18, Progressional offset of 0.30 μm.

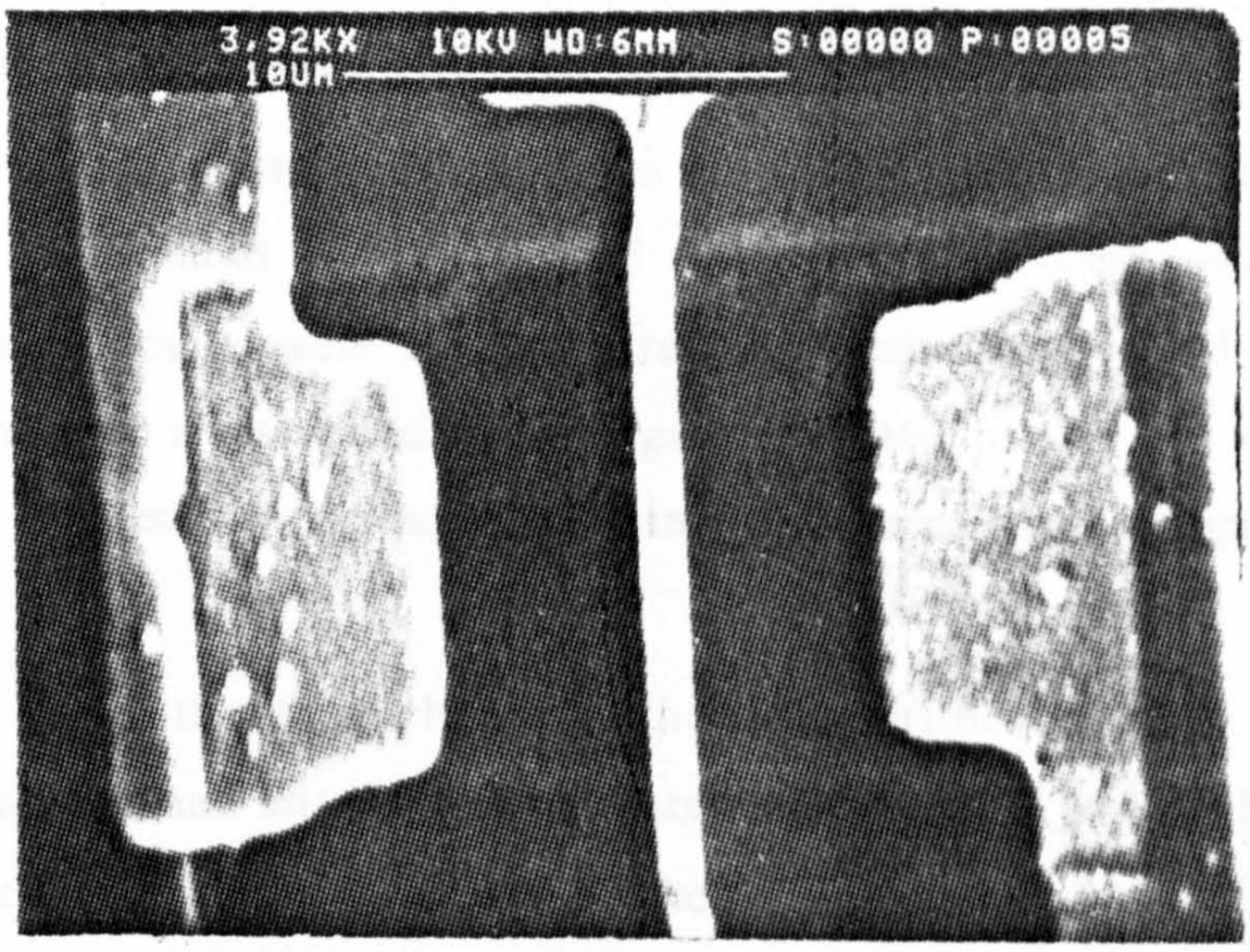


Figure 5.19, Progressional offset of 0.15 μm.

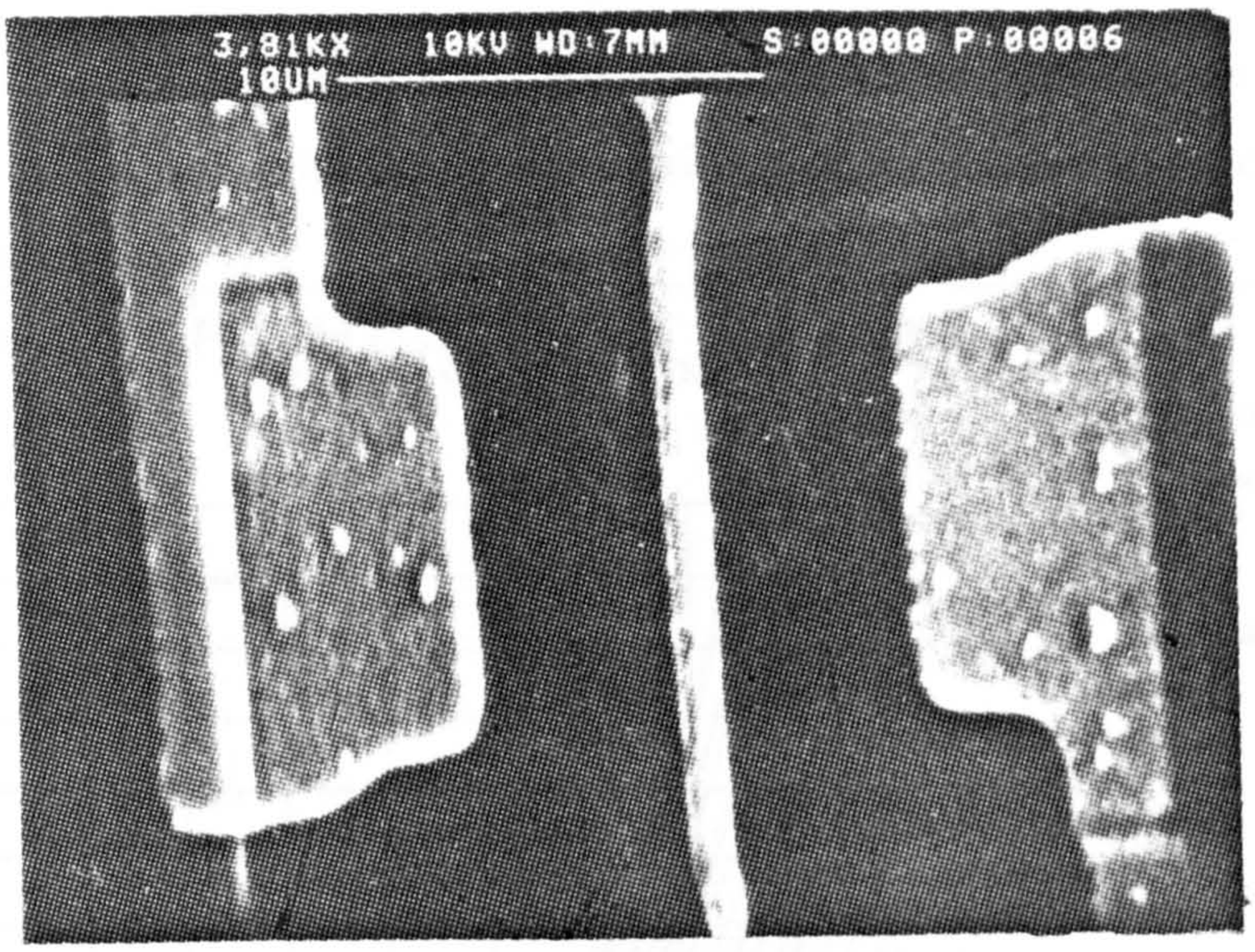


Figure 5.20, Gate to channel alignment in the progressional offset scheme.

5.5. Water Vapour Charging.

The reference transistors tested in the "repaired" half of the batch worked as expected with a threshold voltage of 1.17 volts and good drain voltage saturation behaviour. Similar results were found for reference transistors on other wafers. The first measurement of a test transistor also gave expected results, but when the same transistor was immediately measured again, its characteristics appeared to have changed.

It is now known that those changes in the characteristics were due to the charging of water vapour that had adsorbed to the gate oxide over the part of the channel that was not covered by the gate electrode. At the time the problem was very perplexing, and it took nearly four months until the actual semiconductor characteristics could be confidently and repeatedly measured. Details of the procedures tested are outlined in the following sections.

Characteristic Variation with Measurement.

Figure 5.21 contains a plot of the drain current versus gate voltage (the transconductance plot) for a one micron transistor with half the channel not covered by the gate on the drain side.

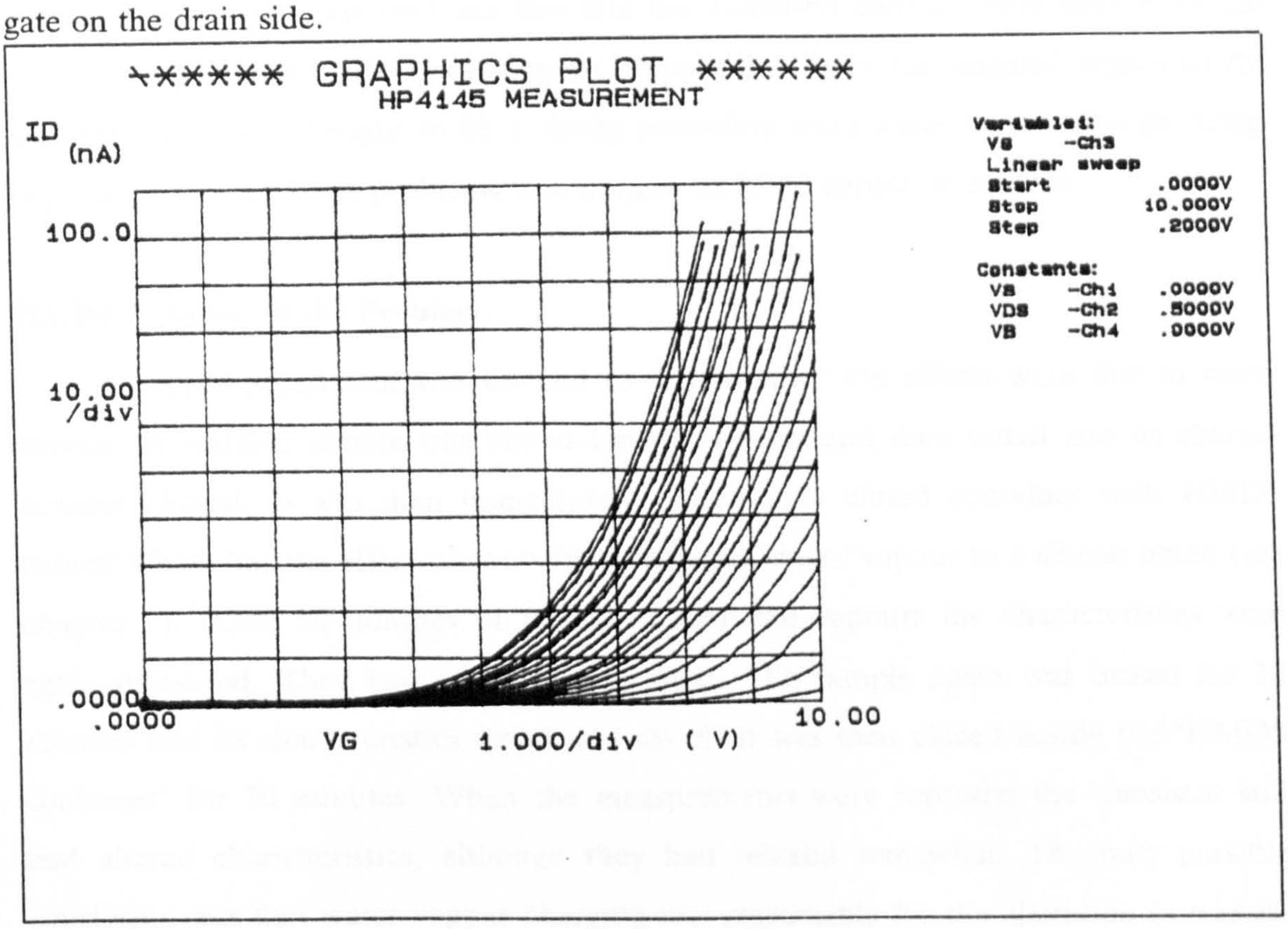


Figure 5.21, Twenty consecutive measurements of a incompletely gated IGFET.

The plot contains twenty consecutive measurements, with the newest being the one on the left. The transistor characteristics clearly were effected by the previous measurements. At first this was thought to be extreme sensitivity to hot electron effects.

The simple test of raising all the surface electrodes to the same bias for a fixed period ruled out hot electrons, since with all the surface electrodes at the same bias no current could have been induced in the transistor but the characteristics were still affected. It was also found that reversing the bias caused the opposite effect on the characteristics. This gave support to the theory that some form of oxide charge was present.

There was fear of mobile oxide charge, but the reference transistors did not exhibit any "fast" form of characteristic modification. The only difference between the test transistor and incompletely gated test transistors was what occurred above the gate oxide. The thought occurred one day that the increased current could have been due to the charged water vapour causing an electric field over the ungated region of the channel. This was thought to be a strong possibility since water vapour charge "fringing" had presented some problems to a project on MOS capacitor analysis.

HMDS Isolation of the Problem.

A simple experiment was devised to test whether the effects were due to water vapour. A suitable sample was biased for 30 minutes and then tested and its characteristics plotted. It was then immediately placed in a closed container with HMDS vapour which has the effect of converting adsorbed water vapour to a silicon oxide (see chapter 2). After 30 minutes of treatment with the vapours the characteristics were again measured. They had reverted to normal. The sample again was biased for 30 minutes and its characteristics tested and saved. It was then placed beside the "HMDS container" for 30 minutes. When the measurements were repeated the transistor still had altered characteristics, although they had relaxed somewhat. The only possible conclusion was that water vapour charging was responsible for the alteration in transistor characteristics.

Photoresist Passivation and Drying.

The first attempt at solving the water vapour problem was to cover the surface with a passivation layer. It was thought that photoresist should be used since it was easy to apply, and could be easily removed if it was not successful.

A mask level for the passivation "window" to allow contact to the pads was easily made by using the database to select all the pad rectangles by name and then copy them to the new level. The size of the rectangles was reduced and then a mask generated by the pattern generator.

A slightly thicker resist but otherwise standard photolithography cell was used to produce the resist passivation layer.

It was immediately evident upon electrical tests being made that the photoresist passivation had made the water vapour charging problem much worse. The reference transistors were again unchanged. It turned out that the photoresist contained a good deal of water. Further hot plate curing of the resist reduced the problem but it certainly did not solve it, as shown in figure 5.22.

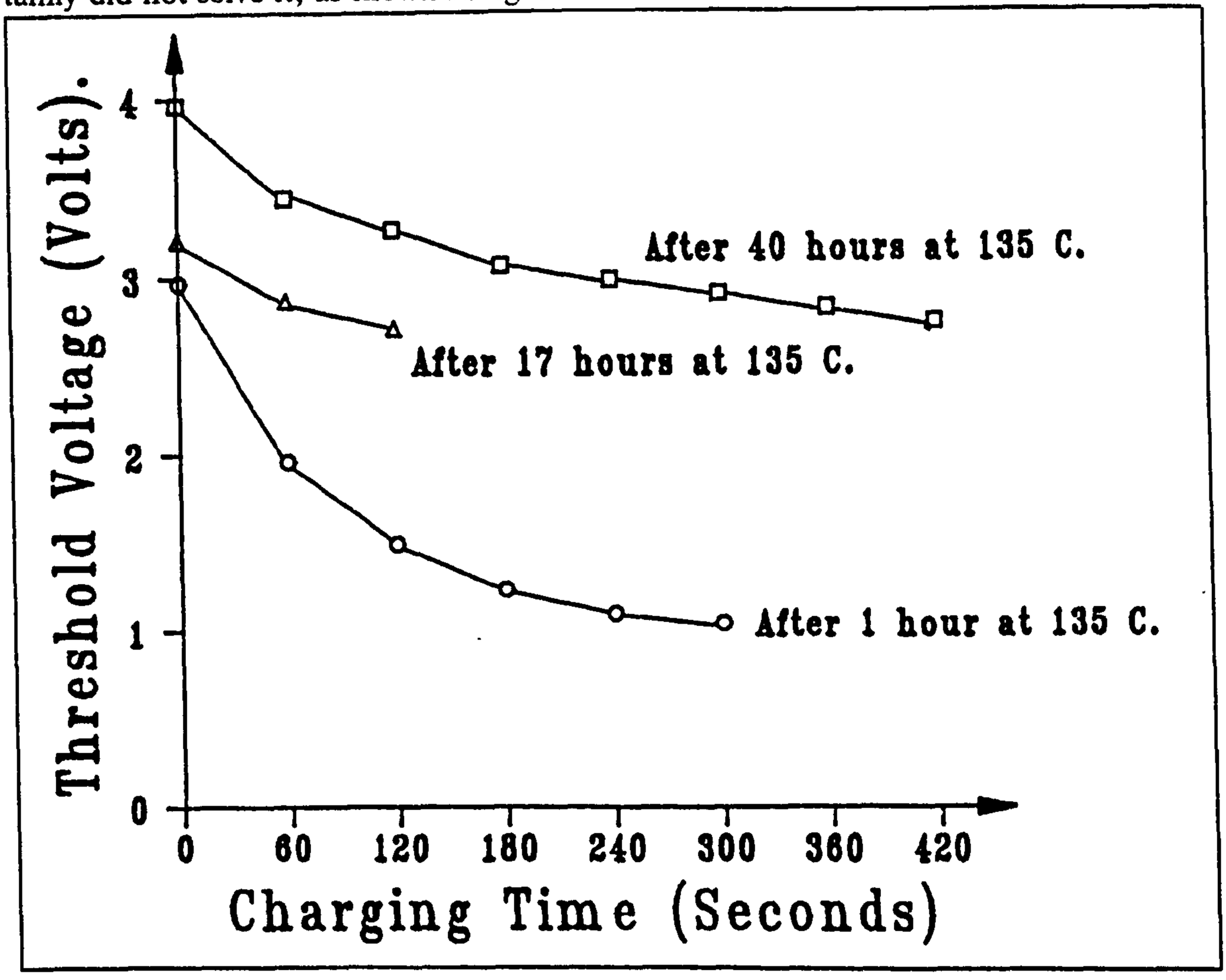


Figure 5.22, The instabilities in threshold voltage due to photoresist water charging.

Pyrolytic Oxide Passivation.

The next solution attempted was to use pyrolytic oxide as a passivation layer. It was applied using the standard production passivation layer process. The charging effect was dramatically reduced with a single measurement not making any perceivable change in the characteristics, but long term biases still did. Although the variations were small on almost completely gated transistors, drain gap transistors with large gaps were still affected too much for reliable characterisation.

Chopped Measurement Strategy.

The final, and effective, solution that was found was the management of the charge added to the water vapour. Earlier experiments had shown that the gate bias was predominately responsible for the charging and discharging of the water vapour. It followed that, if the charge added to the water vapour during a specific duration of a gate bias was removed by a similar duration of the corresponding negative bias, long term stability of the characteristics could be maintained. Figure 5.23 demonstrates that plan, which is similar to chopper stabilisation used for low drift instrumentation amplifiers. The measurement software that achieved this plan is described in the appendix.

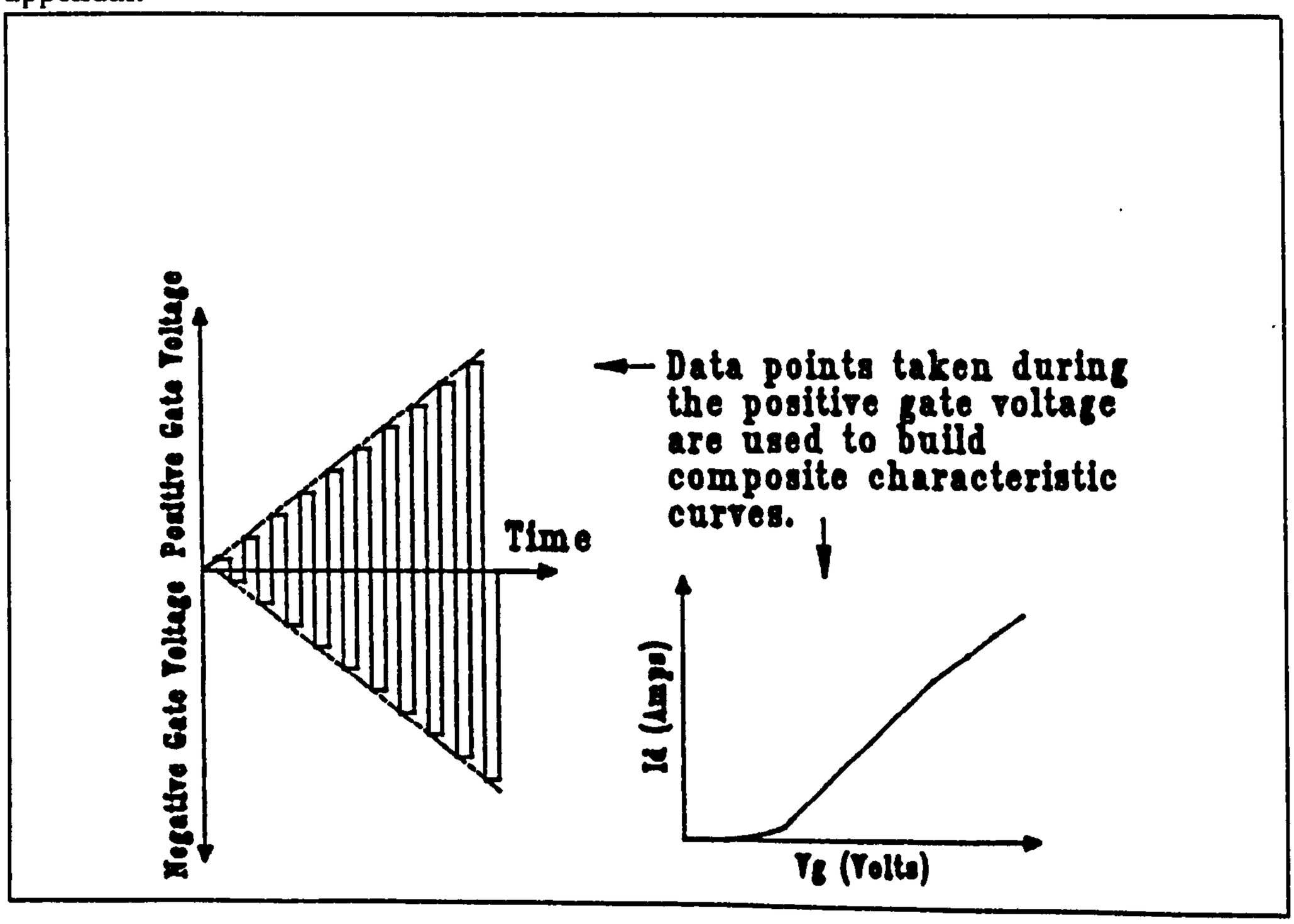


Figure 5.23, Chopper stabilisation of DGT and SGT characteristics.

The combined effect of the pyrox passivation and chopper stabilisation resulted in threshold voltage stability of a drain gap transistor over ten hours that was identical to the reference transistor stability. This enabled reliable characterisation of transistors with a gap in gate-to-channel coverage, and allowed the research into edge effects to continue.

5.6. Electrical Alignment Technique.

In the previous chapter it was said that some form of electrical alignment technique was necessary in the progressional offset technique in order to locate the intended test structure after the various processing effects altered its position in the array. Since the intended structure in this case was an IGFET, a number of electrical tests were performed which could have assessed the alignment of the gate to the channel.

Manually it is easy to differentiate between the curves, but any electrical alignment technique would have had to be able to make quantitative distinctions between the curves. It is important that an electrical alignment technique is accurate, but speed is also a consideration since every wafer contained nearly 13000 transistors. The evaluated techniques are outlined in the following sections.

V_D Threshold Measurement.

One possible technique for detecting alignment would have been to use the drain gapped transistor's drain bias threshold to detect the magnitude of the gap. In order to do so the gate was biased above the threshold voltage for inversion of the channel. Then the drain bias was increased and in so doing the drain depletion region was increased. Once the drain depletion region met the channel inversion region, current would have flown.

In simple theory an abrupt "turn on" at some drain voltage should have occurred, but in practice a gradual increase in current with increasing bias is encountered. The similar problem with the gate threshold voltage extraction is usually solved by taking the intercept of a tangent to the transconductance curve as the threshold voltage, but the drain current-voltage curves for a DGT do not have a suitable shape for that solution. Another method would be to define the drain threshold voltage as the magnitude at which a certain current flows. Table 5.5 shows the results for a column in the progressional array. It can be seen that it was impossible to use a single current definition for the whole array, and therefore the "drain threshold voltage" alignment technique was not a useful tool.

Transistor		Drain Threshold Voltage for Current Level		
Column Position	Gap Size	120 μA (volts)	60 μA (volts)	5 μA (volts)
1	0.65 μm DGT	-	-	4.59
2	0.50 μm DGT	-	4.18	1.42
3	0.35 μm DGT	4.08	2.34	0.71
4	0.20 μm DGT	3.46	0.91	0.00
5	0.05 μm DGT	3.26	0.31	0.00
6	Aligned	3.28	0.20	0.00
7	0.15 μm SGT	-	-	0.00
8	0.30 μm SGT	-	-	0.10
9	0.45 μm SGT	-	-	-

Table 5.5, Drain Voltage Threshold as an alignment technique.

V_G Threshold and Transconductance Measurements.

The channel region of the IGFET that is not covered by the gate in the gapped transistors can be inverted by the electric field fringe from the gate electrode. As a result the gate threshold voltage to achieve channel inversion increases as the gap size increases because the fringing field must be larger to affect the semiconductor farther away from the gate electrode. That phenomenon was tried as an electrical alignment technique with some success. Table 5.6 shows the threshold voltage and transconductance for the same column as in table 5.5. The transconductance is also strongly affected by the electric field magnitude in the "gapped" region, although its extraction can be effected by noise in the measurement since a numerical differentiation is required in its extraction.

Column Position	Gap Size	Transconductance ($\mu A/V$)	Threshold Voltage (Volts)
1	0.65 μm DGT	0.083	4.35
2	0.50 μm DGT	3.65	3.88
3	0.35 μm DGT	13.9	3.19
4	0.20 μm DGT	12.7	1.78
5	0.05 μm DGT	15.3	1.14
6	Aligned	16.9	1.11
7	0.15 μm SGT	12.2	1.54
8	0.30 μm SGT	13.6	1.76
9	0.45 μm SGT	9.9	3.06

Table 5.6, Gate threshold voltage as an alignment technique.

It would appear from that information that the threshold voltage method of electrical alignment can detect the aligned device, but data from another column presented in table 5.7 shows that it was really a subjective measurement since it was unable to determine that all the transistors in the second column were gapped.† Therefore, although the threshold voltage method of alignment detection was better than the drain voltage threshold technique, it still did not have the sensitivity required for detecting slight gaps in gate-to-channel coverage.

---

† The "true" alignment was determined by the source-drain reversed subthreshold curves.

Column Position	Gap Size	Transconductance ($\mu A/V$)	Threshold Voltage (Volts)
1	grossly DGT	0.007	4.34
2	larger DGT	3.11	3.88
3	large DGT	13.6	3.50
4	DGT	17.4	2.13
5	slight DGT	18.2	1.15
6	slight SGT	16.1	1.17
7	SGT	14.1	2.20
8	large SGT	13.1	2.27
9	largest SGT	14.4	3.29

Table 5.7, Gate threshold voltage as an alignment technique.

Gate Capacitance.

Although the other researchers had some success with gate-drain capacitance measurement for evaluating the degree of gate to drain overlap,[6] the capacitance magnitude in this experiment would be only 0.5 fF for the smallest overlaps. The effectiveness of the capacitance alignment technique is limited not only by the small capacitance measurement resolution, but also by effects due to the gate fringing capacitance, other parasitic capacitances, and its inability to detect gapped transistors. It was tried, but the parasitic capacitances prevented any meaningful results from being obtained.

5.6.1. Source-Drain Reversal.

The thought that it was symmetry that differentiated transistors in the progressional offset array, lead to the realisation that if the source and drain terminals of a gapped transistor are interchanged the electrical characteristics would change, but if an aligned transistor had its source and drain reversed its characteristics would remain the same.

Curve Correlations.

The first approach that was taken in comparison of source-drain reversal characteristics was to use statistical methods to compare the transconductance curves and to compare the drain current-voltage curves at some gate voltage.

Two types of comparisons were tried. The correlation coefficient between curves compares the differences between the means of the points in the two curves against the standard deviation of the points in each of the curves. The result, which is usually given the symbol R, was a number that could take a value from 1 to −1, with 1 being a highly correlated curve, 0 being uncorrelated, and −1 being inversely correlated.

The second statistical test applied was the variance of means test. The comparison was taken at each point along the curve by averaging the four surrounding points to get a mean and standard deviation. The means of the two curves were then compared to determine if they were within a combined standard deviation of each other. The result, which is usually labelled F, can take on any value larger than 0. The smaller the number the more likely it was that the two curves were identical.

Table 5.8 shows the results of using the two methods to compare both the transconductance and conductance curves.

Transistor		Gate Characteristic		Drain Characteristic	
Column Position	Gap Size	R	F	R	F
1	0.65 μm DGT	0.9964	119	0.5710	7901
2	0.50 μm DGT	0.9397	887	0.7801	11891
3	0.35 μm DGT	0.9363	887	0.8240	21696
4	0.20 μm DGT	0.9937	928	0.8855	24099
5	0.05 μm DGT	0.9994	158	0.9858	18894
6	Aligned	0.9999	1	0.9992	3717
7	0.15 μm SGT	0.9996	5	0.8543	10115
8	0.30 μm SGT	0.9983	27	0.7912	10662
9	0.45 μm SGT	0.9997	8	0.7276	10469

Table 5.8, Statistical Comparisons between Source-Drain Swapped Characteristics

The correlation coefficients for the gate characteristic curves are all nearly 1.0 because the forward and source-drain-interchanged curves are very similar when there is only a small drain bias. The variance of means test also produces low values for variation between the two configurations for similar reasons. It is clear that comparisons of the gate characteristic curves for the two configurations would not provide a reliable method for detecting gate-to-channel alignment.

Both the correlation coefficient and variance of means test worked better in detecting gate-to-channel mis-alignment when comparing the drain characteristic curves for the two configurations. That was because the drain depletion widening, with increased drain bias, reduced the effect of the gap in gate-to-channel coverage on DGT characteristics but not on SGT characteristics. The drain characteristic comparisons detected the aligned site, but the resolution of the correlation coefficient is poor near the aligned site, and the variance of means test also indicates alignment for badly misaligned sites because of the curves similarity (all noise).

A combination of the two tests might have been successful, but the tests also required a significant amount of computation. Each curve comparison required approximately 10 seconds. If only one column on each die on one wafer were to be analysed, the operation would still have taken 2.7 hours after the measurements were complete. A simpler test was clearly required.

Subthreshold Curves.

The subthreshold current was found to be strongly dependent on the gate electric field at the source side of the channel. Any gap in gate-to-channel coverage near the source-channel junction caused an increase in the gate voltage swing required to reduce the current by one decade. Figures 5.24, 5.25, and 5.26 show the effect on the subthreshold curves caused by reversing the source and drain connections for a source gap transistor (SGT), normally gated transistor (NGT) and for a drain gap transistor (DGT) respectively.

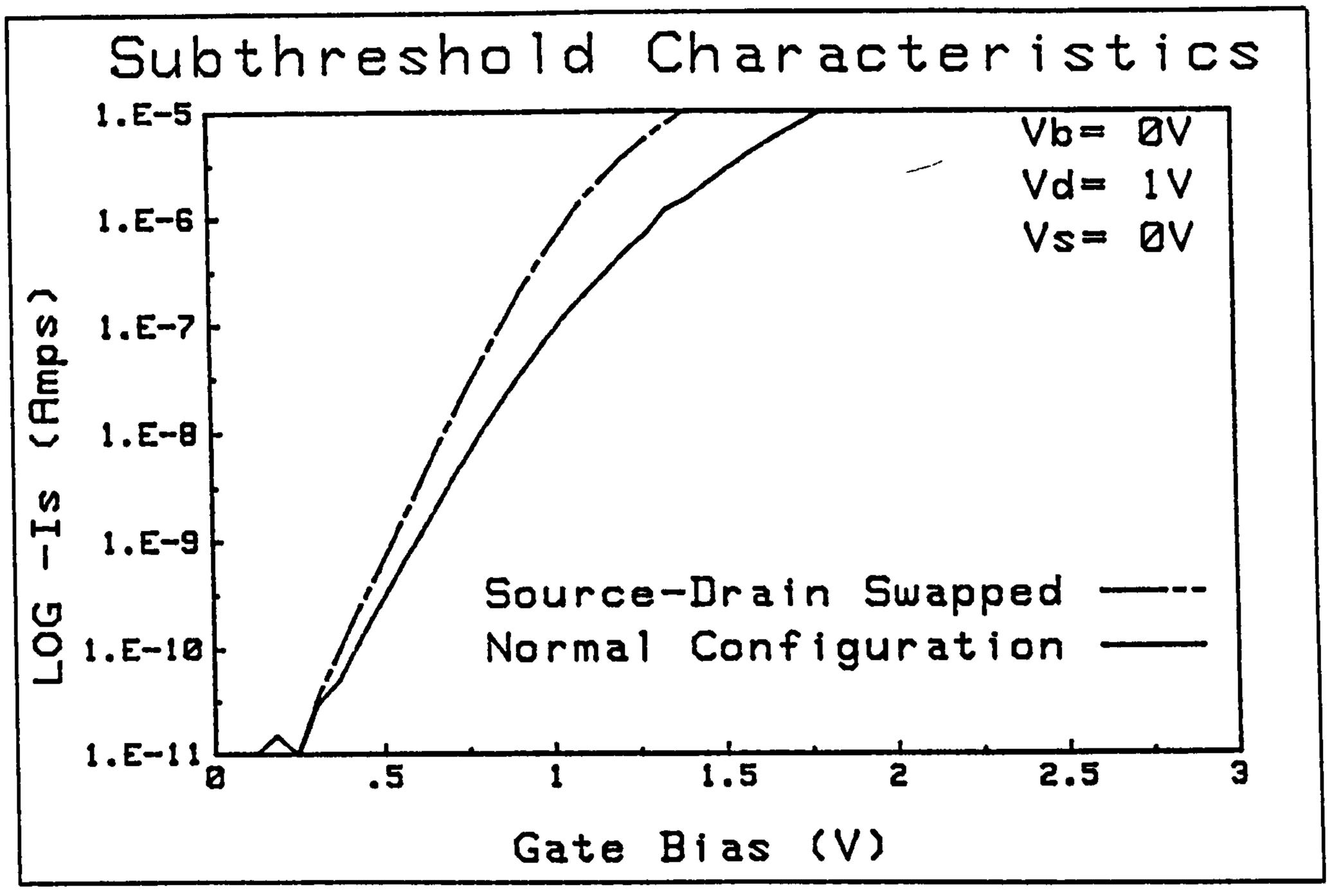


Figure 5.24, Forward and Source-Drain Reversed Subthreshold Characteristics for an SGT.

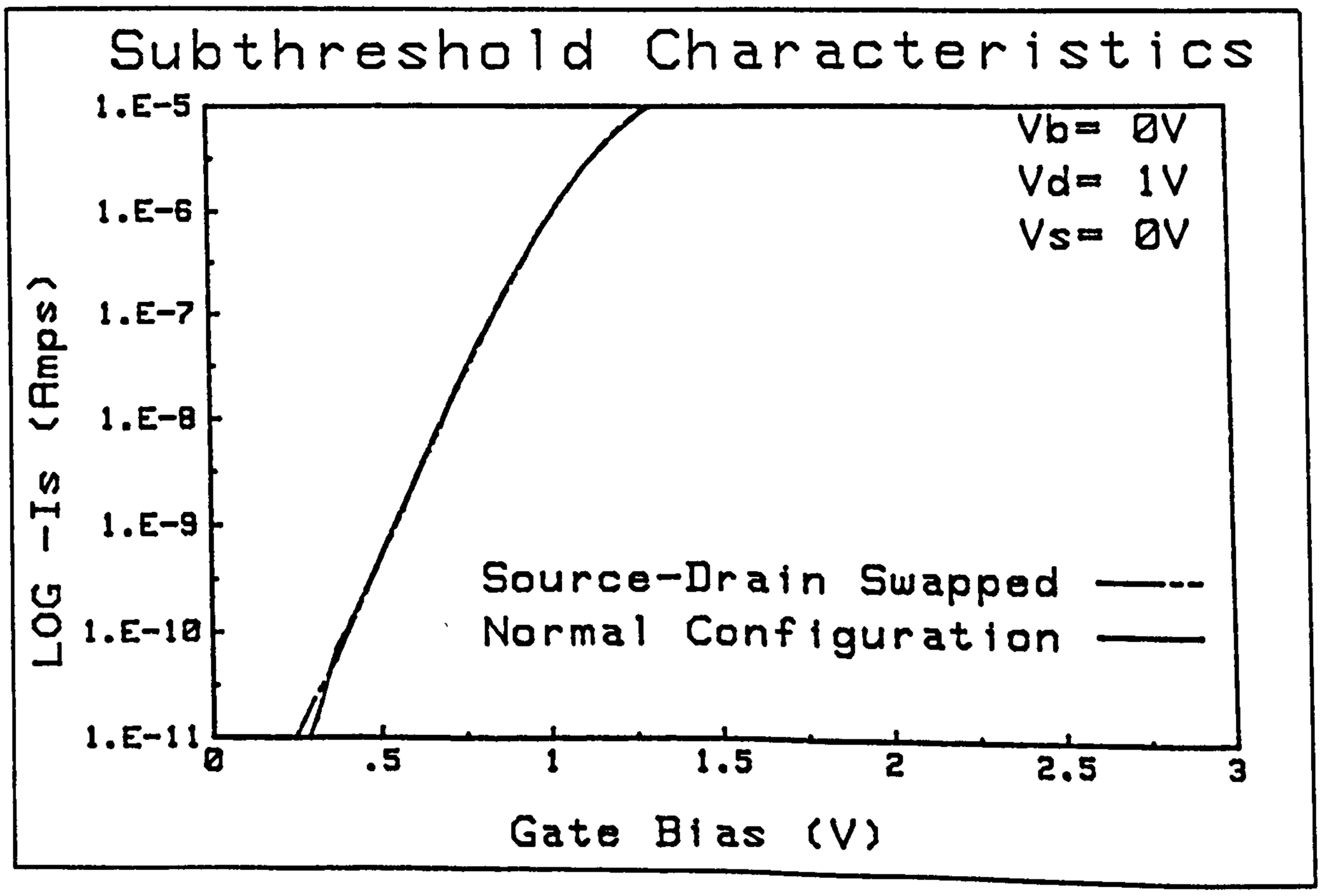


Figure 5.25, Forward and Source-Drain Reversed Subthreshold Characteristics for an NGT.

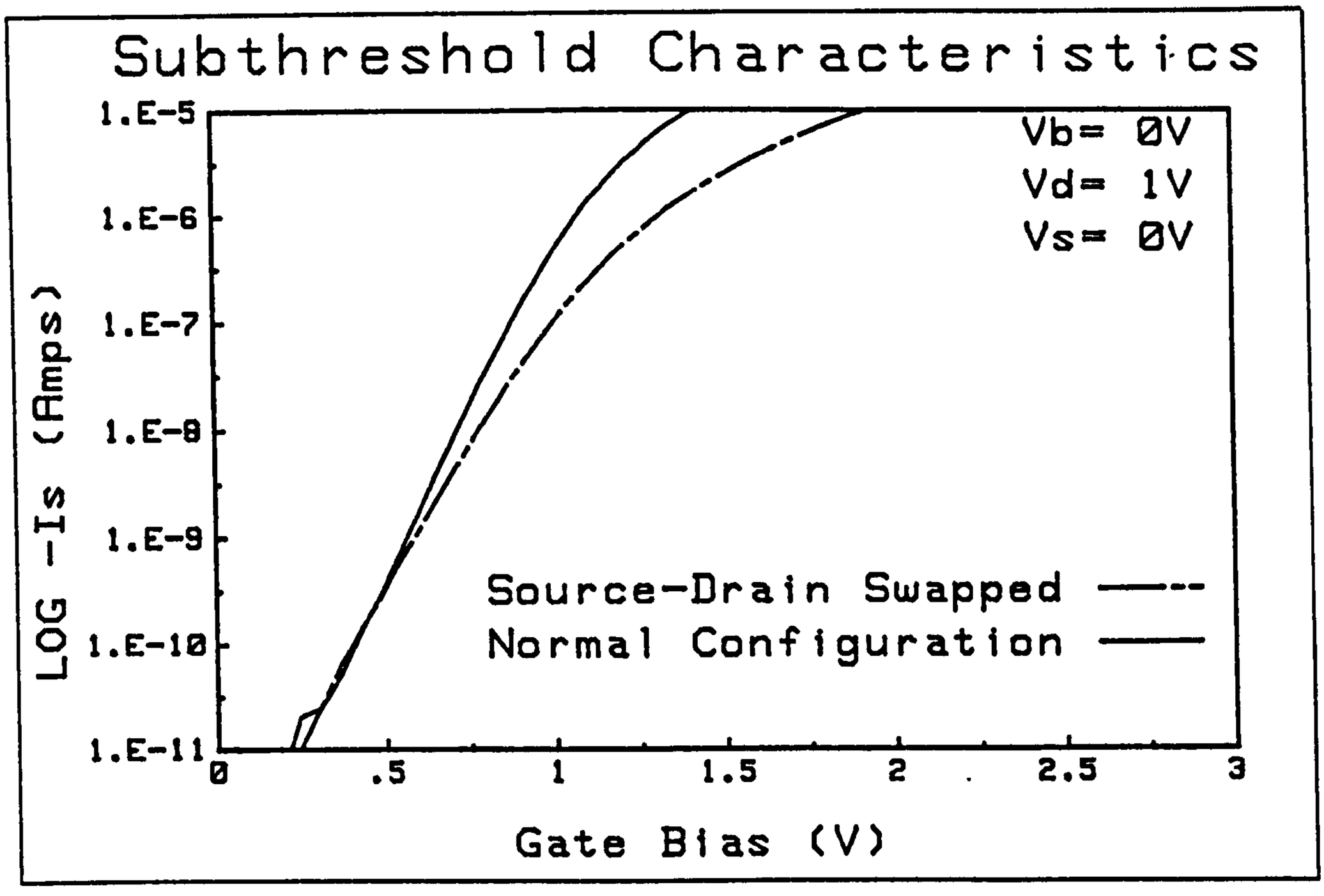


Figure 5.26, Forward and Source-Drain Reversed Subthreshold Characteristics for an DGT.

The opportunity for a simple determination of alignment was clearly available in the subthreshold curves. If the difference between the gate voltage required to reach a specific current in the forward configuration and the gate voltage to reach the same current with the source and drain interchanged was taken, a numerical indication of alignment resulted.

Not only was the alignment technique fast, the result had a sign with physical significance. When the difference ($V_{interchanged} - V_{forward}$) was used, the source gap transistors had a positive difference, the drain gap transistors a negative difference, and the normally gated transistors no difference.

Table 5.9 contains the results for the example column used previously. Note how the voltage difference is related to the gap magnitude including the sign change indicating the shift in gap position from the drain to the source side of the channel. Table 5.10 contains the results for the same column as in table 5.7, which confused the threshold voltage alignment technique.

Column Position	Gap Size	$V_{interchanged} - V_{forward}$ (Volts) for 1 μA.
1	0.65 μm DGT	-
2	0.50 μm DGT	0.73
3	0.35 μm DGT	1.28
4	0.20 μm DGT	0.23
5	0.05 μm DGT	0.06
6	Aligned	0
7	0.15 μm SGT	−0.24
8	0.30 μm SGT	−0.37
9	0.45 μm SGT	−0.42

Table 5.9, Subthreshold Curve Voltage Difference as an alignment technique.

Column Position	Gap Size	$V_{interchanged} - V_{forward}$ (Volts) for 1μA
1	grossly DGT	-
2	larger DGT	0.86
3	large DGT	1.34
4	DGT	0.79
5	slight DGT	0.06
6	slight SGT	−0.12
7	SGT	−0.31
8	large SGT	−0.43
9	largest SGT	−0.67

Table 5.10, Subthreshold Curve Voltage Difference as an alignment technique.

Both tables 5.9 and 5.10 show a decrease in "gap size" according to the alignment technique at column position 2 and do not register a difference for column position 1. The reason behind this is that although the source-channel junction was covered by a gate, the gap in coverage on the drain side of the channel was so large that there was

no efficient "collector" for the charge, so both the forward and source-drain interchange subthreshold characteristics were affected. The subthreshold curves for transistors with smaller gaps were affected less by the gap on the drain side. The "threshold" gap size for a drain gap to effect the subthreshold characteristics is dependent on the drain depletion width, and is a topic of a later chapter. The values presented in the table were measured using a drain bias of 1.0 volt.

The comparison of forward versus source-drain interchanged subthreshold curves as an alignment technique can be used to analyse the symmetry of isolated transistors. The transistor does not need to be part of a progressional alignment array to use this technique, therefore it could be used as a test structure to check sidewall spacer processes.

5.7. Golden Die Selection.

With an accurate method of electrically detecting an aligned transistor it was possible to quickly scan an entire wafer looking for a die that had transistor structures as they were intended in the design. Such "perfect" devices in industry are often called "golden standards".

The forward and source-drain interchanged subthreshold current curves were measured and the voltage difference between them was calculated and stored for each transistor in the same column on every die on a wafer.

The resulting array of 186 groups of nine differences was then available for computer selection of likely candidates. One requirement of the "golden die" was that it included both positive (DGTs) and negative (SGTs) voltage differences, which is essentially the cross-over confirmation mentioned in the last chapter. The second requirement was that at least one die near the center of the column must have had a zero voltage difference (NGT).

When a suitable selection algorithm was run on the data arrays, 30 of the 186 die had such characteristics. They were mostly clustered in a central region of the wafer, which is what one would expect from etching variations. The rejection of many of the other sites was due to a missing zero point, which indicates that the progressional offset step size should have been slightly smaller, or the designed margin for gate to source-

drain overlap should have been larger in order to have more "complete" columns.

The 30 die were then subject to further selection to exclude columns with two completely gated transistors, and those with the "zero point" too close to either end of the array. Four "golden die" locations resulted from that analysis, which then became the experimental transistors for the results in the following chapters.

5.8. Physical Verification.

Two destructive and one non-destructive measurement techniques were used to attempt to provide physical verification of the electrical alignment technique.

Voltage Contrast SEM.

The first technique that was tried was to cleave through an entire column of transistors on one die, a difficult feat in itself. After sufficient practice, or perhaps luck, a cleaved column on a die was available and had bonding wires connected between the appropriate pads and chip package. The cleaved die was then mounted on its edge exposing the cleaved edge and glued in place with electrically conducting epoxy. The whole assembly was then placed in a SEM capable of voltage contrast imaging. Biases were applied to the source and drain in order to highlight them so their position relative to the gate could have been measured. Unfortunately the entire bulk region "lit" up in the image.

Possible reasons for the failure were high bulk contact resistance, or, more likely, that the cleaved surface became inverted under the influence of the SEM beam and leaked current between the junctions.

Cut and Stain SEM.

Another destructive test aimed at confirming the electrical alignment was a cut and stain technique to delineate the source and drain junctions. A sample was cleaved through a column of transistors and the cleaved edge was then subjected to a silicon etch. The doped silicon etched faster than the undoped silicon to reveal the source and drain junctions. Unfortunately on examination in the SEM it was found that the soft aluminium gate was "necked" by the cleaving operation and its position relative to the

channel was altered.

Another attempt was made using a lapping machine to section the transistors in one column, but it too was foiled by the aluminium "smearing" over the cross-section. Attempts to protect the aluminium with photoresist were also tried, but to no avail. The conclusion was that verification through cross-sectional examination would have to wait for a polysilicon gate experiment, since the polysilicon would be less ductile and would allow cleaving or lapping to be used successfully.

Optical Shear Calibration.

Figure 5.27 shows the result of non-destructive optical shearing microscope gate position measurements for a golden die. The relative positions shown were determined primarily by optical shearing microscope data, but also by electrical channel length calculations, and through subthreshold curve alignment information considering the die as a whole.

The channel length calculation used the standard technique of measuring the transconductance for transistors of differing drawn lengths to obtain "delta L", the total sideways diffusion, which could then be used to calculate the channel length.

Then the optical shear microscope was used to obtain five measurements of each gate width, and five measurements of the position of each gate relative to the edge of the active area. The averages and standard deviations were calculated and combined to give the composite error bars that are shown for each transistor.

The relative position of the channels to the group of gates was established from the subthreshold current alignment procedure and the following logic. Site 5,5 was known to exhibit drain gapped transistor characteristics, so the worst case error bar for the gate length must have been to the right of the channel edge in the diagram. Site 5,7 was known to exhibit source gapped transistor characteristics, so the worst case error bar for its gate length must have been to the left of the other side of the channel. This resulted in only one possible position for the gates relative to the channel, which was confirmed by the average position of the gate in site 5,6 agreeing with the observed normally gated transistor characteristics.

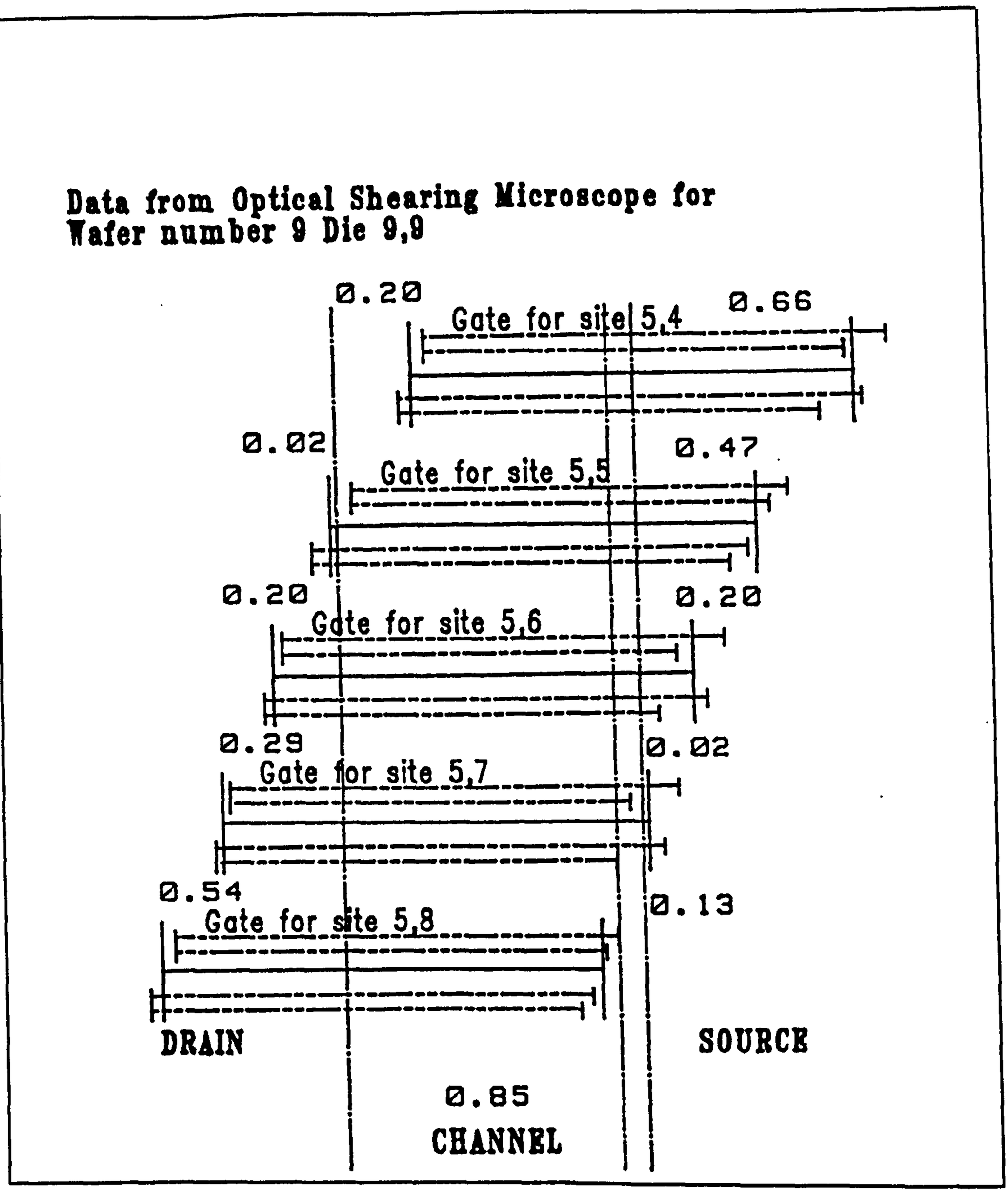


Figure 5.27, Physical Confirmation of Gate-Channel Alignment.

The gathering of those measurements and subsequent analysis took seven hours for that single die. The results however, were crucial to the success of the drain depletion motion study described in chapter seven, and are useful in the next chapter which presents the electrical characteristics of that die.

5.9. Chapter Summary.

In this chapter the background and implementation of a mis-aligned gate experiment using the progressional offset technique was described. The design of a progressional offset array of transistors, with gaps in gate-to-channel coverage, using an object oriented database was presented. The custom non-self-aligned metal gate NMOS process designed to fabricate the array was outlined. A fabrication anomaly and its use in validating the progressional offset technique was detailed. The problem of water vapour charging and the measurement technique of chopper stablisation to reduce its effect on incompletely gated transistor characteristics were reported. Four techniques for electrically detecting the mis-alignment of the gate to the channel were presented. The normal and source-drain-interchanged subthreshold curve comparison was concluded to be the best technique, and one which was ideally suited to an automated scan of a wafer. Finally three methods of physical verification of the degree of gate-to-channel mis-alignment were described.

5.10. References.

1. Ko, P.K., Chan, T.Y., Wu, A.T., and Hu, C., "The Effects of Weak Gate-To-Drain (Source) Overlap on MOSFET Characteristics (Invited Paper).," *IEDM 86*, pp. 292-295, IEEE, 1986.

2. Chan, T.Y., Wu, A.T., Ko, P.K., Hu, C., and Razouk, R.R.;, "Asymmetrical Characteristics in LDD and Minimum-Overlap MOSFET's.," *IEEE Electron Device Letters*, vol. EDL-7, no. 1, pp. 16-19, January 1986..

3. Sun, Y.C., Wordman, M.R., and Laux, S.E., "On the Accuracy of Channel Length Characterization of LDD MOSFET's.," *IEEE Transactions on Electron Devices.*, vol. ED-33, no. 10, pp. 1556-1562, October 1986.

4. Ng, K.K. and Lynch W.T., "Analysis of the Gate-Voltage-Dependent Series Resistance of MOSFET's.," *IEEE Transactions on Electron Devices.*, vol. ED-33, no. 7, pp. 965-972, July 1986..

5. Mikoshiba, H., Horiuchi, T., and Hamano, K., "Comparison of Drain Structures in n-Channel MOSFET's.," *IEEE Transactions on Electron Devices.*, vol. ED-33, no. 1, pp. 140-144, Jan. 1986.

6. Chan, T.Y., Wu, A.T., Ko, P.K., and Hu, C., "A Capacitance Method to Determine the Gate-to-Drain/Source Overlap Length of MOSFET's.," *IEEE Electron Device Letters*, vol. EDL-8, no. 6, pp. 269-271, June 1987.

7. Ng, K.K. and Lynch W.T., "The Impact of Intrinsic Series Resistance on MOSFET Scaling.," *IEEE Transactions on Electron Devices.*, vol. ED-34, no. 3, pp. 503-511, March 1987..

8. Hu, G.J., Chang, C., and Chia Y., "Gate-Voltage-Dependent Effective Channel Length and Series Resistance of LDD MOSFET's.," *IEEE Transactions on Electron Devices.*, vol. ED-34, no. 12, pp. 2469-2475, December 1987..

9. Doyle, B.S., Bourcerie, M., Leclaire, P., and Boudou, A., "Hot Electron Degradation in the Source of Asymmetrical LDD Structures.," *Electronics Letters.*, vol. 23, no. 25, pp. 1356-1357, December 3, 1987..

10. Dars, P., D'Ooville, T., and Mingam, H., *Statistical Analysis of Implant Angles Effects on Asymmetrical NMOSFETs Characteristics and Reliability.*, September 1988. To be presented at ESSDERC'88

11. Yamaguchi, K., Takahashi, S., and Kodera, H., "Theoretical Study of a Channel-Doped Separate Gate Si MOSFET (SG-MOSFET) by Two-Dimensional Computer Simulation.," *IEEE Transactions on Electron Devices.*, vol. ED-28, no. 1, pp. 117-120, January 1981..

12. Yamaguchi, K. and Takahashi, S., "Submicron Gate MOSFET's with Channel-Doped Separate Gate Structure (SG-MOSFET's)," *IEEE Transactions on Electron Devices.*, vol. ED-28, no. 7, pp. 888-890, July 1981..

Chapter 6,
Experimental Asymmetric Transistor Characteristics.

6.1. Introduction.

The previous chapter described a method of selecting a die with the intended progressional offset array. The subthreshold symmetry method was used on the 1.3 μm drawn length transistors on wafer 9. As a result the die located at column 9, row 9, was chosen as the "golden die" for further analysis and presentation.[1,2]

Although the subthreshold symmetry method has reduced the data to be considered, further reduction was required to reach a manageable amount for presentation. Only the electrical characteristics of the transistors in a typical column, containing only one transistor length, will be presented in their "raw" format, but data from other columns will be used in the deeper analyses. The three characteristics plots, subthreshold, gate and drain characteristics, for the 72 test transistors would otherwise require the rest of this volume and part of another to present.

A short hand method of referring to a particular transistor in a batch has been used throughout this research and is worth documenting here. There were five components to the description, the wafer number (1-10), the column position (1-14) and row position (1-14) of the die, and the "xsite" (1-12) and "ysite" (1-9) position of the transistor on the die. In the software these components were kept separate, but a concise method was used for human presentation. Wafer 9 die 9,9 and transistor 11,5 would have been written as 999115.†

6.2. Reference Transistors.

There were 36 reference transistors on each die to act as controls against processing variations. Each reference transistor had a drawn width of 18 μm and length of 5 μm. The subthreshold, gate and drain characteristics for a reference transistor (99946) at the center of the "golden die" are shown in the three plots in figure 6.1.

---

† Although there appears to be the potential for ambiguity in that representation, a unique transistor reference always results. The best way to decode this representation is reading right to left and using logic based on the maximum value of each parameter to delineate between parameters.

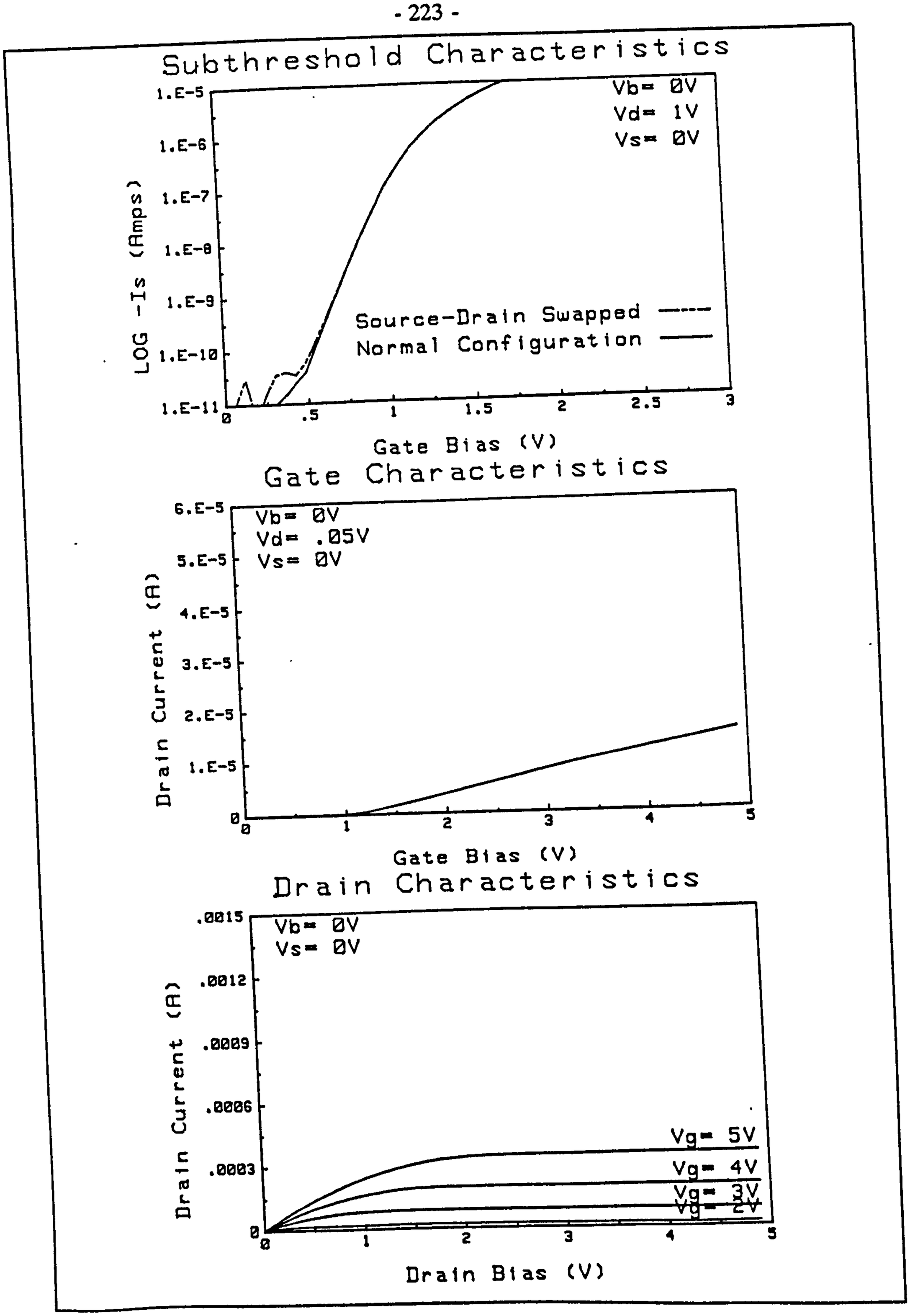


Figure 6.1, Typical characteristics of a reference transistor.

The scale of these plots may appear inappropriate for the reference transistor, but a single consistent format has been used for all the transistors presented.

The unremarkability of these plots is in itself a confirmation that the microfabrication process was successful. The plots presented for the reference transistor are more than just typical of the remainder on the die, the remainder appear identical to it. Deeper analysis through extraction of the threshold voltage, transconductance, linear resistance, and a particular saturation current confirm that there is little process variation across the die. Table 6.1 presents the results for the 36 reference transistors.

Parameter	Mean	Standard Deviation	Percent Variation
Threshold Voltage (V)	1.169	± 0.010	0.86
Transconductance (μS)	4.41	± 0.04	0.91
$I_{D_{sat}}$ $V_G = 5V, V_D = 4V$ (μA)	345	±4.5	0.13
R_{linear} (Ohms)	3502	± 36	1.03

Table 6.1, The consistency of reference transistor across a die.

This indicates that variations in the test transistors are in fact due to the experimental variable of the degree of gate-to-channel coverage.

6.3. Test Transistors.

As mentioned previously there were eight columns of nine test transistors on each die. They were of fixed drawn channel width of 18 μm. The drawn channel lengths were increased along a row from a minimum of 1.00 μm to a maximum of 2.05 μm in steps of 0.15 μm. The drawn alignment of the gate to the channel was shifted from a drain gap of 0.60 μm to a source gap of 0.60 μm down each column in nine steps of 0.15 μm. The results for those transistors are presented here.

6.3.1. Alignment Variations.

Figure 6.2 presents a subjective view of the relative alignment of each column of the die biased on the subthreshold symmetry method of alignment. The column numbering in that figure is such that the shortest gate length is at the top (9992) and

Weak Overlap

Column	← DGT		NGT		SGT →
9992	4	5	6		7
9993	4	5	6		7
9995	4 5		6		7 8
9996	4	5	6	7	8 9
9998	4	5	6	7	8
9999	4	5	6	7	8
99911	4	5	6	7	8 9
99912	4	5	6	7	8 9

Figure 6.2, Alignment across the geometry range on a die.

the longest at the bottom (99912). The numbers across the diagram relate the transistors row position to its gate-to-channel alignment. The column that was used for the selection of the golden die was column 9995 which has the intended progression from DGTs with decreasing gap size through normally gated transistors to SGTs of increasing gap size. One would have expected from the progressional offset theory that the rest of the columns would have a similar pattern. Unfortunately there is a variation of edge position with line width which was not considered in the progressional offset theory. The reason behind this is due to the resolution of the projection lithography system. Although the system is capable of patterning smaller features than those used here, "biases" to increase the mask widths to compensate for resist sidewall notching and other exposure effects must be made. Since no "biasing" of the features was done for this experiment, the smaller gate lengths are also smaller relative to their masking feature than the larger gate lengths. That results in the normally gated transistors covering three sites in the large channel length columns and fewer in the shorter channel

length columns.

The effect is not particularly damaging to the experiment since most comparisons are between transistors in the same column and therefore of the same channel length. It does, however, increase the experimental error in the value of the channel length which is used in some of the following extractions.

6.3.2. Characteristics of a Typical Column.

The column containing transistors with a drawn length of 1.9 μm is presented here because it has all the features found in the other columns. The other columns have slightly different characteristics because of the difference in their channel lengths, but those differences are best presented through the deeper analysis following the "raw" characteristics of the example column. Figures 6.3 to 6.11 show the subthreshold, i_D-v_G, and i_D-v_D characteristics for column 99911.

Subthreshold.

The gap size has a different effect on the subthreshold current for DGTs than for SGTs and the alignment technique calibration is based on this effect. The normal configuration curve from figure 6.3 through to figure 6.11 show the effect of the gap size. Firstly transistors with a large gap on the drain side, as in figure 6.3, have reduced subthreshold current because of the increased series resistance on the drain side of the channel. As the gap is reduced, figure 6.4 and 6.5, the subthreshold current injected at the source is collected more and more effectively by the drain and the "swing" is reduced. Once the size of the drain gap in gate-to-channel coverage is smaller than the drain depletion width, the subthreshold current has a similar curve to that of a completely gated transistor. The subthreshold curves for transistors with variations in the amount of gate overlap of the source and drain (figures 6.7, 6.8, and 6.9) appear identical. Once the gap occurs on the source side of the channel reductions in subthreshold current again result. The reduction is due to increased series resistance and a reduction in injection efficiency due to the weaker gate electric field at the source-to-channel interface. Figure 6.12 summarises the effect of gap size on the subthreshold behaviour. The SGT subthreshold characteristics are not affected by the drain bias as the DGT devices are. This effect is treated in more detail in chapter seven.

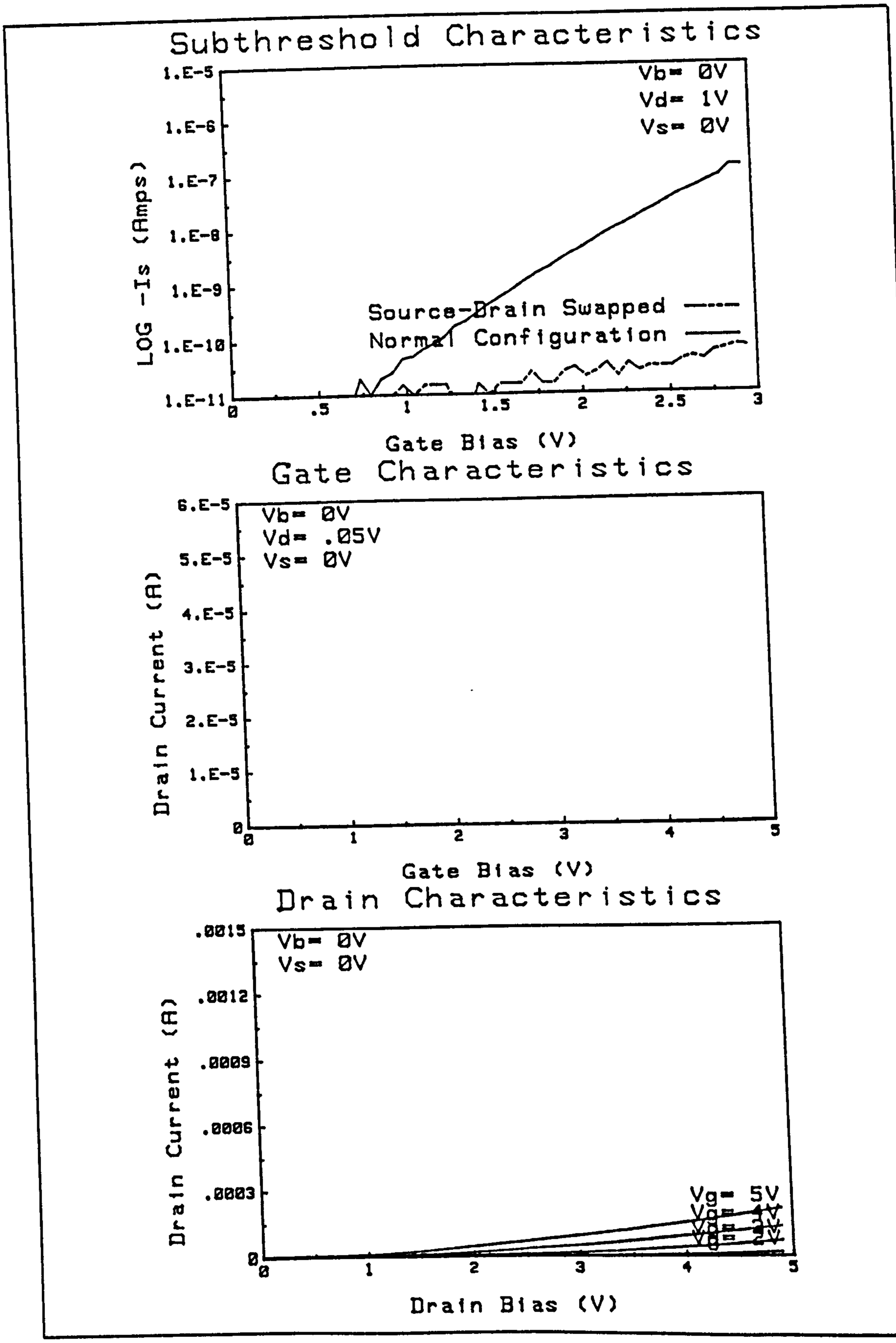


Figure 6.3, Experimental Transistor 999111, Drawn DGT $L_D = 1.9\mu m$ $Gap = 0.60\mu m$.

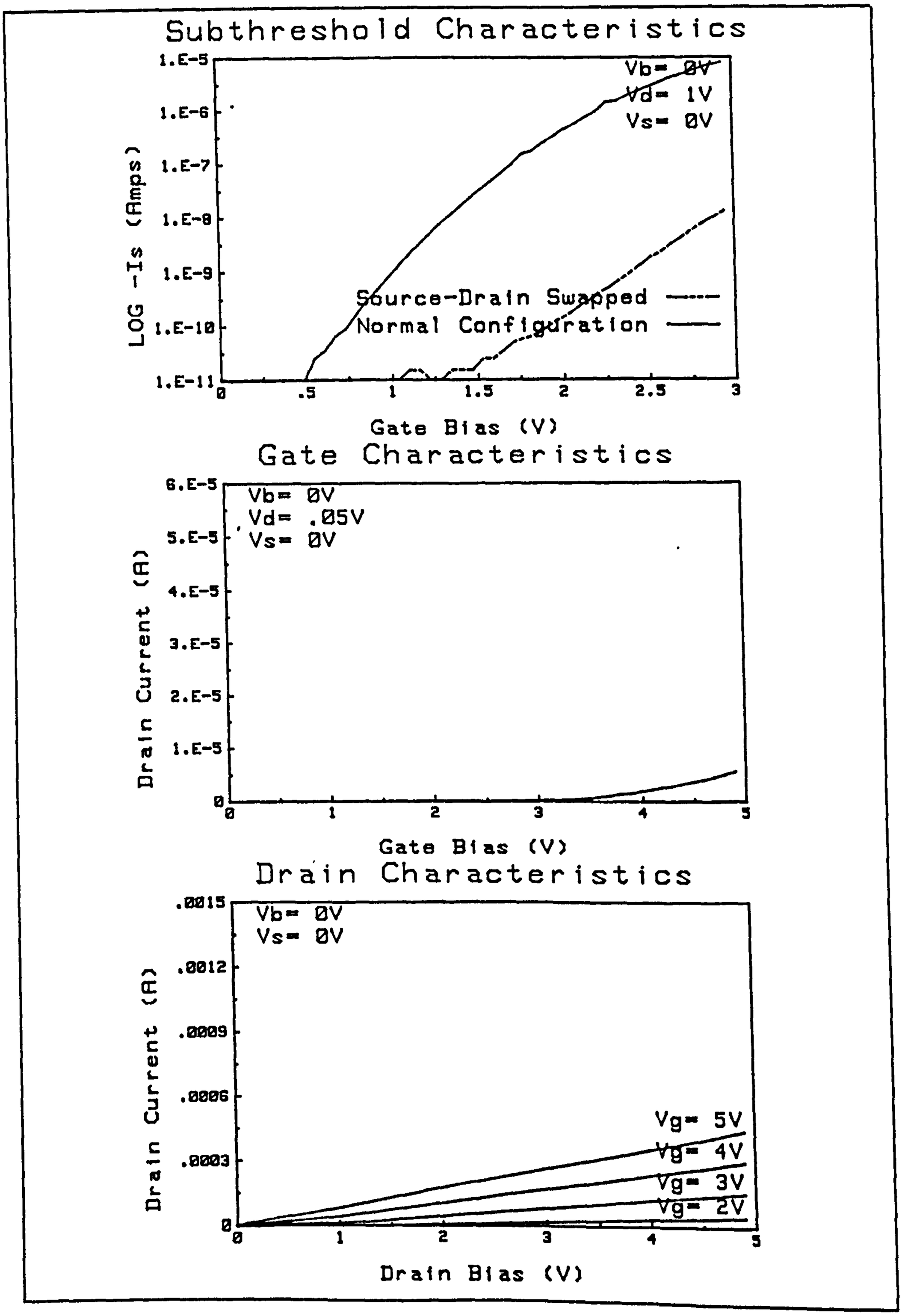


Figure 6.4, Experimental Transistor 999112, Drawn DGT $L_D = 1.9\mu m$ $Gap = 0.45\mu m$.

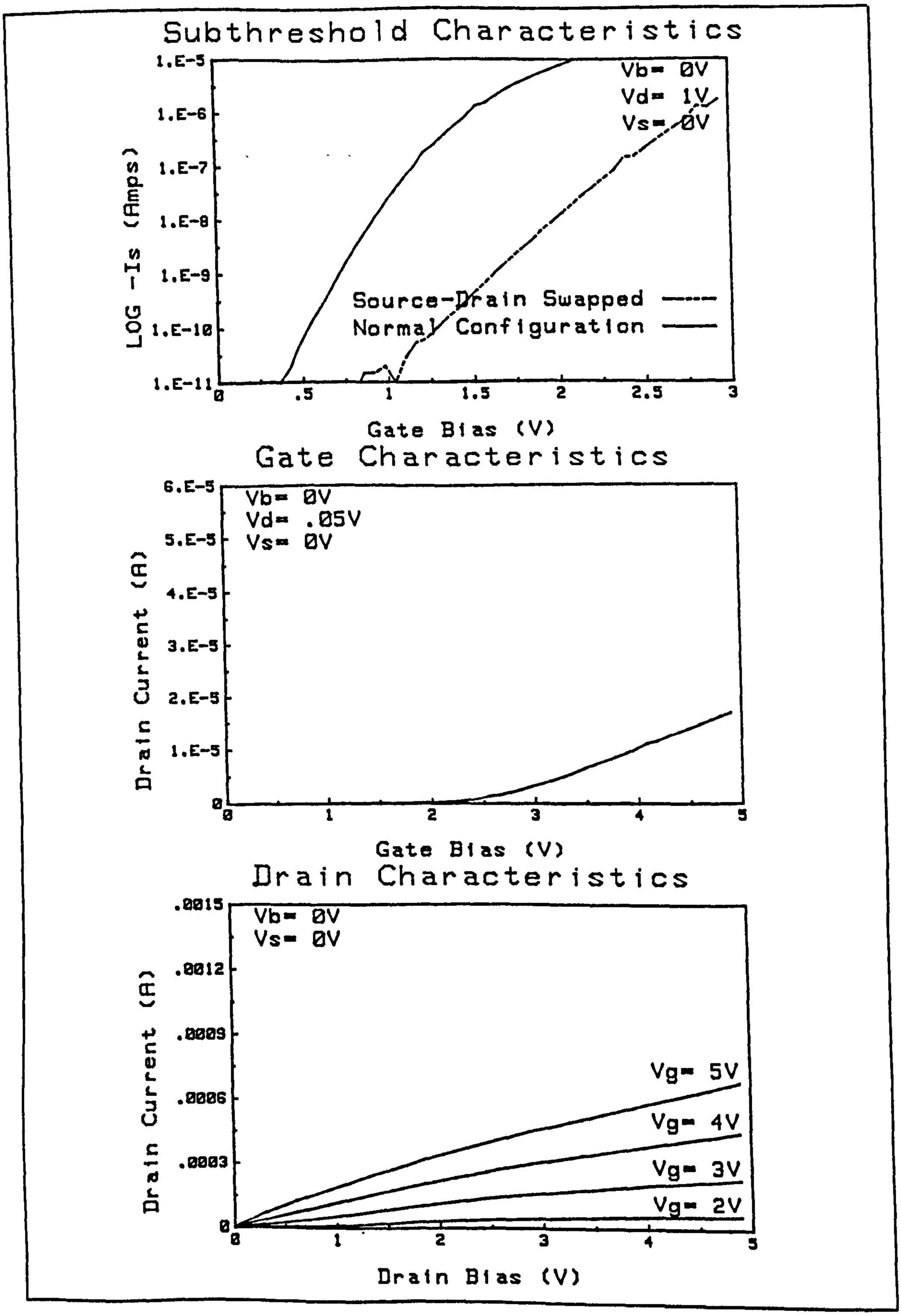


Figure 6.5, Experimental Transistor 999113, Drawn DGT $L_D = 1.9\mu m$ $Gap = 0.30\mu m$.

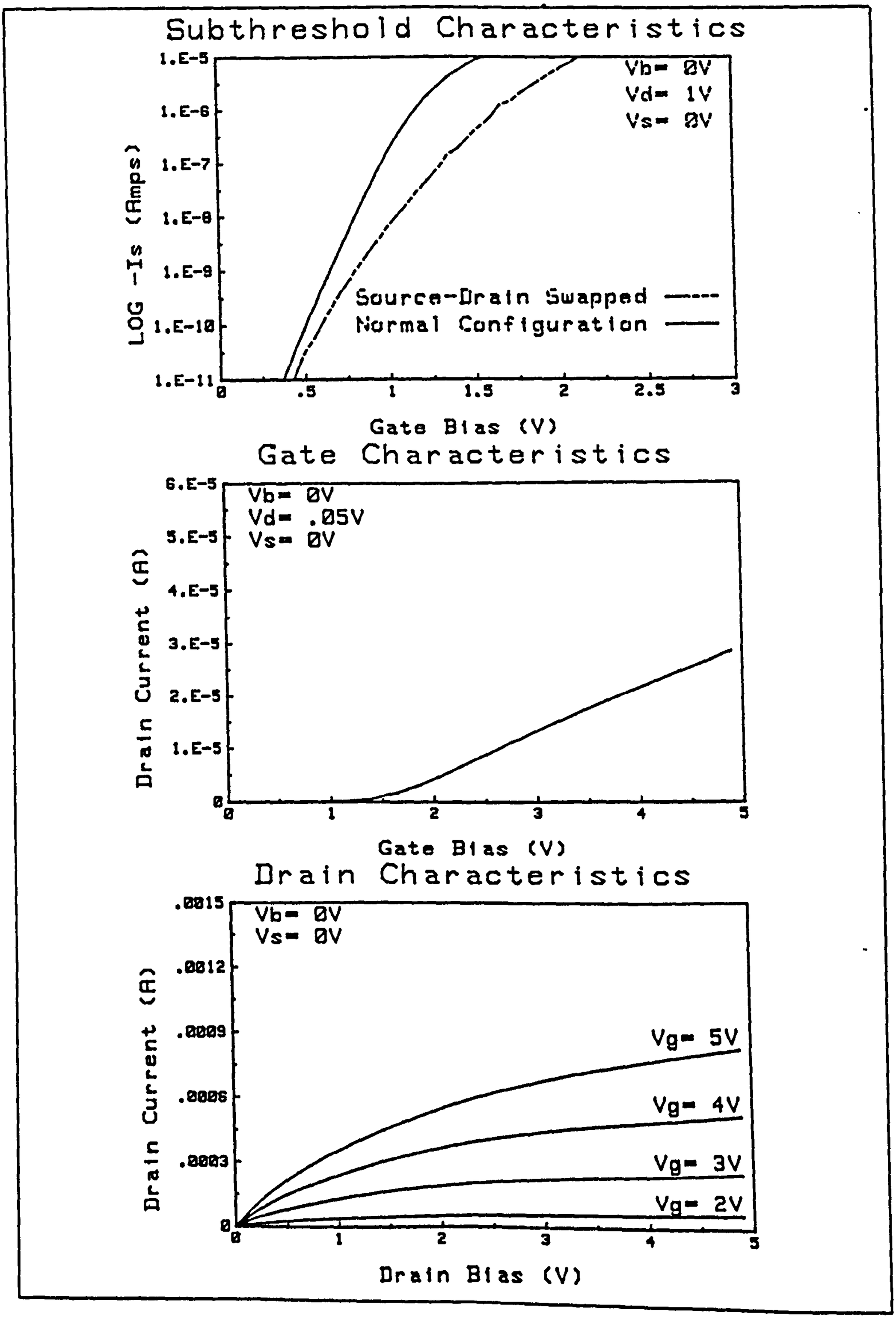


Figure 6.6, Experimental Transistor 999114, Drawn DGT $L_D = 1.9\mu m$ $Gap = 0.15\mu m$.

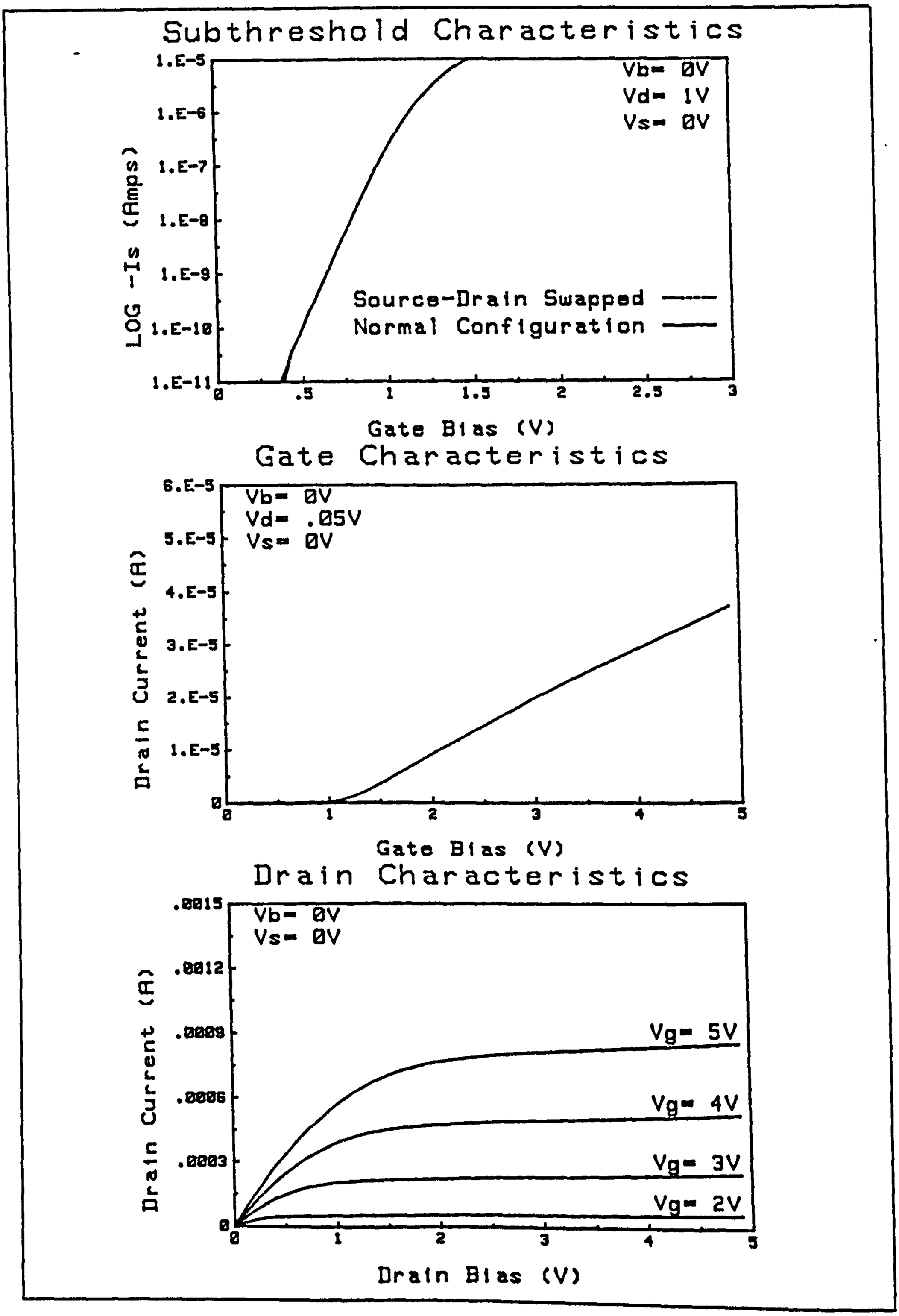


Figure 6.7, Experimental Transistor 999115, Drawn NGT $L_D = 1.9\mu m$

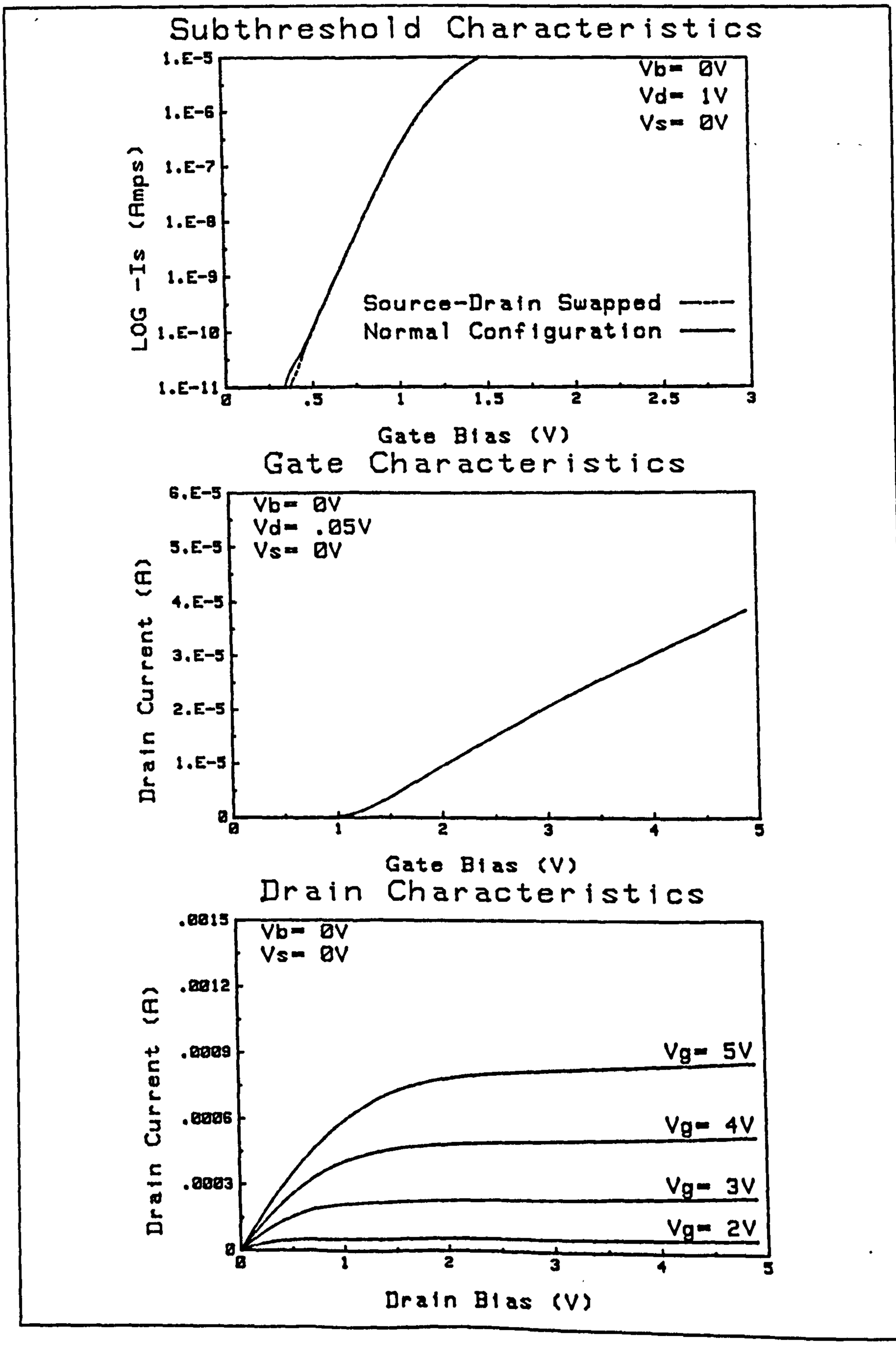


Figure 6.8, Experimental Transistor 999116, Drawn SGT $L_D = 1.9\mu m$ $Gap = 0.15\mu m$.

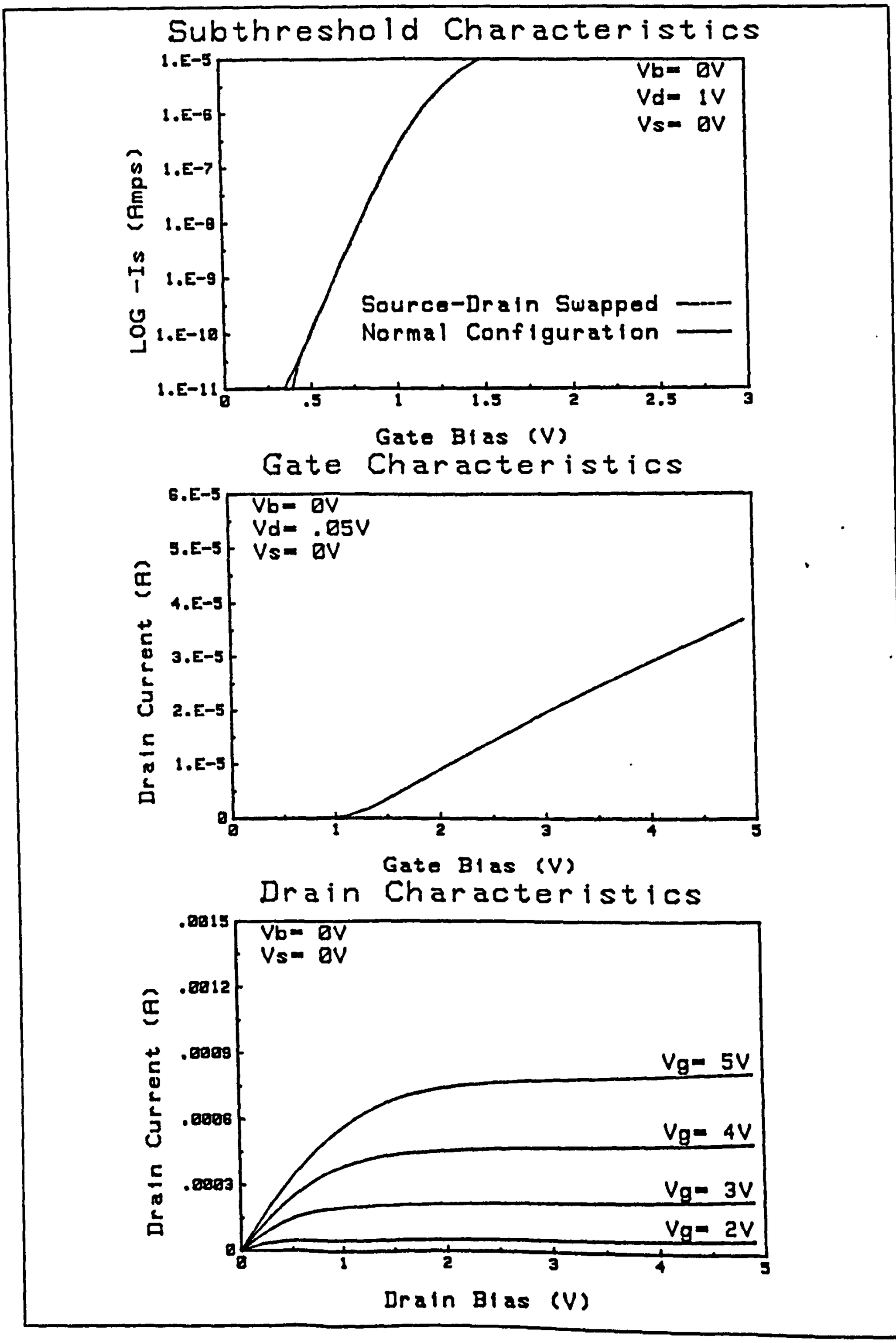


Figure 6.9, Experimental Transistor 999117, Drawn SGT $L_D = 1.9\mu m$ $Gap = 0.30\mu m$.

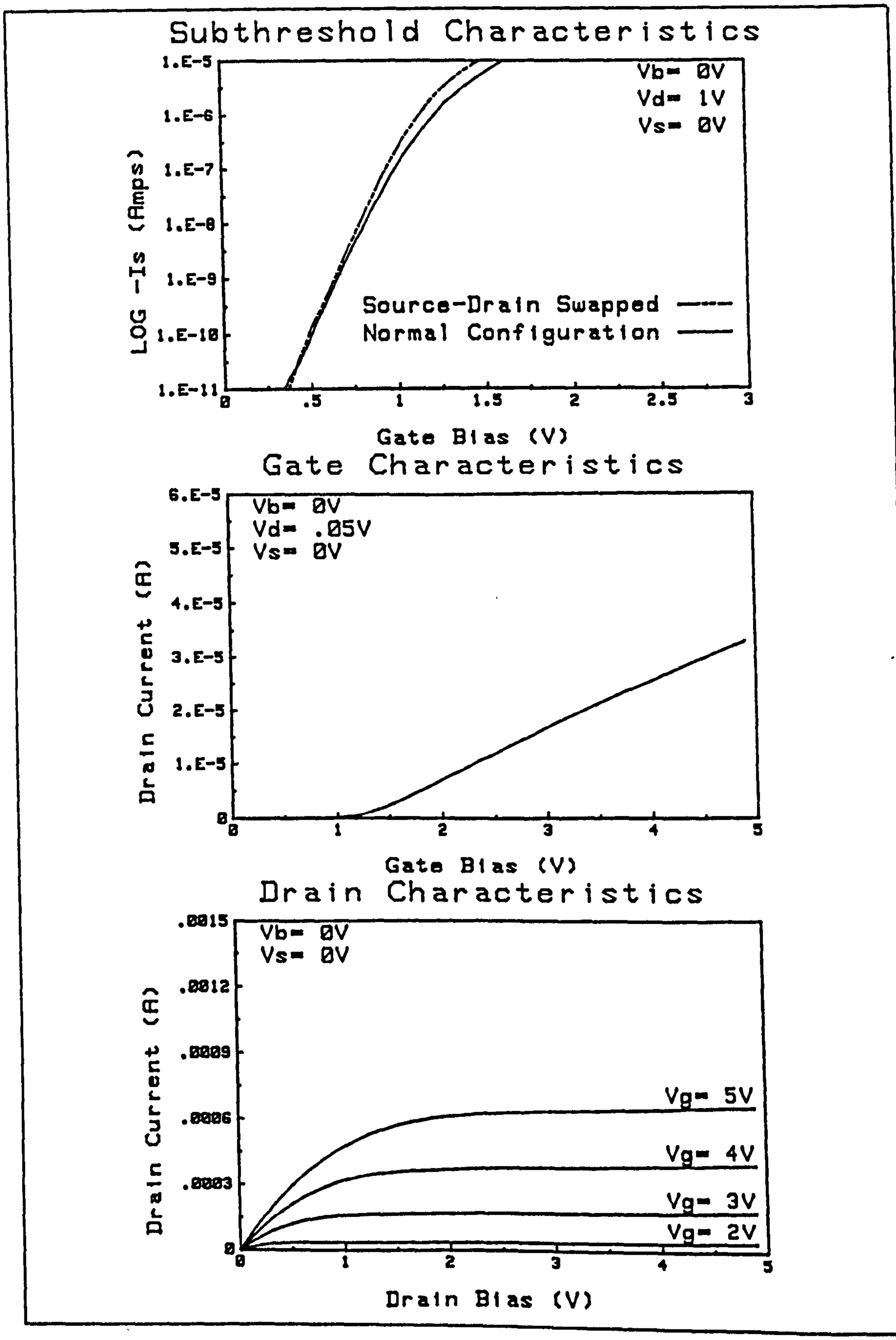


Figure 6.10, Experimental Transistor 999118, Drawn SGT $L_D = 1.9\mu m$ $Gap = 0.45\mu m$.

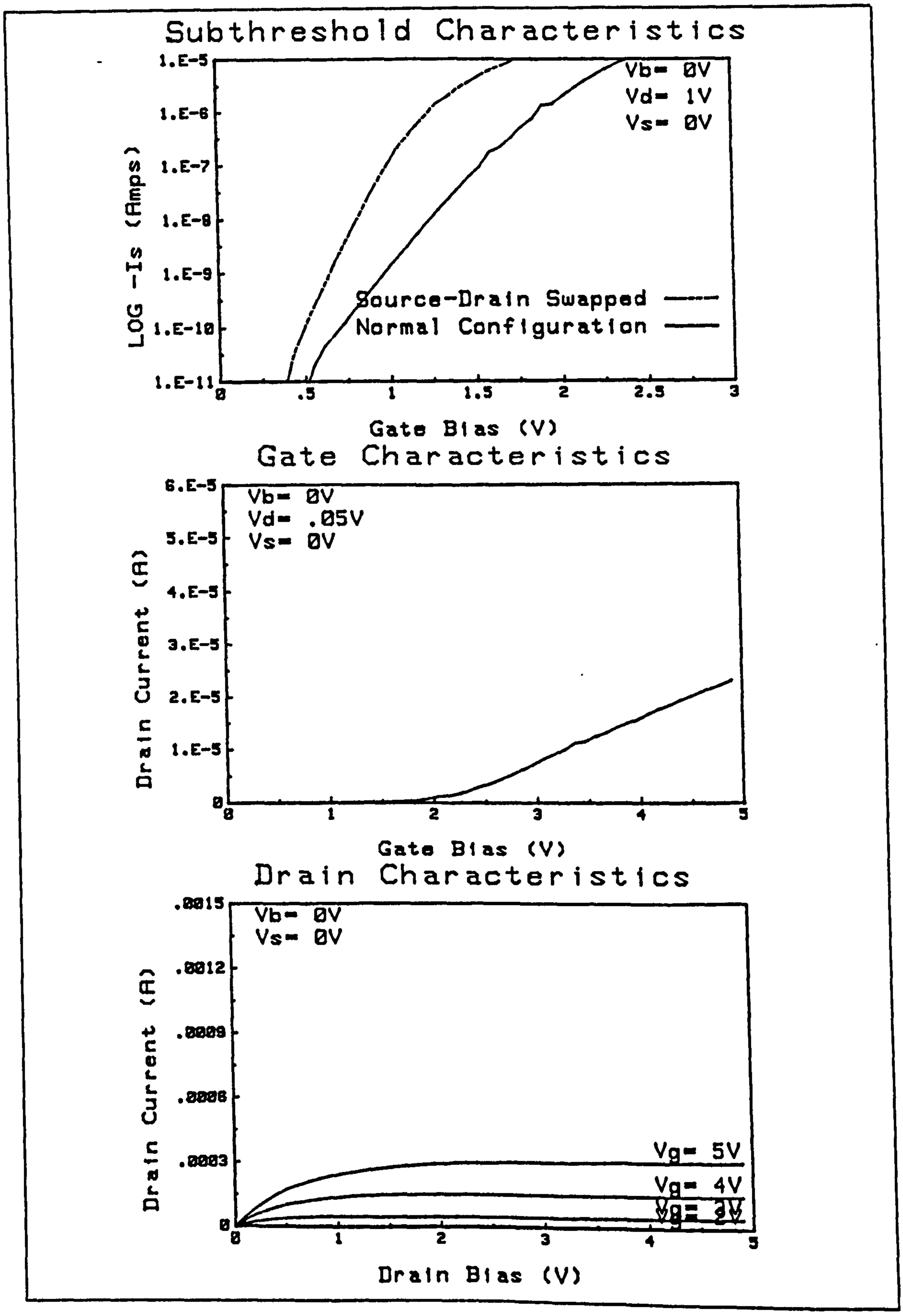


Figure 6.11, Experimental Transistor 999119, Drawn SGT $L_D = 1.9\mu m$ $Gap = 0.60\mu m$.

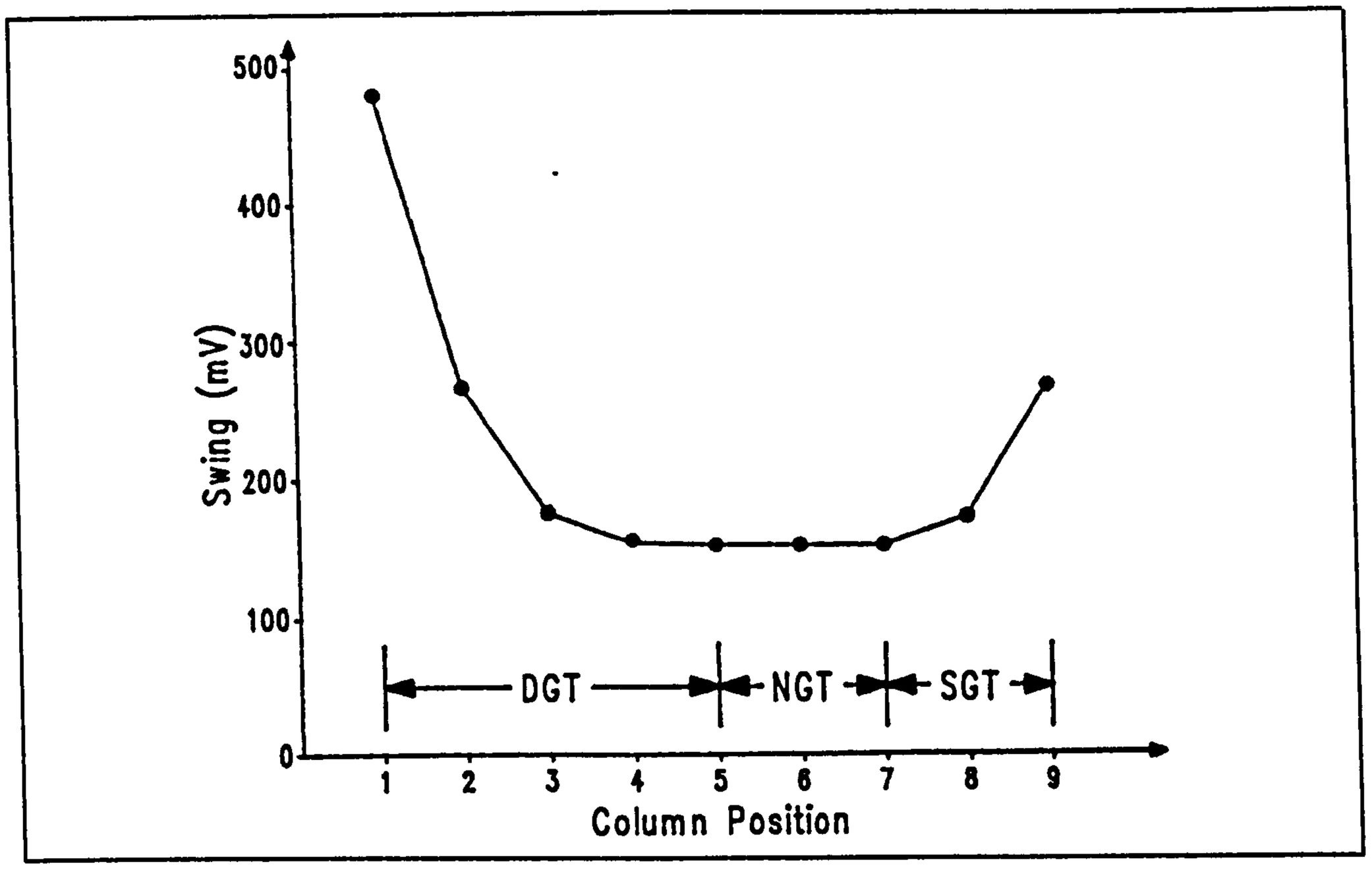


Figure 6.12, Subthreshold swing as a function of column position.

Threshold.

The principle parameters derived from the i_D-v_G characteristic curves is that of the threshold voltage (V_t). In this experimental work V_t is taken as the intercept of the abscissa by a tangent to the curve at a point of its maximum slope. The threshold voltage increases for both DGTs and SGTs as the gap size increases, as noted in the section on electrical alignment techniques in the previous chapter. The ungated region of the channel in both cases must be "inverted" by the fringing electric field from the gate sidewall and top surface. As the gap increases the gate potential must be increased to achieve the same electric field further away from the gate, which results in a higher effective threshold voltage for inversion of the entire channel.

There is also a threshold voltage dependence on channel length for the normally gated transistor, which is predicted by small geometry theory. Figure 6.13 shows this effect for the NGT along with the threshold voltage of the DGTs and SGTs.

There is no good correlation of the DGT and SGT threshold voltages with channel length due to two sources of errors. Firstly the step size of 150 nm, dictated by the metal etching variation, is too coarse to enable selection of transistors with similar gap

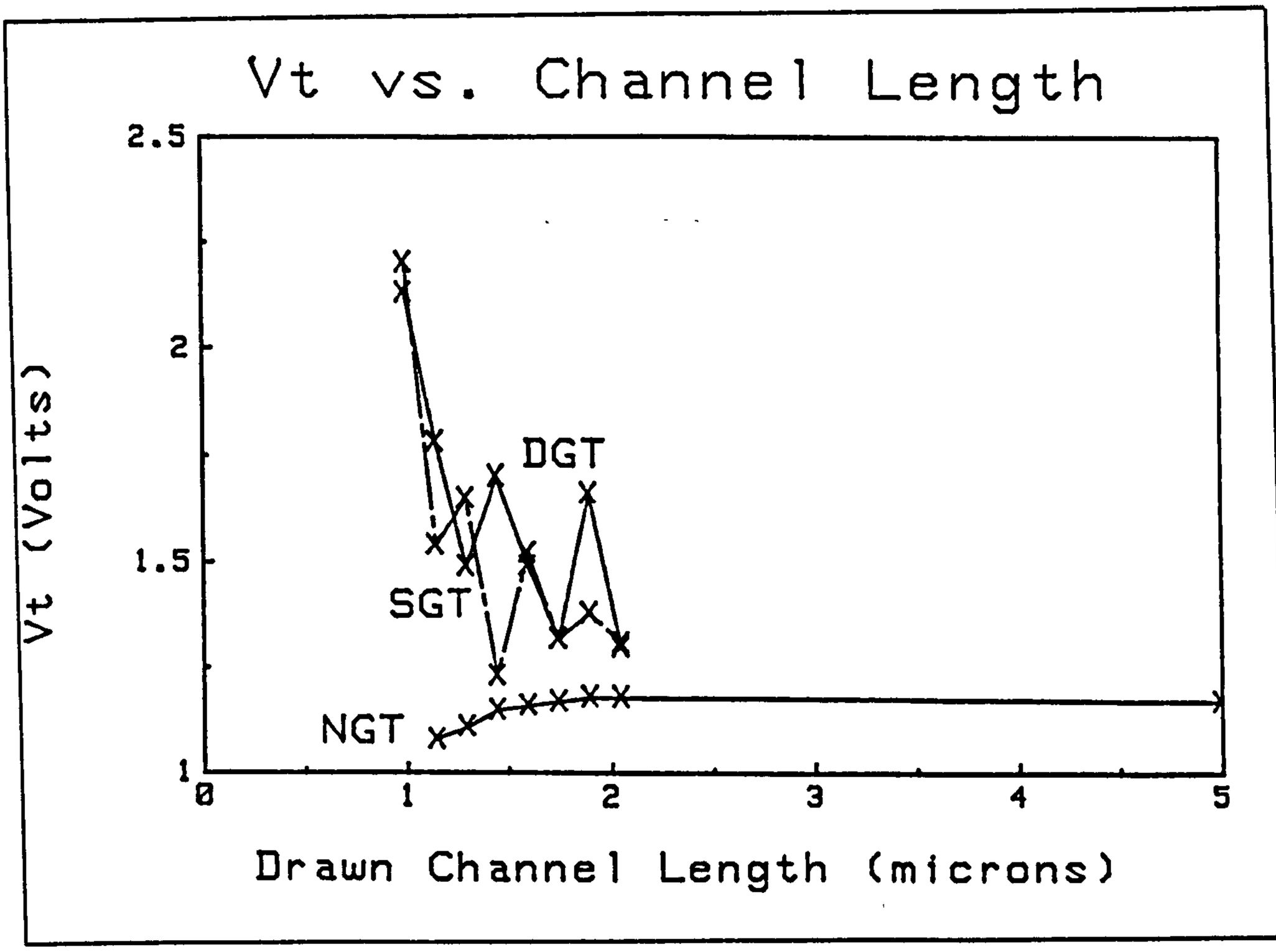


Figure 6.13, Threshold Voltage Dependence on Channel Length.

sizes, and secondly the fringing elcctric field magnitude appears to be affected by more than just the gap size. Variations in the gate sidewall shape due to etching variations are thought to be one possible factor.

Linear Region.

The drain characteristic curves show more dissimilarities between DGTs, NGTs, and SGTs than the rest of the curves. The linear region of operation has a number of differences.

The drain gap transistors in figures 6.3 to 6.6 have an extended linear region, and in fact those with the largest only show linear behaviour. The reason for this is that the drain depletion region is separated from the channel by the drain gap region. As the drain bias is increased the depletion region expands, but in the case of the larger gap sizes never reaches the channel and therefore neither a "pinch off point" or drain current saturation is reached. The DGTs behave like vacuum tube triodes, ie like

a resistor whose value is controlled by the gate voltage. Saturation does start to occur for the smaller gap size (figure 6.6) where the drain depletion region can reach the edge of the gate controlled channel.

Source gap transistors, figures 6.10 and 6.11, have a linear and saturation region similar to the normally gated transistors. The effect of gap size is to increase the series resistance and therefore increase the "linear region" resistance. The larger the gap size the greater the resistance.

The transistors in figures 6.7, 6.8, and 6.9 provide the opportunity to evaluate the effect of gate overlap of the source and drain on the linear region and series resistances. The transistor in figure 6.7 can be considered a "weakly overlapped" transistor on the drain side or a "near DGT" and the transistor in figure 6.9 a "near SGT". The linear region resistance of these three transistors at each gate voltage can be extracted by taking the slope of the tangent to the i_D-v_D curves at the point of their maximum slope.

If those extracted resistances are plotted against the drawn channel lengths for a number of different channel lengths, as was done for the normally gated transistors in figure 6.14, it is possible to draw best fit lines through the group of points for each gate voltage which intersect at a single point. That point contains two useful items of information. Its position along the channel length axis is the total channel length reduction "delta L".

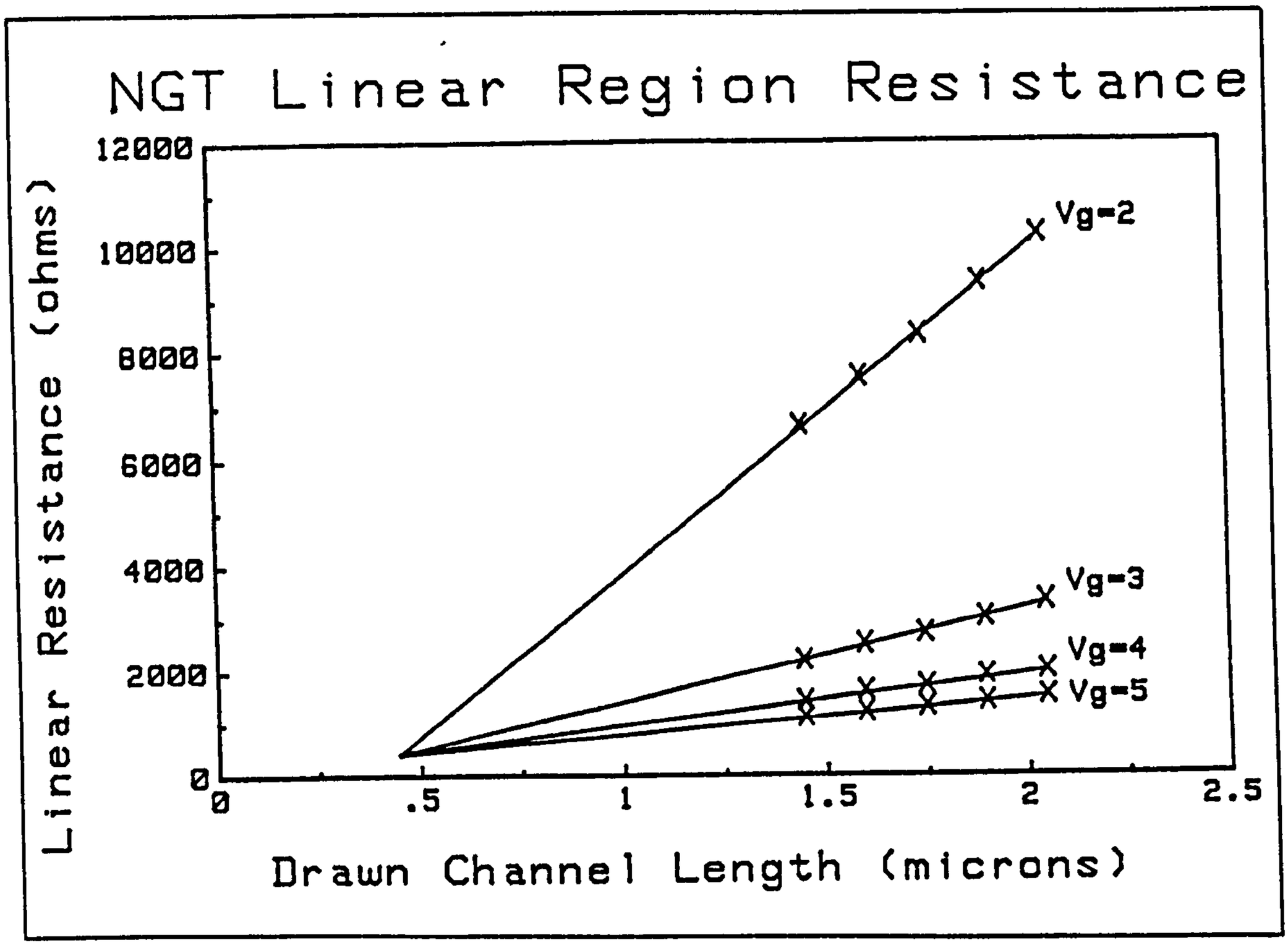


Figure 6.14, Delta L and series resistance extraction for NGTs.

The point's position along the resistance axis is equal to the total series resistance. It is mostly due to source and drain diffusions but also to contacts and tester wiring resistance.

It is possible to extract the values for that point by fitting a "least squares" line to the points, so that

$$R_{total} = b \times L_D + a \tag{6.1}$$

which gives the results in table 6.2.

V_G	Correlation Coefficient r^2	a (Ohms)	b $(Ohms \times \mu m^{-1})$
2	0.999	–2120	6000
3	0.996	–410	1800
4	0.987	–30	1000
5	1.000	133	667

Table 6.2, The linear fit parameters for NGT linear resistance.

The intercepts of those curves can then be taken to obtain six values for the series resistance and delta L which are presented in table 6.3.

Curves	Delta L (μm)	R_S (Ohms)
2 - 3	0.407	323
2 - 4	0.418	388
2 - 5	0.422	414
3 - 4	0.475	445
3 - 5	0.479	452
4 - 5	0.489	459
Mean	0.448	413
Standard Deviation	±0.037	±52

Table 6.3, Intercepts of the fitted curves for NGT linear resistance.

Those results also fit the near drain gap transistors linear resistances as shown in figure 6.15, but clearly a different solution is required for the near SGT linear resistance presented in figure 6.16.

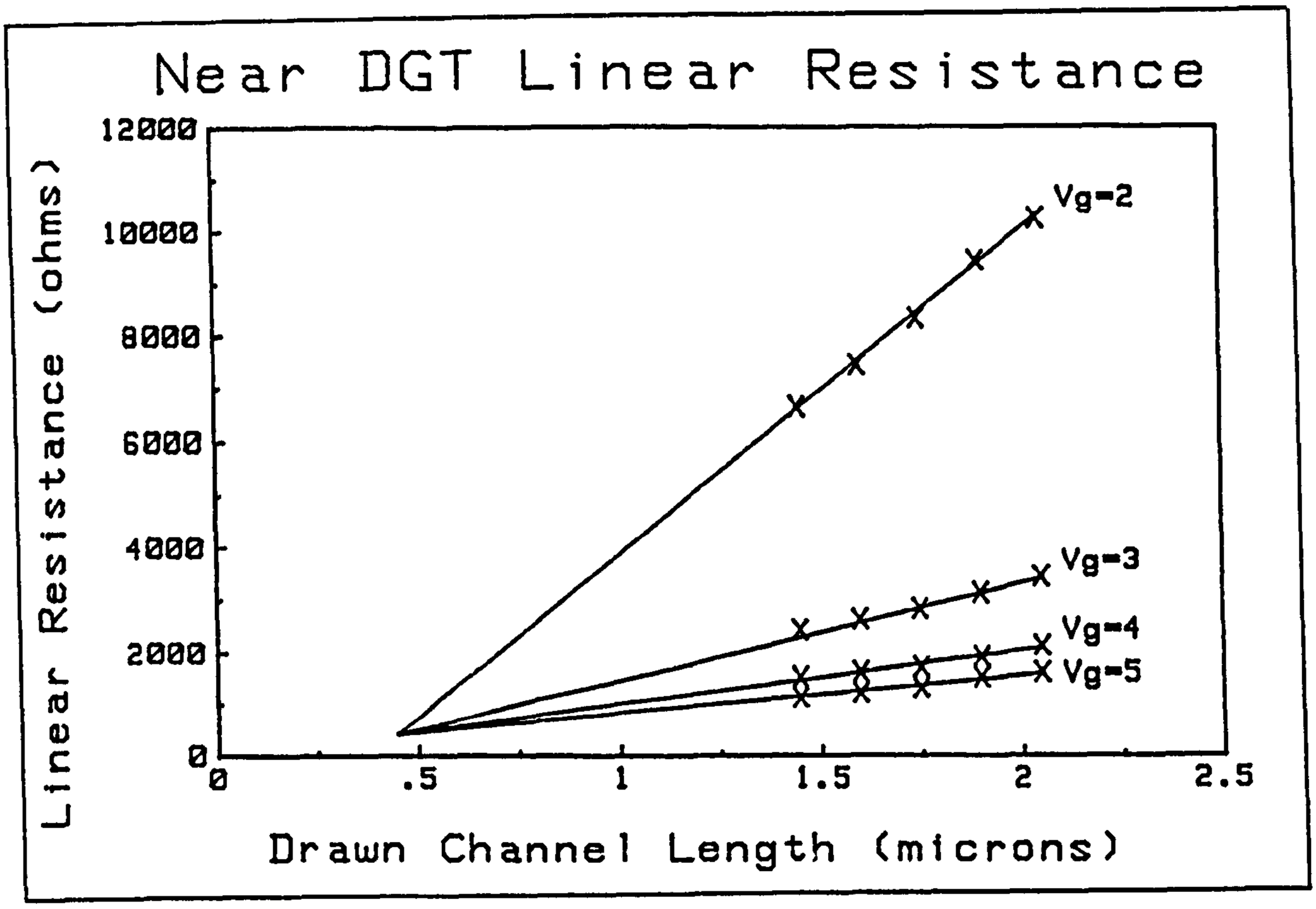


Figure 6.15, Delta L and series resistance extraction for near DGTs.

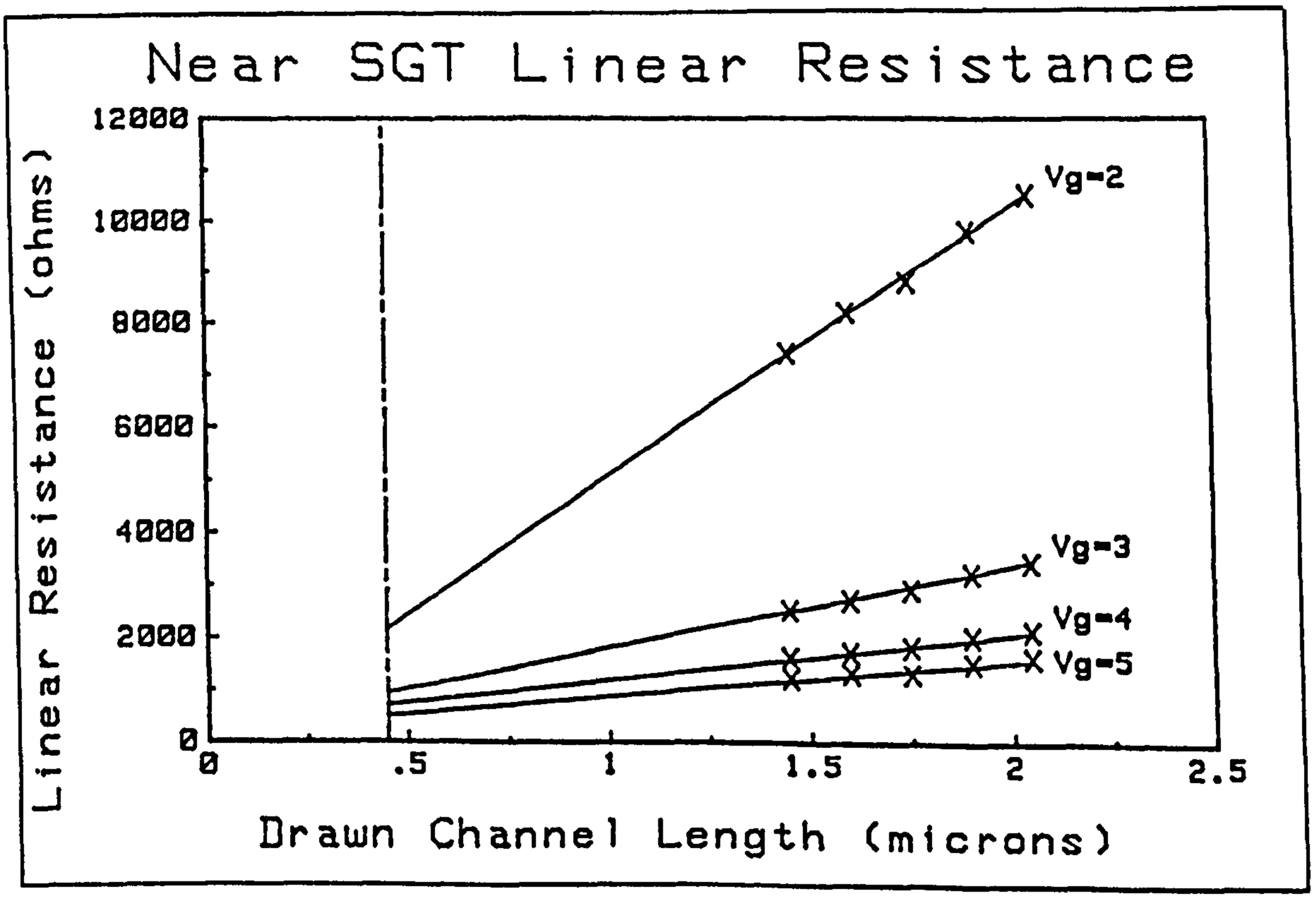


Figure 6.16, Delta L and series resistance extraction for near SGTs.

Since the near source gap transistors channels are made with the same fabrication step as the NGTs, the channel length reduction must be the same. Therefore the near SGTs must have gate voltage dependent series resistance. It is possible to extract the series resistance at each gate voltage by substituting delta L into the fitted curve formulae (equation 6.1). The results were plotted in figure 6.17.

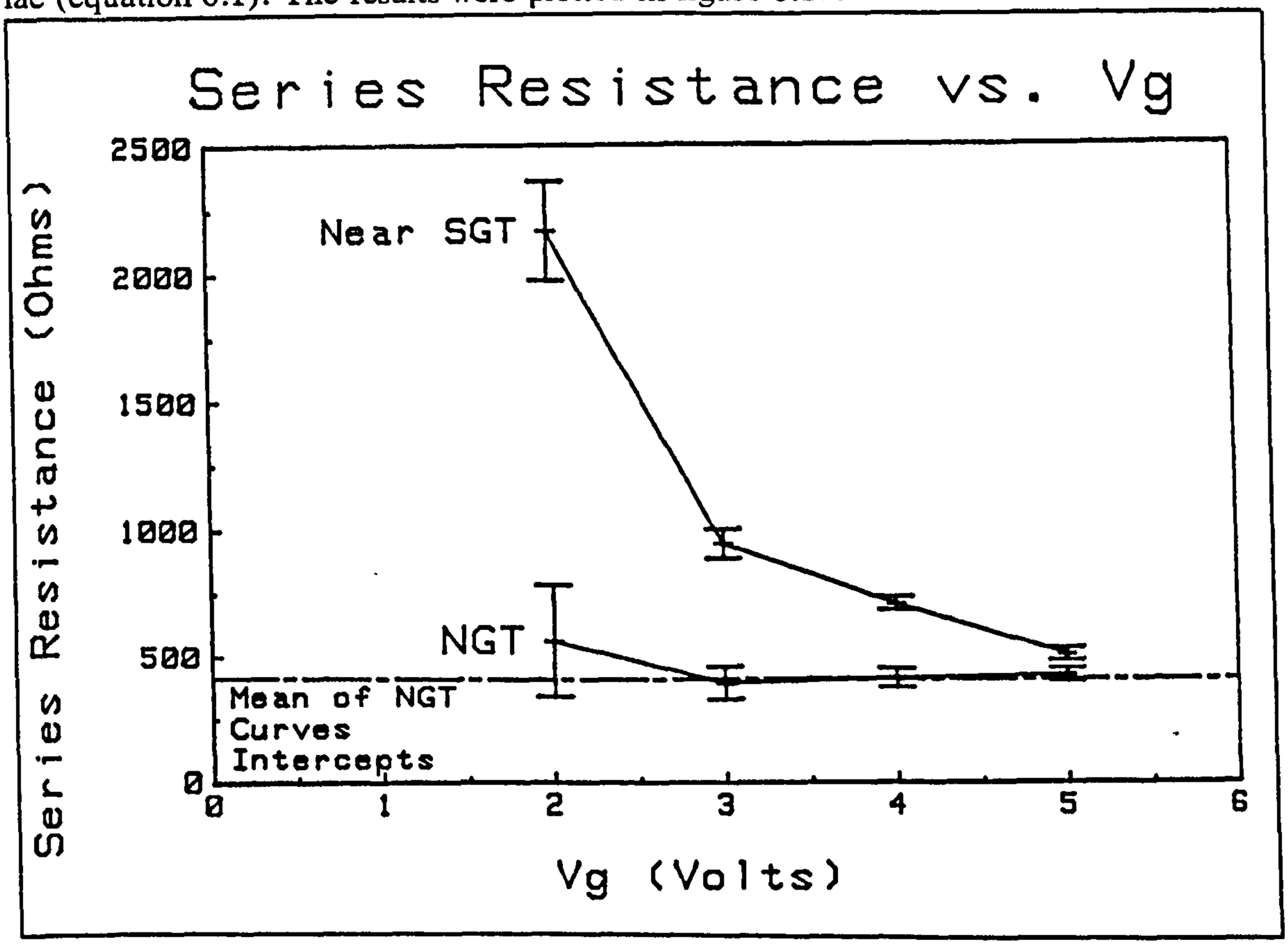


Figure 6.17, Gate voltage dependent series resistance.

An explanation for this effect can be attained by considering the effect of the gate edge roughness. If some parts of the gate edge overlap the source and some parts do not, the entire transistor could be considered as a parallel connection of a number of both NGTs and SGTs. At lower currents, such as the subthreshold region, the NGTs dominate. When higher currents flow the SGTs contribution is more important. The total series resistance would then decrease as the SGTs "turn on". The SGTs depend on the fringing electric field so that as the gate voltage is increased more current flows through them, as more turn on, and the series resistance is reduced.

A similar mechanism should occur for near DGTs, but the drain depletion region moves under the gate as the drain bias is increased and this effectively eliminates the extra series resistance.

Saturation Region.

As mentioned earlier, the DGTs do not show a saturation drain current due to the inability of the drain depletion region to reach the gate controlled channel region.

However, the source gap transistors do show a saturation drain current. Two parameters which can readily be extracted from an i_D-v_G curve are the saturation current at some gate voltage and the slope of that saturation curve with respect to the drain voltage. Figures 6.9, 6.10, and 6.11 show that as the gap increases for SGTs the saturation current for a given gate voltage decreases. This is what one would expect since the source gap region dominates the overall characteristics of the SGT. It is also apparent from the curves that the slope of the saturation curves decrease as the gap size increases. This follows from the domination of overall characteristics by the source gap region transistor whose "channel length" is buffered from drain depletion width shortening by the normally gated region of the channel. A crude test of this theory is shown in figure 6.18 where the slope of the saturation current appears highly related to the magnitude of the saturation current thus supporting the idea that they are both dominated by the indirectly gated region of the channel at the source side.

Hot Carrier Susceptibility.

The sensitivity of the SGTs and DGTs to charge over their gap in gate-to-channel coverage was discussed in chapter five as an experimental irritation. It was, however, realised that the sensitivity could also be used to an advantage as a hot carrier detector.‡ Figure 6.19 shows how a drain gapped transistor could be used to detect both hot electron and hot hole carrier accumulation because the indirectly gated region at the drain side of the channel dominates the DGTs characteristics.

---

† The points for each of the gate voltage curves are from top to bottom from transistors 999117, 999118 and 999119.

‡ This work was presented at the IEE Colloquium on Hot Carriers. [3] At the time of the work the role of hot holes as trap site generators had not yet been discovered.

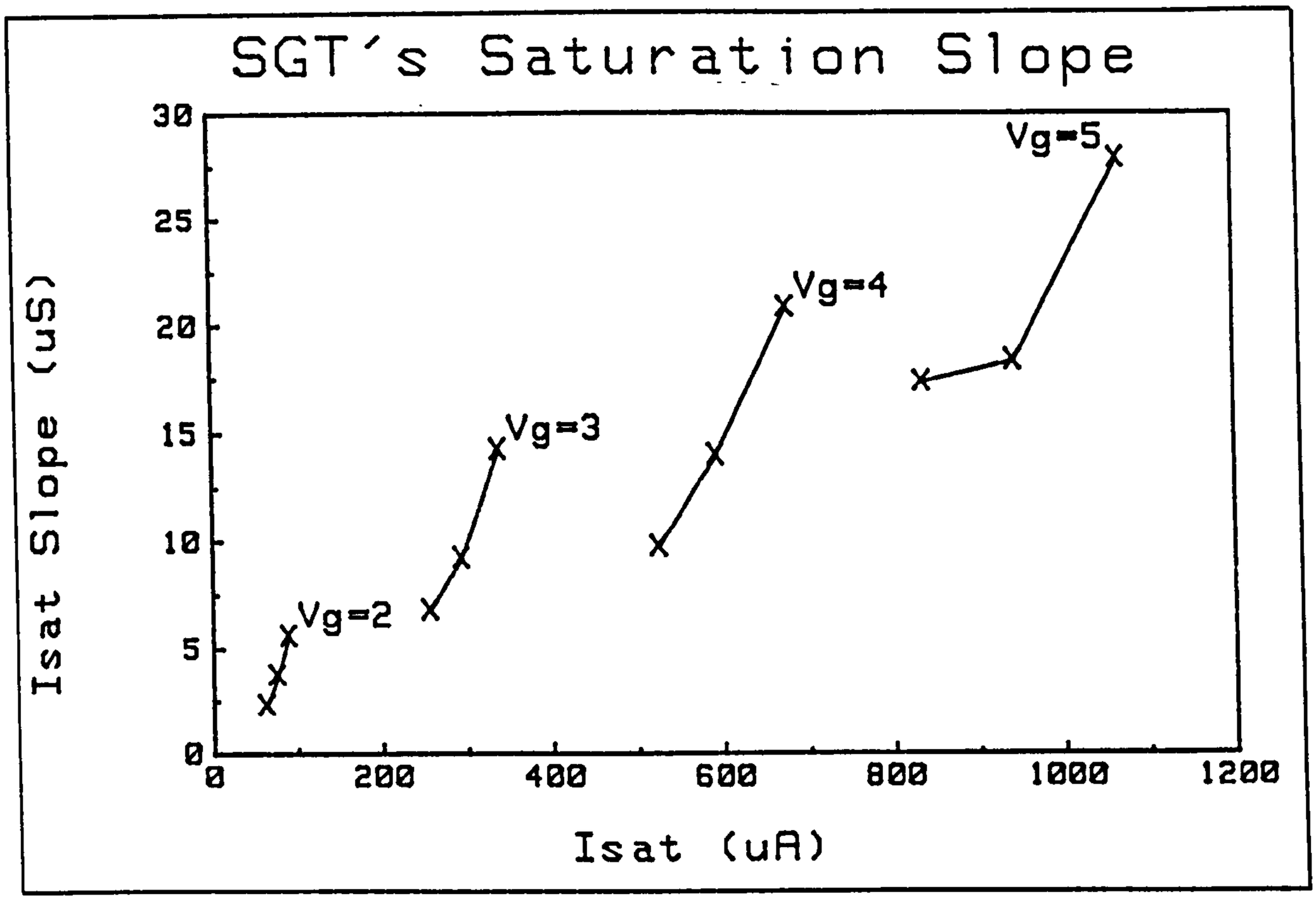
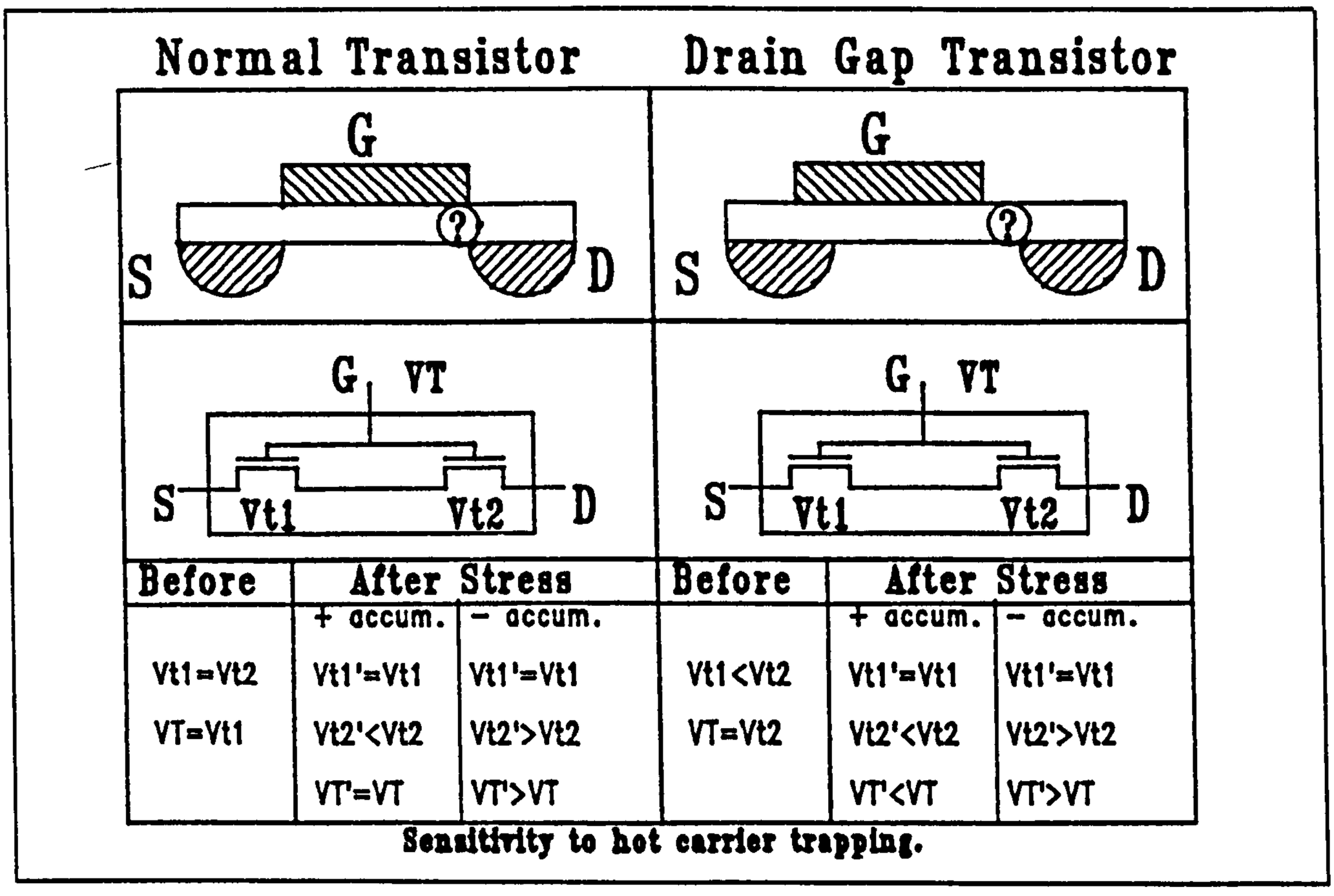


Figure 6.18, The correlation between saturation slope and current for SGTs.†

Although a number of likely biases for hot hole accumulation were applied no negative threshold voltage shifts were observed that could not be explained by temperature fluctuation. The increased sensitivity of the DGT to hot electron degradation was demonstrated by biasing both a NGT and DGT in avalanche hot electron mode ($V_G = V_t + 1\ V, V_D = 8 volts$) as determined by the maximum substrate current. The results, shown in figure 6.20, confirm that the DGT is much more sensitive and in this case 5 times more sensitive than a similar sized NGT.

One general result from this work is that sidewall spacer processes with insufficient design margins to accommodate implanter effects are likely to be more susceptible to hot electron degradation. Another general result is that the increased sensitivity of intentional "gapped" transistors could be used to check a process for susceptibility to hot electrons. If the DGTs made by a process have acceptable threshold voltage shifts, then normally gated transistors made from the same process would have sufficient resistance to hot electron degradation, since they are less sensitive than DGTs.



Before	After Stress		Before	After Stress	
	+ accum.	− accum.		+ accum.	− accum.
Vt1=Vt2	Vt1'=Vt1	Vt1'=Vt1	Vt1<Vt2	Vt1'=Vt1	Vt1'=Vt1
VT=Vt1	Vt2'<Vt2	Vt2'>Vt2	VT=Vt2	Vt2'<Vt2	Vt2'>Vt2
	VT'=VT	VT'>VT		VT'<VT	VT'>VT

Sensitivity to hot carrier trapping.

Figure 6.19, Drain Gap Transistors as Hot Carrier Detectors.

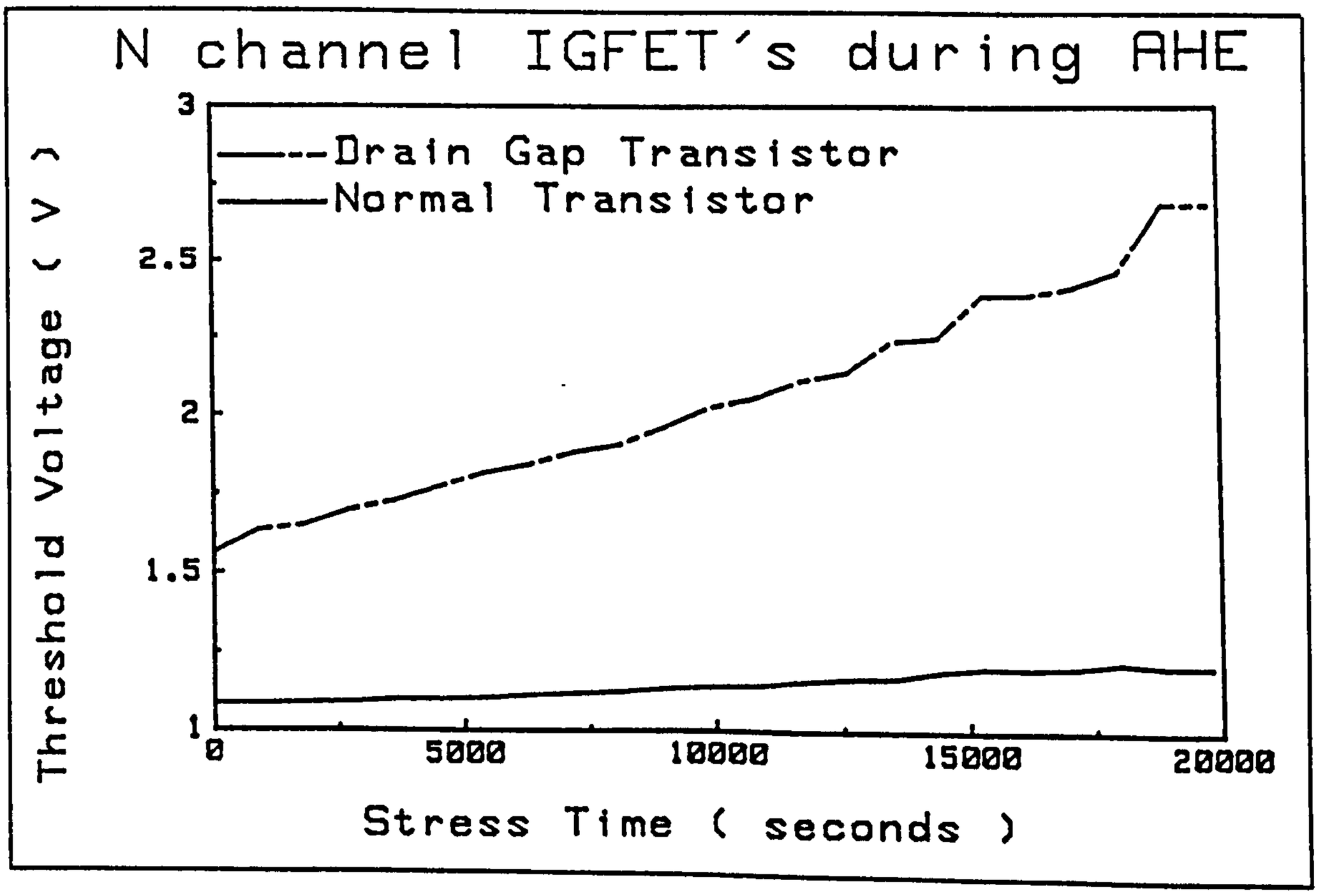


Figure 6.20, The Increased Sensitivity of DGTs to Hot Electron Degradation.

6.4. Chapter Summary.

The characteristics of a series of incompletely gated field effect transistors were presented in this chapter. The basic characteristics of subthreshold current, i_D-v_G, and i_D-v_D curves were shown for a typical reference transistor and for the transistors in a typical column of the progressional offset array. Deeper analysis was presented for; the subthreshold curves, threshold voltage, linear resistance, series resistance, saturation current, slope in saturation, and hot carrier susceptibility of the transistors in the typical column. The hot electron sensitivity of drain gap transistors and the gate voltage dependent series resistance of "near" source gap transistors was of particular interest.

6.5. References.

1. J.A. Serack, A.J. Walton, and J.M. Robertson, "The Effect of Device Geometry on IGFET Characteristics," *ESSDERC 87*, pp. 915-918, Bologna, 14th - 17th Sept 1987. (appended)

2. J.A. Serack, A.J. Walton, and J.M. Robertson, "A Novel Device for Studying Gate and Channel Edge Effects of IGFET's," *Proceedings of the 1988 IEEE International Conference on Microelectronic Test Structures. ICMTS*, pp. 67-72, IEEE Electron Devices Society, Long Beach, California, USA., Feb 22-23, 1988. (appended)

3. J.A. Serack, A.J. Walton, and J.M. Robertson, "The Application of a novel Experimental Technique to Investigate Hot Carriers in MOSFETs," *IEE Colloquim*, no. 1987/15, pp. 4/1-4/4, London, Jan 1987. (appended)

Chapter 7, Drain Depletion Region Boundary Motion.

7.1. Introduction.

Subthreshold Swing is a measure of the change in gate voltage required to reduce the drain current by an order of magnitude. The subthreshold swing of a DGT depends on the magnitude of the gap size and drain bias. Since both normally gated and source gap transistors' swing are unaffected by changes in the drain bias, the DGT's dependence must be due to drain depletion boundary motion.

An experiment using the DGT's swing dependence on drain voltage to study drain depletion edge motion is possible using DGTs with different size gaps. [1] The drain gap transistors from one column in the progressional offset array were used to obtain the results presented in this chapter. Although the experiment was performed before the theory was confirmed, using a two dimensional numerical device simulator, the presentation of this work is more logical if the simulation results are covered first.

7.2. Simulation.

The two dimensional device simulator CANDE, which became available during the course of the research, was used to compare the internal and external characteristics of a DGT in order to confirm that drain depletion boundary motion was responsible for the subthreshold swing improvements seen with increasing drain bias.

Simulated Structures.

Two structures were simulated, one with a normally gated transistor with a 1.0 μm channel length, and another with the same channel but with a gap in gate-to-channel coverage of 0.15 μm on the drain side of the channel. The layout description from the database was used to specify the position of the contacts to the source and drain. Aluminium gates and source-drain contacts were also modelled. The source-drain and channel doping concentrations were entered as equivalent "boxes" with concentration magnitudes taken from the original 1-D simulations. It was felt, based on the experimental results, that the electric field fringing of the gate sidewall had to be modelled to get a realistic simulation. The gate sidewall was modelled by allowing

thick field oxide to extend to both edges of the gate which formed a vertical step in the metalisation to mimic the gate sidewall. These features are shown in the NGT channel section of figure 7.1. The gate oxide and field oxide thicknesses were taken from processing measurements.

The transistor descriptions were then compiled by CANDE which modified the source and drain doping profiles to achieve pseudo-two-dimensional doping profiles. The compiled structure was then saved to allow simulation of the external and internal characteristics of each transistor.

General Behaviour.

The general behaviour of a NGT was simulated first in order to see how realistic the modelling was. The electron carrier concentration, current, and electric potential distribution were simulated for the NGT structure at a number of gate and drain biases. The external characteristics were plotted and appear along with the simulated channel cross-section in figure 7.1.

The curves show similar behaviour to the experimental curves presented in chapter six. The main difference is that the saturation curves are flatter in the experimental characteristics than in the simulated ones. That is likely to be due to errors caused by the simple "box" doping description and the accuracy of CANDE's internal mobility model.

Since the subthreshold region was of main interest and the simulated subthreshold swing was within 15% of the experimental measurement, the model was thought to be sufficient for qualitative purposes.

A similar simulation was performed for the DGT structure. The resulting external characteristics are presented in figure 7.2. The model successfully predicted the forward and source-drain swapped subthreshold curve separation, the increased threshold voltage, and extension of the linear region for a drain gap transistor. On the basis of those results the simulations were thought sufficiently realistic to allow examination of the effect of drain depletion boundary motion on the subthreshold characteristics of a DGT.

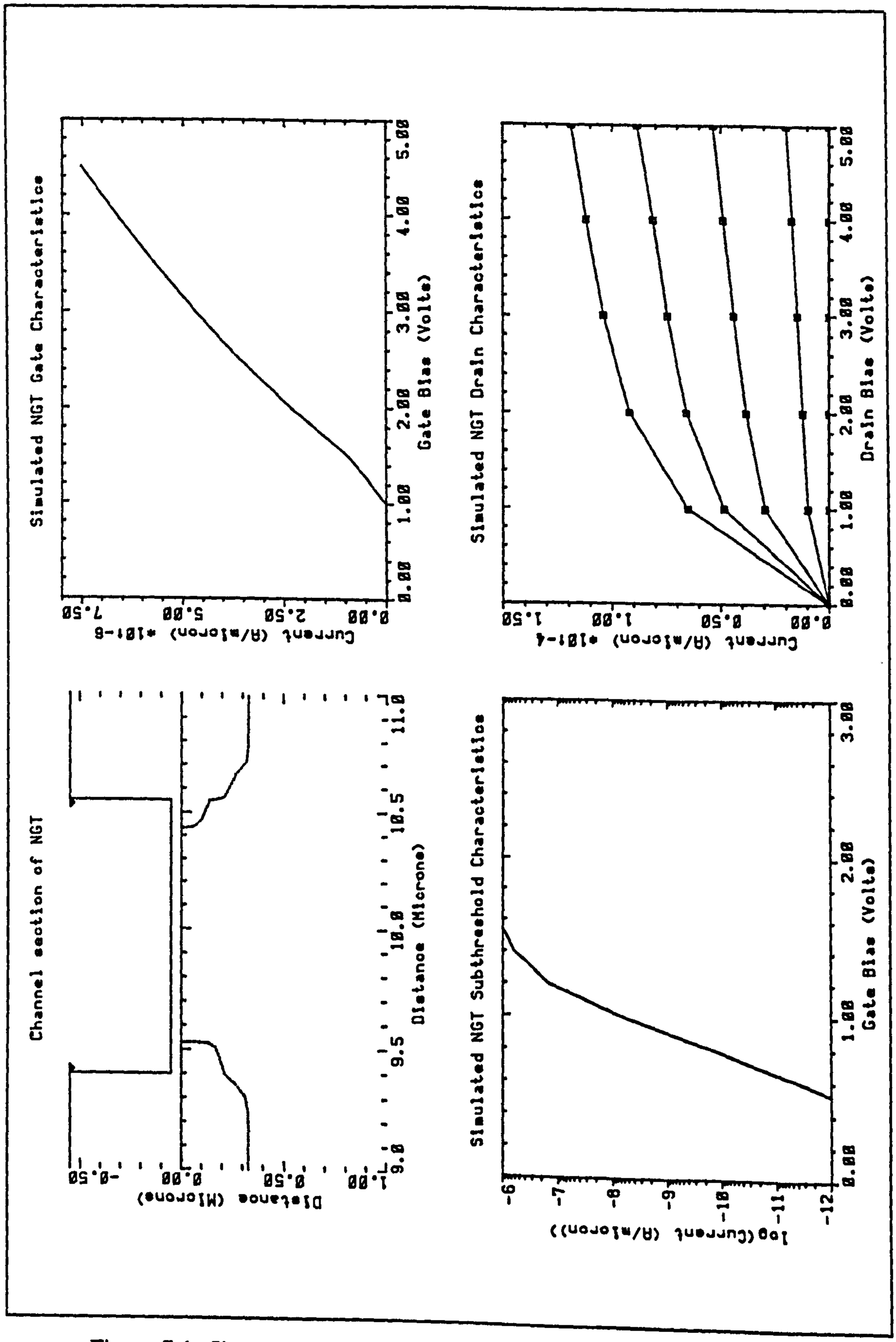


Figure 7.1, Simulated Characteristics for a Normally Gated Transistor.

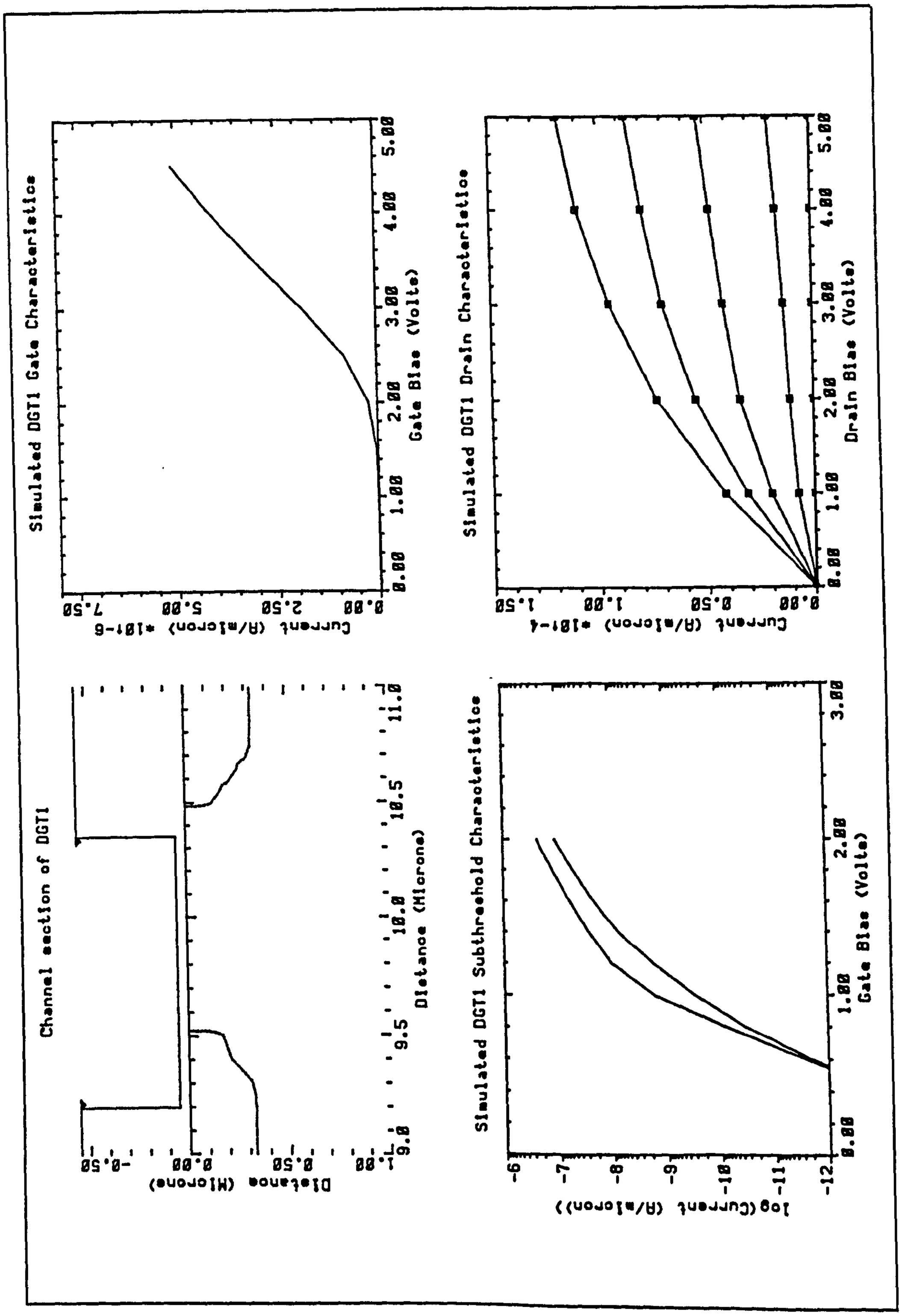


Figure 7.2, Simulated Characteristics for a Drain Gapped Transistor.

DGT Drain Depletion Boundary Motion.

The subthreshold curve was saved during the simulation of the normally gated transistor to act as a reference to compare with the DGT subthreshold curves. The subthreshold curve and drain depletion boundary position were then simulated for the drain gap transistor with a small drain bias. The subthreshold curve of the DGT (solid line) was plotted along with the NGT subthreshold curve in figure 7.3. The corresponding drain depletion boundary for that drain bias is shown in figure 7.4. The zero potential contour shown as a solid line in the plot is a good indication of the extent of the drain depletion region.

The drain bias was increased and the simulation repeated. An improvement in the subthreshold swing can be seen in figure 7.5 as a result of the depletion region expansion towards the gated region of the channel, which is shown in the corresponding figure 7.6. A further increase in the drain bias and resimulation resulted in the drain depletion reaching the edge of the gated channel region at the surface of the semiconductor and an improvement of the subthreshold swing that resulted in a match of the DGT and NGT subthreshold curves. Figures 7.7, and 7.8 show these results.

The knowledge gained from the simulations was that the mechanism for the DGT's swing improvement with increasing drain bias was in fact drain depletion boundary motion, and that a value corresponding to a normally gated transistor was reached once the drain depletion region extended to the edge of the gated portion of the channel.

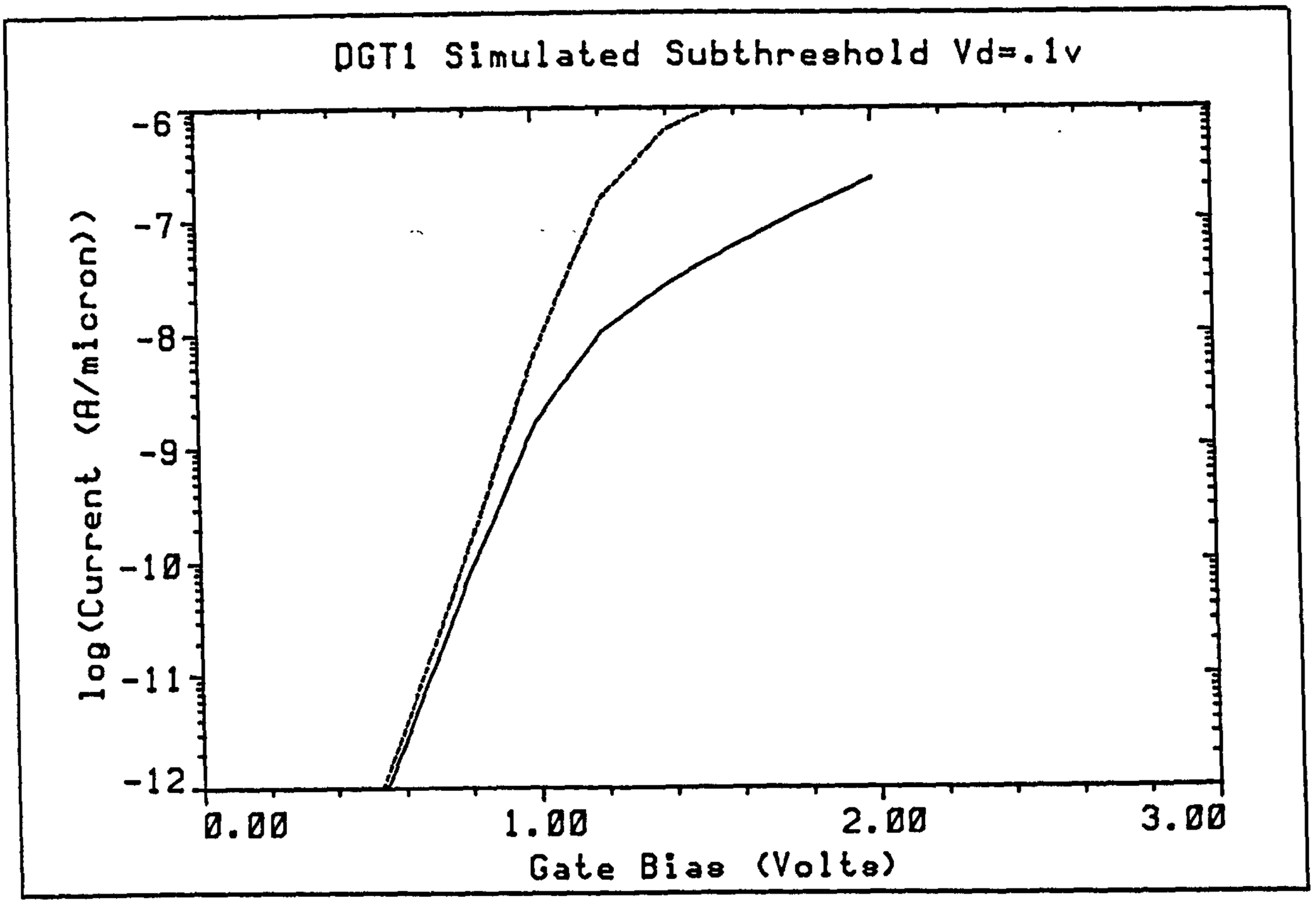


Figure 7.3, The subthreshold curve of a 0.15μm gap DGT, with $V_D = 0.1\ V$.

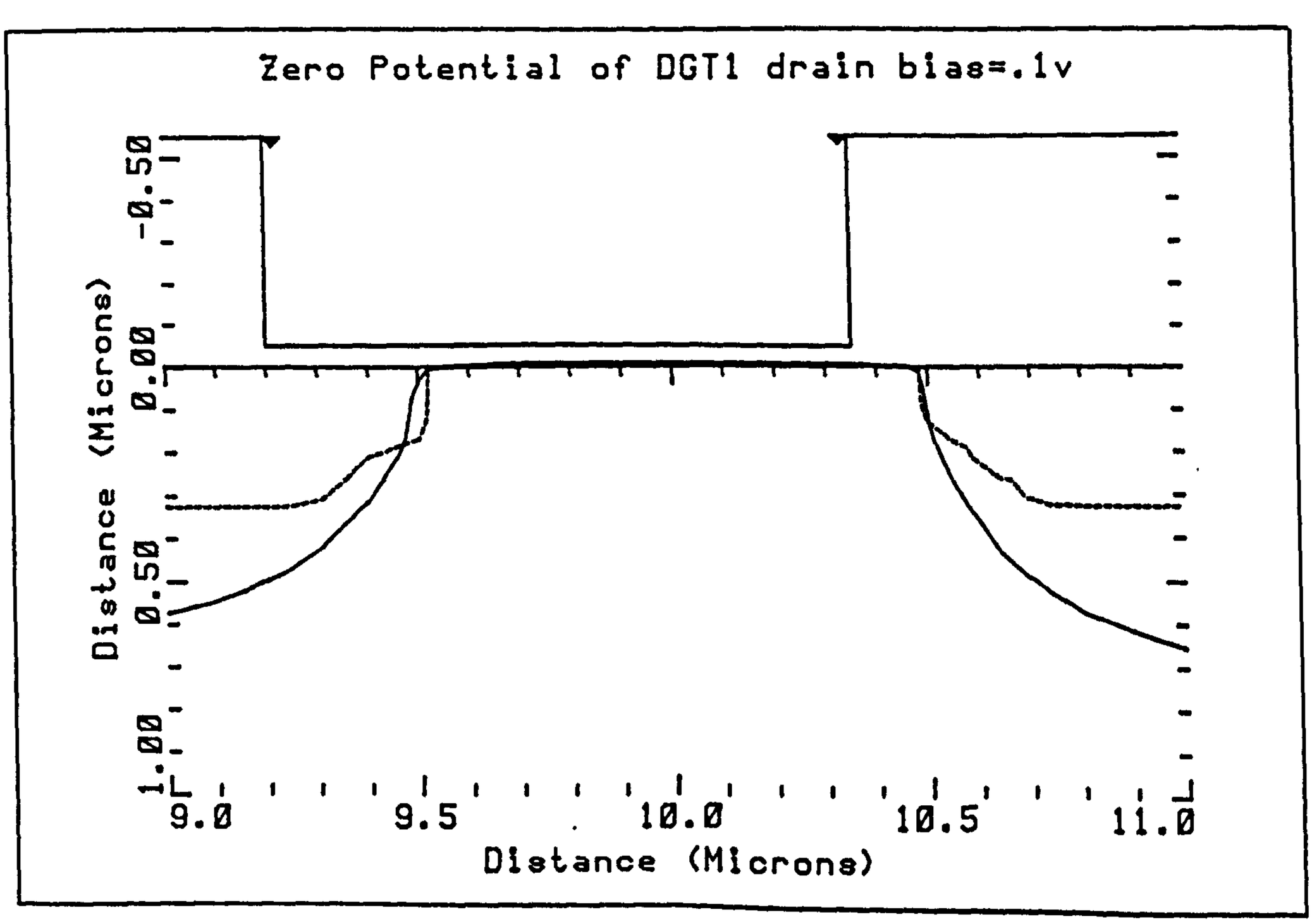


Figure 7.4, The drain depletion width corresponding to the characteristics in figure 7.3.

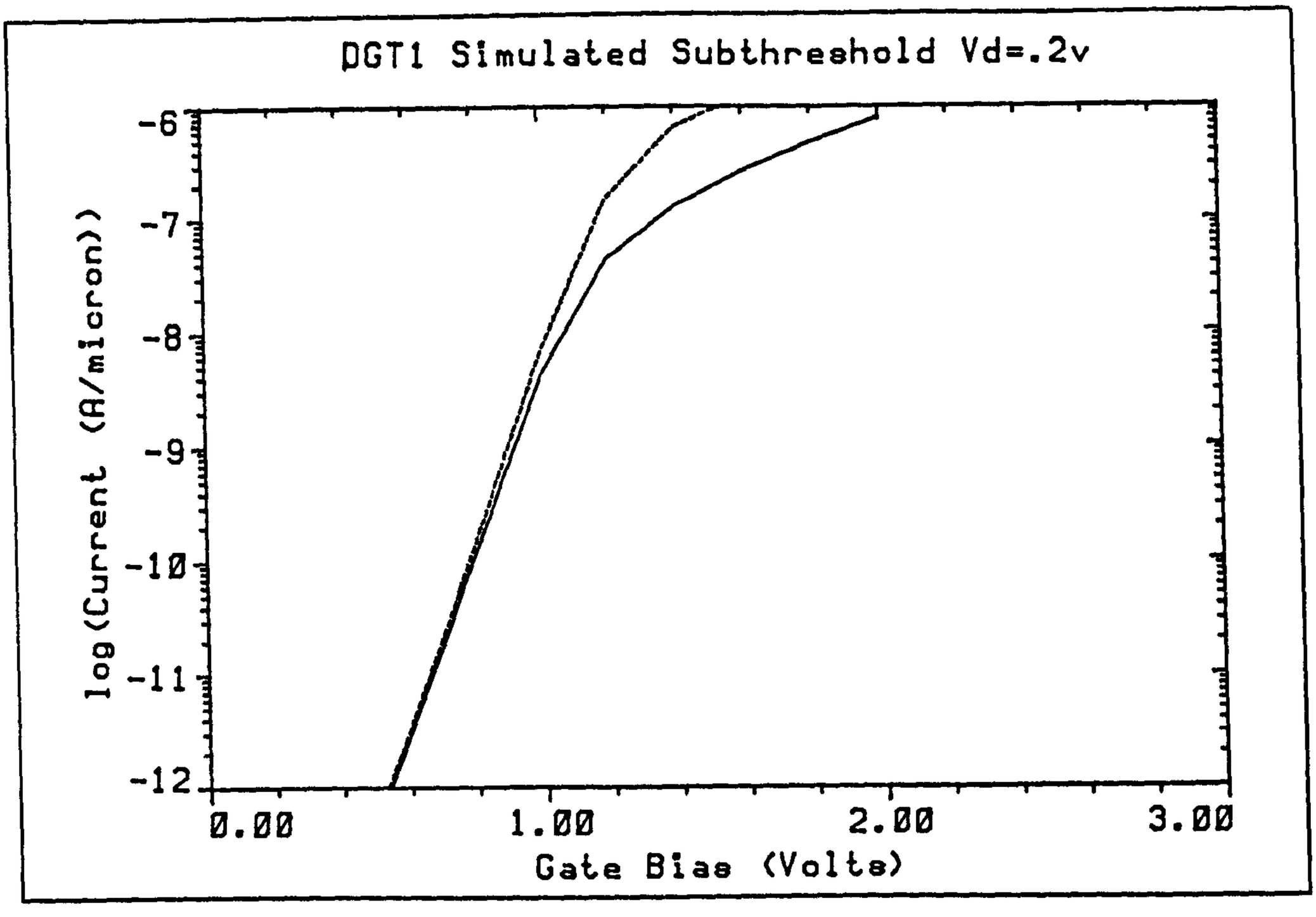


Figure 7.5, The subthreshold curve of a 0.15μm gap DGT, with $V_D = 0.2\ V$.

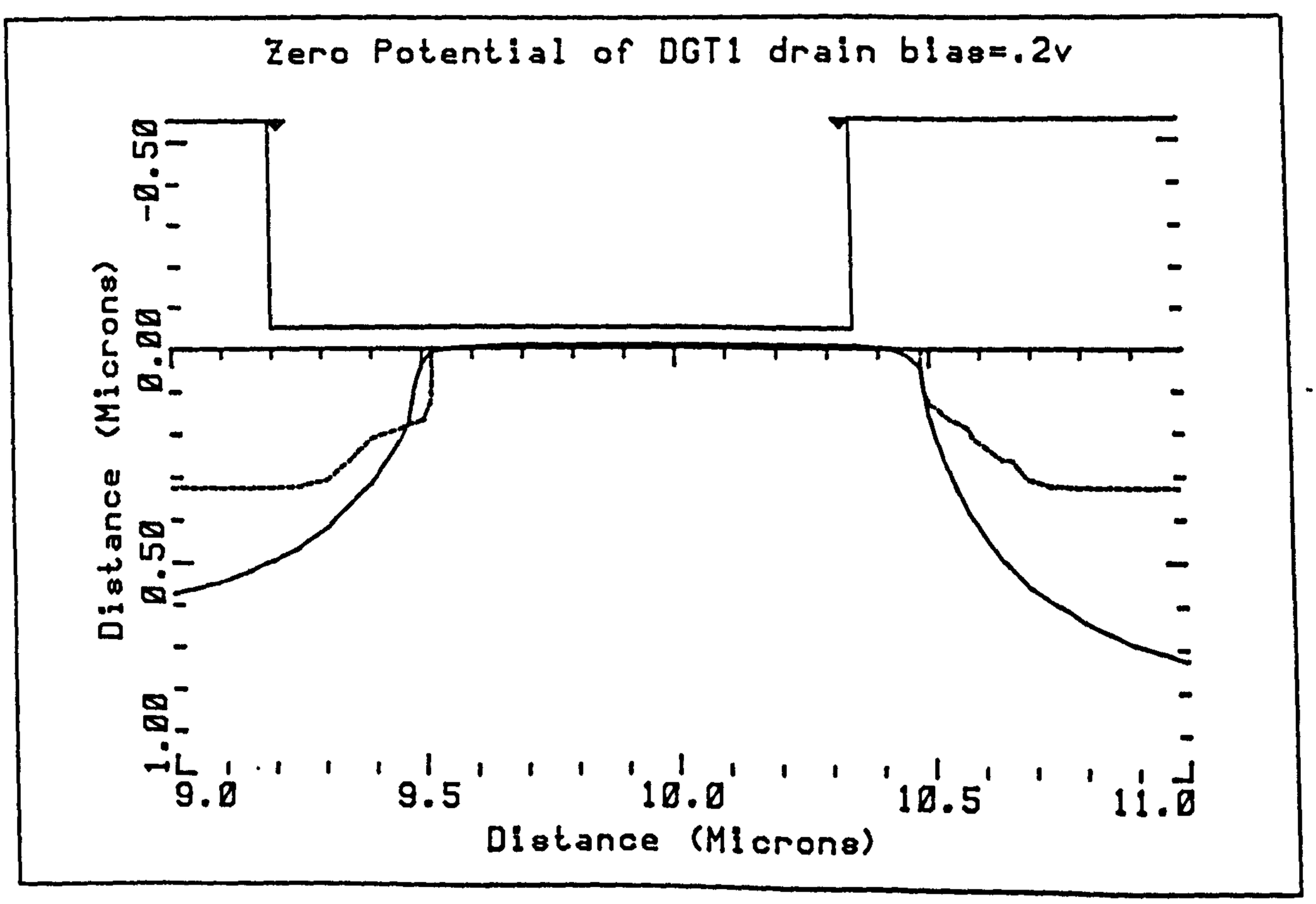


Figure 7.6, The drain depletion width corresponding to the characteristics in figure 7.5.

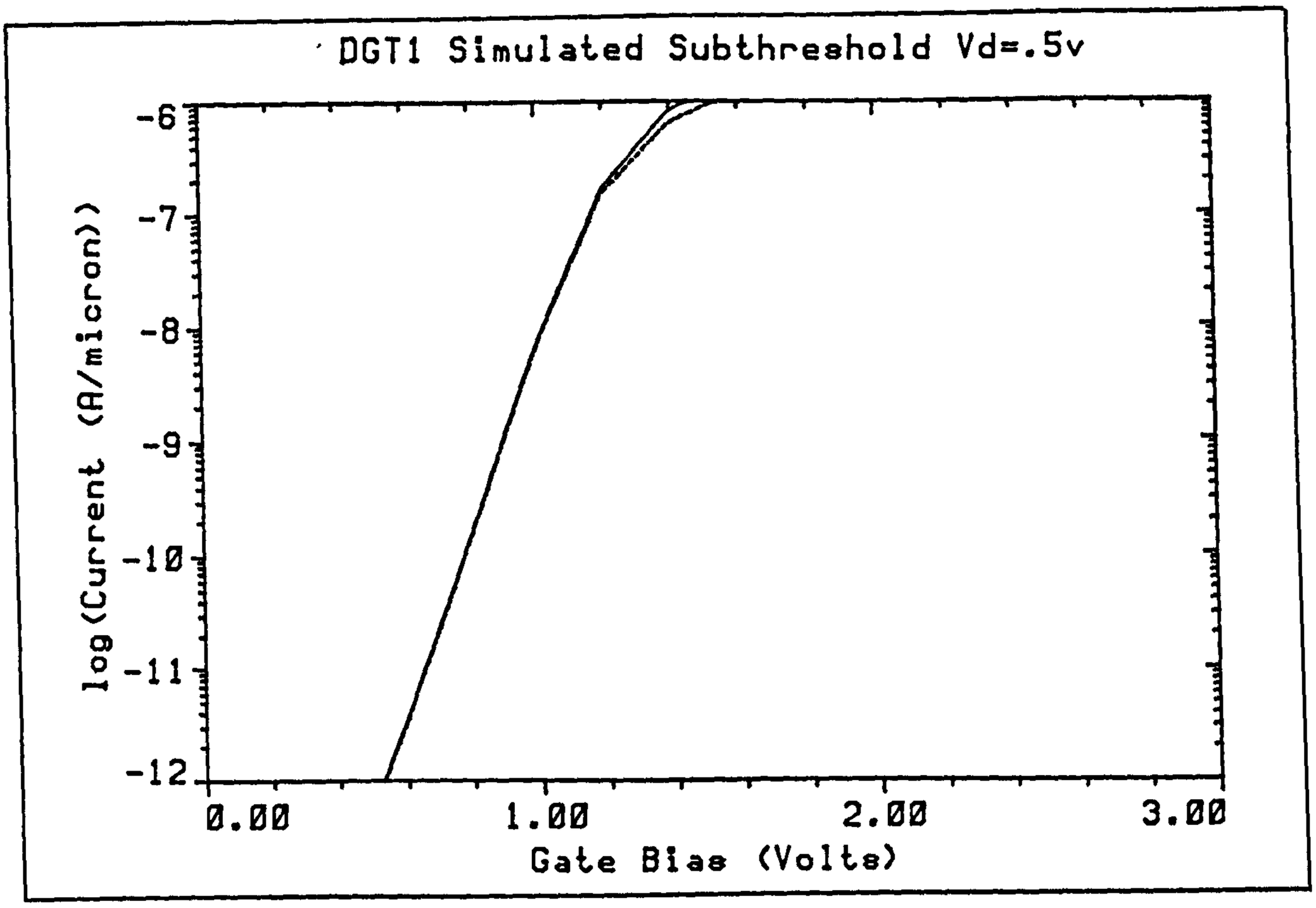


Figure 7.7, The subthreshold curve of a 0.15μm gap DGT, with $V_D = 0.5\ V$.

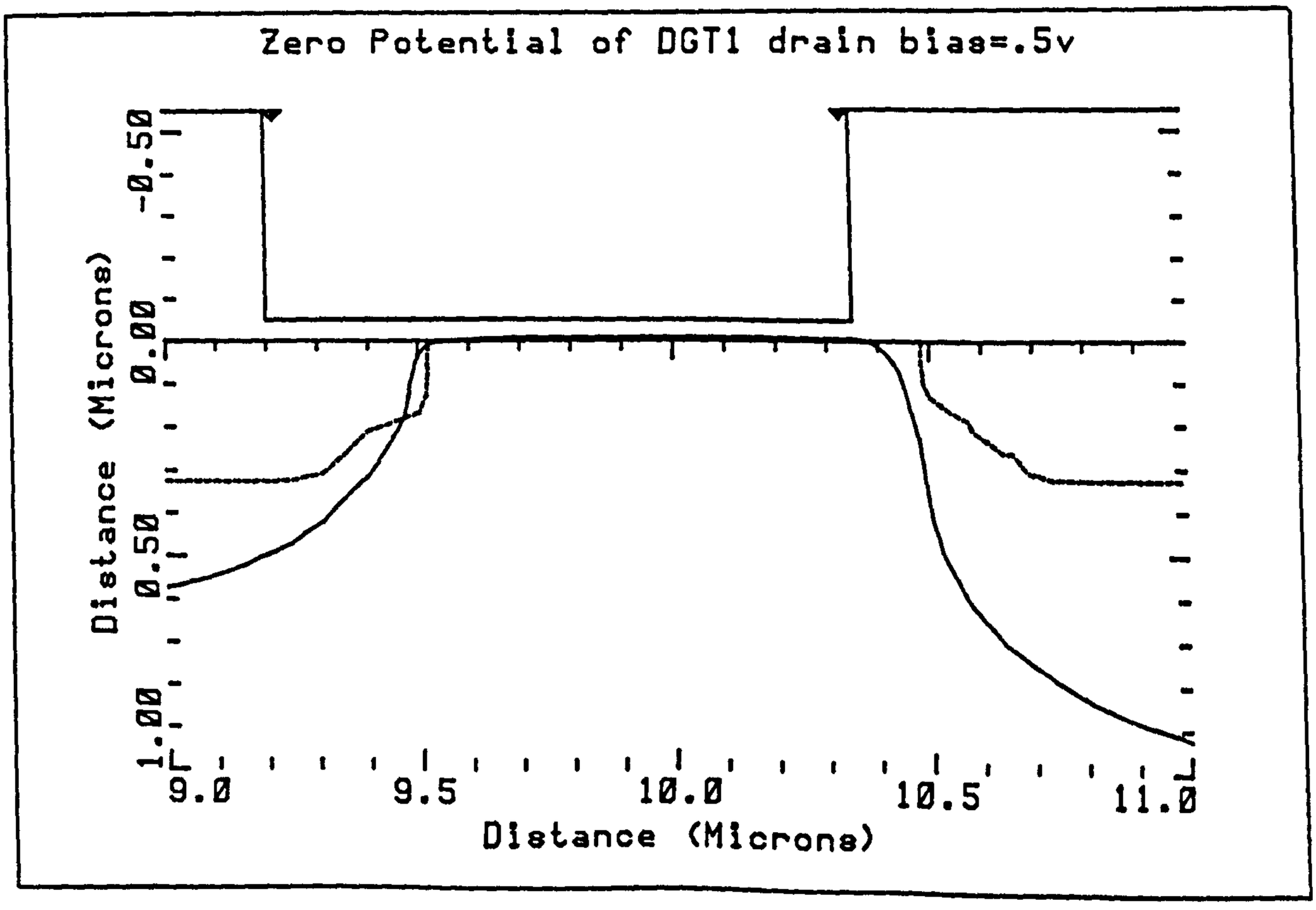


Figure 7.8, The drain depletion width corresponding to the characteristics in figure 7.7.

7.3. Experimental Results.

The column of transistors (9993) used for selection of the "golden die" in chapter six, which also had the magnitudes of its offsets calibrated by the optical shear microscope, was used to experimentally evaluate the DGT subthreshold swing improvements with increasing drain bias.

Subthreshold Swing vs. Drain Bias.

The subthreshold characteristics for each of the transistors in the chosen column were measured with fifty different drain biases each. The subthreshold swing was then automatically extracted at each drain bias for each transistor. Thus, the effective result of over twenty three thousand measurements can be seen in the plot of subthreshold swing versus drain bias for different size gap DGTs shown in figure 7.9.

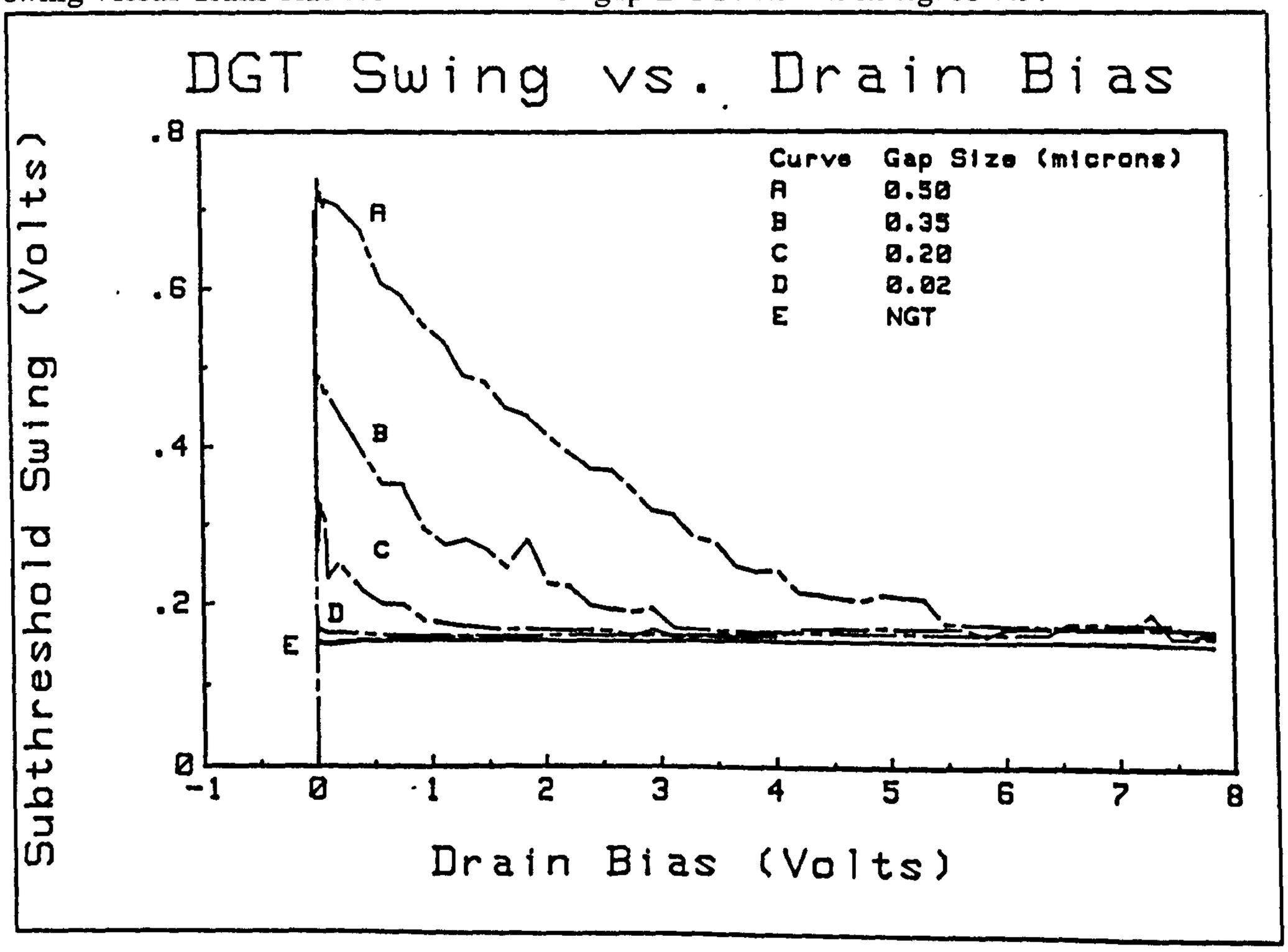


Figure 7.9, The effect of drain bias on DGT subthreshold swing.

It can be seen that the drain gap transistor's subthreshold swing does in fact improve with increasing drain bias up to the limit of the DGT swing being equal to the NGT

swing (For example the NGT intercept for curve "B" in figure 7.9 is about 3.4 V). The drain bias at which the subthreshold swing versus drain bias curve intercepts the NGT curve occurs at a higher value for larger gap sizes.

A confirmation of the responsibility of drain depletion boundary motion for the effect is shown in the SGT subthreshold swing versus drain bias plot in figure 7.10. There is no effect of the drain bias, above the initial $1-e^{-\beta V_{DS}}$ term, on the SGTs subthreshold swing. This is what one would expect because the drain bias does not significantly effect the source depletion width.

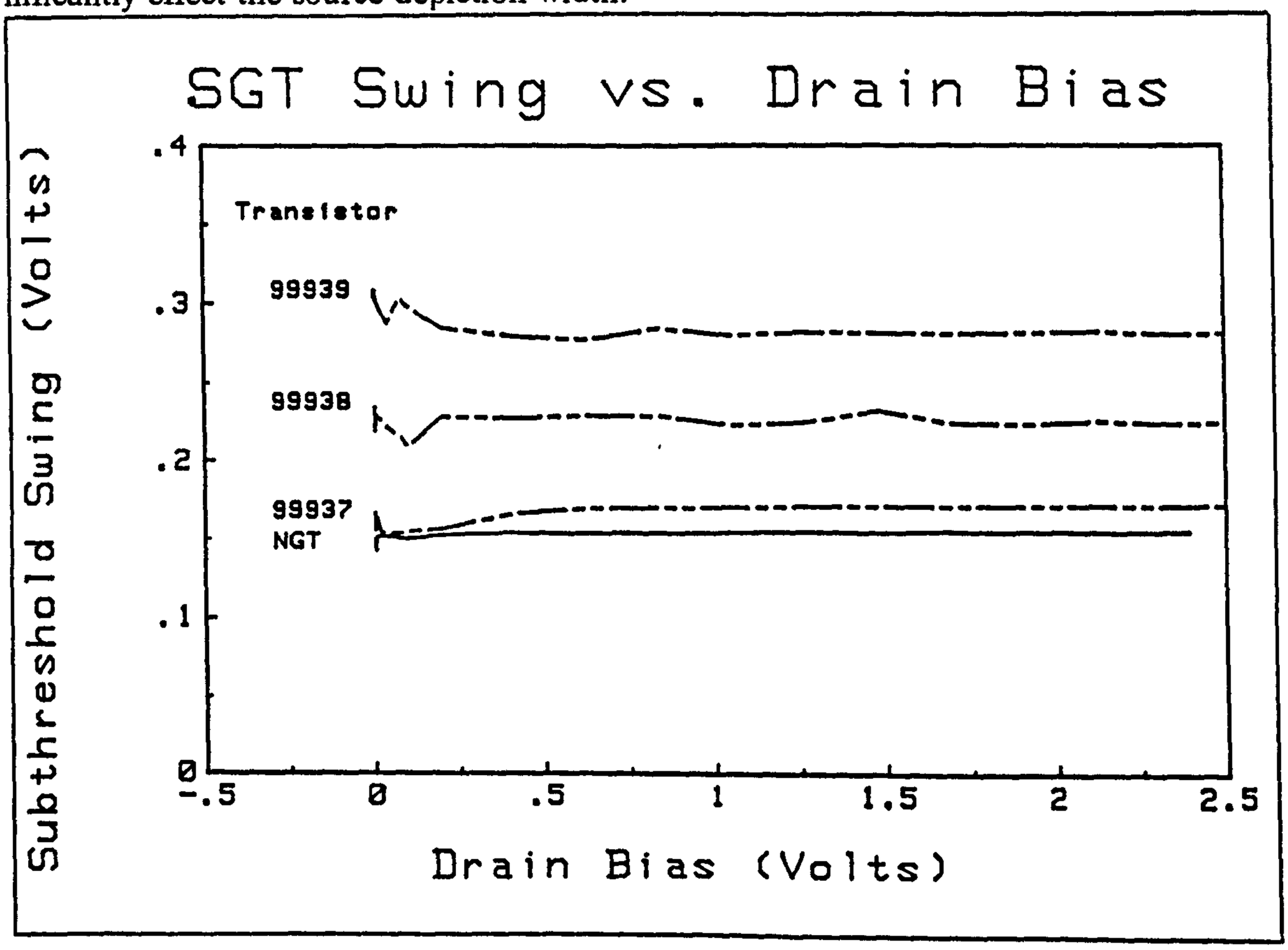


Figure 7.10, The SGT subthreshold swing versus drain bias.

NGT Swing Intercepts vs. Gap Size.

If the drain depletion width can be modelled by the simple abrupt junction formula presented in chapter 3,

$$W = \left[\frac{2\epsilon_s}{qN_B}(V_R + V_{bi}) \right]^{\frac{1}{2}} \tag{3.30}$$

and the NGT subthreshold swing intercept drain voltage (V_D) is taken for a set gap size (y_G) DGT then equation 3.30 can be re-written as,

$$y_G{}^2 = \frac{2\epsilon_S}{qN_A} V_D + \frac{2\epsilon_S}{qN_A}(V_{bi} - \Psi_S) \tag{7.1}$$

Therefore the drain bias at which the DGT's subthreshold swing intercepts the NGT swing should have a linear relationship with the square of the DGT's gap size.

The intercepts of the curves in figure 7.9 were calculated by finding the first drain bias where the swing was within a standard deviation of its final average value for that particular curve. The results are presented in table 7.1 and the resulting plot is shown in figure 7.11. The plot confirms that there is a linear relationship between the intercept drain biases and the square of the gap size.

Site	Gap Size (μm)	Gap^2 (μm^2)	Intercept Bias (Volts)
99956	0	0	Reference
99955	0.02	0.0004	0.73
99954	0.20	0.04	1.45
99953	0.35	0.122	3.43
99952	0.50	0.25	5.72

Table 7.1, The drain bias values required for NGT swing from DGTs.

Extraction of Surface Doping Concentration.

A best fit line was fitted to the data in table 7.1 by linear regression to establish the constants for a fitted equation.

$$y_G{}^2 = bV_D + a \tag{7.2}$$

The resulting constants are presented in table 7.2.

If equation 7.2 is compared to equation 7.1 it is obvious that,

$$b = \frac{2\epsilon_S}{qN_A} \tag{7.3}$$

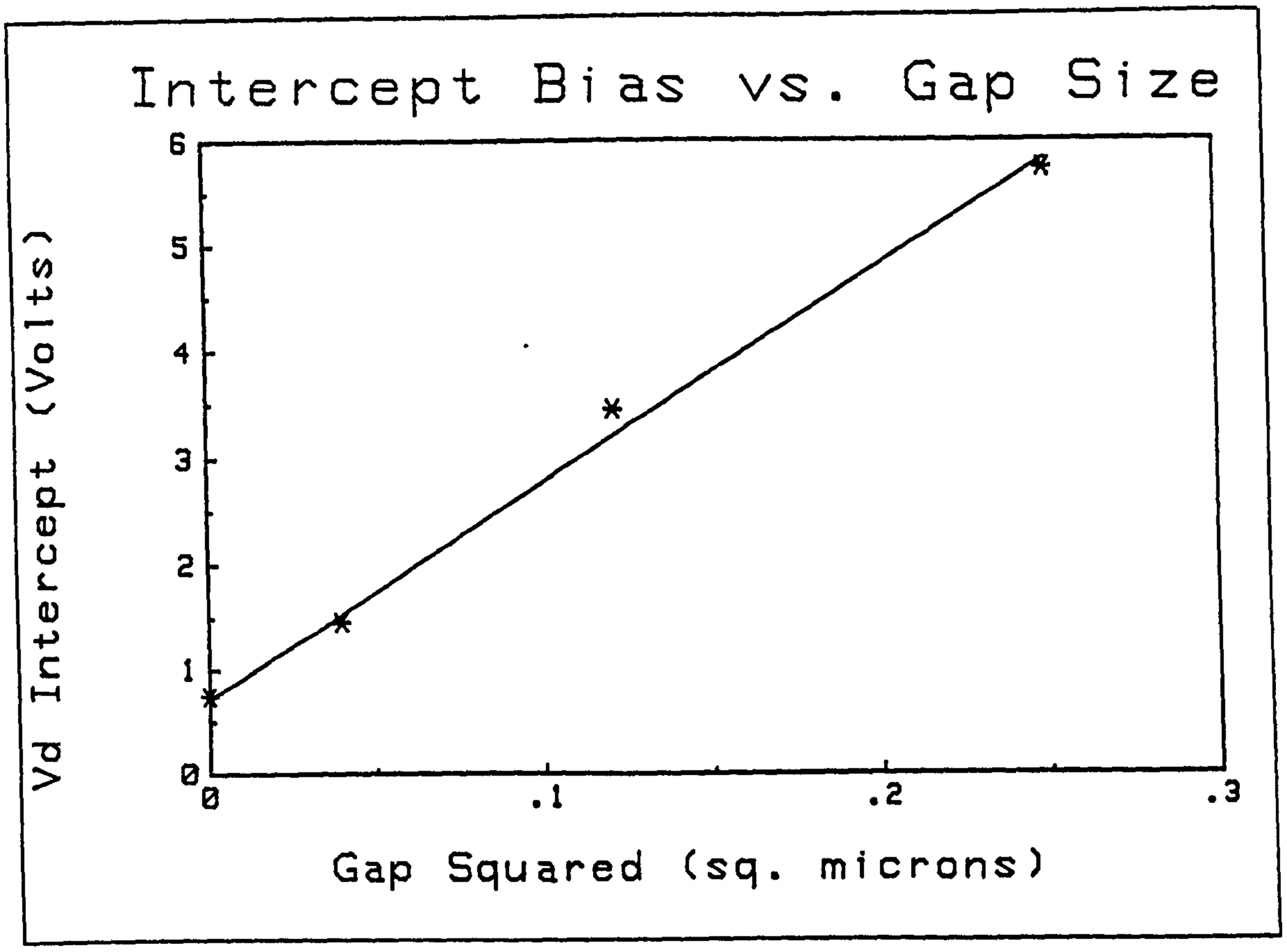


Figure 7.11, NGT swing intercept drain bias versus DGT gap size squared.

Coefficient	Value
Correlation coefficient r^2	0.9955
a (μm^2)	−0.035
b ($\mu m^2 V^{-1}$)	0.049

Table 7.2, The coefficients of a linear fit to the data in table 7.1.

or the surface doping concentration can be extracted by,

$$N_A = \frac{2\epsilon_S}{qb} \tag{7.4}$$

The solution for the surface doping concentration in this case yields $N_A = 2.69 \times 10^{16} atoms\ cm^{-3}$.

If that value is used with the appropriate oxide thickness in equations 3.116 to 3.120, the value for the swing of a NGT is obtained as 150 mV/decade which agrees

well with the observed value of 153 mV/decade.

The "a" coefficient from the linear regression also has a useful physical meaning.

$$a = \frac{2\epsilon_S}{qN_A}(V_{bi} - \Psi_S) \tag{7.5}$$

Equation 7.3 can be used to reduce this to

$$\frac{a}{b} = V_{bi} - \Psi_S \tag{7.6}$$

Ψ_S near inversion can be calculated using equation 3.39. Using the channel doping extracted above, $\Psi_S = 0.728\ V$ and $a/b = -0.72\ V$ which yields a built-in potential of $V_{bi} = 0.008\ V$. One would rightly question such a small value for the built-in potential as too small for that magnitude of doping. However the significance of the result is that at zero drain bias the gap size would also have to be zero, which indicates that the subthreshold alignment technique aligns to the built-in depletion width.

7.4. Chapter Summary.

This chapter presented a study of drain depletion boundary motion through the analysis of subthreshold characteristics of DGTs. The dependence of subthreshold swing on the drain bias of a DGT was modelled by the device simulator CANDE, which confirmed the swing improvements were due to drain depletion boundary motion. Experimental results for the subthreshold swing as a function of drain bias for a number of DGTs with different size gaps were presented. The results were further analysed to extract a value for the surface impurity doping concentration (N_A). The extracted value of N_A was used to calculate the theoretical swing which agreed well with the observed value.

7.5. References.

1. J.A. Serack, A.J. Walton, and J.M. Robertson, "The Effect of Device Geometry on IGFET Characteristics," *ESSDERC 87*, pp. 915-918, Bologna, 14th - 17th Sept 1987. (appended)

Chapter 8, Conclusions and Further Work.

8.1. Progressional Offset Technique.

The logic behind the progressional offset concept was validated by the success of the technique in producing the misaligned gate array. Without the progressional offset technique it would not have been possible to produce the variation in gaps in gate-to-channel coverage on a single wafer which made the misaligned gate experiment so successful. However, the non-linearity of the feature size to mask size caused by the exposure tool and photoresist system increased the error between alignment of features of differing sizes. The next generation of progressional offset theory should take that variation into account.

8.2. Misaligned Gate Experiment.

The misaligned gate experiment was also successful in both the development of a process to implement a progressional offset scheme to create source and drain gap transistors, and in characterising the resulting transistors.

Comparison to other Work.

The constructed metal gate progressional offset array had larger gap sizes than those available with sidewall spacer technologies but it also had more predictable gap sizes. To date the progressional offset technique is the only experimental fabrication technique capable of producing transistors with variations in gate-to-channel alignment with complete experimental controls for all other variables. It is the only technique capable of producing a gap in gate-to-channel coverage on a single size of the channel, and is the only experimental technique capable of producing various intentional gap sizes on a single wafer.

Experiment Design.

The unusual layout methodology of a database was particularly suited to the layout of a progressional offset array. The ability to perform algorithmic adjustment of the gate and channel size, and gate-to-channel alignment resulted in considerable time

savings.

The structures included in the array were satisfactory. The range of offsets in gate-to-channel alignment were sufficient to adsorb the alignment errors introduced by the lithography tools. The range of channel lengths were also satisfactory to enable the extraction of channel length dependent parameters. The reference transistors were very important for solving processing and environment related problems. They were also useful in demonstrating the consistency of the process over a die. The high density of reference transistors to test transistors (1:2) was not necessary, due to the fabrication process consistency, so future arrays could get by with fewer.

The experiment could have benefited from the inclusion of a large gate oxide capacitor to provide more information about oxide quality and silicon surface doping densities. Another structure that would have been useful would have been some form of optical vernier for confirming the alignment offsets between the gate and channel.

The designed microfabrication process was successful but could be improved by using a different channel implant mask (such as high temperature photoresist) or by providing structures for better end-point detection for the etching of the polysilicon channel mask, in order to avoid damage to the gate oxide adjacent to the channel.

Water Vapour Charging.

The water vapour charging problem was certainly one of the most difficult challenges of the entire experiment. Chopper stabilisation along with pyrox passivation brought the effect under control for experimental measurements but another solution would be necessary to obtain stable circuits using DGTs or SGTs.

Electrical Alignment Technique.

The best electrical measurement technique to detect SGTs and DGTs was to compare the subthreshold curve for the transistors in a normal mode against the curve with the source and drain connections interchanged. The voltage difference between the voltages required to achieve a set current level on the two curves gives an immediate indication of the magnitude and position of the gap. The technique is absolute and a totally aligned device has a voltage difference of zero. The sign of the voltage

difference for transistors with gaps in gate-to-channel coverage indicates which side of the channel (source or drain) the gap is on.

The subthreshold method proved to be better in all ways to any method using threshold voltage, transconductance, or comparisons of saturated drain current curves. The subthreshold method is also a general technique that could be used to detect asymmetric transistors that have a gap in gate-to-channel coverage on one side of the channel.

Physical Verification.

Optical verification of relative gate-to-channel positions was possible but had an error approximately equal to the offset step size which reduced the merit of the verification.

An SEM has sufficient resolution to calibrate the gap sizes but unless there is an accidental marking of the channel, or an intentional alignment vernier has been provided, the SEM is unable to locate the channel in a plan view which makes calibration of the gap size impossible.

A cleaved or cut cross-section through an asymmetric channel which has been etched to delineate the source-drain junctions allows an SEM to easily identify the channel section but the cross-sectioning technique often alters the shape and position of aluminium gates and so negates the alignment information. The cross-section technique may be more successful with polysilicon gates because the harder material will maintain the gate shape and position during the cross-sectioning.

Verifying the position of an aluminium gate relative to a channel would be more successful if physical alignment test structures were incorporated in the design.

8.3. Asymmetric Transistor Characteristics.

The electrical behaviour of an IGFET with a gap in gate-to-channel coverage on the drain side of the channel (DGT) is characterised by an increased threshold voltage, reduced transconductance, an extended linear region, drain voltage dependent subthreshold swing and in the case of large gaps, very poor saturation behaviour.

The electrical behaviour of an IGFET with a gap in gate-to-channel coverage on the source side of the channel (SGT) is also characterised by increased threshold voltage and reduced transconductance. It has a markedly increased subthreshold swing but the swing is independent of the drain voltage. The magnitude of the voltage swing is increased with increasing gap size. The SGTs show reduced drain saturation current compared with similar sized normally gated transistors. The slope of the saturation region of an SGT decreases as the gap size increases, which is thought to be due to the dominance of the incompletely gated region of the channel.

SGTs which are nearly completely gated show a gate-voltage-dependent series resistance which is also larger than the series resistance for a similar NGT. Similar DGTs do not exhibit increased series resistance which is thought to be prevented by the expansion of the drain depletion region.

8.4. Drain Depletion Motion.

The subthreshold swing of a DGT is affected by both the magnitude of the drain gap and the drain bias. The larger the gap the greater the swing voltage.

Two dimensional numerical device simulations have shown that the swing of a DGT decreases to the value appropriate to a normally gated transistor once the drain depletion region extends across the gap in gate-to-channel coverage and under the gated portion of the channel.

There was a good experimental fit of the abrupt junction depletion width model by the drain bias required to reach the NGT magnitude of swing, and the square of the gap size, for a number of drain gap transistors. The abrupt pn junction model is therefore useful for the source and drain-to-channel regions.

Further analysis of that data provides a way to extract surface doping concentration. When the extracted value was used in formulae to predict subthreshold swing, a value comparable to the observed value resulted. That allows the conclusion that the experimental method developed here for the analysis of drain depletion boundary motion can be used for both surface doping concentration extraction and gap size calibration, but it does require the measurement and analysis of a large amount of data.

8.5. Recommendations for Further Work.

During this work a number of ideas were generated but they would have required much more time to pursue. They are briefly outlined here as possible research projects that could extend this work.

8.5.1. Applications for the Progressional Offset Technique.

The progressional offset technique clearly has the ability to allow finer tolerances in test structures and it can be applied to many problems. The evolution of the present project indicates that the concept has an important role to play in the future development of test structures.

Source-Drain Edge Effects.

An extension of this work on source-drain edge effects should be possible if polysilicon gates were used. Polysilicon can be etched with a resulting smaller edge variation than aluminium. As a result a finer step size could be used in the progressional offset array. The maximum resolution currently available on pattern generators would allow step sizes smaller than the sidewall spacer widths used in current processing. An array of transistors constructed with that resolution would allow experiments to conclusively determine the minimum overlap to avoid "weakly overlapped" effects with sidewall spacer processes. Smaller gap sizes would also be available to employ in further investigations of drain depletion edge motion.

A polysilicon gate progressional offset array would also allow passivation at high temperature, (eg with LPCVD silicon nitride), of the ungated region of the channel to avoid water vapour charging problems. The dimensions of polysilicon gated progressional offset arrays would also be easier to verify physically because the gates would be more resilient to damage during the cross-sectioning.

Another Ph.D. project has already been set up following this work to implement the polysilicon gate progressional offset array to study LDD transistors.[1] It has a multitude of variations on channel widths, lengths, and offsets, including both self-aligned and multilevel progressional offset gates. Underlying all the variations is a progressional offset step size of 0.10 μm.

"Width" Effects

The progressional offset technique could also be applied to the width dimension of IGFETs. An edge region of considerable importance to the operation of the IGFET is the width edge. The effect of that edge on the transistors characteristics depends on the transistor isolation method.

An on-going Ph.D. investigating the LOCOS isolation "bird's beak" region has taken advantage of the progressional offset technique to design test transistors on only the "bird's beak" region.[2] The intention is to characterise the region directly rather than to try and extract its effect from a normal transistors characteristics.

Another "width" opportunity exists for the progressional offset technique in analysis of edge effects in processes using trench isolation. An experiment is being considered using the progressional offset technique to position a gate edge near the trench-channel interface in order to examine its effect on IGFET characteristics.[3]

8.5.2. Asymmetric Transistor Circuits.

There are a number of characteristics of incompletely gated insulated gate field effect transistors that could result in useful circuit applications.

SGTs and DGTs as Environmental Sensors.

The ungated region of the SGT and DGT channels is extremely sensitive to charge accumulation thereby providing the opportunity for environmental sensors. The most obvious, and naturally occurring sensor is a water vapour detector which could be extended to be a humidity sensor. There would be the scope for a Ph.D. project in the analysis of that effect including both the calibration of the electrical "charging" effects to the amount of water vapour accumulated and the determination of the equilibrium of accumulated water vapour with atmospheric humidity.

Other sensors could also be possible if appropriate coatings were deposited over the ungated region of the channel to establish a charge when they react with some gas or vapour in the environment. Electric field probes and optical detectors should also be possible.

There is a great potential for more research into the area of gate gapped transistors as environmental detectors with the possibility of such research resulting in valuable new devices.

Analogue Circuits.

If a way can be found to eliminate the "charging" problems of incompletely gated field effect transistors without recourse to chopper stabilisation both DGTs and SGTs could be profitably employed in analogue circuits.

Large source gap transistors have a flatness of the saturation curves that rivals transistors many times longer. The minimum slope in saturation of the differential pair in an Op-amp circuit allows a wider common mode rejection ration which is a figure of merit for an Op-amp. Use of large gap SGTs may allow better Op-amps in smaller areas, but the increased series resistance might decrease their merits.

Large drain gap transistors behave like vacuum tube triodes or gate voltage controlled resistors. The DGT's extended linear region may be useful in some circuit designs.

Both DGTs and SGTs could benefit from more research in this area, but first a solution to to the "charging" effects must be found that does not depend on chopper stabilisation.

Alternative Technology for high speed switches.

Consider a transistor with both source and drain gaps in gate-to-channel coverage (SDGT). If such a transistor was used along with a depletion load transistor in the configuration shown in figure 8.1 they would form an inverter circuit.

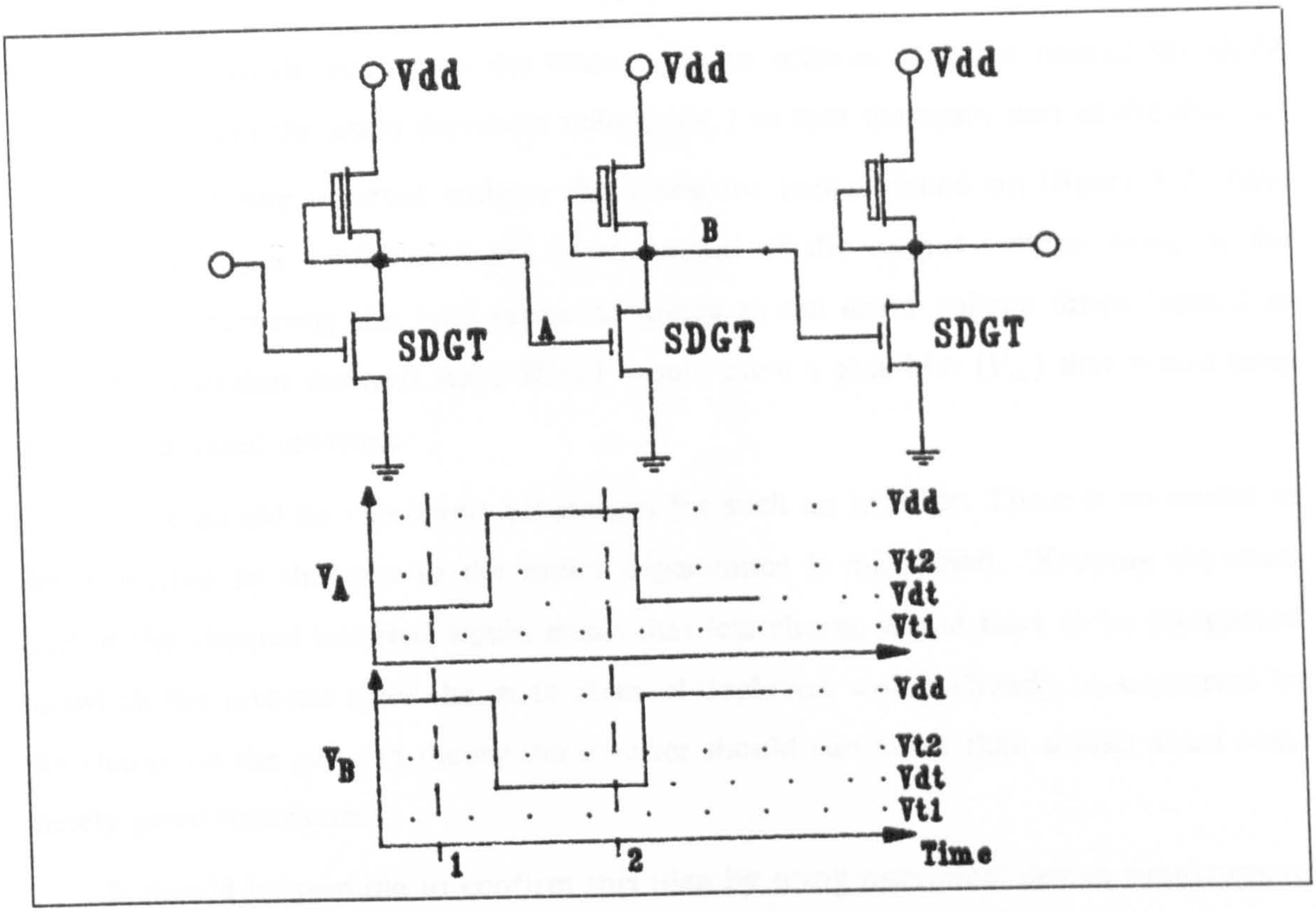


Figure 8.1, An inverter circuit application for a source and drain gap transistor.

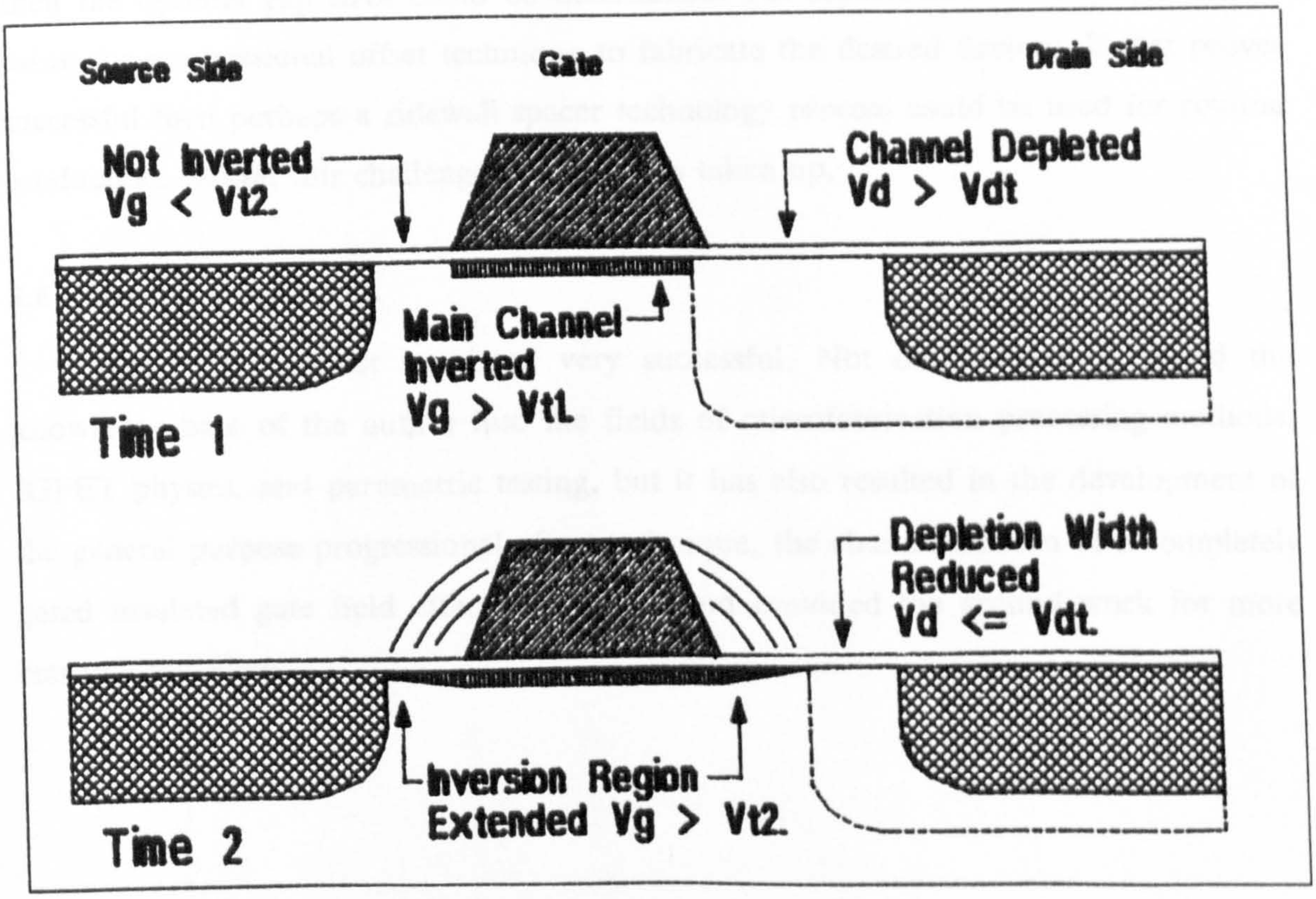


Figure 8.2, The source and drain gap transistor under biases in the inverter circuit.

That inverter circuit would use the source gap to achieve a higher overall threshold voltage (V_{t_2}) than the usual threshold voltage (V_t) so that the main part of the channel could be constantly inverted without the transistor being turned on (figure 8.2, time 1). The drain gap would limit the lower bound of the output voltage swing of the inverter, by increasing the total series resistance as the drain voltage drops (time 2 in figure 8.2) so that the next stage SDGT would have a gate bias (V_{dt}) that would keep the main channel inverted.

There should be two speed advantages for such an inverter. There is no source or drain overlap by the gate so the miller capacitance is minimised. Keeping the main part of the channel inverted would mean that less charge would have to be transferred to switch the inverter since the main channel depletion would already be supported by the charge on the gate. In theory the inverter should run faster than similar sized completely gated transistors.

It should be possible to confirm this idea by using numerical device simulators to model both the SDGT and load devices. If the theory is supported by the simulation then the optimal gap sizes could be determined. An experiment could be performed using the progressional offset technique to fabricate the desired devices. If that proves successful then perhaps a sidewall spacer technology process could be used for routine production. So far this challenge has not been taken up.

8.6. Overall Conclusion.

This Ph.D. project has been very successful. Not only has it expanded the knowledge base of the author into the fields of microfabrication processing methods, IGFET physics, and parametric testing, but it has also resulted in the development of the general purpose progressional offset technique, the characterisation of incompletely gated insulated gate field effect transistors, and provided the ground work for more research to follow.

8.7. References.

1. Neves, H., *First Year Ph.D. Report.*, University of Edinburgh., Edinburgh., June 1988..

2. Fallon, M., *Second Year Ph.D. Poster Session.*, University of Edinburgh., Edinburgh., June 1988..

3. Walton, A.J., Edinburgh., August 1988. Private Communication.

Section 9, Appendix.

The following pages contain some background information on the software written during the course of this Ph.D. Papers published during that research period are also appended.

9.1. Data Base Software.

A general purpose data base program was written during the course of this research. Its purpose was to be a flexible software tool for analysis of processing and transistor characteristics. It became a valuable tool for test chip layout, and processing experiment analysis. It was not used for analysis of transistor characteristics because Hewlett Packard Series 300 personal computers with two megabytes of memory, hard disk data storage, and fast computing speed became available in the EMF test lab. Since these were the same computers that controlled the test equipment, it was sensible to also perform the measurement analysis on them.

9.1.1. The Data Base Concept.

The concept of any data base is to provide the storage interface and data analysis features that are usually required in each analysis program, in a single versatile program. In this way maximum advantage of any programming effort can be had since all applications share the same program.

A versatile data structure is therefore necessary so that many different types of data can be represented. A common data structure is that of object orientation. Each type of data is named as an object and allowed a number of attributes. For example, data about the long term stability of a transistor's threshold voltage could have been stored in an object called threshold. That object could have had the attributes of the value of the threshold voltage, the time the measurement was take, the temperature, the other biases, and information about high field stressing between measurements.

Database commands could have then been used to, generate statistics about the collection of threshold voltage measurements, to select measurements between a certain range of temperatures, and to produce plots of the results.

9.1.2. "dbase" Structure.

The software written during this research was given the name "dbase". It was an object oriented database with objects called ENTITIES and attributes called FIELDS. An adaptive English like language was used to control the functions performed on the data. The data dictionary, or description of the objects to be stored, was treated as data itself and entered into predefined entities. An extensive help facility was also provided, only a summary of the main commands is given here.

9.1.3. Data Input

The only data input mechanism that was implemented was for data typed in from a keyboard, it was called ADD. In the example, it would have been used by typing "add 10 threshold" and the user would have then been prompted for the values of the attributes for ten entries into the threshold entity.

9.1.4. Data Manipulation.

The key to using the data in a data base is to select a set of entries from an entity and then to either alter the data or perform some form of analysis on it.

Finding Sets

Not too surprisingly the command for finding data was simply "FIND". It has qualifying adverbs which allow selection of a particular set of entries from a certain entity. In the example, the following expression "find threshold where temperature >= 19.5"† would have found a set (set1) where the temperature was greater than nineteen and a half degrees. The new set could have then been searched by referring to it by its name "set1". The UNION, INTERSECTION, EXCLUSION, and REMOVE commands provided other ways of combining sets or discarding entries from a set.

Once a set had been found the data contained within could be displayed on the terminal. Since the space required to display all the attributes at once could easily exceed the screen width, a very large virtual screen was used over which the user could

† If care was taken in defining data dictionary names, the abbreviation decoding human interface would have also be able to understand the same command written as "f t w t>=19.5".

move his physical screen. The commands to perform those operations were, LIST, TOP, UP, DOWN, RIGHT, and LEFT which take their English meanings. There were also commands to re-arrange the order of presentation of the attributes and to allow assignment of their values to variables for later use.

Data manipulation.

Data manipulation could be performed on the sets as well. Again the commands functionality was its English meaning. The commands were COPY, CHANGE, DELETE, and SORT. In the case of the CHANGE command quite powerful "adverbs" allow manipulation of the entries. In the example used previously, the temperature attribute could have been converted from Celsius to Kelvin by the command "Change 100 with temperature = temperature + 273". An extensive set of mathematical operations was allowed.

Statistics.

A sub command level could be entered by the command STATISTICS. At that level the mean, maximum, minimum, standard deviation, and root mean square of all numerical fields could be displayed.

9.1.5. Data Output.

Two mechanisms for getting output from the data base were implemented. Plots could be generated from numerical data, and reports could be generated from both numerical and text data.

Graphical Output.

The sub command level PLOT was used to enable graphical output to either a graphics terminal or to a file for later plotting. At that level, commands exist for describing the abscissa and ordinate axes, the graph title, the legend, and most importantly, for which attributes curves were to be drawn.

Report Output.

The report sub command level could produce files with almost any format from selected attributes in the entry. Provisions were included so that a standard "form" could be designed by the user, which the data base would then fill in for each entry in the selected set. In that way output could be tailored for input to other programs or presented in a human readable format.

9.2. Measurement Software.

All of the electrical measurements made during this research were performed using a Hewlett Package HP4062 Semiconductor Parametric Test System. In this section a short description of the HP4062 hardware and software, and a brief outline of the measurement software written for this research will be given.

A semi-automatic documenting package was written to help document the software. Over three hundred pages of documentation resulted, which obviously will not be included here. One subroutine will be included to highlight the measurement software and demonstrate the form of documentation.

9.2.1. HP4062 Description.

A Hewlett Package HP4062 semiconductor parametric test system includes both the hardware and analysis software to analyse most standard semiconductor devices.

Hardware.

The HP4062 hardware includes subsystems for DC measurements, capacitance and conductance measurements, switching and interconnecting, and a computer for system control, data storage, and analysis.

The DC measurement subsystem contains four source-monitor units (SMUs), two extra voltage sources, and two extra voltage monitors. Although there are some restrictions the system was capable of sourcing voltages from ± 1mV to ± 100V and currents from ± 1pA to ± 100mA with an accuracy of better than ± 0.3%. Voltages and currents can be measured over the same range with similar accuracy.

The capacitance measurement subsystem can measure capacitance at a fixed frequency of 1 MHz over a range from 0.001 pF to 1.200 nF with four digit resolution. There are restrictions on the maximum offset capacitance which can be accommodated on each range.

A remote switching matrix, and switching matrix controller, allow computer directed interconnection between any source or measurement unit and up to 48 pins near the device under test. Since the wafer prober used for this work was smaller than the switching matrix head, half metre coaxial cables were used to connect the switching matrix to the probecard.

Initially the system was controlled by a 200 series computer, but at the time of writing the system had evolved to a series 300 computer containing three megabytes of memory. The computer was directly connected to a forty megabyte hard disk drive with built in floppy disk drive, a graphics plotter, and a "Thinkjet" printer. There was also indirect (RS232) connection to the departmental computers and a vast array of peripherals.

Software.

The HP4062 system was supplied with software that could handle most production semiconductor devices. However, due to special needs of the offset gate transistors, only the very basic software could be used in this research.

One of the most important subroutines was the "Connect" command which specified which testing ports were to be connected to which pins. Careful definition of variables allowed statements such as "Connect(SMU1,drain)" which makes interconnection a trivial problem.

After the interconnections were made, the "Force_i" or "Force_v" commands could be used to force a current or apply a bias to a pin on the test transistor. Parameters for, the connection pin, the magnitude of the current or bias, the corresponding compliance level, and the range were required. Similar subroutines allowed the measurement of current or voltage at the other pins. A combined subroutine "Force_meas" allowed the forcing of a current and measurement of the resulting voltage, or the application of a bias and measurement of the resulting current, on a single pin.

A construct for making all connections, or applying all biases, simultaneously was provided for with the "Prepare" and "Execute" subroutines. With them, error conditions, such as accidently forward biasing of source-substrate junctions, could be avoided.

9.2.2. Offset Gate Measurement Software.

The main measurement software was controlled by a basic program which called a number of subroutines which interacted to perform measurements and analyse the data. The following sections briefly describe the software.

Measurement Subroutines.

The measurement subroutines could be classified into three groups, data gathering, data analysis, and data display. Data display could be considered a higher level subroutine because it would call the data gathering, data analysis and plotting subroutines (table 9.1).

Two separate classifications of data gathering subroutines were used. One group consisted of the routines "MSRIDVGDATA", and "MSRIDVDDATA" which measured "sweeps" of drain current verses gate voltage, and drain current verses drain voltage in straight forward manner. The range of the sweep and magnitude of the other biases were controlled by the calling parameters. The second classification provided the same functions but performed the measurements in such a way that any charge applied to water vapour at the edges of the gate was removed during the measurement. The subroutine "MSRIDVGS", whose documentation is presented in figure 9.1, is a good example of this technique. Following the comments through the listing it can be seen that after initialisation, connection setup, and fixed bias application, the gate voltage was applied for one bias and the drain current measured. The gate voltage was then returned to zero bias. The duration of the positive gate potential was then calculated and a mirrored negative gate potential was applied for the same duration. The process was repeated for each bias-current pair required. In that way the no charge accumulated on the water vapour surrounding the gate and the real transistor characteristics could be measured.

Subroutine **Msridvgs**

Description:

This routine gathers data for Id verses Vg curves using chopper stabilisation on the gate. A similar non-stabilised routine is Msridvgdata.

Calling Sequence:

Msridvgs(Vg(*),Id(*),N,Vgmin,Vgmax,Vd,Vb,Idlim,Rc)

Where:

Vg(*)

This array contains the gate voltages applied at each point measured. The units are in volts.

Id(*)

This array contains the measured drain current for each point. The units are in Amps.

N

This variable contains the number of points to generate in the arrays.

Vgmin

This variable should contain the minimum voltage to apply to the gate of the transistor during the gate voltage sweep. The units are in volts.

Vgmax

This variable should contain the maximum voltage to apply to the gate of the transistor during the gate voltage sweep. The units are in volts.

Vd

This variable contains the drain voltage to apply to the device during the gate voltage sweep. The voltage units are volts.

Vb

This variable contains the substrate voltage to apply to the device during th gate voltage sweep. The voltage units are volts.

Idlim

This variable contains the current compliance limit for the drain current while the device is under test. The units are in Amps.

Rc

The return code. For this routine remains unchanged.

Library: MS2

Uses:

Page 1

Figure 9.1, Documentation pages for a charge balancing measurement routine.

Subroutine Msridvgs

COMMON blocks:

Pinnumber

FUNCTIONS:

Smu

No SUBROUTINES.

Note:

Revisions:

Original Nov 22 1986, J. Serack:

Program listing:

```
521  SUB Msridvgs(Ug(*),Id(*),N,Ugmin,Ugmax,Vd,Vb,Idlim,Rc)
522  !------------------------------------------------------ IDVGS
523  ! Subroutine for collecting ID verses VG data
524  ! Uses a symetrical Ug about zero to avoid charging.
525  ! J. Serack NOV 22, 1986.
526  ! J. Serack MAR 2, 1987. Added variable frequency chopping
528  COM /Pinnumber/ INTEGER Source,Drain,Gate,Bulk
529  INTEGER Inttime
531  ! Initialize the system
532  Inttime=2              !Medium integration time
533  IF Idlim<1.E-6 THEN
534       Idrange=Idlim/100000
535  ELSE
536       Idrange=1.E-6
537  END IF
540  Init_system
541  Set_smu(Inttime)
542  ! Connect the pins to the SMUs.
543  Connect(FMSmu(2),Source)
544  Connect(FMSmu(1),Drain)
545  Connect(FMSmu(3),Gate)
546  Connect(FMSmu(4),Bulk)
547  ! Apply the constant voltages
548  Force_v(Bulk,Vb,20,.1)
549  Force_v(Source,0,20,.1)
550  Force_v(Drain,Vd,20,Idlim)
551  Force_v(Gate,0,20,.1)
552  ! Scan the gate voltage and make the measurements
553  FOR I=0 TO N-1
554     Vg(I)=Vgmin+(I-1)*(Vgmax-Vgmin)/(N-1)
555     Vdum=0-Vg(I)
556     Starttime=TIMEDATE
557        Force_v(Gate,Vg(I),20,.1)
558        Measure_i(Drain,Id(I),Idrange)
559     Endtime=TIMEDATE
560        Force_v(Gate,0,20,.1)
561     Endtime2=(Endtime-Starttime)+TIMEDATE
562        Force_v(Gate,Vdum,20,.1)
```

Page 2

Figure 9.1 (continued).

```
563     REPEAT
564     UNTIL Endtime2<=TIMEDATE
565         force_v(Gate,0,20,.1)
566   NEXT I
567   ! Disable sources
568   Disable_port
569   SUBEND
```


Figure 9.1 (continued).

Data analysis could be achieved through routines such as "MSRTHRES", "MSRCOND" and "MSRSWING" which use a measured array of data to determine the threshold voltage, the transconductance, and the subthreshold swing, respectively. Another very specific routine used the voltage difference between, a particular current on the subthreshold curve of a transistor, and the same current on the subthreshold curve with the source and drain pins swapped, to determine the degree of mis-alignment of the gate to the channel.

Data display subroutines provided the array of IGFET curves usually measured by using the data measurement, analysis, and plotting subroutines. Subroutines were written for the usual curves; drain current versus gate voltage, the family of drain current versus drain and gate voltage, subthreshold drain current versus gate voltage, and for unusual curves; drain current versus gate voltage for a number of drain voltages, and subthreshold swing versus drain voltage. Higher level routines also controlled a process of complete characterisation of the transistor and storage of the data onto disk for later analysis by other programs.

Chip Specific Software.

Chip specific software provided the relationship between test structure number and the position on the wafer. It also produced a description of the test structure based on its position, and defined the pin connection to the transistor terminals (which depends on the structure probed).

Plotting Software.

A number of general purpose plotting subroutines were written so that graphs could be displayed on the screen of the measurement controller, plotted to the local plotter, or dumped into a file for subsequent transmission to the departmental computer. A short description of this software is given in table 9.1.

Subroutine	Function
MSRAUTO	Automatic axis scaling.
MSRDATE	Determines the current date.
MSRDEF	Define default plotting parameters.
MSRDRAW	Plots the axes, title, and labels.
MSRFNSH	Finishes the plot. (Puts pens away.)
MSRLABEL	Defines labels to place on the plot.
MSRPFILE	Provides initialisation of a plot file.
MSRPLOT	Plots the data.
MSRTITLE	Defines the graph title.
MSRXAXIS	Defines the attributes of the abscissa.
MSRYAXIS	Defines the attributes of the ordinate.
MSRLOAD	Load a data array from disk.
MSRSAVE	Save a data array to disk.
MSRCORRELATE	Calculates the correlation coefficient between two arrays.
MSRDIFF	Determine the first and second differential of an array.
MSRLINE	Calculates a tangent to a curve at its maximum slope.
MSRVARIANCE	Determines curve matching by the variance of means.

Table 9.1, Software for producing graphs.

9.3. Automatic Prober Control.

At the start of the experimental measurements for this research the Edinburgh Microfabrication Facility had only one fully-automatic wafer prober in the test laboratory.† It was a teledyne TAC prober. It had built in hardware to provide fixed pattern scans of wafers for probecards equipped with edge sensors, but it did not have any mechanism for computer control. Computer control was required for implementation of the experimental tests, so modifications were made. A complete set of

† The others were manual probers.

documentation of the modifications is in the EMF report Eu600, which is available from A.J. Walton.

Hardware Modifications.

The existing control logic of the prober was TTL integrated circuits. The modifications included interception of stepper motor driver control signals, which were replaced by signals from a new board which was under computer control, and interfaces to original clock generation and direction control circuitry. The new board contained six parallel input counters, (three for each axis), six byte wide data latches, and demultiplexing control logic. In all there were twenty additional integrated circuits. The parallel input ports of the counters were attached to the latches which in turn were connected though the multiplexing circuitry to a commercial IEEE to parallel bus interface. The parallel bus interface was also connected to the direction control switches, and had taps to existing motion detection signals. In that way any IEEE compatible computer could control the prober. In this case a Hewlett Packard series 200 or 300 computer was used.

Low level control software.

Three levels of software were written to control the prober. The lowest level was the subroutines responsible for toggling ports on the parallel interface to alter the state of the prober circuitry. Table 9.2 lists these routines.

Subroutine	Function
Prbinit	Initialises software variables, sets up hardware ports.
Prbleft	Signals for a move to the left.
Prbright	Signals for a move to the right.
Prbup	Signals for a move to the front.
Prbdown	Signals for a move to the back.
Prbchuck	Vertical chuck motion control.
Prbxsize	Programs x direction counters.
Prbysize	Programs y direction counters.
Prbink	Signals the prober to firer the die inker.
Prbdh	Converts a decimal integer to a hexadecimal string.
Prbhd	Converts a hexadecimal string to a decimal integer.
Prbdata	Reads a spare four bit data byte.
Prbdone	Checks for move completion.
Prbedge	Reads the edge sensor status.
Prbxpos	Returns the probers x position.
Prbypos	Returns the probers y position.

Table 9.2, Low level prober control routines.

The example of the programming the x-axis step size is typical of these routines. Referring again to an extract from the documentation, in figure 9.2, the program structure can be described. First the required numerical step size was converted to a three character hexadecimal address. Then each of the three counters control latches were first selected, and then the corresponding character from the hexadecimal address was sent. Finally all the latches were de-selected before the routine was exited.

Die and sub-site addressing.

Higher level software was be built on the foundation of the direct control software. The construct of using four integers to specify a particular test structure on a particular chip was used here.

Subroutine Prbxsize

Description:

This routine is used to set the step size in the X direction. The step size is in units of 10 microns and may very from 1 to 1665.

Calling Sequence:

Prbxsize(Size,Rc)

Where:

Size

This variable contains the required step size in units of 10 microns with a range from 1 to 1665.

Rc

Rc is the abreviation for the Return Code which indicates how successful the subroutine was in carrying out its function. The Rc value is set on the return from the subroutine. Some subroutines also use a non zero input Rc to indicate that they are not to write error messages to the screen. Other subroutines may pass an Rc value that is input to lower level routines.

For this routine the rc values have the following meanings:

Non zero entered suppresses printed error messages.
0 normal return code.
1 step size out of range.

Library: PRB

Uses:

COMMON blocks:

Prb

No FUNCTIONS.

SUBROUTINES:

Prbdh

Note:

Prbinit must be called at least once before any other routine.

If the step size is outside the range of 1 to 1665 an error message will be printed unless the subroutine is entered with a non-zero return code.

Revisions:

Original April 28, 1986, J. Serack

Program listing:

Figure 9.2, Documentation pages for example low-level prober software.

Subroutine Prbxsize

```
SUB Prbxsize(Size,Rc)
!-------------------------------XSIZE
! A program to change the X step
! size on the prober
! J. Serack APRIL 28 1986
COM /Prb/ INTEGER Xs,Ys,X1,Y1,Add,Chuck
CALL Prbdh(Size,A$,Rc)
IF Rc<>0 THEN SUBEXIT
OUTPUT Add;"X2"
ENTER A$ USING "10,2X";B$
OUTPUT Add USING "10,10";"L",B$
OUTPUT Add;"X1"
ENTER A$ USING "1X,10,1X";B$
OUTPUT Add USING "10,10";"L",B$
OUTPUT Add;"X0"
ENTER A$ USING "2X,10";B$
OUTPUT Add USING "10,10";"L",B$
OUTPUT Add;"XF"
Xs=Size
SUBEND
```


Figure 9.2 (continued).

All the medium level software in table 9.3 used the variables xdie and ydie to specify which die on the wafer and xsite and ysite to specify a test structure within that die.

Subroutine	Function
Prbmove	Move to the specified structure on the specified die.
Prbset	Allows calibration of chuck position to the wafer position.

Table 9.3, Medium level prober software.

The actual relationship between these parameters and the position on the probers x and y axis depended upon chip specific functions.

Wafer Scanning.

The top level prober software could control a serpentine scan of a wafer. The wafer scanning subroutine used the underlying software and the valid site checker to allow automatic wafer probing without an edge sensor on the probecard. Edge sensors are difficult to implement on sub-die site probing test patterns. Table 9.4 lists the high level software.

Subroutine	Function
Prbscan	Controls a serpentine scan path for testing a wafer.
Prbvalid	Checks a die coordinate for being a valid Optimetrix exposure site.

Table 9.4, High level prober software.

9.4. Published Papers.

The following papers were published during the course of this Ph.D. program with the permission of the program supervisor. The appended papers are photocopies of the original artwork and their inclusion here is with the permission of the principle author and satisfies all copyright requirements.

THE APPLICATION OF A NOVEL EXPERIMENTAL TECHNIQUE TO INVESTIGATE HOT CARRIERS IN MOSFET'S

J. A. Serack, J. M. Robertson, A. J. Walton.

1. Introduction.

Trends in VLSI process design have produced transistors with small dimensions and narrow depletion widths. As a result the so called small geometry transistors have much higher electric fields than the larger ones. These fields are capable of heating the charge carriers to a sufficient energy which enables them to penetrate the silicon-silicon dioxide barrier. Once the carriers enter the gate oxide they may find their way to the gate or get trapped in the gate oxide. Both the passage of carriers through the oxide and trapping cause degradation in transistor performance.

Study of hot carrier generation, transport, and trapping is hampered by the difficultly, if not impossibility, of directly observing the hot carriers. As a result one is forced to come to conclusions about injection mechanisms from observations of transistor characteristic variations. In some configurations, such as hole trapping in an n-channel MOSFET, no overall effect would be observed if only part of the channel is affected.

This presentation addresses a new transistor structure that could be very useful in studying possible hot hole trapping in n-channel MOSFET's.

2. The Drain Gapped Transistor.

The Drain Gapped Transistor (DGT), as shown on the right of figure 1, has a gap in gate to channel coverage in the drain region of the channel. The source is still completely overlapped as are the edges of the transistor. A DGT operates in a similar manner to the normal IGFET except that the area of the channel under the gap has to be turned on by the fringe fields at the side of the gate. Since the magnitude of such a field is much lower than that directly under the gate the gapped region of the channel turns on last, and therefore dominates the transistor characteristics.

Since the DGT characteristics are dominated by the part of the channel near the drain, and this is also the area most effected by the avalanche and channel hot electrons, the DGT is quite suited to the study of hot carriers. In particular a DGT should be able to detect any accumulation of hot holes in the gate oxide whereas this would not be the case for a normal transistor which is explained in the bottom two boxes of figure 1.

J A Serack, J M Robertson and A J Walton are with the EMF, Department of Electrical Engineering University of Edinburgh.

3. Experimental Procedure.

The DGT may appear as a valuable tool to study hot carriers, but without a scheme for manufacture with a close control over the dimension, it remains only a theoretical possibility. It is, however, possible to reliably fabricate short channel DGT's using the following procedure.

The technique is to use a non-self-aligned gate process and a linear array of transistors with systematic steps in gate to channel alignment. Independent of mask alignment, at least one of the transistors will align by chance, and once it is electrically detected the scheme of the array can be used to locate a suitably gapped transistor for the experiment.

Figure 2 shows the transistor array scheme that was used in this experiment. The length of the channels varied along the rows from 1.00 to 2.05 microns in steps of 150 nanometres. Five micron reference transistors were included on every third column. Down a column the drawn alignment of the gate to the channel was varied in steps of 150 nanometres from a drain gap of 0.75 microns to a source gap of 0.6 microns. The reference transistor were drawn with both the source and drain overlapped by 0.3 microns.

The fabrication used a modified n-channel LOCOS process. The starting material was <100>, 15 ohm-cm p-type silicon. Field implants of 2e13 atoms/sq. cm of boron at 130 keV were used and one micron of wet thermal field oxide grown. An implant of 7e11 atoms/sq. cm of boron at 40 keV and punch through implant of 2e12 atoms/sq. cm of boron at 130 keV were implanted through the gate oxide. Polysilicon was deposited and patterned by RIE for a channel stop, and doses of 8e13 and 8e15 of arsenic were implanted at 340 and 170 keV to form the source and drains. The polysilicon was removed, source drain contacts etched chemically and pure aluminium evaporated and patterned using RIE. All photolithography was was performed on a 10:1 Optimetrix wafer stepper from optically produced recticles. The resulting gate oxide thickness was 500 nanometres and the source-drain depths were 0.35 microns.

To detect the aligned device, the forward and reverse (source and drain swapped), subthreshold drain current verses gate voltage curves were compared. Since the electron injection at the source is a strong function of the gate electric field any drain gap causes a marked decrease in current when the source and drain are swapped. Moving down a column until the curves match one can locate the aligned device, then moving back up one can select a DGT with a 0.15, 0.30, 0.45, ... gap.

An HP4062 parameter measurement system and an automatic prober were used for all measurements.

4. Practical Considerations.

There are a number of practical considerations in this type of experiment. One of these is the problem of fabricating the transistor array. It was found that after the poly gate (channel stop) was removed, that the gate oxide was seriously eroded on each side of the channel making gate to source

and drain shorts possible. The problem was traced to the patterning of the poly gate by reactive ion etching and the low 5 to 1 selectivity of poly to gate oxide. The solution was to perform a very short touch up gate oxidation after the poly channel stops were removed. However, a better solution in the future would be to use a high temperature resist directly as the implant channel stop.

Another consideration that can be problematic is that of the DGT showing extreme sensitivity to any charge captured over the gap region. Any water vapour present in that region will quickly charge up under high gate voltages and make the transistor appear almost normal on subsequent runs. The first attempted solution to that problem was the application of photoresist as a passivation layer. The photoresist actually made the problem worse, and although the effect decreased after the resist was baked for a few days it still remained a problem. The next attempt at a solution was the application of pyrox over the wafer. This was more successful than the photoresist but the charging still remained a problem. It should be noted that in all of these trials the normal transistors did not seem affected, demonstrating the sensitivity of the DGT. The final solution and successful one, was to accept the fact that some charging of all materials could be expected and to modify the tests accordingly. A method of chopper stablisation has been employed on the gate. Any positive gate voltage that is applied for stressing or a measurement is matched with the same magnitude of a negative gate voltage. The period of chopping is 100 milliseconds.

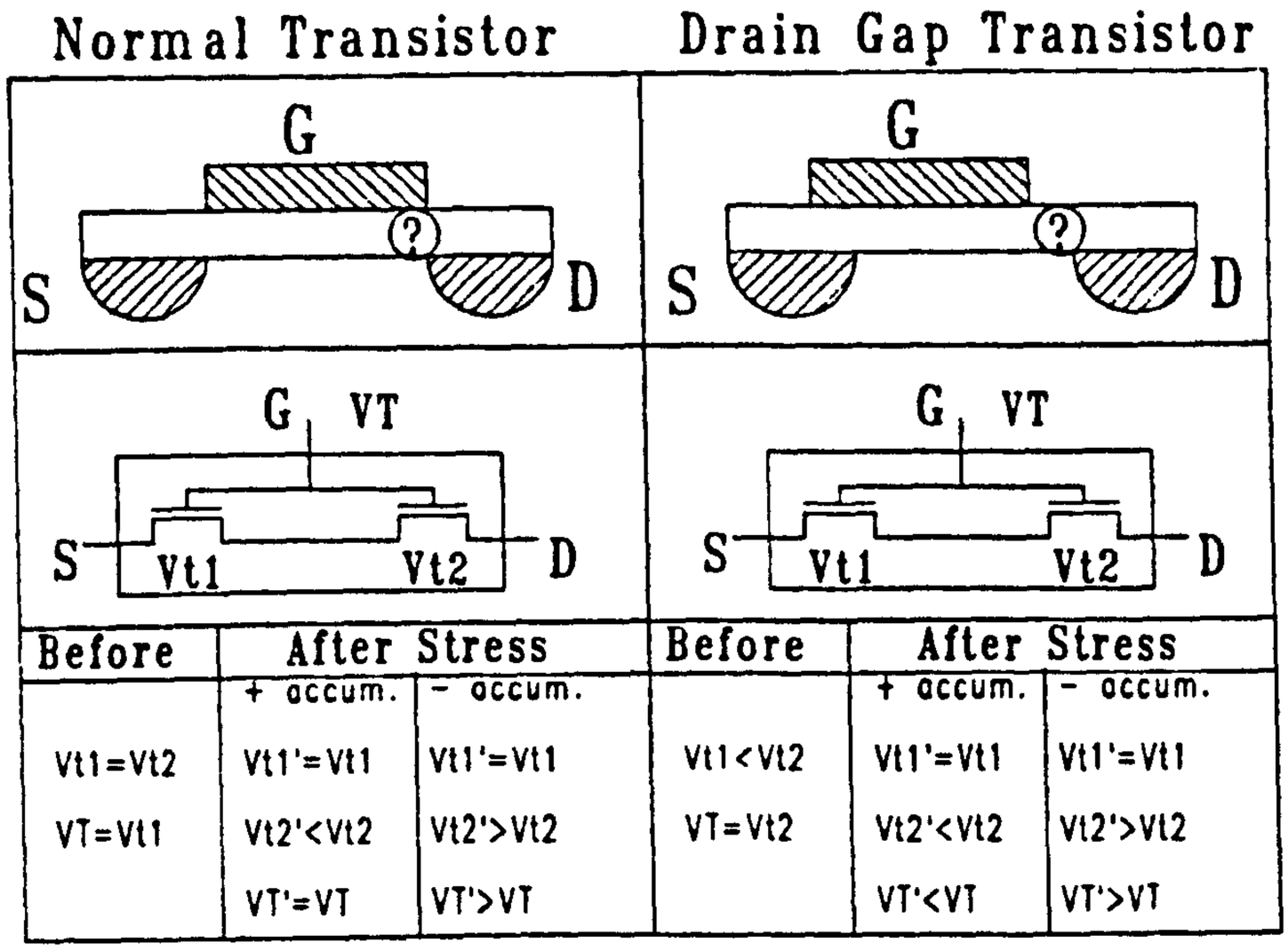
5. Preliminary Experimental Results.

A 1.3 micron long DGT and normal transistor where subjected to stressing in avalanche hot electron mode for 18000 seconds, effective time due to chopper stablisation is 9000 seconds, at Vd=8 volts and Vg=2 volts. That bias position corresponds to the maximum substrate hole current of 5 microamps.

Neither transistor showed any evidence of hot hole capture. Both the normal and the DG transistors showed a positive threshold voltage shift and lowered transconductance. The DGT showed an increase in threshold voltage of over fifty percent where as the normal transistor showed only a ten percent increase. A number of identical runs, on 1.3 and 1 micron devices, resulted in similar stress results.

5. Acknowledgements

This work was funded by SERC. J A Serack would also like to acknowledge financial assistance from Bell Northern Research(Canada), NSERC(Canada) and ORS(UK).



Before	After Stress	
	+ accum.	- accum.
Vt1=Vt2	Vt1'=Vt1	Vt1'=Vt1
VT=Vt1	Vt2'<Vt2	Vt2'>Vt2
	VT'=VT	VT'>VT

Before	After Stress	
	+ accum.	- accum.
Vt1<Vt2	Vt1'=Vt1	Vt1'=Vt1
VT=Vt2	Vt2'<Vt2	Vt2'>Vt2
	VT'<VT	VT'>VT

Figure 1. The DGT and its sensitivity to hot carrier trapping.

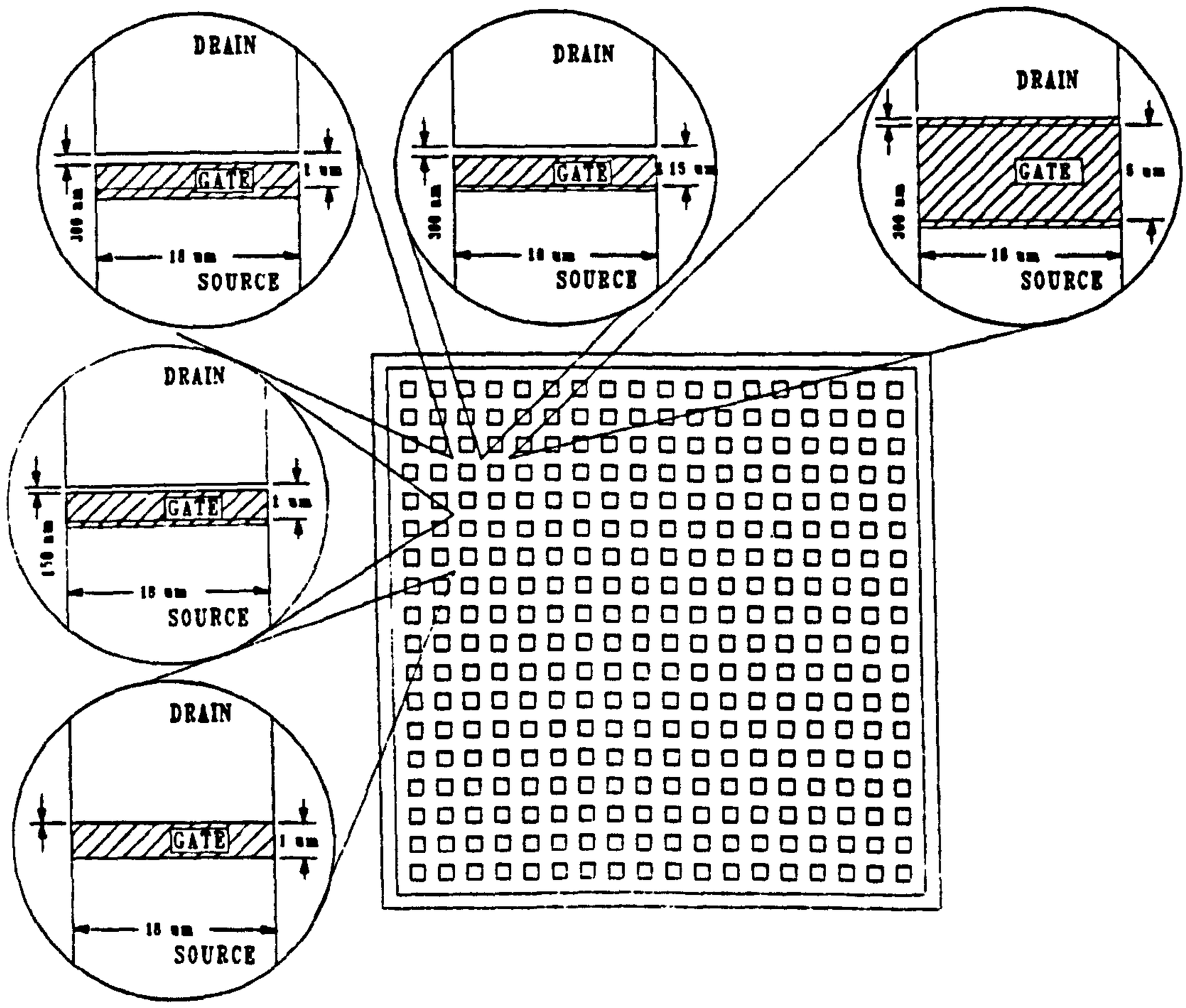


Figure 2. The Experimental Chip. Note the channel is systematically varied along the rows and gate to channel alignment is varied down the columns. Also note the inclusion of the 5 μm reference transistors.

The effect of device geometry on IGFET characteristics.

J. A. Serack, A. J. Walton, J. M. Robertson

Edinburgh Microfabrication Facility, University of Edinburgh, Edinburgh, Scotland.

A novel technique for the fabrication of asymmetrical incompletely gated transistors is described. The subthreshold characteristics of transistors fabricated using the technique are measured and conclusions concerning the mechanisms responsible for the observations are presented.

1. INTRODUCTION

Process designers have been using the primary relationship between gate capacitance and gate overlap of the source and drain for decades to maximise circuit speed by reducing gate capacitance. However little attention has been paid to how the correspondly reduced gate overlaps effect the D.C electrical characteristics of the transistors.

Recently asymmetries in transistor characteristics have been encountered [1-3] when the source and drain connections of the transistor were reversed. Gaps in the gate to channel coverage at either the source or drain end of the channel have been identified as the cause of the asymmetries. The gaps arise from the gate shadow cast during off-axis implantation, (which is universally employed to reduce channeling), coupled with a low thermal budget for the subsequent steps of the process. In order to study the effect that these gaps have on transistor characteristics, sidewall spacers have been used during source-drain implants together with variations in the drive-in times, to obtain transistors with variations in gap size [2].

Unfortunately other drive-in time dependent factors, such as the source and drain depth or the degree of channel implant activation, cause variations in the electrical characteristics of transistors fabricated with the above technique. Unpredictable variations between wafers also makes a controlled comparison between transistors with different magnitudes of gaps or overlaps impossible. The uncontrollable factors inherent in this approach make it only useful for obtaining a qualitative understanding of the effect. This paper presents a novel technique for the fabrication of transistors with a predictable range of gaps and overlaps upon a single wafer. Measurements performed on transistors constructed using this method are free from the factors discussed above.

2. EXPERIMENTAL SET-UP

In order to control the degree of gate overlap in this experiment the convenience of self-aligned gates had to be forgone and separate steps used to define the gate and channel. A fairly typical LOCOS isolation process was employed until after the source and drain implants were completed. What would have then been the polysilicon gate was stripped away and the contact holes cut. Aluminium was patterned by reactive ion etching to form the gate and contacts to the source and drain. The process used <100> silicon wafers and resulted in the following features; arsenic source-drains 0.35 μm deep, a channel impurity concentration of 8×10^{16} atoms/cc of boron, 600 Å thermal gate oxide, and 0.5 μm thick aluminium gates.

Although excellent alignment errors of less than 0.3 μm for the gate to channel were obtained using a 10:1 reduction direct to wafer stepper for the photolithography, the small gate gaps and overlaps are dictated by the chip design.

The chip layout, shown schematically in figure 1, was realised using a relational database to first define a single transistor and then replicate it to form an array of transistors. Then algorithms were applied to adjust both the channel and gate sizes and the gate to channel alignment at each site. This resulted in an array of transistors whose nominal channel lengths vary across a row from 1.15 to 2.05 μm and whose gate to channel alignment varied down a column from 0.75 μm drain gaps to 0.45 μm source gaps, in steps of 0.15 μm. Reference transistors with 5 μm drawn channel lengths and 0.3 μm overlaps were also included. The choice of 0.15 μm as the step size was made after a previous experiment showed the metal edge roughness was of this magnitude for 0.5 μm thick aluminium.

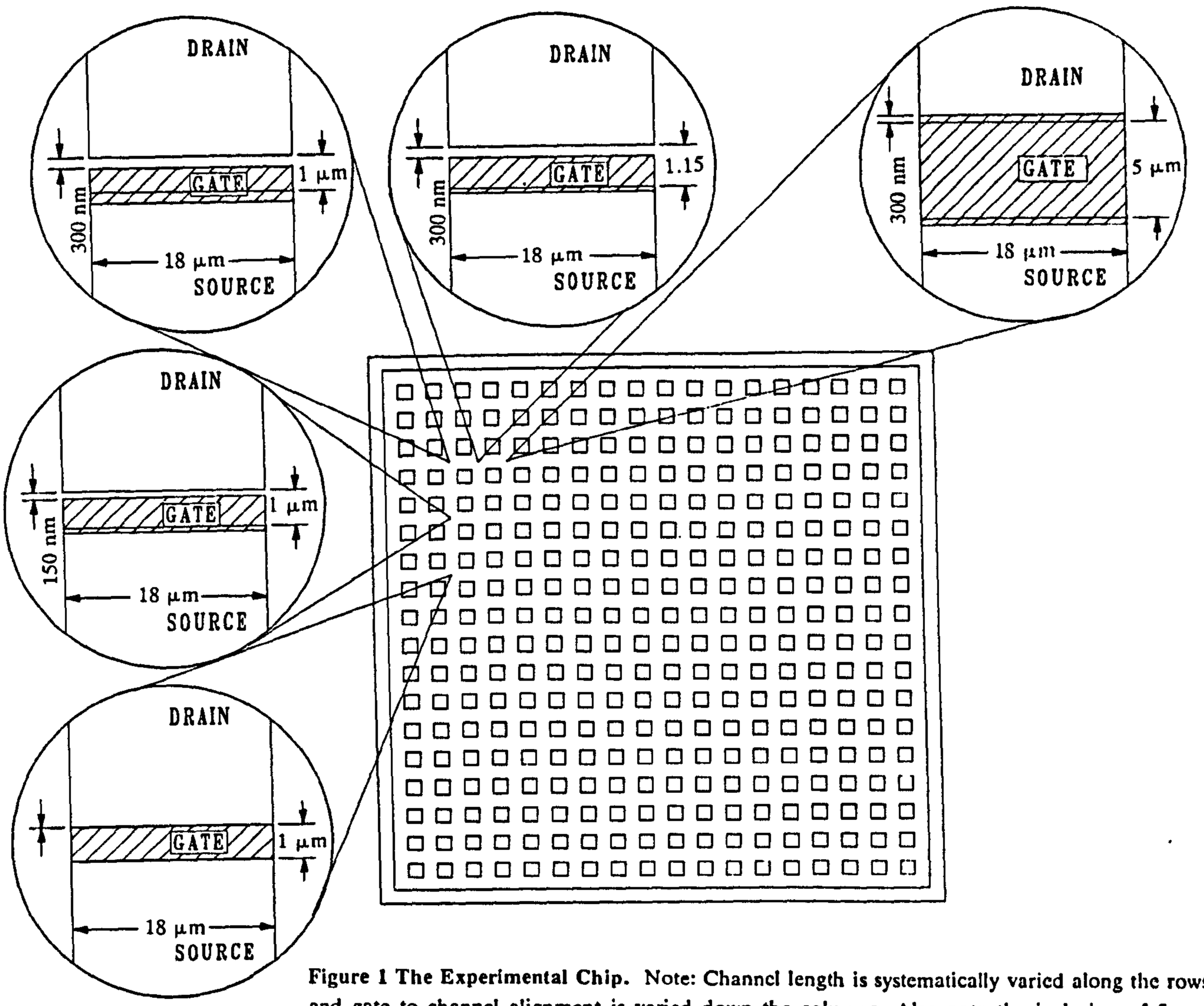


Figure 1 The Experimental Chip. Note: Channel length is systematically varied along the rows and gate to channel alignment is varied down the columns. Also note the inclusion of 5 μm reference transistors.

3. SITE SELECTION and MEASUREMENT CONSIDERATIONS

Once a particular channel length column on a die was selected for characterisation, alignment identification was accomplished by comparing both the normal and source-drain reversed subthreshold curves of each transistor in the column. The transistor or transistors that have coincident curves in both configurations have complete gate coverage of the channel. Figures 2 and 3 show this as well as other characteristics for two transistors from a column. The transistor in figure 2 is incompletely gated with a large drain gap which is evident in the asymmetry of the subthreshold curves (figure 2a). The transistor in figure 3 is from a couple of sites farther down the same column and is aligned. Comparisons of the other characteristics show similar results to that reported in reference [2].

Once the subthreshold curves have been used to determine the first aligned transistor in the column, the structure of the layout can be used to select transistors with a particular gap or overlap for further measurement.

One important consideration when measuring transistors with gaps is their sensitivity to charge accumulation over the gap in gate to channel coverage. If the gap is uncovered water vapour can be charged by the gate and the transistor can appear to change into a non-gapped transistor after a few measurements. Even with a passivation layer the gaps sensitivity to charge requires that any charge added by the gate is removed. This can be accomplished automatically by the measurement software reversing the sense of the gate voltage approximately every quarter second. Of course observations are only recorded during the positive sense.

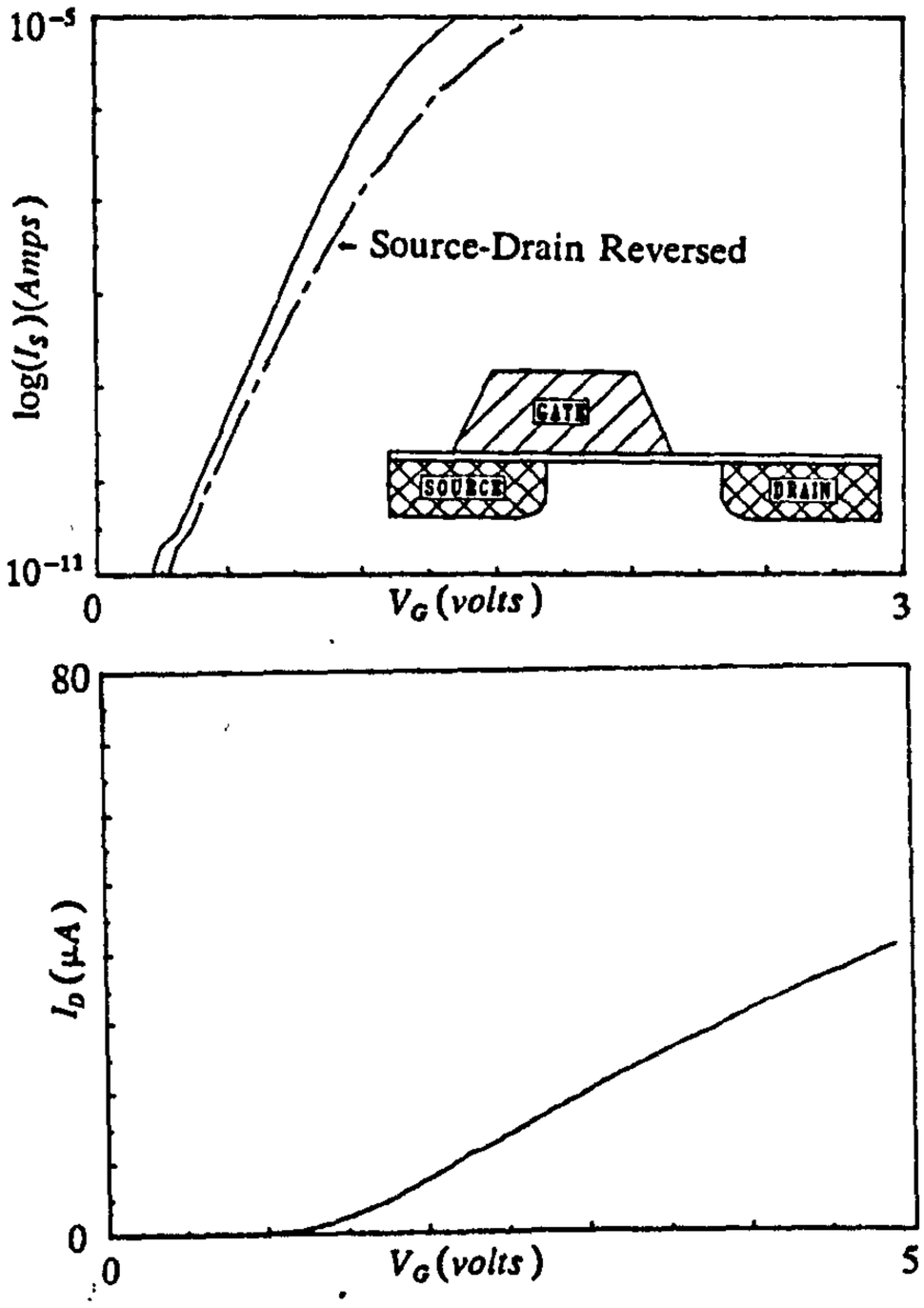


Figure 2 Electrical Characteristics of a 1 μm long transistor with a 0.2 μm drain gap. Figures a-c (top to bottom) are in the subthreshold, threshold and saturation regions.

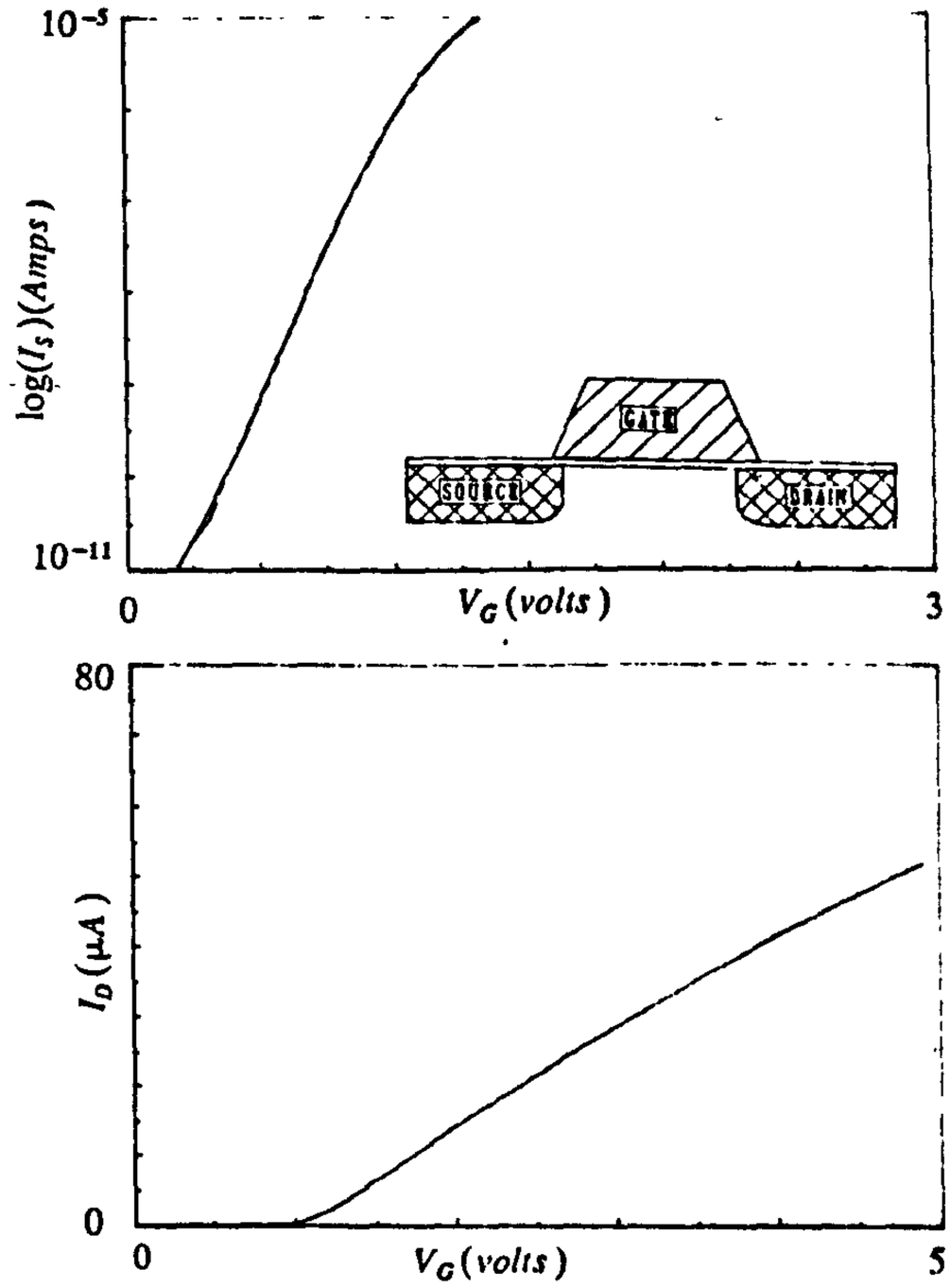


Figure 3 Electrical Characteristics of a 1 μm long transistor with complete channel coverage. Figures a-c (top to bottom) are in the subthreshold, threshold and saturation regions.

4. OBSERVATIONS - SUBTHRESHOLD REGION

An important parameter in the sub-threshold region of operation is the subthreshold swing [4]. It is a measure of how large a change in gate voltage is required to change the channel current by one order of magnitude. Figure 4 is a plot of the sub-threshold swings dependence on the drain voltage for each transistor in an alignment column.

Two generalisations are immediately evident.

Those transistors that have a gap in gate to channel coverage near the source end of the channel have an increased swing that is independent of drain voltage. If the swing for these source gapped transistors (SGT's) is plotted relative to the magnitude of the gap a straight line results. That relationship may be useful in analysis of the gate fringe fields.

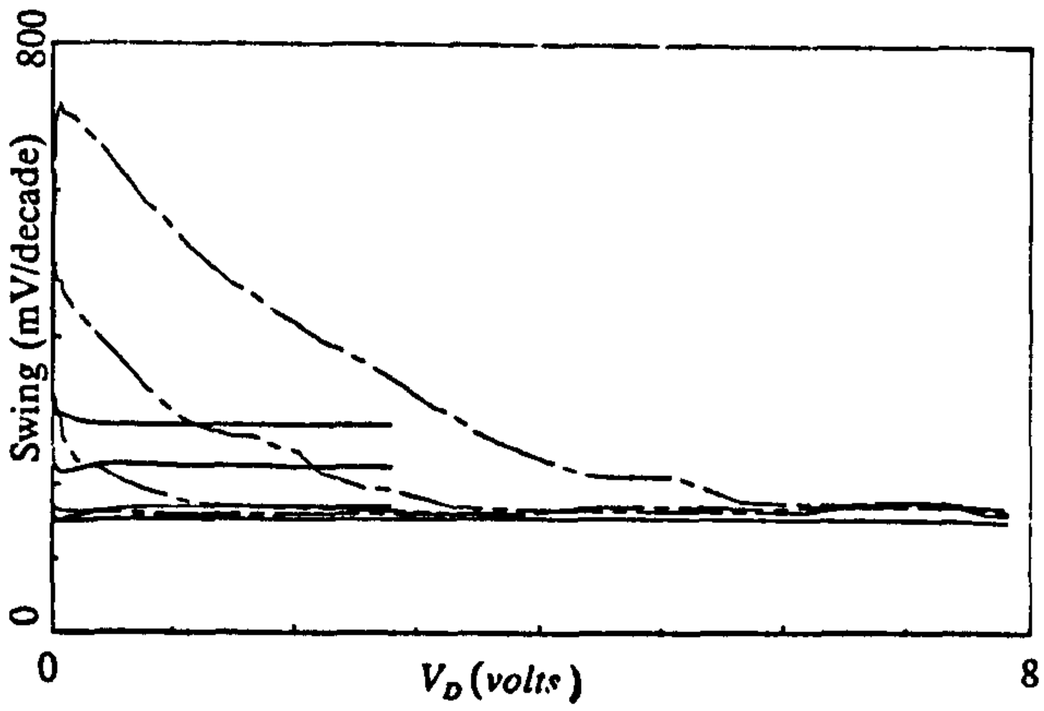


Figure 4 Subthreshold swing as a function of transistor structure and drain voltage. The dashed curves are for DGT's with gaps of (top to bottom) 0.50, 0.35, 0.20, and 0.05 μm. The solid curves are for SGT's with gaps of (top to bottom) 0.35, 0.20, and 0.05 μm. Note: The DGT's are drain voltage dependent.

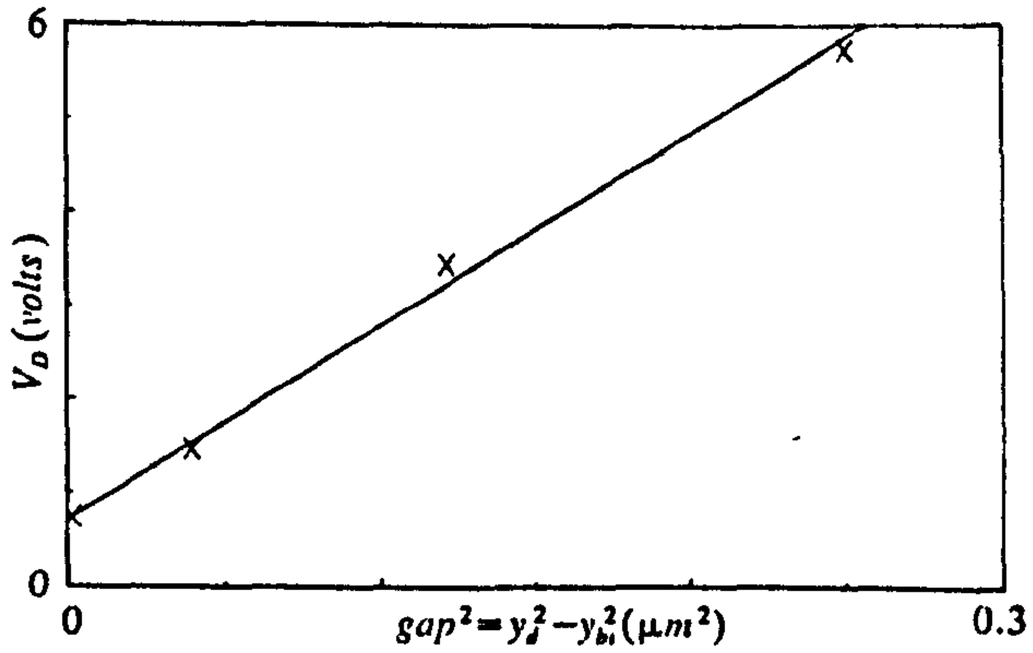


Figure 5 The drain voltage causing constant subthreshold swing as a function of the magnitude of the DGT's gap. Note: The linear dependence supports equation 1.

The majority of the remaining curves in figure 4 are for transistors with a gap in gate to channel coverage near the drain end of the channel. It is obvious that the subthreshold swing of the drain gapped transistors (DGT's) is dependent on the drain voltage. It seems to follow a parabolic dependence on drain voltage until the swing reaches a constant swing magnitude. If the drain voltages at which this first occurs are plotted against the square of the magnitude of the gap size again a linear relationship becomes evident, as illustrated in figure 5. This is simply the motion of the drain depletion region into the channel under the influence of the drain field. The slope of the resulting line can be used to estimate the surface doping concentration using a standard depletion width equation [4].

$$y_d^2 = \frac{2\epsilon_s}{qN_A}(V_{bi} - \Psi_s + V_d) \qquad (1)$$

The predicted swing magnitude calculated using the surface concentration extracted by this method agrees with the observed value.

5. CONCLUSIONS

A novel technique used to manufacture transistors with quantifiable gaps in gate to channel coverage near the source or drain has been described. This technique has a significant advantage over previous methods since transistors with a known gap can be identified and their characteristics measured. These have been used to demonstrate that asymmetries in the subthreshold region of gapped transistors are caused by the requirement for the uncovered channel region to be inverted by the fringing field of the gate. In the SGT case the subthreshold swing is consistently increased and was observed to be linearly dependent on the size of the gap. However the drain voltage dependent swing of the DGT has been shown to be caused by drain depletion widening.

6. FURTHER WORK

This research will continue with consideration of the saturation region. The fabrication of polysilicon gated transistors with smaller step sizes is also planned.

REFERENCES

[1] P.K. Ko et al. 1986 IEDM Tech. Dig. p292.
[2] T.Y. Chan et al. IEEE Elec. Dev. Let. vol. EDL-7 no.1 Jan. 1986 p16.
[3] T.Y. Chan et al. IEEE Elec. Dev. Let. vol. EDL-8 no.6 June. 1987 p269.
[4] S. M. Sze. Physics of Semiconductor Devices (John Wiley & Sons, New York, 2nd Ed. 1981).

ACKNOWLEDGEMENTS

The authors would like to acknowledge the support of DEC and SERC. J. A. Serack is also supported by BNR and NSERC of Canada.

A NOVEL DEVICE STRUCTURE FOR STUDYING GATE AND CHANNEL EDGE EFFECTS IN IGFET'S.

J. A. Serack, A. J. Walton, J. M. Robertson

Edinburgh Microfabrication Facility,
University of Edinburgh,
Edinburgh, EH9 3JL, Scotland.

ABSTRACT: The operation and fabrication of a drain gapped transistor (DGT) is described. Simulated and experimental transistor characteristic curves are presented, and the use of a DGT to study both drain depletion motion, and hot carrier effects is demonstrated.

INTRODUCTION.

Over the last decade, advances in microfabrication have allowed dramatic reductions in the minimum dimensions of IGFET's. However, for a number of reasons, scaling laws were not adhered to for some aspects of modern transistors. Among other features, higher internal electric fields and short high gate electrodes, have made significant changes to transistors' operation and their electrical characteristics.

Due to of a lack of experimental tools for probing internal device operation, engineers and scientists have relied upon computer simulations of transistors. Although careful simulations based on detailed device geometry can predict most transistor characteristics, they don't provide the same insight as good test structures.

A drain gapped transistor (DGT), as shown in figure 7a, is a transistor with a gap in gate electrode coverage at the drain side of the channel. It can be a useful test structure for studying effects related to the electrode edges. It functions in much the same way as a normally gated transistor (NGT) except the uncovered region of the channel must be either inverted by the gate electric fringe field or by charge over the gapped region or depleted by the drain electric field, or some combination of these in order to allow current to pass. Depending on the size of the gapped region, such a structure can be used to explore the field fringe at the side of the gate electrode, or the motion of the drain depletion region under the influence of the lateral electric field.

EXPERIMENTAL METHODOLOGY.

Although DGT's can be made by accident in processes using gate sidewall spacers [1-3], there must be a reliable method for their fabrication if they are to be of use as test structures. The DGT's employed to obtain the results reported here used a non-self-aligned gate process and a linear progressional array of transistors, (schematically shown in figure 1), with systematic steps in gate-to-channel alignment. The length of the channels varied along the rows from 1.0 μm to 2.05 μm in steps of 0.15 μm. Five micron reference transistors were included on every third column. Down a column the drawn alignment of the gate-to-channel was varied in steps of 0.15 μm from a drain gap of 0.75 μm to a source gap of 0.6 μm. The reference transistors were drawn with both source and drain overlaps of 0.3 μm. The choice of 0.15 μm as the step size was based on experimental observations of the edge roughness that was typically obtained in 0.5 μm thick dry etched Al:Si alloy films.

A fairly typical LOCOS isolation process was employed for device fabrication until the source and drain implants were completed. Then, what would have been the polysilicon gate was stripped away and the contact holes cut. Aluminium was patterned by reactive ion etching to form the gate and contacts to the source and drain. The process used <100> silicon wafers and resulted in the following features; arsenic doped source and drain junction depths of 0.35 μm, a channel boron impurity concentration of 8×10^{16} atoms/cm^3, 500 Å thermal gate oxide, and 0.5 μm thick aluminium gates.

Although excellent overlay accuracy between gate and channel lithography steps can be obtained using a 10:1 reduction wafer stepper, there will always be some amount of misalignment. In this experiment the misalignment results in only a displacement of the site within the array where the normally gated transistor is located. It is therefore necessary to have some way of detecting its position.

To detect the aligned transistor, the forward and reverse (source and drain interchanged), subthreshold curves are compared for each transistor in a selected column. Since the electron injection at the source is a strong function of the gate electric field, any drain gapped transistor will have a marked decrease in current when its source and drain are interchanged. The aligned device (fig. 2b) will have similar forward and reverse curves. A useful check is that the device preceding the aligned one (fig. 2a) has better forward than reverse characteristics, and the one following (fig. 2c) has better reverse than forward characteristics.

Once the aligned transistor is detected, the structure of the array allows transistors with any particular gap or overlap to be readily selected.

SIMULATION OF DTGs.

Although this work is centered around experimental measurements, the simulation program CANDE [Technology Modeling Associates] has been useful in obtaining a qualitative understanding of the internal operation of a DGT. Comparing figures 4 to 5 and 6 to 7, one can see good qualitative agreement between the simulation and experimental results. It is important to realise that the simulated structure only approximates the experimental one. The main approximations are in the way the gate sidewall is emulated by a step in oxide, whose surface is covered by the gate electrode (see fig. 4a), and that implant profiles have only been represented by equivalent "boxes".

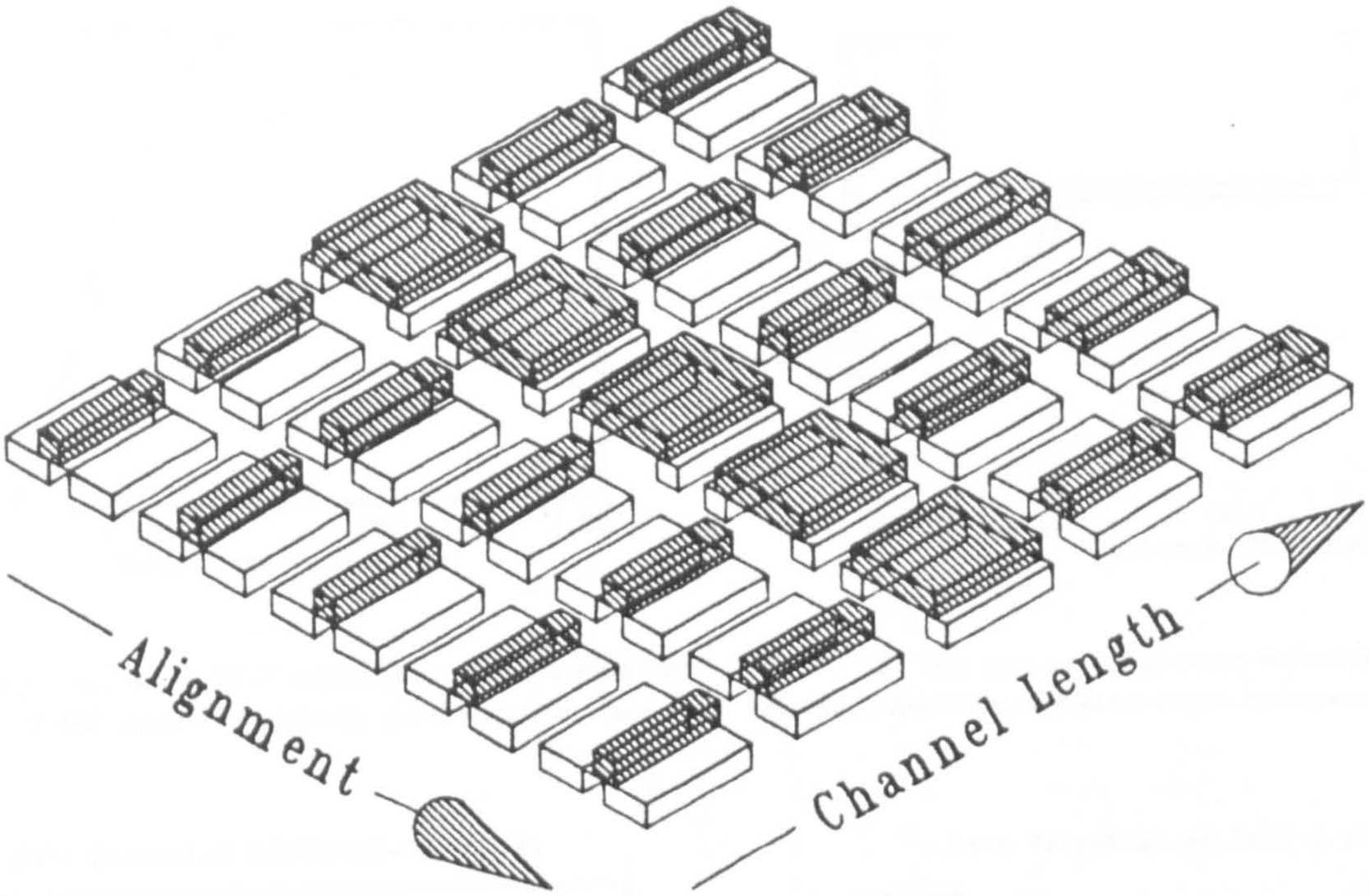


Figure 1. 3-D Schematic of the progressional alignment array.

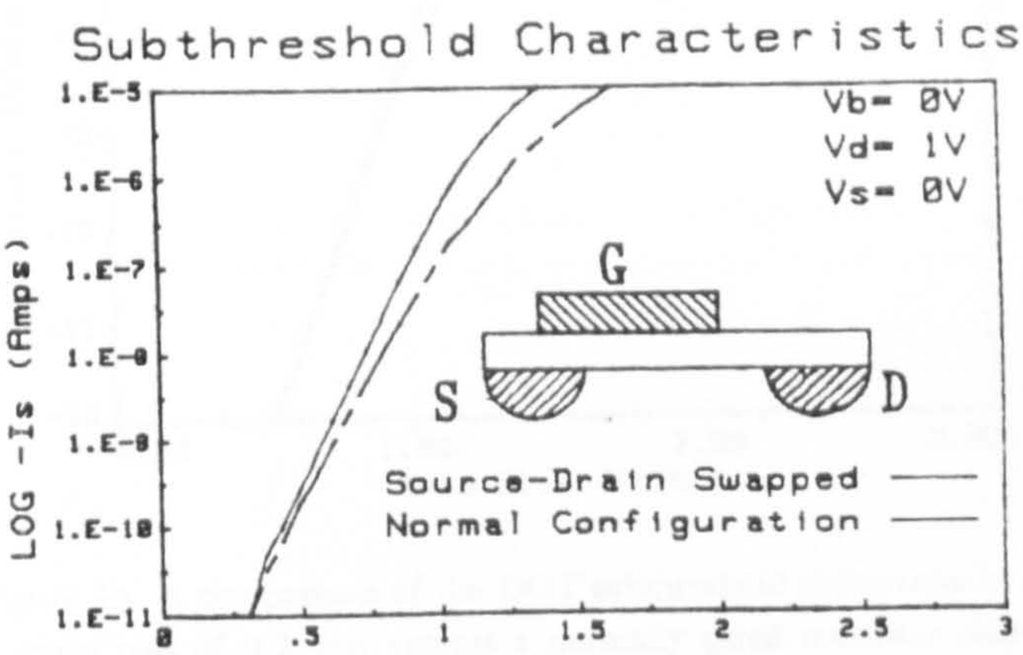


Figure 2a. Measured DGT Subthreshold Characteristics.

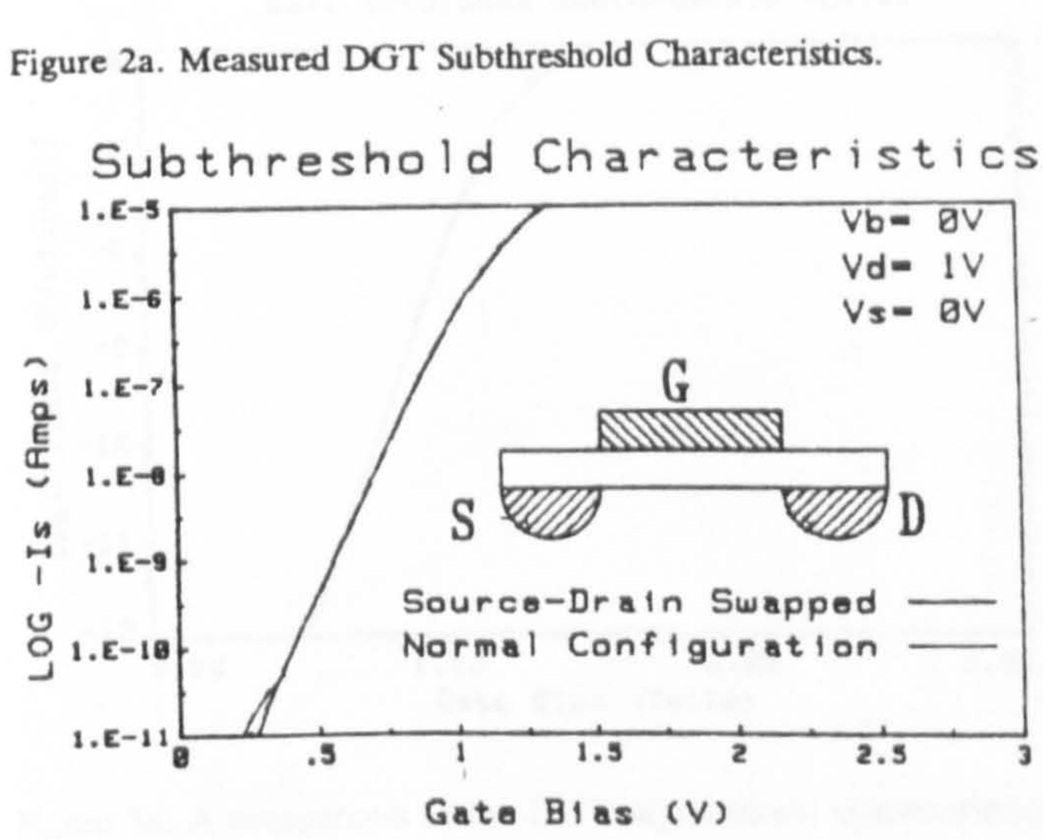


Figure 2b. Measured Normal Subthreshold Characteristics.

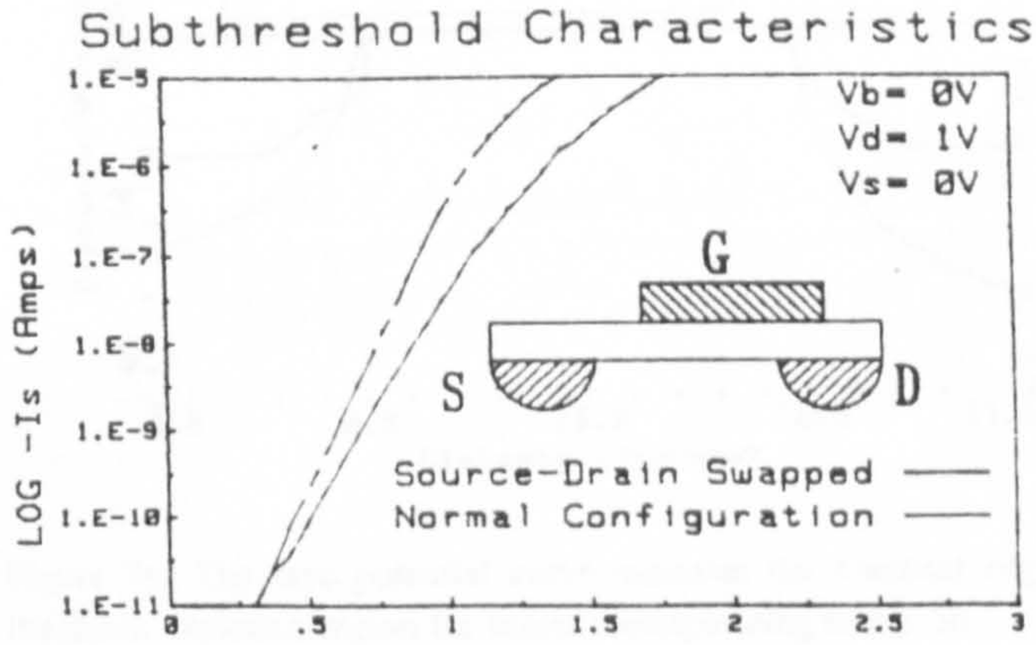


Figure 2c. Measured SGT Subthreshold Characteristics.

APPLICATION OF DGTs AS TEST STRUCTURES.

A DGT could be used as a test structure to study a number of effects, eg. the fringing electric field of the gate electrode (with applications to gate and interconnect capacitance), the effect of high lateral electric fields found in short channel transistors, charge capture in the oxide above the gapped region, and other situations involving interactions at edges of the transistor's electrodes. The application of DGTs to study drain depletion motion and hot electron capture are presented here.

A DGT can be used to measure the drain depletion motion under the influence of the lateral electric field by comparing the subthreshold characteristics to that of an NGT [4]. The simulations in figure 3 help show that as the drain depletion is pushed outward by the lateral field (zero bias contour in 3d,e,f)

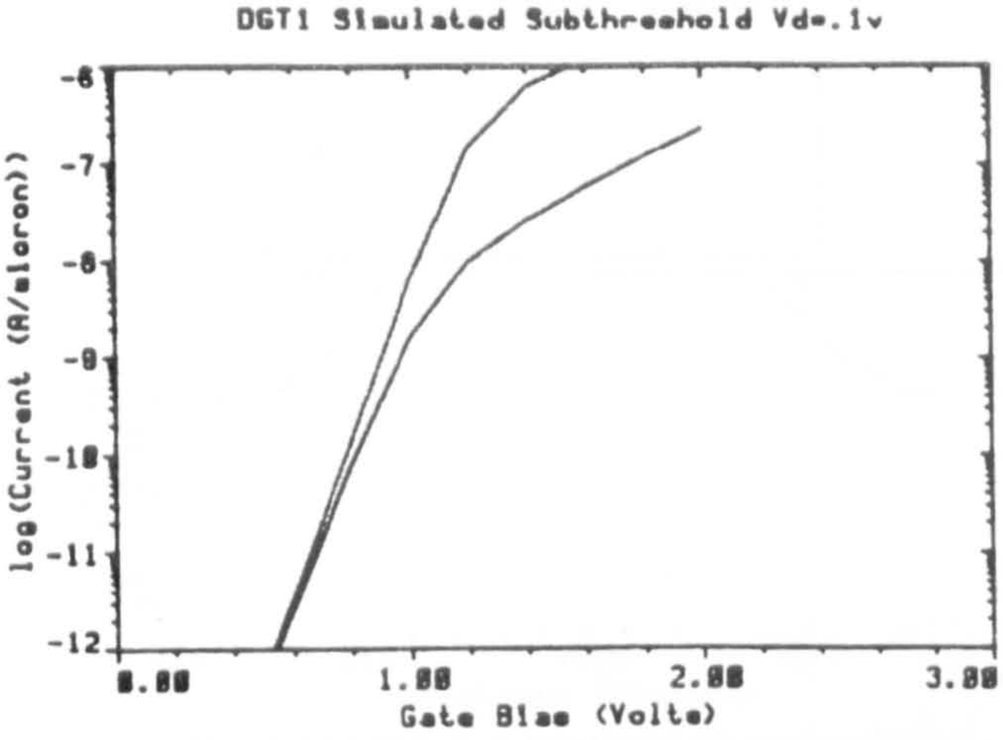


Figure 3a. A comparison of the DGT subthreshold characteristics for a drain bias of 0.1 volt against a normally gated transistor characteristic.

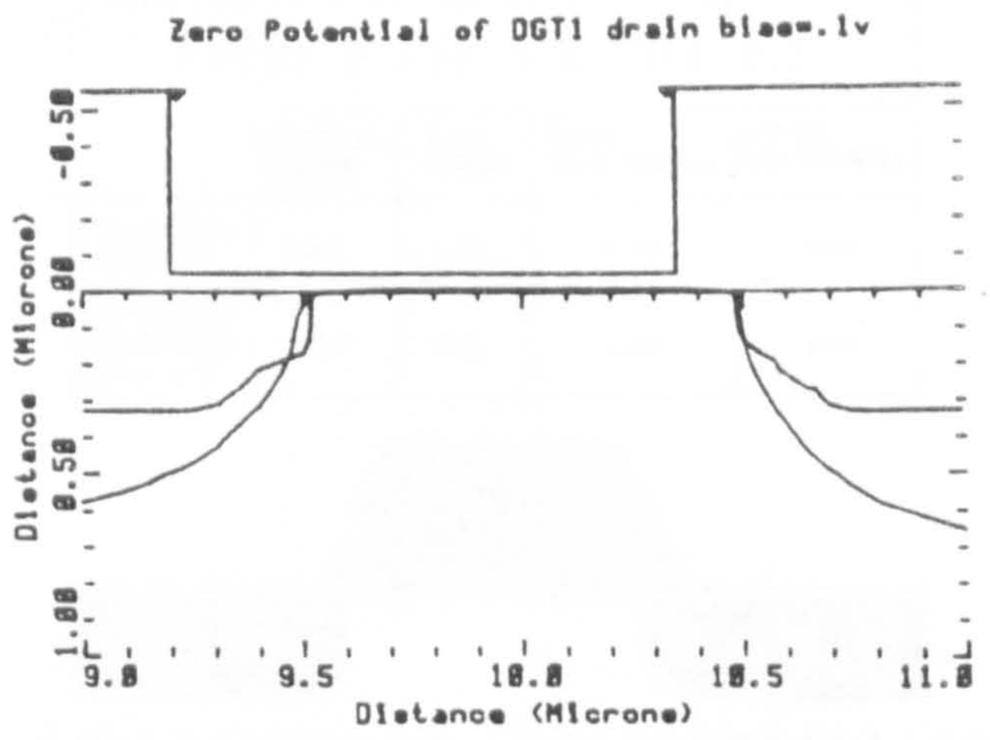


Figure 3d. The zero potential curve indicates the practical edge of the drain depletion region for biases corresponding to fig. 3a.

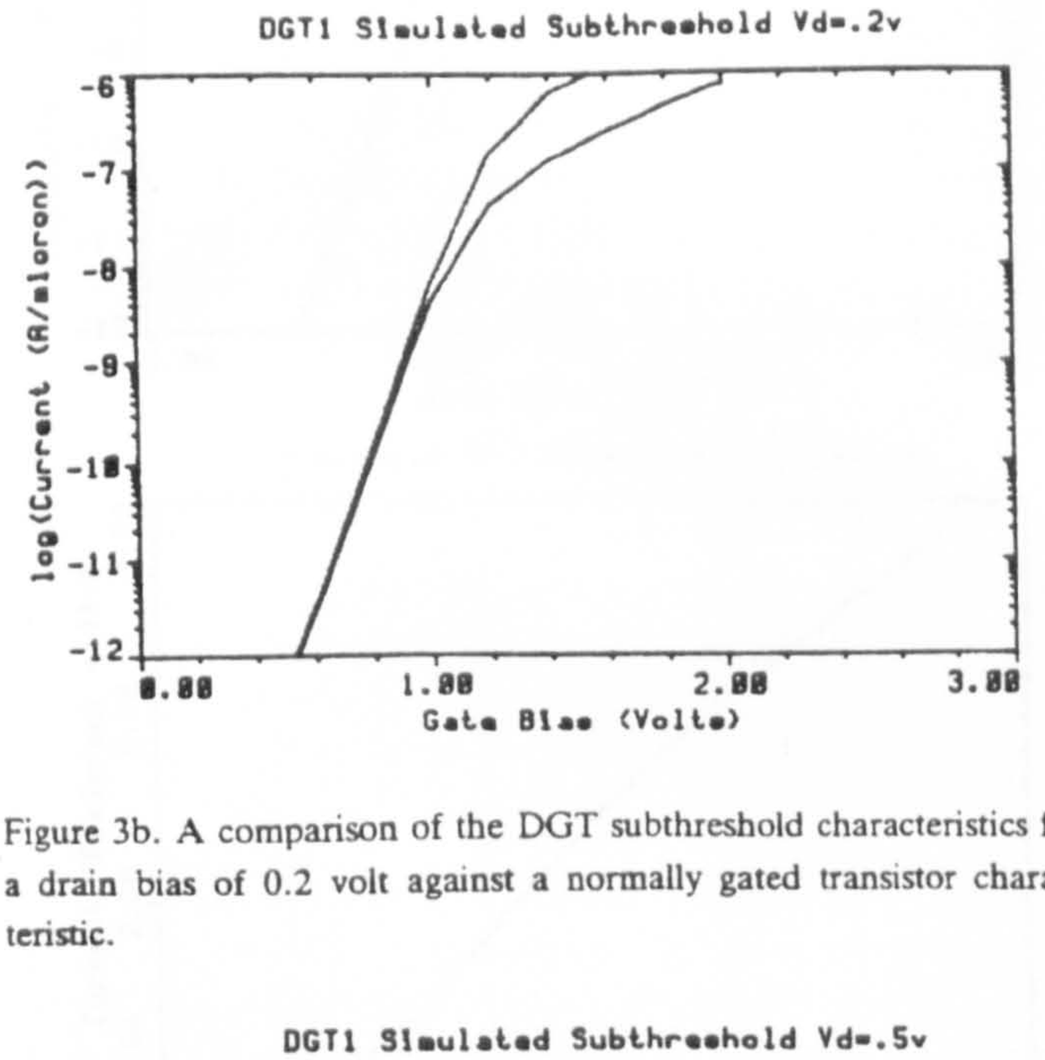


Figure 3b. A comparison of the DGT subthreshold characteristics for a drain bias of 0.2 volt against a normally gated transistor characteristic.

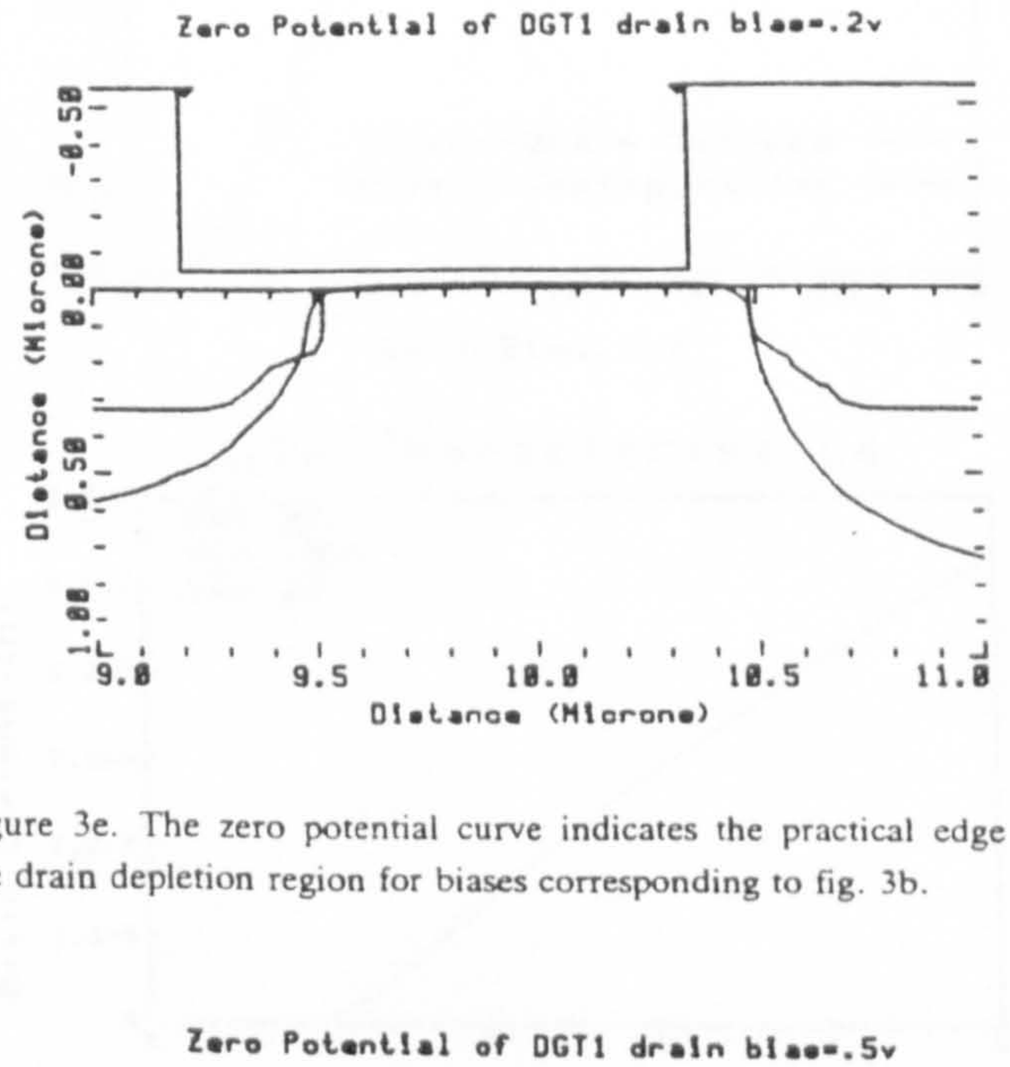


Figure 3e. The zero potential curve indicates the practical edge of the drain depletion region for biases corresponding to fig. 3b.

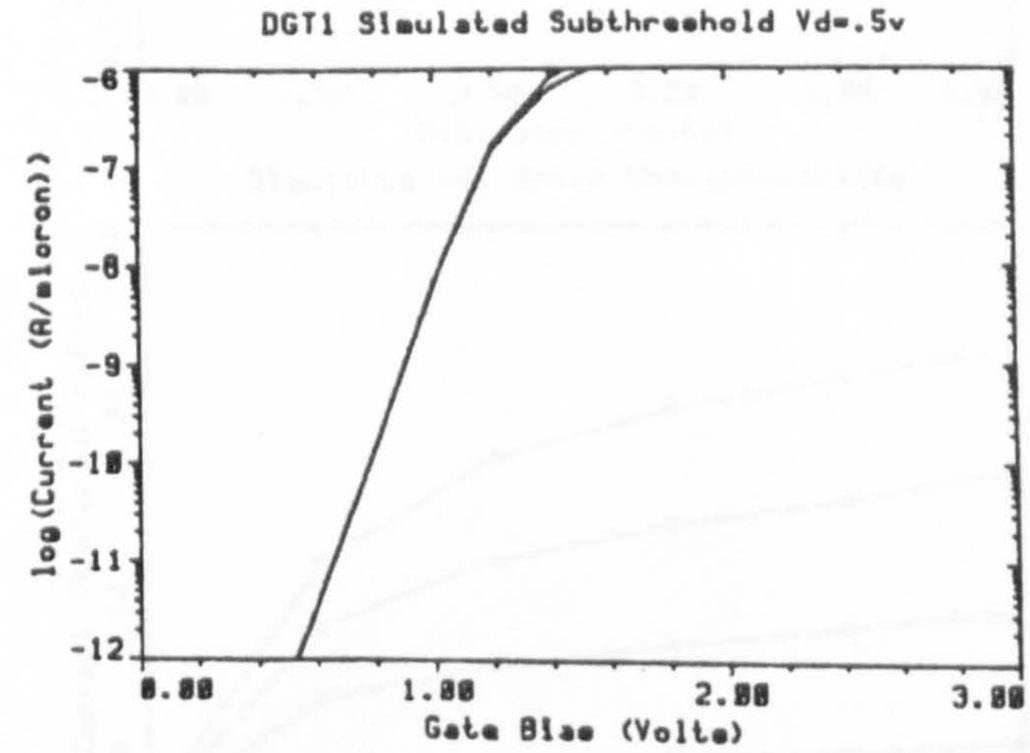


Figure 3c. A comparison of the DGT subthreshold characteristics for a drain bias of 0.5 volt against a normally gated transistor characteristic.

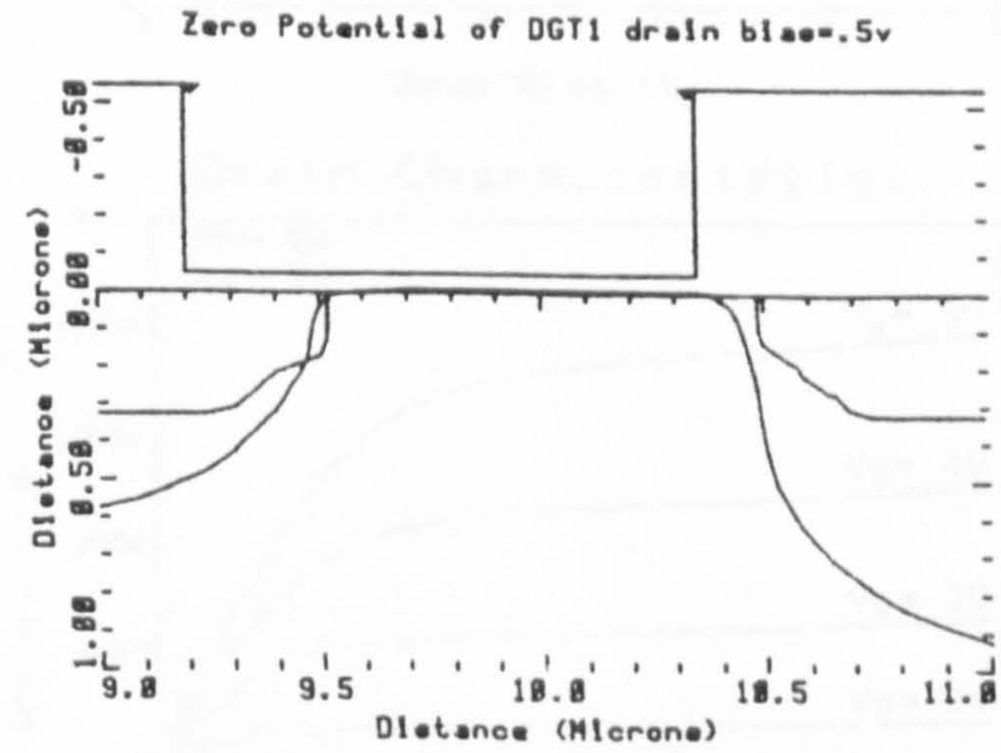
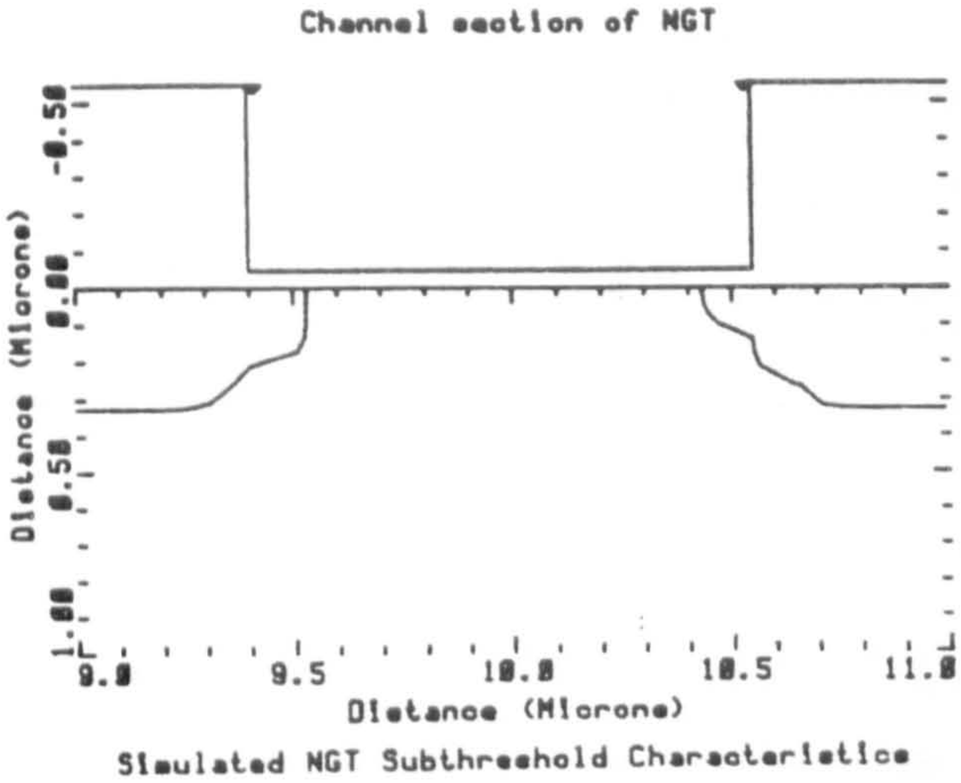
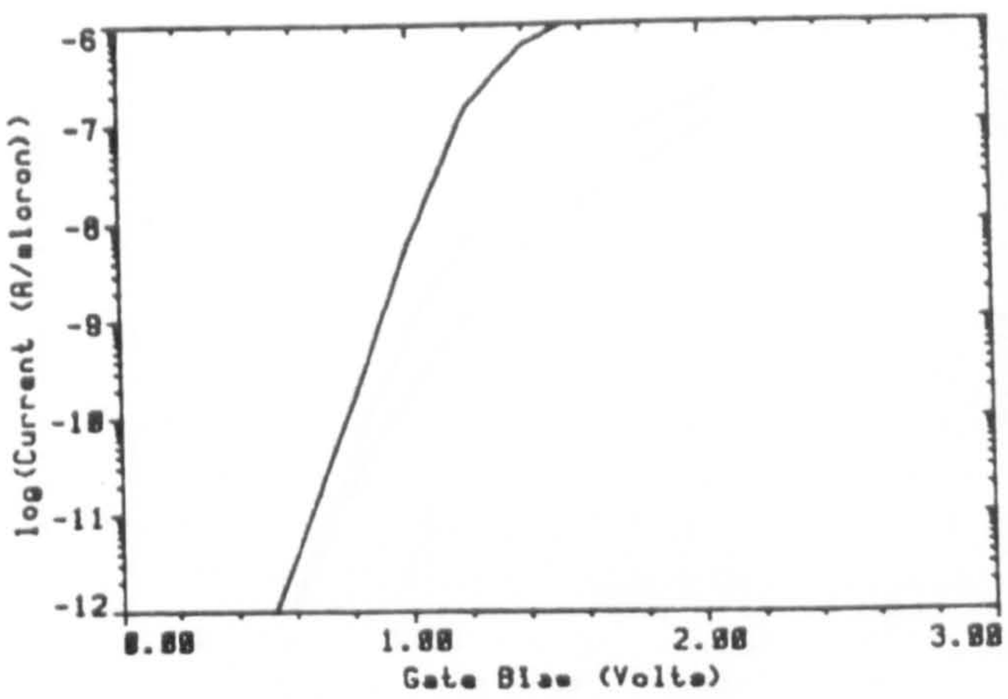


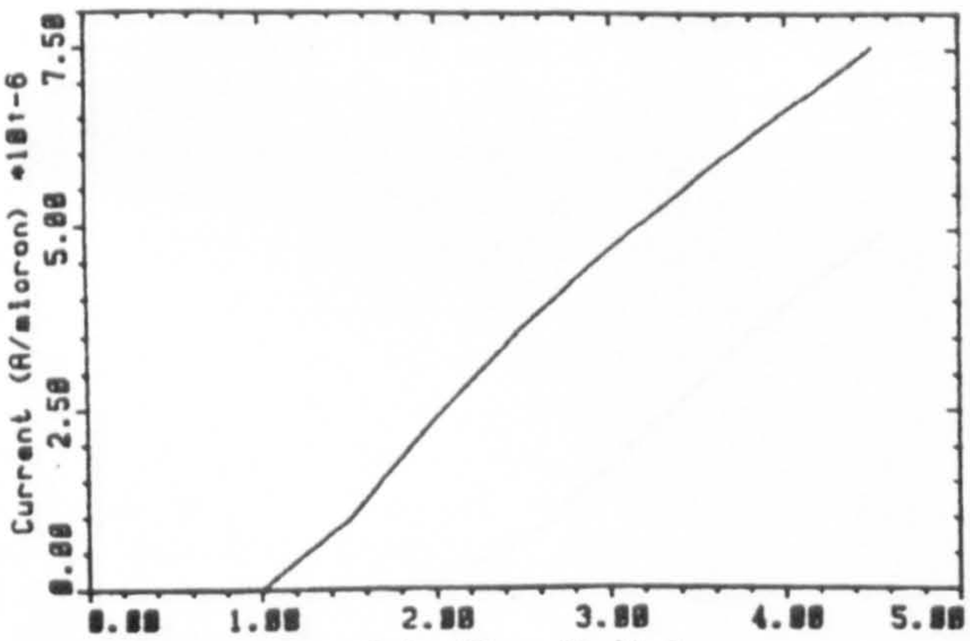
Figure 3f. The zero potential curve indicates the practical edge of the drain depletion region for biases corresponding to fig. 3c.



Simulated NGT Subthreshold Characteristics



Simulated NGT Gate Characteristics



Simulated NGT Drain Characteristics

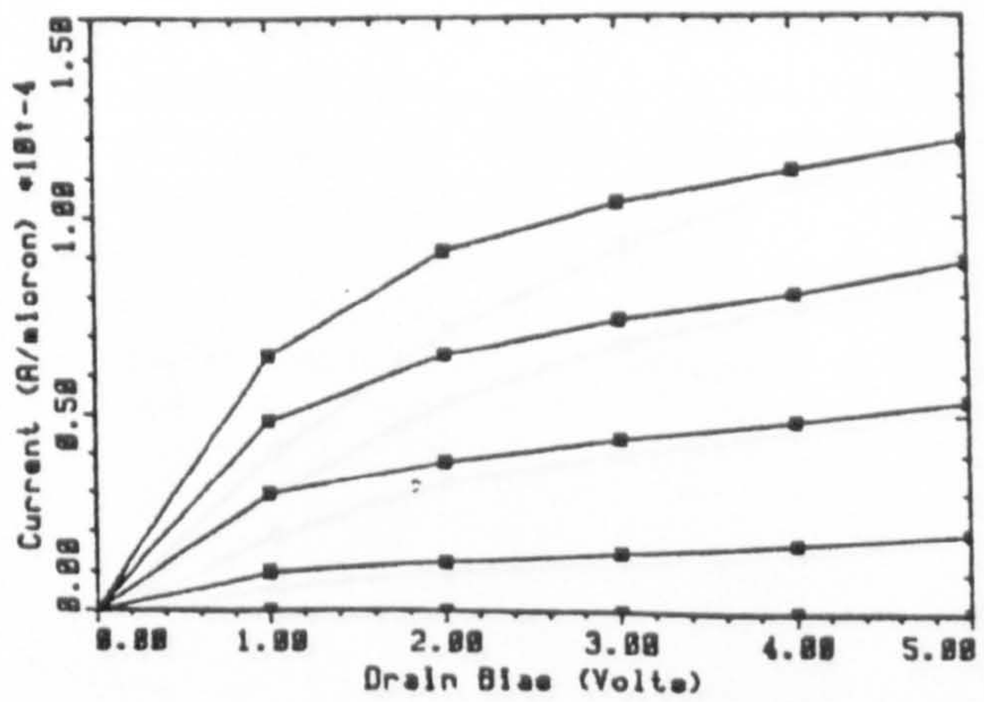
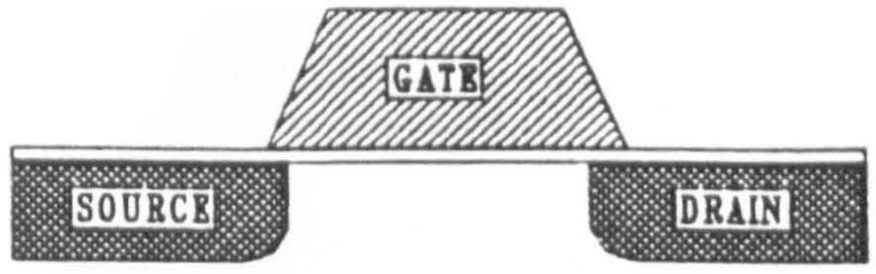
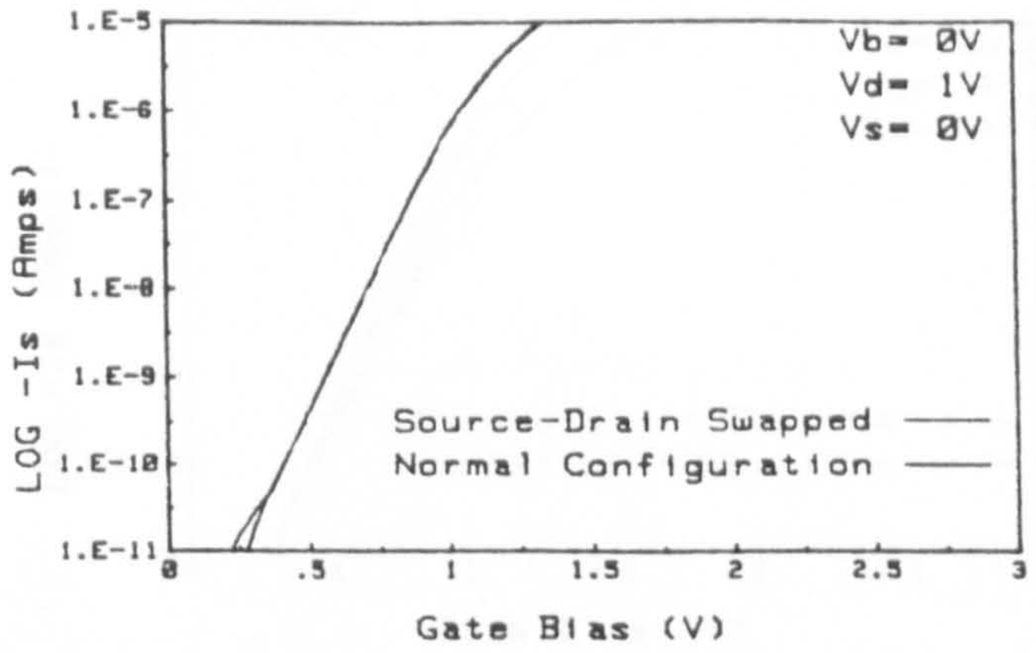


Figure 4a,b,c,d. Simulated transistor (NGT) characteristics.

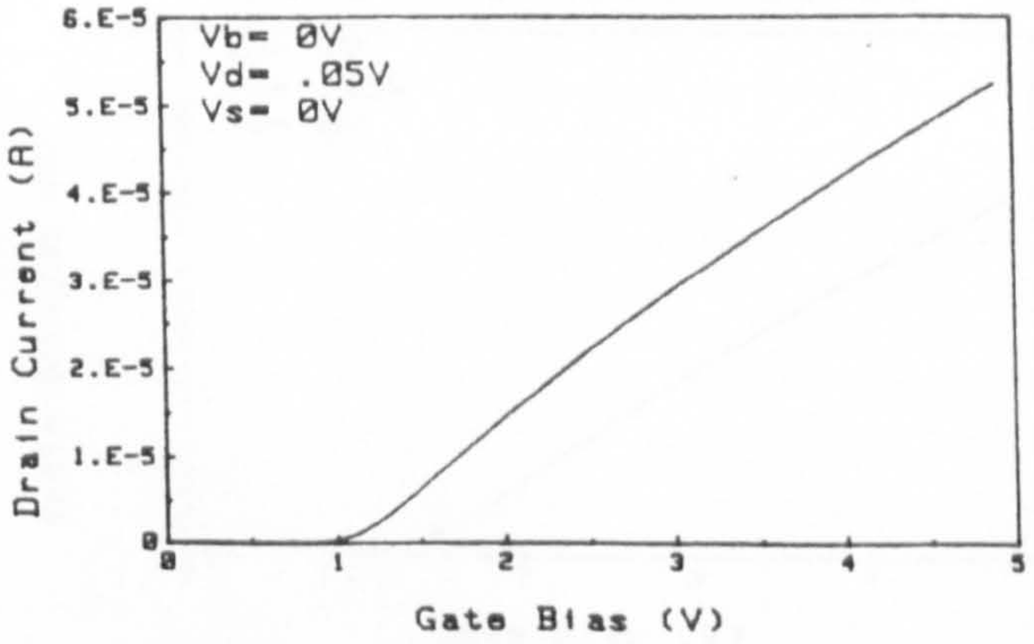
Wafer 9 Die 9,9 Site 5,6				
	Channel Length	Gate Length	Drain Gap/Overlap	Source Gap/Overlap
Drawn Size (microns)	1.30	1.30	0.00	0.00
Effective / Actual Size (microns)	1.05	1.37	0.13	0.07



Subthreshold Characteristics



Gate Characteristics



Drain Characteristics

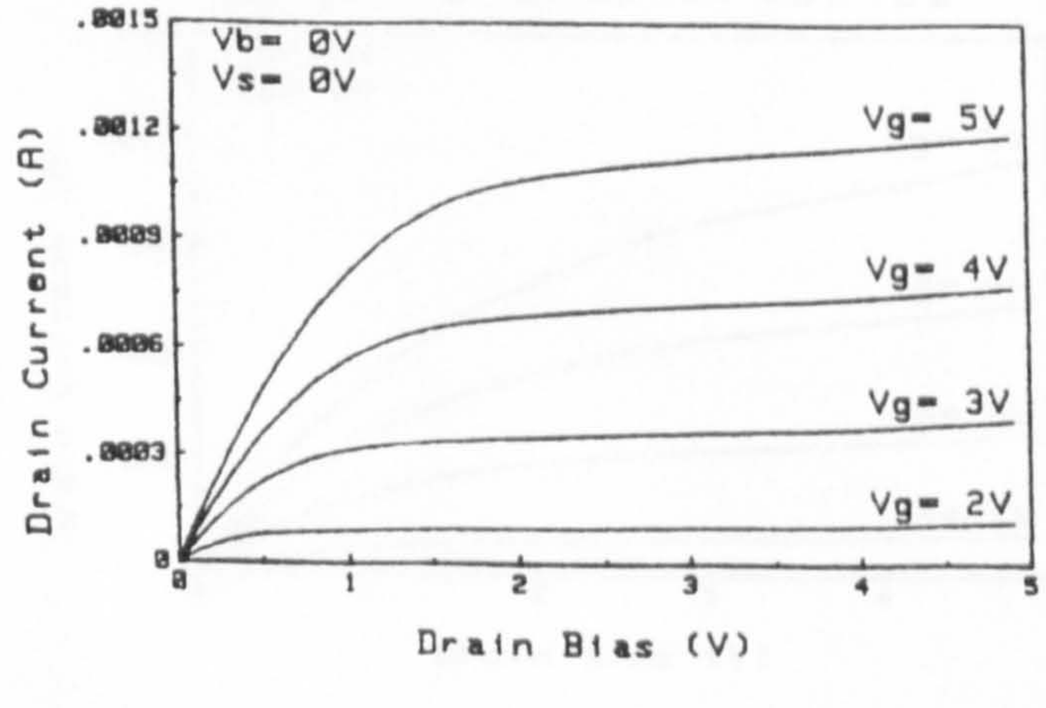
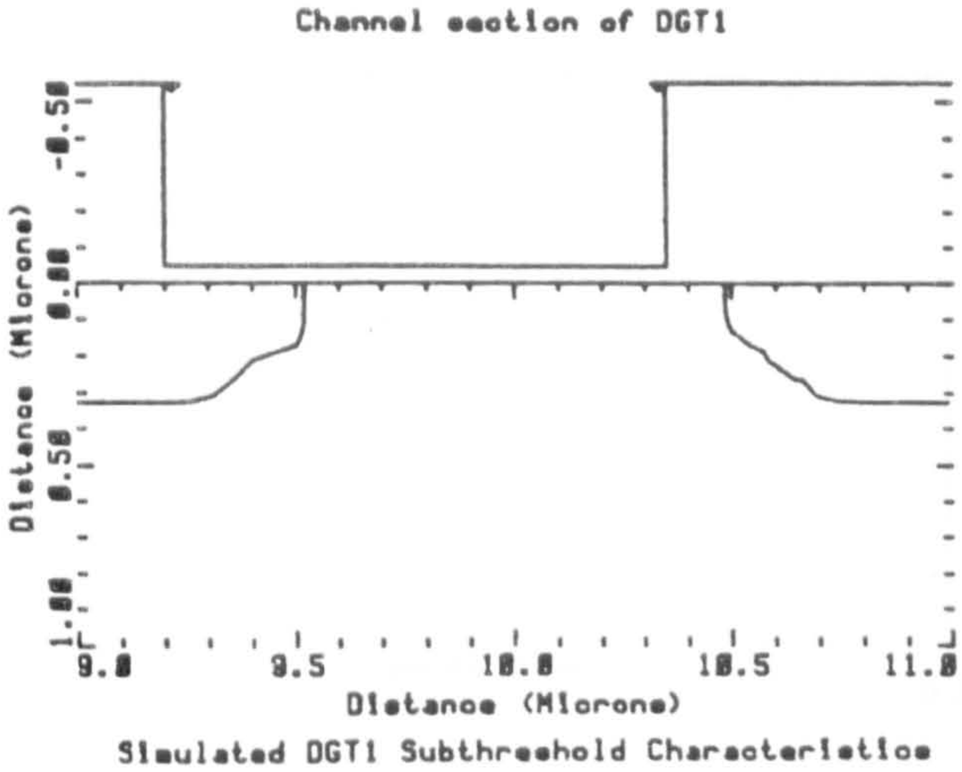
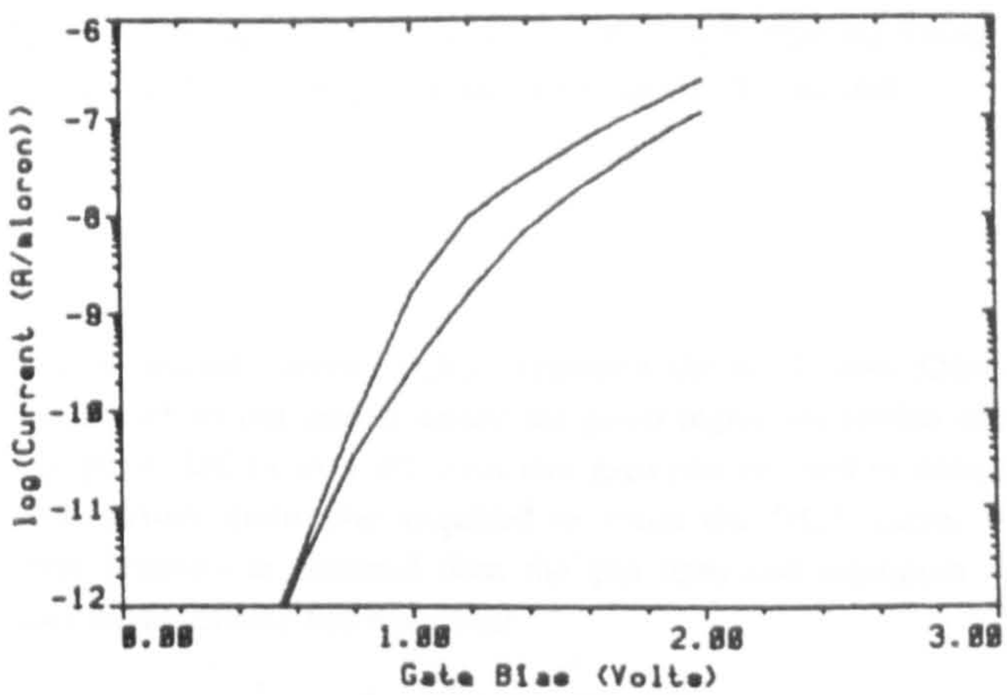


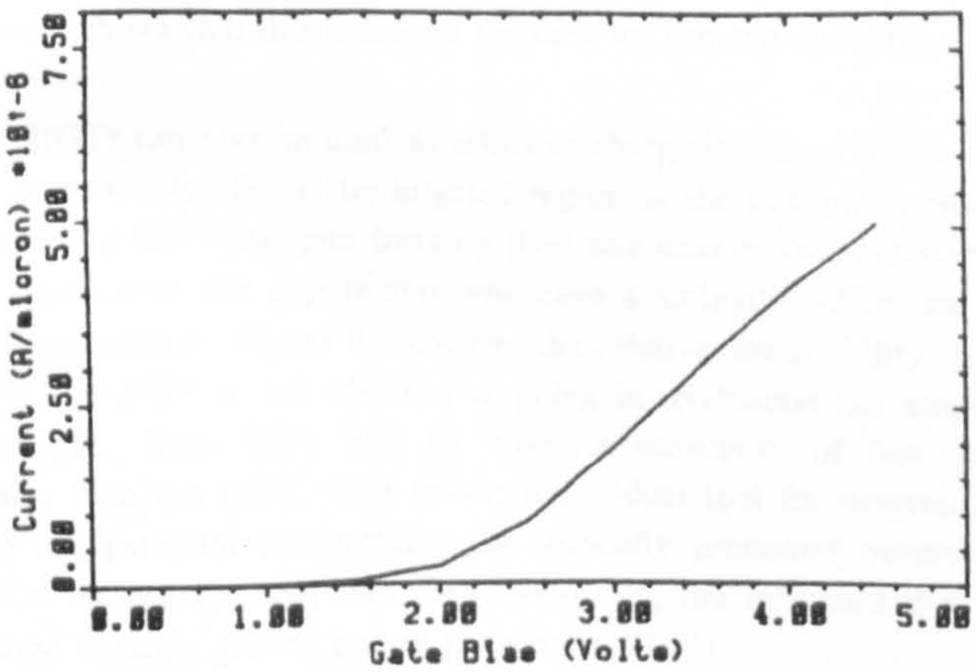
Figure 5a,b,c,d. Measured transistor (NGT) characteristics.



Simulated DGT1 Subthreshold Characteristics



Simulated DGT1 Gate Characteristics



Simulated DGT1 Drain Characteristics

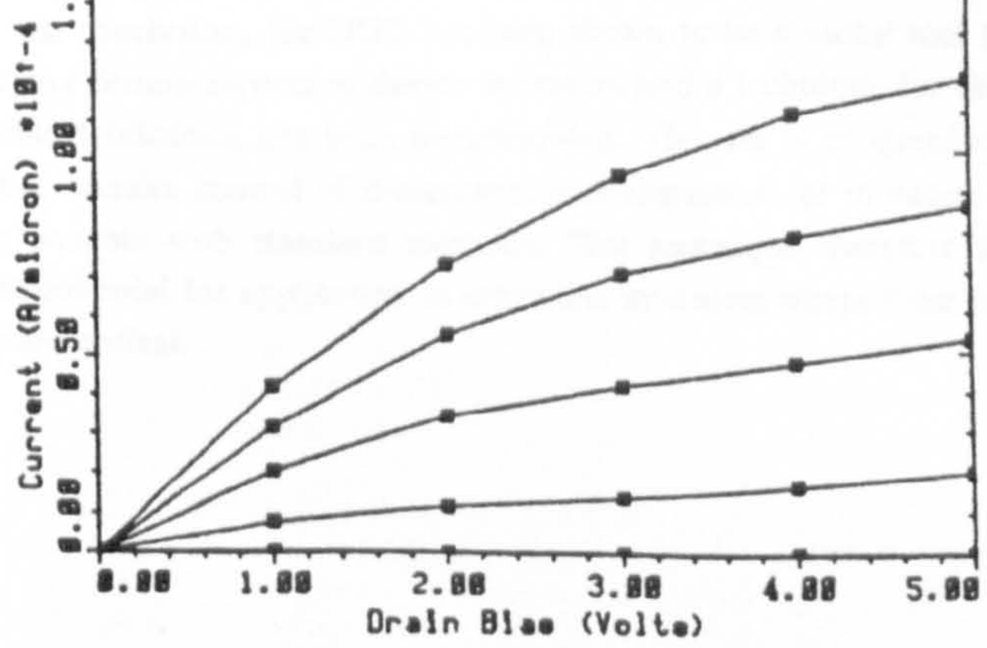
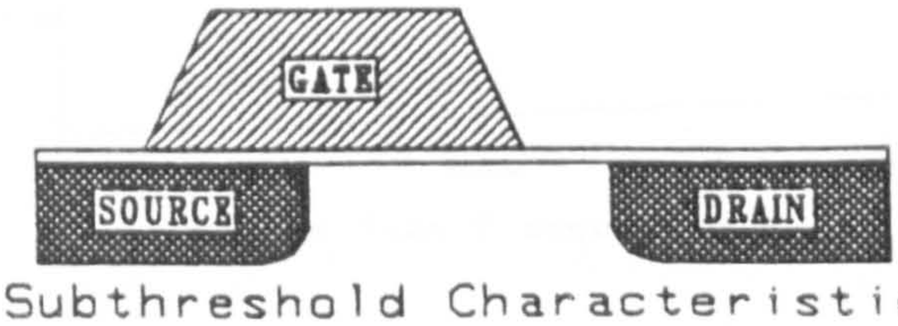
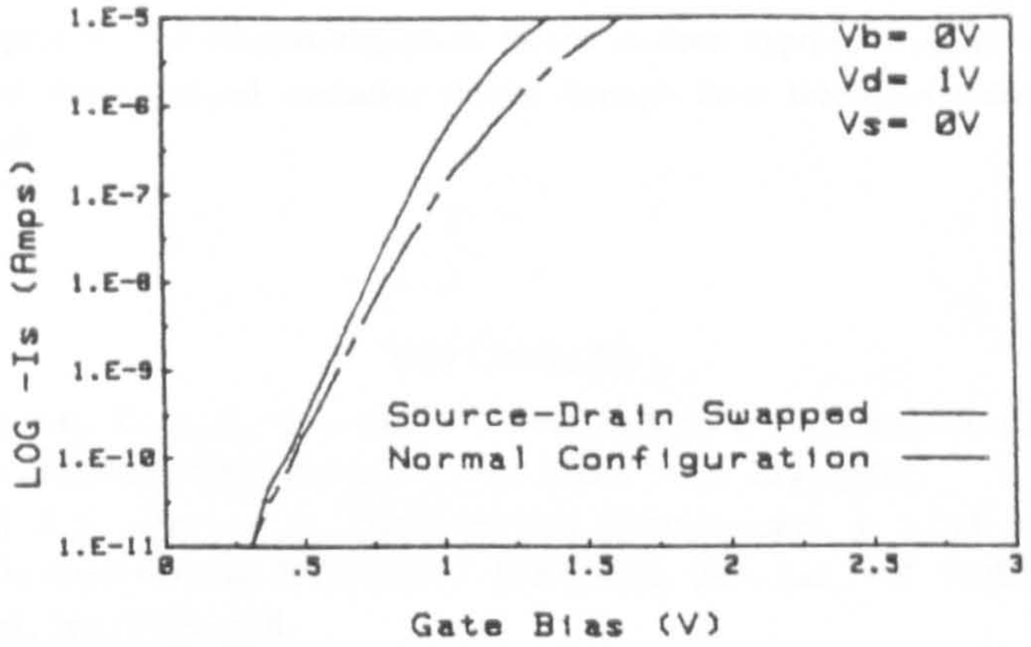


Figure 6a,b,c,d. Simulated transistor (DGT) characteristics.

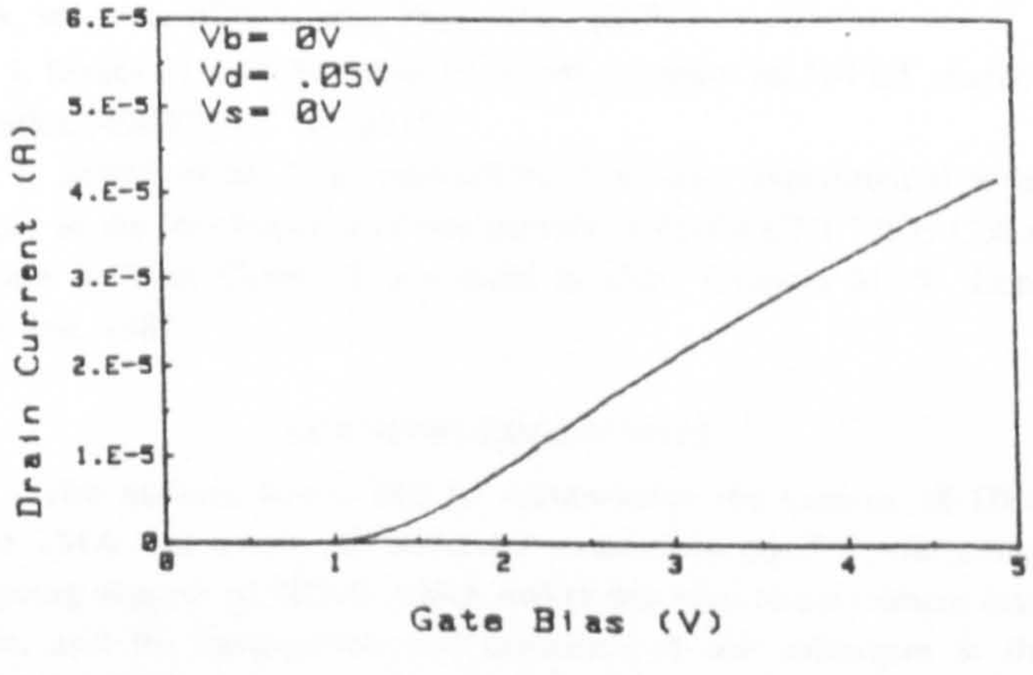
Wafer 9 Die 9,9 Site 5,4				
	Channel Length	Gate Length	Drain Gap/Overlap	Source Gap/Overlap
Drawn Size (microns)	1.30	1.30	0.30	0.30
Effective / Actual Size (microns)	1.05	1.45	0.30	0.57



Subthreshold Characteristics



Gate Characteristics



Drain Characteristics

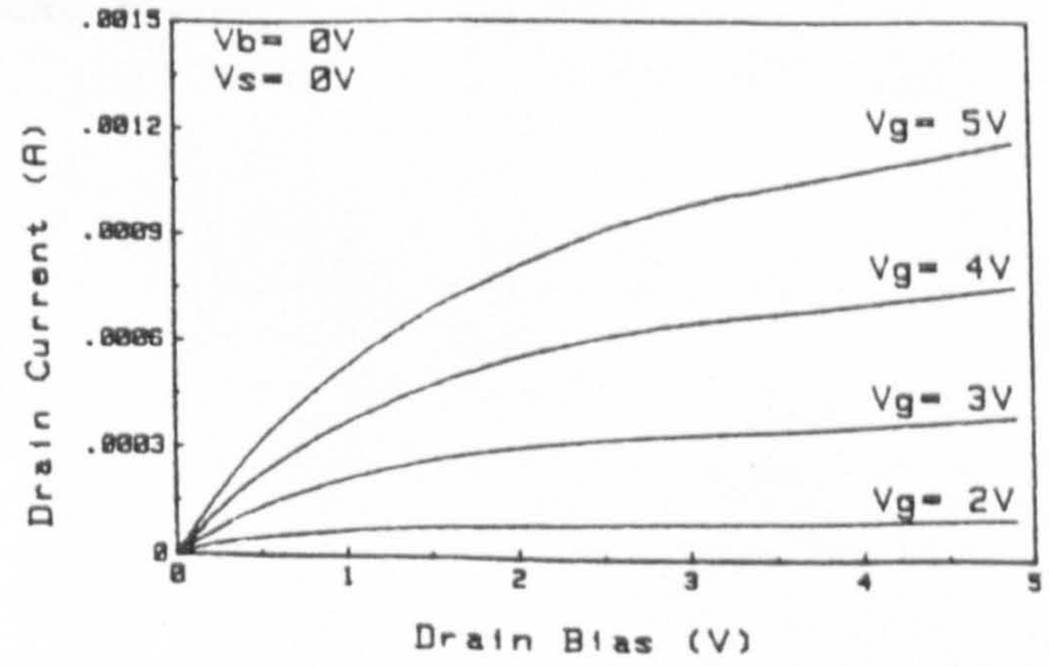


Figure 7a,b,c,d. Measured transistor (DGT) characteristics.

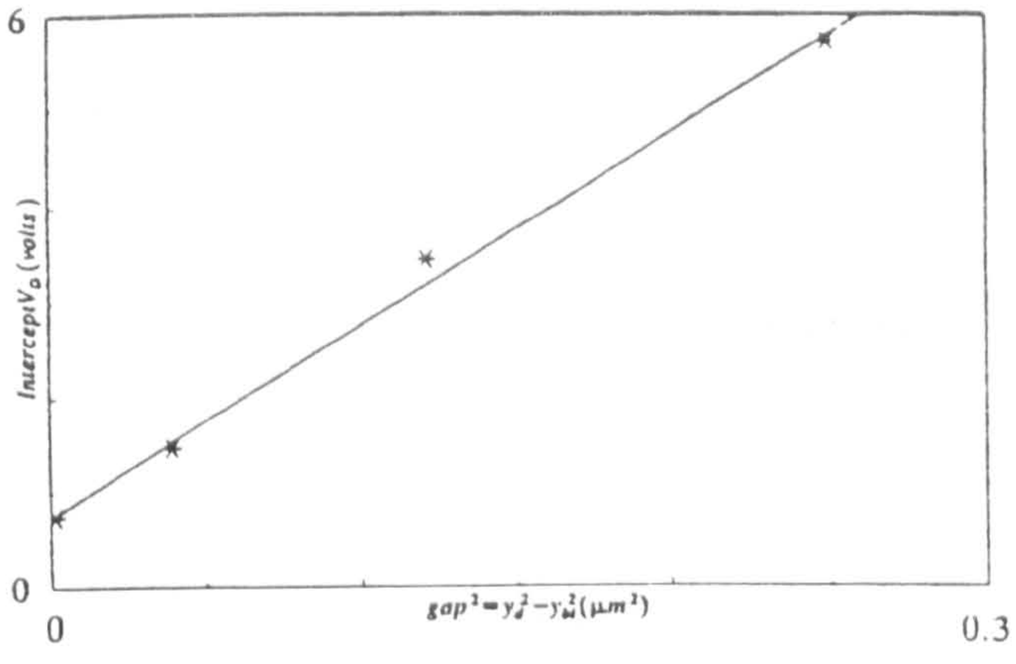


Figure 8. Drain bias to achieve "normal" subthreshold swing as a function of the size of gap in gate coverage of the channel.

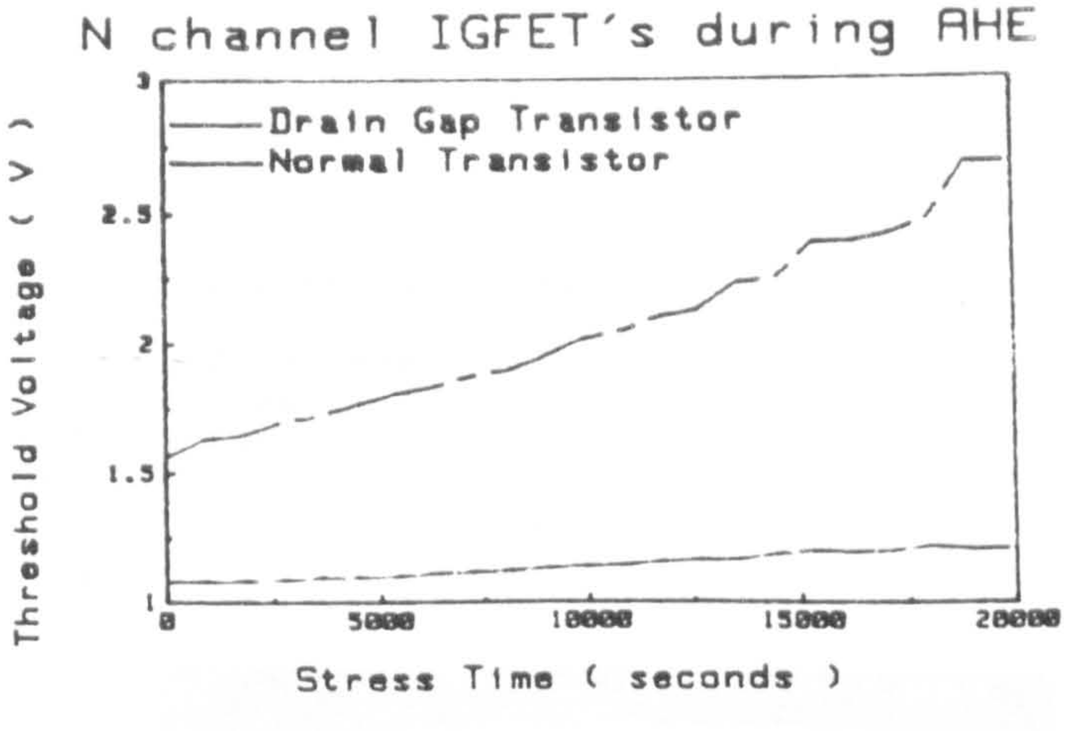


Figure 9. The relative sensitivity to hot electron capture of a normal and drain gapped transistor shown through their threshold voltage shift.

the subthreshold curves (3a,b,c) approach the NGT case. Once the drain depletion has moved under the gated region no further change takes place. DGTs with different size gaps can be used to determine the minimum drain bias required to reach the NGT curve. If an abrupt junction is assumed then the gap sizes and minimum drain biases should satisfy the equation:

$$y_d^2 = \frac{2\epsilon_s}{qN_A}(V_{bi} - \Psi_s + V_d) \qquad (1)$$

Figure 8 shows that this is indeed the case for the experimental measurements.

DGTs can also be used as sensitive charge detectors in hot carrier experiments. Since the ungated region of the channel is usually inverted by the weak gate fringing field any charge accumulation in the oxide over the gap region will have a dramatic effect on the threshold voltage. Figure 9 compares the relative susceptibility of the DGT and NGT to hot electron trapping in avalanche hot electron mode [5]. That DGT had an absolute sensitivity of five times greater than the NGT. This makes it an ideal tool for determining both the potential susceptibility for normally processed transistors, and as an experimental tool for investigating the influence that the method of oxide growth has on hot carrier effects.

CONCLUSION.

In conclusion, the DGT has been shown to be a useful tool for studying certain aspects of device operation and a technique for their reliable fabrication has been demonstrated. The use of progressional arrays, enables control of dimensions and alignment far in excess of that possible with standard methods. This technique therefore has great potential for application to other test structures where these factors are critical.

REFERENCES

[1] P.K. Ko et al. "The effects of weak Gate-to-Drain/Source overlap on MOSFET Characteristics," 1986 IEDM Tech. Dig., p292.

[2] T.Y. Chan et al. "Asymmetrical Characteristics in LDD and Minimum-Overlap MOSFET's," IEEE Elec. Dev. Let., vol. EDL-7, no1, Jan. 1986, p16.

[3] T.Y. Chan et al. "A Capacitance Method to Determine the Gate-to-Drain/Source Overlap Length of MOSFET's," IEEE Elec. Dev. let., vol. EDL-8, no6, June. 1987, p269.

[4] J. Serack et al. "The effect of device geometry on IGFET characteristics," ESSDERC '87, p915.

[5] J. Serack et al. "The application of a novel experimental technique to the investigation of hot carriers in MOSFET's," IEE Colloquium on 'Hot Carrier Degradation in Short Channel MOS', London Jan. 1987.

ACKNOWLEDGEMENTS.

The authors would like to acknowledge the support of DEC and TMA that made the computer simulations possible, the strong ongoing support of SERC which makes this type of experiment realistic, and the cooperation and assistance of our colleagues at the E.M.F. J. Serack appreciates the personal support of BNR and NSERC of Canada.

AN OBJECTIVE METHOD OF ASSESSING METAL PATTERNING QUALITY

J.T.M. Stevenson, J. Gow, J. Serack
Edinburgh Microfabrication Facility
University of Edinburgh
Edinburgh, EH9 3JL, Scotland.

ABSTRACT

Metal patterning is one of the more difficult stages in VLSI microfabrication and conventional methods for assessing its success are tedious. A quick and accurate electrical method is described for evaluation of problems in the photolithography and etching of metal layers.

INTRODUCTION

Metal patterning is a difficult stage in VLSI microfabrication: the reflective, granular surface of the metal can cause notching of tracks and, with monochromatic exposure, standing waves within the resist film can produce terraced sidewalls on the developed profile. The topographical variations which are encountered near the end of a process and the consequent variations in resist thickness make matters worse. These problems are being addressed by resist manufacturers and new photoresist materials specially formulated for use on metal are being introduced.

The ability to evaluate how the behaviour of a photoresist affects the ultimate metal pattern is crucial to the development and control of VLSI processes. In evaluating a photoresist, many factors such as its thickness, pre and post-exposure bake temperatures and times, exposure sensitivity and etch resistance must be considered. It is not realistic to consider the photolithographic stage in isolation. The overall pattern transfer, i.e. photolithography plus etch plus resist stripping must be considered as a whole since the effect of the etchant on the photoresist can be catastrophic, especially when Reactive Ion Etch (RIE) systems based on chlorine chemistry are employed.

Figures 1 and 2 show a typical example where the developed photoresist pattern was satisfactory but its resistance to a RIE plasma was poor.

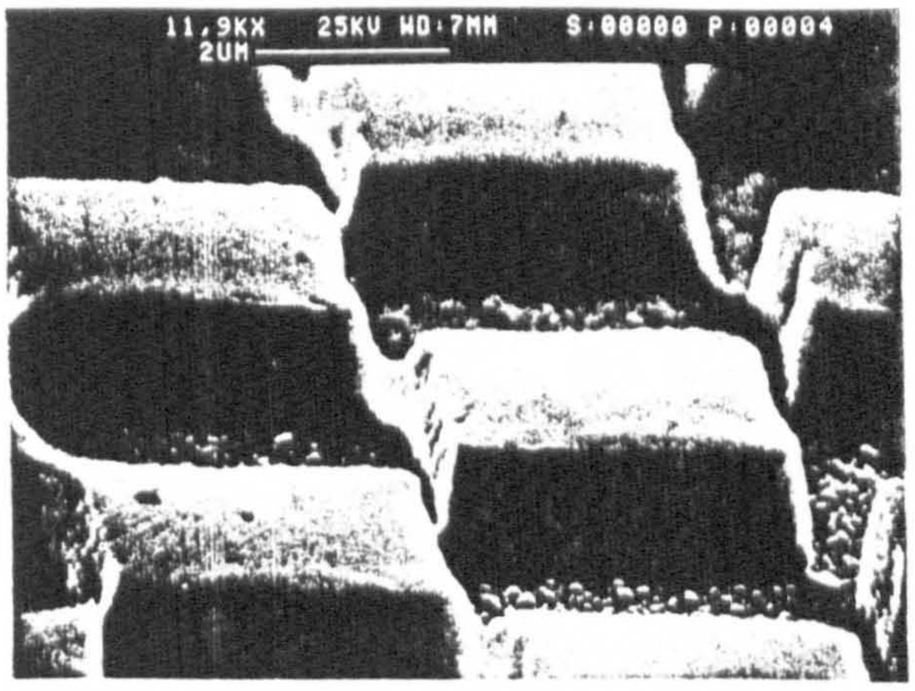


Figure 1 : A developed photoresist pattern in the form of a graded chequer board with 4 micron squares.

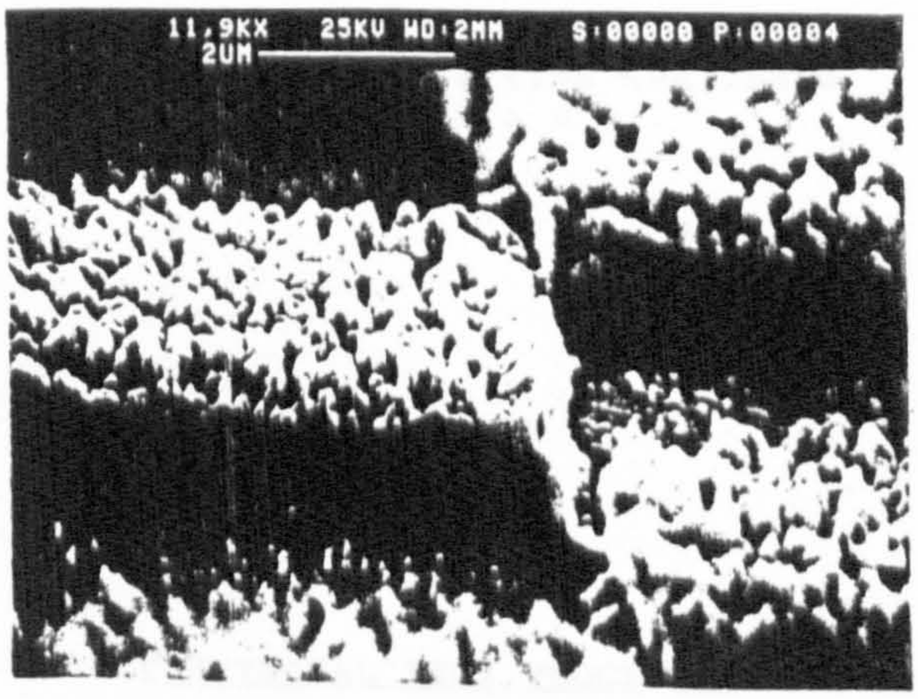


Figure 2 : The effect of the RIE process on the photoresist shown above.

Conventionally, SEM or optical micrographs such as these would be required after each experiment to assess its effect on metal pattern quality. Such methods are tedious, require expert interpretation, and may not detect some of the problems which can arise.

TEST STRUCTURES

An objective method has been developed using electrical conductivity tests on structures designed to highlight photoresist and etch faults. The structures were designed to change electrical continuity should a fault develop, e.g. resist failure could be detected by the breaking of an otherwise complete path and incomplete resist clearing could be detected by bridging between adjacent conductors.

The meandering structures checked for breaks which might be caused by resist failure or by cracks in step coverage or by overetching, whilst the interdigitated fingers checked for bridging which might be caused by underexposure, incorrect focus or underetching. These four test structures were drawn in 10 different sizes on each die, as recorded in Table 1, to allow geometry-dependent tests. For simplicity, the mark-to-space ratio on all patterns was unity.

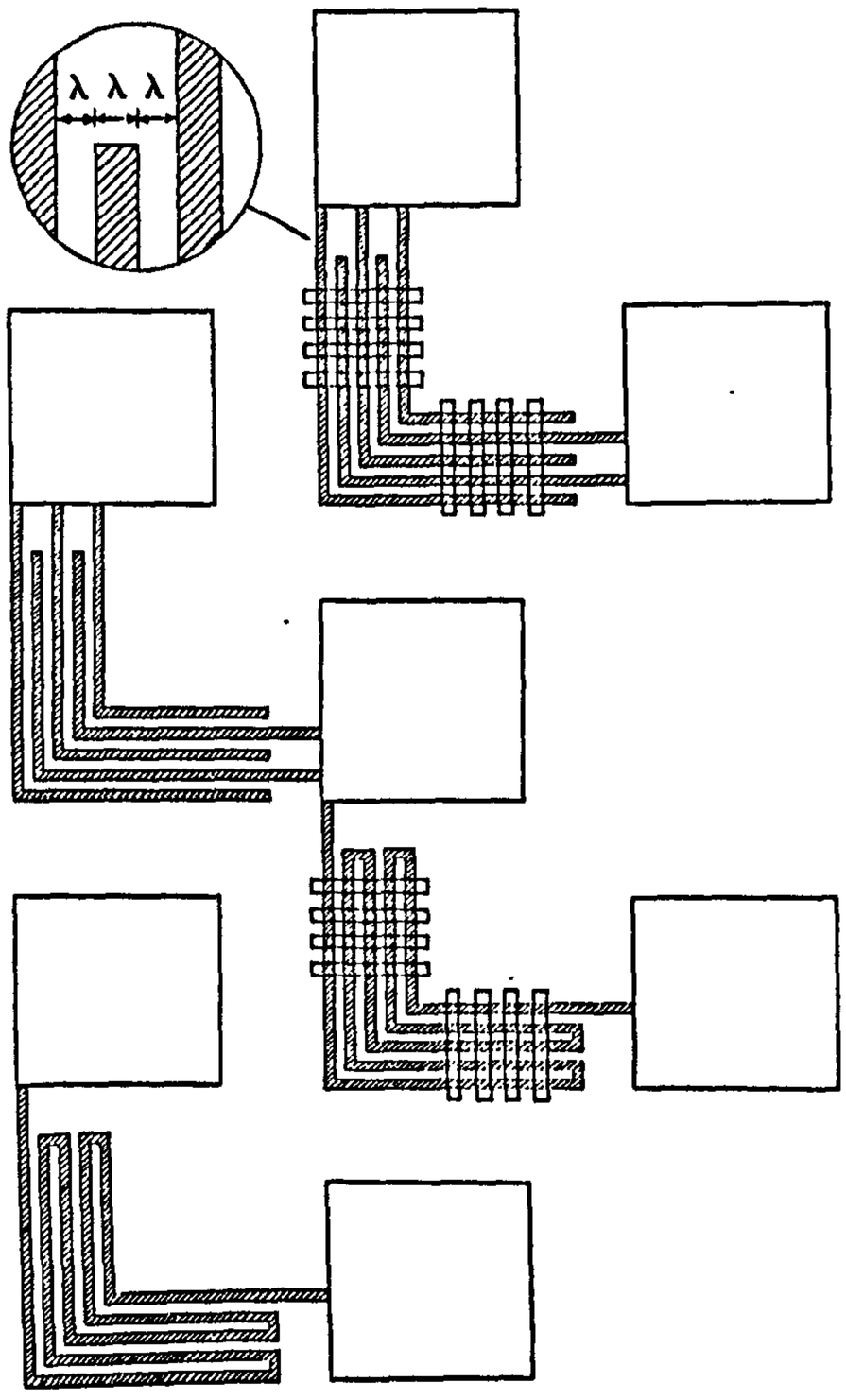


Figure 3 : The test structures.

Geometries	
Structure	λ (microns)
1	0.9
2	1.0
3	1.1
4	1.2
5	1.3
6	1.4
7	1.5
8	2.0
9	2.5
10	5.0

Table 1 : The geometries available within each die.

Four test structures were employed as shown in Figure 3. Two basic structures were printed on a flat surface and over an irregular surface caused by an underlying array of polysilicon bars, each bar being 10 microns wide and approximately 0.5 microns thick. A layer of pyrolytic oxide, approximately 0.8 microns thick, was deposited over these bars and reflowed prior to sputter coating with a 1 micron layer of Al-Si metallisation to simulate the conditions encountered in a typical MOS process.

ELECTRICAL TEST PROCEDURE

Wafers containing etched test structures can be quickly evaluated using an automatic wafer prober under computer control. The die is 4mm square, giving 244 samples on a 3 inch wafer. A selected geometry can be checked on all 244 sites in about 3 minutes, with the results being displayed as a wafer map. The display format is flexible, but the presentation shown in the following figures has been found to be convenient as the predominant cause of failure is easy to see. Clearly, additional data processing could be provided to give a statistical summary of each wafer.

SOME EXPERIMENTAL RESULTS

This technique has been used to compare the process latitude of different photoresists. Test wafers were printed with an exposure/focus matrix using a wafer stepper (Optimetrix 8010, 10X, g-line, 0.32 NA.) The exposure increased from left to right and the focus incremented from bottom to top. In principle, the position of the centre of the patch of good dice on the wafer indicates the optimum exposure and focus and the size of the patch (or the number of good dice) indicates the process tolerance or latitude. Figure 4 shows the results of probing the 2.0 micron tracks on a wafer which was printed with such an exposure/focus matrix. It can be seen that the best focus and exposure occurs near the centre of the wafer. At the left hand side there are shorts between fingers and to the right there are breaks in meanders, probably due to overexposure of the resist.

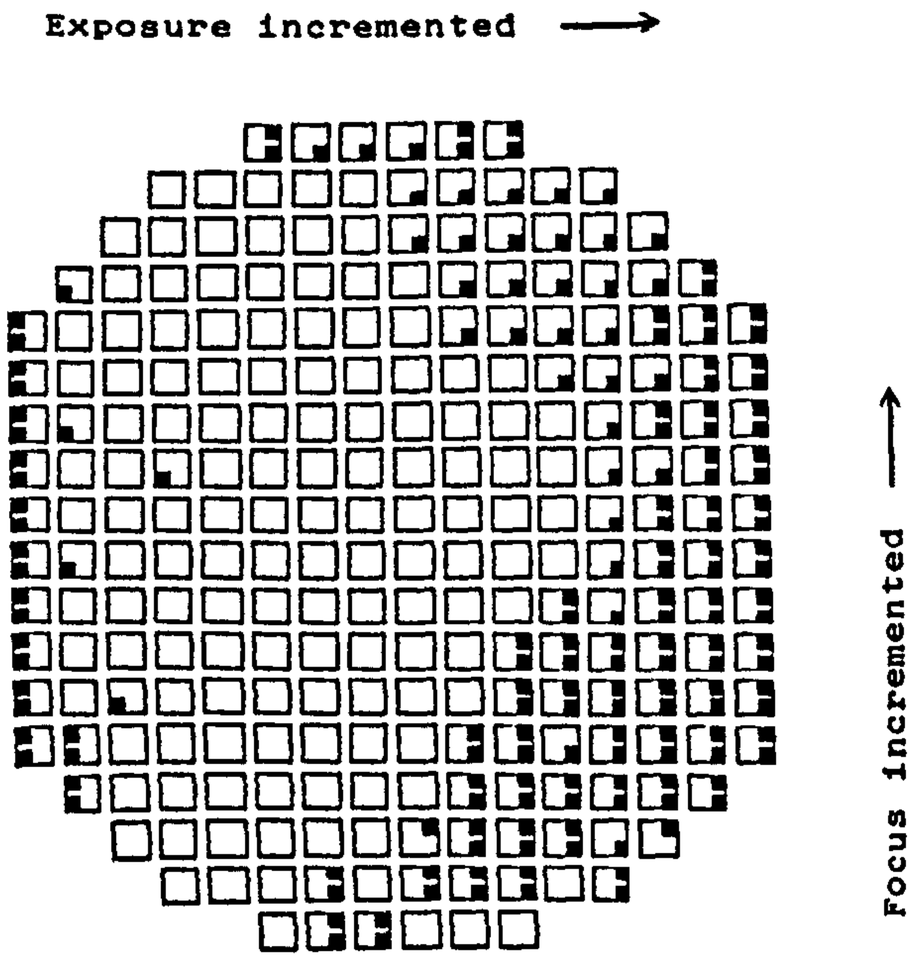


Figure 4 : Map for wafer number 7
For the 2 micron geometry
Exposure / Focus Matrix
Exposure increases from left to right
Focus is incremented from bottom to top

On all wafer maps, faults are marked as:-

- ▣ — fingers over topography shorted
- ▣ — fingers over flat shorted
- ▣ — meanders over flat open
- ▣ — meanders over topography open

In practice, the results from different resists can only be compared if the mean exposure has been chosen to match the sensitivity of the particular material, i.e. the exposure matrix must be normalised. With a monochromatic source, the threshold exposure dose (which just clears the resist after development) is a periodic function of thickness due to the effects of standing waves within the layer, as shown in Figure 5.

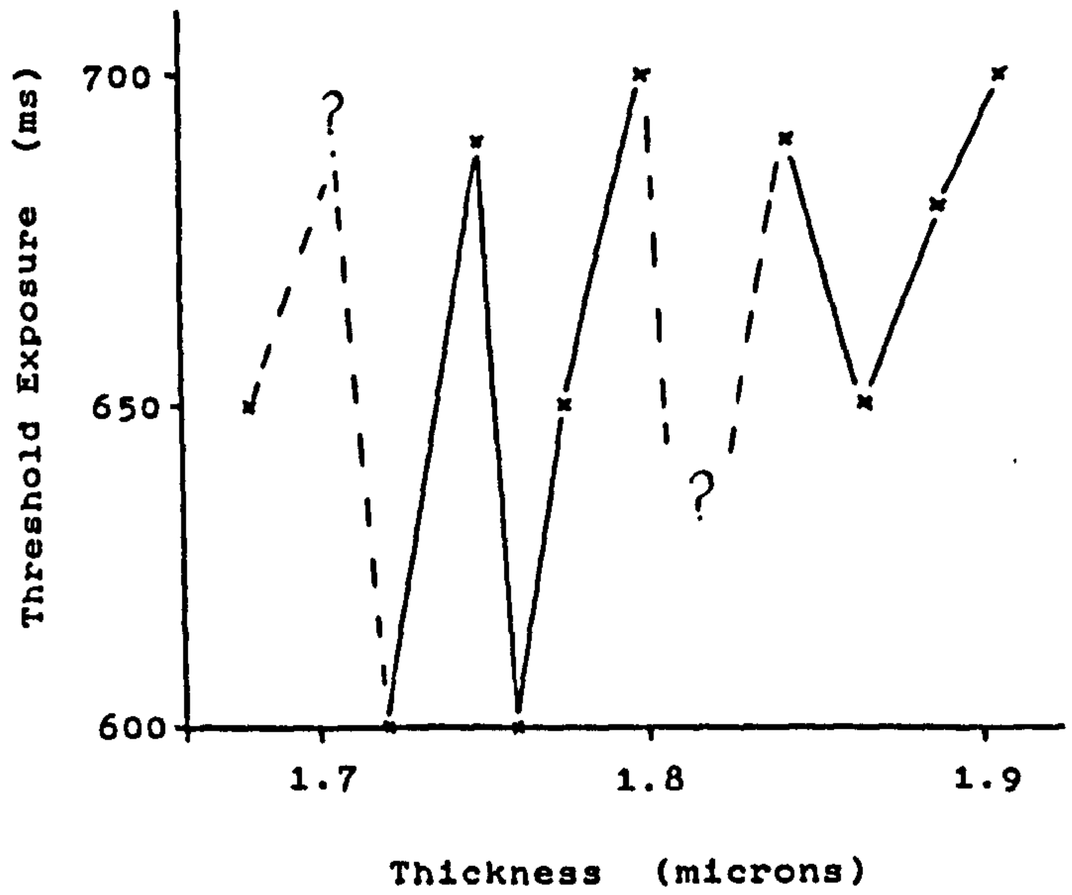


Figure 5 : Threshold Exposure v. Thickness for a dyed photoresist.

The choice of thickness is critical: the best results are achieved if the mean thickness of the resist corresponds to one of the exposure maxima, so that any area with a slightly different thickness will be overexposed, thus ensuring removal in the developer. Figures 6 and 7 show the results from two wafers which were coated with slightly different thicknesses of a dyed resist and then processed under identical conditions. The wafer with the thicker coating has cleared whereas the other shows bridging between fingers over topography.

This method of comparing resists has highlighted two other problems in the patterning of metal. Firstly, the RIE machine gives an etch rate which is not uniform across the wafer: it is lowest in the centre and highest at the periphery. This effect can be clearly seen in Figure 6 where the periphery of the wafer, including the lower exposure areas, does not show bridging between the fingers. Secondly, with anisotropic RIE there appears to be a persistent problem of metal fillets remaining along the edge of the polysilicon bars.

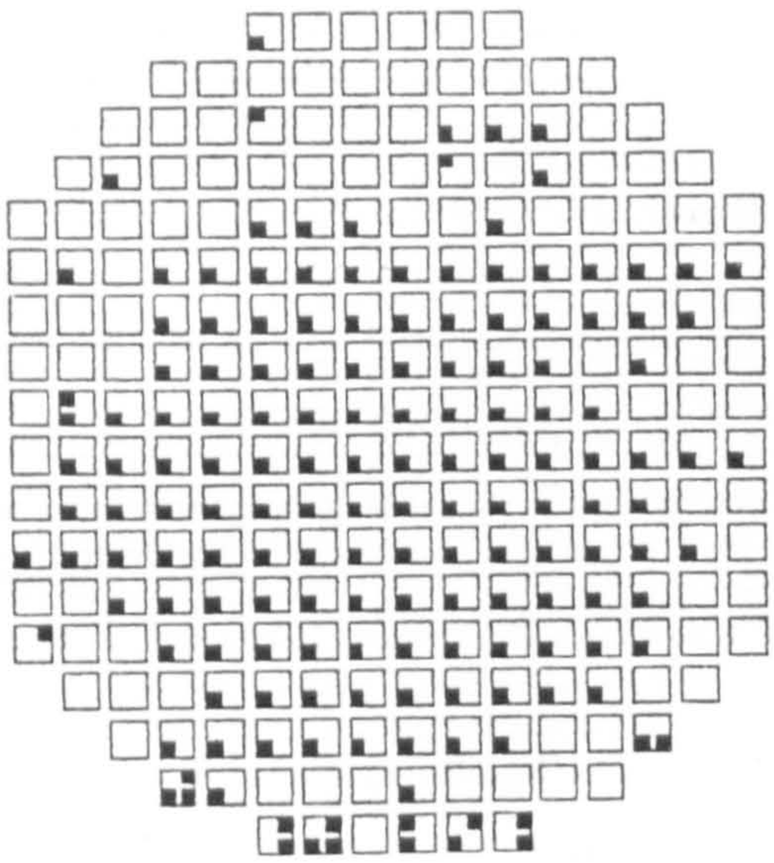


Figure 6 : Map for wafer number 6
For the 5 micron geometry
Resist thickness 1.71 microns
Exposure / Focus Matrix

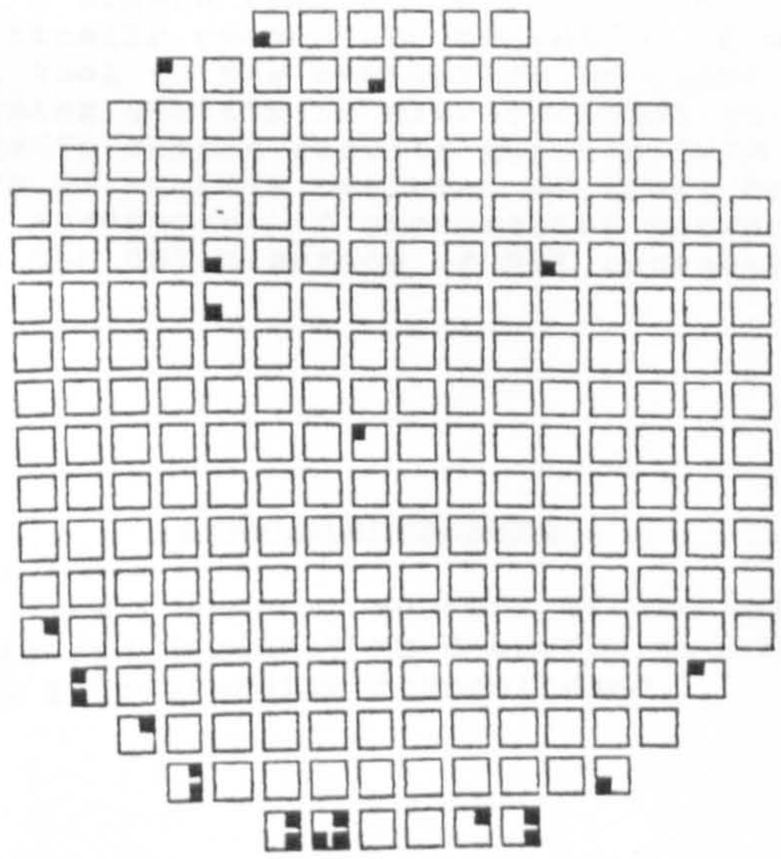


Figure 7 : Map for wafer number 5
For the 5 micron geometry
Resist thickness 1.79 microns
Exposure / Focus Matrix

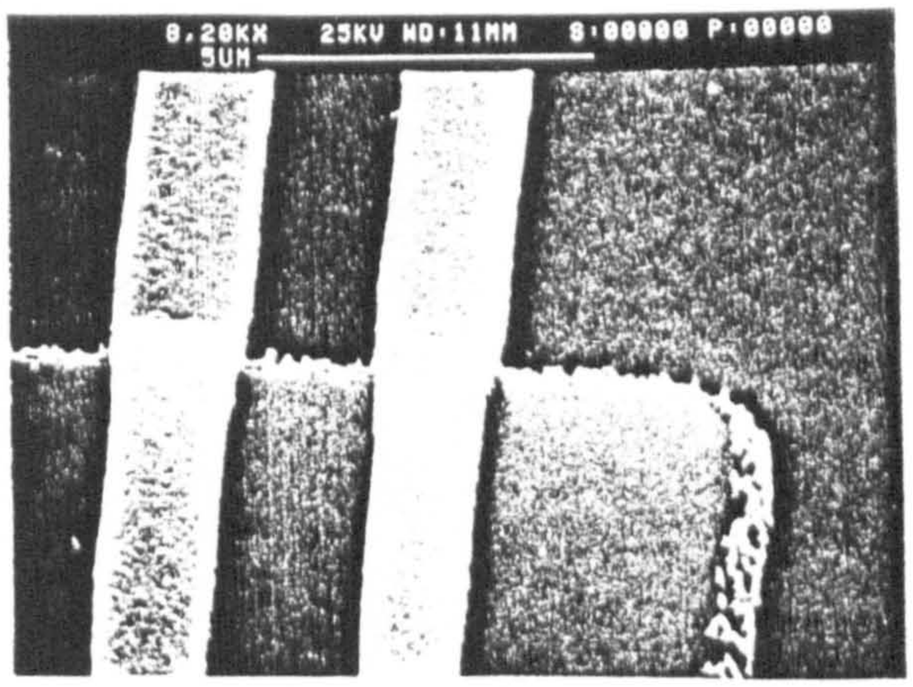


Figure 8 : Residual metal fillets after RIE
causing bridging between tracks

These fillets are very difficult to see on an optical microscope but show clearly on a SEM (Figure 8) and are readily detected by the test structure. Attempts to remove them by extending the etch time resulted in failure of the resist and breaks in meanders. One solution, which has been adopted as an interim measure, is a 15 second dip in isotropic wet etch after RIE. The benefit can be seen by comparing the wafer maps in figures 9 and 10, where the wafer which had a 15 second wet etch after RIE showed much less bridging between tracks. However, this improvement is gained at the expense of poorer control of linewidth and alternative RIE conditions are also being investigated, using this test procedure as an objective measure of any improvements achieved.

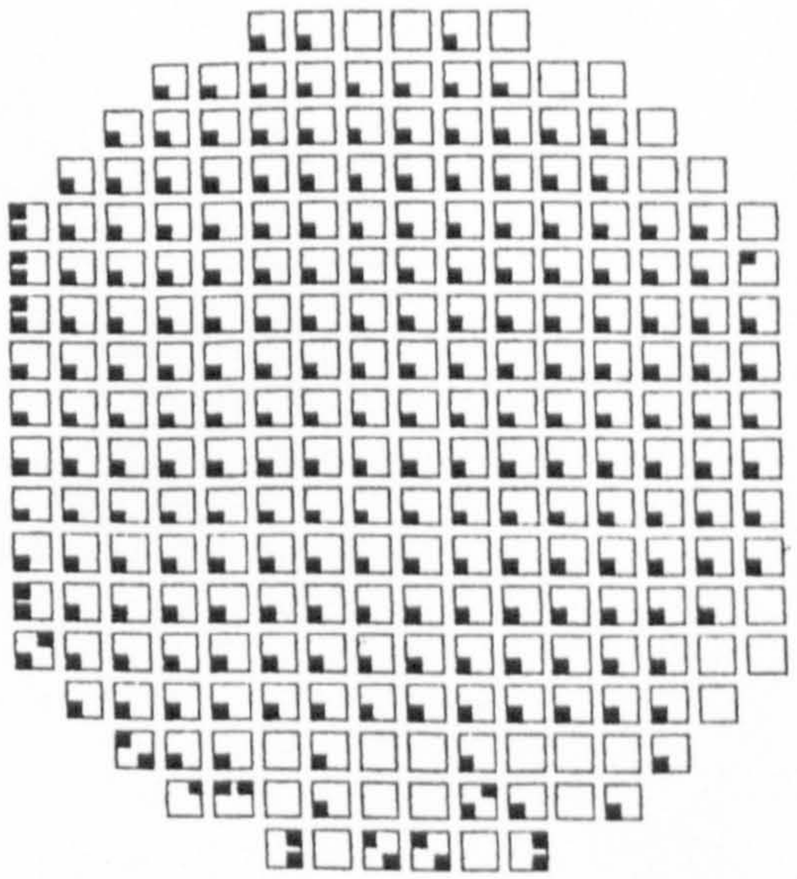


Figure 9 : Map for wafer number 0
For the 5 micron geometry
Exposure / Focus Matrix
Reactive Ion Etched

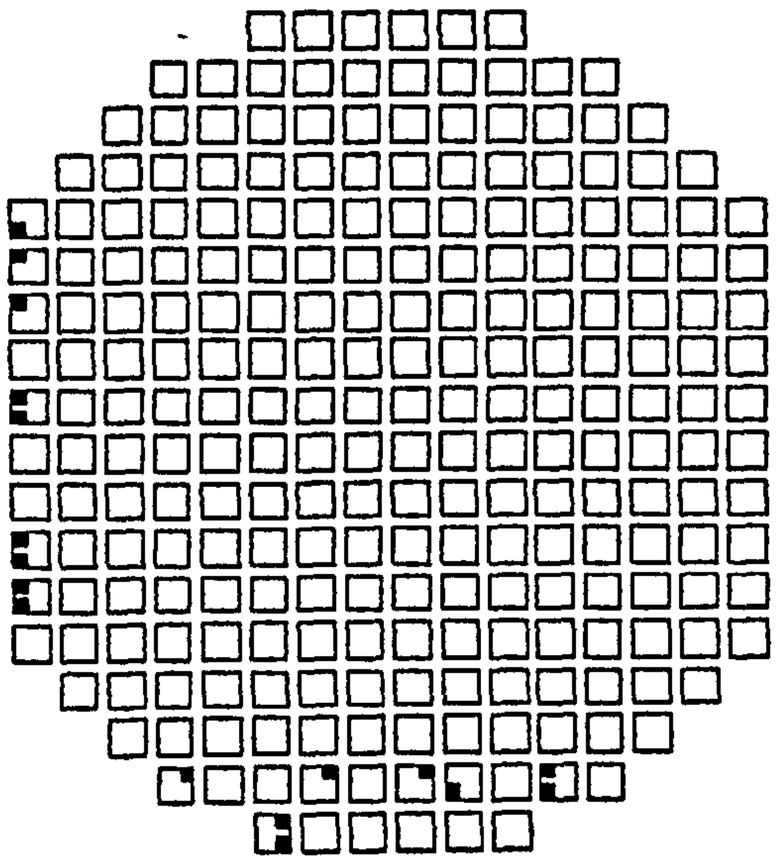


Figure 10 : Map for wafer number 7
For the 5 micron geometry
Exposure / Focus Matrix
RIE + 15 second wet etch

CONCLUSION

A simple test pattern which can be automatically probed has proved to be a useful tool in the evaluation of metal patterning quality in microfabrication. The ability to obtain results quickly from large numbers of samples has been of great benefit in the assessment of photoresist performance and in the optimisation of RIE processes.

ACKNOWLEDGEMENTS

The financial support of the U.K. Science and Engineering Research Council (SERC) is gratefully acknowledged.