ELECTRICAL AND THERMAL MODELLING OF POWER SEMICONDUCTOR DEVICES USING NUMERICAL METHODS

August 1988

Philip Walker

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by Philip Walker.

To the memory of my mother

ROSEMARY JILL WALKER

and to my father

HENRY DAVID WALKER

Abstract

As semiconductor devices continue to evolve the need for more accurate models in device simulation and design becomes greater. Improvements in process technology have meant device dimensions have fallen, putting a greater requirement on the accuracy of device models. This requirement has been fulfilled to a large extent by numerical models, which unlike analytical models can accurately account for the detailed nature of modern VLSI devices. Moreover, the allowable geometries and bias conditions of most analytical models are often too restrictive to provide adequate simulation.

A two dimensional numerical model has been developed, which is capable of modelling a wide variety of different devices over a full range of bias levels. Unlike other numerical models it contains the ability to simulate self-heating and electro-thermal interaction under conditions of high power dissipation.

A highly detailed mathematical model for semiconductor device physics is described. The model serves to couple a solution of Poisson's equation and the current continuity equations with a solution to the heat flow equation. In this way electro-thermal interaction is accurately modelled. The temperature dependence of the physical parameters required by the model has been included wherever possible and novel features include current flow due to temperature gradients, heavy doping effects, carrier-carrier scattering, temperature dependent thermal conductivity and Shockley-Read-Hall and Auger recombination.

The resulting continuous problem was transformed into a discrete one by the application of the classical finite difference method. Two orthogonal co-ordinates were considered, these being cartesian co-ordinates (x,y) and cylindrical co-ordinates (r,y). Very little additional effort is required to obtain the discrete cylindrical problem and yet it accurately defines a device with cylindrical symmetry in all three dimensions. For this reason many of the simulations have concentrated cylindrical type structures.

Having obtained the discrete problem which in general is non-linear, it has been solved using Newton's method. The high power operating condition represents a most severe numerical problem owing to the strong interaction which

exists between the equations that constitute the model, under such conditions. This has led to the requirement for two different solution procedures, these being the decoupled approach, which is most efficient at low injection levels and the coupled approach, which is more efficient at higher injection levels. The application of either approach requires the repeated solution of a large system of simultaneous equations. A number of iterative techniques such as the Strongly Implicit Procedure (SIP), Successive Line Over-Relaxation (SLOR) and Successive Block Over-Relaxation (SBOR), that have proven to be most efficient for solving these systems are described.

Numerical results have been obtained at various stages in the development of the model. Initially a simple Poisson solver was written, and by coupling this with a numerical solution of the ionization integral a very powerful 'off-state' model was obtained. This model is capable of accurately predicting breakdown voltages and it was used to investigate the various techniques for improving the curvature related breakdown voltage of planar diffused junctions. The technique of 'resurf layers' proved to be the most effective and the optimum parameters giving maximum breakdown voltage improvement were obtained.

Results from the off-state model were incorporated into a number of bipolar transistors that have been designed and fabricated as part of this project. Details of these devices are given, together with there characterisation, which was performed with the aid of the model.

Having added the steady state current continuity equations to the model it was used to investigate the consequences of emitter widths on the pulsed power handling capability of interdigitated bipolar transistors. Transient effects were then included in the model, which was also coupled with a collector circuit equation in order to model transistor operation during inductive switch-off. The results of these last two investigations indicated that the trend should be towards finer emitter structures.

Finally a solution to the heat flow equation was incorporated into the model, which was then used to investigate transient self-heating phenomena leading to thermal second breakdown in bipolar transistors. A comparison was made with experiment through the construction of a non-destructive test circuit and the potential benefit to be gained from grading the collector impurity profile was investigated.

Although all the simulations performed in this work have been directed at bipolar transistor operation, the model has been designed in such a way that it would be a very simple matter to alter the model to suit MOST or JFET operation.

Acknowlegments.

I would like to express my sincere thanks to my supervisor, Dr. Keith I. Nuttall, for his continuous support throughout this project. The many stimulating discussions we shared, especially those concerning semiconductor physics, have given me an excellent grounding to my career and will prove invaluable in the years to come. I am also grateful to the Head of the Department of Electrical Engineering and Electronics, Prof. W. Eccleston for providing the facilities required in this project.

I am extremely grateful to a number of colleagues from the Department of Electrical Engineering and Electronics. Firstly, I would like to thank Julian Humphreys, with whom I worked in close collaboration. We have exchanged many interesting thoughts with regard to numerical modelling and device physics, which I have found to be particularly helpful. Special thanks also goes to Dr. Trevor Davies, for supplying me with a copy of his numerical model at the start of my project. This provided me with a very firm foundation upon which I was able to build. Thanks also to Dr. Steve Hall for demonstrating the use of the GAELIC mask drawing facility and also for showing a keen interest in my work. I am also grateful to Andrew Jackson for help with circuit design and to Andrew Demery for all his time spent helping me with my many software problems. Sincere thanks also go to Bill Gould for his technical support and to all the other technical staff for their help and friendship.

I wish to extend my thanks to all my present colleagues in the Discrete Devices Group at Philips Research Laboratories, Surrey. In particular I am very grateful to David Coe and David Paxman for their patience and understanding during the preparation of this thesis.

I would like to thank my parents for the faith and support they gave me throughout my time spent working on this project.

Finally, I wish to acknowledge Ferranti Electronics plc. for providing a CASE award and to the Science and Engineering Research Council for funding this project.

Table of Contents

Chapter 1. Introduction. **1**
1.1 Perspective. 1
1.2 History and Review. 3
1.3 Objectives and Overview. 6
References. 7

Chapter 2. Fundamental Basis For Numerical Modelling. **11**
2.1 Poisson's Equation. 11
2.2 Current Continuity Equations. 12
2.3 Current Transport Equations. 13
2.4 Carrier Concentrations. 20
2.5 Heat Flow Equation. 25
2.6 Boundary and Interface Conditions. 25
2.6.1 Boundary and Interface Conditions for Electrical Model. 28
2.6.2 Boundary and Interface Conditions for Thermal Model. 31
References. 32

Chapter 3. Models for Physical Parameters and Device Attributes. **34**
3.1 Carrier Mobility. 35
3.1.1 Phonon Scattering. 35
3.1.2 Impurity Scattering. 36
3.1.3 Carrier-Carrier Scattering. 43
3.1.4 Velocity Saturation of Carriers. 45
3.2 Intrinsic Carrier Concentration. 52
3.3 Band-gap Narrowing. 52
3.4 Carrier Recombination/Generation. 56
3.4.1 Shockley-Read-Hall Recombination/Generation. 57
3.4.2 Auger Recombination. 59
3.4.3 Carrier Lifetimes. 60
3.5 Thermal Conductivity. 65
3.6 Heat Generation 67
References. 68

Chapter 4. Discretization Of The Governing Equations. **72**
4.1 Discretization Of The Static Equations. 73
4.1.1 Cartesian Co-ordinates. 73
4.1.1.1 Discretization of the Poisson Equation 76
4.1.1.2 Discretization of the Continuity Equations. 78
4.1.1.3 Discretization of the Heat Flow Equation. 88
4.1.1.4 Discretization of the Boundary Conditions. 90
4.1.2 Cylindrical Co-ordinates. 98
4.1.2.1 Discretization of the Poisson Equation. 98
4.1.2.2 Discretization of the Continuity Equations. 99
4.1.2.3 Discretization of the Heat Flow Equation. 100
4.2 Discretization of the Dynamic Equations. 100
4.3 Mesh Generation. 103
References. 105

Chapter 5. The Solution of Large Systems of Non-Linear Algebraic Equations Arising from the Semiconductor Problem. **107**
5.1 Decoupled Solution Procedure. 109
5.1.1 Initialisation Procedure. 111
5.1.2 The Solution of Poisson's Equation. 111
5.1.3 Solution of Electron Current Continuity Equation. 114
5.1.4 Solution of Hole Current Continuity Equation. 116
5.1.5 Solution of the Heat Flow Equation. 117
5.1.6 The Solution of Linear Systems Arising from the Decoupled Procedure. 118
5.1.6.1 Successive Line Over-Relaxation (SLOR). 120
5.1.6.2 The Strongly Implicit Procedure (SIP). 122
5.2 Coupled Solution Procedure. 132
References. 139

Chapter 6. Results. **141**
6.1 An Investigation into the Techniques for Improving the Curvature Related Breakdown Voltage of Diffused Junctions. 142
6.1.1 Introduction. 142
6.1.2 Breakdown Voltage Modelling Considerations. 142
6.1.3 Field Ring Considerations. 150
6.1.4 Results. 153
6.1.5 Conclusion. 156
6.2 Design and Characterization of Bipolar Test Structures. 156
6.2.1 Mask Design. 156
6.2.2 Device Processing. 162
6.2.3 Device Characteristics and Comparison with Model. 167

6.2.4 Summary. 171
6.3 Optimization of Emitter Widths. 172
6.3.1 Introduction. 172
6.3.2 A Comparison of Analytical and Numerical Models for Emitter Pinch. 173
6.3.3 Optimization of Emitter Width. 177
6.3.4 Conclusion. 188
6.4 A Model for Inductive Switching using Bipolar Transistors. 190
6.4.1 Introduction. 190
6.4.2 Coupling of the Numerical Model with the Collector Circuit Equation. 192
6.4.3 Turn-Off of Transistors with Cylindrical Geometry. 192
6.4.4 Discussion. 207
6.4.5 Conclusion. 212
6.5 Electro-thermal Interaction and Thermal Second Breakdown in Bipolar Transistors. 213
6.5.1 Introduction. 213
6.5.2 A Non-Destructive Test Circuit for Pulsing Transistors into Thermal Second Breakdown. 215
6.5.3 Experimental Results. 218
6.5.4 Numerical Results. 222
6.5.5 Comparison of Experimental and Numerical Results. 232
6.5.6 The Consequences of Graded Collector Profiles on Thermal Breakdown. 235
6.5.7 Conclusion. 242
References. 243

Chapter 7. Conclusion. 246
7.1 Summary. 246
7.2 Recommendations. 249
References. 252

Appendix. 253

List of Illustrations

Figure 1. A Bipolar Transistor Structure Illustrating Solution Domains and Boundaries. 27
Figure 2. Mobility of Electrons due to Lattice Scattering in Silicon. 37
Figure 3. Mobility of Holes due to Lattice Scattering in Silicon. 38
Figure 4. Mobility of Electrons due to Lattice and Impurity Scattering in Silicon 41
Figure 5. Mobility of Holes due to Lattice and Impurity Scattering in Silicon 42
Figure 6. Mobility of Electrons due to Carrier-Carrier Scattering 45
Figure 7. Mobility of Holes due to Carrier-Carrier Scattering 46
Figure 8. Mobility of Electrons and Holes Versus Electric Field 50
Figure 9. Drift Velocities of Electrons and Holes Showing Velocity Saturation 51
Figure 10. Intrinsic Carrier Concentration in Silicon from 300K to Melting Point. 53
Figure 11. Band-gap Narrowing as a Function of Total Impurity Concentration 55
Figure 12. Circuit and Voltage Waveform for Lifetime Measurement using the OCVD Technique. 63
Figure 13. Thermal Conductivity of Silicon from 300K to Melting Point. 66
Figure 14. A Typical Mesh. 75
Figure 15. The Five Point Molecule 76
Figure 16. The Growth Function 83
Figure 17. The Growth Function 84
Figure 18. Modulus of the bracketted term in equation (4.42) 85
Figure 19. Illustration of Newton's Method. 109
Figure 20. Flow Diagram Illustrating Decoupled Solution Procedure. 110
Figure 21. Example of a Linear System Resulting from the Decoupled Solution Procedure. 119
Figure 22. Lower and Upper Triangular Matrices for a 4 by 6 Mesh. 124
Figure 23. The LU Product. 125
Figure 24. Mesh in Vicinity of Point i,j. 127
Figure 25. Flow Diagram Illustrating Coupled Solution Procedure. 133
Figure 26. Example of a Linear System Arising from the Coupled Solution Procedure. 137

Figure 27. One-Dimensional Potential Profiles After Successive Iterations of the SIP. 143
Figure 28. Diagram Illustrating the Technique used to Follow a Flux Line Passing Though Mesh Point i,j. 145
Figure 29. Voltage Contours and Flux Lines for a Curved Junction at a Straight Mask Edge 148
Figure 30. Field Contours and Breakdown Locus for a Curved Junction at a Circular Mask Edge 149
Figure 31. Device with a Single Field Limiting Ring 150
Figure 32. Illustration of Technique to Located Field Ring Voltage. 152
Figure 33. Breakdown Voltage as a Function of Resurf Layer Doping. 155
Figure 34. Masks for Devices A and B. 157
Figure 35. Masks for Devices C and D. 158
Figure 36. Masks for Devices E and F. 159
Figure 37. Vertical View of Single Emitter Finger Metalization. 160
Figure 38. Output Characteristics 168
Figure 39. Base and Collector Currents as a Function of Base-Emitter Voltage for Device C 170
Figure 40. A Comparison of Models for Emitter Pinch 175
Figure 41. A Comparison of Models for Emitter Pinch 176
Figure 42. One Dimensional Electric Field Profiles at Various Current Densities 180
Figure 43. Current Flow Profiles Just Prior To Breakdown for Various Emitter Widths 181
Figure 44. Current Densities Along Emitter-Base, Base-Collector and Collector-Substrate Junctions 182
Figure 45. Electric Field Profiles Just Prior to Breakdown 183
Figure 46. Average Collector Current against Emitter Width. 185
Figure 47. Safe Operating Areas for Different Emitter Widths 187
Figure 48. Safe Operating Areas for Different Emitter Widths on a Logarithmic Axis 188
Figure 49. Variation of Current Gain and Base-Emitter Voltage with Collector Current for Various Emitter Widths 189
Figure 50. An Inductive Collector Load with a Free-Wheeling Diode. 191
Figure 51. Flow Diagram of Coupling Procedure. 193
Figure 52. Voltage and Current Transients During Switch-Off. 195
Figure 53. A Schematic of the Collector Load Line 196
Figure 54. Hole Concentration Profiles at Various Times During Switch-Off . . 197
Figure 55. Hole Concentration Profiles at Various Times During Switch-Off. . 198

Figure 56. Vertical Electron Current Densities at Emitter-Base, Base-Collector and Collector-Substrate Junctions During Switch-Off 199
Figure 57. Electron Current Density Profiles Showing Divergence 200
Figure 58. Variation of Electric Fields During Turn-Off for Circular Emitter Structure. 203
Figure 59. Electric Field Profiles at Various Times During Turn-Off. 204
Figure 60. Electric Field Profiles at Various Times During Turn-Off. 205
Figure 61. Variation of Electric Fields During Turn-Off for Ring Emitter Structure. 207
Figure 62. Electric Field Profiles at Various Times Before the Emitter-Base Junction Has Recovered. 208
Figure 63. Vertical Current Densities at Various Times Before the Emitter-Base Junction Has Recovered. 209
Figure 64. Estimated RBSOA for Circular and Ring Emitter Devices. 211
Figure 65. Non-Destructive Test Circuit for Investigating Thermal Second Breakdown. 216
Figure 66. Base Current Waveforms During the Delay Time to Breakdown .. 219
Figure 67. Emitter Current Waveforms During the Delay Time to Breakdown 221
Figure 68. Base Current Waveforms for Graded Collector Profiles 222
Figure 69. Computed and Experimental Base Current Waveforms 224
Figure 70. Temperature Profiles at Various Times Before Breakdown. 225
Figure 71. Temperature Profiles at Various Times Before Breakdown. 226
Figure 72. Contour Maps of Hole Concentration at Various Times Before Breakdown 228
Figure 73. Contour Maps of Hole Concentration at Various Times Before Breakdown 229
Figure 74. Electric Field Profiles Prior To Breakdown. 230
Figure 75. Variation of Peak Temperature and Peak Field due to Self-Heating 232
Figure 76. Vertical Electron Current Densities for the Ring Emitter Structure 233
Figure 77. Base Current Waveforms for Graded Collector Profile 236
Figure 78. Variation of Peak Temperature and Peak Field for Uniform and Graded Collector Profiles. 237
Figure 79. Field Profiles Prior to Breakdown for the Graded Collector 238
Figure 80. Vertical Electron Current Densities Prior to Breakdown for the Graded Collector. 239
Figure 81. Temperature Profiles Prior to Breakdown for the Graded Collector. 240
Figure 82. Contour Maps of Hole Concentration Prior to Breakdown for the Graded Collector 241

List of Tables

Table 1. Lattice Mobility Constants. 36

Table 2. Constants for Mobility due to Impurity Scattering. 39

Table 3. Constants for Mobility due to Combined Effects of Lattice and Impurity Scattering. 40

Table 4. Parameters in Equation (3.14) for Effects of Velocity Saturation of Carriers on Mobility. 48

Table 5. Lifetime Constants. 61

Table 6. Specific Heat and Density of Silicon and Mild Steel. 67

Table 7. Estimated (Old) and Measured/Calculated (New) Parameters. 169

Chapter 1. Introduction.

1.1 Perspective.

The rapid growth of the modern electronics industry was initiated in 1947 with the invention of the point contact bipolar transistor [1.1]. The second major breakthrough was made in 1960 when the first fully operational MOST was demonstrated [1.2]. Soon after this the first integrated circuit became commercially available and rapid advancement has continued so that by today silicon chips are being manufactured which contain over 400,000 devices. As devices have become smaller and more complex there has been a continual need to up-date and improve the models used to represent their structure and operation. Accurate device modelling provides the foundation for the entire semiconductor industry since it is the only way that the device designer can study the detailed internal operation of particular device. In the early classical models for device operation the device was sectioned into various regions, each being treated as a separate entity. These closed form solutions were then matched at the boundaries of the various regions to provide a global solution. Using this procedure it was often necessary to make rather inaccurate assumptions, both within the regions and at their boundaries. Moreover, these models were usually only applicable in one dimension only, with other dimensions often being taken into account by making use of an over simplified distributed one dimensional model.

The use of modern numerical modelling techniques largely overcomes the problems associated with classical approaches. Numerical methods, like classical ones, also require that the device be segmented into separate regions, however, in this instance a self-consistent global solution is obtained, which satisfies each individual region. The solutions are 'automatically' matched at the boundaries between the regions, which are usually made very much smaller than the device dimensions. This allows for very precise evaluation of device operation. The requirement for such a large number of regions imposes an extremely large computational problem and as such numerical solution techniques have only

become feasible in the last 20 years or so, with the advent of powerful mainframe computers.

Power conversion and control provided the very first use for semiconductor devices with applications in digital electronics being realised at a later date. By today a plethora of different types of power devices exist, which use quite different operating principles. At present one specific type of device cannot be singled out as giving a better performance over all the others across the entire power spectrum which ranges up to 10*KV* and 5000*A*. Rather, each particular device has its own advantages and disadvantages in each particular operating region. Power thyristors are at present the only devices that are capable of withstanding the extremely high blocking voltages and conduction currents, but they are only able to achieve this at the cost of speed. The Gate Turn-Off thyristor (GTO), finds its application in the slightly lower voltage range, ie. 1000-2000*V* and is able to handle currents of up to 1000*A*. The power bipolar (BJT) and Darlington transistors are most appropriate between about 250 and 800*V* up to a maximum current rating of 1000*A*. The more recent power MOS devices will find use in high frequency applications up to 1000*V* and 100*A*. The reason for such a low current limit is that these devices suffer from high on-resistance losses since they do not make use of any conductivity modulation effect. However, this very fact means that they are able to operate at ultra high frequencies ($> 500MHz$). Even more recently the invention of the Insulated Gate Transistor (IGT) [1.3] has put the use of conventional bipolar transistors under serious threat. These devices require very little gate drive power and can be driven from low cost integrated electronics, making them extremely attractive indeed. The problem of very high input power dissipation that exists with the bipolar transistor has been alleviated somewhat by the use of the Darlington configuration [1.4], but it is still unable to compete with the IGT in this respect. A number of disadvantages do, however, exist with the IGT the most important being the difficulty of achieving rapid removal of stored charge and the diode voltage drop, which is inherent between source and drain. Nevertheless it has been predicted by some that the power bipolar transistor is likely to be superseded by power MOS devices in the near future [1.5]. This prediction was based on a comparison between IGT's fabricated using modern techniques and conventional bipolar designs. However, a number of bipolar geometries have recently come to light [1.6], [1.7], which have been reported to give extremely impressive operating characteristics. They have been fabricated using Power MOS technology and consist of multiple cellular type emitters. These devices have been reported to be exceptionally fast with the ability to handle very high power. Thus, it would seem fairer to compare this new class of bipolar devices against the IGT, since both use the same state of the art technology. It is

expected that these devices will compare much more favourably with the IGT, and it may well mean that the BJT will maintain its position in the market for many years to come.

1.2 History and Review.

The first devices to exhibit any significant power handling were demonstrated by Hall in 1952 [1.8]. These were rectifiers formed from germanium mesa alloy junctions and had blocking voltages of 200*V* and forward current rating of 35*A*. In the mid 1950's single crystal silicon became available and the larger bandgap of silicon compared with germanium meant a considerable improvement in performance. By the late 1950's diffused junctions were being combined with mesa etching in order to improve voltage blocking capabilities. By 1964 this technique was realising breakdown voltages of up to 2*KV* [1.9]. The first commercially available transistors were manufactured by Texas Instruments in 1954 [1.10] and yet it was not until a decade after this that applications to high power handling began. However, development has been so rapid that by today thyristors are available with current ratings of 6000*A* and blocking voltages of 4000*V*. The development of the power transistor accelerated in the 1960's with the introduction of the planar process by Fairchild [1.11] and the use of photolithography techniques for accurate device definition. The introduction of the epitaxial process [1.12] also proved to be vital. A great deal of effort has since been made to optimise designs in order to relax the trade-off that exists between power capability and speed.

By the early 1980's the state of the art was represented by by extremely large power transistors with diameters of 38*mm* that were capable of delivering a continuous collector current of 200*A* with a collector-emitter open base breakdown voltage (BV_{CEO}) of 850V [1.13]. These transistors were able to conduct a collector current of 100A with a h_{FE} of 8 at a collector emitter saturation voltage of 2*V*. The use of Darlington configured transistors was found to increase the h_{FE} by an order of magnitude [1.14], but resulted in a slower device as stored charge is not as easily removed. The power transistor is continuing to evolve so rapidly that these transistors are already beginning to be replaced by an exciting new type of transistor which is the result of a totally different design methodology. Rather than using very few emitters with large areas as was the case until very recently the current trend is towards using many fine emitters in a single device [1.15]. The development of such fine structures has been made possible by the use of modern MOS processing techniques [1.16]. The advantages of such designs are that stored

charge can be removed more readily and the detrimental effects of emitter current crowding are reduced as for both cases lateral base resistances are low. Such a design has been reported to have very impressive characteristics indeed [1.15]. A typical device was able to switch 50A through an inductive load with a supply voltage up to 1000V. The turn-off time for such an arrangement consisted of a storage time of only $2\mu s$ and a MOSFET like fall time of $30ns$.

Numerical device modelling has evolved over a similar time span as the transistor itself. Since this type of modelling is heavily dependent on computing power this fact is perhaps not too surprising. Details of the first truly numerical model were published by Gummel in 1964 [1.17]. In this paper he presented a one dimensional steady state model for a bipolar transistor, which for the first time used a single algorithm to simulate the entire device as opposed to a particular region within the device. This approach was extended by DeMari [1.18] [1.19], who used it to model steady state and transient operation of single *p-n* junctions. In 1969 Scharfetter and Gummel described a model for a silicon Read diode oscillator [1.20]. In this paper they presented a novel treatment of the current transport equations, which has since proven to be of utmost importance and has been incorporated in all subsequent models of note.

Owing to the limited computer resources that were available in the 1960's these first simulations had to be restricted to one dimension only. Some of the first bipolar transistor simulations in which two dimensions were taken into account were by Slotboom [1.21] and Heimeier [1.22]. During the 1970's and 80's a vast amount of literature has become available on the subject, covering all the important types of MOS and junction devices. For an extensive review of such literature the reader is referred to [1.23]. As far as power BJT's are concerned one of the first notable papers was by Manck et al. [1.24] in which a two dimensional simulation of high injection effects was presented. In order to model these effects it was necessary to use a different solution procedure to that proposed by Gummel in 1964, and which had been used in all previous models. This technique has also proven to be absolutely vital for modelling power devices and has since been widely employed. In a subsequent paper [1.25] Manck and Engl extended their model to include transient capability in order to simulate turn-off dynamics in lateral BJT's. In 1976 Gaur et al. [1.26] described a two dimensional steady state model capable of modelling non-isothermal phenomena. This model represents a primitive version of the model which forms a major part of this project. For the first time a numerical model was used to investigate electrical and thermal interaction due to self-heating. This model was later extended to cater for transient electro-thermal effects [1.27].

In 1978 Turgeon and Navon [1.28] published details of a transient numerical transistor model which was coupled with external base and collector circuit equations consisting of resistors and inductors. This allowed device operation to be evaluated for a specific requirement and in particular for switching inductive loads. An extremely well developed model was presented by Gaur [1.29] in 1977. Many of the effects which are important in characterising power BJT's were accounted for, including avalanche generation, high injection effects and self-heating.

In 1983 an extension for modelling devices with cylindrical symmetry was described [1.30], which has also implemented here. Modelling with cylindrical co-ordinates requires very little extra computational effort over modelling with cartesian co-ordinates and yet it avoids the need to make 2-D approximations, which often leads to significant inaccuracies. In 1986 a 2-D model was used to predict the onset of current mode second breakdown during switch-off with inductive loads [1.31]. The use of numerical modelling provided the necessary accuracy to simulate the detailed two dimensional current flow, which was essential in locating the precise point at which the device was expected to enter breakdown. Recently a similar analysis was reported [1.32], but this time the benefits to gained from altering the emitter structure were also considered.

As device dimensions have continued to fall the validity of the 2-D approximation has become subject to increasing uncertainty. Unfortunately the full 3-D analysis remains a considerable computational problem even for the most powerful modern day computers. As a result of this very few 3-D simulators have been developed and those that have been reported have to be run on extremely fast super-computers to keep run times down to an acceptable level [1.33] [1.34].

Numerical models have also been extensively used to model avalanche breakdown in reverse biased *p-n* junctions. These models are often called 'off-state' models as they do not incorporate the effects of current flow and as such are significantly easier to solve. They are, however, extremely useful for providing optimised data on the various field relief schemes that are used to protect junctions against premature breakdown. The first such model was reported in 1975 by Temple and Adler [1.35]. They considered the curvature related breakdown in single unprotected planar diffused junctions. Since then off-state models have been used to investigate most of the techniques for improving breakdown voltage such as - single field limiting rings [1.36], multiple ring systems [1.37], field plates [1.38], 'resurf' layers [1.39], depletion etch techniques [1.40], p-π-ν structures [1.41], positive angle bevelling [1.42] and negative angle bevelling [1.43]. Such an off-state model has been developed as part of this project [1.39] and close agreement has been obtained between predicted and actual breakdown values. A

number of other authors have also reported close agreement between computed and experimental results eg. [1.37] and [1.40], certifying the excellent accuracy of these models.

Numerical modelling has developed into a well established disciplin and is now widely utilised. A number of text books have become available which cover many aspects of this wide ranging subject [1.23] [1.44] [1.45]. The book by Selberherr [1.23] is extremely helpful and essential for anyone wishing to know more about numerical modelling. Two major conferences specifically concerned with numerical process and device simulation are now held bi-annually [1.46] [1.47] to cater for the interest that now surrounds the subject. In addition, a number of general modelling programs are commercially available, which can be tailored to suit a particular requirement, such as SEDAN [1.48] and BIPOLE [1.49] for 1-D simulations and BAMBI [1.50] for 2-D.

1.3 Objectives and Overview.

The primary aim of this project was to provide the University of Liverpool with a computer model which could accurately reflect the internal physical operation of discrete silicon devices in general. The intention being that the model could be applied to junction type devices eg. diodes, BJT's, thyrisors, as well as MOS devices and also to the more recent combined junction and MOS devices such as MIXFET's and COMFET's. It was necessary to use numerical methods to provide the sufficient degree of generality required to simulate such a diverse range of devices. These methods can be used to solve the universal set of equations that define the operation of all semiconductor devices, making them extremely versatile indeed. Such a solution can be obtained without having to make any initial over-riding assumptions and yet can yield knowledge of internal voltage and temperature profiles, carrier distributions and current flow in the most minute detail. A small amount of numerical modelling had already been carried out at Liverpool prior to the commencement of this project [1.51] and, as such, this work represents an extension to an existing model. However, a great many additions and improvements have been made, which largely outweigh the original work and, thus, for the sake of completeness the model will be described here in its entirety.

An additional objective of this project was to design a set of masks for the fabrication of several different power bipolar transistor structures. A number of circular and interdigitated structures have been subsequently designed, which were fabricated at the Southampton University Microfabrication Centre. For this

project the numerical model has been used almost exclusively to investigate the behaviour and to characterise these devices, and has not been used to model MOS type devices.

Approximately two years elapsed between completion of the mask design and the final fabrication stage, during which time the bulk of the numerical model was being developed. Since these two tasks continued very much in parallel it was extremely difficult to obtain optimised design criteria with the aid of the model prior to device fabrication. However, a number of early results did provide optimum data in sufficient time for it to be included in the processing schedule. This data was concerned with the use of 'resurf layers' to improve the breakdown voltage of the base-collector junction, which has since proven to be extremely effective.

The description of this project is split into a number of major chapters, with each one covering a quite separate and distinct item from the next. The order in which the chapters appear represents the natural progression in the development of the numerical model. In the following chapter a description of the fundamental semiconductor equations will be presented together with the approximations made in their derivation. In chapter 3 the empirical models that have been chosen to represent the physical parameters, which are important in quantifying device operation will be presented. The derivation of the discrete forms of the fundamental semiconductor equations is described in chapter 4. This is the stage that converts the continuous problem into a discrete one. Chapter 5 gives an account of the various techniques that have been employed to solve the discrete problem. Numerical and experimental results are discussed and compared in chapter 6. This chapter is divided into several sub-sections covering the separate studies that have been undertaken as part of the overall project, and finally chapter 7 presents the conclusions drawn from this work.

References.

1.1 J. Bardeen and W. H. Brattain, "The Transistor, a Semiconductor Triode," Phys. Rev., **74**, pp. 230-231 (1948).

1.2 D. Kahng and M. M. Atalla, "Silicon-Silicon Dioxide Field Induced Surface Devices," IRE Solid-State Device Res. Conf., Carnegie Institute of Technology, Pittsburgh, Pa. (1960).

1.3 B. J. Baliga, M. S. Adler, P. V. Gray, R. P. Love and N. Zommer, "The Insulated Gate Rectifier (IGR): A New Power Switching Device," IEDM Tech. Digest, pp. 264-267 (1982).

1.4 P. L. Hower, "Optimum Design of Power Transistor Switches," IEEE Trans. Electron Devices, **ED-20**, pp. 426-435 (1973).

1.5 M. S. Adler, K. W. Owyang, B. J. Baliga, R. A. Kokosa, "The Evolution of Power Device Technology," IEEE Trans. Electron Devices, **ED-31**, pp. 1570-1591 (1984).

1.6 G. Miller, A. Porst and H. Strack, "SIRET, a 1000V Bipolar Transistor with No Two-Dimensional Parasitic Effects," Siemens Res. and Dev. Repts., **17**, pp. 27-34 (1988).

1.7 Y. Nakatani and I. Kuruyu, "An Ultra High Speed - Large Safe Operating Area Switching Power Transistor with New Fine Emitter Structure," INTELEC Tech. Program, pp. 500-507 (1983).

1.8 R. H. Hall, "Power Rectifiers and Transistors," Proc. IRE, **40**, pp. 1512-1518 (1952).

1.9 R. L. Davies and F. E. Gentry, "Control of Electric Field at the Surface of *P-N* Junctions," IEEE Trans. Electron Devices, **ED-11**, pp. 313-323 (1964).

1.10 G. K. Teal, "Some Recent Developments in Silicon and Germanium Materials and Devices," presented at National IRE Conf., Dayton, OH, (1954).

1.11 J. A. Hoerni, "Planar Silicon Transistors and Diodes," presented at IRE Electron Devices Meet., Washington, DC, (1969).

1.12 E. S. Wajda, B. W. Keppenhan and W. H. White, "Epitaxial Growth of silicon," IBM J. Res. Develop., **4**, pp. 288-296 (1960).

1.13 R. J. Basset, "A State of the Art Review of Very High Power Transistors and Their Applications," Power Conversion Int. Proc., pp. 192-201 (1982).

1.14 D. Hartman, K. Owyang and J. Driscoll, "High Current Darlington Transistors," Electronic Eng. pp. 45-55 (August 1982).

1.15 G. Miller, A. Porst and H. Strack, "An Advanced High Voltage Bipolar Power Transistor with Extended RBSOA Using $5 \mu m$ Small Emitter Structures," IEDM Tech. Digest, pp. 142-145 (1985).

1.16 F. Goodenough, "Bipolar Power Transistors Take Their Cue From MOS Technology," Electronic Design, pp. 37-38 (Jan. 26, 1984).

1.17 H. K. Gummel, "A Self-Consistent Iterative Scheme For One-Dimensional Steady State Transistor Calculations," IEEE Trans. Electron Devices, **ED-11**, pp. 455-465 (1964).

1.18 A. DeMari, "An Accurate Numerical Steady-State One-Dimensional Solution of the $P-N$ Junction," Solid State Electron., **11**, pp. 33-58 (1968).

1.19 A. DeMari, "An Accurate One-Dimensional Solution of the $P-N$ Junction Under Arbitrary Transient Conditions," Solid State Electron., **11**, pp. 1021-1053 (1968).

1.20 D. L. Scharfetter and H. K. Gummel, "Large Signal Analysis of a Silicon Read Diode Oscillator," IEEE Trans. Electron Devices, **ED-16**, pp. 64-77 (1969).

1.21 J. W. Slotboom, "Computer-Aided Two-Dimensional Analysis of Bipolar Transistors," IEEE Trans. Electron Devices, **ED-20**, pp. 669-679 (1973).

1.22 H. H. Heimeier, "A Two-Dimensional Numerical Analysis of a Silicon $N-P-N$ Transistor," IEEE Trans. Electron Devices, **ED-20**, pp. 708-714 (1973).

1.23 S. Selberherr, *Analysis and Simulation of Semiconductor Devices*, Springer-Verlag, Wien, 1984.

1.24 O. Manck, H. H. Heimeier and W. L. Engl, "High Injection in a Two-Dimensional Transistor," IEEE Trans. Electron Devices, **ED-21**, pp. 403-409 (1974).

1.25 O. Manck and W. L. Engl, "Two-Dimensional Computer Simulation for Switching a Bipolar Transistor Out of Saturation," IEEE Trans. Electron Devices, **ED-22**, pp. 339-347 (1975).

1.26 S. P. Gaur and D. H. Navon, "Two-Dimensional Carrier Flow in a Transistor Structure Under Nonisothermal Conditions," IEEE Trans. Electron Devices, **ED-23**, pp. 50-57 (1976).

1.27 V. C. Alwin, D. H. Navon and L. J. Turgeon, "Time-Dependent Carrier Flow in a Transistor Structure Under Nonisothermal Conditions," IEEE Trans. Electron Devices, **ED-24**, pp. 1297-1304 (1977).

1.28 L. J. Turgeon and D. H. Navon, "Two-Dimensional Nonisothermal Carrier Flow in a Transistor Structure Under Reactive Circuit Conditions," IEEE Trans. Electron Devices, **ED-25**, pp. 837-843 (1978).

1.29 S. P. Gaur, "Two-Dimensional Analysis of High-Voltage Power Transistors," IBM J. Res. Develop., **21**, pp. 306-314 (1977).

1.30 G. A. Franz, A. F. Franz, S. Selberherr and P. Markowich, "A Quasi Three Dimensional Semiconductor Device Simulation Using Cylindrical Coordinates," Proc. NASECODE III Conf., pp. 122-127 (1983).

1.31 K. Hwang, D. H. Navon and T.-W. Tang and P. L. Hower, Second Breakdown Prediction by Two-Dimensional Numerical Analysis of BJT Turnoff," IEEE Trans. Electron Devices, **ED-33**, pp. 1067-1072 (1986).

1.32 S. A. Higgins, M. K. Johnson, P. A. Gough and J. A. G. Slatter, "Modelling Bipolar Transistor Second Breakdown During Turn-Off by Solution of the Fundamental Device Equations," ESSDERC Tech. Program, pp. 65-69 (1987).

1.33 A. Yoshii, H. Kitazawa, M. Tomizawa, S. Horiguchi and T. Sudo, "A Three-Dimensional Analysis of Semiconductor Devices," IEEE Trans. Electron Devices, **ED-29**, pp. 184-189 (1982).

1.34 T. Toyabe, H. Masuda, Y. Aoki, H. Shukuri and T. Hagiwara, "Three-Dimensional Device Simulator CADDETH with Highly Convergent Matrix Solution Algorithms," IEEE Trans. Electron Devices, **ED-32**, pp. 2038-2043 (1985).

1.35 V. A. K. Temple and M. S. Adler, "Calculation of the Diffusion Curvature Related Avalanche Breakdown in High-Voltage Planar *p-n* Junctions," IEEE Trans. Electron Devices, **ED-22**, pp. 910-916 (1975).

1.36 M. S. Adler, V. A. K. Temple, A. P. Ferro and R. C. Rustay, "Theory and Breakdown Voltage for Planar Devices with a Single Field Limiting Ring" IEEE Trans. Electron Devices, **ED-24**, pp. 107-113 (1977).

1.37 K. R. Whight and D. J. Coe, "Numerical Analysis of Multiple Field Limiting Ring Systems," Solid State Electron., **27**, pp. 1021-1027 (1984).

1.38 A. Rusu and C. Bulucea, "Deep-Depletion Breakdown Voltage of Silicon-Dioxide/Silicon MOS Capacitors," IEEE Trans. Electron Devices, **ED-26**, pp. 201-205 (1979).

1.39 P. Walker, J.T. Davies and K.I. Nuttall, "A Numerical Analysis of the Resurf Diode Structure," IEE Proc., **132**, Pt. I, pp. 285-290 (1985).

1.40 V. A. K. Temple, "Practical Aspects of the Depletion Etch Method in High-Voltage Devices," IEEE Trans. Electron Devices, **ED-27**, pp. 977-982 (1980).

1.41 K. Hwang and D. H. Navon, "Breakdown Voltage Optimization of Silicon p-π-ν Planar Junction Diodes," IEEE Trans. Electron Devices, **ED-31**, pp. 1126-1135 (1984).

1.42 K. P. Brieger, W. Gerlach and J. Pelka, "The Influence of Surface Charge and Bevel Angle on the Blocking Behavior of a High-Voltage p^+-n-n^+ Device," IEEE Trans. Electron Devices, **ED-31**, pp. 733-738 (1984).

1.43 M. S. Adler and V. A. K. Temple, "Maximum Surface and Bulk Electric Fields at Breakdown for Planar and Beveled Devices," IEEE Trans. Electron Devices, **ED-25**, pp. 1266-1270 (1978).

1.44 C. M. Snowden, *Introduction to Semiconductor Device Modelling*, World Scientific, Singapore, 1986.

1.45 M. Kurata, *Numerical Analysis for Semiconductor Devices*, Lexington Books, Lexington, 1982.

1.46 J. J. H. Miller (ed.), Proc. NASECODE I-V Confs., Dublin, Boole Press, 1979-1987.

1.47 K. Board and D. R. J. Owen (eds.), "Simulation of Semiconductor Devices and Processes," Conf. Proc., Vols. 1 and 2, Swansea, Pineridge Press, 1984/86.

1.48 D. C. D'Avanzo, "One Dimensional Semiconductor Device Analysis (SEDAN)," Stanford University, Integrated Circuits Laboratory, Report G-201-5 (1979).

1.49 T. C. Denton, "Validation of BIPOLE," Proc. 2nd International Conf. on Simulation of Semiconductor Devices and Processes, Swansea, Pineridge Press, pp. 169-181, 1986.

1.50 A. F. Franz, G. A. Franz, S. Selberherr, C. Ringhofer, and P. Markowich, "Finite Boxes - A Generalization of the Finite-Difference Method Suitable for Semiconductor Device Simulation," IEEE Trans. Electron Devices, **ED-30**, pp. 1070-1082 (1983).

1.51 J. T. Davies, Ph.D. Thesis, University of Liverpool, 1985.

Chapter 2. Fundamental Basis For Numerical Modelling.

In this chapter a mathematical model, describing the operation of semiconductors in general, will be presented. In subsequent chapters a solution to the equations that constitute this model will be sought using numerical techniques. These equations are often called the basic semiconductor equations and they can be derived from Maxwell's equations together with several relations from solid state physics.

2.1 Poisson's Equation.

Poisson's equation may be derived from Maxwell's third equation, which reads:

$$\operatorname{div} \vec{D} = \rho \tag{2.1}$$

where $\vec{D}$ is the electric flux density vector and ρ is the electric charge density. For semiconductor problems it is usual to represent $\vec{D}$ in terms of the electric field, $\vec{E}$ as follows:

$$\vec{D} = \varepsilon \vec{E} \tag{2.2}$$

where ε is the permittivity. This relation is valid if the permittivity is considered to be time invarient, which is an excellent approximation. It is desirable to introduce the electrostatic potential, ψ in place of the electric field. This may be achieved by setting Maxwell's second equation to zero, which assumes a 'curl' free electric field.

$$\operatorname{curl} \vec{E} = -\frac{\partial \vec{B}}{\partial t} = \vec{0} \tag{2.3}$$

where $\vec{B}$ is the magnetic flux density vector. If the 'curl' of a vector function is zero then it can be represented as the gradient of a scalar function. In this case, therefore, the electric field may be given by:

$$\vec{E} = -\text{grad}\,\psi \tag{2.4}$$

Substituting this equation into (2.2) and then the result into (2.1) gives:

$$\text{div}\,(\varepsilon\,\text{grad}\,\psi) = -\rho \tag{2.5}$$

For the purpose of this work the permittivity, ε has been taken to be a scalar quantity. More precisely though, the permittivity is a tensor of rank two. However, all devices being considered here are assumed to have been fabricated from single crystal silicon, which is not expected to exhibit a significantly anisotropic permittivity. The abrupt change in permittivity at the interfaces between different material layers eg. at an Si-SiO_2 interface, will be treated using Gauss' law of flux continuity as discussed later. In semiconductors the space charge density, ρ is the product of the elementary charge, q and the sum of the various static and mobile charge concentrations.

$$\rho = q\,(N_D - N_A + p - n) \tag{2.6}$$

where N_D and N_A are the magnitudes of the ionized donor and acceptor concentrations, which are positively and negatively charged, respectively. The magnitudes of the mobile carrier concentrations are represented by p and n, which are the positively charged holes and negatively charged electrons, respectively. Equation (2.6) is purely a substitution and no approximations have been required for its derivation, however, a number of assumptions are necessary in calculating the different charge concentrations.

2.2 Current Continuity Equations.

These equations may be derived from Maxwell's first equation, which states:

$$\text{curl}\,\vec{H} = \vec{J} + \frac{\partial \vec{D}}{\partial t} \tag{2.7}$$

where $\vec{H}$ is the magnetic field strength vector and $\vec{J}$ denotes the conduction current density. Applying the 'div' operator to both sides of this equation and remembering that 'div curl' applied to any vector is always zero.

$$\operatorname{div}\operatorname{curl}\vec{H} = \operatorname{div}\vec{J} + \operatorname{div}\frac{\partial \vec{D}}{\partial t} = 0 \qquad (2.8)$$

The last term in this equation can be rewritten by using (2.1) as follows.

$$\operatorname{div}\frac{\partial \vec{D}}{\partial t} = \frac{\partial}{\partial t}(\operatorname{div}\vec{D}) = \frac{\partial \rho}{\partial t} \qquad (2.9)$$

The donor and acceptor concentrations in (2.6) are assumed not to vary with time and, therefore, any impurities that may become electrically active with time, due to temperature fluctuations, for example will be ignored. In addition the current density term, $\vec{J}$ in (2.7) can be split into two separate components, one due to electron flow, $\vec{J}_n$ and another due to the flow of holes, $\vec{J}_p$. On the basis of these considerations equation (2.8) becomes:

$$\operatorname{div}(\vec{J}_n + \vec{J}_p) + q\frac{\partial}{\partial t}(p - n) = 0 \qquad (2.10)$$

This equation states that the time rate of change of charge at a point is dependent or the difference between the inward and the outward current flowing through that point, as would be expected from physical reasoning. Equation (2.10) may be split into two separate continuity equations to allow for a distinct treatment of electrons and holes, by including a quantity, R, which gives:

$$\operatorname{div}\vec{J}_n - q\frac{\partial n}{\partial t} = qR \qquad (2.11)$$

$$\operatorname{div}\vec{J}_p + q\frac{\partial p}{\partial t} = -qR \qquad (2.12)$$

The quantity, R is a function which describes the net recombination and generation of electrons and holes. By its definition a positive value for R implies a net recombination of carriers with a corresponding reduction in their concentrations. Conversely, a negative value for R implies a net generation of electrons and holes. An entirely independent model for R must exist if (2.11) and (2.12) can really be considered as two separate equations. Such a model can be derived from a knowledge of statistical and solid state theories as will be shown in chapter 3.

2.3 Current Transport Equations.

A large number of physical and mathematical concepts must be called upon for the derivation of the current transport equations. A great deal of literature

exists which covers this matter in great depth, so rather than repeat much of this a brief account will be presented here, with a number of references to more detailed studies where appropriate.

In general the electron and hole current densities may be written as the product of the electron or hole concentrations, their velocities, $\vec{v}_n$ or $\vec{v}_p$, and the elementary charge they carry.

$$\vec{J}_n = -q\,n\,\vec{v}_n \tag{2.13}$$

$$\vec{J}_p = q\,p\,\vec{v}_p \tag{2.14}$$

The difficulties arise in relating the average carrier velocities to the electric field, temperature, T and carrier concentrations. This relation may be found, however, by means of a distribution function, f_c, where c is used to represent n for electrons or p for holes. This is a function of phase space, which is the space of spatial co-ordinates $\vec{x} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$, momentum co-ordinates $\vec{k} = k_x\mathbf{i} + k_y\mathbf{j} + k_z\mathbf{k}$ and time, t. Here, **i**, **j** and **k** are conventional unit orthogonal vectors. Phase space consists, therefore, of seven dimensions. The number of particles in a small element of phase space is proportional to $d\vec{k}.d\vec{x}$ and $f_c(\vec{x},\vec{k},t)$ is the constant of proportionality, which may be described as the probable density of particles in phase space. The density of particles in space must include all values of momentum so that [2.1]:

$$c(\vec{x},t) = \int_{V_k} f_c(\vec{x},\vec{k},t).d\vec{k} \tag{2.15}$$

The corresponding current density being given by:

$$\vec{J}_c(\vec{x},t) = \int_{V_k} \frac{q}{m_c^*}.\vec{k}.f_c(\vec{x},\vec{k},t).d\vec{k} = q\,c\,\vec{v}_c \tag{2.16}$$

where m_c^* is the effective mass, assumed to be independent of energy and direction and the integration has been performed over the entire momentum volume V_k. The Liouville theorem states that the time derivative of the distribution function along a particle trajectory, $\vec{x}(t)$, $\vec{k}(t)$ is zero, which results from the need to conserve the number of states [2.2].

$$\frac{d}{dt} f_c(\vec{x}_c(t), \vec{k}_c(t), t) = 0 \tag{2.17}$$

If this derivative is expanded then the implicit form of the Boltzmann transport equation is obtained.

$$\frac{\partial f_c}{\partial t} + \text{grad}_k f_c . \frac{d\vec{k}_c}{dt} + \text{grad}_x f_c . \frac{d\vec{x}_c}{dt} = 0 \tag{2.18}$$

In this equation ' grad_k ' and 'grad_x' are the gradient operators with respect to the $\vec{k}$ and $\vec{x}$ co-ordinates respectively. This equation describes the variation of the distribution function with time due to the movement of particles in spatial and momentum space. The derivative of momentum with respect to time equals the sum of all the forces, $\vec{F}_c$ acting on the particles, and the time derivative of $\vec{x}$ gives the group velocity, $\vec{u}_c$.

$$\frac{d\vec{k}_c}{dt} = \vec{F}_c \tag{2.19}$$

$$\frac{d\vec{x}_c}{dt} = \vec{u}_c \tag{2.20}$$

The force, $\vec{F}_c$ may be divided into forces due to external macroscopic fields, $\vec{F}_{ce}$ and internal localised forces, $\vec{F}_{ci}$, which arise from collisions with impurities and defects. Carriers can also lose energy to thermal lattice vibrations (phonons) [2.3]. The explicit form of the Boltzmann transport equation may now be written.

$$\frac{\partial f_c}{\partial t} + \vec{F}_{ce} . \text{grad}_k f_c + \vec{u}_c . \text{grad}_x f_c = -\vec{F}_{ci} . \text{grad}_k f_c \tag{2.21}$$

The laws of dynamics cannot be used to calculate the effects of the internal forces, $\vec{F}_{ci}$ on the distribution function and statistical laws have, therefore, to be applied. This would make the Boltzmann transport equation amenable to a solution by Monte Carlo methods [2.4], [2.5], whereby the motion of one or more carriers is simulated at a microscopic level. Unfortunately this method requires excessive amounts of computation, making it rather impractical for general application. However, the scattering term on the right hand side of (2.21) may be approximated by supposing that if all external forces are suddenly turned off, giving:

$$\vec{F}_{ce} . \text{grad}_k f_c + \vec{u}_c . \text{grad}_x f_c = 0 \tag{2.22}$$

The distribution function will then return to its equilibrium value due to changes in momentum brought about by collisions, which scatter the carriers. This process can be represented by:

$$\frac{\partial f_c}{\partial t} = -\frac{f_c - f_{co}}{\tau_c} \tag{2.23}$$

where f_{co} is the distribution function in equilibrium and τ_c is time taken to return to equilibrium and is called the relaxation time, which is typically less than a picosecond. The Boltzmann equation may now be written as follows.

$$\frac{\partial f_c}{\partial t} + \vec{F}_{ce}.\text{grad}_k f_c + \vec{u}_c.\text{grad}_x f_c = -\frac{f_c - f_{co}}{\tau_c} \tag{2.24}$$

A number of implicit assumptions have been made in arriving at this equation, the most important of these being:

- All scattering processes are assumed to be elastic and the scattering probability is independent of external forces.
- Carrier-carrier interaction has been neglected. This would affect the scattering term on the right hand side of (2.24).
- All collisions are assumed to be instantaneous so that the duration of a collision is much shorter than the average time between collisions.
- External forces are virtually constant over a distance which is comparable to the physical dimensions of the wave packet associated with carrier motion.
- All effects of degeneracy have been neglected in obtaining the right hand side of (2.24)

The current transport equations may now be obtained by multiplying throughout by the random particle velocity, $\vec{u}_c$ and integrating over the entire momentum space.

$$\frac{\partial}{\partial t}\int_{V_k} \vec{u}_c.f_c.d\vec{k} + \vec{F}_{ce}\int_{V_k} \vec{u}_c.\text{grad}_k f_c.d\vec{k} + \int_{V_k} \vec{u}_c.\left(\vec{u}_c.\text{grad}_x f_c\right).d\vec{k} = -\frac{1}{\tau_c}\int_{V_k} \vec{u}_c.(f_c - f_{co}).d\vec{k} \tag{2.25}$$

Considering each term in sequence, from (2.16),

$$\int_{V_k} \vec{u}_c.f_c.d\vec{k} = c\,\vec{v}_c \tag{2.26}$$

The second integral can be integrated by parts, using the following relation.

$$\int_{V_k} \vec{k}^{i}.\mathrm{grad}_k f_c.d\vec{k} = -i \int_{V_k} \vec{k}^{i-1}.f_c.d\vec{k} \tag{2.27}$$

The integral, therefore, becomes:

$$\int_{V_k} \vec{u}_c.\mathrm{grad}_k f_c.d\vec{k} = -\frac{c}{m_c^{*}} \tag{2.28}$$

The solution to the third integral is not quite as straightforward. Since $\vec{u}_c.\vec{u}_c$ is a scalar and grad_x is independent of $\vec{k}$ space then grad_x can be placed before the integral. The treatment then continues by considering an ideal gas in thermal equilibrium. Since it is known that there is a thermal energy of $\frac{1}{2}kT$ per degree of freedom of motion, then in three dimensions the thermodynamic temperature is obtained by equating the random kinetic energy and the thermal energy.

$$\frac{3kT}{2} = \langle (m_c^{*}.\vec{u}_c - m_c^{*}.\vec{v}_c).(m_c^{*}.\vec{u}_c - m_c^{*}.\vec{v}_c)/2m_c^{*} \rangle \tag{2.29}$$

where $\langle\rangle$ implies an average value for all particles. In semiconductors the average drift velocity, $\vec{v}_c$ is usually much smaller than the random velocity, $\vec{u}_c$ and can be ignored. By approximating T_c to be the carrier temperature in terms of the total energy the third integral can then be defined as follows [2.3].

$$\mathrm{grad}_x \int_{V_k} \vec{u}_c.\vec{u}_c.f_c.d\vec{k} = \frac{1}{m_c^{*}}\mathrm{grad}_x(c\,k\,T_c) \tag{2.30}$$

The integral on the right hand side of (2.25) can be split as follows.

$$\int_{V_k} \vec{u}_c . f_c . d\vec{k} \; - \; \int_{V_k} \vec{u}_c . f_{co} . d\vec{k} \tag{2.31}$$

The equilibrium distribution has zero mean velocity and so the second integral is zero. The first integral is given by (2.26).

It is assumed, here, that the external forces, $\vec{F}_{ce}$ are entirely due to an electric field, $\vec{E}_c$, so that any Lorentz forces due to magnetic induction are neglected, which is also required for the validity of (2.26).

$$\vec{F}_{ne} = -q\,\vec{E}_n , \quad \vec{F}_{pe} = q\,\vec{E}_p \tag{2.32}$$

By substituting for the integrals in (2.25) the following differential equations for the drift velocities of electrons and holes are obtained.

$$\frac{\partial}{\partial t}(n\,\vec{v}_n) + \frac{q}{m_n^*}\,n\,\vec{E}_n + \frac{1}{m_n^*}\,\text{grad}(n\,k\,T_n) = -\frac{n\,\vec{v}_n}{\tau_n} \tag{2.33}$$

$$\frac{\partial}{\partial t}(p\,\vec{v}_p) - \frac{q}{m_p^*}\,p\,\vec{E}_p + \frac{1}{m_p^*}\,\text{grad}(p\,k\,T_p) = -\frac{p\,\vec{v}_p}{\tau_p} \tag{2.34}$$

An approximate solution to these equations can be calculated by firstly defining the effective carrier mobilities, thus:

$$\mu_n = \frac{q\,\tau_n}{m_n^*} \tag{2.35}$$

$$\mu_p = \frac{q\,\tau_p}{m_p^*} \tag{2.36}$$

Equations (2.33) and (2.34) can now be rewritten by using (2.35) and (2.36) and remembering (2.13) and (2.14).

$$\tau_n\,\frac{\partial \vec{J}_n}{\partial t} + \vec{J}_n = q\,\mu_n\,n\,\vec{E}_n + k\,\mu_n\,\text{grad}(n\,T_n) \tag{2.37}$$

$$\tau_p\,\frac{\partial \vec{J}_p}{\partial t} + \vec{J}_p = q\,\mu_p\,p\,\vec{E}_p - k\,\mu_p\,\text{grad}(p\,T_p) \tag{2.38}$$

Since the relaxation times are so small it is usual to ignore the first term in these equations. They are then accurate to the first order in the relaxation times.

From this point it will be assumed that the electron and hole temperatures are equal and that they are in thermal equilibrium with the crystal lattice. Thus, T_n and T_p will both be represented by T. The Einstein relations are valid under these conditions, and they are given by:

$$D_n = \mu_n \frac{k\,T}{q} \tag{2.39}$$

$$D_p = \mu_p \frac{k\,T}{q} \tag{2.40}$$

where D_n and D_p are the diffusion coefficients. The required forms of the current transport equations can now be obtained by expanding the grad operator in equations (2.37) and (2.38).

$$\vec{J}_n = q\,\mu_n\,n\,\vec{E}_n + q\,D_n\,\mathrm{grad}\,n + q\,n\,D_n^T\,\mathrm{grad}\,T \tag{2.41}$$

$$\vec{J}_p = q\,\mu_p\,p\,\vec{E}_p - q\,D_p\,\mathrm{grad}\,p - q\,p\,D_p^T\,\mathrm{grad}\,T \tag{2.42}$$

where D_c^T are the thermal diffusion coefficients, which a.e equal to D_c/T. In addition to the well known drift and diffusion terms, extra terms are included in (2.41) and (2.42) which describe the motion of carriers due to temperature gradients. A number of assumptions have been made in arriving at these equations, the most important of which are as follows.

- The effects of degeneracy have been neglected in obtaining the scattering term due to internal forces.

- All collisions have been assumed to be elastic, and thus, polar optical phonon scattering which is important in GaAs is neglected.

- Spatial variations of collision time and band structure have been ignored, so that a slowly varying impurity concentration over a carrier mean free path is presumed.

- Carrier temperatures have been assumed to be equal to the lattice temperature, which means that the motion of 'hot' carriers [2.6] is incorrectly described. These are particles with energies that are well away from the band edges, and consequently are not in thermal equilibrium with the lattice.

- The Lorentz force due to magnetic induction has been omitted from the term describing the external forces.

- The energy bands are assumed to be parabolic, which also means that degeneracy effects cannot be accounted for.

- The first order current terms in (2.37) and (2.38) have been neglected, which means that time dependent conductivity phenomena such as velocity overshoot cannot be considered.

- The effects of the various boundaries of a device on the distribution function has been ignored and the semiconductor has been assumed to be infinitely large. The distribution function is considerably different at its boundaries eg. $Si\text{-}SiO_2$ interface [2.7].

This derivation has been presented primarily to illustrate the approximations that have to be made in obtaining the current transport equations. Their subsequent use then being restricted to applications where these approximations are valid. As such, this only represents a brief treatment of the Boltzmann equation, however, several texts are available in which much more rigorous procedures are presented eg. [2.8] and [2.9].

A more precise perturbation solution of the Boltmann equation has been performed by Stratton [2.10] in which he showed that for the case where lattice or phonon scattering is the dominant process the thermal diffusion coefficient is exactly half the value obtained here. Since then, however, it has been shown that D_c^T is a sensitive function of impurity scattering [2.11], and the factor of a half can disappear if impurities are present. However, it will be seen that for all devices considered here that the greatest temperature gradients exist in the low doped collector regions where impurity scattering is negligible, and so the halving factor has been included ie. $D_c^T = D_c/(2T)$.

2.4 Carrier Concentrations.

Accurate models for the electron and hole concentrations are of extreme importance if close qualitative and quantitative agreement is to be obtained with practice. Such models are well established and may be obtained by integrating the product of the density of states, $\rho(E)$ and the distribution functions, $f(E)$ across the respective energy bands.

$$n = \int_{E_c}^{E_{top}} \rho_c(E)\, f_n(E)\, dE \qquad (2.43)$$

$$p = \int_{E_{bot}}^{E_v} \rho_v(E)\, f_p(E)\, dE \tag{2.44}$$

where E_c is the lowest energy in the conduction band, E_v is the highest energy in the valence band, and E_{top} and E_{bot} are the energy levels at the top of the conduction band and the bottom of the valence band respectively. The density of available states in the conduction and valence bands are given by [2.12]:

$$\rho_c(E) = \frac{4\,\pi\,(2m_n)^{3/2}}{h^3} \sqrt{E - E_c} \tag{2.45}$$

$$\rho_v(E) = \frac{4\,\pi\,(2m_p)^{3/2}}{h^3} \sqrt{E_v - E} \tag{2.46}$$

where h is Planck's constant, and m_n and m_p are the density of states effective masses of electrons and holes, respectively. The distribution functions can be obtained from statistical analysis and are given by the Fermi-Dirac functions.

$$f_n(E) = \frac{1}{1 + \exp\left(\frac{E - E_{fn}}{k\,T}\right)} \tag{2.47}$$

$$f_p(E) = \frac{1}{1 + \exp\left(\frac{E_{fp} - E}{k\,T}\right)} \tag{2.48}$$

E_{fn} and E_{fp} are the Fermi levels for electrons and holes. Carrying out the integration results in:

$$n = N_c \frac{2}{\sqrt{\pi}} F_{1/2}\left(\frac{E_{fn} - E_c}{k\,T}\right) \tag{2.49}$$

$$p = N_v \frac{2}{\sqrt{\pi}} F_{1/2}\left(\frac{E_v - E_{fp}}{k\,T}\right) \tag{2.50}$$

where N_c and N_v denote the effective density of states in the conduction and valence bands, respectively and are given by:

$$N_c = 2\left(\frac{2\,\pi\,k\,T\,m_n}{h^2}\right)^{3/2} \tag{2.51}$$

$$N_v = 2\left(\frac{2\pi k T m_p}{h^2}\right)^{3/2} \tag{2.52}$$

and $F_{1/2}(x)$ is the Fermi integral of order 1/2, which does not have an analytical solution.

$$F_{1/2}(x) = \int_0^{\infty} \frac{\sqrt{y}}{1+\exp(y-x)}\, dy \tag{2.53}$$

For non-degenerate semiconductors, however, the Fermi integral does have an asymtotic analytical solution.

$$F_{1/2}(x) \simeq \frac{\sqrt{\pi}}{2} \exp x\,, \quad x << -1 \tag{2.54}$$

This is an excellent approximation provided the Fermi levels do not encroach any closer than about $3kT$ from the conduction and valence band edges, and it is consistent with approximations made in the derivation of the current transport equations. The electron and hole concentrations now read:

$$n = N_c \exp\left(\frac{E_{fn} - E_c}{k T}\right) \tag{2.55}$$

$$p = N_v \exp\left(\frac{E_v - E_{fp}}{k T}\right) \tag{2.56}$$

Although non-degenerate statistics have been applied, the combined effects of degeneracy and bandgap narrowing can be incorporated into the model by splitting the band edges into two parts.

$$E_c = E_{co} - \delta E_c \tag{2.57}$$

$$E_v = E_{vo} + \delta E_v \tag{2.58}$$

where E_{co} and E_{vo} denote the band edges for a pure or intrinsic semiconductor, and δE_c and δE_v describe rigid shifts of the band edges due to the influence of dopants.

In order to obtain more suitable forms for n and p it is necessary to introduce some physical considerations concerning intrinsic semiconductors. These are undoped semiconductors in which the electron and hole concentrations are equal and their concentration is called the intrinsic carrier concentration, n_i.

$$n = p = n_i(T) \tag{2.59}$$

The mass action law, which is valid for both intrinsic and extrinsic semiconductors under thermal equilibrium is given by:

$$n\,p = n_i(T)^2 \tag{2.60}$$

An expression for n_i can be obtained setting $E_{fn} = E_{fp} = E_i$, the Fermi level in an intrinsic semiconductor, and then inserting (2.55) and (2.56) into (2.60).

$$n_i(T) = \sqrt{N_c N_v}\ \exp\left(\frac{-E_g}{2\,k\,T} \right) \tag{2.61}$$

where $E_g = E_{co} - E_{vo}$ is the band-gap energy. The intrinsic Fermi level can be obtained by equating (2.55) and (2.56).

$$E_i = \frac{(E_c + E_v)}{2} + \frac{3}{4} k\,T \log_e\left(\frac{m_p}{m_n} \right) \tag{2.62}$$

The small deviation of E_i from the mid-gap point due to the difference in the effective masses is usually negligible and E_i is usually used to specify the mid-gap energy. Thus, for an intrinsic semiconductor (2.55) and (2.56) become:

$$n_i = N_c \exp\left(\frac{E_i - E_{co}}{k\,T} \right) = N_v \exp\left(\frac{E_{vo} - E_i}{k\,T} \right) \tag{2.63}$$

This allows (2.55) and (2.56) to be rewritten in the following form, remembering (2.57) and (2.58).

$$n = n_i \exp\left(\frac{E_{fn} - E_i + \delta E_c}{k\,T} \right) \tag{2.64}$$

$$p = n_i \exp\left(\frac{E_i - E_{fp} + \delta E_v}{k\,T} \right) \tag{2.65}$$

The mid-gap and Fermi energies are now converted into potential form by multiplying by the elementary charge.

$$E_i = -q\,\psi, \quad E_{fn} = -q\,\phi_n, \quad E_{fp} = -q\,\phi_p \tag{2.66}$$

where ψ is the electrostatic potential as considered previously, ϕ_n is the quasi-Fermi potential for electrons and ϕ_p is the quasi-Fermi potential for holes. If in addition the band-gap narrowing terms δE_c and δE_v are converted into units

of electron volts by dividing them by the elementary charge, q then the desired forms for the carrier concentrations are obtained.

$$n = n_i \exp\left(\frac{\psi_n - \phi_n}{V_T} \right) \tag{2.67}$$

$$p = n_i \exp\left(\frac{\phi_p - \psi_p}{V_T} \right) \tag{2.68}$$

where ψ_n, ψ_p and the thermal voltage, V_T are given by:

$$\psi_n = \psi + \delta E_c , \quad \psi_p = \psi - \delta E_v \tag{2.69}$$

$$V_T = \frac{k T}{q} \tag{2.70}$$

The spacial variation of ψ_n gives the slope of the conduction band edge, which represents the driving force for electrons and similarly the slope of the valence band edge is the driving force for holes, which depends on the spatial variation of ψ_p. Thus, in the current transport equations. (2.41) and (2.42) the field terms $\vec{E}_n$ and $\vec{E}_p$ may be given by:

$$\vec{E}_n = - \text{grad}\, \psi_n , \quad \vec{E}_p = - \text{grad}\, \psi_p \tag{2.71}$$

Though not rigorously correct this operational model for band-gap narrowing is extremely attractive as the effects due to shifts of both band edges can be separated. In the past it has been usual to incorporate band-gap narrowing by introducing a so called effective intrinsic concentration [2.13], while assuming equal shifts of both band edges, which is rarely the case. Furthermore, the model presented here is more pertinent with $n_i(T)$ dependent solely on temperature. A model for the doping dependence of band-gap narrowing will be presented in the next chapter. Unfortunately, the proportion of narrowing occurring at each of the band edges is very often not known. However, it has been shown by Adler [2.14] that the effects of this ambiguity on current density is negligible. He compared two extreme cases; one where all the bandgap narrowing was placed in the δE_c term and another where all the narrowing was placed in the δE_v term. He showed, and it has also been observed here, that the electrostatic potential and charge density are altered but the current density is unaffected by such a change. Thus, the external device characteristics do not depend on the difference between δE_c and δE_v but only on their sum. Adler also pointed out that since band-gap narrowing is usually found from transport measurements of the $n.p$ product the combined effects of band-gap narrowing and degeneracy are inherently accounted for.

2.5 Heat Flow Equation.

Since it is intended to investigate device behaviour under operating conditions where considerable power is dissipated it is necessary to account for the associated non-uniform temperature rises and related heat transfer effects. This can be achieved by supplementing the previously obtained electrical model with the heat flow equation. Within the semiconductor heat flows by conduction primarily as a result of phonon interaction. It has been found experimentally that the heat flux, $\vec{f}$ (energy per unit time per unit area) that is conducted in a non-uniform temperature distribution is given by:

$$\vec{f} = -K(T)\,\mathrm{grad}\,T \tag{2.72}$$

where $K(T)$ is the thermal conductivity of the semiconductor material, which itself is dependent upon temperature. In order that energy is conserved the heat flux must obey the following continuity equation.

$$\mathrm{div}\,\vec{f} = Q - \rho\, c\, \frac{\partial T}{\partial t} \tag{2.73}$$

where Q is the thermal generation ($W\,cm^{-3}$) , ρ is the specific mass density ($Kg\,cm^{-3}$) and c is the specific heat capacity ($J\,Kg^{-1}\,K^{-1}$) of the material. The temperature variation of ρ and c can be neglected for practical applications [2.15]. This equation states that the difference between the outward and inward flux flowing through a particular point is equal to the difference between the amount of heat which is generated and that part which is stored causing the temperature to rise at that point. The heat flow equation is obtained by substituting (2.72) into (2.73).

$$\mathrm{div}\,K(T)\,\mathrm{grad}\,T = -Q + \rho\, c\, \frac{\partial T}{\partial t} \tag{2.74}$$

This equation, therefore, accounts for the combined effects of temperature rise and heat transfer via conduction through the semiconducting material.

2.6 Boundary and Interface Conditions.

The previously defined mathematical model can be solved for the set of dependent variables (ψ,n,p,T), which may vary in time and space, only if adequate boundary conditions exist. Thus, the application of the proposed generalised model to a specific problem is achieved by invoking the necessary boundary

conditions. It is often possible to simplify a particular problem without having to make assumptions that would introduce significant error. Many devices, for example, have long finger-like geometries which exhibit homogenous operation along the entire finger length. By modelling the operation over a plane which is orthogonal to the finger and assuming that this is representative of every plane along the finger, only two cartesian co-ordinates are needed.

In general there are two types of boundary conditions, those that represent physical or structural boundaries that are present at the edges of a device and those that are introduced to ease the problem of finding a solution, but do not coincide with actual physical boundaries. These artificial conditions can be imposed on the basis of a certain amount of knowledge of device operation, which is known prior to having obtained a solution. Provided they are correctly applied these boundary conditions should not introduce any error. Many devices consist of a large array of identical structures that are repeated at regular intervals. Furthermore, each of these structures often exhibits symmetrical operating characteristics allowing for further sub-divisions. Artificial boundary conditions enable these building blocks to be isolated from each other and then a solution to only one such block is all that is required to provide a full representation of device operation. Artificial boundaries are also useful for isolating the active region of a device, which is the area of most interest and often takes up only a very small portion of the entire chip area. These isolated regions represent the domains over which the mathematical model formed by the governing equations must be solved.

A second type of condition that must be satisfied is the so-called interface condition. They do not coincide with the boundaries of the domain, but must be applied within the domain itself. These conditions are necessary to account for the abrupt changes in the electric flux, $\vec{D}$ and heat flux, $\vec{f}$ occurring at the interfaces between the various materials that go to make up a device.

The numerical modelling techniques that have been employed are at present restricted to the consideration of two orthogonal spatial dimensions. An extension to three dimensions would in general present too great a computational burden even for the most powerful modern day computers. However, it would be a simple exercise to extend the present model to three dimensions when adequate computational resources do become available.

The various boundary conditions that have been utilized will be described with the aid of a particular example. However, the example that has been chosen is sufficiently general in that it illustrates all the different types of boundary conditions that are commonly used. The example is depicted in Figure 1 and is of bipolar transistor structure which has been fabricated using a conventional double diffused planar process. It can be seen immediately that symmetry has

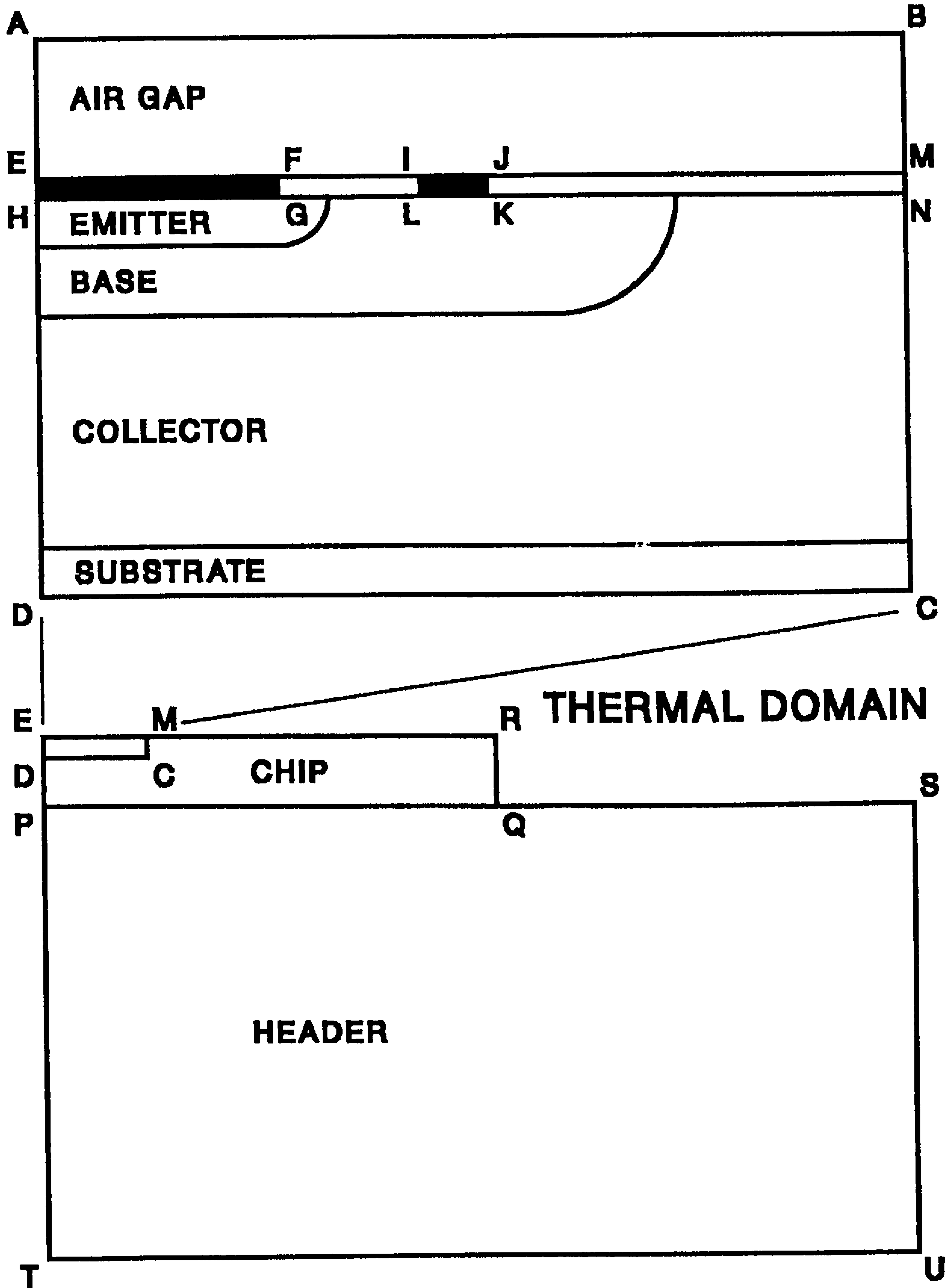


Figure 1. A Bipolar Transistor Structure Illustrating Solution Domains and Boundaries.

been assumed and an artificial boundary (A-T) has been drawn through the centre of the device. Therefore, depending on what co-ordinate system is applied, either a symmetrical long finger-like geometry is assumed (cartesian co-ordinates) or a cylindrically symmetrical geometry with circular emitter and base diffusions (cylindrical co-ordinates) is assumed (cf. chapter 4). The electrical domain represents the region over which Poisson's equation and the current continuity equations (supplemented by the current transport equations) must be solved, and the thermal domain represents the regions over which the heat flow equation is solved. Electro-thermal interaction is inherently accounted for where these two domains coincide. Strictly speaking the electrical domain shown in Figure 1 denotes the area over which only Poisson's equation is solved, with the current continuity equations being relevant only in the semiconductor region (H-N-C-D). The boundary and interface conditions pertaining to the electrical and thermal models will now be considered independently.

2.6.1 Boundary and Interface Conditions for Electrical Model.

In the example three ohmic contacts are present at E-F-G-H, I-J-K-L and D-C. The first two represent real contacts on the top surface of the chip, but the third is an artificial boundary. Since the conductivity of the substrate is usually very high then a negligible voltage will be dropped across it. Thus, for electrical purposes the back contact which in reality exists at the chip-header interface can be shifted to a position just below the collector-substrate junction. This then avoids the need to include most of the substrate, which is usually much thicker than the epi-layer, in the solution domain. However, care must be taken in ensuring that a 'safe' distance exists between the artificial contact and the collector-substrate junction, such that device operation remains unaffected. This depends very much on device operating conditions; for example if the device is operating in saturation the separation should be at least two hole diffusion lengths. Alternatively, under conditions of high collector voltage the separation must at least allow for free depletion of the substrate.

Throughout this work all contacts have been assumed to be voltage controlled, by which it is meant that known potential values are applied at the contacts. It is also possible to specify current controlled contacts, by which a certain current could be made to flow in or out of a contact. The application of this condition unfortunately disrupts the solution procedure [2.1], however, a novel approach which aims to circumvent such problems has been reported recently [2.16]. In all cases considered here though, voltage controlled contacts have proved perfectly adequate. It is well established to assume thermal equilibrium

and space charge neutrality at ohmic contacts. In thermal equilibrium the electron and hole Fermi potentials are equal and the mass action law may be obtained by multiplying (2.67) and (2.68).

$$n\,p \;=\; n_i(T)^2\;\exp\!\left(\frac{\delta E_c+\delta E_v}{V_T}\right) \;=\; n_{ie}(T)^2 \tag{2.75}$$

where $n_{ie}(T)$ is an effective intrinsic concentration. The space charge neutrality condition is simply given from (2.6) by:

$$N \;-\; n \;+\; p \;=\; 0 \tag{2.76}$$

where $N = N_D - N_A$ is the net ionized impurity concentration. Equations (2.75) and (2.76) can now be solved for n and p giving:

$$n \;=\; \frac{\sqrt{N^2 \,+\, 4n_{ie}^2} \;+\; N}{2} \tag{2.77}$$

$$p \;=\; \frac{\sqrt{N^2 \,+\, 4n_{ie}^2} \;-\; N}{2} \tag{2.78}$$

If the applied potential, ψ_A is referenced to the Fermi potential and the 'built-in' contact potential, ψ_b is defined as the difference between the electrostatic (mid-gap) potential, ψ and the Fermi-potential then:

$$\psi \;=\; \psi_A \;+\; \psi_b \tag{2.79}$$

The built-in potential can then be obtained from (2.68) and (2.69). For a contact to n-type semiconductor:

$$\psi_b \;=\; V_T\;\log_e\!\left(\frac{n}{n_i}\right) \;-\; \delta E_c \tag{2.80}$$

and for a contact to p-type semiconductor:

$$\psi_b \;=\; -\,V_T\;\log_e\!\left(\frac{p}{n_i}\right) \;+\; \delta E_v \tag{2.81}$$

At a voltage controlled contact, therefore, the set of dependent variables (ψ, n,p) are known prior to entry into the solution procedure, and furthermore, n and p are dependent only upon the net doping, N, meaning that only ψ needs to be changed with bias. Such a fixed boundary condition, which in this instance holds for Poisson's equation and the current continuity equations is commonly called the Dirichlet boundary condition.

The artificial boundaries A-D and B-C are usually treated by assuming that they are lines of perfect symmetry. This condition is invoked by assuming that the electric flux, and electron and hole current density components normal to these boundaries are zero.

$$\frac{\partial \psi}{\partial x} = 0 \qquad (2.82)$$

$$J_{nx} = 0\,, \quad J_{px} = 0 \qquad (2.83)$$

It may be noted that (2.82) and (2.83) are self-consistent, provided the heat flux and diffusion gradient also vanish at these boundaries, through the current transport equations (2.41) and (2.42). In a strict sense these conditions imply an infinite number of cells whereas only a finite number can exist in practice and some discrepancies are inevitable. However, if the boundary B-C is moved far enough away from the active device region such that any further movement would result in a negligible effect on the solution, then the device can be considered to be a single celled structure. This being formed by the half cell shown in Figure 1 reflected about A-D. The top boundary of the domain for Poisson's equation is separated from the top surface of the chip by an air gap. This has been done to minimise the effects of the imposition of such a boundary on the potential distribution within the chip. The thickness of the air-gap was chosen to be large enough so that the potential distribution along the chip surface was unaffected with the following Dirichlet boundary condition along A-B.

$$\frac{\partial \psi}{\partial y} = 0 \qquad (2.84)$$

That part of the top boundary for the solution of the continuity equations (H-N), which coincides with the Si-SiO_2 interfaces ie. G-L and K-N, is treated by the following condition.

$$J_{ny} = 0\,, \quad J_{py} = 0 \qquad (2.85)$$

This condition states there can be no current flow into the oxide. When solving Poisson's equation the discontinuity of electric field, $\vec{E}$ at the Si-SiO_2 interface and the oxide-air interface (F-I and J-M) can be accounted for by the following two interface conditions, which may be obtained from Gauss' law of flux continuity.

$$\varepsilon_{ox} \frac{\partial \psi}{\partial y} - \varepsilon_{sil} \frac{\partial \psi}{\partial y} = Q_{int} \qquad (2.86)$$

$$\varepsilon_o \frac{\partial \psi}{\partial y} - \varepsilon_{ox} \frac{\partial \psi}{\partial y} = 0 \qquad (2.87)$$

where (2.86) applies to the Si-SiO_2 interface and (2.87) applies to the SiO_2-Air interface. Here Q_{int} is the fixed interfacial oxide charge in $C\,cm^{-2}$ and ε_{sil}, ε_{ox} and ε_o are the permittivities of silicon, SiO_2 and free space respectively.

2.6.2 Boundary and Interface Conditions for Thermal Model.

As was the case for the electrical domain the thermal domain is assumed to be completely symmetrical about its centre line, in which case the following Neumann boundary condition is assumed to hold along E-T.

$$\frac{\partial T}{\partial x} = 0 \qquad (2.88)$$

This is again consistent with boundary condition (2.83), through current transport considerations. The bottom face of the header is assumed to be in intimate thermal contact with a prefect heat sink. Such a heat sink would prevent any rise in temperature at this boundary. It can be seen from equation (2.73) that this is impossible to achieve in practice, since any heat flux entering the heat sink would cause its temperature to rise unless the heat capacity of the heat sink given by the product $\rho\, c$ is infinite. All results presented here, however, are for transient heating effects, and in every case the heat flux generated in the active region of the device did not reach the bottom of the header in the time intervals that were considered. In these cases, therefore, the boundary condition should not affect the results and the following Neumann condition can be applied.

$$T = T_{amb} \qquad (2.89)$$

where T_{amb} is the ambient temperature. Convective heat transfer has been assumed along the remaining portion of the boundary. This can be modelled using Newton's law of cooling [2.17], which states that:

$$\vec{f}.\vec{n} = h\,(T_s - T_{amb}) \qquad (2.90)$$

where $\vec{n}$ is a vector of unit length and is normal to the boundary in question. T_s is the temperature at the surface and h is called the convective heat transfer coefficient or film coefficient. This lumped coefficient depends not only upon the composition of the fluid (air) surrounding the device but also on the nature and geometry of the fluid motion past the surfaces. For unforced convection in air h

takes a value of approximately $1 \times 10^{-3}\, W\, cm^{-2}\, K^{-1}$. Substituting (2.72) into (2.90) gives the following condition for the horizontal boundary portions E-R and Q-S.

$$-K(T)\,\frac{\partial T}{\partial y} = h\,(T_s - T_{amb}) \tag{2.91}$$

and similarly for the vertical portions R-Q and S-U:

$$-K(T)\,\frac{\partial T}{\partial x} = h\,(T_s - T_{amb}) \tag{2.92}$$

The abrupt discontinuity in thermal conductivity that occurs between the chip and header can be treated by assuming continuity of heat flux across the interface, which gives rise to the following interface condition.

$$K(T)\,\frac{\partial T}{\partial y} = K_H\,\frac{\partial T}{\partial y} \tag{2.93}$$

where K_H is the thermal conductivity of the header material.

In this chapter a mathematical model for semiconductor device operation has been proposed, which can be solved subject to a particular set of boundary conditions. However, before a solution can be sought it is necessary to define some empirical models for several physical parameters and attributes pertaining to device operation. For example models must be obtained for mobility, recombination/generation, thermal conductivity and thermal generation. The models that have been employed to simulate such attributes will be presented in the following chapter.

References.

2.1 S. Selberherr, *Analysis and Simulation of Semiconductor Devices*, Springer-Verlag, Wien, 1984.

2.2 P. S. Kireev, *Semiconductor Physics*, MIR Publishers, Moscow, 1978.

2.3 J. E. Carroll, *Rate Equations in Semiconductor Electronics*, Cambridge University Press, Cambridge, 1985.

2.4 C. Moglestue, "A Monte-Carlo Particle Model Study of the Influence of the Doping Profiles on the Characteristics of Field-Effect Transistors," Proc. NASECODE II Conf., pp. 244-249 (1981).

2.5 C. Jacobini and L. Reggiani, "The Monte Carlo Method for the Solution of Charge Transport in Semiconductors With Application to Covalent Materials," Rev. Modern Phys., **55**, pp. 645-705, 1983.

2.6 S. M. Sze, *Physics of Semiconductor Devices*, John Wiley and Sons, New York, 1981.

2.7 A. P. Gnädinger and H. E. Talley, Quantum Mechanical Calculations of the Carrier Distribution and the Thickness of the Inversion Layer of a MOS Field Effect Transistor," Solid State Electron., **13**, pp 1301-1309 (1970).

2.8 R. A. Smith, *Semiconductors*, Cambridge University Press, London, 1979.

2.9 J. L. Moll, *Physics of Semiconductors*, McGraw-Hill Book Co., New York, 1964.

2.10 R. Stratton, "Semiconductor Current-Flow Equations (Diffusion and Degeneracy)," IEEE Trans. Electron Devices, **ED-19**, pp. 1288-1292 (1972).

2.11 J. M. Dorkel, "On Electrical Transport in Non-Isothermal Semiconductors," Solid State Electron., **26**, pp 819-821 (1983).

2.12 E. S. Yang, *Fundamentals of Semiconductor Devices*, McGraw-Hill Book Co., New York, 1978.

2.13 J. W. Slotboom, "The pn-Product in Silicon," Solid State Electron., **20**, pp 279-283 (1977).

2.14 M. S. Adler, "An Operational Method To Model Carrier Degeneracy and Band Gap Narrowing," Solid State Electron., **26**, pp 279-283 (1983).

2.15 C. Kittel, *Introduction to Solid State Physics*, John Wiley and Sons, New York, 1967.

2.16 P. A. Gough, M. K. Johnson, S. A. Higgins, J. A. G. Slatter and K. R. Whight, "Two Dimensional Simulation of Power Devices with Circuit Boundary Conditions," Proc. NASECODE V Conf., pp. 213-218 (1987).

2.17 A. J. Chapman, *Heat Transfer*, The Macmillan Co., New York, 1960.

Chapter 3. Models for Physical Parameters and Device Attributes.

Inherent in the mathematical model derived in the previous chapter are a number of physical parameters, about which very little information has so far been given. In this chapter a model which quantifies each of the parameters will be presented. These models either represent an empirical fit to universal experimental findings or have been obtained on the basis of a physical understanding of the various processes occurring within a device. Usually a combination of both these techniques is employed to obtain a desired result. The overall accuracy of such models is of paramount importance, since not only do they dictate the numerical accuracy of the model against experimentation, but they also define the qualitative nature of device operation. This is especially important if once having obtained good agreement between actual device characteristics and computed results, the computer model is then required to be used to investigate the possible effects of making a change to the device design. In this instance the model is required to account for heating effects and the temperature dependence of the parameters is a primary concern and has been considered, wherever possible and applicable. Since it is intended to use model to investigate phenomena associated with thermal second breakdown the temperature range over which the physical models are to be considered must extend right up to the melting point of silicon (1700K).

In section 3.1 a model will be presented for mobility which depends on a number of scattering processes. Mobility is of obvious importance as the amount of current which flows at any point in a semiconductor is directly proportional to this quantity. In section 3.2 a model for the intrinsic carrier concentration, $n_i(T)$ is proposed which takes account of the temperature dependencies and quantifies the various parameters on the right hand side of equation (2.61). A model for band-gap narrowing due to interactions between carriers and between carriers and ionized impurities at high dopant densities is provided in section 3.3. In section 3.4 a model for recombination/generation resulting from a statistical analysis is presented, and carrier lifetimes, including some experimental measurements are

discussed. A temperature dependent model for thermal conductivity, $K(T)$ is given in section 3.5, and finally an accurate model for thermal generation, Q is presented in section 3.6.

3.1 Carrier Mobility.

During the derivation of the current transport equations the relaxation times, τ_n and τ_p, which describe the average time between scattering events were substituted with a so-called carrier mobility (cf. equations (2.35) and (2.36)). This was done because mobilities are intuitively much easier to imagine than relaxation times, and furthermore they are much more practical quantities for engineering applications. There are a number of scattering mechanisms that affect the relaxation times and, therefore, the carrier mobilities, the most important of which are due to thermal lattice vibrations (phonon scattering), ionized impurities (Coulomb scattering) and the electrons and holes themselves (carrier-carrier scattering). An additional effect which must be accounted for is the saturation of the drift velocity of warm and hot carriers due to lattice vibrations. The effect each of these mechanisms has on the magnitude of the mobility will now be considered in the above sequence.

3.1.1 Phonon Scattering.

Phonon or lattice scattering is the most fundamental process by which carriers in a pure crystal are scattered and is due to their interaction with thermally generated vibrations of the atoms forming the crystal. It is found, because of wave motion, that electrons and holes can actually travel freely through a perfectly periodic crystal without any collisions with the lattice. Thus, there will only be collisions with those atoms that are displaced from their ideal positions in the lattice because of thermal vibrations of the lattice. These vibrations are said to perturb the perfection of the periodic lattice structure and scatter the carriers. Because the atoms in a solid are bound together, the vibration of any one atom is propagated to the other atoms by waves. The particles associated with these waves are called phonons. In silicon there are basically two different scattering mechanisms associated with carrier phonon interaction, these being acoustic phonon scattering and optical phonon scattering. A detailed discussion of these effects is beyond the scope of this work, but further considerations can be found in [3.1] and [3.2].

For the purpose of quantifying these effects it is usual to fit a simple power law to experimentally obtained mobility values. The mobility due solely to phonon scattering is given by:

$$\mu_{n,p}^{L} = \mu_{n,p}^{LO} \left(\frac{T}{300K} \right)^{-\alpha_{n,p}} \quad (3.1)$$

Throughout this chapter the temperature has been normalized by 300K giving a more direct display of room temperature results. Such a fit has been performed by Arora et al. [3.3] on the experimental data of Long [3.4] and Norton et al. [3.5] for electron mobility, and on the data of Li [3.6] for hole mobility. The coefficients which gave the best fit are given in Table 1.

	μ^{LO} ($cm^2V^{-1}s^{-1}$)	α ( )
electrons	1448	2.33
holes	473	2.23

Table 1. Lattice Mobility Constants.

These results for electrons and holes are shown graphically in Figure 2 and Figure 3 respectively. The temperature range has been restricted to 1000K as the model predicts that the carrier mobilities will remain reasonably constant at higher temperatures than this. The model must be subject to some uncertainty at temperatures above 500K since no experimental data is available for validation above this temperature at present. Optical phonon scattering becomes more important at higher temperatures causing carrier transitions between energy valleys in the Brillouin zone [3.1]. These effects may not, therefore be quantified correctly by the empirical model. In fact the experimental results of Li and Thurber [3.7] seem to suggest that the model returns values of electron mobility that are a little low at higher temperatures.

3.1.2 Impurity Scattering.

Impurity or Coulomb scattering is due to the fact that when an electron or a hole travels past a fixed ionized donor or acceptor it will be deflected because of the attractive or repulsive forces set up between the charges. Consequently the role of impurity scattering becomes more important as the total concentration of ionized impurities increases or as the temperature decreases ($< 300K$). The

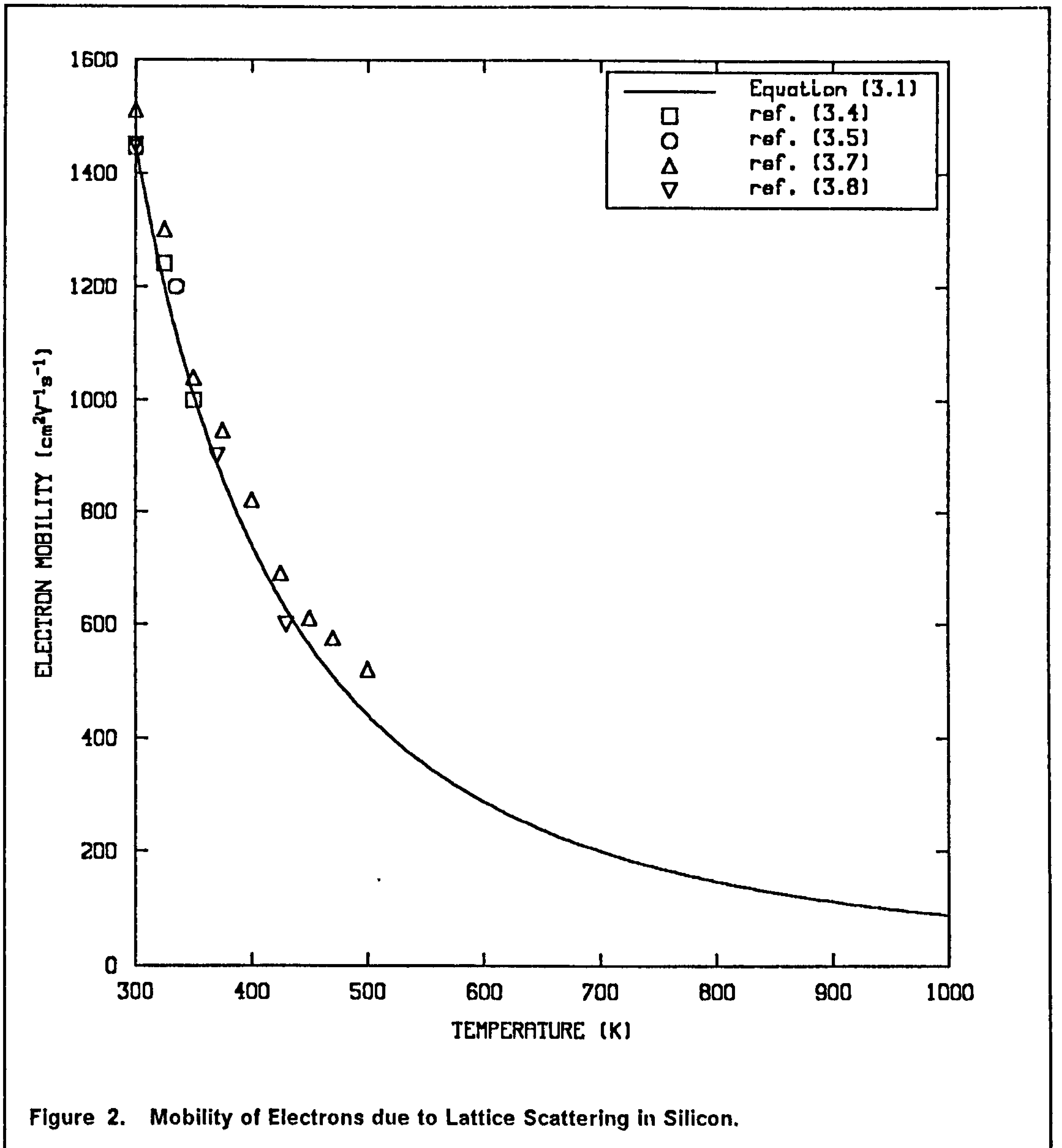


Figure 2. Mobility of Electrons due to Lattice Scattering in Silicon.

ionized impurity scattering mobility, μ^I can be modelled using the modified Brooks-Herring formula [3.6] [3.7], which additionally takes into account anisotropic scattering effects due to the ellipsoidal band structure of silicon.

$$\mu^I_{n,p} = \frac{A}{N_I\,G(b)}\left(\frac{T}{300K}\right)^{1.5} \tag{3.2}$$

where $N_I = N_D + N_A$ is the total number of impurities and $G(b)$ is a function given by:

$$G(b) = \log_e(b+1) - \frac{b}{b+1} \tag{3.3}$$

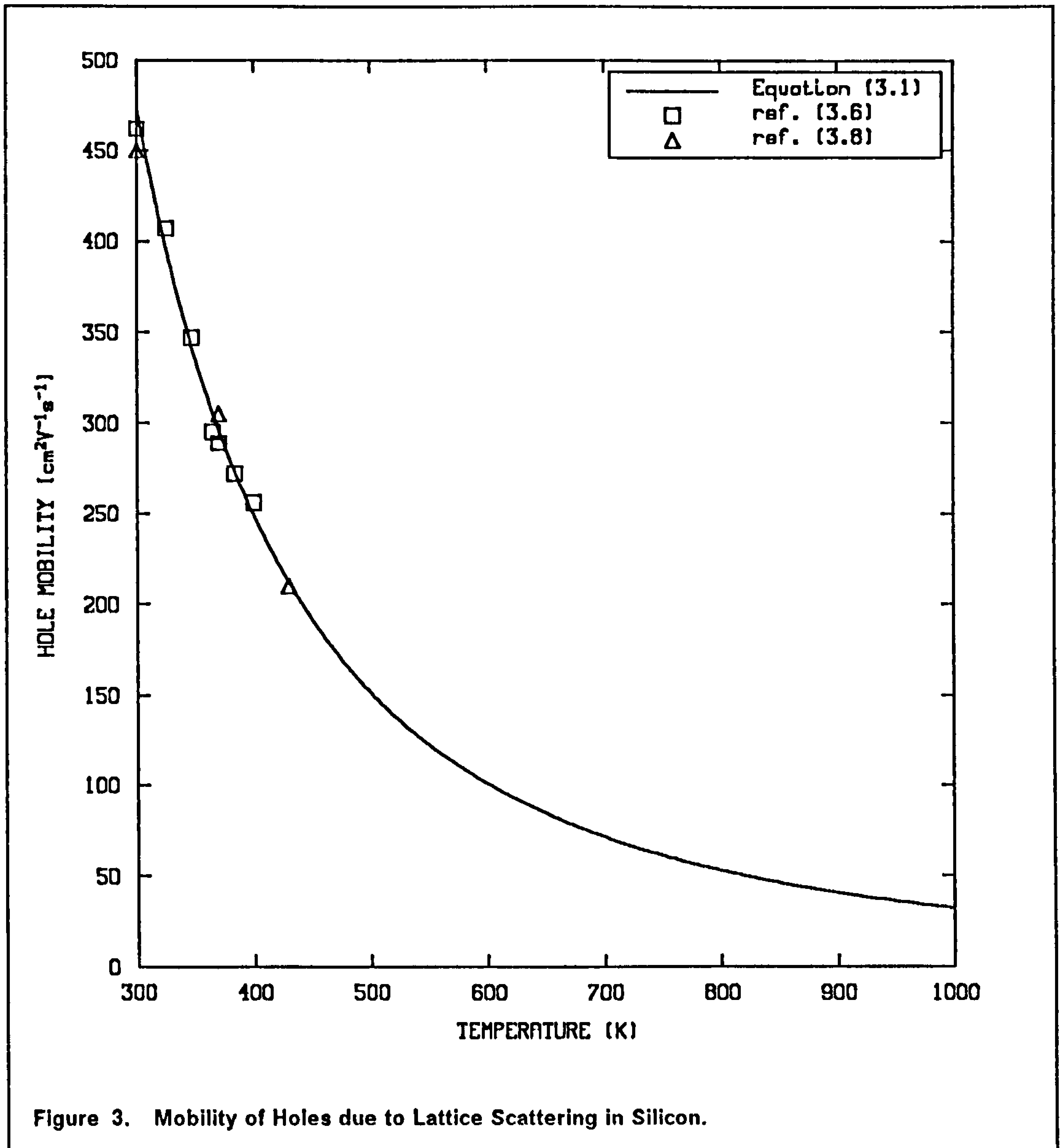


Figure 3. Mobility of Holes due to Lattice Scattering in Silicon.

Assuming all donors and acceptors are ionized, which is an excellent approximation for temperatures above 300K, then b is given by:

$$b = \frac{B}{N_I}\left(\frac{T}{300K}\right)^2 \tag{3.4}$$

The coefficients, A and B for electron and hole mobility are given in Table 2. The function $G(b)$ models the influence of 'neighbouring' ionized impurities which screen each other due to their Coulomb potential and, therefore, are inactive as scattering centres [3.9].

	A $(cm^{-1}V^{-1}s^{-1})$	B (cm^{-3})
electrons	3.793×10^{21}	1.368×10^{20}
holes	2.91×10^{21}	2.25×10^{20}

Table 2. Constants for Mobility due to Impurity Scattering.

So far the formulation has neglected the effect of electron-electron (e-e) and hole-hole (h-h) scattering on the ionized impurity scattering mobility. Although collisions between carriers cannot alter the total momentum, it tends to randomise the way in which this total momentum is distributed amongst carriers with different energies [3.6] [3.7]. Since all donors and acceptors have been assumed to be ionized then the effects of e-e and h-h scattering is to reduce the mobility given by equation (3.2) by a factor of 0.632 for both electron and hole mobility. Having obtained μ^L and μ^I they are then combined according to the mixed scattering formula [3.9].

$$\mu^{LI} = \mu^L\left[1 + x^2\left\{\mathrm{Ci}(x)\ \cos(x) + \sin(x)\left(\mathrm{Si}(x) - \frac{\pi}{2}\right)\right\}\right] \tag{3.5}$$

where $\mathrm{Ci}(x)$ and $\mathrm{Si}(x)$ are the cosine and sine integrals of x respectively and $x = \sqrt{6\mu^L/\mu^I}$. This procedure has been carried out by Arora et al. [3.3] up to a total impurity concentration of $5 \times 10^{18}\ cm^{-3}$ for electrons and $2 \times 10^{18}\ cm^{-3}$ for holes. For electron mobility above $5 \times 10^{18}\ cm^{-3}$ they used the experimental data of Mousty et al. [3.10], Finetti et al. [3.11] and Chapman et al. [3.12] for phosphorus dopant. It is found that at these concentrations the mobility depends on the nature of the dopant and the mobility for phosphorus dopant is on average 10-15% higher than the corresponding value for arsenic dopant [3.13]. Experimental values were also taken for hole mobility above $2 \times 10^{18}\ cm^{-3}$.

These mobility values, both calculated and experimental, were then fitted into the following expression for mobility using an error minimising optimisation technique [3.3].

$$\mu_{n,p}^{LI} = \mu_{n,p}^{MIN} + \frac{\mu_{n,p}^{DIFF}}{1 + \left(N_I/N_{n,p}^{REF}\right)^{\beta_{n,p}}} \tag{3.6}$$

This equation is similar to the well established one used by Caughey and Thomas [3.14]. Here μ^{MIN} is the minimum mobility value expected at the highest dopant densities ($> 10^{20}\,cm^{-3}$), μ^{DIFF} is the difference between the maximum mobility at low dopant densities and the minimum mobility as above. N^{REF} is a reference concentration and β is an exponential factor that controls the slope of the μ^{LI} versus N_I curve around the point where N_I equals N^{REF}. Using their optimisation technique Arora et al. calculated the values of these parameters at temperatures of 200K, 300K, 400K and 500K for electrons, but only at 300K for holes, as an insufficient amount of experimental hole data was available for fitting at high temperatures and concentrations. Thus, the temperature variation of μ^{MIN}, N^{REF} and β for holes was taken to be the same as that for electrons. It was stated that as these three parameters are mainly governed by μ^{I} and since factors affecting the temperature dependence of μ^{I} are the same for electrons and holes (cf. equation (3.2)), then this should be a valid approximation. The resulting form of the equation with temperature dependent parameters reads:

$$\mu_{n,p}^{LI} = \mu_{n,p}^{MINO}\left(\frac{T}{300K}\right)^{-0.57} + \frac{\mu_{n,p}^{DIFFO}\left(\frac{T}{300K}\right)^{-\alpha_{n,p}}}{1+\frac{N_I}{N_{n,p}^{REFO}\left(\frac{T}{300K}\right)^{2.546}}} \quad (3.7)$$

where μ^{MINO}, μ^{DIFFO}, α and N^{REFO} is given in Table 3.

	μ^{MINO} ($cm^2V^{-1}s^{-1}$)	μ^{DIFFO} ($cm^2V^{-1}s^{-1}$)	α ( )	N^{REFO} (cm^{-3})
electrons	88	1252	2.33	1.432×10^{17}
holes	54.3	407	2.23	2.67×10^{17}

Table 3. Constants for Mobility due to Combined Effects of Lattice and Impurity Scattering.

It may be noted that α in equation (3.7) is the same as that in equation (3.1), since μ^{DIFF} is dominated by the lattice mobility. Equation (3.7) is plotted in Figure 4 and Figure 5. Again no experimental data is available above 500K for electrons and 400K for holes. The general agreement between equation (3.7) and experiment is good, though the equation may be returning values of mobility that are slightly too high in the impurity range $10^{16} - 10^{17}cm^{-3}$ for both electrons and holes.

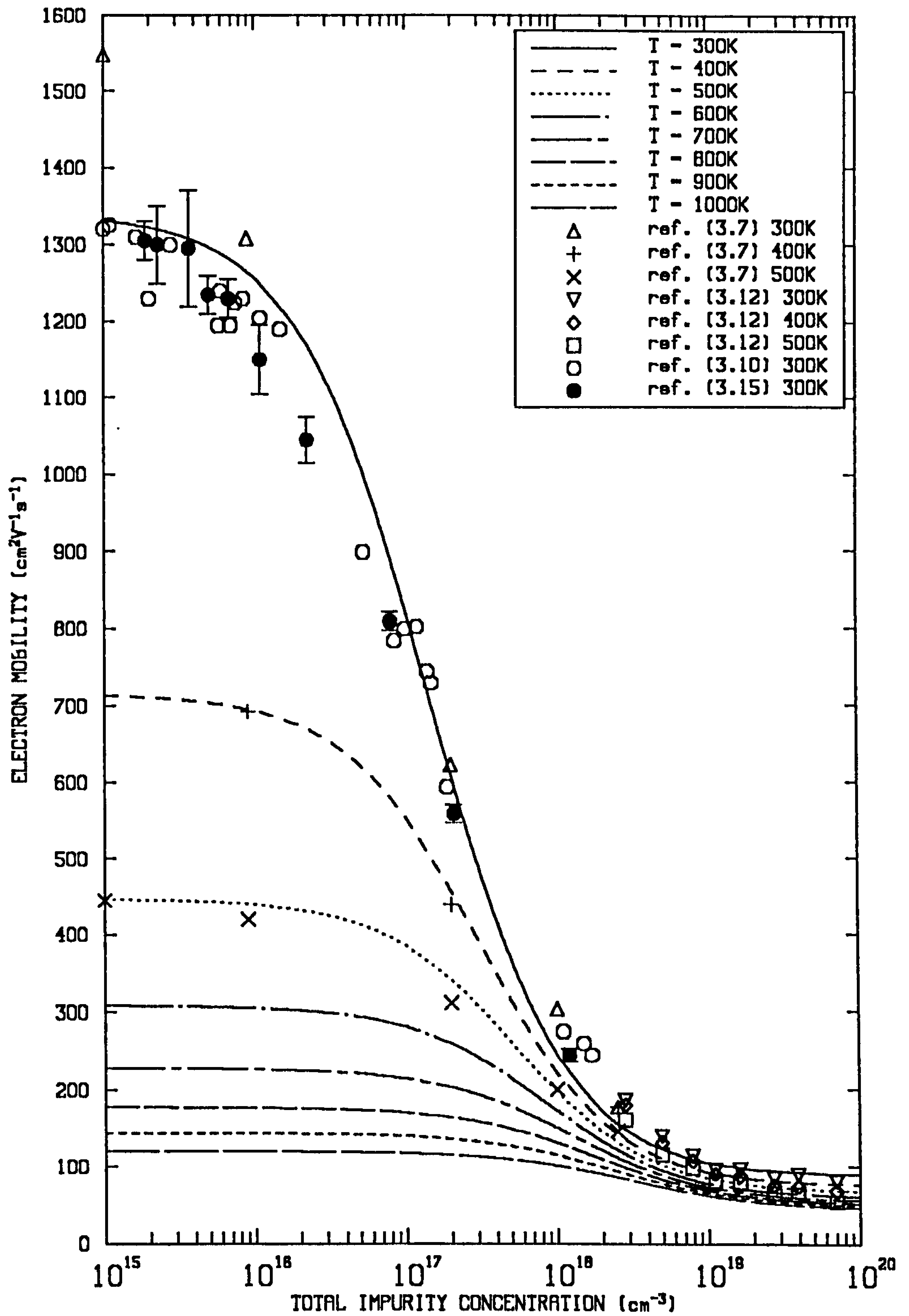


Figure 4. Mobility of Electrons due to Lattice and Impurity Scattering in Silicon: The curves are plotted for equation (3.7) and the points are all experimental results from phosphorus doped samples.

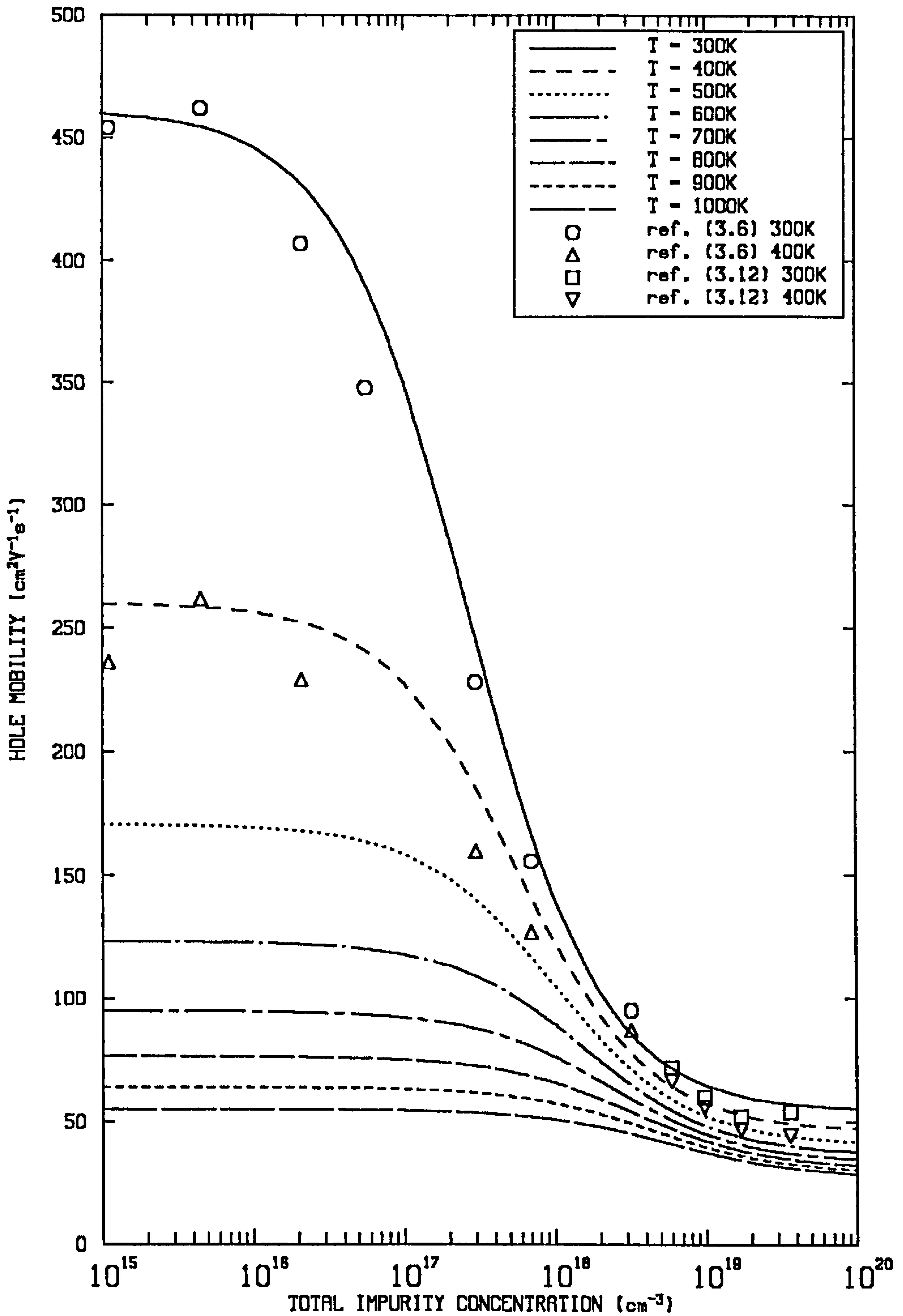


Figure 5. Mobility of Holes due to Lattice and Impurity Scattering in Silicon: The curves are plotted for equation (3.7) and the points are all experimental results from boron doped samples.

The small number of data points at a temperature of 400K in Figure 5 tends to prove the point that the temperature variation of the parameters μ^{MIN}, N^{REF} and β are roughly the same for both electrons and holes. As a final point it must be stated that no difference has be made between the impurity scattering of minority carriers or majority carriers, thus it is assumed for example that an electron in n-type material with a particular impurity concentration will have the same mobility as it would in p-type material with the same total impurity concentration. This could lead significant errors as the scattering characteristics of different dopants need not necessarily be the same [3.16].

3.1.3 Carrier-Carrier Scattering.

Carrier-carrier scattering is a process by which mobile carriers interact with one another. It is similar to ionized impurity scattering, except that mobile particles deflect about a common centre of mass. It causes a further reduction in mobility and unlike the previously considered scattering mechanisms it depends on operating conditions other than temperature. The mobility reduction resulting from carrier-carrier scattering depends upon the amount by which the carrier concentrations exceed their thermal equilibrium values. Thus, it is particularly important in devices where the minority carrier concentration can exceed the background net doping concentration resulting in conductivity modulation. This effect occurs, for example, in the intrinsic region of a P-I-N diode and also in bipolar transistors operating in saturation or under low h_{FE} conditions. However, the effects of carrier-carrier scattering are expected to be negligible in MOSFETs under normal operating conditions as in this case minority carrier concentrations will be very low. However, since the model is intended for use in the simulation of bipolar transistors the effects of carrier-carrier scattering must be accounted for.

The influence of carrier interaction on mobility in silicon were first quantified by Fletcher [3.17] who proposed the use of a formula originally intended to model the interaction between molecules in a non-uniform gas [3.18]. This formula has since been considered by Choo [3.19] who inserted values for the constants in the formula that were left undefined by Fletcher. Having lumped all the constants together the final version of the equation for mobility due to carrier-carrier scattering, μ^C is given as follows.

$$\mu^C_{n,p} = \frac{2.08 \times 10^{21}\, cm^{-1}V^{-1}s^{-1} \left(\frac{T}{300K}\right)^{3/2}}{(\hat{n}+\hat{p})\, \log_e\left\{1 + 1.183 \times 10^{14}\, cm^{-2} \left(\frac{T}{300K}\right)^2 (\hat{n}+\hat{p})^{-2/3}\right\}} \tag{3.8}$$

where $\hat{n}$ and $\hat{p}$ are the excess carrier concentrations. That is, $\hat{n} = n - \bar{n}$, where $\bar{n}$ is the thermal equilibrium value of n calculated from equation (2.77), and $\hat{p} = p - \bar{p}$, where $\bar{p}$ is the thermal equilibrium value of p calculated from (2.78). It may be noted that μ^C has the same value for electrons and holes since it depends only on the sum of the excess carrier concentrations. The mobility, μ^C was combined with the lattice and impurity scattering mobility, μ^{LI} using the simple Mathiessen rule.

$$\mu_{n,p}^{LIC} = \frac{1}{\frac{1}{\mu_{n,p}^{LI}} + \frac{1}{\mu_{n,p}^{C}}} \tag{3.9}$$

This rule is not rigorously correct as the lattice and ionized impurity scattering mechanisms can not be considered to be fully independent from carrier-carrier scattering mechanisms. This is a definite requirement for the applicability of the Mathiessen rule [3.20]. However, it has been observed in the past that more complicated theoretical models do not justify the additional effort for the purpose of simulation [3.21] [3.22].

In nearly all important cases affected by carrier-carrier scattering the excess carrier concentrations, $\hat{n}$ and $\hat{p}$ are equal due to the need to maintain charge neutrality. An exception to this rule in power bipolar transistor operation would occur in the mobile space charge region that is set up in the collector region under conditions of high collector current and voltage. This is commonly known as the Kirk effect [3.23] and in this case charge neutrality does not prevail and the excess carrier concentrations are not equal. However, this effect is due to the velocity saturation of carriers and, therefore, the mobility will be dominated by this phenomenon and the effects of carrier-carrier scattering will be overridden.

Equation (3.9) has been plotted in Figure 6 and Figure 7 for electrons and holes respectively, for a low ionized impurity concentration, N_I of $1 \times 10^{14}\ cm^{-3}$ such that μ^{LI} is dominated by lattice scattering. Bearing in mind the above assumptions the equation has been plotted, at various temperatures, against $\hat{n}$ or $\hat{p}$, which have been assumed to be equal. It is evident from Figure 4 and Figure 6 that carrier-carrier scattering gives rise to a similar reduction in mobility as impurity scattering. This is not too suprising as the relaxation times associated with μ^I and μ^C have similar energy dependent expressions [3.24]. Unfortunately no experimental data is available at present for comparison with these curves. However, a brief comparison has been carried out by Dorkel et al. [3.24] who showed good agreement between the sum of the mobilities, $\mu_n^{LIC} + \mu_p^{LIC}$ calculated from (3.9) at $300K$ and the experimental data of Krausse [3.25] and Danhaüser [3.26], which was taken from measurements on P-I-N diodes. However, a lot more

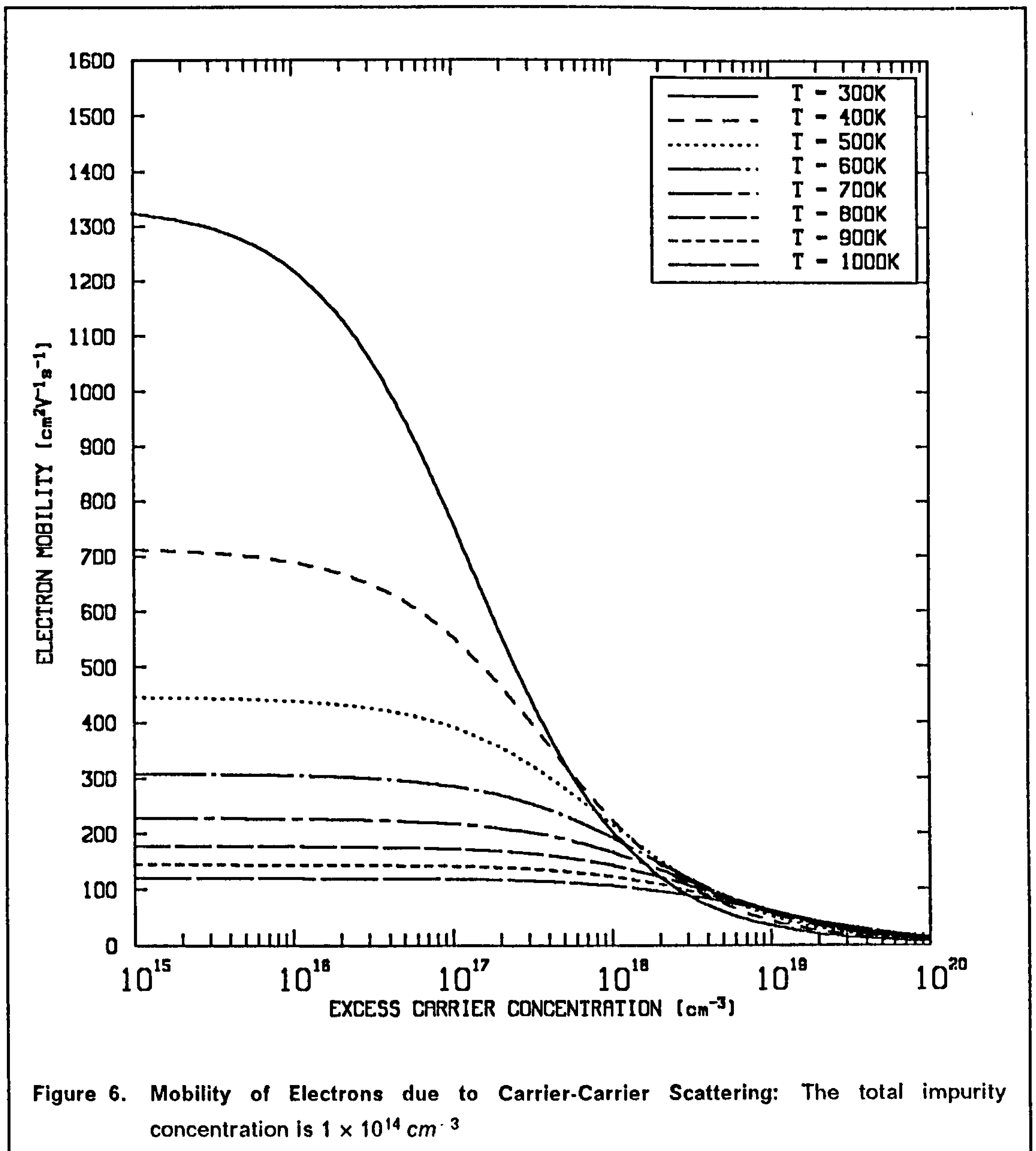


Figure 6. Mobility of Electrons due to Carrier-Carrier Scattering: The total impurity concentration is $1 \times 10^{14}\, cm^{-3}$

work needs to be carried out in order to establish a model that can be used with complete confidence.

3.1.4 Velocity Saturation of Carriers.

At low electric fields the drift velocities of electrons and holes are proportional to the applied electric field and the constant of proportionality is the mobility. However, the carrier velocities can not increase indefinitely with field and they in fact reach a constant value at very high fields ($> 2 \times 10^4 V\, cm^{-1}$). There must, therefore, be a corresponding fall in mobility. The carriers gain energy from

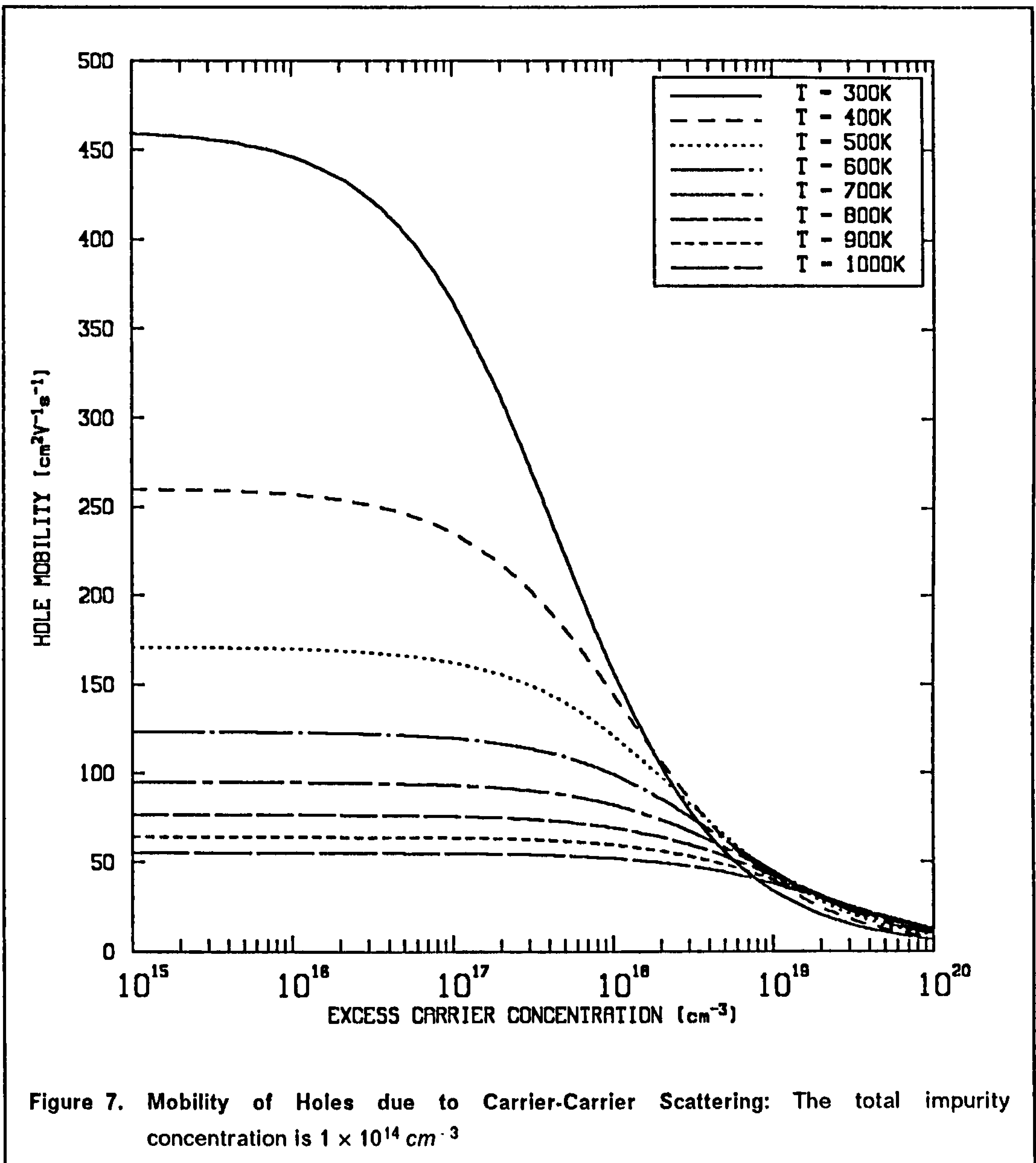


Figure 7. **Mobility of Holes due to Carrier-Carrier Scattering:** The total impurity concentration is $1 \times 10^{14}\ cm^{-3}$

the electric field which increases their temperature over and above that of the crystal lattice, such that they are no longer in thermal equilibrium. The saturation of drift velocity arises because there is a maximum frequency at which the silicon lattice can vibrate which is in the range $10^{13} Hz$ and gives rise to the maximum optical phonon energy. The need to conserve the rate of loss of energy from a particle via optical phonon interaction with the mean rate of supply of energy to a particle from the electric field results in a requirement for a constant drift velocity [3.27].

The magnitude of the drift velocity is usually taken as the product of the mobility and the magnitude of the field component in the direction of current flow, that is:

$$|\vec{v}_n| = \mu_n \frac{\vec{E}.\vec{J}_n}{|\vec{J}_n|} \tag{3.10}$$

$$|\vec{v}_p| = \mu_p \frac{\vec{E}.\vec{J}_p}{|\vec{J}_p|} \tag{3.11}$$

Strictly speaking the driving force for electrons and holes is given by the gradients of their respective quasi-Fermi levels, ϕ_n and ϕ_p. The effects of diffusion gradients on the carrier velocities will then be properly described. Thus (3.10) and (3.11) should be replaced with the following.

$$|\vec{v}_n| = \mu_n \, |\text{grad}\, \phi_n| \tag{3.12}$$

$$|\vec{v}_p| = \mu_p \, |\text{grad}\, \phi_p| \tag{3.13}$$

Here the gradients of the quasi-Fermi levels always point in the direction of current flow. However, in most instances where velocity saturation effects are important the field is equal to the gradient of both quasi-Fermi levels making (3.10) and (3.11) valid. In the region of a forward biased *p-n* junction, though, (3.10) and (3.11) are invalid as in these regions diffusion currents are important. If applied to these regions (3.10) and (3.11) give negative values for the particle speeds! Such problems can be avoided since at forward biased junctions the quasi-Fermi levels are essentially flat having a negligible effect on mobility in these regions. Therefore, the electric field, which is much easier to derive than the gradient of the quasi-Fermi levels can be used in this instance without any loss of accuracy.

The effects of velocity saturation on mobility have been modelled using the formulation of Scharfetter and Gummel [3.28] with the extensions suggested by Thornber [3.29]. The original formulation of Scharfetter and Gummel which accounts for the combined effects of ionized impurity scattering and velocity saturation at 300K is as follows.

$$\mu^{LIE} = \frac{\mu^L}{\sqrt{1 + \frac{N_I}{N^{RE} + N_I/S} + \frac{(E/A)^2}{E/A + F} + \left(\frac{E}{B}\right)^2}} \tag{3.14}$$

where E is the magnitude of the field in the direction of current flow, S is a fitting parameter and N^{RE} is a reference impurity concentration. The remaining parameters A, B and F are defined in Table 4. In this equation the subscript n,p has been omitted from the mobilities and parameters, but it should be noted that (3.14) applies equally to both electron and hole mobility.

	A ($V\,cm^{-1}$)	B ($V\,cm^{-1}$)	F ( )
electrons	3.5×10^3	7.4×10^3	8.8
holes	6.1×10^3	2.5×10^4	1.6

Table 4. Parameters in Equation (3.14) for Effects of Velocity Saturation of Carriers on Mobility.

Since ionized impurity scattering has already been taken into account it is necessary to remove these considerations from equation (3.14). This is achieved by firstly rewriting (3.14) for the low field case:

$$\mu^{LI} = \frac{\mu^L}{\sqrt{1 + \dfrac{N_I}{N^{RE} + N_I/S}}} \qquad (3.15)$$

and then expressing (3.14) in terms of (3.15).

$$\mu^{LIE} = \frac{\mu^{LI}}{\sqrt{1 + (\mu^{LI})^2 \left\{ \dfrac{(E/(\mu^L A))^2}{\mu^L E/(\mu^L A) + F} + \left(\dfrac{E}{\mu^L B} \right)^2 \right\}}} \qquad (3.16)$$

According to Thornber the term $\mu^L B$ in this formula may be interpreted as the carrier saturation velocity v^{sat}. Multiplying μ^L from equation (3.1) at $300K$ with B from Table 4 gives an electron saturation velocity of $1.07 \times 10^7\ cm\ s^{-1}$ and a hole saturation velocity of $1.18 \times 10^7\ cm\ s^{-1}$. These values are quite acceptable, though the value for holes could be considered to be a little high. The term $\mu^L A$ represents the velocity of longitudinal acoustic phonons, v^{ph}, which describes the effects of warm carriers on the carrier drift velocity. According to the detailed scaling considerations of Thornber the term $\mu^L E$ in the denominator of (3.16) should be replaced with $\mu^{LI} E$. Having made these alterations equation (3.16) can now be written in its final form.

$$\mu^{LIE} = \frac{\mu^{LI}}{\sqrt{1 + \left(\frac{\mu^{LI} E}{v^{ph}}\right)^2 \left(\frac{v^{ph}}{\mu^{LI} E + F\, v^{ph}}\right) + \left(\frac{\mu^{LI} E}{v^{sat}}\right)^2}} \tag{3.17}$$

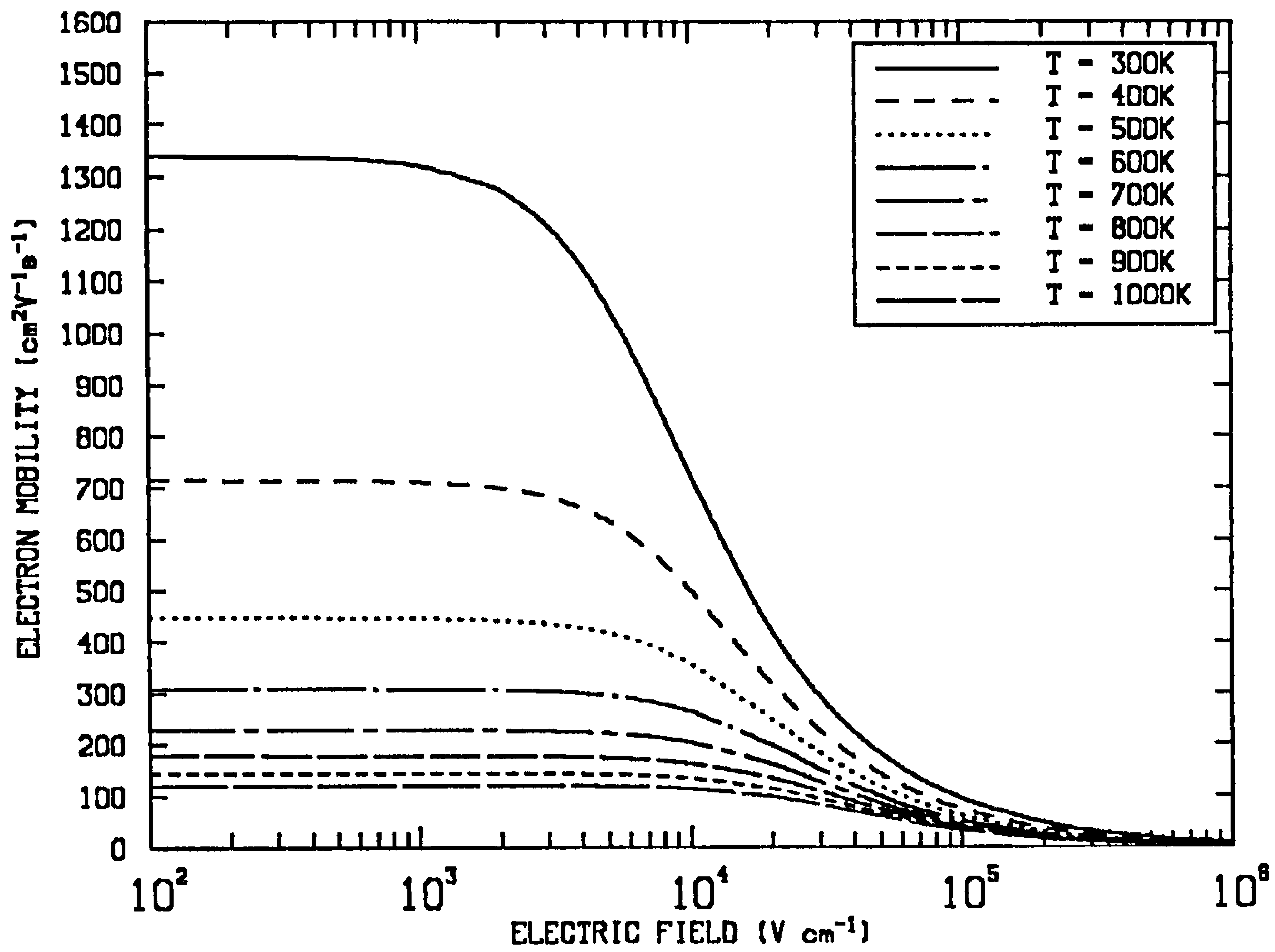
The parameter F does not seem to have any physical significance and can be identified purely as a fitting parameter. At present a temperature dependent model for v^{ph} is not available, however, the saturation velocity is known to fall with temperature and can be modelled with the following expressions [3.22].

$$v_n^{sat} = 1 \times 10^7\, cm\, s^{-1} \left(\frac{T}{300K}\right)^{-0.87} \tag{3.18}$$

$$v_p^{sat} = 8.37 \times 10^6\, cm\, s^{-1} \left(\frac{T}{300K}\right)^{-0.52} \tag{3.19}$$

The effects of carrier-carrier scattering can be incorporated into the above model by simply substituting μ^{LI} in equation (3.17) with μ^{LIC} from equation (3.9). Equation (3.17) has been plotted in Figure 8 for electrons and holes. In this case, however, it is more usual to interpret the results on a drift velocity versus field diagram as is shown in Figure 9. Experimental data has been obtained using the time-of-flight technique [3.30] [3.31], which is based on the well known Haynes-Shockley experiment [3.32]. Good agreement is obtained with experiment for the electron velocity. It can be seen from the two experimental curves at $300K$ that the model is accurate to within the bounds of experimental uncertainty. In the case of hole velocity equation (3.17) predicts rather a low value for the drift velocity, although the experimental curves will be subject to a similar spread as for electrons. Fortunately, for all modelling applications considered here significant hole current is never likely to come under the influence of high fields owing to the way in which the devices have been configured to operate. This would be the case for an *n-p-n* bipolar transistor operating below avalanche breakdown. If *p-n-p* transistor operation were to be considered, however, significant hole current would flow through the collector depletion region and the influence of the field on the hole drift velocity should be carefully considered.

A number of other scattering mechanisms exist, but have not been considered here as they are not expected to influence the operation of the type of devices for which the model was originally conceived. The two most important scattering mechanisms that have been omitted are neutral impurity scattering [3.6] [3.7], which is only important at temperatures below about $77K$, and so-called surface scattering [3.33], which affects the channel mobility in MOS transistors. It



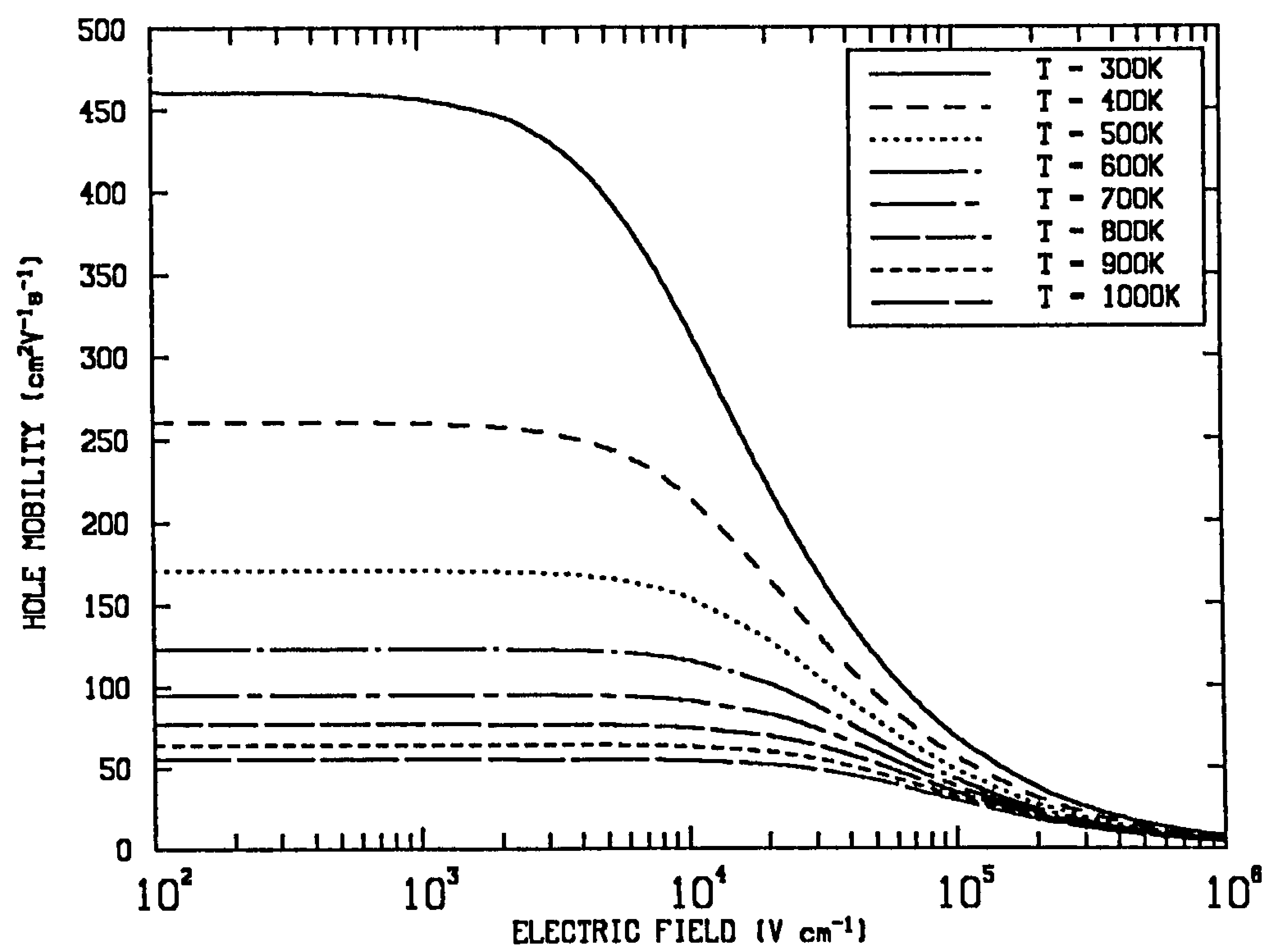
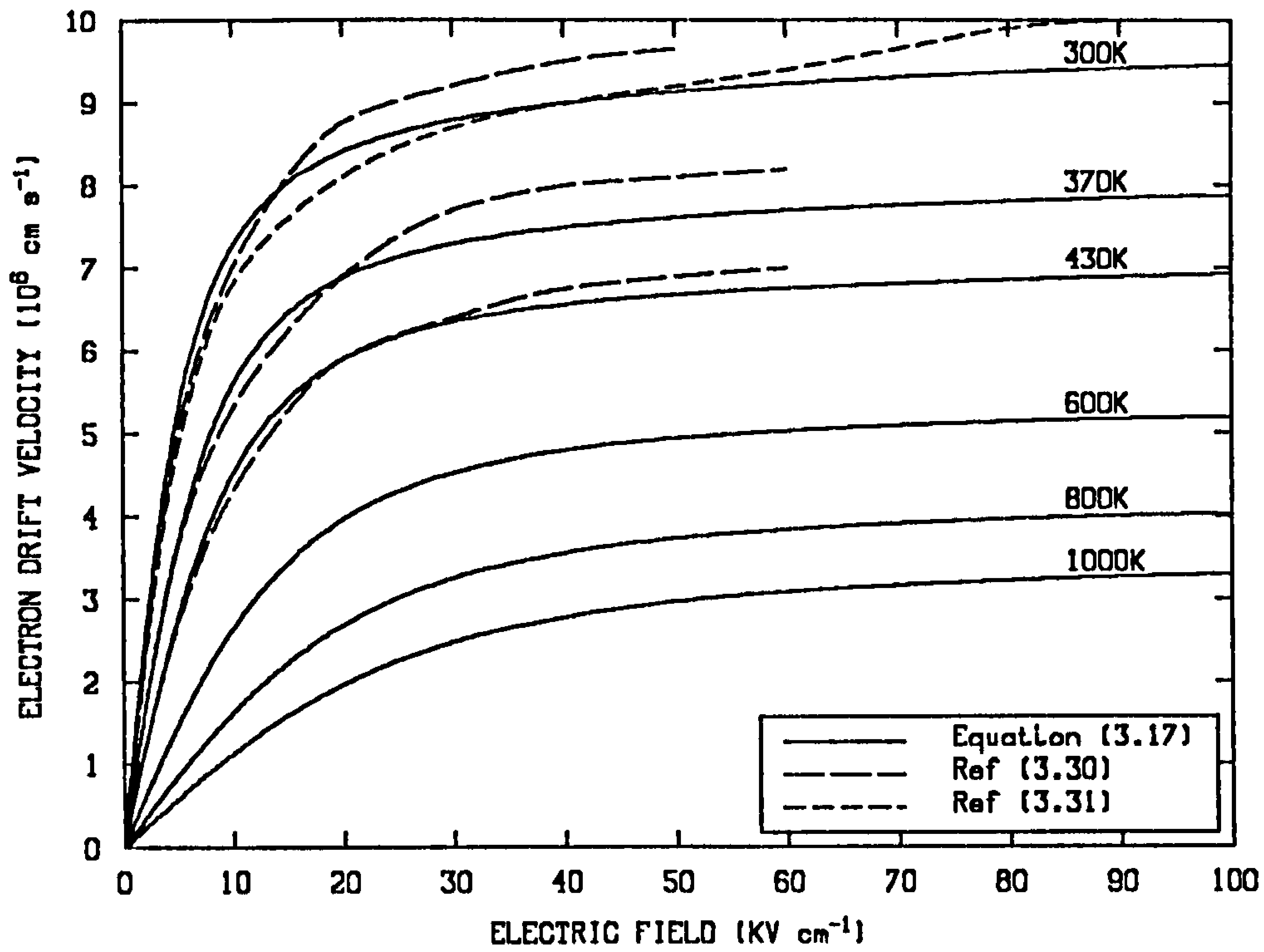


Figure 8. Mobility of Electrons and Holes Versus Electric Field: The effects of impurity and carrier-carrier scattering have been omitted.



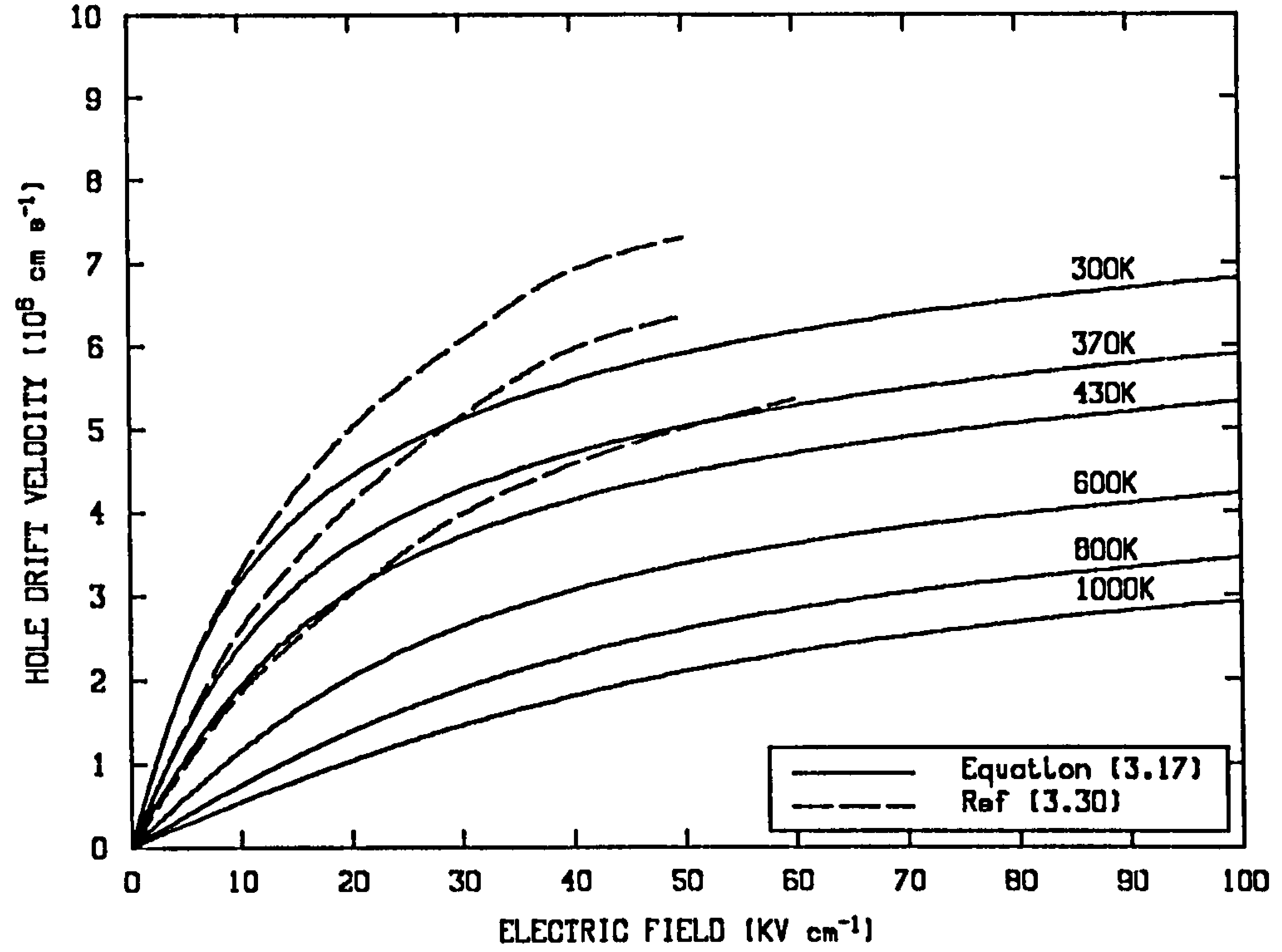


Figure 9. Drift Velocities of Electrons and Holes Showing Velocity Saturation: The effects of impurity and carrier-carrier scattering have been omitted. Dashed lines show experimental data at temperatures of 300K, 370K and 430K.

would be a simple matter to incorporate these considerations into the existing mobility model.

3.2 Intrinsic Carrier Concentration.

The intrinsic carrier concentration, $n_i(T)$, that is the concentration of electrons or holes in a pure semiconductor, can be obtained by inserting equations (2.51) and (2.52) into (2.61). Careful consideration should be taken to ensure that the temperature dependencies of the density-of-states effective masses, $m_n(T)$ and $m_p(T)$, and the band-gap energy, $E_g(T)$ are all accurately accounted for. Such considerations have been made by Barber [3.34], who illustrated the consistency between reported values of m_n, m_p, E_g and n_i through equation (2.61) in the temperature range $200K \rightarrow 700K$. The approach presented by Barber for the calculation of $n_i(T)$ is rather involved for the purposes of device simulation. However, Barber's results were found to agree to within 14% of an empirical model used to fit the earlier measurements of Putley and Mitchell [3.35]. This empirical model was used to fit experimental data obtained in the temperature range $20K \rightarrow 500K$ and is given by the following expression.

$$n_i(T) = 1.61 \times 10^{20}\, cm^{-3} \left(\frac{T}{300K}\right)^{3/2} \exp\left(\frac{-0.603\, eV}{k\, T}\right) \tag{3.20}$$

Although not immediately apparent, the temperature dependencies of m_n, m_p and E_g are all inherently accounted for by this equation, since it has been obtained by direct comparison with measurements of $n_i(T)$. Furthermore, it agrees well with the detailed considerations of Barber. Equation (3.20) is plotted in Figure 10 from room temperature up to the melting point of silicon. The experimental data of Morin and Maita [3.36] and Putley and Mitchell were calculated from conductivity and Hall effect measurements. Those of Putley and Mitchell can be considered most reliable since they were performed on very pure oxygen-free silicon. The data of Herlet [3.37] was taken from measurements on the forward characteristics of *p-n* junctions and are accurate to only 20%. It is estimated that the average deviation of the experimental data of Putley and Mitchell from equation (3.20) is only 5%.

3.3 Band-gap Narrowing.

Nearly all semiconductor devices contain regions where the doping levels are higher than $10^{18}\, cm^{-3}$ and carrier transport through these heavily doped

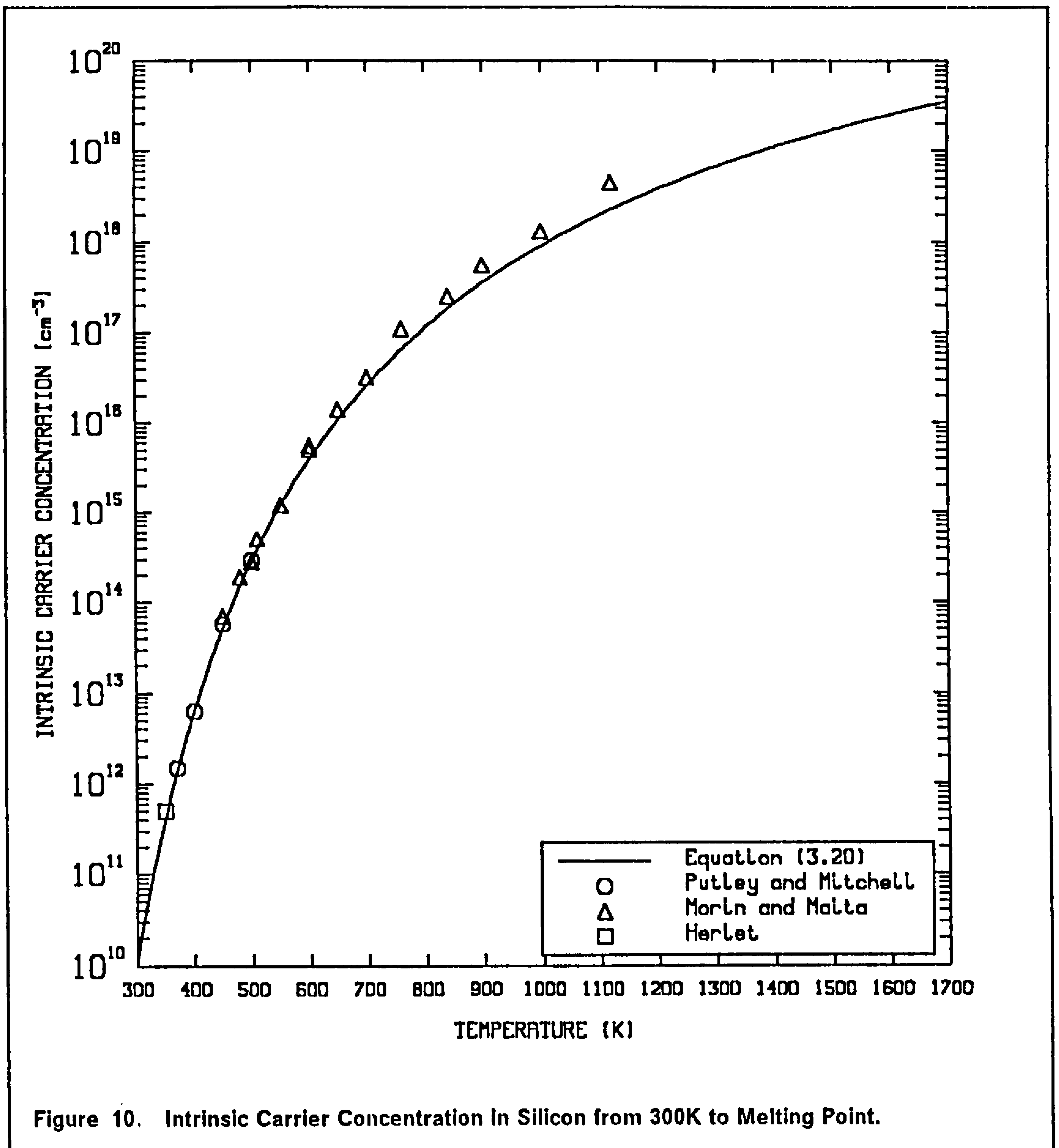


Figure 10. Intrinsic Carrier Concentration in Silicon from 300K to Melting Point.

regions plays an important role in the overall operation of the device. The present understanding of heavily doped semiconductors is fairly limited, but it is generally accepted that the density of states functions for electrons, $\rho_c(E)$ and holes, $\rho_v(E)$ are influenced by essentially two categories of phenomena. The first category are the many-body effects, which consist of interactions between carriers and between carriers and ionized impurities. These effects result in rigid shifts of both the conduction and valence band edges towards each other, but do not alter the parabolic energy distribution of the density of states functions given by (2.45) and (2.46). The second category results from the random distribution of impurities, which causes electrostatic potential fluctuations. The overlapping of the electron wave functions at the impurity states which gives rise to band-tailing and impurity

bands also contributes to this category. This latter category of phenomena result in a distortion of the density of states function such that they can no longer be considered to be parabolic. A detailed survey of these phenomena is beyond the scope of this work, however, a comprehensive review of these effects is given in [3.38].

For practical purposes it is usually acceptable to assume that the parabolic shape of the density-of-states functions does not become distorted, while taking the effects of distortion into account by the use of an effective band-gap narrowing term that assumes only rigid shifts of the band edges. This term can be obtained directly from transport measurements and is a more relevant quantity for general device applications. The best established model that accounts for the quantitative effect of impurity concentration on band-gap narrowing is that of Slotboom and De Graaff [3.39]. Their formula gave the best empirical fit to data obtained from electrical testing of several *n-p-n* transistors with different highly doped base regions, and is given by:

$$\Delta E_g = \delta E_C + \delta E_V = E_1 \left(\log_e\left(\frac{N_I}{N_0}\right) + \sqrt{\left\{ \log_e\left(\frac{N_I}{N_0}\right)\right\}^2 + C} \right) \quad (3.21)$$

Slotboom and De Graaff found that the values, $N_0 = 10^{17}\,cm^{-3}$, $E_1 = 9\,meV$ and $C = 0.5$ gave the best fit to their data, which ranged from $4 \times 10^{15}\,cm^{-3}$ to $2.5 \times 10^{19}\,cm^{-3}$. Since it was first presented this equation has been verified theoretically [3.40] and values of N_0, E_1 and C were derived which were similar to the original values of Slotboom and De Graaff. More recently a model has been reported by Lanyon and Tuft [3.41] that accounts for the effect of temperature on the band-tails and thereby provides a band-gap narrowing model for heavy doping effects which is temperature dependent, as follows.

$$\Delta E_g = \frac{3\,q^3}{16\,\pi\,\varepsilon_{sil}^{3/2}\,k^{1/2}\,T^{1/2}}\sqrt{N_I} \quad (3.22)$$

Inserting values for the physical constants and rearranging gives:

$$\Delta E_g = 22.58\,meV\left(\frac{T}{300K}\right)^{-0.5}\sqrt{\frac{N_I}{10^{18}\,cm^{-3}}} \quad (3.23)$$

Unfortunately this equation is only accurate for non-degenerate material ($N_I < 3 \times 10^{19}\,cm^{-3}$). For the strongly degenerate case ($N_I > 1 \times 10^{20}\,cm^{-3}$) the following equation, which does not contain any temperature dependency has to be applied.

$$\Delta E_g = 162\,meV\left(\frac{N_I}{10^{20}\,cm^{-3}}\right)^{1/6} \tag{3.24}$$

Equations (3.21), (3.23) and (3.24) are plotted as a function of N_I in Figure 11 together with a range of experimental data for comparison. It can be seen that both models provide a good representation of the currently available data, which is subject to a considerable spread. The model of Lanyon and Tuft gives a reduction in high doping effects at higher temperatures owing to an increase in the screening radius [3.41]. An abrupt discontinuity in the gradient of the curves calculated from equation (3.23) is apparent at the point where they meet the asymptote given by equation (3.24). This is not strictly correct and does not

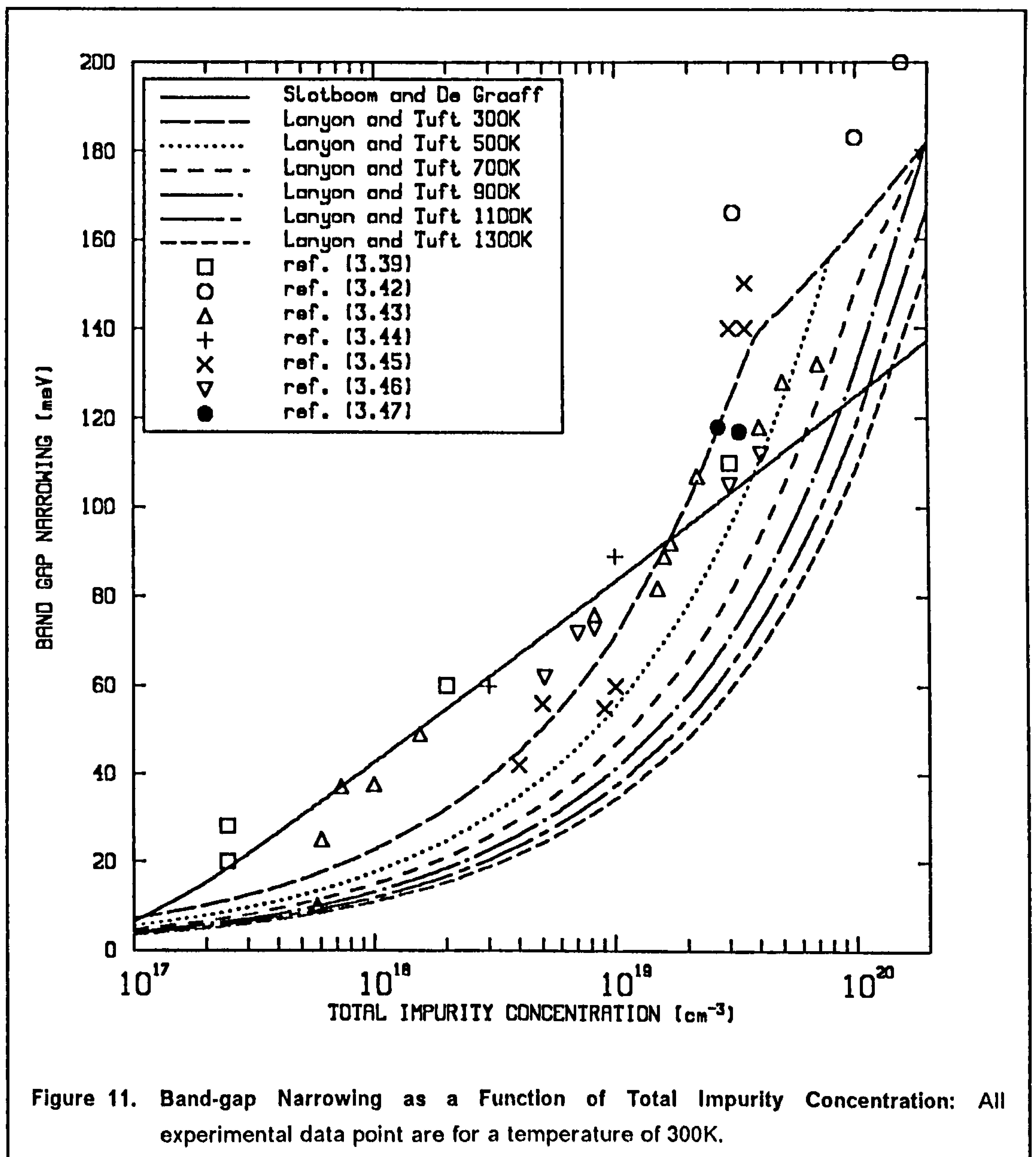


Figure 11. Band-gap Narrowing as a Function of Total Impurity Concentration: All experimental data point are for a temperature of 300K.

exactly follow the model of Lanyon and Tuft who also considered band-gap narrowing in the intermediate doping region between the non-degenerate and strongly degenerate cases. Taking these considerations into account produces a smoother transition between the two extremes given by (3.23) and (3.24). In this intermediate region, however, the calculation of ΔE_g is much more involved, requiring as it does a solution to the Fermi half-integral (2.53). Such a solution cannot be justified for the purposes of general simulation, bearing in mind the overall uncertainty in the experimental data that the models are supposed to represent.

Finally, it must be pointed out that the model of Slotboom and De Graaff was used only in earlier versions of the numerical model and all results presented here were obtained using the temperature dependent model of Lanyon and Tuft.

3.4 Carrier Recombination/Generation.

It was stated in section 2.2 that the splitting of the current continuity equation (2.10) is only possible if an independent model for the recombination/generation term, R is available. It is intended to present such a model in this section, but only a brief account will be given here. For an excellent review of the current understanding of recombination phenomena reference [3.48] is recommended.

Recombination is the rate by which excess electrons and holes in a semiconductor resulting from an external stimulus (eg. ionizing radiation or junction biasing) return to their thermal equilibrium values when the stimulus is removed. Conversely, generation is the rate by which the excess concentrations increase under the influence of the external stimulus. At any point in space and time the carrier concentrations may be written as the sum of their thermal equilibrium and excess values.

$$n = \bar{n} + \hat{n} \tag{3.25}$$

$$p = \bar{p} + \hat{p} \tag{3.26}$$

where $\bar{n}$ and $\bar{p}$ are the thermal equilibrium values and $\hat{n}$ and $\hat{p}$ are the excess values. For the case of an external generating stimulus, G_E the continuity equation for holes (2.12) can be rewritten.

$$\frac{\partial \hat{p}}{\partial t} = -R + G_E - \frac{1}{q}\operatorname{div}\vec{J}_p \tag{3.27}$$

The rate at which the holes try to return to their equilibrium concentration is proportional to the amount by which they exceed their equilibrium concentration, that is their excess concentration. Though it is far from linear the net recombination term, R is often replaced by $\hat{p}/\tau_p$, where τ_p is called the hole lifetime and is not to be confused with the hole dielectric relaxation time. In general τ_p is a function of $\hat{p}$ and equation (3.27) is non-linear. If the semiconductor is uniformly illuminated then $\operatorname{div} \vec{J}_p = 0$ and (3.27) becomes:

$$\frac{\partial \hat{p}}{\partial t} = -\frac{\hat{p}}{\tau_p} + G_E \tag{3.28}$$

The lifetime τ_p can be interpreted in two different ways from this equation. If the external generation is maintained at a constant rate and steady state is reached then τ_p is the ratio of the excess concentration to the generation rate, G_E. This is the generation lifetime. Alternatively if the concentration $\hat{p}$ has reached steady state and the excitation is removed abruptly then:

$$\tau_p = -\frac{\hat{p}}{\partial \hat{p}/\partial t} = \frac{\hat{p}}{R} \tag{3.29}$$

where R is the recombination rate of excess holes. This lifetime is the recombinative lifetime. The two most important mechanisms by which electron-hole pairs can recombine in silicon are by so called Shockley-Read-Hall recombination and Auger recombination. These two mechanisms will now be considered separately.

3.4.1 Shockley-Read-Hall Recombination/Generation.

A model for this mechanism established around the same time by Shockley and Read [3.49] and Hall [3.50]. It is, therefore, most widely known as Shockley-Read-Hall (SRH) recombination/generation (from now on called recombination) In this case recombination occurs via energy states which are situated near the centre of the band-gap, and as such are called deep level recombination centres or traps. These traps are created by impurity atoms other than donors and acceptors which only give rise to energy levels near the band edges. These are called shallow energy levels and are not efficient as recombination centres. The most important impurities that give rise to deep levels in silicon are gold, platinum, copper, iron and oxygen. Each type of impurity leads to more than one deep level, but a single level situated nearest the centre of the

band-gap dominates the carrier recombination in most cases. Deep level traps are also known to occur at crystal defects such as dislocations.

The potential energy of an electron-hole pair is lowered in two stages. the first part of the energy is released when an electron makes a transition from a state in the conduction band to a deep level centre and the remaining energy is released when an electron falls into an empty state in the valence band. The energy is dissipated in the form of phonons. Since silicon is an indirect band-gap semiconductor, that is the electrons near the conduction band edge do not have the same momentum as the holes at the valence band edge, then during the transition momentum as well as energy must be conserved. This is made possible by exchanges of momentum with the deep energy states which have a wide range of momenta. Shockley, Read and Hall derived a model for carrier recombination through a single level recombination centre using non-degenerate statistics, which showed that the net rate of recombination of excess carriers can be written as:

$$R^{SRH} = \frac{np - n_{ie}^2}{\tau_{no}(p + n_{ie}) + \tau_{po}(n + n_{ie})} \tag{3.30}$$

This interpretation of the SRH model assumes that the trap lies exactly at the centre of the band-gap. The parameter τ_{no} is the electron lifetime in highly doped p-type material, τ_{po} is the hole lifetime in highly doped n-type material, and n_{ie} is the effective intrinsic concentration given by equation (2.75). A number of interesting limiting consequences arise from equation (3.30). For instance, in a depletion region $n \simeq p \simeq 0$ and (3.30) can be written.

$$R^{SRH} = -\frac{n_{ie}^2}{\tau_{no} + \tau_{po}} = -\frac{n_{ie}^2}{\tau_{sc}} \tag{3.31}$$

where τ_{sc} is the space charge generation lifetime. Here the recombination term is negative which means that carriers are generated in a depletion region. This accounts in part for the leakage current in reverse biased *p-n* junctions, which increases rapidly with temperature since n_{ie} approximately doubles for every 11°C temperature rise. This phenomenon is extremely important in high power operation.

A second interesting case arises when the excess carrier concentrations greatly exceed their thermal equilibrium values, as is the case under high level injection ($\hat{p} >> \bar{p}$, $\hat{n} >> \bar{n}$). In such situations space charge neutrality also usually applies ($\hat{n} = \hat{p}$) and equation (3.30) reduces to:

$$R^{SRH} = \frac{\hat{n}}{\tau_{no} + \tau_{po}} = \frac{\hat{n}}{\tau_a} \tag{3.32}$$

where τ_a is the ambipolar lifetime. Again the carrier lifetimes are equal to the sum of the electron and hole minority carrier lifetimes. The lifetimes are longer under high level conditions since the electrons and holes are in equal number and half the traps are occupied. The models that have been employed for room temperature lifetimes together with some experimental measurements are presented in section 3.4.3. The second most important recombination mechanism in silicon is Auger recombination and this will now be considered.

3.4.2 Auger Recombination.

This is the mechanism by which carriers recombine whilst interacting with other carriers so that both energy and momentum are conserved when making a transition. There are basically two types of Auger recombination and these are direct band-to-band Auger recombination (BBA) and trap assisted Auger recombination (TAA). The effects of TAA recombination are expected to be insignificant as they are dominated by normal phononic SRH type recombination [3.48] [3.51].

A number of processes are possible via BBA recombination. An electron in the conduction band can, for example, move to the valence band and in doing so it transfers part of its energy and momentum to a second electron in the conduction band, causing this electron to move away from the conduction band edge. Alternatively, the electron could also have transmitted its energy and momentum to a hole in the valence band, which would subsequently move away from the valence band edge. Both these mechanisms are recombinative, however, electron and hole emission can also occur by the consummation of energy from highly energetic electrons in the conduction band or holes in the valence band. The statistics of BBA recombination are well established and the recombination rate has been found to be given by:

$$R^{AUG} = (c_n n + c_p p)(n p - n_{ie}^2) \tag{3.33}$$

where c_n and c_p are the Auger capture coefficients for electrons and holes and are given by $2.8 \times 10^{-31}\,cm^6s^{-1}$ and $1 \times 10^{-31}\,cm^6s^{-1}$ respectively [3.46]. These coefficients have been found to be very weak functions of temperature [3.42] and any temperature dependency has, therefore, been omitted in this work. Auger recombination causes a reduction in minority carrier lifetimes only in regions where the electron concentration or hole concentration or both concentrations exceed about $10^{18}\,cm^{-3}$. Thus, it is restricted to heavily doped regions or regions

that are highly conductivity modulated. For minority holes injected into a heavily doped n-type region the Auger lifetime obtained from (3.29) is as follows.

$$\tau_{low}^{AUG} = \frac{1}{c_n n^2} = \frac{1}{c_n N_D^2} \tag{3.34}$$

In heavily conductivity modulated regions where electrons and holes are in equal quantity and the Auger lifetime is given by:

$$\tau_{high}^{AUG} = \frac{1}{(c_n + c_p) n^2} = \frac{1}{(c_n + c_p) p^2} \tag{3.35}$$

The slightly lower lifetime for this case is reflected by the fact that interactions involving two holes and one electron is probable in this instance as well as interactions between two electrons and one hole. The SRH lifetimes are less than the Auger lifetimes in most regions of a device and they, therefore define the overall lifetime given by: $1/\tau^{TOT} = 1/\tau^{SRH} + 1/\tau^{AUG}$. The SRH lifetimes will now be considered.

3.4.3 Carrier Lifetimes.

The SRH lifetimes τ_{no} and τ_{po} are dependent on a number of quantities and they may be given by the following expressions.

$$\tau_{no} = \frac{1}{\sigma_n v_{th} N_t} \tag{3.36}$$

$$\tau_{po} = \frac{1}{\sigma_p v_{th} N_t} \tag{3.37}$$

where $v_{th} = \sqrt{3kT/m^*}$ is the carrier thermal velocity, $\sigma_{n,p}$ are the electron and hole capture cross-sections and N_t is the density of deep level traps. The temperature dependencies of τ_{no} and τ_{po} are controlled by the temperature dependencies of v_{th} and $\sigma_{n,p}$. Although v_{th} varies as $\sqrt{T}$ the actual lifetimes increase with temperature because at high temperatures the carriers are more energetic and must approach the trap more closely to be captured by it. The temperature dependence of the capture cross-sections account for this. It has been found that $\sigma_{n,p} \propto T^{-b}$ and measurements on the transient behaviour of p-i-n diodes have shown that b has a value of 2.7 for n-type silicon and 3.4 for p-type silicon [3.53]. It is predicted, therefore, that τ_{no} varies as $T^{2.9}$ and τ_{po} as $T^{2.2}$. Unfortunately the temperature dependencies of the carrier lifetimes were not known at the time the numerical model was being developed and room

temperature lifetime were assumed throughout. However, these temperature dependencies should be included in the numerical model for future simulations.

The SRH lifetimes are also dependent upon doping density as well as temperature. A high concentration of dopant atoms introduces defects into the semiconductor crystal causing an increase in the density of recombination centres, N_t. It has also been suggested that the concentration of deep levels could be increased due to the increased solubility of deep level impurities like iron and gold in heavily doped silicon [3.48]. Dislocations can act as deep level SRH recombination centres because the strain field associated with them introduces energy levels in the mid-gap region and also because deep level impurities migrate to these sites due also to the strain they generate in the lattice. An empirical model which describes the variation of the minority carrier SRH lifetimes with dopant density has been proposed by Fossum [3.54].

$$\tau_{no,po} = \frac{\tau_{n,p}^{LOW}}{1 + \dfrac{N_I}{N_{n,p}^{REF}}} \tag{3.38}$$

Values of $\tau_{n,p}^{LOW}$ and $N_{n,p}^{REF}$ that gives the best fit to available measurements of lifetime in substrate material [3.48] are given in Table 5.

	τ^{LOW} (sec)	N^{REF} (cm^{-3})
electrons	3×10^{-4}	7.1×10^{15}
holes	3.95×10^{-4}	7.1×10^{15}

Table 5. Lifetime Constants.

The lifetimes in diffused layers are expected to be much less than those given on the basis of the parameters in Table 5, for substrate material. This is due to the additional crystal damage resulting from ion implantation and subsequent diffusion or oxidation of dopant atoms during processing. Fossum [3.54] estimated a τ_p^{LOW} value for his n^+ diffused layers that was about 500 times less than the value in Table 5. Such a reduction in lifetime is obviously highly dependent on processing. This means that to be able to accurately model recombination in a particular device then the lifetimes should, strictly speaking, be measured from the device that is to be simulated, provided, of course, such a device is available. This presents a rather severe limitation as simulation results are often required in

advance of fabrication. A great deal more data is required on lifetimes in diffused layers and on the variation of lifetimes with processing before these limitations can be removed.

Much of the simulation work reported here has been addressed to several bipolar devices that have been designed and fabricated as part of this project. The design of these devices is outlined in chapter 6. They have been characterised for lifetimes with the aid of the numerical model and also by experimentation. The collector layer of these devices is low doped making them amenable to collector lifetime measurement via open circuit voltage decay (OCVD) techniques [3.55]. For this method the base collector (*p-v-n*) junction is forward biased so that the v region becomes heavily conductivity modulated. The forward bias is then removed and the junction is immediately subject to open circuit conditions. The decay of forward voltage across the junction is measured on a cathode ray oscilloscope (CRO) and the collector lifetime can be calculated from the resulting rate of voltage decay. The circuit that was used to perform this function together with the resulting waveform is shown in Figure 12.

A small signal MOS transistor was used to provide a switch with a high impedance off-state. The *p-v-n* junction is represented by the diode and a unity gain buffer constructed from a 741 operational amplifier was used to isolate the CRO. The applied bias can be altered by changing the supply voltage or series resistance though a single bias point should be sufficient. When the *p-v-n* junction is forward biased the hole concentration in the v region at the edge of the forward biased depletion region associated with the *p-v* junction is:

$$p = \bar{p}\ \exp\left(\frac{q V_1}{k T}\right) \tag{3.39}$$

where V_1 is that part of the forward voltage which falls across this junction. Similarly, the electron concentration in the v region at the edge of the depletion region near the *v-n* junction is:

$$n = \bar{n}\ \exp\left(\frac{q V_2}{k T}\right) \tag{3.40}$$

where V_2 is the voltage across the *v-n* junction. Upon multiplying (3.39) and (3.40) the following is obtained.

$$n\,p = \bar{n}\,\bar{p}\ \exp\left(\frac{q\,(V_1 + V_2)}{k T}\right) = n_i^2\ \exp\left(\frac{q V}{k T}\right) \tag{3.41}$$

where $V = V_1 + V_2$. In high level operation $n \simeq p \simeq \hat{n} \simeq \hat{p} >> N_v$ and:

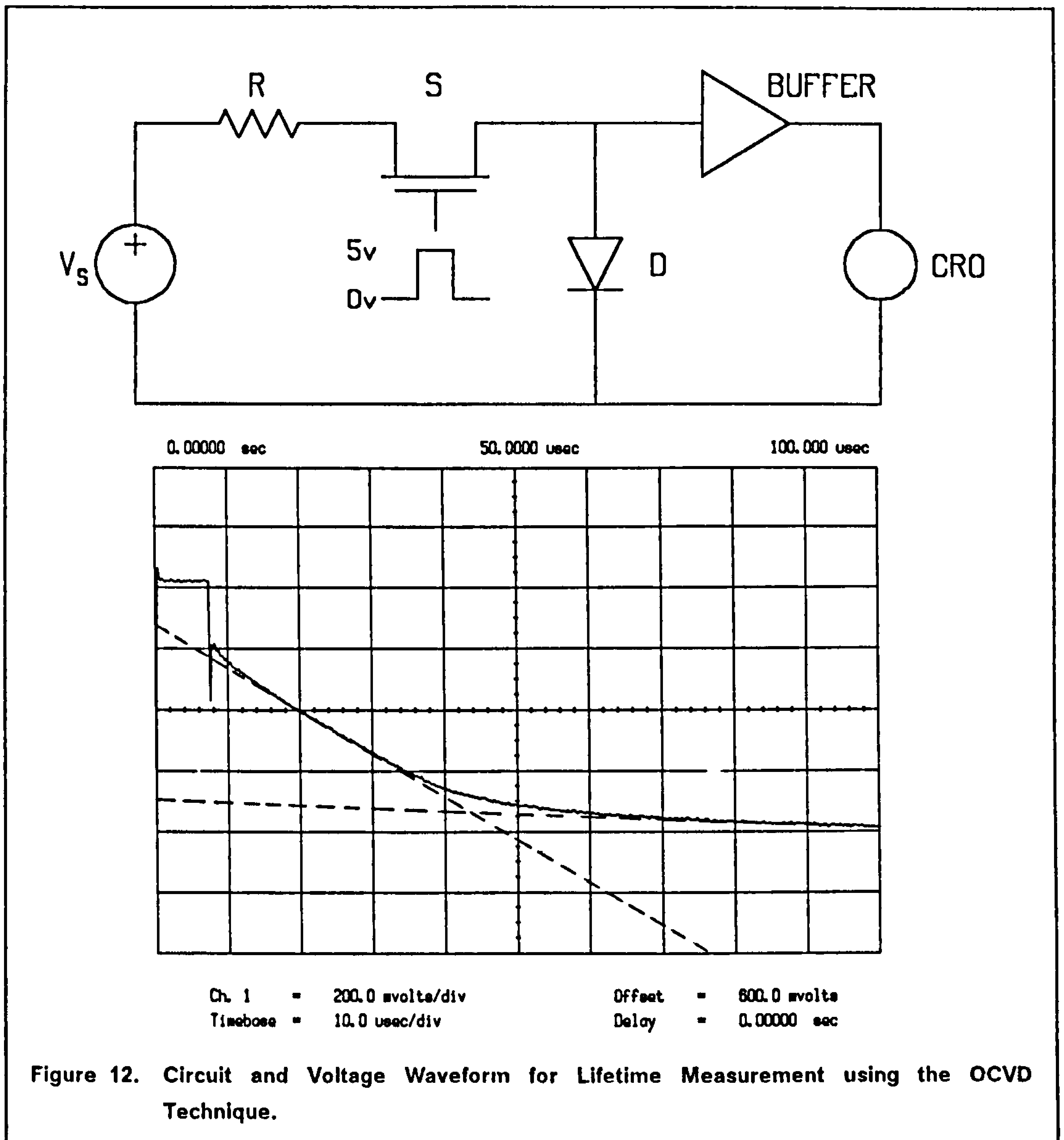


Figure 12. Circuit and Voltage Waveform for Lifetime Measurement using the OCVD Technique.

$$\hat{p} = n_i \exp\left(\frac{q\,V}{2\,k\,T} \right) \tag{3.42}$$

Taking the derivative of this equation with respect to time gives:

$$\frac{\partial \hat{p}}{\partial t} = \frac{q}{2\,k\,T}\,\hat{p}\,\frac{\partial V}{\partial t} \tag{3.43}$$

Rearranging this equation and remembering expressions (3.29) and (3.32) gives the required result.

$$\tau_a = -\frac{\hat{p}}{\partial \hat{p}/\partial t} = -\frac{2\,k\,T}{q}\,\frac{1}{\partial V/\partial t} \tag{3.44}$$

The voltage waveform in Figure 12 exhibits a sudden drop in voltage prior to two distinct linear voltage decay ranges. The sudden fall in potential is an ohmic potential drop which occurs immediately the forward bias is removed. This is followed by a rapid linear voltage decay giving way to a relatively slow voltage decay. Detailed numerical measurements have have shown that the lifetime in the ν region can be calculated using the above technique from the slope in the less rapidly varying potential region [3.56] [3.57]. The initial steep slope is due to recombination in the end regions. As previously explained the lifetime in these regions is considerably shorter, and this results in a rapid reduction of the stored charge at the edges of the ν region. Since the charge in the centre of the ν region is relatively unaffected, large diffusion gradients are set up which result in diffusion of carriers into the end regions. Eventually the carrier profiles become flat and diffusion into the end regions can be ignored. Carrier decay is then governed by recombination in the ν region which has a much longer time constant giving a slower forward voltage decay. For the example shown in Figure 12 this condition does not arise until the forward voltage has dropped to 0.25v, which corresponds to low level conditions, where the concentration of holes in the ν region is much less than the background donor concentration, N_ν. In this situation the entire forward voltage is dropped across the p-ν junction and taking the derivative of equation (3.39) with respect to time gives a value for the low level lifetime that is exactly half the value given by (3.44). From the slope of the waveform in Figure 12 the collector lifetime is evaluated to be $30\mu s$. This is an acceptable value, although it is an order of magnitude lower than that given by (3.38) using the values in Table 5. However, as previously stated these values are only valid for bulk material. The base lifetime has been calculated from the slope of the initial rapid voltage decay of the waveform in Figure 12, assuming low level injection in the highly doped end regions. This gives a value of $2\mu s$, which is expected to be rather high since diffusion into the end regions will tend to sustain the voltage dropped across the junctions. In modern day bipolar transistors operating in low level conditions the base recombination component of the base current is usually negligible as base transit times are much shorter than the base lifetime. For this case, therefore, the choice of base lifetime is not critical. In saturation or high level operation, however, the base recombination component can become significant, although in most cases recombination in the collector should dominate owing to its larger volume compared with the base.

The emitter minority carrier lifetime has been obtained by directly comparing experimentally and numerically obtained current gain (h_{FE}) values at low injection levels. The transistors were found to have a h_{FE} of 33 at a collector-emitter bias of 10 v and a collector current of 8 mA. A value for the hole

lifetime in the emitter of 165 ns was required to repeat these values with the numerical model. This lifetime value has been found to give acceptable current gains for a range of bias points.

It has been pointed out by Adler et al. [3.58] that SRH recombination and band-gap narrowing are the dominant mechanisms limiting the emitter injection efficiency, which for an n-type emitter is defined as the ratio of the electron current density to the total current density flowing across the emitter-base junction. The injection efficiency governs the h_{FE} of the transistor since the base transport factor, which is defined as the ratio of the electron current density leaving the base to that entering the base, is in general very close to unity. Band-gap narrowing was found to dominate the injection efficiency for shallow diffused emitters of about 1 μm and SRH recombination dominated for emitters deeper than about 4 μm, but the effects of Auger recombination were negligible in all cases. Full agreement has been obtained with these conclusions in all simulations carried out for this project. The Auger lifetime given by (3.34) is much larger than the SRH lifetime within a diffusion length from the forward biased emitter-base depletion region on the emitter side. This is because the ionized impurity density within this length is too low to give a significant reduction in τ^{AUG}, since the emitter impurity profile is diffused and not abrupt. The SRH lifetime within the diffusion length, therefore, governs the hole injection into the emitter together with a drift term due to band-gap narrowing.

These values of lifetime have been assumed to be constant within their specified regions and any variations that may occur due to non-uniform impurity distribution have been neglected. This section completes the characterisation of the electrical parameters and the parameters that affect the thermal characteristics of a silicon device will now be considered.

3.5 Thermal Conductivity.

The thermal conductivity of low doped silicon and germanium has been measured by Glassbrenner and Slack [3.59] from 3K to melting point. From theoretical reasoning they found that their data could be fitted with the following formula.

$$K(T) = \frac{1}{a + b\,T + c\,T^2} \tag{3.45}$$

where $a = 0.03\ W^{-1}\ cm\ K$, $b = 1.56 \times 10^{-3}\ W^{-1}\ cm$ and $c = 1.65 \times 10^{-6}\ W^{-1}\ cm\ K^{-1}$. This equation has been plotted in Figure 13 and is gives values that are within 5% of the data of Glassbrenner and Slack up to a temperature of 1000 K.

In relatively pure semiconductors ($< 10^{18}\ cm^{-3}$ impurities) virtually all the heat is conducted by phonons up to a temperature of about 600 K [3.60]. At higher temperatures ($> 1000\ K$) photons, electron-hole pairs and the electrons and holes themselves make a significant contribution to heat conduction. As the temperature is raised above the Debye temperature considerable numbers of carriers are generated and the thermal conductivity due to electron-hole pairs (ambipolar diffusion) is the same order of magnitude as the phonon thermal conductivity [3.58].

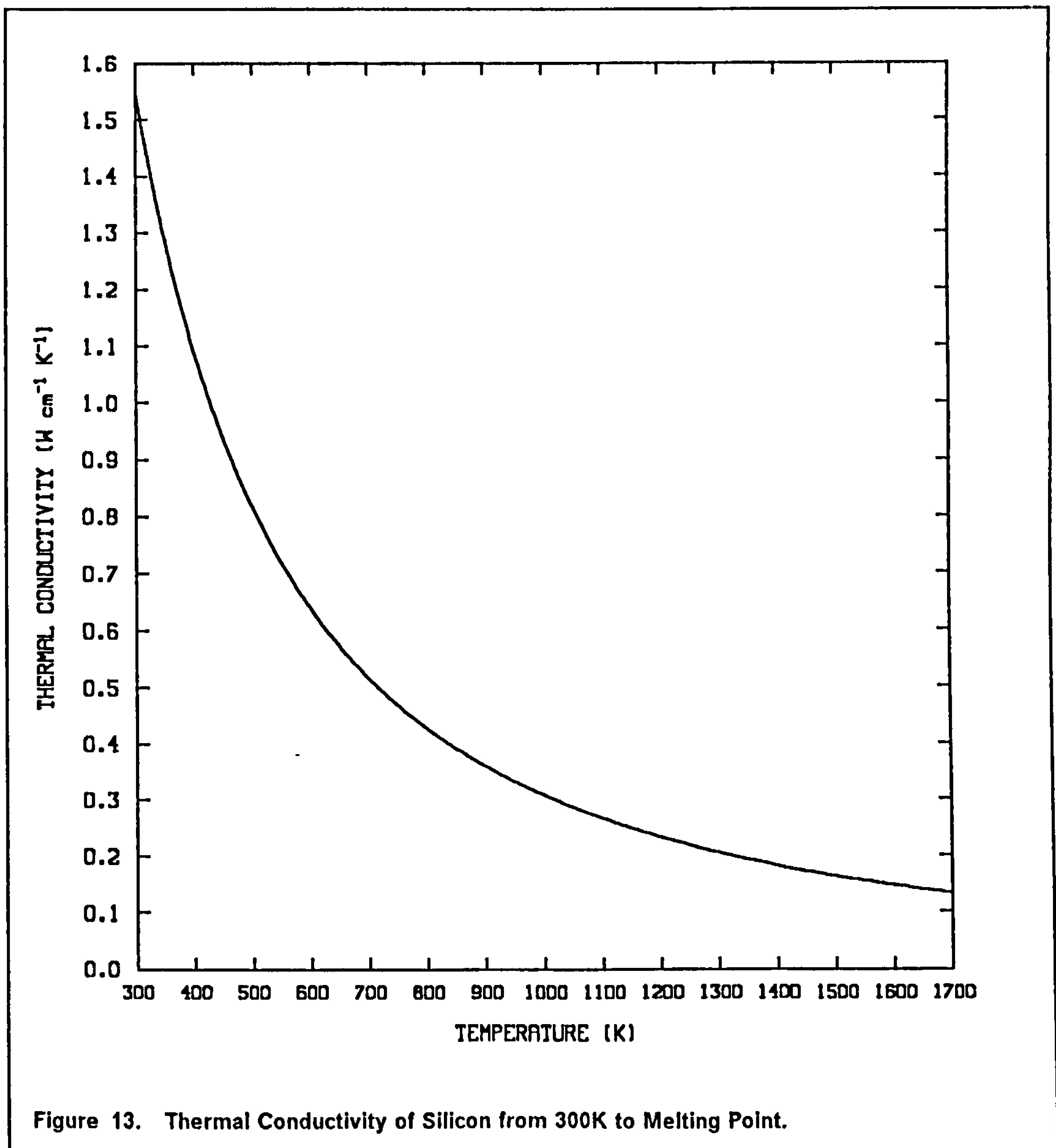


Figure 13. Thermal Conductivity of Silicon from 300K to Melting Point.

The thermal conductivity of highly doped silicon ($> 10^{19}\, cm^{-3}$) is found to be approximately 20% less than for pure material despite the larger number of carriers that are available for transporting heat. Theory indicates that the thermal conductivity is decreased as a result of carrier-phonon scattering. However, owing to the lack of available information on this subject the effects of heavy doping on thermal conductivity has been omitted from the numerical model.

The headers of the devices under consideration are made of mild steel. In metals heat is conducted primarily by electrons and the thermal conductivity is a much weaker function of temperature. Mild steel has a thermal conductivity of $0.45\, W\, cm^{-1}\, K^{-1}$ at room temperature and this value has been assumed to be independent of temperature for the purposes of modelling. The specific heat capacity, c and density, ρ of both silicon and steel are also weak functions of temperature and have been assigned the constant values given in Table 6.

	c ($J\, Kg^{-1}\, K^{-1}$)	ρ ($Kg\, cm^{-3}$)
silicon	703	2.328×10^{-3}
steel	460	7.884×10^{-3}

Table 6. Specific Heat and Density of Silicon and Mild Steel.

3.6 Heat Generation

A model for heat generation which is physically very sound has been suggested by Adler [3.61].

$$Q = \mathrm{div}\left(\frac{E_c}{q} \vec{J}_n + \frac{E_v}{q} \vec{J}_p \right) \tag{3.46}$$

where the band edges E_c and E_v are in Joules. Expanding the 'div' operator gives:

$$Q = \frac{\vec{J}_n}{q} .\mathrm{grad}\, E_c + \frac{\vec{J}_p}{q} .\mathrm{grad}\, E_v + R\, E_g \tag{3.47}$$

where R is the recombination/generation rate and E_g is the band-gap energy in Joules. Without any loss of accuracy equation (3.47) can be converted to a form more suitable for computer implementation which is given by:

$$Q = \vec{J}_n.\vec{E}_n + \vec{J}_p.\vec{E}_p + R\,E_g \tag{3.48}$$

where $\vec{E}_n$ and $\vec{E}_p$ are given by equations (2.71). The last term in this equation takes account of the heat gained or lost when carriers recombine or when they are generated.

References.

3.1 C. Kittel, *Introduction to Solid State Physics*, John Wiley and Sons, New York, 1976.

3.2 J. L. Moll, *Physics of Semiconductors*, McGraw-Hill Book Co., New York, 1964.

3.3 N. D. Arora, J. R. Hauser and D. J. Roulston, "Electron and Hole Mobilities in Silicon as a Function of Concentration and Temperature," IEEE Trans. Electron Devices, **ED-29**, pp. 292-295 (1982).

3.4 D. Long, "Scattering of Conduction Electrons by Lattice Vibrations in Silicon," Phys. Rev., **120**, pp. 2026-2032 (1960)

3.5 P. Norton, T. Braggins and H. Levinstein, "Impurity and Lattice Scattering Parameters as Determined from Hall and Mobility Analysis in *n*-Type Silicon," Phys. Rev., **B8**, pp. 5632-5653 (1973).

3.6 S. S. Li, "The Dopant Density and Temperature Dependence of Hole Mobility and Resistivity in Boron Doped Silicon," Solid State Electron., **21**, pp. 1109-1117 (1978).

3.7 S. S. Li and W. R. Thurber, "The Dopant Density and Temperature Dependence of Electron Mobility and Resistivity in *n*-Type Silicon," Solid State Electron., **20**, pp. 609-616 (1977).

3.8 C. Canali, G. Majni, R. Minder and G. Ottaviani, "Electron and Hole Drift Velocity Measurements in Silicon and Their Empirical Relation to Electric Field and Temperature," IEEE Trans. Electron Devices, **ED-22**, pp. 1045-1047 (1975).

3.9 H. Brooks, "Scattering by Ionized Impurities in Semiconductors," Phys. Rev., **83**, p. 879 (1951).

3.10 F. Mousty, P. Ostoja and L. Passari, "Relationship Between Resistivity and Phosphorus Concentrations in Silicon," J. Appl. Phys., **45**, pp. 4576-4580 (1974).

3.11 M. Finetti and A. M. Mazzoue, "Impurity Effects on Conduction in Heavily Doped *n*-type Silicon," J. Appl. Phys., **48**, pp. 4597-4600 (1977).

3.12 P. W. Chapman, O. N. Tufte, J. D. Zook and D. Long, "Electrical Properties of Heavily Doped Silicon," J. Appl. Phys., **34**, pp. 3291-3295 (1963).

3.13 V. I. Fistul, *Heavily Doped Semiconductors*, Plenum, New York, 1969.

3.14 D. M. Caughey and R. E. Thomas, "Carrier Mobilities in Silicon Empirically Related to Doping and Field," Proc. IEEE, **52**, pp. 2192-2193 (1967).

3.15 M. G. Buehler, "Semiconductor Measurement Technology," NBS Spec. Publ. 400-22 (1976); M. G. Buehler and W. R. Thurber, IEEE Trans. Electron Devices, **ED-23**, p. 968 (1976).

3.16 H. S. Bennett, "Improved Concepts for Predicting the Electrical Behavior of Bipolar Structures in Silicon," IEEE Trans. Electron Devices, **ED-30**, pp. 920-927 (1983).

3.17 N. D. Fletcher, "The High Current Limit for Semiconductor Junction Devices," Proc. IRE, **45** pp. 862-872 (1957).

3.18 S. Chapman and T. G. Cowling, *The Mathematical Theory of Non-Uniform Gases*, Cambridge University Press, Cambridge, pp. 177-179, 1952.

3.19 S. C. Choo, "Theory of a Forward-Biased Diffused-Junction P-L-N Rectifier - Part I: Exact Numerical Solutions," IEEE Trans. Electron Devices, **ED-19**, pp. 954-966 (1972)

3.20 K. Seeger, *Semiconductor Physics*, Springer, Wien-New York, 1973.

3.21 C. T. Sah, P. C. H. Chan, C.-K. Wang, R. L. Y. Sah, K. A. Yamakawa and R. Lutwack, "Effect of Zinc Impurity in Silicon Solar-Cell Efficiency," IEEE Trans. Electron Devices, **ED-28**, pp. 304-313 (1981).

3.22 S. Selberherr, *Analysis and Simulation of Semiconductor Devices*, Springer-Verlag, Wien-New York, 1984.

3.23 C. T. Kirk, "A Theory of Transistor Cutoff Frequency at High Current Densities," IEEE Trans. Electron Devices, **ED-9**, pp. 164-174 (1966).

3.24 J. M. Dorkel and Ph. Leturcq, "Carrier Mobilities in Silicon Semi-Empirically Related to Temperature, Doping and Injection Level," Solid State Electron., **24**, pp. 821-825 (1981).

3.25 J. Krausse, "Die Abhängigkeit der Trägerbeweglichkeit in Silizium von der Freien Ladungsträger - II," Solid State Electron., **15**, pp. 1377-1381 (1972).

3.26 F. Danhaüser, "Die Abhängigkeit der Trägerbeweglichkeit in Silizium von der Freien Ladungsträger - I," Solid State Electron., **15**, pp. 1371-1375 (1972).

3.27 J. E. Carroll, *Physical Models for Semiconductor Devices*, Edward Arnold Ltd., London, 1974.

3.28 D. L. Scharfetter and H. K. Gummel, "Large-Signal Analysis of a Silicon Read Diode Oscillator," IEEE Trans. Electron Devices, **ED-16**, pp. 64-77 (1969).

3.29 K. K. Thornber, "Relation of Drift Velocity to Low-Field Mobility and High-Field Saturation Velocity," J. Appl. Phys., **51**, pp. 2127-2136 (1980).

3.30 C. Jacoboni, C. Canali, G. Ottaviani and A. A. Quaranta, "A Review of Some Charge Transport Properties of Silicon," Solid State Electron., **20**, pp. 77-89 (1977).

3.31 P. Smith, M. Inoue and J. Frey, "Electron Velocity in Si and GaAs at Very High Electric Fields," Appl. Phys. Lett., **37**, pp. 797-798 (1980).

3.32 J. R. Haynes and W. Shockley, "The Mobility and Life of Injected Holes and Electrons in Germanium," Phys. Rev., **81**, pp. 835-843 (1951).

3.33 S. C. Sun and J. D. Plummer, "Electron Mobility in Inversion and Accumulation Layers on Thermally Oxidized Silicon Surfaces," IEEE Trans. Electron Devices, **ED-27**, pp. 1497-1508 (1980).

3.34 H. D. Barber, "Effective Mass and Intrinsic Concentration in Silicon," Solid State Electron., **10**, pp. 1039-1051 (1967).

3.35 E. H. Putley and W. H. Mitchell, "The Electrical Conductivity and Hall Effect of Silicon," Proc. Phys. Soc. London, **A72**, p. 193 (1958).

3.36 F. J. Morin and J. P. Maita, "Electrical Properties of Silicon Containing Arsenic and Boron," Phys. Rev., **96**, p. 28 (1954).

3.37 A. Herlet, Z. Angew. Phys., **9**, p. 155 (1957).

3.38 D. S. Lee and J. G. Fossum, "Energy-Band Distortion in Highly Doped Silicon," IEEE Trans. Electron Devices, **ED-30**, pp. 626-634 (1983).

3.39 J. W. Slotboom and H. C. de Graaff, "Measurements of Bandgap Narrowing in Si Bipolar Transistors," Solid State Electron., **19**, pp. 857-862 (1976).

3.40 S. R. Dhariwal and V. N. Ojha, "Band Gap Narrowing in Heavily Doped Silicon," Solid State Electron., **25**, pp. 909-911 (1982).

3.41 H. P. D. Lanyon and R. A. Tuft, "Bandgap Narrowing in Moderately to Heavily Doped Silicon," IEEE Trans. Electron Devices, **ED-26**, pp. 1014-1018 (1979).

3.42 A. Neugroschel, S. C. Pao and F. A. Lindholm, "A Method for Determining Energy Gap Narrowing in Highly Doped Semiconductors," IEEE Trans. Electron Devices, **ED-29**, pp. 894-902 (1982).

3.43 D. D. Tang, "Heavy Doping Effects in p-n-p Bipolar Transistors," IEEE Trans. Electron Devices, **ED-27**, pp. 563-570 (1980).

3.44 H. E. J. Wulms, "Base Current of I^2L Transistors," IEEE J. Solid-State Circuits, **SC-12**, pp. 143-150 (1977).

3.45 R. P. Mertens, J. L. Van Meerbergen, J. F. Nijs and R. J. Van Overstraeten, "Measurement of Minority Transport Parameters in Heavily Doped Silicon," IEEE Trans. Electron Devices, **ED-27**, pp. 949-955 (1980).

3.46 A. W. Wieder, "Emitter Effects in Shallow Bipolar Devices: Measurements and Consequences," IEEE Trans. Electron Devices, **ED-27**, pp. 1492-1497 (1980).

3.47 J. L Van Meerbergen, J. F. Nijs, R. P. Mertens and R. J. Van Overstraeten, "Measurement of Bandgap Narrowing and Diffusion Length in Heavily Doped Silicon," Rec. 13th IEEE Photovoltaic Specialists Conf., (1978).

3.48 M. S. Tyagi and R. Van Overstraeten, "Minority Carrier Recombination in Heavily Doped Silicon," Solid State Electron., **26**, pp. 577-597 (1983).

3.49 W. Shockley and W. T. Read, "Statistics of the Recombinations of Holes and Electrons," Phys. Rev., **87**, pp. 835-842 (1952).

3.50 R. N. Hall, "Electron-Hole Recombination in Germanium," Phys. Rev., **87**, p. 387 (1952).

3.51 J. G. Fossum, R. P. Mertens, D. S. Lee and J. F. Nijs, "Carrier Recombination and Lifetime in Highly Doped Silicon," Solid State Electron., **26**, pp. 569-576 (1983).

3.52 J. Dziewior and W. Schmid, "Auger Coefficients for Highly Doped and Highly Excited Silicon," Appl. Phys. Lett., **31**, pp. 346-348 (1977).

3.53 I. V. Grekhov, N. N. Korobkov and A. E. Otblesk, Soviet Phys. Semicond., **12**, pp. 184 (1977).

3.54 J. G. Fossum, "Computer-Aided Numerical Analysis of Silicon Solar Cells," Solid State Electron., **19**, pp. 269-277 (1976).

3.55 L. W. Davies, "The use of *P-L-N* Structures in Investigations of Transient Recombination from High Injection Levels in Silicon," Proc. IEEE, **51**, pp. 1637-1642 (1963).

3.56 M. J. B. Hamouda and W. Gerlach, "Determination of the Carrier Lifetime from the Open-Circuit Voltage Decay of p-i-n Rectifiers at High Injection Levels," IEEE Trans. Electron Devices, **ED-29**, pp. 953-955 (1982).

3.57 H. Schlangenotto and W. Gerlach, "On the Post-Injection Voltage Decay of *p-s-n* Rectifiers at High Injection Levels," Solid State Electron., **15**, pp. 393-402 (1972).

3.58 M. S. Adler, B. A. Beatty, S. Krishna, V. A. K. Temple and M. L. Torreno, Jr., "Limitations on Injection Efficiency in Power Devices," IEEE Trans. Electron Devices, **ED-23**, pp. 858-863 (1976).

3.59 C. J. Glassbrenner and G. A. Slack, "Thermal Conductivity of Silicon and Germanium from 3K to Melting Point," Phys. Rev., **134**, pp. A1058-A1069 (1964).

3.60 P. D. Maycock, "Thermal Conductivity of Silicon, Germanium, III-V Compounds and III-V Alloys," Solid State Electron., **10**, pp. 161-168 (1967).

3.61 M. S. Adler, "Accurate Calculations of the Forward Drop and Power Dissipation in Thyristors," IEEE Trans. Electron Devices, **ED-25**, pp. 16-22 (1978).

Chapter 4. Discretization Of The Governing Equations.

In general the system of partial differential equations that govern device operation cannot be solved by analytical methods. Only a limited number of special types of elliptic equations have been solved analytically; these solutions being restricted to problems involving domains with simple shapes and boundary conditions that can be easily satisfied. In most cases, therefore, a solution must be sought using numerical methods.

The major difference between analytical and numerical methods lies in the underlying assumptions which have to be made in both methods. Assumptions in analytical approaches are usually made on the basis of some prior knowledge of device operation. The numerical approach, however, is to solve a set of discrete equations, which is itself an approximation to the governing equations. A solution can be calculated without having to make many prior assumptions of device operation. This makes it superior for solving complex problems, where an analytical approach would require many over-simplifying assumptions.

A solution to a problem using numerical methods involves essentially two stages. Firstly, the domain must be partitioned into a number of subdomains. The differential equations are then approximated within each of the subdomains by algebraic equations. This stage serves to reduce the continuous problem into one which involves only values of the continuous dependent variables (ψ,n,p,T) at discrete points in the domain, and is called the discretization stage. Secondly, the resulting system of algebraic equations, which is in general non-linear must be solved for the discrete dependent variables. This stage will be treated separately in Chapter 5.

Discretization of the governing equations is most commonly achieved using either finite difference [4.1] or finite element [4.2] methods. The difference method uses a local approximation to the differential operator, whereas the finite element method applies a collection of shape functions as trial functions to approximate the solution globally. The finite difference method has be chosen for use in preference to the finite element method as it has a number of advantages. Firstly, the

formulation of the difference method is better established, moreover it has been adapted especially to suit the semiconductor problem. Secondly, calculation of a desired set of algebraic equations requires significantly less effort by the difference method. Finally, a problem arises when the finite element method is applied to the continuity equations [4.3]. This is due to the difficulty of fitting shape functions, which are represented by low order polynomials, to the exponentially varying carrier concentrations. The finite element method does, however, have the advantage that it allows more efficient partitioning of the domain into sub-domains. This results in a smaller set of equations than would be obtained with an equally accurate difference method. Less effort would then be required in solving this reduced set of equations. However, this saving is somewhat offset by the greater effort required to obtain the discrete approximation by the finite element method in the first place.

4.1 Discretization Of The Static Equations.

Initially the discretization of the static equations, only, will be considered. These equations describe device operation for the case where the boundary conditions for electrostatic potential are made time invariant, and are obtained from the governing equations by setting the partial derivatives of the carrier concentrations and temperature with respect to time to zero, thus:

$$\text{div}\,\text{grad}\,\psi + \frac{q}{\varepsilon_{sil}}(N_D - N_A + p - n) = 0 \tag{4.1}$$

$$\frac{1}{q}\,\text{div}\,\vec{J}_n - R = 0 \tag{4.2}$$

$$\frac{1}{q}\,\text{div}\,\vec{J}_p + R = 0 \tag{4.3}$$

$$\text{div}\,K(T)\,\text{grad}\,T + Q = 0 \tag{4.4}$$

Discretization of these equations will now be carried out using both cartesian and cylindrical co-ordinate systems.

4.1.1 Cartesian Co-ordinates.

Only two orthogonal space dimensions will be considered. A third dimension can be added by simply extending the two dimensional formulation. The three dimensional problem, however, will not be considered as it results in

an extremely large set of algebraic equations, the solution of which requires computational resources well beyond those available at present. In many semiconductor devices (eg. interdigitated structures) current flow, electrostatic potential and temperature vary only in two dimensions over much of the device. Operation of these devices can, therefore, be accurately modelled using two space dimensions only.

The first stage towards a solution using the classical method of finite differences is to generate a mesh, which is a series of mesh lines drawn parallel to the co-ordinate axes. For example, if NX lines are drawn parallel to the y-axis and NY lines parallel to the x-axis, such that the first and last lines coincide with the boundaries of the domain, then the resulting mesh will have $NX.NY$ points of intersection. These points are usually called nodes or mesh points. A typical mesh for the example problem is shown in Figure 14. Discretization will be performed using the classical five point method and the nomenclature to be used throughout is given in Figure 15.

Taylors theorem provides the basis for all difference equations and it states that if a function, u and its derivatives are single valued, finite and continuous functions of x and y, then:

$$u_{i+1,j} = u_{i,j} + h_i \left.\frac{\partial u}{\partial x}\right|_{i,j} + \frac{h_i^2}{2!} \left.\frac{\partial^2 u}{\partial x^2}\right|_{i,j} + \frac{h_i^3}{3!} \left.\frac{\partial^3 u}{\partial x^3}\right|_{i,j} + \frac{h_i^4}{4!} \left.\frac{\partial^4 u}{\partial x^4}\right|_{i,j} + O(h^5) \tag{4.5}$$

$$u_{i-1,j} = u_{i,j} - h_{i-1} \left.\frac{\partial u}{\partial x}\right|_{i,j} + \frac{h_{i-1}^2}{2!} \left.\frac{\partial^2 u}{\partial x^2}\right|_{i,j} - \frac{h_{i-1}^3}{3!} \left.\frac{\partial^3 u}{\partial x^3}\right|_{i,j} + \frac{h_{i-1}^4}{4!} \left.\frac{\partial^4 u}{\partial x^4}\right|_{i,j} + O(h^5) \tag{4.6}$$

$$u_{i,j+1} = u_{i,j} + k_j \left.\frac{\partial u}{\partial y}\right|_{i,j} + \frac{k_j^2}{2!} \left.\frac{\partial^2 u}{\partial y^2}\right|_{i,j} + \frac{k_j^3}{3!} \left.\frac{\partial^3 u}{\partial y^3}\right|_{i,j} + \frac{k_j^4}{4!} \left.\frac{\partial^4 u}{\partial y^4}\right|_{i,j} + O(k^5) \tag{4.7}$$

$$u_{i,j-1} = u_{i,j} - k_{j-1} \left.\frac{\partial u}{\partial y}\right|_{i,j} + \frac{k_{j-1}^2}{2!} \left.\frac{\partial^2 u}{\partial y^2}\right|_{i,j} - \frac{k_{j-1}^3}{3!} \left.\frac{\partial^3 u}{\partial y^3}\right|_{i,j} + \frac{k_{j-1}^4}{4!} \left.\frac{\partial^4 u}{\partial y^4}\right|_{i,j} + O(k^5) \tag{4.8}$$

Discretization is carried out by the simple application of these series to the continuous problem. Initially, only the inner points ($1 < i < NX$, $1 < j < NY$) will be considered, and the boundary conditions will be treated in a later section. Each of

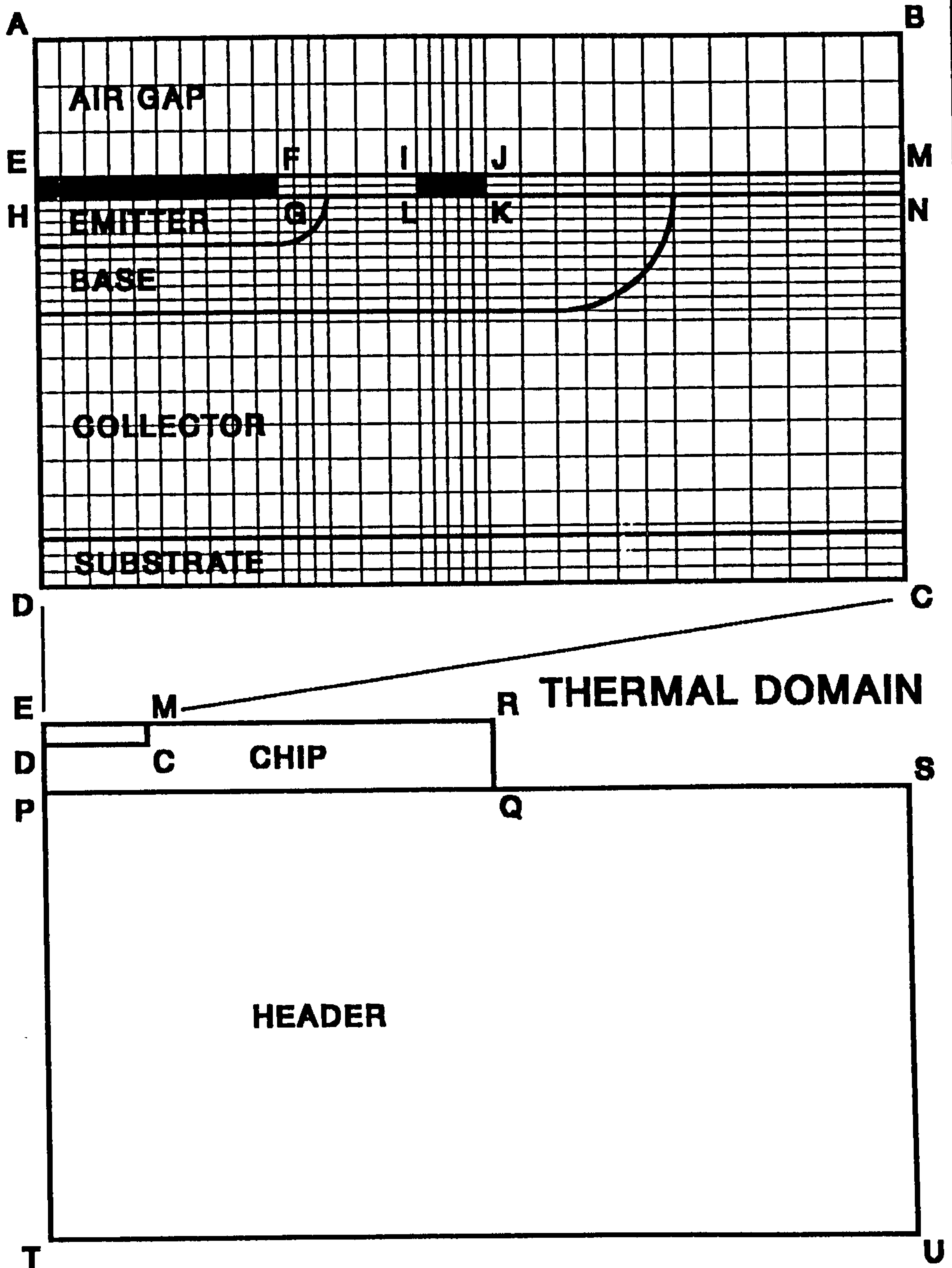


Figure 14. A Typical Mesh.

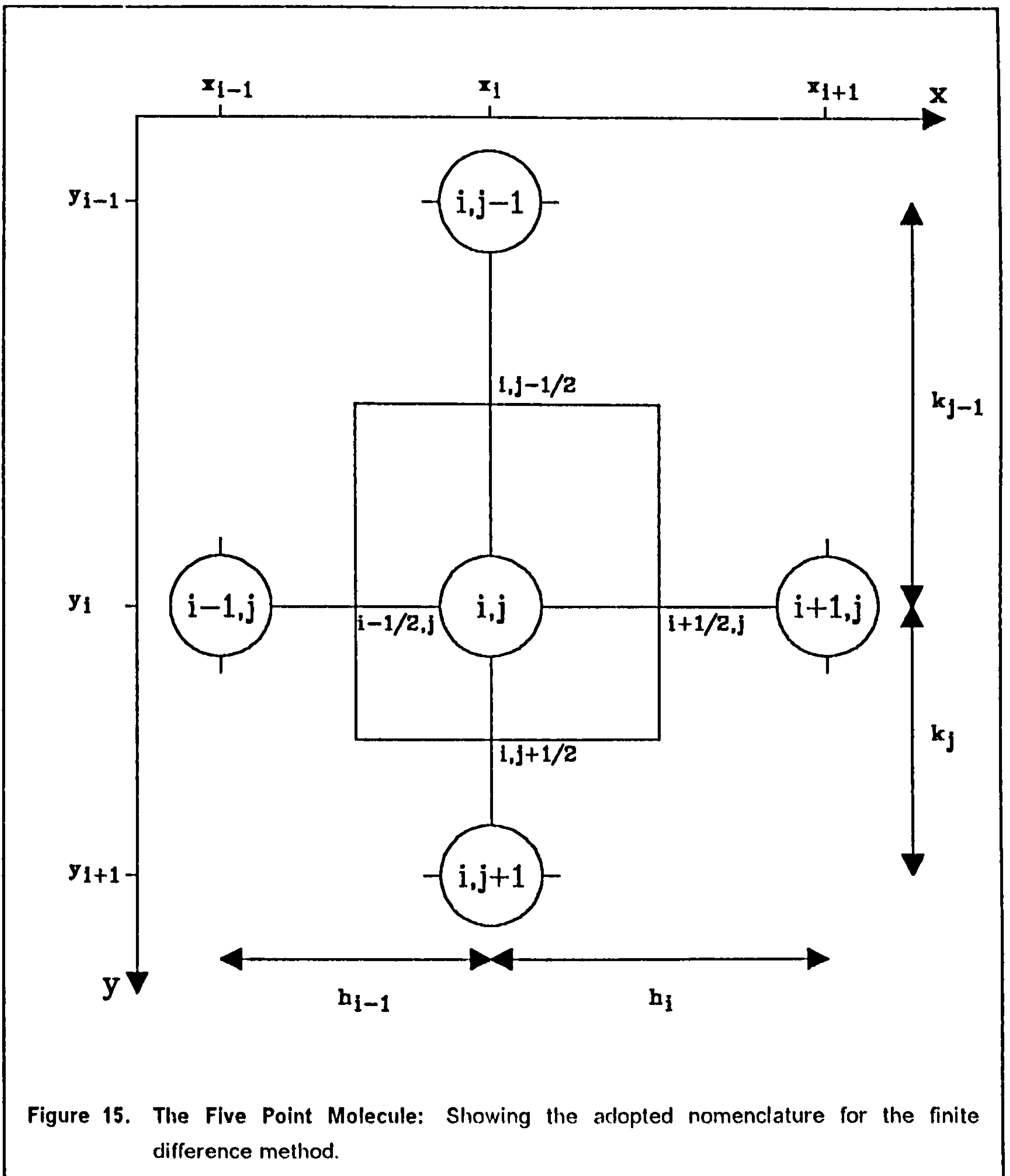


Figure 15. The Five Point Molecule: Showing the adopted nomenclature for the finite difference method.

the partial differential equations will be treated separately, beginning with the Poisson equation.

4.1.1.1 Discretization of the Poisson Equation

Expansion of the 'div' and 'grad' operators in the Poisson equation (4.1) results in:

$$\frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} + \frac{q}{\varepsilon_{sil}}(N_D - N_A + p - n) = 0 \tag{4.9}$$

The partial derivatives can now be represented at each of the inner points using the Taylor series expansions. Eliminating the first order derivatives from series' (4.5) and (4.6) and ignoring terms of fifth and higher order gives:

$$\left.\frac{\partial^2\psi}{\partial x^2}\right|_{i,j} = \frac{\dfrac{\psi_{i+1,j}-\psi_{i,j}}{h_i} - \dfrac{\psi_{i,j}-\psi_{i-1,j}}{h_{i-1}}}{0.5\,(h_{i-1}+h_i)} + T_{\psi x_{i,j}} \tag{4.10}$$

The local truncation error, $T_{\psi x_{i,j}}$ is given by:

$$T_{\psi x_{i,j}} = \frac{h_{i-1}-h_i}{3}\left.\frac{\partial^3\psi}{\partial x^3}\right|_{i,j} - \frac{(h_i-h_{i-1})^2+h_{i-1}h_i}{12}\left.\frac{\partial^4\psi}{\partial x^4}\right|_{i,j} \tag{4.11}$$

The local truncation error is the difference between the exact value of the second derivative at (i,j) and its discrete approximation. Here it is seen that for a uniform mesh the truncation error is proportional to the square of the mesh spacing, weighted with the fourth order partial derivative. However, a much finer mesh would be required to define the electrostatic potential in regions where it varies rapidly, compared to that required in regions of slowly varying potential. Hence, the use of a strongly non-uniform mesh is often obligatory. A truncation error which is of first order in mesh spacing weighted with the third order derivative is, therefore, the best that can be achieved.

Similarly, for the derivatives parallel to the y-axis, elimination of the first order derivatives from (4.7) and (4.8) results in:

$$\left.\frac{\partial^2\psi}{\partial y^2}\right|_{i,j} = \frac{\dfrac{\psi_{i,j+1}-\psi_{i,j}}{k_j} - \dfrac{\psi_{i,j}-\psi_{i,j-1}}{k_{j-1}}}{0.5\,(k_{j-1}+k_j)} + T_{\psi y_{i,j}} \tag{4.12}$$

where,

$$T_{\psi y_{i,j}} = \frac{k_{j-1}-k_j}{3}\left.\frac{\partial^3\psi}{\partial y^3}\right|_{i,j} - \frac{(k_j-k_{j-1})^2+k_{j-1}k_j}{12}\left.\frac{\partial^4\psi}{\partial y^4}\right|_{i,j} \tag{4.13}$$

The final discrete form of the Poisson equation is obtained by substituting equations (4.10) and (4.12) into equation (4.9), thus:

$$\frac{\dfrac{\psi_{i+1,j}-\psi_{i,j}}{h_i} - \dfrac{\psi_{i,j}-\psi_{i-1,j}}{h_{i-1}}}{0.5\,(h_{i-1}+h_i)} + \frac{\dfrac{\psi_{i,j+1}-\psi_{i,j}}{k_j} - \dfrac{\psi_{i,j}-\psi_{i,j-1}}{k_{j-1}}}{0.5\,(k_{j-1}+k_j)} \tag{4.14}$$

$$+ T_{\psi x_{i,j}} + T_{\psi y_{i,j}} + \frac{q}{\varepsilon_{sil}}(N_{D_{i,j}} - N_{A_{i,j}} - n_{i,j} + p_{i,j}) = 0$$

Collecting all terms in the discrete potentials and ignoring the local truncation error results in the following equation, which is more suitable for computer implementation.

$$A_{i,j}\psi_{i,j-1}+B_{i,j}\psi_{i-1,j}+C_{i,j}\psi_{i,j}+D_{i,j}\psi_{i+1,j}+E_{i,j}\psi_{i,j+1}+\frac{q}{\varepsilon_{sil}}(N_{D_{i,j}}-N_{A_{i,j}}-n_{i,j}+p_{i,j})=0 \tag{4.15}$$

The coefficients of the discrete potentials are given by:

$$A_{i,j}=\frac{2}{k_{j-1}(k_{j-1}+k_j)} \tag{4.16}$$

$$B_{i,j}=\frac{2}{h_{i-1}(h_{i-1}+h_i)} \tag{4.17}$$

$$D_{i,j}=\frac{2}{h_i(h_{i-1}+h_i)} \tag{4.18}$$

$$E_{i,j}=\frac{2}{k_j(k_{j-1}+k_j)} \tag{4.19}$$

$$C_{i,j}=-(A_{i,j}+B_{i,j}+D_{i,j}+E_{i,j}) \tag{4.20}$$

This completes treatment of the Poisson equation. It remains to note that the discrete Laplace equation is obtained by simply equating the space charge term in equation (4.15) to zero.

4.1.1.2 Discretization of the Continuity Equations.

Initially, consideration will be given to finding discrete approximations to the carrier transport equations, (2.41) and (2.42). The desired result will then be obtained by substituting these approximations into discrete versions of the continuity equations, (4.2) and (4.3).

The formulation that has been adopted for the discretization of the transport equations was first suggested by Scharfetter and Gummel [4.4], and has since been extended to include the effects of temperature gradients by McAndrew et. al. [4.5]. This formulation has become the very foundation for all accurate numerical solutions of the semiconductor problem, and it has been used to good effect by many workers in the past eg. [4.6], [4.7]. It gains tremendous advantages over conventional discretization schemes, because the assumptions made in its derivation have been chosen to reflect the physics of carrier transport in semiconductors. The main assumption being made is that the components of the

electron and hole current densities parallel to the x and y axes are conserved along the line between adjacent mesh points. This is equivalent to assuming that both recombination/generation processes and divergence of current are negligible between adjacent mesh points. The resulting formulae predict an exponential variation of the carrier concentrations between mesh points. This, therefore, gives a very good representation of the distribution of carriers in regions of particular interest and importance, for example in diffused layer regions and in the vicinity of forward biased junctions. Conventional discretization schemes, that assume a linear variation of carrier concentration between mesh points, are at a distinct disadvantage, as they would require a much finer mesh in order to resolve the concentrations to the same accuracy.

Only the electron transport equation will be treated as the procedure for holes is completely analogous. However, results for both electrons and holes will be given for completeness. Firstly, the current density between two mesh points is approximated by the following truncated Taylor expansions.

$$J_{nx}(x \in [x_i, x_{i+1}], y_j) = J_{nx_{i+1/2,j}} + \left(x - x_i - \frac{h_i}{2}\right) \frac{\partial J_{nx}}{\partial x}\bigg|_{i+1/2,j} \tag{4.21}$$

$$J_{ny}(x_i, y \in [y_j, y_{j+1}]) = J_{ny_{i,j+1/2}} + \left(y - y_j - \frac{k_j}{2}\right) \frac{\partial J_{ny}}{\partial y}\bigg|_{i,j+1/2} \tag{4.22}$$

Similar expressions can be written for the current densities in the intervals, $(x \in [x_{i-1}, x_i], y_j)$ and $(x_i, y \in [y_{j-1}, y_j])$.

Substitution of equation (4.21) into the electron transport equation, (2.41), for example, yields the following relation for the electron concentration in the interval $(x \in [x_i, x_{i+1}], y_j)$.

$$kT\mu_n \frac{\partial n}{\partial x} - q\mu_n n \frac{\partial \psi_n}{\partial x} + \frac{k}{2}\mu_n n \frac{\partial T}{\partial x} = J_{nx_{i+1/2,j}} + \left(x - x_i - \frac{h_i}{2}\right) \frac{\partial J_{nx}}{\partial x}\bigg|_{i+1/2,j} \tag{4.23}$$

Here the Einstein relation, (2.39) is assumed to hold for the diffusion coefficient and mobility and the thermal diffusion coefficient, D_n^T has been assumed to be given by $D_n/(2T)$. In the initial treatment the last term in equation (4.23) will be omitted. The equation can then be written in the form stated by McAndrew et. al. [4.5], which is suitable for solving.

$$\frac{\partial (nT)}{\partial x} - \left(\frac{1}{2T}\frac{\partial T}{\partial x} + \frac{q}{kT}\frac{\partial \psi_n}{\partial x}\right)(nT) = \frac{J_{nx_{i+1/2,j}}}{k\mu_n} \tag{4.24}$$

A number of assumptions must be made before a solution can be sought. Firstly, the partial derivative of the electrostatic potential must be assumed to be constant in the interval under consideration. This assumption has already been invoked in the discretization of the Poisson equation. Similarly, the partial derivative of the electron temperature must also be assumed to be constant. As will be seen in section 4.1.1.3 this is consistent with the assumptions to be made in the discretization of the heat flow equation. Finally, the carrier mobility must also be assumed to be constant along the path of integration. Having made these assumptions the equation becomes a first order non-homogenous partial differential equation in the variable (nT) with a single parameter, $J_{n_{i+1/2,j}}$. It can be solved subject to the following boundary conditions.

$$n(x_i, y_j) = n_{i,j}, \quad T(x_i, y_j) = T_{i,j} \tag{4.25}$$

$$n(x_{i+1}, y_j) = n_{i+1,j}, \quad T(x_{i+1}, y_j) = T_{i+1,j} \tag{4.26}$$

Following the usual solution procedure for such an equation the integrating factor is calculated, thus:

$$I.F. = \exp\left\{ -\int \left(\frac{1}{2T} \frac{\partial T}{\partial x} + \frac{q}{kT} \frac{\partial \psi}{\partial x} \right) \partial x \right\} \tag{4.27}$$

The subscript of ψ_n has been omitted at this stage, in order to preserve clarity. It must be noted that in the treatment of the electron transport equation all occurrences of ψ should, therefore, be replaced by ψ_n, and similarly in the treatment of the hole transport equation, all occurrences of ψ should be replaced by ψ_p. Application of the chain rule to equation (4.27) gives:

$$I.F. = \exp\left\{ -\frac{1}{2} \int \frac{\partial T}{T} - \frac{q}{k} \int \frac{\partial \psi}{\partial x} \frac{\partial x}{\partial T} \frac{\partial T}{T} \right\} \tag{4.27}$$

Since $\frac{\partial \psi}{\partial x}$ and $\frac{\partial T}{\partial x}$ have been assumed to be constant then:

$$I.F. = T^E \tag{4.29}$$

where the exponent, E is given by:

$$E = -\frac{1}{2} - \frac{q}{k} \frac{\partial \psi}{\partial T} \tag{4.30}$$

Equation (4.24) is then multiplied throughout by its integrating factor, giving:

$$\frac{\partial(nT^{E+1})}{\partial x} = \frac{J_{nx_{i+1/2,j}}}{k\,\mu_{n_{i+1/2,j}}}\;T^{E} \tag{4.31}$$

Separating the variables and integrating results in:

$$\int \partial(nT^{E+1}) = \frac{J_{nx_{i+1/2,j}}}{k\,\mu_{n_{i+1/2,j}}}\;\frac{\partial x}{\partial T}\int T^{E}\partial T \tag{4.32}$$

Carrying out the integration results in:

$$nT^{E+1} = \frac{J_{nx_{i+1/2,j}}}{k\,\mu_{n_{i+1/2,j}}}\;\frac{\partial x}{\partial T}\;\frac{T^{E+1}}{E+1} + C \tag{4.33}$$

The parameter, $J_{nx_{i+1/2,j}}$ and the constant of integration, C may be found by substituting the boundary conditions, (4.25) and (4.26) into equation (4.33), and then solving the resulting pair of equations. Following this procedure results in:

$$J_{nx_{i+1/2,j}} = \frac{k\,\mu_{n_{i+1/2,j}}}{h_i L(T_{i,j},T_{i+1,j})}\left(n_{i+1,j}B\{F(\psi_{i+1,j},\psi_{i,j},T_{i+1,j},T_{i,j})\} - n_{i,j}B\{F(\psi_{i,j},\psi_{i+1,j},T_{i,j},T_{i+1,j})\}\right) \tag{4.34}$$

and,

$$C = \frac{n_{i+1,j} - n_{i,j}}{\exp\left\{\dfrac{F(\psi_{i+1,j},\psi_{i,j},T_{i+1,j},T_{i,j})\log_e T_{i+1,j}}{\log_e(T_{i+1,j}/T_{i,j})}\right\} - \exp\left\{\dfrac{F(\psi_{i+1,j},\psi_{i,j},T_{i+1,j},T_{i,j})\log_e T_{i,j}}{\log_e(T_{i+1,j}/T_{i,j})}\right\}} \tag{4.35}$$

where the Bernoulli function, B and the functions F and L are given by:

$$B(x) = \frac{x}{\exp x - 1} \tag{4.36}$$

$$F(\psi_1,\psi_2,T_1,T_2) = \left(\frac{T_2 - T_1}{2} - \frac{q}{k}\,(\psi_2 - \psi_1)\right).L(T_1,T_2) \tag{4.37}$$

$$L(T_1,T_2) = \frac{\log_e(T_2/T_1)}{T_2 - T_1} \tag{4.38}$$

Substituting for $J_{nx_{i+1/2,j}}$ and C in equation (4.33) gives the desired solution for the electron concentration, which is:

$$n(x \in [x_i, x_{i+1}], y_j) = n_{i,j} + G(x, \psi, T)\,(n_{i+1,j} - n_{i,j}) \tag{4.39}$$

where the growth function, $G(x, \psi, T)$ is given by,

$$G(x, \psi, T) = \frac{1 - \exp\left\{F(\psi_{i+1,j}, \psi_{i,j}, T_{i+1,j}, T_{i,j})\dfrac{\log_e(1 + (T_{i+1,j} - T_{i,j})(x - x_i)/T_{i,j}h_i)}{\log_e(T_{i+1,j}/T_{i,j})}\right\}}{1 - \exp\{F(\psi_{i+1,j}, \psi_{i,j}, T_{i+1,j}, T_{i,j})\}} \tag{4.40}$$

A fully analogous treatment of the hole transport equation results in:

$$p(x \in [x_i, x_{i+1}], y_j) = p_{i,j} + G(x, -\psi, T)\,(p_{i+1,j} - p_{i,j}) \tag{4.41}$$

The growth function is plotted in Figure 16 and Figure 17 in a normalised interval for different values of $(\psi_{i+1,j} - \psi_{i,j})$, $T_{i+1,j}$ and $T_{i,j}$. The shape of the growth function arises as a direct consequence of the condition of constant current density imposed in the derivation. Since the diffusion gradient of the carrier concentrations is given by the gradient of the growth function it is evident the growth function reduces to a simple linear function if the field and temperature gradient are zero.

If the previous calculation is performed without omitting the last term in equation (4.23) then an additional term must be included on the right hand side of equation (4.34) for the internodal current density. This term is given by:

$$T_{n_{i+1/2,j}} = h_i\left\{\frac{0.5 - T_{i+1,j}/(T_{i+1,j} - T_{i,j} + F(\psi_{i,j}, \psi_{i+1,j}, T_{i,j}, T_{i+1,j})/L(T_{i,j}, T_{i+1,j}))}{\exp\{F(\psi_{i+1,j}, \psi_{i,j}, T_{i+1,j}, T_{i,j})\} - 1}\right. \tag{4.42}$$

$$\left. - \frac{0.5 + T_{i,j}/(T_{i+1,j} - T_{i,j} + F(\psi_{i,j}, \psi_{i+1,j}, T_{i,j}, T_{i+1,j})/L(T_{i,j}, T_{i+1,j}))}{\exp\{F(\psi_{i,j}, \psi_{i+1,j}, T_{i,j}, T_{i+1,j})\} - 1}\right\} \left.\frac{\partial J_{nx}}{\partial x}\right|_{i+1/2,j}$$

The bracketted term in equation (4.42) has been plotted in Figure 18 against $(\psi_{i+1,j} - \psi_{i,j})$ for various combinations of temperature. It may be seen that the absolute value of $T_{n_{i+1/2,j}}$ never exceeds 0.5. Thus at worst, the truncation error is of first order in mesh spacing weighted with the first derivative of the current density.

The equation for the vertical component of the electron current density in the interval $(x_i, y \in [y_j, y_{j+1}])$ can now be written by simply adusting equation (4.34), thus:

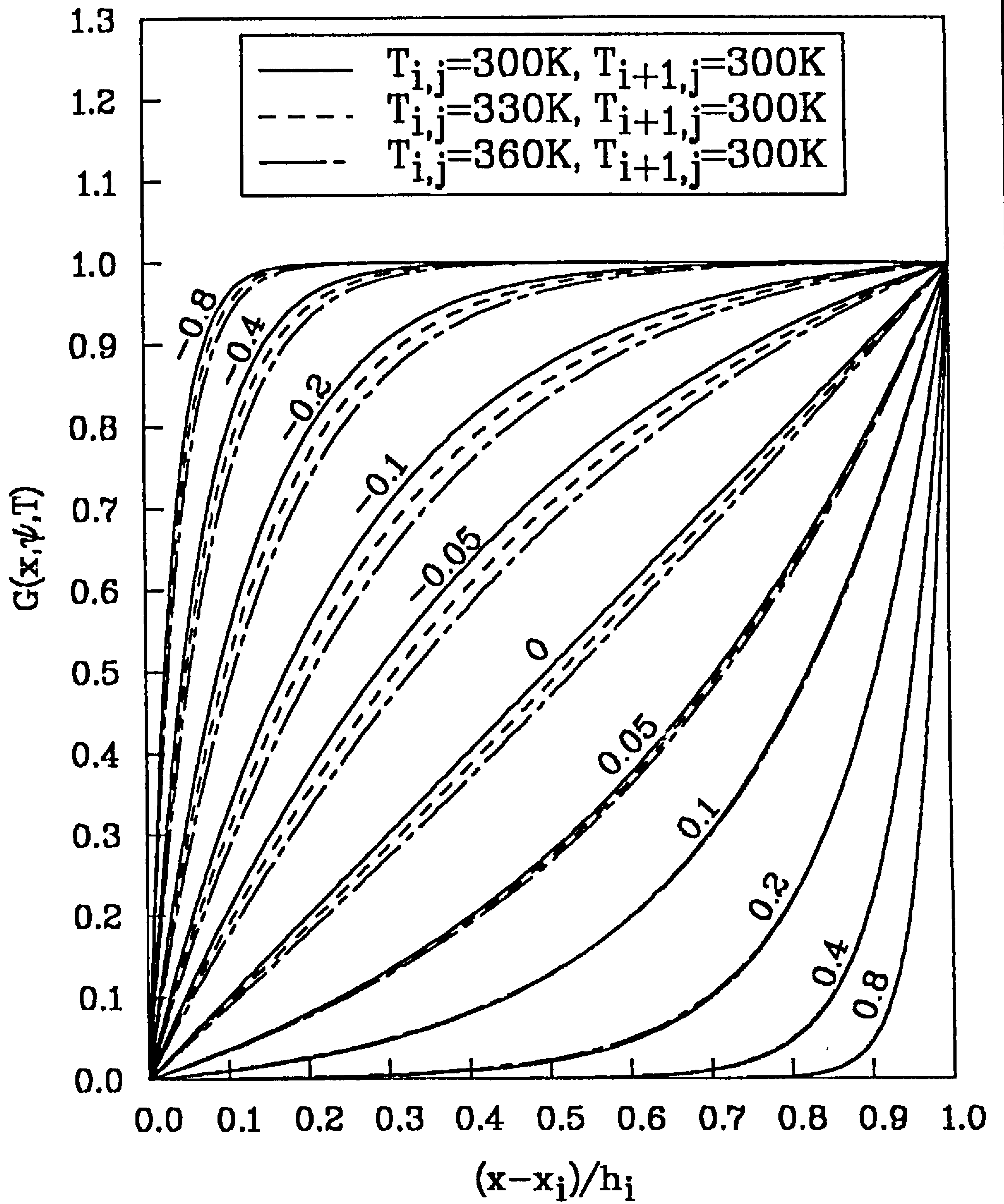


Figure 16. The Growth Function: Plotted in a normalised interval for various values of $(\psi_{i+1,j} - \psi_{i,j})$.

$$J_{ny_{i,j+1/2}} = \frac{k\,\mu_{n_{i,j+1/2}}}{k_j L(T_{i,j}, T_{i,j+1})} \left(n_{i,j+1} B\{F(\psi_{i,j+1}, \psi_{i,j}, T_{i,j+1}, T_{i,j})\} - n_{i,j} B\{F(\psi_{i,j}, \psi_{i,j+1}, T_{i,j}, T_{i,j+1})\} \right) \tag{4.43}$$

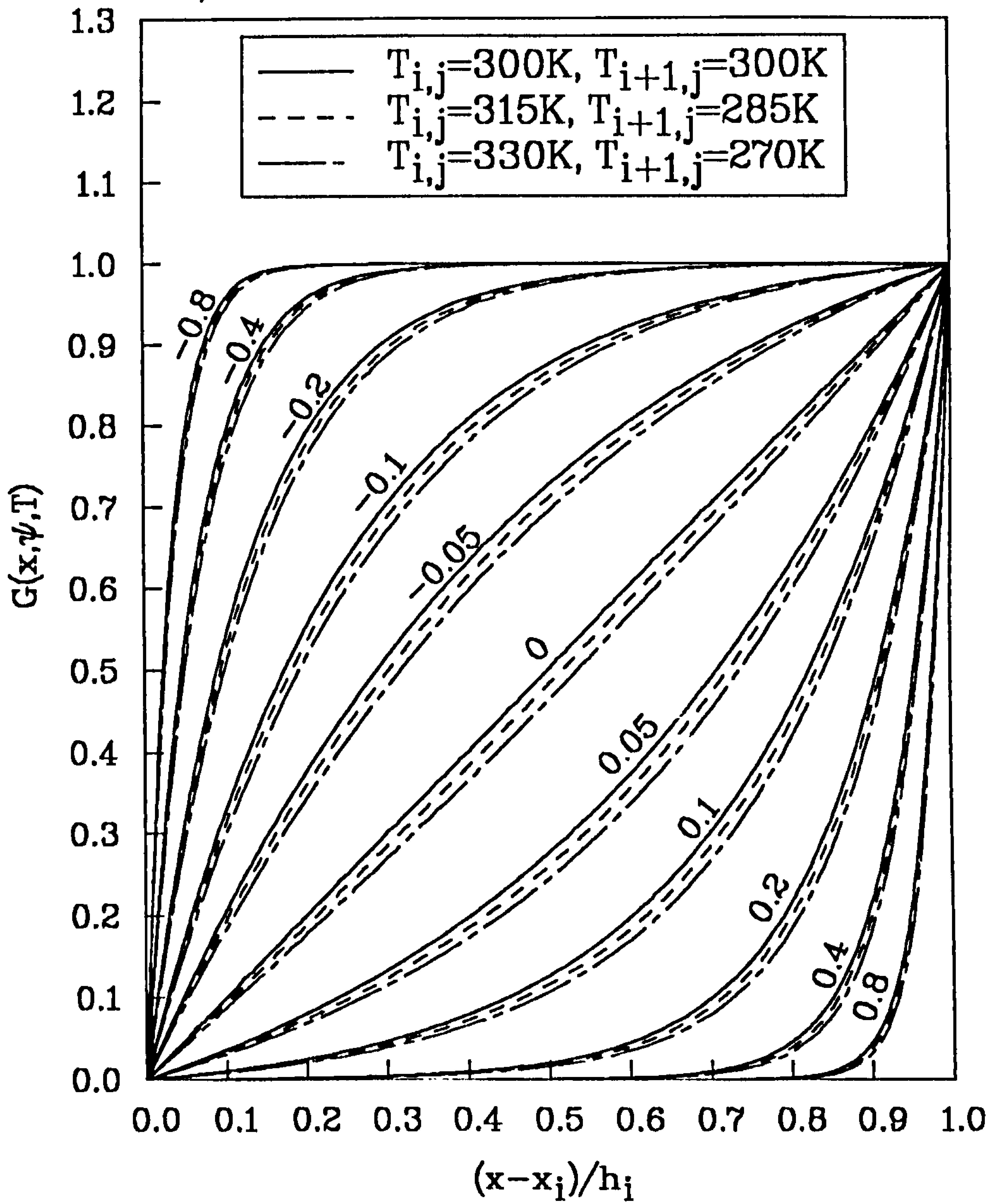


Figure 17. The Growth Function: Plotted in a normalised interval for various values of $(\psi_{i+1,j} - \psi_{i,j})$.

The corresponding hole current density components may be calculated by treating the hole transport equation (2.42) in exactly the same way. These components are given by:

$$J_{px_{i+1/2,j}} = \frac{k\,\mu_{p_{i+1/2,j}}}{h_i L(T_{i,j}, T_{i+1,j})} \left(p_{i,j} B\{F(\psi_{i+1,j}, \psi_{i,j}, T_{i,j}, T_{i+1,j})\} - p_{i+1,j} B\{F(\psi_{i,j}, \psi_{i+1,j}, T_{i+1,j}, T_{i,j})\}\right) \quad (4.44)$$

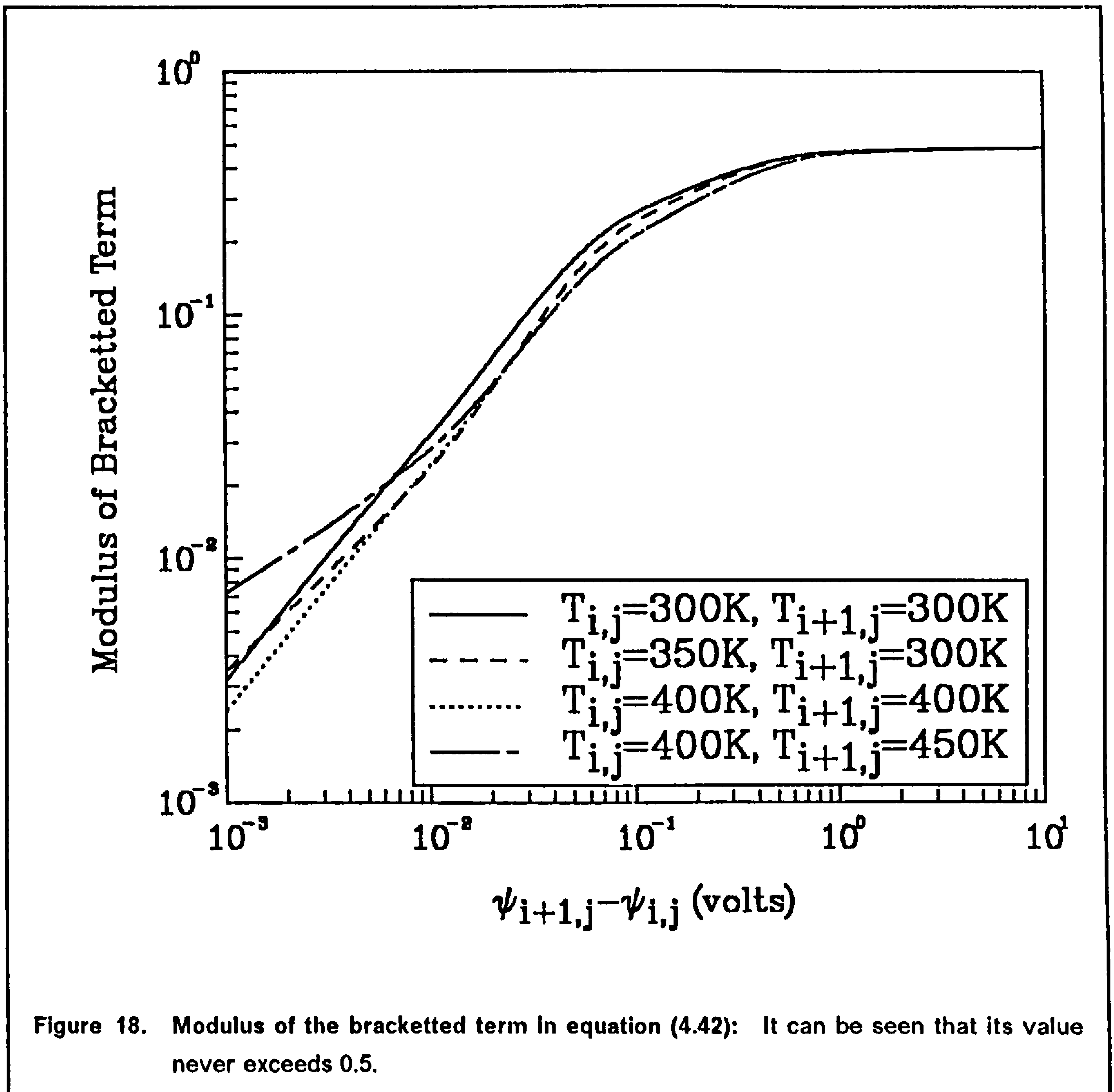


Figure 18. Modulus of the bracketted term in equation (4.42): It can be seen that its value never exceeds 0.5.

$$J_{py_{i,j+1/2}} = \frac{k\,\mu_{p_{i,j+1/2}}}{k_j L(T_{i,j}, T_{i,j+1})} \left(p_{i,j} B\{F(\psi_{i,j+1}, \psi_{i,j}, T_{i,j}, T_{i,j+1})\} - p_{i,j+1} B\{F(\psi_{i,j}, \psi_{i,j+1}, T_{i,j+1}, T_{i,j})\}\right) \tag{4.45}$$

In order to discretize the continuity equations, (4.2) and (4.3), the 'div' operator must first be expanded, thus:

$$\frac{1}{q}\left(\frac{\partial J_{nx}}{\partial x} + \frac{\partial J_{ny}}{\partial y}\right) - R = 0 \tag{4.46}$$

$$\frac{1}{q}\left(\frac{\partial J_{px}}{\partial x} + \frac{\partial J_{py}}{\partial y}\right) + R = 0 \tag{4.47}$$

Discretized versions of the partial derivatives may be obtained by, firstly writing Taylor series' for the x and y directed current densities at the central node (i,j). This stage is similar to that outlined in the discretization of the Poisson equation (cf. (4.5) to (4.8)). However, rather than eliminate the first order derivatives as was done in the Poisson equation, here, the current densities at the central node are eliminated. The following relations are subsequently obtained.

$$\left.\frac{\partial J_{nx}}{\partial x}\right|_{i,j} = \frac{J_{nx_{i+1/2,j}} - J_{nx_{i-1/2,j}}}{0.5(h_i + h_{i-1})} + \frac{h_{i-1} - h_i}{4} \left.\frac{\partial^2 J_{nx}}{\partial x^2}\right|_{i,j} \tag{4.48}$$

$$\left.\frac{\partial J_{ny}}{\partial y}\right|_{i,j} = \frac{J_{ny_{i,j+1/2}} - J_{ny_{i,j-1/2}}}{0.5(k_j + k_{j-1})} + \frac{k_{j-1} - k_j}{4} \left.\frac{\partial^2 J_{ny}}{\partial y^2}\right|_{i,j} \tag{4.49}$$

The discrete version of the continuity equation for electrons is obtained by adding (4.48) to (4.49) and substituting the result into equation (4.46). By ignoring the second term in equations (4.48) and (4.49) the following relation is obtained.

$$\frac{2}{q}\left(\frac{J_{nx_{i+1/2,j}} - J_{nx_{i-1/2,j}}}{h_i + h_{i-1}} + \frac{J_{ny_{i,j+1/2}} - J_{ny_{i,j-1/2}}}{k_j + k_{j-1}}\right) - R_{i,j} = 0 \tag{4.50}$$

Similarly for holes,

$$\frac{2}{q}\left(\frac{J_{px_{i+1/2,j}} - J_{px_{i-1/2,j}}}{h_i + h_{i-1}} + \frac{J_{py_{i,j+1/2}} - J_{py_{i,j-1/2}}}{k_j + k_{j-1}}\right) + R_{i,j} = 0 \tag{4.51}$$

Substituting the discretized current transport equations (eg. (4.34) and (4.43)) into equation (4.50), and collecting all the terms in the discrete electron concentrations results in the final discrete version of the electron continuity equation, which is given by:

$$F_{i,j}\, n_{i,j-1} + G_{i,j}\, n_{i-1,j} + H_{i,j}\, n_{i,j} + I_{i,j}\, n_{i+1,j} + L_{i,j}\, n_{i,j+1} - R_{i,j} = 0 \tag{4.52}$$

where,

$$F_{i,j} = \frac{2\,k\,\mu_{n_{i,j-1/2}}\, B\{F(\psi_{i,j-1}, \psi_{i,j}, T_{i,j-1}, T_{i,j})\}}{q\, L(T_{i,j-1}, T_{i,j})\, k_{j-1}\,(k_j + k_{j-1})} \tag{4.53}$$

$$G_{i,j} = \frac{2\,k\,\mu_{n_{i-1/2,j}}\, B\{F(\psi_{i-1,j}, \psi_{i,j}, T_{i-1,j}, T_{i,j})\}}{q\, L(T_{i-1,j}, T_{i,j})\, h_{i-1}\,(h_i + h_{i-1})} \tag{4.54}$$

$$H_{i,j} = \frac{-2k}{q}\left(\frac{\mu_{n_{i,j-1/2}} B\{F(\psi_{i,j}, \psi_{i,j-1}, T_{i,j}, T_{i,j-1})\}}{L(T_{i,j-1}, T_{i,j})\, k_{j-1}\,(k_j + k_{j-1})} + \frac{\mu_{n_{i-1/2,j}} B\{F(\psi_{i,j}, \psi_{i-1,j}, T_{i,j}, T_{i-1,j})\}}{L(T_{i-1,j}, T_{i,j})\, h_{i-1}\,(h_i + h_{i-1})} + \frac{\mu_{n_{i+1/2,j}} B\{F(\psi_{i,j}, \psi_{i+1,j}, T_{i,j}, T_{i+1,j})\}}{L(T_{i,j}, T_{i+1,j})\, h_i\,(h_i + h_{i-1})} + \frac{\mu_{n_{i,j+1/2}} B\{F(\psi_{i,j}, \psi_{i,j+1}, T_{i,j}, T_{i,j+1})\}}{L(T_{i,j}, T_{i,j+1})\, k_j\,(k_j + k_{j-1})}\right) \tag{4.55}$$

$$I_{i,j} = \frac{2k\,\mu_{n_{i+1/2,j}} B\{F(\psi_{i+1,j}, \psi_{i,j}, T_{i+1,j}, T_{i,j})\}}{q\,L(T_{i,j}, T_{i+1,j})\, h_i\,(h_i + h_{i-1})} \tag{4.56}$$

$$L_{i,j} = \frac{2k\,\mu_{n_{i,j+1/2}} B\{F(\psi_{i,j+1}, \psi_{i,j}, T_{i,j+1}, T_{i,j})\}}{q\,L(T_{i,j}, T_{i,j+1})\, k_j\,(k_j + k_{j-1})} \tag{4.57}$$

A similar treatment of the hole continuity equation (4.51) results in the following equation in the discrete hole concentration values.

$$M_{i,j}\,p_{i,j-1} + W_{i,j}\,p_{i-1,j} + X_{i,j}\,p_{i,j} + S_{i,j}\,p_{i+1,j} + U_{i,j}\,p_{i,j+1} + R_{i,j} = 0 \tag{4.58}$$

where,

$$M_{i,j} = \frac{-2k\,\mu_{p_{i,j-1/2}} B\{F(\psi_{i,j}, \psi_{i,j-1}, T_{i,j-1}, T_{i,j})\}}{q\,L(T_{i,j-1}, T_{i,j})\, k_{j-1}\,(k_j + k_{j-1})} \tag{4.59}$$

$$W_{i,j} = \frac{-2k\,\mu_{p_{i-1/2,j}} B\{F(\psi_{i,j}, \psi_{i-1,j}, T_{i-1,j}, T_{i,j})\}}{q\,L(T_{i-1,j}, T_{i,j})\, h_{i-1}\,(h_i + h_{i-1})} \tag{4.60}$$

$$X_{i,j} = \frac{2k}{q}\left(\frac{\mu_{p_{i,j-1/2}} B\{F(\psi_{i,j-1}, \psi_{i,j}, T_{i,j}, T_{i,j-1})\}}{L(T_{i,j-1}, T_{i,j})\, k_{j-1}\,(k_j + k_{j-1})} + \frac{\mu_{p_{i-1/2,j}} B\{F(\psi_{i-1,j}, \psi_{i,j}, T_{i,j}, T_{i-1,j})\}}{L(T_{i-1,j}, T_{i,j})\, h_{i-1}\,(h_i + h_{i-1})} + \frac{\mu_{p_{i+1/2,j}} B\{F(\psi_{i+1,j}, \psi_{i,j}, T_{i,j}, T_{i+1,j})\}}{L(T_{i,j}, T_{i+1,j})\, h_i\,(h_i + h_{i-1})} + \frac{\mu_{p_{i,j+1/2}} B\{F(\psi_{i,j+1}, \psi_{i,j}, T_{i,j}, T_{i,j+1})\}}{L(T_{i,j}, T_{i,j+1})\, k_j\,(k_j + k_{j-1})}\right) \tag{4.61}$$

$$S_{i,j} = \frac{-2k\,\mu_{p_{i+1/2,j}} B\{F(\psi_{i,j}, \psi_{i+1,j}, T_{i+1,j}, T_{i,j})\}}{q\,L(T_{i,j}, T_{i+1,j})\, h_i\,(h_i + h_{i-1})} \tag{4.62}$$

$$U_{i,j} = \frac{-2\,k\,\mu_{p_{i,j+1/2}}\,B\{F(\psi_{i,j},\,\psi_{i,j+1},\,T_{i,j+1},\,T_{i,j})\}}{q\,L(T_{i,j},\,T_{i,j+1})\,k_j\,(k_j + k_{j-1})} \tag{4.63}$$

The discrete continuity equations have a local truncation error that is dependent upon the local variations in potential, temperature and current density. It is, therefore, desirable to increase the mesh density in regions where these quantities are expected to be significant. Since the absolute value of the bracketed term in equation (4.42) never exceeds 0.5, then the truncation error must obey the following inequality.

$$\begin{aligned} T_{n_{i,j}} < \; & \frac{h_i}{h_i + h_{i-1}} \left| \frac{\partial J_n}{\partial x} \bigg|_{i+1/2,j} \right| + \frac{h_{i-1}}{h_i + h_{i-1}} \left| \frac{\partial J_n}{\partial x} \bigg|_{i-1/2,j} \right| \\ & + \frac{k_j}{k_j + k_{j-1}} \left| \frac{\partial J_n}{\partial y} \bigg|_{i,j+1/2} \right| + \frac{k_{j-1}}{k_j + k_{j-1}} \left| \frac{\partial J_n}{\partial y} \bigg|_{i,j-1/2} \right| \end{aligned} \tag{4.64}$$

4.1.1.3 Discretization of the Heat Flow Equation.

The heat flow equation is treated in a similar way as the Poisson equation. Firstly, the 'div' and 'grad' operators in equation (4.4) are expanded to give:

$$\frac{\partial}{\partial x}\left(K(T)\frac{\partial T}{\partial x}\right) + \frac{\partial}{\partial y}\left(K(T)\frac{\partial T}{\partial y}\right) + Q = 0 \tag{4.65}$$

Initially, consideration will be given to the first term in this equation. The outer, first order partial derivative is replaced using:

$$\frac{\partial u}{\partial x}\bigg|_{i,j} = \frac{u_{i+1/2,j} - u_{i-1/2,j}}{0.5\,(h_i + h_{i-1})} + \frac{h_{i-1} - h_i}{4}\,\frac{\partial^2 u}{\partial x^2}\bigg|_{i,j} \tag{4.66}$$

This gives:

$$\begin{aligned} \frac{\partial}{\partial x}\left(K(T)\frac{\partial T}{\partial x}\right)\bigg|_{i,j} = \; & \frac{K_{i+1/2,j}\dfrac{\partial T}{\partial x}\bigg|_{i+1/2,j} - K_{i-1/2,j}\dfrac{\partial T}{\partial x}\bigg|_{i-1/2,j}}{0.5\,(h_i + h_{i-1})} \\ & + \frac{h_{i-1} - h_i}{4}\,\frac{\partial^2}{\partial x^2}\left(K(T)\frac{\partial T}{\partial x}\right)\bigg|_{i,j} \end{aligned} \tag{4.67}$$

The mid-interval derivatives with respect to temperature are given by:

$$\frac{\partial T}{\partial x}\bigg|_{i+1/2,j} = \frac{T_{i+1,j} - T_{i,j}}{h_i} - \frac{h_i^2}{24}\,\frac{\partial^3 T}{\partial x^3}\bigg|_{i+1/2,j} \tag{4.68}$$

$$\left.\frac{\partial T}{\partial x}\right|_{i-1/2,j} = \frac{T_{i,j} - T_{i-1,j}}{h_{i-1}} - \frac{h_{i-1}^2}{24}\left.\frac{\partial^3 T}{\partial x^3}\right|_{i-1/2,j} \tag{4.69}$$

Substitution into equation (4.67) results in:

$$\left.\frac{\partial}{\partial x}\left(K(T)\frac{\partial T}{\partial x}\right)\right|_{i,j} = \frac{K_{i+1/2,j}\dfrac{(T_{i+1,j} - T_{i,j})}{h_i} - K_{i-1/2,j}\dfrac{(T_{i,j} - T_{i-1,j})}{h_{i-1}}}{0.5\,(h_i + h_{i-1})} + T_{Tx_{i,j}} \tag{4.70}$$

This relation is then substituted into equation (4.65) together with the corresponding relation for the partial derivatives in the y-direction. Upon collection of all the terms in the discrete temperatures, the required result is obtained as follows.

$$T1_{i,j}\,T_{i,j-1} + T2_{i,j}\,T_{i-1,j} + T3_{i,j}\,T_{i,j} + T4_{i,j}\,T_{i+1,j} + T5_{i,j}\,T_{i,j+1} + Q_{i,j} = 0 \tag{4.71}$$

where the coefficients are given by:

$$T1_{i,j} = \frac{2\,K_{i,j-1/2}}{k_{j-1}(k_j + k_{j-1})} \tag{4.72}$$

$$T2_{i,j} = \frac{2\,K_{i-1/2,j}}{h_{i-1}(h_i + h_{i-1})} \tag{4.73}$$

$$T4_{i,j} = \frac{2\,K_{i+1/2,j}}{h_i(h_i + h_{i-1})} \tag{4.74}$$

$$T5_{i,j} = \frac{2\,K_{i,j+1/2}}{k_j(k_j + k_{j-1})} \tag{4.75}$$

$$T3_{i,j} = -\,(T1_{i,j} + T2_{i,j} + T4_{i,j} + T5_{i,j}) \tag{4.76}$$

The local truncation error is given by:

$$\begin{aligned} T_{T_{i,j}} = {} & \frac{K_{i-1/2,j}\,h_{i-1}^2}{12\,(h_i + h_{i-1})}\left.\frac{\partial^3 T}{\partial x^3}\right|_{i-1/2,j} - \frac{K_{i+1/2,j}\,h_i^2}{12\,(h_i + h_{i-1})}\left.\frac{\partial^3 T}{\partial x^3}\right|_{i+1/2,j} \\ & + \frac{h_{i-1} - h_i}{4}\left.\frac{\partial^2}{\partial x^2}\left(K(T)\frac{\partial T}{\partial x}\right)\right|_{i,j} + \frac{K_{i,j-1/2}\,k_{j-1}^2}{12\,(k_j + k_{j-1})}\left.\frac{\partial^3 T}{\partial y^3}\right|_{i,j-1/2} \\ & - \frac{K_{i,j+1/2}\,k_j^2}{12\,(k_j + k_{j-1})}\left.\frac{\partial^3 T}{\partial y^3}\right|_{i,j+1/2} + \frac{k_{j-1} - k_j}{4}\left.\frac{\partial^2}{\partial y^2}\left(K(T)\frac{\partial T}{\partial y}\right)\right|_{i,j} \end{aligned} \tag{4.77}$$

The truncation of error of the heat flow equation is of first order in the mesh spacing multiplied with the third order temperature derivative. This is true regardless of whether the mesh points are uniformly or non-uniformly spaced, which is in contrast to the case for the Poisson equation (cf. (4.11), (4.13)) and may be attributed to the temperature dependent thermal conductivity.

4.1.1.4 Discretization of the Boundary Conditions.

Having treated the governing equations at all the inner mesh points, attention will now turned to the points which coincide with the boundaries of the domain. Firstly the Dirichlet boundary condition is considered. It is a simple matter to obtain the discrete form of this condition since only values at the central node (i,j) are involved. Equations (2.77), (2.78) and (2.79) are simply written at each node as follows.

$$\psi_{i,j} = \psi_{A_{i,j}} + \psi_{b_{i,j}} \tag{4.78}$$

if $N_{i,j} > 0$ then:

$$n_{i,j} = \frac{\sqrt{N_{i,j}^2 + 4\, n_{ie_{i,j}}^2} + N_{i,j}}{2} \tag{4.79}$$

$$p_{i,j} = \frac{n_{ie_{i,j}}^2}{n_{i,j}} \tag{4.80}$$

else:

$$p_{i,j} = \frac{\sqrt{N_{i,j}^2 + 4\, n_{ie_{i,j}}^2} - N_{i,j}}{2} \tag{4.81}$$

$$n_{i,j} = \frac{n_{ie_{i,j}}^2}{p_{i,j}} \tag{4.82}$$

This technique to obtain the discrete electron and hole concentrations is preferable to directly applying (2.77) and (2.78) to each node as it avoids any inherent problems associated with numerical round off and cancellation. For the example shown in fig. (4.1) the built-in potential and applied potential at each node of the various contacts, assuming an *n-p-n* transistor are as follows.

$$\text{Emitter } (1 \leq i \leq 10,\, 4 \leq j \leq 6),\quad \psi_{A_{i,j}} = V_e,\quad \psi_{b_{i,j}} = \frac{kT_{i,6}}{q} \log_e\left(\frac{n_{i,6}}{n_{i_{i,6}}}\right) - \delta E_{C_{i,6}} \tag{4.83}$$

Base $(17 \leq i \leq 22, 4 \leq j \leq 6)$, $\psi_{A_{i,j}} = V_b$, $\psi_{b_{i,j}} = -\frac{kT_{i,6}}{q} \log_e\left(\frac{p_{i,6}}{n_{i_{i,6}}}\right) + \delta E_{V_{i,6}}$ (4.84)

Collector $(1 \leq i \leq 34, j = 26)$, $\psi_{A_{i,j}} = V_c$, $\psi_{b_{i,j}} = \frac{kT_{i,j}}{q} \log_e\left(\frac{n_{i,j}}{n_{i_{i,j}}}\right) - \delta E_{C_{i,j}}$ (4.85)

In the case of the emitter and base contacts the built-in potentials are calculated using the values of n, p, T, δE_C and δE_V at $j = 6$ since this row is the upper boundary for the solution of the current continuity and heat flow equations. The carrier concentrations at these contacts, given by equations (4.79) to (4.82), need only be calculated, therefore, at $j = 6$.

As an example of the discretisation of the Neumann boundary condition (eg.(2.82),(2.83) and (2.88)), the boundary A-D (ie. $i = 1$, $1 \leq j \leq 26$) will be considered, as all the governing equations must obey this condition along this boundary. In this instance the unit normal vector is parallel to the x axis and the equations evaluate to:

$$\left.\frac{\partial \psi}{\partial x}\right|_{i,j} = 0 \quad (4.86)$$

$$J_{nx_{i,j}} = 0 \quad (4.87)$$

$$J_{px_{i,j}} = 0 \quad (4.88)$$

$$\left.\frac{\partial T}{\partial x}\right|_{i,j} = 0 \quad (4.89)$$

For the case of the electrostatic potential discretization is most simply performed by rewriting the Taylor series (4.5) in terms of ψ and setting the first order derivative to zero (cf. (4.86)). The resulting equation for the second order differential becomes,

$$\left.\frac{\partial^2 \psi}{\partial x^2}\right|_{i,j} = \frac{2\,(\psi_{i+1,j} - \psi_{i,j})}{h_i^2} - \frac{h_i}{3} \left.\frac{\partial^3 \psi}{\partial x^3}\right|_{i,j} \quad (4.90)$$

This equation is then substituted into the Poisson equation, (4.9) in place of equation (4.10). The coefficients $B_{i,j}$ and $D_{i,j}$ of the discrete Poisson equation, (4.15) now become:

$$B_{i,j} = 0 \quad (4.91)$$

$$D_{i,j} = \frac{2}{h_i^2} \tag{4.92}$$

The coefficients $A_{i,j}$, $E_{i,j}$ and $C_{i,j}$ are given once again by (4.16), (4.19) and (4.20) respectively. It may be noted from equation (4.90) that the local truncation error of the discrete Neumann condition is of the same order of magnitude as that of the discrete Poisson equation at the inner mesh points. Thus, accuracy is maintained. Derivation of the discrete boundary conditions of the remaining equations is achieved by utilising an extremely effective technique called 'mirror imaging'. Firstly, the following interpolation formula, for a uniform mesh ($h_i = h_{i-1}$) is obtained from Taylor series'.

$$u_{i,j} = \frac{u_{i+1/2,j} + u_{i-1/2,j}}{2} - \frac{h_i^2}{8} \left. \frac{\partial^2 u}{\partial x^2} \right|_{i,j} \tag{4.93}$$

If the quantities J_{nx}, J_{py} and $\frac{\partial T}{\partial x}$ are successively substituted for u in (4.93), then, bearing in mind the boundary conditions, (4.87) to (4.89), the following relations are obtained:

$$J_{nx_{i-1/2,j}} = -J_{nx_{i+1/2,j}} + \frac{h_i^2}{4} \left. \frac{\partial^2 J_{nx}}{\partial x^2} \right|_{i,j} \tag{4.94}$$

$$J_{px_{i-1/2,j}} = -J_{px_{i+1/2,j}} + \frac{h_i^2}{4} \left. \frac{\partial^2 J_{px}}{\partial x^2} \right|_{i,j} \tag{4.95}$$

$$K_{i-1/2,j} \left. \frac{\partial T}{\partial x} \right|_{i-1/2,j} = -K_{i+1/2,j} \left. \frac{\partial T}{\partial x} \right|_{i+1/2,j} + \frac{h_i^2}{4} \frac{\partial^2}{\partial x^2} \left. \left(K(T) \frac{\partial T}{\partial x} \right) \right|_{i,j} \tag{4.96}$$

Although the quantities defined by (4.94) to (4.96) lie outside the solution domain, they can now be used to provide the discrete boundary conditions, which will have truncation errors of the same order as the corresponding discrete approximations at the inner points. If (4.94) is substituted into (4.50), (4.95) into (4.51) and (4.96) into (4.67), then the following approximations are obtained.

$$\frac{2}{q} \left(\frac{J_{nx_{i+1/2,j}}}{h_i} + \frac{J_{ny_{i,j+1/2}} - J_{ny_{i,j-1/2}}}{k_j + k_{j-1}} \right) - R_{i,j} = 0 \tag{4.97}$$

$$\frac{2}{q} \left(\frac{J_{px_{i+1/2,j}}}{h_i} + \frac{J_{py_{i,j+1/2}} - J_{py_{i,j-1/2}}}{k_j + k_{j-1}} \right) + R_{i,j} = 0 \tag{4.98}$$

$$2\left(\frac{K_{i+1/2,j} \frac{\partial T}{\partial x} \Big|_{i+1/2,j}}{h_i} + \frac{K_{i,j+1/2} \frac{\partial T}{\partial y} \Big|_{i,j+1/2} - K_{i,j-1/2} \frac{\partial T}{\partial x} \Big|_{i,j-1/2}}{k_j + k_{j-1}} \right) + Q_{i,j} = 0 \tag{4.99}$$

The desired result is then obtained by substituting the discrete approximations for inter-nodal current densities and temperature gradients into these equations. In the case of electron current current continuity, the coefficients $F_{i,j}$ and $L_{i,j}$ are unchanged from their values given by (4.53) and (4.57). However, the remaining coefficients now become:

$$G_{i,j} = 0 \tag{4.100}$$

$$H_{i,j} = \frac{-2\,k}{q} \left(\frac{\mu_{n_{i,j-1/2}} B\{F(\psi_{i,j}, \psi_{i,j-1}, T_{i,j}, T_{i,j-1})\}}{L(T_{i,j-1}, T_{i,j})\, k_{j-1}\,(k_j + k_{j-1})} + \frac{\mu_{n_{i+1/2,j}} B\{F(\psi_{i,j}, \psi_{i+1,j}, T_{i,j}, T_{i+1,j})\}}{L(T_{i,j}, T_{i+1,j})\, h_i^2} + \frac{\mu_{n_{i,j+1/2}} B\{F(\psi_{i,j}, \psi_{i,j+1}, T_{i,j}, T_{i,j+1})\}}{L(T_{i,j}, T_{i,j+1})\, k_j\,(k_j + k_{j-1})} \right) \tag{4.101}$$

$$I_{i,j} = \frac{2\,k\,\mu_{n_{i+1/2,j}} B\{F(\psi_{i+1,j}, \psi_{i,j}, T_{i+1,j}, T_{i,j})\}}{q\,L(T_{i,j}, T_{i+1,j})\, h_i^2} \tag{4.102}$$

For hole current continuity $M_{i,j}$ and $U_{i,j}$ are given by (4.59) and (4.63) respectively, and:

$$W_{i,j} = 0 \tag{4.103}$$

$$X_{i,j} = \frac{2\,k}{q} \left(\frac{\mu_{p_{i,j-1/2}} B\{F(\psi_{i,j-1}, \psi_{i,j}, T_{i,j}, T_{i,j-1})\}}{L(T_{i,j-1}, T_{i,j})\, k_{j-1}\,(k_j + k_{j-1})} + \frac{\mu_{p_{i+1/2,j}} B\{F(\psi_{i+1,j}, \psi_{i,j}, T_{i,j}, T_{i+1,j})\}}{L(T_{i,j}, T_{i+1,j})\, h_i^2} + \frac{\mu_{p_{i,j+1/2}} B\{F(\psi_{i,j+1}, \psi_{i,j}, T_{i,j}, T_{i,j+1})\}}{L(T_{i,j}, T_{i,j+1})\, k_j\,(k_j + k_{j-1})} \right) \tag{4.104}$$

$$S_{i,j} = \frac{-2\,k\,\mu_{i+1/2,j} B\{F(\psi_{i,j}, \psi_{i+1,j}, T_{i+1,j}, T_{i,j})\}}{q\,L(T_{i,j}, T_{i+1,j})\, h_i^2} \tag{4.105}$$

In the case of the discrete Heat Flow equation, (4.71) the coefficients $T1_{i,j}$, $T3_{i,j}$ and $T5_{i,j}$ are given once again by (4.72), (4.76) and (4.75) respectively. However, $T2_{i,j}$ and $T4_{i,j}$ now become:

$$T2_{i,j} = 0 \tag{4.106}$$

$$T4_{i,j} = \frac{2\,K_{i+1/2,j}}{h_i^2} \tag{4.107}$$

In order to obtain the discrete Poisson equation at the silicon-silicon dioxide interface G-L ($12 \le i \le 16, j = 6$) and K-N ($23 \le i \le 34, j = 6$), Gauss' law must be taken into account. For a mesh point at this interface the unit normal vector is parallel to the y-axis and Gauss' law (2.86) states that:

$$\varepsilon_{ox} \frac{\partial \psi}{\partial y}\Big|_{ox_{i,j}} - \varepsilon_{sil} \frac{\partial \psi}{\partial y}\Big|_{sil_{i,j}} = Q_{int} \tag{4.108}$$

By substituting ψ for u in the Taylor series (4.8) the following discrete representation for the potential gradient in the silicon dioxide is obtained.

$$\frac{\partial \psi}{\partial y}\Big|_{ox_{i,j}} = \frac{\psi_{i,j} - \psi_{i,j-1}}{k_{j-1}} + \frac{k_{j-1}}{2}\frac{\partial^2 \psi}{\partial y^2}\Big|_{ox_{i,j}} - \frac{k_{j-1}^2}{6}\frac{\partial^3 \psi}{\partial y^3}\Big|_{ox_{i,j}} \tag{4.109}$$

Similarly from series (4.7) the potential gradient in the silicon is obtained.

$$\frac{\partial \psi}{\partial y}\Big|_{sil_{i,j}} = \frac{\psi_{i,j+1} - \psi_{i,j}}{k_j} - \frac{k_j}{2}\frac{\partial^2 \psi}{\partial y^2}\Big|_{sil_{i,j}} - \frac{k_j^2}{6}\frac{\partial^3 \psi}{\partial y^3}\Big|_{sil_{i,j}} \tag{4.110}$$

The quantities defined by (4.109) and (4.110) are now substituted into (4.108) to give:

$$\begin{aligned} \varepsilon_{ox}k_{j-1}\frac{\partial^2 \psi}{\partial y^2}\Big|_{ox_{i,j}} + \varepsilon_{sil}k_j\frac{\partial^2 \psi}{\partial y^2}\Big|_{sil_{i,j}} &= 2\,Q_{int} + 2\varepsilon_{ox}\frac{(\psi_{i,j-1} - \psi_{i,j})}{k_{j-1}} + 2\varepsilon_{sil}\frac{(\psi_{i,j+1} - \psi_{i,j})}{k_j} \\ &+ \frac{\varepsilon_{ox}k_{j-1}^2}{3}\frac{\partial^3 \psi}{\partial y^3}\Big|_{ox_{i,j}} - \frac{\varepsilon_{sil}k_j^2}{3}\frac{\partial^3 \psi}{\partial y^3}\Big|_{sil_{i,j}} \end{aligned} \tag{4.111}$$

Discretization now proceeds by adding the following term to the left and right hand sides of (4.111).

$$\varepsilon_{ox}k_{j-1}\frac{\partial^2 \psi}{\partial x^2}\Big|_{ox_{i,j}} + \varepsilon_{sil}k_j\frac{\partial^2 \psi}{\partial x^2}\Big|_{sil_{i,j}} \tag{4.112}$$

Remembering that the discrete version of the second derivative with respect to the x-axis is given by (4.10) then the following result is obtained.

$$\varepsilon_{ox}\, k_{j-1}\left(\frac{\partial^2 \psi}{\partial x^2}\bigg|_{ox_{i,j}} + \frac{\partial^2 \psi}{\partial y^2}\bigg|_{ox_{i,j}} \right) + \varepsilon_{sil}\, k_j\left(\frac{\partial^2 \psi}{\partial x^2}\bigg|_{sil_{i,j}} + \frac{\partial^2 \psi}{\partial y^2}\bigg|_{sil_{i,j}} \right)$$

$$= \left(\varepsilon_{ox}\, k_{j-1} + \varepsilon_{sil}\, k_j\right)\left(\frac{\frac{\psi_{i+1,j} - \psi_{i,j}}{h_i} - \frac{\psi_{i,j} - \psi_{i-1,j}}{h_{i-1}}}{0.5\,(h_{i-1} + h_i)} \right) \tag{4.113}$$

$$+ \, 2\, Q_{int} + 2\, \varepsilon_{ox}\, \frac{(\psi_{i,j-1} - \psi_{i,j})}{k_{j-1}} + 2\, \varepsilon_{sil}\, \frac{(\psi_{i,j+1} - \psi_{i,j})}{k_j}$$

$$+ \left(\varepsilon_{ox}\, k_{j-1} + \varepsilon_{sil}\, k_j\right) \frac{(h_{i-1} - h_i)}{3}\, \frac{\partial^3 \psi}{\partial x^3}\bigg|_{i,j} + \left(\frac{\varepsilon_{ox} k_{j-1}^2 - \varepsilon_{sil} k_j^2}{3} \right) \frac{\partial^3 \psi}{\partial y^3}\bigg|_{i,j}$$

In order to continue the Laplace and Poisson equations are now recalled. These are the governing equations of potential in the oxide and semiconductor regions, respectively.

$$\frac{\partial^2 \psi}{\partial x^2}\bigg|_{ox_{i,j}} + \frac{\partial^2 \psi}{\partial y^2}\bigg|_{ox_{i,j}} = 0 \tag{4.114}$$

$$\frac{\partial^2 \psi}{\partial x^2}\bigg|_{sil_{i,j}} + \frac{\partial^2 \psi}{\partial y^2}\bigg|_{sil_{i,j}} = -\frac{q}{\varepsilon_{sil}}\left(N_{D_{i,j}} - N_{A_{i,j}} + p_{i,j} - n_{i,j}\right) \tag{4.115}$$

The bracketted terms on the left hand side of (4.113) are now replaced with (4.114) and (4.115). After collecting all the terms in the discrete potentials, the resulting equation is found to be of the same form to that obtained at the points within the silicon (cf. (4.15)), and is given by:

$$A_{i,j}\, \psi_{i,j-1} + B_{i,j}\, \psi_{i-1,j} + C_{i,j}\, \psi_{i,j} + D_{i,j}\, \psi_{i+1,j} + E_{i,j}\, \psi_{i,j+1}$$
$$+ \frac{q}{\varepsilon_{sil}}\left(N_{D_{i,j}} - N_{A_{i,j}} - n_{i,j} + p_{i,j} + \frac{2\, Q_{int}}{q\, k_j} \right) = 0 \tag{4.116}$$

The coefficients are given by:

$$A_{i,j} = \frac{2\, \varepsilon_{ox}}{\varepsilon_{sil}\, k_j\, k_{j-1}} \tag{4.117}$$

$$B_{i,j} = \frac{2\left(\varepsilon_{ox}\, k_{j-1} + \varepsilon_{sil}\, k_j\right)}{\varepsilon_{sil}\, k_j\, h_{i-1}\, (h_{i-1} + h_i)} \tag{4.118}$$

$$D_{i,j} = \frac{2\left(\varepsilon_{ox} k_{j-1} + \varepsilon_{sil} k_j\right)}{\varepsilon_{sil} k_j h_i (h_{i-1} + h_i)} \tag{4.119}$$

$$E_{i,j} = \frac{2}{k_j^2} \tag{4.120}$$

$$C_{i,j} = -(A_{i,j} + B_{i,j} + D_{i,j} + E_{i,j}) \tag{4.121}$$

The discrete interface condition could have been obtained more simply by substituting (4.109) and (4.110) into (4.108). However, the truncation error resulting from this technique would contain second order derivatives of potential. This is extremely undesirable, because in obtaining the discrete Poisson equation it is the second order derivatives which must be resolved (cf. (4.10)). The above method has therefore been chosen in preference, as it can be seen from (4.113) that the associated truncation error is of first order in mesh spacing multiplied by the third order partial derivative. This is, once again, of the same order as the truncation error at the inner points.

The discrete equation at the air-oxide interface, F-I ($12 \le i \le 16$, $j = 4$) and J-M ($23 \le i \le 34$, $j = 4$), is obtained by simply comparing (4.108) with the corresponding equation for Gauss' law at the air-oxide interface (2.87), which is:

$$\varepsilon_o \left.\frac{\partial \psi}{\partial y}\right|_{air_{i,j}} - \varepsilon_{ox} \left.\frac{\partial \psi}{\partial y}\right|_{ox_{i,j}} = 0 \tag{4.122}$$

The coefficients $A_{i,j}$ to $E_{i,j}$ are, therefore, obtained by substituting ε_o for ε_{ox} and ε_{ox} for ε_{sil} in (4.117) to (4.120). Also, since no charges are assumed to be present at this interface the space charge term on the left hand side of (4.116) would be zero.

A completely analogous procedure can be applied to the heat flow equation at the silicon-header boundary (P-Q). In this case the heat flux continuity equation (2.93) must be obeyed. The resulting discrete equation is given, once again by (4.71), with the heat generation, Q set to zero. However, the coefficients are now given by:

$$T1_{i,j} = \frac{2\, K_{i,j-1/2}}{k_{j-1}\, k_j} \tag{4.123}$$

$$T2_{i,j} = \frac{2(k_{j-1}\, K_{i-1/2,j} + k_j\, K_H)}{k_j\, h_{i-1}(h_i + h_{i-1})} \tag{4.124}$$

$$T4_{i,j} = \frac{2(k_{j-1}\, K_{i+1/2,j} + k_j\, K_H)}{k_j\, h_i(h_i + h_{i-1})} \tag{4.125}$$

$$T5_{i,j} = \frac{2\, K_H}{k_j^2} \tag{4.126}$$

and $T3_{i,j}$ is once again given by (4.76)

The convective boundary condition for the heat flow equation will now be considered. The boundary R-Q will be taken as an example and for a node on this boundary equation (2.92) can be written:

$$K(T)\frac{\partial T}{\partial x}\bigg|_{i,j} = -h\,(T_{i,j} - T_{amb}) \tag{4.127}$$

Using (4.93) the term on the left hand side of (4.127) can be replaced as follows.

$$K(T)\frac{\partial T}{\partial x}\bigg|_{i,j} = \frac{K_{i+1/2,j}\frac{\partial T}{\partial x}\Big|_{i+1/2,j} + K_{i-1/2,j}\frac{\partial T}{\partial x}\Big|_{i-1/2,j}}{2} \tag{4.128}$$

The term which lies outside the solution domain can now be replaced with the following:

$$K_{i+1/2,j}\frac{\partial T}{\partial x}\bigg|_{i+1/2,j} = -K_{i-1/2,j}\frac{\partial T}{\partial x}\bigg|_{i-1/2,j} - 2\,h\,(T_{i,j} - T_{amb}) \tag{4.129}$$

This term can now be substituted into (4.67) in place of (4.68), resulting in:

$$\frac{\partial}{\partial x}\left(K(T)\frac{\partial T}{\partial x}\right)\bigg|_{i,j} = -2\,K_{i-1/2,j}\frac{(T_{i,j} - T_{i-1,j})}{h_{i-1}^2} - 2\,h\,\frac{(T_{i,j} - T_{amb})}{h_{i-1}} \tag{4.130}$$

This relation is then substituted into the heat flow equation together with the corresponding term in the y direction, which remains unaltered from that of the inner points. Having gathered together all terms in the discrete temperatures the result becomes:

$$T1_{i,j}\,T_{i,j-1} + T2_{i,j}\,T_{i-1,j} + T3_{i,j}\,T_{i,j} + T5_{i,j}\,T_{i,j+1} + Q_{i,j} + \frac{2\,h\,T_{amb}}{h_{i-1}} = 0 \tag{4.131}$$

where:

$$T2_{i,j} = \frac{2\,K_{i-1/2,j}}{h_{i-1}^2} \tag{4.132}$$

$T1_{i,j}$ and $T5_{i,j}$ are given again by (4.72) and (4.75) and $T3_{i,j}$ is given by:

$$T3_{i,j} = -\left(T1_{i,j} + T2_{i,j} + T4_{i,j} + T5_{i,j} + \frac{2\,h}{h_{i-1}} \right) \tag{4.133}$$

Although the higher order terms have been omitted from this analysis the truncation error of (4.131) is of the same order as the truncation error at the inner mesh points.

4.1.2 Cylindrical Co-ordinates.

Devices that exhibit cylindrical symmetry are most efficiently modelled using a cylindrical co-ordinate system. Such devices may be fabricated by employing circular or annular masking techniques. The three cylindrical co-ordinates are radius r, displacement y and rotation θ. If the device to be modelled is assumed to be perfectly symmetrical about its centre point then the co-ordinate θ becomes redundant and can be omitted. Modelling in all three space dimensions is, therefore, achieved using only two co-ordinates (r,y). Thus, if only two orthogonal space co-ordinates are considered, then models for devices that can be assumed to be cylindrically symmetrical are inherently more accurate than models for devices without this symmetry. It will be seen that only a small amount of additional effort is required to discretize the equations on a cylindrical mesh, which makes modelling of cylindrical devices very efficient indeed.

4.1.2.1 Discretization of the Poisson Equation.

Expanding the 'div' and 'grad' operators in the Poisson equation gives [4.6]:

$$\frac{\partial^2 \psi}{\partial r^2} + \frac{\partial^2 \psi}{\partial y^2} + \frac{1}{r}\frac{\partial \psi}{\partial r} + \frac{q}{\varepsilon_{sil}}(N_D - N_A + p - n) = 0 \tag{4.134}$$

The x axis of the cartesian co-ordinate system must now be considered to be the radial axis of the cylindrical co-ordinate system, so that the h values now represent the mesh spacings in the radial direction. Having redefined this axis the only additional term that must be considered is the third term in (4.134). The discrete version of this term is obtained by eliminating the second order terms in the Taylor series' (4.5) and (4.6).

$$\frac{1}{r_i}\frac{\partial\psi}{\partial r}\bigg|_{i,j} = \frac{h_{i-1}}{r_i\, h_i(h_i + h_{i-1})}(\psi_{i+1,j} - \psi_{i,j}) - \frac{h_i}{r_i\, h_{i-1}(h_i + h_{i-1})}(\psi_{i-1,j} - \psi_{i,j}) - \frac{h_i\, h_{i-1}}{6r_i}\frac{\partial^3\psi}{\partial r^3}\bigg|_{i,j} \tag{4.135}$$

The truncation error is again of the same order as that for the discrete versions of the second order derivatives. For cylindrical co-ordinates, therefore, the coefficients $B_{i,j}$ given by (4.17) must be multiplied by a factor $M1_i$ and the coefficients $D_{i,j}$ given by (4.18) must be multiplied by a factor $M2_i$, where:

$$M1_i = \left(1 - \frac{h_i}{2r_i}\right) \tag{4.136}$$

$$M2_i = \left(1 + \frac{h_{i-1}}{2r_i}\right) \tag{4.137}$$

The remaining coefficients are then given by (4.16), (4.17) and (4.20).

4.1.2.2 Discretization of the Continuity Equations.

In this case the only alteration that is required is the addition of an extra term in the divergence operator of the continuity equations (4.2) and (4.3) which now become [4.7]:

$$\frac{1}{q}\left(\frac{\partial J_{nr}}{\partial r} + \frac{\partial J_{ny}}{\partial y} + \frac{1}{r}J_{nr}\right) - R = 0 \tag{4.138}$$

$$\frac{1}{q}\left(\frac{\partial J_{pr}}{\partial r} + \frac{\partial J_{py}}{\partial y} + \frac{1}{r}J_{pr}\right) + R = 0 \tag{4.139}$$

A discrete version of the extra term is found by eliminating the first order derivatives from the Taylor series' (4.5) and (4.6).

$$\frac{1}{r}J_{nr}\bigg|_{i,j} = \frac{h_{i-1}}{r_i\, h_i(h_i + h_{i-1})}J_{nr}\bigg|_{i+1/2,j} + \frac{h_i}{r_i\, h_{i-1}(h_i + h_{i-1})}J_{nr}\bigg|_{i-1/2,j} \quad \frac{h_i\, h_{i-1}}{2r_i}\frac{\partial^2 J_{nr}}{\partial r^2}\bigg|_{i,j} \tag{4.140}$$

A similar equation may be written for hole current. Here, once again, the truncation error is of similar magnitude to that for the discrete partial derivatives of the current densities (cf. (4.48)). Upon substituting the discrete transport equations for the internodal current densities in (4.140), it becomes apparent that

the coefficients $G_{I,J}$ and $I_{I,J}$ of the discrete electron current continuity equation (4.52) need to be multiplied by $M1_I$ and $M2_I$ respectively. Similarly, in the case of the hole continuity equation (4.58) the coefficients $W_{I,J}$ and $S_{I,J}$ must also be multiplied by $M1_I$ and $M2_I$ respectively.

4.1.2.3 Discretization of the Heat Flow Equation.

The heat flow equation (4.4) may be rewritten for the case of cylindrical co-ordinates as follows.

$$\frac{\partial}{\partial r}\left(K(T)\frac{\partial T}{\partial r}\right)+\frac{\partial}{\partial y}\left(K(T)\frac{\partial T}{\partial y}\right)+\frac{1}{r}K(T)\frac{\partial T}{\partial r} + Q = 0 \qquad (4.141)$$

The discrete version of the extra term is calculated in a similar way to that for the Poisson equation (cf. (4.135)).

$$\frac{1}{r_I}\frac{\partial T}{\partial r}\bigg|_{I,J} = K_{I,J}\frac{h_{I-1}}{r_I h_I(h_I+h_{I-1})}(T_{I+1,J}-T_{I,J}) - K_{I,J}\frac{h_I}{r_I h_{I-1}(h_I+h_{I-1})}(T_{I-1,J}-T_{I,J}) - \frac{K_{I,J}\,h_I\,h_{I-1}}{6r_I}\frac{\partial^3\psi}{\partial r^3}\bigg|_{I,J} \qquad (4.142)$$

In this case the coefficients $T2_{I,J}$ and $T4_{I,J}$ of the discrete heat flow equation, given by (4.73) and (4.74) must be multiplied by the factors $MT2_{I,J}$ and $MT4_{I,J}$ respectively. These factors are given by:

$$MT2_{I,J} = \left(1 - K_{I,J}\frac{h_I}{2\,K_{I-1/2,J}\,r_I}\right) \qquad (4.143)$$

$$MT4_{I,J} = \left(1 + K_{I,J}\frac{h_{I-1}}{2\,K_{I+1/2,J}\,r_I}\right) \qquad (4.144)$$

4.2 Discretization of the Dynamic Equations.

In order to accurately simulate transient device operation the full dynamic semiconductor equations must be solved. Such a solution is desired for the case where the boundary conditions for electrostatic potential at the contacts are time variant. The dynamic equations may be written in shorthand form as follows.

$$F_\psi = 0 \qquad (4.145)$$

$$F_n - \frac{\partial n}{\partial t} = 0 \qquad (4.146)$$

$$F_p + \frac{\partial p}{\partial t} = 0 \qquad (4.147)$$

$$F_T + \rho c \frac{\partial T}{\partial t} = 0 \qquad (4.148)$$

Here, F_ψ, F_n, F_p and F_T have been used to represent the set of algebraic equations resulting from the spatial discretization of the Poisson, electron current continuity, hole current continuity and heat flow equations, respectively. In this section time discretization, only, will be considered. The following notation will be used to represent a single discrete time interval.

$$d_m = t_{m+1} - t_m \qquad (4.149)$$

A number of different schemes are available for time discretization of the parabolic systems (4.139), (4.140) and (4.141), the simplest of which is the fully explicit (forward Euler) method, which may be written:

$$F_\psi(\psi_{m+1}, n_m, p_m) = 0 \qquad (4.150)$$

$$F_n(\psi_{m+1}, n_m, p_m, T_m) - \frac{n_{m+1} - n_m}{d_m} = 0 \qquad (4.151)$$

$$F_p(\psi_{m+1}, n_{m+1}, p_m, T_m) + \frac{p_{m+1} - p_m}{d_m} = 0 \qquad (4.152)$$

$$F_T(T_m, Q_{m+1}) - \rho c \frac{T_{m+1} - T_m}{d_m} = 0 \qquad (4.153)$$

Each set is solved in succession at each point in time using the 'best available' values for the dependent variables. It is a simple matter to calculate the variables at time $m+1$. This being achieved by the solution of the linear system (4.150) for ψ_{m+1} followed by the calculation of n_{m+1}, p_{m+1} and finally T_{m+1} from (4.151), (4.152) and (4.153) respectively, by simple substitution.

The fully explicit scheme is extremely attractive, but unfortunately it possess a serious drawback in that the size of the time step d_m must be made very small in order to guarantee stability of the numerical solution [4.8]. Stability is ensured provided that the following inequality is obeyed.

$$d_m \leq \frac{(h_{\min}^2 + k_{\min}^2)}{2} \qquad (4.154)$$

This restriction is in general so severe that it renders the fully explicit method unacceptable for practical purposes. Several semi-implicit time

discretization schemes have been presented by Mock [4.9] [4.10], but a number of intrinsic problems has meant limited success. By far the most secure scheme is the fully implicit (backward Euler) method. For this scheme the discrete equations read:

$$F_{\psi}(\psi_{m+1}, n_{m+1}, p_{m+1}) = 0 \tag{4.155}$$

$$F_{n}(\psi_{m+1}, n_{m+1}, p_{m+1}, T_{m+1}) - \frac{n_{m+1} - n_m}{d_m} = 0 \tag{4.156}$$

$$F_{p}(\psi_{m+1}, n_{m+1}, p_{m+1}, T_{m+1}) + \frac{p_{m+1} - p_m}{d_m} = 0 \tag{4.157}$$

$$F_{T}(T_{m+1}, Q_{m+1}) - \rho\, c\, \frac{T_{m+1} - T_m}{d_m} = 0 \tag{4.158}$$

This method is known to be unconditionally stable for any time step. The accuracy of this scheme as time advances is readily monitored by the local truncation error associated with the backward difference approximation, with can be obtained from the Taylor expansion.

$$T_{u_{m+1}} = \frac{d_m}{2} \left. \frac{\partial^2 u}{\partial t^2} \right|_{m+1} \tag{4.159}$$

where u denotes n, p or T. The main disadvantage of the fully implicit scheme is apparent from the large system of non-linear algebraic equations that has to be solved at every point in time. However, since this is the only feasible method available at present it has been employed for all transient simulations described in chapter 6.

An accurate value for the truncation error $T_{u_{m+1}}$ cannot be calculated having obtained a solution at time $m+1$ as a solution at time $m+2$ would also be required in order to calculate the second order derivative in (4.159) (cf. (4.10)). However, the value of the second order derivative at time m will in general provide a close approximation to the value at time $m+1$, and this value can be calculated from an equation of the form of (4.10). Using this approximation the time step was chosen such that the following condition was obeyed at every node.

$$\frac{T_{u_{m+1}}}{\left(\dfrac{u_{m+1} - u_m}{d_m}\right)} \times 100 \simeq \frac{\dfrac{d_m}{2} \left. \dfrac{\partial^2 u}{\partial t^2} \right|_{m}}{\left(\dfrac{u_{m+1} - u_m}{d_m}\right)} \times 100 < 0.1\% \tag{4.160}$$

Thus, the truncation error is constrained to be less than 0.1% of the difference approximation to the first order time derivative. If this condition is not satisfied the time step is halved and the solution at time $m+1$ is recalculated. Otherwise the time step is successively increase by a factor of 10%. The minimum time step is 1 *ns* since in general silicon devices do not operate above 1 *GHz*.

For the sake of completeness the discrete equations for the transient case at the inner mesh points will now be written in full

$$A_{i,j}\,\psi_{i,j-1,m+1} + B_{i,j}\,\psi_{i-1,j,m+1} + C_{i,j}\,\psi_{i,j,m+1} + D_{i,j}\,\psi_{i+1,j,m+1} + E_{i,j}\,\psi_{i,j+1,m+1} + \frac{q}{\varepsilon_{sil}}\left(N_{i,j} - n_{i,j,m+1} + p_{i,j,m+1}\right) = 0 \tag{4.161}$$

$$F_{i,j,m+1}\,n_{i,j-1,m+1} + G_{i,j,m+1}\,n_{i-1,j,m+1} + H_{i,j,m+1}\,n_{i,j,m+1} + I_{i,j,m+1}\,n_{i+1,j,m+1} + L_{i,j,m+1}\,n_{i,j+1,m+1} - R_{i,j,m+1} - \frac{n_{i,j,m+1} - n_{i,j,m}}{d_m} = 0 \tag{4.162}$$

$$M_{i,j,m+1}\,p_{i,j-1,m+1} + W_{i,j,m+1}\,p_{i-1,j,m+1} + X_{i,j,m+1}\,p_{i,j,m+1} + S_{i,j,m+1}\,p_{i+1,j,m+1} + U_{i,j,m+1}\,p_{i,j+1,m+1} + R_{i,j,m+1} + \frac{p_{i,j,m+1} - p_{i,j,m}}{d_m} = 0 \tag{4.163}$$

$$T1_{i,j,m+1}\,T_{i,j-1,m+1} + T2_{i,j,m+1}\,T_{i-1,j,m+1} + T3_{i,j,m+1}\,T_{i,j,m+1} + T4_{i,j,m+1}\,T_{i+1,j,m+1} + T5_{i,j,m+1}\,T_{i,j+1,m+1} + Q_{i,j,m+1/2} - \rho\, c\, \frac{T_{i,j,m+1} - T_{i,j,m}}{d_m} = 0 \tag{4.164}$$

4.3 Mesh Generation.

From previous considerations it is apparent that the numerical solution is highly dependent upon the characteristics of the mesh. Unfortunately an optimised mesh cannot be designed prior to having obtained a solution since the solution itself is required in order to provide an estimate for the local truncation errors. This problem can be circumvented by making an initial estimate for the optimum mesh, calculating a solution on this mesh and then redesigning the mesh on the basis of the local truncation errors obtained. In this way the mesh is repeatedly improved until the desired accuracy is achieved. Such a procedure clearly involves considerable computational effort.

The initial mesh should be chosen so that the device geometry and doping profiles are adequately resolved. The mesh lines that define the interfaces between the various different regions of a particular geometry and those that define the boundaries of the simulation domain are mandatory and must remain fixed throughout the adaptive meshing sequence. For example, in Figure 14 the column $i = 11$ defines the edge of the emitter contact and also in this case the

emitter mask edge. A second example is row $j = 6$, which defines the $Si\text{-}SiO_2$ interface and also the points where the emitter and base metalisation comes into contact with the silicon. A relatively fine mesh has been chosen to resolve the diffused emitter and base impurity profiles. Also, mesh lines have been positioned as close as possible to the metallurgical junctions. Therefore, the mesh depicted in Figure 14 represents a good initial design.

Having obtained a solution on this mesh the local truncation errors can then be estimated using finite differences. Calculation of the truncation errors for the Poisson equation requires estimates for the third and fourth order derivatives with respect to potential (cf. (4.11) and (4.13)). The third order derivatives can be obtained by firstly calculating the second order derivatives from (4.10) and (4.12) and then taking the first order derivative of these values by inserting them into an equation similar to (4.135). In a similar manner the fourth order derivatives can be calculated by the double application of (4.10) and (4.12). The third and fourth order derivatives of potential are equivalent to the first and second order derivatives of space charge and, therefore, the truncation error will be largest near junctions or near the edges of depletion regions. The mesh spacing in these regions must be adjusted accordingly.

The local truncations errors for the current continuity equations can be obtained from equations of the form of (4.42), which represents one of four components for electron current. The values of the current density derivatives at the mid-points between nodes can be obtained by interpolating the results of (4.48) or (4.49). It is important to note from equation (4.64) that the truncation error of the discrete continuity equations is of the same order of magnitude as the difference approximation itself if a large potential difference exists between mesh points (cf. Figure 18). Consequently a fine mesh will be required in regions where $\operatorname{div}\vec{J}$ is large. In the example of the BJT of Figure 14 when the base-emitter junction is forward biased the vertical electron current density along this junction will peak sharply at the periphery of the emitter ($i = 11$) owing to the effects of emitter pinch (cf. chapter 6). The divergence of electron current will then be large in the base and collector regions directly beneath the emitter periphery. Thus, a fine mesh is required in these regions so that accurate simulation of current flow within the device is ensured.

The heat flow equation can be treated for truncation errors (4.77) using a similar method to that outlined above for the Poisson equation. The third order derivative of T in this instance is proportional to the first order derivative of heat generation, Q in steady state or to the difference between heat generation and storage if heating takes place (cf. (2.74)). A fine mesh should therefore be utilised where the net generation is rapidly variant.

When adding new mesh lines the solution has to be interpolated onto the new mesh points to provide an initial estimate for subsequent computation. An interpolation scheme which has been recommended is to use directly the difference scheme for the newly introduced mesh points with the previous solution at the 'old' mesh points being held fixed [4.11]. Further details on adaptive meshing can be found [4.11] and [4.12]

The above described adaptive meshing feature has not yet been implemented in the numerical model and is left for future consideration. The initial mesh which was chosen on the basis of device geometry and impurity profiles has also been assumed to be adequate for the definition of the solution. It must, therefore, be stated that the truncation error may be considerably large at certain locations within the solution domain and the simulation results could be affected. However, upon moving away from these regions the truncation error will decay and the effect on the overall solution is often relatively small [4.12]. As a result the calculated contact currents can be expected to be comparatively accurate. Having completed the discretization stage the techniques that have been used to solve the resulting large set of non-linear algebraic equations (4.161)-(4.164) will be presented in the following chapter.

References.

4.1 G. E. Forsythe and W. R. Wasow, *Finite Difference Methods for Partial Differential Equations*, John Wiley and Sons, New York, 1960.

4.2 O. C. Zienkiewicz, *The Finite Element Method*, McGraw-Hill, New York, 1977.

4.3 W. L. Engl, H. K. Dirks, "Numerical Device Simulation Guided by Physical Approaches," Proc. NASECODE I Conf., Boole Press, Dublin, pp. 65-93 (1979).

4.4 D. L. Scharfetter and H. K. Gummel, "Large-Signal Analysis of a Silicon Read Diode Oscillator," IEEE Trans. Electron Devices, **ED-16**, pp. 64-77 (1969).

4.5 C. C. McAndrew, K. Singhal and E. L. Heasell, "A Consistent Nonisothermal Extension of the Scharfetter-Gummel Stable Difference Approximation," IEEE Trans. Electron Dev. Lett., **EDL-6**, pp. 446-447 (1985).

4.6 S. Selberherr, A. Schütz and H. W. Pötzl, "MINIMOS - a Two-Dimensional MOS Transistor Analyser," IEEE Trans. Electron Devices, **ED-27**, pp. 1540-1550 (1980).

4.7 W. L. Engl, H. K. Dirks and B. Meinerzhagen, "Device Modeling," Proc. IEEE, **71**, pp. 10-33 (1983).

4.8 G. D. Smith, *Numerical Solution of Partial Differential Equations: Finite Difference Methods*, Clarendon Press, Oxford, 1978.

4.9 M. S. Mock, *Analysis of Mathematical Models of Semiconductor Devices*, Boole Press, Dublin, 1982.

4.10 M. S. Mock, "The Charge-Neutral Approximation and time Dependent Simulation," Proc. NASECODE I Conf., Boole Press, Dublin, pp. 120-135 (1979).

4.11 S. Selberherr, *Analysis and Simulation of Semiconductor Devices*, Springer-Verlag, Wien-New York, 1984.

4.12 P. A. Markowich, *The Stationary Semiconductor Device Equations*, Springer-Verlag, Wien-New York, 1986.

Chapter 5. The Solution of Large Systems of Non-Linear Algebraic Equations Arising from the Semiconductor Problem.

The main result of the preceding chapter was that discretization of the governing equations produces a large system of non-linear algebraic equations. This chapter is concerned with the solution of this system for the discrete dependent variables (ψ,n,p,T). For both static and transient problems the set of discrete equations may be described with the aid of the following notation.

$$\mathbf{F}(\mathbf{v}) = \mathbf{0} \tag{5.1}$$

$$\mathbf{F}(\mathbf{v}) = \begin{bmatrix} \mathbf{F}_\psi(\mathbf{v}) \\ \mathbf{F}_n(\mathbf{v}) \\ \mathbf{F}_p(\mathbf{v}) \\ \mathbf{F}_T(\mathbf{v}) \end{bmatrix} \tag{5.2}$$

$$\mathbf{v} = \begin{bmatrix} \psi \\ \mathbf{n} \\ \mathbf{p} \\ \mathbf{T} \end{bmatrix} \tag{5.3}$$

In these equations $\mathbf{F}$ is a vector function of rank four which itself consists of the vector functions $\mathbf{F}_\psi(\mathbf{v})$, $\mathbf{F}_n(\mathbf{v})$, $\mathbf{F}_p(\mathbf{v})$ and $\mathbf{F}_T(\mathbf{v})$ that are all of rank ($NX.NY$). These vector functions represent, in listed form, the discrete approximations to the Poisson equation, the continuity equations and the heat flow equation respectively. The vector of unknowns, $\mathbf{v}$ also consists of four vectors, which represent the discrete dependent variables. The total scalar rank of $\mathbf{F}$ and $\mathbf{v}$ is, therefore, ($4.NX.NY$).

The values of the elements of the vector function, $\mathbf{F}$ are dependent upon the values of the unknowns, $\mathbf{v}$ (cf. (4.52)) and so the problem is non-linear. In general such systems can only be solved by iterative techniques. By far the most widely used technique is Newton's method. Using this method the equations are

repeatedly linearised and solved until a solution of the form (5.1) is obtained. Newton's method will initially be considered for the case of a single unknown and these considerations will then be extended to the multi-dimensional case in a formal manner.

Newton's method can be applied to find the zero or root, x^r of a function, f(x) provided a close initial approximation is available and f(x) has a continuous first derivative. Let x^0 be an initial approximation to a root, x^r of function, f(x). The function is then approximated by its tangent at the point x^0; that is, the function f(x) is approximated by the linear polynomial:

$$\hat{f}(x) = f(x^0) + (x - x^0)\, f'(x^0) \tag{5.4}$$

and the next approximation to the root of f is determined by the root of this linear approximation, which is:

$$x^1 = x^0 - \frac{f(x^0)}{f'(x^0)} \tag{5.5}$$

Repetition of the above step gives rise to the general formula for Newton's method.

$$x^{k+1} = x^k - \frac{f(x^k)}{f'(x^k)} \tag{5.6}$$

A graphical representation of Newton's method is given in Figure 19. If the initial approximation, x^0 is not sufficiently close to the solution this method may diverge. However, provided the x^k values are sufficiently close to x^r then 'quadratic convergence' is obtained [5.1], that is:

$$\|x^{k+1} - x^r\| \leq \alpha\, \|x^k - x^r\|^2 \tag{5.7}$$

This technique can be immediately extended to the multi-dimensional case given by:

$$\mathbf{v}^{k+1} = \mathbf{v}^k - \mathbf{F}'(\mathbf{v}^k)^{-1}\, \mathbf{F}(\mathbf{v}^k) \tag{5.8}$$

The matrix $\mathbf{F}'(\mathbf{v}^k)$ is called the Jacobian matrix of $\mathbf{F}(\mathbf{v}^k)$. In order to avoid costly inversion of the Jacobian matrix equation (5.8) may be re-expressed as follows.

$$\mathbf{F}'(\mathbf{v}^k)\, \delta\mathbf{v}^k = -\mathbf{F}(\mathbf{v}^k) \tag{5.9}$$

where:

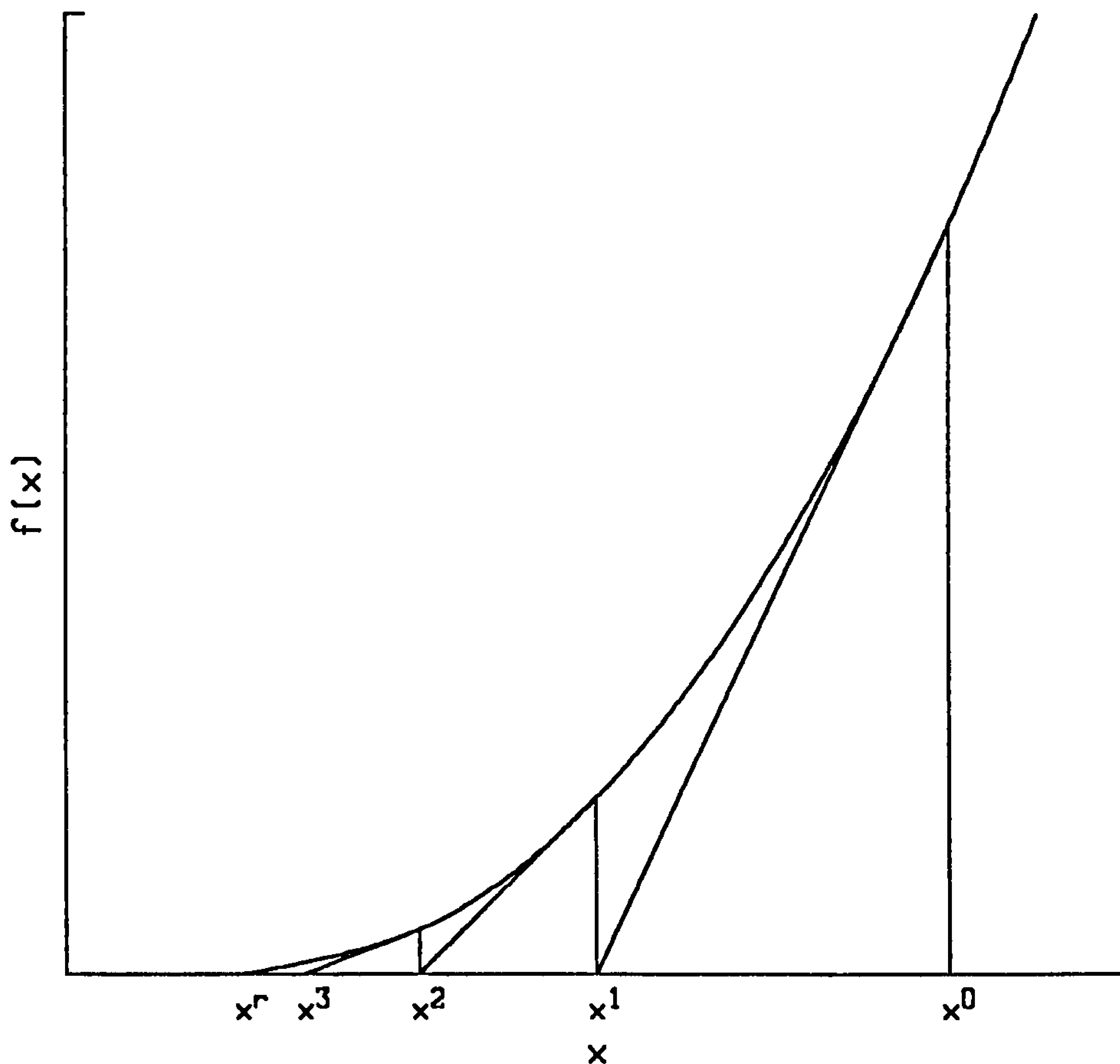


Figure 19. Illustration of Newton's Method.

$$\delta v^k = v^{k+1} - v^k \tag{5.10}$$

The linear system of equations (5.9) is relatively easy to solve as it allows the use of iterative procedures, which can take into account the sparse nature of the Jacobian. Newton's method has been used extensively to solve the semiconductor problem in the past. Two different solution procedures have proven to be particularly well suited to the solution of the semiconductor equations, namely the decoupled and coupled procedures. Both make use of Newton's method, but they differ in their implementation of it. These techniques will now be considered separately and then their relative merits will then be discussed.

5.1 Decoupled Solution Procedure.

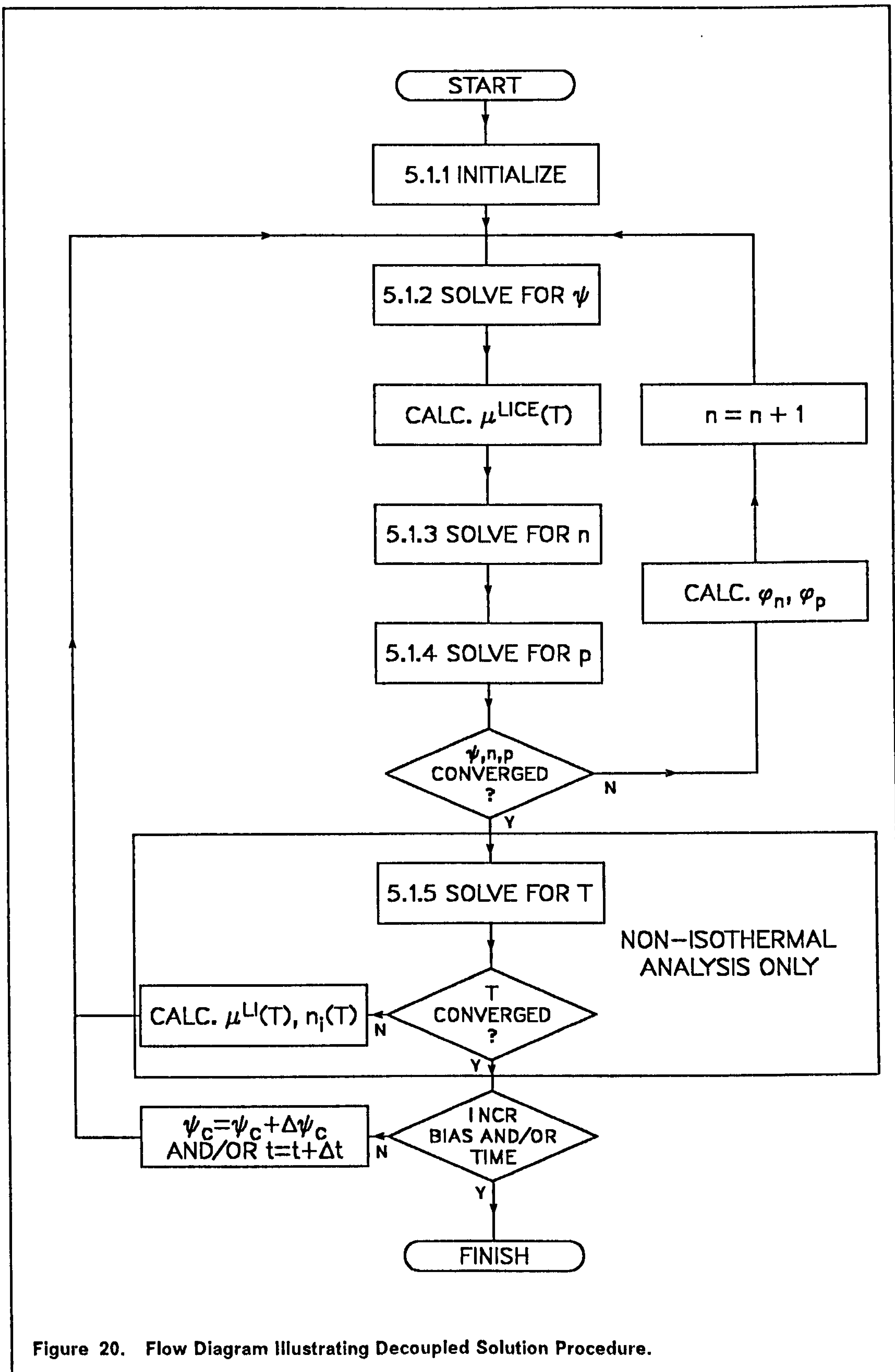


Figure 20. Flow Diagram Illustrating Decoupled Solution Procedure.

This procedure was first suggested by Gummel [5.2] in 1964 and is often referred to as Gummel's method. He suggested that each of the governing equations be solved individually and in a cyclical manner until convergence is achieved. This method will be described with the aid of the flow diagram of Figure 20. In the decoupled approach each of the governing equations is solved individually at each iteration. The non-isothermal problem is solved using two iteration levels, with (ψ,n,p) being solved in an inner iteration and temperature, T being solved in an outer iteration. This iteration procedure is continued until a consistent set of discrete dependent variables is obtained that satisfies expression (5.1).

5.1.1 Initialisation Procedure.

The initialization stage is required to provide a full and unambiguous definition of the problem. The device geometry within the solution domain must be specified together with the boundary conditions which are to be applied. Certain physical attributes such as impurity profiles, SRH lifetimes and mobility, μ^{LI} must also be assigned. A mesh must be designed for the subsequent discretization stage on the basis of device geometry, impurity profiles and any prior knowledge of device operation. Newton's method also requires initial estimates for the discrete dependent variables. The initial values of ψ, n and p at all mesh points in the semiconductor are calculated by assuming space charge neutrality and thermal equilibrium, as was assumed for the contact boundary condition (cf. (4.78)-(4.82)). The applied potential ψ_A throughout each region of the device (eg. base, emitter and collector) is initially assumed to be the same as that applied at their respective contacts (cf. (4.83)-(4.85)). The quasi-Fermi levels $\phi_{n_{i,j}}$ and $\phi_{p_{i,j}}$ are set equal to the applied potential, $\psi_{A_{i,j}}$ at each mesh point.

5.1.2 The Solution of Poisson's Equation.

In the case of Poisson's equation a solution of the following form is desired at each iteration (index n) shown in Figure 20.

$$\mathbf{F}_{\psi}(\mathbf{v}^{n}) = 0 \tag{5.11}$$

Recalling the discrete Poisson equation (4.15), it then becomes apparent that system (5.11) is non-linear as the electron and hole concentrations are dependent upon potential through the Maxwell-Boltzmann approximations (2.67) and (2.68). Substituting these approximations into (4.15) results in the following:

$$A_{i,j}\,\psi^n_{i,j-1} + B_{i,j}\,\psi^n_{i-1,j} + C_{i,j}\,\psi^n_{i,j} + D_{i,j}\,\psi^n_{i+1,j} + E_{i,j}\,\psi^n_{i,j+1}$$

$$+\frac{q}{\varepsilon_{sil}}\left\{N_{i,j} - n_{i_{i,j}}\exp\left(\frac{\psi^n_{i,j} - \phi^{n-1}_{n_{i,j}} + \delta E_{c_{i,j}}}{V_{T_{i,j}}}\right)\right. \tag{5.12}$$

$$\left. + n_{i_{i,j}}\exp\left(\frac{\phi^{n-1}_{p_{i,j}} - \psi^n_{i,j} + \delta E_{v_{i,j}}}{V_{T_{i,j}}}\right)\right\} = 0$$

It may be noted that the quasi-Fermi levels are fixed at their values from the previous iteration. It was suggested by Gummel [5.2] that a stable solution scheme could be obtained by introducing a second inner iteration (index k) to calculate the discrete potentials at each nth or outer iteration.

$$A_{i,j}\,(\psi^k_{i,j-1})^n + B_{i,j}\,(\psi^k_{i-1,j})^n + C_{i,j}\,(\psi^k_{i,j})^n + D_{i,j}\,(\psi^k_{i+1,j})^n + E_{i,j}\,(\psi^k_{i,j+1})^n$$

$$+\frac{q}{\varepsilon_{sil}}\left\{N_{i,j} - n_{i_{i,j}}\exp\left(\frac{(\psi^k_{i,j})^n - \phi^{n-1}_{n_{i,j}} + \delta E_{c_{i,j}}}{V_{T_{i,j}}}\right)\right. \tag{5.13}$$

$$\left. + n_{i_{i,j}}\exp\left(\frac{\phi^{n-1}_{p_{i,j}} - (\psi^k_{i,j})^n + \delta E_{v_{i,j}}}{V_{T_{i,j}}}\right)\right\} = 0$$

This equation is subsequently linearized by the application of Newton's method, that is:

$$\left(\frac{\partial \mathbf{F}_{\psi}(\mathbf{v}^k)}{\partial \psi}\,\delta\psi^k \;=\; -\,\mathbf{F}_{\psi}(\mathbf{v}^k)\right)^n \tag{5.14}$$

where:

$$\delta\psi^k \;=\; \psi^{k+1} - \psi^k \tag{5.15}$$

The corresponding linearized form of (5.13) is then obtained by applying (5.14) and is given by:

$$A_{i,j}\,(\delta\psi^k_{i,j-1})^n + B_{i,j}\,(\delta\psi^k_{i-1,j})^n + (\hat{C}^k_{i,j})^n\,(\delta\psi^k_{i,j})^n + D_{i,j}\,(\delta\psi^k_{i+1,j})^n + E_{i,j}\,(\delta\psi^k_{i,j+1})^n$$

$$= \;-\Big[A_{i,j}\,(\psi^k_{i,j-1})^n + B_{i,j}\,(\psi^k_{i-1,j})^n + C_{i,j}\,(\psi^k_{i,j})^n + D_{i,j}\,(\psi^k_{i+1,j})^n + E_{i,j}\,(\psi^k_{i,j+1})^n \tag{5.16}$$

$$+\frac{q}{\varepsilon_{sil}}\left\{N_{i,j} - (n^k_{i,j})^n + (p^k_{i,j})^n\right\}\Big]$$

In this equation $(n_{i,j}^{k})^{n}$ and $(p_{i,j}^{k})^{n}$ have been re-substituted for the Maxwell-Boltzmann terms in (5.13) and the modified coefficient, $(\hat{C}_{i,j}^{k})^{n}$ is given by:

$$(\hat{C}_{i,j}^{k})^{n} = C_{i,j} - \frac{q}{\varepsilon_{sil} V_{T_{i,j}}} \left\{ (p_{i,j}^{k})^{n} + (n_{i,j}^{k})^{n} \right\} \tag{5.17}$$

it may be noted that the values of the coefficients A, B, C, D and E are not related to the dependent variables (cf. (4.16)-(4.20)) and, therefore, they need only be calculated once at the initialisation stage. Also, in order that the Dirichlet boundary condition at the contacts is not violated it is necessary to ensure that all these coefficients plus the residuals, $\mathbf{F}_{\psi}$, are set to zero and the coefficient $\hat{C}$ is set to unity at mesh points which lie on the contacts. The actual solution of the linear system given by (5.16) for the values of the $\delta\psi$'s will be considered separately in section 5.1.6. Having calculated these values at the kth Newton iteration they are then added to the existing potentials according to (5.15). The values of the carrier concentrations are then recalculated from the updated potentials using the Maxwell-Boltzmann approximations and then the linear system (5.16) is subsequently resolved. This process is repeated until a self-consistent solution is obtained and convergence has been achieved. Since Maxwell-Boltzmann statistics are applied then the following condition must be adhered to.

$$\psi_{i,j} - \phi_{n_{i,j}} \leq \frac{E_{g_{i,j}}}{2} - \delta E_{c_{i,j}} \tag{5.18}$$

If the solution of the linear system returns a value of potential that violates this condition then the potential is amended to the following before recalculating the electron concentration values.

$$\psi_{i,j} = \frac{E_{g_{i,j}}}{2} - \delta E_{c_{i,j}} + \phi_{n_{i,j}} \tag{5.19}$$

Similarly if the condition:

$$\phi_{p_{i,j}} - \psi_{i,j} \leq \frac{E_{g_{i,j}}}{2} - \delta E_{v_{i,j}} \tag{5.20}$$

is violated then the potential is amended to the following before recalculating the hole concentrations.

$$\psi_{i,j} = -\frac{E_{g_{i,j}}}{2} + \delta E_{v_{i,j}} + \phi_{p_{i,j}} \tag{5.21}$$

These considerations do not only eliminate numerical overflow during computation but they also serve to improve the rate of convergence. The criterion for convergence is given by:

$$\max |\delta\psi_{i,j}^{k}| < \delta\psi_{\max} \quad \forall\, i,j \tag{5.22}$$

A value of $0.01 \times V_T(300K)$ *volts* was chosen for $\delta\psi_{\max}$, which ensures that the solution is sufficiently close to the root, ψ^r to be within the region of quadratic convergence, and furthermore, any further reduction in $\delta\psi_{\max}$ together with consistent reductions in the convergence criteria for electrons and holes (cf. sections 5.1.3 and 5.1.4) only produced a change of at most, 0.1% in any of the calculated contact currents. Finally, it must be pointed out that for the transient case the above procedure is repeated at each point in time (cf. (4.161)).

Before entering the routines to solve the continuity equations it is necessary to take into consideration the effects of carrier-carrier scattering (3.8) and carrier velocity saturation (3.14) on the electron and hole mobilities. The fields are calculated from the updated potentials arising from the solution of Poisson's equation at the nth outer iteration. The effects of carrier-carrier scattering are calculated using the latest concentration values also obtained from the solution of Poisson's equation.

5.1.3 Solution of Electron Current Continuity Equation.

Following a similar procedure to that for Poisson's equation the iteration in this case is defined by:

$$\left(\frac{\partial \mathbf{F}_n(\mathbf{v}^k)}{\partial \mathbf{n}} \delta \mathbf{n}^k = -\mathbf{F}_n(\mathbf{v}^k) \right)^n \tag{5.23}$$

In analogy with the Poisson equation the discrete electron continuity equation, (4.52) is recalled so that the constituent equation of the above system can be written as follows.

$$\begin{aligned} & F_{i,j}^n (\delta n_{i,j-1}^k)^n + G_{i,j}^n (\delta n_{i-1,j}^k)^n + (\hat{H}_{i,j}^k)^n (\delta n_{i,j}^k)^n + I_{i,j}^n (\delta n_{i+1,j}^k)^n + L_{i,j}^n (\delta n_{i,j+1}^k)^n \\ = & -\left[F_{i,j}^n (n_{i,j-1}^k)^n + G_{i,j}^n (n_{i-1,j}^k)^n + H_{i,j}^n (n_{i,j}^k)^n + I_{i,j}^n (n_{i+1,j}^k)^n + L_{i,j}^n (n_{i,j+1}^k)^n - (R_{i,j}^k)^n \right] \end{aligned} \tag{5.24}$$

where:

$$(\hat{H}_{i,j}^k)^n = H_{i,j}^n - \left(\frac{\partial R_{i,j}^k}{\partial n_{i,j}} \right)^n \tag{5.25}$$

the derivative of the recombination term, which includes SRH (3.30) and Auger recombination (3.33) is given by:

$$\frac{\partial R}{\partial n} = \frac{a\,p - b\,\tau_{po}}{a^2} + (c_n\,n + c_p\,p)\,p + b\,c_n \tag{5.26}$$

where:

$$a = \tau_{po}\,(n + n_{ie}) + \tau_{no}\,(p + n_{ie}) \tag{5.27}$$

$$b = n\,p - n_{ie}^2 \tag{5.28}$$

If any non-linearities associated with the carrier mobilities are ignored then the coefficients F, G, H, I and L can be considered to be independent of the value of electron concentration. Rather, they depend only upon the discrete potentials and temperatures. These coefficients need only be calculated once for each outer iteration, following the solution of Poisson's equation. As before care must be taken to comply with the Dirichlet condition at ohmic contacts and also in insulating regions, including any air gaps that may be present. The linearised continuity equation (5.24) is of the same form as the linearised Poisson equation, and the solution of these systems will be considered in section 5.1.6.

In the transient case a solution is sought at time $m+1$ by making use of knowledge of the concentrations at the previous or m^{th} point in time (cf.(4.156)). The modified coefficient $\hat{H}_{i,j}$ and the residual $F_{n_{i,j}}$ must be altered to take into account the rate of change of electron concentration.

$$\hat{H}_{i,j,m+1} = \hat{H}_{i,j} - \frac{1}{d_m} \tag{5.29}$$

$$F_{n_{i,j,m+1}} = F_{n_{i,j}} - \frac{n_{i,j,m+1} - n_{i,j,m}}{d_m} \tag{5.30}$$

A good initial guess for the electron concentration at t_{m+1} is provided by the value at the previous time, t_m. Once the δn values have been calculated at the kth inner iteration they are then added to the existing concentration values. The residuals, $F_{n_{i,j,m+1}}$ and modified coefficients $\hat{H}_{i,j,m+1}$ are then recalculated and the linear system is resolved for the next set of δn values. The criterion for convergence of this procedure is given by:

$$\max \left| \frac{\delta n_{i,j}}{n_{i,j}} \right| \times 100 < \delta n_{\max} \quad \forall\, i,j \tag{5.31}$$

The value of δn_{max} is chosen to be consistent with that of $\delta\psi_{max}$ via the Maxwell-Boltzmann approximation.

$$\delta n = n_i \exp\left(\frac{\psi + \delta\psi - \phi_n + \delta E_c}{V_T} \right) - n_i \exp\left(\frac{\psi - \phi_n + \delta E_c}{V_T} \right) \tag{5.32}$$

which reduces to:

$$\delta n = n \left\{ \exp\left(\frac{\delta\psi}{V_T} \right) - 1 \right\} \tag{5.33}$$

thus:

$$\delta n_{max} = \max \left| \frac{\delta n}{n} \right| \times 100 = \left\{ \exp\left(\frac{\delta\psi_{max}}{V_T} \right) - 1 \right\} \times 100 \tag{5.34}$$

Therefore, for $\delta\psi_{max} = 0.01 \times V_T$ the corresponding value for δn_{max} is 1%.

5.1.4 Solution of Hole Current Continuity Equation.

The procedure for hole is completely analogous to that for electrons, and the iteration is defined by:

$$\left(\frac{\partial \mathbf{F}_p(\mathbf{v}^k)}{\partial \mathbf{p}} \delta \mathbf{p}^k = -\mathbf{F}_p(\mathbf{v}^k) \right)^n \tag{5.35}$$

Recalling the discrete hole continuity equation (4.58), then the constituent equation is:

$$\begin{aligned} & M_{i,j}^n (\delta p_{i,j-1}^k)^n + W_{i,j}^n (\delta p_{i-1,j}^k)^n + (\hat{X}_{i,j}^k)^n (\delta n_{i,j}^k)^n + S_{i,j}^n (\delta p_{i+1,j}^k)^n + U_{i,j}^n (\delta p_{i,j+1}^k)^n \\ = & -\left[M_{i,j}^n (p_{i,j-1}^k)^n + W_{i,j}^n (p_{i-1,j}^k)^n + X_{i,j}^n (p_{i,j}^k)^n + S_{i,j}^n (p_{i+1,j}^k)^n + U_{i,j}^n (p_{i,j+1}^k)^n + (R_{i,j}^k)^n \right] \end{aligned} \tag{5.36}$$

where:

$$(\hat{X}_{i,j}^k)^n = X_{i,j}^n + \left(\frac{\partial R_{i,j}^k}{\partial p_{i,j}} \right)^n \tag{5.37}$$

The derivative of the recombination term is given by:

$$\frac{\partial R}{\partial p} = \frac{a\, n - b\, \tau_{no}}{a^2} + (c_n\, n + c_p\, p)\, n + b\, c_p \tag{5.38}$$

where a and b are defined by (5.27) and (5.28). In the transient case the following alterations must be made as before.

$$\hat{X}_{i,j,m+1} = \hat{X}_{i,j} + \frac{1}{d_m}$$
$$F_{p_{i,j,m+1}} = F_{p_{i,j}} + \frac{p_{i,j,m+1} - p_{i,j,m}}{d_m} \tag{5.39}$$

The δp values are calculated at each iteration until the percentage change in hole concentration at all mesh points is less than $\delta p_{max} = 1\%$.

The solution for the set (ψ,n,p) at the end of each outer iteration is considered to be converged if the maximum changes in potential and carrier concentrations at any node do not exceed the specified maximum allowable values previously defined by $\delta\psi_{max}$, δn_{max} and δp_{max} for the inner iterations. If this condition is not satisfied, however, the electron and hole quasi-Fermi levels are recalculated from the Maxwell-Boltzmann approximations using the newly updated carrier concentrations.

$$\phi^n_{n_{i,j}} = \psi^n_{i,j} + \delta E_{c_{i,j}} - V_{T_{i,j}} \log_e\left(\frac{n^n_{i,j}}{n_{i_{i,j}}}\right) \tag{5.40}$$

$$\phi^n_{p_{i,j}} = \psi^n_{i,j} - \delta E_{v_{i,j}} + V_{T_{i,j}} \log_e\left(\frac{p^n_{i,j}}{n_{i_{i,j}}}\right) \tag{5.41}$$

These updated quasi-Fermi levels are then used for solving the Poisson equation at the next outer iteration.

5.1.5 Solution of the Heat Flow Equation.

For problems that involve significant power dissipation within the device it may be necessary to solve the heat flow equation in order to take temperature rises into account. Before a solution to the discrete heat flow equation, (4.164) can be found the heat generated at each node must be calculated by the simple application of equation (3.48). In contrast to the previously outlined solution procedures Newton's method has not been employed for the solution of the discrete heat flow equation. A procedure which has proven to be much simpler and more effective is obtained by initially assuming equation (4.164) to be linear. The non-linearity arising from the temperature dependence of the thermal conductivity (cf. (3.45)) is initially neglected and the inter-nodal thermal conductivities are calculated from the temperatures arising from a previous

iteration. The coefficients $T1 - T5$ are then obtained from these thermal conductivity values. Having solved the resulting system, which is assumed to be linear the thermal conductivities are then re-evaluated from the newly obtained temperatures. The assumed linear system is subsequently resolved with the updated coefficients. This procedure was repeated until the maximum change made in a single iteration at all nodes is reduced to below 0.001 K.

5.1.6 The Solution of Linear Systems Arising from the Decoupled Procedure.

The solution of large systems of simultaneous equations can be achieved either by direct methods or by iterative methods. Direct methods solve the system of equations in a known number of arithmetic operations, and any errors incurred in the solution arise entirely from rounding errors introduced during computation. Basically these direct methods are elimination methods of which the best known examples are the systematic Gaussian elimination method and the triangular decomposition method [5.3]. In contrast, iterative methods seek to obtain a solution by the repeated application of a particular algorithm in order to make successive improvements to some initial guess at the solution. This cyclical procedure is repeated until a solution is obtained that is considered to be sufficiently close to the exact solution of the system. The direct approach is the most efficient method available for small sets of equations, but it is not for large sets. This method requires $2n^2$ arithmetic operations to solve n equations of the type being considered. However, due to the sparse (ie. large number of zero elements) nature of the Jacobian resulting from the five point difference scheme the most elementary iterative method available requires only $9n$ operations per iteration. Thus, for larger sets of equations ($n \gtrsim 100$) iterative methods become the most efficient. In addition to computational efficiency, iterative procedures possess other advantages over elimination methods. They require much less memory for storage of intermediate data and they are very easy to program. They are also usually more applicable to non-linear sets of equations than direct methods. This is because very few iterations are required near the root of the system when the changes being made are small. Because of these advantages iterative methods have been preferred in the past for solving the linear systems arising from the decoupled approach [5.4] [5.5]. Two such methods have been employed here; namely the Successive Line Over-Relaxation method (SLOR) [5.6] and the Strongly Implicit Procedure (SIP) [5.7]. The SLOR method proved to be

more efficient in solving the heat flow equation, whereas the electrical model was better suited to a solution by the SIP.

Each of the linear systems may be expressed in matrix notation as follows.

$$\mathbf{A}\,\mathbf{z} = \mathbf{b} \tag{5.42}$$

The constituent equation of this system has the following form.

$$a_{i,j}\, z_{i,j-1} + f_{i,j}\, z_{i-1,j} + c_{i,j}\, z_{i,j} + d_{i,j}\, z_{i+1,j} + e_{i,j}\, z_{i,j+1} = b_{i,j} \tag{5.43}$$

An example of such a system for a small 4×6 mesh is shown in Figure 21. The equations are ordered using natural ordering, that is from the top node to the bottom node of every column from left to right. Only the non-zero elements of matrix **A** are shown, which can be seen to have five non-zero diagonals. Iterative methods can ease the solution of such a system by taking advantage of its sparse nature.

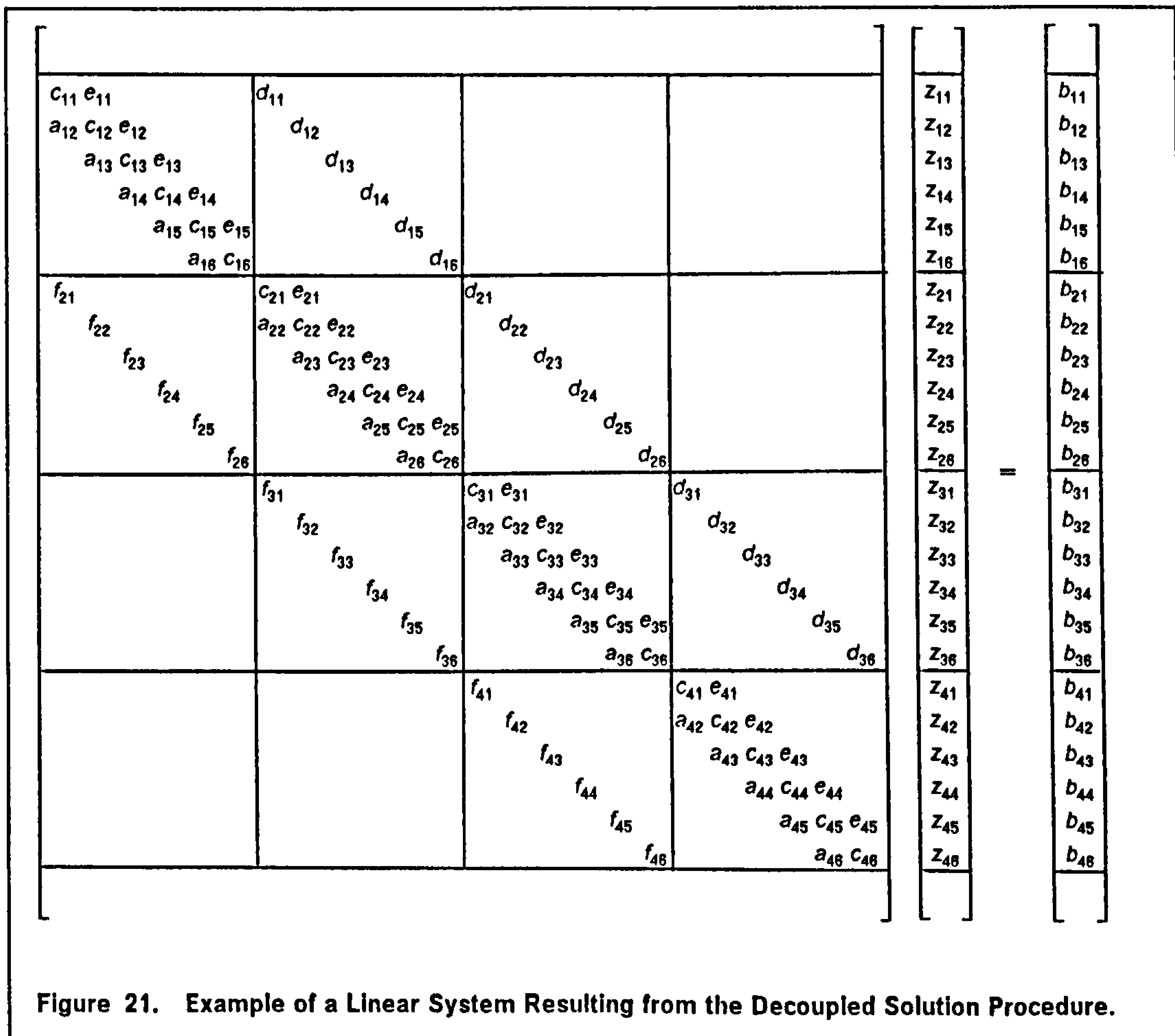


Figure 21. Example of a Linear System Resulting from the Decoupled Solution Procedure.

5.1.6.1 Successive Line Over-Relaxation (SLOR).

This method belongs to the family of block iterative methods in which groups of mesh points are treated as single units [5.6]. The values $z_{i,j}$ at each mesh point within the same group are modified simultaneously. This will involve the simultaneous solution of a sub-system of equations. Consequently individual components are implicitly defined in terms of other components in the same group. In the case of SLOR the unit is considered to be a mesh row or column. For the example shown in Figure 21 the system may be partitioned in the following manner.

$$\begin{bmatrix} \mathbf{D}_1 & \mathbf{E}_1 & & \\ \mathbf{C}_2 & \mathbf{D}_2 & \mathbf{E}_2 & \\ & \mathbf{C}_3 & \mathbf{D}_3 & \mathbf{E}_3 \\ & & \mathbf{C}_4 & \mathbf{D}_4 \end{bmatrix} \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \\ \mathbf{z}_3 \\ \mathbf{z}_4 \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \mathbf{b}_3 \\ \mathbf{b}_4 \end{bmatrix} \tag{5.44}$$

In this example columns have been taken as units and $\mathbf{C}_i$, $\mathbf{D}_i$ and $\mathbf{E}_i$ are $NY \times NY$ sub-matrices, $\mathbf{D}_i$ being tridiagonal and $\mathbf{C}_i$ and $\mathbf{E}_i$ being simple diagonal matrices. The block successive over-relaxation iterative method, which can now be rigorously applied to the system (5.42) takes the following form.

$$\mathbf{D}_i \hat{\mathbf{z}}_i^{m+1} = \mathbf{b}_i - \mathbf{C}_i \mathbf{z}_{i-1}^{m+1} - \mathbf{E}_i \mathbf{z}_{i+1}^{m} \tag{5.45}$$

and:

$$\mathbf{z}_i^{m+1} = \mathbf{z}_i^{m} + \omega (\hat{\mathbf{z}}_i^{m+1} - \mathbf{z}_i^{m}) \tag{5.46}$$

A single iteration (index m) consists of applying (5.45) followed by (5.46) to each column in succession, starting at column 1 and ending at column NX. Rather than taking the vector $\hat{\mathbf{z}}_i$ to be the solution at the ith column the SLOR method seeks to exaggerate the modifications made along each column by introducing an acceleration parameter, ω. This has been found to drastically reduce the number of iterations required for convergence. The optimum value of ω lies between 1 and 2. This technique is called over-relaxation and it is known to diverge for $\omega \geq 2$ [5.3]. In order to illustrate this technique equation (5.45) has been applied to column 2 of the example problem of Figure 21 and the result is given by:

$$\begin{bmatrix} c_{21} & e_{21} & & & & \\ a_{22} & c_{22} & e_{22} & & & \\ & a_{23} & c_{23} & e_{23} & & \\ & & a_{24} & c_{24} & e_{24} & \\ & & & a_{25} & c_{25} & e_{25} \\ & & & & a_{26} & c_{26} \end{bmatrix} \begin{bmatrix} \hat{z}_{21}^{m+1} \\ \hat{z}_{22}^{m+1} \\ \hat{z}_{23}^{m+1} \\ \hat{z}_{24}^{m+1} \\ \hat{z}_{25}^{m+1} \\ \hat{z}_{26}^{m+1} \end{bmatrix} = \begin{bmatrix} b_{21} - f_{21} z_{11}^{m+1} - d_{21} z_{31}^{m} \\ b_{22} - f_{22} z_{12}^{m+1} - d_{22} z_{32}^{m} \\ b_{21} - f_{23} z_{13}^{m+1} - d_{23} z_{33}^{m} \\ b_{21} - f_{24} z_{14}^{m+1} - d_{24} z_{34}^{m} \\ b_{21} - f_{25} z_{15}^{m+1} - d_{25} z_{35}^{m} \\ b_{21} - f_{26} z_{16}^{m+1} - d_{26} z_{36}^{m} \end{bmatrix} \qquad (5.47)$$

This result is of the general form:

$$\mathbf{D}_i \, \hat{\mathbf{z}}_i^{m+1} = \mathbf{v}_i \qquad (5.48)$$

This tridiagonal system can be solved by the application of the so called Thomas algorithm [5.8], which uses triangular decomposition. Initially the sub-matrix $\mathbf{D}_i$ is factorised into $\mathbf{D}_i = \mathbf{LU}$, where **L** and **U** are lower and upper triangular matrices respectively. They are both bidiagonal matrices with the matrix **L** having non-zero elements in the diagonals corresponding to the *a*- and *c*-diagonals of $\mathbf{D}_i$, and matrix **U** having non-zero elements in those diagonals corresponding to c and e with those corresponding to c being everywhere equal to unity. The **LU** product for the example problem is as ıollows.

$$\mathbf{LU} = \begin{bmatrix} c_1' & & & & & \\ a_2' & c_2' & & & & \\ & a_3' & c_3' & & & \\ & & a_4' & c_4' & & \\ & & & a_5' & c_5' & \\ & & & & a_6' & c_6' \end{bmatrix} \begin{bmatrix} 1 & e_1' & & & & \\ & 1 & e_2' & & & \\ & & 1 & e_3' & & \\ & & & 1 & e_4' & \\ & & & & 1 & e_5' \\ & & & & & 1 \end{bmatrix} \qquad (5.49)$$

The non-zero elements of **L** and **U** may be evaluated by equating the **LU** product to $\mathbf{D}_i$ and using forward substitution. The following is obtained for column *i*.

$$a_j' = a_{i,j} \qquad (5.50)$$

$$c_1' = c_{i,1}, \quad c_j' = c_{i,j} - a_{i,j}\, e_{j-1}' \quad (2 \leq j \leq NY) \qquad (5.51)$$

$$e_1' = \frac{e_{i,1}}{c_{i,1}}, \quad e_j' = \frac{e_{i,j}}{c_{i,j} - a_{i,j}\, e_{j-1}'} \quad (2 \leq j \leq NY - 1) \qquad (5.52)$$

Having factored the sub-matrix $\mathbf{D}_i$ into **L** and **U** the problem is written as follows.

$$\mathbf{LU} \, \hat{\mathbf{z}}_i^{m+1} = \mathbf{v}_i \qquad (5.53)$$

Following the established procedure an intermediate vector is introduced such that:

$$\mathbf{L}\,\mathbf{w} = \mathbf{v}_i \tag{5.54}$$

Setting corresponding elements of the vectors represented by the left and right hand sides of this equation equal permits the elements of **w** to be calculated as follows.

$$w_1 = \frac{v_1}{c_{i,1}}\,, \quad w_j = \frac{v_j - a_{i,j}\,w_{j-1}}{c_{i,j} - a_{i,j}\,e'_{j-1}} \quad (2 \le j \le NY) \tag{5.55}$$

Equations (5.53) and (5.54) are then combined to eliminate $\mathbf{v}_i$ and the result is pre-multiplied by $\mathbf{L}^{-1}$ to give:

$$\mathbf{U}\,\hat{\mathbf{z}}_i^{m+1} = \mathbf{w} \tag{5.56}$$

The components $\hat{z}_{i,j}^{m+1}$ of the solution vector $\hat{\mathbf{z}}_i^{m+1}$ are then given recursively by:

$$\hat{z}_{i,NY}^{m+1} = w_{NY}\,, \quad \hat{z}_{i,j}^{m+1} = w_j - e'_j\,\hat{z}_{i,j+1}^{m+1} \quad (NY-1 \ge j \ge 1) \tag{5.57}$$

In summary, therefore, the solution vector is obtained by firstly evaluating the two intermediate vectors **e'** and **w** from (5.52) and (5.55) and then applying (5.57). Finally, equation (5.46) is applied in order to introduce some over-relaxation into the solution procedure.

A problem of paramount importance associated with the SLOR method is the determination of an optimum value, ω_{opt} of ω, which minimises the number of iterations required to attain a converged solution. As pointed out by Smith [5.8] ω_{opt} can be simply related to the spectral radius of the Jacobi iteration matrix associated with matrix **A** of the linear system (5.42). However, evaluation of the spectral radius is an extremely difficult task for problems with a non-uniform mesh and non-simple boundary conditions. In light of this the value of ω_{opt} was determined, more simply, by running a number of sample problems for several different values of ω $(1 \le \omega < 2)$ and in each case noting the number of iterations required for convergence. The SLOR method was employed only for the solution of the heat flow equation (4.164) and in this case ω_{opt} was found to be 1.2.

5.1.6.2 The Strongly Implicit Procedure (SIP).

This method, first suggested by Stone [5.7], is more implicit than the SLOR technique. That is, each step of the SIP is more closely related to a direct solution

by elimination than a comparable step of SLOR, and it can therefore be expected to exhibit higher convergence rates. Basically, the SIP involves choosing a matrix **B** which is a good approximation to **A** in (5.42), with the prerequisite that **B** can be factored into upper and lower triangular matrices significantly more easily than **A**. The following iterative scheme, obtained by adding **B z** to the left and right hand sides of (5.42) can then be expected to be rapidly convergent.

$$\mathbf{B}\,\mathbf{z}^{m+1} = (\mathbf{B}-\mathbf{A})\,\mathbf{z}^{m} + \mathbf{b} \qquad (5.58)$$

As previously pointed out the direct factorization of **A** requires excessive computational effort. In this case the lower triangular matrix will be the same size as **A**, but it will have non-zero elements in every diagonal from the diagonal corresponding to the *f*- diagonal of matrix **A** through to the diagonal corresponding to the *c*- diagonal of matrix **A**. Similarly, the upper triangular matrix will have non-zero elements in every diagonal from the *c*- diagonal through to the *d*- diagonal. In the elimination procedure each of these elements must be calculated and stored for later use. Thus, for each mesh point the number of elements that must be computed is $2(NY+1)$. The generation of such a large number of intermediate coefficients makes direct elimination very slow. In order to alleviate these problems Stone suggested that the matrix **A** be modified in order to make the system (5.42) amenable to a direct solution. It was proposed that the triangular matrices **L** and **U** should be restricted to having only three non-zero diagonals. The non-zero diagonals of the lower triangular matrix, **L** are designated to coincide with the *f*-, *a*- and *c*- diagonals of **A**, and the non-zero diagonals of the upper triangular matrix, **U** to coincide with the *e*- and *d*- diagonals of **A**, with each element of its principle diagonal being equal to unity. Such matrices resulting from the 4×6 mesh are depicted in Figure 22. Following a similar procedure to that outlined in section 5.1.6.1 for factorising tridiagonal matrices, the product of **L** and **U** is performed, the result of which is shown in Figure 23. Each element of the resulting matrix is then equated to the corresponding element of **A** and this gives the following set of relations for each mesh point.

$$f'_{i,j} = f_{i,j} \qquad (5.59a)$$

$$f'_{i,j}\,e'_{i-1,j} = 0 \qquad (5.59b)$$

$$a'_{i,j} = a_{i,j} \qquad (5.59c)$$

$$c'_{i,j} + f'_{i,j}\,d'_{i-1,j} + a'_{i,j}\,e'_{i,j-1} = c_{i,j} \qquad (5.59d)$$

$$c'_{i,j}\,e'_{i,j} = e_{i,j} \qquad (5.59e)$$

c'_{11} $a'_{12}\ c'_{12}$ $a'_{13}\ c'_{13}$ $a'_{14}\ c'_{14}$ $a'_{15}\ c'_{15}$ $a'_{16}\ c'_{16}$			
f'_{21} f'_{22} f'_{23} f'_{24} f'_{25} f'_{26}	c'_{21} $a'_{22}\ c'_{22}$ $a'_{23}\ c'_{23}$ $a'_{24}\ c'_{24}$ $a'_{25}\ c'_{25}$ $a'_{26}\ c'_{26}$		
	f'_{31} f'_{32} f'_{33} f'_{34} f'_{35} f'_{36}	c'_{31} $a'_{32}\ c'_{32}$ $a'_{33}\ c'_{33}$ $a'_{34}\ c'_{34}$ $a'_{35}\ c'_{35}$ $a'_{36}\ c'_{36}$	
		f'_{41} f'_{42} f'_{43} f'_{44} f'_{45} f'_{46}	c'_{41} $a'_{42}\ c'_{42}$ $a'_{43}\ c'_{43}$ $a'_{44}\ c'_{44}$ $a'_{45}\ c'_{45}$ $a'_{46}\ c'_{46}$

$1\ e'_{11}$ $1\ e'_{12}$ $1\ e'_{13}$ $1\ e'_{14}$ $1\ e'_{15}$ 1	d'_{11} d'_{12} d'_{13} d'_{14} d'_{15} d'_{16}		
	$1\ e'_{21}$ $1\ e'_{22}$ $1\ e'_{23}$ $1\ e'_{24}$ $1\ e'_{25}$ 1	d'_{21} d'_{22} d'_{23} d'_{24} d'_{25} d'_{26}	
		$1\ e'_{31}$ $1\ e'_{32}$ $1\ e'_{33}$ $1\ e'_{34}$ $1\ e'_{35}$ 1	d'_{31} d'_{32} d'_{33} d'_{34} d'_{35} d'_{36}
			$1\ e'_{41}$ $1\ e'_{42}$ $1\ e'_{43}$ $1\ e'_{44}$ $1\ e'_{45}$ 1

Figure 22. Lower and Upper Triangular Matrices for a 4 by 6 Mesh.

$\eta_{11}\ \lambda_{11}$ $\gamma_{12}\ \eta_{12}\ \lambda_{12}$ $\gamma_{13}\ \eta_{13}\ \lambda_{13}$ $\gamma_{14}\ \eta_{14}\ \lambda_{14}$ $\gamma_{15}\ \eta_{15}\ \lambda_{15}$ $\gamma_{16}\ \eta_{16}$	σ_{11} $\xi_{12}\ \sigma_{12}$ $\xi_{13}\ \sigma_{13}$ $\xi_{14}\ \sigma_{14}$ $\xi_{15}\ \sigma_{15}$ $\xi_{16}\ \sigma_{16}$		
$\alpha_{21}\ \beta_{21}$ $\alpha_{22}\ \beta_{22}$ $\alpha_{23}\ \beta_{23}$ $\alpha_{24}\ \beta_{24}$ $\alpha_{25}\ \beta_{25}$ α_{26}	$\eta_{21}\ \lambda_{21}$ $\gamma_{22}\ \eta_{22}\ \lambda_{22}$ $\gamma_{23}\ \eta_{23}\ \lambda_{23}$ $\gamma_{24}\ \eta_{24}\ \lambda_{24}$ $\gamma_{25}\ \eta_{25}\ \lambda_{25}$ $\gamma_{26}\ \eta_{26}$	σ_{21} $\xi_{22}\ \sigma_{22}$ $\xi_{23}\ \sigma_{23}$ $\xi_{24}\ \sigma_{24}$ $\xi_{25}\ \sigma_{25}$ $\xi_{26}\ \sigma_{26}$	
	$\alpha_{31}\ \beta_{31}$ $\alpha_{32}\ \beta_{32}$ $\alpha_{33}\ \beta_{33}$ $\alpha_{34}\ \beta_{34}$ $\alpha_{35}\ \beta_{35}$ α_{36}	$\eta_{31}\ \lambda_{31}$ $\gamma_{32}\ \eta_{32}\ \lambda_{32}$ $\gamma_{33}\ \eta_{33}\ \lambda_{33}$ $\gamma_{34}\ \eta_{34}\ \lambda_{34}$ $\gamma_{35}\ \eta_{35}\ \lambda_{35}$ $\gamma_{36}\ \eta_{36}$	σ_{31} $\xi_{32}\ \sigma_{32}$ $\xi_{33}\ \sigma_{33}$ $\xi_{34}\ \sigma_{34}$ $\xi_{35}\ \sigma_{35}$ $\xi_{36}\ \sigma_{36}$
		$\alpha_{41}\ \beta_{41}$ $\alpha_{42}\ \beta_{42}$ $\alpha_{43}\ \beta_{43}$ $\alpha_{44}\ \beta_{44}$ $\alpha_{45}\ \beta_{45}$ α_{46}	$\eta_{41}\ \lambda_{41}$ $\gamma_{42}\ \eta_{42}\ \lambda_{42}$ $\gamma_{43}\ \eta_{43}\ \lambda_{43}$ $\gamma_{44}\ \eta_{44}\ \lambda_{44}$ $\gamma_{45}\ \eta_{45}\ \lambda_{45}$ $\gamma_{46}\ \eta_{46}$

$$\alpha_{i,j} = f'_{i,j}$$
$$\beta_{i,j} = f'_{i,j}\, e'_{i-1,j}$$
$$\gamma_{i,j} = a'_{i,j}$$
$$\eta_{i,j} = c'_{i,j} + f'_{i,j}\, d'_{i-1,j} + a'_{i,j}\, e'_{i,j-1}$$
$$\lambda_{i,j} = c'_{i,j}\, e'_{i,j}$$
$$\xi_{i,j} = a'_{i,j}\, d'_{i,j-1}$$
$$\sigma_{i,j} = c'_{i,j}\, d'_{i,j}$$

Figure 23. The LU Product.

$$a'_{i,j}\, d'_{i,j-1} = 0 \qquad (5.59f)$$

$$c'_{i,j}\, d'_{i,j} = d_{i,j} \qquad (5.59g)$$

By combining (5.59*a*) and (5.59*b*) it may be shown that $e'_{i,j} = 0$ for all (i,j). This would require from (5.59*e*) that $e_{i,j} = 0$ for all (i,j) which is obviously not the case. Therefore, these seven relationships cannot be satisfied.

As pointed out by Stone the simplest definition of the approximation, **B** that can be factored into **L** and **U** would result from ignoring (5.59*b*) and (5.59*f*) and using the five remaining relations to obtain the five dashed coefficients at each mesh point. Following this procedure yields a matrix **B** of the following form.

$$\mathbf{B} = \mathbf{LU} = \mathbf{A} + \mathbf{N} \tag{5.60}$$

The matrix **N** consists of two non-zero diagonals which result as a consequence of disregarding the equalities (5.59*b*) and (5.59*f*). These diagonals will be represented by:

$$g_{i,j} = f'_{i,j}\, e'_{i-1,j} \tag{5.61}$$

$$h_{i,j} = a'_{i,j}\, d'_{i,j-1} \tag{5.62}$$

It was found, however, that the above definition of matrix **B** could not be used as the basis of a rapidly convergent iterative scheme, as it defines a particularly poor approximation to **A**. In order to obtain an improved approximation it is necessary to reduce the magnitude of matrix **N** in (5.60). This quantity is defined as the sum of the magnitudes of all the elements of **N**. Stone proposed that this could be achieved by altering the original matrix, **A**. The altered matrix may be derived with the aid of Figure 24, which shows the mesh in the vicinity of point (i,j).

The original difference equation (5.43) at this point has non-zero coefficients for the unknown z values at (i,j) and its four nearest neighbours, as indicated by the solid dots in Figure 24. The modified equation corresponding to the approximate matrix **B** of Figure 23 has non-zero coefficients not only for these z- values, but also for the values at the points indicated by the crosses in Figure 24; that is points $(i-1,j+1)$ and $(i+1,j-1)$. In order to reduce the influence of these new terms, Stone suggested that approximately equal terms be subtracted from them. Suitable terms were obtained by firstly writing the Taylor series for point $(i-1,j+1)$ in the vicinity of point (i,j).

$$\begin{aligned} z_{i-1,j+1} = z_{i,j} - h_{i-1} \left.\frac{\partial z}{\partial x}\right|_{i,j} + k_j \left.\frac{\partial z}{\partial y}\right|_{i,j} + \frac{h_{i-1}^2}{2!} \left.\frac{\partial^2 z}{\partial x^2}\right|_{i,j} + \frac{k_j^2}{2!} \left.\frac{\partial^2 z}{\partial y^2}\right|_{i,j} \\ - h_{i-1} k_j^2 \left.\frac{\partial^2 z}{\partial x \partial y}\right|_{i,j} \end{aligned} \tag{5.63}$$

Similar Taylor expansions for points $(i-1,j)$ and $(i+1,j)$, which are given by the series' (4.6) and (4.7) respectively are then subtracted from (5.63). If the last

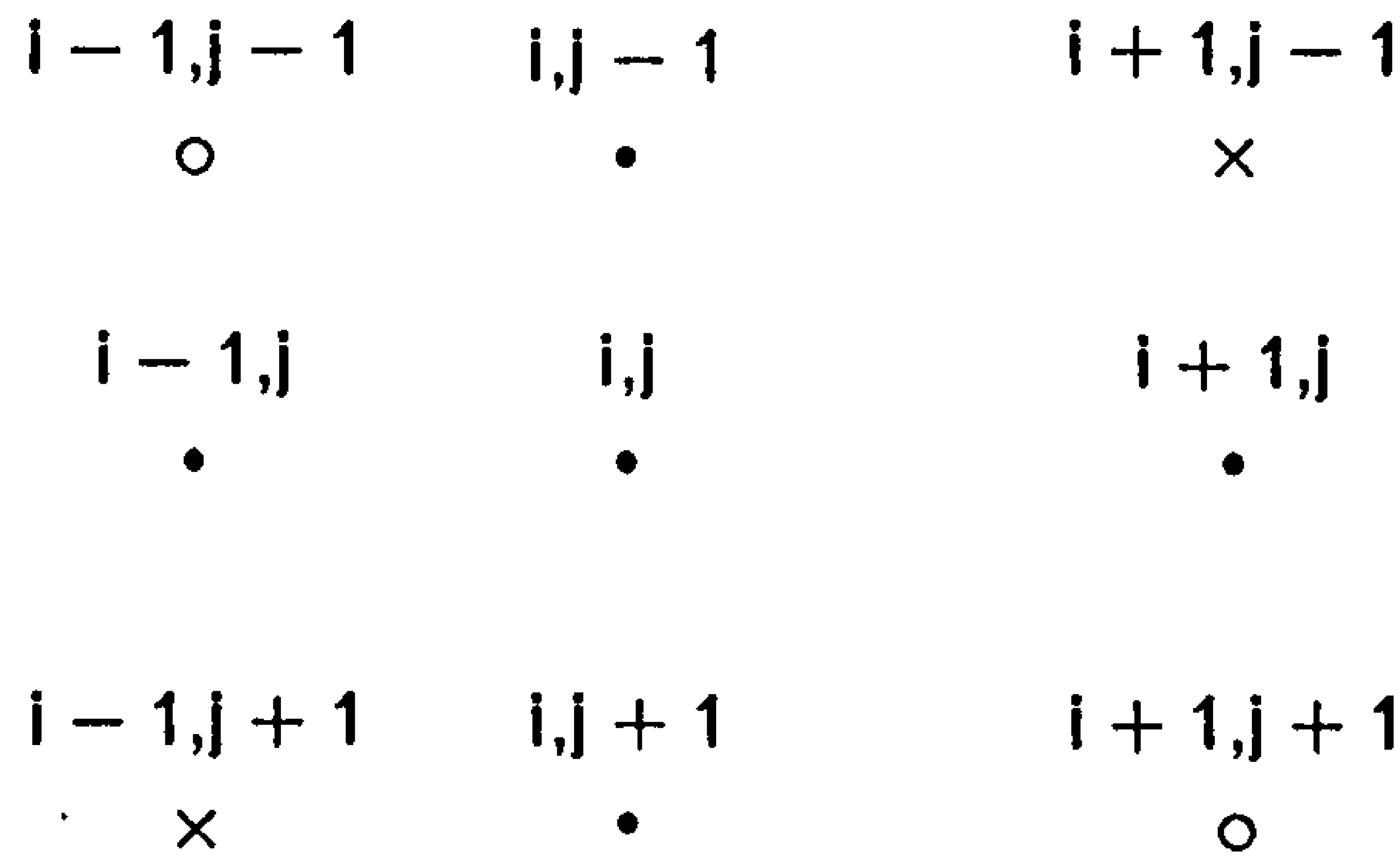


Figure 24. Mesh in Vicinity of Point i,j.

term in (5.63) is neglected together with all third and higher order terms then the following relation is obtained.

$$z_{i-1,j+1} \simeq -z_{i,j} + z_{i,j+1} + z_{i-1,j} \tag{5.64}$$

$$z_{i+1,j-1} \simeq -z_{i,j} + z_{i+1,j} + z_{i,j-1} \tag{5.65}$$

Provided z varies slowly across the mesh then equations (5.64) and (5.65) are good approximations. At this stage a certain amount of flexibility is introduced by multiplying the right hand sides of (5.64) and (5.65) by a variable iteration parameter, α before considering them to be approximations for $z_{i-1,j+1}$ and $z_{i+1,j-1}$. If α lies in the range $0 < \alpha < 1$ then whenever $z_{i-1,j+1}$ and $z_{i+1,j-1}$ are introduced during factorisation then they are partially reduced by subtracting a fraction of the right hand sides of (5.64) and (5.65). Adopting this approach results in the following modified version of the original equation (5.43).

$$\begin{aligned} &a_{i,j} z_{i,j-1} + f_{i,j} z_{i-1,j} + c_{i,j} z_{i,j} + d_{i,j} z_{i+1,j} + e_{i,j} z_{i,j+1} \\ &\quad + g_{i,j} \{z_{i-1,j+1} - \alpha(-z_{i,j} + z_{i,j+1} + z_{i-1,j})\} \\ &\quad + h_{i,j} \{z_{i+1,j-1} - \alpha(-z_{i,j} + z_{i+1,j} + z_{i,j-1})\} = b_{i,j} \end{aligned} \tag{5.66}$$

The top line of (5.66) is simply the left hand side of the original equation and the terms which are multiplied by $g_{i,j}$ and $h_{i,j}$ are the partially reduced z- values. In an analogous procedure to that used in obtaining equations (5.59*a*) through (5.59*g*), each of the terms in the nodal z values of (5.66) are collected and equated to the corresponding element of the **LU** product shown in Figure 23. Carrying this out results in the following set of simultaneous equations.

$$f'_{i,j} = f_{i,j} - \alpha\, g_{i,j} \tag{5.67a}$$

$$a'_{i,j} = a_{i,j} - \alpha\, h_{i,j} \qquad (5.67b)$$

$$c'_{i,j} + f'_{i,j}\, d'_{i-1,j} + a'_{i,j}\, e'_{i,j-1} = c_{i,j} + \alpha\, g_{i,j} + \alpha\, h_{i,j} \qquad (5.67c)$$

$$c'_{i,j}\, e'_{i,j} = e_{i,j} - \alpha\, g_{i,j} \qquad (5.67d)$$

$$c'_{i,j}\, d'_{i,j} = d_{i,j} - \alpha\, h_{i,j} \qquad (5.67e)$$

These equations together with (5.61) and (5.62) have been found to define a matrix, **B** which may be used as the basis of the iterative scheme given by (5.56). The dashed coefficients of **L** and **U** are firstly obtained at node (1,1). At this node the coefficients $g_{1,1}$ and $h_{1,1}$ are both zero and the remaining coefficients are given by

$$\begin{aligned} f'_{1,1} &= 0 \\ a'_{1,1} &= 0 \\ c'_{1,1} &= c_{1,1} \\ e'_{1,1} &= e_{1,1}/c_{1,1} \\ d'_{1,1} &= d_{1,1}/c_{1,1} \end{aligned} \qquad (5.68)$$

At the remaining mesh points along the first column $(2 \le j \le NY)$ $g_{1,j}$ is zero and $h_{i,j}$ is evaluated by eliminating the unknown, $a'_{1,j}$ from (5.62) and (5.67*b*). This, and the other coefficients are then given by:

$$\begin{aligned} h_{1,j} &= \frac{d'_{1,j-1}\, a_{1,j}}{1 + \alpha\, d'_{1,j-1}} \\ f'_{1,j} &= 0 \\ a'_{1,j} &= a_{1,j} - \alpha\, h_{1,j} \\ c'_{1,j} &= c_{1,j} - a'_{1,j}\, e'_{1,j-1} + \alpha\, h_{1,j} \\ e'_{1,j} &= e_{1,j}/c'_{1,j} \\ d'_{1,j} &= (d_{1,j} - \alpha\, h_{1,j})/c'_{1,j} \end{aligned} \qquad (5.69)$$

Similarly, at the remaining mesh points along the top row $(2 \le i \le NX)$ $h_{i,1}$ is zero and $g_{i,1}$ is obtained by eliminating $f'_{i,1}$ from (5.61) and (5.67*a*) . The coefficients are then given by:

$$\begin{aligned}
g_{i,1} &= \frac{e'_{i-1,1}\, f_{i,1}}{1 + \alpha\, e'_{i-1,1}} \\
f'_{i,1} &= f_{i,1} - \alpha\, g_{i,1} \\
a'_{i,1} &= 0 \\
c'_{i,1} &= c_{i,1} - f'_{i,1}\, d'_{i-1,1} + \alpha\, g_{i,1} \\
e'_{i,1} &= (e_{i,1} - \alpha\, g_{i,1})/c'_{i,1} \\
d'_{i,1} &= d_{i,1}/c'_{i,1}
\end{aligned} \tag{5.70}$$

The coefficients at the remaining nodes ($2 \leq i \leq NX$, $2 \leq j \leq NY$) can now be calculated by application of the full set of equations given by:

$$\begin{aligned}
g_{i,j} &= \frac{e'_{i-1,j}\, f_{i,j}}{1 + \alpha\, e'_{i-1,j}} \\
h_{i,j} &= \frac{d'_{i,j-1}\, a_{i,j}}{1 + \alpha\, d'_{i,j-1}} \\
f'_{i,j} &= f_{i,j} - \alpha\, g_{i,j} \\
a'_{i,j} &= a_{i,j} - \alpha\, h_{i,j} \\
c'_{i,1} &= c_{i,1} - f'_{i,j}\, d'_{i-1,j} - a'_{i,j}\, e'_{i,j-1} + \alpha\, g_{i,j} + \alpha\, h_{i,j} \\
e'_{i,j} &= (e_{i,j} - \alpha\, g_{i,j})/c'_{i,j} \\
d'_{i,j} &= (d_{i,j} - \alpha\, h_{i,j})/c'_{i,j}
\end{aligned} \tag{5.71}$$

Having calculated all the elements of **L** and **U** they can then be used in the iterative scheme of (5.58), which can be rewritten as follows.

$$\mathbf{B}\, \delta\mathbf{z}^m = \mathbf{b} - \mathbf{A}\, \mathbf{z}^m \tag{5.72}$$

where:

$$\delta\mathbf{z}^m = \mathbf{z}^{m+1} - \mathbf{z}^m \tag{5.73}$$

Equation (5.72) is then re-expressed as:

$$\mathbf{LU}\, \delta\mathbf{z}^m = \mathbf{u} \tag{5.74}$$

where the residual **u** is given by:

$$\mathbf{u} = \mathbf{b} - \mathbf{A}\, \mathbf{z}^m \tag{5.75}$$

The solution vector $\delta\mathbf{z}^m$ can then be found by introducing an intermediate vector, **w** as was done for the SLOR method, thus:

$$\mathbf{L}\,\mathbf{w} = \mathbf{u} \tag{5.76}$$

The components of **w** are calculated using forward substitution by the application of the following expressions in the order they are specified below.

$$\begin{aligned}
w_{1,1} &= u_{1,1}/c'_{1,1} \\
w_{1,j} &= (u_{1,j} - a'_{1,j}\, w_{1,j-1})/c'_{1,j} \quad (2 \le j \le NY) \\
w_{i,1} &= (u_{i,1} - f'_{i,1}\, w_{i-1,1})/c'_{i,1} \quad (2 \le i \le NX) \\
w_{i,j} &= (u_{i,j} - a'_{i,j}\, w_{i,j-1} - f'_{i,j}\, w_{i-1,j})/c'_{i,1} \quad (2 \le i \le NX,\ 2 \le j \le NY)
\end{aligned} \tag{5.77}$$

As for SLOR the solution is then obtained by solving:

$$\mathbf{U}\,\delta\mathbf{z}^m = \mathbf{w} \tag{5.78}$$

This equation is solved recursively by backward substitution using the following expressions.

$$\begin{aligned}
\delta z^m_{NX,NY} &= w_{NX,NY} \\
\delta z^m_{NX,j} &= w_{NX,j} - e'_{NX,j}\, w_{NX,j+1} \quad (NY - 1 \ge j \ge 1) \\
\delta z^m_{i,NY} &= w_{i,NY} - d'_{i,NY}\, w_{i+1,NY} \quad (NX - 1 \ge i \ge 1) \\
\delta z^m_{i,j} &= w_{i,j} - d'_{i,j}\, w_{i+1,j} - e'_{i,j}\, w_{i,j+1} \quad (NX - 1 \ge i \ge 1,\ NY - 1 \ge j \ge 1)
\end{aligned} \tag{5.79}$$

Finally the δz values are added to the corresponding z values. In summary, therefore, the solution procedure consists of applying (5.68)-(5.71) to find the dashed coefficients followed by the repeated application of (5.77), (5.79) and (5.73) to obtain the updated solution. It was suggested by Stone, however, that for every other iteration (ie. for even numbered iterations) the above algorithm should be applied with i varying as $NX, NX - 1 \ldots 1$ rather than $1, 2 \ldots NX$. This re-ordering of the grid points has the effect of making the non-zero coefficients appear for the values $z_{i-1,j-1}$ and $z_{i+1,j+1}$ as indicated by the open circles in Figure 24, rather than for the values $z_{i-1,j+1}$ and $z_{i+1,j-1}$ as in the odd numbered steps. As stated by Stone, this variation is not always essential for convergence, but it often increases the rate of convergence dramatically. This alteration can be implemented by reordering the original matrix, **A**. This is achieved by 'reflecting' coefficients a, c and e about the vertical centre line of the mesh and also exchanging coefficients a and e, before re-applying the above scheme; that is:

$$\begin{aligned}
a_{i,j} &\equiv e_{NX+1-i,j} \\
e_{i,j} &\equiv a_{NX+1-i,j} \\
c_{i,j} &\equiv c_{NX+1-i,j}
\end{aligned} \tag{5.80}$$

Having obtained the δz values at the even numbered steps they should also be reflected so that they are correctly positioned before being added to their corresponding z values.

$$\delta z_{NX+1-i,j} \equiv \delta z_{i,j} \tag{5.81}$$

For this procedure it is noted that a new set of dashed coefficients must be calculated prior to each application of (5.77) and (5.79).

The convergence rate of this method depends on the value of α. Rather than use a fixed value for every iteration Stone suggested using a range of α values ranging between 0 and α_{max} to improve convergence. The best maximum parameter was found to depend upon the particular problem being solved and a suitable value is given by:

$$\alpha_{max} = 1 - \min\left[\frac{2\,h_i^2}{1 + h_i^2/k_j^2}\,,\ \frac{2\,k_j^2}{1 + k_j^2/h_i^2}\right] \tag{5.82}$$

which for most applications gives a value just below unity. The individual parameters, α_n should be geometrically spaced as follows.

$$1 - \alpha_n = (1 - \alpha_{max})^{n/N} \qquad n = 0,1 \ldots N \tag{5.83}$$

It was recommended by Stone that a minimum of four parameters are used and in this instance a total of eight have been used ($N = 7$). Each individual parameter is used for one forward and reverse sweep of the algorithm and the parameters are applied in a cyclical manner. The order of application of the parameters is not critical. While α_{max} defined by (5.82) proved to be the optimum value for the solution of the linearised Poisson equation (5.16), experimentation has shown that the optimum value for the linearised continuity equations (5.24) and (5.36) is zero. In this case, therefore, all the α_n values are zero and reflection of the coefficients becomes redundant. This value is probably justified by the fact that the carrier concentrations vary exponentially over certain regions within the simulation domain and in this case (5.64) and (5.65) can be expected to be very poor approximations.

It was found that only one single forward or reverse sweep of the SIP was required to provide a sufficiently accurate solution to each of the linearised equations (5.16), (5.24) and (5.36) at each Newton iteration. Thus, the modified coefficient and residual (right hand side) of these equations are recalculated before each sweep of the SIP. In addition the solution vector, $\mathbf{z}^1$ was initialized at $\mathbf{z}^0 = \mathbf{0}$ prior to each application of the SIP.

5.2 Coupled Solution Procedure.

The coupled procedure was first reported for one dimensional steady state problems by Gokhale [5.9], and was later extended to transient problems [5.10] and then to two dimensional problems [5.11] [5.12]. For problems where the potential distribution is a sensitive function of current density (eg. current induced base widening in BJT's) the coupled procedure gives much faster convergence than the decoupled procedure. In these situations the strong electrical coupling between the Poisson and continuity equations is reflected by the way in which the equations are solved. The coupled approach does, however, possess a number of disadvantages compared to the decoupled procedure. It is more involved with regard to program structure as it requires the use of additional coefficients over and above those of the decoupled approach. A more serious limitation is that this procedure requires a close initial guess to the solution before it can be applied otherwise divergence will result. This was not a prerequisite for the decoupled approach.

In this particular application the electrical model which is comprised of Poisson's equation and the continuity equations is solved in a coupled manner, whilst the heat flow equation has been solved separately as in the decoupled approach. Although this may not be as attractive from a mathematical viewpoint as solving the complete system in a coupled manner, it does allow for greater flexibility as the effects of self-heating are often negligible or not required. The exclusion of the heat flow equation also simplifies the solution quite dramatically. A flow diagram for the coupled procedure is shown in Figure 25. Notice that the outer iteration (index *n*) is not required in the coupled approach and the treatment of the heat flow equation is exactly the same as for the decoupled approach. The decoupled method is more efficient at low-level injection and it was used to initialise the coupled method. The base bias was then incremented in small steps (0.05*v*) while the collector voltage was kept constant. In this way high level injection could be simulated. The Newton method as applied to the electrical model may be written as follows.

$$\begin{bmatrix} \dfrac{\partial F_\psi}{\partial \psi} & \dfrac{\partial F_\psi}{\partial n} & \dfrac{\partial F_\psi}{\partial p} \\ \dfrac{\partial F_n}{\partial \psi} & \dfrac{\partial F_n}{\partial n} & \dfrac{\partial F_n}{\partial p} \\ \dfrac{\partial F_p}{\partial \psi} & \dfrac{\partial F_p}{\partial n} & \dfrac{\partial F_p}{\partial p} \end{bmatrix}^k \begin{bmatrix} \delta\psi^k \\ \delta n^k \\ \delta p^k \end{bmatrix} = - \begin{bmatrix} F_\psi(\psi^k, n^k, p^k) \\ F_n(\psi^k, n^k, p^k) \\ F_p(\psi^k, n^k, p^k) \end{bmatrix} \tag{5.84}$$

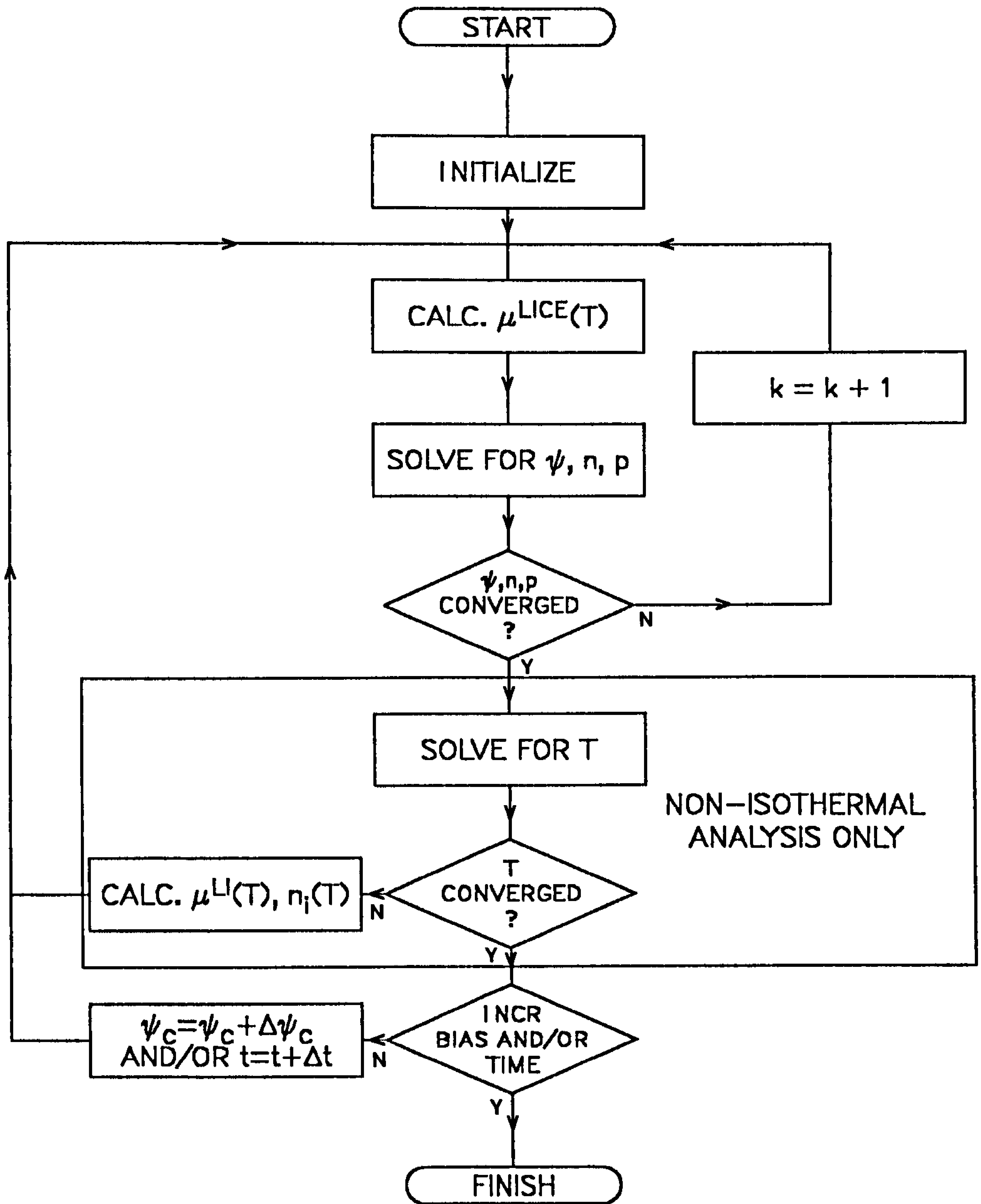


Figure 25. Flow Diagram Illustrating Coupled Solution Procedure.

The size of the Jacobian is now $3.NX.NY \times 3.NX.NY$ as opposed to $NX.NY \times NX.NY$ for the decoupled approach. The corresponding nodal equations at the inner mesh points are evaluated by applying (5.84) to the discrete Poisson equation (4.15) and the continuity equations (4.52) and (4.58). The linearised Poisson equation is given by:

$$\begin{aligned} A_{i,j}\,\delta\psi^k_{i,j-1} + B_{i,j}\,\delta\psi^k_{i-1,j} + C_{i,j}\,\delta\psi^k_{i,j} + D_{i,j}\,\delta\psi^k_{i+1,j} + E_{i,j}\,\delta\psi^k_{i,j+1} \\ \overline{A}_{i,j}\,\delta n^k_{i,j} + \overline{B}_{i,j}\,\delta p^k_{i,j} = -F_\psi(\psi^k, n^k, p^k) \end{aligned} \tag{5.85}$$

where the derivatives with respect to the carrier concentrations are given by:

$$\overline{A}_{i,j} = -\frac{q}{\varepsilon_{sil}}\,, \quad \overline{B}_{i,j} = \frac{q}{\varepsilon_{sil}} \tag{5.86}$$

Applying (5.84) to the electron current continuity equation gives:

$$\begin{aligned} \overline{F}^k_{i,j}\,\delta\psi^k_{i,j-1} + \overline{G}^k_{i,j}\,\delta\psi^k_{i-1,j} + \overline{H}^k_{i,j}\,\delta\psi^k_{i,j} + \overline{I}^k_{i,j}\,\delta\psi^k_{i+1,j} + \overline{L}^k_{i,j}\,\delta\psi^k_{i,j+1} \\ + F^k_{i,j}\,\delta n^k_{i,j-1} + G^k_{i,j}\,\delta n^k_{i-1,j} + \hat{H}^k_{i,j}\,\delta n^k_{i,j} + I^k_{i,j}\,\delta n^k_{i+1,j} + L^k_{i,j}\,\delta n^k_{i,j+1} \\ + \overline{C}^k_{i,j}\,\delta p^k_{i,j} = -F_n(\psi^k, n^k, p^k) \end{aligned} \tag{5.87}$$

The derivative of the discrete equation with respect to $\psi_{i,j-1}$ is denoted by $\overline{F}_{i,j}$ and may be written as follows.

$$\overline{F}_{i,j} = \frac{\partial F_{n_{i,j}}}{\partial \psi_{i,j-1}} = \frac{\partial\left(F_{i,j}\,n_{i,j-1} + H_{i,j}\,n_{i,j}\right)}{\partial \psi_{i,j-1}} \tag{5.88}$$

The following is obtained upon substitution of the coefficients $F_{i,j}$ and $H_{i,j}$ with (4.53) and (4.55).

$$\begin{aligned} \overline{F}_{i,j} = \frac{2\,k\,\mu_{n_{i,j-1/2}}}{q\,L(T_{i,j-1}, T_{i,j})\,k_j\,(k_j + k_{j-1})} \Bigg(& n_{i,j-1}\frac{\partial B\{F(\psi_{i,j-1}, \psi_{i,j}, T_{i,j-1}, T_{i,j})\}}{\partial \psi_{i,j-1}} \\ & - n_{i,j}\frac{\partial B\{F(\psi_{i,j}, \psi_{i,j-1}, T_{i,j}, T_{i,j-1})\}}{\partial \psi_{i,j-1}} \Bigg) \end{aligned} \tag{5.89}$$

The derivatives can be evaluated with the aid of the chain rule as follows.

$$\frac{\partial B(x)}{\partial \psi} = \frac{\partial B(x)}{\partial x}\,\frac{\partial x}{\partial \psi} = P(x)\,\frac{\partial x}{\partial \psi} \tag{5.90}$$

where $P(x)$ is the derivative of the Bernoulli function, (4.36) and is given by:

$$P(x) = \frac{(1-x)\exp(x) - 1}{(\exp(x) - 1)^2} \tag{5.91}$$

The coefficient $\overline{F}_{i,j}$ is finally given by:

$$\overline{F}_{i,j} = \frac{2\,\mu_{n_{i,j-1/2}}}{k_{j-1}\,(k_j + k_{j-1})}\,(n_{i,j-1}P\{F(\psi_{i,j-1}, \psi_{i,j}, T_{i,j-1}, T_{i,j})\} + n_{i,j}P\{F(\psi_{i,j}, \psi_{i,j-1}, T_{i,j}, T_{i,j-1})\}) \tag{5.92}$$

The remaining coefficients of the $\delta\psi$'s are given as follows.

$$\overline{G}_{i,j} = \frac{2\,\mu_{n_{i-1/2,j}}}{h_{i-1}\,(h_i + h_{i-1})}\,(n_{i-1,j}P\{F(\psi_{i-1,j}, \psi_{i,j}, T_{i-1,j}, T_{i,j})\} + n_{i,j}P\{F(\psi_{i,j}, \psi_{i-1,j}, T_{i,j}, T_{i-1,j})\})$$

$$\overline{I}_{i,j} = \frac{2\,\mu_{n_{i+1/2,j}}}{h_i\,(h_i + h_{i-1})}\,(n_{i+1,j}P\{F(\psi_{i+1,j}, \psi_{i,j}, T_{i+1,j}, T_{i,j})\} + n_{i,j}P\{F(\psi_{i,j}, \psi_{i+1,j}, T_{i,j}, T_{i+1,j})\})$$

$$\overline{L}_{i,j} = \frac{2\,\mu_{n_{i,j+1/2}}}{k_j\,(k_j + k_{j-1})}\,(n_{i,j+1}P\{F(\psi_{i,j+1}, \psi_{i,j}, T_{i,j+1}, T_{i,j})\} + n_{i,j}P\{F(\psi_{i,j}, \psi_{i,j+1}, T_{i,j}, T_{i,j+1})\})$$

$$\overline{H}_{i,j} = -\,(\overline{F}_{i,j} + \overline{G}_{i,j} + \overline{I}_{i,j} + \overline{L}_{i,j}) \tag{5.93}$$

In taking the derivatives with respect to potential any non-linearities due to the field dependent mobility have been ignored. This non-linearity is not expected to be significant as the variation of mobility with potential is of second order importance compared with variation of field with potential. Thus, the effect on the convergence rate of the Newton method should be small. The coefficient $\overline{C}_{i,j}$ originates from the recombination term and is given by (5.38). The discrete hole current continuity equation can be treated in exactly the same way. Applying (5.84) to the hole continuity equation gives:

$$\begin{aligned} &\overline{M}^k_{i,j}\,\delta\psi^k_{i,j-1} + \overline{W}^k_{i,j}\,\delta\psi^k_{i-1,j} + \overline{X}^k_{i,j}\,\delta\psi^k_{i,j} + \overline{S}^k_{i,j}\,\delta\psi^k_{i+1,j} + \overline{U}^k_{i,j}\,\delta\psi^k_{i,j+1} \\ &+ M^k_{i,j}\,\delta p^k_{i,j-1} + W^k_{i,j}\,\delta p^k_{i-1,j} + \hat{X}^k_{i,j}\,\delta p^k_{i,j} + S^k_{i,j}\,\delta p^k_{i+1,j} + U^k_{i,j}\,\delta p^k_{i,j+1} \\ &+ \overline{D}^k_{i,j}\,\delta n^k_{i,j} = -F_p(\psi^k, n^k, p^k) \end{aligned} \tag{5.94}$$

The coefficients of the $\delta\psi$'s are as follows.

$$\overline{M}_{i,j} = \frac{2\,\mu_{p_{i,j-1/2}}}{k_{j-1}\,(k_j + k_{j-1})}\,(p_{i,j-1}P\{F(\psi_{i,j},\psi_{i,j-1},T_{i,j-1},T_{i,j})\} + p_{i,j}P\{F(\psi_{i,j-1},\psi_{i,j},T_{i,j},T_{i,j-1})\})$$

$$\overline{W}_{i,j} = \frac{2\,\mu_{p_{i-1/2,j}}}{h_{i-1}\,(h_i + h_{i-1})}\,(p_{i-1,j}P\{F(\psi_{i,j},\psi_{i-1,j},T_{i-1,j},T_{i,j})\} + p_{i,j}P\{F(\psi_{i-1,j},\psi_{i,j},T_{i,j},T_{i-1,j})\})$$

$$\overline{S}_{i,j} = \frac{2\,\mu_{p_{i+1/2,j}}}{h_i\,(h_i + h_{i-1})}\,(p_{i+1,j}P\{F(\psi_{i,j},\psi_{i+1,j},T_{i+1,j},T_{i,j})\} + p_{i,j}P\{F(\psi_{i+1,j},\psi_{i,j},T_{i,j},T_{i+1,j})\})$$

$$\overline{U}_{i,j} = \frac{2\,\mu_{p_{i,j+1/2}}}{k_j\,(k_j + k_{j-1})}\,(p_{i,j+1}P\{F(\psi_{i,j},\psi_{i,j+1},T_{i,j+1},T_{i,j})\} + p_{i,j}P\{F(\psi_{i,j+1},\psi_{i,j},T_{i,j},T_{i,j+1})\})$$

$$\overline{X}_{i,j} = -\,(\overline{M}_{i,j} + \overline{W}_{i,j} + \overline{S}_{i,j} + \overline{U}_{i,j}) \tag{5.95}$$

and $\overline{D}_{i,j}$ is the negative of the value given by (5.26).

The additional coefficients generated in the coupled method disrupt the penta-diagonal structure of the Jacobian resulting from the decoupled method. A new linear equation solution technique is, therefore, required which is suitable for solving the particular system generated in the coupled method. An iterative scheme that has proven to be highly convergent will now be described.

Firstly the system is ordered in such a way that the bandwidth of the Jacobian is minimised. This ordering also serves to place nine out of a total of fifteen non-zero diagonals at or directly adjacent to the principle diagonal of the Jacobian. This gives a significant improvement in the convergence rate of the iterative scheme to be employed. More specifically equations (5.85), (5.87) and (5.94) are ordered consecutively for each node starting from (1,1), then going down each column, moving from left to right. An example of the resulting linear system for a small 3×3 mesh is given in Figure 26. It is readily apparent that the system can be partitioned in the following manner.

$$\begin{bmatrix} \mathbf{Y}_1 & \mathbf{Z}_1 & & & & & \\ \mathbf{X}_2 & \mathbf{Y}_2 & \mathbf{Z}_2 & & & & \\ & \bullet & \bullet & \bullet & & & \\ & & \mathbf{X}_i & \mathbf{Y}_i & \mathbf{Z}_i & & \\ & & & \bullet & \bullet & \bullet & \\ & & & & \bullet & \bullet & \mathbf{Z}_{NX-1} \\ & & & & & \mathbf{X}_{NX} & \mathbf{Y}_{NX} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Delta}_1 \\ \boldsymbol{\Delta}_2 \\ \bullet \\ \boldsymbol{\Delta}_i \\ \bullet \\ \bullet \\ \boldsymbol{\Delta}_{NX} \end{bmatrix} = - \begin{bmatrix} \mathbf{F}_1 \\ \mathbf{F}_2 \\ \bullet \\ \mathbf{F}_i \\ \bullet \\ \bullet \\ \mathbf{F}_{NX} \end{bmatrix} \tag{5.96}$$

The sub-matrices, $\mathbf{X}_i$, $\mathbf{Y}_i$ and $\mathbf{Z}_i$ are of dimension $3NY \times 3NY$ and $\boldsymbol{\Delta}_i$ and $\mathbf{F}_i$ are vectors of length $3NY$. The method of SLOR has been utilised to solve this equation. The particular scheme which has been used resembles closely that

$C_{11}\ \bar{A}_{11}\ \bar{B}_{11}$ $\bar{H}_{11}\ \hat{H}_{11}\ \bar{C}_{11}$ $\bar{X}_{11}\ \bar{D}_{11}\ \hat{X}_{11}$	E_{11} $\bar{L}_{11}\ L_{11}$ $\bar{U}_{11}\ U_{11}$		D_{11} $\bar{I}_{11}\ I_{11}$ $\bar{S}_{11}\ S_{11}$						$\delta\psi_{11}$ δn_{11} δp_{11}		$F_{\psi 11}$ F_{n11} F_{p11}
A_{12} $\bar{F}_{12}\ F_{12}$ $\bar{M}_{12}\ M_{12}$	$C_{12}\ \bar{A}_{12}\ \bar{B}_{12}$ $\bar{H}_{12}\ \hat{H}_{12}\ \bar{C}_{12}$ $\bar{X}_{12}\ \bar{D}_{12}\ \hat{X}_{12}$	E_{12} $\bar{L}_{12}\ L_{12}$ $\bar{U}_{12}\ U_{12}$		D_{12} $\bar{I}_{12}\ I_{12}$ $\bar{S}_{12}\ S_{12}$					$\delta\psi_{12}$ δn_{12} δp_{12}		$F_{\psi 12}$ F_{n12} F_{p12}
	A_{13} $\bar{F}_{13}\ F_{13}$ $\bar{M}_{13}\ M_{13}$	$C_{13}\ \bar{A}_{13}\ \bar{B}_{13}$ $\bar{H}_{13}\ \hat{H}_{13}\ \bar{C}_{13}$ $\bar{X}_{13}\ \bar{D}_{13}\ \hat{X}_{13}$			D_{13} $\bar{I}_{13}\ I_{13}$ $\bar{S}_{13}\ S_{13}$				$\delta\psi_{13}$ δn_{13} δp_{13}		$F_{\psi 13}$ F_{n13} F_{p13}
B_{21} $\bar{G}_{21}\ G_{21}$ $\bar{W}_{21}\ W_{21}$			$C_{21}\ \bar{A}_{21}\ \bar{B}_{21}$ $\bar{H}_{21}\ \hat{H}_{21}\ \bar{C}_{21}$ $\bar{X}_{21}\ \bar{D}_{21}\ \hat{X}_{21}$	E_{21} $\bar{L}_{21}\ L_{21}$ $\bar{U}_{21}\ U_{21}$		D_{21} $\bar{I}_{21}\ I_{21}$ $\bar{S}_{21}\ S_{21}$			$\delta\psi_{21}$ δn_{21} δp_{21}		$F_{\psi 21}$ F_{n21} F_{p21}
	B_{22} $\bar{G}_{22}\ G_{22}$ $\bar{W}_{22}\ W_{22}$		A_{22} $\bar{F}_{22}\ F_{22}$ $\bar{M}_{22}\ M_{22}$	$C_{22}\ \bar{A}_{22}\ \bar{B}_{22}$ $\bar{H}_{22}\ \hat{H}_{22}\ \bar{C}_{22}$ $\bar{X}_{22}\ \bar{D}_{22}\ \hat{X}_{22}$	E_{22} $\bar{L}_{22}\ L_{22}$ $\bar{U}_{22}\ U_{22}$		D_{22} $\bar{I}_{22}\ I_{22}$ $\bar{S}_{22}\ S_{22}$		$\delta\psi_{22}$ δn_{22} δp_{22}	= −	$F_{\psi 22}$ F_{n22} F_{p22}
		B_{23} $\bar{G}_{23}\ G_{23}$ $\bar{W}_{23}\ W_{23}$		A_{23} $\bar{F}_{23}\ F_{23}$ $\bar{M}_{23}\ M_{23}$	$C_{23}\ \bar{A}_{23}\ \bar{B}_{23}$ $\bar{H}_{23}\ \hat{H}_{23}\ \bar{C}_{23}$ $\bar{X}_{23}\ \bar{D}_{23}\ \hat{X}_{23}$			D_{23} $\bar{I}_{23}\ I_{23}$ $\bar{S}_{23}\ S_{23}$	$\delta\psi_{23}$ δn_{23} δp_{23}		$F_{\psi 23}$ F_{n23} F_{p23}
			B_{31} $\bar{G}_{31}\ G_{31}$ $\bar{W}_{31}\ W_{31}$			$C_{31}\ \bar{A}_{31}\ \bar{B}_{31}$ $\bar{H}_{31}\ \hat{H}_{31}\ \bar{C}_{31}$ $\bar{X}_{31}\ \bar{D}_{31}\ \hat{X}_{31}$	E_{31} $\bar{L}_{31}\ L_{31}$ $\bar{U}_{31}\ U_{31}$		$\delta\psi_{31}$ δn_{31} δp_{31}		$F_{\psi 31}$ F_{n31} F_{p31}
				B_{32} $\bar{G}_{32}\ G_{32}$ $\bar{W}_{32}\ W_{32}$		A_{32} $\bar{F}_{32}\ F_{32}$ $\bar{M}_{32}\ M_{32}$	$C_{32}\ \bar{A}_{32}\ \bar{B}_{32}$ $\bar{H}_{32}\ \hat{H}_{32}\ \bar{C}_{32}$ $\bar{X}_{32}\ \bar{D}_{32}\ \hat{X}_{32}$	E_{32} $\bar{L}_{32}\ L_{32}$ $\bar{U}_{32}\ U_{32}$	$\delta\psi_{32}$ δn_{32} δp_{32}		$F_{\psi 32}$ F_{n32} F_{p32}
					B_{33} $\bar{G}_{33}\ G_{33}$ $\bar{W}_{33}\ W_{33}$		A_{33} $\bar{F}_{33}\ F_{33}$ $\bar{M}_{33}\ M_{33}$	$C_{33}\ \bar{A}_{33}\ \bar{B}_{33}$ $\bar{H}_{33}\ \hat{H}_{33}\ \bar{C}_{33}$ $\bar{X}_{33}\ \bar{D}_{33}\ \hat{X}_{33}$	$\delta\psi_{33}$ δn_{33} δp_{33}		$F_{\psi 33}$ F_{n33} F_{p33}

Figure 26. Example of a Linear System Arising from the Coupled Solution Procedure.

described in section 5.1.6.1 for the decoupled approach. However, the most basic element of the Jacobian is now taken to be a 3×3 sub-matrix rather than a single scalar quantity as was previously the case. The SLOR iterative scheme is completely defined by:

$$\mathbf{Y}_l\,\hat{\Delta}_l^{m+1} = -\mathbf{F}_l - \mathbf{X}_l\,\Delta_{l-1}^{m+1} - \mathbf{Z}_l\,\Delta_{l+1}^{m} \tag{5.97}$$

$$\Delta_l^{m+1} = \Delta_l^m + \omega(\hat{\Delta}_l^{m+1} - \Delta_l^m) \tag{5.98}$$

These equations are of exactly the same form as (5.45) and (5.46). As before a single iteration of SLOR consists of calculating the Δ values simultaneously for each column from column 1 through to NX. Equation (5.97) can be expanded as follows:

$$
\begin{bmatrix}
\mathbf{C}_{i,1} & \mathbf{E}_{i,1} & & & & \\
\mathbf{A}_{i,2} & \mathbf{C}_{i,2} & \mathbf{E}_{i,2} & & & \\
& \bullet & \bullet & \bullet & & \\
& & \mathbf{A}_{i,j} & \mathbf{C}_{i,j} & \mathbf{E}_{i,j} & \\
& & & \bullet & \bullet & \bullet \\
& & & & \bullet\ \bullet & \mathbf{E}_{i,NY-1} \\
& & & & \mathbf{A}_{i,NY} & \mathbf{C}_{i,NY}
\end{bmatrix}
\begin{bmatrix}
\hat{\delta}^{m+1}_{i,1} \\ \hat{\delta}^{m+1}_{i,2} \\ \bullet \\ \hat{\delta}^{m+1}_{i,j} \\ \bullet \\ \bullet \\ \hat{\delta}^{m+1}_{i,NY}
\end{bmatrix}
=
\begin{bmatrix}
-\mathbf{f}_{i,1} - \mathbf{F}_{i,1}\,\delta^{m+1}_{i-1,1} - \mathbf{D}_{i,1}\,\delta^{m}_{i+1,1} \\
-\mathbf{f}_{i,2} - \mathbf{F}_{i,2}\,\delta^{m+1}_{i-1,2} - \mathbf{D}_{i,2}\,\delta^{m}_{i+1,2} \\
\bullet \\
-\mathbf{f}_{i,j} - \mathbf{F}_{i,j}\,\delta^{m+1}_{i-1,j} - \mathbf{D}_{i,j}\,\delta^{m}_{i+1,j} \\
\bullet \\
\bullet \\
-\mathbf{f}_{i,NY} - \mathbf{F}_{i,NY}\,\delta^{m+1}_{i-1,NY} - \mathbf{D}_{i,NY}\,\delta^{m}_{i+1,NY}
\end{bmatrix}
\tag{5.99}
$$

where the sub-matrices and sub-vectors are given by:

$$
\mathbf{A}_{i,j} = \begin{bmatrix} A_{i,j} & & \\ \overline{F}_{i,j} & F_{i,j} & \\ \overline{M}_{i,j} & & M_{i,j} \end{bmatrix},\quad
\mathbf{C}_{i,j} = \begin{bmatrix} C_{i,j} & \overline{A}_{i,j} & \overline{B}_{i,j} \\ \overline{H}_{i,j} & \hat{H}_{i,j} & \overline{C}_{i,j} \\ \overline{X}_{i,j} & \overline{D}_{i,j} & \hat{X}_{i,j} \end{bmatrix},\quad
\mathbf{E}_{i,j} = \begin{bmatrix} E_{i,j} & & \\ \overline{L}_{i,j} & L_{i,j} & \\ \overline{U}_{i,j} & & U_{i,j} \end{bmatrix}
$$

$$
\mathbf{F}_{i,j} = \begin{bmatrix} B_{i,j} & & \\ \overline{G}_{i,j} & G_{i,j} & \\ \overline{W}_{i,j} & & W_{i,j} \end{bmatrix},\quad
\mathbf{D}_{i,j} = \begin{bmatrix} D_{i,j} & & \\ \overline{I}_{i,j} & I_{i,j} & \\ \overline{S}_{i,j} & & S_{i,j} \end{bmatrix},\quad
\delta_{i,j} = \begin{bmatrix} \delta\psi_{i,j} \\ \delta n_{i,j} \\ \delta p_{i,j} \end{bmatrix},\quad
\mathbf{f}_{i,j} = \begin{bmatrix} F_{\psi_{i,j}} \\ F_{n_{i,j}} \\ F_{p_{i,j}} \end{bmatrix}
\tag{5.100}
$$

Equation (5.99) is similar to (5.47) except that the elements of the matrix are now 3×3 sub-matrices and the elements of the vectors are 1×3 sub-matrices. This system can be solved by formally extending the Thomas algorithm described in section 5.1.6.1 to a more general block tridiagonal form. The elemental 3×3 matrices, $\mathbf{E}'_j$ of the upper triangular matrix (cf. equation (5.49)) can be obtained from the generalised version of (5.51) which is given by:

$$
\mathbf{E}'_1 = \mathbf{C}^{-1}_{i,1}\,\mathbf{E}_{i,1},\quad \mathbf{E}'_j = (\mathbf{C}_{i,j} - \mathbf{A}_{i,j}\,\mathbf{E}'_{j-1})^{-1}\,\mathbf{E}_{i,j} \quad (2 \leq j \leq NY-1) \tag{5.101}
$$

Similarly the elemental 1×3 sub-vectors $\mathbf{w}_j$ are obtained from the extension of (5.55).

$$
\mathbf{w}_1 = \mathbf{C}^{-1}_{i,1}\,\mathbf{v}_1,\quad \mathbf{w}_j = (\mathbf{C}_{i,j} - \mathbf{A}_{i,j}\,\mathbf{E}'_{j-1})^{-1}\,(\mathbf{v}_j - \mathbf{A}_{i,j}\,\mathbf{w}_{j-1}) \quad (2 \leq j \leq NY-1) \tag{5.102}
$$

where

$$
\mathbf{v}_j = -\mathbf{f}_{i,j} - \mathbf{F}_{i,j}\,\delta^{m+1}_{i-1,j} - \mathbf{D}_{i,j}\,\delta^{m}_{i+1,j} \tag{5.103}
$$

The $\hat{\delta}_{i,j}$ values are then given recursively using the following extension of (5.57).

$$
\hat{\delta}^{m+1}_{i,NY} = \mathbf{w}_{NY},\quad \hat{\delta}^{m+1}_{i,j} = \mathbf{w}_j - \mathbf{E}'_j\,\hat{\delta}^{m+1}_{i,j+1} \quad (NY-1 \geq j \geq 1) \tag{5.104}
$$

Some over-relaxation is then introduced by applying:

$$\delta_{i,j}^{m+1} = \delta_{i,j}^{m} + \omega\,(\hat{\delta}_{i,j}^{m+1} - \delta_{i,j}^{m}) \tag{5.105}$$

where:

$$\omega = \begin{bmatrix} \omega_\psi & & \\ & \omega_n & \\ & & \omega_p \end{bmatrix} \tag{5.106}$$

To summarize, therefore, a single iteration of block SLOR consists of calculating $\mathbf{E}'_j$ from (5.101) and $\mathbf{w}_j$ from (5.102) followed by $\delta_{i,j}^{m+1}$ from (5.104) and (5.105), starting at column 1 and ending at column NX. Inversion of the 3×3 matrices, $\mathbf{C}_{i,1}$ and $(\mathbf{C}_{i,j} - \mathbf{A}_{i,j}\mathbf{E}'_{j-1})$ in (5.101) and (5.102) is most efficiently accomplished using cofactors as follows.

$$\begin{bmatrix} a_{11}\, a_{21}\, a_{31} \\ a_{12}\, a_{22}\, a_{32} \\ a_{13}\, a_{23}\, a_{33} \end{bmatrix}^{-1} = \frac{1}{\det} \begin{bmatrix} a_{22}a_{33} - a_{32}a_{23} & a_{31}a_{23} - a_{21}a_{33} & a_{21}a_{32} - a_{31}a_{22} \\ a_{32}a_{13} - a_{12}a_{33} & a_{11}a_{33} - a_{31}a_{13} & a_{31}a_{12} - a_{11}a_{32} \\ a_{12}a_{23} - a_{22}a_{13} & a_{21}a_{13} - a_{11}a_{23} & a_{11}a_{22} - a_{21}a_{12} \end{bmatrix} \tag{5.107}$$

where the determinant is given by:

$$\det = a_{11}(a_{22}a_{33} - a_{32}a_{23}) + a_{21}(a_{32}a_{13} - a_{12}a_{33}) + a_{31}(a_{12}a_{23} - a_{22}a_{13}) \tag{5.108}$$

The use of (5.106) allows for different acceleration parameters to be specified for the three equations being solved. Experimentation has shown that convergence is most rapid for $\omega_\psi = 1.6$ and $\omega_n = \omega_p = 1.0$. The iteration is repeated until the changes being made to the δ values becomes negligible, at which point ψ, n and p are updated for the next Newton iteration. This procedure is repeated until the maximum changes in $\psi_{i,j}$, $n_{i,j}$ and $p_{i,j}$ are less than $\delta\psi_{max}$, δn_{max} and δp_{max} as defined for the decoupled procedure. The overall accuracy should, therefore, be equivalent in both procedures.

This completes the description of the numerical model and it will now be used to provide information about the internal operation of a number of different power devices.

References.

5.1 J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

5.2 H. K. Gummel, "A Self-Consistent Iterative Scheme for One-Dimensional Steady State Transistor Calculations," IEEE Trans. Electron Devices, **ED-11**, pp. 455-465 (1964).

5.3 G. D. Smith, *Numerical Solution of Partial Differential Equations: Finite Difference Methods*, Clarendon Press, Oxford, 1985.

5.4 H. H. Heimeier, "A Two-Dimensional Numerical Analysis of a Silicon N-P-N Transistor," IEEE Trans. Electron Devices, **ED-20**, pp. 708-714 (1973).

5.5 J. W. Slotboom, "Computer-Aided Two-Dimensional Analysis of Bipolar Transistors," IEEE Trans. Electron Devices, **ED-20**, pp. 669-679 (1973).

5.6 R. S. Varga, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, New Jersey, 1962.

5.7 H. L. Stone, "Iterative Solution of Implicit Approximations of Multidimensional Partial Differential Equations," SIAM J. Numer. Anal., **5**, pp. 530-558 (1968).

5.8 W. F. Ames, *Nonlinear Partial Differential Equations in Engineering*, Academic Press, New York, 1965.

5.9 B. V. Gokhale, "Numerical Solutions for a One-Dimensional Silicon *n-p-n* Transistor," IEEE Trans. Electron Devices, **ED-17**, pp. 594-602 (1970).

5.10 G. D. Hatchtel, R. C. Joy and J. W. Cooley, "A New Efficient One-Dimensional Analysis Program for Junction Device Modeling," Proc. IEEE, **60**, pp. 86-98 (1972).

5.11 O. Manck, H. H. Heimeier and W. L. Engl, "High Injection in a Two-Dimensional Transistor," IEEE Trans. Electron Devices, **ED-21**, pp. 403-409 (1974).

5.12 E. M. Buturla and P. E. Cottrell, "Simulation of Semiconductor Transport using Coupled and Decoupled Solution Techniques," Solid State Electron., **23**, pp. 331-334 (1980).

Chapter 6. Results.

Several studies which have been carried out with the aid of the numerical model will be described in this chapter. In addition to the development of the model a certain amount of time has been devoted to the design of a number of different bipolar transistor geometries. The mask design, fabrication and eventual characterisation of these devices will also be discussed in this chapter. The period required for fabrication extended over many months during which time much of the numerical model was being developed. The masks had to be designed at a very early stage in the project which meant that little time was available for optimizing the mask designs on the basis of predictive results from the model. The various items that have been covered will be described in chronological order and the way in which each item leads on from the last will then become apparent.

The first task was to make an investigation into the curvature related avalanche breakdown of diffused junctions near a mask edge. Of particular interest are the several field relieving techniques that can serve to substantially improve the reverse breakdown voltage. This study was made possible by the use of a small section of the full numerical model which was the first part to be developed. Having completed this, the mask set for the bipolar transistors was designed. Some considerations made on the basis of the previous analysis of the field relieve schemes were included in the mask design. In this way the voltage handling capability of the devices could be maximized. Upon completion of the device design the full electro-thermal model was developed. The model was initially used to indicate the optimum power bipolar geometry for maximum pulsed power handling. This was followed by an analysis of transistor switching under inductive loading conditions. Such an analysis was made possible by the coupling of the numerical model with the collector circuit equation. In a final study the coupled electrical and thermal phenomena associated with thermal second breakdown were modelled.

6.1 An Investigation into the Techniques for Improving the Curvature Related Breakdown Voltage of Diffused Junctions.

6.1.1 Introduction.

A wide variety of techniques are available to protect curved junctions, which arise at mask edges, against premature reverse breakdown. As stated in chapter 1 these techniques include field plates, field limiting rings, 'resurf' layers and *p-π-n* structures. In the following section a number of additions and subtractions to the existing model are described which will allow it to be used to calculate breakdown voltages. This part of the model only requires a solution to Poisson's equation and as such is comparatively simple, and yet it is extremely powerful as it can be applied to the most complicated geometry without necessarily incurring any loss of accuracy.

6.1.2 Breakdown Voltage Modelling Considerations.

The voltage profile of a reverse biased junction may be accurately modelled by solving Poisson's equation as described in section 5.1.2 for the decoupled approach. A solution to the current equations can be avoided by assuming that the quasi-Fermi levels extend without variation through the device with their values being given by the applied voltage at the appropriate contact. Thus, the electron quasi-Fermi level is set equal to the voltage applied at the contact to the *n*-type side of the junction, and the hole quasi-Fermi level is equal to the voltage applied to the *p*-type side. Once initialized, the quasi-Fermi levels are held fixed for the duration of the solution procedure. The electrostatic potential is initialized to give space-charge neutrality at all mesh points through the use of the Maxwell-Boltzmann approximations.

The quasi-Fermi levels are, therefore, separated by the applied potential difference across the junction, and consequently the condition of thermal equilibrium ($pn = n_i^2$) in the neutral regions is violated. Such a positioning of the Fermi-levels results in a *pn*-product that is very much less than n_i^2. The consequences of this on Recombination/Generation processes and leakage currents have not been assessed as the continuity equations are not solved. Its effect on the space charge term is negligible because the minority carrier concentrations are in general always too small to significantly affect the space

charge in neutral regions at low injection levels. The solution for electrostatic potential has been found to be almost entirely unaffected by the assumption of flat Fermi levels [6.1] and such 'off-state' models have been used to good effect by many workers in the past (cf. chapter 1).

A typical run is shown in Figure 27 for the case of a simple p^+-n junction in one dimension, with a p^+ region that is assumed to have a Gaussian impurity distribution. Here the potential profile has been plotted after every Newton iteration of the decoupled procedure, which requires a single iteration of the SIP as stated in section 5.1.6.2. It may be observed that the slope of the initial step profile becomes smaller with each iteration until the converged solution is obtained in approximately 20 iterations.

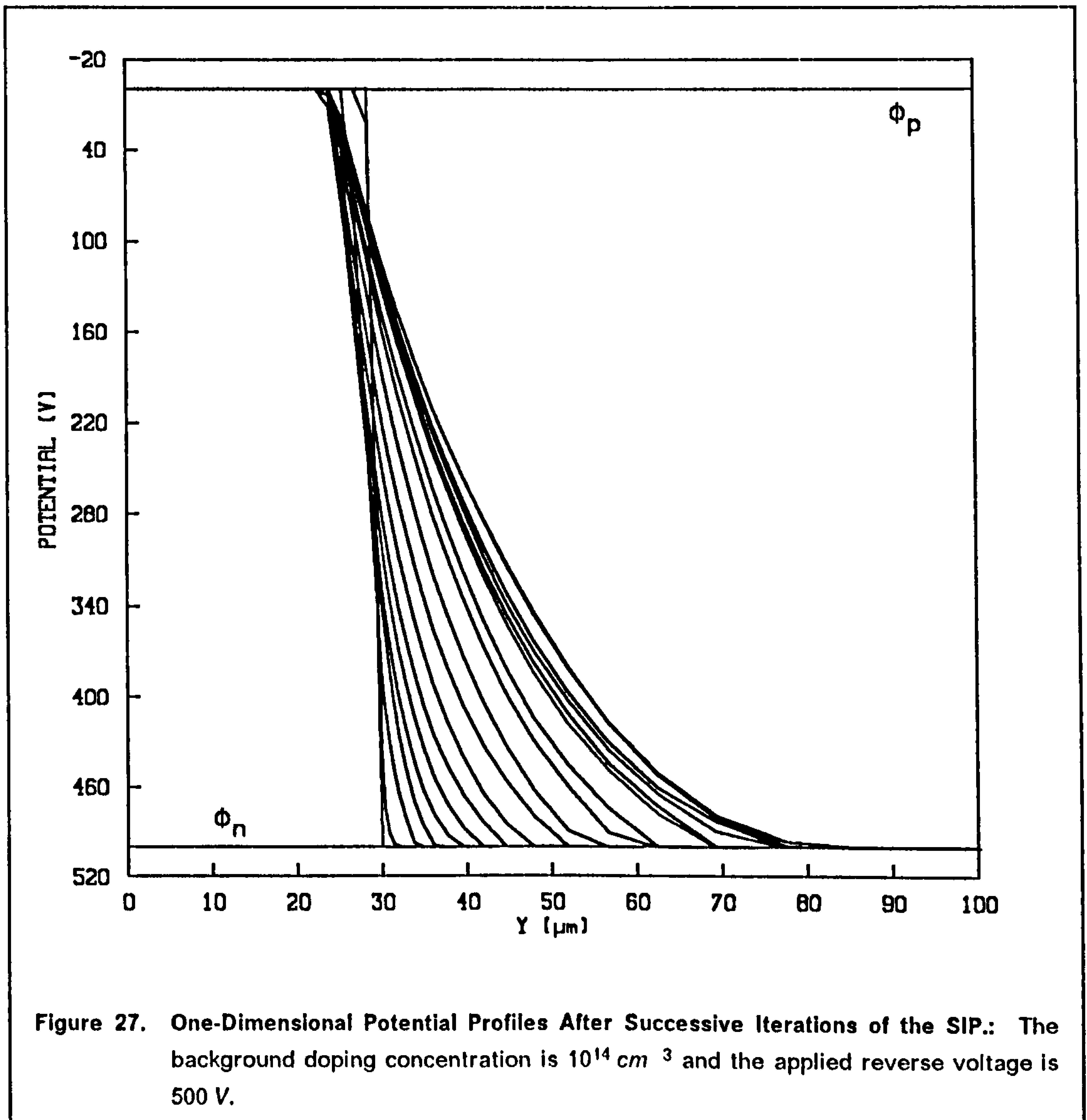


Figure 27. One-Dimensional Potential Profiles After Successive Iterations of the SIP.: The background doping concentration is 10^{14} cm^{-3} and the applied reverse voltage is 500 V.

Having obtained the converged potential distribution it is then necessary to solve the ionization integral in order to ascertain as to whether or not the applied reverse voltage is above or below that to cause avalanche breakdown. The ionization integral for pure hole injection may be written as follows.

$$I_p = 1 - \frac{1}{M_p} = \int_{z_p}^{z_n} \alpha_p \exp\left(\int_{z_p}^{z} (\alpha_n - \alpha_p)\, dz' \right) dz \qquad (6.1)$$

where z is directed along a field (or flux) line, M_p is the hole multiplication factor $(= J(z_n)/J(z_p))$, z_p and z_n are the locations, along the flux line, of the depletion layer edges on the p-type and n-type sides and α_n and α_p are the ionisation rates for electrons and holes respectively. Breakdown occurs when $M_p \rightarrow \infty$ in which case $I_p = 1$. This condition applies regardless of whether multiplication results from carriers that enter the depletion layer by diffusion from the end regions or from carriers produced by charge generation at deep levels within the depletion layer. The field dependencies of the ionisation rates have been taken into account using Chynoweth's law according to the measurements of Van Overstraeten and De Man [6.2], as follows.

$$\alpha_n(E) = 7.03 \times 10^5\, cm^{-1} \exp\left(\frac{-1.231 \times 10^6\, V\, cm^{-1}}{|E|} \right) \qquad (6.2)$$

in the range $1.75 \times 10^5\, V\, cm^{-1} \leq |E| \leq 6 \times 10^5\, V\, cm^{-1}$ and

$$\alpha_p(E) = 1.582 \times 10^6\, cm^{-1} \exp\left(\frac{-2.036 \times 10^6\, V\, cm^{-1}}{|E|} \right) \qquad (6.3)$$

in the range $1.75 \times 10^5\, V\, cm^{-1} \leq |E| \leq 4 \times 10^5\, V\, cm^{-1}$. For the purposes of modelling the range of validity of these equations has been assumed to be extrapolated down to a field of $5 \times 10^4\, V\, cm^{-1}$. Such an extrapolation is required as significant multiplication can occur down to this field value. There seems to be no physical evidence to suggest that the ionization rates will deviate considerably from their values given by (6.2) and (6.3) in this sightly lower field range [6.2].

From physical considerations the upper limit of I_p is unity, however, evaluation of the double line integral can yield values greater than this if the applied voltage exceeds the breakdown voltage. Although physically implausible, this fact can be used to good advantage since the magnitude of the I_p value provides an indication as to how far the applied voltage is from breakdown. It has

been found that the flux line along which most multiplication takes place is nearly always the one that passes through the point of maximum electric field. This fact is due to the exponentially varying nature of the ionization coefficients with field. The exception to this rule occurs when high fields are sustained over substantial regions of a device. In this case care must be taken to locate the flux line giving maximum multiplication.

Before the ionization integral can be solved the variation of α_n and α_p with z must be found. This has been done by initially obtaining values of α_n and α_p at discrete points along the flux line, which will then allow the ionization integral to be evaluated numerically. The technique used to follow the flux line will be described with the aid of Figure 28

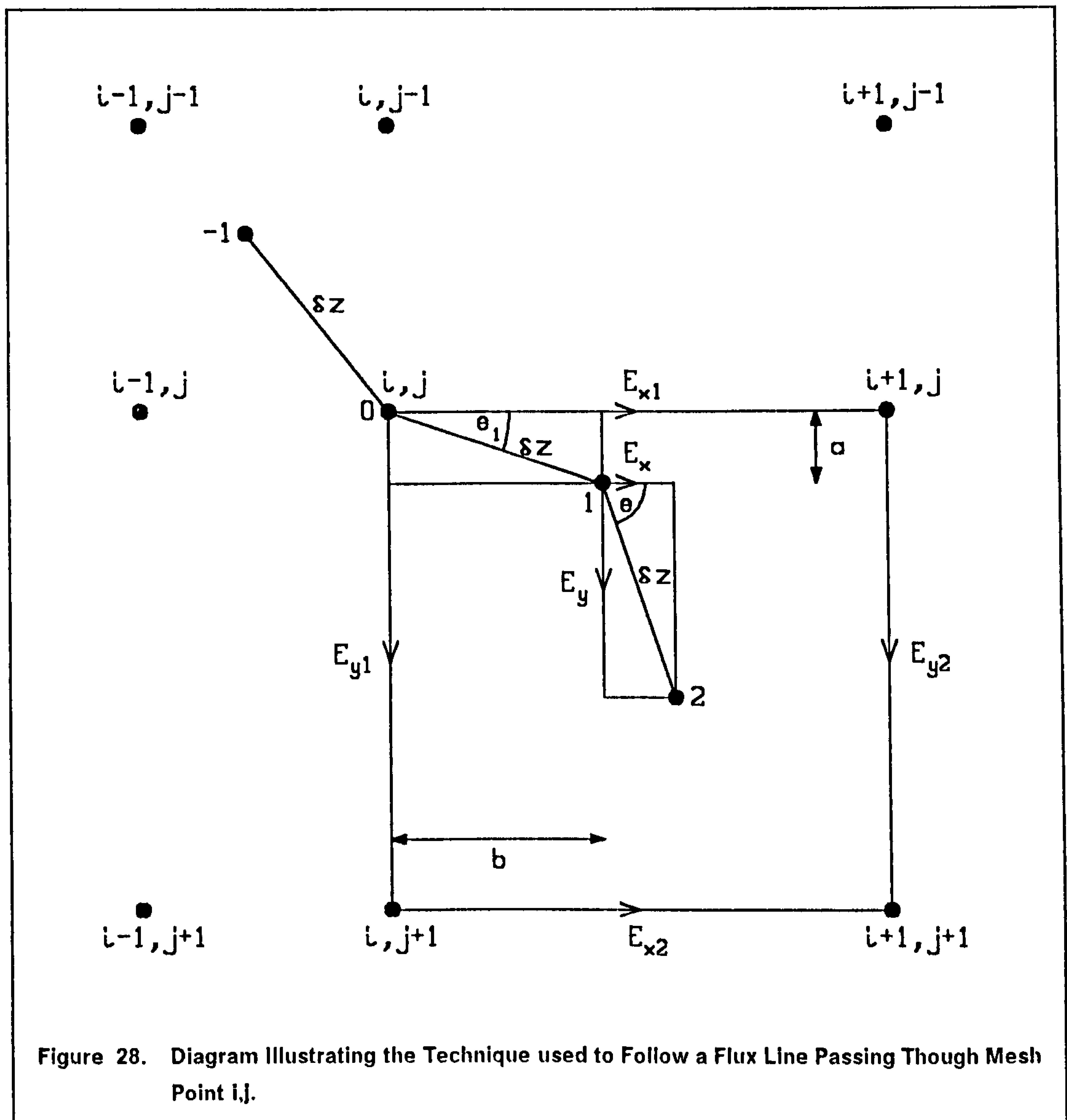


Figure 28. Diagram Illustrating the Technique used to Follow a Flux Line Passing Though Mesh Point i,j.

Firstly the ionization rates are calculated from (6.2) and (6.3) at the mesh point coinciding with the maximum field in the device, which is represented by point 0 on the flux line in Figure 28. In solving Poisson's equation it was assumed that the internodal electric field values were constant and so the fields E_{x1}, E_{x2}, E_{y1} and E_{y2} shown in Figure 28 are given by:

$$\begin{aligned} E_{x1} &= (\psi_{i,j} - \psi_{i+1,j})/h_i , \quad & E_{x2} &= (\psi_{i,j+1} - \psi_{i+1,j+1})/h_i \\ E_{y1} &= (\psi_{i,j} - \psi_{i,j+1})/k_j , \quad & E_{y2} &= (\psi_{i+1,j} - \psi_{i+1,j+1})/k_j \end{aligned} \tag{6.4}$$

The angle θ is given simply by $\tan^{-1}(E_{y1}/E_{x1})$. The procedure then continues by calculating $\alpha_n(E)$ and $\alpha_p(E)$ at equally spaced intervals, δz along the flux line. The nominal value of δz was taken to be one thousandth of the width of the parallel plane depletion layer, W_{pp} associated with the junction. This is simply given by:

$$W_{pp} = \sqrt{\frac{2\,\varepsilon_{sil}\,\varepsilon_o\,V_R}{q\,N_B}} \tag{6.5}$$

where N_B is the impurity concentration on the low doped side of the junction. Any further reduction in δz would in general cause a change only in the third decimal place of I_p. The location of the first point along the flux line (labelled 1 in Figure 28) is given by

$$x_L = x_i + \delta z\ \cos\theta , \qquad y_L = y_j + \delta z\ \sin\theta \tag{6.6}$$

This point does not in general coincide with the mesh and in order to interpolate the fields the four mesh points which form a box surrounding this point must be found. If the flux line has entered a new box then the fields given by (6.4) must be recalculated for the new box. These fields are then linearly interpolated on to the flux line using the following.

$$E_x = E_{x1} + \frac{(E_{x2} - E_{x1})}{k_j}\,a , \quad E_y = E_{y1} + \frac{(E_{y2} - E_{y1})}{h_i}\,b \tag{6.7}$$

with a and b being defined in Figure 28. Having calculated the ionization rates at this point the location of the next point is then found by applying:

$$\theta = \tan^{-1}\left(\frac{E_y}{E_x}\right), \quad x_L \Rightarrow x_L + \delta z\ \cos\theta , \quad y_L \Rightarrow y_L + \delta z\ \sin\theta \tag{6.8}$$

This procedure of stepping along the flux line, locating the surrounding mesh points, interpolating the fields and then calculating α_n and α_p is repeated until the field drops below the minimum value required for significant ionization ($\simeq 5 \times 10^4\ V\,cm^{-1}$). The whole procedure is then repeated while following the flux

line in the opposite direction, starting from the maximum field point. Care must taken to ensure that the flux line does not enter the silicon dioxide layer. This can be achieved by constraining y_L to be greater than or equal to the distance of the $Si\text{-}SiO_2$ interface from the top boundary of the simulation domain. Also, if the flux line should coincide with the $Si\text{-}SiO_2$ interface then the y-directed field component should be set to zero if it is directed upwards, since this component does not contribute to any multiplication.

Once $\alpha_n(i)$ and $\alpha_p(i)$ have been obtained at all points, i along the flux line then the ionization integral can be evaluated by simple summation of strips.

$$I_p = \sum_{i=i_{min}}^{i_{max}} \frac{\alpha_p(i)}{\alpha_n(i) - \alpha_p(i)} \left[\exp\left(\sum_{i_{min}}^{i} (\alpha_n(i) - \alpha_p(i))\delta z \right) - \exp\left(\sum_{i_{min}}^{i-1} (\alpha_n(i) - \alpha_p(i))\delta z \right) \right] \tag{6.9}$$

This can be implemented in Fortran with a single DO-loop as follows.

```
  SUM1 = 0.0
  SUM2 = 0.0
  DO 5 I = IMIN, IMAX
     ANMAP = AN(I) - AP(I)
     SUMR = SUM1
     SUM1 = SUM1 + ANMAP * DZ
     SUM2 = SUM2 + AP(I)/ANMAP * ( EXP(SUM1) - EXP(SUMR) )
5 CONTINUE
```

The value of I_p is given by SUM2. More elaborate numerical integration techniques such as the Trapezoidal rule or Simpson's rule proved to be less efficient as they require the use of nested DO-loops. In order to gain confidence in the above scheme the electron ionization integral was also evaluated using a similar technique. This integral is given as follows.

$$I_n = 1 - \frac{1}{M_n} = \int_{z_p}^{z_n} \alpha_n \exp\left(\int_{z_p}^{z} (\alpha_p - \alpha_n)\, dz' \right) dz \tag{6.10}$$

This integral is not such a strong function of voltage as I_p because the electron multiplication factor, M_n is a much weaker function of reverse voltage than

M_p [6.3]. However, I_n and I_p were both found to be unity at breakdown. As a further test the left and right hand sides of the following equality [6.4] were calculated numerically:

$$\int_{z_p}^{z_n} (\alpha_n - \alpha_p)\, dz \;=\; \log_e\!\left(\frac{M_n}{M_p}\right) \tag{6.11}$$

and were indeed found to be equal to within the bounds of computational accuracy limits, giving further prove that the model had been implemented correctly.

Two example results are shown in Figure 29 and Figure 30 for single unprotected diffusions. Figure 29 shows the voltage contours at breakdown in the free space region immediately above the device as well as in the silicon. The p^+ diffusion is assumed to have a Gaussian vertical profile with error function lateral

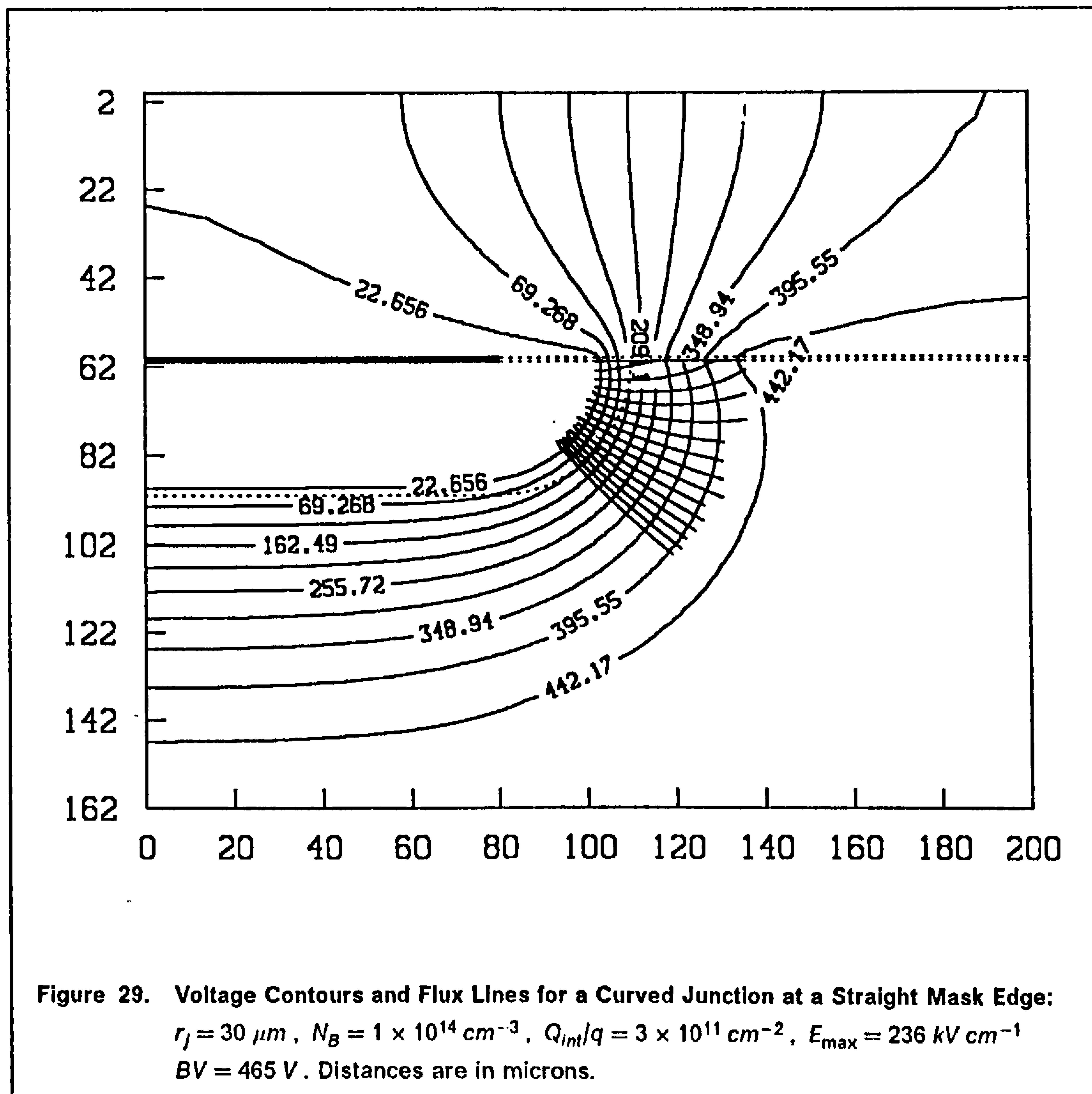


Figure 29. Voltage Contours and Flux Lines for a Curved Junction at a Straight Mask Edge: $r_j = 30\ \mu m$, $N_B = 1 \times 10^{14}\ cm^{-3}$, $Q_{int}/q = 3 \times 10^{11}\ cm^{-2}$, $E_{max} = 236\ kV\ cm^{-1}$ $BV = 465\ V$. Distances are in microns.

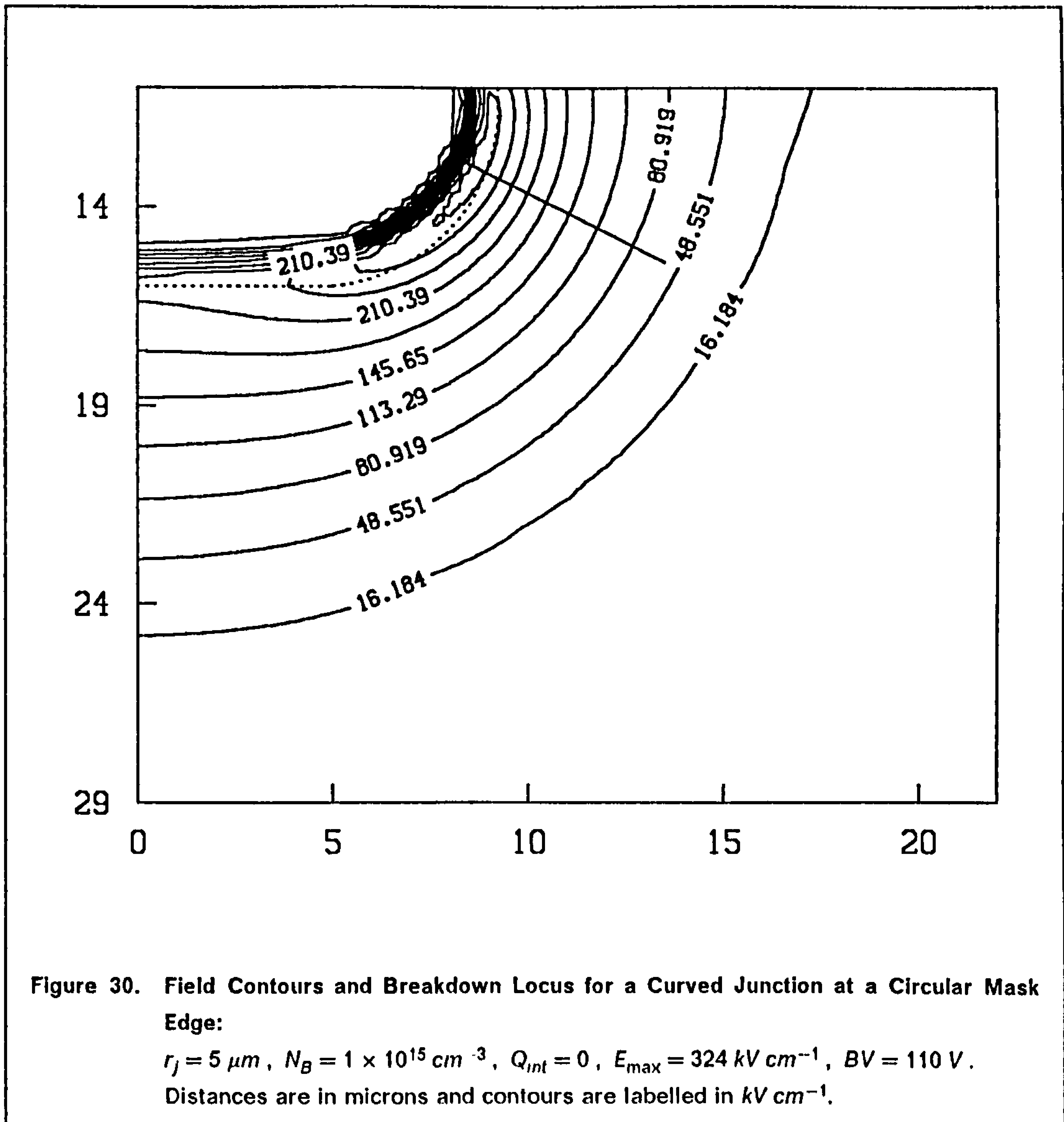


Figure 30. Field Contours and Breakdown Locus for a Curved Junction at a Circular Mask Edge:
$r_j = 5\ \mu m$, $N_B = 1 \times 10^{15}\ cm^{-3}$, $Q_{int} = 0$, $E_{max} = 324\ kV\ cm^{-1}$, $BV = 110\ V$.
Distances are in microns and contours are labelled in $kV\ cm^{-1}$.

roll off. A more detailed description of such profiles will be given in section 6.2. The depletion region is constricted at the Si-SiO_2 interface owing to the presents of positive interfacial oxide charge, Q_{int}.

A number of flux lines are drawn which are normal to the equipotentials apart from those which are constrained to run along the Si-SiO_2 interface. Breakdown occurs along the flux line that is coincident with the Si-SiO_2 interface over its entire length. This is due directly to the presents of positive interface charge which causes the peak field and breakdown point to move from the bulk towards the interface. Figure 30 shows the field contours and breakdown locus for a diffusion which is assumed to have been defined by a circular mask of radius $5 \mu m$, as in this instance cylindrical co-ordinates have been used. Since no

interface charges are assumed to be present the breakdown locus is in the bulk. In both examples the breakdown locus passes through the point of maximum field.

6.1.3 Field Ring Considerations.

The use of field limiting rings was first proposed by Kao and Wolley [6.5]. They consist of additional diffusions positioned adjacent to the main junction and are usually formed at the same processing step as the main junction itself. This is achieved by simply making extra openings in the mask defining the main junction and no extra processing is, therefore required, making field rings very attractive indeed. A diode with a single field ring is illustrated in Figure 31. This is the situation at breakdown where the depletion region encompasses the field ring, which assumes a potential that is intermediate to the anode and cathode voltages, as no external contact is made to the ring. In this case the spacing between the field ring mask edge and the main junction mask edge is $42 \mu m$ and this is the optimum distance giving maximum breakdown improvement for a single field ring. The ionisation integral along both the flux lines is unity at the same reverse voltage of $660 V$. If the ring was further away from the main junction breakdown

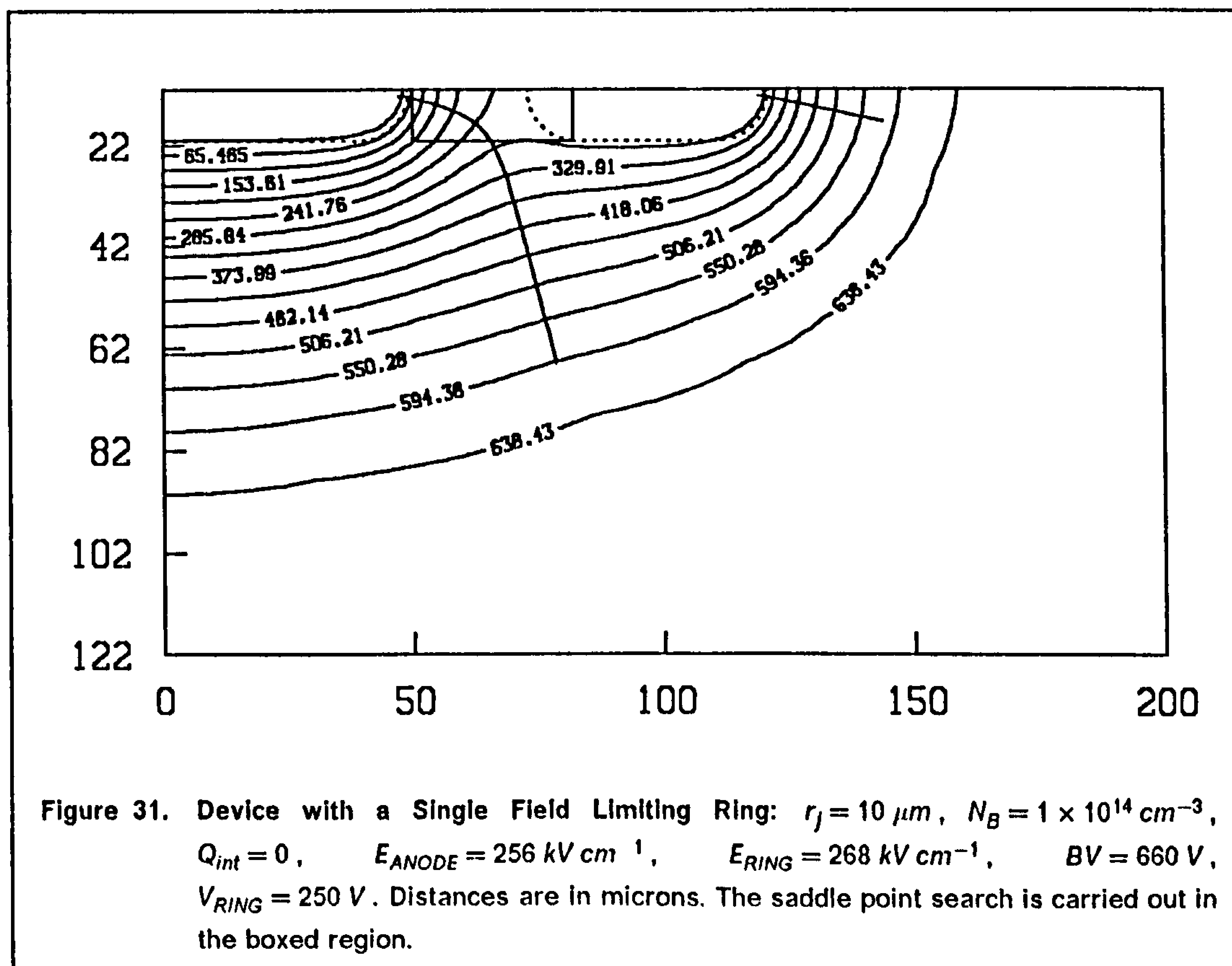


Figure 31. Device with a Single Field Limiting Ring: $r_j = 10\ \mu m$, $N_B = 1 \times 10^{14}\ cm^{-3}$, $Q_{int} = 0$, $E_{ANODE} = 256\ kV\ cm^{-1}$, $E_{RING} = 268\ kV\ cm^{-1}$, $BV = 660\ V$, $V_{RING} = 250\ V$. Distances are in microns. The saddle point search is carried out in the boxed region.

would have occurred at the main junction at some lower voltage. Too small an anode-ring separation will on the other hand result in premature breakdown at the ring junction. The breakdown voltage of 660*V* represents an improvement of 38% over the unprotected junction breakdown voltage.

The potential assumed by the ring is that which results in zero net current flow into the ring under steady state conditions [6.6]. In the case of a p^+ ring, any hole current originating from deep levels in the depletion region or from diffusion into the depletion region, which then flows into the ring must be balanced out by hole current leaving the ring at some slightly forward biased point on the ring junction. Strictly speaking, therefore, a solution to the current continuity equations is required in addition to Poisson's equation, so that the correct ring potential may be obtained. However, this rather drastic measure can be avoided by using the algorithm now to be described.

Initially the ring potential, that is the hole quasi-Fermi level in the metallurgical ring region is set to the voltage applied to the p^+ anode and then Poisson's equation is solved. Having obtained the potential distribution a search is then made for the minimum nodal potential value along the ring junction. This potential is always located on the side of the ring nearest the main junction and the search is restricted to the box shown in Figure 31. The minimum need not necessarily coincide exactly with the metallurgical junction and it should really be defined as the minimum potential on the crest of a vertical ridge of potential that arises between the ring and main junction. If no interface charges are present this minimum will be located at the $Si\text{-}SiO_2$ interface. However, in the presence of interface charge the minimum is relocated at a slightly lower position in the bulk silicon and in this case the minimum lies at the 'saddle' point of the ridge.

The hole quasi-Fermi level in the ring is then reset to the potential at the minimum or saddle point minus the built-in voltage at this point, and the potential in the ring is re-initialized for space charge neutrality using the updated Fermi-level, ϕ_p. This technique implies that the ring is not actually forward biased at the minimum, but that it is under zero bias. Tests have shown that because only a small forward bias is required to account for the low reverse leakage current entering the ring, this approximation results in a ring voltage that may be in error by only a fraction of a volt. The effect on the potential and field distribution is negligible.

The Poisson equation is then resolved and the new ring potential is calculated as described above. This procedure is repeated until the ring voltage converges, which typically takes ten iterations. The technique is illustrated in Figure 32, which shows the potential along the $Si\text{-}SiO_2$ interface of the example

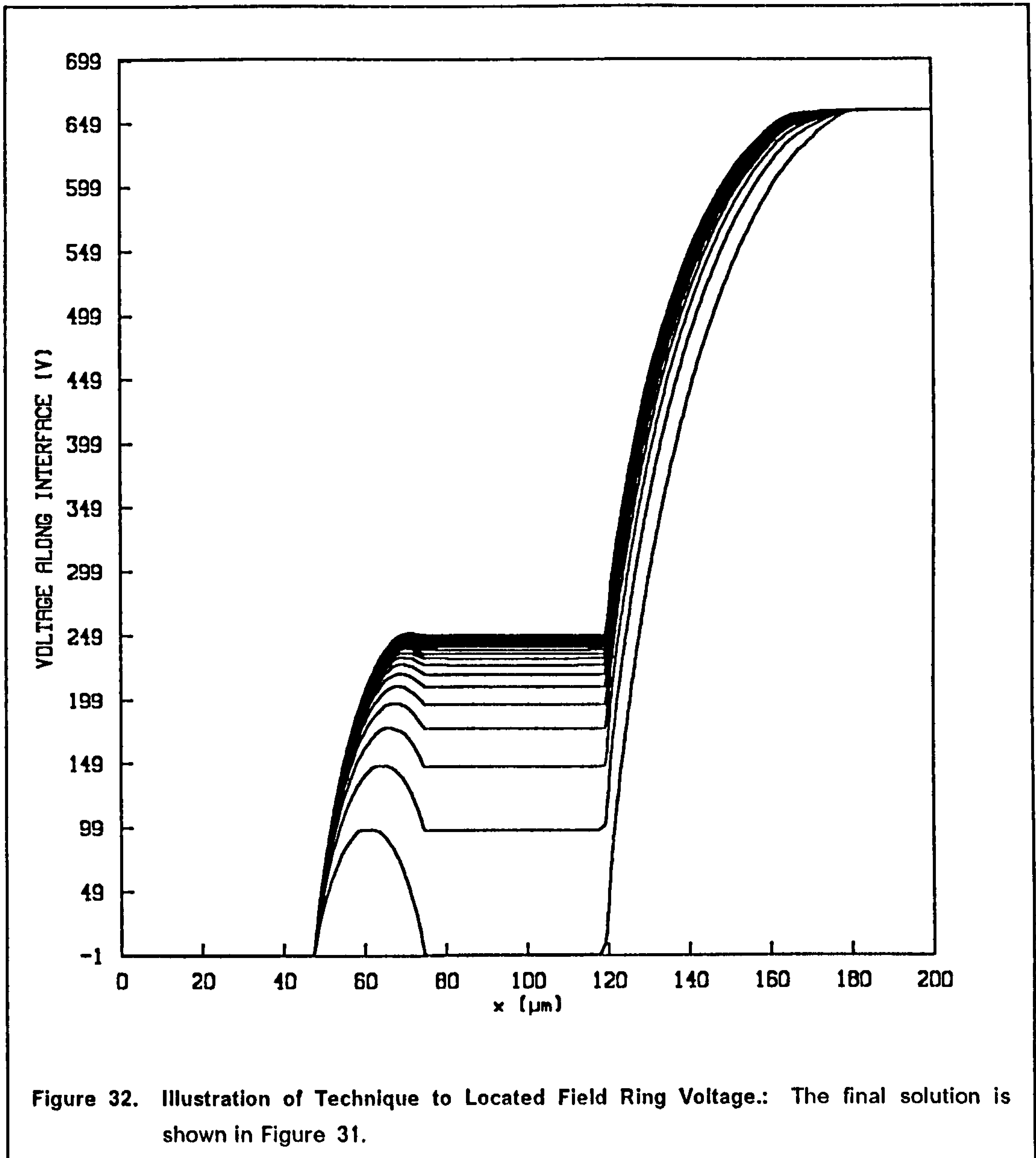


Figure 32. Illustration of Technique to Located Field Ring Voltage.: The final solution is shown in Figure 31.

shown in Figure 31 directly after each Poisson solution. The ring voltage increases rapidly at first and then gradually converges to its final value.

Although this technique for locating the ring voltage was developed independently as part of this project a published article [6.7] has since come to light, which describes a very similar technique. In this paper a comparison was made with ring voltages obtained from experimental ring devices, which showed that the algorithm described above provides accurate ring voltage values. It was also shown that the algorithm can be formally extended to multiple ring systems.

6.1.4 Results.

The 'off-state' model has been utilized in an investigation of various techniques for reducing field enhancement at curved junction regions to improve the breakdown voltage. A fully comprehensive review of the results obtained from this investigation have been published [6.8], and a facsimile of this paper entitled 'A Numerical Analysis of the Resurf Diode Structure' has been included in the Appendix at the end of this thesis. The paper describes virtually all the results obtained, and consequently only a brief description of the results will be made here, with the aid of the figures in the paper. For a detailed account of the results the reader is referred to the original work in the appendix.

The basic structure of the 'resurf' diode is illustrated in Figure 1 of the paper. In addition to the main p^+ junction are are a p^- resurf (standing for REduced SURface Field) layer and an n^+ channel stopper. The p^- layer can be formed either by a second epitaxial process or by ion implantation, which is rapidly becoming the standard technique for introducing dopant into silicon. The latter process is probably the more feasible as it relatively inexpensive and uninvolved. Ion implantation also provides precise control over the amount of dopant introduced into the silicon, which is a prime requisite of the resurf method. For the purposes of simulation the resurf layer has been assumed to be uniformly doped right up to the p^--n^- junction, which is obviously impossible to achieve in practice with either method. Some experimentation is inevitable especially bearing in mind dopant diffusion and complex dopant segregation processes at the Si-SiO_2 interface during subsequent thermal annealing [6.9].

Figure 3 of the paper shows that a significant improvement over the unprotected junction breakdown voltage is possible by a suitable choice of resurf layer doping. A range of possible interfacial oxide charges can be accommodated for by slightly increasing the resurf layer doping. It is significant that the same breakdown voltage improvement can be obtained with or without interface charge. In this respect the resurf layer serves to passify the positive interface charges by providing negative acceptors in the immediate vicinity of the interface. This does not apply to other protection techniques such as field rings and field plates [6.10] where the presence of interface charge results in a reduction in breakdown voltage even after reoptimization. The mechanism by which the layer operates is illustrated in Figure 4 which shows the equi-potentials at breakdown for three different layer doping densities. As the doping is increased the breakdown point moves from the main junction, at or near the Si-SiO_2 interface to a sightly deeper position in the bulk silicon and finally to the resurf layer-channel stopper junction. All results are based on the assumption that the resurf layer and main junction

depths are equal. This condition has been found to give maximum breakdown voltage improvement without affecting the parallel-plane portion of the main junction. The sensitivity of the breakdown voltage to resurf layer depth is indicated by Figure 5, which shows that the resurf depth should be within $1\mu m$ of the main junction depth of $10\mu m$ to guarantee maximum benefit from resurf action. Figures 6, 7 and 8 provide resurf doping data that ensures optimum bulk breakdown regardless of interface charge density. Data is provided for three different background dopings of $1 \times 10^{14}\,cm^{-3}$, $3 \times 10^{14}\,cm^{-3}$ and $1 \times 10^{15}\,cm^{-3}$ for which respective breakdown voltages of up to $1400V$, $600V$ and $250V$ are attainable. The breakdown voltage improvement possible by the use of resurf layers is illustrated by the universal curves of Figure 11. A single universal curve can be plotted that describes the breakdown voltage of unprotected or protected junctions for all junction curvatures and background dopings. This is achieved by normalising the breakdown voltage by the corresponding parallel plane breakdown voltage and plotting this against the junction radius of curvature normalised to the parallel plane depletion width at breakdown. It can be seen that resurf layers give a particularly good improvement in breakdown voltage, which for most cases is over 90% of parallel plane value. This is significantly better than what can be obtained with a single field ring and a good improvement over double and triple field ring systems is obtained for shallow junctions in lightly doped material. In Figure 12 a comparison has been made between the resurf and single field ring techniques. For the resurf case the effect of the interface charge on the breakdown voltage can be effectively nullified by a slight increase in resurf doping, whereas for the ring structure a readjustment of the ring spacing in the presence of interface charge results in a breakdown, which at best is 30% lower than for $Q_{int} = 0$.

Having obtained such encouraging results for the resurf technique it was decided that they should be incorporated into the transistor designs to be fabricated, to protect the base-collector junction. The particular devices in question were required to have an n^- collector layer thickness of $37\mu m$ and an impurity concentration of $1 \times 10^{14}\,cm^{-3}$. The diffused p^+ base junction depth was specified as $6\mu m$. The starting material for this process was obtained by a double epitaxial process with a $6\mu m$ p^- layer being grown on top of a $37\mu m$ epitaxial collector layer. A number of different surface geometries were designed including interdigitated and circular types. Further details of processing and device design are given in section 6.2. Although this collector impurity concentration corresponds to that for which optimum resurf doping data is available from Figure 6 of the paper, unfortunately this data is not applicable as these results assume that the n^- layer is wide enough to allow for free depletion. The model has, therefore, been re-run for the specific geometry being considered and the result, which is similar to

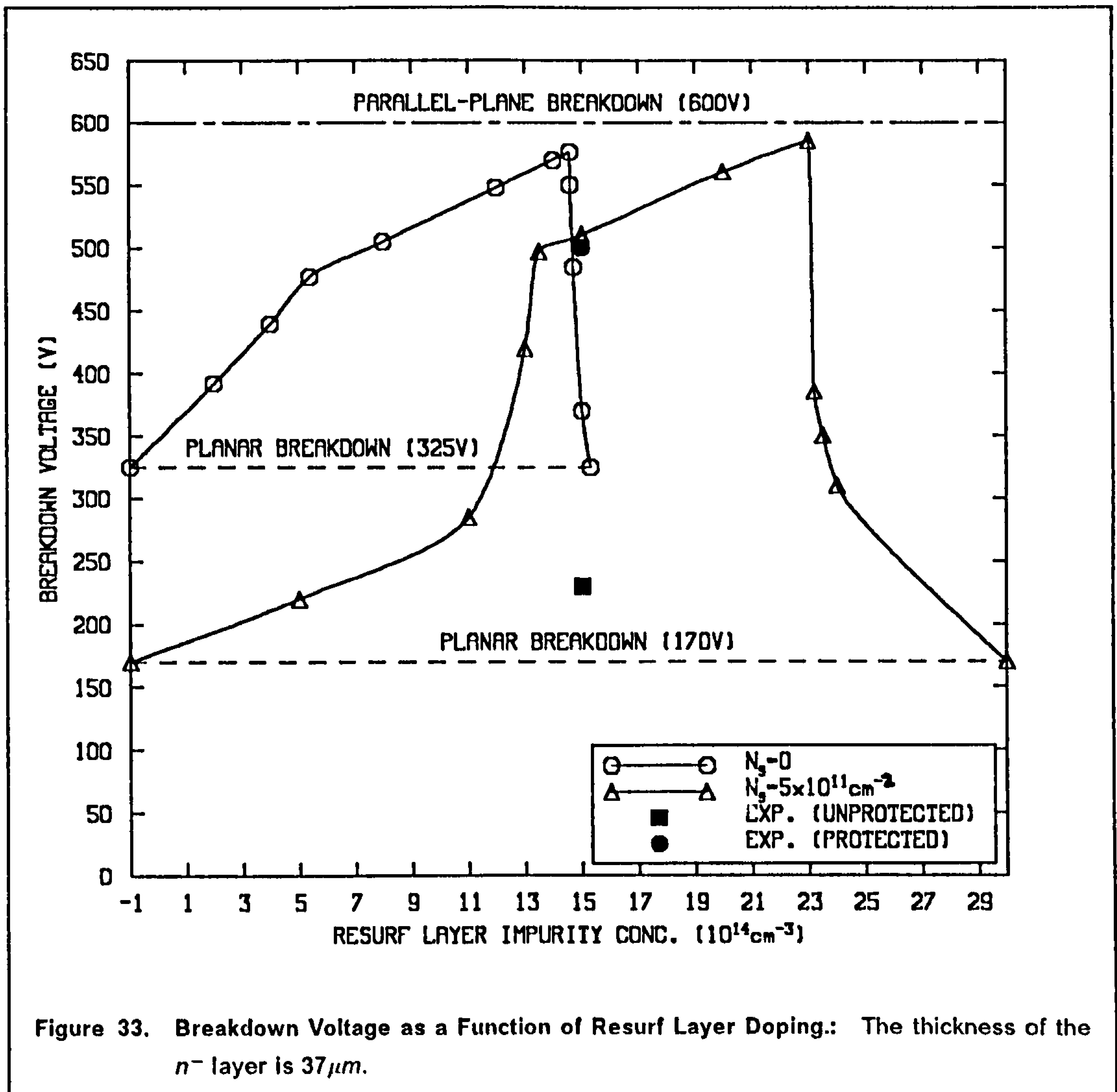


Figure 33. Breakdown Voltage as a Function of Resurf Layer Doping.: The thickness of the n^- layer is $37 \mu m$.

Figure 3 of the paper, is illustrated in Figure 33. On the basis of the computer generated curves a value of $1.5 \times 10^{15}\,cm^{-3}$ was specified for the resurf doping. Devices were fabricated both with and without the resurf layer and the experimental points in Figure 33 indicate that the resurf layer produces a favourable improvement in breakdown voltage. The measured values were consistently obtained to within 5% of the indicated points, and furthermore, there was no notable difference between devices with different surface geometries. The rather low experimental breakdown voltage obtained for the unprotected junction suggests that some interface charge may be present. Although, it could equally be due to differences in junction depth and/or differences in the lateral to vertical junction depth ratio, which has been assumed to be unity for the purposes of simulation.

6.1.5 Conclusion.

An extremely powerful model has been developed, which can be used to predict the breakdown voltage of even the most complicated geometries. The model has been used to characterize the technique of resurf layers and obtain optimum resurf doping levels for maximum breakdown improvement. A suitable resurf layer to be incorporated into a number bipolar devices was designed with the aid of the model, and measurements on these devices have shown that the resurf layer gives a considerable improvement in breakdown voltage.

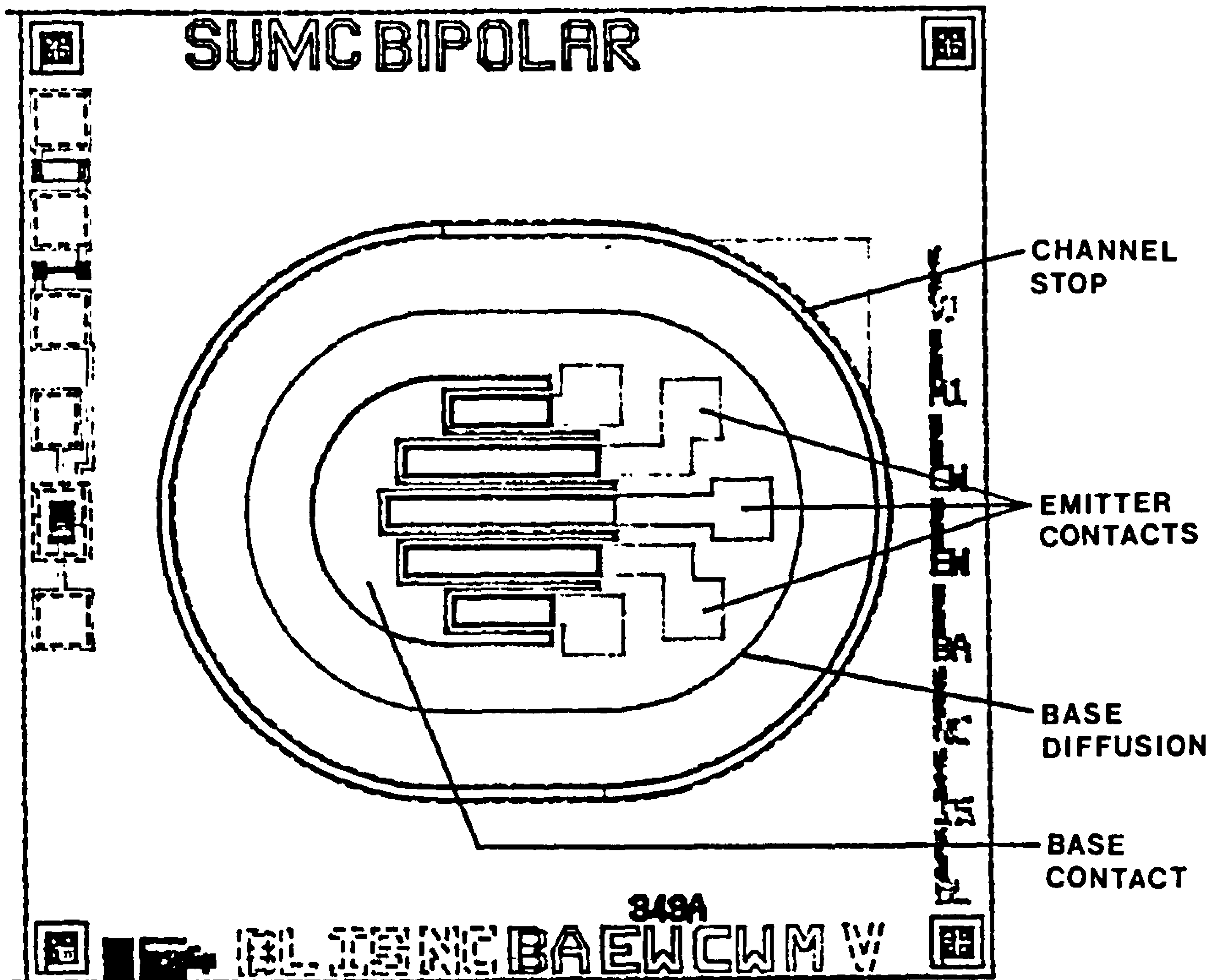
6.2 Design and Characterization of Bipolar Test Structures.

In this section an explanation of the design, fabrication and characterization of the several bipolar geometries will be given. The entire design and development of these devices has been undertaken as part of this project and as a result a great deal is known about them. This makes them ideal for use in verifying the accuracy of the model.

6.2.1 Mask Design.

The lithographic masks for defining device topology were designed using the GAELIC computer graphic facility provided by the Rutherford Appleton Laboratories. Six different geometries were designed, labelled A to F. Each geometry was fabricated on a separate chip and the chips were arranged in a 2 by 3 block, which was repeated across the silicon wafer. The various topologies are illustrated in Figure 34 to Figure 36. The minimum feature size is $5\mu m$ and the bond pads were required to be at least $120\mu m$ square.

Device A has interdigitated base contact windows and emitter fingers. This is a common technique used in power devices to reduce the effects of emitter pinch (cf. section 6.3). Three different emitter finger lengths of $460\mu m$, $390\mu m$ and $200\mu m$ are included and all emitters are $60\mu m$ wide. Each emitter has its own bond pad which allows for characterization of each individual emitter. This arrangement can also be used to investigate current redistribution between emitters, which is a consequence of self-heating under conditions of high power dissipation. The aspect ratio, R (length/width) of the longest emitter was chosen to be similar to that of a commercially available BSX59 transistor for which the aspect ratio is eight. This is an extremely important parameter in the operation of interdigitated



0 2mm

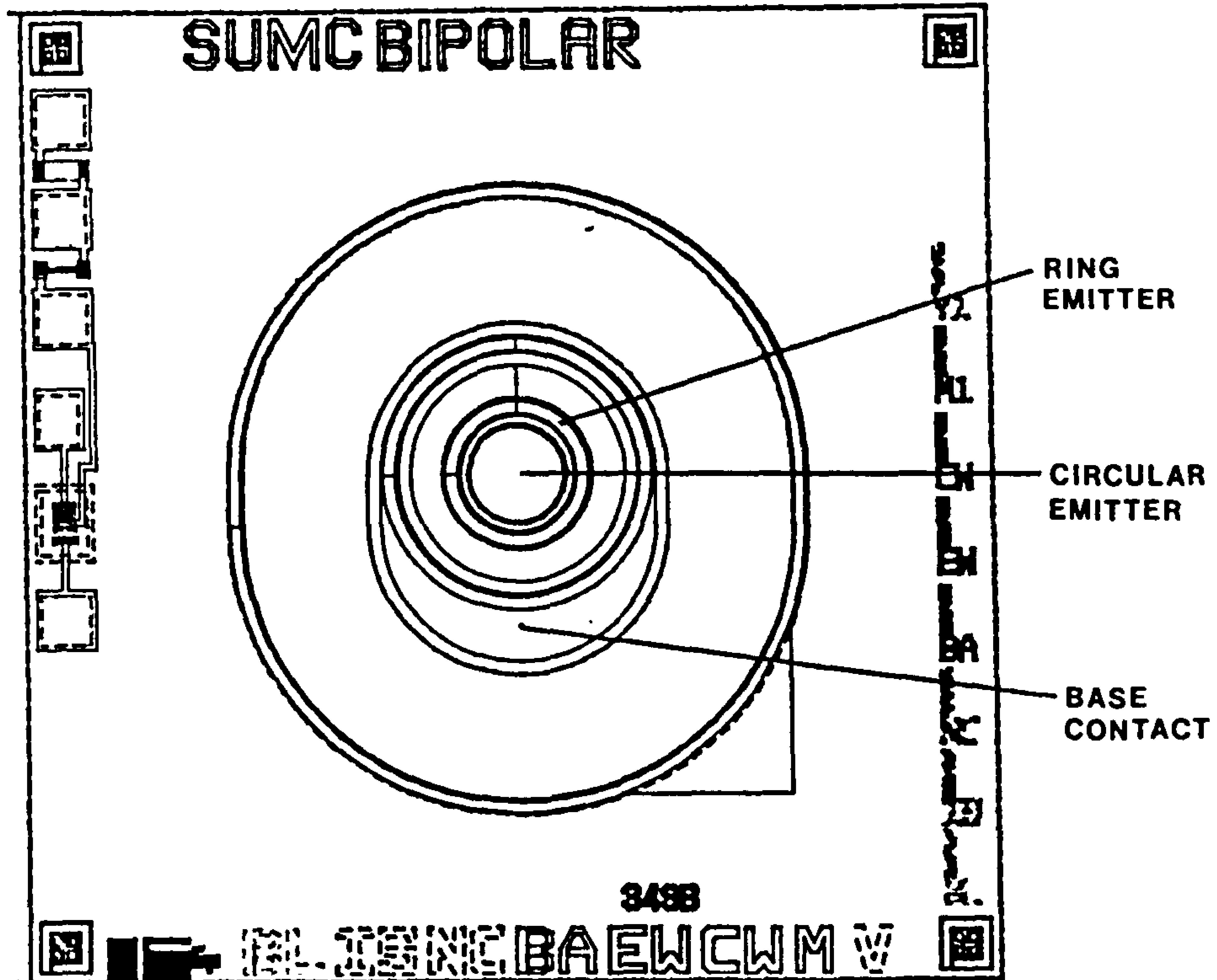
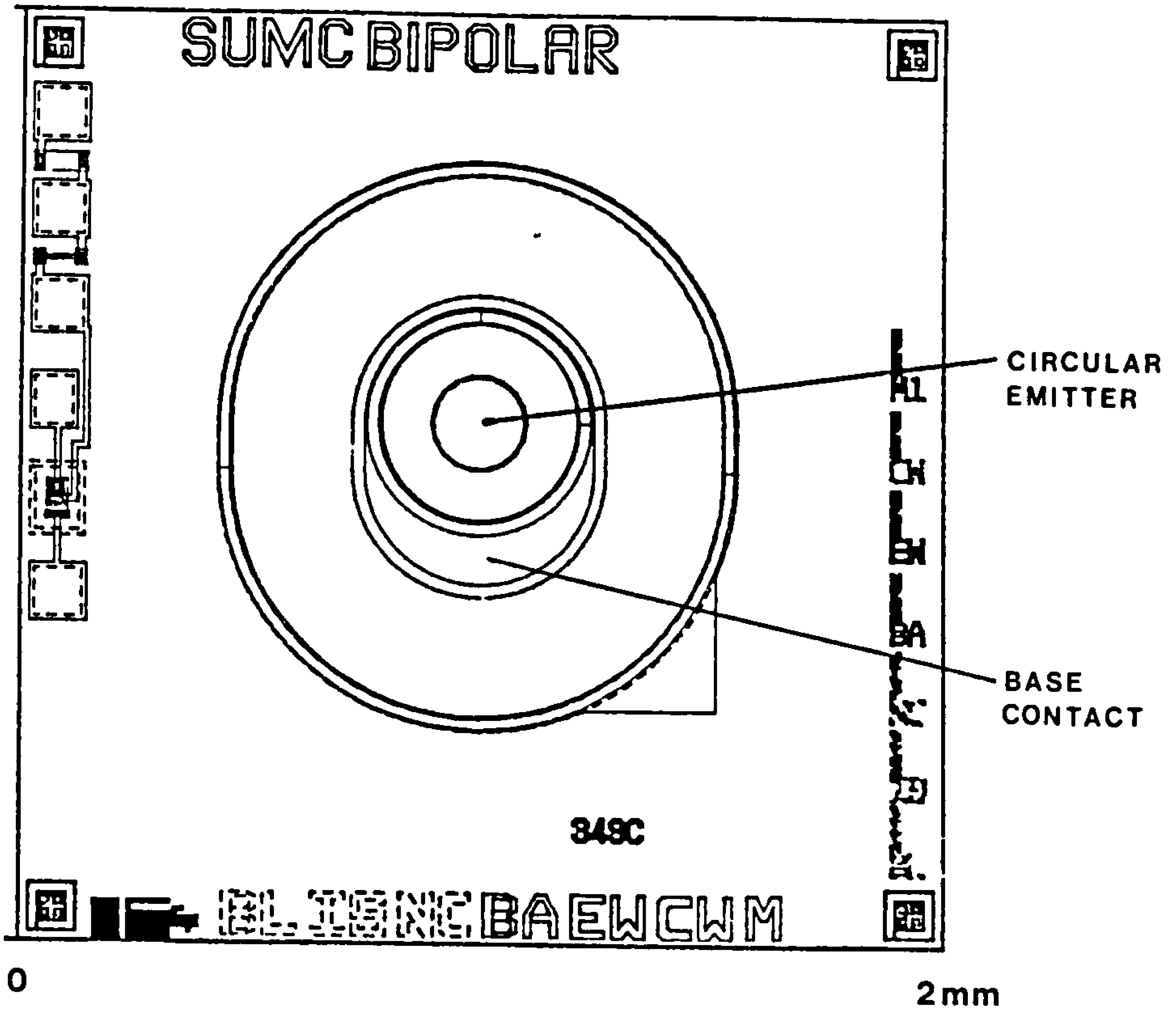


Figure 34. Masks for Devices A and B.



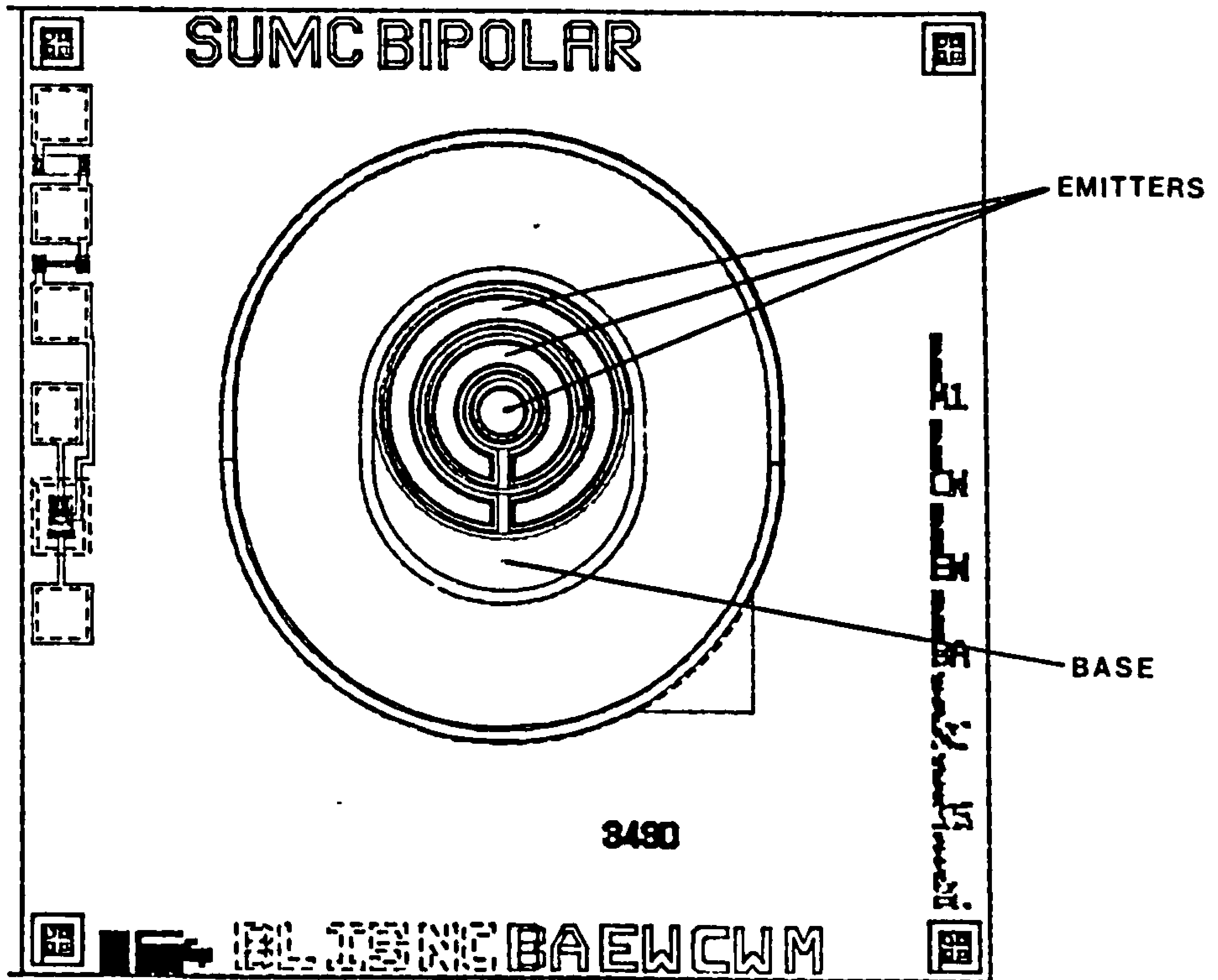


Figure 35. Masks for Devices C and D.

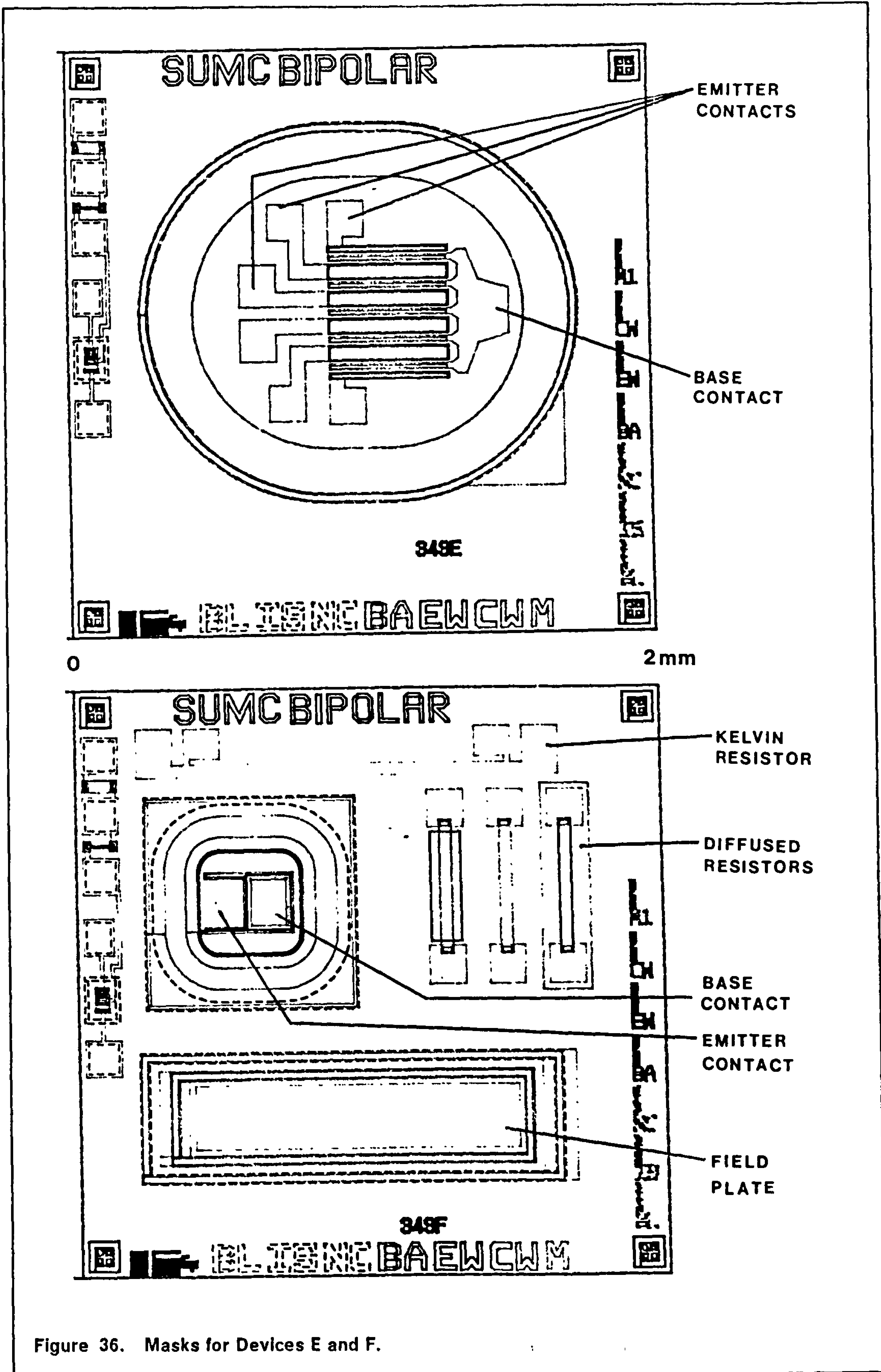


Figure 36. Masks for Devices E and F.

devices. Too large an aspect ratio can result in very uneven current injection from the emitter, with most current flowing from the end of the emitter nearest the bond pad. This is due to the finite resistance of the emitter metalization and the following simple calculation will be used to illustrate the effect. The calculation is performed with the aid of Figure 37, which shows a single emitter finger.

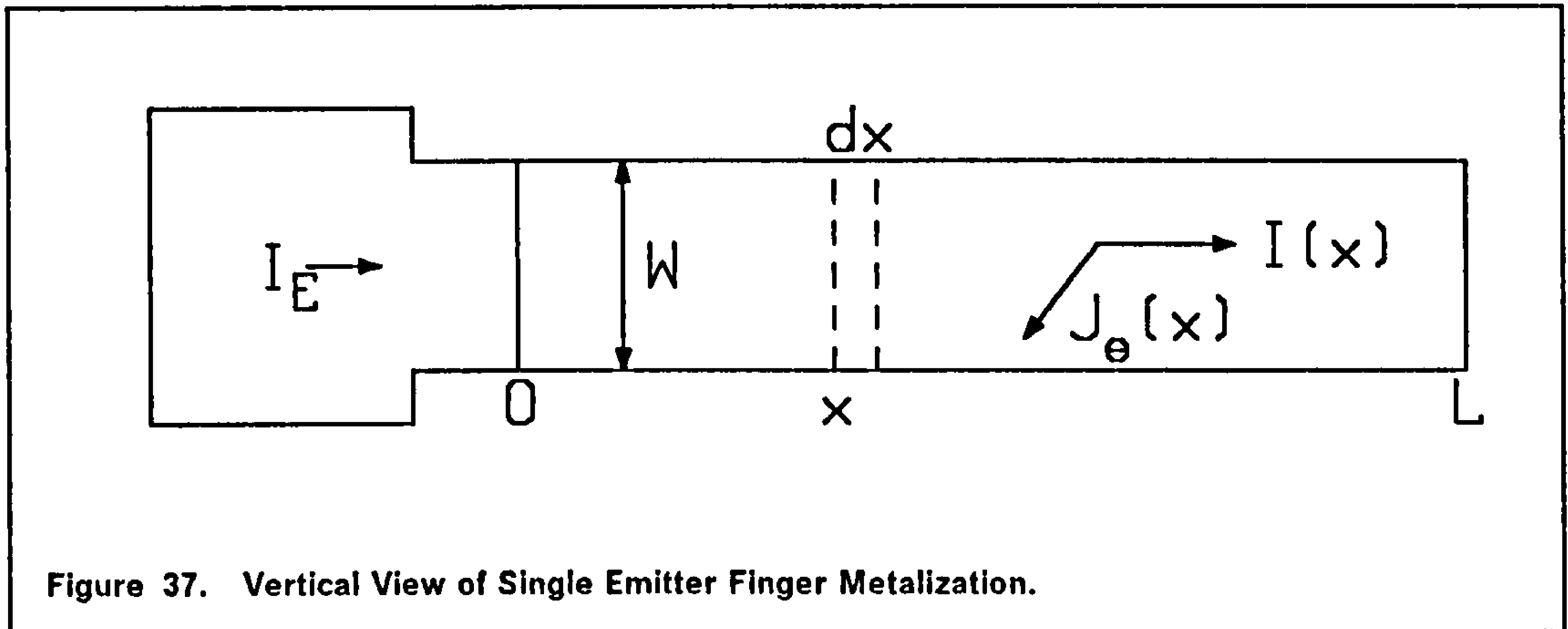


Figure 37. Vertical View of Single Emitter Finger Metalization.

The distance x along the finger is measured f om the end of the emitter diffusion nearest the bond pad. The vertical current density into the semiconductor, $J_e(x)$ can be written:

$$J_e(x) = J_{eo} \exp\left(\frac{q\, V_{BE}(x)}{k\, T}\right) \tag{6.12}$$

where $V_{BE}(x)$ is the base-emitter voltage the current density J_{eo} will be assumed to be constant in this analysis. $J_e(x)$ is defined as the current density at a distance x along the finger, averaged across the width of the finger. The following two equations also hold.

$$dV_{BE} = -\frac{\rho\, I(x)}{W\, t}\, dx \tag{6.13}$$

$$dI(x) = -W\, J_e(x)\, dx \tag{6.14}$$

where ρ is the resistivity of aluminium, $I(x)$ is the current flowing along the contact and t is the thickness of the contact. Inserting (6.12) into (6.14) and substituting for dx from (6.13) into (6.14) gives:

$$dI(x) = \frac{W^2\, t}{\rho\, I(x)}\, J_{eo}\, exp\left(\frac{q\, V_{BE}(x)}{k\, T}\right) dV_{BE}(x) \tag{6.15}$$

Separating the variables and integrating between 0 and L gives:

$$\int_{I_E}^{0} I(x)\, dI(x) = \frac{W^2\, t\, J_{eo}}{\rho} \int_{V_{BE}(0)}^{V_{BE}(L)} \exp\left(\frac{q\, V_{BE}(x)}{k\, T} \right) dV_{BE}(x) \qquad (6.16)$$

Carrying out the integration:

$$-\frac{I_E^2}{2} = \frac{W^2 t}{\rho}\, \frac{k\, T}{q}\, (J_e(L) - J_e(0)) \qquad (6.17)$$

The average current density, J_{av} over the entire emitter area is $I_E/(LW)$ and this allows (6.17) to be rewritten as follows.

$$I_E = \frac{2\, t}{R\, \rho}\, \frac{k\, T}{q} \left(\frac{J_e(0) - J_e(L)}{J_{av}} \right) \qquad (6.18)$$

where R is the aspect ratio (L/W). For maximum current spreading it is desirable to minimize the ratio $(J_e(0) - J_e(L))/J_{av}$ at a given emitter current and this is achieved by decreasing the aspect ratio.

Electrical measurements taken from a test metalization pattern have shown that, in this case, the thickness of the aluminium is $0.477 \mu m$. As an example the longest finger of device A can be expected to handle a maximum current of around $200 mA$. Taking the resistivity of aluminium to be $2.67\ \mu\Omega cm$ at room temperature gives a value of 1.58 for the bracketed term in (6.18) at this current.

As a final point it must be stated that equation (6.18) will overestimate the amount of current crowding if the transistor is biased into high level injection. The pre-exponential term, J_{eo} in (6.12) will also then depend on x and will in fact increase with increasing x. Therefore, equation (6.18) strictly applies to low level operation only. From (6.18) it would seem desirable to make the emitter fingers very short and wide. However, distributed base resistances, which also result in current crowding will impose a limitation on the maximum width of the finger (cf. section 6.3).

Device B has a cylindrical geometry with a central circular emitter encircled by a ring emitter, but with no base contact between them. Again separate contacts are provided for each emitter. In devices A to E the emitter and base metalization was designed not to extend beyond the base-collector junction to avoid the possibility of any inadvertent field plate action. In the case of device B this meant that the base diffusion had to be stretched to make room for the base contact, and as a result the device is not completely cylindrical. The annular emitter has the same area as the central emitter, to allow for a fair comparison between their characteristics. Such a device could be used to investigate the effects of different

types of emitters on device ruggedness, especially with a view to switching inductive loads (cf. section 6.4).

Device C is similar to devicė B except that in this case the ring emitter has been omitted. This is an important device as it is of simple construction and its operation can be modelled using cylindrical co-ordinates as described in chapter 4. The simulations should be in close agreement with device characteristics provided the correct device attributes have been inserted into the model. Simulations of cross-sections through an interdigitated device using cartesian co-ordinates are likely to be less accurate, bearing in mind the variations along the fingers as previously discussed. Thus, the most thorough test for the model would be a comparison with electrical measurements from device C.

Device D has a circular central emitter with two annular emitters. An interlaced base contact is included to minimize emitter pinch. A small section of the annular emitters had to be masked off to allow for this. Once again the individual emitters can be bonded as required. The topology of device E is based on the commercially available BSX59 transistor, the only difference being that each emitter now has a separate bond pad. The four central emitter fingers are $50\mu m$ wide and $400\mu m$ long, whilst the two outer emitters are the same length, but are only $20\mu m$ wide.

Device F is based on a commercially available Ferranti MPSA42 transistor. The construction requires that a portion of the emitter be masked off to allow for a contact to the base. The emitter contact surrounds the base contact and extends over the emitter-base and base-collector junctions to provide field plate protection. The channel stopper contact is also extended over the oxide to give dual field plate action. Several test structures have also been included on this chip. The metalization pattern along the top of the chip was included to provide an estimate for the aluminium thickness. The resistance of the metal strip was obtained by passing a current between the two outer contact pads while measuring the voltage developed across the two inner pads with a high impedance volt meter. This four point probe technique is used to eliminate unwanted lead and contact resistances. The resistance was found to be 1.4Ω, which translates to an aluminium thickness of $0.477\mu m$. Other test patterns include diffused resistors and a large capacitor, which could be used to measure interfacial oxide charge.

6.2.2 Device Processing.

Fabrication was carried out jointly between the Southampton University Microfabrication Centre (SUMC) and Ferranti Electronics plc. The initial stage was performed at SUMC where all diffusions together with the top metalization were

made. Subsequent back preparation of the wafers and final packaging was carried out Ferranti. The accuracy with which the devices can be modelled depends upon a precise knowledge of the doping profiles in the base and emitter regions.

The *p*-type base was formed by ion implantation of boron with a dose of $8 \times 10^{14}\, cm^{-2}$ at an implantation energy of $50\, KeV$. This was then annealed at a temperature of 1150°C firstly in dry oxygen for 10 *mins*, then in an inert ambient for 240 *mins* and finally in wet oxygen for 20 *mins*. Unfortunately, an accurate estimate of the resulting profile cannot obtained from analytical techniques and numerical methods must be employed. A number of models that are designed to simulate device processing in silicon are commercially available. Such models include ICECREM [6.11] and the incredibly well developed SUPREM III [6.12]. These two models can only be used to model processing along a single line through a device, that is they are one-dimensional. More recently two-dimensional models such as SUPREM IV [6.13] and COMPOSITE [6.14] have become available that can be used to obtain dopant distribution near mask edges. However, such models were not available to simulate the process and the vertical doping profile is assumed to be Gaussian. The profile was obtained trom the curves supplied by Irvin [6.15] on the basis of a target junction depth of $6 \mu m$ and a sheet resistance of $112 \Omega / square$, with a background doping of $1 \times 10^{14}\, cm^{-3}$. The data was then fitted to the Gaussian distribution given by:

$$N(y) = N_S \exp\left(- \frac{y^2}{\sigma^2} \right) \tag{6.19}$$

where $\sigma = \sqrt{D\, t}$, $N(y)$ and D are the concentration and diffusivity of the diffusing specie respectively, N_S is the surface concentration and t the total anneal time. The values of N_S and σ which gave the best fit were $4 \times 10^{18}\, cm^{-3}$ and $1.84\, \mu m$ respectively.

An oxide layer of thickness $1700 Å \pm 50 Å$ was grown prior to implantation and this is close to the projected range of the implantation. The peak boron concentration immediately after implantation, therefore, lies close to the silicon surface. The Pearson IV type impurity profile resulting from ion implantation can be closely approximated by Gaussian distributions, especially at low implantation energies such as $50\, KeV$. [6.9]. An important property of the Gaussian profile is that it maintains its Gaussian nature as diffusion progresses. Thus, if the initial profile is Gaussian then to a first order the final profile will also be Gaussian. This is especially true in this case as intrinsic boron diffusivity can be assumed, which is independent of dopant concentration. This is a further requirement for the

validity of (6.19). The use of a Gaussian distribution to approximate the base impurity profile is, therefore, very well founded.

The emitter is formed from an initial phosphorus predeposition at 1000°C for 30 *mins* followed by a drive-in at 1150°C in nitrogen for 30 *mins* and then wet oxygen for 20 *mins*. In this case $\sqrt{Dt}$ for the drive-in should be considerably larger than $\sqrt{Dt}$ for the predeposition and the vertical profile can again be assumed to be Gaussian. From the process schedule the estimated emitter depth was $3\mu m$ and the sheet resistance was $10\Omega/square$. With the aid of Irvin's curves [6.15] the best fit to (6.19) was obtained with $N_S = 7.5 \times 10^{19}\, cm^{-3}$ and $\sigma = 1.27\mu m$.

Ideally the impurity profiles would be measured from the actual manufactured devices, which is most commonly achieved from spreading resistance measurements, capacitance-voltage measurements or secondary-ion mass spectrometry (SIMS) [6.9]. All three techniques have limitations and are very difficult to perform. However, a simple electrical test can be carried out, which provides valuable information about the total charge in the base. This is an extremely important quantity in transistor operation. The classical relation between collector current and base-emitter voltage for current flow in a single direction is given by:

$$I_C = \frac{\overline{\mu}_n k T n_i^2 A}{G(V_{CB})} \exp\left(\frac{q V_{BE}}{k T}\right) \qquad (6.20)$$

where $\overline{\mu}_n$ is the average electron mobility across the base, A is the emitter area and G is called the Gummel number [6.16] and is defined as the number of holes in one square centimetre of the base layer, and is given by:

$$G(V_{CB}) = \int_{y_e}^{y_b} p(y)\, dy \qquad (6.21)$$

where y_e is the emitter junction depth and y_b is the base junction depth. If the collector voltage is increased the depletion region will extend further into the base as well as the collector. and G will fall by an amount ΔG causing I_C to rise by an amount ΔI_C. If V_{BE} is kept constant and $\overline{\mu}_n$ is assumed not to vary with V_{CE} then from (6.20):

$$\Delta I_C = \frac{I_C\, G}{\Delta G} \bigg|_{V_{BE}=const} \qquad (6.22)$$

rearranging gives:

$$G = \frac{\Delta I_C \, \Delta G}{I_C} \Bigg|_{V_{BE}=const} \tag{6.23}$$

The ΔG term can be approximated by assuming that the base-collector junction is a one sided step junction and that the collector impurity concentration is known, thus:

$$\Delta G \simeq \sqrt{\frac{2\,\varepsilon_{sil}\,\Delta V_{CB}\,N_{DC}}{q}} \tag{6.24}$$

Although this equation assumes that the depletion region does not extend into the base, which is inconsistent with the principle behind this technique, quantitatively this is of little consequence, and (6.24) only slightly over estimates ΔG. This analysis was performed under low level operating conditions ($V_{BE} = 0.6V$) to avoid modulation of the base charge and the Gummel number was found to be $4.2 \times 10^{12}\,cm^{-2}$. The profiles described above, when inserted into the model give a value for G of $1.3 \times 10^{13}\,cm^{-2}$. This was reduced to a more acceptable value of $4.5 \times 10^{12}\,cm^{-2}$ by assuming a slightly lower boron concentration at the surface of $1.5 \times 10^{18}\,cm^{-3}$, whilst maintaining the base depth at $6\mu m$. This requires new σ values for the base and emitter of $1.93\mu m$ and $1.19\mu m$ respectively. When annealing in an oxidizing ambient it is found that boron will segregate freely into the oxide [6.9]. It is, therefore, quite possible that the boron concentration is somewhat less than expected, especially at the $Si\text{-}SiO_2$ interface. It is equally probable that the final emitter depth is greater than $3\mu m$, which would also give a reduction in the Gummel number. However, this approach was found to give simulations that were not in such close agreement with device output characteristics (cf. section 6.2.3).

According to Kennedy and O'Brien [6.17] the two dimensional profile arising near a mask edge was modelled using an error function lateral profile as follows.

$$N(x,y) = \frac{N_S}{2}\,\exp\left(-\frac{y^2}{\sigma^2}\right)\left\{1 + \mathrm{erf}\left(\frac{x_{MASK} - x}{\sigma}\right)\right\} \tag{6.25}$$

where x_{MASK} is the location of the mask edge. This equation applies near a right hand window edge. Near a left hand window edge x_{MASK} and x in this expression must be interchanged.

The collector layer thickness and doping was chosen to give an open base breakdown voltage, BV_{CEO} of 250V. It is desirable when designing power bipolar transistors to minimize the collector resistance in order to reduce the associated quasi-saturation effects [6.18]. This can be achieved by utilizing a combination of

high doping levels and thin collector layers, which is able to support the required BV_{CEO}. The resulting specifications are such that the collector will become fully depleted just prior to breakdown and the required donor concentration in the collector is given by:

$$N_{DC} = \frac{2\,\varepsilon_{sil}\,E_C}{q\,W_C} - \frac{2\,\varepsilon_{sil}\,BV_{CEO}}{q\,W_C^2} \tag{6.26}$$

where E_C is the peak electric field at breakdown and W_C is the collector thickness. The resistance of a unit area of the collector material is given by:

$$R = \frac{W_C}{q\,\mu_n\,N_{DC}} \tag{6.27}$$

The resistance of the collector is a minimum when:

$$\frac{dR}{dW_C} = 0 \tag{6.28}$$

This gives,

$$W_C = \frac{3\,BV_{CEO}}{2\,E_C} \tag{6.29}$$

The corresponding optimum doping level is:

$$N_{DC} = \frac{4\,\varepsilon_{sil}\,E_C^2}{9\,q\,BV_{CEO}} \tag{6.30}$$

This gives an optimized collector resistance of:

$$R_{min} = \frac{27\,BV_{CEO}^2}{8\,E_C^3\,\mu_n\,\varepsilon_{sil}} \tag{6.31}$$

The exact value of E_C has been shown to depend on h_{FE} and BV_{CEO} [6.20], and is given by:

$$E_C = 6.4 \times 10^5\,[BV_{CEO}(h_{FE}+1)]^{-1/6} \quad V\,cm^{-1} \tag{6.32}$$

For all practical cases E_C varies between $2 \times 10^5\,V\,cm^{-1}$ for low values of $BV_{CEO}(h_{FE}+1)$ and $1 \times 10^5\,V\,cm^{-1}$ for high values of $BV_{CEO}(h_{FE}+1)$. Hence, in order to ensure that BV_{CEO} is at least 250V the value of E_C was taken to be $1 \times 10^5\,V\,cm^{-1}$. Inserting this value into (6.29) and (6.30) gives $1 \times 10^{14}\,cm^{-3}$ for N_{DC} and 37.5 μm for W_C. This combination provides a minimum on-resistance of 0.145 Ωcm^2.

A number of other collector layers with linearly graded doping profiles were designed. The impurity concentration was required to increase linearly from an initial low value at the base-collector junction to a high value at the collector-substrate junction. This was achieved using epitaxial growth. By grading the doping the critical collector current density at which avalanche injection is triggered can be vastly increased. As the collector current density is increased the electric field profile progressively transfers across the epitaxial layer in a controlled manner, and the maximum sustainable collector-emitter voltage will not fall below the zero current value until the field profile reaches the collector-substrate junction. The consequences of graded collector profiles are fully covered elsewhere [6.20], and their thermal properties will be briefly considered here. The collector profile is given by:

$$N_{DC}(y) = N_{D0} + a\,(y - y_b) \tag{6.33}$$

where N_{D0} is the impurity concentration at the base-collector junction and a is the gradient of the profile. The two specifications considered here were both designed to give a BV_{CEO} of 250V. Both had the same N_{D0} value of $5 \times 10^{13}\,cm^{-3}$ and a value of $1.1 \times 10^{17}\,cm^{-4}$, but they had different thickness's of 50μm and 100μm. Unless otherwise stated the collector specification is assumed to be uniformly doped at $1 \times 10^{14}\,cm^{-3}$ with a thickness of 37.5μm.

Table 7 on page 169 summarizes the values of various parameters used to model the impurity profile, and also the Shockley-Read-Hall lifetimes in the emitter, base and collector regions. Two sets of values are given. The 'new values' are those that have been obtained on the basis of the physical considerations and measurements described above and in section 3.4.3. The 'old values' are estimated values that have been used in all the simulations to be described in sections 6.3, 6.4 and 6.5. These estimates were made prior to a detailed knowledge of the process schedule and also before any lifetime or Gummel number measurements had been made.

6.2.3 Device Characteristics and Comparison with Model.

Some typical output characteristics obtained from the devices are shown in Figure 38. In this case all emitters have been bonded together. The quasi-saturation region [6.18], which is due to the presence of a lowly doped collector layer is clearly visible. The gradient of the line joining the points on each curve at which operation leaves quasi-saturation as shown for device A gives the collector resistance, which in this case is 150Ω. Assuming the vertical electron current density to be uniform over the entire emitter area and that no current

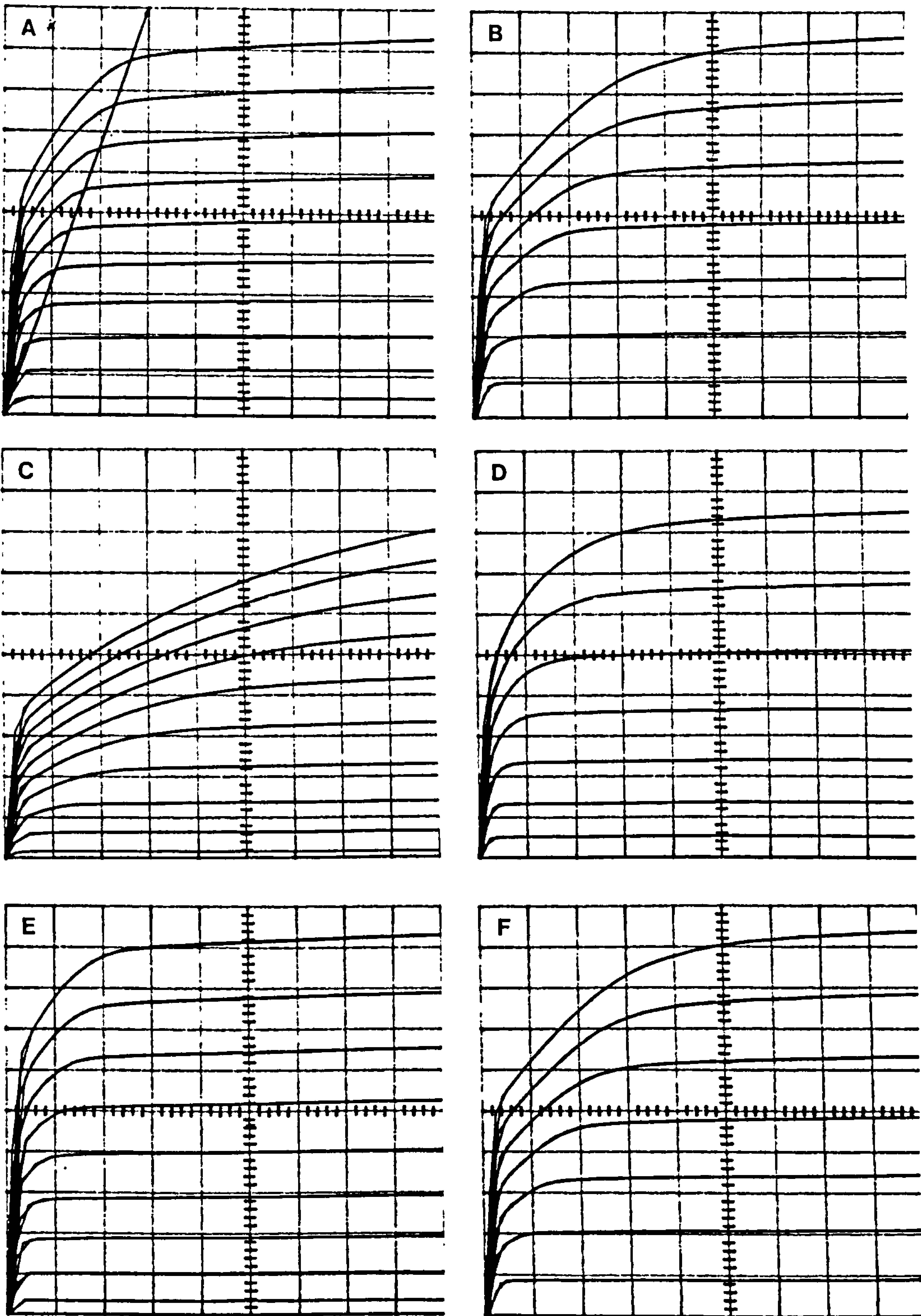


Figure 38. Output Characteristics: The vertical scale is 2 *mA/div* and the horizontal scale is 1 *V/div*. The base current steps are 50 μA except for device A where they are 20 μA.

spreading occurs, then the collector resistance is given by $W_C/(q\mu_{nc}N_{DC}A_E)$, where A_E is the emitter area and μ_{nc} is the low field electron mobility in the collector. Inserting values from Table 7 gives a resistance of 170Ω. The differences can be attributed to current spreading in the collector, which will tend to reduce the resistance, and also to the uncertainty associated with locating the correct locus from the output characteristics.

	'old values'	'new values'
Boron Surface Concentration	$7 \times 10^{17}\ cm^{-3}$	$1.5 \times 10^{18}\ cm^{-3}$
Base Depth	$6\mu m$	$6\mu m$
σ for Boron in (6.25)	$2\mu m$	$1.93\mu m$
Arsenic Surface Concentration	$2 \times 10^{20}\ cm^{-3}$	$7.5 \times 10^{19}\ cm^{-3}$
Emitter Depth	$2.5\mu m$	$3.0\mu m$
σ for Arsenic in (6.25)	$0.934\mu m$	$1.19\mu m$
Collector Doping Concentration	$1 \times 10^{14}\ cm^{-3}$	$1 \times 10^{14}\ cm^{-3}$
Collector Layer Width	$37.5\mu m$	$37.5\mu m$
Emitter Lifetime	$165ns$	$165ns$
Base Lifetime	$2\mu s$	$2\mu s$
Collector Lifetime	$2\mu s$	$30\mu s$

Table 7. Estimated (Old) and Measured/Calculated (New) Parameters.

Figure 39 shows the base and collector currents as a function of V_{BE} for device C with a V_{CE} of $10V$. Both experimental and simulated data has been obtained at temperatures of 27°C and 150°C. These simulations were performed using the 'new values' listed in Table 7. In general all 'on-state' simulations are carried out by firstly applying the 'off-state' model as described in section 6.1 which sets the voltage profile for zero current flow. The decoupled solution procedure is then applied for a low base-emitter bias. The base-emitter voltage is then incremented in steps of $0.05V$, with all subsequent steps being solved with the coupled technique. The use of such small base steps ensures that the solution procedure remains within the radius of convergence of the Newton method. The reason for switching to the coupled technique is that it is much more efficient under high injection conditions as stated in chapter 5.

The error bars on the experimental data in Figure 39 at 27°C indicate the range of data obtained from several devices. Only one set of data was obtained at 150°C and the error bars associated with these points can be expected to be of similar magnitude to those at 27°C. The overall agreement between experiment and simulation is good at 27°C, provided the effects of carrier-carrier scattering given by equation (3.8) are omitted. The close agreement between simulated and experimental collector current curves at 27°C signifies that the Gummel number

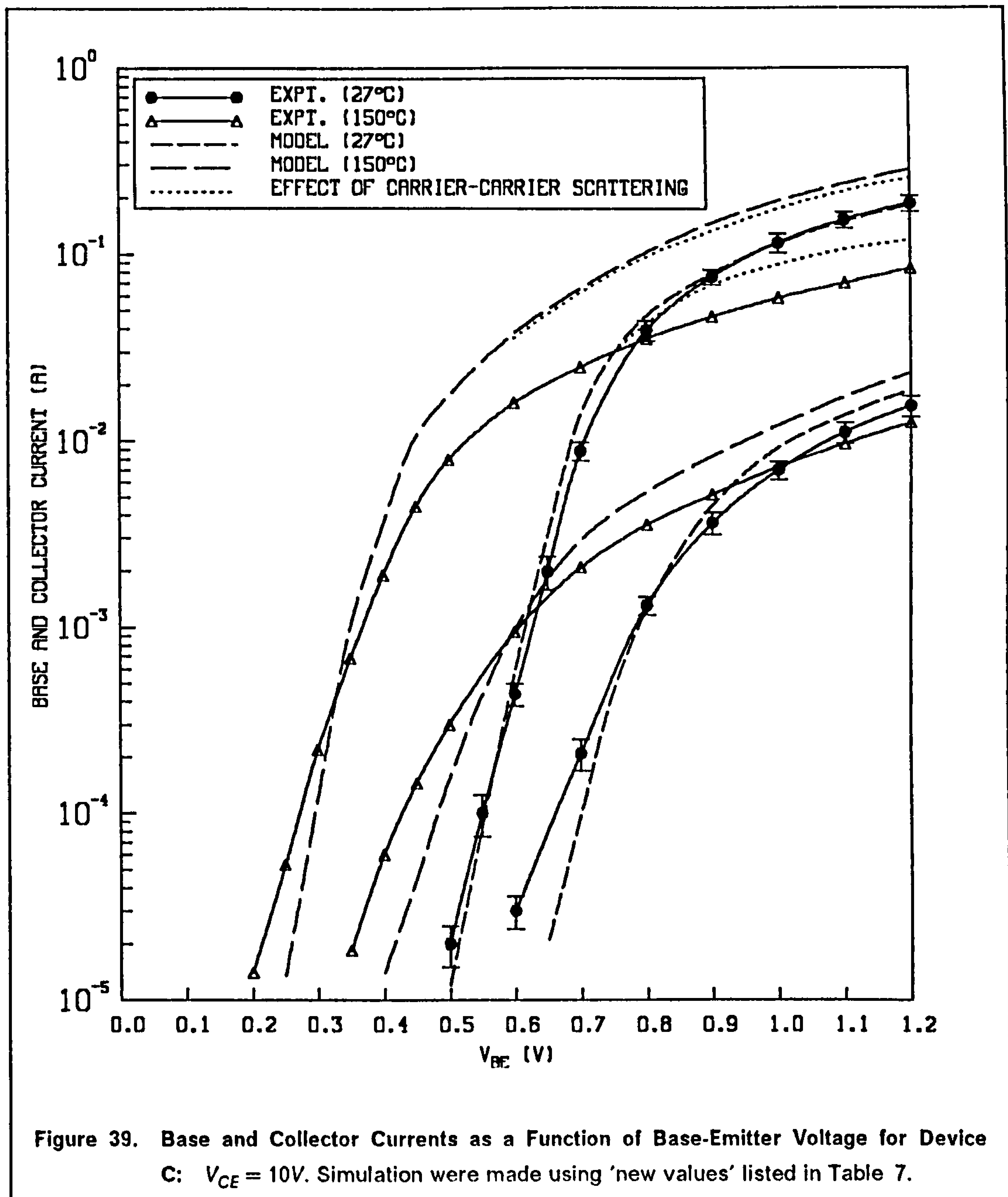


Figure 39. Base and Collector Currents as a Function of Base-Emitter Voltage for Device C: $V_{CE} = 10V$. Simulation were made using 'new values' listed in Table 7.

measurement made earlier is valid. The deviation of simulated and measured base currents at low levels of operation indicates differences between the carrier lifetimes, especially in the region of the base-emitter junction, since at low currents recombination in this region dominates the base current.

The experimental and simulated results at 150°C are not as close. At medium to high injection levels the simulated base and collector currents are too high. In this operating region the amount of base and collector current that flows at a given base voltage is heavily influenced by the lateral base resistance between the base contact and the emitter-base junction. A considerable proportion

of the base voltage is dropped across this resistance leaving less voltage to actively bias the emitter-base junction. It is apparent, therefore, that the model does not predict a large enough increase in base resistance with increasing temperature. This leads one to question the validity of equation (3.7) at higher temperatures. It may be observed from Figure 5 on page 42 that only a very small amount of experimental data is available for comparison with the model, which must be viewed with a considerable amount of uncertainty.

The effect of including carrier-carrier scattering into the mobility model is also shown in Figure 39. Although the collector current is significantly reduced by carrier-carrier scattering at high injection levels the reduction in base current was found to be negligibly small. This is because the base current is dominated by the lateral base resistance between the outer edge of the emitter and the base contact, which is ohmic and is not influenced by high injection. Carrier-carrier scattering can be seen to have a smaller effect at higher temperatures, and this is because the mobility is already reduced because of increased lattice scattering (cf. Figure 6 on page 45). The effect of carrier-carrier scattering is to disrupt the good agreement that would otherwise be obtained between the model and experiment at 27°C , and in this respect the mobility reduction given by (3.8) would seem to be too severe. For this reason, together with the fact that in general very little data is available which reliably quantifies carrier-carrier scattering, equation (3.8) has been omitted from the mobility model in the studies to be described in the following sections. As already stated the 'old values' in Table 7 have been used to describe the devices in the following sections. These values were chosen to obtain a good fit to measured device characteristics at 27°C such as those illustrated in Figure 39. The simulated results calculated with the 'old values' are very similar to those calculated with the new values for steady state problems at room temperature.

6.2.4 Summary.

In this section the design and fabrication of several different bipolar transistor geometries has been described. The devices have been characterized and results obtained from the model have been compared against experimental measurements. Close quantitative agreement was obtained with room temperature measurements, but only qualitative agreement was obtained at higher temperatures. This is reflected by the uncertainty associated with the models for the physical parameters eg. mobility, at higher temperatures.

6.3 Optimization of Emitter Widths.

6.3.1 Introduction.

The numerical model has been used to investigate the potential benefit to be gained in power handling capability from altering the emitter width of an interdigitated bipolar transistor geometry. The simulated devices will be assumed to be biased in the common-emitter configuration, under conditions of high collector voltage and high collector current density. Although considerable self-heating is likely to occur under such conditions a solution to the heat flow equation has been omitted. The inclusion of thermal effects into the solution procedure is not expected to affect the general conclusions of this investigation. Moreover, a solution to the full electro-thermal model requires considerable computational resources because several solutions of the electrical and thermal models are generally required to obtain a self-consistent solution for each bias point or at each instant of time. Thus, if such a solution can be avoided then it should significantly ease the task of device optimization.

Hence, in a strict sense the results to be presented here are only valid for pulsed operation, which would minimize energy dissipation and related self-heating. In addition the mark to space ratio of the pulse chain should be made sufficiently large to avoid any thermal accumulation effects. However, the results obtained from the electrical model were found to favour geometries in which the heat generation is more evenly spread over a large volume of the device. This would tend to minimize both the heating rate and the final steady state temperature attained should the device not fail. These thermal considerations tend to reinforce the recommendations arising from this analysis.

For optimum performance an *n-p-n* transistor should be designed so that the vertical electron current density is as uniform as possible as it flows from the emitter to the collector. This would minimize on-resistance and maximize the power that could be handled safely, as will be shown. Uniformity of current flow is disrupted by lateral resistances across the emitter metalization as described in section 6.2 and also lateral resistances associated with the base layer beneath the emitter. The lateral base current flows through this region tending to increase the forward base-emitter bias at the periphery of the emitter with respect to its centre. This effect is called 'emitter pinch' and it results in a significant difference between the current density at the emitter edge compared with its centre.

The effects of emitter metallization resistances can be alleviated by using a thicker layer of aluminium. However, it is not such a straightforward matter to

decrease the resistance of the base layer, as this would require an increase in the base impurity concentration, which would give rise to a corresponding fall in current gain through its influence on the Gummel number. Hence the need for interdigitation. As a preliminary exercise a comparison will be made between the numerical model and an analytical model [6.19] for emitter pinch. This will serve to illustrate the problem and also to illustrate the need for a numerical model to tackle such a problem.

6.3.2 A Comparison of Analytical and Numerical Models for Emitter Pinch.

Hauser [6.19] showed that the variation in emitter current density from the emitter centre ($x = 0$) to its edge ($x = W_E$), due to the effects of distributed base potential, could be given by,

$$J_E(x) = J_E(W_E) \frac{\cos^2 Z}{\cos^2(Z\, x/W_E)} \tag{6.34}$$

where W_E is half the width of the emitter finger, $J_E(W_E)$ is the peak emitter current density and Z is a dimensionless quantity which may be determined from,

$$Z \tan Z = \frac{I_E}{I_X} \frac{W_E}{L_E} \tag{6.35}$$

where I_E is the total emitter current, L_E is the total length of the emitter finger(s) and I_X has the dimensions of current and is given by,

$$I_X = \frac{2\, V_T\, W_B\, (1 + h_{FE})}{\rho_B} \tag{6.36}$$

where V_T is the thermal voltage (kT/q), W_B is the base width, h_{FE} is the transistor current gain (I_C/I_B) and ρ_B is the reciprocal of the average conductivity in the base region beneath the emitter, that is,

$$\rho_B = \frac{1}{\dfrac{q}{W_B} \displaystyle\int_0^{W_B} \mu_p(y)\, N_B(y)\, dy} = R_S\, W_B \tag{6.37}$$

where $N_B(y)$ is the net impurity concentration in the base and R_S is the sheet resistance of the base layer. The base width can be measured using the angle-lapping and staining technique [6.9] and the sheet resistance can be obtained using the Van der Pauw technique [6.21]. Both these measurements

require extensive preparation and the resulting estimate for ρ_B does not include the effects of base conductivity modulation and base widening, which occur at high injection levels. This represents a major limitation of the analytical model as it stands.

Hauser also indicated that the peak emitter current density could be obtained from the value of the average emitter current density, thus,

$$\begin{aligned} J_E(W_E) &= \overline{J_E(x)} \, \frac{Z}{\sin(Z)\,\cos(Z)} \\ &= \frac{I_E}{2\,W_E\,L_E} \, \frac{Z}{\sin(Z)\,\cos(Z)} \end{aligned} \qquad (6.38)$$

The comparison has been made for a device with an emitter half-width of 100 μm, an emitter length of 1 cm , and no de-biasing is assumed to occur due to the finite emitter metallization resistance. The 'old values' in Table 7 have been used for the impurity profiles and lifetimes and this combination results in a ρ_B of 0.9 Ωcm. The same total emitter current has been assumed for both analytical and numerical cases. The results are shown in Figure 40 and Figure 41 for low and high current conditions, respectively. The agreement is seen to be particularly good for the low current case with the slight differences occurring at the emitter edge being attributed to two dimensional effects, which are not taken into account by the analytical model. More specifically the 'kink' in the numerical curve is due to the injection of base current into the emitter side wall in the numerical model. In the analytical model all the base current is assumed to flow vertically upwards from beneath the emitter. At the emitter edge, however, a certain proportion of the hole current back injected into the emitter originates from a lateral position. Thus, at the emitter edge the vertically injected hole current is less than in the analytical model. This results in an increase in the lateral base current flowing through the base layer towards the centre of the emitter. Therefore the vertical emitter current density near the emitter edge falls off more rapidly with distance than in the analytical model (cf. Figure 40).

The ratio of the vertical back injected hole current density to the vertical forward injected electron current density was found to be reasonably uniform along the emitter-base junction up to within 3μm of the emitter edge. This ratio is much lower at the emitter edge due to sidewall injection as described above. As a result of this the lateral base current falls off more rapidly as it flows towards the emitter centre than in the analytical model, which assumes a value for this current density ratio that is averaged across the entire emitter width. Consequently at locations further than about 3μm away from the emitter edge the base-emitter voltage does not fall off as rapidly as in the analytical model. This results in a

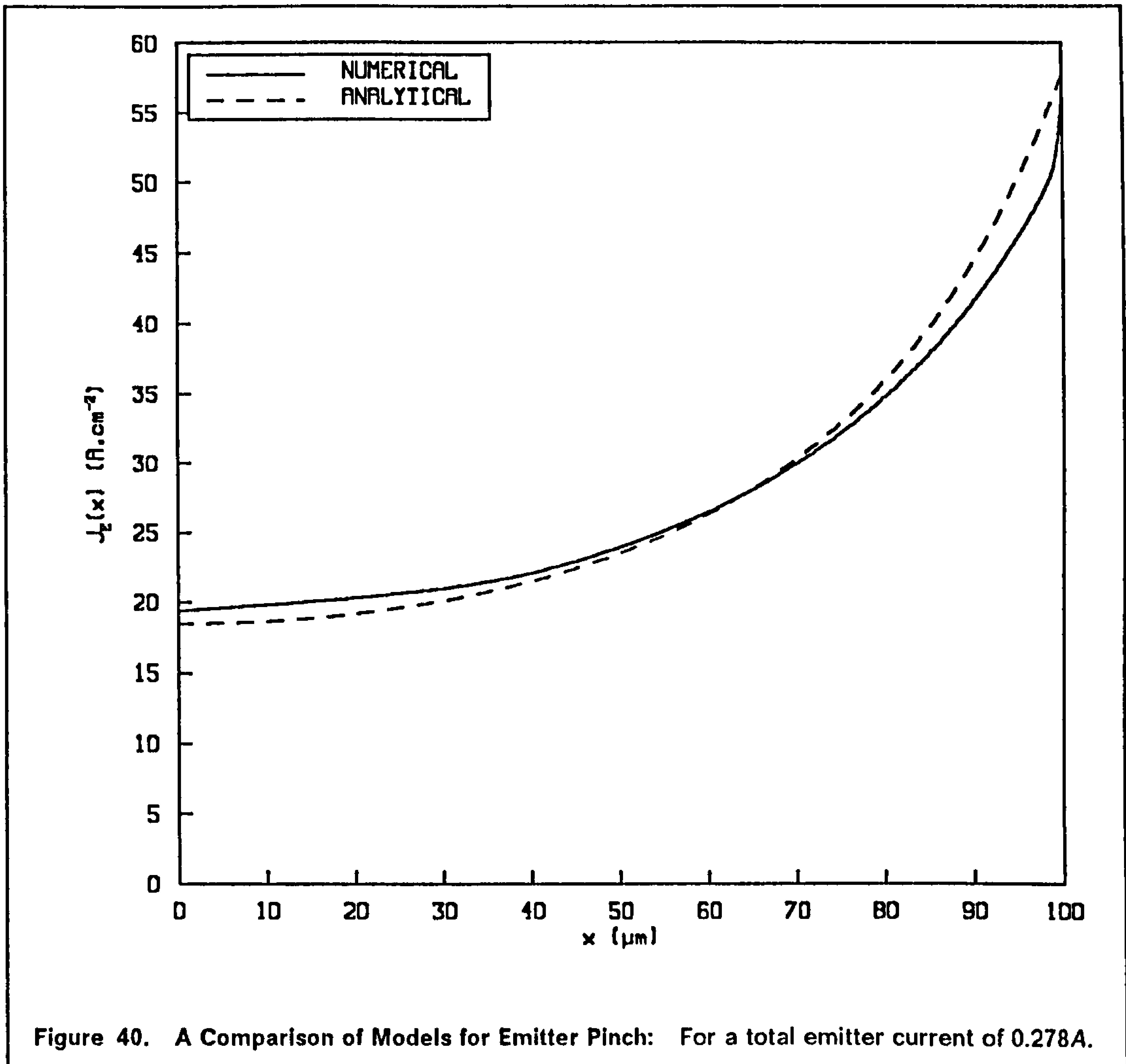


Figure 40. A Comparison of Models for Emitter Pinch: For a total emitter current of 0.278*A*.

vertical electron current density at the emitter centre, which is higher in the numerical model (cf. Figure 40).

For the high current case however the agreement is not as good. The effects of emitter pinch are now more severe in both cases, because the base current is higher causing a greater voltage drop along the base-emitter junction. At these injection levels the current density in the vicinity of the emitter edge is sufficiently large to cause the minority carrier concentration in the base to become comparable with majority carrier concentration there. This results in a reduction in the injection efficiency, $J_E(x)/J_B(x)$ at high emitter current densities. In the analytical model an average value is used for the injection efficiency related to the $(1 + h_{FE})$ term in equation (6.36). In the numerical model, which takes into account high injection effects, the injection efficiency near the emitter edge will be less than the average value. The lateral base current, therefore, falls off more rapidly in moving from the emitter edge to its centre, than is predicted by the analytical

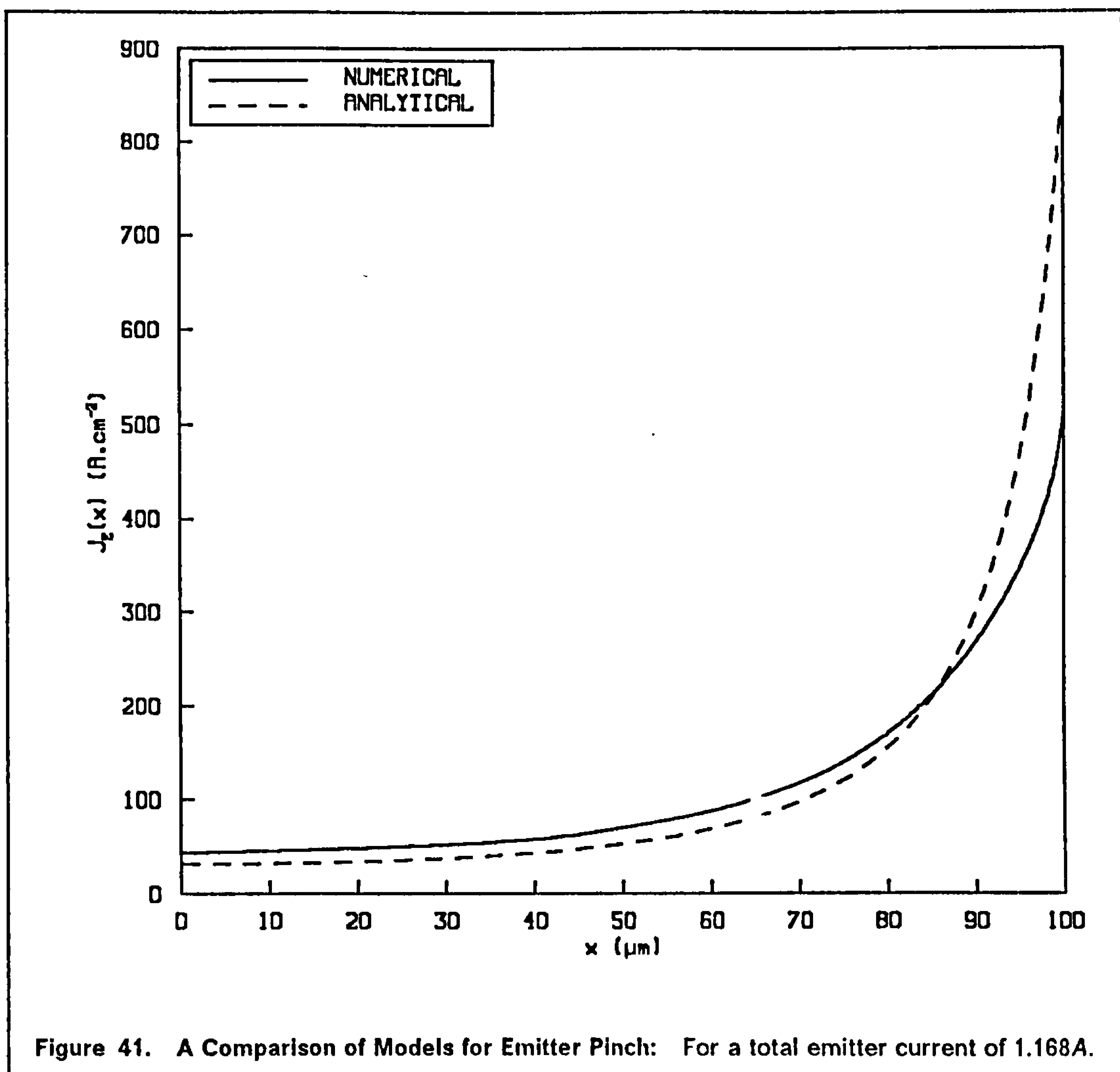


Figure 41. A Comparison of Models for Emitter Pinch: For a total emitter current of 1.168*A*.

model. At the emitter centre the injection efficiency is greater than the average value, and consequently the current density at this point is greater than that predicted analytically. High injection also results in base conductivity modulation resulting in a reduction of the lateral base resistance in regions of high emitter current density. Once base conductivity modulation occurs, emitter pinch tends to remain constant on the basis that a higher J_C causes a higher J_B, but the lateral voltage drop this extra current produces is partially cancelled by the higher base conductivity. Both this effect and the variation of the emitter efficiency along the emitter-base junction tend to moderate the amount of emitter pinching that would otherwise occur.

These high injection effects cannot be easily incorporated into the analytical model, because of the difficulties in measuring sheet resistance and effective base width in non-equilibrium conditions. Even if these values could be found, the variation of these parameters together with the emitter efficiency as a

function of x would also be required. In conclusion, therefore, the full numerical model should be employed to correctly simulate emitter pinch, especially in high level operation, where the analytical model drastically over estimates current crowding.

6.3.3 Optimization of Emitter Width.

The model will now be used to investigate the consequences of changing a particular design parameter, with a view to optimizing device performance. In this section the effect of varying the emitter finger width will be investigated in an attempt to improve the device power rating, which is of primary importance in all power semiconductors.

The importance of emitter pinch in power bipolar transistors was indicated in the previous section. The non-uniform current flow which results as a consequence of emitter pinch can lead to a significant degradation of the combined current and voltage handling capability of the device. It would be extremely desirable to minimize these effects and so obtain more uniform current flow. Moreover, if current flow could be made sufficiently uniform then device operation could be modelled using essentially one dimensional models, which would significantly ease the task of device design.

Since the concentration of injected electrons peaks at the edge of the emitter, then a lateral concentration gradient exists, which will tend to cause the current to diverge as it passes through the base and collector. If the combined thickness of the base and collector layers is sufficient, and the emitter fingers are thin enough, then the vertical current density can become uniform at some point prior to the high field region in the collector. If this can be achieved then it will serve to increase the pulsed voltage and/or current required to initiate avalanche injection.

The 'old values' have been used in this investigation and the cell half-width is taken to be $W_E + 15\,\mu m$, whilst the base contact half-width was taken to be $5\,\mu m$. These values were chosen as they represent the minimum dimensions that are commonly used in conventional bipolar technology. It is desirable to minimize these dimensions so that the emitter area is maximized.

For this analysis the collector to emitter voltage was held constant and the base voltage was incremented in small steps (0.05V), whilst noting the currents and peak fields at each step. This procedure was repeated until the critical field, E_C was reached, at which point the device was considered to enter avalanche breakdown. The critical field was taken to be $1 \times 10^5\,V\,cm^{-1}$, which under estimates the value given by (6.32), resulting in breakdown voltages that are less

than what should be possible experimentally. This test was carried out on a number of designs, all with different emitter widths.

The dependence of the peak field on the collector current can be attributed to the well known Kirk effect [6.22]. At sufficiently high current densities the concentration of mobile electrons in the collector can exceed the background donor concentration there. This occurs because the fields in the collector depletion region are high enough ($> 10^4\, V cm^{-1}$) to cause the velocity of the electrons passing through it to saturate at their limiting drift velocity, v_n^{sat}. If the collector current density, J_C is assumed to be uniform throughout the collector then a simple one dimensional analysis gives:

$$J_C = q\, n\, v_n^{sat} \tag{6.39}$$

where n is the electron concentration in the collector. The diffusion component is omitted, since at the fields under consideration the drift term dominates. The field is then obtained by solving Poisson's equation, which for this case is given by:

$$\frac{dE}{dx} = \frac{q}{\varepsilon_{sil}}\left(N_{DC} - \frac{J_C}{q\, v_n^{sat}}\right) \tag{6.40}$$

This equation is solved subject to the condition that at the base collector junction ($x = 0$) the field is given by $E(0)$. Thus:

$$E(x) = E(0) + \frac{q\, x}{\varepsilon_{sil}}\left(N_{DC} - \frac{J_C}{q\, v_n^{sat}}\right) \tag{6.41}$$

$E(0)$ can be calculated from:

$$\int_0^{W_C} E(x)dx = -V_{CB} \tag{6.42}$$

where V_{CB} is the applied collector-base voltage, which is much larger than the contact potential between the base and collector contacts, which has therefore been omitted. Solving gives:

$$E(0) = -\frac{q\, W_C}{2\, \varepsilon_{sil}}\left(N_{DC} - \frac{J_C}{q\, v_n^{sat}}\right) - \frac{V_{CB}}{W_C} \tag{6.43}$$

Substituting this back into (6.41) gives the required result.

$$E(x) = \frac{q}{\varepsilon_{sil}}\left(N_{DC} - \frac{J_C}{q\, v_n^{sat}}\right)\left[x - \frac{W_C}{2}\right] - \frac{V_{CB}}{W_C} \tag{6.44}$$

If the width of the space charge region is less than the collector width then W_C must be replaced with the width of the space charge region given by:

$$W = \sqrt{\left| \frac{2\, \varepsilon_{sil}\, V_{CB}}{q\left(N_{DC} - \dfrac{J_C}{q\, v_n^{sat}}\right)} \right|} \qquad (6.45)$$

Equation (6.44) has been plotted in Figure 42 for the optimised 250V device and for a collector voltage of 187.5V, which is 3/4 times BV_{CEO} and a reasonable voltage at which the device could be expected to operate. Results are illustrated for various collector current densities. At a current density of 460 $A\,cm^{-2}$ the peak field reaches the critical value for avalanche injection leading to second breakdown. Thus, if the collector current density can be considered to be uniform, then this current density multiplied by the active device area represents the maximum current a given device can handle. Any local increase in current density will give rise to a corresponding increase in the local field, which in turn will initiate avalanche injection at a lower total collector voltage.

A non-uniformity in collector current becomes especially apparent for devices with wider emitter fingers, where the effects of emitter pinch are more pronounced. This is illustrated in Figure 43, which shows how the current flows in four cells with different emitter widths. The collector voltage in all cases is 187.5V and the peak field is $1 \times 10^5\ V\,cm^{-1}$. The devices are, therefore, considered to be operating at a point just prior to breakdown. Each contour in Figure 43 represents a line, such that a constant fraction of the total current flows between it and the emitter centre line. Each contour is labelled with the corresponding current in $A\,m^{-1}$ (Amps per metre length of emitter finger). Equivalently, it may be stated that equal proportions of the total current is constrained between adjacent lines.

The diagram for $W_E/2 = 10\mu m$ shows that for narrow emitter widths very little pinching occurs and the contours are reasonably equally spaced along the emitter-base junction. Once the electrons have been injected into the base they diffuse laterally into the region of low electron concentration beneath the base contact. At the same time they drift vertically towards the substrate under the influence of the electric field. Divergence of current continues until the lateral diffusion gradient of electrons disappears, which is seen to occur approximately half way down the collector. Current flow through the remaining part of the collector is essentially uniform. The electron current leaving the emitter is constrained to flow within the cell half-width due to the zero current boundary conditions imposed along the vertical boundaries, and the current will redistribute itself between these two lines. If the right hand boundary is moved away from the

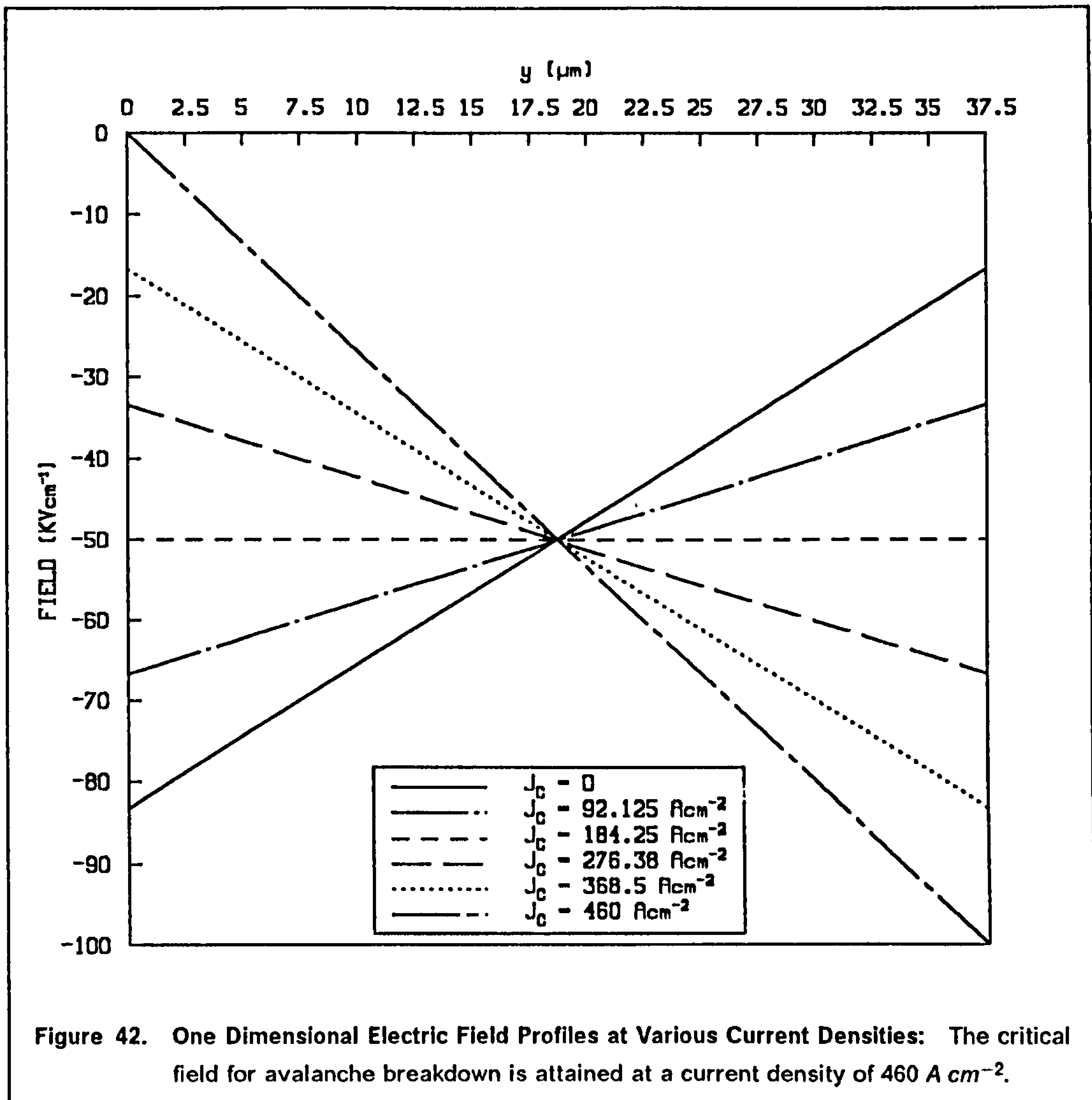


Figure 42. One Dimensional Electric Field Profiles at Various Current Densities: The critical field for avalanche breakdown is attained at a current density of 460 $A\ cm^{-2}$.

emitter a limit would be reached at which the uniformity of collector current is lost. It is therefore necessary to ensure that the gap between the emitters is narrow enough to avoid this problem.

In contrast, the device with an emitter half-width of $50\mu m$ exhibits a significant amount of emitter-pinch, with the contour separation becoming smaller towards the emitter edge. However, a sufficient amount of divergence causes the current density to become virtually uniform in the collector layer as before. For an $80\mu m$ emitter half-width the pinching is greater still and the cell width is much greater than the collector layer thickness. In this case, therefore, the current density is unable to become uniform in the time it takes for the electrons to traverse the base and collector layers. As stated previously this can cause a significant reduction in the pulsed (or surge) power handling capability of the device. A more extreme example of this is shown for $W_E/2 = 120\mu m$ where the

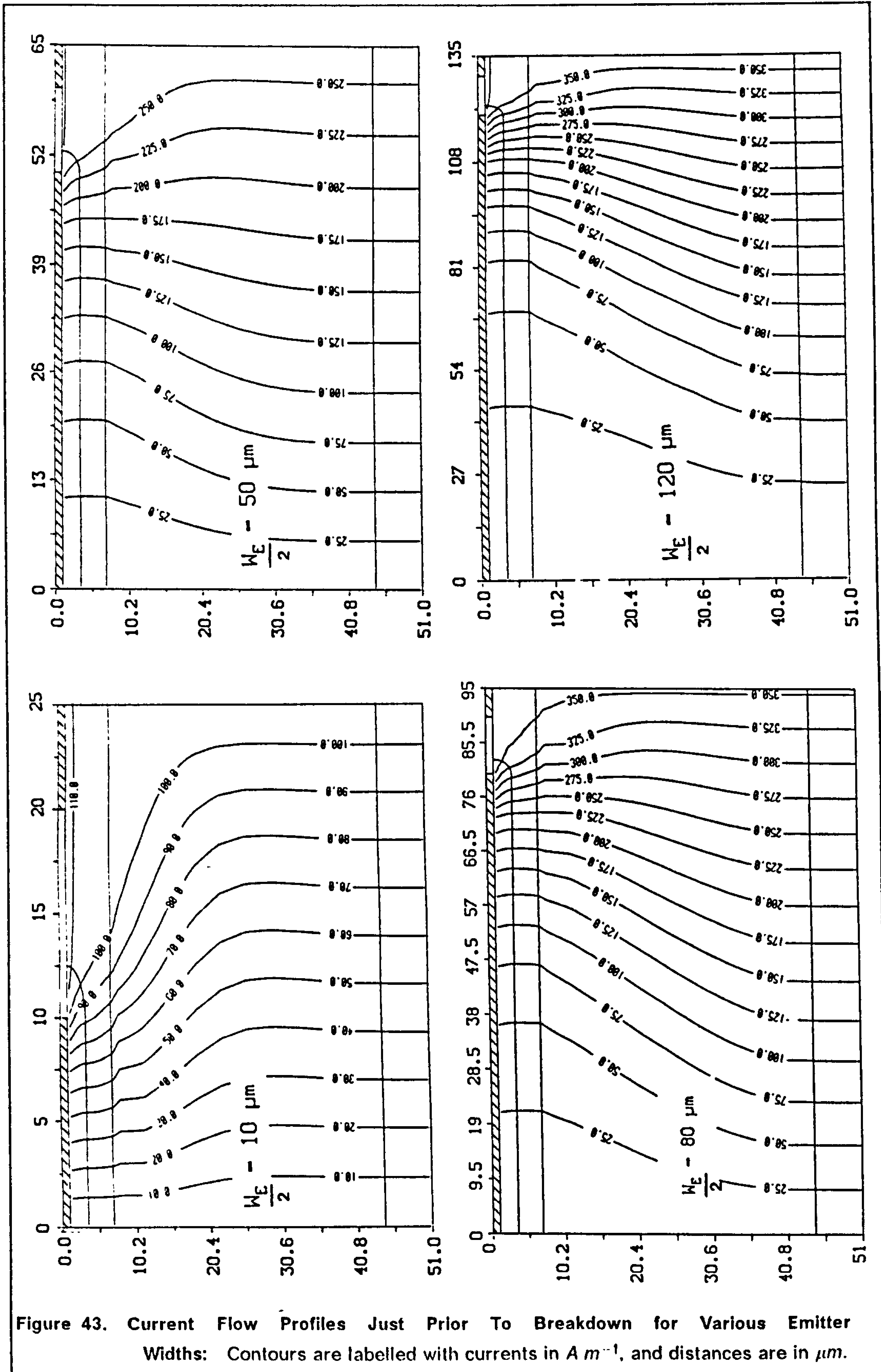


Figure 43. Current Flow Profiles Just Prior To Breakdown for Various Emitter Widths: Contours are labelled with currents in $A\,m^{-1}$, and distances are in μm.

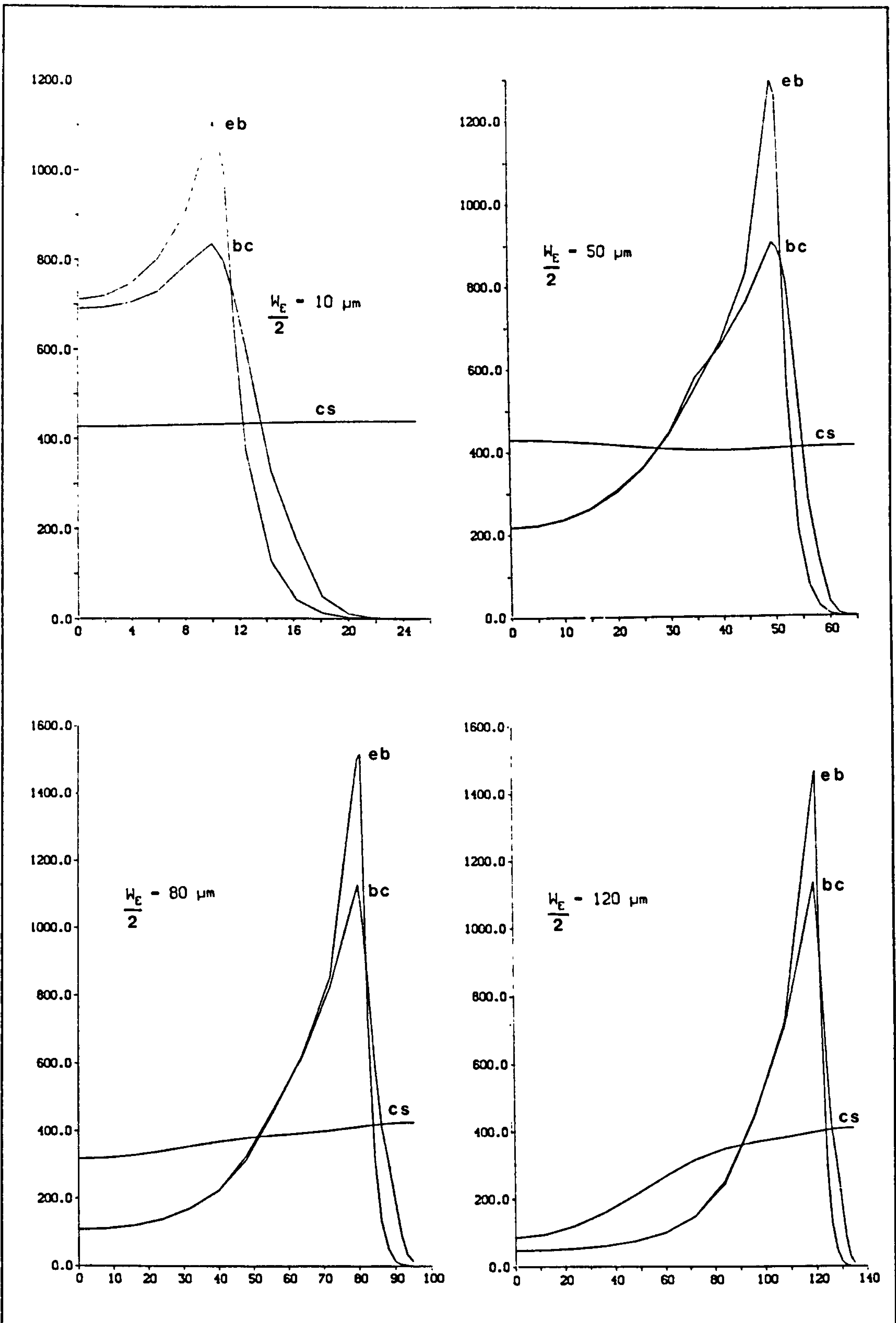


Figure 44. Current Densities Along Emitter-Base, Base-Collector and Collector-Substrate Junctions: Current densities are in $A\ cm^{-2}$ and distances are in μm.

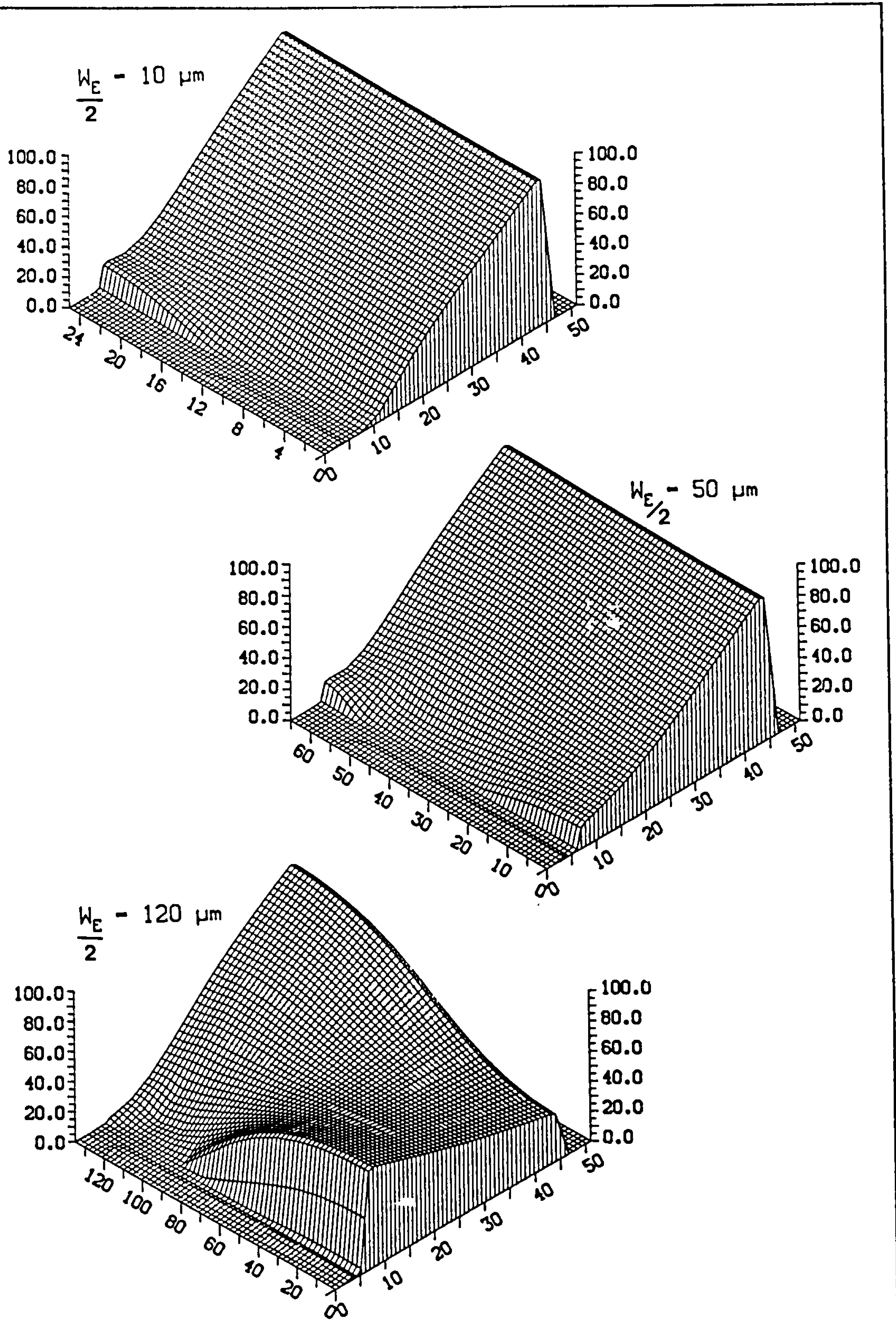


Figure 45. Electric Field Profiles Just Prior to Breakdown: Fields are in $KV cm^{-1}$ and distances are in μm.

current density at the collector-substrate junction is approximately four times higher under the base contact than under the emitter centre. This can lead to extreme reduction in surge capability.

The corresponding electron current densities across the emitter-base, base-collector and collector-substrate junctions for all four cases are shown in Figure 44. For $W_E/2 = 50\mu m$ two slight peaks are apparent in the current density profile along the collector-substrate junction, at the left and right hand edges of the cell, with a shallow minimum located below the emitter edge. This occurs despite the fact that the current initially peaks at the emitter edge, and is due to the existence of lateral fields in the collector forcing the electron current towards the left and right hand boundaries. These lateral fields are caused by the fact that the base is widest beneath the emitter edge, which can be seen from careful inspection of the field profile in Figure 45. The non-uniformity of base width is itself a consequence of the uneven injection from the emitter, resulting from emitter pinch. This effect is also present for the $10\mu m$ emitter half-width, where the current density at the right hand boundary becomes slightly larger than that at the left hand boundary. In short the lateral field introduces an additional drift component that aids the divergence process, but it also results in slightly non-uniform current flow. It is readily apparent that the pinching is more severe for wider emitters. The non-uniformity of current density at the collector-substrate junction is evident at a width of $80\mu m$ and the peak current density at the this junction in all cases is $430\, A\, cm^{-2}$, which corresponds very closely to the value required for breakdown calculated using the simple one dimensional model. Thus, devices with narrower emitters which have uniform collector current densities can be accurately modelled with respect to avalanche injection using the simple one dimensional model.

The electric field profiles which result as a consequence of the mobile space charge associated with the non-uniform current densities in the collector layer are shown in Figure 45. For the narrower emitters the field is uniform across the whole of the collector-substrate junction, as expected. Owing to emitter pinch, however, at higher emitter widths the field falls off towards the emitter centre in accordance with the current density. In the extreme case ($W_E/2 = 120\mu m$) the concentration of injected carriers from the emitter centre is so small that it does not exceed the thermally ionized donor concentration in the collector, and as a consequence the field peaks at the base-collector junction.

In order to investigate the effect of emitter width on current handling capability, the average current density at breakdown was calculated by dividing the total collector current required to produce the critical field by the cell width. This value was then plotted against the corresponding emitter half width as shown

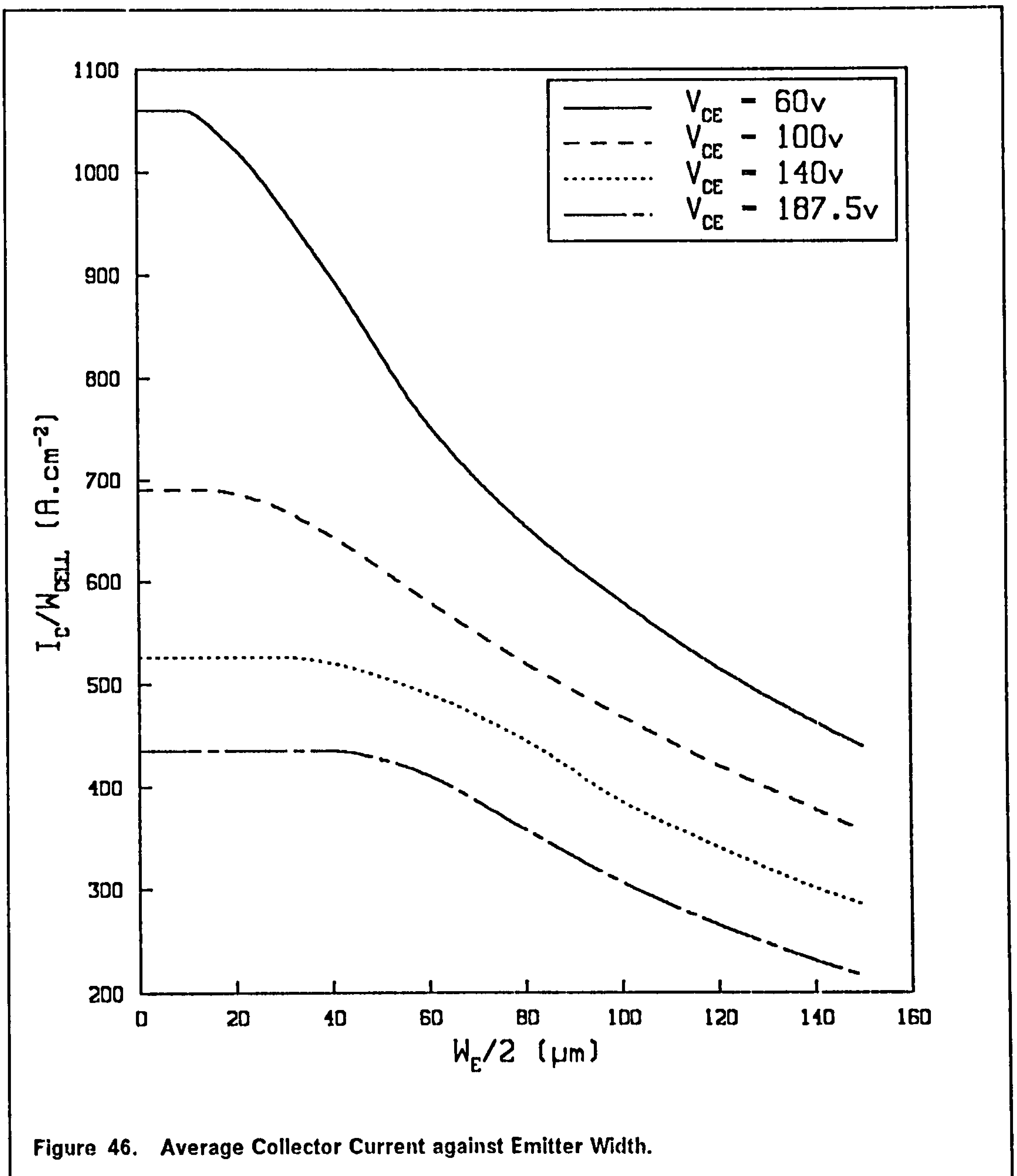


Figure 46. Average Collector Current against Emitter Width.

in Figure 46. In this diagram a number of different curves have been plotted for various collector breakdown voltages. Thus, if the average current density values of the ordinate axis are multiplied by the total active chip area of a device with a particular emitter width, then the curves indicate the maximum current pulse that the device can handle. For a collector voltage of 187.5*v* the surge capability becomes impaired once the emitter half-width exceeds 40*μm*, since the collector current density becomes non-uniform at widths greater than this.

At lower collector voltages, greater collector currents are required to push the field up to its critical value. Because emitter pinch is more severe at higher

currents it is necessary to make a further reduction in the emitter width in order to recover the desired uniform current density in the collector. Thus, at a collector voltage of 60V the emitter half-width must be made less than about 10μm so that the collector layer is fully utilized. It is evident that if the emitter half-width is designed to be less than or equal to this value, this will then ensure maximum current handling at all collector voltages.

A safe operating area (SOA), which defines the combinations of current and voltage that are acceptable for stable operation, can now be constructed for devices with different emitter widths. It is constructed by noting the average current densities at each of the collector voltages for a particular emitter half-width and then plotting the currents against the voltages. Thus, the parameter is transformed from collector voltage to emitter half-width. This procedure has been carried out for a number of widths and the results are plotted in Figure 47. It can be seen that the biggest improvement to be gained from utilizing narrow emitters is in the low voltage, high current operating regime. This is because of the acute pinching at high current as previously stated. It may also be noted that a device with 20μm wide emitters can handle approximately twice the current of a device with 280μm wide emitters at any particular collector voltage.

The safe operating areas have been replotted in more conventional form on logarithmic axes in Figure 48. The maximum safe operating limit for an emitter half-width of 10μm or less is in close agreement with the simple one dimensional model previously outlined. This predicts that the maximum voltage that can be sustained is proportional to the reciprocal of the peak current which can be handled at that voltage. The slope of the 10μm curve in Figure 48 is, therefore, very close to -1. In practice, however, the critical field, E_C is not constant in the operating range under consideration. Its value tends to rise at higher currents and lower voltages due to a corresponding fall in h_{FE} [6.20]. As a consequence, the slope of the safe operating area will be greater than unity and is usually in the range -1.5 to -2.

From the point of view of power handling, therefore, it seems beneficial to use very narrow emitter fingers. The consequences of such a design on the normal operation of the device has been investigated. Figure 49 illustrates how the h_{FE} varies with average collector current density. The fall of h_{FE} with collector current may be attributed to high injection and base widening effects [6.22]. For very narrow emitters the h_{FE} is low and it then rises as the emitter is widened. It reaches a peak at an emitter half-width of 20μm and then falls with further increases in width. For narrow emitters the h_{FE} is low because the emitter area is small, and consequently the emitter current density must be pushed up to account for the necessary collector current. This tends to enhance the high level effects

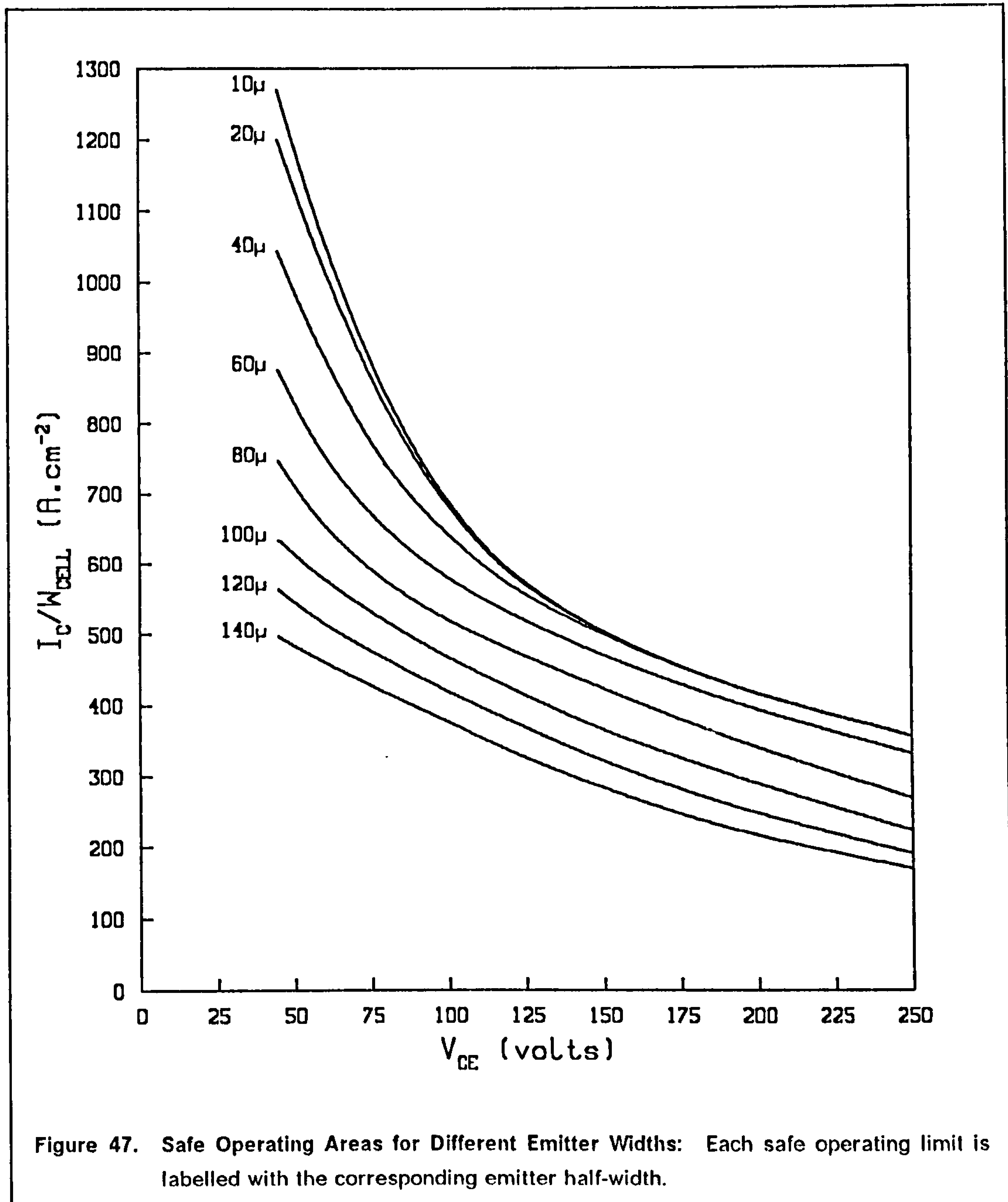


Figure 47. Safe Operating Areas for Different Emitter Widths: Each safe operating limit is labelled with the corresponding emitter half-width.

resulting in a premature reduction in h_{FE} . At higher emitter widths the greater pinching effects are known to cause the h_{FE} to fall off more rapidly [6.23].

A similar situation exists for the emitter-base voltage requirement, which is also shown in Figure 49. At very low widths a slightly higher V_{BE} is required to account for the higher emitter current densities. Whereas for very wide emitters the increased pinching must be offset by a corresponding increase in V_{BE}. Though not illustrated here these trends were also evident at collector voltages of 60, 100 and 140V.

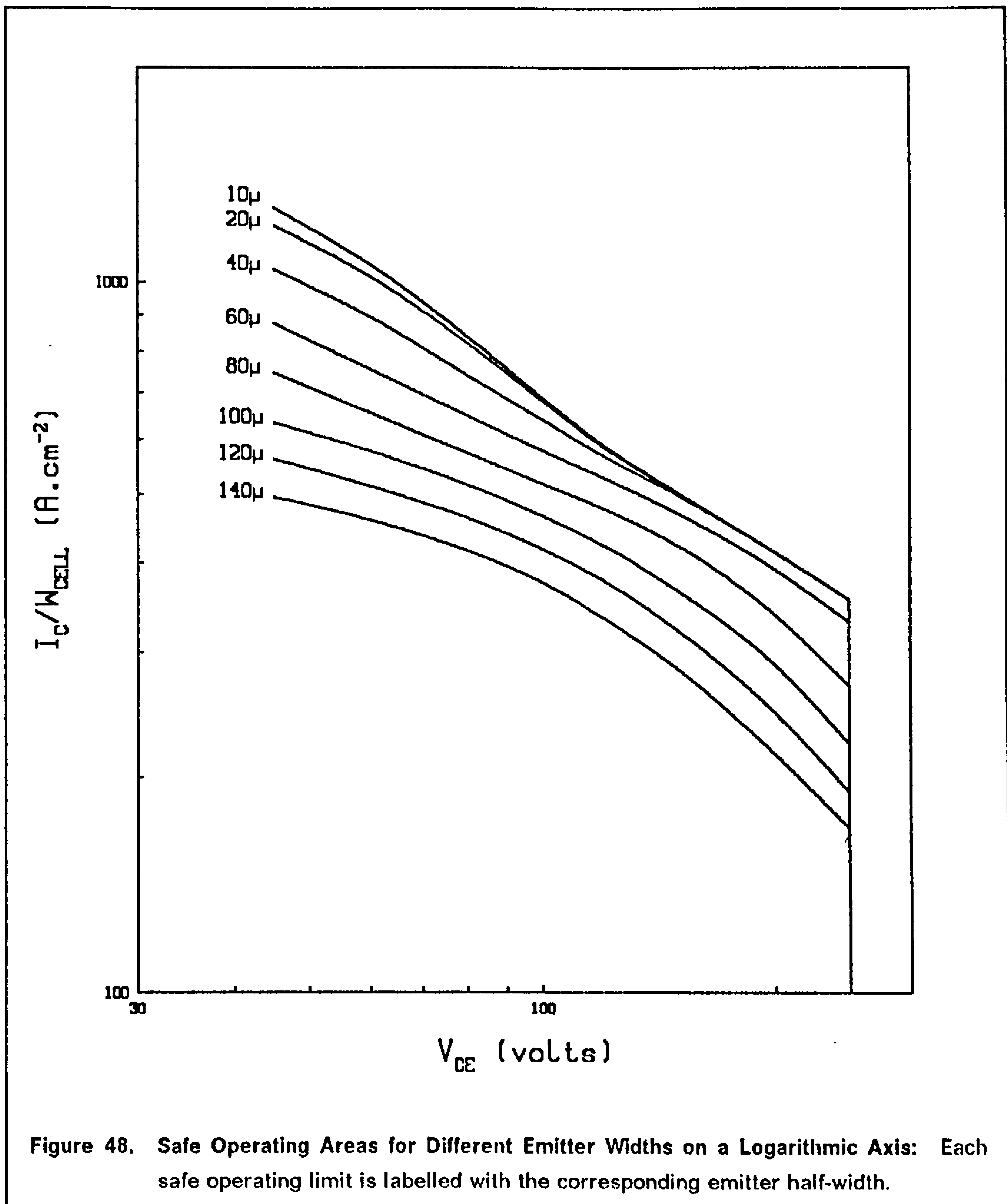
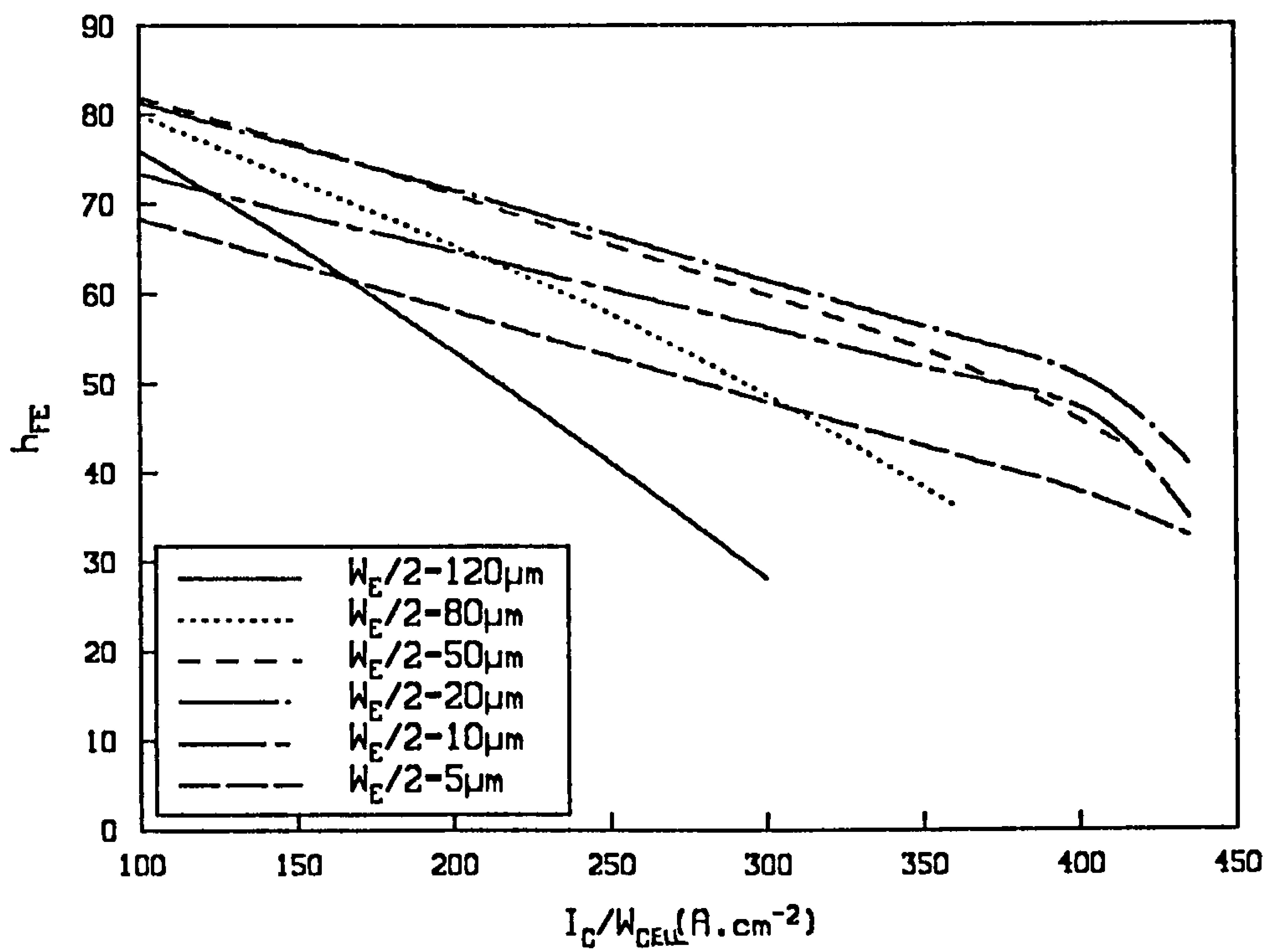


Figure 48. Safe Operating Areas for Different Emitter Widths on a Logarithmic Axis: Each safe operating limit is labelled with the corresponding emitter half-width.

Making the emitter width too small or too wide, therefore, puts a greater requirement on the base drive circuitry resulting in less efficient power conversion. For a device with this particular specification, therefore, an emitter half-width of approximately $20\mu m$ can be recommended. This choice of width would provide optimum power handling and power conversion.

6.3.4 Conclusion.



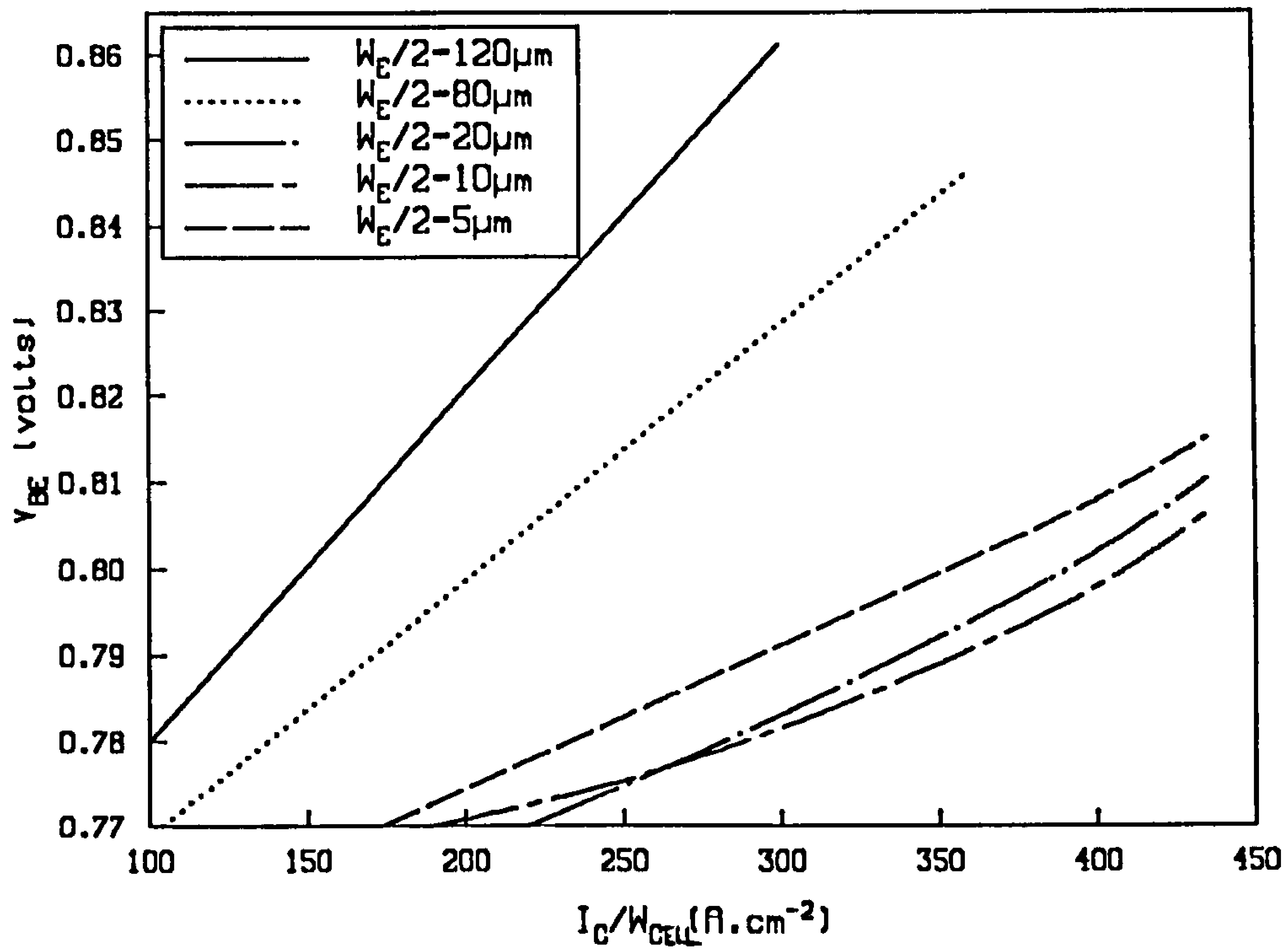


Figure 49. Variation of Current Gain and Base-Emitter Voltage with Collector Current for Various Emitter Widths: All curves are for a collector voltage of 187.5v.

This investigation has illustrated the need to carefully choose the emitter stripe width so that the entire active chip area is fully utilized. Although the analysis was restricted to one particular case, its conclusions can be extended to all interdigitated type structures. In this analysis the collector specification was calculated to give minimum on-resistance, however, the collector could, just as easily, have been designed to give maximum collector current density for breakdown. This specification results in a marginally wider collector layer with a lower resistivity, resulting in a slightly higher current handling capability.

Both optimization procedures give rise to the need for thicker collector layers at higher BV_{CEO} requirements. In this case it is, therefore, expected that the emitter-half width could be extended above $20\mu m$, while maintaining maximum current handling. The only merit for widening the emitter seems to be to reduce debiasing along the emitter stripe. A properly designed power device should, however, incorporate a reasonably thick metalization pattern and emitter debiasing should not be a problem. Previous results have shown that it is, in fact, extremely undesirable to extend the emitter width as this aids emitter pinch. The overall conclusion of this investigation is that the emitter width should be minimized to obtain maximum pulsed power handling, but if the emitter width is made too small then the power conversion efficiency will be impaired. This statement should hold regardless of lateral base resistance and collector specification.

6.4 A Model for Inductive Switching using Bipolar Transistors.

6.4.1 Introduction.

There is an increasing demand in industry for high power transistors that are capable of efficiently switching inductive loads. A prime example of this is in inverter circuits, which convert dc power to ac power and have, for example, applications in ac locomotives, dc power transmission lines and induction heating. Such circuits conventionally use thyristors, however, new power transistor designs offering higher switching speed and lower switching losses are now finding applications in this field. Furthermore, complicated and expensive commutating circuitry is not required when using transistors. A typical building block of an inverter circuit is shown in Figure 50, where L and R represent the primary winding of a transformer.

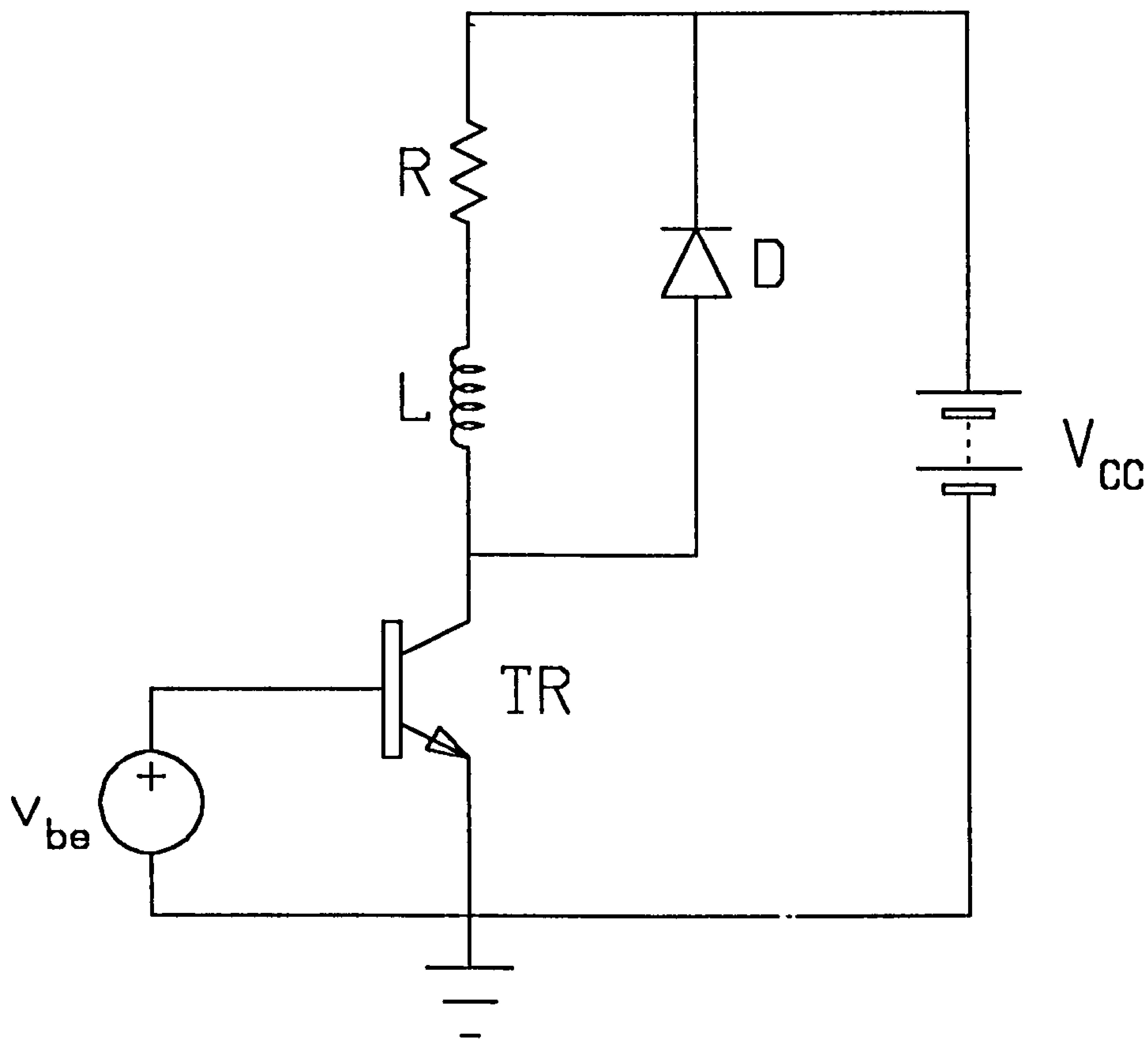


Figure 50. An Inductive Collector Load with a Free-Wheeling Diode.

The most severe condition that the transistor must handle arises during switch-off with an inductive load. The voltage developed across the inductor is proportional to the rate of change of current through it. Thus, during switch-off, the collector current falls creating extremely large collector voltages. The freewheeling diode, *D* is present in order to prevent the collector voltage from reaching dangerously high values. It starts to conduct once the collector voltage reaches the supply voltage, V_{CC} plus a forward diode voltage drop. The collector voltage is then clamped at this value and the inductor discharges through the diode. During turn-off, therefore, a high voltage/high current condition exists, which as previously shown can initiate avalanche injection leading to second breakdown.

The transient device simulator has been coupled with the collector circuit equation in order to predict transistor switching dynamics during switch-off. The effect of altering the emitter geometry on the switching performance was then investigated.

6.4.2 Coupling of the Numerical Model with the Collector Circuit Equation.

Kirchoff's second law applied to the collector circuit of Figure 50 gives:

$$F = v_{ce} - V_{CC} - R\,i_c - L\,\frac{di_c}{dt} = 0 \qquad (6.46)$$

Using backward differencing for the time derivative results in:

$$F = v_{ce}^{m+1} - V_{CC} - R\,i_c^{m+1} - L\,\frac{i_c^{m+1} - i_c^{m}}{d_m} = 0 \qquad (6.47)$$

In order to find a compatible collector voltage at time $m+1$, which satisfies both the transistor model and equation (6.47) an initial guess is made at the collector voltage, v_{ce}. A good initial guess can be obtained by using a linear extrapolation from the two previous discrete points in time. The collector current calculated from the transistor model is then used to calculate F in (6.47). If F is not sufficiently close to zero then Newton's method is applied to (6.47) giving:

$$\delta v_{ce}^{n} = -\frac{F^{n}}{\dfrac{dF^{n}}{dv_{ce}}} \qquad (6.48)$$

where,

$$\frac{dF^{n}}{dv_{ce}} = 1 - R\,\frac{di_c^{n}}{dv_{ce}} - \frac{L}{d_m}\,\frac{di_c^{n}}{dv_{ce}} \qquad (6.49)$$

The derivatives of collector current with respect to collector voltage can be approximated using the method of secants, thus:

$$\frac{di_c^{n}}{dv_{ce}} = \frac{i_c^{n} - i_c^{n-1}}{v_{ce}^{n} - v_{ce}^{n-1}} \qquad (6.50)$$

The numerical model is then re-solved using the updated collector voltage. Condition (6.46) is subsequently checked and the whole procedure is repeated until F is sufficiently close to zero. In this case the absolute value of F was required to be less than 0.1% of V_{CC} for convergence. This method is outlined in the form of a flow diagram in Figure 51.

6.4.3 Turn-Off of Transistors with Cylindrical Geometry.

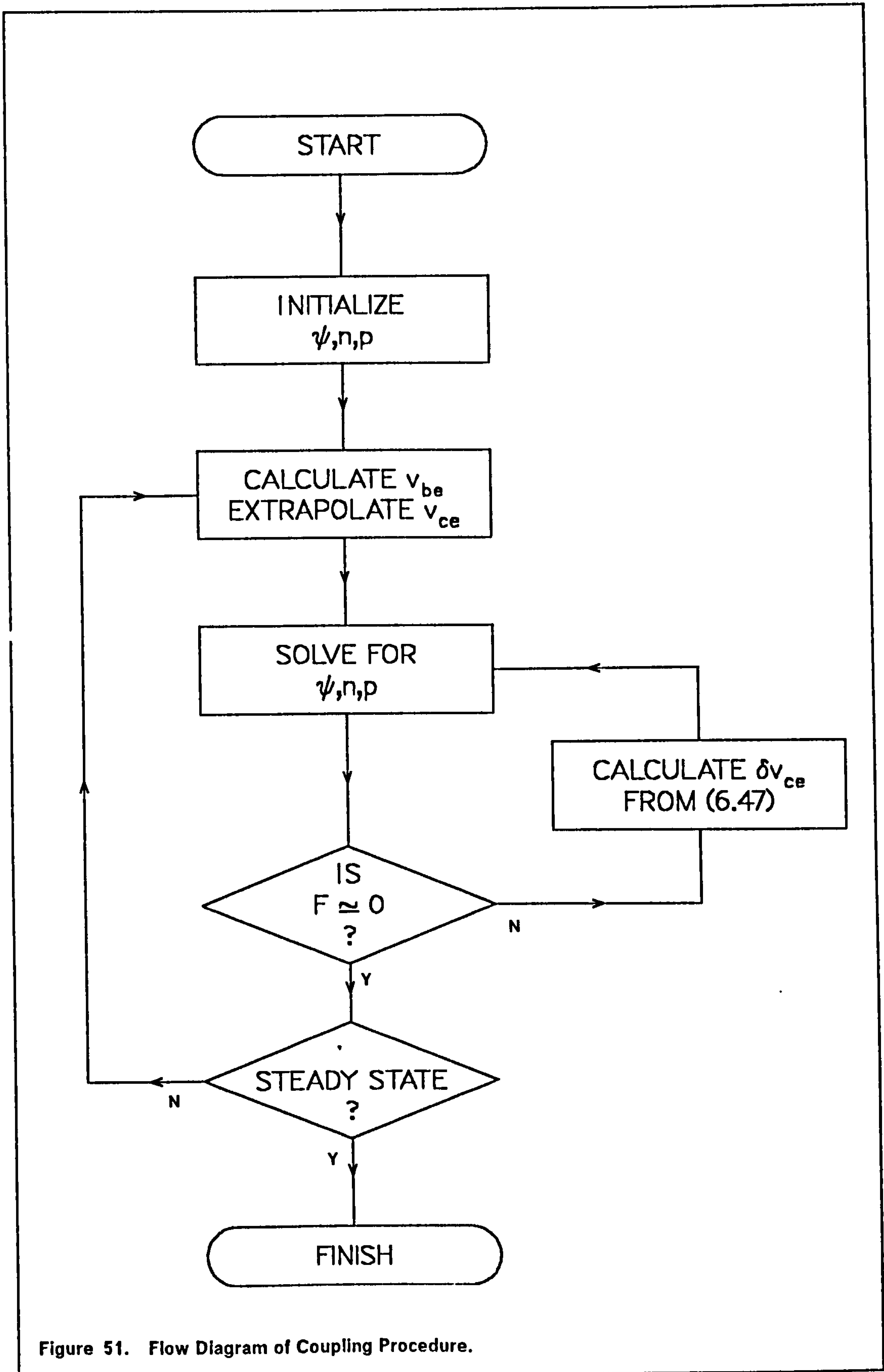


Figure 51. Flow Diagram of Coupling Procedure.

Turn-off of the type C device incorporating an inductive collector load has been modelled, and the 'old values' in Table 7 on page 169 have once again been taken to represent the device. The energy dissipated in the transistor is comparatively small during turn-on and the transistor is safe from breakdown. Consequently, the transient analysis was initialized with the result of a steady state simulation in which the device was biased into saturation with fixed collector and base voltages. In this analysis the collector voltage was initialized at 0.2V and the base voltage at 1V. The collector current was found to be 46.3mA for this biasing arrangement. The supply voltage, V_{CC} was assumed to be 100V and the collector resistance was subsequently calculated to be 2157Ω. Finally, the value of the inductor was taken to be 1mH, which is typical of many inverter applications.

The device was turned off by reducing the base voltage to zero over a period of 220ns. This should be typical of what can be achieved with the available low impedance pulse generators ($R_O = 10\Omega$). The base voltage should not be reduced more rapidly because the reverse base current that arises during switch-off generates a voltage across this impedance. This is apparent from Figure 52, which shows the computer generated base voltage and current waveforms and the corresponding collector voltage and current waveforms.

The collector load line is shown in Figure 53. The device comes out of saturation in approximately 460ns and then enters the quasi-saturation region [6.18]. The operating point then moves directly into the high field region at about 620ns, as the collector voltage increases. This mode of operation is similar to that described in section 6.2, where a current induced base is set up due to the velocity saturation of electrons. The voltage eventually gets clamped at 100V and the collector current decays to zero. It must be pointed out that operation at a particular point on the load line will not correspond exactly with the coincident operating point on the output characteristic. This is due to a number of transient effects, and therefore, Figure 53 only represents schematically what happens during switch-off.

The reverse base current is initially very large, since holes are at first removed from directly below the base contact and series resistances are very low. As time proceeds holes at a further distance from the base contact come under the influence of a reduced base potential. The series resistance between these more remote charges and the contact is larger and as a result the reverse base current falls. The variation in the hole distribution with time is shown in Figure 54 and Figure 55.

Initially the entire collector region is saturated with holes. By 200ns a significant number of holes have been removed from the collector region directly beneath the base contact. After 462ns the collector region has become even more

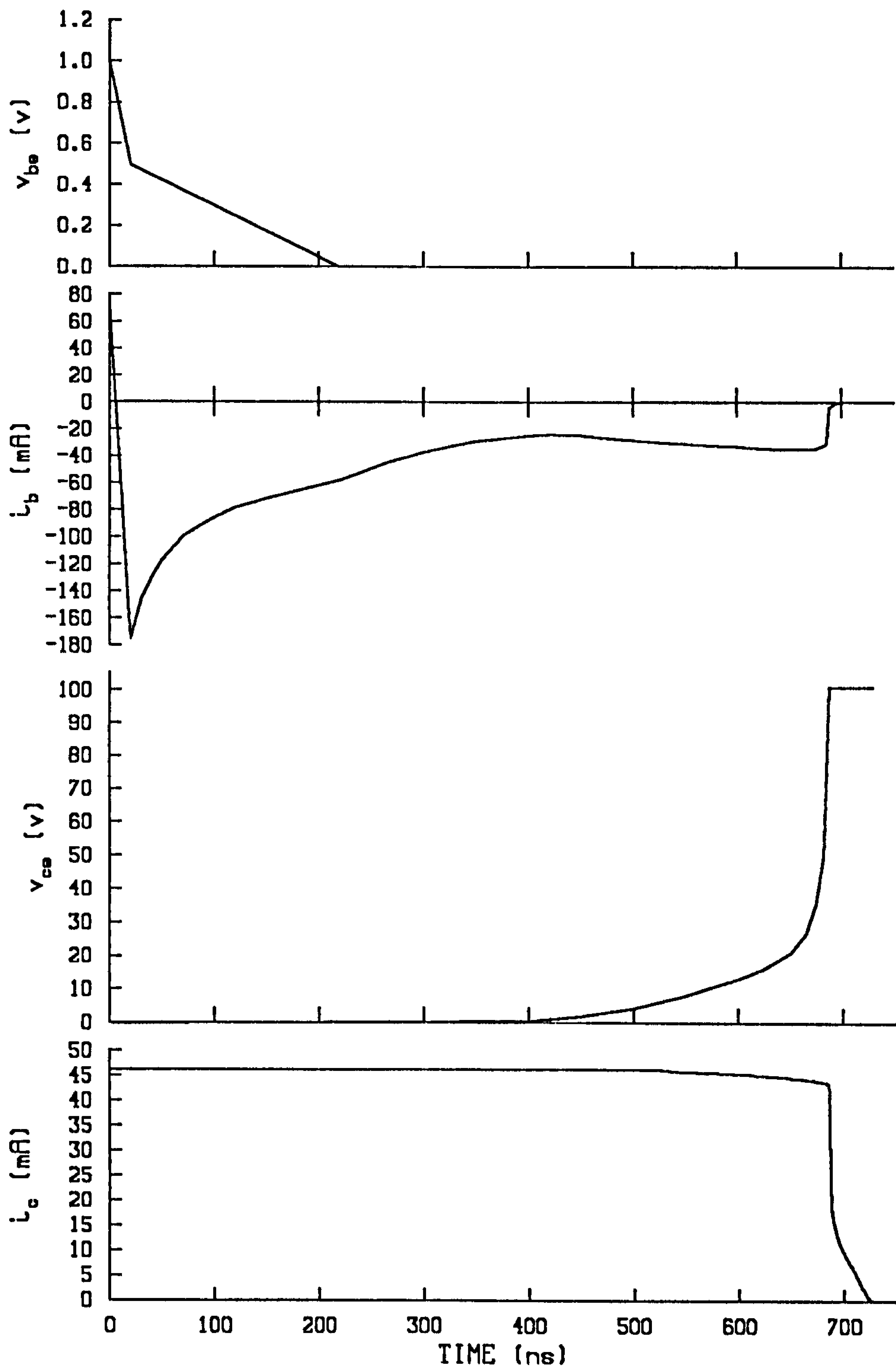


Figure 52. Voltage and Current Transients During Switch-Off.

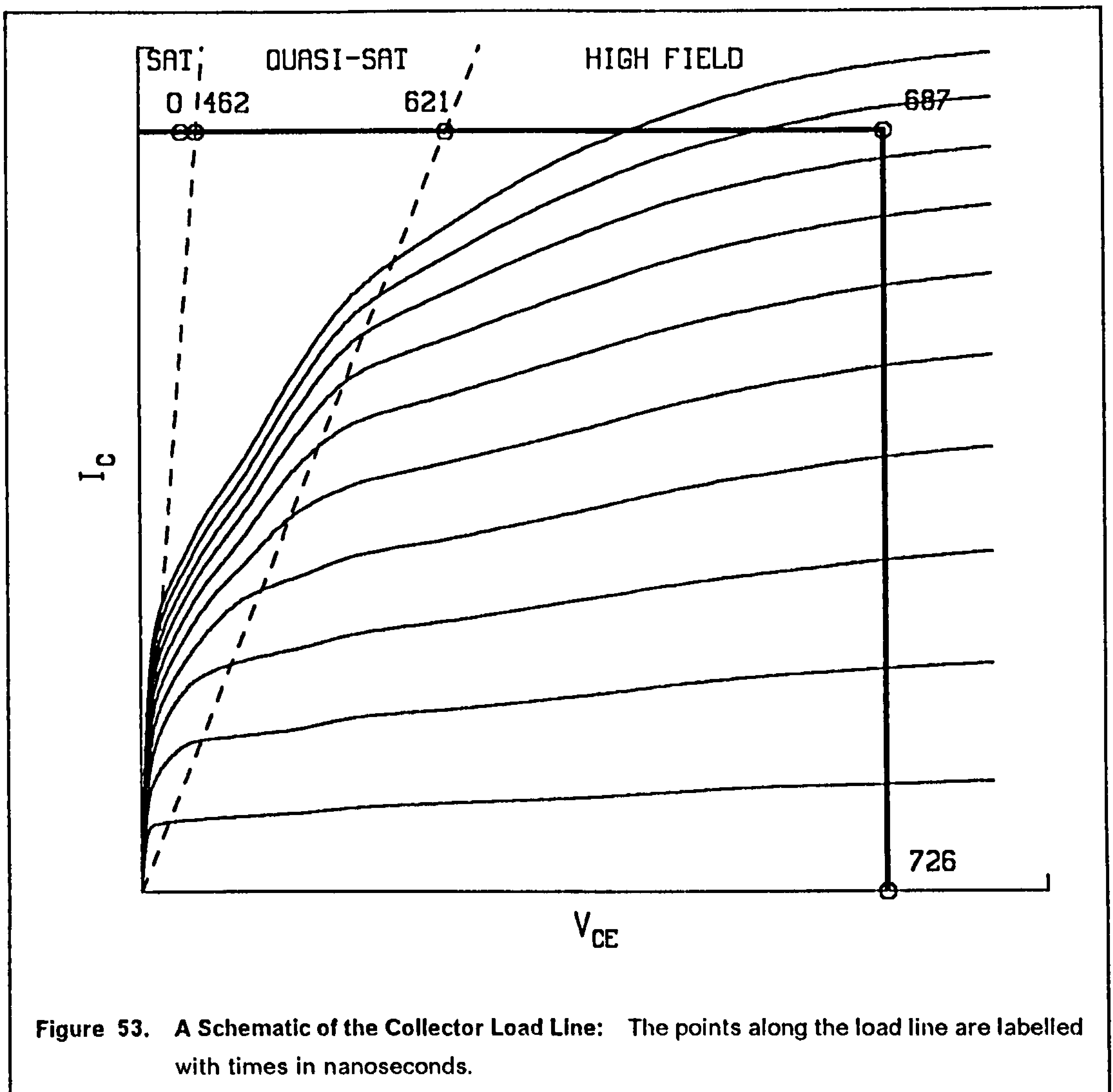


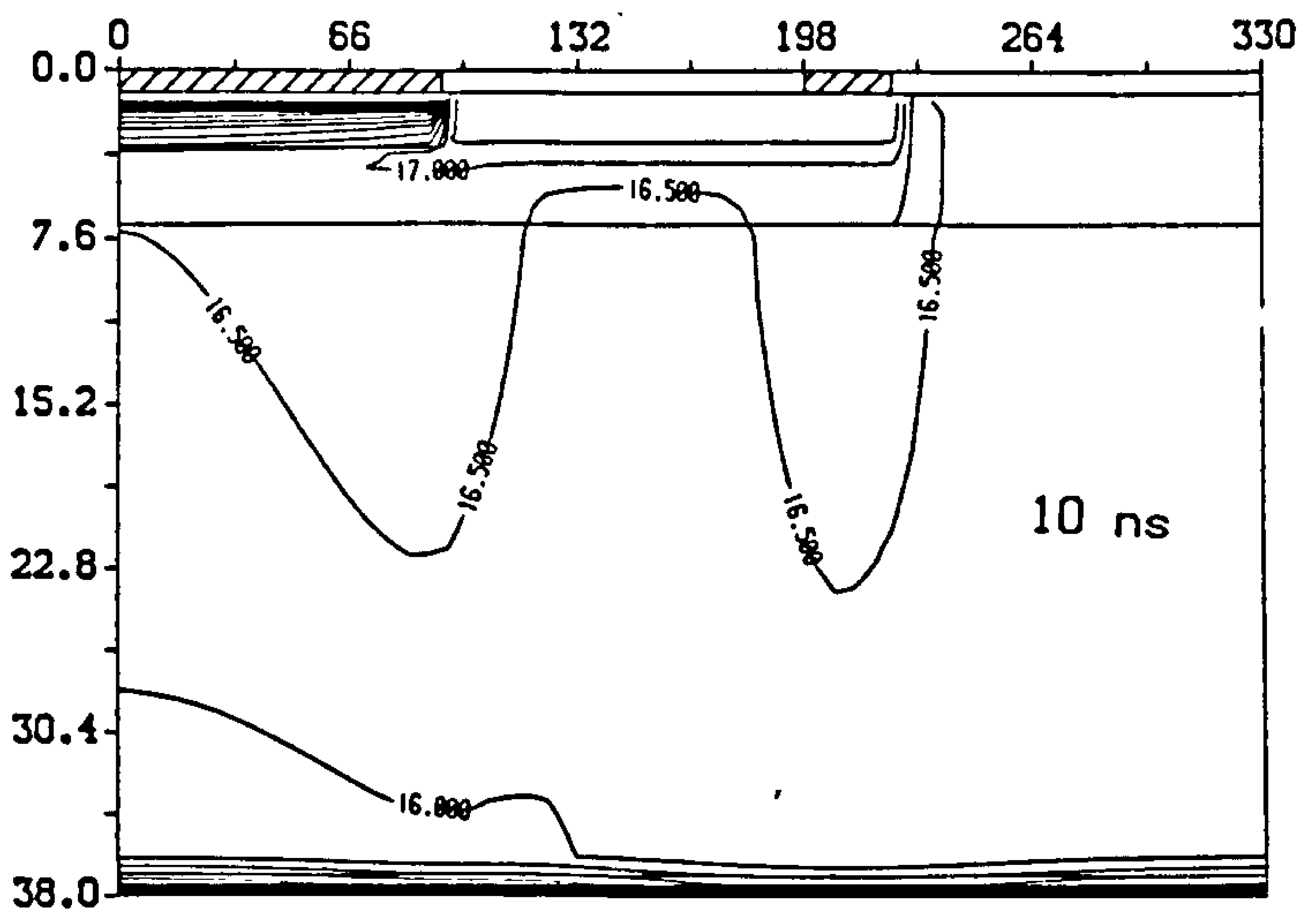
Figure 53. A Schematic of the Collector Load Line: The points along the load line are labelled with times in nanoseconds.

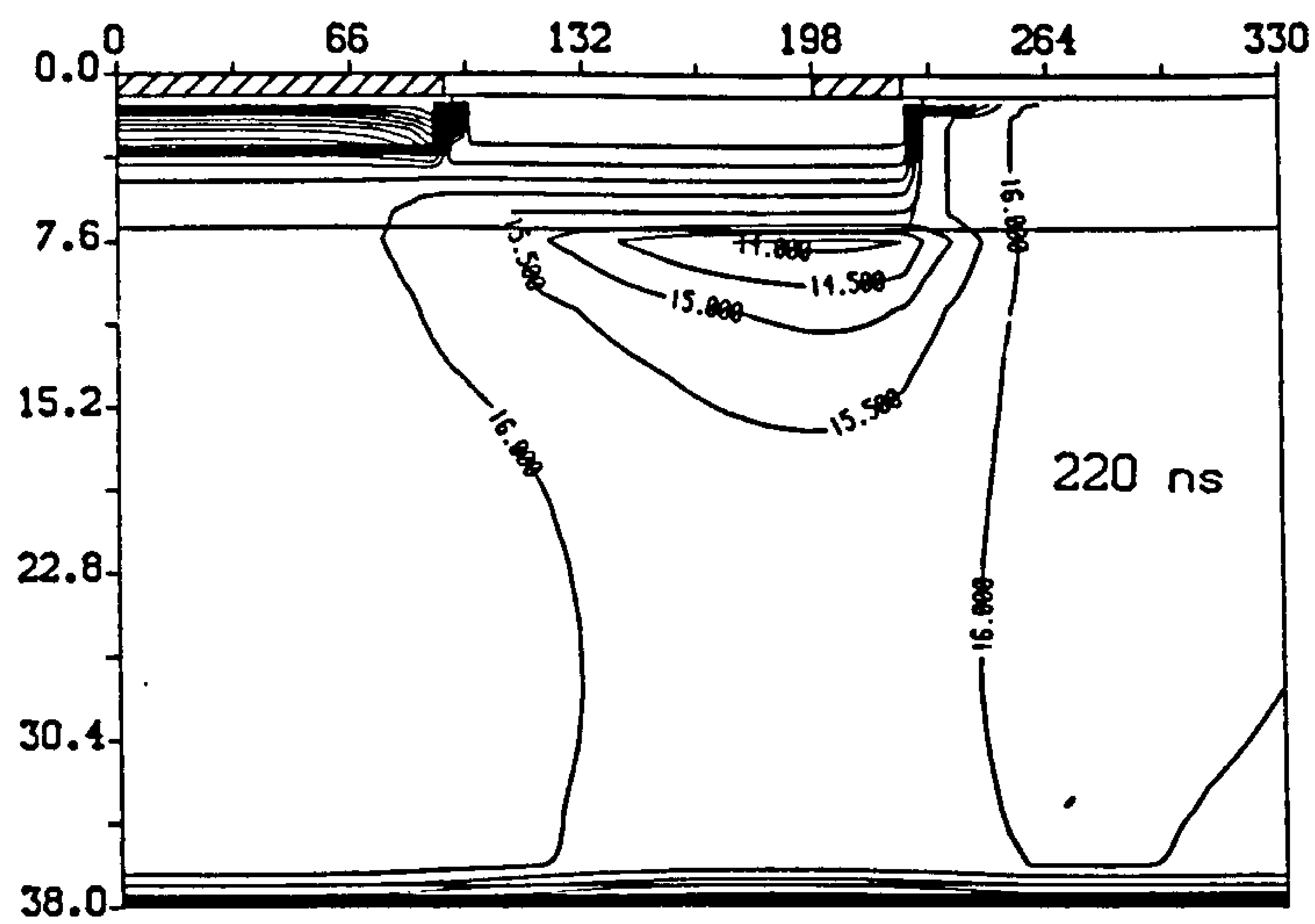
depleted of charge. It is also seen that the emitter region closest to the base contact has become depleted of holes. This indicates that this part of the emitter has recovered and is no longer injecting electrons into the base.

However the total emitter current was was found not to decrease in proportion with the conducting area lost, which can be explained as follows. Equation (6.46) can be rearranged to give:

$$- v_{ce} = V_{CC} + R\, i_c + L \frac{di_c}{dt} \tag{6.51}$$

If i_c should fall during switch-off then if L is large enough v_{ce} will rise, which in turn tends to maintain the collector current. For this reason it is found that the collector current is virtually constant during switching and the effect of the inductance is to ensure that the collector voltage rises in such a way that i_c remains approximately constant despite the reduction in the conducting emitter





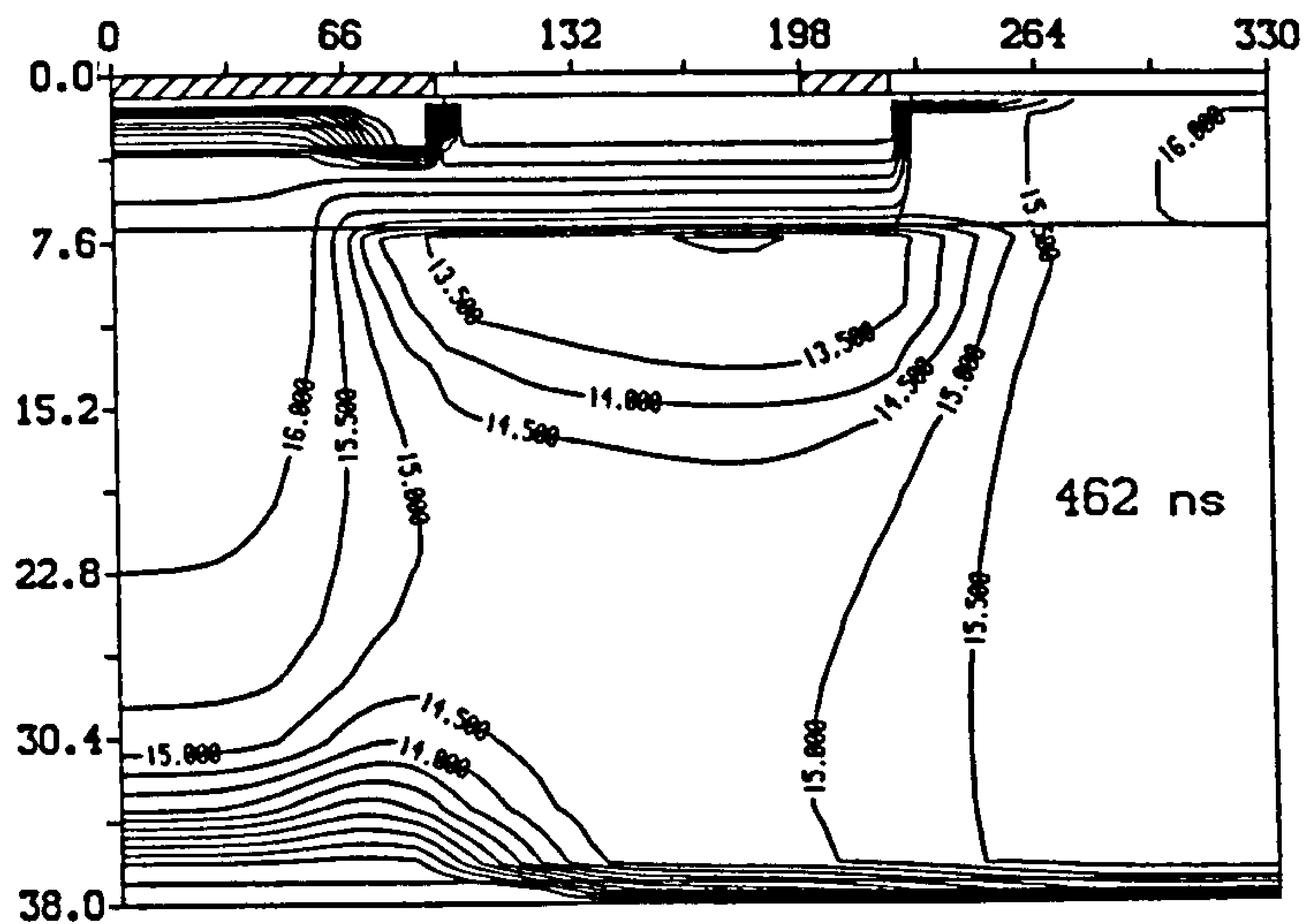
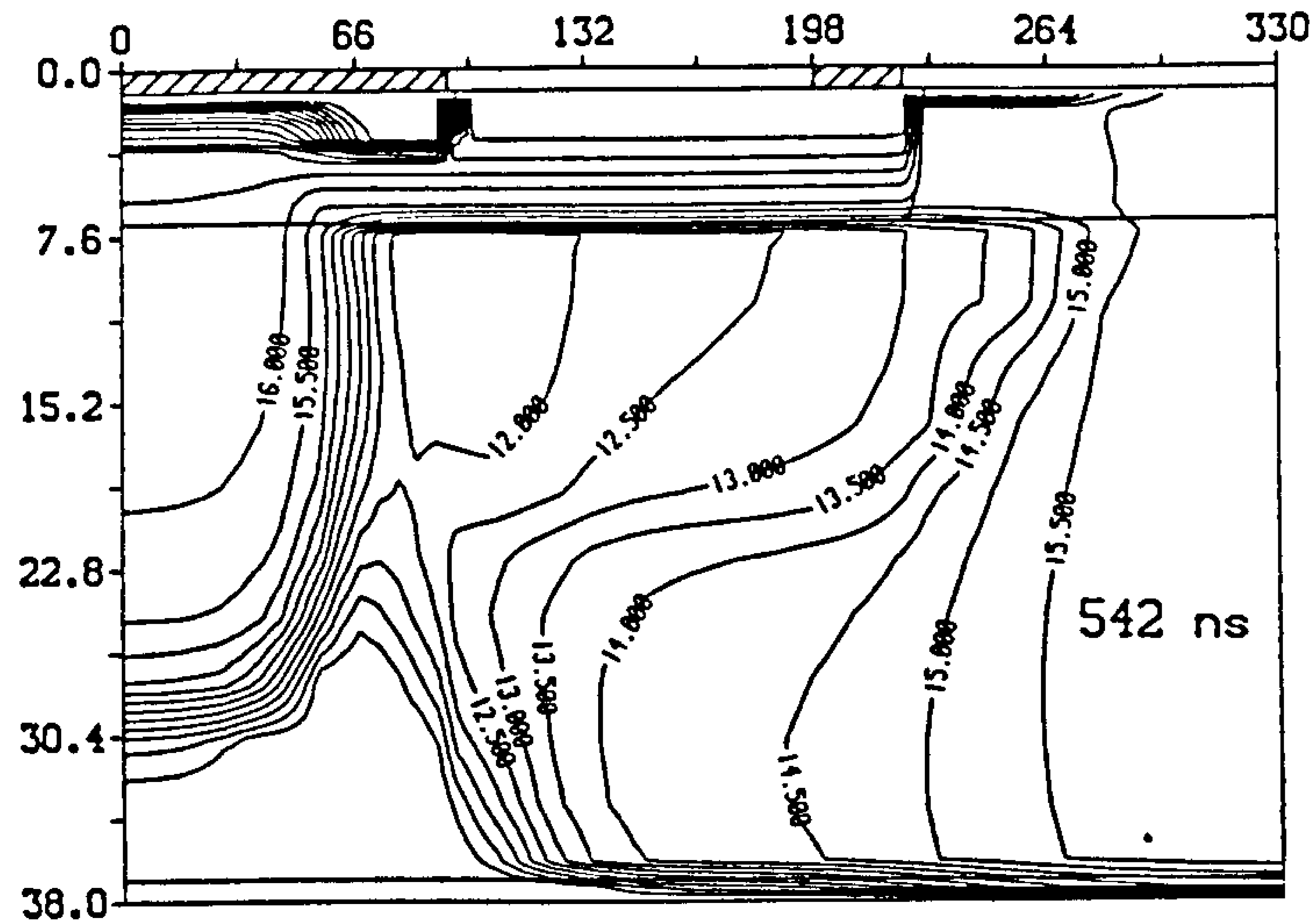
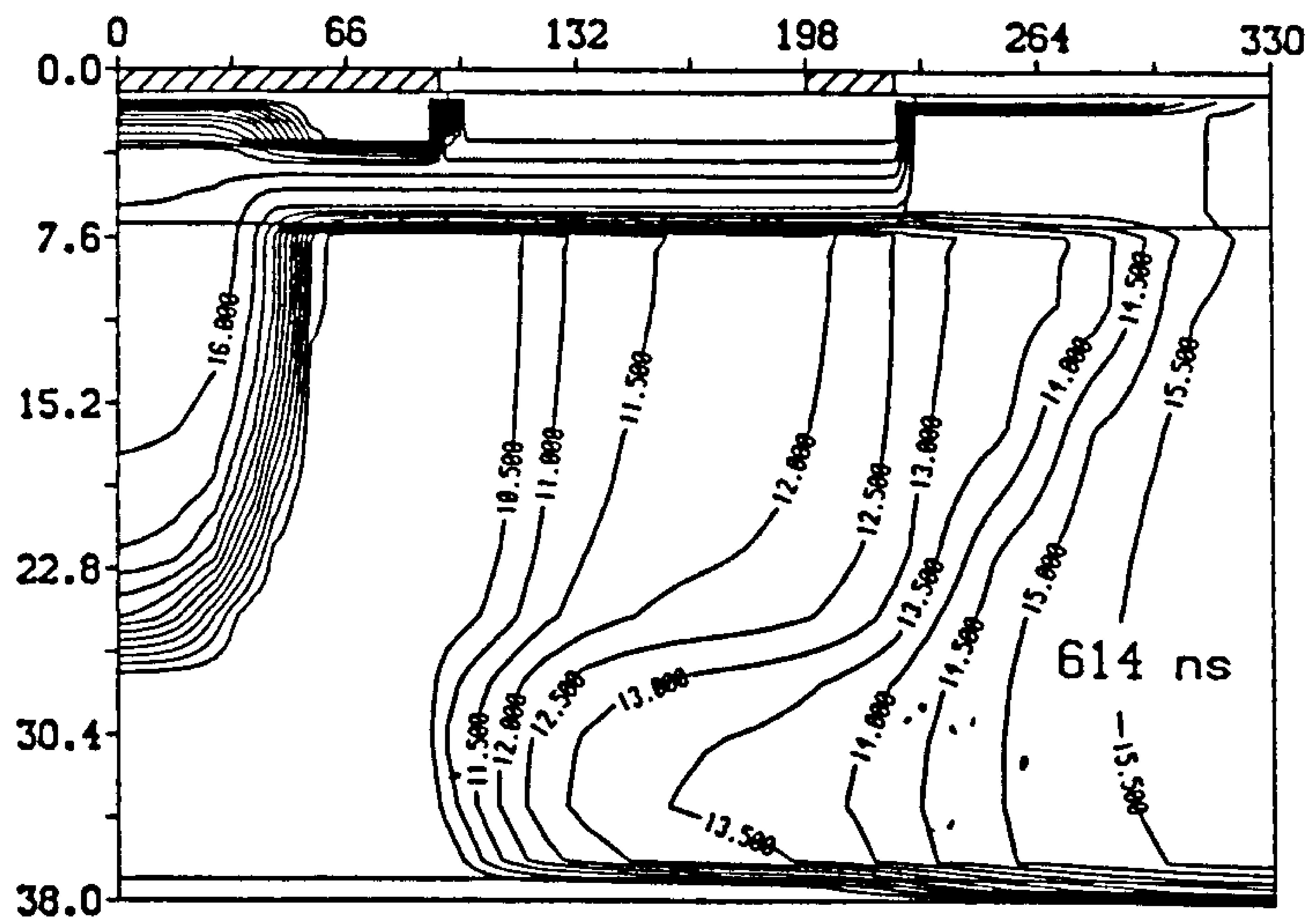


Figure 54. Hole Concentration Profiles at Various Times During Switch-Off: Contours are labelled with $\log_{10}$ values and distances are in μm.





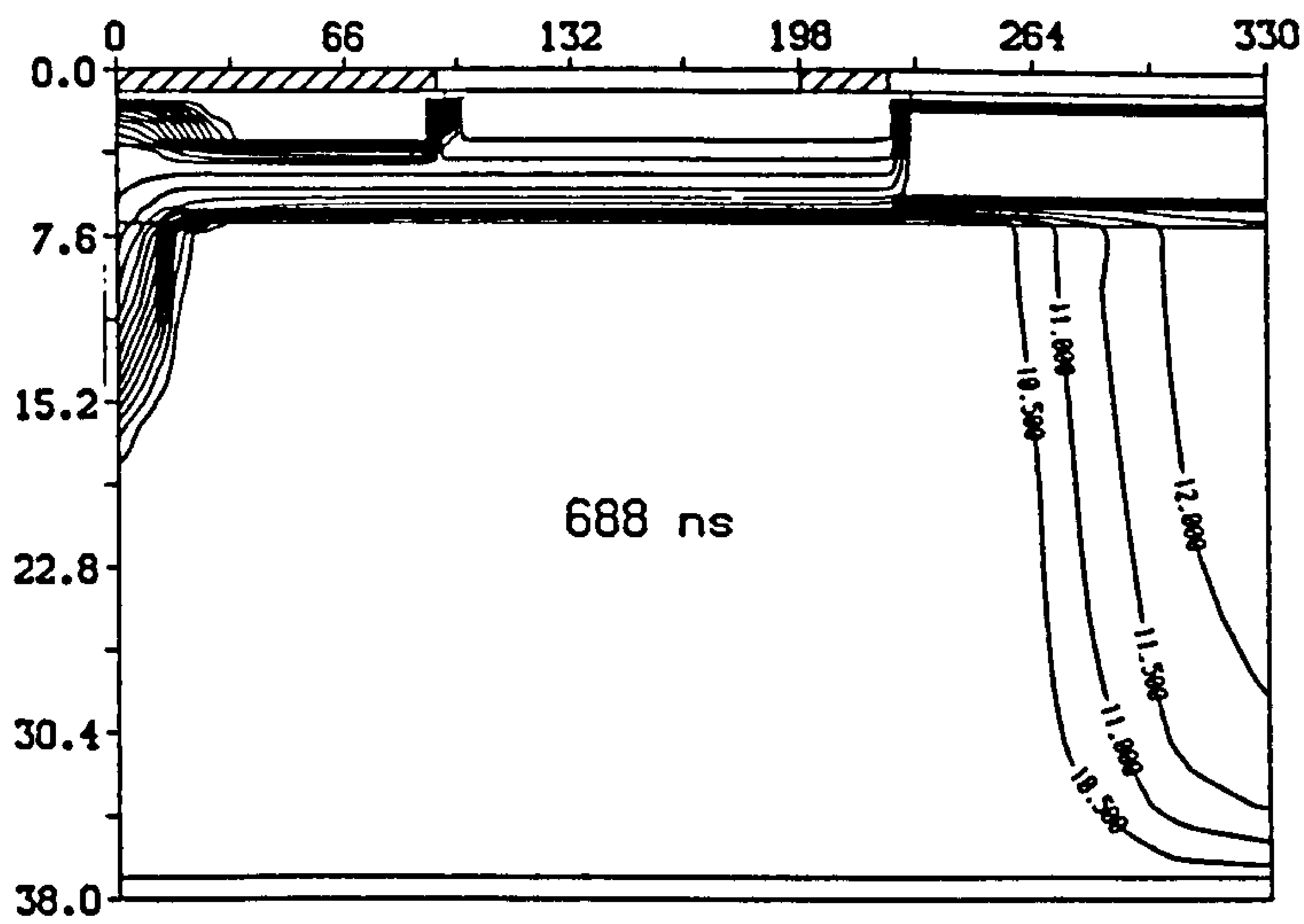


Figure 55. Hole Concentration Profiles at Various Times During Switch-Off.: Contours are labelled with $\log_{10}$ values and distances are in μm.

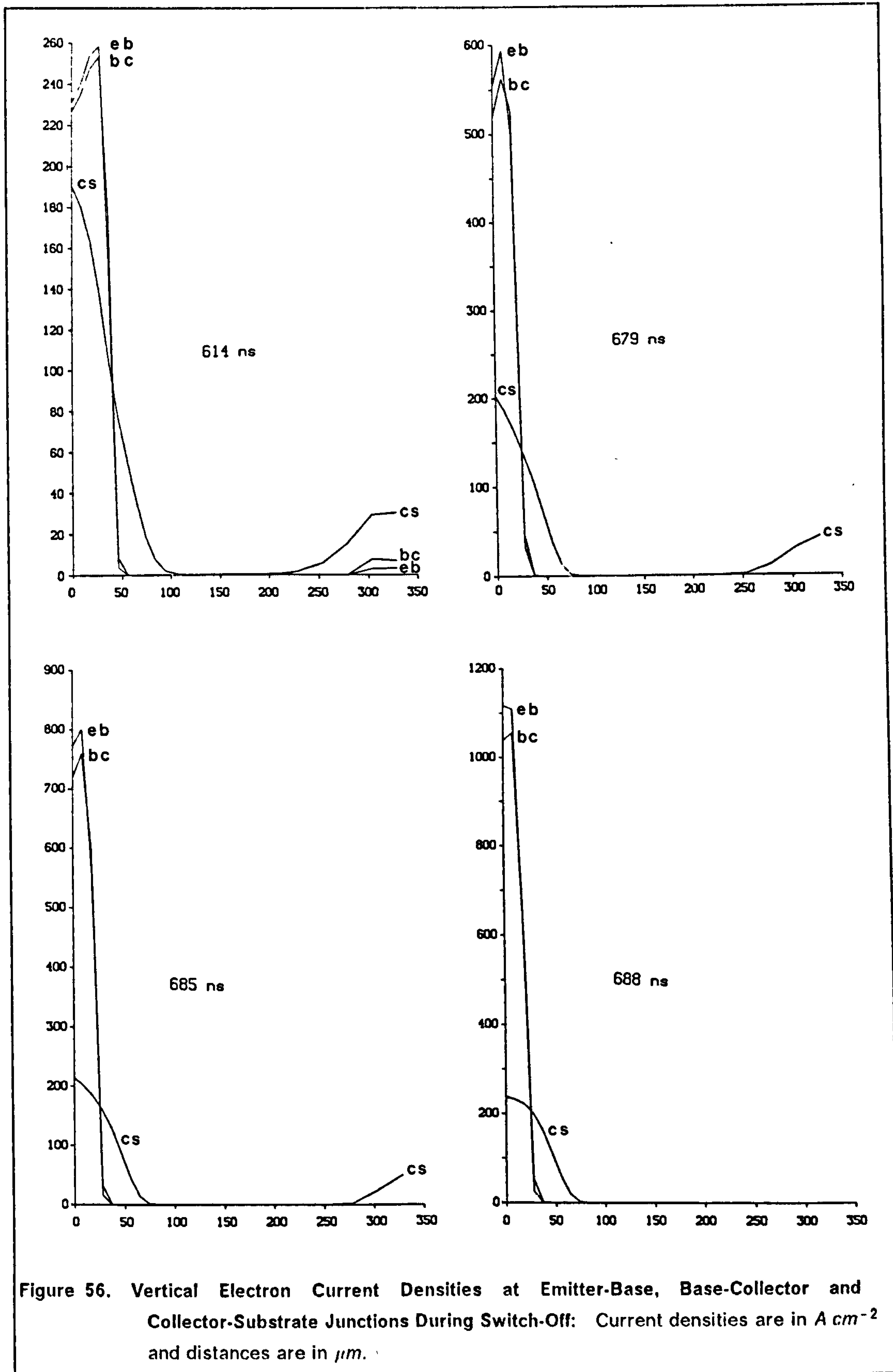
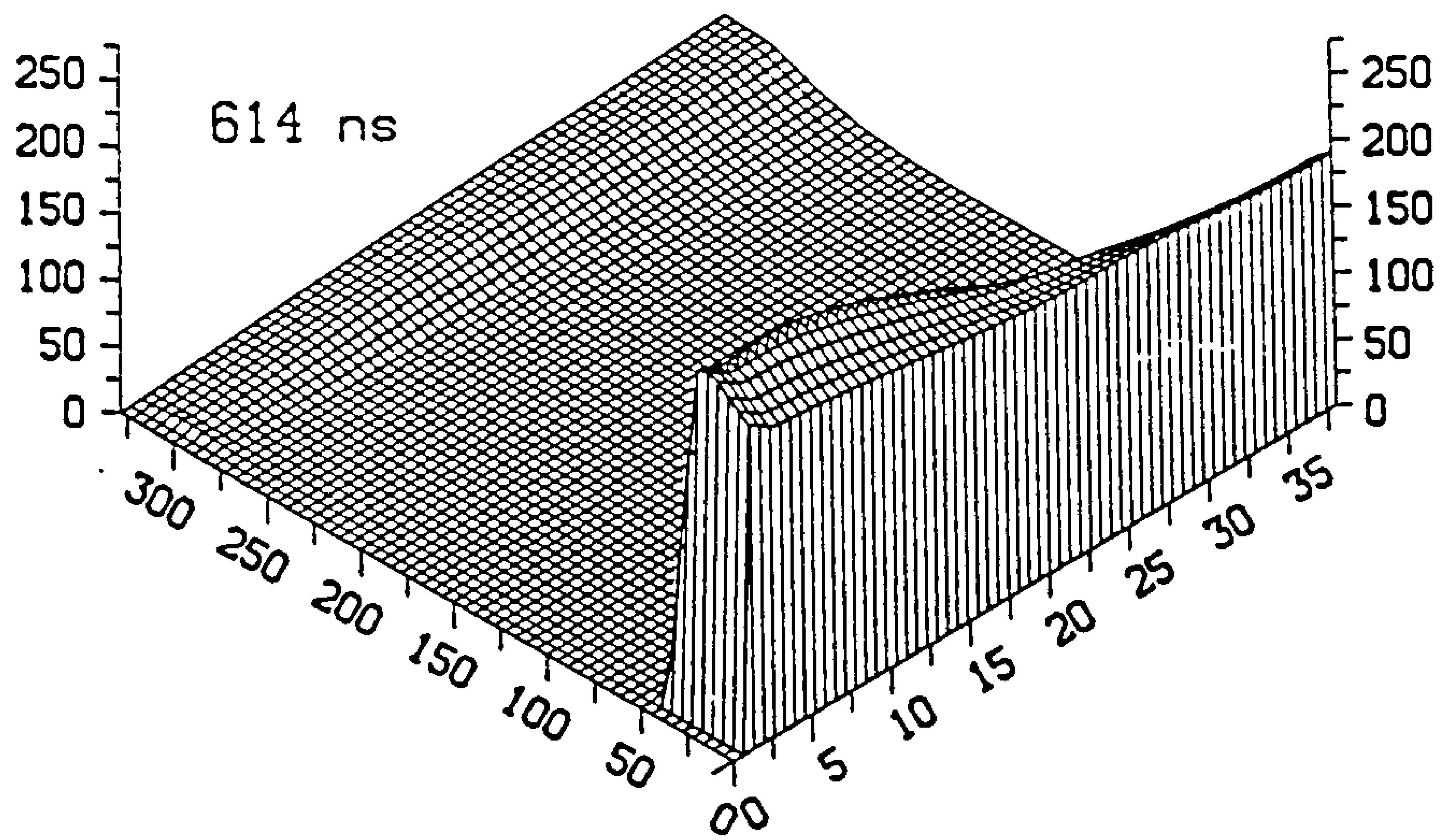
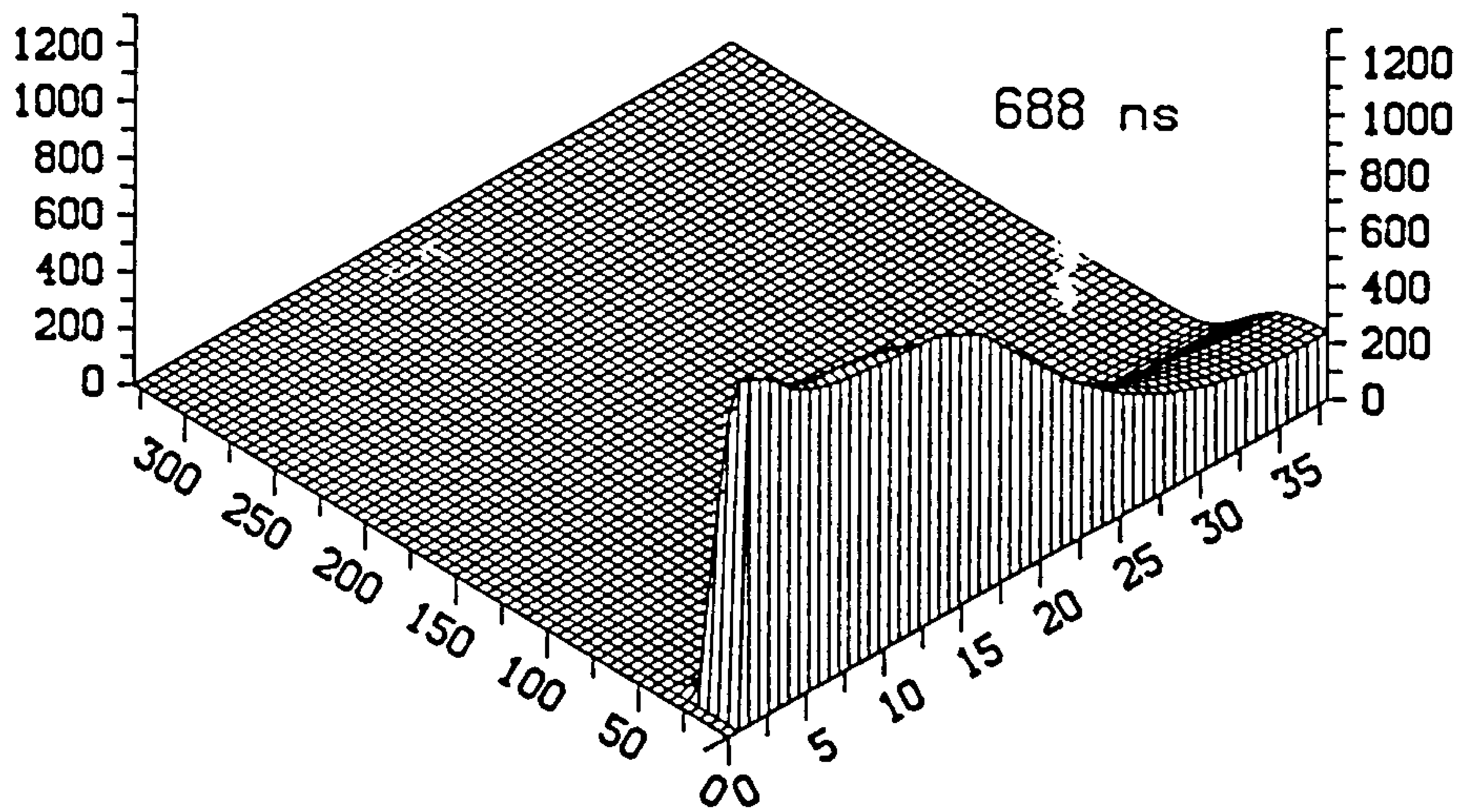


Figure 56. Vertical Electron Current Densities at Emitter-Base, Base-Collector and Collector-Substrate Junctions During Switch-Off: Current densities are in $A\ cm^{-2}$ and distances are in μm.





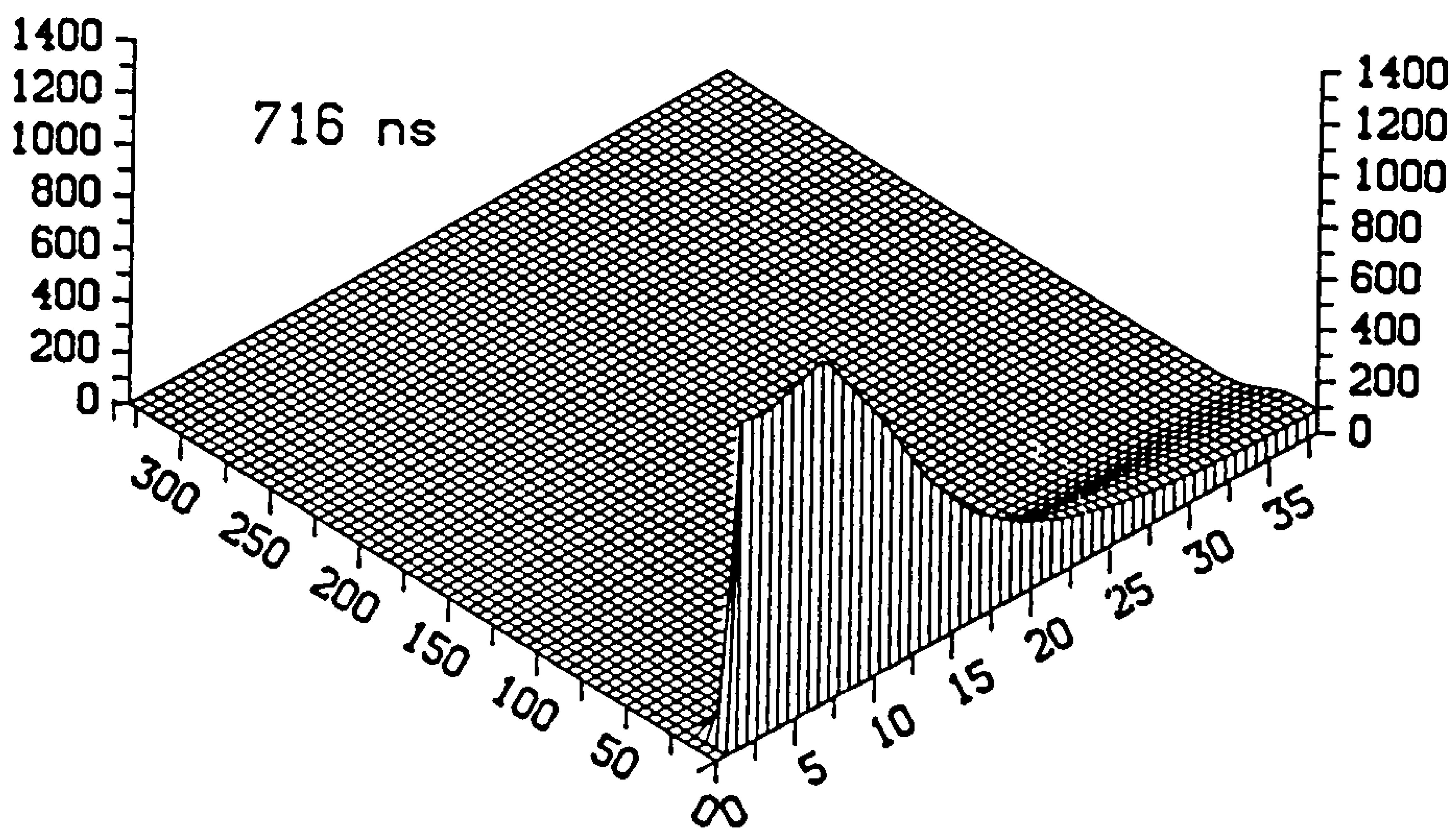


Figure 57. Electron Current Density Profiles Showing Divergence: Current densities are in $A\ cm^{-2}$ and distances are in μm.

area. Thus, it can be seen from Figure 52 at 462*ns* that the collector voltage is now starting to rise. By this time much of the charge in the collector has either been extracted from the device or lost via recombination. The collector current consists of two separate components; the first being the electron current originating from the emitter, which flows across the base to the collector by normal transistor action. The second results from the removal of electrons from the charge stored in the collector region that does not lie beneath the emitter. By 462*ns* the stored charge at the base-collector junction has dropped well below the background donor concentration. Thus, a space charge region is formed, the junction begins to recover and the collector voltage starts to rise. This rise in collector voltage causes operation to enter the quasi-saturation region shown in Figure 53. It can seen from Figure 54 that the hole concentration profile has receeded away from the collector-substrate junction beneath the emitter. This reveals an small ohmic region of the collector and a current induced base is now evident. Both these phenomena are associated with quasi-saturation. Hence, beneath the emitter the developing collector voltage is supported across an ohmic resistance and beneath the base contact the voltage is supported by a depletion region.

At a time of 542*ns* a greater portion of the emitter edge has recovered. The collector voltage has risen to 7*V* causing the depletion region beneath the base contact to widen. The current induced base has receded further revealing a larger ohmic collector region, and the device remains in quasi-saturation. By 614*ns* the emitter is only conducting over half its radius, the base-collector depletion region has expanded further, and the current induced base width continues to decrease. However, the collector voltage has now risen to 12*V* creating a field which exceeds the value required to cause the velocity of the electrons passing through it to saturate ($\simeq 20KVcm^{-1}$) . Thus, a mobile negative space charge region is set up, which supports the collector voltage and the operating point moves into the high field region as shown in Figure 53. At 688*ns* the collector voltage is clamped at the supply voltage by the freewheeling diode and the emitter injects only at its centre. The current induced base now has a very small radius and the collector charge has been almost entirely removed. It can be seen from Figure 52 that all the charge is recovered by 726*ns* and the collector current is reduced to zero.

The vertical electron current densities at the emitter-base, base-collector and collector-substrate junctions are illustrated in Figure 56 at various times after operation has entered the high field region. At 614*ns* the emitter can be seen to conduct only over half its radius as stated previously. The electrons diverge as they pass to the collector contact causing the peak current density to fall and the current density profile to spread out. However, approximately two thirds of the collector current is supplied from the charge stored in the collector. Although the

current density emanating from the stored charge is much less than that flowing down the centre of the device, it flows over a much greater area because of the cylindrical geometry and, therefore, constitutes a larger current.

At 679*ns* the current density has risen at the emitter owing to the increasing collector voltage. However, the rise in peak current density at the collector-substrate is very small in comparison. This is explained by noting that the radius of the current filament is smaller. Thus, the large lateral diffusion gradients existing at the edge of the filament are able to penetrate to the centre of the filament in the time it takes for the electrons to flow from the emitter to the substrate. This effect is nicely illustrated in Figure 57, which shows that as time advances and the filament radius becomes smaller the effects of divergence are more pronounced. At 614*ns* the current filament is sufficiently thick at the emitter to prevent a large reduction of the current density at the centre. However, at 688*ns* the filament is sufficiently thin such that the current density starts to fall off at $y \simeq 15\mu m$. By 716*ns* the filament radius is only about $5\mu m$ and the current density falls off soon after the electrons have been injected from the emitter ($y \simeq 6\mu m$). The current component originating from the stored collector charge cannot be maintained at 688*ns*, and a sharp fall in the collector current is apparent from Figure 52, together with a corresponding sharp rise in collector voltage.

The variation in the peak field at the base-collector and collector-substrate junctions is shown in Figure 58, and a number of field profiles at various points in time are given in Figure 59 and Figure 60. The peak field at the collector-substrate junction is situated directly beneath the centre of the emitter and it reaches a maximum value of $68KV\,cm^{-1}$ just before the collector voltage becomes clamped. This value is just below that required for breakdown. However, the existence of such high fields is obviously extremely undesirable. If the on-state collector current and/or the clamp voltage were slightly higher then the field at turn-off could have been large enough to induce breakdown. The value of the load resistor would need to be decreased to obtain a higher on-state current, for a given steady state base-emitter voltage or base current. Hence, the transistor would not be as heavily saturated and the ratio of electron current flowing from the emitter to collector to that which would be required to supply the collector plasma would be increased. Thus, the smaller plasma could be removed well before the emitter recovers and the current induced base would be much wider as the collector voltage rises sharply. This would inevitably lead to higher fields at breakdown. From the point of view of inductive switching, therefore, it which seem beneficial to drive the transistor into heavy saturation prior to switch-off, though this would result in longer storage times.

Since the high fields developed during turn-off are a direct result of current pinching towards the centre of the emitter, it would seem desirable to try and eliminate this effect. The most obvious technique would be to remove the centre portion of the emitter. This is easily accomplished by masking off the required region during processing.

In order to investigate such a design alteration the above transient analysis was repeated, but this time a circular centre portion of the emitter with a radius equal to half the original emitter radius was masked off. All other device attributes were left unchanged, and the device was biased into saturation with the same initial base and collector voltages. The initial base and collector currents were found to be almost identical to those for the original circular emitter structure. This is accounted for by the emitter pinch effects discussed in section 6.2. This effect

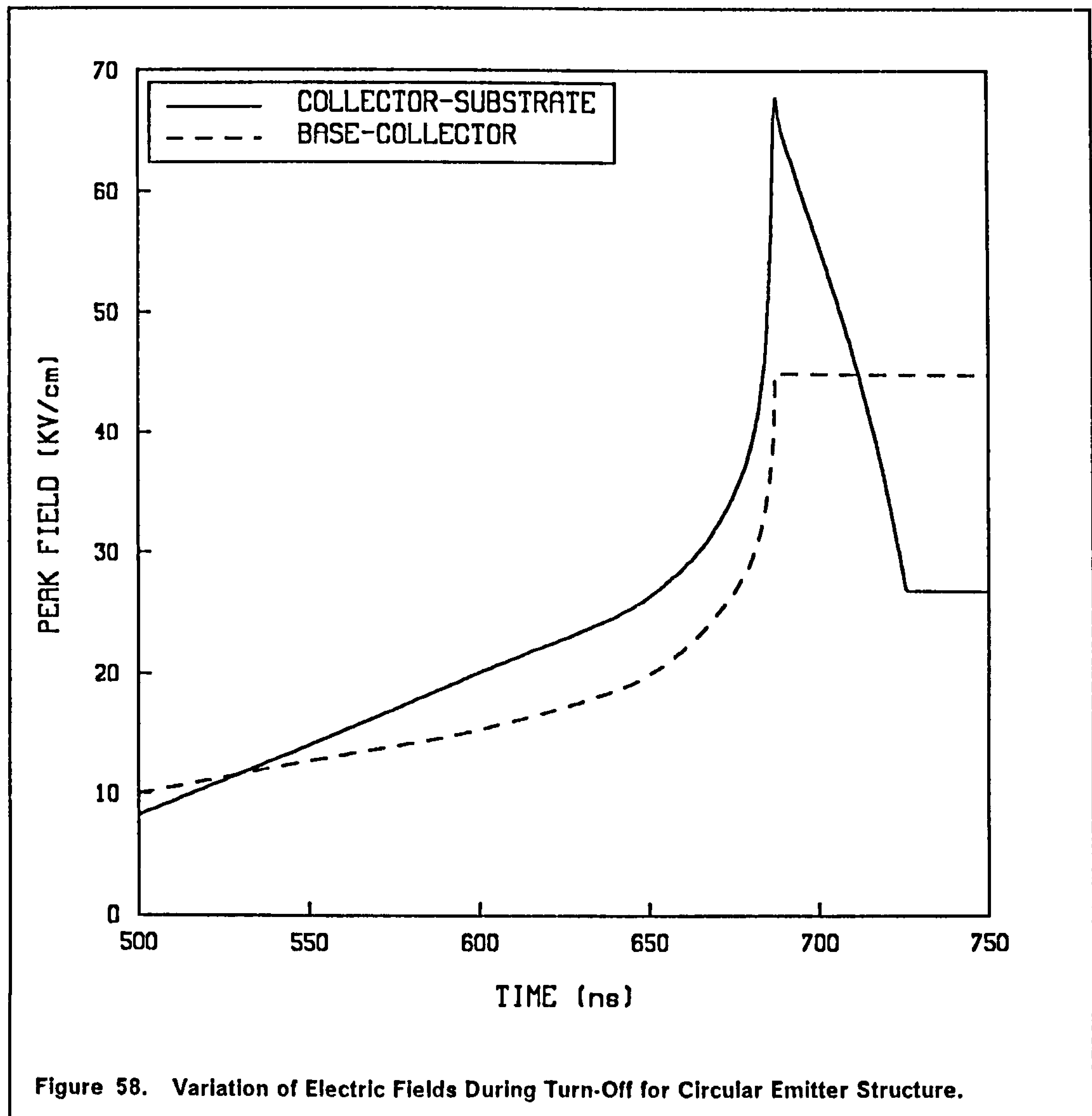
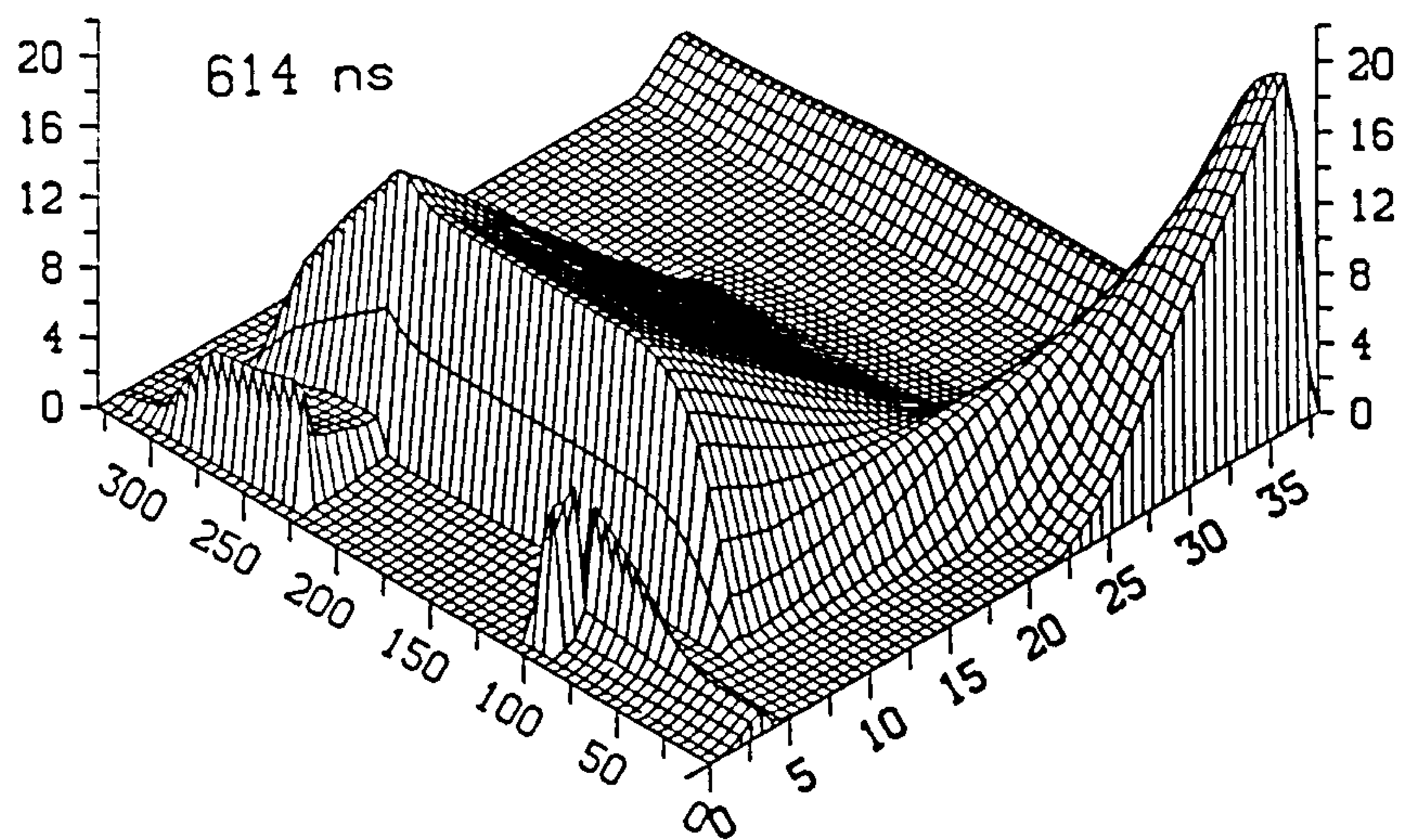
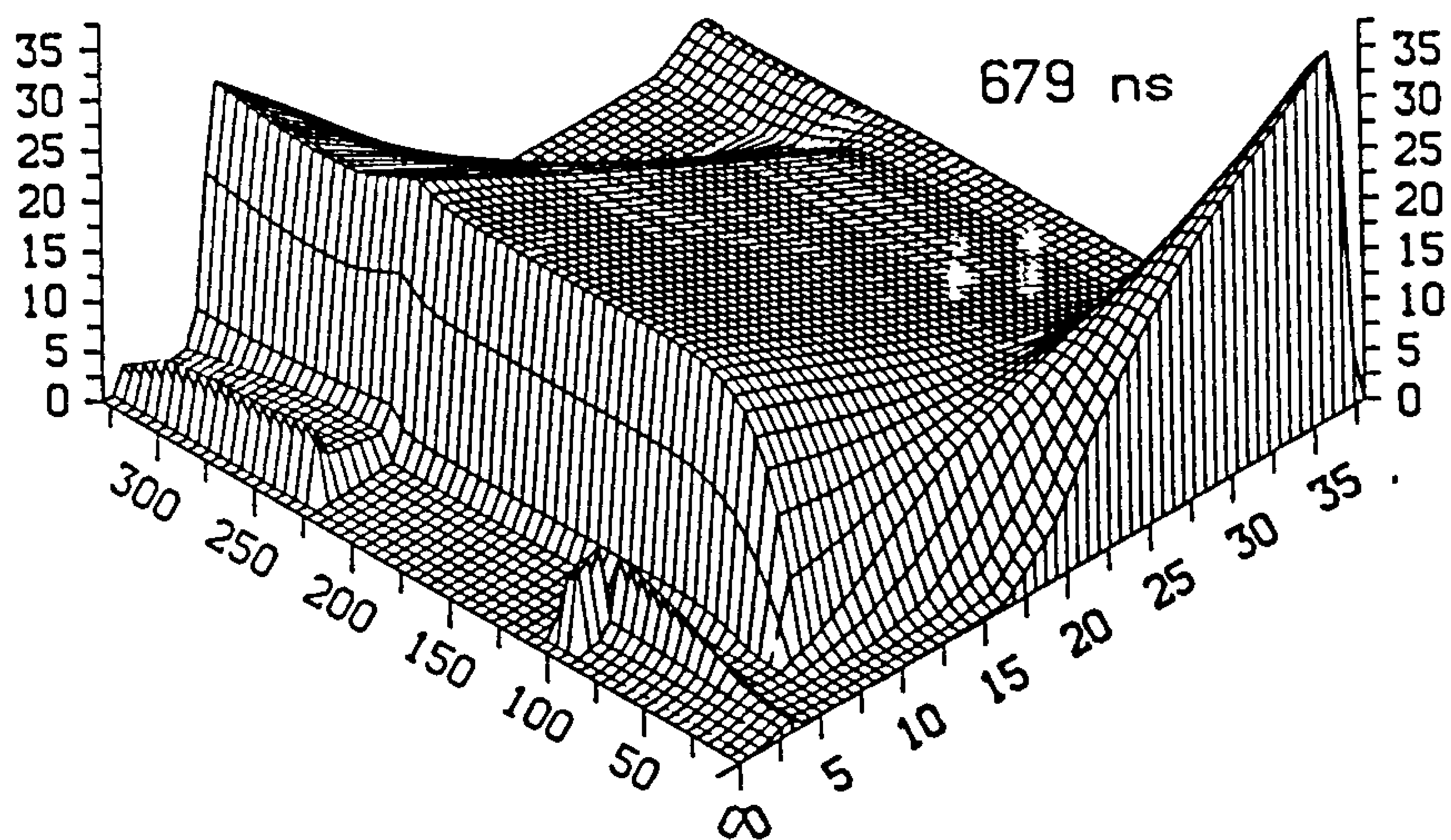


Figure 58. Variation of Electric Fields During Turn-Off for Circular Emitter Structure.





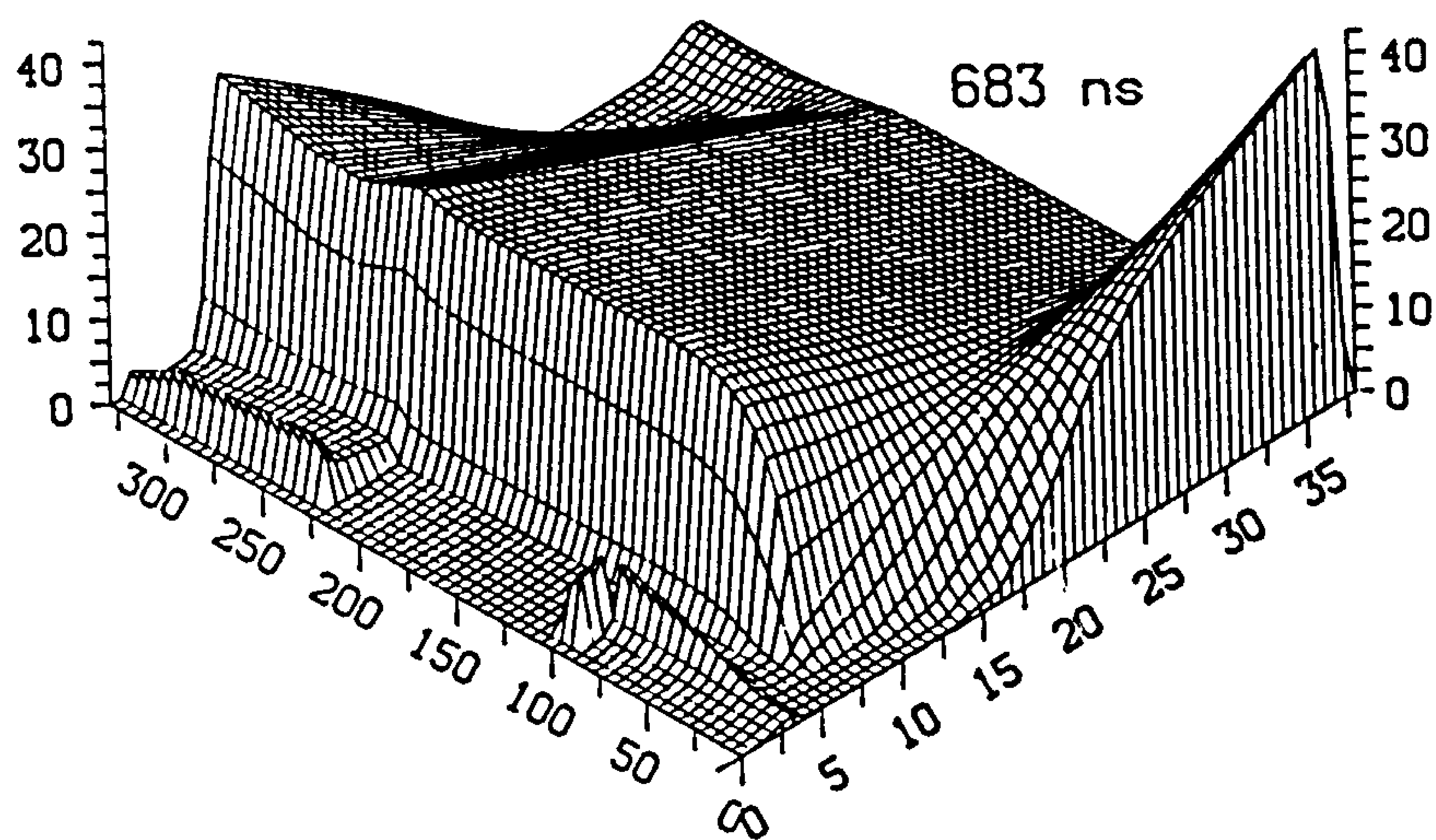
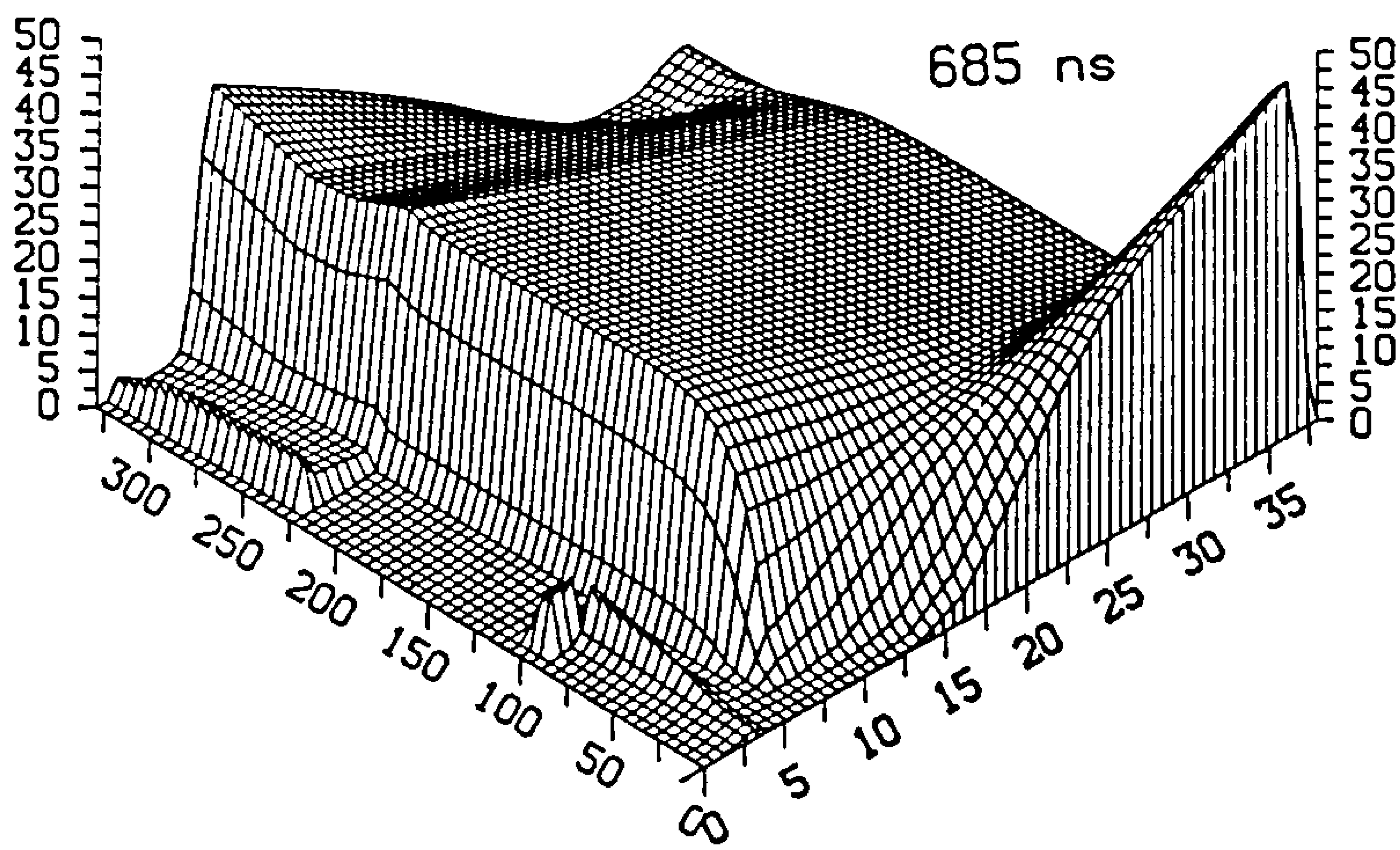
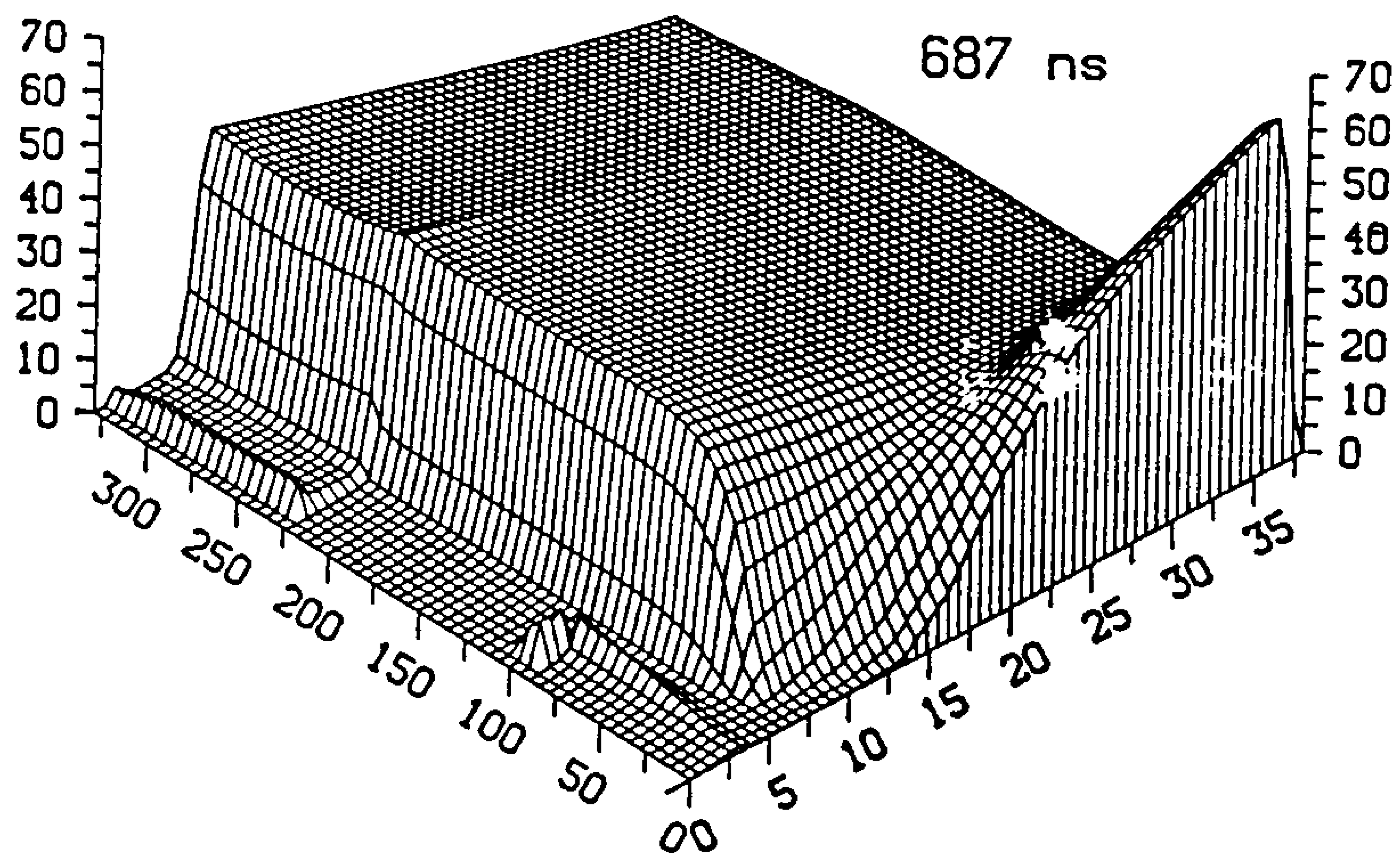


Figure 59. Electric Field Profiles at Various Times During Turn-Off.: Fields are in $KV\,cm^{-1}$ and distances are in μm.





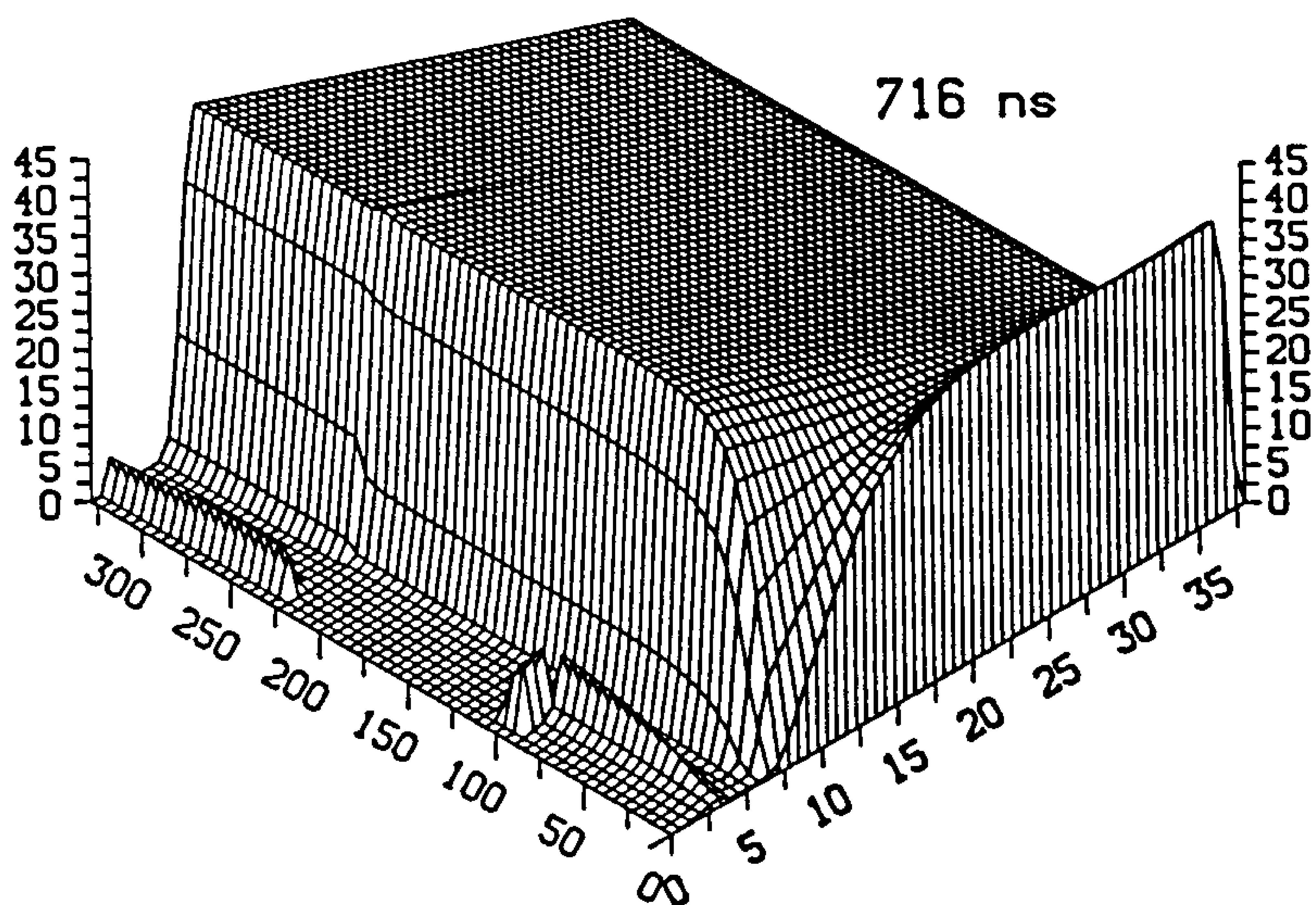


Figure 60. Electric Field Profiles at Various Times During Turn-Off.: Fields are in $KV\,cm^{-1}$ and distances are in μm.

is present even under saturation conditions where the base and collector are heavily conductivity modulated. Although heavily modulated, the resistance of the base region is still adequate to give significant debiasing of the emitter centre. Thus, the emitter centre is virtually redundant under all operating conditions except during turn-off. As a result the load resistance value is virtually the same.

The base voltage was then reduced with the same waveform as in Figure 52. It was found that the base current, collector voltage and collector current waveforms were very similar to those for the circular emitter. This is not too surprising when it is considered that the charge stored in the collector dominates the turn-off characteristics. Since the base and collector specifications have not been altered then the amount of stored charge should be close to that of the original device.

The variation of the peak fields at the base-collector and collector-substrate junctions for the ring emitter structure are shown in Figure 61. The x and y axes are the same as in Figure 58 to allow accurate comparison. In this case it can be seen that the peak field at the collector-substrate junction stays below the peak field at the base-collector junction for the entire turn-off transient. This is in complete contrast to the case for the circular emitter. The peak field at the collector substrate junction initially rises very slowly. At this stage the collector voltage has started to rise and the ring emitter is still injecting around its inner radius. The region around the inner radius is in quasi-saturation and the peak field at the collector-substrate junction lies directly below the inner radius. Field profiles at various times before the emitter stops injecting are given in Figure 62, and the vertical current densities are shown in Figure 63. The field peak reaches a maximum at 620*ns* and then falls as the emitter-base junction around the inner radius starts to recover. Meanwhile the collector current is maintained by the supply of electrons from the charge stored in the collector. The emitter-base junction is fully recovered by about 650*ns* and it is significant that this occurs before the collector voltage has risen to a high value. The peak field as a result of current flow from the emitter, therefore, only reaches a value of 13 $KV\,cm^{-1}$. In contrast the emitter-base junction of the circular emitter device does not recover until after the collector voltage has reached the supply potential, resulting in the rather severe field spike shown in Figure 58.

In the case of the ring emitter structure the fields at both junctions rise to their steady state values as the collector voltage reaches the supply potential. The field at the collector-substrate junction never exceeds its off state value in the switching duration. The collector current falls off much faster than is shown in Figure 52 once the collector voltage is clamped, since in this case the emitter-base junction has already recovered.

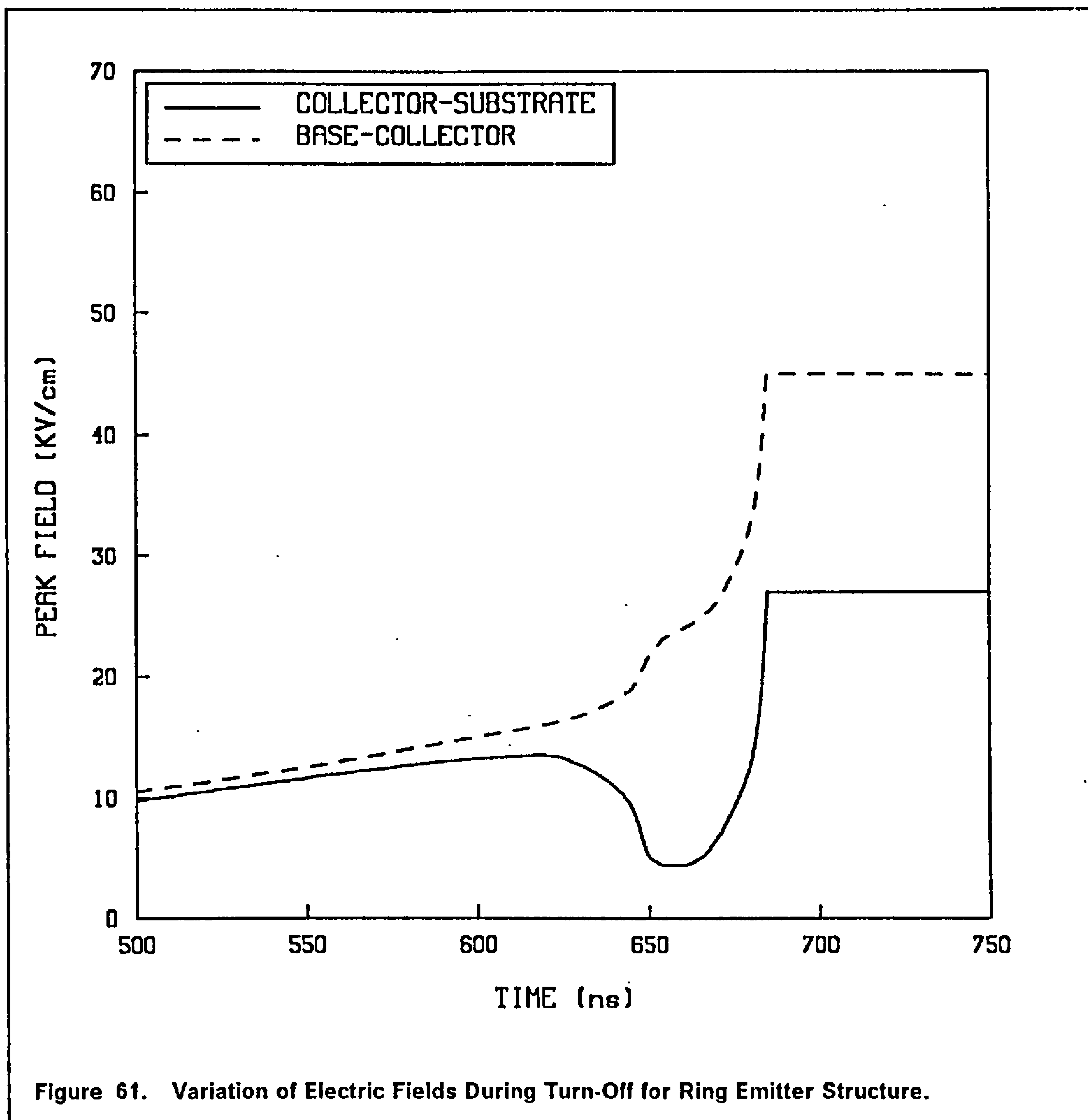
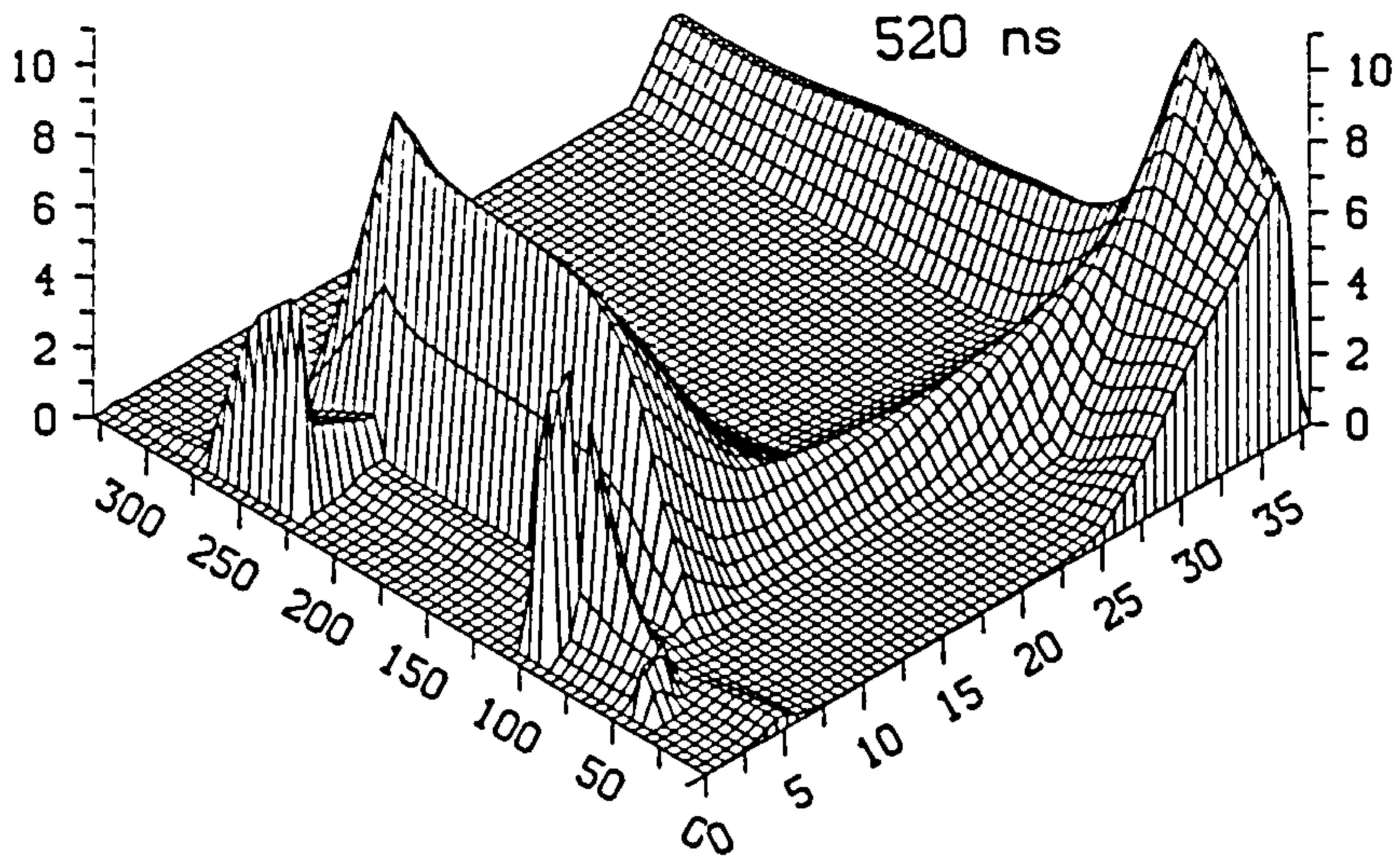
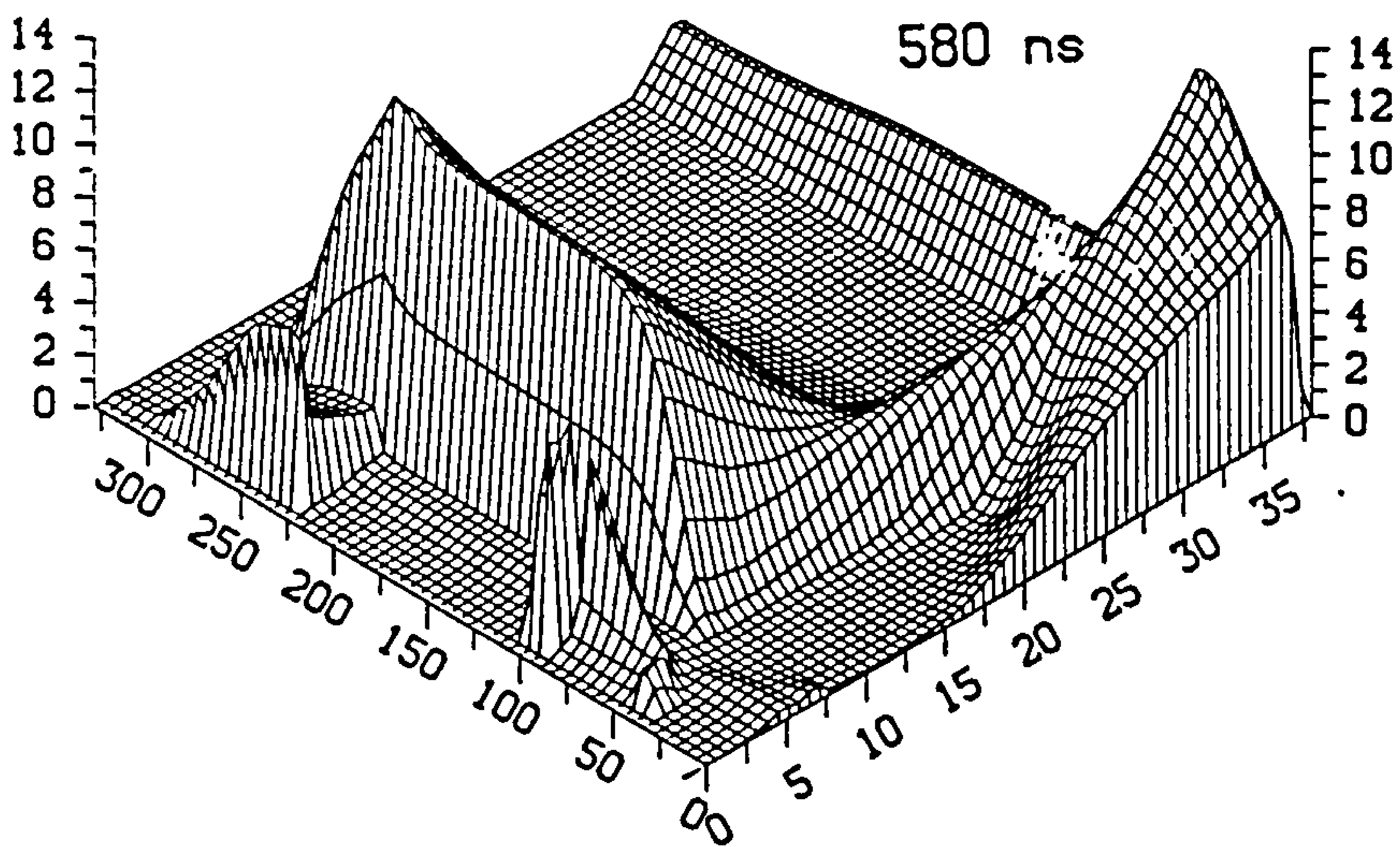


Figure 61. Variation of Electric Fields During Turn-Off for Ring Emitter Structure.

6.4.4 Discussion.

The locus of current and voltage values which arise during turn-off and which are large enough to cause breakdown defines the so called Reverse Bias Safe Operating Area (RBSOA) of the device. The series inductor in Figure 50 is usually large enough to give a rapid increase of collector voltage for only a small fall in collector current. In this case, therefore, the RBSOA defines the maximum on-current and corresponding maximum clamp voltage. An estimated RBSOA for the circular and ring emitter structure is shown in Figure 64. A different RBSOA exists for a different value of initial forward base current and/or a different base turn-off waveform. The maximum allowable collector current at low collector voltages is limited by the current handling capability of the bonding wire, bonds or emitter metalization. The maximum allowable voltage can rise above BV_{CEO}





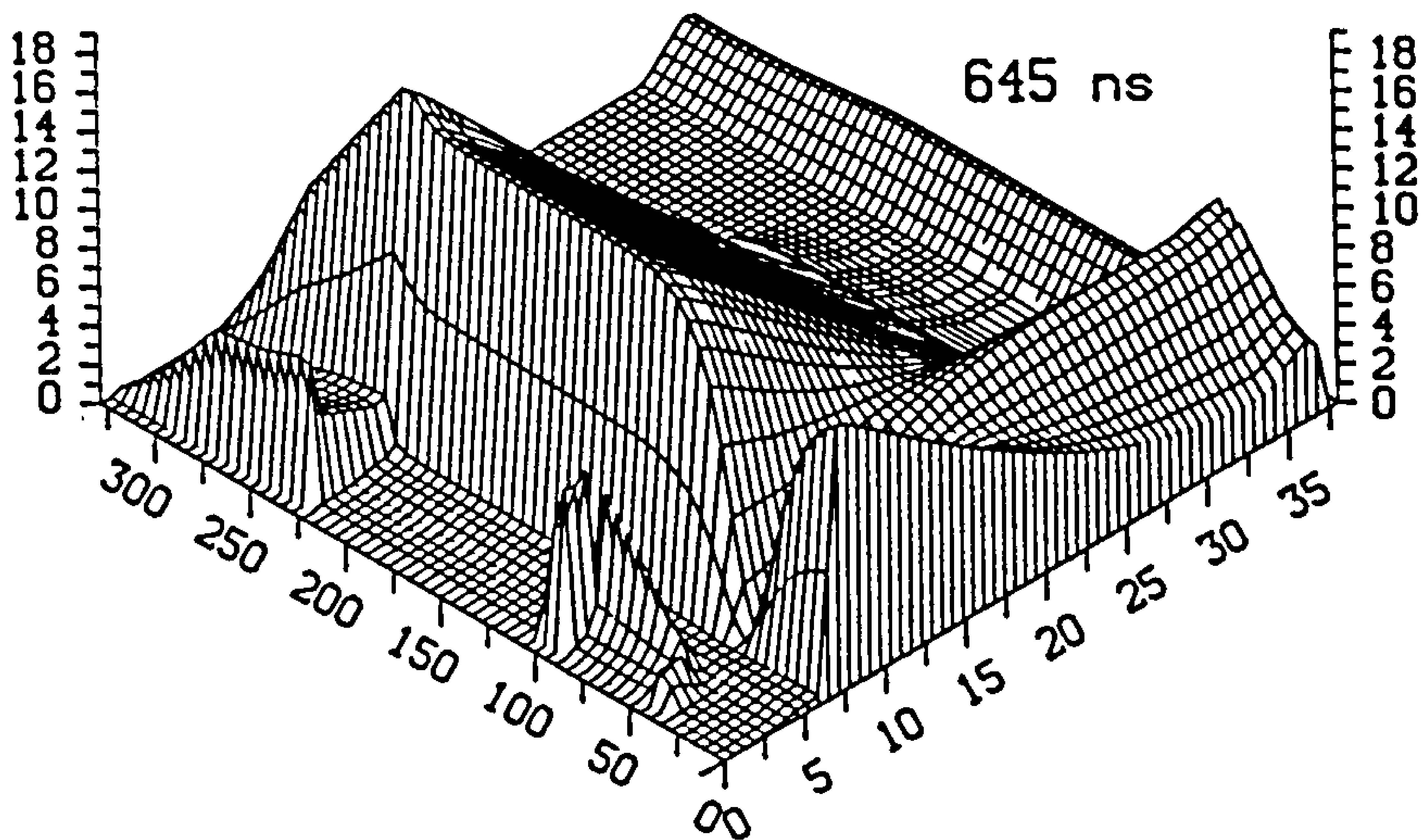


Figure 62. Electric Field Profiles at Various Times Before the Emitter-Base Junction Has Recovered.: Fields are in $KV\,cm^{-1}$ and distances are in μm.

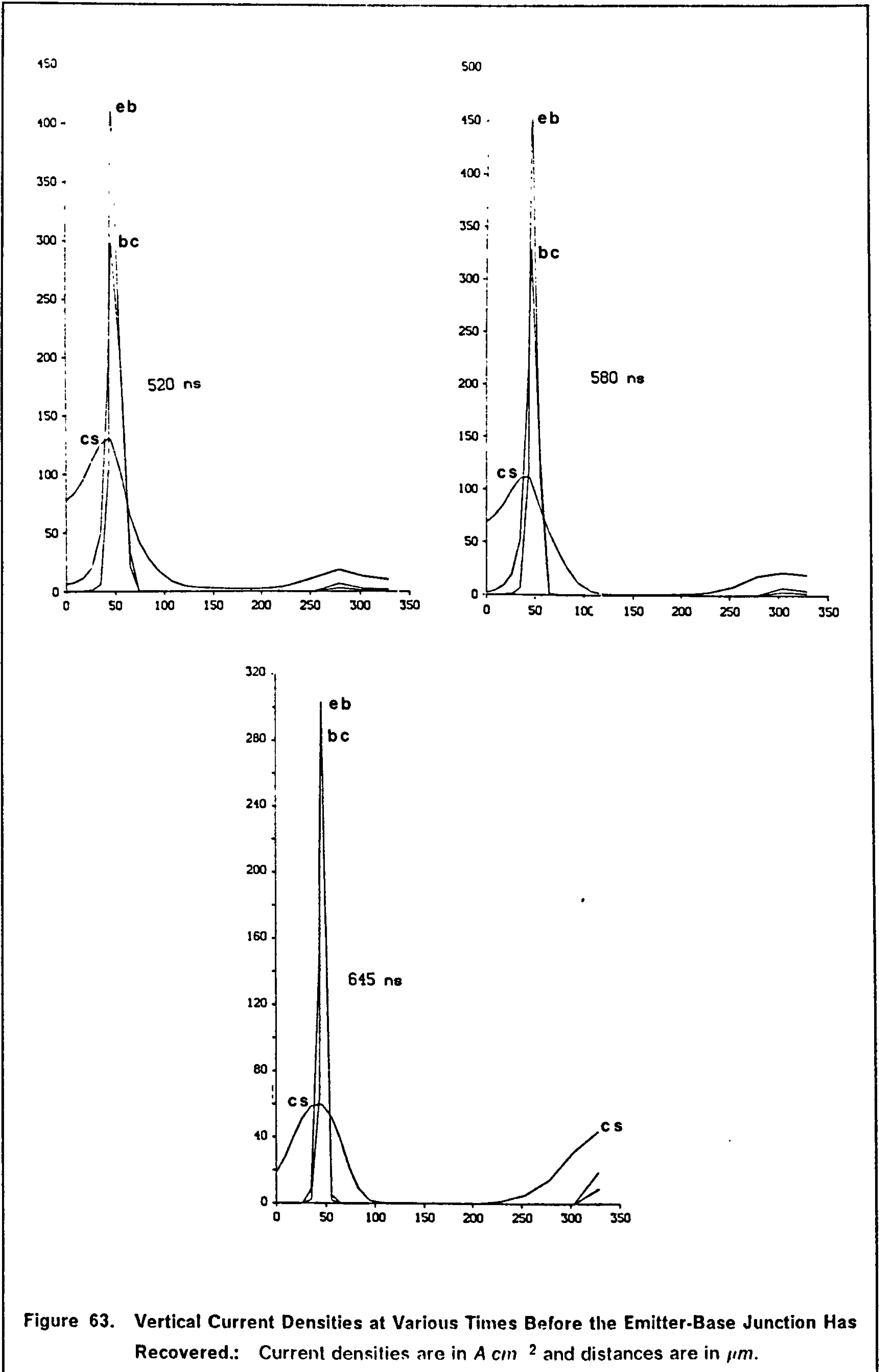


Figure 63. Vertical Current Densities at Various Times Before the Emitter-Base Junction Has Recovered.: Current densities are in $A\ cm^{-2}$ and distances are in μm.

provided the holes generated by avalanche ionisation can be removed via the base contact. For large enough reverse base current the maximum allowable voltage can even reach the collector-base junction breakdown voltage, BV_{CBO}. The curved part of the RBSOA represents the limit above which avalanche injection occurs. On the basis of the previous results it is expected that for the ring emitter structure this limit can be extended well above that for the circular emitter structure. In fact it is quite possible that the RBSOA could even be rectangular since no excess fields seem to be induced during turn-off.

A number of power bipolar transistors have been fabricated which utilize the hollow emitter concept [6.24] [6.25] [6.26]. In the first case, [6.24] an interdigitated structure has been reported, which incorporates an emitter stripe having a shallow, low doped n-type region along its centre. Consequently this centre region did not contribute any injected electrons. In this case only a 10% improvement in current handling at a clamp voltage of 100V was reported in comparison with a similar device with a full emitter. This could be due to the removal of an insufficient width of emitter, and thus, significant pinching could still have occurred. In this case it is probable that the collector voltage increased before recovery of the emitter-base junction, resultinq in little improvement in RBSOA performance.

In [6.25] a cellular type geometry has been fabricated and tested. Two similar structures with ring like emitters have been compared with a device that did not have its centre masked off. The two ring emitter devices showed a dramatic 100% improvement in the current that could be handled at any particular voltage all the way up to a BV_{CBO} value of 1200V. Another cellular design has been constructed [6.26], which uses a double layer metalization technique to separately contact the base and emitter. In this case the emitters were square shaped with a large square centre portion masked off, such that the emitter width was only 5μm. The first metalization layer was formed into a grid to contact the base region around each emitter. The upper metalisation was insulated from the first by an oxide layer and made contact with the emitter via a polysilicon ballast resistor. For this structure a fully rectangular RBSOA was reported, which extended all the way up to a BV_{CBO} of 1000V. In this case it is almost certain that the emitter recovers before all the stored charge is removed since the emitters are extremely narrow.

In [6.27] a numerical model has been used in a similar analysis to that described here. However, in this case a single cell of an interdigitated structure has been modelled, and the free-wheeling diode has been omitted. Again the collector current could not be maintained by the stored charge and the field at the collector-substrate junction increased rapidly during turn-off. For a striped

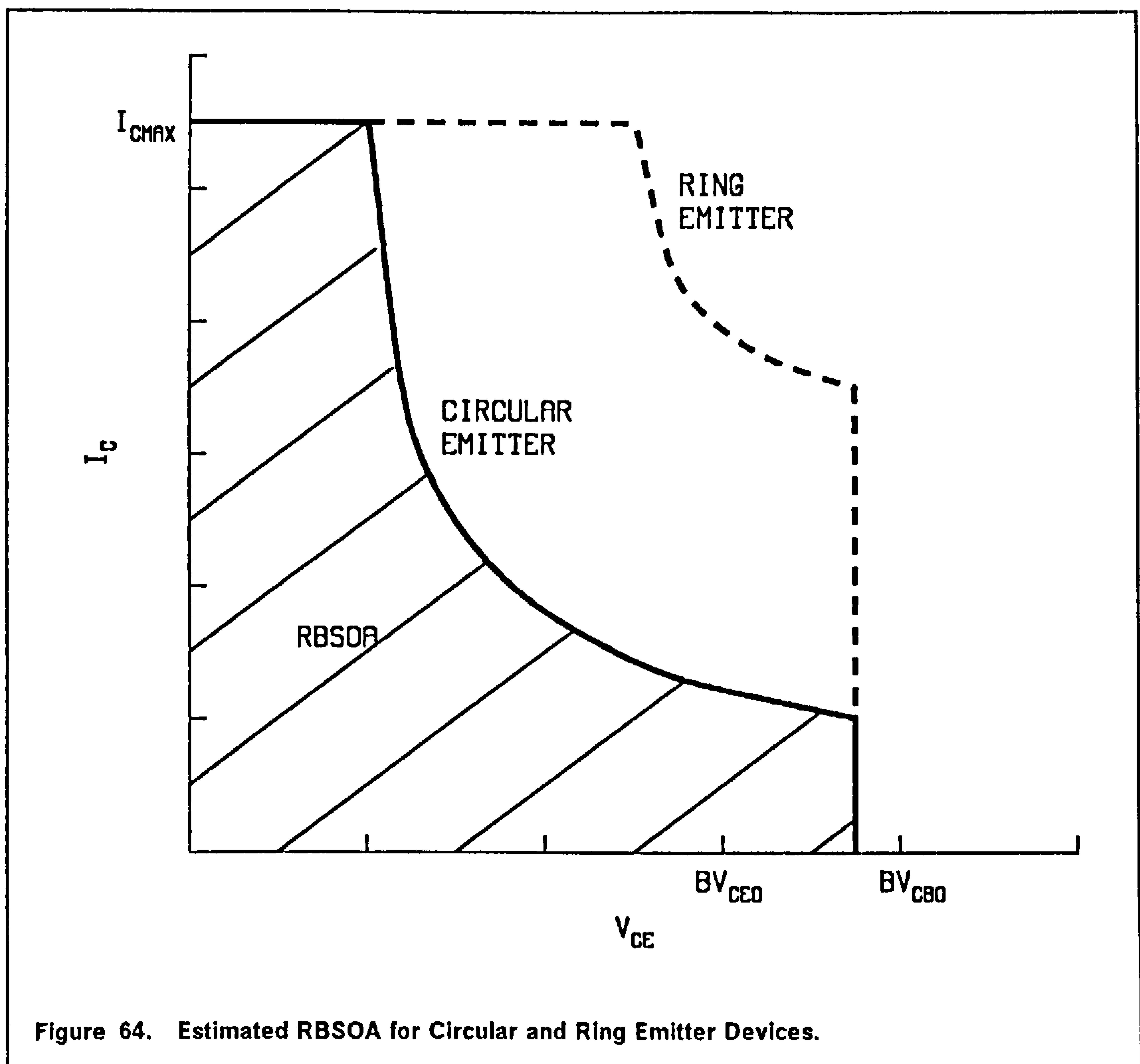


Figure 64. Estimated RBSOA for Circular and Ring Emitter Devices.

geometry a sharp ridge of high current density arises at turn-off, rather than a current filament in the case of the circular emitter. Comparison was made with a device where the centre of the stripe had been masked off. This causes the current ridge to be split into two smaller ones, running down the inside edge of each emitter finger. Thus, the mobile space charge is reduced and the fields are smaller at a particular collector voltage as the voltage rises. A field reduction of approximately 20% at any particular voltage up to 1000v was reported in comparison with the full emitter geometry.

A second interesting comparison has been made in [6.25], which illustrates the effect of reducing the emitter size on the RBSOA. A geometry was proposed which would consist of a uniform array of solid emitter squares surrounded by a base contact grid. Three such designs were fabricated, each with a different size of emitter square, but with approximately the same over all active chip area. It was found that the design with the smallest emitter gave over a hundred percent improvement in current handling capability at any particular collector voltage,

relative to the design with the largest emitter. This can be explained by noting that the device with the smaller emitters can contain many more emitter squares per unit area. Thus, should current filamentation arise during switch-off then the current density within each filament should be less as a greater number of filaments will exist, provided that the current is shared equally between emitters. In addition divergence of current can result in a uniform current density within the collector layer, which minimizes the peak field at the collector-substrate junction (cf. section 6.2). Susceptibility to avalanche injection should therefore be minimized. It is also quite possible, for a fine enough emitter, that emitter injection will have terminated before a significant voltage rise, as the previous results have shown.

The use of very fine emitters together with high packing density seems to be a very attractive technique for improving the RBSOA. However, problems arise, particularly in interdigitated devices, if the emitter contacts are made very narrow. This is because the resistance along the contact stripe becomes significant, causing considerable de-biasing of the emitter-base junction at the opposite end from the contact pad. Recently, the problem of simultaneously achieving low contact resistance and high packing density has been achieved by utilizing double layer metalization, which is a more modern technique usually employed in power MOS fabrication. Such techniques also lend themselves very well to the fabrication of cellular structures.

It is expected that devices with fine emitters and high packing, be they stripe or cellular geometry, will in general exhibit improved RBSOA performance compared to the hollow type emitter structure modelled here. This is because for the hollow emitter case the emitter must be made reasonably large since a sizeable hollow region is required to prevent filamentation, and this consumes valuable chip area. Moreover, the results given here, have shown that the removal of the emitter centre does not affect the on-state operation because of predominant emitter pinching. This suggests that the emitter width was too large and its full area was not being utilized in the first place. However, if many fine emitters could be packed into an equivalent chip area then the electron current would be more evenly spread and could even become uniform before reaching the collector-substrate junction. This will result in optimum combined current and voltage handling.

6.4.5 Conclusion.

It has been shown that if enough charge is stored in the collector region to maintain the collector current until after the emitter base junction has fully

recovered then a dramatic improvement in the RBSOA will result. This leads to a desire to minimize emitter dimensions and maximize packing density, but this implies reducing the volume available to store the charge that allows filamentation to cease before voltage recovery. However, if the emitters dimensions are small they should recover sooner in the turn-off transient and the amount of stored charge required should be reduced. Even if the collector current could not be maintained by the stored charge, the resulting uniformity of current would tend to minimize fields. In this respect this conclusion is consistent with that obtained in section 6.3 to maximize the forward bias safe operating area. Both indicate that the trend should be towards finer emitter geometries to improve device robustness, provided that the problem of contacting to such a design can be overcome.

6.5 Electro-thermal Interaction and Thermal Second Breakdown in Bipolar Transistors.

6.5.1 Introduction.

Since it was first reported in 1958 [6.28] the second breakdown phenomenon has received extensive coverage in the literature. Some of the more notable publications are [6.29], [6.30] and [6.31]. Unfortunately second breakdown remains rather a grey area despite all this effort. However, it is generally accepted that second breakdown can be induced either by thermal means (thermal second breakdown or TSB) or by subjecting a device to avalanche injection conditions (current mode second breakdown or CSB). Both these mechanisms lead to a sudden collapse of the collector to emitter voltage, which is indicative of second breakdown. CSB is initiated when the local current density at a particular applied voltage is of sufficient magnitude to give rise to avalanche injection. This results in filamentary current flow and the voltage collapses to a value that is sustained by the mobile space charge in the multiplication zone at the collector-substrate junction [6.32]. This process does not require any heating and depends only upon the instantaneous current/voltage combination applied to the device. Transistors are particularly susceptible to CSB when used for inductive switching as discussed in Section 6.4. The voltage sustained in CSB never varies very far from 20V and is relatively independent of device dimensions and doping. The high currents associated with the voltage collapse can cause a rapid temperature rise allowing operation to enter the TSB failure mode, which can result in permanent damage

to the device. In order to avoid this the duration of the low voltage state must be carefully controlled.

Entry into TSB depends not only on the current/voltage combination, but also on length of time for which this combination is applied. Thus, a delay time exists between the instant that the power is applied and the point at which TSB occurs. This delay is known as the triggering time and is an important quantity in accurately defining the transistor safe operating area. TSB is initiated via a so called lateral thermal instability, which arises as a result of the positive temperature coefficient of the forward current across the emitter-base junction. Thus, if the temperature varies laterally along this junction, then preferential injection will arise where the temperature is highest. This will increase the power dissipation and, therefore, the temperature at this location and the mechanism will become regenerative. Eventually, the local temperature will exceed the intrinsic temperature of the material [6.33] resulting in significant thermal generation of electron-hole pairs, especially in the collector space charge region where power dissipation is highest. A plasma of electrons and holes forms a low resistance bridge across the collector space charge region, which is known as a mesoplasma. The voltage sustained in the low voltage state then depends on the ohmic potential drops across the spreading resistances to the plasma [6.34]. This voltage is typically $10V$ and is also a fairly weak function of device dimensions and doping. A regenerative self-heating mechanism can also bring about the conditions necessary to induce CSB, prior to entry into TSB resulting in a reduction in the triggering time. Temperatures within the mesoplasma can easily reach the melting point of silicon (1412°C), but it is more likely that the eutectic temperature of the aluminium-silicon system at the contacts will be attained first [6.33].

In this section the combined electrical and thermal phenomena are simulated with the aid of the full numerical model as described in Chapters 4 and 5. Since avalanche multiplication effects are not included in the recombination/generation model the effects leading to CSB cannot be taken into account. However, the regenerative effects leading to TSB should be accurately reflected by the model. In order to reduce the possibility of the occurrence of avalanche breakdown prior to thermal breakdown, the initial current/voltage combination applied to the device was chosen such that the peak field was well below the value required to induce multiplication.

In this section the combined electrical and thermal phenomena are simulated with the aid of the full numerical model, as described in Chapters 4 and 5. Attention will be confined to the cylindrical type geometries ie. device types B and C, so that the two dimensional approximation can be avoided. This

approximation is expected to be particularly inaccurate now the thermal model is to be included. This is because localized temperature rises will cause the current to redistribute unevenly between and along the emitter fingers. Thus, device operation cannot be modelled by simply considering a single cell half-width, though this could be overcome by modelling several emitter fingers at once. Devices B and C are relatively easy to model because they do not have many emitters and heat flow is inherently modelled in all directions.

In addition, a non-destructive test circuit has been constructed for the purpose of pulsing the devices into thermal second breakdown. As soon as the voltage collapse is detected the transistor under test (TUT) is turned off, which protects the device from permanent damage. A description of this circuit together with some typical results will now be presented. Some numerical results will then be considered and a comparison with experimental results will be made.

6.5.2 A Non-Destructive Test Circuit for Pulsing Transistors into Thermal Second Breakdown.

The circuit shown in Figure 65 was developed to study the behaviour of bipolar transistors during the delay time prior to entry into TSB. The circuit was designed to maintain both a constant collector current and a constant collector voltage for the entire duration of the delay time. This is an extremely desirable feature as it allows the transistor safe operating area to be defined on the conventional axes of collector current against collector voltage. As soon as the voltage collapse is detected the current flowing through the TUT is switched off to protect the device from extreme overheating. In the following description of circuit operation it is initially assumed that the DMOS transistor in series with the TUT is fully switched on.

When the input (I/P) to the circuit shown in Figure 65 is at $0V$ the resistor network formed by R1, R2 and R3 gives a voltage of approximately $2V$ at the inverting input of OP1. Feedback resistor R4 and R5 form a potential divider giving a voltage at the non-inverting input that is about one hundredth of the collector voltage of the TUT. Therefore, provided the collector voltage is below $200V$ then the output voltage of OP1 will be $-15V$ Diode D1 conducts through R6 and a small negative potential exists at the base of T1, which consequently does not conduct and the TUT is turned off. The circuit will remain in this off state provided the supply voltage, V_{CC} does not exceed $200V$.

If a negative going pulse ($\simeq -3V$) is applied at the input, so that the voltage at the inverting input of OP1 drops below that at the non-inverting input, then the

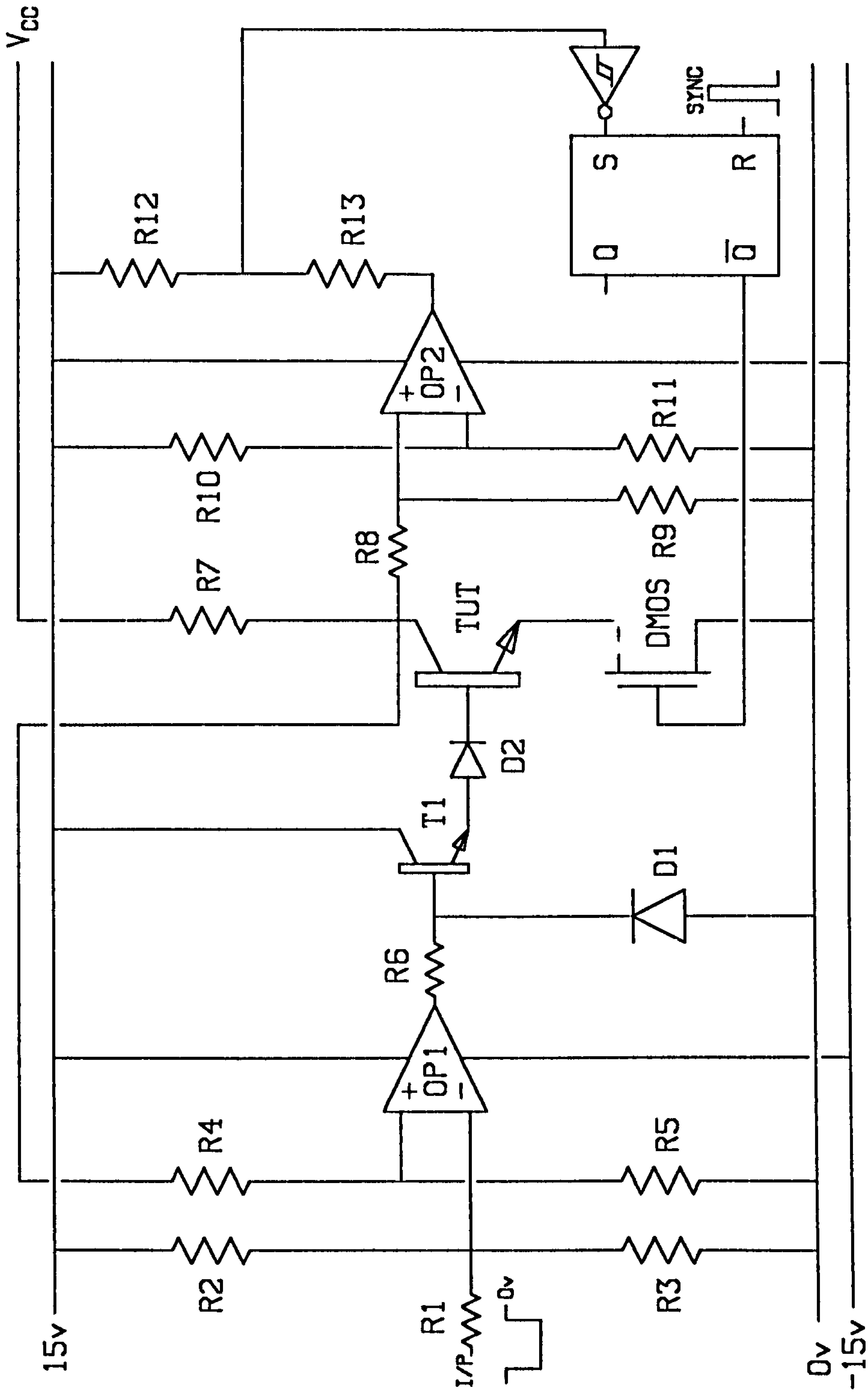


Figure 65. Non-Destructive Test Circuit for Investigating Thermal Second Breakdown.

Component Values	
R1,R3,R5,R6,R9,R11	1K
R2	2.7K
R4,R8,R12,R13	100K
R7	47Ω
R10	47K
D1,D2	IN4001
T1	BFX85
OP1,OP2	LM741
DMOS	Mullard BUZ64
Schmitt Inverter	CMOS 40106
SR Flip-Flop	CMOS QUAD NAND 4011

voltage at the non-inverting input will follow the voltage at the inverting input. The output of OP1 goes positive turning T1 on, which supplies base current to the TUT. The collector voltage of the TUT is now fixed at one hundred times the value at the non-inverting input of OP1 by R4 and R5, which in turn depends upon the amplitude of the input pulse. The collector current is simply given by the difference between V_{CC} and the collector voltage divided by the resistance, R7.

Since the collector current and voltage values are fixed for the duration of the input pulse, then any alteration in the operating characteristics of the TUT due to self-heating must be accommodated for by a change in base current and voltage. It was necessary to include transistor T1 so that sufficient base current could be supplied to the TUT, which operates under conditions of low h_{FE}. The base current requirement is, therefore, often well above the maximum output current rating of OP1.

Should the device enter TSB then the associated collapse in collector voltage is detected by the comparator circuit formed by R8, R9, R10, R11 and OP2. The resistor values were chosen such that if the collector voltage drops below 30V then the output of OP2 switches from +15V to −15V. Resistors R12 and R13 serve to limit the switching range from +15V to 0V to provide the required logic levels for the Schmitt inverter, which interfaces the CMOS SR flip-flop with the analogue circuitry. Thus, when the voltage collapse is detected the Set input of the flip-flop switches to logic 1 (+15V), the $\overline{Q}$ output switches to logic 0 (0V) and the DMOS transistor in series with the TUT is switched off. This inhibits any current flow from collector to emitter and in addition diode D2 prevents the flow of any

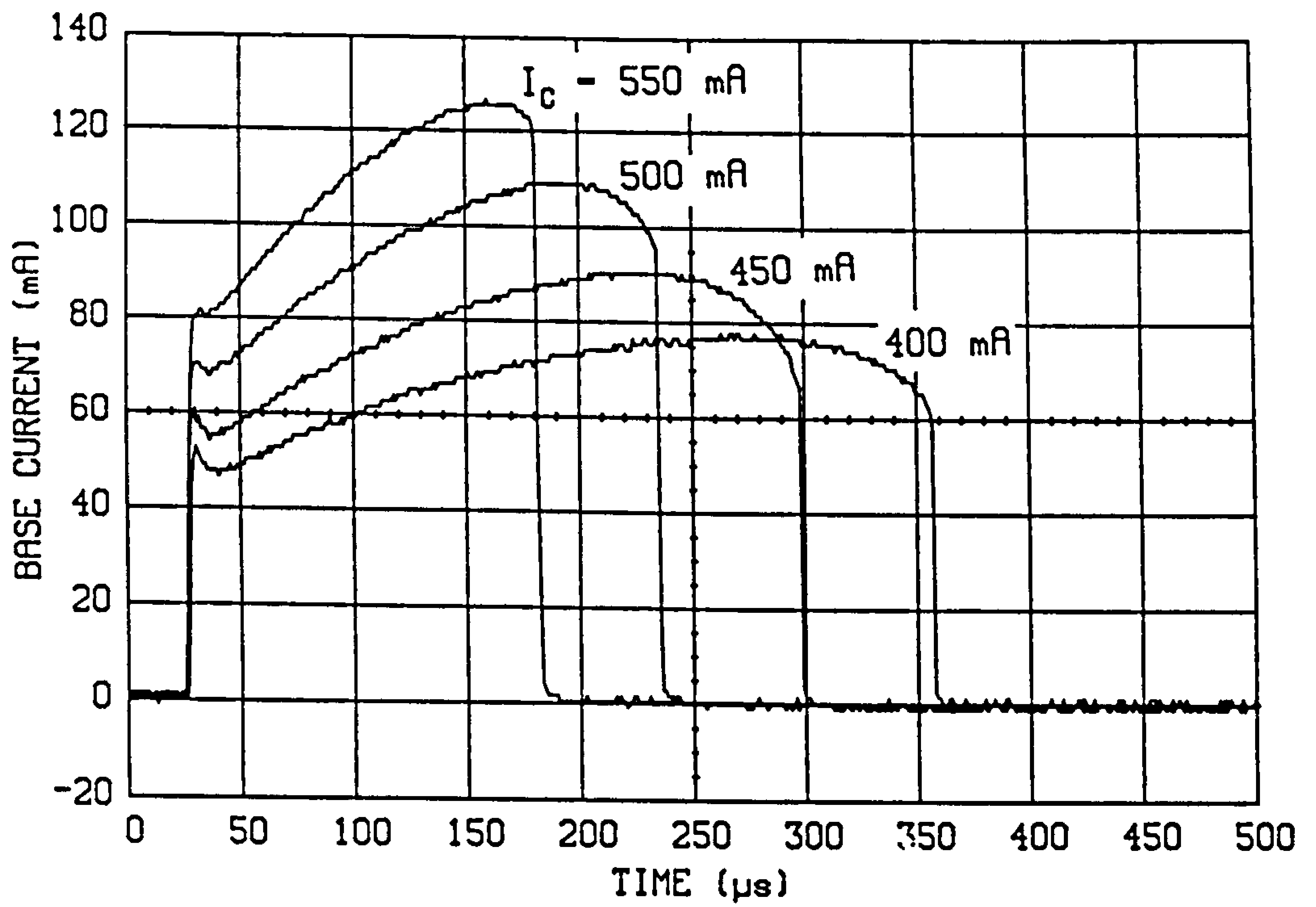
reverse base current. Therefore, no power is dissipated by the TUT, which subsequently cools with a corresponding recovery of collector voltage. The DMOS transistor will remain latched in the off-state until the SR flip-flop is Reset. An appropriate Reset pulse is provided by the prepulse or 'SYNC' output of the input pulse generator.

6.5.3 Experimental Results.

In this investigation only single pulses have been considered, although the circuit could also be used to examine effects such as thermal accumulation under continuous pulsed operation. Base current waveforms have been measured using a current probe and stored on a digital storage oscilloscope. The waveforms illustrated in Figure 66 were taken from Device C ($BV_{CEO} = 250V$, $W_{epi} = 37.5\mu m$, $N_{epi} = 1 \times 10^{14} cm^{-3}$) and in all cases the collector voltage was held fixed at $100V$, while the collector current was varied. It can be seen that the triggering time falls from $7ms$ to $150\mu s$ as the collector current is increased from $200mA$ to $550mA$. Only a brief explanation of the general shape of the base current waveforms will be given here; a more detailed account will be given in the following section.

The initial rise of the base current may be attributed to the temperature dependence of the limiting drift velocity of the electrons forming the mobile space charge region, which supports the collector voltage. Equation (3.18) predicts that the velocity will fall with increasing temperature. This will result in an increase in the concentration of the mobile electron space charge causing further expansion of the current induced base. This results in an increase in the Gummel number of the device and since the collector current is inversely proportional to this number then the base drive must be increased to maintain the specified collector current. The increase in base current is some what offset by the carrier lifetimes, which increase with temperature as discussed in chapter 3.

Eventually the base current waveform levels off and starts to fall. This is because the hole current components originating from thermal generation in the high field region and diffusion from the substrate both increase with temperature. The intrinsic carrier concentration increases rapidly with temperature, approximately doubling every 11°C. and the pn product is very small in the mobile space charge region. The equations for SRH and Auger recombination, therefore, predict an increase in the carrier generation rate with temperature in the high field region. The holes thus generated are swept into the base region to constitute a component of the base current. The second component is due to the hole concentration gradient in the substrate at the edge of the high field region. Holes diffuse upwards under the influence of this gradient and into the high field region



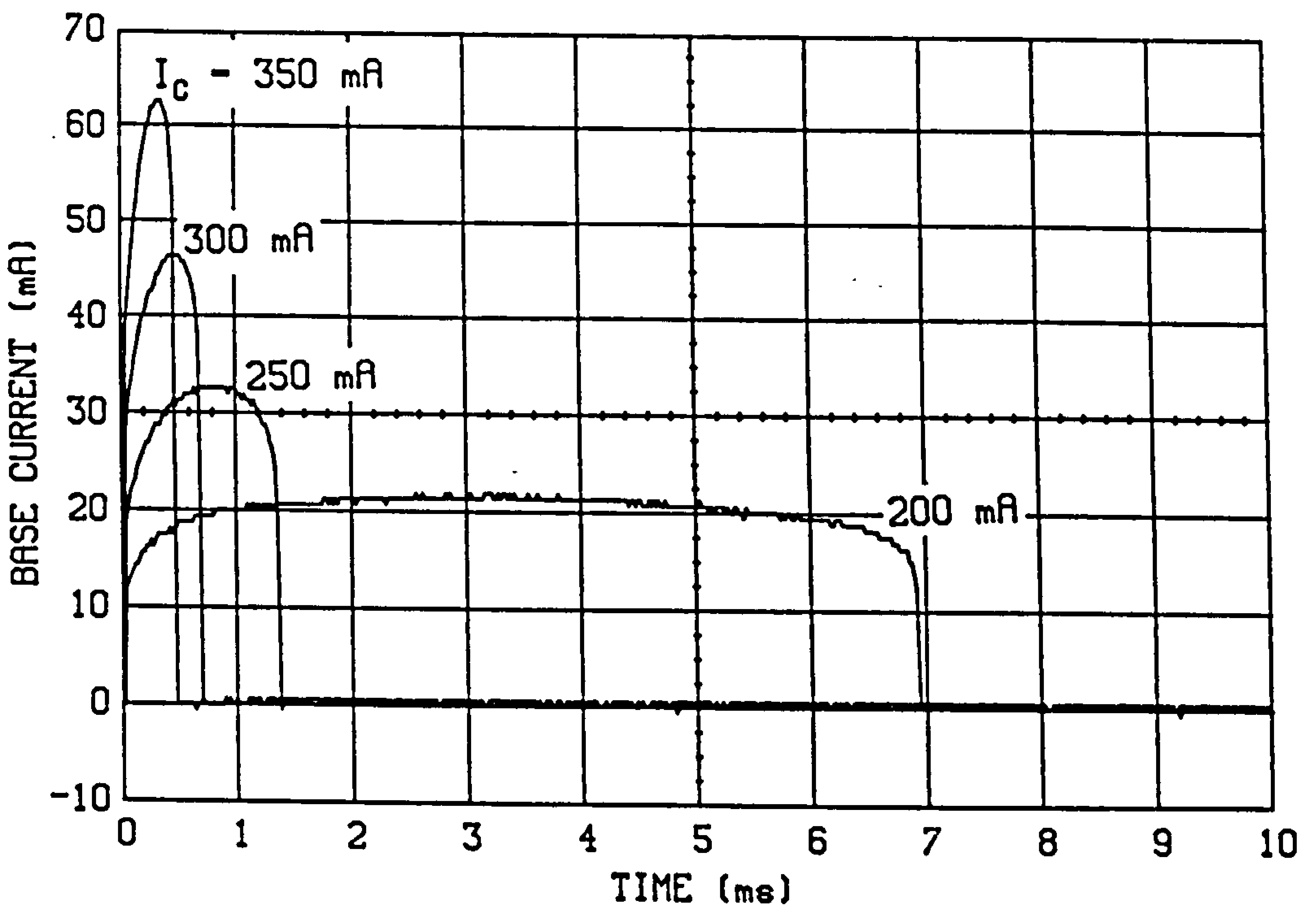


Figure 66. Base Current Waveforms During the Delay Time to Breakdown: All results were taken from Device C and the collector voltage is 100V.

(mobile space charge region) whence they are swept into the base. The equilibrium concentration of holes in the substrate increases with temperature as n_i increases, and the hole concentration gradient becomes steeper resulting in a greater diffusion component. Therefore, in order to maintain a constant collector current the externally supplied base current requirement must be correspondingly reduced.

Close inspection of the base current waveforms for the higher collector currents reveals an abrupt discontinuity, with a rapid fall in current after a period of slowly falling base current. This indicates that avalanche multiplication has been induced, resulting in extremely rapid carrier generation over and above thermal generation. Immediately this occurs operation is triggered into second breakdown. This effect is expected to be more prominent at higher currents since the initial peak field prior to any self-heating will be higher at higher currents. Thus, only a small increment in the field due to self-heating would be required to attain be avalanche breakdown condition.

At collector currents greater than $400mA$ (upper graph) the triggering time can be seen to decrease linearly with input power. The triggering time depends on the time it takes the self-heating effects to give enough field enhancement to initiate avalanche breakdown. However, at lower currents (lower graph), the triggering time decreases much more rapidly with the input power. At these currents the peak field throughout much of the duration of the delay time is expected to stay well below the value required for significant multiplication and breakdown will be dominated by thermal generation. The heat flow reaches the chip-header interface in approximately $200\mu s$ and the bottom of the header in approximately $4ms$. Operation was found to be stable for currents below $200mA$. In this situation the steady state chip temperature attained is well below that required for TSB. The steady state condition arises when the peak chip temperature reaches a value such that all the power dissipated as heat is removed from the chip and header via the heat transfer processes of conduction, convection and radiation.

A second interesting result is shown in Figure 67, which shows the separate currents flowing in both emitters of device B ($BV_{CEO} = 250V$, $W_{epi} = 36\mu m$, $N_{epi} = 1 \times 10^{14} cm^{-3}$) for a collector current of $350mA$ and a collector voltage of $100V$. It is clearly seen that the current redistributes itself between the emitters as the device heats up, with the centre emitter conducting a greater portion of the total current as time advances. Current redistribution towards the centre emitter occurs despite the fact that initially a greater portion of the current flows through the outer emitter. This observation is in good agreement with the numerical results, which are described in the next section.

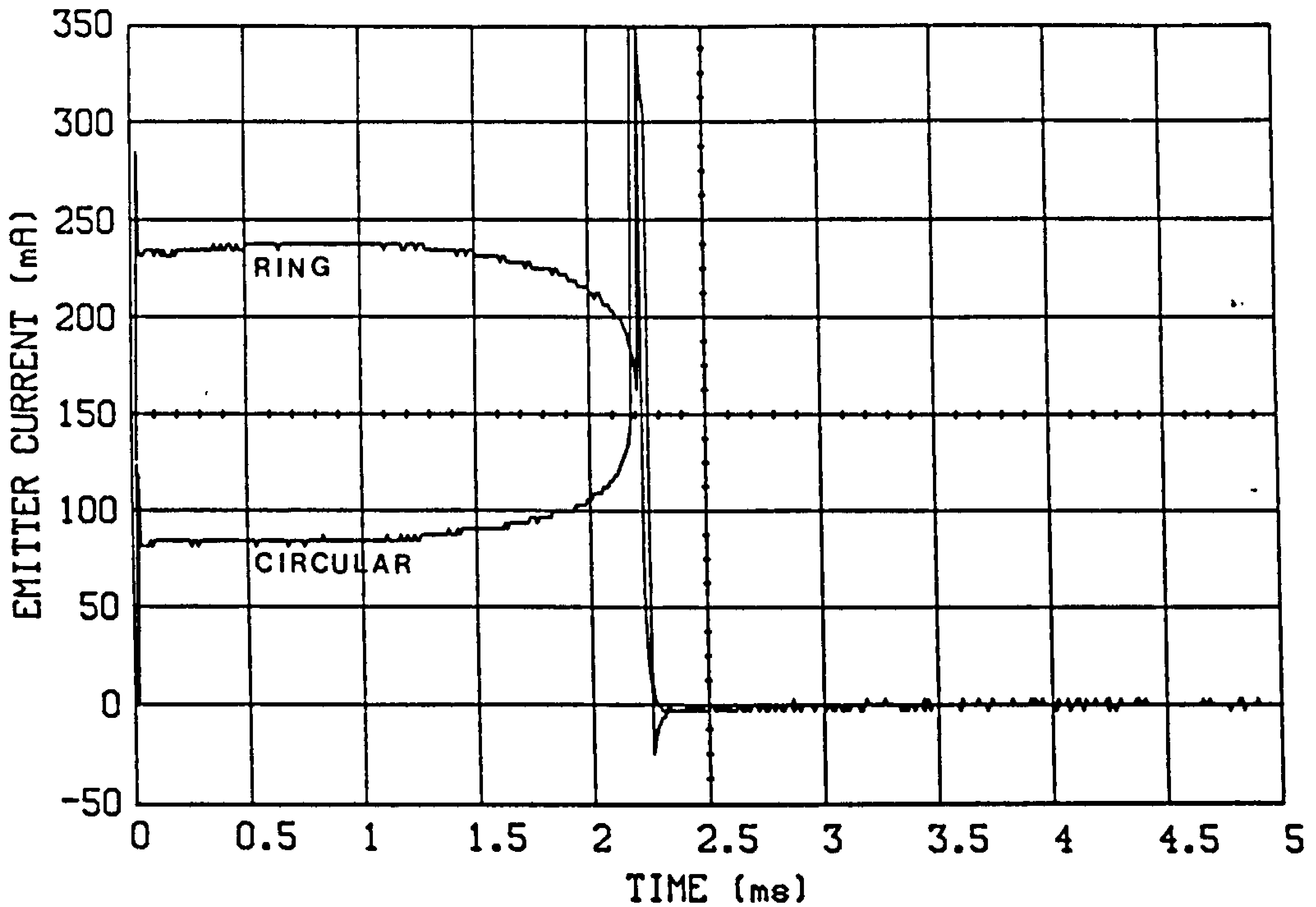


Figure 67. Emitter Current Waveforms During the Delay Time to Breakdown: The collector current is 300mA.

Finally a brief investigation into the potential benefit to be gained from the use of graded collector impurity profiles was carried out. Figure 68 shows the base waveforms for device C geometry with the 50μm and 100μm thick graded collector layers described in section 6.2.2. The collector current was again fixed at 350mA and the collector voltage at 100V. Although these collector layers together with the uniformly doped layer have been designed to give a BV_{CEO} of 250V the triggering time of the 50μm layer is ten times longer than that for the uniformly doped collector, and the triggering time of the 100μm layer is nearly twice as long again. The effect of the grading is to prevent the field from rising as the saturation velocity decreases and the base widens with increasing temperature. The field profile gradually transfers from the base-collector junction to the collector-substrate junction as the saturation velocity falls. The peak field starts to rise only when the field profile reaches the collector-substrate junction. Hence, there is a delay before the peak field starts to rise and furthermore the peak field should not increase as rapidly as for the uniform collector, because the collector doping near the collector-substrate junction is higher than for the uniformly doped collector (cf. section 6.2.2). Hence, the localised heating in the peak field region is less severe for the graded collector. A more detailed account

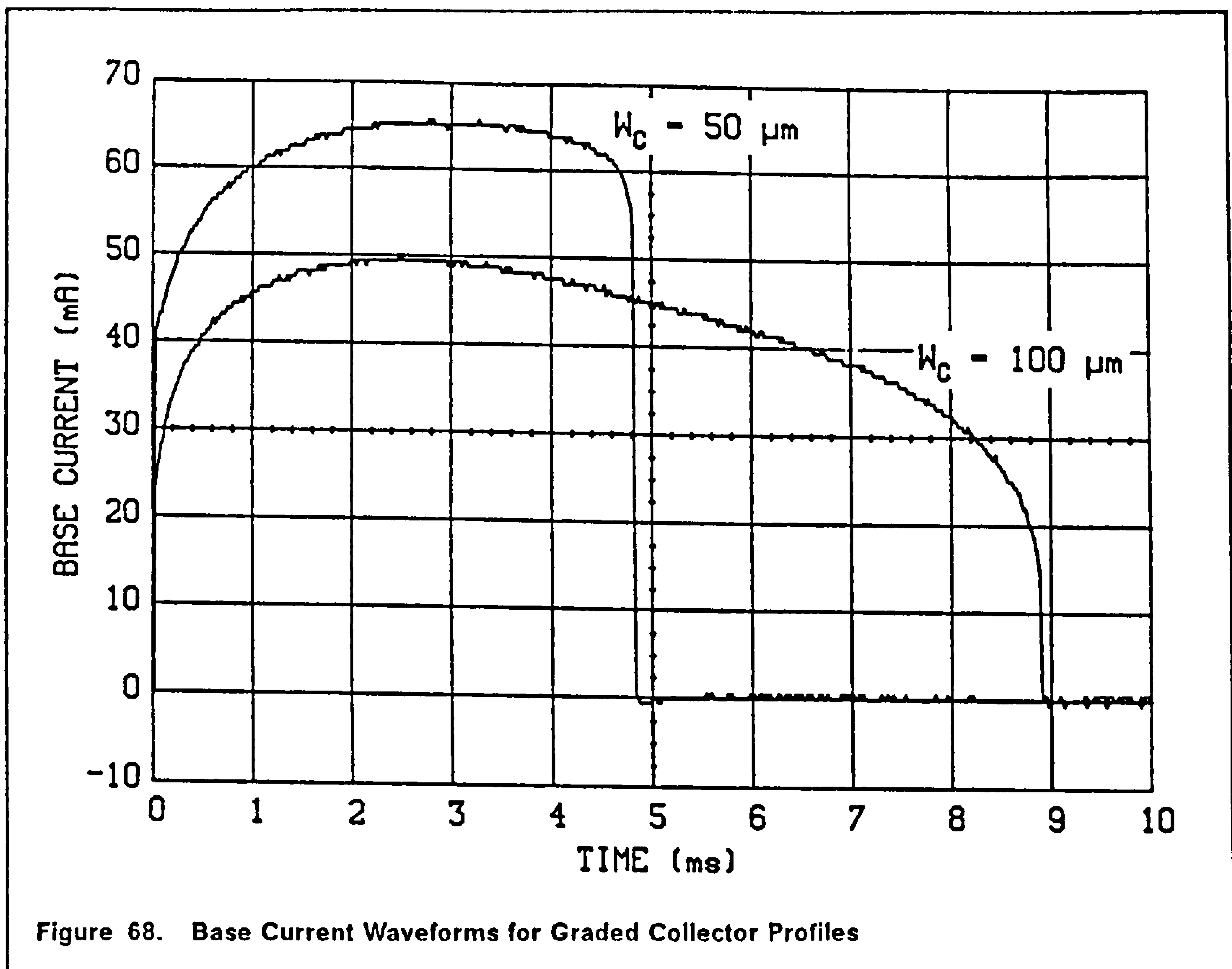


Figure 68. Base Current Waveforms for Graded Collector Profiles

of these phenomena will be given in the following section with the aid of a number of numerical results.

6.5.4 Numerical Results.

The full numerical model for electro-thermal interaction has been utilized in this analysis. The energy dissipated in the bipolar transistor during switch-on is insignificant in comparison with that which is dissipated when the transistor is in the on-state. Thus, the time taken to charge the junction and diffusion capacitances is very short in comparison with the time it takes for significant heating. By making use of this fact the transient analysis could be initialized with ψ, n and p distributions taken from a steady state analysis at the required collector current and voltage with the temperature fixed at ambient 300K. This eliminates a great deal of unnecessary computation and will not affect the accuracy of the results in the time ranges being considered. The collector current is subsequently maintained at a constant value by adjusting the base emitter voltage in an equivalent manner to the experimental test circuit. At each point in time an initial guess is made for v_{be} and the numerical model is solved for a self-consistent set

of ψ, n, p and T. If the resulting collector current is not sufficiently close to the required value then v_{be} is altered by assuming I_C to be of the following form.

$$I_C = I_{CO} \exp\left(\frac{q\, v_{be}}{m\, k\, T} \right) \tag{6.52}$$

Taking the derivative of I_C with respect to v_{be} and approximating the derivatives by differences gives the following estimate for the required base voltage adjustment.

$$\Delta v_{be} = \frac{m\, k\, T}{q} \; \frac{I_{CR} - I_C}{I_C} \tag{6.53}$$

Here, I_{CR} is the required collector current, the dimensionless factor m can be obtained from the steady state characteristics calculated for the initialization and the temperature, T was taken to be the average temperature along the emitter-base junction. This procedure is repeated at every instant of time until the calculated current lies within the required tolerance limits. As an example it takes typically three such iterations for a collector current of $350mA$ with an allowable tolerance of $\pm 2mA$.

The results of such an analysis for device C with uniform collector doping are shown in Figure 69. Once again the device attributes are taken to be the 'old values' in Table 7, but numerical results are also shown for the situation where the base lifetime is equal to the emitter lifetime and also for a ring emitter structure, which as before was formed by removing a centre portion of the emitter with half the radius of the full emitter. The equivalent experimentally obtained waveform is also plotted for comparison. Notice that because avalanche generation has not been included in the model the numerically obtained waveforms decay smoothly to zero without any discontinuities.

By reducing the base lifetime the triggering time has been increased by 50%. This is primarily due to the greater base current requirement which allows for more thermal carrier generation prior to TSB. The improvement to be gained from the use of ring emitters arises because this structure is less susceptible to the formation of localized regions of high current density.

A number of temperature profiles at various instants of time for the full emitter structure ($\tau_b = 2\mu s$) are given in Figure 70 and Figure 71. It may be seen that the current density is high enough such that the peak power dissipation and, therefore, the peak temperature is located close to the collector-substrate junction. The full chip thickness is shown so that the lower boundary of these contour diagrams coincides with the chip-header interface. After $5\mu s$ the peak temperature has risen to $360K$ and is located beneath the outer perimeter of the emitter,

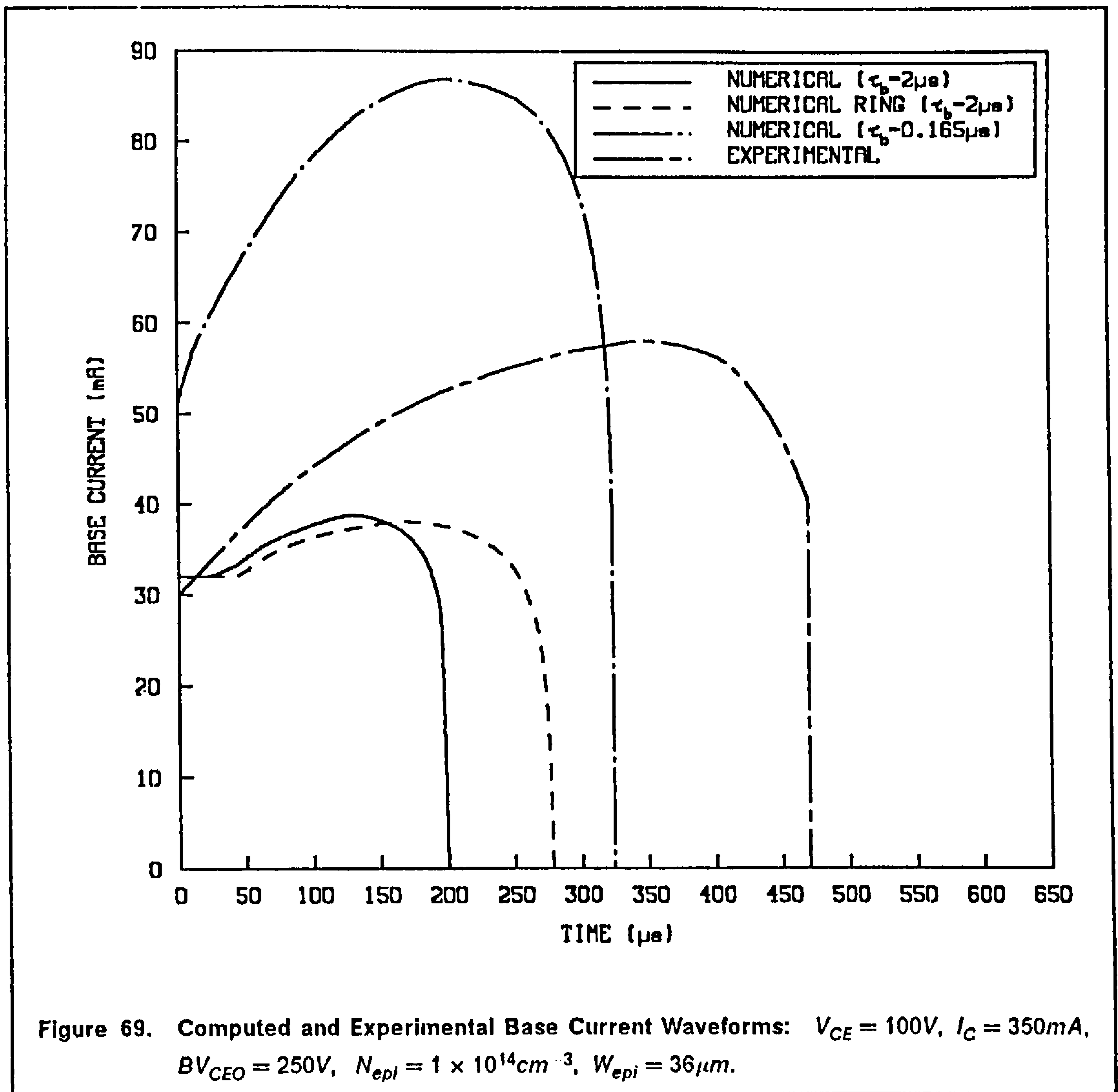
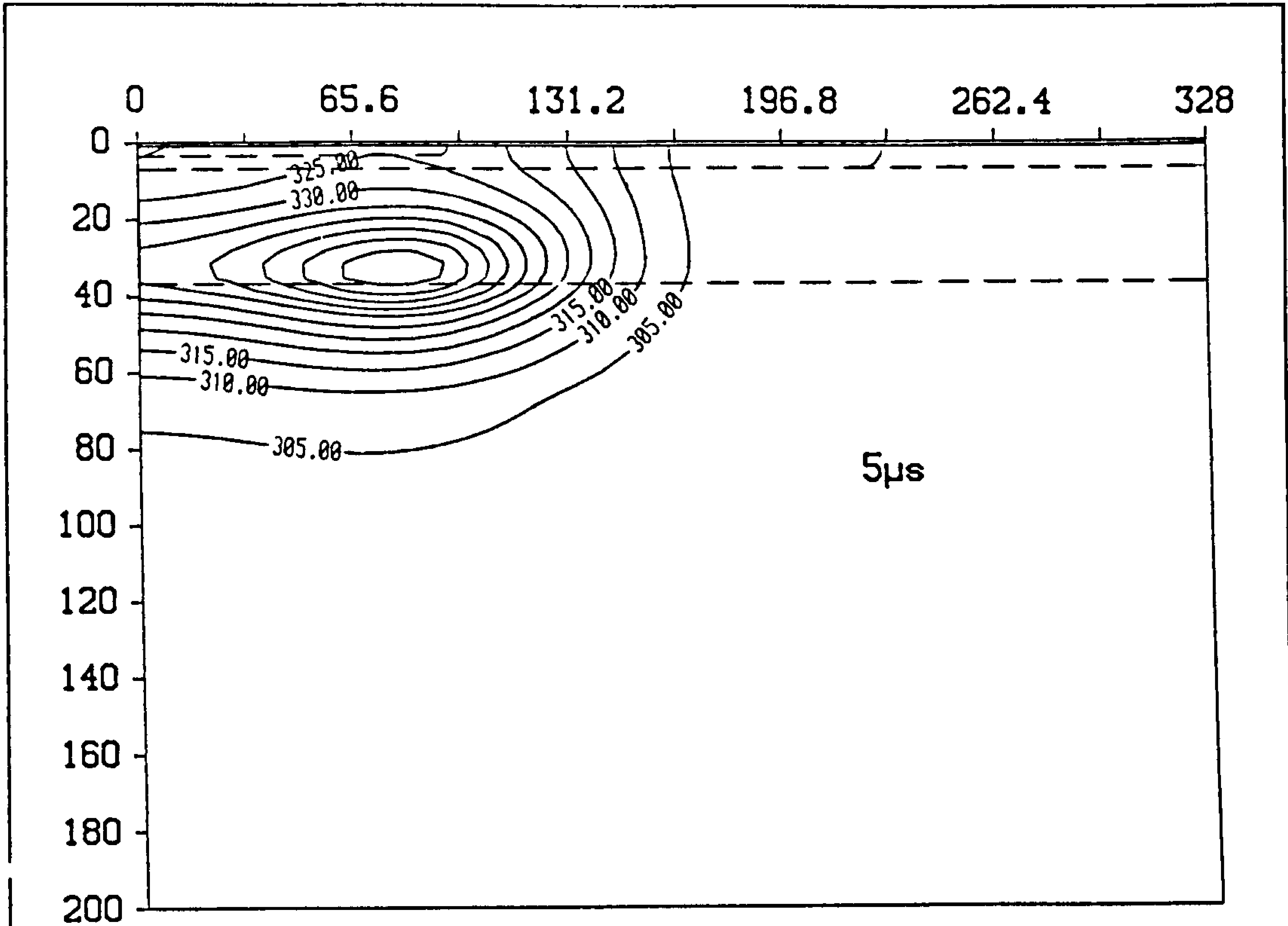


Figure 69. Computed and Experimental Base Current Waveforms: $V_{CE} = 100V$, $I_C = 350mA$, $BV_{CEO} = 250V$, $N_{epi} = 1 \times 10^{14} cm^{-3}$, $W_{epi} = 36\mu m$.

because of emitter pinching. However, the effects of emitter pinch are alleviated to some extent by current divergence in the collector layer. For example at this point the ratio of the current density at the collector-substrate junction underneath the emitter perimeter to that under the emitter centre is only 1.75.

By $70\mu s$ the peak temperature has risen to $650K$ and its location has shifted slightly towards the centre. The current density profile remains relatively unchanged and the point of peak temperature does not, therefore, coincide with the position of the peak heat generation which remains under the emitter edge. Since the temperature at the emitter centre is 50°C higher than at its edge, the current might be expected to redistribute towards the centre. However, the mobility in the base falls with increasing temperature and the corresponding increase in the lateral base resistance was found to balance out this effect leaving the current density profile relatively unchanged.



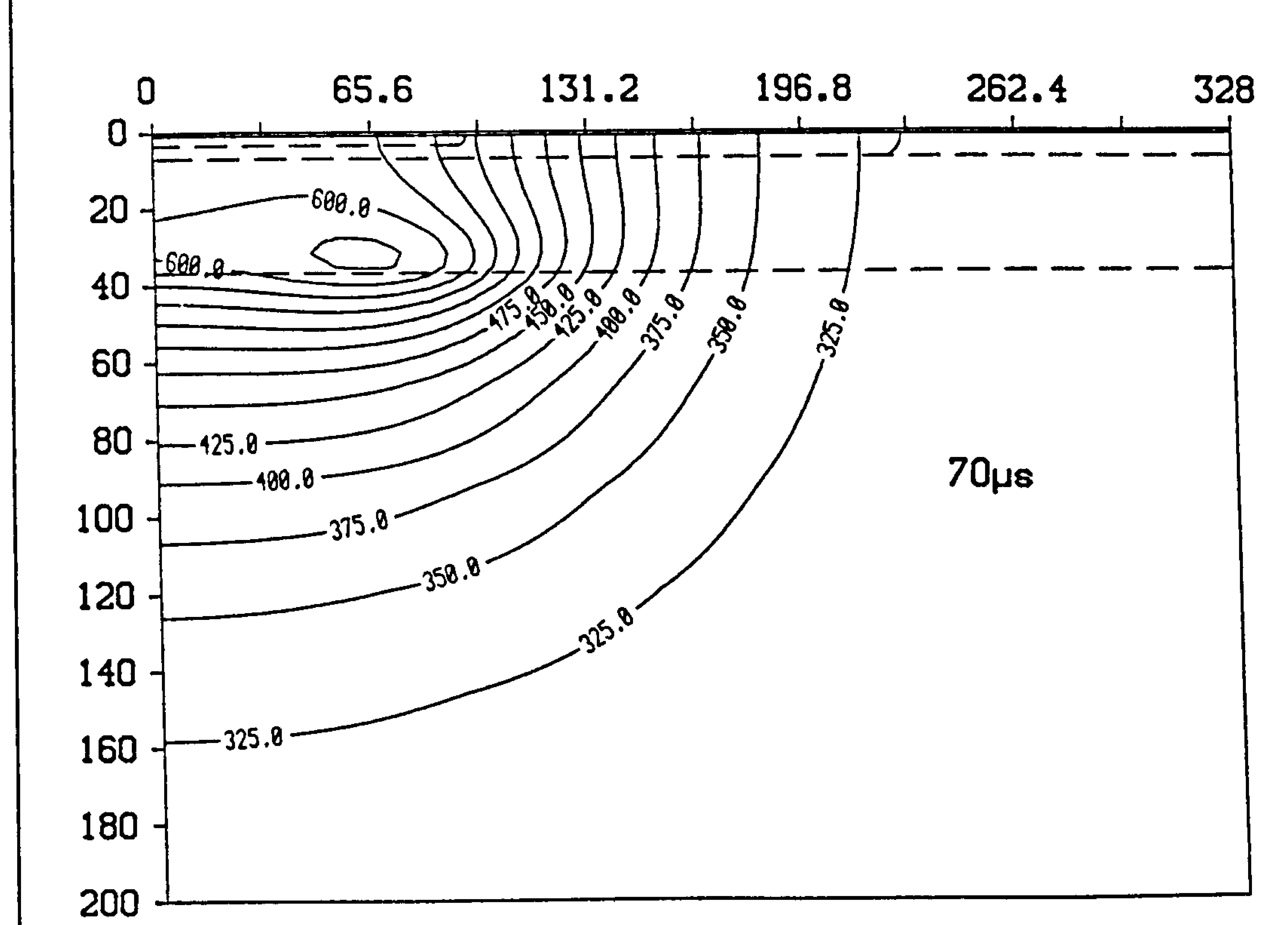
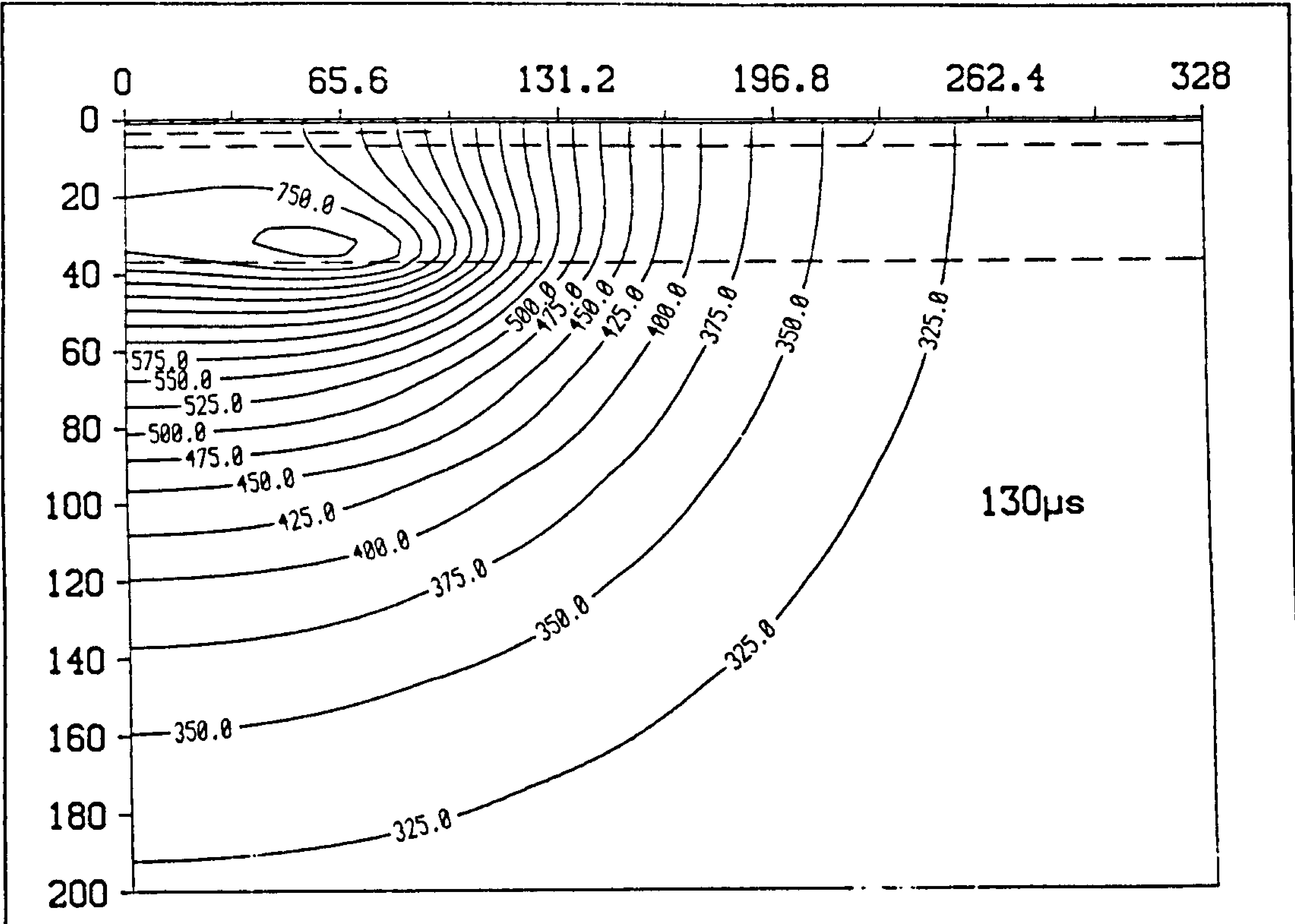


Figure 70. Temperature Profiles at Various Times Before Breakdown.



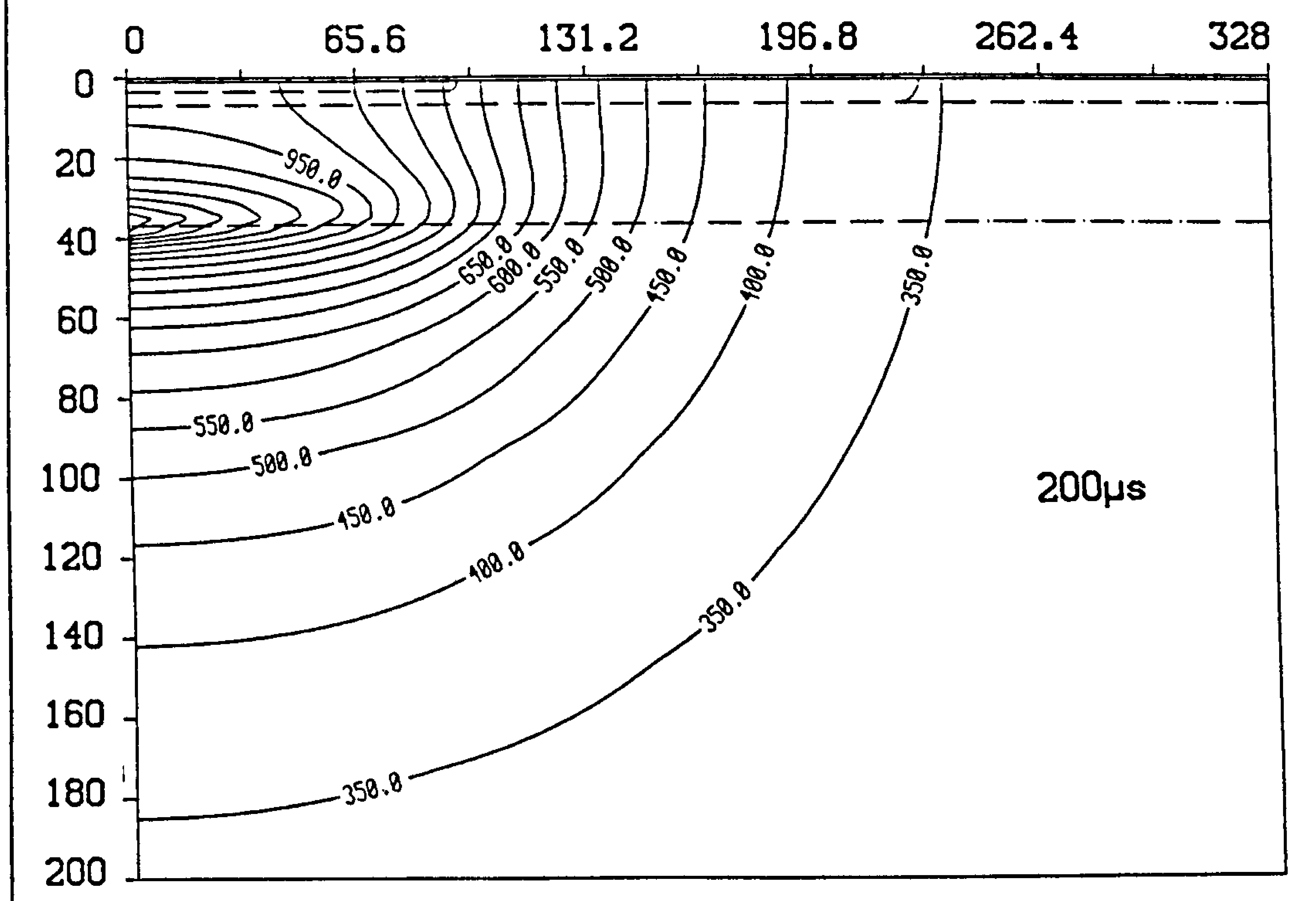


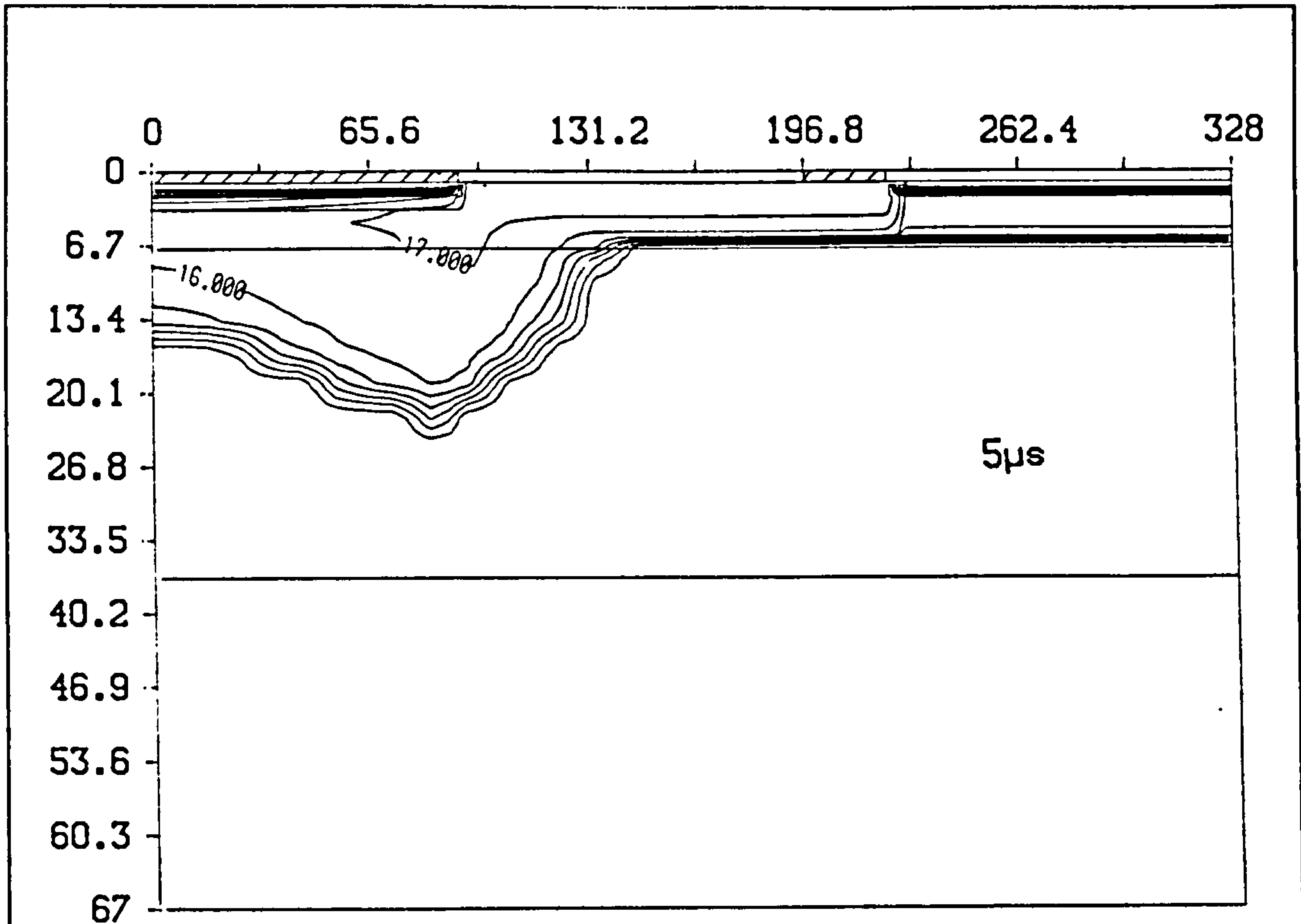
Figure 71. Temperature Profiles at Various Times Before Breakdown.

After $130\mu s$ the peak temperature is $800K$ and has again moved slightly nearer the centre. Sufficient time has now elapsed such that the heat flow wavefront has reached the silicon-header interface. Since the thermal conductivity of mild steel is about seven times less than that of silicon then the temperature can now be expected to rise at a slightly faster rate. In this case, however, the temperature build-up is so rapid that very little time is allowed for conduction of heat away from the active region and conditions are almost adiabatic. Consequently the effects of the header should be negligible in this case.

At $200\mu s$ the terminal base current has dropped to zero and the temperature profile has changed significantly. A mesoplasma has begun to form and the entire device current is shorted through it. The peak temperature in the mesoplasma is $1300K$, which is still well below the melting point of silicon ($1700K$). However, the peak temperature of $940K$ at the emitter contact is well above the minimum eutectic point of the aluminium-silicon system [6.33], which has a value of $850K$ for a system with 10% silicon and 90% aluminium (atomic percentages). Thus, melting can occur even prior to any voltage collapse.

Figure 72 and Figure 73 show the variation of hole concentration with time. The effects of emitter pinch on the width of the current induced base are clearly visible at $5\mu s$. By $70\mu s$ the current induced base has widened even further. This is entirely due to the fall in the limiting drift velocity of electrons with temperature as predicted by equation (3.18). The subsequent increase in the concentration of electrons which constitute the mobile space charge region will result in a thinner high field region given by (6.45) and, therefore, a wider current induced base. This causes an increase in the charge stored in the base and a corresponding reduction in emitter efficiency. Also, since the base is wider, more base recombination can be expected giving a reduction in the base transport factor. The resulting h_{FE} reduction is offset by an increase in base current in order to maintain a constant collector current.

After $130\mu s$ the holes that are thermally generated in the substrate and space charge regions become apparent. Those that are generated in the space charge region are swept into the base under the influence of the high electric field. Those that are generated in the substrate within about two diffusion lengths of the high field region come under the influence of the hole concentration gradient and diffuse into the high field region. From Figure 71 the temperature over the space charge region is reasonably uniform with a value of $760K$, which gives an $n_i(T)$ of $6 \times 10^{16}\,cm^{-3}$. From Figure 73 the hole concentration in the space charge region can be seen to be much smaller than this. It has also been found that the electron concentration in the space charge region is much less than n_i. Thus, the space charge generated base current component can be calculated by approximating n



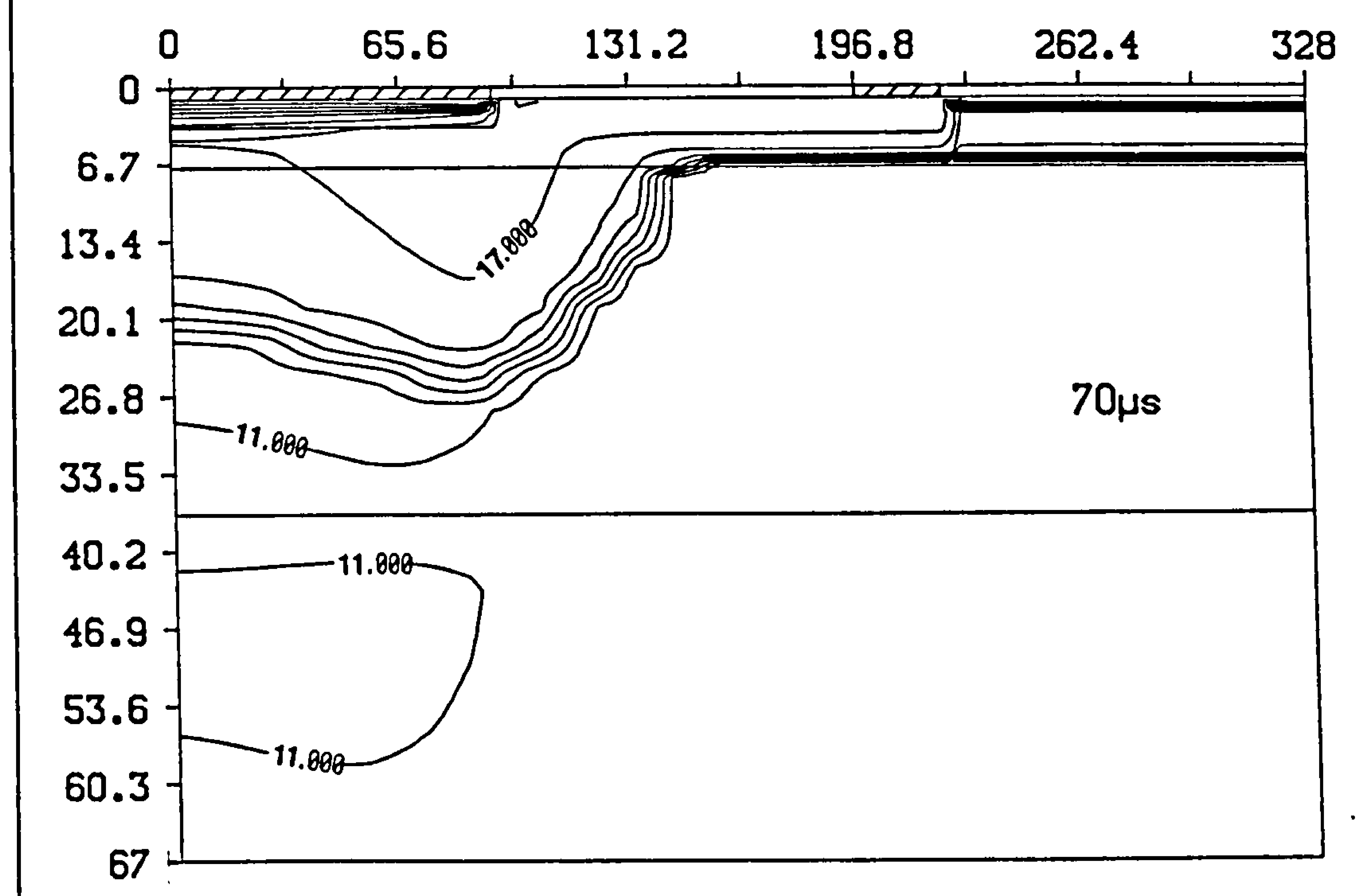


Figure 72. Contour Maps of Hole Concentration at Various Times Before Breakdown: The contours are labelled with $\log_{10}$ values of concentrations in units of cm^{-3}.

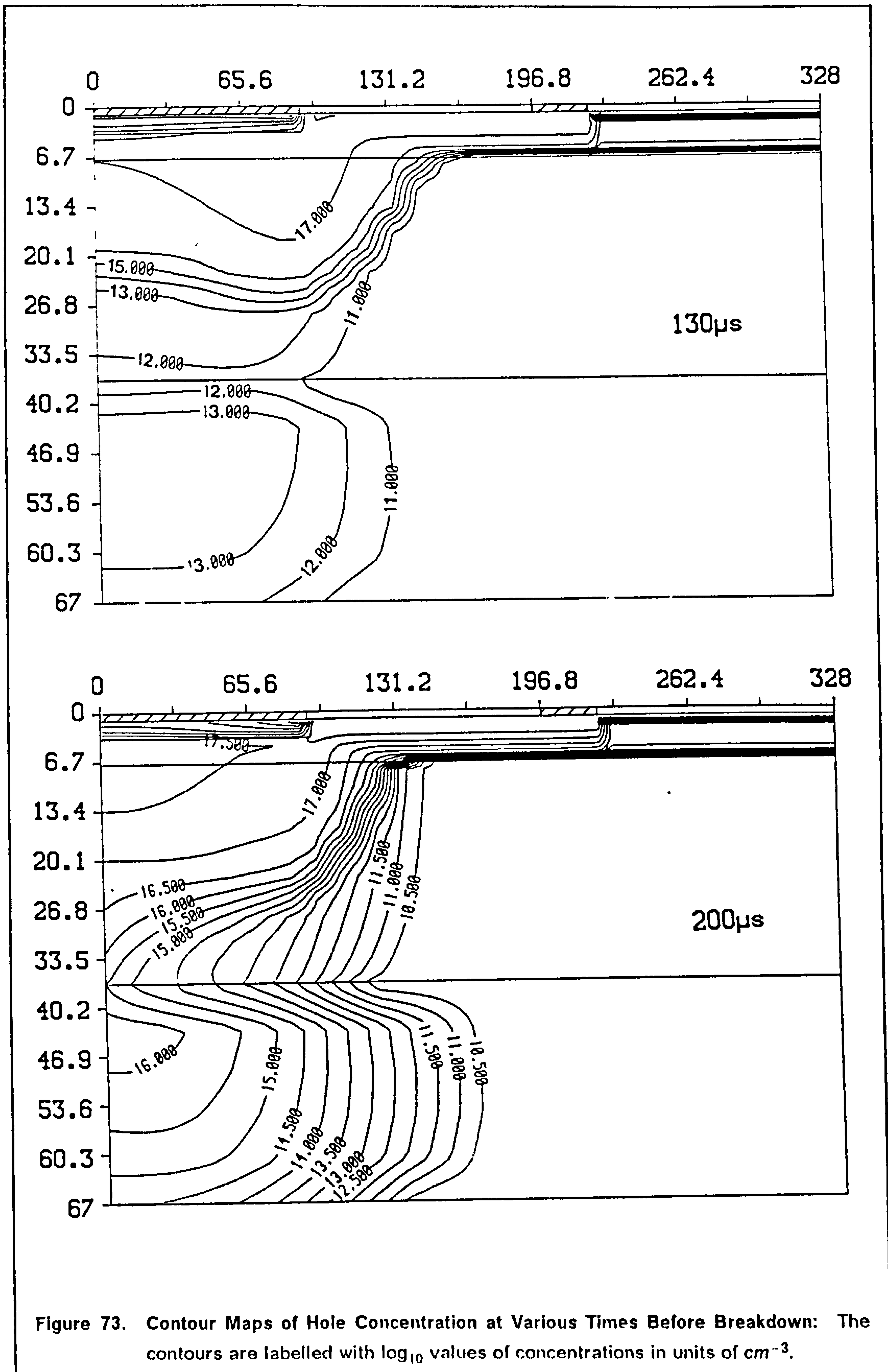


Figure 73. Contour Maps of Hole Concentration at Various Times Before Breakdown: The contours are labelled with $\log_{10}$ values of concentrations in units of cm^{-3}.

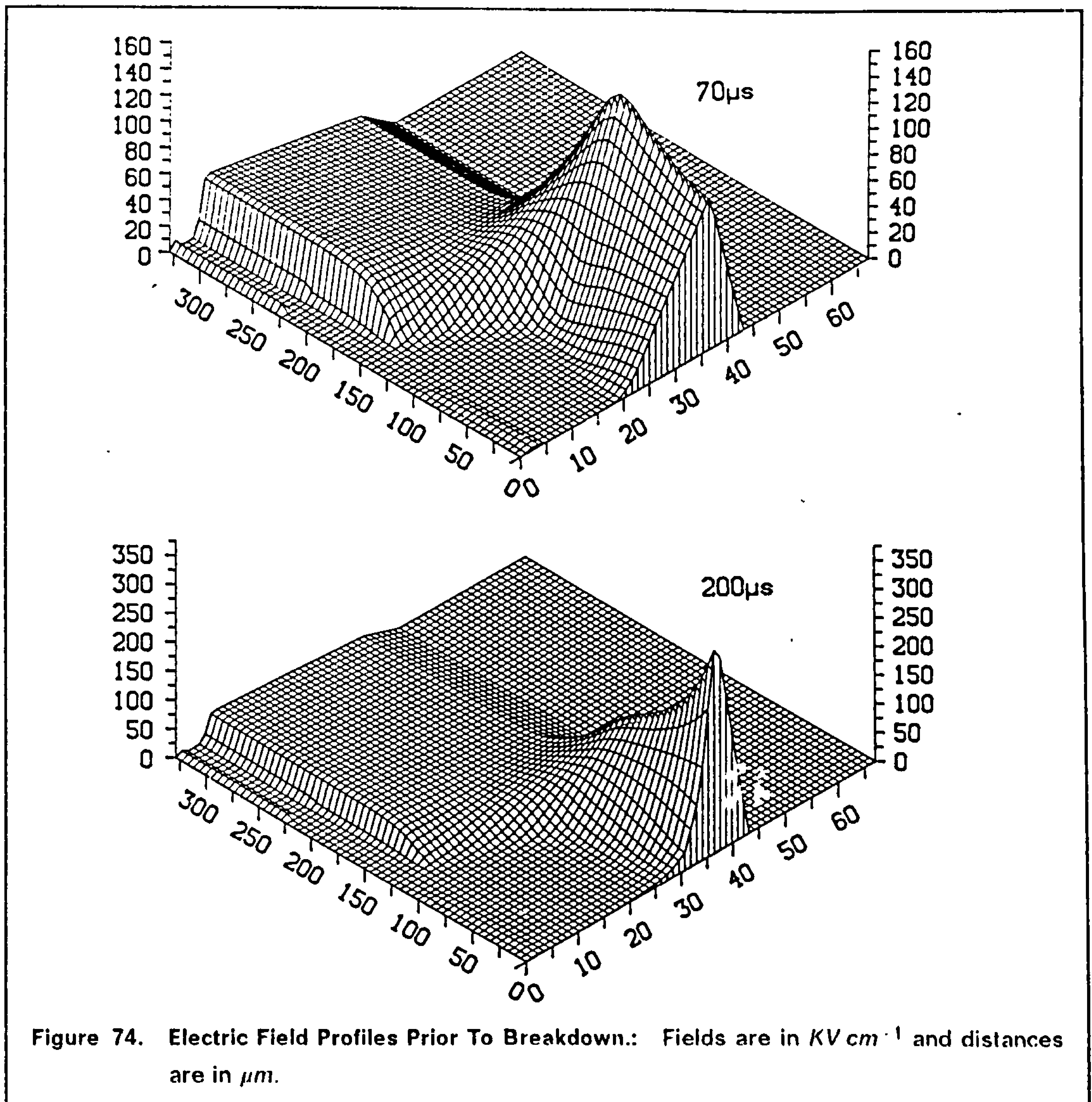


Figure 74. Electric Field Profiles Prior To Breakdown.: Fields are in $KV cm^{-1}$ and distances are in μm.

and p in equation (3.30) to zero. The total generated base current is the sum of that generated in the space charge region and that diffusing from the substrate and is given by:

$$I_{B_{gen}} = I_{gen} + I_{diff} = \frac{q\,A\,W\,n_i(T)}{\tau_{n0} + \tau_{p0}} + q\,A\,\frac{L_p(T)}{\tau_{p0}}\,\frac{n_i(T)^2}{N_{DS}} \tag{6.54}$$

where W is the width of the space charge region, A is its area, which is roughly equal to the emitter area, N_{DS} is the substrate donor concentration and $L_p(T)$ is the hole diffusion length in the substrate ($= \sqrt{D_p(T)\,\tau_{p0}}$). At $130\mu s$ for $\tau_{p0} = \tau_{n0} = 2\mu s$ the space charge generation component is roughly $0.38mA$ and the diffusion component is roughly $0.23mA$. Both these components will grow with time to account for the levelling off and eventual rapid decline of the base current waveform.

By $200\mu s$ the collector current is self-sustaining and all control over device operation is lost. The downward curvature of the hole concentration contours at the centre of the device in Figure 73 indicate the initial stages of mesoplasma formation, and this represents the picture just prior to the voltage collapse.

The electric field profiles at $70\mu s$ and $200\mu s$ are illustrated in Figure 74. It may be seen that like the peak temperature the peak field shifts from beneath the emitter periphery to beneath the emitter centre. The vertical field profile along the emitter centre line also changes from one with a constant gradient as predicted by the previously outlined one dimensional theory to one with an increasing gradient. This occurs because of the alteration in the space charge density caused by the thermally generated electrons and holes. As a result, the negative space charge density increases from the edge of the current induced base to the collector-substrate junction and the field profile changes accordingly. In both cases the peak field is well above the previously accepted value for the critical field for breakdown of $100\,KV\,cm^{-1}$. It is quite possible, therefore, that avalanche breakdown could have been induced prior to thermal breakdown in practice.

The time variation of the peak field and temperature for the full emitter and ring emitter structures is shown in Figure 75. A number of workers [6.35], [6.36] have suggested that TSB occurs when the peak temperature reaches the intrinsic temperature of the collector layer. That is, the temperature at which the intrinsic concentration becomes equal to the ionized impurity concentration of the collector. However, this suggests a value of $500K$ for the $1 \times 10^{14}\,cm^{-3}$ collector considered here, which is well below the calculated value of $1360K$. However, better agreement has been obtained with a number of other workers [6.31], [6.37] who have calculated the temperature to be at or near the melting point of silicon ($1700K$). More specifically though, the results described here seem to disagree with the concept of a triggering temperature. Rather, the results indicate that entry into TSB will occur as soon as the point is reached where the entire base current requirment is supplied by thermally generated holes. Any further generation will result in a base current that exceeds the value required to maintain a constant collector current. The collector current is then forced to rise and rapid mesoplasma formation will ensue.

Current filamentation is prevented in the case of the ring emitter structure and instead the current redistributes towards the inner edge of the emitter forming a tube of high current density. Figure 76 shows that the effects of emitter pinch are reduced with time as the thermally generated base current component, which does not generate any significant lateral voltage drop, becomes more significant. The effect of removing the centre of the emitter is to spread the current around the inner emitter edge and so reduce the heat generation. However, this effect is

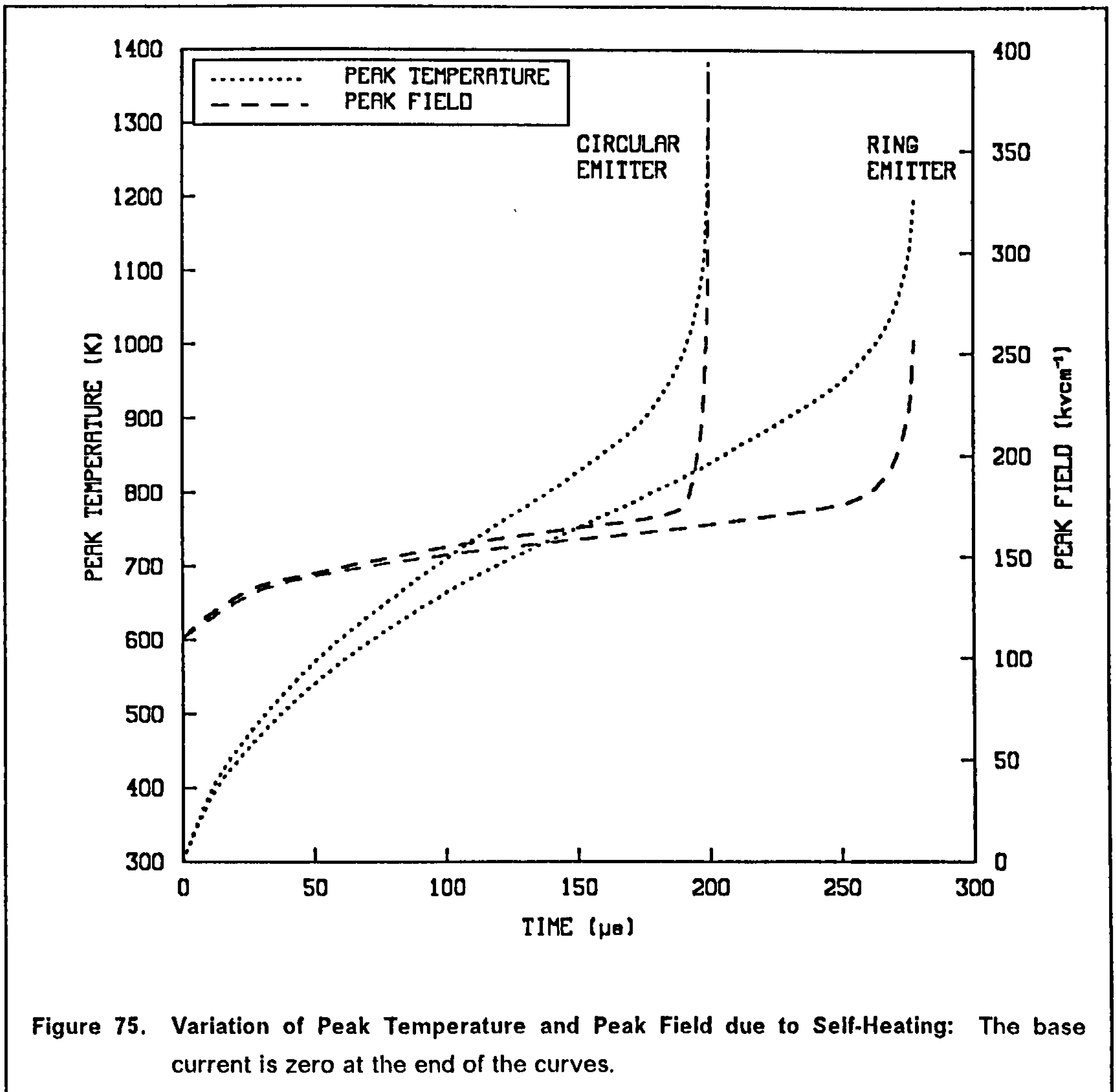


Figure 75. Variation of Peak Temperature and Peak Field due to Self-Heating: The base current is zero at the end of the curves.

somewhat offset by the additional consequence of spreading the heat generation over a larger volume. As may be expected the peak field and temperature accordingly moves from below the outer emitter edge to below its inner edge.

6.5.5 Comparison of Experimental and Numerical Results.

It is clear from Figure 69 that the computed results underestimate the triggering time to breakdown. The differences may be due to inaccuracies in the models used to represent the temperature dependent physical parameters or they could be due to deficiencies in the existing model. Although most of the important parameters which govern device operation have been accurately quantified at room temperature a considerable amount of uncertainty still exists as to their variation with temperature.

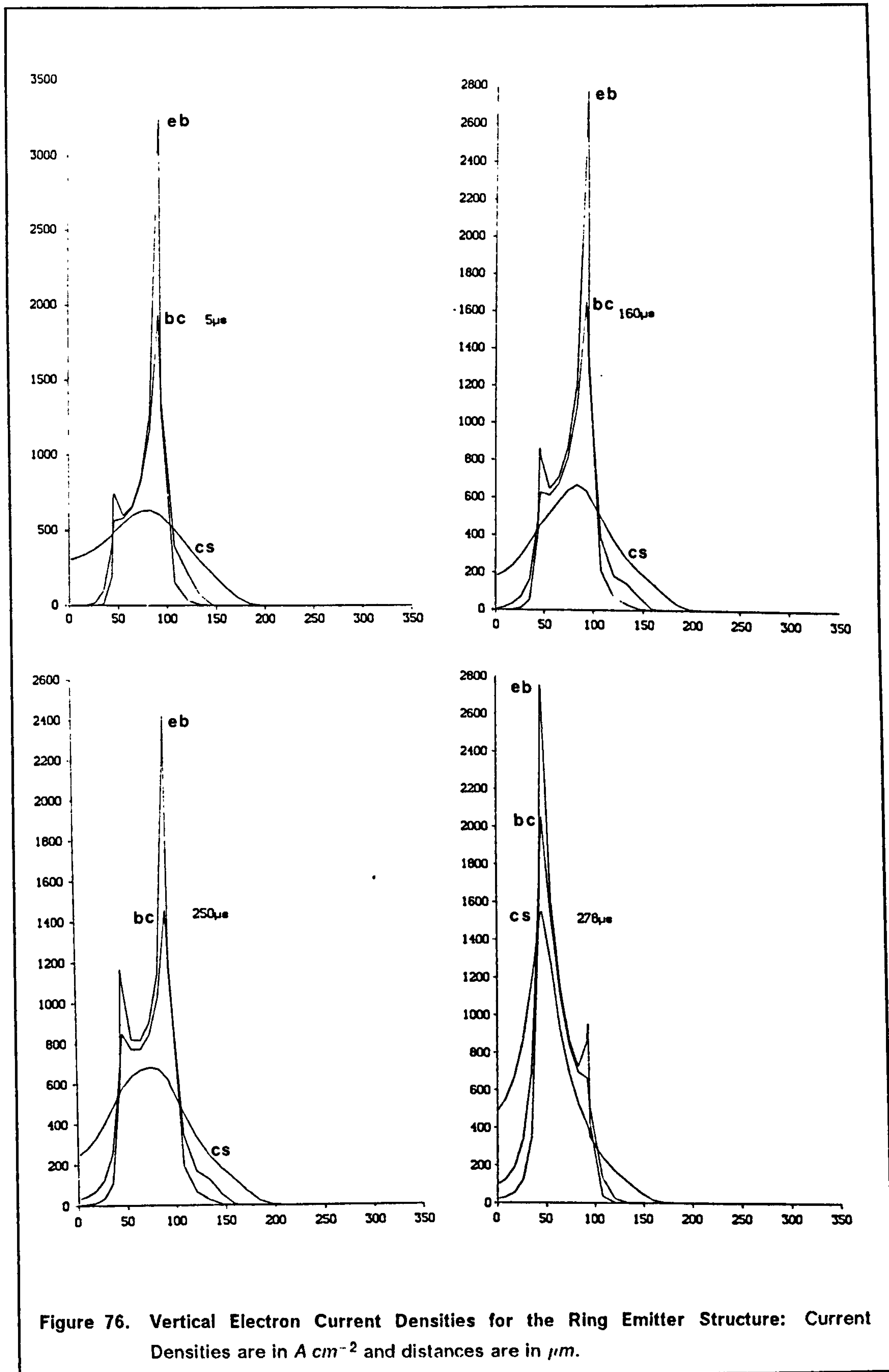


Figure 76. Vertical Electron Current Densities for the Ring Emitter Structure: Current Densities are in $A\ cm^{-2}$ and distances are in μm.

The computed base current waveform is seen not to rise as rapidly as the equivalent experimental waveform. Since this effect seems to be due entirely to the temperature dependence of the electron saturation velocity then the model used in the computations would seem to be at fault. The experimental waveform would seem to suggest that the velocity decreases at a faster rate with temperature than predicted by equation (3.18). A greater amount of experimental data than that given in Figure 9 on page 51 is required to justify this model.

The hole current component originating from the high field region is proportional to $n_i(T)$ and that originating from the substrate is proportional to $n_i(T)^2$, and n_i is, therefore, an extremely critical quantity. From Figure 10 on page 53 it is seen that the model may infact underestimate n_i at higher temperatures, if only by a small amount. However, the early measurements of Morin and Maita should be viewed with caution as they were obtained from rather impure material.

In $200\mu s$ the temperature profile only just reaches the header and so no heat will be lost from the system through the header or heat sink if attached. However, the heat does reach the top of the chip and the model assumes unforced convection to be the only heat transfer process by which heat is allowed to escape from the top surface of the chip. A considerable amount of heat can escape by means of conduction through the emitter contact metalization and subsequently through the bonding wire. In addition, at the higher surface temperatures arising just prior to breakdown a significant amount of energy can escape by radiative heat transfer. For example the Stefan-Boltzmann law [6.38] predicts that approximately six times more energy is lost via radiation compared with unforced convection at a surface temperature of $900K$.

A more important difference between the model and the devices seems to be the collector SRH lifetimes. From Table 7 on page 169 the measured lifetime is $30\mu s$ and the value used in the model is $2\mu s$. Since the current generated in the high field region varies as $1/\tau_{p0}$ and that generated in the substrate as $1/\sqrt{\tau_{p0}}$ the overall predicted thermal generation will be about an order of magnitude greater at a particular temperature than it should be. This may well account for the observed differences.

It is equally probable that the differences are due to the existence in practice of finite resistances in series with the emitter, which act as efficient stabilization against thermal runaway [6.39]. The most important contribution to this resistance comes from the emitter bond wire and contact metallisation. The devices were bonded with $25\mu m$ diameter gold wire which was found to be capable of carrying at least $2A$ of steady-state current. Assuming the bond wire to be $0.3cm$ long then at 900°C the resistivity of gold is $11.8\,\mu\Omega cm$ and the corresponding resistance of the bond wire is 0.7Ω. For an emitter current of

400mA the voltage drop across this resistance is 0.3V, which still leaves 99.7V across the device. Therefore, it is unlikely that this would account for any of the differences. However, as previously pointed out the aluminium-silicon eutectic and also the aluminium itself (melting point 660°C) can melt well before any voltage collapse. Although the resistivity of aluminium, even in its molten state would not account for a significant resistance, the electrical properties of such a molten contact will be rather uncertain. For instance the existence of voids in the aluminium would serve to reduce the area of current flow and so increase the series resistance.

Evidence that the aluminium melts before TSB was obtained by observing the emitter metallisation through an optical microscope before and after thermal testing. The aluminium forming the emitter contact was found to have a slightly lighter colouration after testing, whereas the base metallisation remained unaffected. Despite this most of the tested devices remained fully operational, with their electrical characteristics being unaffected. In some extreme cases the gold bond wire was found to fuse, indicating that the temperature had risen above the melting point of gold which is 1064°C.

In summary the emitter resistances together with the different collector lifetimes probably account for the differences in triggering time. From the above analysis it is more likely that it is the difference in collector lifetimes that is responsible for the observed discrepancy. The model should be re-run with a collector lifetime of 30μs to clarify this matter.

6.5.6 The Consequences of Graded Collector Profiles on Thermal Breakdown.

Two computed base current waveforms for type C geometry with 50μm thick graded collector layers are illustrated in Figure 77. As before the collector voltage is 100V and the collector current is 350mA. The collector layer specifications were chosen according to the curves provided in [6.20], as described in section 6.2.2. The corresponding experimental result is provided for comparison. Numerical results have been calculated not only for the intended optimum profile, but also for the experimentally obtained profile, which was measured using the spreading resistance technique [6.9]. It was found that the measured profile was linear but the slope, a and the initial doping, N_{D0} in equation (6.33) were both higher than intended, having values of $2.2 \times 10^{17}\ cm^{-4}$ and $3 \times 10^{14}\ cm^{-3}$ respectively.

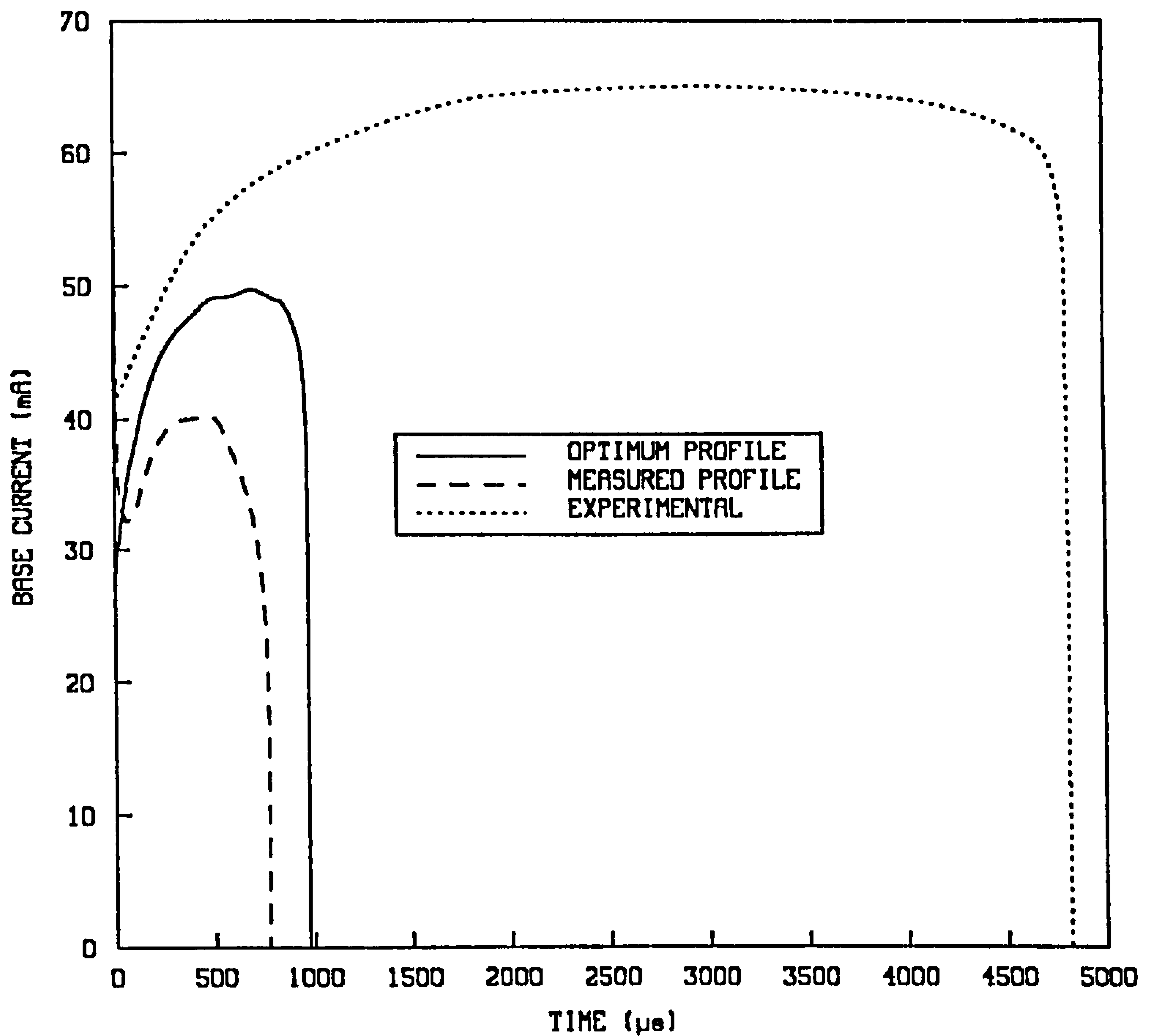


Figure 77. Base Current Waveforms for Graded Collector Profile: $V_{CE} = 100V$, $I_C = 350mA$, $BV_{CEO} = 250V$, $N_{DO} = 5 \times 10^{13} cm^{-3}$, $a = 1.1 \times 10^{17} cm^{-4}$, $W_{epi} = 56\mu m$.

The consequences of the grading on the peak field and temperature are illustrated in Figure 78. These results have been calculated for the optimum $50\mu m$ profile and the corresponding results for the $37\mu m$ uniformly doped collector are given for comparison.

The field is clearly kept much lower by the grading. The slow but steady rise of the peak field indicates that the field profile is intially in 'contact' with the collector-substrate junction as otherwise its value would remain constant because of the grading. The effect of the grading is to impede the rapid development of the field spike and the associated highly localized temperature rise. As a result the triggering time has been increased by five times that of the conventional device. It may also be seen that the field is kept well below the critical field for avalanche injection over almost the entire delay time and it is, therefore, unlikely that the triggering time will be significantly shortened due to these effects.

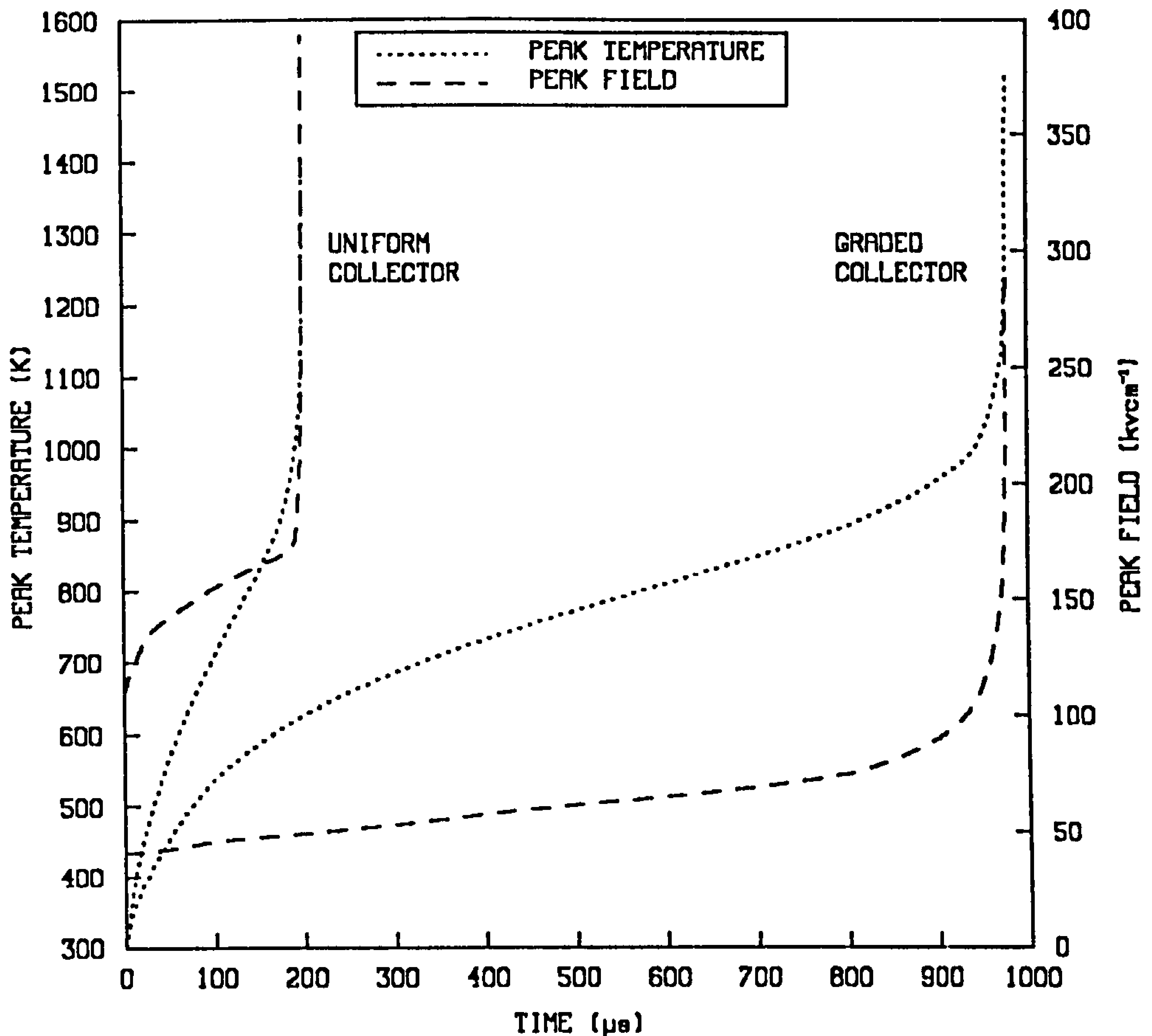
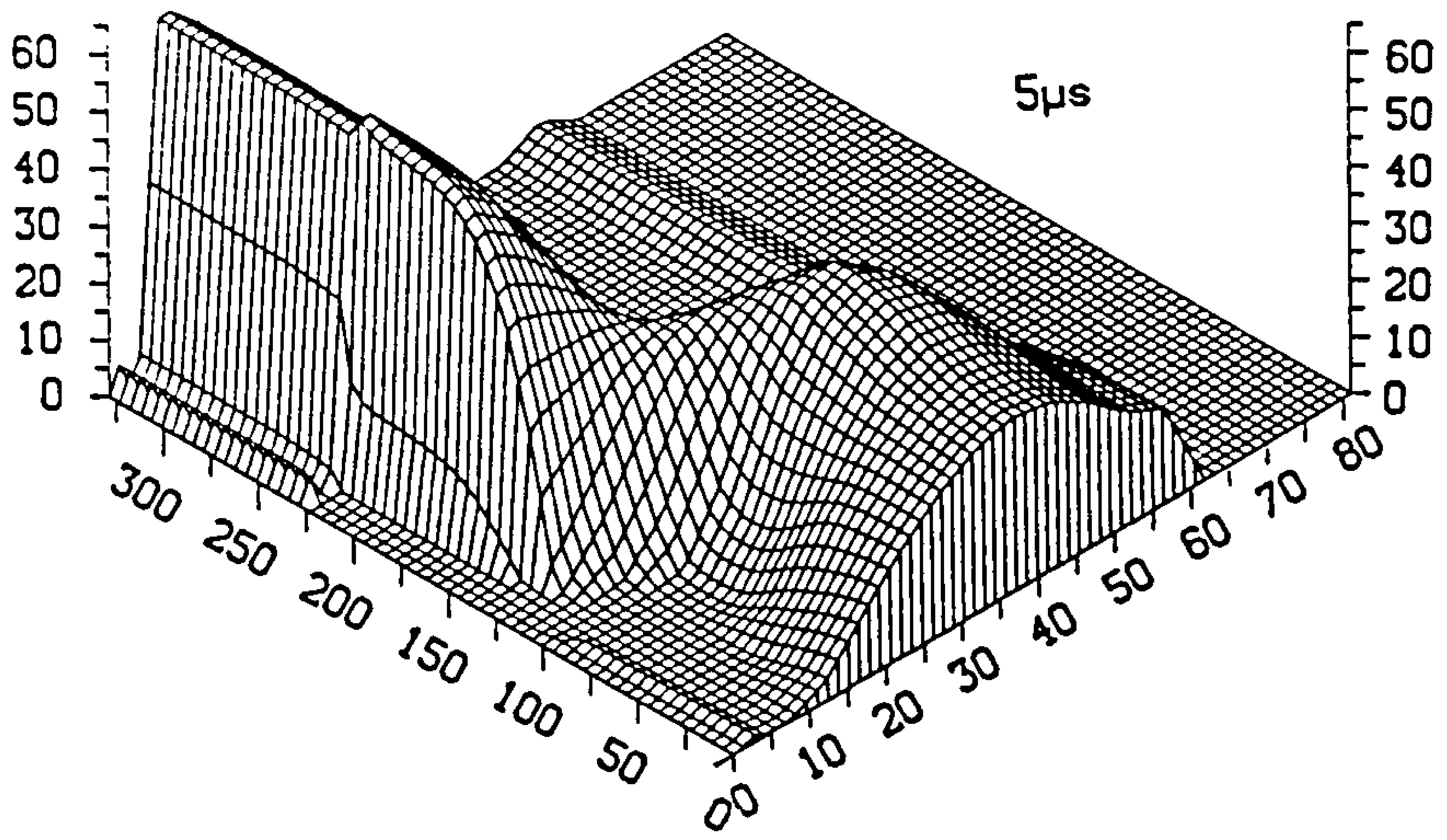
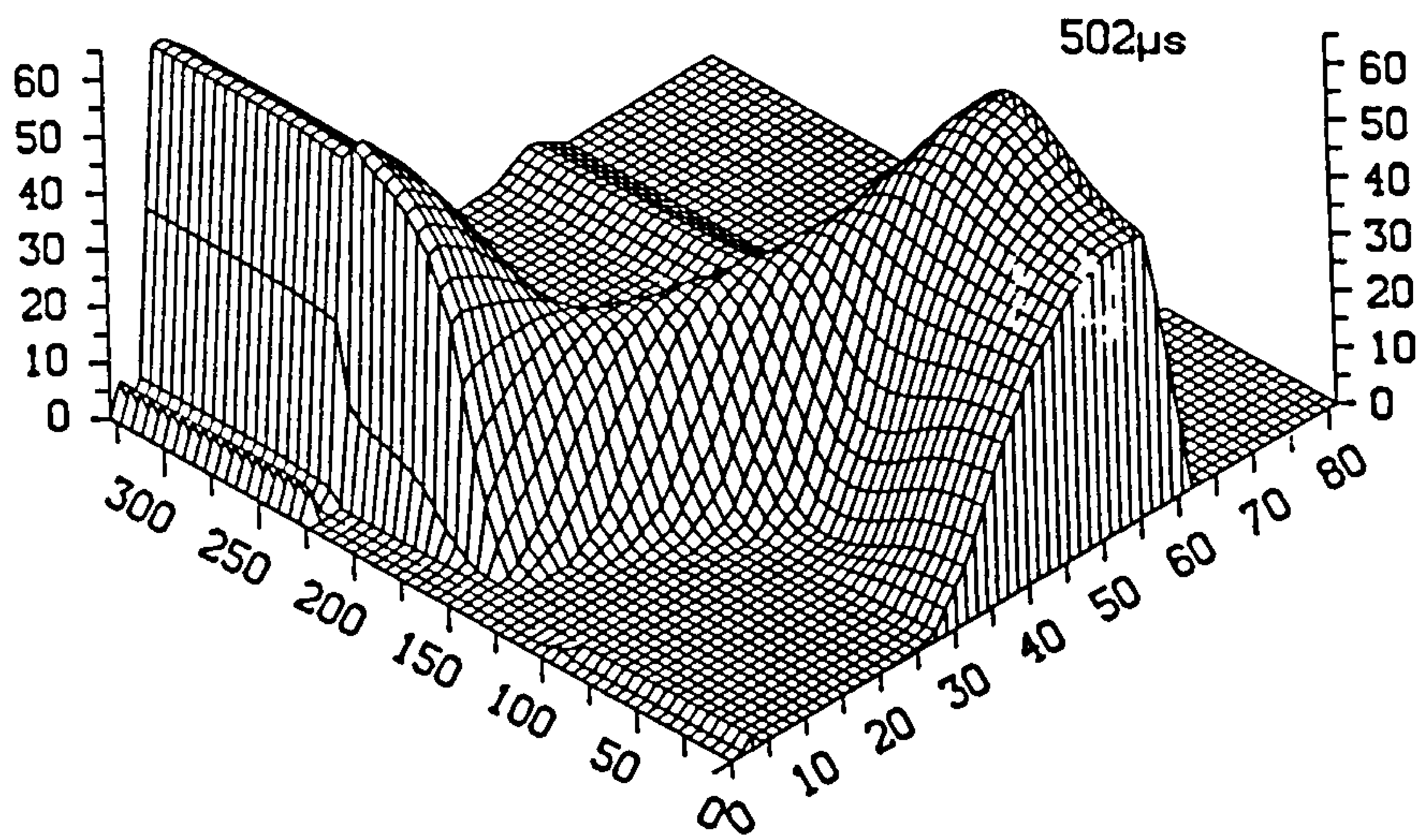


Figure 78. Variation of Peak Temperature and Peak Field for Uniform and Graded Collector Profiles.

The field profiles at various times prior to TSB are given in Figure 79 and the vertical electron current densities are given in Figure 80. The curvature of the vertical field profile is evident as is the downward movement of the field profile with time. This downward movement is explained by the reduction in v_n^{sat} with temperature causing an increase in the electron concentration in the space-charge region. The temperature profiles are shown in Figure 81 and the hole concentration contour diagrams in Figure 82. It can be seen that the temperature initially peaks half way down the collector at the position of the peak field. The peak then moves downwards and towards the emitter centre before TSB is entered. A well developed mesoplasma is evident from Figure 82 at 977μs. At this point in time the plasma concentration is of the order of the peak impurity concentration in the base.





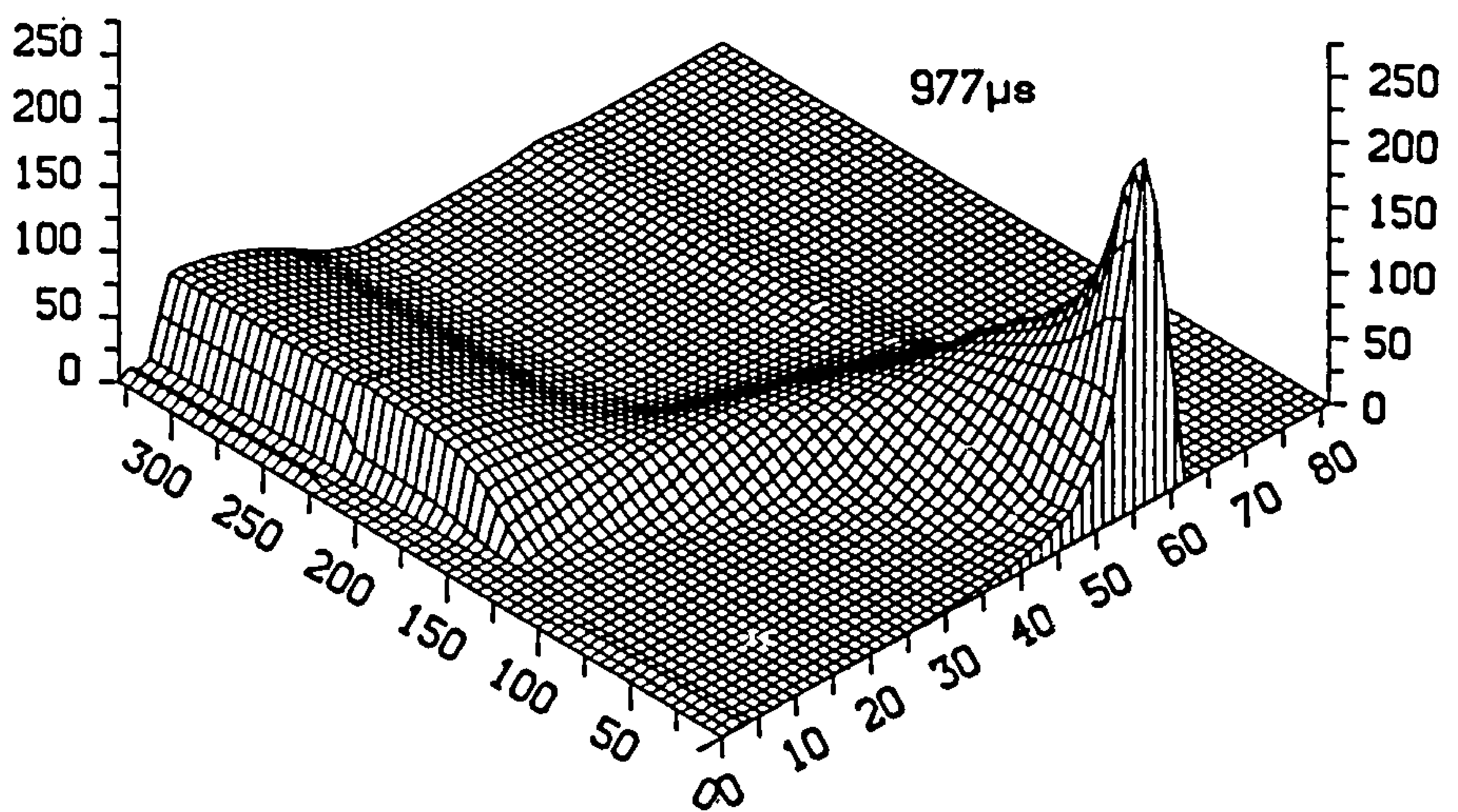


Figure 79. Field Profiles Prior to Breakdown for the Graded Collector: Fields are in $KV\,cm^{-1}$ and distances are in μm.

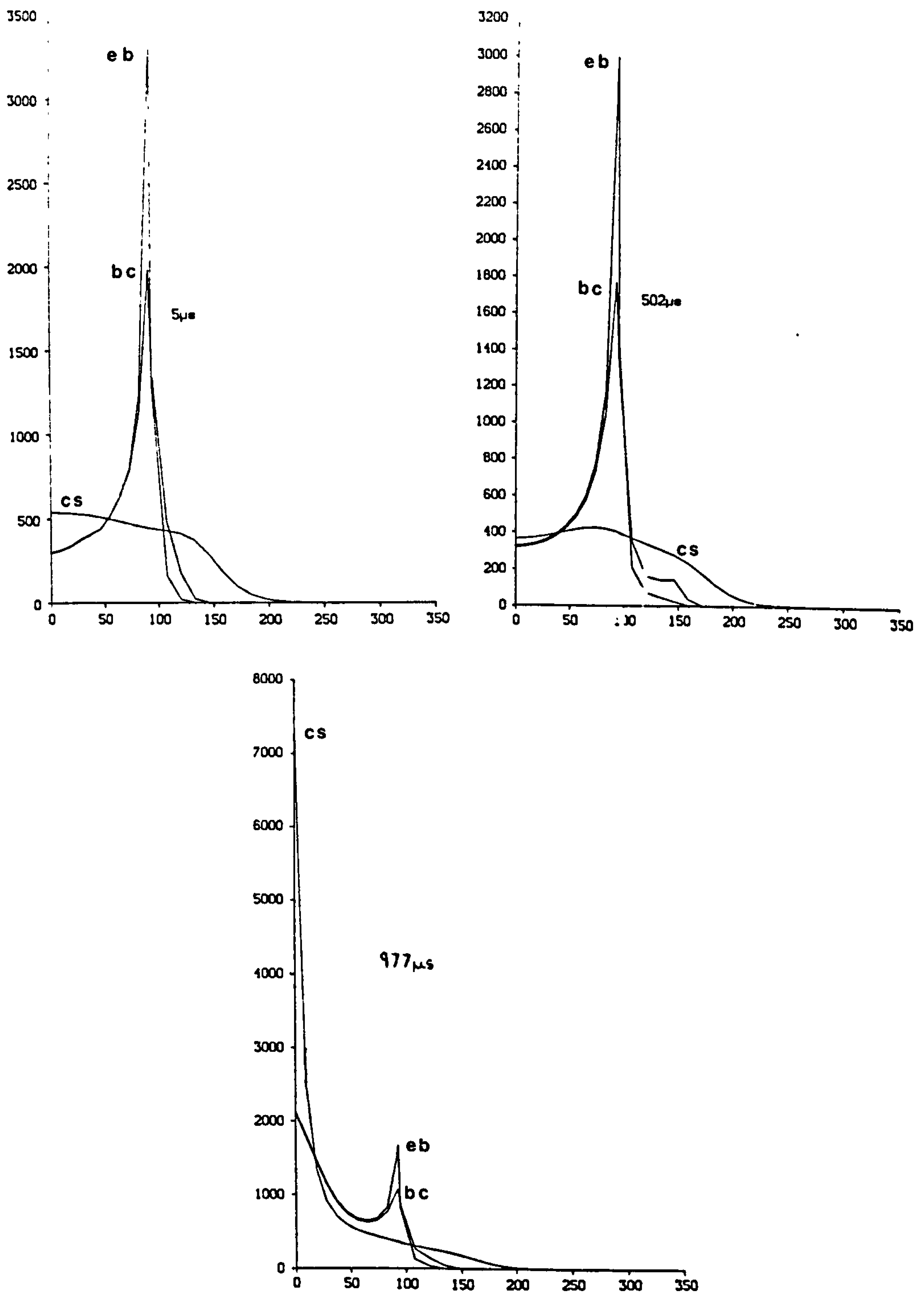
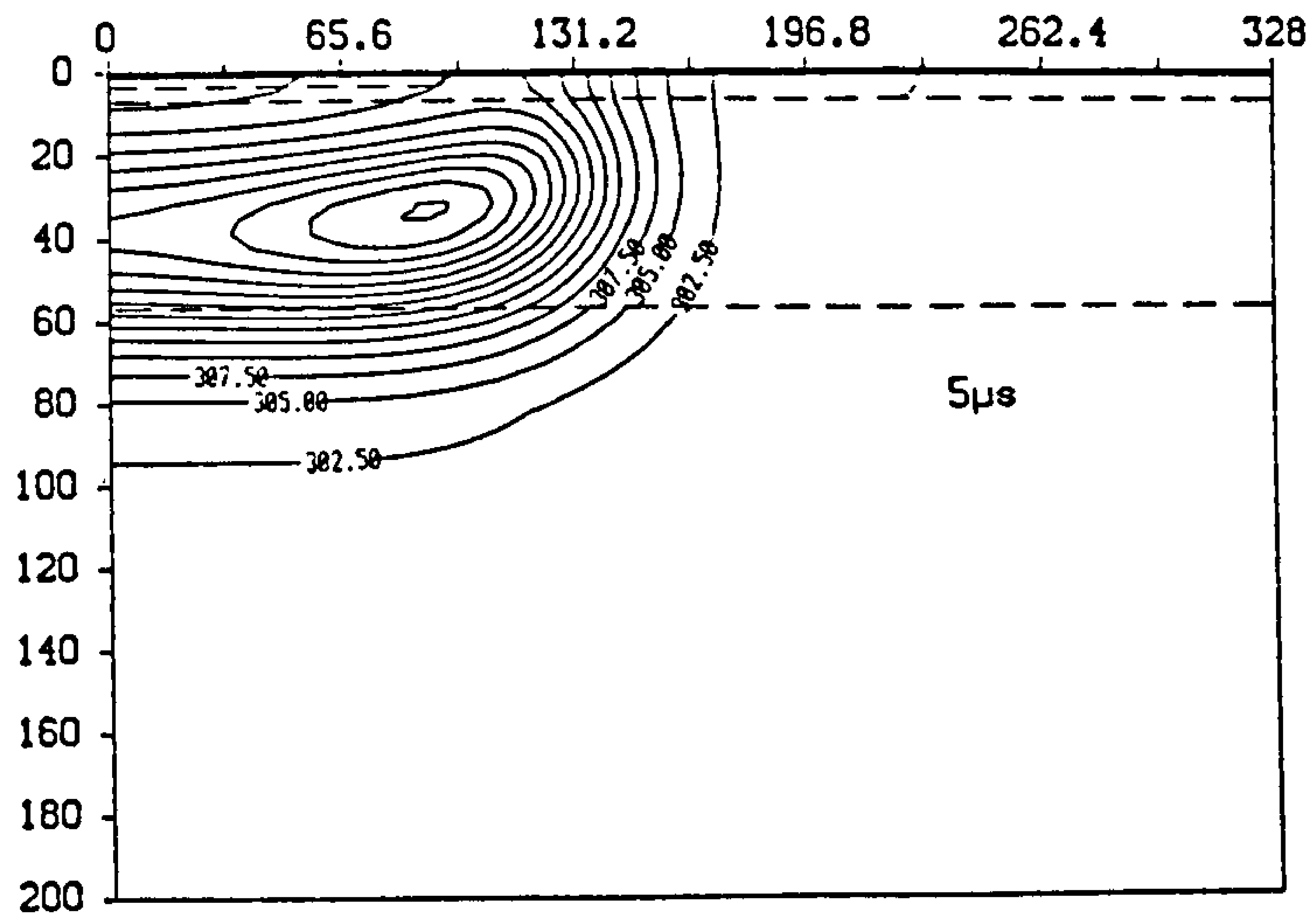
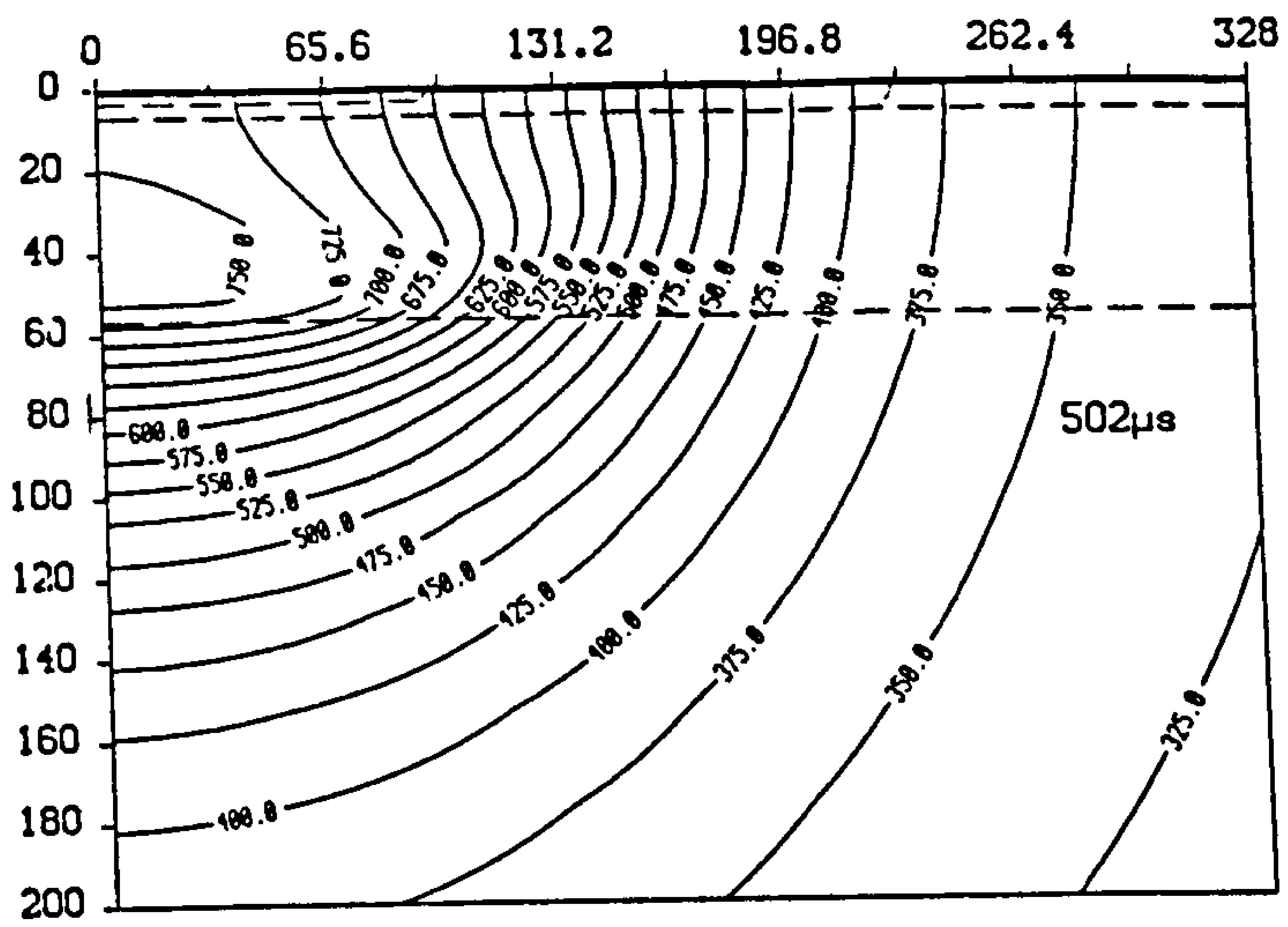


Figure 80. Vertical Electron Current Densities Prior to Breakdown for the Graded Collector.: Current densities are in $A\,cm^{-2}$ and distances are in μm.





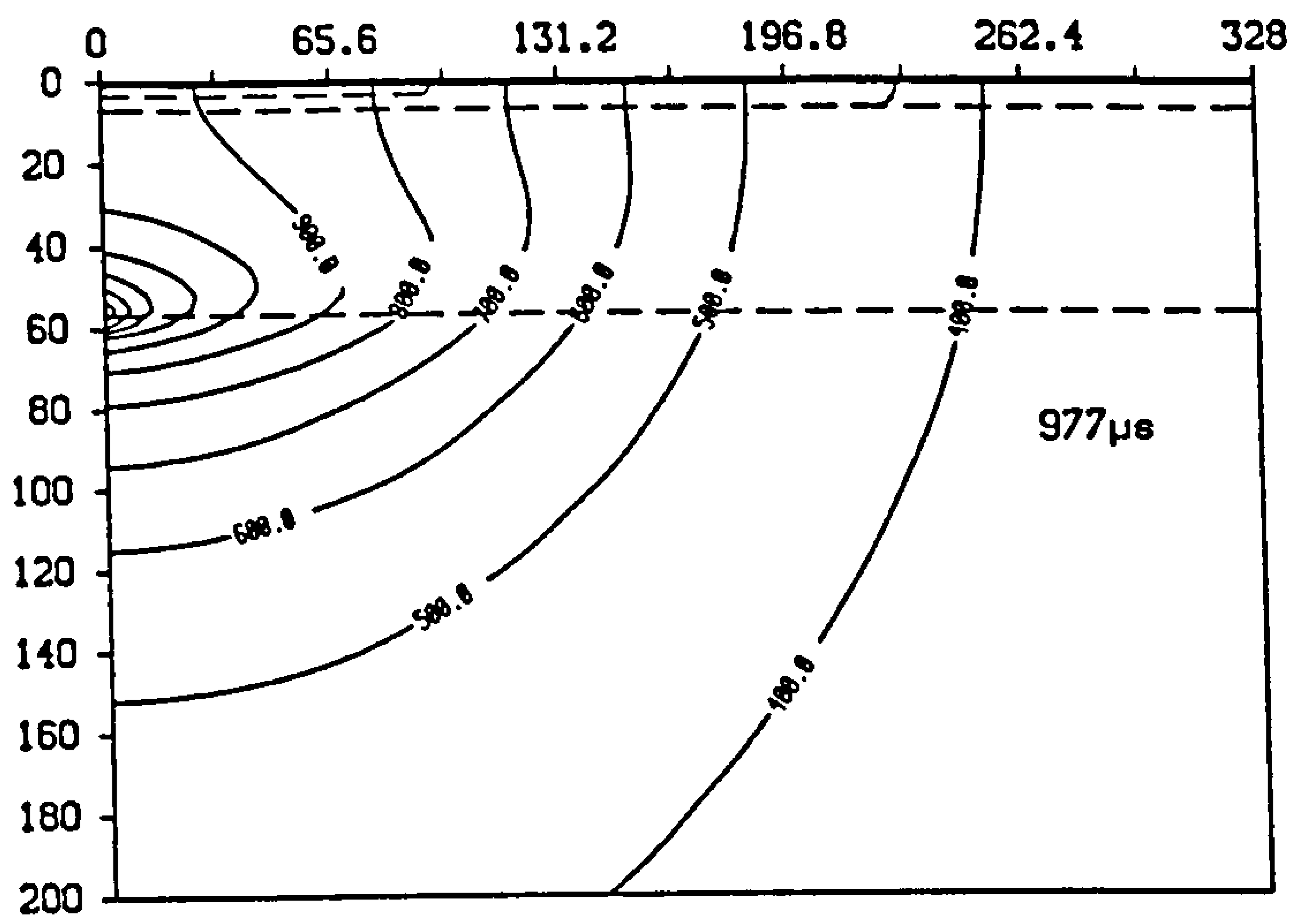
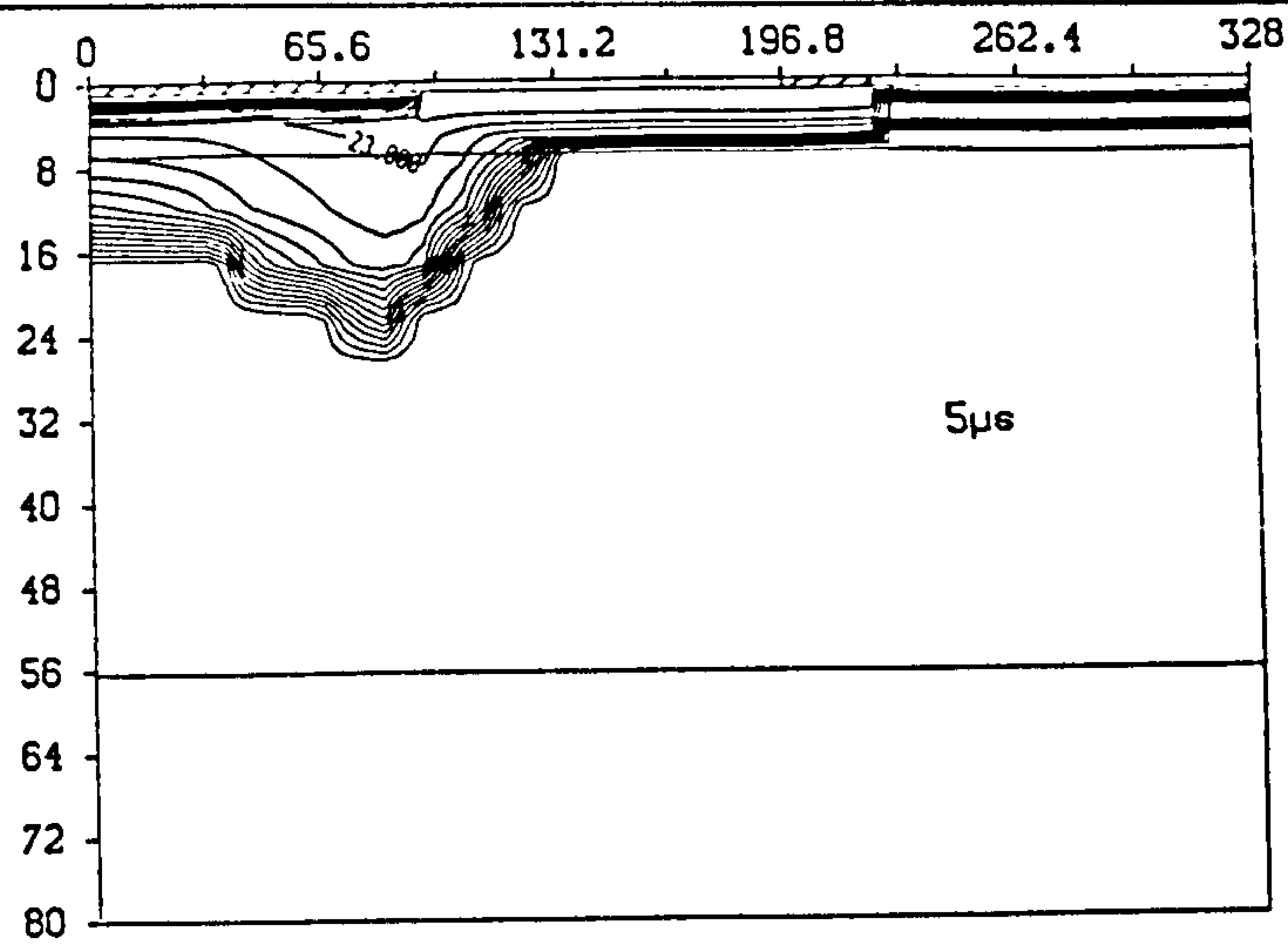
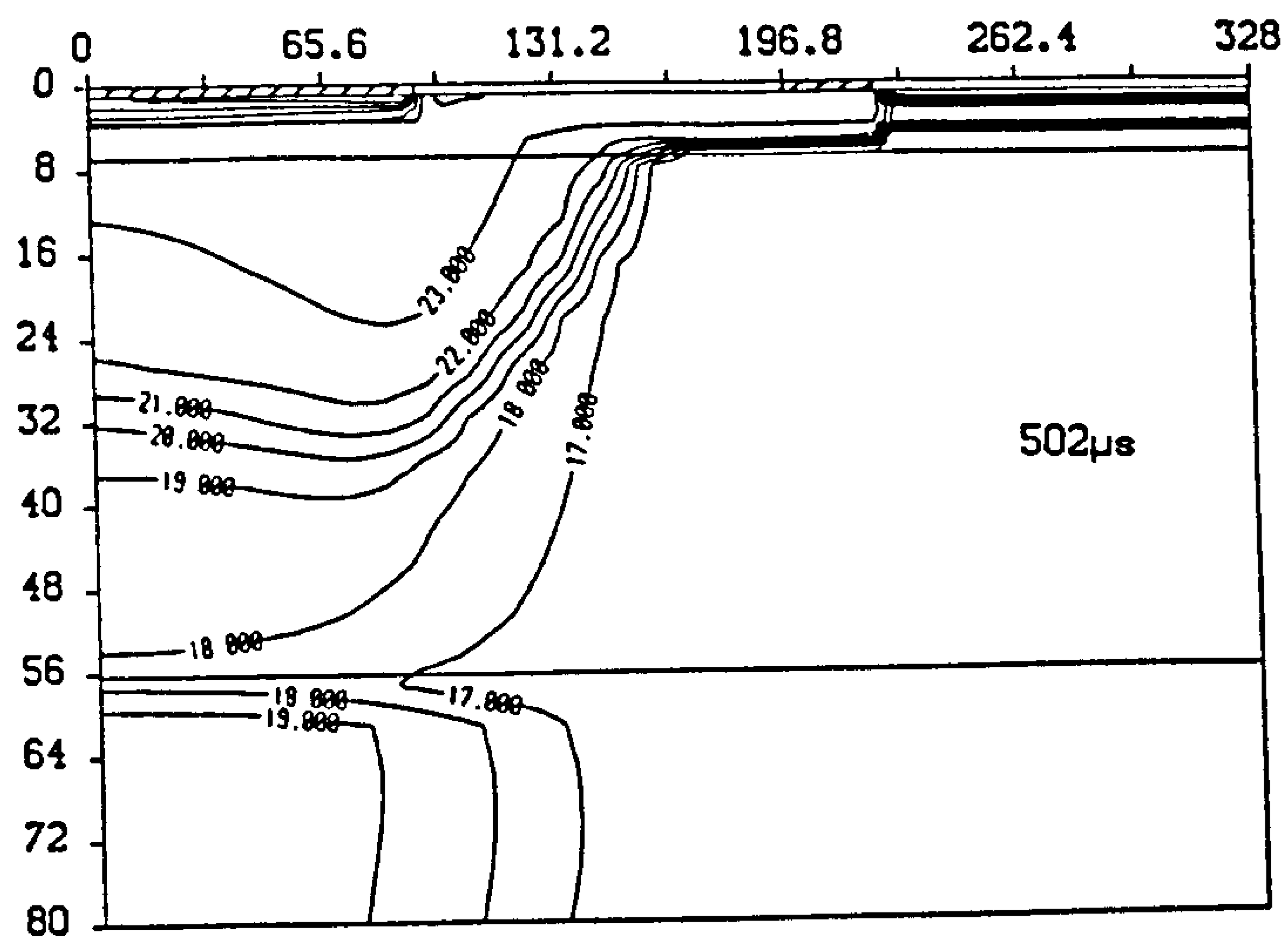


Figure 81. Temperature Profiles Prior to Breakdown for the Graded Collector.





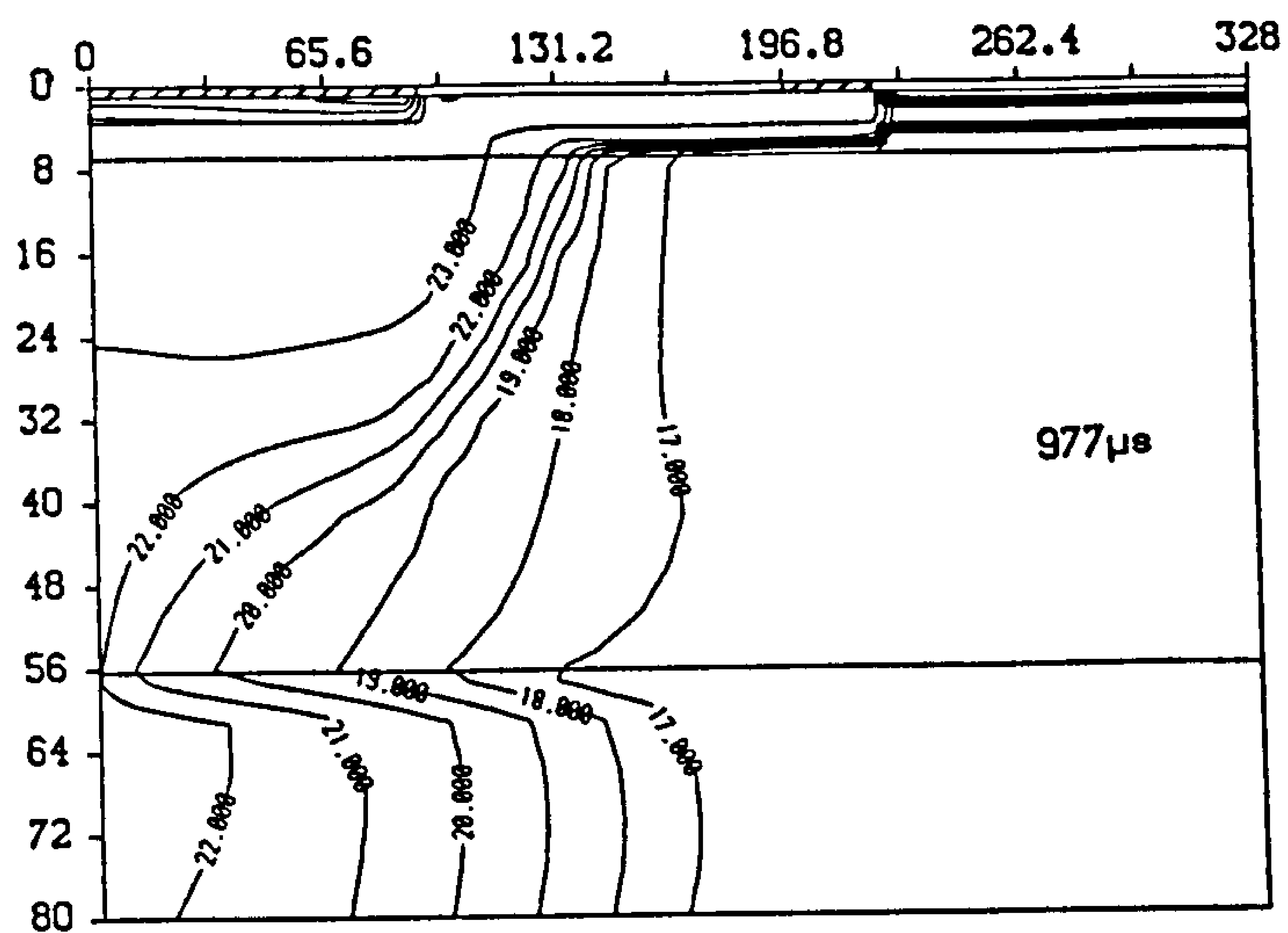


Figure 82. Contour Maps of Hole Concentration Prior to Breakdown for the Graded Collector: The contours are labelled with $\log_{10}$ values of concentrations in units of m^{-3}.

The differences between the numerically and experimentally obtained waveforms may once again be attributed to the previously described differences. An additional factor affecting the 50μm device is that the measured base current prior to any self-heating is higher than the computed value. This suggests that variations in the base doping and perhaps lifetimes have occurred between separate process batches. The modelled profile, however, was not varied so that accurate comparison could be made between computed results for uniform and graded collector layers.

The computed triggering time for a device with the measured collector profile is seen to be less than that for the optimum profile. The measured profile was found to have a higher initial doping and steeper slope than the optimum. This gives in a higher peak field at a given voltage, resulting in a more rapid local temperature rise. A particularly interesting observation can be made by comparing the experimental current waveforms in Figure 68 for the 50μm and 100μm graded collector devices. The base waveform of the 100μm layer exhibits a long slow fall off before entry into TSB. This is because the collector layer is wider and more heating is required to shift the peak field to the collector substrate junction. The peak field does not rise with temperature until the field profile reaches the collector-substrate junction. Thus, the heating is slow resulting in a slow fall in base current.

6.5.7 Conclusion.

The mechanism by which self-heating and eventual second breakdown occurs has been demonstrated with the aid of the numerical model and qualitative agreement has been obtained with experiment. The results have shown that both the peak field and peak temperature shift laterally from below the emitter edge to below the emitter centre prior to the formation of a high current filament through the emitter centre. The results do not support the concept of a critical temperature, rather the device is found to enter TSB when $I_B = 0$, which depends on the over all temperature profile. The differences between computed and measured triggering times can be attributed to the different collector lifetimes and to the fact that series emitter resistances were not included in the model. It has also been demonstrated that the triggering time to TSB can be improved by utilizing a ring emitter structure or a graded collector impurity profile. Device ruggedness is, therefore, improved and its Safe Operating Area (SOA) is extended.

References.

6.1 K. Hwang and D. H. Navon, "Breakdown Voltage Optimization of Silicon p-π-ν Planar Junction Diodes," IEEE Trans. Electron Devices, **ED-31**, pp. 1126-1135 (1984).

6.2 R. Van Overstraeten and H. De Man, "Measurement of the Ionization Rates in Diffused Silicon *p-n* Junctions," Solid State Electron., **13**, pp. 583-608 (1970).

6.3 N. R. Howard, "Avalanche Multiplication in Silicon Junctions," J. Electron. Control, **13**, pp. 537-544 (1962).

6.4 C. A. Lee, R. A. Logan, R. L. Batdorf, J. J. Kleimack and W. Wiegmann, "Ionization Rates of Holes and Electrons in Silicon," Phys. Rev., **134**, pp. A761-A773 (1964).

6.5 Y. C. Kao and E. D. Wolley, "High-Voltage Planar *p-n* Junctions," Proc. IEEE, **55**, pp. 1409-1414 (1967).

6.6 M. S. Adler, V. A. K. Temple, A. P. Ferro and R. C. Rustay, "Theory and Breakdown Voltage for Planar Devices with a Single Field Limiting Ring," IEEE Trans. Electron Devices, **ED-24**, pp. 107-113 (1977).

6.7 K. R. Whight and D. J. Coe, "Numerical Analysis of Multiple Field Limiting Ring Systems," Solid State Electron., **27**, pp. 1021-1027 (1984).

6.8 P. Walker, J.T. Davies and K.I. Nuttall, "A Numerical Analysis of the Resurf Diode Structure," IEE Proc., **132**, Pt. I, pp. 285-290 (1985).

6.9 S. M. Sze, *VLSI Technology*, McGraw Hill, New York, 1988.

6.10 F. Conti and M. Conti, "Surface Breakdown in Silicon Planar Diodes Equipped with Field Plate," Solid State Electron., **15**, pp. 93-105 (1972).

6.11 H. Ryssel, K. Haberger, K. Hoffmann, K. Muller and R. Henkelmann, "Simulation of Doping Processes," IEEE Trans. Electron Devices, **ED-27**, pp. 1484-1492 (1980).

6.12 C. P. Ho, J. D. Plummer, S. E. Hansen and R. W. Dutton, "VLSI Process Modeling-SUPREM3," IEEE Trans. Electron Devices, **ED-30**, pp. 1438-1453 (1983).

6.13 M. E. Law, C. S. Rafferty and R. W. Dutton, "SUPREM IV Users Manual," Technical Report, Integrated Circuits Laboratory, Department of Electrical Engineering, Stanford University, July 1986.

6.14 J. Lorenz, J. Pelka, H. Ryssel, A. Sachs, A. Seidl and M. Svoboda, "COMPOSITE-A Complete Modeling Program of Silicon Technology," IEEE Trans. Electron Devices, **ED-32**, pp. 1977-1985 (1985).

6.15 J. C. Irvin, "Resistivity of Bulk Silicon and of Diffused Layers in Silicon," Bell Sys. Tech. J., **41**, pp. 387-410 (1962).

6.16 H. K. Gummel, "Measurement of the Number of Impurities in the Base Layer of a Transistor," Proc. IRE, **49**, p. 834 (1961).

6.17 D. P. Kennedy and R. R. O'Brien, "Analysis of the Impurity Atom Distribution Near the Diffusion Mask for a Planar *p-n* Junction," IBM Journal, **9**, pp. 179-186 (1965).

6.18 W. J. Chudobiak, "The Saturation Characteristics of *n-p-v-n* Transistors," IEEE Trans. Electron Devices, **ED-17**, pp. 843-852 (1970).

6.19 R. J. Hauser, "The Effects of Distributed Base Potential on Emitter-Current Injection Density and Effective Base Resistance for Stripe Transistor Geometries," IEEE Trans. Electron Devices, **ED-11**, pp. 238-242 (1964).

6.20 M. J. Humphreys and K.I. Nuttall, "Control of Avalanche Injection in Bipolar Transistors through the use of Graded Collector Impurity Profiles," IEE Proc., **132**, Pt. I, pp. 285-290 (1985).

6.21 L. J. Van der Pauw, "A Method of Measuring Specific Resistivity and Hall Effect of Discs of Arbitrary Shape," Philips Res. Repts., **13**, p. 1 (1958).

6.22 C. T. Kirk, Jr., "A Theory of Transistor Cutoff Frequency Falloff at High Current Densities," IEEE Trans. Electron Devices, **ED-9**, pp. 164-174 (1966).

6.23 P. L. Hower and W. G. Einthoven, "Emitter Current-Crowding in High-Voltage Transistors," IEEE Trans. Electron Devices, **ED-25**, pp. 465-471 (1978).

6.24 K. Owyang and P. Shafer, "A New Power Transistor Structure for Improved Switching Performances," IEDM Tech. Digest, pp. 667-670 (1978).

6.25 Y. Nakatani and I. Kuruyu, "An Ultra High Speed - Large Safe Operating Area Transistor with New Fine Emitter Structure," INTELEC Tech. Program, pp. 500-507 (1983).

6.26 G. Miller, A. Porst and H. Strack, "An Advanced High Voltage Bipolar Power Transistor with Extended RBSOA using $5\mu m$ Small Emitter Structures," IEDM Tech. Digest, pp. 142-145 (1985).

6.27 S. A. Higgins, M. K. Johnson, P. A. Gough and J. A. G. Slatter, "Modelling Bipolar Transistor Second Breakdown During Turn-Off by Solution of the Fundamental Device Equations," ESSDERC Tech. Program, pp. 65-69 (1987).

6.28 C. G. Thornton and C. D. Simmons, "A New High Current Mode of Transistor Operation," IRE Trans. Electron Devices, **ED-5**, pp. 6-10 (1958).

6.29 H. A. Schafft, "Second Breakdown-A Comprehensive Review," Proc. IEEE, **55**, pp. 1272-1288 (1967).

6.30 P. L. Hower and V. G. K. Reddi, "Avanache Injection and Second Breakdown in Transistors," IEEE Trans. Electron Devices, **ED-17**, pp. 320-335 (1970).

6.31 W. B. Smith, D. H. Pontius and P. P. Budenstein, "Second Breakdown and Damage in Junction Devices," IEEE Trans. Electron Devices, **ED-20**, pp. 731-744 (1973).

6.32 I. Dunn and K. I. Nuttall, "An Investigation of the Voltage Sustained by Epitaxial Bipolar Transistors in Current Mode Second Breakdown," Int. Jnl. Electron., **45**, pp. 353-372 (1978).

6.33 S. K. Ghandi, *Semiconductor Power Devices*, John Wiley & Sons, New York, 1977.

6.34 B. S. Khurana, T. Sugano, H. Yanai, "Thermal Breakdown in Silicon $p-n$ Junction Devices," IEEE Trans. Electron Devices, **ED-13**, pp. 763-770 (1966).

6.35 H. Melchior and M. J. O. Strutt, "Secondary Breakdown in Transistors," Proc. IEEE, **52**, pp. 439-440 (1964).

6.36 T. Agatsuma, T. Kohisa, and A. Sugiyama, "Turnover Phenomenon of N^+NN^+ Plate Contact Silicon Devices and Second Breakdown in Transistors," Proc. IEEE, **53**, p. 95 (1965).

6.37 R. A. Sunshine and M. A. Lampert, "Second Breakdown Phenomenon in Avalanching Silicon-on-Sapphire Diodes," IEEE Trans. Electron Devices, **ED-19**, pp. 873-885 (1972).

6.38 A. J. Ede, *An Introduction to Heat Transfer Principles and Calculations*, Pergamon Press, Oxford, 1967.

6.39 R. P. Arnold and D. S. Zoroglu, "A Quantitative Study of Emitter Ballasting," IEEE Trans. Electron Devices, **ED-21**, pp. 385-391 (1974).

Chapter 7. Conclusion.

The multi-dimensional nature of semiconductor device operation demands the use of numerical simulation in device design. Analytical models simply cannot account for the for operation in three dimensions. Furthermore, analytical models usually only apply to a particular aspect of overall device operation and the intimate interaction between the different processes within a device is usually neglected. Such problems are largely overcome by the use of numerical modelling, which can be used to accurately predict the operation of practically realisable devices that are beyond purely academic examples. The accuracy is such that the number of design and fabrication cycles required for device development is drastically reduced, giving a significant saving of both time and money.

A major limitation currently affecting the use of numerical methods is that they require considerable computational resources. Until recently numerical models could only be implemented on powerful mainframe computers, which in general has meant that their use has been restricted to the research type environment. Consequently there has been a relatively slow acceptance of numerical modelling by the commercial establishments. However, all this is likely to change in the near future with the advent of cheap but powerful workstations together with parallel processing systems.

7.1 Summary.

The intention of project has been to develop and substantiate a numerical model for semiconductor device operation. Numerical modelling was chosen for its inherent accuracy and because it can be easily adapted to suit a wide range of situations. It has been shown that numerical methods can be used to solve the most complicated problems such as that arising from semiconductor physics. Hence, the problem need not be simplified in order to make it amenable to a solution. Here lies the fundamental difference between analytical and numerical approaches. In analytical methods the full problem is simplified in order to ease

the solution, where as in numerical methods the full problem is solved using a discrete approximation.

The mathematical model described in chapter 2 is fundamental to most semiconductors and it only becomes specific to a given device upon the application of the boundary conditions. A diverse range of problems can be specified therefore, by simply altering the boundary condition, giving a high degree of flexibility. The model caters for a wide range of physical phenomena not normally accounted for such as current spreading, current flow due to temperature gradients, electro-thermal interaction and high doping effects.

In order to quantify the model for a particular semiconducting material a number of physical parameters within the model must be defined. In this case the the material is silicon and the empirically obtained expressions used to describe the physical parameters are presented in chapter 3. A temperature dependent mobility model was presented, which accounts for all the important scattering mechanisms in silicon. An expression for the variation of the intrinsic carrier concentration with temperature was also presented that has proven to be critical when modelling self-heating. The carrier concentrations have been modelled using the Boltzmann approximation with heavy doping effects being treated by assuming rigid shifts of the parabolic density of states functions, through the use of an effective band gap narrowing term. The model for carrier recombination/generation accounts for phonon assisted transitions via deep level traps (SRH) and band to band transitions resulting from interactions between three carriers (Auger). These are by far the most important recombination mechanisms in silicon. Careful consideration has been given to the SRH lifetimes, which are extremely difficult to quantify. This is because unlike many other parameters they are not only functions of known quantities such as temperature and doping, but they also depend upon device processing conditions and material purity. Since the model was to be compared against a particular set of devices, which were all fabricated with the same process, then it was decided that the lifetimes to be inserted into the model should obtained from these devices. The values were either measured directly from these devices, or obtained by comparison of computed and measured electrical data. Also included in chapter 3 is an expression which describes the temperature dependence of the thermal conductivity in silicon, together with an accurate model for heat generation.

Having formulated and fully defined the problem the mathematical model was then solved using numerical methods, which represent the only feasible approach to solving a problem of such complexity. Chapter 4 describes the finite difference method for transforming the continuous problem into a discrete one. Care has been taken to include truncation errors wherever possible, so that the

approximations being made are readily apparent. In addition the truncation errors can themselves be estimated using finite differences and on the basis of these estimates an optimum mesh can be designed.

The discretisation stage generates a large system of non-linear algebraic equations, which presents a considerable computational problem. Here lies the major drawback of numerical modelling. Chapter 5 describes two separate approaches to solving such a system. Two approaches are generally required since one operates more efficiently at low injection levels, while the other is far superior for high level operation. Both approaches make use of Newton's method, which involves repeatedly linearising and solving the system. The solution of the linear system is computational intensive requiring large amounts of processing time. A number of iterative methods for performing this function are described in chapter 5. Iterative techniques were preferred to direct solvers as they are more efficient for sparse banded systems.

Finally in chapter 6 the results of a number of investigations that were performed with the aid of the model were described, together with the design and characterisation of several bipolar transistor geometries. Development of the model was carried out in stages and simulations were made after each development phase. In this way each version of the model could be thoroughly tested for programming errors and numerical accuracy before adding a further stage. Hence, in section 6.1 a comparatively simple 'off-state' model was presented, which provided the foundation for the full model. In this case only Poisson's equation needed to be solved, and by calculating the ionization integral along the resulting field lines it is has been shown that it is possible to obtain an accurate value for reverse biased breakdown voltage. The model was then extended to a full isothermal electrical model for steady state simulation. This involved a solution of the coupled Poisson and current continuity equations for time invariant applied voltages. Electrical results using this model are described in sections 6.2 and 6.3. The steady state model was then extended to cater for transient operation, and at the same time the model was coupled with an inductive collector circuit equation as described in section 6.4. Finally, the full transient electro-thermal model was completed, and it was used to investigate thermal second breakdown phenomena in bipolar transistors, which was discussed in section 6.5.

In many cases comparison has been made between simulation and electrical measurements obtained from the bipolar test structures described in section 6.2. These devices were designed and fabricated as part of this project, and were found to be fully operational. They provided the data against which the

model was calibrated, and favourable agreement was obtained between measured and computed results.

In summary the main objectives of this project have largely been achieved, in that a powerful facility is now available at Liverpool for modelling, in detail, the operation of silicon devices. Although the applications considered here have concentrated on bipolar transistors, the model can just as easily be applied to any other type of device eg. MOSFET's and JFET's. Moreover, the model has been written in such a way that only a small number of simple input routines need to be altered to completely specify a totally different type of device.

7.2 Recommendations.

A number of outstanding recommendations can be made with regard to the future development of the model, which will now be listed in a vague order of preference.

1. The recombination/generation term, R should be extended to include the effects of avalanche generation, G, given by

$$G = \frac{1}{q}\left(\alpha_n \left|J_n\right| + \alpha_p \left|J_p\right|\right) \tag{7.1}$$

   where α_n and α_p are the ionisation rates and are given by (6.2) and (6.3) respectively. Since this is a purely generative process expression (7.1) should be subtracted from R. Once this amendment is made, the contribution avalanche generation makes to the current flow can be simulated. This avoids the need to assume a critical field for breakdown, which was necessary in chapter 6. Apart from providing a more accurate breakdown prediction the addition of (7.1) to the model allows the phenomena leading to breakdown to be investigated. Breakdown voltages can be obtained under a full range of operating conditions, and not just for reverse biased junction breakdown as obtained from the 'off-state' model.
2. Consideration should be given to providing a user interface to the model. This would allow the user to define a particular problem by creating an input file using an input language, which must be capable of being interpreted by the input routines. The input data is then used to initialise the model. The input data file should contain a statement of the discretization mesh together with details of device geometry and attributes. The geometry can be defined using REGION statements followed by the range of mesh points that define the region eg. REGION CONTACT (X 1 5 Y 6 8). Each region block should contain

details of various attributes pertaining to that region eg. applied voltage, doping, relative permittivity, thermal conductivity, lifetimes etc. Any attributes that may vary within a region should be specified using function statements eg. doping, lifetime. The input language would, therefore, be capable of defining any arbitrary geometry including MOS and bipolar transistors.

3. The fact that numerical methods are capable of solving the most difficult problems means that more emphasis is put on the fine details of the physical parameters described in chapter 3. The empirical models used to represent the physical parameters should be reviewed whenever new models or new experimental results become available. Although the models may be well defined at room temperature, some uncertainty exists at higher temperatures due to a lack of experimental data. The model for carrier-carrier scattering is particularly doubtful and this should be given immediate attention. The measurements upon which the mobility model is based are all for majority carrier mobility, that is electron (hole) mobility in *n*-type (*p*-type) material. However, it has been found that minority carrier mobility can be significantly different from this [7.1]. Ionized impurity scattering is found to be heavily dependent on the type of impurity present, especially at high impurity concentrations. The carrier lifetimes should also be carefully considered. Although the doping dependence of lifetime in bulk silicon has been reasonably well quantified, very little is known about the influence of device processing on these values. Thermal diffusion and oxidation together with epitaxy and ion implantation are likely to bring about significant lifetime reductions in the various layers of a device. Some measurements on phosphorus diffused layers are given in [7.2]

4. A major problem with the conventional finite difference method is that it requires mesh lines to be extended right up to the boundaries of the simulation domain. In many cases these mesh lines will extend into regions that are of little interest or regions were the truncation errors are low. A better meshing algorithm would allow for the termination of mesh lines within the domain. This can be achieved using the technique of finite boxes [7.3]. A more general technique [7.4] allows for a more arbitrary positioning of mesh points. In this case the mesh points do not have to be positioned along mesh lines and this approach seems to be particularly well suited to the semiconductor problem. Here the mesh is made up of triangles and mesh refinement is achieved by fitting successively smaller triangles within each other. Although the use of such a mesh requires more 'book-keeping' than for a conventional difference mesh, this should be more than offset by the resulting reduction in problem size. Careful consideration should be given to mesh design so that

truncation errors are minimised and investigations should be made into the use of adaptive meshing as discussed in section 4.3.

5. More effort is required to improve the linear solvers described in chapter 5, especially with regard to solving the system resulting from the coupled approach. The convergence rate of the successive block over-relaxation method is slow even for optimum acceleration parameters. In this case direct *LU* factorisation may be more efficient, though storage requirements will be drastically increased. Furthermore, *LU* decomposition requires the computation of an extremely large number of intermediate coefficients, which can be very time consuming. Careful comparison must be made to identify the most efficient technique. An excellent review of the current state of the art methods for solving the linear systems arising from the semiconductor equations is given in [7.5] In this paper a method called conjugate gradients squared (CGS) is described for solving the linear system resulting from the coupled approach. This method was shown to give a considerable reduction in the number of operations required to obtain a solution compared with *LU* decomposition.
6. The accuracy of the simulated results are critically dependent upon device geometry and impurity profiles. Although such information can be measured from actual devices this would require pre-processing and the model could not be used as a predictive tool. However, the process could be modelled in order to obtain an estimate of device structure and doping. Process modelling is very involved and requires as much effort as device modelling itself. It is, therefore, recommended that the device model be coupled to an existing commercially available process model such as ICECREM, SUPREM III, COMPOSITE or SUPREM IV as described in section 6.2.2. If the model is coupled to a 1D process simulator an error function lateral roll off of the impurity profile would be required to approximate the lateral profile near a mask edge. It would be more desirable to couple the model with a 2D process simulator as this would allow a closed form solution for a given process.
7. Throughout this work all contacts have been assumed to be voltage controlled as the boundary condition for current controlled contacts disrupts the banded nature of the coefficient matrices as stated in chapter 2. However, the novel approach suggested in [7.6] avoids this problem and should be subject to further investigation.
8. As a final point it should be stated that if MOS type devices are to be simulated then the effects of surface scattering of carriers in channel regions should be incorporated into the mobility model. Some models for this are described in [7.7] and [7.8].

References.

7.1 H. S. Bennet, "Improved Concepts for Predicting the Electrical Behavior of Bipolar Structures in Silicon," IEEE Trans. Electron Devices, **ED-30**, pp. 920-927 (1983).

7.2 D. J. Roulston, N. D. Arora and S. G. Chamberlain, "Modeling and Measurement of Minority-Carrier Lifetime versus Doping in Diffused Layers of n^+-p Silicon Diodes," IEEE Trans. Electron Devices, **ED-29**, pp. 284-291 (1982).

7.3 A. F. Franz, G. A. Franz, S. Selberherr, C. Ringhofer, P. Markowich, "Finite Boxes - A Generalization of the Finite Difference Method Suitable for Semiconductor Device Simulation," IEEE Trans. Electron Devices, **ED-30**, pp. 1070- 1082 (1983).

7.4 C. H. Price, *Two-Dimensional Numerical Simulation of Semiconductor Devices*, Ph. D. Thesis, Stanford University, 1982.

7.5 S. J. Polak, C Den Heijer, W. H. A. Schilders and P. Markowich, "Semiconductor Device Modelling From The Numerical Point of View," Int. J. for Num. Meth. in Eng., **24**, pp. 763-838 (1987).

7.6 P. A. Gough, M. K. Johnson, S. A. Higgins, J. A. G. Slatter and K. R. Whight, "Two Dimensional Simulation of Power Devices with Circuit Boundary Conditions," Proc. NASECODE V Conf., pp. 213-218 (1987).

7.7 K. Yamaguchi, "A Mobility Model for Carriers in the MOS Inversion Layer," IEEE Trans. Electron Devices, **ED-30**, pp. 658-663 (1983).

7.8 S. Selberherr, A. Schütz and H. Pötzl, "Two Dimensional MOS-Transistor Modeling," In: Process and Device Simulation for Integrated Circuit Design, The Hague, Martinus Nijhoff, pp. 490-581 (1983).

Appendix.

Copies of two publications arising from this project are included here. The first, entitled 'A Numerical Analysis of the Resurf Diode Structure' is described in section 6.1. The second, entitled 'Optimisation of VDMOS Power Transistors for Minimum On-State Resistance' has not been considered in this thesis, as a detailed account of the results in this paper has already been given in J. T. Davies' Ph. D. Thesis, The University of Liverpool, 1985.

A numerical analysis of the resurf diode structure

P. Walker, J.T. Davies and K.I. Nuttall, Ph.D.

Indexing terms: *Diodes, Semiconductor devices and materials, Numerical analysis*

Abstract: The results of a 2-dimensional numerical analysis of medium and high-voltage diode structures that incorporate a 'resurf' field reduction layer are presented. The work illustrates the effect of surface charge on the optimisation of the design and indicates the requirements that will ensure bulk breakdown for a wide range of surface charge densities. The results are used to assess the analytical design equation presented by Appels *et al.*, modified to take account of surface charge. A comparison is also made with results obtained from an analysis of the field limiting ring technique, and the relative performance of the two methods is assessed.

1 Introduction

The production of high-voltage devices requires special consideration to be given to the effects of junction curvature and surface charge. Both of these can lead to a significant reduction of the breakdown voltage when compared to the theoretical maximum possible with a given substrate resistivity. Various schemes have been proposed to alleviate these effects, including surface contouring [1], diffused guard rings [2] and field plate structures [3]. More recently, the field limiting ring system originally proposed by Kao *et al.* [4] has received further attention [5, 6], and a new technique employing 'resurf' layers has been suggested [7]. Because 2-dimensional effects are fundamental to a detailed appreciation of these techniques, particularly in the presence of surface charge, computer analysis methods are being increasingly employed to assess their relative value and to produce design guidelines. Both single and multiple ring systems have been analysed recently in this way [5, 6] and design rules for these are emerging. Unfortunately, variations in the surface charge density pose a particular problem concerning the practical implementation of the structures and, in many cases, its influence on the optimised design has yet to be assessed in a quantifiable manner.

This paper presents the results of a computer analysis of diode structures employing 'resurf' layer protection techniques. The results illustrate how surface charges modify the design parameters, and indicate the requirements that will ensure bulk breakdown for a wide range of surface charge densities. A comparison is made with results obtained from an analysis of the field limiting ring system, and the relative performance of the two methods is assessed.

2 Analysis

A 2-dimensional solution of Poisson's equation using the finite difference method with a nonuniform grid has been obtained for the typical diode structure shown in Fig. 1.

Local space charge has been assumed to be given by

$$\varrho = -q(N_D - N_A + p - n)$$

where $p = n_i \exp\{q(\psi_i - \phi_p)/KT\}$
$n = n_i \exp\{q(\phi_n - \psi_i)/KT\}$
ψ_i = local potential
ϕ_n, ϕ_p = quasi-Fermi levels for electrons and holes respectively,
and N_D, N_A are the local donor and acceptor dopant concentrations.

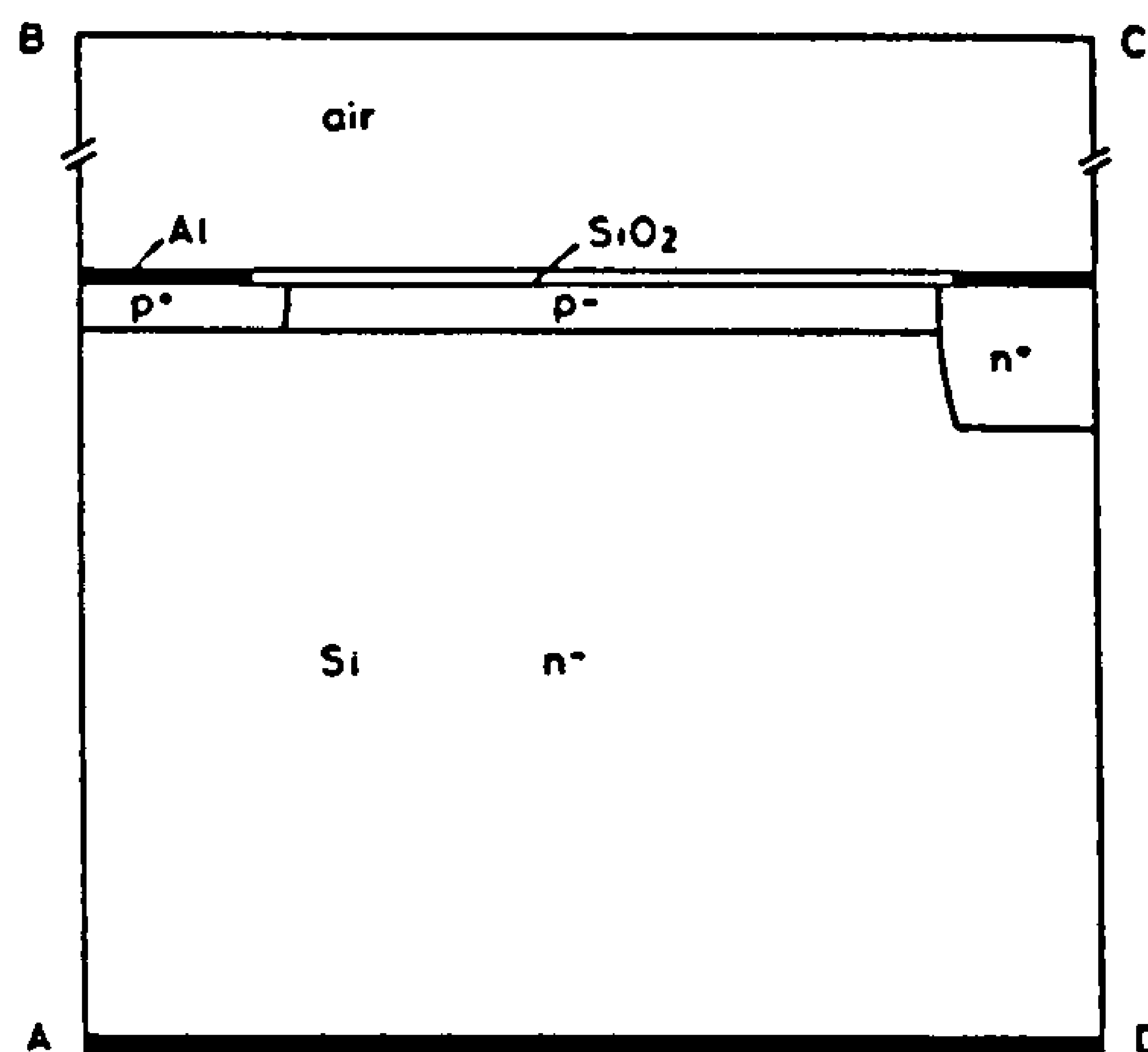


Fig. 1 *The geometry of the resurf structure studied in this work*
The construction assumed is that of a uniformly doped p^- epitaxial layer on an n^- substrate with p^+ and n^+ Gaussian diffusions

Uniform doping has been assumed for the n^- substrate and p^- expitaxial layer, with Gaussian profiles to approximate the p^+ and n^+ diffusions. An assessment of leakage currents has not been included in this analysis, and the positions of the quasi-Fermi levels have therefore been assumed to be determined by the condition of electrical neutrality at the appropriate ohmic contact and to extend without variation throughout the remainder of the silicon [8].

A condition of symmetry has been applied to the two sides *AB*, *CD* of the region analysed. Both the contact to the p^+ diffused region and the top boundary delineating the air space above the chip were fixed at zero potential, and the appropriate reverse bias applied to a contact along the lower face *AD*. The material parameters used are those appropriate to silicon, capped with a 1μm SiO_2 layer and aluminium metallisation where applicable. Equal vertical and horizontal diffusion depths have been assumed for the p^+ and n^+ diffusions, but the vertical scale of Fig. 1 has been enlarged by a factor of ten for presentation purposes.

The breakdown condition has been determined by evaluating the ionisation integral

$$1 - 1/M_p = \int_{x_n}^{x_p} \alpha_p \exp \int_{x_n}^{x} (\alpha_n - \alpha_p)\, dx'\, dx \quad (1)$$

Paper 4251I (E3, P6), received 31st May 1985

Mr Walker and Dr. Nuttall are, and Mr Davies was formerly, with the Department of Electrical Engineering and Electronics, University of Liverpool, Brownlow Hill, PO Box 147, Liverpool L69 3BX, United Kingdom. Mr. Davies is now with British Aerospace PLC, Warton Aerodrome, Warton, Lancs., United Kingdom

along several flux lines, commencing with that passing through the region of maximum electric field strength. Although that flux line often resulted in the largest value for the integral, care was needed when substantial regions of moderate field strength existed. The applied bias was steadily increased until the peak value for the ionisation integral became unity. Although eqn. 1 assumes the multiplication to be dominated by minority carrier injection from the n^- region, the breakdown voltage obtained is not dependent on that assumption.

The ionisation rate data used was based on the results obtained by Van Overstraeten and DeMan [13] for fields in the range $1.75 \times 10^5 < E < 4.0 \times 10^5$ V/cm according to the relationships

$$\alpha_n = 7.03 \times 10^5 \exp\{-1.231 \times 10^6/E\} \text{ cm}^{-1}$$

$$\alpha_p = 1.582 \times 10^6 \exp\{-2.036 \times 10^6/E\} \text{ cm}^{-1}$$

Calculation of the breakdown voltage appropriate to the plane diffused p^+n junction along the boundary AB confirmed that these coefficients satisfactorily predicted the measured breakdown voltages of plane p^+n step junctions published previously by various authors (Fig. 2). The deviation noted at lower breakdown voltages can be attributed

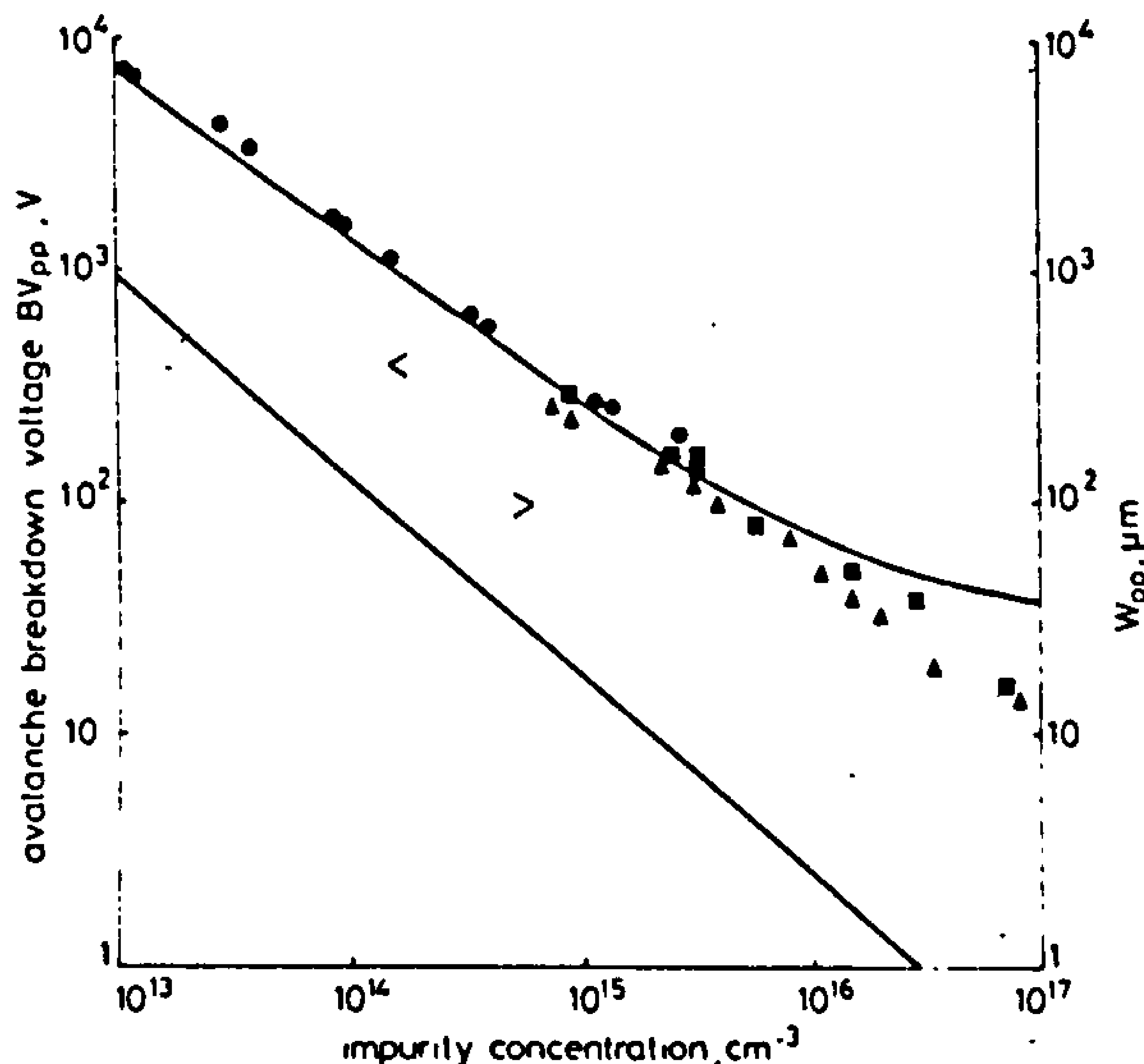


Fig. 2 *Avalanche breakdown voltage and depletion layer width at breakdown against substrate concentration for a plane parallel p^+n diffused junction*

The solid lines representing breakdown voltage and junction width have been calculated using the ionisation rate data of Van Overstraeten and De Man [13]

—— this work
● Kokosa and Davies [10]
■ Miller [11]
▲ McKay [12]

to the diffused impurity profile of the junction analysed, whilst the slight underestimation noted for high voltage devices is consistent with the conclusions of recent authors [9].

3 Results

The procedure described above has been applied to devices with the geometry shown in Fig. 1, using a range of p^+ junction depths (x_j) and p^- epitaxial layer and n^- substrate resistivities. Tests were also conducted for different p^- epitax layer thicknesses, but in this work the layer thickness has been limited to the depth of the p^+ diffusion to allow a fair comparison with alternative schemes and to minimise the possibility of interference with the normal operation of the particular device protected. Hence the peak breakdown voltage realised in these calculations cannot exceed that established by the plane p^+n^- interface, BV_{pp}. In the absence of any protection, simulated by replacing N_{epi} by $-N_{sub}$, the planar breakdown voltage BV_{planar} is obtained. These represent the upper and lower bounds for which breakdown voltages have been determined. Surface charge densities of 0 and 5×10^{11} charges/cm^2 have also been considered, encompassing the range typically encountered in practice [14]. In addition, the breakdown voltage will be dependent on the width of the p^- resurf layer, but the sensitivity is weak at all resurf doping levels provided that it exceeds that of the plane parallel depletion layer. Accordingly, this dimension has been chosen to be considerably wider than would normally be necessary, to remove the consequences of a further adjustable parameter.

Fig. 3 shows the result obtained on a device with a substrate doping concentration of 10^{14} cm^{-3} and junction

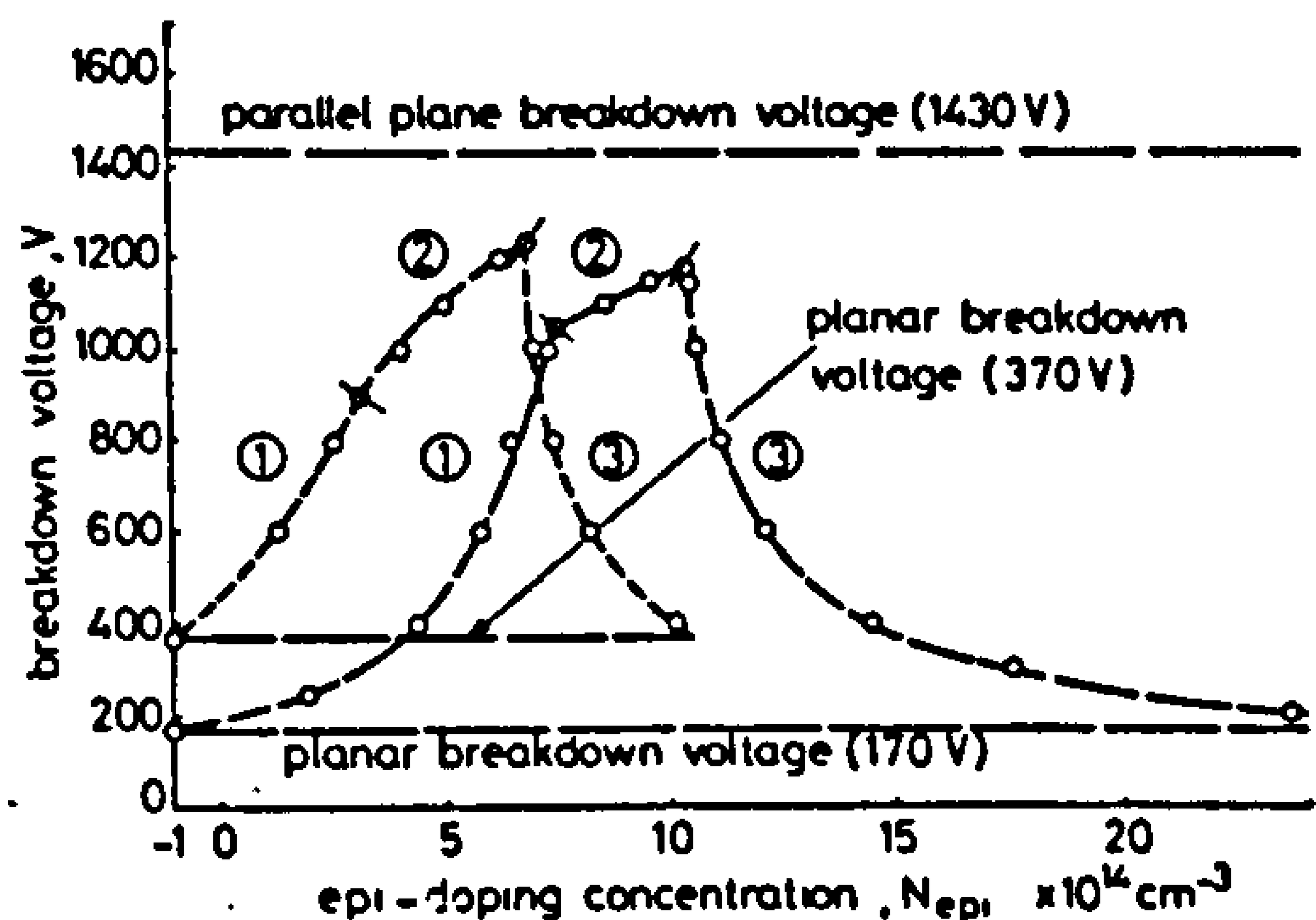
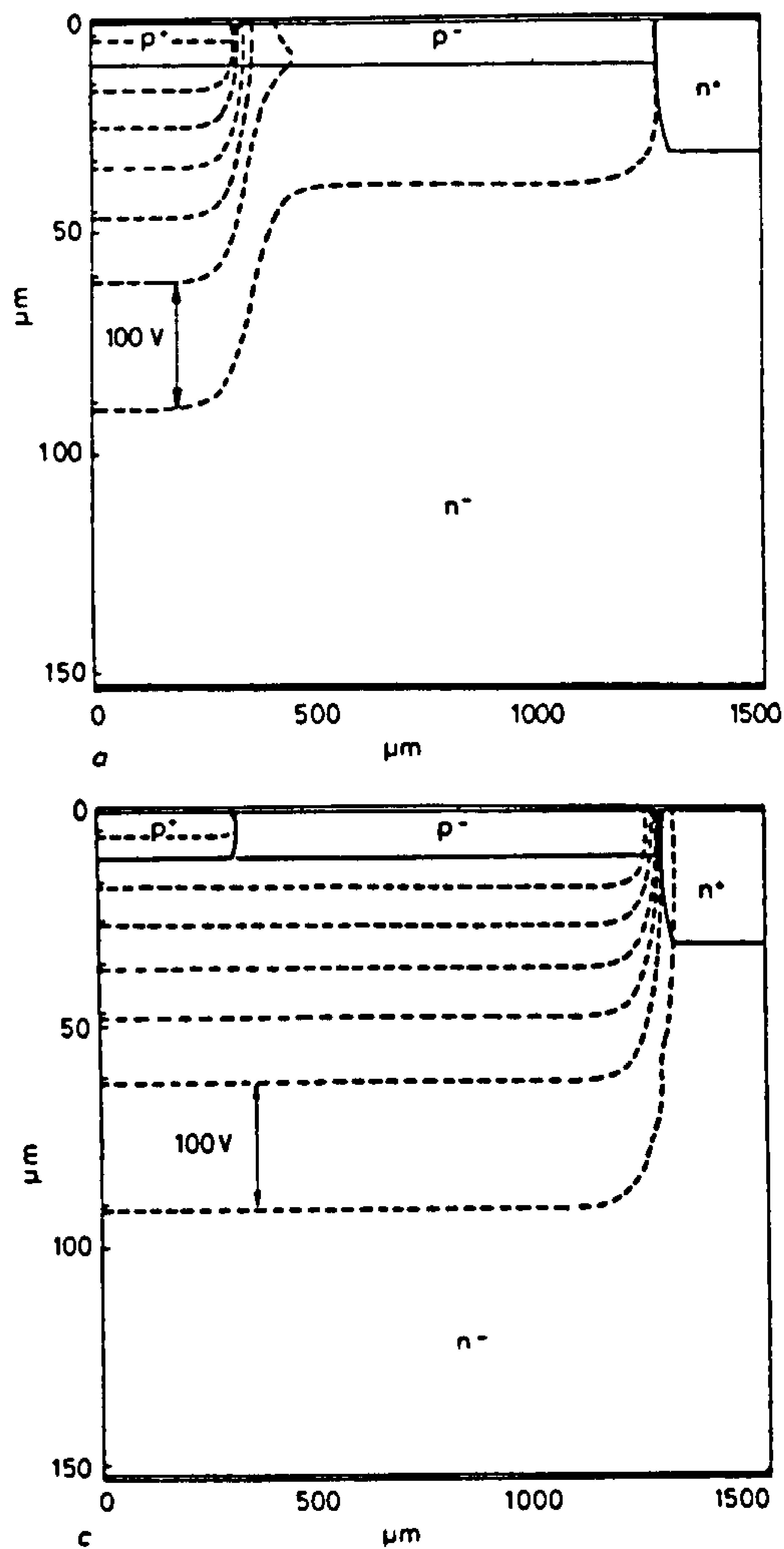


Fig. 3 *Breakdown voltage against impurity concentration in the p^- epitaxial resurf layer*

$x_j = 10$ µm, $N_{sub} = 10^{14}$ cm^{-3}, $N_{ss} = 0$ and 5×10^{11} cm^{-2}
—— $N_{ss} = 0$
– – – $N_{ss} = 5 \times 10^{11}$ cm^{-2}

depth 10 µm. The breakdown voltage for a plane junction, BV_{pp}, and that of the corresponding planar junction, BV_{planar}, (closely approximating a p^+n step junction at this value of substrate doping) are also indicated. Junction curvature effects cause BV_{planar} to be less than BV_{pp} and surface charges lower BV_{planar} still further. The results show that the substantial reduction in breakdown voltage due to junction curvature can, to a very large extent, be prevented by using the resurf layer, even in the presence of significant surface charge. Three portions of the graph may be identified, labelled 1-3 in Fig. 3, each corresponding to breakdown at a different location. Potential contour maps are shown in Fig. 4 to illustrate the sequence that is typical of this structure. The initial rise of breakdown voltage with p^- epitax doping (region 1 in Fig. 3) is associated with surface breakdown at the p^+p^- interface, Fig. 4*a*. As the epitax doping is increased, the breakdown location eventually transfers to the lower portion of the curved p^+ diffusion boundary in the bulk, giving rise to region 2 in Fig. 3. The breakdown voltage is often relatively insensitive to epitax resistivity in this region, but it is abruptly terminated by region 3 when the doping is sufficient to allow breakdown at the p^-n^+ interface, Fig. 4*c*. A relatively small range of epitax doping levels may therefore be identified as representing a near optimum device, outside of which the breakdown voltage decreases sharply. The severity of the fall is greater than that expected on the basis of a



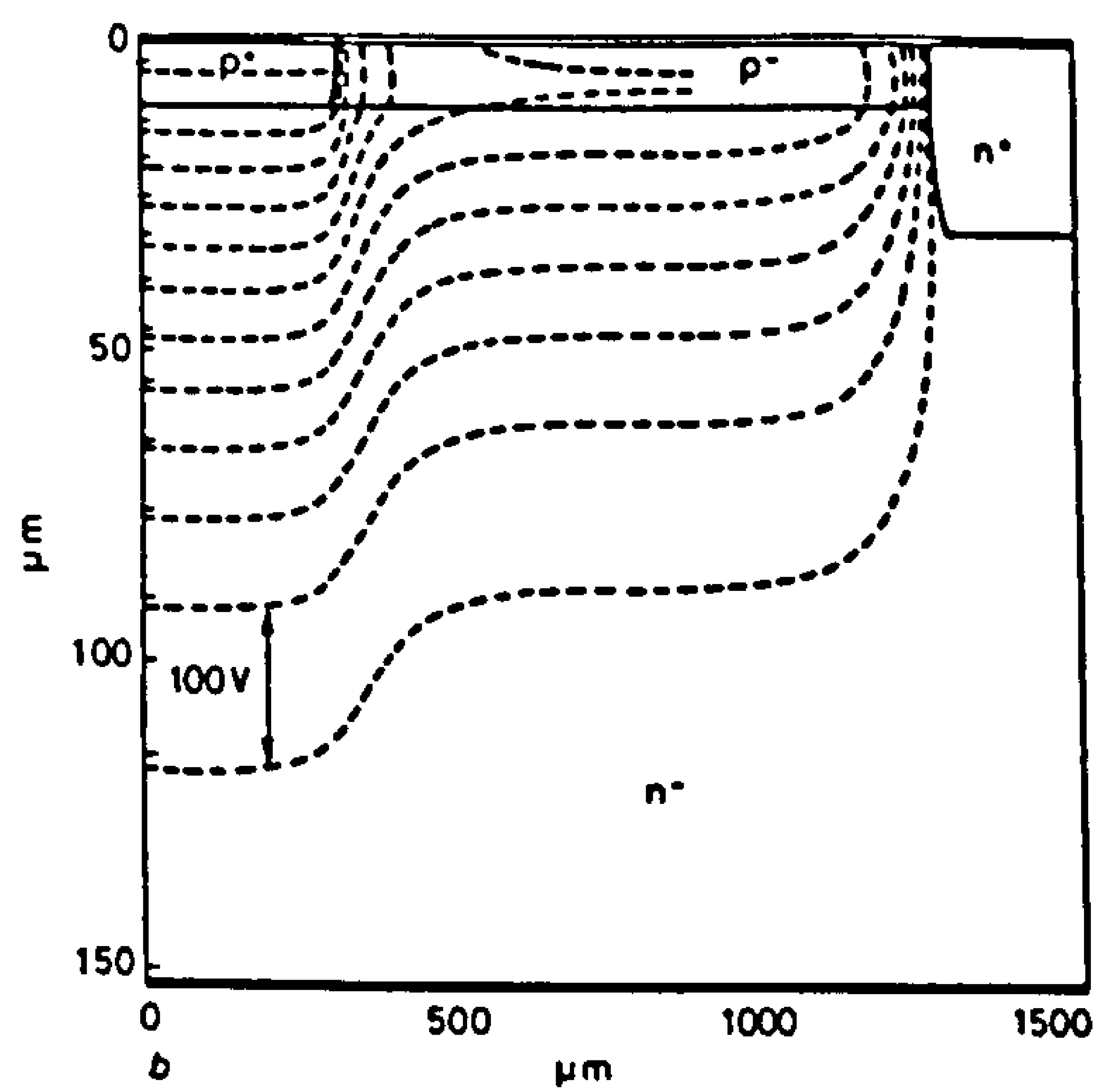


Fig. 4 *Equipotential contours*
$N_s = 5 \times 10^{11}$ cm^{-2}
$x_j = 10$ µm
$N_{sub} = 10^{14}$ cm^{-3}
a device corresponding to region 1 of Fig. 3
$N_{epi} = 5.8 \times 10^{14}$ cm^{-3}
$V_B = 600$ V
b device corresponding to region 2 of Fig. 3
$N_{epi} = 8.7 \times 10^{14}$ cm^{-3}
$V_B = 1100$ V
c device corresponding to region 3 of Fig. 3
$N_{epi} = 1.2 \times 10^{15}$ cm^{-3}
$V_B = 600$ V

resistivity change alone because the potential adopted by the *p*-type resurf layer, and hence the proportion of the total voltage being supported by the respective parts of the device, also changes with the epitax doping. The data presented in Fig. 3 relate to a device for which the depth of the p^+ diffusion and p^- epitax layer are equal. It is necessary to ensure that the depth of the p^- layer is not significantly less than that of the p^+ diffusion for the method to be properly effective. Fig. 5 illustrates the consequences of a reduced epitax layer thickness on the breakdown voltage of a typical device optimised for bulk breakdown. If other considerations permit, the layer could be allowed to extend beyond the junction depth, resulting in the structure previously considered by Appels *et al.* [7]. However, apart from Fig. 5, all the data presented in this work apply to the condition $x_j = d_{epi}$.

The effect of surface charge will depend upon the asymmetry of the junction considered. For the p^+n device, positive surface charges will lower the breakdown voltage when breakdown occurs at the p^+p^- interface (low N_{epi} values) and raise the voltage when located at the p^-n^+ junction (high N_{epi} values). As a consequence, the optimum epitax doping range is shifted to higher values as shown in Fig. 3. For n^+p^- structures, the opposite will occur and the optimum epitax doping range will be moved to lower values by the surface charge. The results shown in Fig. 3 relate to the p^+n configuration and show that, if high breakdown voltages are to be realised, the epitax doping specification must not only be tightly controlled but also selected according to the surface charge density present.

The extent to which the observed sensitivity to surface charge can be eased by using deeper diffusions is shown in Fig. 6. The upper and lower limits to the range of N_{epi} values for which breakdown is confined to the bulk have been identified and plotted against $1/x_j$ for each of two

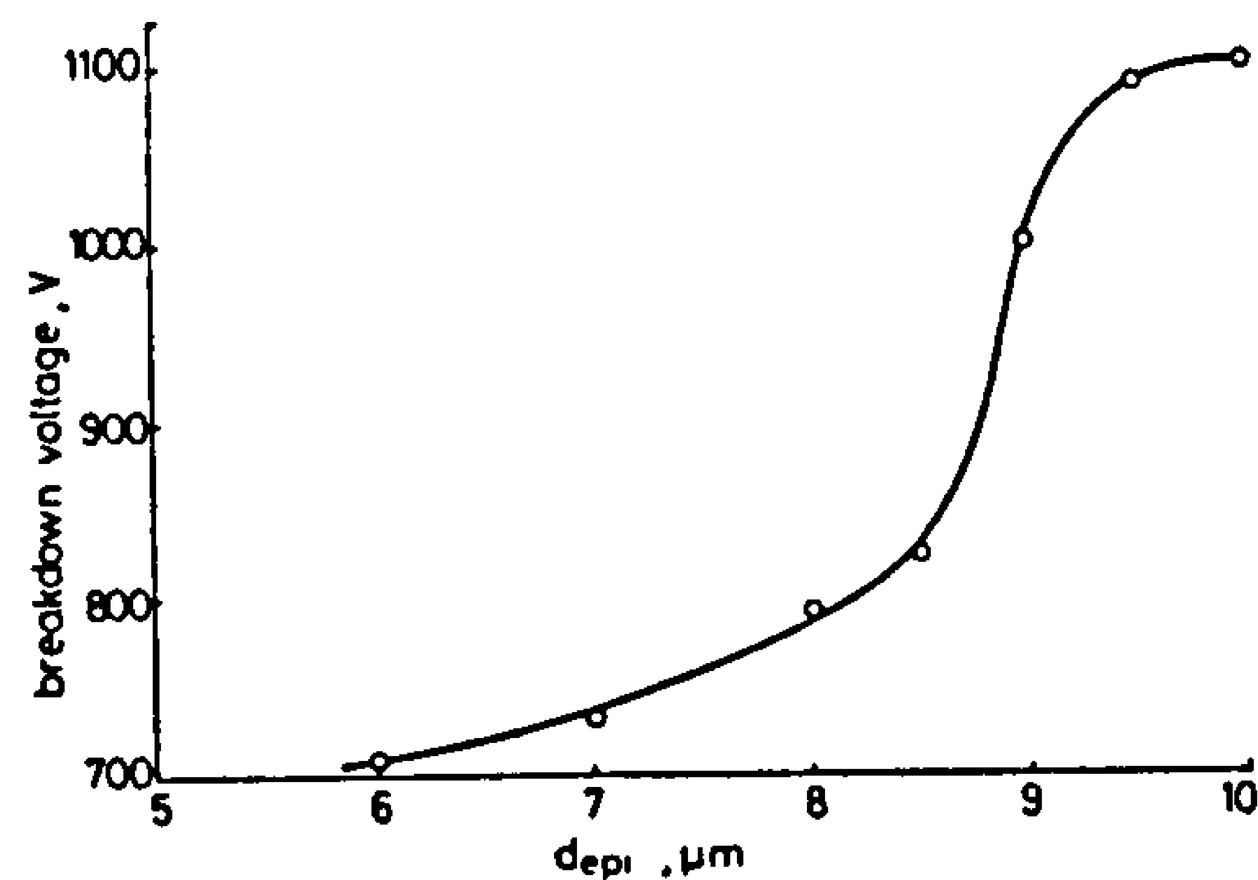


Fig. 5 *Breakdown voltage versus p^- resurf layer depth (d_{epi})*
$N_s = 5 \times 10^{11}$ cm^{-2}, $x_j = 10$ µm, $N_{epi} = 8.7 \times 10^{14}$ cm^{-3}, $N_{sub} = 10^{14}$ cm^{-3}

values of surface charge density, $N_s = 0$ and 5×10^{11} charges/cm^2 respectively. The breakdown voltage achieved

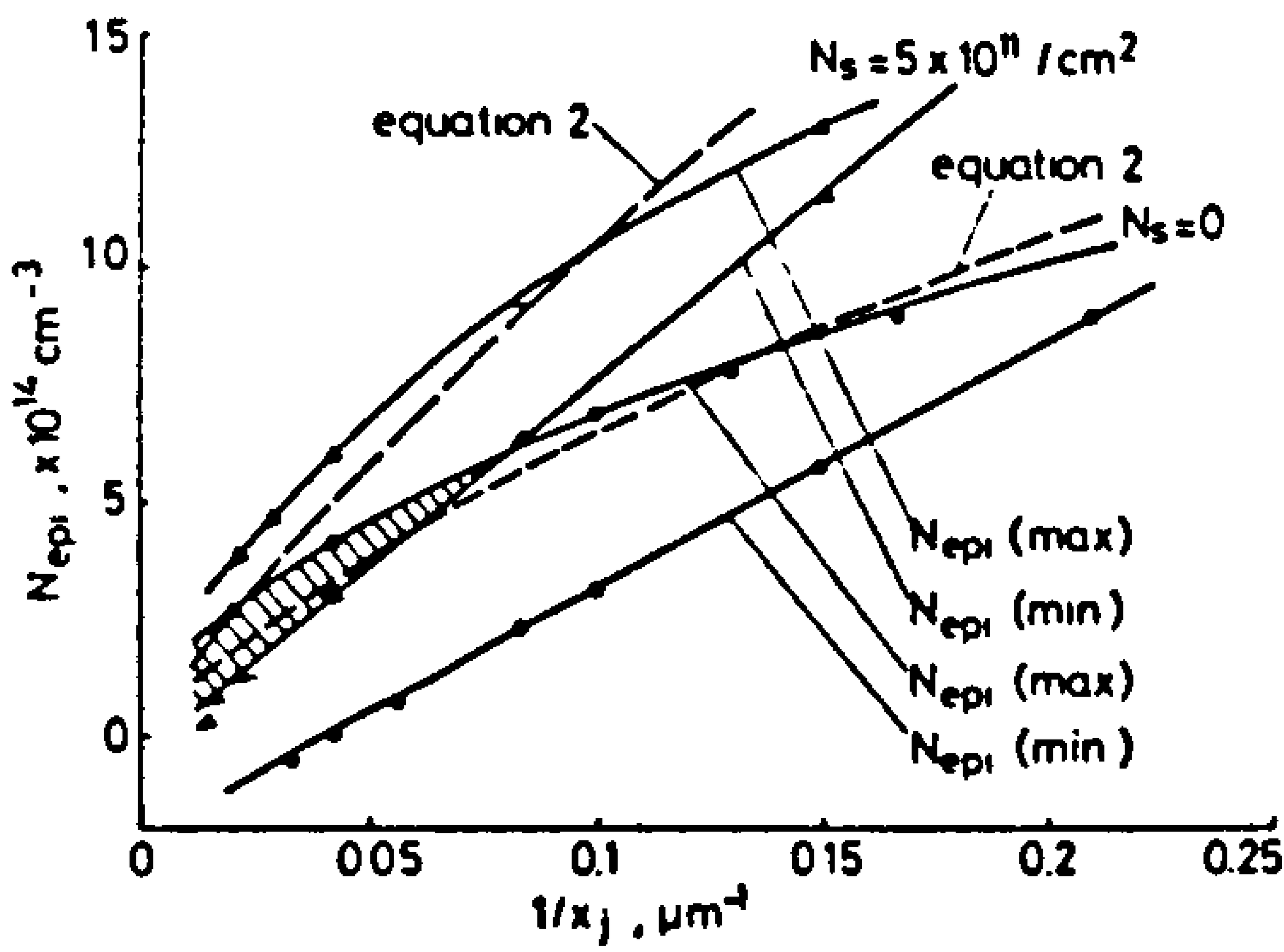


Fig. 6 *Upper and lower limits for N_{epi} that ensure bulk breakdown, against reciprocal junction depth*
$N_s = 0$ and 5×10^{11} cm^{-2}, $N_{sub} = 10^{14}$ cm^{-3}, $d_{epi} = x_j$

within these ranges of N_{epi} rises with both N_{epi} and x_j as an increase in either parameter acts to reduce the divergence of flux in the vicinity of the curved p^+ boundary, which is where the breakdown is located. The design that results in maximum breakdown voltage therefore corresponds to the upper limits of N_{epi} shown in Fig. 6 for each value of surface charge density and x_j. However, the breakdown voltage variation across the range shown for N_{epi} is not usually great, and above a junction depth of approximately 12 μm the bands overlap, revealing combinations of N_{epi} and x_j values for which the breakdown location will be relatively independent of surface charge density. The appropriate region is shown shaded in Fig. 6.

The results have been extended to include lower voltage structures as shown in Figs. 7 and 8. In each case, the

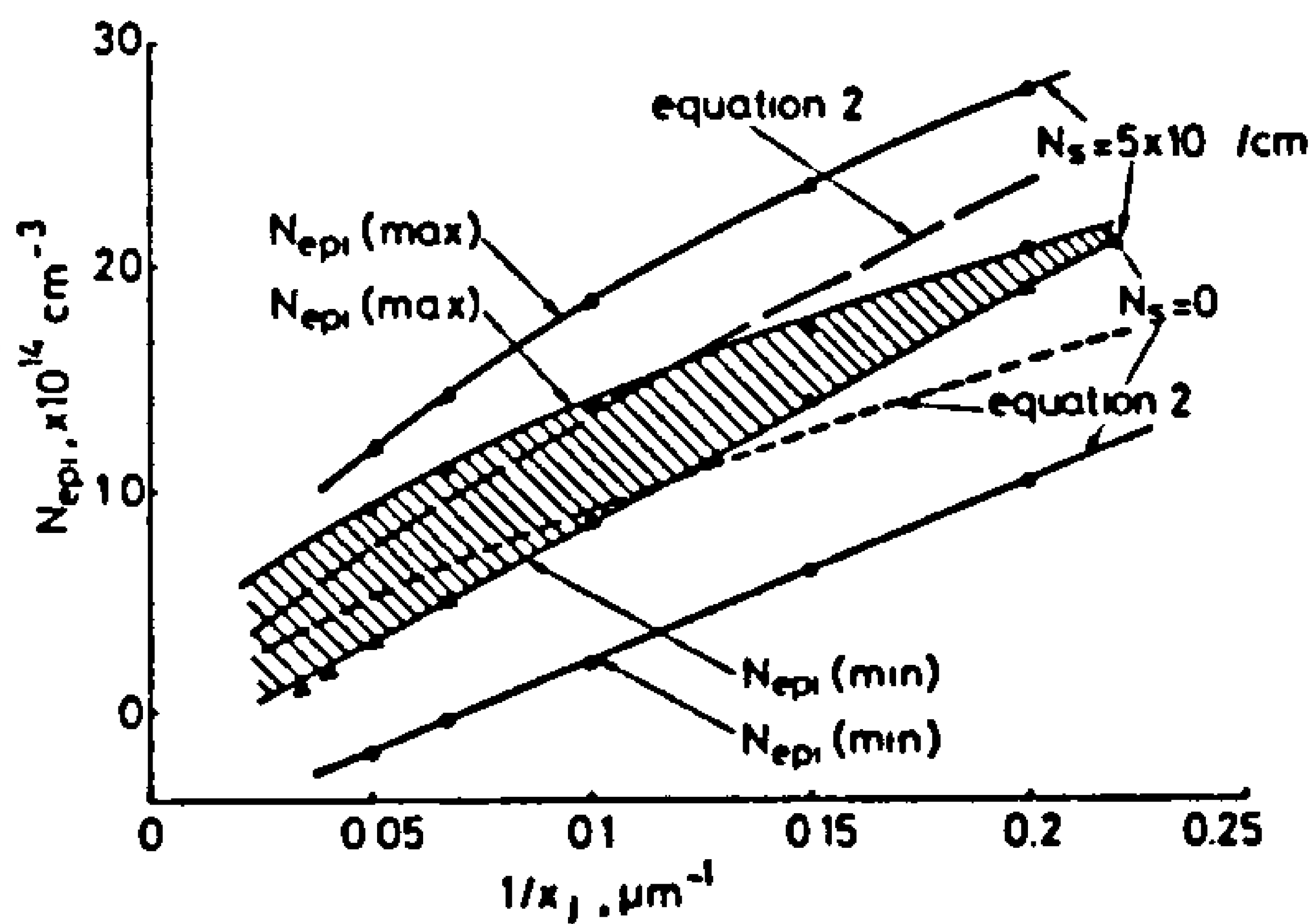


Fig. 7 *Upper and lower limits for N_{epi} that ensure bulk breakdown, against reciprocal junction depth*
$N_s = 0$ and 5×10^{11} cm^{-2}, $N_{sub} = 3 \times 10^{14}$ cm^{-3}, $d_{epi} = x_j$

lower limit of N_{epi} for which breakdown occurs in the bulk agrees well with an equation of the form

$$N_{epi}(min) = \frac{A}{x_j} + B$$

and the upper limit may be described by the equation

$$N_{epi}(\max) = C x_j^D$$

Appropriate values for A, B, C and D are given in Table 1 and these equations are represented by the solid lines in Figs. 6, 7 and 8.

An expression that allows an estimate for N_{epi} to be obtained, given values for d_{epi} and N_{sub}, has been derived analytically by Appels *et al.* The analysis proceeds on the

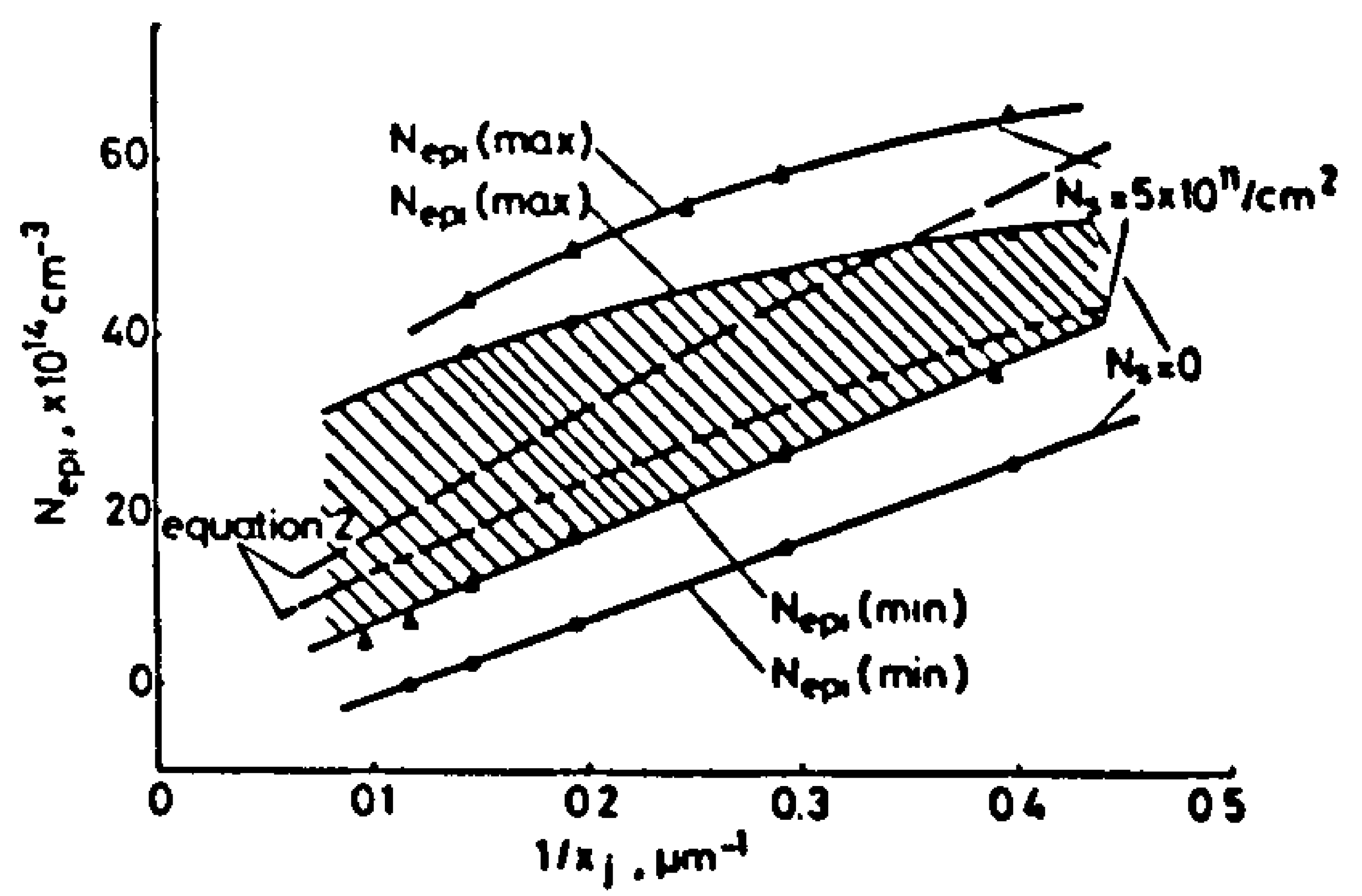


Fig. 8 *Upper and lower limits for N_{epi} that ensure bulk breakdown, against reciprocal junction depth*
$N_s = 0$ and 5×10^{11} cm^{-2}, $N_{sub} = 10^{15}$ cm^{-3}, $d_{epi} = x_j$

basis that the p^- resurf layer should be allowed to become fully depleted before breakdown occurs at the vertical p^-n^+ junction. In the presence of surface charge it is necessary to extend this concept to allow for surface depletion. Provided that the length of the resurf layer is sufficient, the electric field distribution near the centre of the layer may be considered to be essentially 1-dimensional and normal to the surface. Under these conditions, the undepleted depth of the resurf layer will be reduced by an amount N_s/N_{epi}, where N_s is the density of surface charges per unit area. The relationship then becomes

$$N_{epi}(d_{epi} - N_s/N_{epi}) < \frac{\varepsilon_0 \varepsilon_{si} E_c}{q(1 + N_{epi}/N_{sub})^{1/2}} \quad (2)$$

This expression requires an estimate for E_c, the critical field for breakdown, which also depends on the doping level (N_{epi} in this case). The relation

$$E_c = 2.75 \times 10^3 \times N_i^{0.135} \text{ V/cm} \quad (3)$$

conforms to the peak field at breakdown for the computed data of Fig. 2 and has been used in eqn. 2 to determine the relationship between N_{epi} and d_{epi} included in Figs. 6, 7 and 8 and for which $d_{epi} = x_j$. Although the equation does not provide an optimum design, it does ensure bulk breakdown for a wide range of junction depths and makes an appropriate allowance for surface charge. The general validity of the results obtained by the equation is good, particularly in view of approximations involved in its deri-

Table 1: Coefficients used in the empirical relationships for N_{epi}(max) and N_{epi}(min).

N_{sub}, cm^{-3}	A	B	C	D
		$N_s = 0$		
10^{14}	5.4×10^{11}	-2.2×10^{14}	1.1×10^{13}	-0.57
3×10^{14}	8.2×10^{11}	-6.0×10^{14}	2.7×10^{13}	-0.57
10^{15}	9.1×10^{11}	-1.1×10^{15}	3.6×10^{14}	-0.32
		$N_s = 5 \times 10^{11}$ cm^{-2}		
10^{14}	8.1×10^{11}	-0.4×10^{14}	1.6×10^{13}	-0.60
3×10^{14}	1.1×10^{12}	-2.3×10^{14}	2.9×10^{13}	-0.60
10^{15}	1.0×10^{12}	-3.5×10^{14}	2.8×10^{14}	-0.38

The dimensions for N_{epi} and x_j are cm^{-3} and cm, respectively.

vation, but some reservations on its value when applied to shallow junctions in high resistivity materal are apparent.

The results of Figs. 6, 7 and 8 show that, for the normal range of surface charge densities, it is possible to identify certain values for N_{epi} and x_j for which breakdown is confined to the bulk, regardless of the precise density of surface charge present. However, for medium- to high-voltage devices the appropriate ranges become very restrictive and require progressively deeper junctions. The prospect of alleviating this difficulty by the use of field plates to control the influence of surface charge on the breakdown location has also been investigated. The previous structure was modified by the addition of field plates connected to the p^+ and n^+ diffusions and extended across the p^- resurf layer for a distance of 150 μm over the 1 μm oxide layer. The consequences of these additions on the breakdown voltage for a device with substrate doping of 10^{14} cm^{-3} and surface charge density $N_s = 5 \times 10^{11}$ charges/cm^2 are shown in Fig. 9. The voltage corresponding to $N_{epi} = -10^{14}$ cm^{-3} provides the breakdown of a

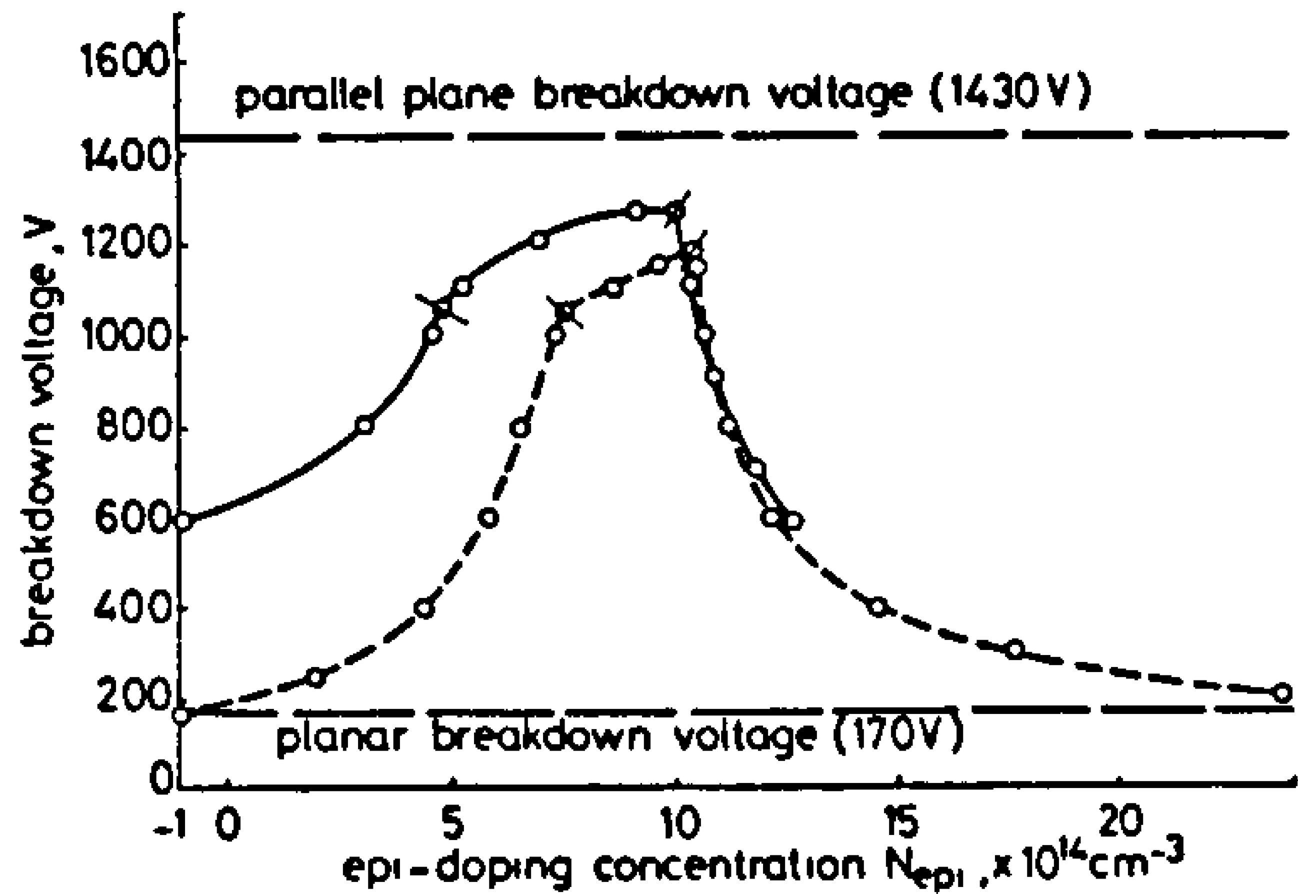


Fig. 9 *Breakdown voltage against impurity concentration in the p^- epitaxial resurf layer, illustrating the effect on the breakdown voltage of adding field plates to the basic structure shown in Fig. 2*

$x_j = d_{epi} = 10$ μm, $N_{sub} = 10^{14}$ cm^{-3}, $N_s = 5 \times 10^{11}$ cm^{-2}
—— without field plate
- - - with field plate

planar diode in the absence of a resurf layer. Although the field plates by themselves make a significant contribution towards raising the planar breakdown voltage, it is still well below that made possible by the addition of a resurf layer of appropriate specification. The field plates raise the voltage when breakdown is located near the p^+ diffusion, but an improvement when located at the n^+ diffusion only becomes apparent with shallower junctions than that shown in Fig. 9. The effect of the field plates is summarised in Fig. 10 where the range of N_{epi} values that ensure bulk

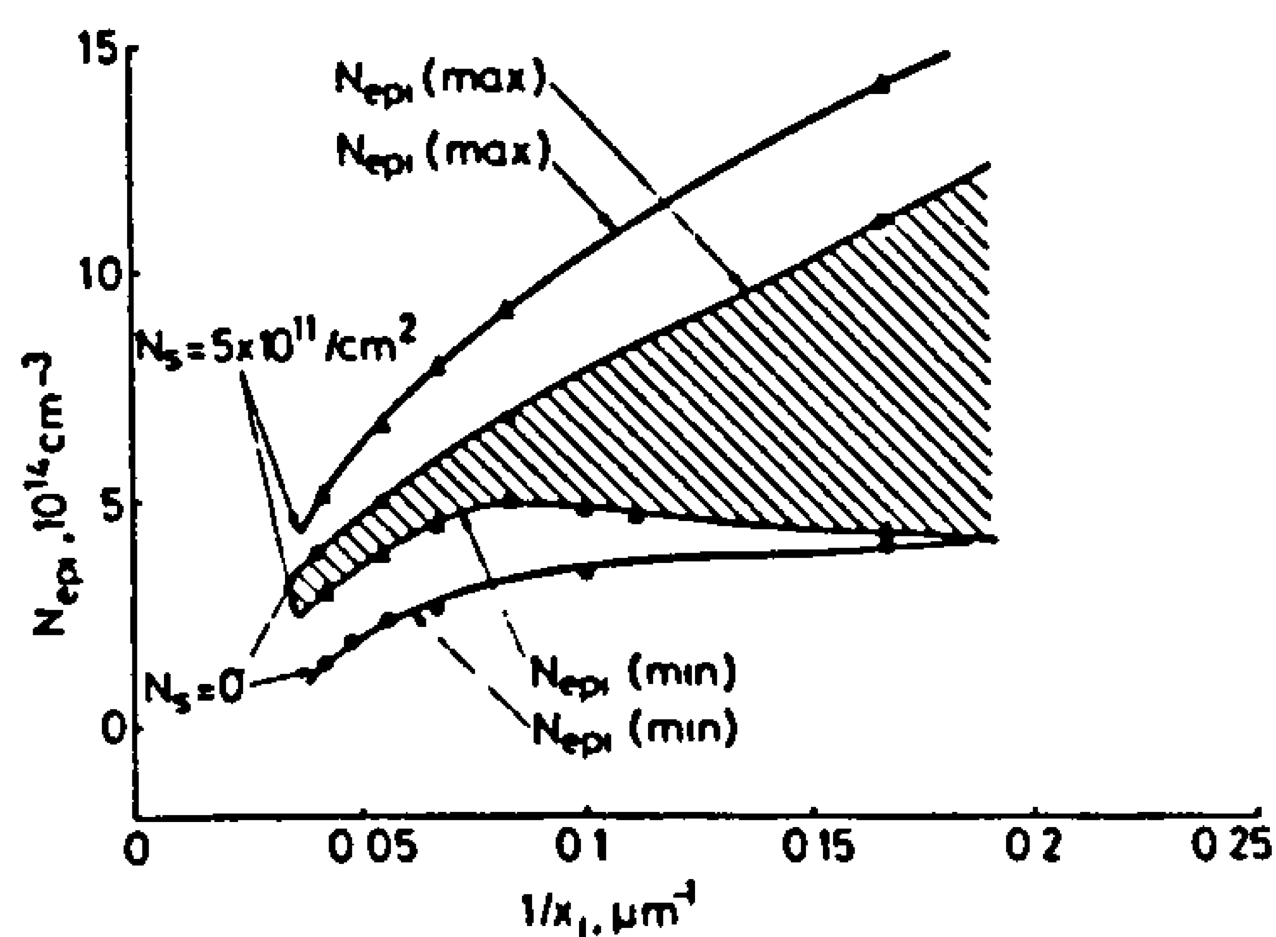


Fig. 10 *Upper and lower limits for N_{epi} that ensure bulk breakdown for a resurf structure with field plates, against reciprocal junction depth*

$N_s = 0$ and 5×10^{11} cm^{-2}, $N_{sub} = 10^{14}$ cm^{-3}, $d_{epi} = x_j$

breakdown have been plotted as a function of x_j^{-1}. The field plates are particularly effective in shallow-diffused, high-voltage resurf structures as a means of ensuring bulk breakdown despite the presence of a wide range of surface charge densities.

The improvement in breakdown voltage that can be achieved by the addition of an optimally designed resurf layer is shown in Fig. 11, and may be compared with that

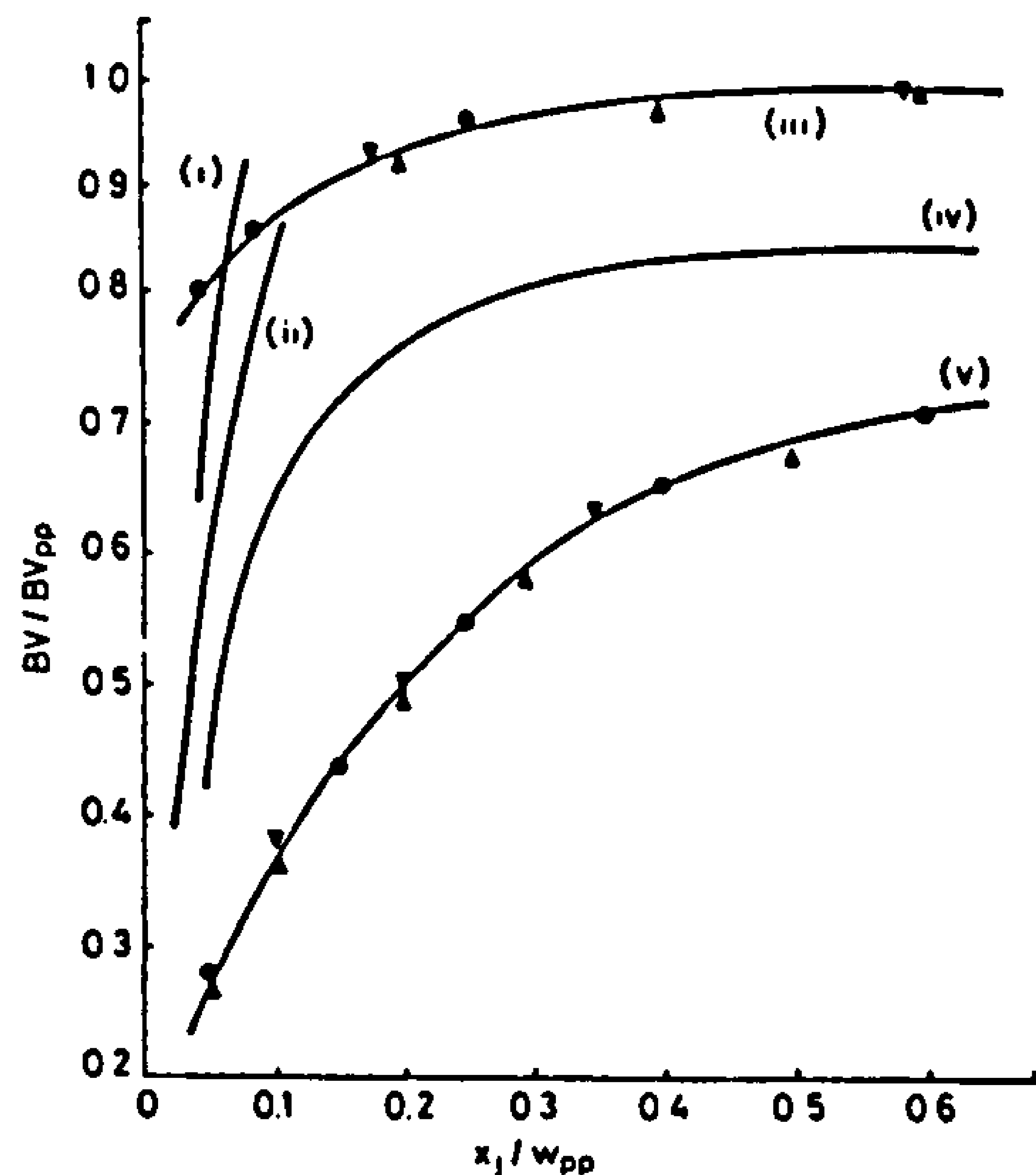


Fig. 11 *Normalised breakdown voltage against normalised junction depth for an unprotected planar junction and an optimised resurf structure*

Also included for comparison are the corresponding data for field ring structures. The multiple ring data were presented by Brieger et al [6] from an analytical solution which has a limited validity range of x_j/W_{pp} ratios
(i) Triple field ring (Brieger et al [6])
(ii) Double field ring (Brieger et al [6])
(iii) 'Resurf' layer
(iv) Single field ring (Adler et al [5])
(v) Planar junction
● $N_B = 1 \times 10^{14}$ cm^{-3}
▲ $N_B = 3 \times 10^{14}$ cm^{-3}
▼ $N_B = 1 \times 10^{15}$ cm^{-3}

obtained using both single and multiple field limiting rings. The multiple ring data has been calculated using the appropriate design guidelines established by Brieger *et al.* [6], the validity of which are limited to the range x_j/w_{pp} shown. The resurf technique is able to achieve breakdown voltages that compare favourably with those obtained using ring structures. However, a disadvantage of ring structures not apparent in Fig. 11 is the sensitivity of the design to surface charge density, the compensation for which requires a reduction in ring spacing and an increase in the number of rings. A computer solution for the single ring structure was undertaken to demonstrate the effect and a typical example is shown in Fig. 12. The upper graph on the left-hand side of the Figure shows the breakdown voltage as a function of junction to ring spacing for a 30 μm diffused junction in the absence of surface charge. The optimum spacing of 37 μm is critical and produces a peak breakdown voltage that represents little more than 70% of the plane parallel voltage. The addition of surface charge reduces the breakdown voltage achieved at that ring spacing to such an extent that for $N_s = 5 \times 10^{11}$ charges/cm^2, the ring has little beneficial effect. A reduction in junction to ring spacing down to 26 μm is required to restore an optimum design, but the breakdown voltage remains significiantly reduced. These results may be contrasted with the behaviour of an identical junction protected by the resurf technique for the same variation of

surface charge, shown on the right hand side of Fig. 12. Breakdown voltages greater than 95% of BV_{pp} can be

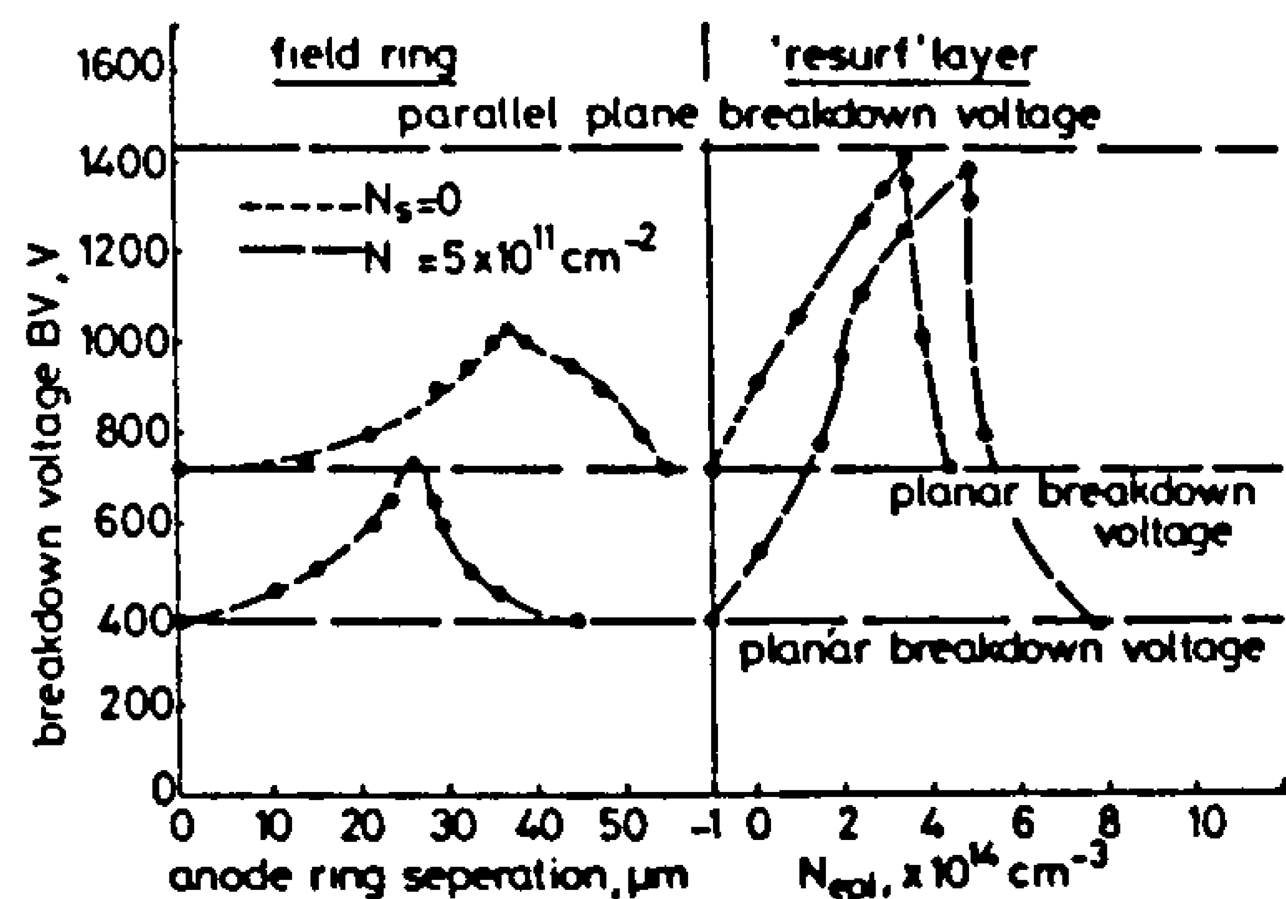


Fig. 12 *Breakdown voltage for a planar diode with a single field limiting ring compared to that of the same diode employing a resurf layer, as a function of ring separation and resurf layer doping concentration, respectively*

$N_s = 0$ and 5×10^{11} cm^{-2}, $N_{sub} = 10^{14}$ cm^{-3}, $x_j = 30$ μm

achieved with and without surface charge, provided that appropriate adjustments are made to the doping level of the resurf layer. Even in the absence of any such adjustment, the breakdown voltage remains well above the best that can be achieved with a single ring system.

4 Conclusions

The results of a 2-dimensional numerical analysis of p^+n diffused diode structures with resurf protection have been presented. The allowed tolerance in the resurf layer doping concentration that ensures bulk breakdown has been obtained over a range of junction depths, and the consequences of surface charge quantified. The work has shown that certain combinations of values for the resurf doping concentration and depth result in bulk breakdown for a wide range of surface charge densities. However, a certain minimum junction depth is required that becomes particularly restrictive in medium- to high-voltage structures. This problem can be eased by the use of field plates that also serve to stabilise charge migration over the surface. A brief comparison with field limiting ring structures applied under similar circumstances demonstrates the superiority of the resurf technique, particularly in the presence of surface charge.

5 Acknowledgment

The authors wish to acknowledge the continued interest and related discussions held with Dr. J. Turner and Mr. N. Harrison of Ferranti Electronics plc, Discrete Transistor Division. This work forms part of a project funded by the UK Science & Engineering Research Council.

6 References

1 DAVIES, R.L., and GENTRY, F.E.: 'Control of electrical field at the surface of p-n junctions', *IEEE Trans.*, 1964, ED-11, pp. 313-321
2 GHANDI, S.K.: 'Semiconductor power devices' (Wiley Interscience, New York, 1977)
3 CONTI, F., and CONTI, M.: 'Surface breakdown in silicon planar diodes equipped with field plate', *Solid-State Electron.*, 1972, 15, pp. 93-105
4 KAO, Y.C., and WOLLEY, E.D.: 'High-voltage planar p-n junctions', *Proc. IEEE*, 1967, 55, (8)
5 ADLER, M.S., TEMPLE, V.A.K., FERRO, A.P., and RUSTAY, R.C.: 'Theory and breakdown voltage for planar devices with a single field limiting ring', *IEEE Trans.*, 1977, ED-24, (2), pp. 107-113
6 BRIEGER, K.P., GERLACH, W., and PELKA, J.: 'Blocking capability of planar devices with field limiting rings', *Solid-State Electron.*, 1983, 26, (8), pp. 739-745
7 APPELS, J.A., COLLET, M.G., HART, P.A.H., VAES, H.M.J., and VERHOEVEN, J.F.C.M.: 'Thin layer high-voltage devices (resurf devices)', *Philips J. Res.*, 1980, 35, (1), pp. 1-3
8 YASUDA, S., and KURATA, M.: 'Two-dimensional field distribution analysis of reverse biased pn junction devices' *Solid-State Electron.*, 1980, 23, pp. 1077-1084
9 HWANG, K., and NAVON, D.H.: 'Breakdown voltage optimisation of silicon p-π-ν planar junction diodes', *IEEE Trans.*, 1984, ED-31, (9), pp. 1126-1135
10 KOKOSA, R.A., and DAVIES, R.L.: 'Avalanche breakdown of diffused silicon pn junctions', *ibid.*, 1966, ED-13, (12), pp. 874-881
11 MILLER, S.L.: 'Ionisation rates for holes and electrons in silicon', *Phys. Rev.*, 1957, 105, pp. 1246-1249
12 McKAY, K.G.: 'Avalanche breakdown in silicon', *ibid.*, 1954, 94, pp. 877-884
13 VAN OVERSTRAETEN, R., and DE MAN, H.: 'Measurement of the ionization rates in diffused silicon p-n junctions', *Solid-State Electron.*, 1970, 13, pp. 583-608
14 DEAL, B.E., SKLAR, M., GROVE, A.S., and SNOW, E.H.: 'Characteristics of the surface-state charge (Qss) of thermally oxidized silicon', *J. Electrochem. Soc.*, 1967, 114, p. 226

Optimisation of VDMOS power transistors for minimum on-state resistance

J.T. Davies
P. Walker
K.I. Nuttall

Indexing terms: Transistors, Optimisation, Metal-oxide-semiconductor structures

Abstract: A 2-dimensional numerical simulation program has been applied to the power VDMOS structure to determine design guidelines for minimum specific on-state resistance subject to a given breakdown voltage requirement. The entire cell has been modelled to take full account of contributions from the inversion and accumulation layers as well as the effect of cell spacing on the breakdown voltage. Optimised cell width, epitaxial thickness and doping concentration are presented for a range of breakdown voltages. The results show that the optimum body diffusion spacing increases with the breakdown voltage rating up to approximately 400 V, giving an approximately linear relationship between the on-state resistance-area product and breakdown voltage for a constant body width of 15 μm.

1 Introduction

The on-state resistance of the VDMOS transistor is its most serious limitation when assessed against the performance offered by a bipolar transistor of comparable voltage rating. It is of special concern in low-frequency (<20 kHz) power applications, for which the on-state losses become dominant. The desire to minimise these losses has led to several previous studies [1–4] that have identified three principal contributions to the total resistance, all of which are dependent upon the bias conditions (Fig. 1). The first component, the inversion channel resistance, is sensitive to the gate voltage. The second, the surface accumulation layer resistance, is dependent on both the gate and drain voltages, and the third, the drain resistance, is affected by the drain voltage through parasitic JFET action produced by the body diffusions. To minimise the effect of the bias conditions and allow an assessment of the basic cell design, it is normal to evaluate and compare the minimum resistance produced under conditions of very low drain voltage and high gate voltage. Even under these conditions, it has been found that the channel and accumulation layer resistances represent a significant proportion of the total resistance in devices designed for low voltage applications (<150 V). In contrast, high-voltage devices are dominated by the parasitic resistance of the drain.

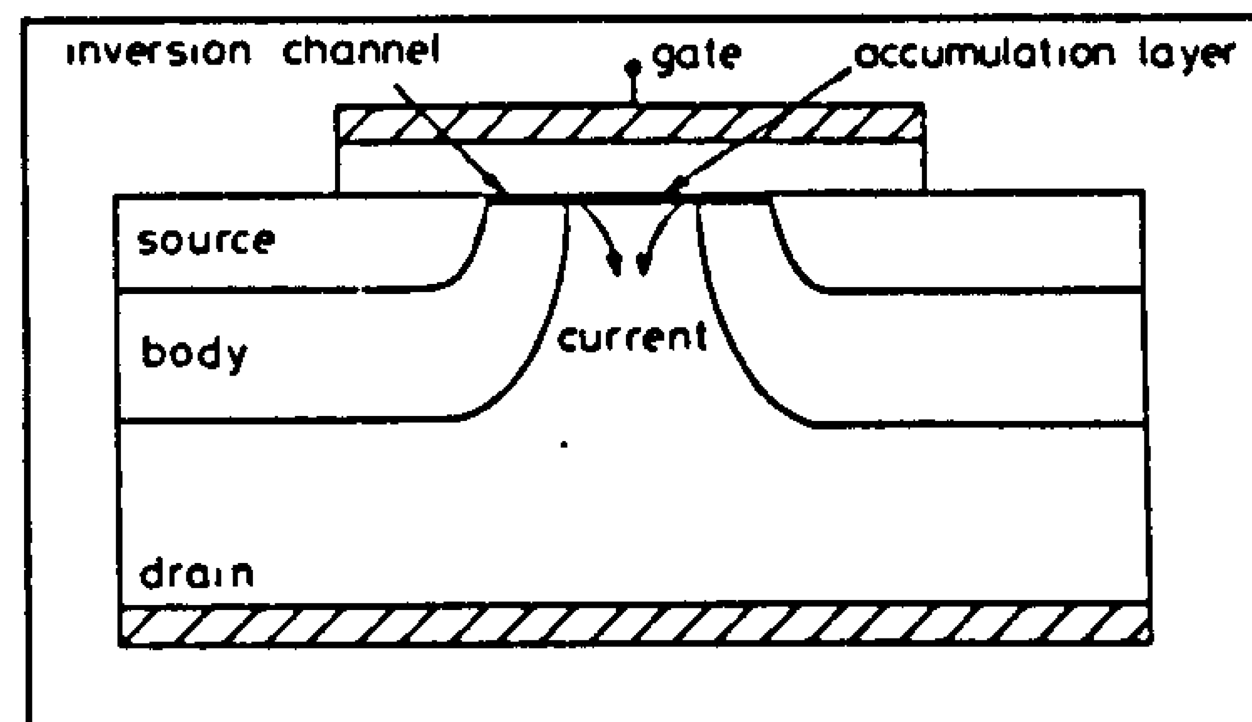


Fig. 1 *Cross-section of the VDMOS transistor*

The earlier work has ranged from an analytical approach [1] through to a 2-dimensional numerical analysis confined to the drain region [4, 7]. Although this earlier work produces a valuable insight into the origin and relative magnitude of the various components comprising the total resistance, much of it has been concerned with non-optimised devices. Tamar *et al.* [2] performed a 2-dimensional numerical analysis of both breakdown voltage and on-state resistance for VDMOS, VMOS and UMOS devices with breakdown voltages of 100, 550 and 1000 V. However, the effect of the surface accumulation layer on the current distribution in the drain region was ignored in the VDMOS structure, which has been shown to be in error [3]. The main value of the work was in the comparison between the on-state resistances of the three types of structure studied, although it is noted that the cell widths used were chosen on an apparently arbitrary basis and no attempt was made to optimise each structure in this respect. Byrne and Board [3] have adopted a more satisfactory approach, but the effect of cell halfwidth on breakdown voltage was neglected. Their choices of epitaxial layer thickness and doping concentrations were based on the specifications given by Tamar *et al.*, which relate to a constant nonoptimised cell spacing. In their later work [4] they adopted values based on the planar breakdown voltage with a correction for gate overlap effects, but gave little guidance on the optimum design for different breakdown voltages. A still later contribution [7] identified the importance of this effect, and presented an optimised design for a breakdown voltage of 1000 V, but again based on a simplified model with the solution confined to the drain region.

Paper 5361I (E3), first received 28th November 1986 and in revised form 17th February 1987

P. Walker and K.I. Nuttall are, and J.T. Davies was formerly, with the Department of Electrical Engineering and Electronics, University of Liverpool, Brownlow Hill, PO Box 147, Liverpool L69 3BX United Kingdom

J.T. Davies is now with British Aerospace, Preston, Lancs, United Kingdom

This work is concerned with the medium voltage range over which all conducting regions of the device can be expected to contribute to the on-state resistance. A full solution of the entire cell has therefore been used to provide a more rigorous assessment of the optimum structure in the voltage range where cell spacing is important. Full account is taken of cell width, epitaxial layer thickness and resistivity on both the on-state resistance and breakdown voltage so that the true optimum structure is identified for a range of breakdown voltages. Design guidelines are provided for the appropriate cell spacing and epitaxial layer specifications, although in practice the breakdown voltage may be established by an additional diffusion through the central portion of the body diffusion. This prevents lateral body current from activating the parasitic *npn* transistor produced by the source, body and drain which would otherwise reduce the breakdown voltage to its BV_{ceo} value. An appropriate design procedure in this case would be to optimise the device first without the additional diffusion, but for a slightly higher breakdown voltage than ultimately required, and to subsequently add the controlling diffusion to produce the desired breakdown voltage. In this way a properly optimised structure should result with the guarantee of breakdown in the plane part of the junction. The work assumes that appropriate field termination measures are taken to prevent premature breakdown at the edges of the device.

The optimisation procedure adopted here is similar to that of Tamar *et al.* except that cell width has been included as a variable and the results obtained from a more comprehensive numerical treatment of the device. Specific resistance values have been calculated on the assumption of a linear cell geometry. The geometric advantages of cellular designs will produce lower values.

2 Results and discussion

The analysis is based on a 5-point finite difference computer program that solved both the Poisson and electron current continuity equations for the entire device. The grid was selectively refined in critical areas to ensure computational accuracy. Particular attention was necessary in the region of the surface inversion and accumulation layers. The breakdown voltages were obtained using the ionisation coefficients presented by Van Overstraeten and DeMan [5], and surface mobilities from results obtained by Sun and Plummer [6]. Device symmetry allows the analysis to be confined to the half cell shown in Fig. 2, which also defines the terms used here.

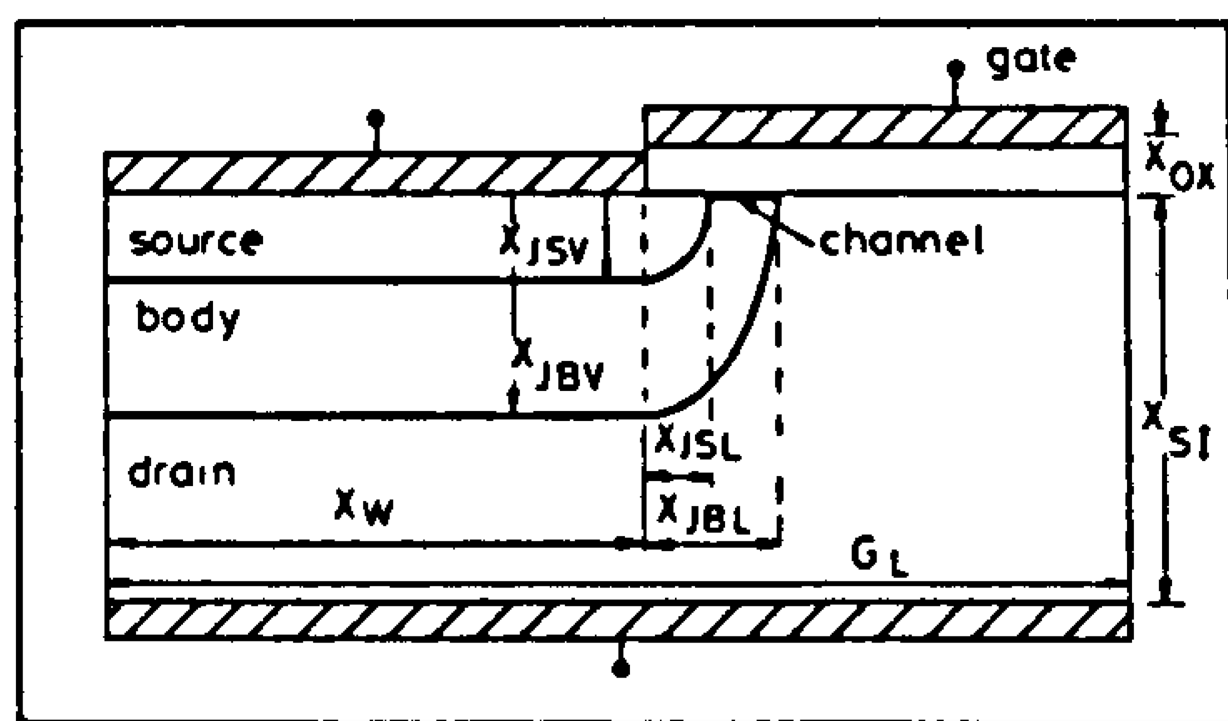


Fig. 2 *Half-cell cross-section used in the simulation program*
The vertical edges mark lines of symmetry in the full transistor cell.

The simulation was first verified by a comparison with measurements obtained on nominal 30 V and 90 V devices (Fig. 3), the specification is included in Table 1. As the on-state resistance of low voltage devices includes a significant component due to the surface accumulation and inversion layers, this comparison represents a more

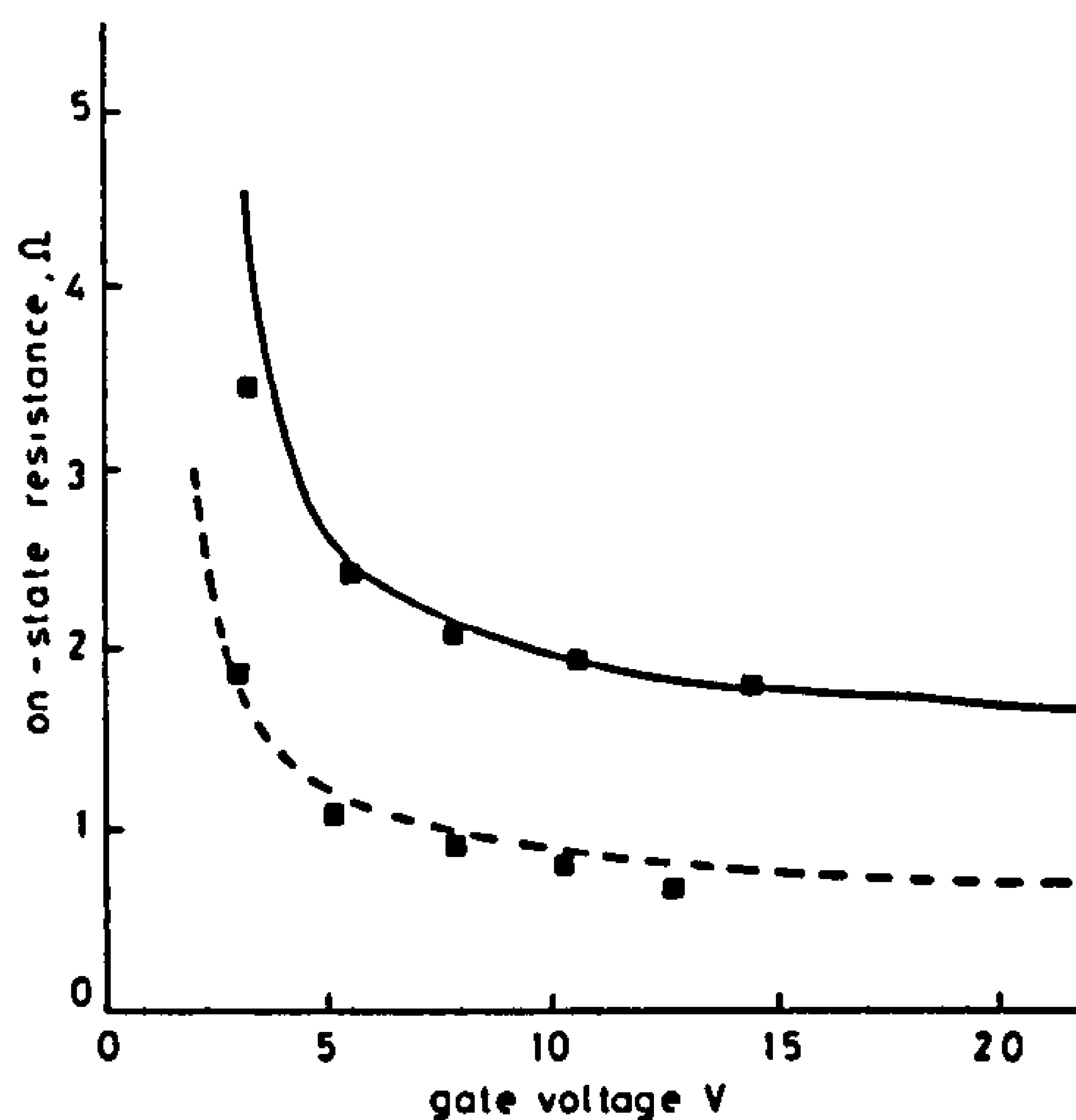


Fig. 3 *On-state resistance as a function of gate voltage for two VDMOS transistors used for confirmation of the computer simulation*
■ simulation
—— 90 V } experiment
- - - - 30 V } experiment

Table 1 : Device parameters

<table>
<tr><th></th><th>30 V device</th><th>90 V device</th></tr>
<tr><td>X_{SI}, μm</td><td>11.0</td><td>15.5</td></tr>
<tr><td>N_{DO}, cm^{-3}</td><td>11.5×10^{15}</td><td>2.2×10^{15}</td></tr>
<tr><td>G_L, μm</td><td>13.65</td><td>14.65</td></tr>
<tr><td>X_{JBV}, μm</td><td>4.82</td><td>5.4</td></tr>
<tr><td>X_{JBL}, μm</td><td>2.64</td><td>3.57</td></tr>
<tr><td>X_{JSV}, μm</td><td colspan="2">1.55</td></tr>
<tr><td>X_{JSL}, μm</td><td colspan="2">1.32</td></tr>
<tr><td>X_{ox}, nm</td><td colspan="2">89</td></tr>
<tr><td>X_w, μm</td><td colspan="2">7.4</td></tr>
<tr><td>gate width, cm</td><td colspan="2">3.76</td></tr>
</table>

comprehensive test of the model than would be obtained on higher voltage structures in which the resistance is dominated by that of the drain region. The on-state resistances shown in Fig. 3 are the asymptotic values obtained as the drain voltage approaches zero and demonstrate that the contribution from the surface layers persist to high gate voltages. In this work, as in others, the representative limiting on-state resistance is taken to be that obtained at $V_G - V_T = 15$ V. The results also illustrate the strong dependence of on-state resistance on the drain doping level, which becomes even more pronounced in higher-voltage devices. This is attributed to the increased resistance of the lightly doped drain, but with a compensatory effect due to current spreading from the surface accumulation layer, which has been shown to take place over a longer distance in high-resistivity material [4].

The foregoing illustrates the problem of simultaneously achieving low forward voltage drop and a significant reverse voltage capability in VDMOS structures. Apart form considerations relating to the epitaxial layer specification, the resistance of a VDMOS transistor cell can be reduced by increasing the separation between the body diffusions. However, this increases the width of the cell and unless the resistance is reduced by proportion-

ately more than the device area increases, the resistance per unit area will increase. At very small body spacings, the current from the surface inversion layer is injected approximately uniformly into the drain material by the accumulation layer (Fig. 1), and the resistance varies approximately inversely with the body separation. For a fixed body diffusion width, the cell resistance therefore varies much faster than the rate at which the total cell area changes and consequently the on-state resistance-area product falls as the cell width is increased. However, at larger separations the rate at which the active area between the body diffusions increases starts to slow relative to the total cell area. Also the accumulation layer eventually becomes unable to maintain a uniform injection into the drain as its length increases so that still less benefit is to be obtained by extending the cell width further. The resistance per unit area therefore passed through a minimum at an optimum cell width and then begins to increase. The rate at which current is lost from the accumulation layer is greater for a high conductivity drain and thus the effective 'injection distance' is shorter. Consequently its effects become noticeable at smaller cell widths in low breakdown voltage devices and the optimum cell width is correspondingly smaller.

Because of the effects described above and also the interaction that takes place between the body depletion layers at high voltages, both the breakdown voltage and on-state resistance will, in general, be functions of gate width, drain epitaxial layer thickness and resistivity. The requirement is therefore to choose the combination of these three parameters that produces the lowest possible on-state resistance-area product, and at the same time achieving some target breakdown voltage. To satisfy this requirement, contours of constant breakdown voltage were first produced by variation of the drain epitaxial layer thickness over a range of epitaxial layer doping levels, with gate width as an incremented parameter. Values of on-state resistance were then calculated at a number of points along the constant breakdown voltage contours. The calculations were performed using the 2-dimensional model described above, incorporating Gaussian impurity profiles for the source and body diffusions. For a given process schedule, a change of drain doping also affects the net body doping concentration and diffusion depth. An allowance was made for these effects based on measurements taken on sectioned VDMOS devices. A factor of 0.7 was also found to be appropriate for the ratio of lateral/vertical diffusions. These and other details of the geometry chosen for simulation are summarised in Table 2.

Table 2: Device parameters used in the computer simulations

X_{JBV}	$= 76\ N_{DD}^{-0.076}$ μm (N_{DD} in cm^{-3})
X_{JBL}/X_{JBV}	$= X_{JSL}/X_{JSV} = 0.7$
X_{ox}	= 1000 A
X_w	= 7.5 μm
V_G	= 15 V
V	= 1 mV

Gaussian impurity profiles were assumed for both body and source diffusions.

A constant body diffusion width of 15 μm has been chosen, based on that of the test devices used to verify the model. It is very desirable to reduce this dimension to a minimum to make most use of the available chip area, consistent with an acceptable device yield. The benefits to be gained in this respect are greatest for low-voltage devices on account of the thinner optimum epitaxial layers used, which results in less efficient use of the material under the body diffusion. The optimisation of the design with respect to this parameter is based on technological considerations and is not considered an adjustable parameter for the purposes of this work.

Results have been obtained for devices with breakdown voltages of 100, 175, 260 and 350 V, of which Fig. 4 is representative. In this diagram, the solid lines represent contours of constant breakdown voltage and the discrete numbered points the appropriate on-state resistance-area products. For the purposes of presentation, values have only been assigned to those points that identify the minimum resistance location. At low values of doping concentration all the curves for a given breakdown voltage tend to converge and the voltage is determined solely by the thickness of the epitaxial layer. This arises because under these conditions the layer is fully depleted and breakdown occurs in the plane part of the junction beneath the body diffusion. As the doping level increases, the location of breakdown begins to move towards the curved portion of the body diffusion and the breakdown voltage becomes independent of the drain epitaxial layer thickness. The proximity of the adjacent body diffusions then becomes important. For small cell spacings, the depletion layers interact strongly and reduce the effect of junction curvature. Higher drain doping levels can then be used to achieve the same breakdown voltage. As the cell spacing is increased, the interaction reduces and eventually becomes insignificant. Breakdown is then determined by the curvature of the body diffusions and causes the constant breakdown voltage contours to bunch together.

The results for on-state resistance show that there is an optimum combination of cell width, epitaxial layer thickness and epitaxial layer doping concentration which produces the lowest on-state resistance for each breakdown voltage. The minimum resistance-area products and corresponding values for G_L, X_{SI} and N_{DD} are plotted in Figs. 5 and 6. Tolerances have been drawn on the results for G_L that correspond to the increments used for this parameter in the calculations. It is clear from these results that for design voltages of less than 400 V, the optimum cell size (G_L) is different for each breakdown voltage, the reduction being more significant than that suggested by Byrne and Board, and a feature that was omitted entirely from Tamer's work. The results of Fig. 5 show that the 5 μm body spacing half width used by Tamar is appropriate for a 100 V device, (corresponding to an approximate G_L value of 7.5 + 3.1 + 5 = 15.6 μm for the cell geometry modelled here), but higher voltages require larger values. A true optimum can therefore only be found using a procedure that takes into account all of the dependent variables.

There is some evidence that the optimum G_L value tends to saturate with increasing breakdown voltage. Inspection of Fig. 5 and comparison with the results of earlier work concerned with higher voltages suggests that a G_L value of 23 μm (corresponding to a body spacing of approximately 22 μm), is appropriate for devices with breakdown voltage rating in excess of 400 V. The value compares favourably with that deduced as optimum by Byrne and Board [4] for a 400 V device, but is larger than that indicated by the later work relating to a 1000 V device [7]. Fortunately, the resistance is insensitive to body spacing in this higher voltage range, so that such discrepancies should have little practical importance. This feature is convenient, however, because it allows a

single mask set to be used to provide a wide range of breakdown voltages. Different breakdown voltages can then be accommodated simply by appropriate adjustments to the values of X_{SI} and N_{DD}.

The observed dependence of the optimum X_{SI} and N_{DD} on drain voltage can be represented by

$$X_{SI} = 0.064\,(BV - V_{bod_s}) + X_{JBi}\ (\mu m)$$

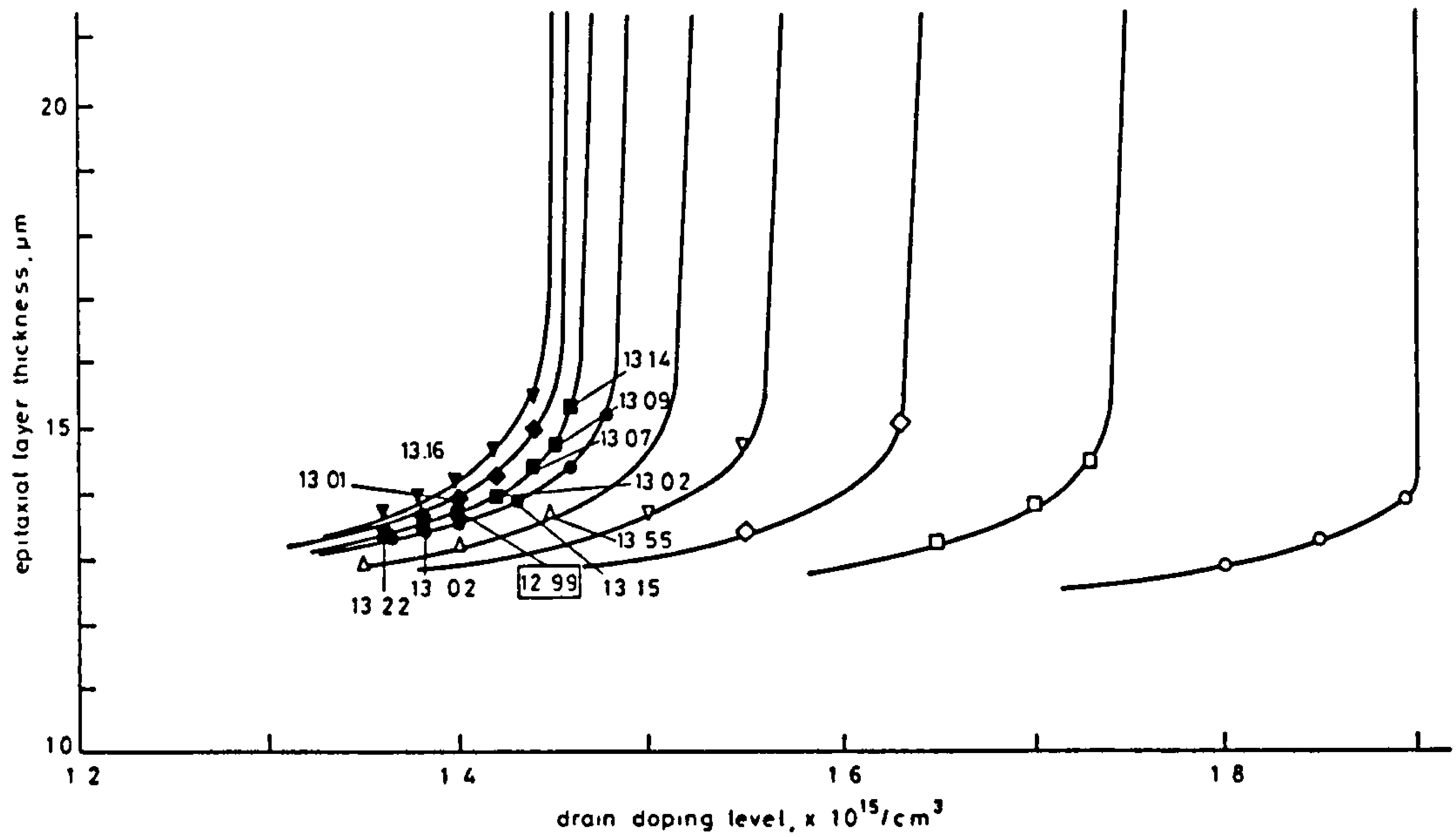


Fig. 4 *Constant 175 V breakdown contours and calculated on-state resistances (mΩ cm²)*

▼ $G_L = 20\ \mu m$ ▽ $G_L = 15\ \mu m$
◆ $G_L = 19\ \mu m$ ◇ $G_L = 14\ \mu m$
■ $G_L = 18\ \mu m$ □ $G_L = 13\ \mu m$
● $G_L = 17\ \mu m$ ○ $G_L = 12\ \mu m$
△ $G_L = 16\ \mu m$

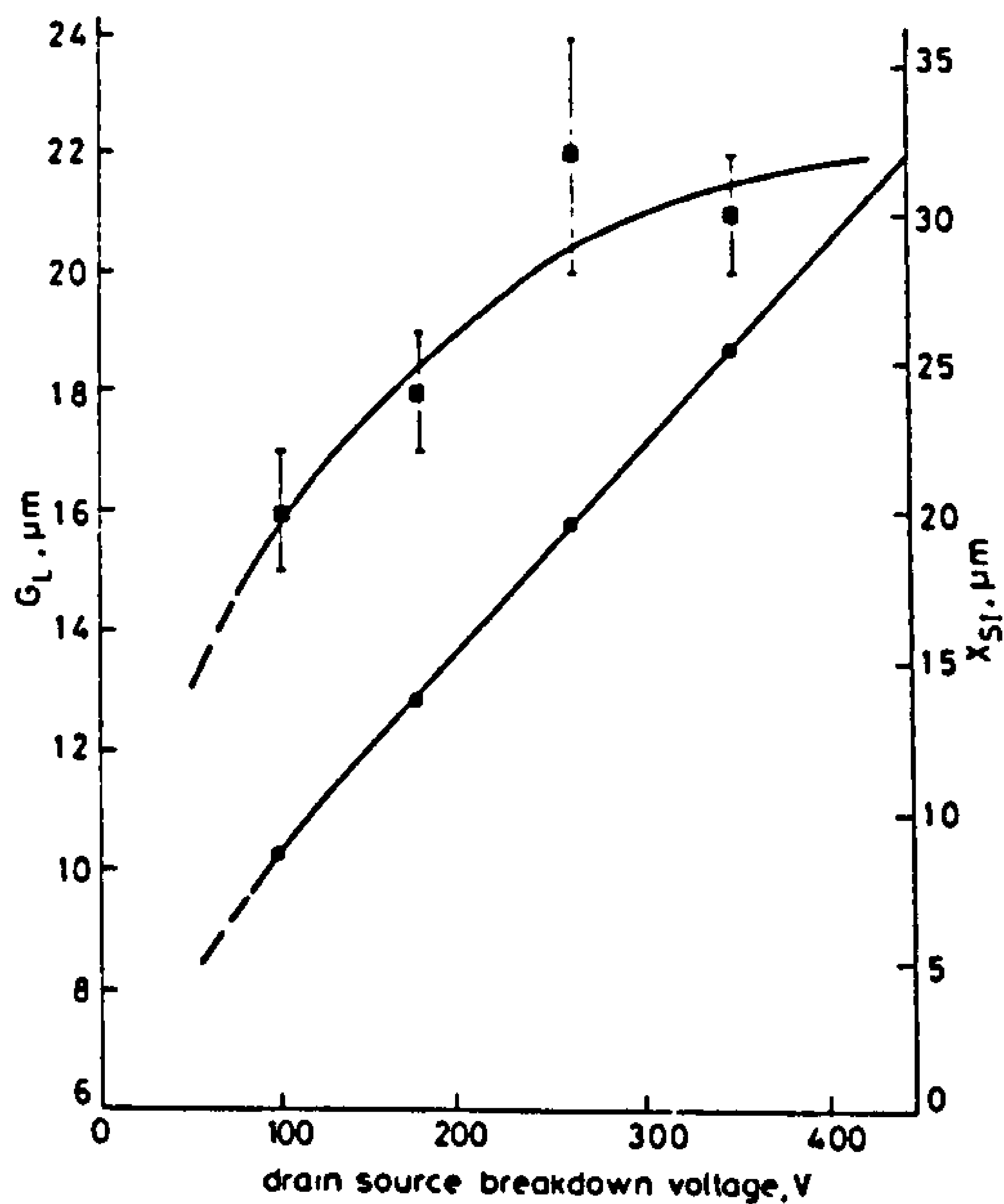


Fig. 5 *Optimum G_L and X_{SI} as a function of breakdown voltage*

■ G_L
● X_{SI}

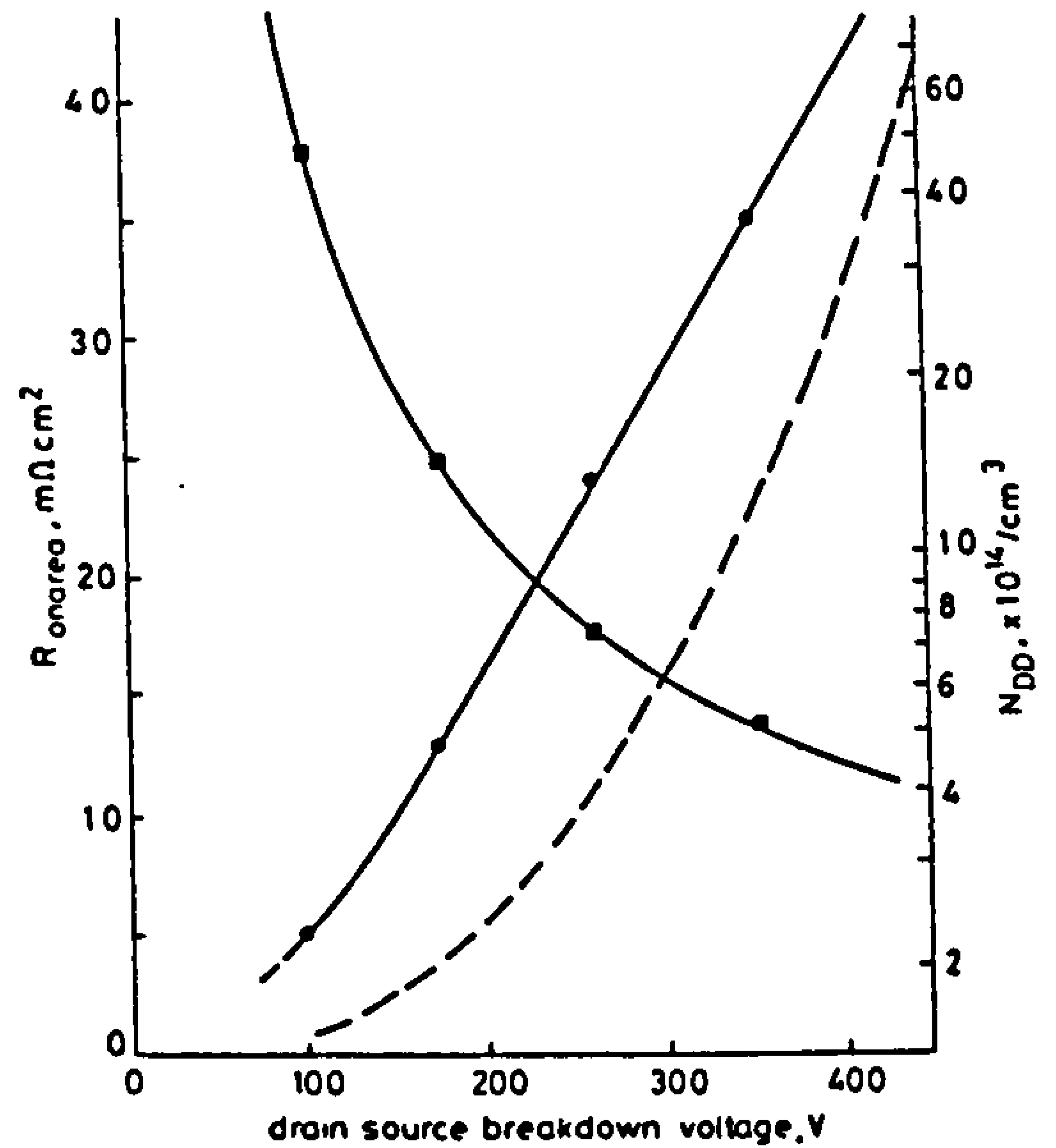


Fig. 6 *Optimum N_{DD} and $R_{on\,area}$ product as a function of breakdown voltage*

---- ideal $R_{on\,area}$ for uniform current flow and continuous top and bottom contacts
● $R_{on\,area}$
■ N_{DD}

and

$$N_{DD} = 10^{18}(BV - V_{body})^{-4/3} \text{ (cm}^{-3})$$

where V_{body} is the voltage contained within the body diffusion. Although this varies slightly with drain doping, a constant value of 45 V allows a good agreement with the results of this work. These relationships between breakdown voltage, N_{DD} and X_{SI} suggest an absolute limit to the resistance-area product given by:

$$R_{on\,area} = 3.0 \times 10^{-8}(BV - V_{body})^{7/3} + 2.3 \times 10^{-6}(BV - V_{body})^{4/3}$$

based on a constant body diffusion depth (X_{JBV}) of 5 μm, and continuous top and bottom contacts to a region with the optimised thickness and doping concentration X_{SI} and N_{DD}. In practice, the finite source and body dimensions limit the available resistance-area product to higher values. The relationship is included in Fig. 6 and a comparison with the calculated values for $R_{on\,area}$ proves a measure for the chip area use efficiency. In practice the geometric advantages of cellular, rather than linear, geometry devices leads to improvements in this comparison.

3 Conclusions

The results of a 2-dimensional analysis of the VDMOS transistor show that the interbody spacing of an optimised structure decreases for design breakdown voltages less than 400 V. At higher voltages, a constant value appears to be appropriate so that only the epitaxial layer specification needs to be adjusted to accommodate a range of breakdown voltages. Over the range 150–400 V, the minimum $R_{on\,area}$ product that can be achieved with a constant body width of 15 μm has been found to increase approximately linearly with the device breakdown voltage rating.

4 References

1 SUN, S.C., and PLUMMER, J.D.: 'Modeling of the on-resistance of LDMOS, VDMOS and VMOS power transistors', *IEEE Trans.*, 1980, ED-27, (2), pp. 356–367

2 TAMAR, A.A., RAUCH, K., and MOLL, J.L.: ibid., 1983, ED-30, (1), pp. 73–76

3 BYRNE, D.J., and BOARD, K.: 'Minimisation of on-resistance of VDMOS power FETs', 1983, *Electron. Lett.*, 19, (14), pp. 519–521

4 BOARD, K., BYRNE, D.J., and TOWERS, M.S.: *IEEE Trans.*, 1984, ED-31, (1), pp. 75–80

5 VAN OVERSTRAETEN, R., and DE MAN, H.: 'Measurement of the ionization rates in diffused silicon *p-n* junctions', *Solid-State Electron.*, 1970, 13, p. 583–608

6 SUN, S.C., and PLUMMER, J.D.: 'Electron mobility in inversion and accumulation layers on thermally oxidised silicon surfaces', *IEEE Trans.*, 1980, ED-27, (8), pp. 1497–1508

7 DARWISH, M.N., and BOARD, K.: *ibid.*, 1984, ED-31, (12), pp. 1769–1773