This item is held in Loughborough University's Institutional Repository (https://dspace.lboro.ac.uk/) and was harvested from the British Library's EThOS service (http://www.ethos.bl.uk/). It is made available under the following Creative Commons Licence conditions.

Attribution-NonCommercial-NoDerivs 2.5

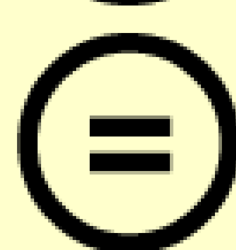
You are free:

- to copy, distribute, display, and perform the work

Under the following conditions:

Attribution. You must attribute the work in the manner specified by the author or licensor.

Noncommercial. You may not use this work for commercial purposes.



No Derivative Works. You may not alter, transform, or build upon this work.

- For any reuse or distribution, you must make clear to others the license terms of this work.
- Any of these conditions can be waived if you get permission from the copyright holder.

Your fair use and other rights are in no way affected by the above.

This is a human-readable summary of the Legal Code (the full license).

Disclaimer

For the full text of this licence, please go to:
http://creativecommons.org/licenses/by-nc-nd/2.5/

Electrical Overstress and Electrostatic Discharge Failure in Silicon MOS Devices

by

Martin James Tunnicliffe B.Eng.(Hons.)

A Doctoral Thesis

Submitted in partial fulfilment of the requirements
for the award of the Degree of Doctor of Philosophy
of Loughborough University of Technology

January 1993

© by M.J.Tunnicliffe, 1993.

Still round the corner there may wait

a new road or a hidden gate

And though we pass them by today,

tomorrow we may come this way

And take the hidden paths that run

towards the Moon or to the Sun.

J.R.R.Tolkien

The Fellowship of the Ring

Abstract

This thesis presents an experimental and theoretical investigation of electrical failure in MOS structures, with a particular emphasis on short-pulse and ESD failure. It begins with an extensive survey of MOS technology, its failure mechanisms and protection schemes. A program of experimental research on MOS breakdown is then reported, the results of which are used to develop a model of breakdown across a wide spectrum of time scales. This model, in which bulk-oxide electron trapping/emission plays a major role, prohibits the direct use of *causal* theory over short time-scales, invalidating earlier theories on the subject.

The work is extended to ESD stress of both polarities. Negative polarity ESD breakdown is found to be primarily oxide-voltage activated, with no significant dependence on temperature of luminosity. Positive polarity breakdown depends on the rate of surface inversion, dictated by the Si avalanche threshold and/or the generation speed of light-induced carriers. An analytical model, based upon the above theory is developed to predict ESD breakdown over a wide range of conditions.

The thesis ends with an experimental and theoretical investigation of the effects of ESD breakdown on device and circuit performance. Breakdown sites are modelled as resistive paths in the oxide, and their distorting effects upon transistor performance are studied. The degradation of a damaged transistor under working stress is observed, giving a deeper insight into the latent hazards of ESD damage.

Acknowledgements

I am indebted all my colleagues in the International Electronics Reliability Institute (IERI), past and present, for their technical assistance and stimulating conversation over the years. In particular, I would like to thank my supervisor Dr. Vince Dwyer for his help on matters mathematical, quantum mechanical and otherwise. Thanks are also due to Professor David Campbell, former director of IERI, Mr. David Shore of the Ministry of Defence (the sponsor of this research), Dr. Andy Franklin of National Semiconductor and Dr. Ajith Amerasekera of Texas Instruments, all of whom have provided useful suggestions and criticisms. Thanks also to Jeff Jones and Alan Schwarzenberger, whose formidable computer skills have often saved the day.

There are many other people, almost too numerous to mention, whose friendship has kept me going through my years at Loughborough University. My thanks go out to Neil, Tim, Farzin, Shyam, Paul, Jonathan, Joe, Roger, Jane, Rosemary, Danny, Susan, Christine, Muhammed, Nick, Linda and many more.

Finally I would like to thank all my family, above all my mother and father, for their love and support throughout so many years. It is to them that this work is dedicated.

Contents

Page No.

Abstract (i)

Acknowledgements (ii)

Symbols and Nomenclature (ix)

Chapter 1: INTRODUCTION 1

- **1.1 Historical Background** 1
- **1.2 Operation of the MOSFET** 3
- **1.3 The Future of MOS Technology** 4
- **1.4 Study Synopsis** 5
- **1.5 References** 7

Chapter 2: MOS TECHNOLOGY IN THEORY AND PRACTICE 9

- **2.1 Introduction** 9
- **2.2 Physics of Silicon Dioxide** 9
  - 2.2.1 Properties and Applications of SiO_2 9
  - 2.2.2 Atomic Structure 10
  - 2.2.3 Electronic Structure 11
  - 2.2.4 Bulk Oxide Defects 13
  - 2.2.5 The Si-SiO_2 Defects 15
  - 2.2.6 The Gate-SiO_2 Interface 15
- **2.3 Theory of MOSFET Operation** 16
  - 2.3.1 Inversion Layer Formation 16
  - 2.3.2 First Order MOSFET Model 18
- **2.4 Miniaturisation of the MOSFET** 20
  - 2.4.1 Device Scaling 20
  - 2.4.2 Second Order Effects 21
  - 2.4.3 Second Order Model 22
- **2.5 MOS Technologies** 24
  - 2.5.1 Planar Bulk-Silicon Technologies 24
    - 2.5.1.1 PMOS and NMOS 24
    - 2.5.1.2 Complimentary MOS (CMOS) 25
    - 2.5.1.3 High Performance MOS (HMOS) 26

2.5.2 Insulating Substrate Technologies 27
2.5.3 Ultra-Small Dimension MOS Technologies 28
2.5.3.1 Double Diffused MOS (DMOS) 28
2.5.3.2 V-Groove MOS (VMOS) 29
2.5.4 Charge Transfer Devices (CCDs) 29
2.6 Summary 31
2.7 References 31

Chapter 3: MOS FAILURE MECHANISMS 34

3.1 Introduction 34
3.2 Sources of Event Dependent Failure 34
3.2.1 In-Circuit Electrical Overstress (EOS) 35
3.2.2 Electrostatic Discharge (ESD) Pulse Stress 36
3.2.3 Electromagnetic Pulse (EMP) Stress 39
3.3 Dielectric Breakdown 39
3.3.1 Phenomenology of Dielectric Breakdown 39
3.3.2 Types of Dielectric Breakdown 40
3.3.3 Field-Dependent or Avalanche Breakdown 41
3.3.4 Thermal Dielectric Breakdown 44
3.3.5 Causal or Time Dependent Dielectric Breakdown (TDDB) 45
3.3.5.1 Dielectric 'Erosion' Models 48
3.3.5.2 Trapped Charge Related Models 49
3.3.5.3 Other Models 51
3.3.6 Intrinsic and Extrinsic Dielectric Breakdown 51
3.3.7 Fast Transient and ESD Dielectric Breakdown 53
3.4 Hot Carrier Injection 55
3.4.1 Field Stimulation of Hot Carriers 56
3.4.2 Modelling Hot Electron Injection 57
3.4.3 Avalanche Breakdown and Injection in MOS Capacitors 59
3.4.4 Hot Carriers in MOSFETs 60
3.5 Other MOS Failure Mechanisms 61
3.5.1 Parasitic Bipolar Effects 61
3.5.2 Electromigration 61
3.5.3 Electro-Thermomigration ('Spiking') 62
3.5.4 Dielectric Instability 62
3.5.5 Radiation Damage 62
3.6 Latent or 'Walking Wounded' Failure 63
3.7 Failure Mechanisms in Different MOS Technologies 65
3.8 Methods of Improving Reliability 65
3.8.1 Anti Static Precautions 66
3.8.2 Protection Circuits 66
3.8.3 Screening 68
3.9 Conclusions 69
3.10 Summary 69
3.11 References 70

Chapter 4: EXPERIMENTAL APPARATUS AND TEST SAMPLES 83

4.1	**Introduction**	83
4.2	**Apparatus**	83
	4.2.1 Wafer Chuck and Microprober	83
	4.2.2 Electrical Instruments	85
	4.2.2.1 Parametric Analyzer	85
	4.2.2.2 Programmable Electrometer	86
	4.2.2.3 Programmable Voltage Source	86
	4.2.2.4 L-C-R Meter	86
	4.2.2.5 Digital Oscilloscope	86
	4.2.3 Stress Generation	88
	4.2.3.1 Human Body ESD Pulse Stress	88
	4.2.3.2 Constant Voltage Generation	92
	4.2.3.3 Ramp Voltage Generation	92
	4.2.3.4 Constant Current Generation	95
4.3	**Test Samples**	96
	4.3.1 Three Inch NMOS Bulk Silicon Wafers	97
	4.3.2 Three Inch CMOS Bulk Silicon Wafers	97
	4.3.3 Four Inch HMOS Bulk Silicon Wafers	98
	4.3.4 Four Inch Bulk Silicon MOS Capacitor Wafers	98
	4.3.5 Four Inch Silicon-on-Sapphire (SOS) Wafer	98
	4.3.6 Past History of Samples	99
4.4	**Summary**	99
4.5	**References**	99

Chapter 5: EXPERIMENTAL PROCEDURES AND RESULTS 101

5.1	**Introduction**	101
5.2	**Electrostatic Discharge Experiments**	101
	5.2.1 Uniformity of Experimental Samples	102
	5.2.2 Dimensional Dependence of ESD Breakdown	106
	5.2.3 Temperature Dependence of ESD Breakdown	107
	5.2.4 ESD Breakdown as a Function of Luminous Intensity	109
	5.2.5 ESD Breakdown as a Function of Discharge Capacitance	113
	5.2.6 ESD Breakdown as a Function of Discharge Resistance	115
	7.2.7 Conclusions	117
5.3	**Ramp Voltage Experiments**	117
	5.3.1 Negative Polarity	117
	5.3.2 Positive Polarity	118
5.4	**Constant Voltage Experiments**	118
	5.4.1 Temperature and Dimensional Dependencies	119
	5.4.2 Illumination Dependence	120
	5.4.3 Doping Dependence	122
	5.4.4 Electric Field Dependence	123

5.5 **Constant Current Injection Experiments** 124
5.5.1 Dimensional Dependence 124
5.5.2 Current Dependence 125
5.6 **Summary** 125
5.7 **References** 127

Chapter 6: DISCUSSION AND FURTHER EXPERIMENTS 128

6.1 **Introduction** 128
6.2 **Qualification of ESD Data** 128
6.2.1 AutoZap Voltage Resolution 128
6.2.2 Analysis of Breakdown Criterion 129
6.2.2.1 NMOS and HMOS Devices 130
6.2.2.2 CMOS Devices 131
6.2.2.3 SOS Devices 132
6.2.3 Cumulative Damage 132
6.2.4 Ideal Test System 134
6.3 **Analysis of ESD Results** 134
6.3.1 Breakdown as a Function of Illumination 134
6.3.2 Breakdown as a Function of Temperature 134
6.3.3 Breakdown as a Function of Device Size 135
6.3.4 Breakdown as a Function of C_1 and R_2 136
6.4 **Analysis of Constant Voltage and Current Data** 139
6.4.1 Region III (Short Time-Scale) 139
6.4.1.1 Fast Transient Tunnelling Experiments 140
6.4.1.2 Theoretical Analysis of Tunnelling Time Delay 143
6.4.2 Region II (Moderate Time-Scale) 146
6.4.3 Region I (Long Time-Scale) 148
6.4.4 Analysis of Constant Current Data 150
6.5 **Analysis of Ramp Voltage Data** 151
6.5.1 Negative Polarity 152
6.5.2 Positive Polarity 155
6.5.3 Analysis of ESD Transient Data 158
6.5.3.1 Negative Polarity Analysis 158
6.5.3.2 Positive Polarity Analysis 159
6.6 **Qualitative Model of Breakdown** 161
6.6.1 Space-Charge Evolution in Oxide 162
6.6.2 Mechanisms of Oxide Wearout 165
6.6.3 Model of ESD Breakdown 168
6.6.3.1 Negative Polarity Model 168
6.6.3.2 Positive Polarity Model 170
6.7 **Summary** 176
6.8 **References** 177

Chapter 7: MATHEMATICAL MODEL OF OXIDE WEAROUT 181

7.1 Introduction 181
7.2 Analytical Model of Oxide Wearout 182
7.3 Analytical Model of Constant Voltage/Constant Current Breakdown 183
7.4 Analytical Model of ESD Breakdown 185
7.4.1 Low-Resistance Breakdown 186
7.4.1.1 Breakdown Voltage Threshold 186
7.4.1.2 Time-to-Breakdown 190
7.4.2 High-Resistance Breakdown 191
7.4.2.1 Voltage Pinning Model 191
7.4.2.2 Field-Dependent (FIM) Breakdown 194
7.4.3 Comparison with Experimental Data 194
7.4.3.1 Empirical Fitting of Models to ESD Data 194
7.4.3.2 Comparison with Constant Current/Voltage Data 195
7.4.4 'Enhanced Injection' Model of High Resistance Breakdown 199
7.5 Summary and Conclusions 203
7.6 References 204

Chapter 8: LATENT FAILURE AND PARAMETRIC DRIFT 205

8.1 Introduction 205
8.2 Apparatus 207
8.3 Test Samples 207
8.4 Sub-Breakdown Latency and Parametric Drift 208
8.4.1 Experiment 208
8.4.2 Theoretical Analysis 210
8.5 Latent Damage and Parametric Drift in Damaged-Oxide Devices 211
8.5.1 Experimental Procedure 211
8.5.2 Results 212
8.5.3 Discussion and Modelling 216
8.5.3.1 Qualitative Analysis of Failure Characteristics 216
8.5.3.2 PSpice Model of Failed Devices 217
8.5.3.3 Effect of Degradation upon CMOS Operation 217
8.6 Summary and Conclusions 221
8.7 References 223

Chapter 8: DISCUSSIONS AND CONCLUSIONS 224

9.1 Introduction 224
9.2 General Aims of This Study 224
9.3 Literature Survey Chapters 225
9.4 Experimental Studies of Oxide Breakdown 226
9.5 Theoretical Analysis of Oxide Breakdown 227
9.6 Studies of Parametric Degradation 228

9.7 Final Comments 229
9.8 Suggestions for Future Research 230
9.9 Summary 230
9.10 References 231

Appendix A: MATHEMATICAL ANALYSIS OF THE MOS STRUCTURE 232

A.1 Introduction 232
A.2 Oxide Capacitance in Small Dimension Devices 232
A.3 Fowler-Nordheim Tunnelling Theory 235
A.4 Image-Force Barrier Lowering 239
A.5 The Effect of Cathode Electron Concentration 240
A.6 The Effect of Cathode Doping 243
A.7 Effect of Cathode Surface Quantization 246
A.8 References 246

Appendix B: ABSTRACTS OF PROGRESS REPORTS 248

B.1 Progress Report No.6 248
B.2 Progress Report No.7 248
B.3 Progress Report No.8(a) 249
B.4 Progress Report No.8 250

Appendix C: REPRODUCTIONS OF PUBLISHED PAPERS 251

Paper I (Proc. CERT, 1990) 252
Paper II (Proc. 12th. EOS/ESD Symp., 1990) 263
Paper III (Proc. Euro. ESD Symp., 1991) 271
Paper IV (Proc. 14th. EOS/ESD Symp., 1992) 281

Symbols and Nomenclature

I: Units and Prefixes

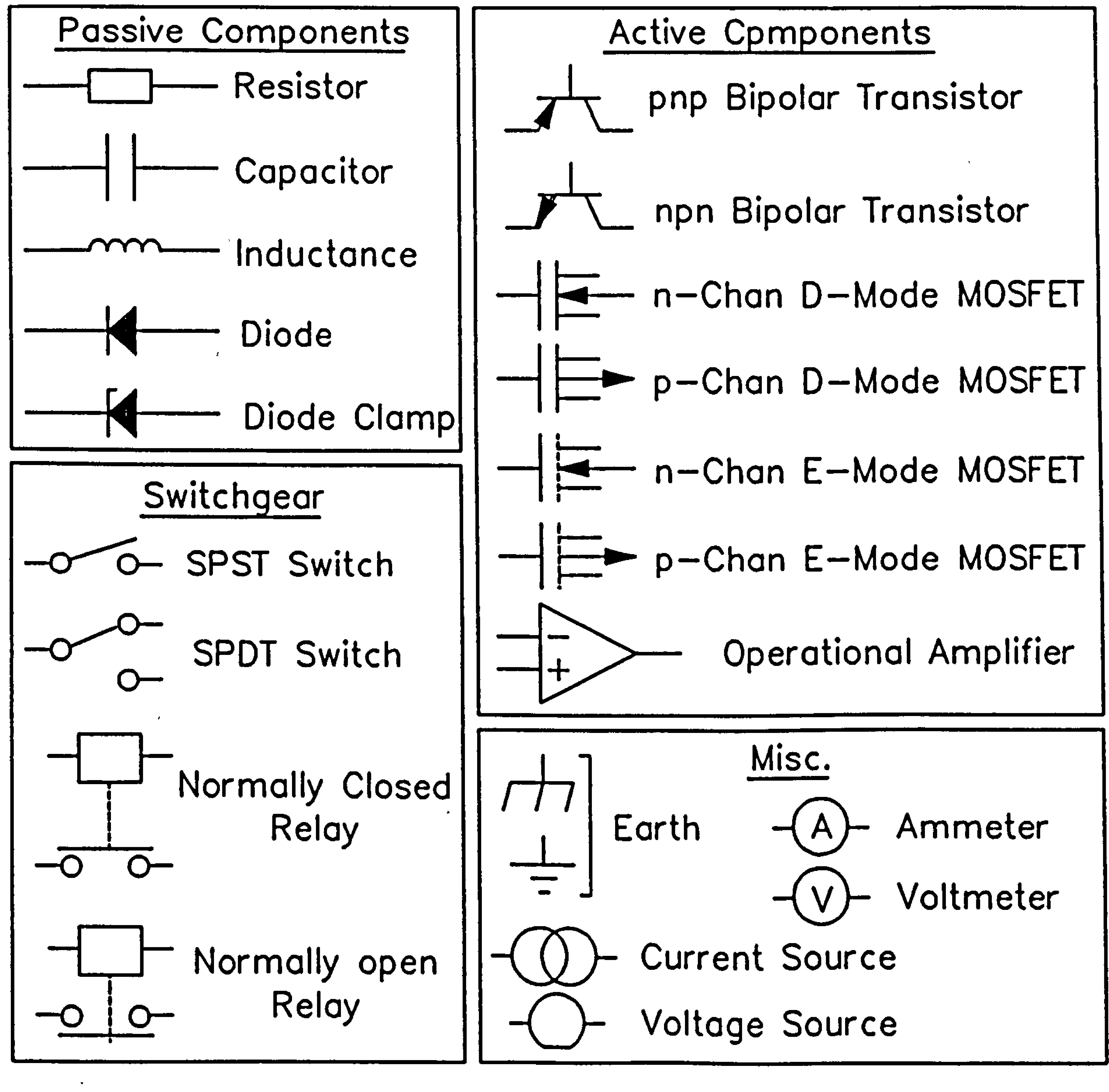
A	Ampere.
Å	Angstrom (10^{-8}cm).
C	Coulomb.
°C	Centigrade.
cal.	Calorie.
cm	Centimetre.
eV	Electron-volt.
F	Farad.
G	Giga (10^{9}).
H	Henry.
Hz	Hertz.
J	Joule.
K	(1) Kilo (1024), (2) Kelvin.
k	Kilo (10^{3}).
M	Mega (10^{6}).
m	Metre, Milli (10^{-3}).
mol.	Mole.
n	Nano (10^{-9}).
p	Pico (10^{-12}).
s/sec.	Second.
V	Volt.
Ω	Ohm.

II: Chemical Symbols and Formulae

Al	Aluminium.
As	Arsenic.
B	Boron.
Cl	Chlorine.
Cr	Chromium.
Ga	Gallium.
Ge	Germanium.
H	Hydrogen.
III/IV/V	Groups in periodic table.
In	Indium.
K	Potassium.
N	Nitrogen.
Na	Sodium.
O	Oxygen.
P	Phosphorus.
Si	Silicon.

AlAs	Aluminium Arsenide.
Al_2O_3	Aluminium Oxide (Sapphire).
GaAs	Gallium Arsenide.
HCl	Hydrochloric Acid.
InP	Indium Phosphate.
Si_3N_4	Silicon Nitride.
SiO	Silicon Monoxide.
SiO_2	Silicon Dioxide.

III: Electrical Circuit Symbols



IV: Acronyms and Abbreviations

a.c. (or A.C.)	Alternating current.
BBII	Band-to-band impact ionization.
CAT	Coulombic attractive trap.
CBM	Charged board model.
CCD	Charge coupled device.
CDM	Charged device model.
CMOS	Complementary metal oxide semiconductor.
CORS	Circuit oxide reliability simulator.
CRT	Coulombic repulsive trap.
C/V	Capacitance vs. voltage.
CVD	Chemical vapour deposition.
CT1	Tektronix current transformer probe.
D-Mode	Depletion mode.
DMOS	Double diffused MOS.
DUT	Device under test.
d.c. (or D.C.)	Direct current.
E′	E-Prime centre (i.e. a dangling half-filled orbital).
EMI	Electromagnetic interference.
E-Mode	Enhancement mode.
EMP	Electromagnetic pulse.
EOS	Electrical overstress.
E^2PROM	Electrically erasable programmable read only memory.
ESD	Electrostatic discharge.
FET	Field effect transistor.
FIM	Field induced model.
F-N	Fowler-Nordheim.
HBM	Human body model.
HMOS	High performance MOS.
HPIB	Hewlett-Packard interface bus (also called GPIB and an IEEE-488 bus).
HTRB	High temperature reverse bias.
I.C. (or i.c.)	Integrated circuit.
IGFET	Insulated gate field effect transistor.
I/II/III	Regions in the log(t_{bd}) vs. F curve.
I/V	Current vs. voltage.
L.C.D.	Liquid crystal diode.
L-C-R	Inductance-capacitance-resistance.
LO	Longitudinal optic (phonon mode).
LSI	Large scale integration.
MESFET	Metal semiconductor field effect transistor.
MM	Machine model.
MOSFET	Metal oxide semiconductor field effect transistor.
MSI	Medium scale integration.
n-Type	Semiconductor with majority electrons.
n^+-Type	Heavily n-Type.

NAND	Not-AND logic gate.
NMOS	Technology based on n-channel MOSFET.
MNOS	Metal nitride oxide silicon.
NOR	Not-OR logic gate.
NOT	Invertor logic gate.
NTS	Neutral trap state.
p-Type	Semiconductor with majority holes.
p^+-Type	Heavily p-Type.
p.c.b.	Printed circuit board.
RAM	Random access memory.
RC1/RC2	Relay coils in voltage ramp generator.
RL1	I/V vs. HBM switchover relay.
PMOS	Technology based on p-channel MOSFET.
SCR	Silicon controlled rectifier (thyristor).
SEM	Scanning electron microscope.
SHA	Safe handling area.
SMU	Source/measure unit.
SOI	Silicon on insulator.
SOS	Silicon on sapphire.
SPDT	Single pole double throw.
SPST	Single pole single throw.
SSI	Small scale integration.
s.t.p.	Standard temperature and pressure.
TBII	Trap-to-band impact ionization.
TCM	Tunnelling current microscopy.
TDDB	Time dependent dielectric breakdown.
ULSI	Ultra large scale integration.
VDU	Visual display unit.
VLSI	Very large scale integration.
VMOS	V-groove MOS.
w.r.t.	With respect to.
ZMR	Zone melting/recrystallization.

V: Mathematical Symbols

A	Oxide area.
B	Exponential constant in Fowler-Nordheim equation.
C	Capacitance.
C_c	Coupling capacitor in t_d measurement apparatus.
C_{cg}	Chuck-ground capacitance.
C_{in}	Input capacitance (of logic gate).
C_k	Charge collection capacitor placed in series with DUT.
C_{ox}	Oxide capacitance.
C_p	Oscilloscope probe capacitance.
$C_p(A)$, $C_p(B)$	Capacitances of probes A and B.
C_{par}	Capacitance directly in parallel with oxide.

C_{pc}	Probe-chuck capacitance.
C_{sys}	Total capacitance of ESD pulse system (excluding C_{ox}).
C_t	Capacitance of pulse transmission medium.
C_x	Parasitic capacitance of discharge module.
C_1	HBM discharge (body) capacitance.
C_2	Parasitic capacitance in parallel with R_2.
D(F)	Density of oxide defects activated at field F.
$D(E_x)$	Tunnelling barrier transmission coefficient.
D_n	Electron diffusion coefficient.
E	Energy.
E_a	Activation energy.
E_{bd}	Electrostatic energy required to support ESD breakdown.
E_c	Energy of conduction band edge.
E_f	Fermi level.
E_i	Intrinsic Fermi level.
E_{ion}	Ionization energy.
E_v	Energy of valance band edge.
E_x	Energy component normal to Si-SiO_2 interface.
E^*	Critical ionization energy for electron acceleration under LO phonon scattering.
F	Electric field.
f	Frequency.
f(E)	Ferm-Dirac function.
F_a,F_b	Field modulation constants.
F_{an}, F_{cat}	Anode and cathode electric fields.
F_{bd}	Field threshold for dielectric breakdown.
F_x,F_y	Transverse & longitudinal channel fields.
g_m	Transconductance (or 'mutual conductance').
G_1, G_2	Constants in high-resistance breakdown model.
H	(1) Heaviside function, (2) exponential constant in $\alpha(F)$ function, (3) constant equal to (γ-B).
h	Plank's constant.
$\hbar$	Reduced Plank's constant ($h/2\pi$)
I	Oxide injection current.
I(t)	Oscilloscope current waveform after processing.
I_b	Base current.
I_c	Collector current.
I_{cr}	'Critical' current at brittle/ductile transition.
I_d	Drain current.
I_{df}	First order drain current.
I_g	Gate current.
I_{max}	Maximum injection current during ESD pulse.
I_{ox}^*	Critical oxide current required to support ESD oxide breakdown.
$I_{sc}(t)$	'Raw' current waveform from oscilloscope.
$\Im$	Imaginary part of complex number.
J	Current density.

J_{cr}	Critical current density between the 'brittle' and 'ductile' failure modes.
J_{max}	Maximum injection current density during ESD pulse.
J_{ox}	Oxide injection current density.
J_{ox}^*	Critical oxide injection current required to support ESD oxide breakdown.
J_0	Pre-exponential constant in simplified Fowler-Nordheim equation.
k	(1) Wave vector, (2) Boltzmann's constant, (3) Pre-exponential constant in Fowler-Nordheim equation.
L	(1) Inductance, (2) capacitive loading factor, (3) MOSFET channel length.
L_d	Length of defect site in channel.
L_D	Debye length.
L_{ds}	Drain-source length (i.e. channel length).
$L_{ds(eff)}$	Effective drain-source length.
ℓ	Channel length modulation coefficient.
M	Multiplication factor.
m_0	Electron rest mass.
m_{ox}^*	Effective electron mass in SiO_2.
m_{si}^*	Effective electron mass in Si.
m_1,m_2	Constants in high-resistance breakdown model.
n	Electron density.
n(0)	Electron density at Si-SiO_2 interface.
n_1	Density of electrons in NTS traps.
n_2	Density of electrons in CRT traps.
$N(E_x)$	Electron supply function.
N_1	Density of NTS traps.
N_2	Density of CRT traps.
N_a (or N_A)	Acceptor dopant density.
N_c	Effective density of states for conduction band.
N_d (or N_D)	Donor dopant density.
p_x,p_y,p_z	Orthogonal momentum components.
P(F)	Probability of oxide failure at field F.
Q	Charge.
q	Electron charge.
Q_{bd}	Charge to breakdown.
Q_D	Depletion layer charge.
Q_{ox}	Effective oxide space-charge density, total oxide injected charge.
Q_p^*	Minimum density of trapped hole charge required to support breakdown.
Q_{ss}	Surface-state charge density.
r	Voltage ramp rate (dV_2/dt).
r^*	Critical ramp-rate required for breakdown.
R	Resistance.
R_b	Resistance of bulk silicon.
R_c	Coupling resistor in t_d measurement apparatus.

R_{cg} Chuck/ground resistance.
R_g Parasitic gate resistance.
R_{gen} Resistive circuit element, used to model inversion-charge generation.
R_p Oscilloscope probe resistance.
R_{pr} Ditto.
R_f Breakdown filament resistance.
R_s (1) External resistor connected in series with DUT in ramp voltage test, (2) Resistance of bulk silicon.
R_1 HBM charging resistor.
R_2 HBM discharge (body) resistor.
R^* Effective Richardson constant.
$\Re$ Real part of complex number.
s(t) Oxide damage function.
S Defect clustering coefficient.
Sigma Standard deviation.
S_{seq} Total damage inflicted by ESD pulse sequence.
S_t Total damage inflicted by ESD pulse.
T Temperature.
t Time.
t(y) Pre-exponential barrier lowering function.
t_{bd} Time to breakdown.
T_c Boundary between low and high temperature dielectric breakdown.
t_d Time delay to start of ESD pulse.
T_e Electron temperature.
T_{eff} Effective oxide thickness.
t_{fall} Fall-time.
T_{ox} Oxide thickness.
t_{rise} Rise-time.
T_{sc} Time domain width of oscilloscope.
T_t Transit time for electrons in tunnelling barrier.
V Voltage.
V(t) Processed oscilloscope voltage waveform.
v Carrier velocity.
v(y) Exponential barrier lowering function.
V_a Total voltage across R_s and DUT.
V_{av} Semiconductor avalanche voltage.
V_b Voltage across R_s.
V_{bd} ESD breakdown voltage threshold.
V_{bd}^f Field-induced breakdown voltage.
V_{dep} Voltage across depletion layer.
v_d Saturation velocity of channel electrons.
v_d Drift velocity.
V_{dd} Drain supply voltage.
V_{dep} Voltage across depletion layer.
V_{ds} Drain-source voltage.
$V_{ds(sat)}$ Drain saturation voltage.

V_e	Effective ESD pulse voltage.
V_e^*	Value of V_e at breakdown.
V_{FB}	Flat-band potential.
V_{FN}	Fowler-Nordheim potential.
V_{gb}	Gate-substrate voltage.
V_{gs}	Gate-source voltage.
V_{in}	Logic gate input voltage.
V_k	Voltage across C_k.
V_{out}	Logic gate output voltage.
V_{ox}	Voltage across oxide.
V_{ox}^*	Oxide breakdown voltage.
$V_{ox}^{*(seq)}$	Oxide breakdown voltage in ESD pulse sequence.
V_{ox}^{max}	Maximum oxide voltage during ESD pulse.
V_{ox}^p	Peak oxide voltage in ESD pulse sequence.
V^p	ESD pulse voltage.
$V_{sc}(t)$	'Raw' voltage waveform measured by oscilloscope.
V_{ser}	Voltage in series with oxide in MOS device.
V_{seq}	Constant for converting single-pulse breakdown thresholds to their pulse-sequence equivalents.
V_{ss}	Source supply voltage.
V_T (or V_t)	Strong-inversion threshold potential.
v_{th}	Thermal velocity.
V_0	Applied ESD pulse voltage.
V_1	Voltage across C_1.
V_2	Output voltage of ESD generator.
$V_1^{(0)}$	Value of V_1 at the exact moment V_{ox} reaches V_{FN}.
W	Channel width.
W_d	(1) Depletion layer width, (2) width of channel defect site.
x	Position coordinate.
x(t)	V_b/V_a
x_t	Centroid position of trapped charge.
y(t)	$\int_0^t V_b dt/V_a$.
Z_d	A.C. impedance of MOS device.
α	Dimensional scaling factor.
$\alpha(F)$	Hole generation efficiency.
α_c	Field-induced ESD coupling coefficient.
α_0	Pre-exponential constant in $\alpha(F)$ function.
γ	Exponential constant in $\tau(F)$ function.
ΔJ	Current resolution.
ΔL	Channel modulation length.
ΔV	ESD pulse voltage resolution.
ΔV_{lat}	Width of latent damage window.
ΔV_{ox}	Oxide voltage resolution.
$\Delta\phi_{ms}$	Effective work function difference.
ϵ_0	Permittivity of free space.
ϵ_{ox}	Dielectric constant of SiO_2.
ϵ_{si}	Dielectric constant of Si.

η	Hole trapping efficiency.
$\Theta(J)$	Charge-to-breakdown function.
λ	Mean free path length.
μ	Carrier mobility.
μ_e (or μ_n)	Electron mobility.
μ_0	Bulk Si mobility.
π	3.14159
σ	(1) Trap capture cross section, (2) standard deviation.
σ_1	Capture cross section of NTS traps.
σ_2	Capture cross section of CRT traps.
$\tau(F)$	Time-to-breakdown function.
τ_c	Time constant of ESD pulse.
τ_e	Trap emission time constant.
τ_{maj}	Majority carrier response time.
τ_{min}	Minority carrier response time.
τ_0	Pre-exponential constant in $\tau(F)$ function.
ϕ_f	Bulk Si Fermi potential.
ϕ_s	Si surface potential.
ϕ_{ms} (or ϕ)	Effective Si-SiO_2 work function.
χ_{ms}	Effective electron affinity.
$\psi(x)$	Electron wavefunction.
ψ	Potential.
ψ_s	Si-SiO_2 surface potential.

Chapter 1

Introduction

1.1 Historical Background

Active electronic components form the basis of computers, communication networks, spacecraft control systems and all other modern electronic hardware. Such devices can amplify electrical signals and perform switching operations in logic circuits. The history of the active device can be traced back to the 1880s when Edison discovered that a hot cathode passed an electric current across a vacuum to a neighbouring anode. This led to the invention the thermionic 'diode' valve by Fleming in 1904. In 1907, de Forest improved Fleming's design, producing the 'Audion' or 'triode' valve, in which the current between a hot cathode and a cold anode was controlled by a voltage applied to an intervening 'grid'. The triode was the earliest active electronic component. Variations on this design dominated the market until the 1950s when they were gradually superseded by the cheaper and less cumbersome 'transfer-resistor' or 'transistor', an active device based on the emerging solid-state technology [1].

The earliest transistor design was patented in the United States and Canada by J.E.Lillenfeld in 1925 [2]. Its proposed operation was not unlike that of de Forest's triode. Current flowing in a copper sulphide channel between gold 'drain' and 'source' electrodes was controlled by a voltage applied to an aluminium 'gate' electrode. In 1928, Lillenfeld patented the earliest insulated gate field effect transistor (IGFET) design, in which the conductivity of a copper sulphide channel was controlled by the voltage on a gate, insulated from the channel by a layer of aluminium oxide [2,3]. The IGFET was also patented by Heil in Great Britain in 1935 [4]. These designs were the forerunners of the modern MESFET and MOSFET transistors. It is extremely doubtful, however, that any workable transistors were constructed at this time.

The first known operational transistor, a germanium bipolar device, was developed in 1947 by Shockley, Bardeen and Brattain [5,6]. By the late 1950s, small scale bipolar integrated circuits[1] (SSI) were manufactured but these were limited by high power dissipation and complicated fabrication processes.

---

[1]. This thesis deals exclusively with monolithic (as opposed to hybrid) integrated circuits.

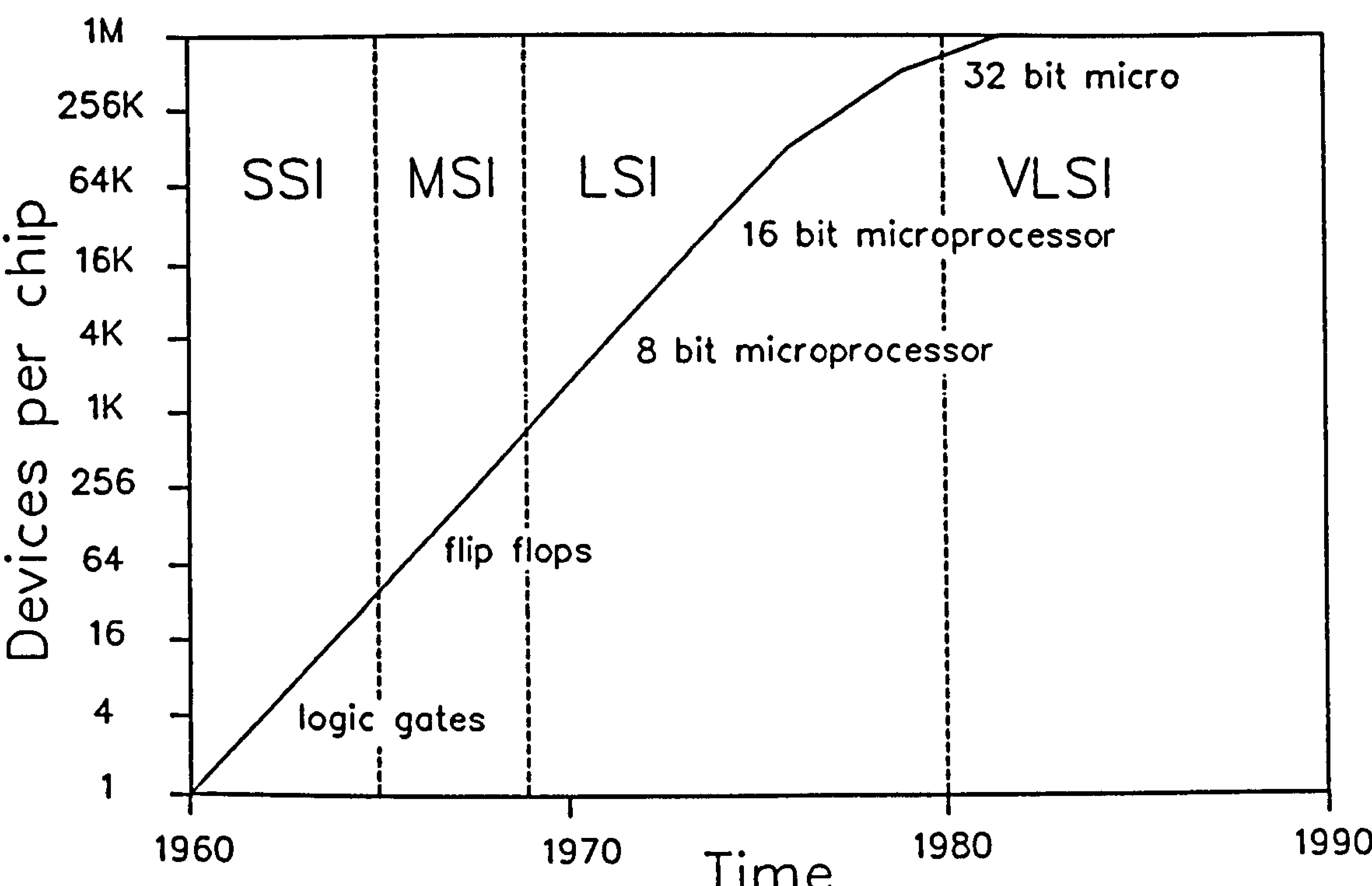


Figure 1.1: Integrated Circuit Complexity as a Function of Time (after G.Moore [11]). [Note: Various definitions of LSI, VLSI etc. are employed by different authors. Here, the definitions used by A.D.Milne [16] are adopted.]

In 1948, Shockley and Pearson performed a successful (albeit crude) practical demonstration of Lillenfeld's IGFET principle [7]. However, it was not until 1960, with the advent of silicon planar technology, that the first modern 'metal-oxide-semiconductor-field-effect-transistor' or MOSFET was developed [8]. Although germanium and gallium arsenide based MOSFETs were developed [9,10], silicon soon proved the best material for these devices, due to the suitability of its high quality native oxide (SiO_2) for a gate insulator. Although the early MOSFETs suffered from dielectric instability [8], the advancement of silicon oxidation technology permitted the practical application of the MOSFET by the late 1960s.

The simple structure and low power dissipation of the MOSFET allowed a greater number of components to be fabricated per unit area of silicon. Medium scale integration (MSI) followed in the mid to late 60s and large scale integration (LSI) in the 70s. This latter category includes the 8 and 16 bit microprocessors and RAM chips up to 64 kilobytes. Very large scale integration (VLSI) followed in the 1980s, with the fabrication of 32 bit microprocessors and 256K RAMs. Such scales of integration can approach one million devices per chip. The growth of integrated circuit complexity over the 70s and 80s is illustrated in Fig.1.1 [11].

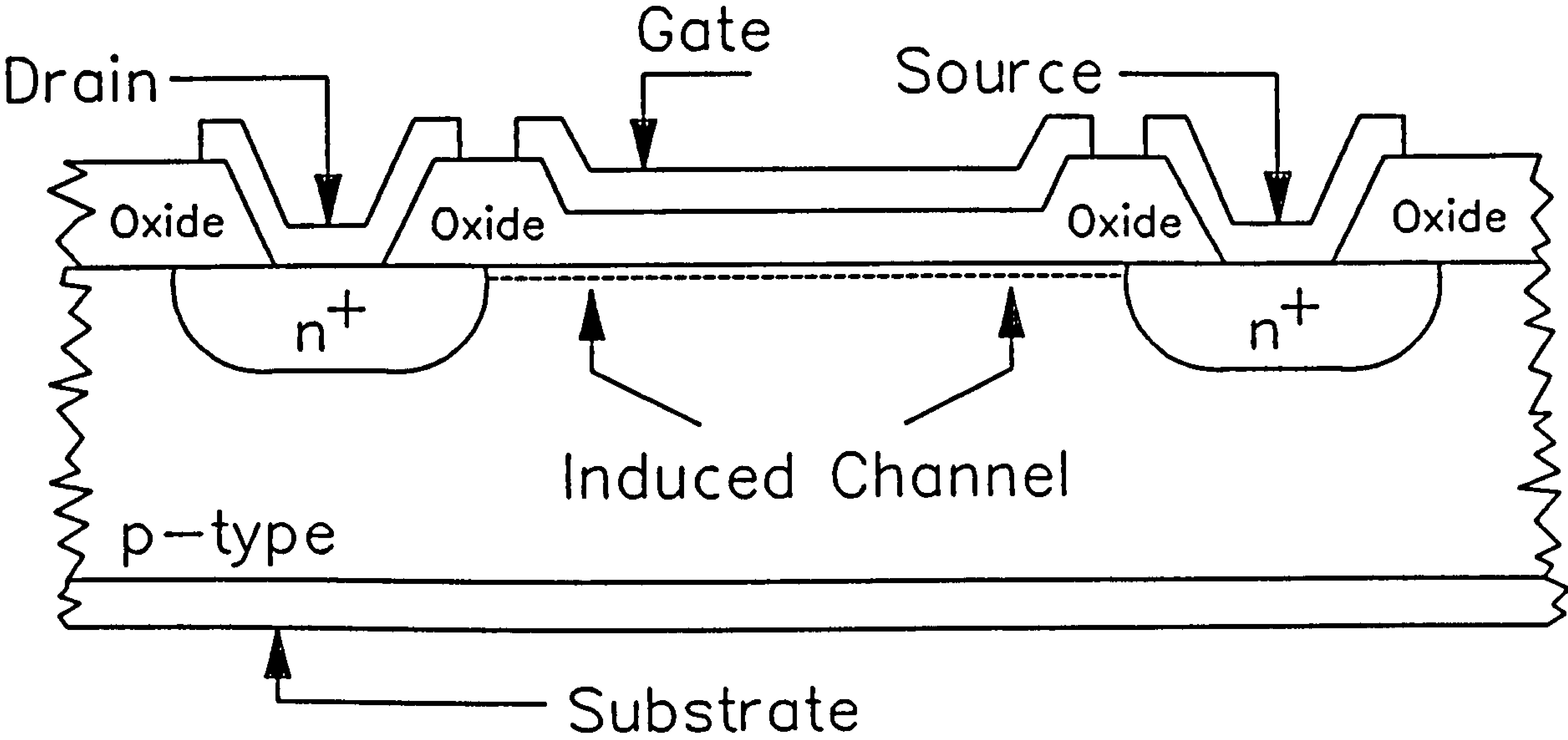


Figure 1.2: Cross Section of Enhancement Mode n-Channel MOSFET

1.2 Operation of the MOSFET [e.g.12,13]

The MOSFET is amongst the simplest and most logical designs for an active device. A simple n-channel enhancement mode (E-mode) MOSFET is shown in Fig.1.2. It consists of a p-type silicon substrate with an etched field-oxide forming a controlled gate dielectric. The gate is formed by a deposited metal (or degenerately doped polycrystalline silicon) layer upon this oxide. Heavy n-type implantations form the drain and source contacts. A p-channel device can be visualised by reversing the polarity of doping in the drain, source and channel regions. The device is essentially a four terminal structure, its contacts being the gate, source, drain and substrate (or back contact). The latter may be given a reference voltage or may be left floating.

This discussion is confined to transistor operation in the 'common source' mode, i.e. the source terminal will be used as the voltage reference. If the gate-source voltage V_{gs} is held at zero, the channel material remains p-type and an n^+-p-n^+ structure appears between the drain and source. Thus (apart from junction leakage current), the drain current I_d remains zero, irrespective of the drain-source voltage V_{ds}. If V_{gs} is made sufficiently positive, minority electrons are attracted into the channel, forming an n-type 'inversion' layer between drain and source. In this condition, an n^+-n-n^+ structure appears between drain and source and current flows in the drain. I_d depends upon the inversion layer charge, which is in turn controlled by V_{gs}.

In a depletion mode (D-mode) device, the situation is reversed. The conductive channel between drain and source is implanted during fabrication and a drain current flows under zero V_{gs}. If V_{gs} becomes negative, electrons are expelled from the channel and the drain current falls. Both E-mode and D-mode MOSFETs are used extensively in modern microelectronics. The characteristics of the MOSFET are discussed more fully in Chapter Two.

1.3 The Future of MOS Technology

For reasons of economy and speed of operation, the further miniaturisation and increased packing density of electronic devices is desirable. Ultra large scale integration (ULSI) is now being envisaged, in which tens of millions of devices will be fabricated per chip [14]. Some authorities predict that ULSI will be available by the year 2000 [15], permitting the fabrication of an entire computer system upon a single integrated circuit. This will greatly reduce the cost and unreliability associated with external connections.

Since a ten million transistor chip can be rendered useless by the failure of a single transistor, the fabrication of reliable devices is of increasing importance. Furthermore, micrometer-scale MOSFETs, fabricated by electron-beam or X-ray lithography [16] introduce failure mechanisms associated with their small dimensions. For example, 'hot' electrons, accelerated by the high electric fields in the channel region, can enter the oxide and become trapped, causing parametric drift during working life. Additionally the dissipation of spurious electromagnetic pulses (EMP) and electrostatic discharges (ESD) in small structures can cause thermal dissociation of the semiconductor material or dielectric breakdown of the thin SiO_2 gate dielectrics [17]. ESD has now become so important that specialist annual conferences in Europe and the U.S.A. are devoted to its study. Another difficulty is introduced when two neighbouring MOSFETs interact to form an SCR (silicon controlled rectifier), which can be triggered (or 'latched') by an ESD or EMP event. This latter mechanism is termed 'latchup' [18].

In view of these problems, the long-term future of the MOSFET is uncertain. Alternative technologies have been proposed, including high-T_c superconducting FETs [19], nonlinear optical logic and molecular electronics [20]. However, since these technologies are far from technical maturity, the MOSFET's major short-term competitors are based on less exotic principles. Recently, much attention has been focused upon the III/V group of compounds (e.g. GaAs, InP, AlAs) which, with advances in processing techniques, can be readily produced on an industrial scale [21]. A common example of a III/V device is the GaAs MESFET. The extremely high electron mobility of GaAs (8500 $cm^2V^{-1}s^{-1}$ compared

with 1350 $cm^2V^{-1}s^{-1}$ in Si [12]) renders this device ideal for high speed switching circuits. Furthermore, modern fabrication techniques allow complex integrated circuits to be realised in GaAs. The GaAs MESFET may therefore prove to be a major competitor to the MOSFET in the 21st. century, particularly if the advantageous properties of GaAs can be combined with those of silicon [22]. Alternative high-mobility devices may be constructed using Si/SiGe superlattice structures, whose electronic band structures can be tailored for specific applications [23].

The study of the reliability of these structures is, however, still in its early stages and it is still possible that enhanced-reliability MOS devices will prevail.

1.4 Study Synopsis

The study presented in this thesis relates to electrically induced damage in MOS devices with a particular emphasis on electrostatic discharge failure. Experimental and theoretical work is reported from which an accurate analytical model is developed.

The remainder of thesis is divided into the following eight chapters.

Chapter 2 outlines the properties of SiO_2, the metal-oxide-semiconductor system and the MOS transistor. The effects of miniaturising the MOSFET are then described and a 'second-order' model of a miniature device is discussed. The chapter ends with a brief overview of the various MOS technologies currently available.

Chapter 3 describes the various reliability concerns relevant to MOS technology. The various causes of event-related failure (ESD, EOS, EMP etc.) are introduced and discussed. This is followed by a review of the physical mechanisms of MOS failure (eg. oxide dielectric breakdown, hot electron injection and latchup) and a brief summary of the techniques used to protect circuits from failure. These investigations show that relatively little research has been performed on SiO_2 dielectric breakdown under fast transient and electrostatic discharge (ESD) pulse stress, and this area is chosen as the main topic of the thesis.

Chapters 2 and 3 present the technological background to the thesis. Chapters 4 to 9 present the author's own original research into MOS reliability.

Chapter 4 describes the apparatus and test samples used for the experimental work of the thesis. The apparatus is designed to study MOS oxide breakdown under voltage, current and electrostatic discharge (ESD) stress conditions. The experimental MOS structures, which include a wide range of MOS technologies, are then described.

Chapter 5 presents the results of a preliminary program of experimental research, performed using the apparatus and test samples of Chapter 4.

Chapter 6 discusses the experimental results of Chapter 5 and presents additional diagnostic experiments designed to shed further light on the processes involved. A qualitative model is developed, which presents a unified picture of breakdown across a wide spectrum of experimental conditions, including constant voltage, constant current and ESD.

Chapter 7 extends the breakdown theory of Chapter 6 into the realms of quantitative simulation and develops an accurate mathematical model of ESD oxide breakdown under both polarities.

Chapter 8 studies the effect of ESD oxide breakdown upon device and circuit performance. The problem of ESD induced *latent* damage is examined both experimentally and theoretically.

Chapter 9 draws conclusions from the results of the thesis and shows how they may be used to improve reliability and quality control in the semiconductor industry.

The experimental facility has been under development for several years and earlier versions have been described by Amerasekera [17] and Franklin [24]. The current system, including several innovations introduced by the author, is fully described in this work. The experimental samples, which were obtained from several suppliers, are also described in detail.

Some of the original work in Chapters 4 to 8 has already been published as a series of papers in fully refereed specialist conferences [25-28] and as a series of progress reports to the Ministry of Defence [29-31] by whom this research was funded. Abstracts to the progress reports are given in Appendix B, and all the published papers are reproduced in Appendix C.

1.5 References

1. *Encyclopedia Britannica.*
2. Gosling,W, Townsend,W.G., Watson,J, *"Field-Effect Electronics"*, Butterworth, London, 1971.
3. Lillenfeld,J.E., U.S. Patent No.1745175, 1930.
4. Heil,O, British Patent No.439137, 1935.
5. Bardeen,J, Brattain,W.H., *"The Transistor, A Semiconducting Triode"*, Physical Review, **74**, p.230, 1948.
6. Shockley,W, *"The Theory of p-n Junctions in Semiconductors and p-n Transistors"*, Physical Review, **80**, p.467, 1950.
7. Shockley,W, Pearson,G.L., *"Modulation of Conductance of Thin Films of Semiconductor by Surface Charges"*, Physical Review, 74, p.232, 1948
8. Kahng,D, Atalla,M.M., *"Silicon-Silicon Dioxide Field Induced Surface Devices"*, IRE Solid State Device Research Conference, Carnegie Institute of Technology, Pittsburgh, PA, 1960.
9. Chang,L.L., Yu,H.N., *"The Germanium Insulated-Gate Field-Effect Transistor (FET)"*, Proc.IEEE, **53**, p.316, 1965.
10. Wilmsen,C.W., Szpak,S, *"MOS Processing for III-IV Compound Semiconductors: Overview and Bibliography"*, Thin Solid Films, **46**, p.17, 1977.
11. Moore,G, *"VLSI: Some Fundamental Challenges"*, IEEE Spectrum, 16(4), p.30, 1979.
12. Streetman,B.G., *"Solid State Electronic Devices"*, Prentice/Hall International, London, pp.294-323, 1980.
13. Yang,E.S., *"Fundamentals of Semiconductor Devices"*, McGraw Hill Book Co., New York, pp.197-224, 1978.
14. Sasaki,H, *"Directions and Strategies to Achieve ULSI"*, 1988 Symposium on VLSI Technology, Digest of Technical Papers, pp.3-6, 1988.
15. Neppl,F, Pfleiderer,H-J, *"CMOS-Technology - Status, Trends and Applications"*, Journal de Phisique, Colloque C4, Supplément au no.5, Tome 49, Sept. 1988.
16. Milne,A.D., (Editor), *"MOS Devices: Design and Manufacture"*, Edinburgh University Press, Edinburgh, 1982.
17. Amerasekera,E.A., *"Failure Mechanisms in MOS Devices"*, Ph.D. Thesis, Loughborough University of Technology, 1986.
18. Ochoa,A, Dawes,W, *"Latch-up Control in CMOS Integrated Circuits"*, IEEE Trans. Nuclear Science, **NS26**, pp.5065-8, 1979.
19. Moore,D, *"Competitive High-T_c FETs Move Closer"*, Physics World, 5(1), pp.21-2, 1992.
20. Bloor,D, *"Breathing New Life into Electronics"*, Physics World, 4(11), pp.36-40, 1991.
21. Shur,M, *"GaAs Devices and Circuits"*, Pelenum Press, New York, 1987.
22. Dettmer,R, *"GaAs on Si: The Ultimate Wafer?"*, IEE Review, 35(4), 136-7, 1989.
23. Abstreiter,G, *"Engineering the Future of Electronics"*, Physics World, 5(3), pp.36-9, 1992.
24. Franklin,A.J., *"Electrical Overstress Failure in GaAs MESFET Structures"*, Ph.D. Thesis, Loughborough University of Technology, 1990.
25. Tunnicliffe,M.J., Dwyer,V.M., Campbell,D.S., *"Slow Transient Voltage and ESD*

Breakdown in Unprotected MOS Gate Oxides", Proc. 1st. Components Engineering, Reliability and Test Conference (CERT), pp.78-87, 1990.

26. Tunnicliffe,M.J., Dwyer,V.M., Campbell,D.S., "*Experimental and Theoretical Studies of EOS/ESD Oxide Breakdown in Unprotected MOS Structures*", Proc. 12th. EOS/ESD Symposium, pp.162-8, 1990.
27. Tunnicliffe,M.J., Dwyer,V.M., Campbell,D.S., "*A Model for ESD Oxide Breakdown in MOS Transistors*", European ESD Symposium, Eindhoven, Netherlands, 3-5 Sept. 1991.
28. Tunnicliffe,M.J., Dwyer,V.M., Campbell,D.S., "*Parametric Drift in Electrostatically Damaged MOS Transistors*", accepted for 14th. EOS/ESD Symposium, Dallas, TX, 16-18 Sept., 1992.
29. Tunnicliffe,M.J., Dwyer,V.M., "*Failure Mechanisms in Semiconductor Devices: Progress Report No.6*", Contract No. (LUT)ELJB7/(MOD)A5a/1412, 1988.
30. Tunnicliffe,M.J., Dwyer,V.M., "*Failure Mechanisms in Semiconductor Devices: Progress Report No.7*", Contract No. (LUT)ELJB7/(MOD)A5a/1412, 1990.
31. Tunnicliffe,M.J., Dwyer,V.M., "*Failure Mechanisms in Semiconductor Devices: Progress Report No.8(a):- Failure Mechanisms in MOS Technology: a Literature Survey*", Contract No. (LUT)ELJB7/(MOD)A5a/1412, 1992.

Chapter 2

MOS Technology in Theory and Practice

2.1 Introduction

This chapter examines the nature of the metal-oxide-silicon (MOS) system and shows how the structure is practically utilised. It begins by describing the structure and properties of thermal silicon dioxide and goes on to develop a simple 'first order' d.c. MOSFET model. Since the SiO_2 structure is highly complex and poorly understood by even the best authorities, the treatment is highly simplified and priority is given to those matters directly relevant to this thesis.

The 'second order' effects of miniaturisation are then considered and appropriate modifications to the first order model are introduced. The chapter ends with a discussion of some of the major variants of MOS technology.

2.2 The Physics of Silicon Dioxide

2.2.1 Properties and Applications of SiO_2

SiO_2 or 'silica' is the native oxide of crystalline silicon. It is chemically identical to crystalline quartz and is the major constituent of most glasses. It grows rapidly on clean silicon surfaces to a depth of about 2nm on exposure to room temperature air. Since such thin layers have little practical use, thicker oxides are grown at about 1000°C in oxidation furnaces over periods of time between 15 minutes to 2 hours [1]. Although oxide growth can be achieved by room temperature anodization [1,2], chemical vapour deposition (CVD) [2] or via an oxygen plasma [2], the present text is confined to thermal oxidation. The oxidising ambient can be pure O_2 for 'dry' oxides or steam for 'wet' oxides, the wet oxide having a faster growth rate at the expense of electrical stability [1]. The Si/wet SiO_2 surface is also much smoother than the Si/dry SiO_2 interface [3]. Dry oxidation is usually followed by annealing in N_2 or H_2 between 350 and 1000°C [1,2] in order to improve its electrical stability and dielectric strength [4]. Full reviews of the silicon oxidation process are given by Rigo [2], Mott [5], Verwey et al. [6] and Kooi [7].

SiO_2 plays two major roles: Firstly as a 'passivating' layer and secondly as a MOSFET gate dielectric. Passivating SiO_2 is usually thick (of the order of $1\mu m$) and confers the following advantages:

1. It reduces the density of interface states on the silicon surface which seriously affect device performance (Section 2.2.5).
2. It protects the silicon from unwanted electrical contact, allowing aluminium tracks to be deposited upon the wafer.
3. It provides a chemical barrier, preventing foreign atoms from entering the underlying silicon [7].
4. The thickness and high permittivity of SiO_2 protect the silicon from the field effect of itinerant ions on the wafer surface.
5. Since SiO_2 has a low density compared to silicon, it provides a degree of mechanical protection [8].

Since the electrical stability of passivation is not normally of paramount importance, 'wet' oxides are usually used since they have a rapid growth rate [1].

The second major application of SiO_2 is as a gate insulator in MOS devices. Although thermally grown SiO_2 is the best known material for a gate insulator, its properties are far from perfect. 'Dry' oxidation is usually used, since the gate oxide integrity and stability is of importance to the circuit performance and reliability [1]. Strategies have been developed to improve the electrical properties of gate oxides, including the addition of HCl in the oxidising ambient [9] and the deposition of an additional yttrium oxide or silicon nitride dielectric layer [10,11].

2.2.2 Atomic Structure

When grown under controlled conditions, the silica exhibits a vitreous quality, i.e. the atomic arrangement contains a degree of short-range order, although without any long-range organisation. For this reason, the vitreous SiO_2 structure is termed a 'network' rather than a 'lattice', the latter term being reserved for crystalline structures [12].

Each silicon atom shares its outer $3sp^3$ orbitals with four neighbouring oxygen atoms, forming a SiO_4 tetrahedral unit cell (see Fig.2.1). Each oxygen atom is shared between two such cells, forming a 'bridge' between neighbouring Si atoms. In vitreous silica, the angle between the bonds can vary between 110 and 180° [12], with a mean value of approximately 153° [5]. The silicon atoms are separated by approximately 0.3nm [6,12], while oxygen

atoms are about 0.265nm apart [12]. The bonds themselves are largely covalent, containing an ionic component of about 20% of the total bond energy [6]. The resulting structure is fairly open, giving a high mobility to both electrons and ions [1,5].

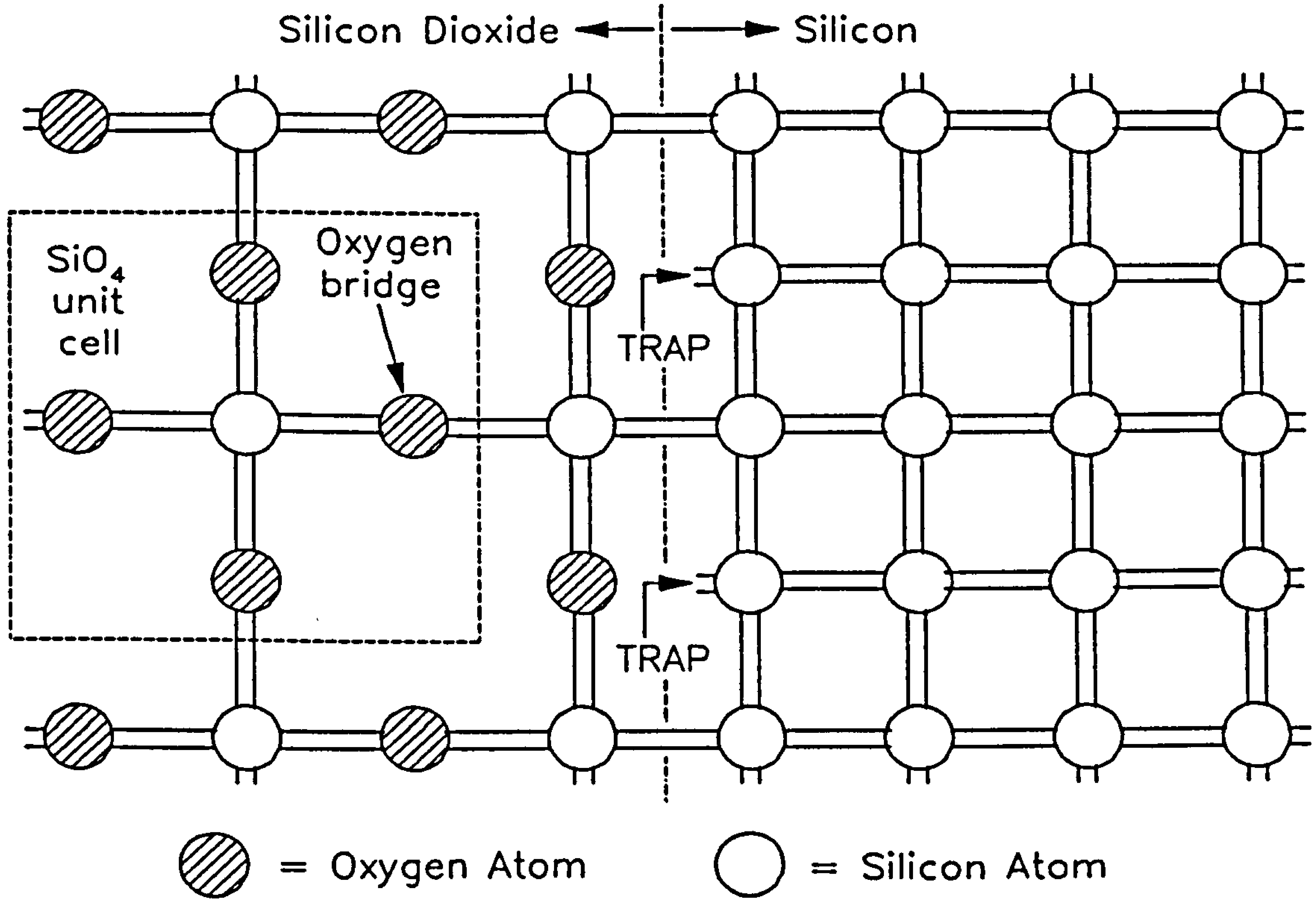


Figure 2.1: Schematic 2-dimensional representation of Si-SiO_2 Interface.

2.2.3 Electronic Structure

Despite the lack of long-range order, vitreous SiO_2 has sufficient short-range regularity to diffract the electron wave functions into a valance and a conduction band. Fig.2.2 shows a simplified density of states function for SiO_2. Its main feature is its 8.8eV forbidden gap [12], which is amongst the widest occurring anywhere in nature. Thermal electron-hole pair generation across this energy gap is practically non-existent at room temperature, rendering the material a phenomenally good insulator. However, the structure can conduct electricity by electron injection into the conduction band or hole injection into the valence band.

The valence and conduction band edges of vitreous SiO_2 are poorly defined in the energy domain and 'tails' of localised and semi-localised states extend approximately 4.7meV into the forbidden gap [5,13]. These states are introduced by the natural angular variation and elongation of bonds associated with a non-crystalline structure [12]. They form part of the

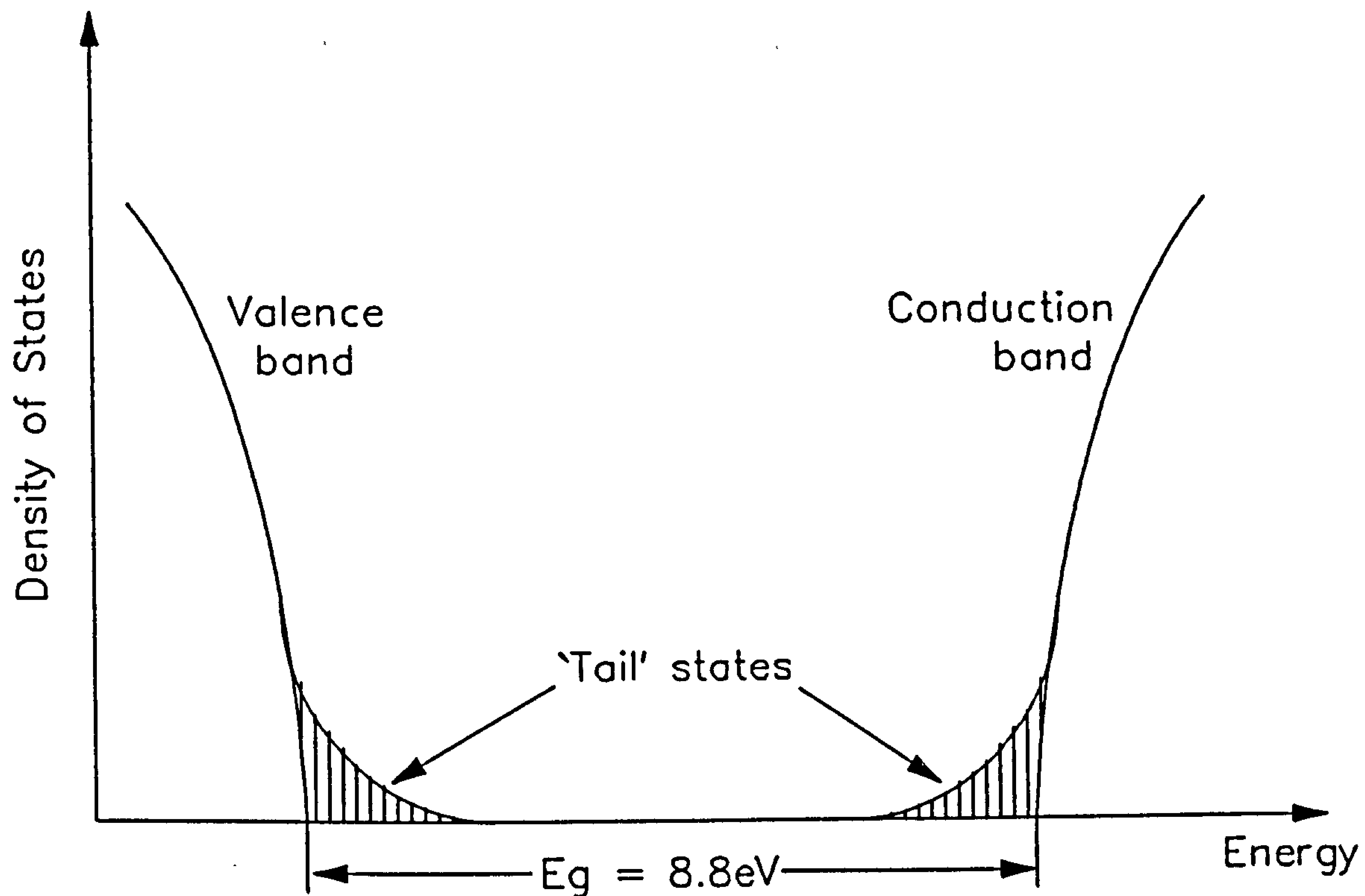


Figure 2.2: Density-of-States Function for Vitreous SiO_2.

group of *Intrinsic* defects discussed later the next section.

The conduction band is believed to be approximately parabolic for low energy electrons. The dispersion relationship between an electron's wave-number k and energy E for a parabolic band is given by

$$k^2 = \frac{2a}{\hbar^2}(E - E_c) \qquad 2(1)$$

where a is a constant. The effective electron mass $m_{ox}^* = \hbar^2/(d^2E/dk^2)$ is equal to a for all values of E above and below E_c. This simple model suffers from the disadvantage that it predicts no valence band, giving imaginary values of k for all energies below E_c. This is remedied by using the Franz 'two-band' dispersion model [14]:

$$k^2 = \frac{2a}{\hbar^2}(E - E_c)\left(1 + \frac{E - E_c}{E_g}\right) \qquad 2(2)$$

where E_g is the width of the forbidden gap. In this case m_{ox}^* is only equal to a at the conduction band edge, its value increasing for $E>E_c$ and decreasing for $E<E_c$. At the energy-gap midpoint $E_i=E_c-E_g/2$, the effective mass changes sign and becomes negative. (It should be noted that the Franz model is purely empirical, with no justification other than the fact that it predicts two energy bands.)

Experiments show that low energy electrons (<2eV) in SiO_2 have an effective mass of approximately $1.3m_0$ falling to $1.0m_0$ at higher energies [6]. These electrons drift freely under the influence of an electric field with a mobility of about $20cm^2/V/sec$ [5,15,16], which decreases with increasing temperature [5]. There are several conflicting estimates of the electron mean free path length in SiO_2, most of them varying between 0.1nm and 3nm [6].

Holes are believed to be localised in the non-bonding 2p oxygen orbitals which form the top of the valence band, distorting the equilibrium positions of nearby nuclei and thereby forming small polarons [5,17,18]. Hopping of holes between neighbouring orbitals is characterised by the small polaron formation time ($\sim 10^{-12}$s) and the hole mobility is therefore low. According to Hughes [17], the hole mobility is $2.10^{-5}cm^2/V/sec$, and it appears to increase with temperature [5]. Estimates of effective hole mass vary between $1m_0$ and $10m_0$ [6].

2.2.4 Bulk Oxide Defects

Structural defects in SiO_2 play an extremely important role in the behaviour of MOS structures. Several species of defect exist, which may be classified into three main groups: *point defects* which exist at single atomic sites, *complex defects* which are clusters of point defects, and *Microheterogenities* which are devitrified regions in an otherwise non-crystalline oxide [12].

The 'point defect' category can be subdivided into two sub-species: *Intrinsic* and *Extrinsic*. An intrinsic defect can be defined as any feature which disturbs the periodic nature of the silica structure, and may include a broken, strained or elongated bond, a non-bridging oxygen atom or a silicon/oxygen vacancy (sometimes called an *interstitial*) [12]. These defects reflect the oxide quality, and depend not only upon the oxidation and annealing conditions but also the effects of the subsequent processing steps [1,6].

A single electron on an unconnected bond (known as an E' centre [5]) is clearly able to accept or donate an electron and appears as a *trap* state in the forbidden gap. A trap with a tendency to donate an electron to the conduction band, becoming positively charged, is said to be *donor* type. Similarly a trap which accepts an electron to become negatively charged is called an *acceptor* [1]. These traps have been experimentally studied using a number of techniques including capacitance-voltage (C-V) profiling [19] and transient photodepopulation [20].

Trap occupation kinetics can be modelled by the first order rate equation [6,12]

$$\frac{\partial n}{\partial t} = \frac{J\sigma}{q}\frac{v_{th}}{v_d}(N-n) \qquad 2(3)$$

where N is the trap density, n is the trapped electron density, J is the current density in the oxide conduction band, v_{th} is the thermal electron velocity, v_d is the electron drift velocity and σ is the trap capture cross section. The latter parameter, which represents the area surrounding a trap through which an electron must pass in order to be captured, depends upon the electron kinetic energy and hence the electric field. It also depends upon whether the trap is electrically neutral, negatively charged (coulombic repulsive) or positively charged (coulombic attractive) [6]. Typical σ values are given in Table 2.1, together with their relative field sensitivities.

Table 2.1: Data on Capture Cross Sections [6]

Type of Trap	Capture Cross Section σ (cm^2)	Relative Field Dependence of σ
Coulombic Attractive Trap (CAT)	10^{-12} - 10^{-14}	High
Neutral Trap State (NTS)	10^{-14} - 10^{-18}	Medium
Coulombic Repulsive Trap (CRT)	10^{-18} - 10^{-21}	Low

Extrinsic defects are associated with the presence of foreign ions, typically Na^+, appearing as positive charge in the oxide [12]. These ions are relatively mobile and can drift through the oxide under the influence of an applied field. This ionic charge is usually modelled in terms of an 'effective' oxide charge density Q_{ox} (C/cm^2) appearing at the silicon-oxide interface.

Trap states can significantly affect the electronic properties of the oxide. If the trap density is high, carriers can tunnel or 'hop' between traps, providing an additional mechanism for electronic conduction [13,21,22]. If the traps are very closely spaced, their wave functions may overlap, forming a 'metallic' energy band within the insulator [13]. Trapped electrons can also be thermally emitted into the conduction band under a high electric field, an effect known as Poole-Frenkel emission [13].

2.2.5 The Si-SiO_2 Defects

A cleaved Si crystal has a high density of surface trap states introduced by incomplete or 'dangling' orbitals. These behave as interband surface states or traps and have the same areal density as the surface atoms (i.e. $\sim 10^{15}$traps/cm^2). Although a thermally grown SiO_2 layer passivates the surface, completing most of the surface orbitals, the periodic mismatch between the Si and SiO_2 networks implies that some Si bonds remain incomplete (see Fig.2.1). Stretched, twisted and dangling bonds do indeed appear on passivated Si-SiO_2 interfaces, introducing an 'interface' trap density of the order of 10^{12}traps/cm^2 [23]. Extrinsic defects associated with ionic contaminants can also exist [24] and although both donor and acceptor species exist at the interface, the donors are more numerous and the net surface charge is positive [6]. Hence the extrinsic surface state charge can be modelled by a positive charge density Q_{ss}.

Interface state density is affected by the type and density of Si dopants, boron giving a higher trap density than phosphorus or arsenic due to an introduced oxygen deficiency [6]. Interface states can sometimes be annealed away by post-oxidation low-temperature thermal processing [6], although this can *increase* rather than decrease the defect density [25,26]. This is in some cases due to void formation under the deoxidation reaction ($Si+SiO_2 \rightarrow 2SiO\uparrow$) across the Si-$SiO_2$ interface. It can be prevented by increasing the O_2 content of the annealing ambient, thereby causing compensatory reoxidisation [27]. Interface traps are also generated under electric field stress, although these may have slightly different properties from the pre-existing defects [28].

The charged interface states, together with unoxidised Si^+ ions, present a charge storage capacity at the interface, which can be detected using charge-pumping analysis [29], a.c. conductance measurement [30] or deep-level transient spectroscopy [31]. Traps also cause enhanced scattering in the MOSFET inversion layer, and their density can therefore be measured in terms of the thermal 1/f noise [32]. The trapped interface charge density is denoted Q_{ss} (C/cm^2).

2.2.6 The Gate-SiO_2 Interface

The most common materials used for gate fabrication are aluminium and polysilicon (polycrystalline silicon). Both materials are applied to the SiO_2 by vapour deposition, and exhibit a polycrystalline structure. The polysilicon/SiO_2 junction can be considered to display similar properties to the thermal Si/SiO_2 interface. The energy barrier between the polysilicon

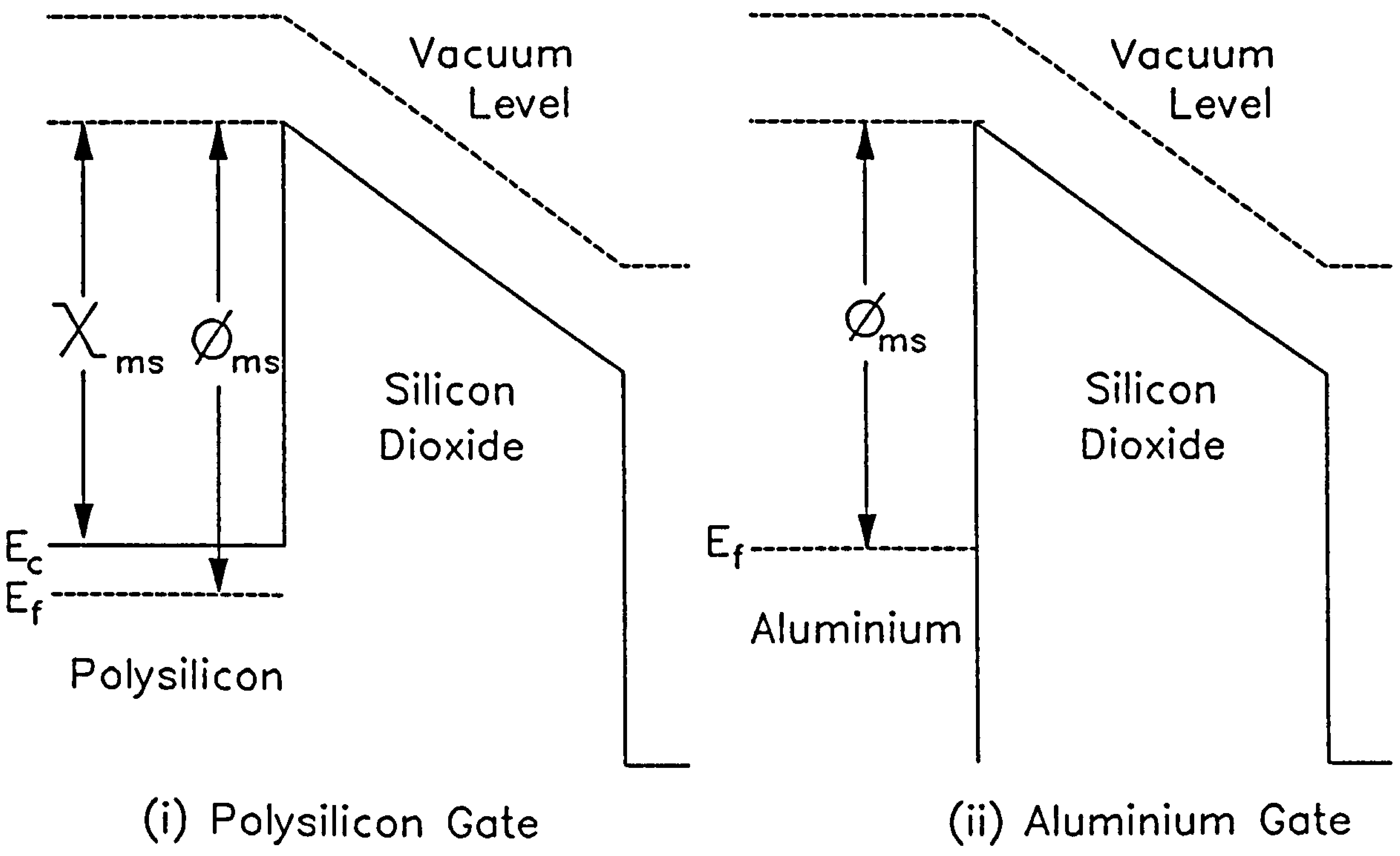


Figure 2.3: Energy band diagrams for PolySi/SiO_2 and Al/SiO_2 MOS Structures.

gate and the SiO_2 can be characterised in terms of the effective work function ϕ_{ms}, i.e. the potential difference between the silicon Fermi-level and the oxide conduction band. Alternatively the effective electron affinity χ_{ms}, i.e. the potential difference between the oxide and silicon conduction band edges can be used. Both parameters, shown schematically in Fig.2.3(i), can be measured using the internal photoemission technique [6,16]. The Si-SiO_2 interface appears to have a ϕ_{ms} of approximately 3.15eV, a value which varies slightly with the Si crystal orientation [6]. The quantum mechanical tunnelling experiments of Lenzlinger and Snow [14] suggest an additional dependence of ϕ_{ms} upon temperature. If the gate material is metallic [Fig.2.3(ii)], the effective electron affinity is meaningless since the conduction band edge bears little relation to the energy of cathode electrons. The effective work function is therefore the only meaningful parameter, equalling 4.1V for an aluminium gate material.

2.3 Theory of MOSFET Operation

2.3.1 The Inversion Layer [1,8,33,34]

Fig.2.4 shows a simplified energy band diagram of the silicon-oxide interface in an n-channel MOSFET. The surface is assumed to be in thermal equilibrium, and can therefore

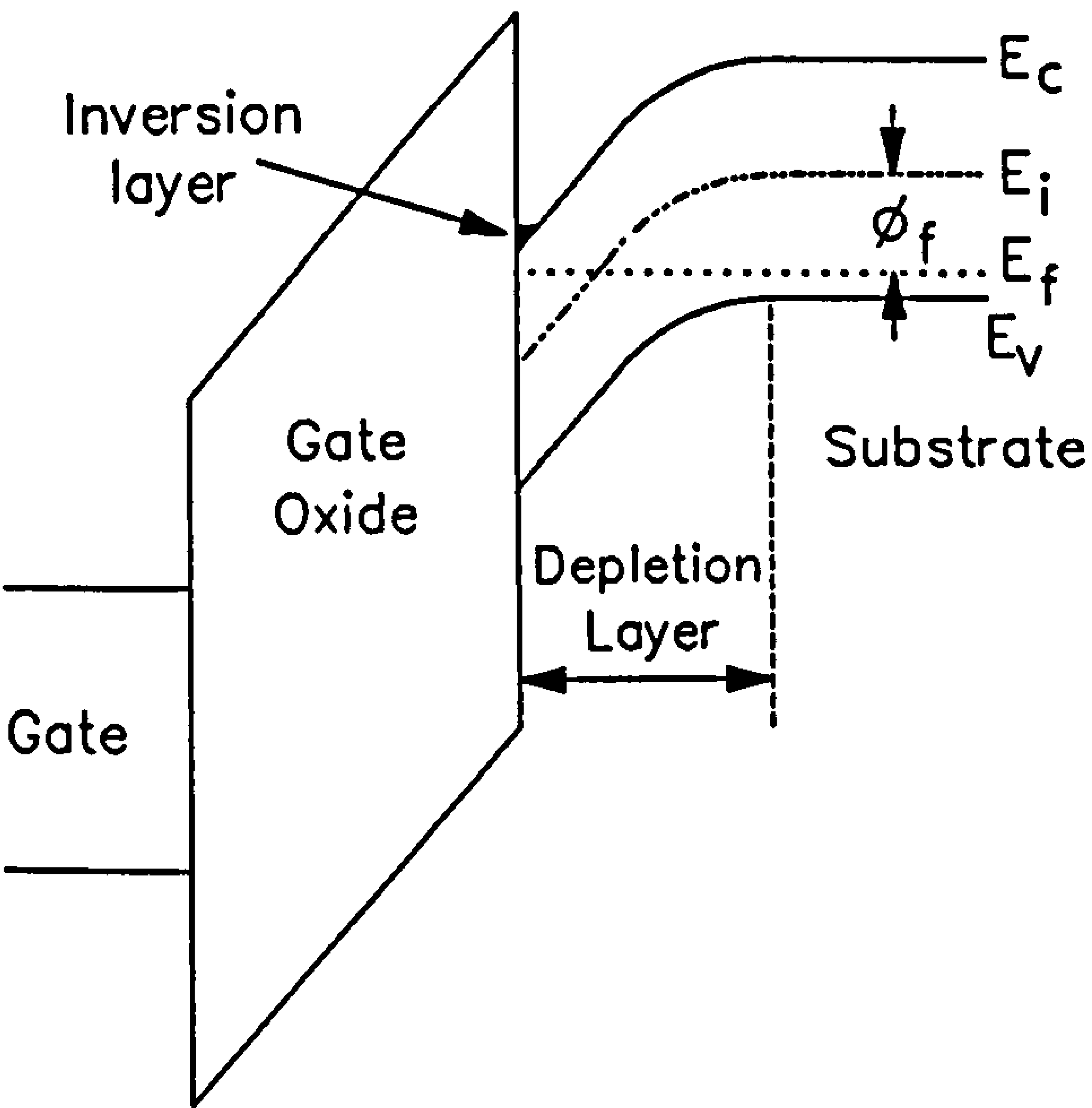


Figure 2.4: Formation of Depletion and Inversion Layers in an MOS System.

be modelled by a single Fermi-level E_f. A positive gate bias bends the silicon valence band edge E_v away from E_f, reducing the valence band hole density and creating a negatively charged depletion layer of width W_d. Further increases in the gate bias bring E_c close to E_f, inducing conduction band electrons and forming an inversion layer. 'Strong' inversion is defined to occur when the surface electron density exceeds the bulk hole density, i.e. when the surface potential $\phi_s > 2\phi_f$, where ϕ_f is the bulk Fermi potential ($=[kT/q]\log_e[N_a/n_i]$). Once strong inversion is achieved, the depletion layer width ceases to increase with increasing gate bias, and saturates at a value $W_{d(max)}$. The gate bias for which $\phi_s = 2\phi_f$ is called the *strong inversion threshold voltage* and is denoted V_T. In n-channel devices, V_T is positive for an E-mode device and negative for a D-mode device (the reverse is true for a p-channel device). V_T is given by

$$V_T = \frac{\Delta \Phi_{ms}}{q} - \frac{Q_{ox}}{C_{ox}} - \frac{Q_{ss}}{C_{ox}} - \frac{Q_d}{C_{ox}} + 2\Phi_f \qquad 2(4)$$

where $\Delta\phi_{ms}$ is the difference between the gate and substrate effective work functions, Q_{ss} is the surface state charge density, Q_d is the maximum depletion layer charge density and Q_{ox} is the effective internal oxide charge density due to ions and bulk-oxide traps, referred to the Si-SiO_2 surface. According to the depletion approximation, $Q_d = -qN_aW_{d(max)} = -2(q\epsilon_0\epsilon_{si}N_a\phi_f)^{1/2}$ (Section 2.3.2).

Another commonly used parameter is the *flat-band voltage* V_{FB}, i.e. the voltage which must be applied in order to make the oxide energy bands flat and the oxide voltage zero. V_{FB} is given by

$$V_{FB} = \frac{\Delta \phi_{ms}}{q} - \frac{Q_{ox} + Q_{ss}}{C_{ox}} \qquad 2(5)$$

The inversion layer does not appear instantaneously as the bias voltage is applied. It is gradually built up from electron-hole pairs thermally generated in the depletion layer or electrons wandering into the depletion region from the bulk silicon. Inversion layer growth is governed by the *minority carrier response time* τ_{min}, which is usually somewhere between 10msec and 1sec [33].

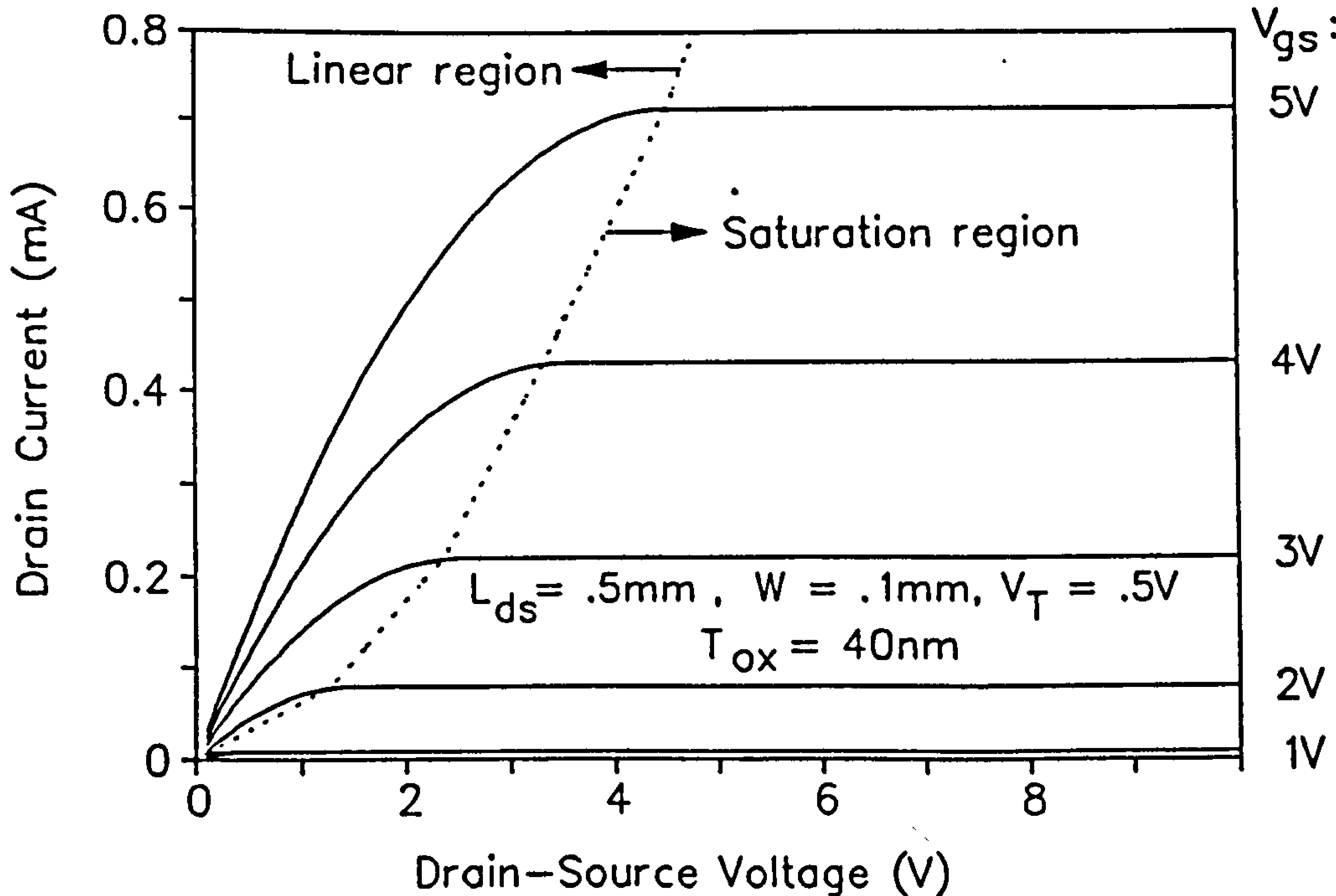


Figure 2.5: MOS Characteristics Predicted by First-Order Model

2.3.2 First Order MOSFET Model [1,8,34,35]

MOSFETs are often characterised in terms of their drain-source characteristics, in which drain current I_d is plotted against V_{ds} for constant values of V_{gs} (Fig.2.5). The characteristics can be divided into two sections: the 'linear' region, in which I_d increases with increasing V_{ds} and the 'saturation' region in which I_d remains constant.

Fig.2.6 shows a cross section of a MOSFET. L_{ds} is the channel length and x is the distance of any point along the channel from the source. Let V(x) and Q(x) represent the voltage and the inversion-charge profiles across $0<x<L_{ds}$. Below the drain saturation voltage $V_{ds(sat)}$, the characteristics can be modelled within the *gradual channel approximation*.

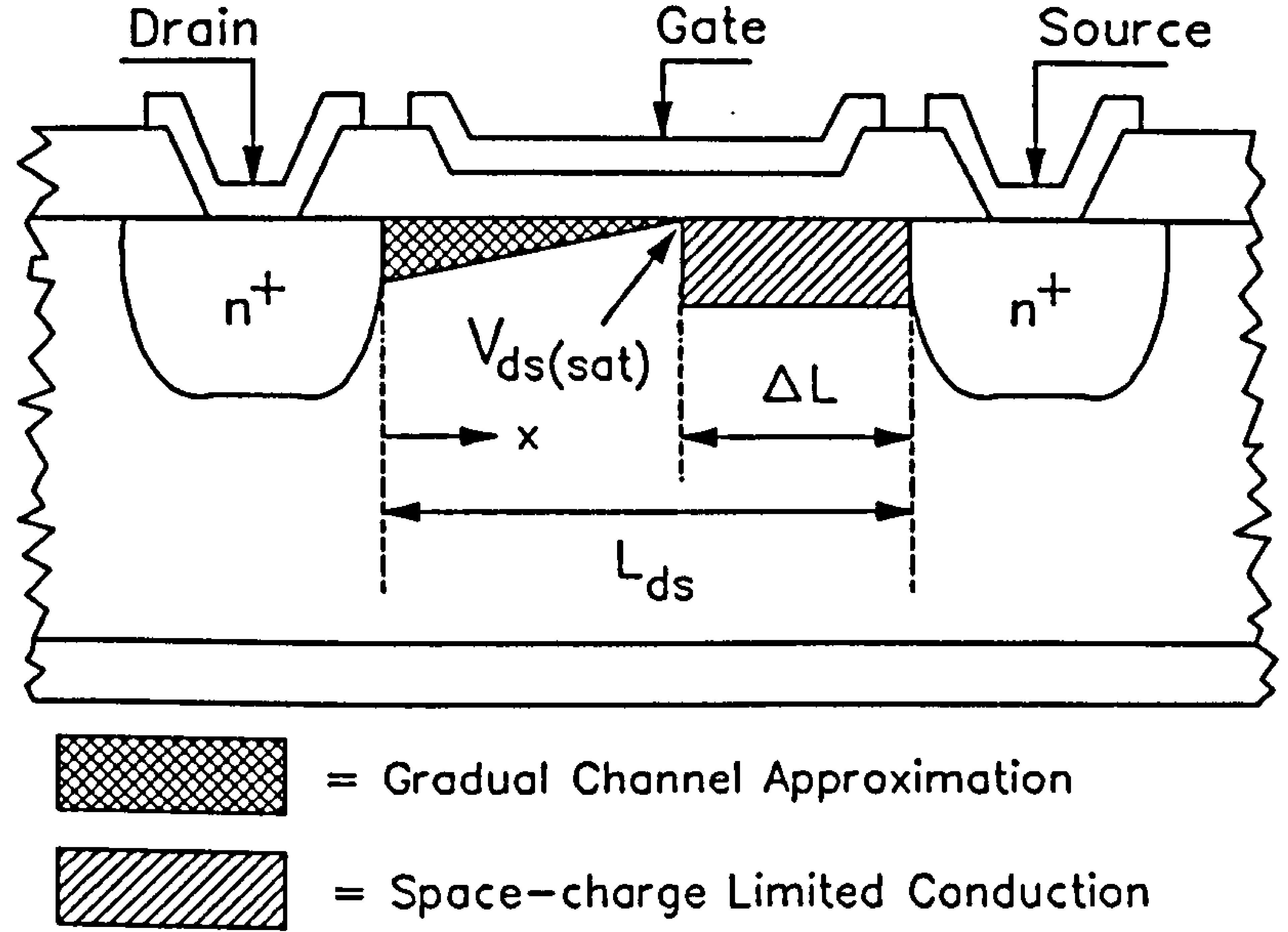


Figure 2.6: MOSFET biased in saturation mode.

According to Ohm's law, the drain current I_d is related to Q(x) (C/cm^2) by the equation

$$I_d dx = \mu W Q(x)\, dV(x) \qquad 2(6)$$

where μ is the Si surface carrier mobility and W is the channel width. Q(x) can be determined by considering that the gate voltage V_{gs} must be equal to the sum of the flat-band voltage V_{FB}, the Si surface potential $\phi_s(x)$ and the oxide voltage $V_{ox}(x)$. In strong inversion, $\phi_s(x)=2\phi_f+V(x)$ and $V_{ox}(x)=-[Q_d(x)+Q(x)]/C_{ox}$. Furthermore, the depletion approximation requires that $Q_d(x)=-(2q\epsilon_0\epsilon_{si}N_a[2\phi_f-V(x)])^{1/2}$. Combining all these expressions and solving for Q(x) yields

$$Q(x)=C_{ox}\left[V_{gs} - \frac{\Delta\phi_{ms}}{q} + \frac{Q_{ox}+Q_{ss}}{C_{ox}} - 2\phi_f - V(x) - \frac{\sqrt{2q\,\epsilon_{si}N_a(2\phi_f+V(x))}}{C_{ox}}\right] \qquad 2(7)$$

The final term in Eqn.2(7) represents the depletion layer charge density Q_d. If the effect of V(x) upon Q_d is ignored, then Eqns.2(4) and 2(7) can be combined to yield

$$Q(x) \approx C_{ox}\left[V_{gs} - V_T - V(x)\right] \qquad 2(8)$$

This approximation shall be used for the remainder of this chapter. Inserting Eqn.2(8) into Eqn.2(7) and integrating yields the relationship between V_{ds}, V_{gs} and I_d. In the linear region, the inversion layer extends throughout the channel, and the integration can be performed from x=0 to $x=L_{ds}$. In a large dimension device in which the electric fields are low, μ has a constant value μ_0 and the equation becomes

$$I_d = \frac{\mu_0 W C_{ox}}{L_{ds}} V_{ds}\left[V_{gs} - V_T - \frac{V_{ds}}{2}\right] \; ; \quad 0 \le V_{ds} \le V_{ds(sat)} \qquad 2(9)$$

When V_{ds} reaches the saturation voltage $V_{ds(sat)}=(V_{gs}-V_T)$, insufficient gate-channel voltage exists at the drain end of the channel to support inversion. This condition, termed *pinch-off*, is characterised by a saturation drain current $I_{d(sat)}$, given by

$$I_{d(sat)} = \frac{W \mu_0 C_{ox}}{2 L_{ds}} (V_{gs} - V_T)^2 \; ; \quad V_{ds} > V_{ds(sat)} \qquad 2(10)$$

Above $V_{ds(sat)}$, the inversion layer recedes from the source, reducing the effective channel length (see Fig.2.6). In large dimension MOSFETs, this effect is negligible and the drain current saturates at $I_{d(sat)}$. This is called the *saturation* region, and it forms the operating condition for most analogue circuits. Saturation performance is characterised by the *transconductance* g_m, defined as $\partial I_{d(sat)}/\partial V_{gs}$. Differentiation of Eqn.2(10) w.r.t. V_{gs} yields the following expression for g_m:

$$g_m = \frac{W \mu_0 C_{ox}}{L_{ds}} (V_{gs} - V_T) \qquad 2(11)$$

Fig.2.5 shows a typical set of characteristics predicted by Eqns.2(9) and 2(10).

2.4 Miniaturisation of the MOSFET

2.4.1 Device Scaling [8,34]

The 'first-order' model developed in the previous section can be used to predict the effects of device miniaturisation. If the device voltages, currents and dimensions (lateral and horizontal) are multiplied by a factor α (<1) and the substrate doping is multiplied by $1/\alpha$ then the power per device becomes multiplied by α^2 and the speed-power product (a figure of merit equal to the product of the gate time delay and the gate power dissipation) by α^3. Miniaturisation therefore improves not only the production costs (less silicon per device), but also the power dissipation and the operating speed.

The α^2 and α^3 factors are not always attainable in practice. Firstly, there are restrictions on the scaling of the supply voltage, which must be significantly higher than the thermal potential kT/q. The supply voltage V_{dd} should ideally be +5V in order to maintain

compatibility with other ic technologies (e.g. TTL). Holding V_{dd}=5V causes the speed-power product to decrease by a factor α^2 instead of α^3, increasing the power dissipation per unit area of silicon by a factor of $1/\alpha$ (>1) and thereby introducing a heat-removal limitation.

2.4.2 Second Order Effects [8,34]

Channel lengths below 5μm introduce problems at a more fundamental level. One such problem is *channel length modulation* which occurs during saturation when the pinched-off channel region ΔL (which increases with increasing V_{ds}) becomes comparable to the drain-source length L_{ds}. This makes the effective channel length $L_{ds(eff)}=L_{ds}-\Delta L$ decrease with increasing V_{ds}, causing $I_{d(sat)}$ to increase after pinch-off. Also, if V_{dd}=5V is maintained for small dimensions, the large electric fields cause a high degree of inversion-layer band-bending, resulting in the quantisation of carrier wave-functions into two-dimensional sub-bands [5]. Due to Si-SiO_2 surface roughness and/or surface charge, the carriers are less mobile in the sub-bands than they are in the bulk silicon [5] and their mobility decreases with increasing electric field [36]. These effects invalidate the simple first order MOSFET model upon which the above scaling theory is based.

Further limitations are imposed by the *"punch through"* mechanism [37]: if $L_{ds(eff)}$ becomes zero, the entire channel becomes pinched off, causing a rapid increase in I_d. Another dilemma is the *second gate* effect, whereby the field induced by V_{ds} can invert the channel even when $V_{gs}<V_T$. This effect increases with increasing T_{ox}/L_{ds}, thereby inhibiting the reduction of L_{ds}. The threshold voltages of short-channel structures are very difficult to control during fabrication, due to fringing fields and the relatively large source and drain depletion regions. Miniaturisation also introduces problems of reliability; the high electric fields rendering the gate oxide more prone to dielectric breakdown. High fields also produce 'hot' inversion layer electrons which can be injected into the oxide conduction band, charging the oxide and causing instability in V_T. Additional size restrictions are imposed by the wavelength of light in photolithographic fabrication.

Finally, if the drain and source in a MOSFET are in very close proximity, a parasitic bipolar transistor can be induced, as shown in Fig.2.7(a). In this simple model, I_c represents the 'collector' current of the parasitic transistor while R_1, R_2 and R_3 represent the bulk substrate. Under large voltages, the impact ionisation in the drain region creates a substrate hole current I_c(M-1) (where M is the multiplication factor). Some of these holes flow laterally and reach the 'base' (via R_2) where they recombine with electrons and increase the 'collector' current. This clearly creates a condition for positive feedback. The subsequent collapse of

the drain-source voltage with increasing drain current is known as *snapback* (Fig.2.7(b)).

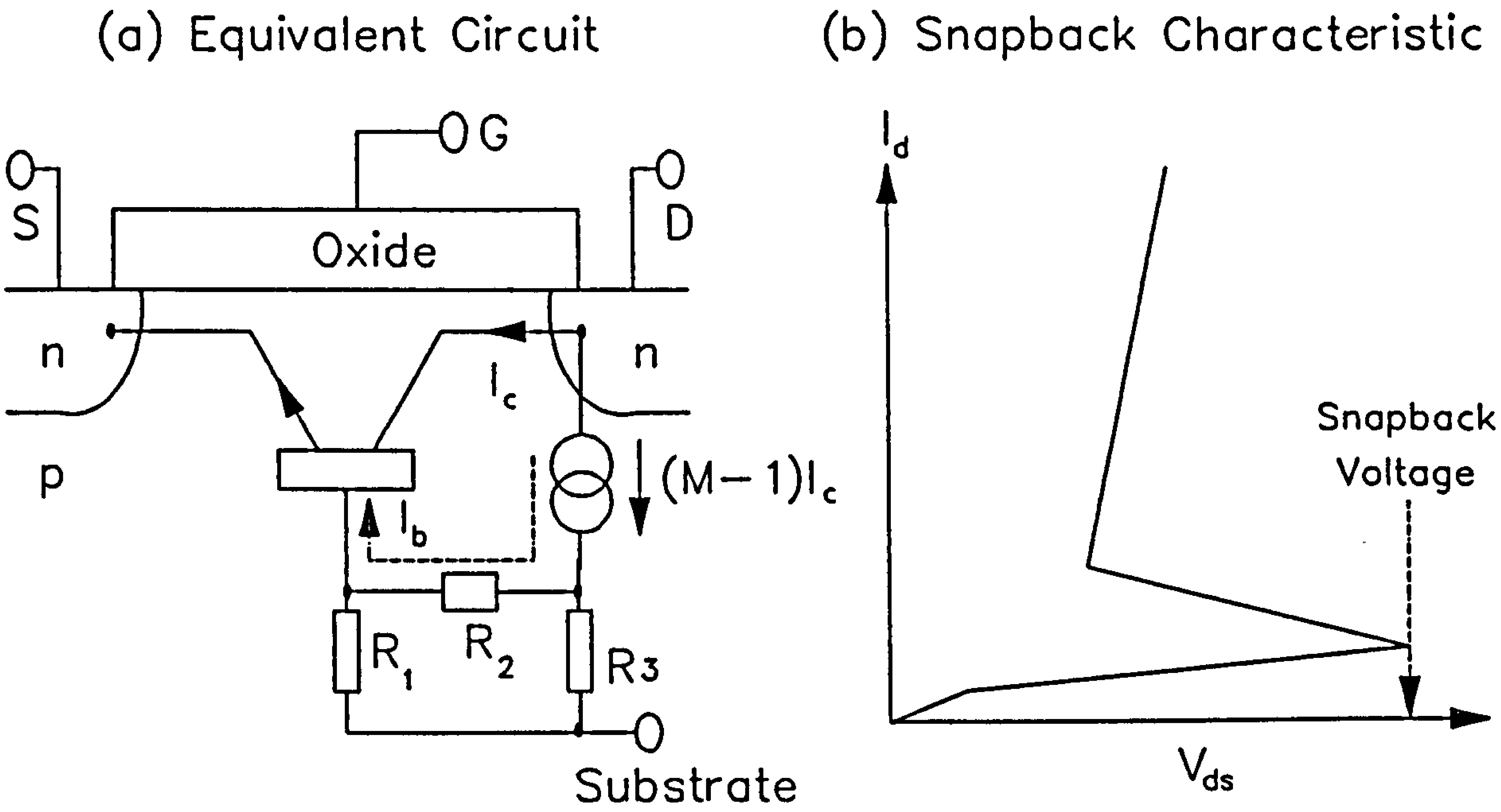


Figure 2.7: (a) Circuit model of bipolar action in small dimension MOSFET, (b) Schematic representation of snapback characteristic.

2.4.3 Second Order MOSFET Models [35-41]

Numerous second-order analytical and semi-analytical techniques have been developed, many of which are concerned with the modelling of surface inversion-layer mobility under high electric fields. The field in the channel has two orthogonal components: a transverse component F_y, associated with V_{gs}, and a longitudinal component F_x, caused by V_{ds}. The mobility can be modelled by the empirical equation

$$\mu = \frac{\mu_0}{\left[1 + \frac{F_y}{F_a}\right]\left[1 + \frac{F_x}{F_b}\right]} \qquad 2(12)$$

where μ_0 is the bulk Si mobility and F_a and F_b are constants. An advantage of this model is that it permits an analytical solution of the gradual channel approximation Assuming for simplicity that $F_y=(V_{gs}-V_T)/T_{ox}$ throughout the channel, Eqn.2(12) can be combined with Eqns.2(6-8), yielding the following expression for drain current in the linear region

$$I_d = \frac{I_{df}}{\left[1 + \frac{(V_{gs} - V_T)}{F_a T_{ox}}\right]\left[1 + \frac{V_{ds}}{L_{ds} F_b}\right]} \qquad 2(13)$$

where I_{df} is the equivalent value of I_d predicted by the first order model (Fig.2.8).

The decrease in mobility with increasing field makes the electron velocity v saturate at a constant value v_d for high V_{ds}. The saturation voltage $V_{ds(sat)}$ is defined as the drain voltage at which v_d is attained at the drain end of the channel. For $V_{ds} > V_d$, the $v=v_d$ point moves a distance ΔL towards source, reducing the effective channel length to $(L_{ds}-\Delta L)$ and causing channel length modulation. Channel-length modulation in the saturation-region can be modelled numerically [39], or using Schichman's empirical technique [35] of multiplying $I_{d(sat)}$ by $[1 + \ell V_{ds}]$ (ℓ is an adjustable fitting parameter). Several techniques have been employed to calculate the saturation current as a function of V_{gs} and V_{ds} [eg.39], most of them requiring lengthly numerical solutions. The most noticeable feature of the characteristics predicted by these models is the almost constant g_m which, unlike that of the first-order model, is independent of V_{gs}.

More accurate models of short-channel structures employ numerical finite element techniques [34]. Several computer packages such as BAMBI and MINIMOS are available for such simulations.

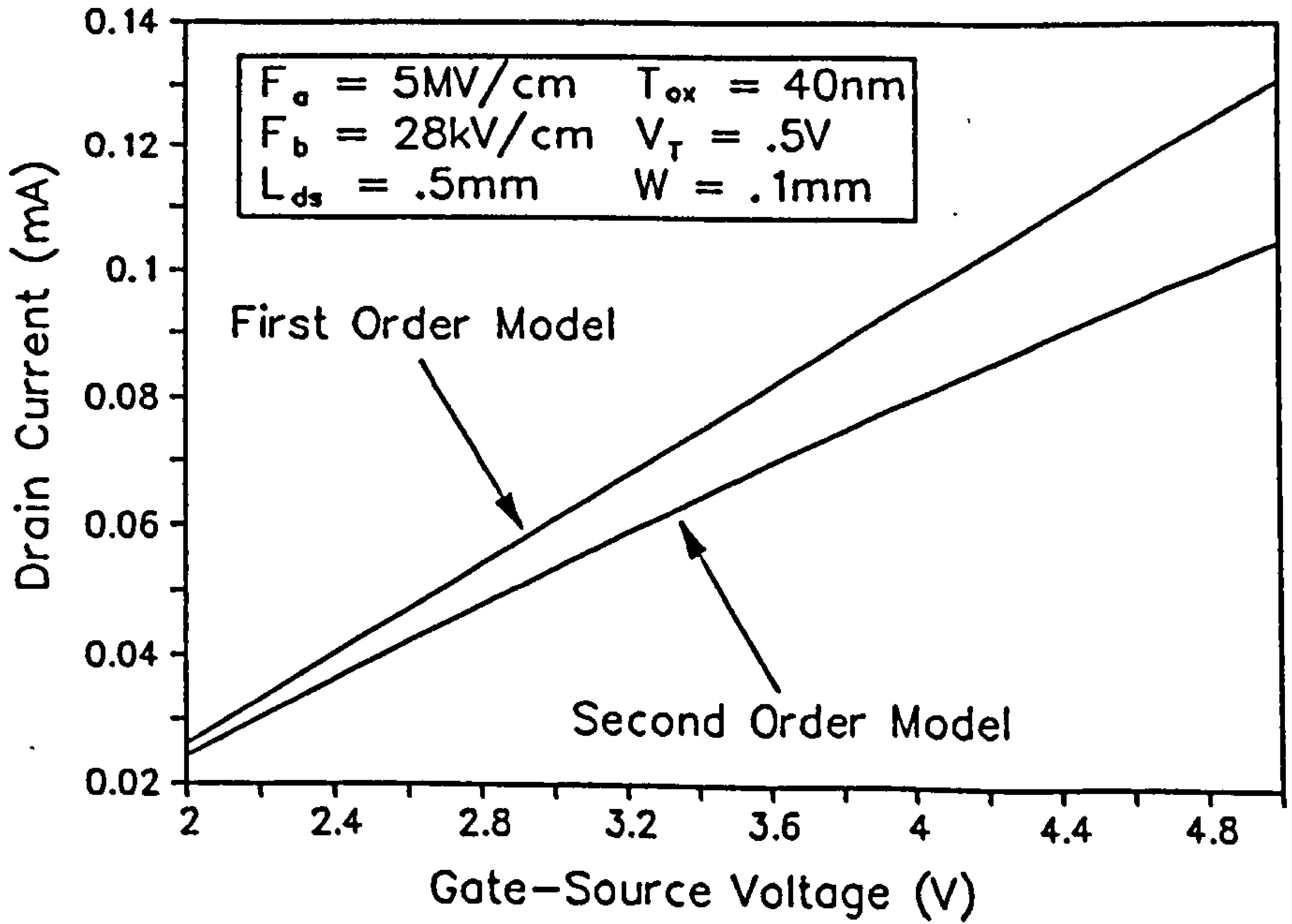


Figure 2.8: Transfer Characteristics of small dimension MOSFET biased in Linear Region, as predicted by first and second-order models.

2.5 MOS Technologies

The previous section examined the benefits of miniaturisation in terms of price, power consumption and operating speed. The development of MOS technology has therefore involved a thrust towards smaller devices, producing the 'Moore's law' exponential increase in integrated circuit complexity with time [8].

During this push towards miniaturisation, numerous device structure and layout philosophies have been employed, each imposing its own advantages and limitations in size and performance. This section briefly describes some the major technologies and discusses their advantages and drawbacks.

2.5.1 Planar Bulk-Silicon Technologies

Most MOS devices are fabricated on the surfaces of bulk silicon 'wafers', coated in photoresist and exposed to optical radiation via a 'mask' in order to define the device geometries. This process is known as *photolithography*. The wafers themselves are slices of a cylindrical single-crystal ingot pulled from a silicon melt (the Czochralski method). A diamond tipped saw blade is then used to divide the ingot into wafers which are finely polished, producing an optically flat surface.

A problem with bulk silicon in high density integrated circuits is caused by possible interference between neighbouring devices via parasitic inversion layers. This can be prevented by increasing the field-oxide thickness and surface doping or by surrounding the devices by bars of heavy doping called *guard rings* or *channel stops*. Although such strategies are undesirable (they consume valuable chip area which might otherwise be used for active circuitry), they are often unavoidable unless an insulating substrate is used.

2.5.1.1 PMOS and NMOS

The earliest form of MOS technology used in IC fabrication was PMOS, which utilised the p-channel MOSFET. PMOS circuits are readily fabricated without the aid of ion implantation and enjoy a high noise immunity (at the expense of a large supply voltage). However, their operating speed is limited by the low Si hole mobility. In the early 1970s, the market became dominated by the NMOS process. This technology, offering a reduced supply voltage and higher switching speed employs the n-channel MOSFET. Fabrication uses a self-aligned polysilicon gate structure and a number of ion implantation stages. The voltage

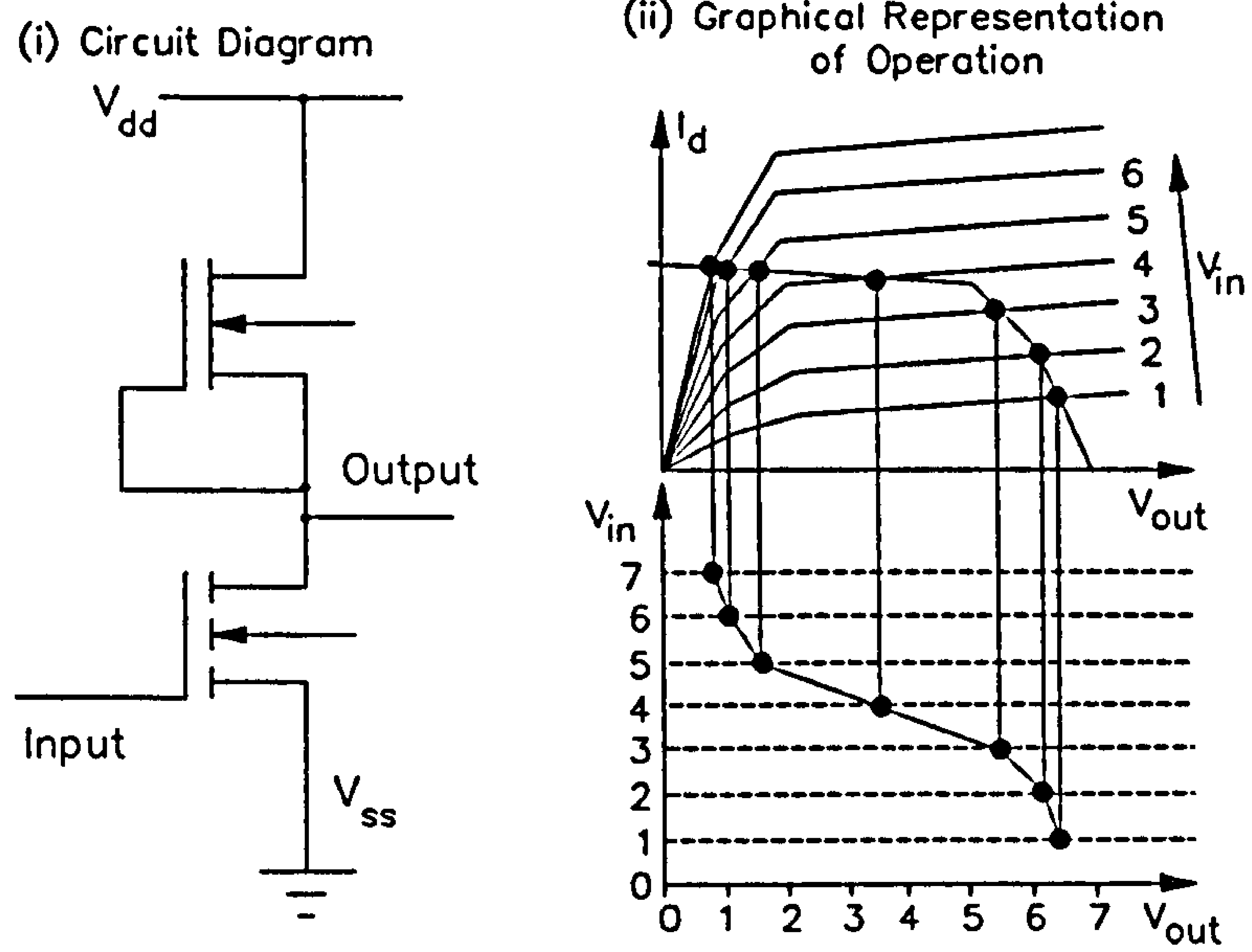


Figure 2.9: Circuit for NMOS invertor gate and logic symbol.

supply can be reduced to +5V, allowing compatibility with bipolar circuitry.

NMOS and PMOS can be used to fabricate logic gates and dynamic random access memories (DRAMs). Fig.2.9(i) shows a simple NMOS inverter (NOT) gate consisting of two n-channel transistors, one of which is E-Mode and the other D-Mode (the D-Mode device is permanently biased in the 'on' state). Zero input voltage (logic '0') causes zero drain current to flow in the E-mode transistor, causing the voltage across the D-mode device to be zero and the output voltage to follow V_{dd} (logic '1'). A positive input voltage ('1') switches the E-Mode transistor on, causing a voltage to appear across the D-Mode transistor, thereby reducing the output voltage to logic '0'. Fig.2.9(ii) shows how the gate transfer function can be constructed from the superimposition of the two transistor characteristics. It is clear that negligible current is consumed when the output is '1', but a sizable I_d flows when the output is '0'.

2.5.1.2 Complementary MOS (CMOS)

A CMOS integrated circuit consists of adjacent p and n-channel E-mode MOSFETs. This is achieved by creating a lightly doped p-type 'tub' or 'well' in an otherwise n-type wafer by boron ion implantation and subsequent diffusion. Fig.2.10(i) shows two such devices coupled to form a CMOS inverter gate (NAND and NOR gates can be produced by more elaborate series/parallel arrangements). Both devices are enhancement mode, i.e. V_T is positive for the n-channel device and negative for the p-channel device. A zero input voltage (logic '0') clearly switches the n-channel MOSFET off and the p-channel MOSFET

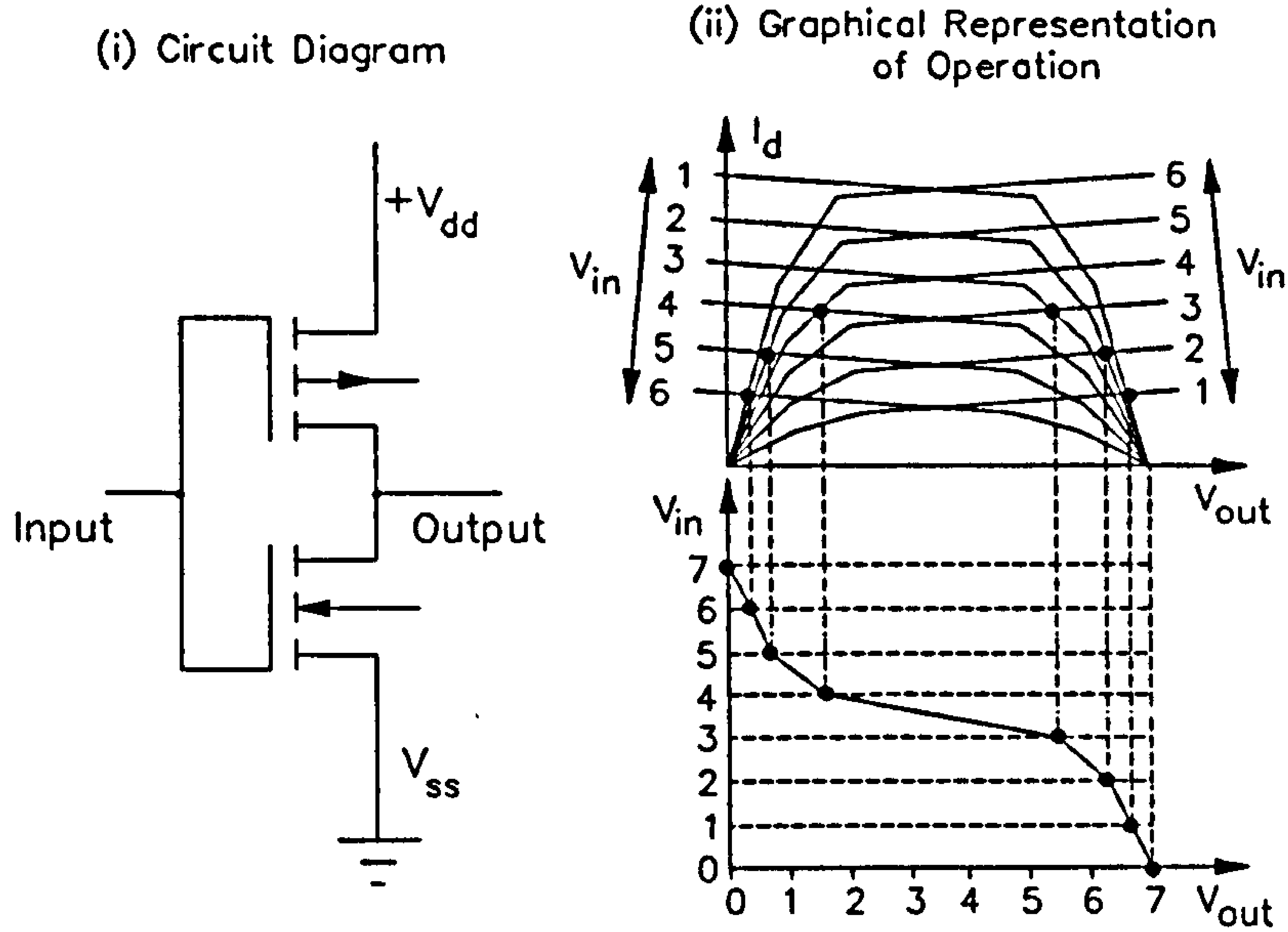


Figure 2.10: Circuit for CMOS invertor gate.

on. The output therefore follows the positive supply voltage, giving a logic '1'. If a positive input voltage is applied the situation is reversed, giving an output '0'. Fig.2.10(ii) shows the construction of the transfer function from a superimposition of the two transistor characteristics. The most important feature of this arrangement is that one transistor is always off and there is never a direct connection between the supply rails (which there is for NMOS when the gate output is '0'). The supply current and power dissipation are therefore very much smaller than PMOS, NMOS and bipolar logic. Another advantage of CMOS is a wide range of supply voltages can be used, ranging between 1.5 and 20V. CMOS also possesses a high noise immunity.

Unfortunately, CMOS is inherently slower than bipolar logic due to the large gate-oxide capacitances, and its greater complexity means a larger chip area per gate. However, improvements in processing have recently made NMOS and CMOS comparable in complexity and CMOS is consequently the most widely used technology in modern integrated circuits.

2.5.1.3 High Performance MOS (HMOS) [8,34]

High Performance MOS technology involves the straightforward miniaturisation of the above structures by means of carefully controlled processing (ie. better mask alignment, cleaner oxides etc.) and multiple channel implant techniques to minimise punch-through and short-channel effects (Section 2.4.2). This has permitted the miniaturisation of standard planar structures down to the 2μm region with the expense of a larger financial investment in fabrication plant. A typical HMOS logic gate has a delay time of approximately 1ns and

a speed-power product of 1pJ. The major disadvantage is the requirement for a lower supply voltage, typically 2-4V, causing incompatibility with standard 5V TTL logic.

The size of HMOS devices is limited by the photolithographic fabrication outlined by Sze [34]. As the feature size approaches the wavelength of visible light (380-780nm), diffraction fringes are produced by the mask, blurring the edges of the regions of exposed photoresist. This problem can be solved by using shorter wavelength radiation such as X-ray or an electron beam, permitting channel lengths in the $0.2\mu m$ region.

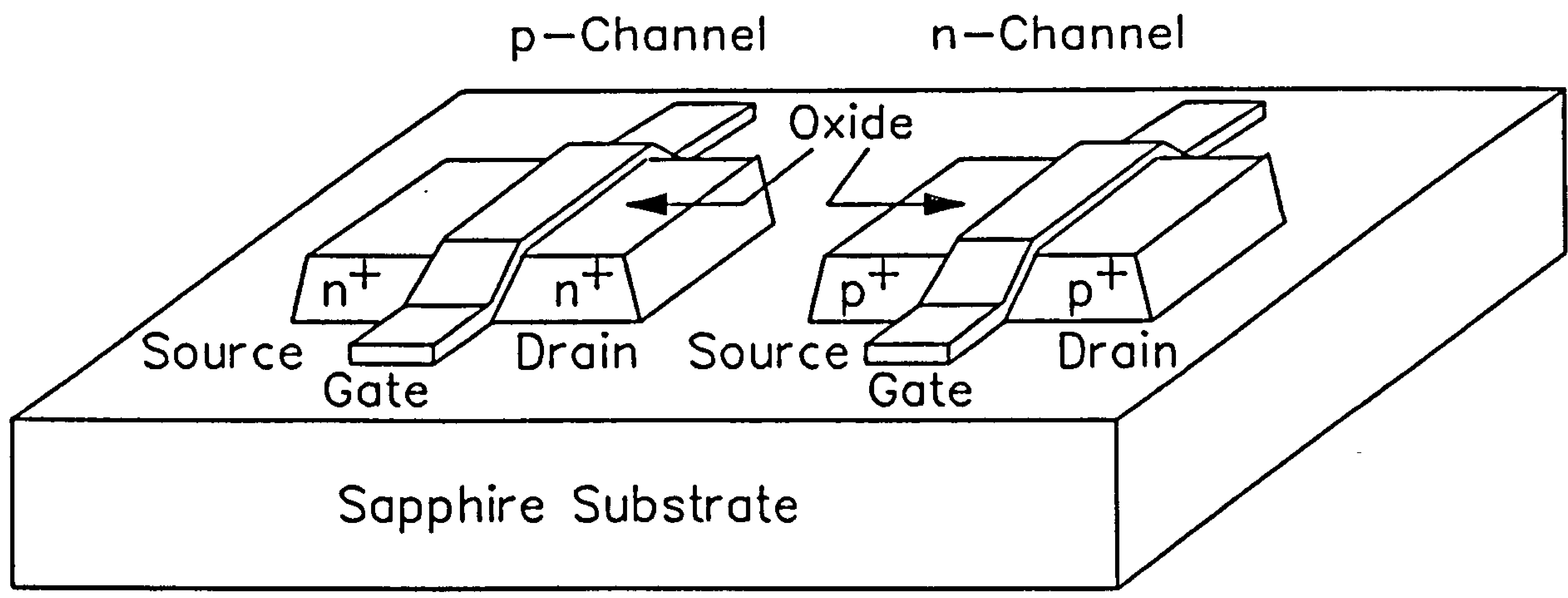


Figure 2.11: Schematic diagram of SOS CMOS structures.

2.5.2 Insulating Substrate Technologies

Various advantages can be gained by fabricating MOS devices as single-crystal silicon *islands* upon an insulating substrate [42]. Such circuits have extremely high temperature tolerance [43], are immune from latchup (see Chapter 3) and are less prone to radiation damage [42]. The Si islands are produced by the vacuum deposition of polysilicon which is later recrystallised using zone-melting recrystallisation (ZMR) [44]. The substrate must have a thermal expansion to match that of the silicon and two such materials are sapphire (Al_2O_3) and silicon dioxide. The corresponding technologies are called *silicon-on-sapphire (SOS)* and *silicon-on-insulator (SOI)*.

Fig.2.11 shows typical p and n-channel SOS transistors in cross section. The islands are composed of p-type silicon, into which p^+ and n^+ source and drain regions are diffused. The gate oxide and gate electrode are fabricated on top of the intermediate p-type channel. Although the p-channel device is constructed as though it were D-mode, a lightly doped channel produces a negative V_T due to the large value of $\phi_{ms}-(Q_{ox}+Q_{ss})/C_{ox}$ relative to $2\phi_f-Q_d/C_{ox}$. While the depletion layer extends throughout the Si island under zero gate bias

(*deep depletion*), $V_{gs} < V_T$ induces a p-type inversion layer and a subsequent drain current. The other structure behaves as an ordinary E-mode, n-channel MOSFET, the channel inverting when $V_{gs} > V_T$. The devices can therefore be coupled to form a CMOS gate.

Since the devices are not electrically connected, the necessity for isolating guard rings is removed, making SOS highly economical in chip area. The limited depth of the Si islands (about $1\mu m$) eliminates the large junction capacitance encountered with bulk silicon CMOS and the capacitance between the interconnection tracks and the bulk silicon is also eliminated. Junction leakage to the substrate is also eliminated, giving a lower power dissipation even than CMOS. The speed-power product is thereby limited to about 0.2pJ, which is the lowest value currently available. The relative simplicity of SOS fabrication gives it a potentially better yield than bulk-silicon CMOS. Additionally the insulating substrate provides protection against radiation damage, making it ideal for military and space applications. Unfortunately the sapphire substrate is ten times more expensive than bulk silicon substrate, but this is easily compensated by the improved performance.

Silicon-on-Insulator (SOI) technology closely resembles SOS. SOI devices are also fabricated as islands of silicon, but the insulating substrate is a thick layer of SiO_2, thermally grown upon a silicon wafer.

2.5.3 Ultra-Small-Dimension MOS Technologies

The aforementioned problems of ultra-small photolithographic fabrication can be solved by using smaller wavelength radiation (*fine-line lithography*), shorter channels can be achieved while maintaining the standard photolithographic techniques. The techniques described below allow the channel length to be defined by dopant diffusion or epitaxial growth while maintaining a high punch-through voltage. Although they have been superseded by advanced HMOS and SOS, these devices may possible play an important role in the integrated circuits of the future.

2.5.3.1 Double Diffused MOS (DMOS) [8,34]

Fig.2.12 shows the cross section of a DMOS transistor. It resembles a planar E-Mode NMOS structure with a lightly doped p-type channel. A region of much denser p-doping is achieved by lateral diffusion via the source opening in the field oxide prior to the diffusion of the n^+-source. Since the overall characteristics are determined by the more densely doped region, the effective channel length is dictated by the width of the p-type diffusion and the

benefits of a short-channel MOSFET are achieved. The more lightly doped region reduces the drain-source capacitance and prevents punch-through. The structure suffers however from the disadvantage of a poorly controlled threshold voltage V_T.

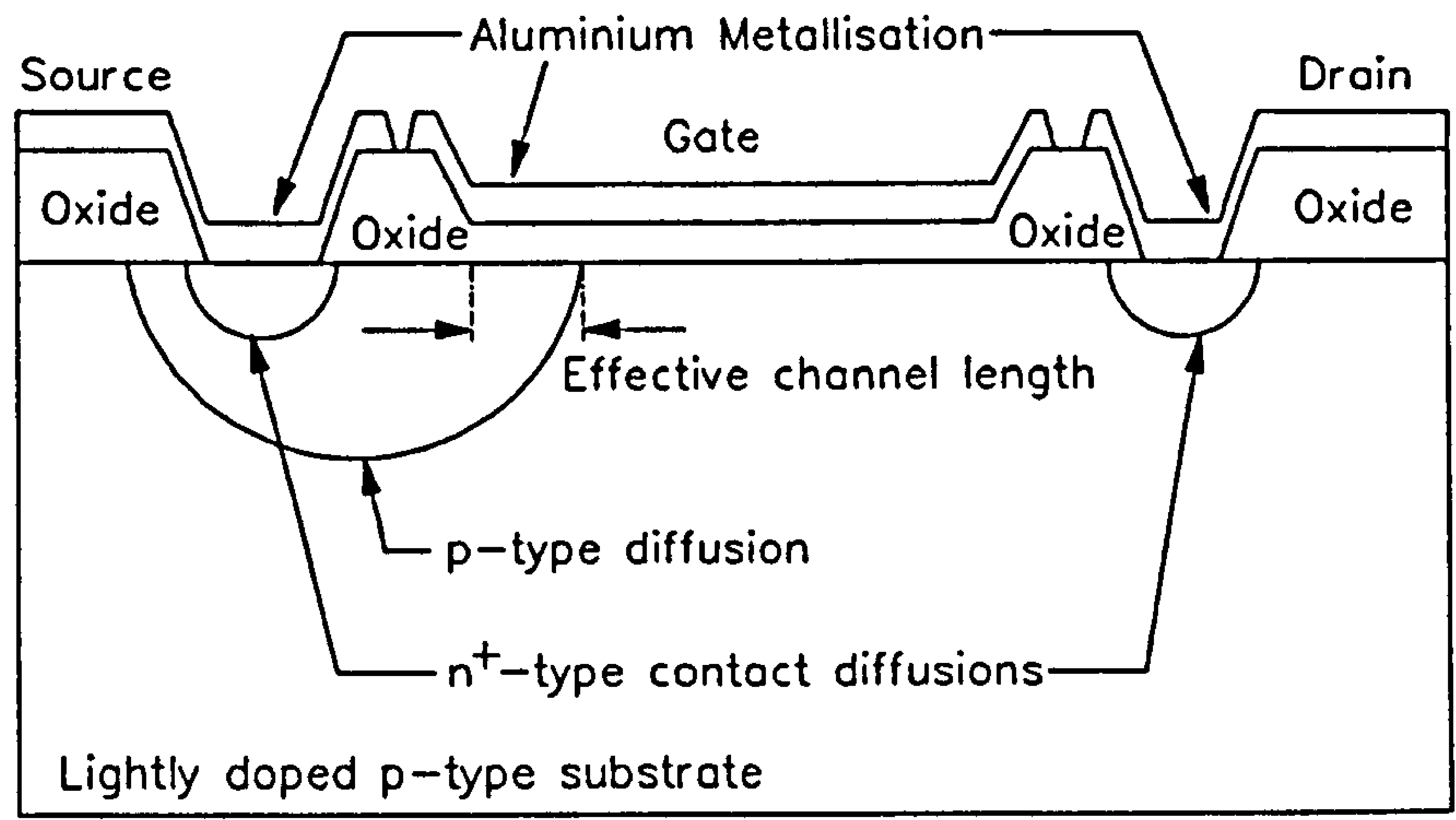


Figure 2.12: Schematic cross section of DMOS structure.

2.5.3.2 V-Groove MOS (VMOS) [8,34,45]

Preferential etching allows a regular V-shaped groove or pyramidal crater to be cut into <100> silicon, the exposed surfaces being <111> planes upon which the gate structure is fabricated. Several variants of VMOS make use of this groove, one of which is illustrated in Fig.2.13. The device is an n^+-p-n^+ sandwich, formed by epitaxial growth, the drain and source contacts being the n^+ layers. If the p-type layer is penetrated by the groove, the channel length is dictated by its depth, which is defined by the epitaxial growth process. As with the DMOS transistor, VMOS can have a very low drain-source capacitance and a high punch-through voltage, with the added advantage of an easily controlled threshold voltage ($\pm 15\%$ for $V_T = 1V$). The geometry permits a very large channel width, giving a high transconductance and allowing the device to sink large currents. The gate-delay for a typical NMOS structure is approximately 1ns, with a speed-power product of approximately 1pJ.

2.5.4 Charge Transfer Devices

The most important member of the family of charge transfer devices is the *charge coupled device* or CCD. Fig.2.14 shows the original CCD structure developed by Boyle and

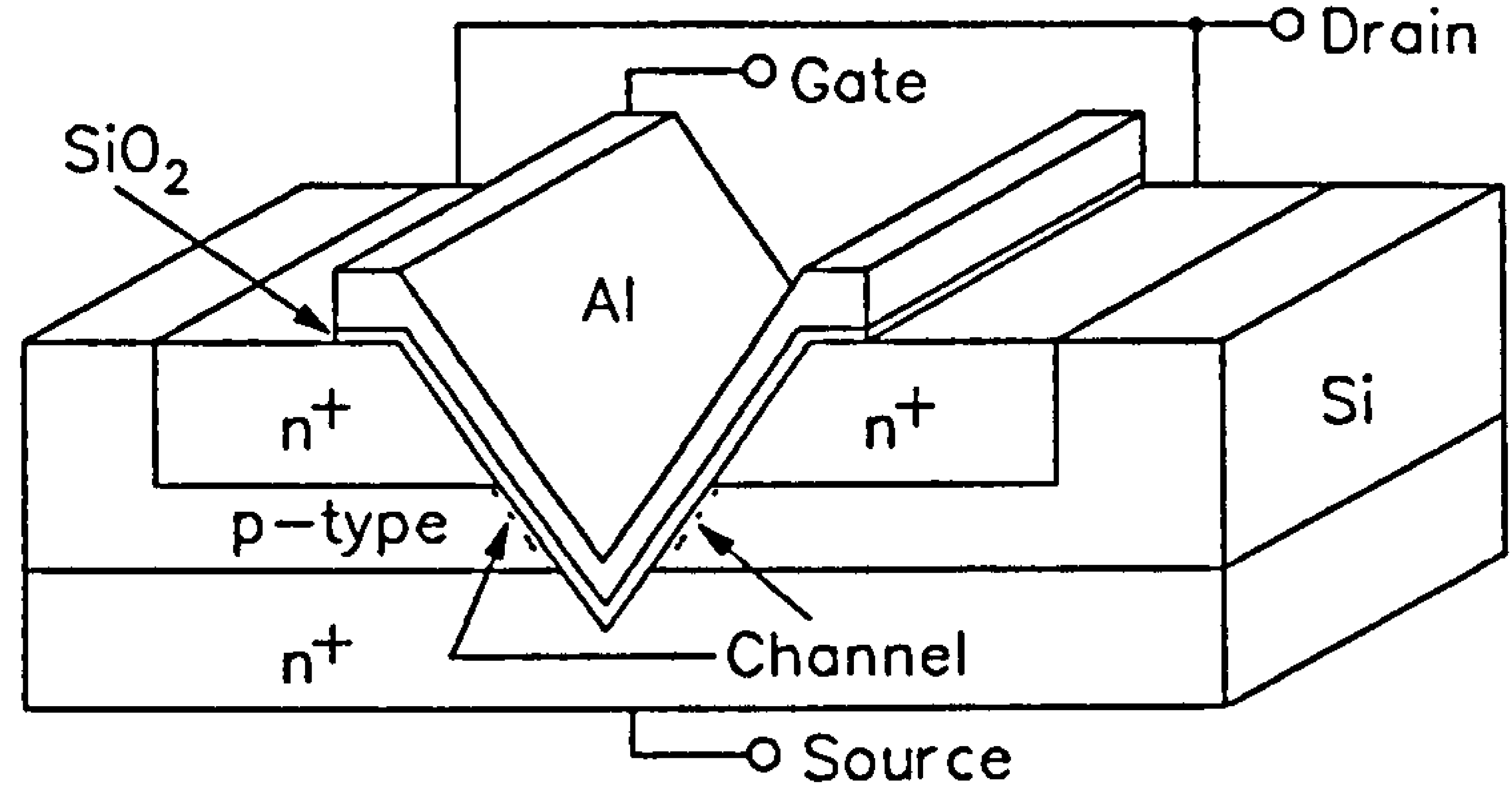


Figure 2.13: Schematic diagram of V-Groove MOS structure.

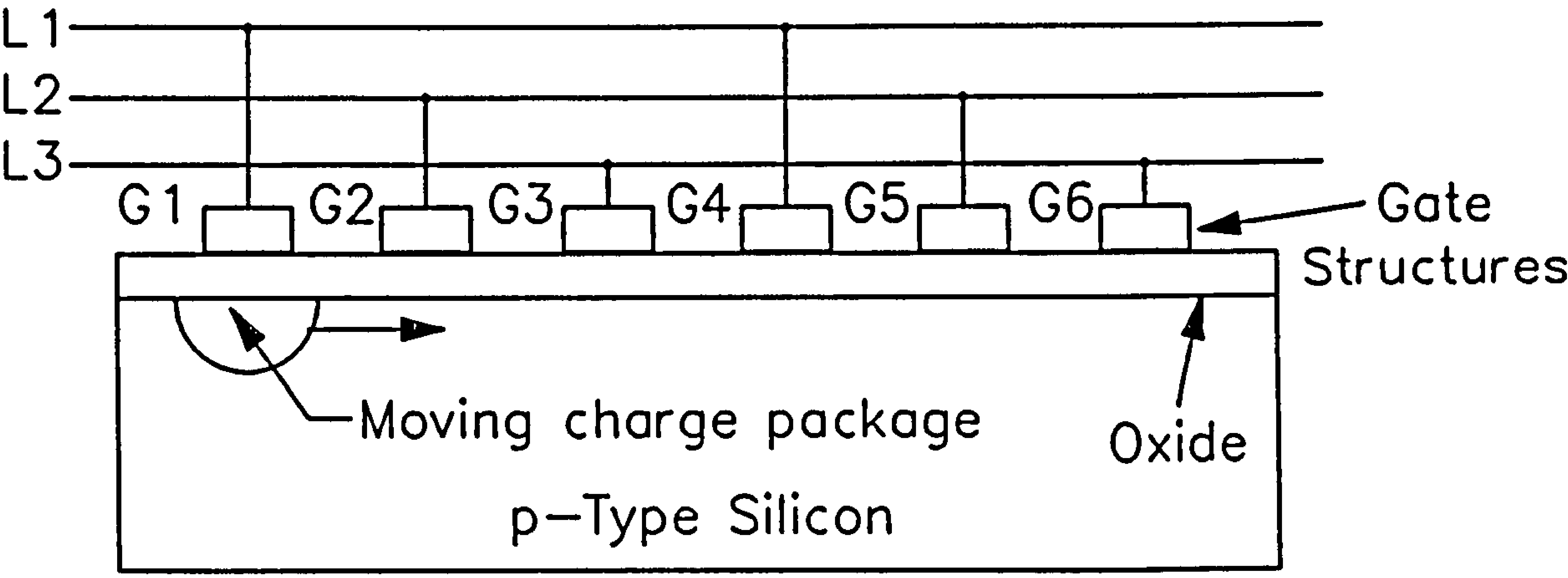


Figure 2.14: Charge Coupled Device (CCD) Structure

Smith in 1969 [34] (several variants have since been developed). It consists of a row of closely spaced MOS capacitors upon a p-type substrate. A positive voltage pulse applied to gate G_1 initially produces a *deep-depletion* layer in the underlying silicon, followed by the gradual development of an inversion layer of mobile electrons, induced by thermal generation. Since this latter effect is undesirable in CCD operation, the pulse width is kept well below the thermal relaxation time.

In deep depletion, the silicon surface below gate G_1 becomes a kind of *potential well* into which a 'packet' of mobile electrons can be injected via a diode. Application of an appropriate pulse sequence to L_1, L_1 and L_3 transfers the charge packet from device to device in a 'bucket brigade' fashion toward a detector at the opposite end.

CCDs have many applications in imaging, signal processing and in certain memory structures. Although the CCD is not strictly speaking a MOSFET structure, it finds its place amongst the general family of MOS devices.

2.6 Summary

1. The atomic and electronic structure of silicon dioxide has been discussed, together with the physics of the Si-SiO_2 and Gate-SiO_2 interfaces.

2. The operating theory of the MOS transistor has been presented. The standard 'first order' d.c. gradual-channel model of a MOSFET has been presented, together with simulated drain/source characteristics. The model is d.c. only.

3. The theory of MOSFET miniaturisation has been described, together with its advantages and drawbacks. A second order model of a miniature MOSFET has been presented, and its predictions have been compared with those of the first order model. Again, the model is limited to d.c. effects.

4. The following MOS variant technologies have been described:
   a) Planar PMOS, NMOS, CMOS and HMOS.
   b) Insulating substrate SOS and SOI.
   c) Small dimension DMOS and VMOS.
   d) Charge coupled devices (CCD).

The discussion has not included MOS-based nonvolatile memory devices such as FAMOS, FLOTOX and MNOS [34].

2.7 References

1. Davis,J.R., *"Instabilities in MOS Devices"*, Electrocomponent Science Monographs Volume 1, Gordon and Breach, New York, 1981.
2. Rigo,S, *"Silica Films on Silicon"*, Chapter 1 (pp.5-100) in *"Instabilities in Silicon Devices: Silicon Passivation and Related Instabilities, Vol.1"*, Eds. Barbottin,G, Vapaille,A, North-Holland, Amsterdam, 1986.
3. Niwa,M, Onoda,M, Matsumoto,M, Iwasaki,H, Sinclair,R, *"SiO_2/Si Interfaces Studied by STM and HRTEM (II)"*, Jap. J. Appl. Phys., 29(11), pp.2665-70, 1990.
4. Arienzo,L, Dori,L, Szabo,T.N., *"Effect of Post-Oxidation Anneal on Ultrathin SiO_2 Gate Oxides"*, Appl. Phys. Lett., 49(16), pp.1040-2, 1986.
5. Mott,N, *"Properties of Non-Crystalline Materials"*, Oxford University Press, Oxford, 1987.
6. Verwey,J.F., Amerasekera,E.A., Bisschop,J, *"The Physics of SiO_2 Layers"*, Rep. Prog. Phys., 53, pp.1297-331, 1990.
7. Kooi,E, *"The Surface Properties of Oxidised Silicon"*, Phillips Technical Library,

Netherlands, 1967.

8. Milne,A.D. (Editor), *"MOS Devices: Design and Manufacture"*, Edinburgh University Press, Edinburgh, 1982.
9. Hamasaki,M, *"Effect of Adding HCl to the Oxidizing Ambient on Characteristics of Metal-Oxide-Semiconductor Capacitors"*, J. Appl. Phys., 52(5), pp.3484-90, 1981.
10. Manchanda,L, Gurvitch,M, *"Yttrium Oxide/Silicon Dioxide: A New Dielectric Structure for VLSI/ULSI Circuits"*, IEEE Electron Dev. Lett., 9(4), pp.180-2, 1988.
11. Moslehi,M.M., Saraswat,K.C., *"Thermal Nitridation of Si and SiO_2 for VLSI"*, IEEE Trans. Electron Dev., ED32(2), pp.106-23, 1985.
12. Balland,B, *"Defects in Silica Films"*, Chapter 2 (pp.101-53) in *"Instabilities in Silicon Devices: Silicon Passivation and Related Instabilities, Vol.1"*, Eds. Barbottin,G, Vapaille,A, North-Holland, Amsterdam, 1986.
13. Jonscher,A.K., *"Electronic Properties of Amorphous Dielectric Films"*, Thin Solid Films, 1, pp.213-34, 1967.
14. Lenzlinger,M., Snow,E.H., *"Fowler-Nordheim Tunnelling into Thermally Grown SiO_2"*, J. Appl. Phys., 40(1), pp.278-83, 1969.
15. Hughes,R.C., *"Charge-Carrier Transport Phenomena in Amorphous SiO_2: Direct Measurement of the Drift Mobility and Lifetime"*, Phys. Rev. Lett., 30(26), pp.1333-6, 1973.
16. Williams,R, *"Photoemission of Electrons from Silicon into Silicon Dioxide"*, Phys. Rev., 140(2A), pp.A569-575, 1965.
17. Hughes,R.C., *"Time-Resolved Hole Transport in α-SiO_2"*, Phys. Rev. B, 15(4), pp.2012-20, 1977.
18. McLean,F.B., Boesch,H.E., McGarrity,J.M., *"Hole Transport and Recovery Characteristics of SiO_2 Gate Insulators"*, IEEE Trans. Nuc. Sci., NS23(6), pp.1506-12, 1976.
19. Nissan-Cohen,Y, Shappir,J, Frohman-Bentchkowsky,D, *"Determination of SiO_2 Trapped Charge Distribution by Capacitance-Voltage Analysis of Undoped Polycrystalline Silicon-Oxide-Silicon Capacitors"*, Appl. Phys. Lett., 44(4), pp.417-9, 1984.
20. Thomas,J.H., *"Transient Photodepopulation Measurements of Electron Trap Distributions in Thin SiO_2 Films on Silicon"*, J. Appl. Phys., 44(2), pp.811-4, 1973.
21. Hill,R.M., *"Single Carrier Transport in Thin Dielectric Films"*, Thin Solid Films, 1, pp.39-68, 1967.
22. Jonscher,A.K., *"Energy Losses in Hopping Conduction at High Electric Fields"*, J. Phys. C: Solid St. Phys., 4, pp.1331-40.
23. Goetzberger,A, Klausmann,E, Schulz,M.J., *"Interface States on Semiconductor/ Insulator Surfaces"*, CRC Critical Reviews in Solid State Science, 6, pp.1-43, 1976.
24. Rosencher,E, Coppard,R, *"Transient Capacitance Spectroscopy of Na^+-Induced Surface States at the Si/SiO_2 Interface"*, J. Appl. Phys., 55(4), pp.971-9, 1984.
25. Reed,M.L., Plummer,J.D., *"Si-SiO_2 Interface Trap Production by Low-Temperature Thermal Processing"*, Appl. Phys. Lett., 51(7), pp.514-6, 1987.
26. Burte,E.P., Matthies,P, *"Time-Resolved Thermal Annealing of Interface Traps in Aluminium Gate - Silicon Oxide - Silicon Devices"*, IEEE Trans. Nuc. Sci., NS35(5), pp.1113-20, 1988.
27. Hofmann,K, Rubloff,G.W., Liehr,M, Young,D.R., *"High Temperature Reaction and Defect Chemistry at the Si/SiO_2 Interface"*, Appl. Surf. Sci., 30, pp.25-31, 1987.

28. Vuillaume,D, Goguenheim,D, Bourgoin,J.C., *"Nature of the Defects Generated by Electric Field Stress at the Si-SiO_2 Interface"*, Appl. Phys. Lett., 58(5), pp.490-2, 1991.
29. Elliot,A.B., *"The Use of Charge Pumping Currents to Measure Surface State Densities in MOS Transistors"*, Sol. State Electron., 19, pp.241-7, 1976.
30. Noras,J.M., *"Parameter Estimation in MOS Conductance Studies"*, Sol. State Electron, 31(5), pp.981-7, 1988.
31. Bauza,D, Morfouli,P, Pananakakis,G, *"Detection of Interface and Volume Traps in Very Thin Oxide MOS Structures using DLTS, Quazi-Static and Conductance Measurements"*, Sol. State Electron., 43(9), pp.933-6, 1991.
32. Çelik-Butler,Z, Hsiang,T.Y., *"Determination of Si-SiO_2 Interface Trap Density by 1/f Noise Measurements"*, IEEE Trans. Electron Dev., ED35(10), pp.1651-5, 1988.
33. Nicollian,E.H., Brews,J.R., *"MOS (Metal Oxide Semiconductor) Physics and Technology"*, John Wiley and Sons, New York, 1982.
34. Sze,S.M., *"Physics of Semiconductor Devices"* (Second Edition), John Wiley & Sons, New York, pp.431-510, 1981.
35. Shichman,H, Hodges,D.A., *"Modelling and Simulation of Insulated-Gate Field-Effect Transistor Switching Circuits"*, IEEE Journal of Solid State Circuits, SC3(3), pp.285-9, 1968.
36. Lin,M-S, *"A Better Understanding of the Channel Mobility of Si MOSFET's Based on the Physics of Quantized Subbands"*, IEEE Trans. Electron Dev., 35(12), pp.2406-11, 1988.
37. Hsu,F-C, Muller,R.S., Hu,C, Ko,P-K, *"A Simple Punchthrough Model for Short-Channel MOSFET's"*, IEEE Trans. Electron Dev., ED30(10), pp.1354-9, 1983.
38. Emrani,A, Ghibaudo,G, Balestra,F, *"On the Modelling of Mobility in Silicon MOS Transistors"*, Proc. 1990 European Solid State Device Research Conference (ESSDERC), pp.497-500, 1990.
39. El Banna,M, El Nokali,M, *"A Pseudo-Two-Dimensional Analysis of Short Channel MOSFETs"*, Solid-State Electronics, 31(2), pp.269-274, 1988.
40. Sabnis,A.G., *"Impact of Advances in Technology on the Properties of Si/SiO_2 Interface"*, Proc. 22nd. IEEE Reliability Physics Symp., pp.156-60, 1984.
41. Miura-Mattausch,M, Jacobs,H, *"Analytical Model for Circuit Simulation with Quarter Micron MOSFETs: Subthreshold Characteristics"*, Proc. 1990 European Solid State Device Research Conference (ESSDERC), pp.153-6, 1990.
42. Stanley,T.D., *"The State-of-the-Art in SOI Technology"*, IEEE Trans. Nuc. Sci., 35(6), pp.1346-9, 1988.
43. Fleetwood,D.M., Thome,F.V., Tsao,S.S., Dressendorfer,P.V., Dandini,V.J., Schwank,J.R., *"High-Temperature Silicon-on-Insulator Electronics for Space Nuclear Power Systems: Requirements and Feasibility"*, IEEE Trans. Nuc. Sci., NS35(5), pp.1099-112, 1988.
44. Haisma,J, *"SOI Technologies: Their Past, Present and Future"*, Journal de Physique, Colloque C4, Supplément au no 9, Tome 49, pp.3-12, Sept.1988.
45. Holmes,F.E., Salama,C.A.T., *"VMOS - A New MOS Integrated Circuit Technology"*, Sol. St. Elect., 17, pp.791-7, 1974.

Chapter 3

MOS Failure Mechanisms

3.1 Introduction

Electronic failure mechanisms can be classified into two basic types: *time dependent* and *event dependent* and each category may be subdivided into *intrinsic* and *extrinsic* mechanisms. Time dependent failures are the result of continuous wearout during the component's working life, while event dependent failures are caused by specific external events. Intrinsic failures are associated with the inherent structure of the circuit or device while extrinsic failures are due to structural defects (e.g. crystallographic flaws, poorly grown or ion-contaminated oxides). Intrinsic and extrinsic failures are responsible for the *wear out* (or *senile decay*) and *infant mortality* features of the reliability 'bathtub' curve (Fig.3.1). Extrinsic failure is most noticeable in immature technologies (such as the GaAs MESFET) and the improvement of manufacturing technology and experience should eventually bring these failures to a minimum.

This thesis is primarily concerned with the event dependent failures in MOS circuitry. The present chapter examines the sources of event related failure and discusses the physical mechanisms involved, many of which are relevant to both time and event dependent failure. Latent or 'walking wounded' failure is then discussed and the chapter ends with an examination of reliability improvement techniques.

3.2 Sources of Event Dependent Failure

There are a number of electrical mechanisms which may induce event dependent failure, most of them associated with poor circuit design or careless handling. These mechanisms can be grouped under the broad category of *Electromagnetic Interference (EMI)* [1], although the various sub-categories are given different terms and definitions throughout the literature. The definitions employed in this thesis are given in this section.

Manufacturers and consumers estimate that between 60 and 80% of field failures are due to electrically related stress, electrostatic discharge (ESD) being the main culprit [2]. Failure studies by NASA indicate that ESD is responsible for up to 70% of the electronic

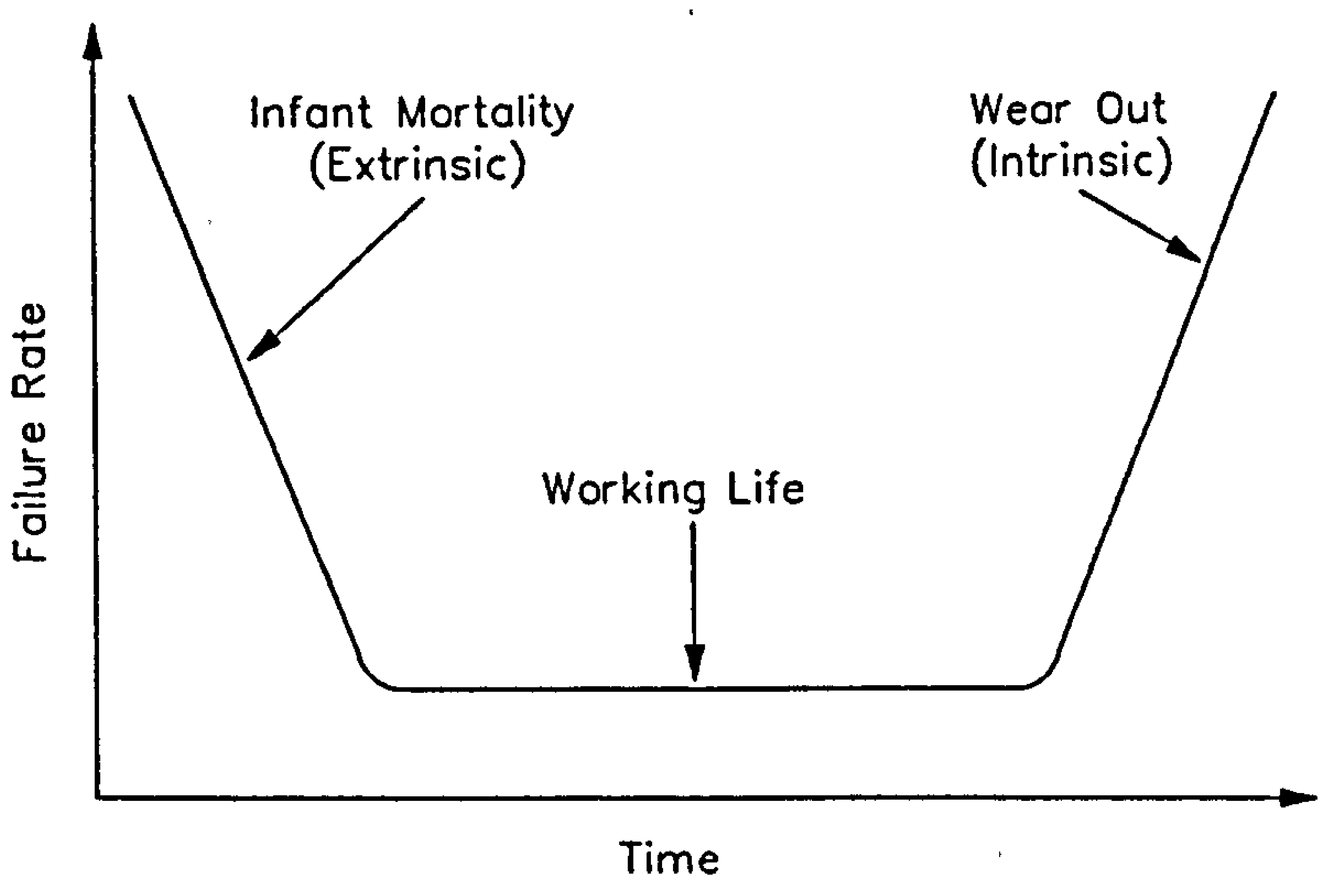


Figure 3.1: Reliability 'bathtub' curve. Failure rate (% failure per unit time) is highest when the component is young (due to extrinsic defects) and very old (due to intrinsic wearout or 'senile decay'). Time scale is logarithmic.

failures encountered throughout the U.S. space program [3]. However, this figure may not be accurate since the effects of ESD and other electrical failure mechanisms are not easily distinguished by failure analysis [4]. The major categories of electrical stress are described below:

3.2.1 In-circuit Electrical Overstress (EOS)

In any non-ideal environment, components can be exposed to spurious high current/voltage excursions, which are generally termed *electrical overstress* or EOS [2]. Types of EOS include noise bursts or electrical field transients (EFT) and high energy excursions (SURGE) [1]. Negligent exposures to conditions exceeding the manufacturers' recommended maxima are not included in the EOS category [2].

EOS may have a number of effects upon devices and circuits. High power pulses dissipated in small dimension structures can create localised high temperature regions ('hot-spots'), causing silicon meltdown, electromigration or electro-thermomigration leading to short- or open-circuit failure [2]. Thermal junction meltdown has been found to be the major ESD failure mechanism in GaAs MESFET circuitry [5]. Additionally, high voltage transients appearing across MOS gate oxides may support dielectric breakdown [6]. These problems clearly become more serious as device dimensions diminish [8].

3.2.2 Electrostatic Discharge (ESD) Pulse Stress

In any working environment, a human operator can accumulate considerable electrostatic potentials which are generally negative with respect to ground [8]. These voltages are usually a result of *triboelectrification*, whereby abrasion between materials with different electron affinities causes an exchange of charge. The relative efficiency with which different materials transfer charge to each other is characterised by the *triboelectric series* [9,10,11]. Thus a person's shoes may acquire charge from contact with the floor, causing charge separation within the body. In this way, 12kV may be accumulated by walking across a vynal floor and 35kV from crossing a synthetic carpet. Alternatively the movement of synthetic clothing can generate charge by the same mechanism [12], or charge can be transferred from another charged object [13] (e.g. a VDU screen). A standing person can accumulate anything up to about 40kV, above which corona discharges introduce a clamping effect [3,14]. The charge magnitude and duration vary inversely with relative humidity [14,15], causing the amount of static electricity to fluctuate on an annual basis [12].

The static voltage V is related to the accumulated charge Q by the equation $V=Q/C$ where C is the capacitance to ground. The latter quantity varies according to the person's position with respect to the ground plane and an operator may therefore magnify the static voltage by the act of sitting. If he or she then touches an i.c. terminal (either an input or an output [16]), an electrostatic discharge (ESD) pulse is injected into the circuit. A heavily charged person may not even need to touch the device, since current can pass by an air discharge. The resulting power transient can be several kilowatts [8], which can easily damage the circuit. Since static discharges below 3-4kV are undetected by human nerves [3,8,10], the operator is unlikely to be aware of the damage he or she causes.

Until relatively recently, the popular belief that only MOS devices are susceptible to ESD was still widely accepted [10]. However, the recent reduction of component dimensions has increased ESD sensitivity throughout the whole range of devices, both passive and active [17,18]. This has increased the overall awareness of ESD as well as the availability of numerous antistatic gadgets for use by semiconductor manufacturers and consumers. Table 3.1 shows the three categories into which ESD sensitivities have been classified.

Table 3.1: Categories of ESD Sensitivity [9,10,11]

Category 1 (<1kV)	e.g. Unprotected MOS, Small Dimension Bipolar Schottky Logic, GaAs MESFETs [5].
Category 2 (1-4kV)	e.g. Protected MOS, Standard Bipolar Schottky Logic, Linear Bipolar Integrated Circuits.
Category 3 (4-15kV)	e.g. Standard Bipolar Logic.

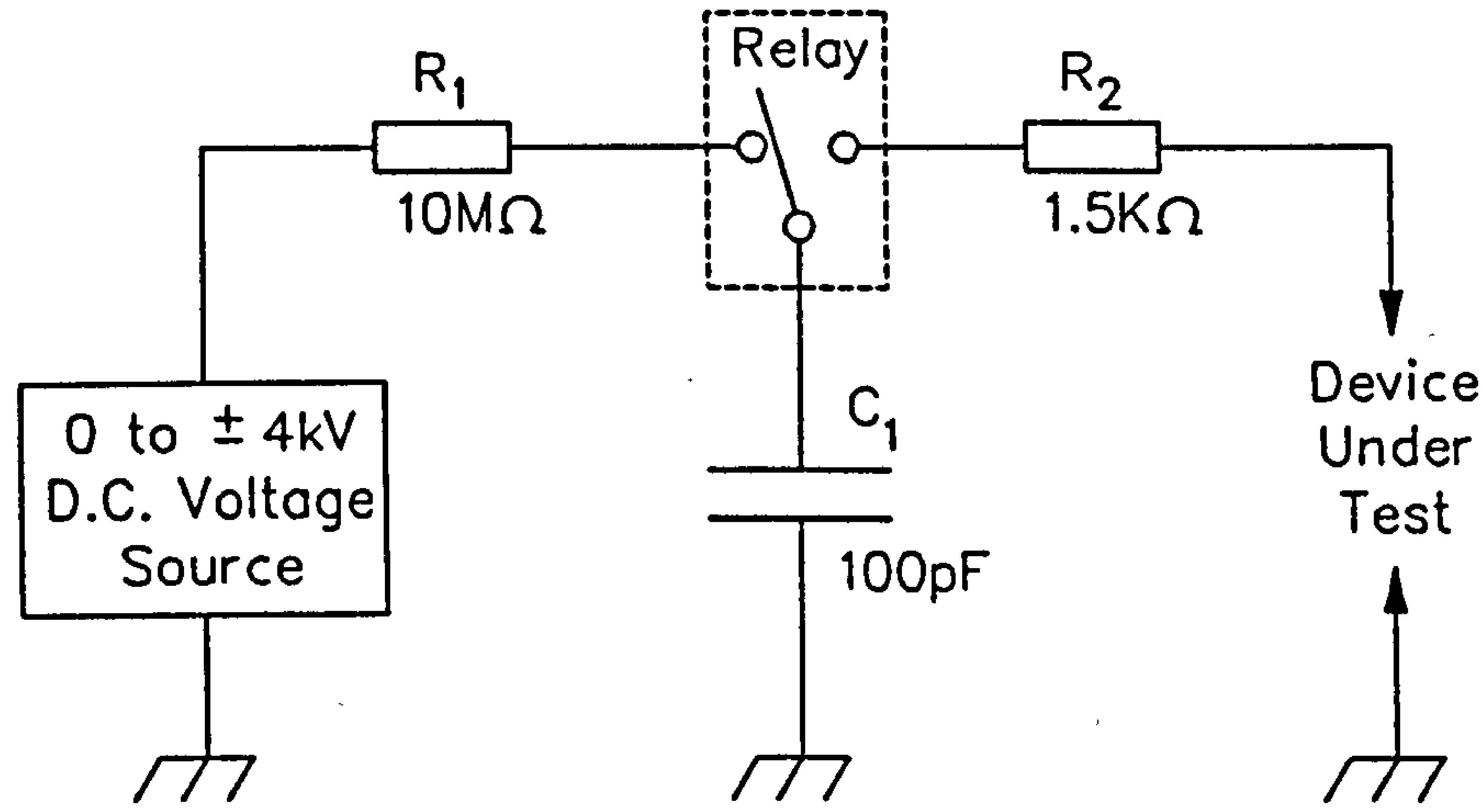


Figure 3.2: Human Body Model (HBM) Equivalent Circuit.

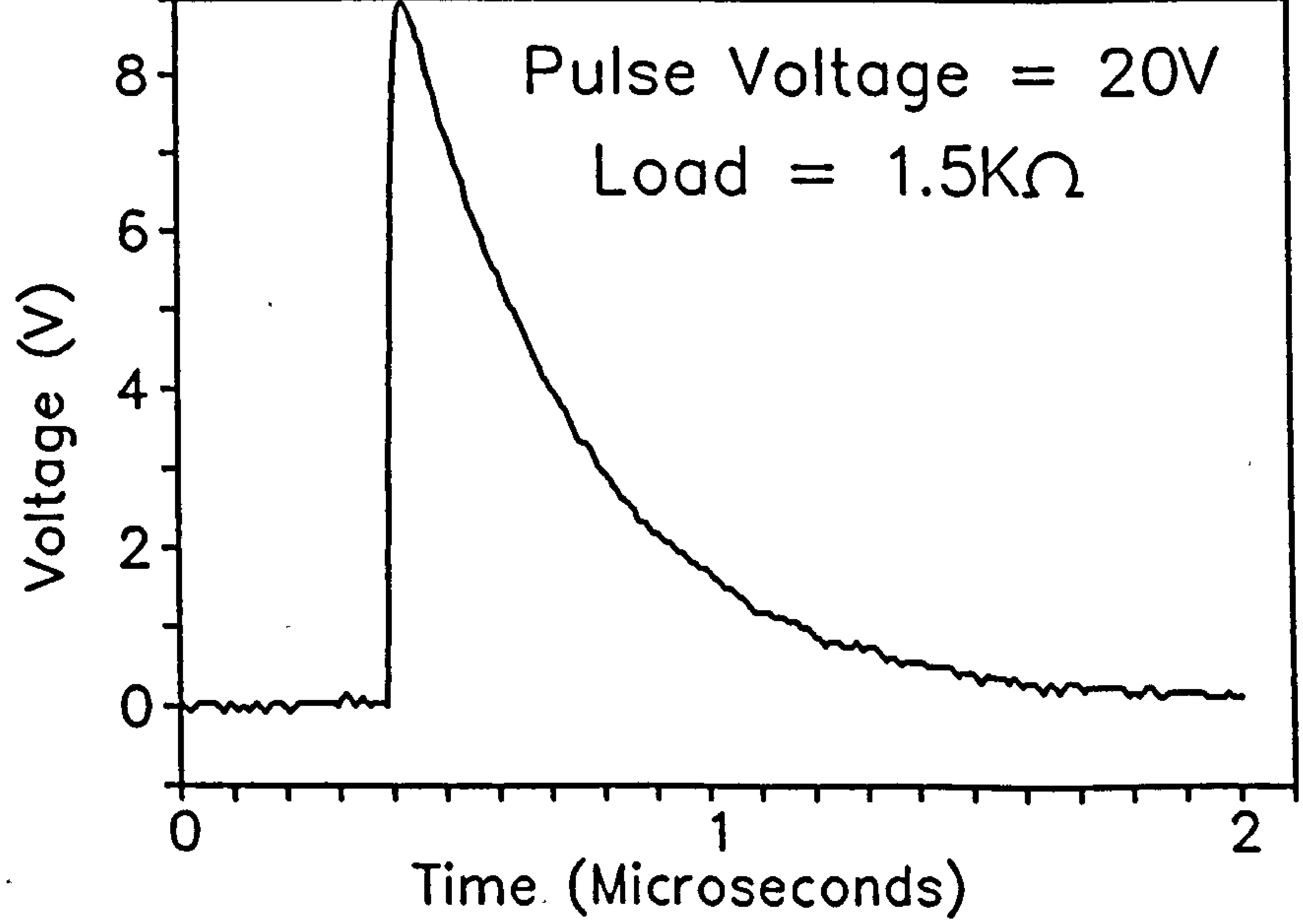


Figure 3.3: Typical Human Body Model (HBM) Pulse Waveform.

Unprotected MOS devices are highly vulnerable to gate oxide dielectric breakdown under ESD [6] and therefore appear near the bottom of Category 1 (CCDs are particularly sensitive [19]). The common failure mechanism in non-MOS devices is Joule heating by the dissipating static charge generating localised high temperature regions (hot spots). This can cause meltdown or evaporation in metallisation, thin-film resistors and active device junctions [2]. All of these are described as *hard* failures since they involve damage to the electronic hardware. *Soft* failure, i.e. the injection of spurious information via an ESD pulse lies outside the scope of this thesis (although work on this is in progress).

In the 1960s and 70s it was widely accepted that components mounted in a p.c.b. are safe from electrostatic damage, since the limited ESD charge is dissipated by the large assembly capacitance causing voltage attenuation. This myth was exposed in the 1980s [10,20] when it was shown that the voltage attenuation is not necessarily significant. It has even been suggested that components in an assembly are in *greater* danger, since many devices may be destroyed by an ESD pulse on a single connector pin [20].

ESD events are readily simulated by the *Human Body Model* or HBM (MIL-STD-883C [21]), which consists of a 'body' (or skin) capacitance C_1=100pF and a 'body' resistance R_2=1.5KΩ (Fig.3.2). Corresponding values for actual human beings can vary in the ranges 50-450pF and 0.1-100kΩ depending on body size, posture and skin moisture [3,10,13]. The standard HBM model assumes direct contact between the device and the human skin (although another model includes the effects of a corona discharge [22]). C_1 is charged to the required stress voltage V_0 and the pulse is applied to the device under test (DUT) via the relay. The pulse charge and energy are given by C_1V_0 and $\frac{1}{2}C_1V_0^2$ respectively, and the pulse time-constant is of the order of C_1R_2=150ns. Standard apparatus (notably the Hartley 'AutoZap' and the KeyTek 'ZapMaster') has been developed to simulate the HBM pulse. In order to conform to MIL-STD-883C, the output wave of such a simulator must have a risetime below 10ns and ringing below 15% of the maximum current [21]. Fig.3.3 shows a typical voltage profile of a +20V HBM pulse discharging to ground via a 1.5KΩ resistive load.

There are several alternatives to the HBM test. The so-called *True-ESD* [23] resembles the HBM with an additional short time constant to represent the individual discharge from the operator's hand. Additionally, since a component is rarely grounded as in the case of the HBM, an additional 20pF capacitance between the DUT and the ground plain has been proposed [13]. Other workers [22] have suggested that the 100pF 'skin' capacitor of the HBM should be supplemented by a much larger 'bulk body' capacitance, from which the 100pF capacitor re-charges after an ESD event without the need for further triboelectrification.

The *Machine Model* (MM), developed in Japan to provide a more stringent version

of the HBM, uses a body capacitance of 200pF and together with zero series resistance [1]. The MM pulse rise-time is governed by the circuit inductance which is usually about 150nH [24]. As its name suggests, the machine model provides a good representation of ESD from a charged 'machine' such as a VDU screen [25]. An alternative 'worst case' human body model was introduced by Fisher in 1989 [26], which included inductive effects in the human body.

The *Charged Device Model* (CDM) [27] represents situations when a device, charged by induction [28] or by the triboelectric charging from packaging materials [29], is discharged into the ground plane. As human operators have gradually been replaced by robotic assemblers, the relative importance of the CDM has increased relative to the HBM [30,31]. A variation of the CDM called the *Charged Board Model* (CBM) [32] represents the effects of a charged printed circuit card.

Yet another form of ESD is the *Field Induced Model* (FIM) [33] which occurs when charge is redistributed in a component due to the strong electric field which may be induced by a human body up to half a meter away [3]. For this reason, sensitive devices should not be brought within the vicinity of any charged object [34] and should ideally remain within a conductive 'Faraday cage' enclosure. Field induced damage may be caused by internal dielectric breakdown [8,9] by discharge to the ground [8,35].

3.2.3 Electromagnetic Pulse (EMP) Stress

A spurious current or voltage pulse can also be induced in the circuit by a local electromagnetic disturbance [1]. This may include a nearby human-body electrostatic discharge [36], a lightening strike or a nuclear explosion [1]. The former is often called *Indirect ESD*. The latter, usually termed *electromagnetic pulse* (EMP), is clearly an important concern in the design of military hardware.

3.3 Dielectric Breakdown

3.3.1 Phenomenology of Dielectric Breakdown

Electrical breakdown in the SiO_2 gate dielectric is the most well known MOS failure mechanism. The dielectric strength of silicon dioxide is usually between 7 and 14MV/cm, depending on thickness [37,38], method of preparation [38], Si doping [38], electrode

material [39] and testing conditions. A higher dielectric strength is achieved by a number of techniques including the introduction of water or HCl [40] during oxidisation. Breakdown is accelerated by any mechanism which enhances the oxide field, for example a rough cathode/SiO_2 interfaces with many asperities and protuberances is more prone to breakdown than a smooth surface [41,42]. Breakdown is also affected by the post-oxidation annealing time/temperature, having a similar functional dependence to the Si/SiO_2 interface trap density described in Chapter 2 [43].

At breakdown a large localised current flows in the dielectric, evaporating a region of gate material and producing a visible crater [44]. The breakdown site can be located by electron microscopy [45,46] or electroluminescence under current injection [47]. If the gate electrode is thin then the heat dissipated by breakdown may evaporate the gate material in the vicinity of the crater, isolating the defect and *self healing* the oxide [45]. However, either the molten gate material or an air discharge may form a conduction path between the gate and the underlying silicon, sustaining the breakdown current and causing damage propagation across the gate area [45]. Alternatively the current path across the oxide may fuse under further stress, providing another mechanism for self healing [48].

Although the definitions of what constitutes a damaged oxide varies between publications, a typical example is a leakage current of 10μA at 5V bias [49]. However, the rapid resistivity change at breakdown (typically from 1000GΩ to 1KΩ in less than 1μs) appears to render the exact definition somewhat unimportant. Some recent experiments on ultra-thin SiO_2 (about 55Å) show that nondestructive multiple breakdown events can occur prior to final breakdown as the oxide switches between these high and low conductance levels [50]. Later studies have uncovered the existence of multiple conduction levels in the oxide, between which the dielectric may switch prior to breakdown [51,52].

Breakdown mechanisms differ for SiO_2 and silicon nitride (Si_3N_4) films due to different conduction properties [53]. Although silicon oxide/nitride double dielectrics are widely used in MNOS technology, Si_3N_4 breakdown differs fundamentally from SiO_2 breakdown and is therefore excluded from this study.

3.3.2 Types of Dielectric Breakdown

The physics of dielectric breakdown has been under some dispute for most of this century. Early theories focused on direct field-induced distortion of the dielectric network, enhanced by random thermal motion. Large electric fields were believed to distort the structure beyond its elastic limit, leading to mechanical fracture. Other theories invoked Zener tunnelling of electrons between bands [54], rapidly increasing the conductivity at high

fields. These and other early theories are reviewed by Whitehead [55], although most of them have now been discredited and are of historic interest only.

Some years ago a distinction between *thermal* and *electronic breakdown* was identified [55]. While electronic breakdown is associated directly with mobile electrons and takes place in about 10^{-8}seconds [56], thermal breakdown is much slower and is associated with Joule heating of the atomic structure (i.e. if heat is generated faster than it can be dissipated then the melting temperature is reached and the material dissociates) [55]. In 1974, Hamano [57] discovered three modes of dielectric breakdown in SiO_2, two of which were electronic and the other thermal. The importance of thermal breakdown was found to wane as film thickness decreased, electronic breakdown dominating in oxides thinner than 100nm. It should be noted that the terms *thermal* and *electronic* refer to the mechanism *triggering* breakdown, rather than the resulting destruction. The latter is *always* thermal, fuelled by the discharge current in the punctured oxide, causing local temperatures exceeding 2000°C. This has been confirmed by correlating of the size of the damage crater to the available electrostatic energy [58]. Other modes of breakdown have also been explored. For example a dielectric containing a large density of 'voids' is prone to *discharge* breakdown [56], i.e. air breakdown within a cavity. Additionally a dielectric exposed to a field for a long period of time can gradually change from an insulating to a conducting state, ultimately reaching a condition in which electronic or thermal breakdown can occur [56].

3.3.3 Field-Dependent or Avalanche Breakdown

The electron-avalanche model has long been accepted for gaseous breakdown (occurring at about 30MV/cm at s.t.p.). Impact ionisation of atoms by field-accelerated electrons release further free electrons, which themselves gain enough energy from the field to cause impact ionisation. The inverse proportionality between the breakdown field F_{bd} and mean free path length λ (Paschen's law [59]) implies that an electron needs to gain a specific energy E_{ion} between collisions in order to cause ionisation and subsequent breakdown [60]. If $F<F_{bd}$ then most electrons undergo inelastic collisions and are unable to attain E_{ion} over any number of path lengths. The energy lost in these collisions is absorbed by electronic 'excitation', causing the gas to glow.

In 1937, von Hippel [60] extended this model to solids, predicting F_{bd}=115MV/cm for an ideal gas compressed to solid-state density (10^{23}molecules/cm^3) at room temperature. This value was significantly larger than the experimentally observed 1-10MV/cm. He therefore proposed that the shielding effect of neighbouring atoms reduced E_{ion} to below the

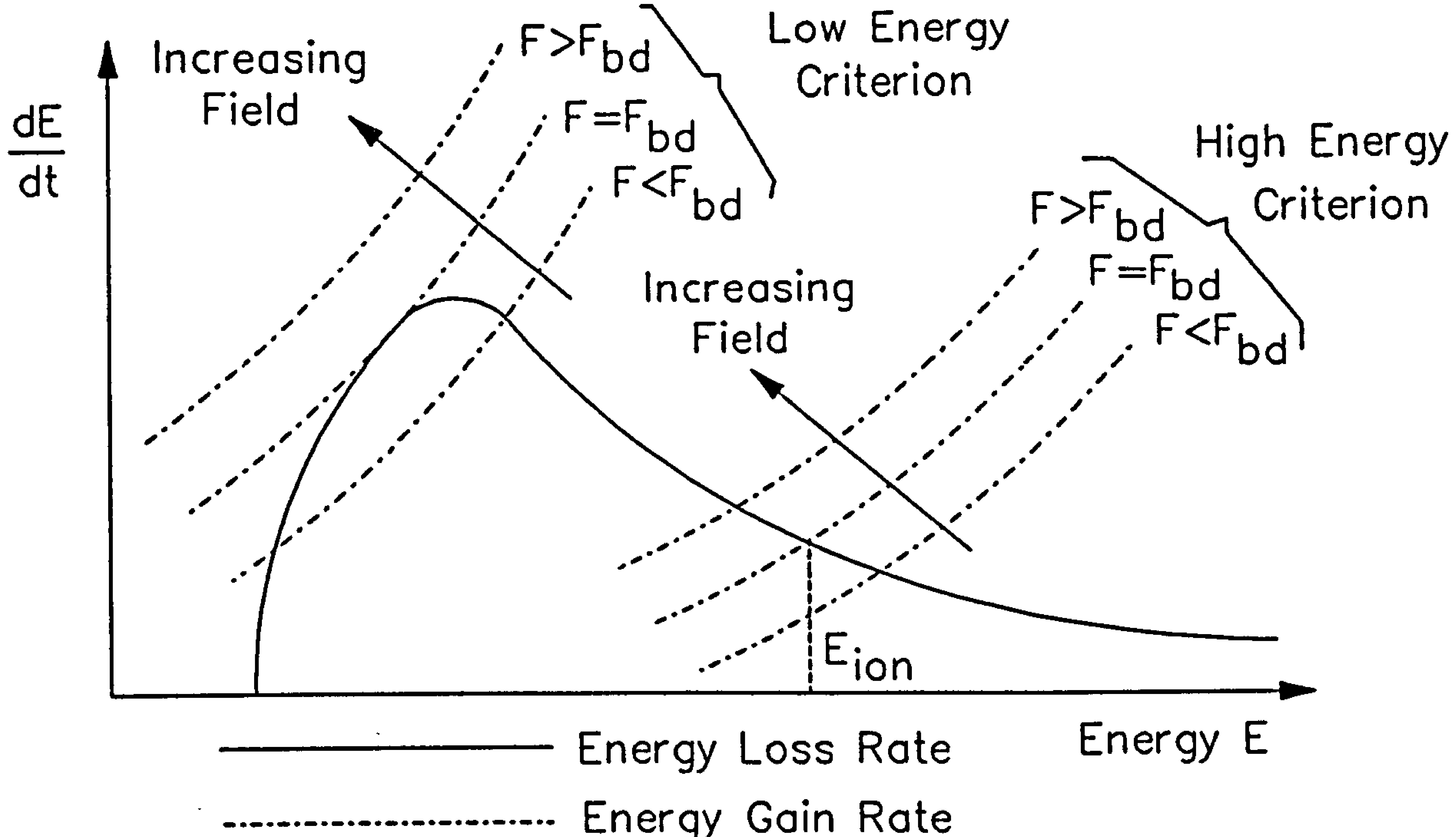


Figure 3.4: Energy transfer rated for electron-LO phonon interaction, showing high and low energy breakdown criteria.

excitation energy, thereby removing the excitation 'barrier'. Electron acceleration in a solid is therefore governed not by electronic excitation but by interactions with lattice vibrations or *phonons*.

There are two phonon categories, *optical* and *acoustic*, both of which can appear in *longitudinal* and *transverse* modes. Fröhlich [61] reasoned that the longitudinal optic (LO) phonon is dominant in ionic crystals, in which the rates of electron energy gain (field-to-electron) and energy loss (electron-to-phonon energy transfer) vary as shown in Fig.3.4. Above a critical energy E^*, related inversely to the field F, an electron accelerates discontinuously. Fröhlich proposed the *'high energy criterion'* [61], specifying that a self-sustaining avalanche requires $E^* < E_{ion}$, such that electrons below E_{ion} might be accelerated to cause impact ionisation. Other theoreticians proposed the *'low energy criterion'* [62,63], stating that E^* must lie below the maximum of the energy loss curve, allowing universal acceleration for the entire electron gas.

The above models, predicting increasing breakdown field with increasing temperature, describe *low temperature breakdown*. They are applicable only up to a certain critical temperature T_c, above which the conduction electron concentration is much greater due to thermal trap ionisation and the electron-electron interaction dominates the electron-phonon interaction [64,65]. This is sometimes termed *high temperature breakdown* and is characterised by a decreasing breakdown field with increasing temperature. T_c decreases with

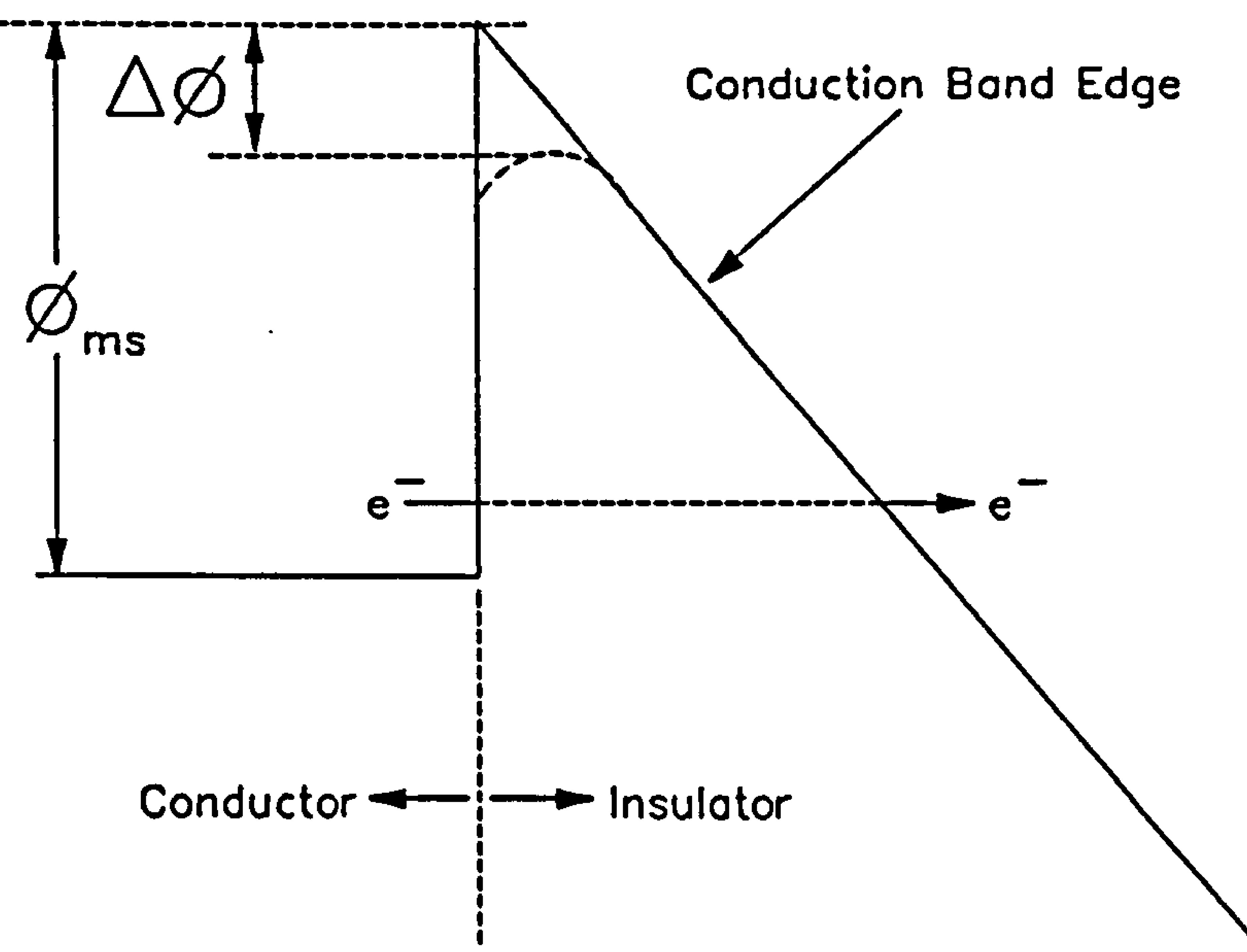


Figure 3.5: Fowler-Nordheim Tunnelling Mechanism (the existence of $\Delta\phi$, the classical image-force potential barrier lowering is under dispute).

increasing atomic disorder such that while a crystal may exhibit low temperature behaviour above 100°C, an amorphous material may undergo high temperature breakdown at room temperature. Seitz [54] calculated that 40 generations of impact ionisation are necessary to overcome the binding energy of the dielectric lattice (approx. 230 kcal./mol.) and cause permanent structural damage. These models are largely field dependent, with little or no time dependence (breakdown usually occurs within 10^{-8} seconds [56]). Such behaviour is sometimes described as *intrinsic*, although this term is given a slightly different meaning in the present text (see Section 3.3.6 below).

There has been some disagreement about the application of these models to SiO_2. Firstly, since SiO_2 is only partly ionic, the acoustic phonon is believed to dominate transport at high fields [66,67]. Ridley [68] performed some simple calculations using both optical and acoustic phonon modes, showing a large discrepancy for wide bandgap insulators, e.g. 61MV/cm for SiO_2 compared to the experimental ~9MV/cm. However, Ridley's model was rather simplistic and the more elaborate analysis of Ferry [69] predicted an SiO_2 breakdown field between 7 and 8MV/cm.

These avalanche-breakdown models have been improved by the inclusion of three other mechanisms: Firstly, electrons are injected into the dielectric by Fowler-Nordheim (F-N) tunnelling [70]. This is illustrated in Fig.3.5, which shows how the thinning of the Cathode/SiO_2 potential barrier under high electric fields allows quantum-mechanical tunnelling of electrons into the oxide conduction band. The tunnelling current depends on the barrier transmission coefficient (which is a function of the electric field) and the electron

supply-rate to the SiO_2 surface (which depends upon the cathode electron concentration [7]). Secondly, the relatively immobile holes generated by impact ionisation distort the internal dielectric field, affecting the ionisation and electron injection rates [71-73]. Thirdly, the hole-electron generation rate is offset by the rate of hole-electron recombination [73]. All these mechanisms have been combined in a unified model [68,73-76] which correlated well with experimental data on SiO_2 breakdown [77].

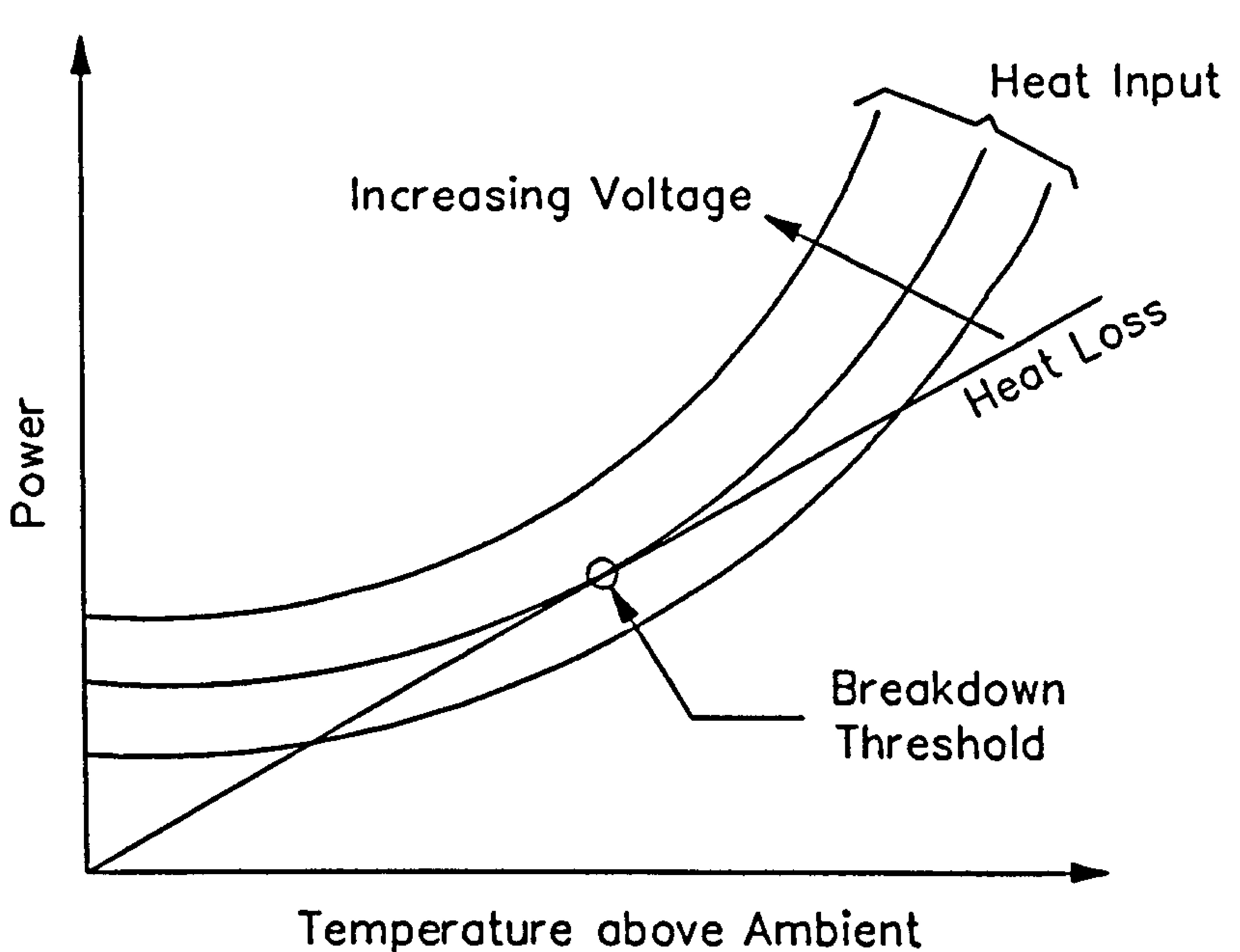


Figure 3.6: Illustration of Joule Heating Breakdown Model (after Klein & Gafni [44]).

3.3.4 Thermal Dielectric Breakdown

The most important features of the avalanche breakdown model are its field dependence and time independence. Experiments have shown that dielectric breakdown is not exclusively field dependent and has a measurable time or 'history' dependence. For example, Klein and Gafni [45] found that oxide breakdown could occur under a decreasing field, an observation clearly inconsistent with any field-dependent model.

Several workers have interpreted this as evidence of thermal breakdown: Filamentary current injection causes Joule heating and increased local temperature. This causes increased local conductivity which accelerates the heating [45,78]. Above a critical field, heat is generated faster than it can disperse and the dielectric reaches its melting temperature T_{melt} and dissociates [79]. This process, illustrated in Fig.3.6, has been applied successfully to polymeric insulators [80]. Tatsuuma [81] suggested that in an MOS structure, electrons in the oxide conduction band lose energy as they are emitted at the anode, causing a buildup

of heat and subsequent oxide meltdown.

Recently the idea that *all* dielectric breakdown is thermal was questioned by Wolters et al. [58], who calculated that under certain conditions, the energy dissipated in the dielectric is unable to cause meltdown even if it were confined to a single atom. However, Hamano [57] claims to have identified both electronic and thermal SiO_2 failure modes and it is possible that what Wolters observed was one of the electronic modes. Alternatively a temperature increase, insufficient to cause damage, may reduce the dielectric strength (as in the case of the 'high temperature breakdown' above), triggering electronic breakdown.

3.3.5 Causal or Time Dependent Dielectric Breakdown (TDDB)

One of the main features of the thermal breakdown model is its 'elasticity', i.e. when an oxide fails to break down, it cools back to its virgin condition when stress is removed. However, in many cases breakdown has been shown to be caused by a cumulative and irreversible 'wearout' to which all previous stress (however long ago) has contributed. For example, many early workers observed a continuous electrochemical deterioration in chloronaphthalene capacitors, leading to dielectric breakdown [82]. Several mechanisms were suggested for SiO_2 wearout, many of which were associated with mobile ion drift [83].

The present consensus maintains that SiO_2 wearout is driven by Fowler-Nordheim tunnelling injection of electrons from the cathode into the dielectric. Such wearout tends to be characterised either by the time-to-breakdown t_{bd} or the charge Q_{bd} injected prior to breakdown (although some authors dispute the significance of the latter [84]). Fig.3.7 illustrates the causal nature of wearout. It shows the Fowler-Nordheim current J plotted against the injected charge Q for two oxides, one stressed at a continuous 12MV/cm, the other stressed first at 12MV/cm and then at 11MV/cm after an indefinite interruption. Clearly only a small additional charge is required to push the oxide into breakdown at 11MV/cm after the injection it has already sustained. This is called *causal wearout* or *time-dependent dielectric breakdown* (TDDB). The current decay shown in Fig.3.7 may indicate electron trapping, making the oxide more negative and reducing the cathode field supporting tunnelling. Occasionally a current increase has been observed [85,86], possibly caused by mobile positive ions drifting toward the cathode and increasing the cathode field [86]. Causal wearout can be modelled by the equation

$$\int_0^{t_{bd}} \frac{dt}{\tau(F)} = 1 \qquad 3(1)$$

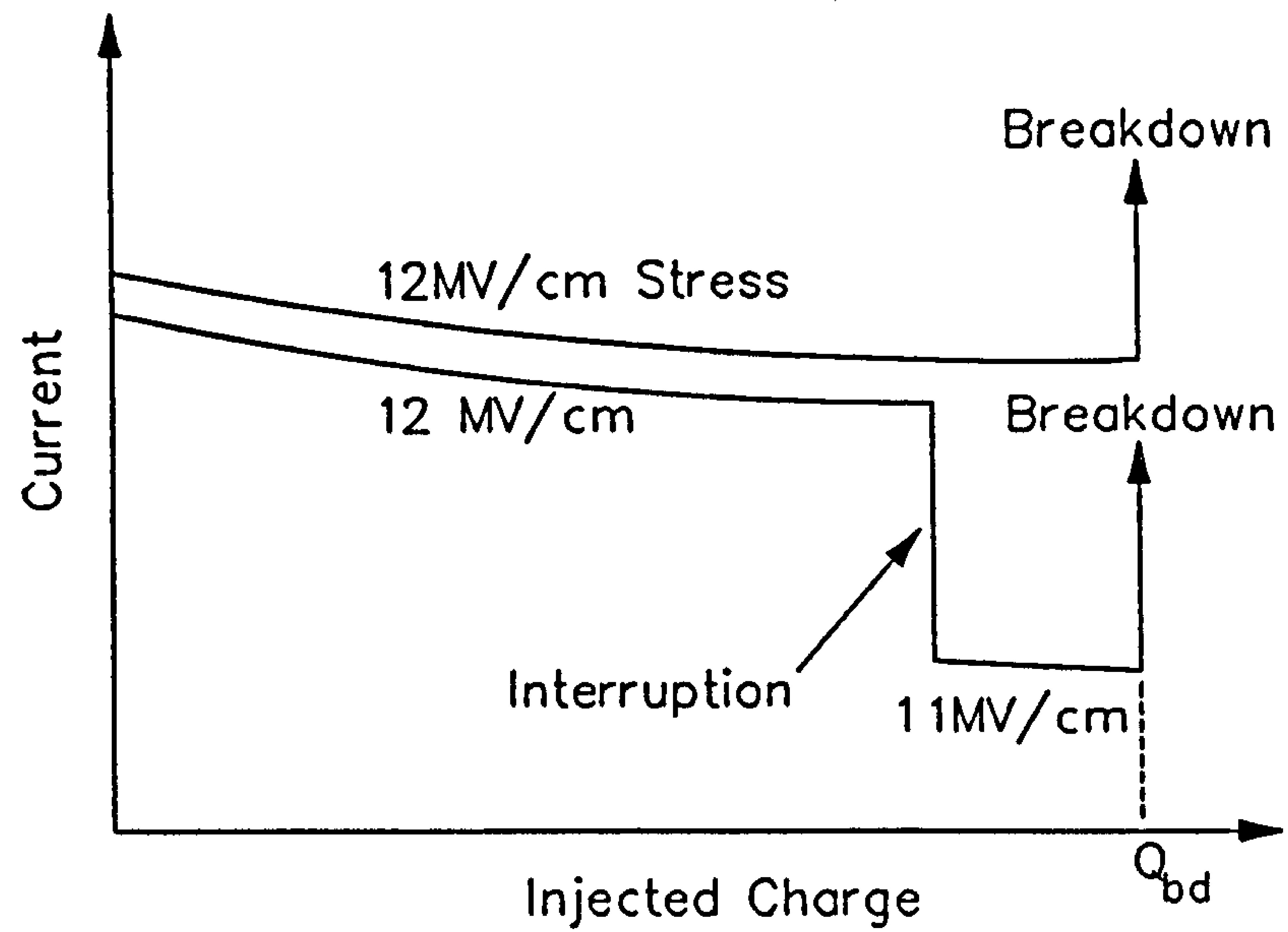


Figure 3.7: Illustration of Causal Dielectric Wearout (after Wolters et al. [57]).

where $\tau(F)$ is the *time-to-breakdown-function*, i.e. the value of t_{bd} under a constant field F. The empirical (or semi-empirical) forms used for $\tau(F)$ are usually proportional to $e^{-\alpha F}$ [87,88], $e^{\beta/F}$ [89,90], or $t^{-\gamma}$ [86,83], α, β and $\gamma(\approx 1/4)$ being constants. Breakdown is also polarity dependent (t_{bd} being generally 2 to 3 times shorter under negative gate bias than under positive bias [90,91,92]) and thickness dependent ($\tau(F)$ for a given field decreasing with increasing thickness).

The temperature dependence of TDDB has been found to follow the Arrhenius equation:

$$t_{bd} \propto e^{\frac{E_a}{kT}} \qquad 3(2)$$

where the activation energy E_a varies with experimental conditions. McPherson and Baglee [93] developed a reaction rate model of TDDB, predicting a field-dependent activation energy. This theory was later extended [94] to produce an accurate unified model of the field/time/temperature dependence of TDDB. It was discovered [94] that E_a is a function of both field and oxide quality, explaining the wide variety of values reported throughout the literature. Tatsuuma et al. [81] found that the activation energy has different values above and below 125°C, suggesting that two different mechanisms are involved.

One of the most interesting properties of TDDB was discovered by Wolters and van der Schoot in 1985 [58]. A close analogy was found to exist between dielectric breakdown and mechanical fracture in solids. Table 3.2 shows the corresponding properties of mechanical and dielectric systems. For example, as charge enters the SiO_2 by Fowler-Nordheim tunnelling, the oxide becomes negatively charged and resists further injection. This

corresponds to the strain hardening of metals, whereby a permanently deformed object resists further strain. Also, the strain required for fracture remains constant under slow elongation (small injection current J). This is known as 'ductile fracture'. Under fast elongation (large J), the fracture strain (Q_{bd}) falls dramatically - a condition known as 'brittle' fracture. This is shown by the data reproduced in Fig.3.8. The critical current between 'ductile' and 'brittle' failure is usually denoted J_c

Several TDDB models have been developed in conjunction with observed parametric drift in stressed dielectrics. The major theories are discussed in the following sections.

Table 3.2: Analogy between Mechanical and Dielectric Systems

(after Wolters et al. [58])

Mechanical System	Dielectric System
Stress (Force)	Electric Field
Strain (Elongation)	Charge Displacement
Elastic Deformation	Dielectric Polarisation
Plastic Deformation	Fowler-Nordheim Tunnelling
Strain Hardening	Electron Trapping
Mechanical Fracture	Dielectric Breakdown

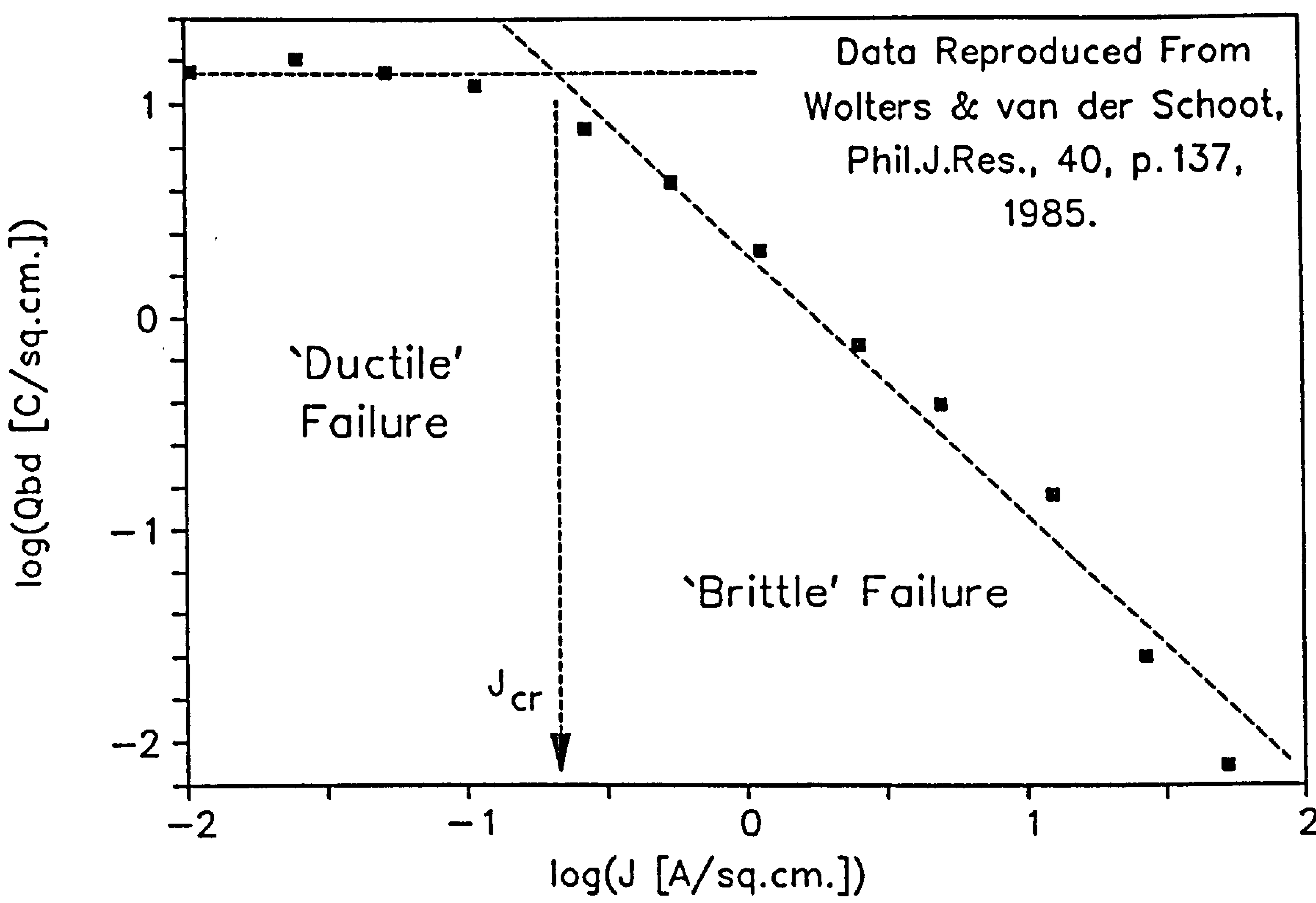


Figure 3.8: Logarithmic Q_{bd} vs. J curve showing 'brittle' and 'ductile' failure modes.

3.3.5.1 Dielectric 'Erosion' Models

The non-reversible nature of oxide wearout suggests a physical 'erosion' of the SiO_2 network over a period of time. However, Jonscher [95] argued that few electrons are capable of gaining sufficient energy from the field to damage an intact SiO_2 structure, a statement subsequently confirmed by calculation [58]. Damage therefore propagates by the extension of pre-existing defects, forming defect *clusters* or *trees* which eventually span the dielectric, supporting breakdown. This scheme, sometimes known as *treeing*, has been developed by several authors, some of whom have used fractals to model propagation [96-98]. However, such models have limited application in ultra-thin dielectrics where self-similarity disappears. Tree growth can also be observed in thick dielectrics in terms of accompanying light emission [99]. Wolters et al. [58] suggested that the rapid releases of energy which occur as electrons are emitted at the anode can cause the generation of traps. Electrons entering these traps also release energy, generating further traps and providing a mechanism for tree growth from the anode to the cathode. Budenstein [100] suggested that the tree structure evaporates at breakdown, forming a plasma-channel across the dielectric which then supports the breakdown current. Several quantitative models have been developed which correlate the rate of trap generation in the tree to the rate of electron injection at the cathode [101,102]. Breakdown is deemed to have occurred when the trap density reaches a critical level.

3.3.5.2 Trapped Charge Related Models

Most of the current models of oxide breakdown are based upon the trapping of charge within the dielectric. Harari [103] observed the generation of a high density of electron traps in SiO_2 under a large tunnelling current. It was proposed that the trapped electron charge enhances the internal oxide field to a critical value at which breakdown occurs. This critical field may be the threshold for avalanche breakdown [104] or the field at which the Si-O bonds dissociate [103]. However, the existence of oxide fields sufficient to break the Si-O bonds has been questioned by other workers [105].

An alternative scheme was suggested by Holland, Chen and Hu in 1984 [106] which has received considerable attention since. It was suggested that holes, released by band-to-band impact ionisation (BBII) in the oxide, drift toward the cathode where some of them are trapped. The resulting enhancement of the cathode field and tunnelling current lead, eventually, to unstable conduction and breakdown. Hole trapping is believed to occurs in small 'weak' areas of oxide, where the interface trap density is unusually high. The different trapping properties of the gate-oxide and Si-oxide interfaces introduce a polarity dependence of the time-to-breakdown [107]. A mathematical version of this model [89], yielded the following form for the hole-charge density

$$Q_p^* = \int_0^{t_{bd}} \eta \, J \int_{T_t}^{T_{ox}} \alpha(F) \, dx \, dt \qquad 3(3)$$

where η is the hole-trapping efficiency (i.e. the probability that a generated hole is trapped at the cathode) and α(F) is the field-dependent impact-ionisation coefficient. According to the Fowler-Nordheim equation, the tunnelling current J is approximately proportional to $\exp(-\beta/F)$, while α(F) is given by

$$\alpha(F) = \alpha_0 e^{-\frac{H}{F}} \qquad 3(4)$$

where α_0 and H are constants. Solving these equations for a constant field F yields the following form for the time-to-breakdown function:

$$t_{bd} = \tau(F) = \tau_0 e^{\frac{\gamma}{F}} \qquad 3(5)$$

where τ_0 and γ are constants. The hole trapping efficiency is a very strong function of the processing parameters, increasing with oxidation and annealing temperature with a subsequent effect on t_{bd}.

However, theory predicts that band-to-band impact ionisation in SiO_2 requires a minimum of 12V across the dielectric [72,105] while TDDB was observed below 10V in ultra thin oxides [105]. One explanation is that ionisation interactions take place between the tail states of the vitreous SiO_2 rather than the mobile bands themselves [108], reducing the effective threshold energy for ionisation. Alternatively, holes might be generated by processes other than BBII. For example, as electrons leave the SiO_2 at the anode they may lose their 3eV excess energy in the creation of hot holes [109], which may tunnel into the oxide [110,111]. Theoretical studies have suggested that the energy transfer between the hot electrons and the electron-hole pairs take place via the emission and decay of surface plasmons [66]. However, the fact that the anode/oxide barrier height does not affect breakdown rather contradicts this theory [108].

The question as to whether electron or hole trapping initiates breakdown remains largely unanswered. The results of Dutoit et al. [112] and Haddad et al. [113] suggest that breakdown is associated with negative charge and Kusaka [114] successfully used a quantitative electron trap generation model to simulate SiO_2 wearout. Tzou et al. [115] observed that the trapped hole charge and t_{bd} both decrease with increasing temperature, while Nissan-Cohen et al. [116] showed that the presence of hydrogen affects t_{bd} but not the hole charge density. Since Eqn.3(3) implies an increase in t_{bd} with decreasing hole charge, these observation tend to contradict the hole trapping model.

On the other hand, a hole current emanating from the oxide cathode has been positively identified [111,117], suggesting that holes are definitely generated in the SiO_2. (The current has the wrong field dependence to have been generated *at* the cathode surface by Fowler-Nordheim tunnelling of valence-band electrons). The integral of this hole current between 0 and t_{bd} is approximately constant irrespective of bias conditions [111]. If the hole trapping efficiency η is constant then the trapped hole density to breakdown is constant also, supporting Eqn.3(3). Furthermore, oxide breakdown has been observed with negligible electron trapping [108], suggesting that its occurrence is unconnected with breakdown.

In view of these contradicting claims, it seems unlikely that a universal breakdown mechanism exists and that the actual wearout process depends on oxide properties and test conditions. However, the trapped hole theory has been found to give excellent quantitative agreement with experimental data, and is often used empirically for reliability modelling [90,118].

3.3.5.3 Other Models

Other wearout models invoke mobile ion migration [119]. According to Ridley [68], the enhanced Fowler-Nordheim current at a Si-SiO_2 interface asperity causes filamentary Joule heating and subsequent activation of Na^+,H^+ and H_3O^+ ions. These ions drift in the field, enhancing the cathode field and injection current, causing acceleration towards breakdown. Further evidence of such a mechanism was provided in 1973 by Raider [86], who observed time evolution of the Fowler-Nordheim tunnelling current consistent with a mobile charge sheet drifting toward the cathode.

Attempts have also been made to correlate oxide breakdown with the generation of interface states during stress [120], although the results have tended to be rather confused [121]. Tzou [115] discovered (using an empirical relationship between trap density and transconductance) that the interface-trap generation time is directly proportional to t_{bd}.

In 1983, Ricco et al. [105] suggested a novel breakdown mechanism for MOS oxides. They proposed that defect sites in the SiO_2 can *resonate* with mobile electron states in the cathode, causing a unity transmission coefficient in the tunnelling barrier. The current in the resulting short-circuit path causes thermal dissociation. The probability of such resonant tunnelling is a function of the interface field which in turn depends upon the trapped charge density. Self consistent solution of trapping, field evolution and resonant tunnelling yields a model of TDDB. The presence of multiple nondestructive breakdown events prior to catastrophic breakdown [51] suggests that resonant tunnelling truly exists in MOS structures. Both this and the hole-trapping model [106] predict a breakdown dependence upon the cathode/oxide interface quality. Such a dependence has been observed by Olivo et al. [84] using oxides whose surfaces were artificially damaged by implantation.

It is clear from the foregoing discussion that no unified model of dielectric breakdown has yet emerged, and the field is still open for new investigations.

3.3.6 Intrinsic and Extrinsic (Defect-Related) Dielectric Breakdown

The existence of intrinsic and extrinsic dielectric breakdown has been identified by numerous workers. Klein for instance found localised breakdown at voltages below the maximum voltage at which the entire oxide was destroyed [45]. These low-field 'extrinsic' breakdowns are most common for very thin dielectrics [38] and might therefore be due to surface non-uniformities causing variations in the effective thickness of the sample. Since breakdown is local, the defect sites could undergo self healing, eventually allowing the

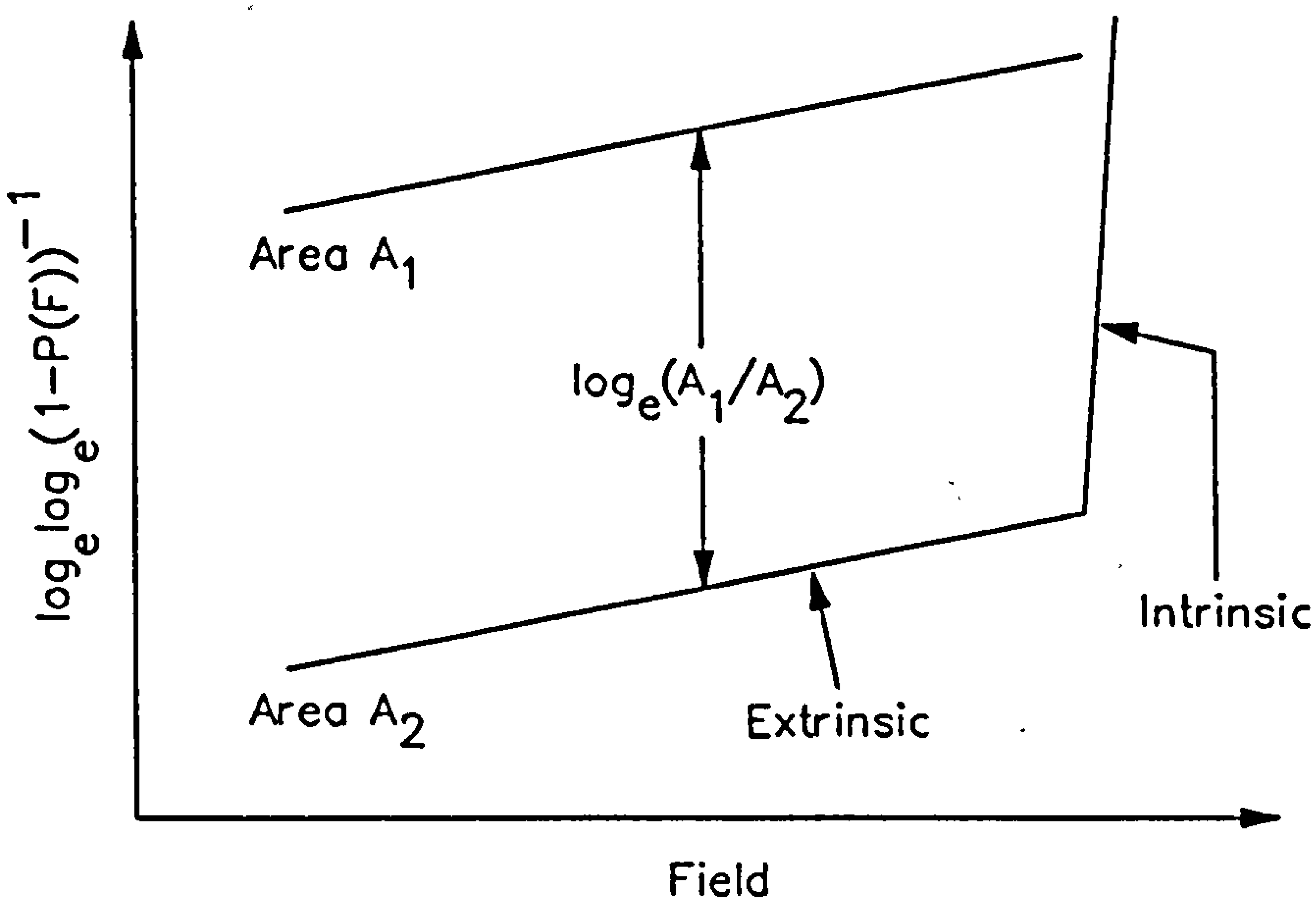


Figure 3.9: Extremal probability graphs for MOS oxides with two different areas (after Wolters et al [57]).

voltage to reach the threshold for intrinsic breakdown. Osburn [122] showed how self healing can undermine attempts to measure the extrinsic defect distribution: Extrinsic breakdown sites, activated above a certain threshold field, may self heal, thereby going undetected and registering as intrinsic. Thus the extrinsic breakdowns can appear to be concentrated into a tight distribution, when in reality they are spread continuously across a field range [e.g.58]. This problem is overcome by using an edge-triggered device to detect breakdown events as they occur [122].

Wolters and van der Schoot [58] showed that if defects are randomly distributed across the oxide area, then the probability of failure P(F) under a field F is given by

$$P(F) = 1 - \exp[-A \cdot D(F)] \qquad 3(6)$$

where A is the oxide area and D is the density of defects activated by a field less than or equal to F. Plotted on an extremal probability format ($\log_e \log_e (1-P(F))^{-1}$ vs. F), the data appears as two straight lines indicating intrinsic and extrinsic breakdown. The position of the extrinsic line varies linearly with area (Fig.3.9) in accordance with Eqn.3(6).

Whilst the intrinsic distribution might be expected to be concentrated at a single field, in reality it has a finite width (Fig.3.9). Since the distribution is between two and five times too wide to be explained by macroscopic thickness variations [122], it has been suggested that the so-called 'intrinsic' breakdown is in fact due to closely distributed defects [66].

The extrinsic defects may be regions of thin oxide or oxide pinholes (diameter .1-1μm) into which gate material migrates, causing localised high-current regions. Alternatively

they may be the result of ionic contamination causing localised barrier lowering [123,124] or oxide decomposition ($Si+SiO_2 \rightarrow 2SiO\uparrow$) [125]. Falster [126] identified two different extrinsic breakdown modes, suggesting that the defects may be of more than one type. Extrinsic defect density has also been correlated with particulate contamination during oxidation [127]. The defects are usually studied by measuring the breakdown distributions obtained from a large number of oxides. Alternatively the non-destructive detection of defects has been achieved using tunnelling current microscopy (TCM) [128].

Whatever the exact physical nature of these defects may be, they can be modelled as localised regions of thin oxide. This 'Effective Thickness Theory' was developed by Lee et al. [129,130] in conjunction with the causal model of oxide breakdown. In 1989 Coleman extended Lee's model, showing that it is independent of the form given to the $\tau(F)$ function [131].

Translating Eqn.3(6) to the effective thickness model, the probability of an oxide containing a defect of effective thickness equal to or less than T_{eff} is [90]

$$P\left(\le T_{eff}\right) = 1 - \exp\left[-A \cdot D\left(T_{eff}\right)\right] \qquad 3(7)$$

If the defects are not perfectly random, then Eqn.3(7) can be written [129]

$$P\left(\le T_{eff}\right) = 1 - [1 + ADS]^{-1/S} \qquad 3(8)$$

where S is a parameter which characterises the degree of defect clustering (S=0 for a perfectly random distribution).

This effective thickness theory of extrinsic defects has been used to develop *CORS* (Circuit Oxide Reliability Simulator), a pre/post-processor for the Berkley *SPICE* electronic circuit simulator, allowing oxide reliability to be modelled for any given circuit structure [90].

3.3.7 Fast Transient and ESD Dielectric Breakdown

Some workers have chosen to apply Eqn.3(1) to very short duration ($<1\mu s$) transient stress [132] in order to predict breakdown under any arbitrary pulse waveform F(t). However, insufficient data is available to determine whether or not the causal model is applicable under these conditions. Other workers have investigated the effect of bipolar a.c. pulses (positive polarity followed by negative polarity) upon ultrathin SiO_2 films, simulating the conditions under which E^2PROM tunnelling oxides are operated. The resulting Q_{bd} is much higher than under d.c. stress, suggesting that opposing stress polarities have opposing

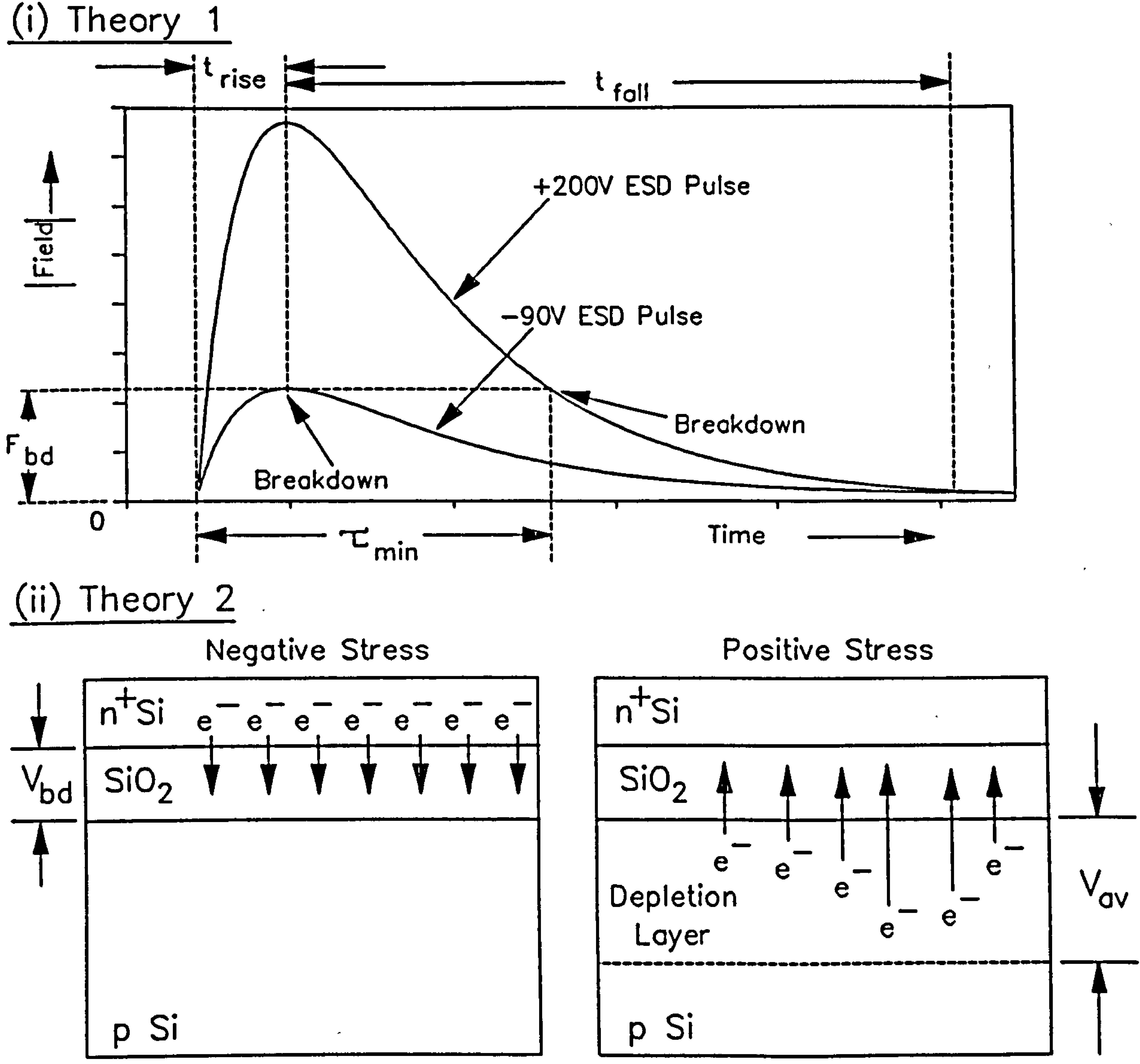


Figure 3.10: Amerasekera & Campbell's Models of Polarity Dependent ESD Breakdown in p-Substrate MOS Structures.

effects on the oxide.

Oxide breakdown under HBM pulse stress has been extensively studied by Amerasekera and Campbell [7,49], who showed that the positive polarity breakdown voltage threshold in a p-type MOS capacitor is considerably larger than the negative threshold. Two alternative models were proposed to explain this behaviour, both of which are illustrated in Fig.3.10. According to the earlier model [7] (Fig.3.10(i)), breakdown requires the simultaneous existence of an oxide field F_{bd} and a cathode inversion layer of injectable electrons. An ESD pulse profile with rise-time $t_{rise} \sim 500$ns and fall-time $t_{fall} \sim 1$ms is also assumed. The electron-rich polysilicon gate serves as cathode under negative stress, and breakdown is therefore governed by the peak oxide field only. Under a positive pulse however, inversion layer growth is governed by the *minority carrier response time* τ_{min}

(~ms) and a supply of injectable electrons may not exist until well into the t_{fall}. At this point the field has fallen to well below its peak value and the magnitude of the breakdown voltage magnitude therefore exceeds its negative counterpart.

The second model [49] (Fig.3.10(ii)), examines the behaviour of the oxide prior to inversion, when the Si in a state of deep depletion (i.e. the positive charge on the gate is entirely compensated by a negative depletion layer in the bulk silicon). The model assumes that ESD breakdown requires inversion within the 500ns pulse risetime and such a rapid inversion rate can only be supported by a depletion layer avalanche. When an avalanche is induced, the oxide is bombarded with high energy electrons and immediately breaks down.

Close examination reveals that neither of these models is applicable to unprotected MOSFET structures, in which the oxide itself provides the only discharge path for the pulse charge. The pulse fall-time is therefore governed by the oxide tunnelling current, which depends in turn upon the cathode surface inversion, producing an inherent contradiction. ESD polarity dependence is therefore an open subject for research.

ESD breakdown thresholds were also shown to be considerably higher than the corresponding slow-ramp thresholds [133] (although this was partly explained by the capacitive loading of the oxide upon the ESD source [49]). This, together with the temperature independence of ESD [134] suggested that ESD breakdown is due to a different mechanism from continuous voltage wearout. In view of the short ESD time duration, electron avalanche breakdown was proposed as the most likely mechanism [7]. It was later shown that the electrostatic energy $\frac{1}{2}C_1V_{bd}^2$ required to support negative polarity breakdown is approximately constant, suggesting a thermal breakdown mechanism [49]. Other workers have identified the snap-back mode (see Chapter 2) as an ESD hazard in output MOSFETs. A high voltage drain pulse can easily trigger a device into the snapback mode [135], leading to hot hole injection into the oxide and subsequent gate-drain dielectric breakdown [136].

3.4 Hot Carrier Injection

The previous section described degradation due to Fowler-Nordheim tunnelling of electrons into SiO_2 under high electric fields. However, electrons can also enter the oxide under much lower fields by gaining sufficient energy to surmount the Si-SiO_2 potential barrier. These *hot electrons* can be thermally or optically excited, or induced by *field stimulation*. Field stimulation involves field-induced excitation of the mobile carrier gas to a mean energy greater than that under equilibrium conditions. This is the major cause of hot electron instability in MOS devices.

Once in the oxide, electrons with energies above 3.7eV can generate interface traps

[137]. These traps are believed to be hydrogen related, and are formed by the breaking of Si-H bonds [138]. The subsequent space-charge evolution causes the device parameters to drift with time, and the increased interface state density enhances 1/f noise in the circuit [139]. Although hot carrier injection cannot itself support dielectric breakdown [140], it has been shown to reduce an oxide's resistance to future high-field stress [140-142].

3.4.1 Field Stimulation of Hot Carriers

Mobile electrons have two components of motion: a random 'thermal' motion and a motion due to macroscopic drift or diffusion. The latter is characterised by a 'drift velocity' $v_d=-\mu_e F$ (or a 'diffusion velocity' $=D_n.dn/dx$). The carrier gas can be assigned a temperature T_e such that the mean electron kinetic energy due to the random thermal motion is equal to $1.5kT_e$. Hence

$$\frac{3}{2}kT_e = < \frac{1}{2}m_0(v-v_d)^2 > \qquad 3(9)$$

where v is *total* carrier velocity [137] (< > denotes mean value). Thus if the electrons form a Maxwellian energy distribution, a few 'hot' carriers exist at energies well above $1.5kT_e$.

An electric field accelerates the carriers, increasing their mean kinetic energy. Subsequent interaction with lattice vibration modes scatters the carriers, causing them to lose energy (electron-phonon interaction). An equilibrium condition exists at which the average energy gain and loss are equal and the mean energy for a given field remains constant. The equilibrium energy, the electron temperature T_e and hence the proportion of 'hot' carriers increase with increasing field.

The hot carrier distribution clearly depends upon the scattering mechanisms in the semiconductor. Let p(r).dr represent the probability that a given carrier travels a distance between r and r+dr prior to scattering. Assuming the existence of a single electron-phonon interaction mechanism with a single mean free path length λ, the interaction probability is proportional to the distance travelled giving

$$p(r) = \frac{1}{\lambda}\exp\left[-\frac{r}{\lambda}\right] \qquad 3(10)$$

and the probability of a mean free path greater than d is given by

$$P(r>d) = \int_d^\infty p(r)\,dr = \exp\left(-\frac{d}{\lambda}\right) \qquad 3(11)$$

If several mechanisms exist with mean free paths λ_1,λ_2...etc. then the overall mean free path $\lambda=1/(1/\lambda_1+1/\lambda_2+...)$ [143].

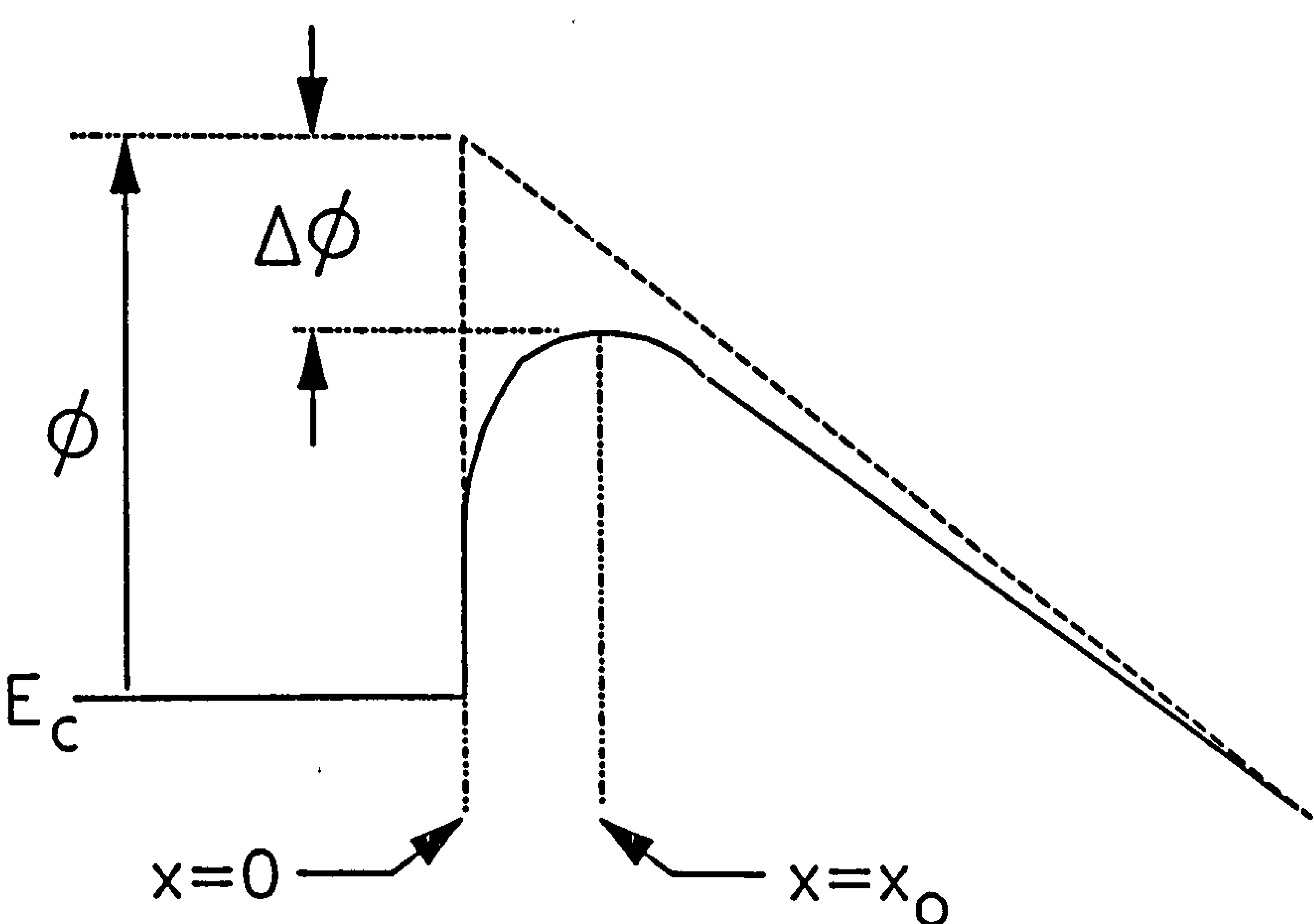


Figure 3.11: Schottky barrier-lowering effect.

3.4.2 Modelling Hot Carrier Injection

In order to enter the oxide, a hot electron must surmount the $Si\text{-}SiO_2$ potential barrier. Once in the insulator, the electron induces a positive 'mirror image' charge +q in the conductive cathode, the proximity of which lowers the electron's potential energy. This reduces the $Si\text{-}SiO_2$ barrier height by $\Delta\phi=(q^3F/4\pi\epsilon_0\epsilon_{ox})^{½}=\beta F^{½}$ (where $\beta=2.6.10^{-}eV^{½}cm^{½}$) at a distance $x_0=(q^3/16\pi\epsilon_0\epsilon_{ox}F)^{½}$ from the interface (see Fig.3.11). This is called the Schottky image-force effect. In principle, an electron with an x-component of momentum (normal to the interface) equal to or greater than $p_c=(2m_e^*[\phi-\Delta\phi])^{½}$ reaches x_0 and is thereafter drawn into the oxide by the field.

The energy distribution of field-excited carriers is known to be Maxwellian and the injection current density J is therefore given by standard thermionic emission theory:

$$J = R^{*}T_{e}^{2}\exp\left[-\frac{\phi - \Delta\phi}{kT_{e}}\right] \qquad 3(12)$$

where R^{*} is the effective Richardson constant. At low electric fields, J is slightly smaller than the value predicted by Eqn.3(12), due to inelastic interactions between 0 and x_0 [143]. This can be corrected by including a factor $\exp(-x_0/\lambda_{ox})$ [i.e. the probability that a carrier with a mean free path in the oxide of λ_{ox} will travel a distance x_0 before an interaction].

Alternatively the *Lucky Electron Model* may be used. This is closely analogous to Shockley's impact ionisation model [143], ignoring the concept of an overall temperature T_e and concentrating exclusively on those electrons who gain the critical energy $[\phi\text{-}\Delta\phi]$. The central assumption is that the electrons which attain $[\phi\text{-}\Delta\phi]$ do so in a single flight, evading all interactions along the way. Since the flight path required to attain $[\phi\text{-}\Delta\phi]$ may be several times the mean free path length, these carriers are termed *lucky electrons*. In order to enter the oxide, these electrons must be accelerated over a distance d **normal** to the Si-SiO_2 interface such that $\int_d^0 F(x).dx = [\phi - \Delta\phi]/q$. The probability of a carrier travelling a distance d without undergoing any interactions is given by Eqn.3(11). Although the simplicity of this model renders it attractive, the assumption that *only* lucky electrons are injected may be unsound. However, a fairly close agreement with experiment can be obtained by including a pre-exponential factor P_0 in Eqn.3(11). Thus the probability P of an electron entering the oxide can be written [143]

$$P = P_0 \exp\left(-\frac{d}{\lambda}\right) \qquad 3(13)$$

The lucky electron model implies a Maxwellian distribution of electron energies. Although such a distribution is not universally consistent with the results of Monte Carlo simulations, it can be shown to be accurate in the high energy tail of the distribution and therefore suitable for hot electrons [144]. The only adjustable parameters in the model are P_0 and λ which are independent of silicon doping and the applied voltages.

Although models have been developed which include the injection of carriers which attain the energy $[\phi\text{-}\Delta\phi]$ over several interactions, these require lengthly numerical computation. Within a restricted range of experimental conditions, Eqn.3(13) agrees with the numerical models to within one order of magnitude and in the light of the other inaccuracies this approximation is more than adequate. The model is only applicable to one-dimensional structures, i.e. those in which the electric field is normal to the Si-SiO_2 interface. When the field is parallel to the interface, injection must be modelled by the thermionic emission equation.

3.4.3 Avalanche Breakdown and Injection in MOS Capacitors

The hot electron injection rate depends on the electric field and the semiconductor carrier density. Although field and carrier density are usually related inversely, a high field and a high carrier density can co-exist in the presence of avalanche conduction. Hot carrier injection under these conditions is called *avalanche injection*. The mechanism, first proposed in 1960 by Nicollian and Goetzberger [145], has been shown to support current densities up to 10A/cm^2 [146].

Avalanche injection can be readily observed in a p-substrate MOS capacitor under *avalanche breakdown* conditions: A positive potential applied to the gate structure drives the silicon into depletion until an inversion layer forms at the oxide surface. The growth of the inversion layer is governed by the thermal *minority carrier response time* τ_{min}, which is typically 100-1000ms. For a fast pulsed gate bias ($<$1ms), the inversion layer is practically non-existent and the depletion layer can expand beyond its equilibrium width W_m. When the depletion layer voltage reaches the avalanche threshold V_{av} it becomes pinned and an inversion layer forms from electrons supplied by the avalanche plasma.

Under certain doping conditions, the high carrier density of the avalanche plasma is accompanied by a sufficiently high field over a sufficient distance to cause avalanche injection over the Si-SiO_2 potential barrier. Since the system is one-dimensional (the avalanching carriers travel normal to the Si-SiO_2 interface) it can be modelled by the lucky electron model.

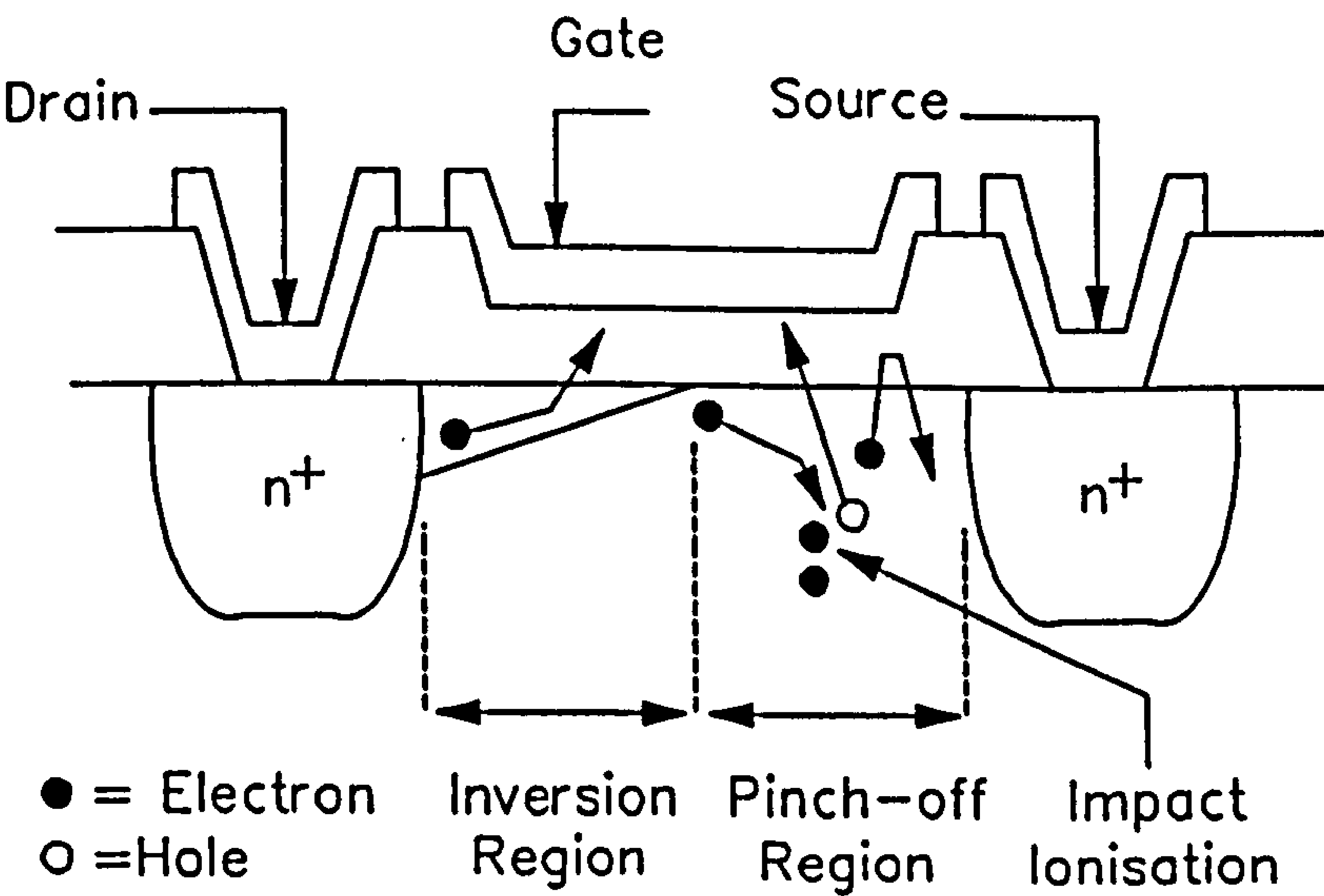


Figure 3.12: Hot carrier injection mechanisms in a MOSFET.

3.4.4 Hot Carriers in MOSFETs

In the following discussion, an n-channel MOSFET shall be considered, although hot carrier damage in p-channel devices is comparable (with the roles of electrons and holes reversed) [147]. The *channel hot carrier effect*, occurring principally when a transistor is biased beyond the pinch-off point, is the major injection mechanism. Fig.3.12 shows the major injection processes: The longitudinal electric field in the inverted channel causes field excitation of hot electrons and subsequent injection into the SiO_2. In the 'pinch-off' region between the inversion layer and the gate, electrons repelled from the surface by the field can generate holes by impact ionisation which can be heated and injected into the SiO_2 (a few electrons may also be injected in this region but most will be reflected by the field). This 'drain avalanche stress' has been found to degrade the oxide much faster than the channel hot-carriers [148]. The effects of hot carrier stressing generally increase with decreasing dimensions but decrease for channel lengths below 0.2μm [137].

Some of these injected carriers are trapped in the SiO_2, causing space charge evolution [149]. Thus the value of V_T changes with time, causing the device to drift from E-mode to D-mode or vice versa. The breakdown voltages of the source and drain junctions can also be affected and an uneven charge distribution along the channel can even affect the transconductance g_m [149]. Variations in V_T and g_m have been found to vary linearly with the number of generated traps [150]. The severity of hot carrier damage increases with temperature, although a greater number hot carriers are injected at low temperatures [151,152]. Hot-carrier sensitivity can be reduced by the incorporation of F or Cl in the SiO_2 [153,154]. Hot electron damage can also be annealed at room temperature by grounding the drain contact for 600-900s [155].

Several links have been established between hot carrier injection and ESD. Sub-catastrophic ESD damage has been shown to affect the hot electron reliability [156] and ESD induced snap-back behaviour (see Section 2.4.2) can cause catastrophic oxide breakdown by the injection of hot holes from the impact-ionization [112,140].

Transistor lifetime under hot-carrier stress can be predicted using the charge-pumping technique to monitor interface states [157] and electron injection can be observed in-situ by detecting optical emissions [158]. Alternatively, the inclusion of Ge impurities in the channel region can increase the scattering, thereby reducing the number of lucky electrons [159]. However, the reduction of hot electron effects often conflicts with strategies to reduce the second-order transistor phenomena [160].

3.5 Other MOS Failure Mechanisms

3.5.1 Parasitic Bipolar Effects

The effect of a parasitic bipolar structure on a miniaturised MOSFET has been discussed in Chapter 3. When the 'snapback' condition is reached, the drain presents a resistance whose value decreases with increasing current. Unless the current in such a structure is limited, this *negative differential resistance* may lead to current runaway and subsequent thermal burnout [161]. Snapback also provides a supply of hot holes for injection into the oxide [140,142].

Parasitic bipolar transistors can also form between adjacent MOSFETS in a CMOS circuit. These devices are arranged so as to act as a 'thyristor' or 'silicon controlled rectifier' (SCR). The SCR can be triggered by a spurious pulse (e.g. an ESD event) or by exposure to ionising radiation, causing a short circuit between V_{ss} and V_{dd} and subsequent thermal burnout [162]. Parasitic SCRs and latchup can also be induced in NMOS as a result of the extra junctions induced by junction spiking [163] (discussed below). Latchup can also be self-propagating, spreading from location to location across the chip. This propagation can be observed using infra-red [164] or emission microscopy [163]. Latchup susceptibility has also been shown to degrade with increasing temperature [165].

A number of techniques are available to combat latchup. For example, careful design can keep the parasitic transistor gain [162] to a minimum, increasing the latchup triggering voltage. This can be achieved geometrically, or by using gold doping in order to reduce the minority carrier lifetime [166]. Alternatively, trenched barriers can be fabricated in order to separate adjacent n-channel and p-channel structures [167]. An excellent review of latchup is given in the book by Troutman [168].

3.5.2 Electromigration [2]

Since the atoms in a metallic crystal have a Maxwellian energy distribution, a proportion of them will have an energy greater than the bonding energy of the lattice and free themselves from their lattice positions. The number of these 'mobile' atoms is therefore related to the lattice temperature by the Arrhenius equation. Under zero current conditions the motion of the mobile atoms is random so that the material maintains its original shape. However, if a current flows, the electron 'wind' will cause a preferential 'migration' of material in the direction opposite to that of the current. Furthermore, the rate of migration

increases with increasing current density, causing it to be highest at constricted points in a conductive track or wire. Hence these 'bottlenecks' become increasingly narrow and ultimately fuse due to Joule heating. Thus the end result is an open circuit failure.

3.5.3 Electro-thermomigration ('Spiking')

Contact injection of minority carriers at the ohmic contacts of a MOSFET can generate an electron-hole plasma in the drain (or source) diffusion [169]. The resulting Joule heating can provide a condition for electromigration and thermal diffusion of aluminium or polysilicon into the silicon substrate [170]. This results in a highly conductive filament or 'spike' of silicon penetrating the diffusion and forming a junction short. Depending upon bias conditions, these spikes may point downward into the substrate or they may penetrate the diffusion side wall. This may cause p-n junctions to behave in an Ohmic fashion, or it may set up the initial condition for latchup [163].

3.5.4 Dielectric Instability

The charge distribution in the oxide can alter during a device's working life, causing threshold voltage drift and resulting parametric instability (see Eqns.2(9-11)). The charge may be situated in surface states at the Si-SiO_2 interface, or in bulk traps within the SiO_2 volume. Mechanisms of bulk-oxide charging include hot carrier injection and high field Fowler-Nordheim tunnelling. Both these mechanisms have similar effects on dielectric stability, causing V_T to shift with time [149]. The oxide charge is usually positive and is identified by a negative shift in the capacitance vs. voltage (C-V) curve [171]. Other causes of dielectric instability include mobile ion migration [172] or Si-SiO_2 interface instability [173].

3.5.5 Radiation Damage

Ionising radiation (e.g. X-ray, γ-ray) can have a number of adverse effects in MOS circuitry. For example, particle irradiation (e.g. α-particles emitted by radioactive trace elements in the packaging material) can generate electron-hole pairs and subsequent electrical pulses, which may trigger soft memory errors [174] and/or latchup [175]. Radiation can also

affect MOS oxide stability in a similar manner to high electric field stress, i.e. the generation of defect states in the oxide [176,177]. These states, which are also generated by exposure to ultraviolet light [178], may be associated with the breaking of Si-H or Si-OH bonds [179] and may exist as both donors and acceptors [180]. Thus these states can be either positively or negatively charged, introducing V_T drift of either polarity [177]. The radiation tolerance or *hardness* depends heavily upon processing parameters such as Si doping and oxidation/annealing conditions [181]. The introduction of small quantities of fluorine into the oxide has been shown to improve the radiation hardness [182].

The similarity between electrical and radiation-induced damage allows radiation hardness to be assessed using Fowler-Nordheim tunnelling by the hot-carrier avalanche injection technique (see Section 3.4.3) [183].

3.6 Latent or 'Walking Wounded' Failure

Latent or *walking wounded* failure is a combination of time and event dependent phenomena. It results from the cumulative nature of damage which allows time and event dependent degradation to combine to cause catastrophic failure. A typical example is the partial evaporation of a metallisation track caused by an ESD event. Such damage may well pass an electrical (i.e. non-visual) inspection and enter the field [184]. However, the current constriction at the damage site causes accelerated electromigration, resulting in premature failure during working life. The possible consequences could be devastating for manufacturers and consumers alike. While the cost of a walking wounded component detected during manufacture may be negligible, it could run into millions of pounds when in service (e.g. in a satellite in orbit). More frightening still, a latent failure in a military application could cost millions of lives.

It is therefore surprising that until the end of the 1970s, very few workers had considered latent failure, many still doubting its existence. In 1980, McAteer et al. [185] unveiled an experimental program to identify latency problems in the proposed Trident missile system. The results, published in 1982 [186], showed that the problem was very real. More alarming still was the discovery that a damaged device may pass visual as well as electrical inspection, the damage only becoming apparent under chemical etch, SEM analysis [186] or infra-red microscopy [187].

Latent damage can sometimes be observed as sub-catastrophic performance degradation. Observing the effects of ESD on polysilicon-gate NMOS performance, Amerasekera and Campbell [134] identified several modes of characteristic degradation, the main hallmark of breakdown being a reduced transconductance. 'Reduced transconductance'

devices are of particular importance since they behave as operational transistors and may therefore pass a functional inspection. Such degradation could be the result of charge-trap buildup in the oxide [188] or in a local region of broken-down oxide [134]. Amerasekera [7], Holmes [189] and Fong and Hu [132] reported negative threshold voltage shifts in MOSFETs stressed with ESD pulses below the oxide breakdown threshold (although Holmes also reported time-dependent recovery), indicating the storage of positive charge which may be associated with degradation. Analogous degradation in GaAs MESFET [190] and bipolar [186] devices has also been identified.

Alternatively a device may show no detectable signs of latent damage prior to catastrophic failure. These may correspond to gate oxides which, although still intact, have been stressed almost to the point of critical wearout. Models have been developed based upon the cumulative growth of microcracks produced by thermoelastic stress relief [191]. This echoes McAteer's warning that a walking wounded device may remain deceptively within its specification until failure [185].

Experimental data relating to the MOS latent hazard is varied and contradictory. McClullough [192], Schwank [193] and Woodhouse et al. [194] found little evidence of degradation in MOS integrated logic subjected to sub-critical ESD damage and Bowers [195] found gate oxide latency in discrete MOSFETs almost impossible to achieve. Furthermore Roberts [196] discovered no cumulative damage from rectangular pulses below 70% of the damage power threshold. Latent failure was, however, observed in LSI NMOS by Enoch et al. [16], in MOS memory by Strauss et al. [197] and in HMOS by Crockett et al. [198], although the latter was attributed to polysilicon resistor burnout. Krakaur & Mistry [199] showed that trapped holes, induced in the oxide during ESD snap-back could be transformed into interface states under working conditions, causing parametric drift. Aur [136] showed that ESD damage reduced the lifetime of an NMOS transistor under hot-electron stress (although interestingly enough the reverse is not true: hot electron stress has no affect on ESD performance). Hellström [200] successfully used third harmonic distortion (THD) analysis to detect latent damage in MOSFETs, although manufacturing defects were discovered to produce similar results.

The opposite of latent failure, i.e. latent recovery, has also been recorded and devices whose characteristics have degraded out of specification can sometimes recover under further stress. This has been attributed to dielectric self-healing [48,78] or the re-melting/re-solidification of junction spikes [16].

3.7 Failure Mechanisms in Different MOS Technologies

Some of the different MOS technologies covered in Chapter 3 are more prone to certain failure mechanisms than others [2]. For example, CMOS is by its very nature prone to latchup. This is partially remedied by the use of insulated substrate technology, isolating the n and p-channel MOSFETs and preventing them from interacting to form an SCR. However, a single transistor latch mechanism has been reported [201], to which SOI technology has no immunity. Although SOI and SOS are prone to the same dielectric instability mechanisms as ordinary CMOS [202], they exhibit a greatly improved radiation hardness [203].

The relative sensitivity to different failure mechanisms also depends on device size, although early experiments on HMOS reliability gave similar results to standard NMOS [204]. As for VMOS, the major failure mechanism appears to be dielectric breakdown [205], the apex of the V-groove being particularly sensitive due to the high electric field [206]. The failure rates of TDDB in VMOS has been measured at approximately 1% per 100,000 hours [207].

The similarity between charge coupled devices and other MOS structures renders them prone to similar failure processes [2], although they have been found to be specially sensitive to dielectric instability [19,121,208].

3.8 Methods of Improving Reliability

Numerous strategies for the elimination of event-related failure have been developed over the years. These strategies fall into two categories: firstly on-chip protection circuitry can be implemented at the design stage to limit device voltage and absorb the energy of a spurious pulse. Sensible chip design also minimises ESD hazards [31]. Secondly anti-static precautions can be implemented at the manufacture, transportation and handling stages.

Time dependent failure under working conditions can also be minimised by the screening of components after manufacture, weeding out the weak elements from a component batch. All these strategies are described in the following sections.

3.8.1 Anti-Static Precautions

Strategies for the minimisation of electrostatic charge were introduced by Freeman [209] in 1974 and have been considerably improved since. These strategies include the use of *static dissipators* and *anti-static materials* (i.e. materials which dissipate charge and resist triboelectrification) in packaging and handling materials as well as in the floors and furnishings of the working environment. Components should only be handled in a safe handling area (SHA) equipped with equipment, floors (or mats), conveyer belts [12], benches and tools all grounded [210]. Since the fast dissipation of static can itself cause damage, static dissipators should have moderate (rather than low) resistivity [34] and the grounded via a 1MΩ resistance [11]. Although a high relative humidity is beneficial, it provides only a slight advantage [211] and it is usually limited by other factors [8]. Materials can also be treated with 'topical antistats' which are both inexpensive and effective [212]. Operators should wear non-synthetic clothing, conductive shoes, wrist (or heel) straps and should be trained in anti-static methodology [12,8,184,213]. The SHA can also be neutralised by the use of an air ioniser [214], although components should not be brought too close to the ion source [34]. Since this occasionally causes operator discomfort [215], it is rarely used except for the most highly static sensitive devices. Kolyer [34] recommends that conductor and insulator voltages should not exceed 50V and 300V respectively.

The dramatic benefits introduced by such precautions are illustrated in a paper by Gilby [12], showing a 30% reduction in the failure rate of television receivers. Furthermore Giusti [211] calculated that in the absence of anti-static protection, the failure probability for an NMOS device handled at a work bench can be as high as 87% (20% relative humidity). However, care should be exercised in the choice, implementation and maintenance of anti-static equipment, weighing capital expenditure against effectiveness [12,19,216]. Incorrect implementation can undermine the whole anti-static campaign [33] and can even increase rather than decrease the static problem [19]. The most important factor in the elimination of static has been shown to be the operator's awareness of the problem [34].

3.8.2 Protection Circuits

Protection circuitry is built into most modern integrated circuits to protect them from event related failure. With the recent increase in ESD sensitivity, considerable research into protection circuitry has been performed. Protection circuits are required to protect the DUT not only from over-voltage but also from latchup [217]. Specification for the ideal protection

circuit is as follows: it must (i) limit the voltage across device inputs, (ii) shunt the excess current with minimal heating, (iii) not affect the device's performance, (iv) require minimal additional processing and (v) occupy minimal chip area. With the recent reduction in device dimensions however, this ideal has become increasingly difficult to attain.

Numerous protection schemes have been developed over the years [218]. A very simple example might consist of a current-limiting resistor and a voltage-limiting Zener diode, the breakdown voltage of which is set below the damage threshold of the device. Both the diode and the resistor can be fabricated together by diffusing a p-type resistor into an n-type substrate. When a large spurious voltage appears at the input, the diode pins the device voltage at the Zener threshold and the excess voltage is dropped across the resistor R. Another common design replaces the Zener diode with two ordinary diodes, connected to the supply rails. One or other of these diodes becomes forward biased when the input voltage goes above V_{dd} or below V_{ss}, thereby clamping the device voltage within these limits. Avalanche, punchthrough, spark-gap, varistor and snapback structures have also been used to the same effect [219-222], although the snapback devices have too slow a response to protect against CDM ESD stress [13,223].

Although the circuit's ability to limit the device voltage requires a large value of R/(diode series resistance), making R too large causes frequency response problems associated with the time constant $C_{in}R$ (C_{in}=input capacitance). Thus a trade off exists between the diode dimensions and its voltage limiting ability and most protection circuits are only effective up to 2kV.

However, most failures in protected MOS circuits are caused not by inadequate voltage limitation but by heat removal (except for CDM on snapback protected devices [13]), analogous to the thermal dielectric breakdown [224]. The energy absorbed by the diode junction during a fast pulse may not disperse in time to prevent the local temperature from reaching the silicon melting point. Thermal breakdown of diode junctions is not only relevant to protected MOS but also to bipolar and GaAs MESFET circuitry and several mathematical models have been developed to simulate it [e.g.225-228].

The protection circuit resistor is also vulnerable to Joule heating and thermal meltdown. Deposited polysilicon resistors are more sensitive than diffused resistors since they are thermally insulated by the passivation [13].

The above strategies protect against *direct* ESD applied to the terminals of a device. *Indirect* ESD, i.e. the electromagnetic disturbance produced by a local ESD event can be protected against by the use of conductive shields [229].

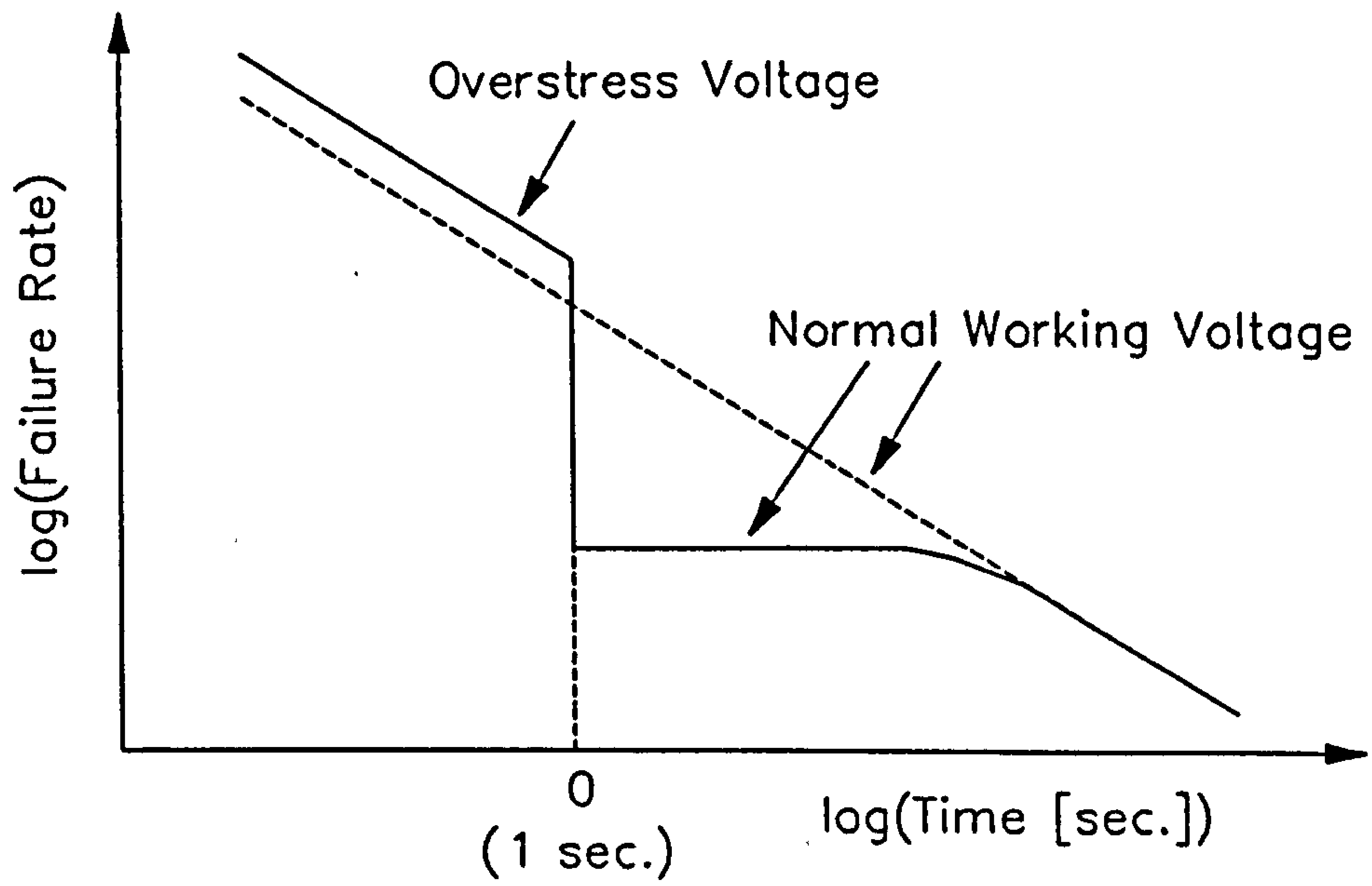


Figure 3.13: Failure rate as a function of time for screened and unscreened components (after Crook [201]).

3.8.3 Screening

Screening is a technique of reducing extrinsic defects from a component batch, thereby eliminating the 'infant mortality' section of the bathtub curve. After manufacture, the devices are stressed above their rated voltage or temperature in order to accelerate the failure mechanisms (screening under temperature stress is also called *burn-in*). In this way the devices with the more severe extrinsic defects fail very quickly and are eliminated from the batch. The survivors can be shown to have a much lower failure rate than the original batch.

Fig.3.13 shows the failure rate vs. time curve for unscreened MOS oxides together with the corresponding curve for screened oxides during and after screening [230]. Although the failure rate is higher during the screen, it is dramatically lower than that of the unscreened components after stress. Only towards the far right of the graph do the failure rates merge.

The major problem is to devise means of activating the extrinsic defects without severely affecting the intrinsic reliability of the device. One example is *High Temperature Reverse Bias (HTRB)* screening [231], in which oxides are subjected to simultaneous high voltage (max. volts + 50%) and high temperature (150°C) in order to activate any ionic contaminants. ESD pulsing has also been found to provide an adequate screen-stress for CMOS logic gates [192]. Although burn-in is a proven technique for MOS screening, the devices are more sensitive to ESD at the elevated temperature [210] (this is because of the protection diode's temperature sensitivity).

In theory, a latent ESD damage may be regarded as an extrinsic defect and thereby eliminated by a screening process. In practice however this is extremely difficult [186]. Indeed the increased handling during screening can increase the probability of ESD [19].

3.9 Conclusions

This chapter has reviewed the current understanding of failure mechanisms in MOS devices. The main topics of research can be summarised as follows:

1. Hot electron instability.
2. Parasitic bipolar mechanisms (Latchup & Snapback).
3. Thermal breakdown in protection diode structures.
4. Radiation damage.
5. Metallisation failure (e.g. Electromigration).
6. Dielectric wearout under long time-scale stress.

A considerable body of literature is available on areas 1 to 5, and although research on these topics is still proceeding, the underlying physics is fairly well understood. A large number of papers have also been published on long-term dielectric wearout, although there is still no general consensus as to which physical mechanisms are responsible. However, reasonably accurate empirical models have been developed for use by reliability engineers (Sections 3.3.5 & 3.3.6).

By comparison, very little work has been published on short time-scale oxide wearout. The mechanisms of ESD-induced oxide breakdown are still very poorly understood, particularly with regard to the observed polarity and voltage dependencies (Section 3.3.7). Since an improved understanding of these mechanisms may contribute to the design of better protection circuits, the remainder of this thesis is devoted to the topic.

3.10 Summary

1. The possible sources of event-dependent failure have been described, with a particular emphasis on electrostatic discharge (ESD) stress.

2. The various theories of dielectric breakdown have been covered, including avalanche,

thermal and time-dependent breakdown. Several models of time dependent dielectric breakdown (TDDB) were discussed, including the treeing, electron-trapping and the hole-trapping models. Arguments for and against these theories were presented.

3. The distinction between 'intrinsic' and 'defect-related' dielectric breakdown was discussed, and the statistical 'effective thickness' model of extrinsic breakdown was discussed.

4. Hot-carrier instability has been discussed. The coverage included the 'lucky electron' model of electron-gas heating, the avalanche-injection mechanism and the processes of hot-carrier injection in MOSFETs.

5. Other failure mechanisms were briefly described, including CMOS latchup, electromigration, electro-thermomigration and dielectric instability.

6. The dangers of 'latent' or 'walking wounded' failure were described, and practical examples were cited.

7. The various strategies designed to improve reliability were examined. The discussion included anti-ESD precautions, protection circuitry and component screening.

8. The chapter ended with a brief summary of the current state of MOS failure research. Fast transient/ESD oxide breakdown was chosen as the main topic for this thesis.

3.11 References

1. *"The Pulsed EMI Handbook"* (Second Edition), KeyTek Instrument Corp., Wilmington, MASS, 01887, United States, 1991.
2. Amerasekera,E.A., Campbell,D.S., *"Failure Mechanisms in Semiconductor Devices"*, John Wiley & Sons (UK), 1987.
3. Bell,K, Pinkerton,A.S., *"The Fatal Spark"*, IEE Review, 38(2), pp.51-3, 1992.
4. Noel,P.H., *"EOS or ESD: Can Failure Analysis Tell the Difference"*, Proc. 5th. EOS/ESD Symp., pp.154-7, 1983.
5. Franklin,A.J., Dwyer,V.M., *"Electrostatic Discharge Thermal Failure in GaAs Devices"*, Proc. 1st. Components Engineering, Reliability and Test Conference (CERT), pp.71-7, 1990.
6. Amerasekera,E.A., Campbell,D.S., *"An Investigation of the Nature and Mechanisms of ESD Damage in NMOS Devices"*, Sol. St. Elect., 32(3), pp.199-206, 1989.

7. Amerasekera,E.A., Campbell,D.S., *"ESD Pulse and Continuous Voltage Breakdown in MOS Capacitor Structures"*, Proc. 8th. EOS/ESD Symposium, Las Vegas, Nevada, Sept.23-5, 1986. pp.208-13, 1986.
8. Enoch,R.D., *"Dealing with Electrostatic Problems on Future Equipment"*, Proceedings of The Institute of Quality Assurance Electrostatic Damage to Sensitive Devices and its Prevention Conference (2 June 1982), pp.41-54, 1982.
9. Amerasekera,E.A., Campbell,D.S., *"Failure Mechanisms in Semiconductor Devices"*, John Wiley & Sons, U.K., pp.12-20, 1986.
10. Moss,R.Y., *"Caution - Electrostatic Discharge at Work!"*, IEEE Transactions on Components, Hybrids and Manufacturing Technology, CHMT5(4), pp.512-5, 1982.
11. Gould,D, *"The Handling and Packaging of Electrostatic Sensitive Devices"*, Proceedings of The Institute of Qualtiy Assurance Electrostatic Damage to Sensitive Devices and its Prevention Conference (2 June 1982), pp.1-19, 1982.
12. Gilby,D.B., *"Before and After - a practical Antistatic Precautions Campaign"*, Proceedings of The Institute of Qualtiy Assurance Electrostatic Damage to Sensitive Devices and its Prevention Conference (2 June 1982), pp.30-40, 1982.
13. Avery,L.R., *"Electrostatic Discharge (ESD) Tutorial"*, EOS/ESD Symposium Tutorials, pp.4.1-4.5, 1988.
14. Calderbank,J.M., Overbay,D.E., Snyder,H.Z., *"The Effects of High Humidity Environments on Electrostatic Generation and Discharge"*, Proc. 2nd. EOS/ESD Symp., pp.12-16, 1980.
15. Blinde,D, Lavoie,L, *"Quantitative Effects of Relative & Absolute Humidity on ESD Generation/Suppression"*, Proc. 3rd. EOS/ESD Symposium, pp.9-13, 1981.
16. Enoch,R.D., Shaw,R.N., Taylor,R.G., *"ESD Sensitivity of NMOS LSI Circuits and their Failure Characteristics"*, Proc. 5th. EOS/ESD Symp., pp.185-97, 1983.
17. Pancholy,R.K., *"The Effects of Scaling on EOS/ESD Failure Threshold"*, Proc. 3rd. EOS/ESD Symposium, pp.85-9, 1981.
18. McAteer,O, *"ESD: A Decade of Progress"*, Proc. 10th. EOS/ESD Symp., pp.1-6, 1988.
19. Whitehead,A.P., Brydon,G.M., *"Electrostatic Damage in the Development of a High Rel Device"*, Proceedings of The Institute of Qualtiy Assurance Electrostatic Damage to Sensitive Devices and its Prevention Conference (2 June 1982), pp.20-29, 1982.
20. Thompson,W.H., *"ESD Damage, Does it Happen on PCBs"*, Proc. 6th. EOS/ESD Symp., pp.22-4, 1984.
21. MIL-STD-883C Notice 7, Method 3015.6, *"Electrostatic Discharge Sensitivity Classification"*, pp.1-5, 1988.
22. Hyatt,H, Calvin,H, Mellberg,H, *"A Closer Look at the Human ESD Event"*, Proc. 3rd. EOS/ESD Symp., pp.1-8, 1981.
23. *"ESD Testing for ICs"*, KeyTek Instrument Corp., Wilmington, MASS, U.S.A., 01887, 1990.
24. Avery, L.R., *"Electrostatic Discharge (ESD) Tutorial"*, EOS/ESD Symposium Tutorials, pp.4.1-4.5, 1988.
25. Pon,H, Maloney,T, *"High Current ESD Damage to MOS I/O Structures Caused by Charged Video Monitor Surfaces and Castings"*, Proc. 11th. EOS/ESD Symp., pp.78-83, 1989.
26. Fisher,R.J., *"A Severe Human ESD Model for Safety and High Reliability System Qualification Testing"*, Proc. 11th. EOS/ESD Symp., pp.55-8, 1989.

27. Bossard,P.R., Chemelli,R.G., Unger,B.A., *"ESD Damage from Triboelectrically Charged i.c. Pins"*, Proc. 2nd. EOS/ESD Symp., pp.17-22, 1980.
28. Chemelli,R.G, Unger,B.A., Bossard,P.R., *"ESD by Static Induction"*, Proc. 5th. EOS/ESD Symp., pp.29-36, 1983.
29. Unger,B.A., Hart,D.L., *"Triboelectric Characterisation of Packaging Materials"*, Proc. 7th. EOS/ESD Symp., pp.107-10, 1985.
30. Avery,L.R., *"Charged Device Model Testing: Trying to Duplicate Reality"*, Proc 9th. EOS/ESD Symp., pp.88-92, 1987.
31. Maloney,T.J., *"Designing MOS Inputs and Outputs to Avoid Oxide Failure in the Charged Device Model"*, Proc. 10th. EOS/ESD Symp., pp.220-27, 1988.
32. Shaw,R.N., Enoch,R.D., *"An Experimental Investigation of ESD-Induced Damage to Integrated Circuits on Printed Circuit Boards"*, Proc. 7th. EOS/ESD Symp., pp.132-40, 1985.
33. Dangelmayer,G.T., *"A Realistic and Systematic ESD Control Plan"*, Proc. 6th. EOS/ESD Symp., pp.1-6, 1984.
34. Kolyer,J.M., Anderson,W.E., Watson,D.E., *"Hazards of Static Charges and Fields at the Work Station"*, Proc. 6th. EOS/ESD Symp., pp.7-19, 1984.
35. Greason,W.D., Peter Castle,G.S., *"The Effects of Electrostatic Discharge on Microelectronic Devices"*, IEEE Transactions on Industry Applications, IA20(2), pp.247-52, 1984.
36. Boverie,B, *"Coupling of ESD-Generated EMP to Electronics"*, Proc. 10th. EOS/ESD Symp., pp.173-6, 1988.
37. Scatterfield,H.W., Werber,R.F., *"Statistical and Graphical Analysis of Oxide Thickness and ESD Failure Modes"*, Proc. 9th.EOS/ESD Symp., pp.258-64, 1987.
38. Osburn,C.M., Ormund,D.W., *"Dielectric Breakdown in Silicon Dioxide Films, Part II: Influence of Processing and Materials"*, J. Electrochem. Soc., 119(5), pp.597-603, 1972.
39. Sarrabayrouse,G, Prom,J.L., Kassmi,K, *"Oxide Breakdown in a Metal-SiO_2-Si Capacitor: Influence of the Metal Electrode"*, IEE Proc., 137(G6), pp.475-8, 1990.
40. Wang,S.T., Harari,E, Nielsen,W.Y., *"Improved Reliability and Yield in Thin Thermal SiO_2"*, Proc. 16th. IEEE Reliability Physics Symp. (IRPS), pp.137-9, 1978.
41. Marcus,R.B., Sheng,T.T., Lin,P, *"Polysilicon/SiO_2 Interface Microtexture and Dielectric Breakdown"*, J. Electrochem. Soc., 129(6), pp.1282-89, 1982.
42. Klein,N, Nevanlinna,O, *"Lowering of the Breakdown Voltage of Silicon Dioxide by Asperities and at Spherical Electrodes"*, Sol. St. Elect., 26(9), pp.883-92, 1983.
43. Lassig,S, Liang,M.-S., *"Time-Dependent Degradation of Thin Gate Oxide Under Post-Oxidation High-Temperature Anneal"*, IEEE Electron Dev. Lett., EDL8(4), pp.160-1, 1987.
44. Bridgwood,M.A., *"Breakdown Mechanisms in MOS Capacitors following Electrical Overstress"*, Proc. 8th. EOS/ESD Symp., pp.200-7, 1985.
45. Klein,N, Gafni,H, *"The Maximum Dielectric Strength of Thin Silicon Oxide Films"*, IEEE Trans. Electron Dev., ED12(12), pp.281-9, 1966.
46. Klein,N, *"Switching and Breakdown in Films"*, Thin Solid Films, 7, pp.149-77, 1971.
47. Hansel,G, Villareal,E, *"Cost Effective Failure Analysis Method for Detecting Failure Site associated with Extremely Small Leakage"*, Proc. 9th. EOS/ESD Symp., pp.71-7, 1987.

48. Taylor,R.G., Woodhouse,J, *"Junction Degradation and Dielectric Shorting: Two Mechanisms for ESD Recovery"*, Proc. 8th. EOS/ESD Symp., pp.92-9, 1986.
49. Amerasekera,E.A., Campbell,D.S., *"Charge Limited Breakdown in MOS Capacitors"*, Proc. 17th. Annual Solid State Device Research Conference, pp.733-41, 1987.
50. Neri,B, Olivo,P, Riccò, *"Low-Frequency Noise in Silicon-Gate Metal-Oxide-Silicon Capacitors Before Oxide Breakdown"*, Appl. Phys. Lett., 51(25), pp.2167-9, 1987.
51. Suñé,J, Farrés,E, Placencia,I, Barniol,N, Martín,F, Aymerich,X, *"Nondestructive Multiple Breakdown Events in Very Thin SiO_2 Films"*, Appl. Phys. Lett., 55(2), pp.128-30, 1989.
52. Saletti,R, Neri,B, Olivo,P, Modelli,A, *"Correlated Fluctuations and Noise Spectra of Tunneling and Substrate Currents Before Breakdown in Thin-Oxide MOS Devices"*, IEEE Trans. Electron Dev., ED37(11), pp.2411-3, 1990.
53. Davis,J.R., *"Instabilities in MOS Devices"*, Electrocomponent Science Monographs Vol.1, Gordon & Breach, New York, pp.133-49, 1981.
54. Seitz,F, *"On the Theory of Electron Multiplication in Crystals"*, Phys. Rev., 76(9), pp.1376-93, 1949.
55. Whitehead,S, *"Dielectric Breakdown of Solids"*, Clarendon Press, Oxford, pp.12-23, 1951.
56. Anderson,J.C., *"Dielectrics"*, Chapman & Hall Ltd., London, pp.98-120, 1964.
57. Hamano,K, *"Breakdown Characteristics in Thin SiO_2 Film"*, Jap. J. Appl. Phys., 13(7), pp.1085-92, 1974.
58. Wolters,D.R., van der Schoot,J.J., *"Dielectric Breakdown in MOS Devices"*, Phillips Journal of Research, 40(3), pp. 115-92, 1985.
59. Meek,M.J., Craggs,J.D., *"Electrical Breakdown of Gases"*, Clarendon Press, Oxford, 1953.
60. von Hippel,A, *"Electric Breakdown in Solid and Liquid Insulators"*, J.Appl.Phys., 8, pp.815-32, 1937.
61. Fröhlich,H, *"Theory of Electrical Breakdown in Ionic Crystals"*, Proc. Roy. Soc., A160, pp.230-41, 1930.
62. O'Dwyer,J.J., *"The Theory of Dielectric Breakdown of Solids"*, Clarendon Press, Oxford, 1964.
63. Callen,H.B., *"Electric Breakdown in Ionic Crystals"*, Phys. Rev., 79(9), pp.1394-1402, 1949.
64. Fröhlich,H, *"On the Theory of Dielectric Breakdown in Solids"*, Proc. Roy. Soc., A188, pp.521-32, 1947.
65. Fröhlich,H, *"Energy Distribution and Stability of Electrons in Electric Field"*, Proc. Roy. Soc., A188, pp.532-41, 1947.
66. Fischetti,M.V., *"High Field Electron Transport in SiO_2 and Generation of Positive Charge at the Si-SiO_2 Interface"*, Insulating Films on Semiconductors, pp.181-9, 1986.
67. DiMaria,D.J., Fischetti,M.V., *"Hot Electrons in Silicon Dioxide: Ballistic to Steady State Transport"*, Appl. Surf. Sci., 30, pp.278-97, 1987.
68. Ridley,B.K., *"Mechanism of Electrical Breakdown in SiO_2 Films"*, J.Appl.Phys., 46(3), pp.998-1007, 1975.
69. Ferry,D.K., *"Electron Transport and Breakdown in SiO_2"*, J.Appl.Phys., 50(3), pp.1422-7, 1979.

70. Lenzlinger,M, Snow,E.H., "*Fowler-Nordheim Tunneling into Thermally Grown SiO_2*", J. Appl. Phys., 40(1), pp.278-83, 1969.
71. DiStefano,T.H., Shatzkes,M, "*Dielectric Instability and Breakdown in Wide Bandgap Insulators*", J. Vac. Sci. Technol., 12(1), pp.37-46, 1975.
72. DiStefano,T.H., Shatzkes,M, "*Impact Ionisation Model for Dielectric Instability and Breakdown*", Appl. Phys. Lett., 25(12), pp.685-7, 1974.
73. Klein,N, Solomon,P, "*Current Runaway in Insulators Affected by Impact Ionisation and Recombination*", J. Appl. Phys., 47(10), pp.4364-72, 1976.
74. O'Dwyer,J.J., "*Theory of Dielectric Breakdown in Solids*", J.Electrochem.Soc., 116(2), pp.239-42, 1969.
75. O'Dwyer,J.J., "*Theory of High Field Conduction in a Dielectric*", J.Appl.Phys., 40(10), pp.3887-90, 1969.
76. Agarwal,V.K., Srivastava,V.K., "*Thickness Dependence of Breakdown Field in Thin Films*", Thin Solid Films, 8, pp.377-81, 1971.
77. Osburn,C.M., Weitzman,E.J., "*Electrical Conduction and Dielectric Breakdown in Silicon Dioxide Films on Silicon*", J. Electrochem. Soc., 119(5), pp.603-9, 1972.
78. Klein,N, "*Electrical Breakdown in Thin Dielectric Films*", J.Electrochem.Soc., 116(7), pp.963-72, 1969.
79. Shousha,A.H.M., "*Filamentary Thermal Breakdown in Thin Dielectrics*", J. Appl. Phys., 43(1), pp.15-18, 1972.
80. Hikita,M, Hirose,T, Ito,Y, Mizutani,T, Ieda,M, "*Investigation of Electrical Breakdown of Polymeric Insulating Materials Using a Technique of Pre-Breakdown Current Measurements*", J. Phys. D: Appl. Phys., 23, pp.1515-27, 1990.
81. Tatsuuma,K, Sugimoto,M, Ajiki,T, "*The New Concept of Thin Oxide Breakdown Mechanism*", 1988 Symposium on VLSI, Digest of Technical Papers, pp.43-4, 1988.
82. Whitehead,S, "*Dielectric Breakdown of Solids*", Clarendon Press, Oxford, pp.236-44, 1951.
83. Osburn,C.M., Raider,S.I., "*The Effect of Mobile Sodium Ions on Field Enhancement Dielectric Breakdown in SiO_2 Films on Silicon*", J. Electrochem. Soc., 120(10), pp.1369-76, 1973.
84. Olivo,P, Ricco,B, Nguyen,T.N., Kuan,T.S., Jeng,S.J., "*Evidence of the Role of Defects Near the Injecting Interface in Determining SiO_2 Breakdown*", Appl. Phys. Lett., 51(26), pp.2245-7, 1987.
85. Osburn,C.M., Weitzman,E.J., "*Electrical Conduction and Dielectric Breakdown in Silicon Dioxide Films on Silicon*", J.Electrochem.Soc., 119, pp.603-9, 1972.
86. Raider,S.I., "*Time-Dependent Breakdown of Silicon Dioxide Films*", Appl.Phys.Lett., 23(1), pp.34-6, 1973.
87. Hill,R.M., Dissado,L.A., "*Theoretical Basis for the Statistics of Dielectric Breakdown*", J. Phys. C: Sol. State. Phys., 16, pp.2145-56, 1983.
88. Anolick ,E.S., Nelson,G.R., "*Low Field Time Dependent Dielectric Integrity*", Proc. **17th.** IEEE Reliability Physics Symp., pp.8-12, 1979.
89. Chen,I-C, Holland,S.E., Hu,C, "*Electrical Breakdown in Thin Gate and Tunneling Oxides*", IEEE Trans. Electron Dev., ED32(2), pp.413-22, 1985.
90. Rosenbaum,E, Lee,P.M., Moazzami,R, Ko,P.K., Hu,C, "*Circuit Reliability Simulator - Oxide Breakdown Module*", 1989 International Electronic Device Meeting (IEDM) Technical Digest, pp.331-4, 1989.
91. Hokari,Y, "*Stress Voltage Polarity Dependence of Thermally Grown Thin Gate Oxide*

Wearout", IEEE Trans. Electron Dev., ED35(8), pp.1299-304, 1988.

92. Hokari,Y, "*Dielectric Breakdown Wearout Limitation of Thermally-Grown Thin-Gate Oxides*", Sol. St. Elect., 33(1), pp.75-8, 1990.
93. McPherson,J.W., Baglee,D.A., "*Acceleration Factors for Thin Gate Oxide Stressing*", Proc. 23th. IEEE Reliability Physics Symp., pp.1-6, 1985.
94. Moazzami,R, Lee,J.C., Hu,C, "*Temperature Acceleration of Time-Dependent Dielectric Breakdown*", IEEE Trans. Electron Dev., ED36(11), pp.2462-5, 1989.
95. Jonscher,A.K., "*On a Cumulative Model for Dielectric Breakdown in Solids*", IEEE Trans. Electrical Insulation, EI19(6), pp.567-77, 1984.
96. Niemeyer,L, et al, "*Fractal Dimension of Dielectric Breakdown*", Phys. Rev. Lett., 52(12), pp.1033-6, 1984.
97. Willming,D.A., Wu,C.H., "*A Stochastic Model for Dielectric Breakdown in Thin Capacitors*", J. Appl. Phys., 63(2), pp.456-9, 1988.
98. Barclay,A.L., Sweeney,P.J., Dissado,L.A., Stevens,G.C., "*Stochastic Modelling of Electrical Treeing: Fractal and Statistical Characteristics*", J. Phys. D: Appl. Phys., 23, pp.1536-45, 1990.
99. Forster,E.O., "*Progress in the Understanding of Electrical Breakdown in Condensed Matter*", J. Phys. D: Appl. Phys., 23, pp.1506-14, 1990.
100. Budenstein,P.B., "*On the Mechanism of Dielectric Breakdown of Solids*", IEEE Trans. Electrical Insulation, IE15(3), pp.225-40, 1980.
101. Avni,E, Shappir,J, "*A Model for Silicon-Oxide Breakdown under High Field and Current Stress*", J.Appl. Phys., 64(2), pp.743-8, 1988.
102. Placencia,I, Suñé,J, Barniol,N, Farrés,E, Aymerich,X, "*Oxide Degradation and Breakdown in Stressed MOS Capacitors*", Journal de Physique, Colloque C4, Supplément au n°9, Tome 49, 1988.
103. Harari,E, "*Dielectric Breakdown in Electrically Stressed Thin Films of Thermal SiO_2*", J.Appl.Phys., 49(4), pp.2478-89, 1978.
104. Meyer,W.K., Crook,D.L., "*Model for Oxide Wearout due to Charge Trapping*", Proc. 21st. IEEE Reliability Physics Symp., pp.242-7, 1983.
105. Ricco,B, et al, "*Novel Mechanism for Tunnelling and Breakdown of Thin SiO_2 Films*", Phys. Rev. Lett., 51(9), pp.1795-8, 1983.
106. Holland,S, Chen,I-C, Hu,C, "*On Physical Models for Gate Oxide Breakdown*", IEEE Electron Dev. Lett., EDL5(8), pp.302-5, 1984.
107. Dumin,D.J., Dickerson,K.J., Hall,M.D., "*Polarity Dependence of Thin Oxide Wearout*", Proc. 27th. IEEE Reliability Physics Symp., pp.28-33, 1989.
108. Chen,I-C, Holland,S, Hu,C, "*Hole Trapping and Breakdown in Thin SiO_2*", IEEE Electron Dev. Lett., EDL7(3), pp.164-7, 1986.
109. Jones,R.E., Abram,R.A., "*Hot Hole Creation due to Impact Ionisation in the Metal Electrode of a Forward Biased MIS Structure*", Sol. St. Elect., 31(5), pp.989-97, 1988.
110. Holland,S, Chen,I-C, Hu,C, "*Ultra-Thin Silicon-Dioxide Breakdown Characteristics of MOS Devices with n^+ and p^+ Polysilicon Gates*", IEEE Electron Dev. Lett., EDL8(12), pp.572-5, 1987.
111. Chen,I-C, Holland,S, Young,K.K., Chang,C, Hu,C, "*Substrate Hole Current and Oxide Breakdown*", Appl. Phys. Lett., 49(11), pp.669-71, 1986.
112. DuToit,M, Fazan,P, Benjeloun,A, Ilegems,M, "*Charge Trapping and Breakdown in Thin SiO_2 Polysilicon-Gate MOS Capacitors*", Appl. Surf. Sci., 30, pp.333-8, 1987.

113. Haddad,S, Liang,M-S, *"The Nature of Charge Trapping Responsible for Thin-Oxide Breakdown Under a Dynamic Field Stress"*, IEEE Electron Dev. Lett., EDL8(11), pp.524-7, 1987.
114. Kusaka,T, Ohji,Y, Mukai,K, *"Time-Dependent Dielectric Breakdown of Ultra-Thin Silicon Oxide"*, IEEE Electron Dev. Lett., EDL8(2), pp.61-3, 1987.
115. Tzou,J.J., Yao,C.C., Cheung,R, Chan,H, *"Temperature Dependence of Charge Generation and Breakdown in SiO_2"*, IEEE Electron Dev. Lett., EDL7(7), pp.446-8, 1986.
116. Nissan-Cohen, Y, et al, *"High Field Current Induced Positive-Charge Transients in SiO_2"*, J.Appl.Phys., 54(10), pp.5793-800, 1983.
117. Toriumi,A, Iwase,M, *"Substrate Hole Current Generation due to Electron Tunneling in Metal-Oxide-Semiconductor System"*, Ext. Abst. 19th. Conf. on Sol. State Dev. and Mat., pp.351-4, 1987.
118. Hokari,Y, Baba,T, Kawamura,N, *"Reliability of 6-10nm Thermal SiO_2 Films Showing Intrinsic Dielectric Integrity"*, IEEE Trans. Electron Dev., ED32(11), pp.2485-91, 1985.
119. Li,S.P., Prussin,S, Maserjian,J, *"Model for MOS Field-Time-Dependent Breakdown"*, Proc. 16th. IEEE Reliability Physics Symp. (IRPS), pp.132-6, 1978.
120. Ozawa,Y, Iwase,M, Toriumi,A, *"Interface State Generation due to Electron Tunneling into Thin Oxides"*, Proc. 27th. IEEE Reliability Physics Symp., pp.22-7, 1989.
121. Kerber,M, Schwalke,U, *"Interface Degradation and Dielectric Breakdown of Thin Oxides due to Homogeneous Charge Injection"*, Proc. 27th. IEEE Reliability Physics Symp., pp.17-21, 1989.
122. Osburn,C.M., Ormund,D.W., *"Dielectric Breakdown in Silicon Dioxide Films, Part I: Measurement and Interpretation"*, J. Electrochem. Soc., 119(5), pp.591-7, 1972.
123. Li,S.P., Bates,E.T., Maserjian,J, *"Time Dependent MOS Breakdown"*, Sol. State Elect., 19, pp.235-9, 1976.
124. Osburn,C.M., Ormond,D.W., *"Sodium-Induced Barrier-Height Lowering and Dielectric Breakdown in SiO_2 Films on Silicon"*, J. Electrochem. Soc., 121(9), pp.1195-8, 1974.
125. Hofmann,K, Rubloff,G.W., Young,D.R., *"Role of Oxygen in Defect-Related Breakdown in Thin SiO_2 Films on Si(100)"*, J. Appl. Phys., 61(9), pp.4584-8, 1987.
126. Falster,R, *"The Phenomenology of Dielectric Breakdown in Thin Silicon Dioxide Films: Al Cathodes and p-Type Si Anodes"*, J. Appl. Phys., 66(7), pp.3355-70, 1989.
127. Monkowski,J.R., Zahour,R.T., *"Failure Mechanism in MOS Gates Resulting from Particulate Contaminatiom"*, Proc. 20th. IEEE Reliability Physics Symp., pp.244-8, 1982.
128. Dallmann,A, *"Localization of Defects in Gate Oxides by means of Tunneling Current Microscopy"*, Proc. 24th. IEEE Reliability Physics Symp., pp.95-8, 1986.
129. Lee,J, Chen,I-C, Hu,C, *"Statistical Modeling of Silicon Dioxide Reliability"*, Proc. 26th. IEEE Reliability Physics Symp., pp.131-8, 1988.
130. Lee,J.C., Chen,I-C, Hu,C, *"Modelling and Characterisation of Gate Oxide Reliability"*, IEEE Trans. Electron Dev., ED35(12), pp.2268-78, 1988.
131. Coleman,D.J., Hunter,W.R., Brown,G.A., Chen,I-C, *"Extensions of Effective Thickness Theory of Oxide Breakdown"*, Proc. 27th. IEEE Reliability Physics Symp., pp.39-42, 1989.

132. Fong,Y, Hu,C, *"The Effects of High Electric Field Transients on Thin Gate Oxide MOSFETs"*, Proc. 9th. EOS/ESD Symp., pp.252-7, 1987.
133. Amerasekera,E.A., Campbell,D.S., *"Oxide Breakdown in MOS Structures under ESD and Continuous Voltage Stress Breakdown"*, Proc. 1986 European Conf. on Rel., Copenhagen, pp.325-30, 1986.
134. Amerasekera,E.A., Campbell,D.S., *"Electrostatic Pulse Breakdown in NMOS Devices"*, Qual. & Rel. Eng. Int., 2, pp.107-116, 1986.
135. Mistry,K.R., Krakauer,D.B., Doyle,B.S., *"Impact of Snapback-Induced Hole Injection on Gate Oxide Reliability of N-MOSFET's"*, IEEE Electron Dev. Lett., EDL11(10), pp.460-2, 1990.
136. Aur,S, Chatterjee,A, Polgreen,T, *"Hot Electron Reliability and ESD Latent Damage"*, Proc. 26th. IEEE Reliability Physics Symp., pp.15-18, 1988.
137. Henrickson,L, Peng,Z, Frey,J, Goldsman,N, *"Enhanced Reliability in Si MOSFETs with Channel Lengths under 0.2 Micron"*, Sol. State Electronics, 33(10), pp.1275-8, 1990.
138. Hu,C, Tam,S.C., Hsu,F-C, Ko,P-K, Chan,T-Y, Terrill,K.W., *"Hot-Electron-Induced MOSFET Degradation - Model, Monitor, and Improvement"*, IEEE Trans. Electron Dev., 32(2), pp.375-85, 1985.
139. Pimbley,J.M., Gildenblat,G, *"Effect of Hot-Electron Stress on Low Frequency MOSFET Noise"*, IEEE Electron Dev. Lett., EDL5(9), pp.345-7, 1984.
140. Chen,I.C., Choi,J.J., Chan,T.Y., Ong,T.C., Hu,C, *"The Effect of Channel Hot Carrier Stressing on Gate Oxide Integrity in MOSFET"*, Proc. 26th. IEEE Reliability Physics Symp., pp.1-7, 1988.
141. Davis,M, Lahri,R, *"Gate Oxide Charge-to-Breakdown Correlation to MOSFET Hot-Electron Degradation"*, IEEE Electron Dev. Lett., EDL9(4), pp.183-5, 1988.
142. Chen,I-C, Choi,J.Y., Chan,T-Y, Hu,C, *"The Effect of Channel Hot-Carrier Stressing on Gate-Oxide Integrity in MOSFET's"*, IEEE Electron Dev. Lett., EDL35(12), pp.2253-7, 1988.
143. Garrigues,M, Balland,B, *"Hot Carrier Injection into SiO_2"*, Chapter 8 (pp.403-40) in *"Instabilities in Silicon Devices"*, eds. G.Barbottin & A.Vapaille, North-Holland (Amsterdam), 1986.
144. Goldsman,N, Henrickson,L, Frey,J, *"Reconciliation of a Hot-Electron Distribution Function with the Lucky Electron-Exponential Model in Silicon"*, IEEE Electron Device Letters, 11(10), pp.472-4, 1990.
145. Nicollian,E.A., Goetzberger,A, *"Avalanche Injection Currents and Trapping Phenomena in Thermal SiO_2"*, IEEE Trans. Electron Dev., p.686, 1968.
146. Verwey,J.F., de Maagt,B.J., *"Avalanche-Injected Electron Currents in SiO_2 at High Injection Densities"*, Sol. State Electron., 17, pp.963-71, 1974.
147. Hsu,C.C.-H., Wen,D.S., Wordeman,M.R., Taur,Y, Ning,T.H., *"A Comparative Study of Hot-Carrier Instabilities in P- and N-Type Poly Gate MOSFETs"*, IEDM Technical Digest, pp.75-8, 1989.
148. Fu,K-Y, Teng,K-W, *"Comparison of Channel Current and Drain Avalanche Current-Induced Hot-Carrier Effects in MOSFET's"*, IEEE Electron Dev. Lett., EDL8(4), pp.132-4, 1987.
149. Davis,J.R., *"Instabilities in MOS Devices"*, Electrocomponent Science Monographs Vol.1, Gordon & Breach, New York, pp.109-32, 1981.
150. Hsu,F-C, Tam,S, *"Relationship Between MOSFET Degradation and Hot-Electron-*

Induced Interface-State Generation", IEEE Electron Dev. Lett., EDL5(2), pp.50-2, 1984.

151. Hsu,F.-C., Chiu,K.-Y., "*Temperature Dependence of Hot-Electron-Induced Degradation in MOSFET's*", IEEE Electron Dev. Lett., EDL5(5), pp.148-50, 1984.
152. Heremans,P, Van der bosch,G, Bellens,R, Groeseneken,G, Maes,H.E., "*Understanding of the Temperature Dependence of Channel Hot-Carrier Degradation in the Range 77K to 300K*", IEDM Technical Digest, pp.67-70, 1989.
153. MacWilliams,K.P., Halle,L.F., Zeitlow,T.C., "*Improved Hot-Carrier Resistance with Fluorinated Gate Oxides*", IEEE Electron Dev. Lett., EDL11(1), pp.3-5, 1990.
154. Nishioka,Y, Da Silva,E.F., Wang,Y, Ma,T-P, "*Dramatic Improvement of Hot-Electron-Induced Interface Degradation in MOS Structures Containing F or Cl in SiO_2*", IEEE Electron Dev. Lett., 9(1), pp.38-40, 1988.
155. Doyle,B.S., Bourcerie,M, Marchetaux,J.-C., Boudou,A, "*Relaxation Effects in NMOS Transistors After Hot-Carrier Stressing*", IEEE Electron Dev. Lett., EDL8(5), pp.234-6, 1987.
156. Aur,S, Chatterjee,A, Polgreen,T, "*Hot-Electron Reliability and ESD Latent Damage*", IEEE Trans. Electron Dev., ED35(12), pp.2189-93, 1988.
157. Bellens,R, Heremans,P, Groeseneken,G, Maes,H.E., "*A New Procedure for Lifetime Prediction of n-Channel MOS-Transistors using the Charge Pumping Technique*", Proc. 26th. IEEE Reliability Physics Symp., pp.8-14, 1988.
158. Khurana,N, Chiang,C-L, "*Analysis of Product Hot Electron Problems by Gated Emission Microscopy*", 24th. IEEE Reliability Physics Symp., pp.189-94, 1986.
159. Ng,K.K., Pai,C-S, Mansfield,W.M., Clarke,G.A., "*Suppression of Hot-Carrier Degradation in Si MOSFET's by Germanium Doping*", IEEE Electron Dev. Lett., EDL11(1), pp.45-7, 1990.
160. Bauer,F, Jain,S.C., Lauer,V, Offenberg,M, Balk,P, "*Incompatibility of Requirements for Optimizing Short Channel Behaviour and Long Term Stability in MOSFET's*", Sol. St. Elect., 31(1), pp.27-33, 1988.
161. Chen,K-L, "*The Effects of Interconnect Process and Snapback Voltage on the ESD Failure Threshold of NMOS Transistors*", IEEE Trans. Electron Dev., ED35(12), pp.2140-50, 1988.
162. Ochoa,A, Dawes,W, Estreich,D, "*Latch-Up Control in CMOS Integrated Circuits*", IEEE Trans. Nuc. Sci., NS26, pp.5065-8, 1979.
163. Lim,S-C, Tan,E-G, "*Detection of Junction Spiking and its Induced Latch-Up by Emission Microscopy*", Proc. 26th. IEEE Reliability Physics Symp., pp.119-125, 1988.
164. Khurana,N, "*Pulsed Infra-Red Microscopy for Debugging Latchup on CMOS Products*", Proc. 22nd. IEEE Reliability Physics Symp. (IRPS), pp.122-7, 1984.
165. Dooley,J.G., Jaeger,R.C., "*Temperature Dependence of Latchup in CMOS Circuits*", IEEE Electron Dev. Lett., EDL5(2), pp.41-3, 1984.
166. Dawes,W.R., Derbenwick,G.F., "*Prevention of CMOS Latch-Up by Gold Doping*", IEEE Trans. Nuc. Sci., NS23(6), pp.2027-30, 1976.
167. Sangiorgi,E, Swirhun,S, "*Trenched Schottky Barrier PMOS for Latchup Resistance*", IEEE Electron Dev. Lett., EDL5(8), pp.293-5, 1984.
168. Troutman,R.R., "*Latchup in CMOS Technology: The Problem and Its Cure*", Kluwer Academic Publishers, Boston, 1986.
169. Maloney,T.J., "*Contact Injection: A Major Cause of ESD Failure in Integrated*

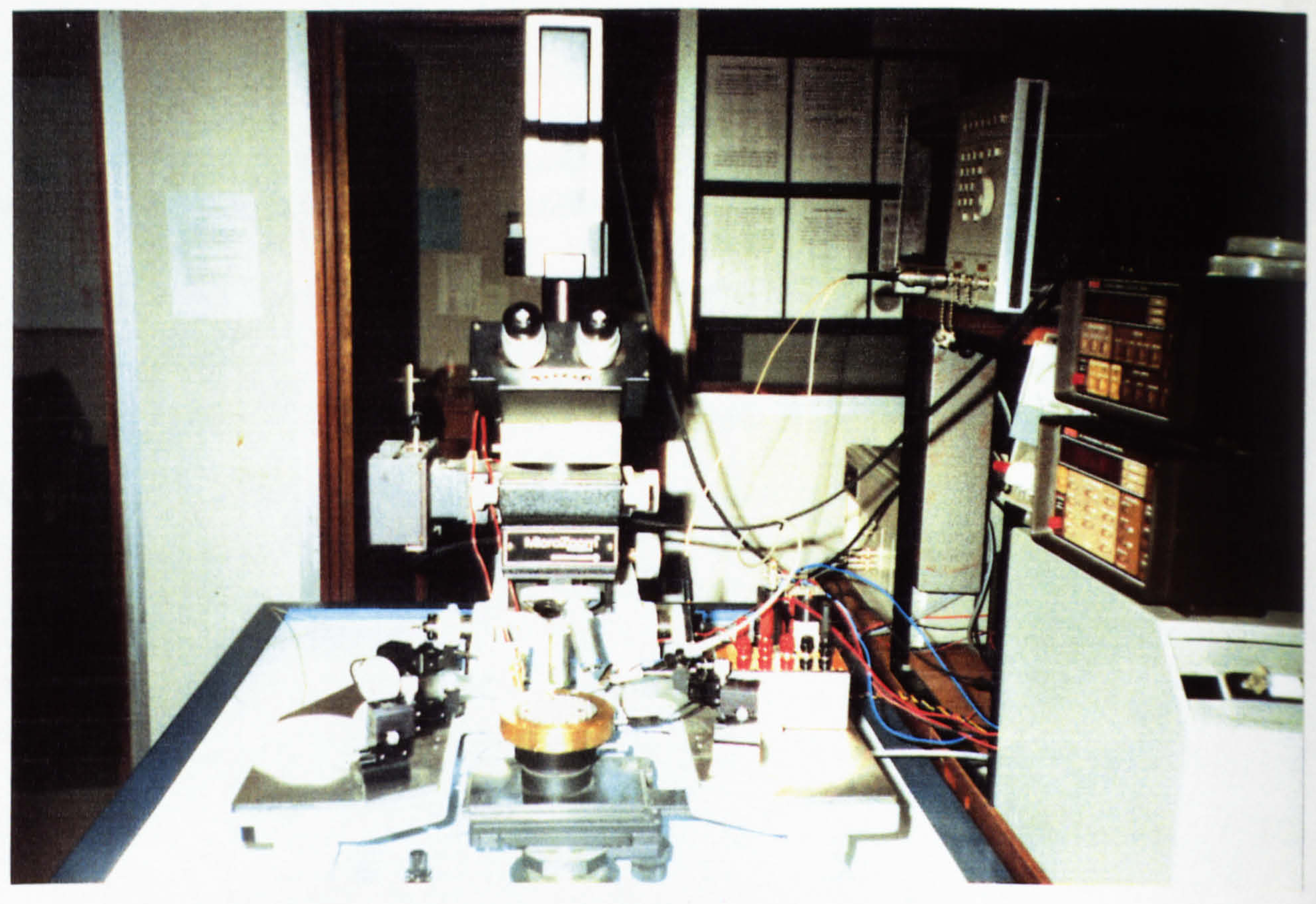
Circuits", Proc. 8th. EOS/ESD Symp., pp.166-72, 1986.
170. Pierce,D.G., "*Electro-Thermomigration as an Electrical Overstress Failure Mechanism*", Proc. 7th. EOS/ESD Symp., pp.67-76, 1985.
171. Breed,D.J., "*Non-Ionic Room Temperature Instabilities in MOS Devices*", Sol. State Elect., 17, pp.1229-43, 1974.
172. Kriegler,R.J., "*Ion Instabilities in MOS Structures*", Proc. 12th. IEEE Reliability Physics Symp. (IRPS), pp.250-8, 1974.
173. Eisele,I, Vitanov,P, Fischer,F, Schwabe,U, "*Turn-on Voltage Degradation of Short-Channel MOSFETs due to Generation of Interface States*", Proc. 22nd. IEEE Reliability Physics Symp., pp.210-13, 1984.
174. Sivo,L.L., Peden,J.C., Brettschneider,M, Price,W, Pentecost,P, "*Cosmic Ray-Induced Soft Errors in Static MOS Memory Cells*", IEEE Trans. Nuc. Sci., NS26(6), pp.5042-7, 1979.
175. Adams,J.R., Sokel,R.J., "*Neutron Irradiation for Prevention of Latch-Up in MOS Integrated Circuits*", IEEE Trans. Electron Dev., NS26(6), pp.5069-73, 1979.
176. Warren,W.L., Lenahan,P.M., "*A Comparison of Positive Charge Generation in High Field Stressing and Ionising Radiation on MOS Structures*", IEEE Trans. Nuc. Sci., NS34(6), pp.1355-8, 1987.
177. Mott,D.B., Buchner,S.P., "*Comparison of Interface Positive Charge Generated in Metal-Oxide-Silicon Devices by High-Field Electron Injection and X-Ray Irradiation*", J. Appl. Phys., 51(20), pp.1643-5, 1987.
178. Kassabov,J, Dimitrov,D, Grueva,A, "*Acceptor States at the Si-SiO$_2$ Interface Generated by UV and their Effect on Electron Mobility*", Sol. State Electron., 31(1), pp.49-51, 1988.
179. Schwank,J.R., Fleetwood,D.M., Winokur,P.S., Dressendorfer,P.V., Turpin,D.C., Sanders,D.T., "*The Role of Hydrogen in Radiation-Induced Defect Formation in Polysilicon Gate MOS Devices*", IEEE Trans. Nuc. Sci., NS34(6), pp.1152-8, 1987.
180. McWhorter,P.J., Winokur,P.S., Pastorek,R.A., "*Donor/Acceptor Nature of Radiation-Induced Interface Traps*", IEEE Trans. Nuc. Sci., NS35(6), pp.1154-9, 1988.
181. Sanders,T.J., "*Silicon Gate CMOS on Bulk Silicon for Ionizing Radiation Environments*", IEE Trans. Nuc. Sci., NS26(6), pp.5056-7, 1979.
182. da Silva,E.F., Nishioka,Y, Ma,T.-P., "*Radiation Response of MOS Capacitors Containing Fluorinated Oxides*", IEEE Trans. Nuc. Sci., NS34(6), pp.1190-5, 1987.
183. Boesch,H.E., McGarrity,J.M., "*An Electrical Technique to Measure the Radiation Susceptibility of MOS Gate Insulators*", IEEE Trans. Nuc. Sci., NS26(6), pp.4814-8, 1979.
184. Berbeco,G.R., "*Passive Static Protection: Theory and Practice*", Proc. 2nd. EOS/ESD Symposium, pp.1-11, 1980.
185. McAteer,O.J., Twist,R.E., Walker,R.C., "*Identification of Latent ESD Failures*", Proc. 2nd. EOS/ESD Symposium, pp.54-7, 1980.
186. McAteer,O.J., Twist,R.E., Walker,R.C., "*Latent ESD Failures*", Proc. 4th. EOS/ESD Symp., pp.41-8, 1982.
187. Stephens,C.E., Amos,C.T., "*A Study of EOS in Microcircuits Using the Infra-Red Microscope*", Proc. 8th. EOS/ESD Symp., pp.219-23, 1986.
188. deChiaro,L.F., "*Latency in ESD Damage of Semiconductor Devices*", Proc. 9th. EOS/ESD Symp., p.291, 1987.

189. Holmes,G.C., *"An Investigation into the Effects of Low Voltage ESD Transients on MOSFETs"*, Proc. 7th EOS/ESD Symp., pp.170-4, 1985.
190. Franklin,A.J., Amerasekera,E.A., Campbell,D.S., *"A Comparison Between GaAs MESFET and Si NMOS ESD Behaviour"*, Proc. 1986 ERA Seminar on ESD in Electronics, pp.4.3.1-14, London, 1986.
191. Neelakantaswamy,P.S., Sarkar,T.K., Turkman,I.R., *"Residual Fatigues in Microelectronic Devices due to Thermoelastic Strains caused by Repetitive Electrical Overstressings: A Model for Latent Failures"*, Proc. 7th. EOS/ESD Symp., pp.77-83, 1985.
192. McCullough,D.T., Lane,C.H., Blore,R.A., *"Reliability of EOS Screened Gold Doped 4002 CMOS Devices"*, Proc. 1st. EOS/ESD Symp., pp.41-44, 1979.
193. Schwank,J.R., Baker,R.P., Armendariz,M.G., *"Suprising Patterns of CMOS Susceptibility to ESD and Implications on Long-Term Reliability"*, Proc. 2nd. EOS/ESD Symp., pp.104-11, 1090.
194. Woodhouse,J, Lowe,K.D., *"ESD Latency: A Failure Analysis Investigation"*, Proc. 10th. EOS/ESD Symp., pp.47-52, 1988.
195. Bowers,J.S., Rossi,M.G., Beall,J.R., *"A Study of Latent Defects in Semiconductors"*, Proc. 5th. EOS/ESD Symp., pp.198-204, 1983.
196. Roberts,B.C., *"Determination of Threshold Energies and Damage Mechanisms in Semiconductor Devices Subjected to Voltage Transients"*, Proc. 6th. EOS/ESD Symp., pp.157-64, 1984.
197. Strauss,M.S., Lin,D.L., Welsher,T.L., *"Variations in Failure Modes and Cumulative Effects Produced by Commercial Human-Body-Model Simulators"*, Proc. 9th. EOS/ESD Proc., pp.56-63, 1987.
198. Crockett,R.G.M., Smith,J.C., Hughes,J.F., *"ESD Sensitivity and Latency Effects of some HCMOS Integrated Circuits"*, Proc. 6th. EOS/ESD Symp., pp.196-201, 1984.
199. Krakauer,D.B., Mistry,K.R., *"On Latency and the Physical Mechanisms Underlying Gate Oxide Damage during ESD Events in n-Channel MOSFETs"*, Proc. 11th. EOS/ESD Symp., pp.121-6, 1989.
200. Hellström,S, Welander,A, Eklöf,P, *"Studies and Revelation of Latent ESD-Failures"*, Proc. 8th. EOS/ESD Symp., pp.81-91, 1986.
201. Bunyan,R.J.T., Uren,M.J., Thomas,N.J., Davis,J.R., *"Degradation in Thin-Film SOI MOSFET's Caused by Single-Transistor Latch"*, IEEE Electron Dev. Lett., 11(9), pp.359-61, 1990.
202. Spialter,E.B., Brandewie,J.V., Kjar,R.A., *"An Ion Inplant Induced Instability Mechanism in CMOS/SOS Device"*, Proc. 22nd. IEEE Reliability Physics Symp., pp.161-4, 1984.
203. Matloubian,M, Zorinsky,E.J., Spratt,D.B., *"Total Dose Radiation Characteristics of SOI MOSFETs Fabricated Using Islands Technology"*, IEEE Trans. Nuc. Sci., NS35(6), pp.1650-2, 1988.
204. Rosenberg,S, Crook,D, Euzent,B, *"H-MOS Reliability"*, Proc. 16th. IEEE Reliability Physics Symp. (IRPS), pp.19-22, 1978.
205. McAteer,O.J., Twist,R.E., *"Analysis of Electrostatic Discharge Failures"*, Proc. 3rd. EOS/ESD Symp., pp.14-20, 1981.
206. Bhatti,I.S., Fuller,E, Jenne,F.B., *"VMOS Electrostatic Protection"*, Proc. 16th. IEEE Reliability Physics Symp. (IRPS), pp.140-5, 1978.
207. Edwards,J.R., Fuller,E, Allman,G, Malwah,M.L., *"VMOS Reliability"*, Proc. 16th.

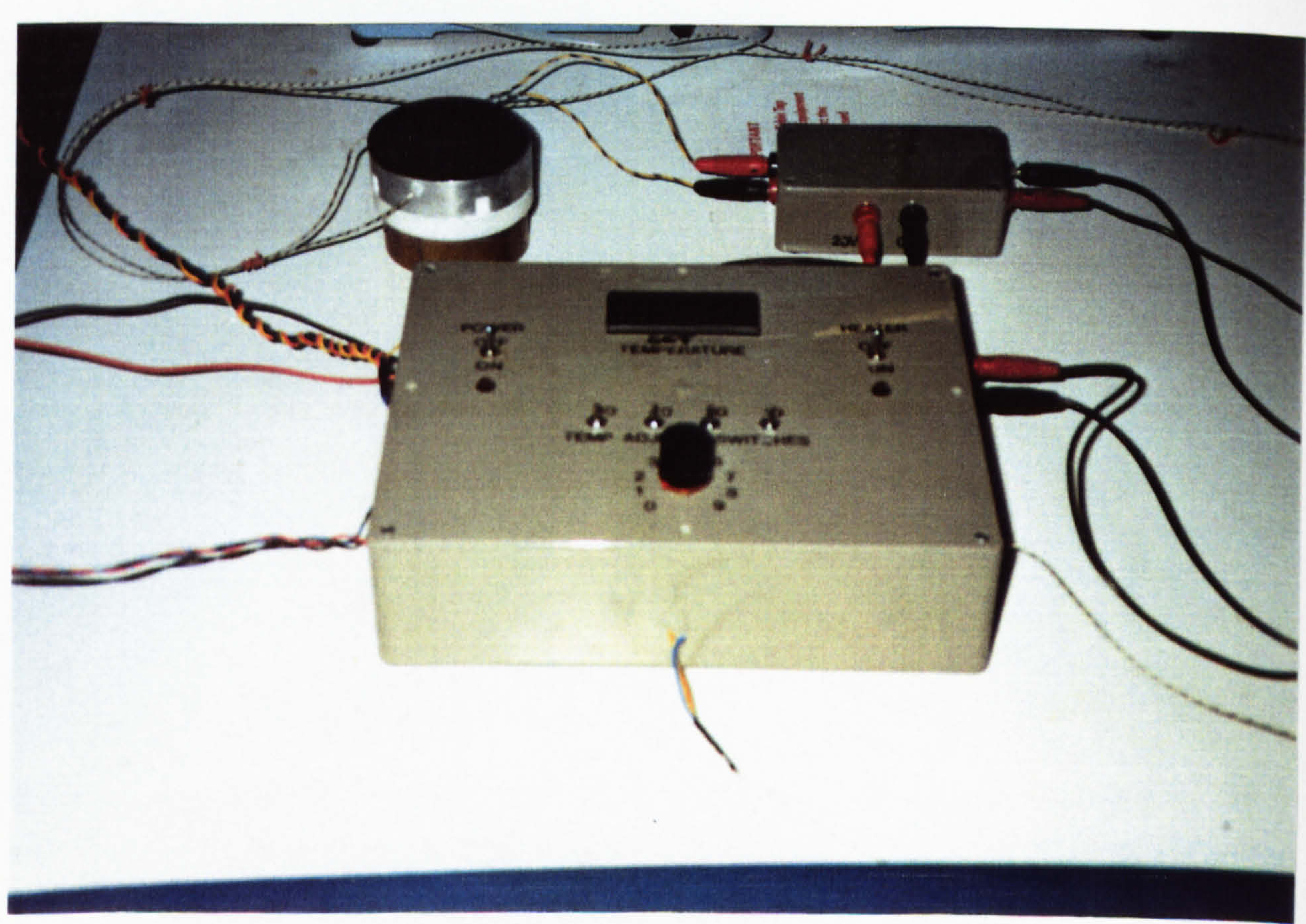
IEEE Reliability Physics Symp. (IRPS), pp.23-7, 1978.

208. Whitehead,A.P., Lynch,J.T., *"The Effect of ESD on CCD Reliability"*, Proc. 5th. EOS/ESD Symp., pp.147-53, 1983.
209. Freeman,E.R., Beall,J.R., *"Control of Electrostatic Discharge Damage to Semiconductors"*, Proc. 12th. IEEE Reliability Physics Symp. (IRPS), pp.304-12, 1974.
210. Youn,S.Y., Hartdegen,N, Sharp,M, *"ESD Minimisation Technique for MOS Manufacturing Final Test Area"*, Proc. 4th., EOS/ESD Symp., pp.157-64, 1982.
211. Giusti,J, *"The Probability of an ESD Failure in an Unprotected Environment"*, Proc. 8th. EOS/ESD Symp., pp.62-8, 1986.
212. Halperin,S.A., *"Static Control using Topical Antistats"*, Proc. 1st. EOS/ESD Symposium, pp.13-21, 1979.
213. McAteer,O.J., *An Effective ESD Awareness Training Program"*, Proc. 1st. EOS/ESD Symposium, pp.1-3, 1979. (The same paper appears in Proc. 2nd. EOS/ESD Symp., pp.189-91, 1980.)
214. Mykkanen,C.F., Blinde,D.R., *"The Room Air Ionization System, a Better Alternative than 40% Relative Humidity"*, Proc. 5th. EOS/ESD Symposium, pp.67-75, 1983.
215. Carlton,R.M., *"Effects of Air Ions and Electric Fields on Health and Productivity"*, Proc. 6th. EOS/ESD Symp., pp.34-9, 1984.
216. Fuqua,N.B., *"A 'Tailorable' ESD Control Program for the Manufacturing Environment"*, Proc. 8th. EOS/ESD Symp., pp.19-29, 1986.
217. Maene,N, Vandenbroeck,J, Allaert,K, *"Optimisation of ESD Protection Structures for Latch-Up and Minimum Area for Use in 1.2μm CMOS Telecommunication Circuits"*, Proc. 1st. Components Engineering, Reliability and Test Conference (CERT), pp.88-94, 1990.
218. Shenton,G, *"MOS and the Static Problem"*, Electronic Production, Sept.1976, pp.52-7, 1976.
219. Keller,J.K., *"Protection of MOS Integrated Circuits from Destruction by Electrostatic Discharge"*, Proc. 2nd. EOS/ESD Symposium, pp.73-80, 1980.
220. Duvvury,C, Rountree,R.N., White,L.S., *"A Summary of Most Effective Electrostatic Discharge Protection Circuits for MOS Memories and their Observed Failure Modes"*, Proc. 5th. EOS/ESD Symp., pp.181-4, 1983.
221. Korn,S, *"Improved Transient Suppression Made Possible by Advances in Varistor Technology"*, Proc. 7th. EOS/ESD Symp., pp.41-4, 1985.
222. Fong,Y, Hu,C, *"High-Current Snapback Characteristics of MOSFET's"*, IEEE Trans. Electron Dev., 37(9), pp.2101-3, 1990.
223. Krieger,G, *"The Dynamics of Electrostatic Discharge Prior to Bipolar Action Related Snap-Back"*, Proc. 11th. EOS/ESD Symp., pp.136-44, 1989.
224. Rountree,R.N., Hutchins,C.L., *"NMOS Protection Circuitry"*, IEEE Trans. Electron Dev., ED32(5), pp.910-7, 1985.
225. Wunsch,D.C., Bell,R.R., *"Determination of Threshold Failure Levels of Semiconductor Diodes and Transistors due to Pulsed Voltages"*, IEEE Trans. Nuc. Sci., NS15, pp.244-59, 1968.
226. Tasca,D.M., *"Pulse Power Failure Modes in Semiconductors"*, IEEE Trans. Nuc. Sci., NS17, pp.364-72, 1970.
227. Dwyer,V.M., Franklin,A.J., Campbell,D.S., *"Thermal Failure in Semiconductor Devices"*, Sol. State Elect., 33(5), pp.553-60, 1990.

228. Beltman,R.A.M., Krabbenborg,B, Mouthaan,A.J., *"Modelling of Electro-Thermal Effects in Integrated Circuits during ESD Events"*, European ESD Symposium, 1991.
229. Tabata,Y, Matsuo,Y, Yoshimochi,H, *"Shield Effect of Electrically Conductive Materials against Electromagnetic Waves Radiated by Electrostatic Discharge"*, Proc. 9th. EOS/ESD Symp., pp.153-7, 1987.
230. Crook,D.L., *"Method for Determining Reliability Screens for Time Dependent Dielectric Breakdown"*, Proc. 17th. IEEE Reliability Physics Symp., pp.1-7, 1979.
231. Stojadinovic,N.D., *"Failure Physics of Integrated Circuits - a Review"*, Microelectronics & Reliability, 23, pp.609-707, 1983.



1: Photograph of microscope stage with wafer-chuck and microprobers.



2: Photograph of wafer temperature controller, showing control console, heated chuck, thermocouples and power-transistor unit.

Plate 1: Photographs of Experimental Apparatus.

Chapter 4

Experimental Apparatus and Test Samples

4.1 Introduction

This chapter describes the apparatus and test structures used for the practical work of the thesis. The experimental facility has been under continuous development for several years and earlier versions have been described by Amerasekera [1] and Franklin [2]. This section describes the present state of the system in its entirety, including several innovations. Plate 1 (opposite) shows a photograph of the system is its present state.

The test structures, supplied on unscribed wafers, were obtained from four different semiconductor manufacturers. They cover a wide range of MOS integrated circuit technologies, including NMOS, CMOS and SOS (see Section 2.5). The experimental procedures and results are presented in Chapter 5.

4.2 Apparatus

4.2.1 Wafer Chuck and Microprober System

Two different wafer chucks were used to probe the silicon wafers. The first, intended for 4" wafers, was made from brass and held the wafer in place by means of a vacuum pump system. The second type, designed for 3" wafers, was made from aluminium and held the wafers in position by a system of mechanical clips arranged around the wafer periphery. The 3" chuck was heated by a series of wire-wound resistors, embedded in the body of the chuck, while its temperature was monitored by a Cr/Al thermocouple. The loop was completed by an electronic proportional controller, developed as an undergraduate project [3]. This system accurately stabilised the chuck (and hence the wafer) at any user-defined temperature between ambient and 200°C. Temperature was independently monitored via a second thermocouple by a Comark Type 1602-2 electronic thermometer. Since these thermocouples acted as antennae for surrounding field perturbations, they were temporarily removed during electrical measurements. Plate 1 (opposite) shows a photograph of the temperature control system.

Electrical contact to the device structures was made using adjustable 20μm tip

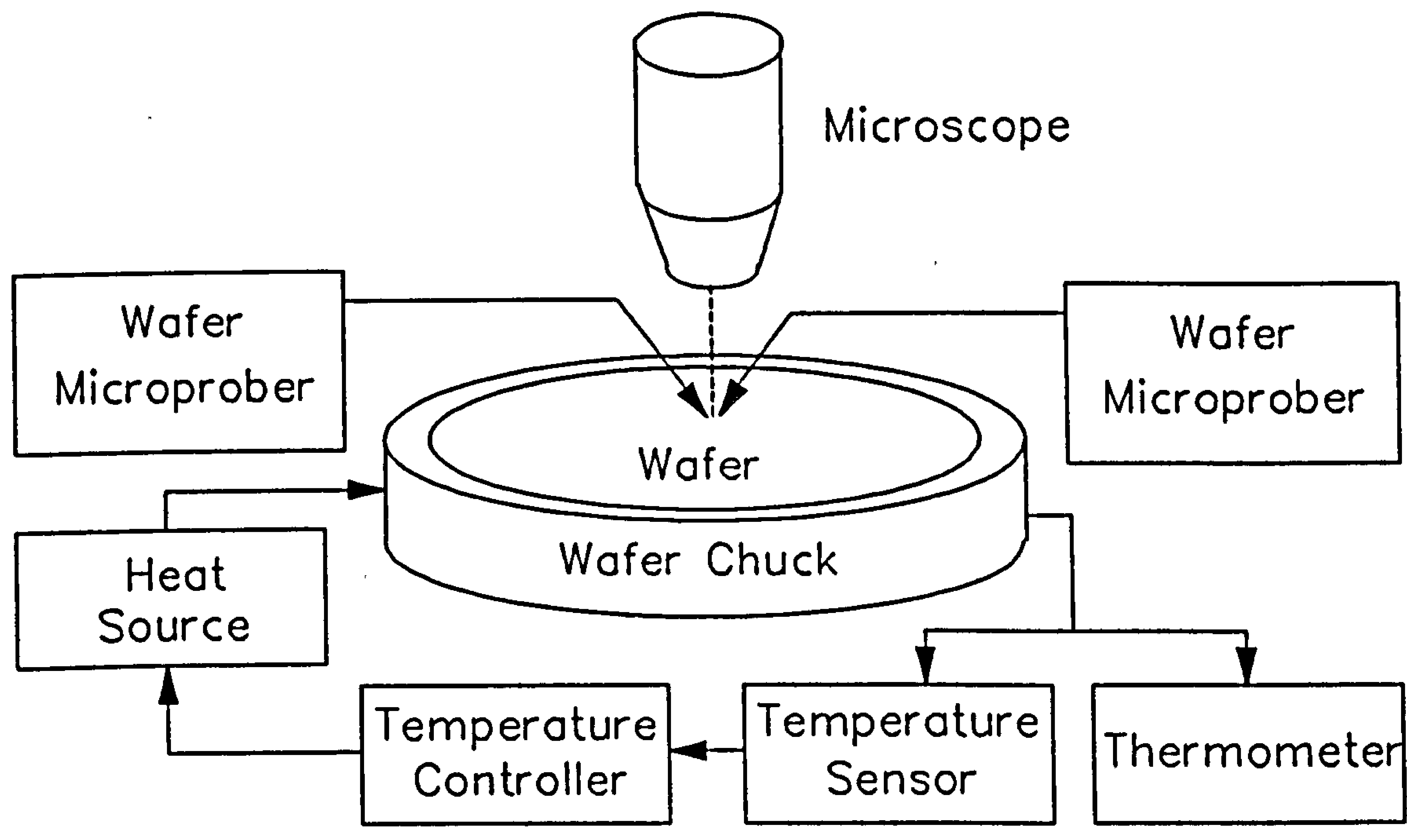


Figure 4.1: Wafer microprober station (together with temperature control system).

microprobes, by which the device bond pads were manually probed under microscopic observation. Substrate contact was made using the chuck itself, which was in contact with the metallised wafer base. The overall arrangement is shown in Fig.4.1.

The microscope sample illuminator had three levels of intensity denoted 1, 2 and 3. Since some of the experiments in this thesis are light sensitive, three standard levels of luminous intensity were defined. These are described in Table 4.1.

Table 4.1: Definition of Illumination Levels.

Illumination Level	Description
LOW	Microscope illumination off, laboratory lighting off.
MODERATE	Microscope illumination on Level 1, laboratory lighting on.
HIGH	Microscope illumination on Level 3, laboratory lighting on.

4.2.2 Electrical Instruments

4.2.2.1 Parametric Analyser

Parametric analysis of devices was performed using a Hewlett Packard HP4145B semiconductor parameter analyser, which is capable of performing a number of different functions. The HP4145B is basically a programmable d.c. measurement system consisting of two voltage sources, two voltage monitors and four source-monitor units (SMUs). The latter are two-mode devices which can source current and monitor voltage or source voltage and monitor current. The arrangement is shown schematically in Fig.4.2.

The parametric analyser can measure the I/V characteristics of a device or it can measure current or voltage as a function of time. Accuracy is ultimately limited by system noise, and measurements to within 10mV/10nA are readily attainable. The instrument also has a range of analytical functions (e.g. measurement of the slope/intercept of a tangent to an I/V characteristic), enabling the user to extract device parameters from the measured characteristics.

Measured data were stored on 3½" floppy disks for future retrieval. Hard copies of characteristics were also obtained using a Hewlett Packard HP7475A plotter.

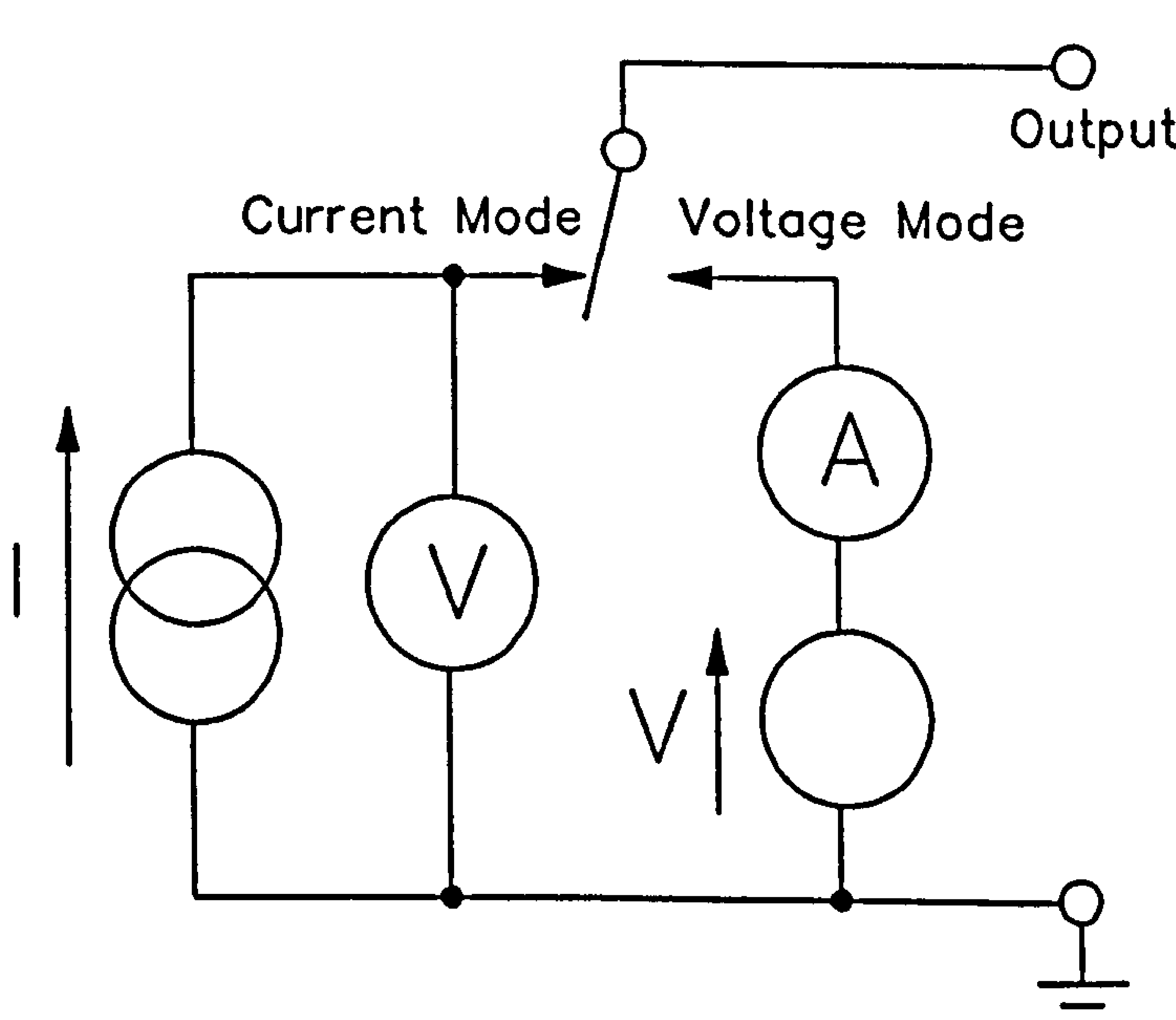


Figure 4.2: Schematic Diagram of Source-Measure Unit (SMU).

4.2.2.2 Programmable Electrometer

Electrical measurements were also made using the Keithley 617 programmable electrometer. Very high accuracies were available: resistance scales range between 2kΩ and 200GΩ with a 50 parts per million resolution. The instrument has a "suppress" function, allowing the user to eliminate stray circuit elements in the test leads prior to taking any measurements. It also contains a voltage source which can be set between 0 and 102V in both polarities with a resolution of 50mV and a maximum output current of 2mA.

4.2.2.3 Programmable Voltage Source

The Keithley 230 programmable voltage source provides precision d.c. voltages between 0 and 100V in both polarities with a resolution below 10mV and an output current limit between 2mA and 100mA. The unit has an internal program memory, allowing it to generate a user-defined voltage waveform.

4.2.2.4 L-C-R Meter

The Wayne-Kerr 4210 L-C-R bridge was used to measure high frequency circuit elements. This instrument measures resistance, capacitance and inductance in parallel or series configuration using sampling frequencies of 100Hz, 1kHz and 10kHz. It also incorporates a "suppress" function, similar to that of the Keithley 617 electrometer, which allows the parasitic circuit elements of the test probes and leads to be eliminated from the measurements.

4.2.2.5 Digital Oscilloscope

The Hewlett Packard 54111D digital oscilloscope can observe single-shot events and transfer the captured waveform to a computer via an HPIB interface. The controlling software, developed by Dr.A.J.Franklin [4], runs on a Walters 286 AT compatible desktop computer and allows the waveform data to be stored on a disk or transferred to the Mathsoft *MathCAD* mathematical software system for analysis (Fig.4.3). The maximum sampling speed of the oscilloscope is 10^9 samples per second, allowing the measurement of frequency components up to 500MHz.

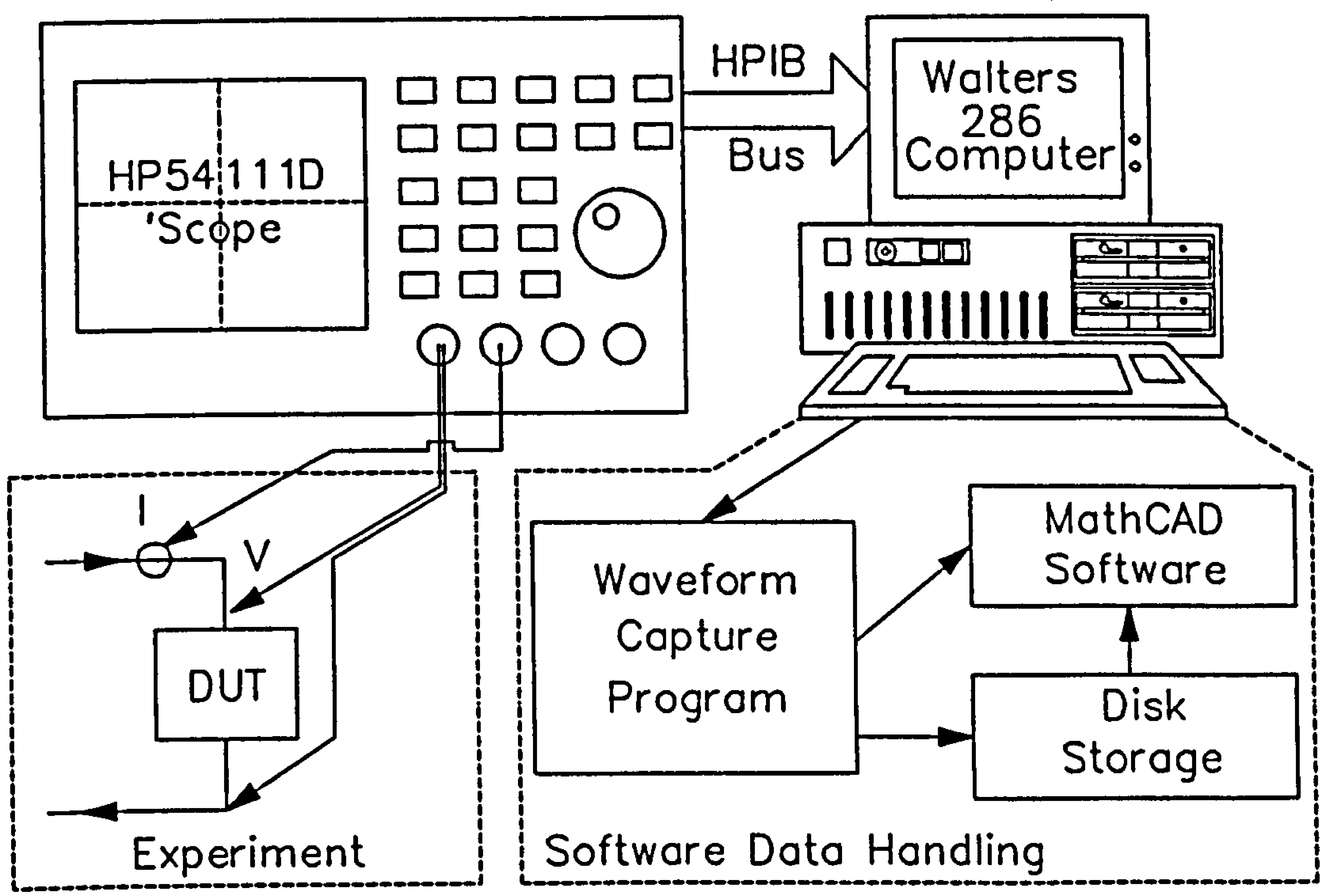


Figure 4.3: Waveform Capture and Analysis System.

The signals monitored by the oscilloscope were either voltages, measured by the Hewlett Packard 10431A (1MΩ/6.5pF) and 10440A (10MΩ/2.5pF) voltage probes, or electrical currents, measured by the Tektronix CT1 current-transformer probe. The HP10431A and HP10440A probes were labelled A and B respectively for identification.

The signals from the voltage and current probes needed very little in the way of mathematical processing. Although probe attenuations were corrected by the oscilloscope itself, a problem was experienced with the zero level, whose value tended to drift. For this reason, the computer was programmed to calculate its own zero-voltage level by averaging the leading tenth of the waveform (prior to the beginning of the pulse) and then subtracting this value from the signal, i.e.

$$V(t) = V_{sc}(t) - \frac{10}{T_{sc}} \int_0^{T_{sc}/10} V_{sc}(\tau)\, d\tau \qquad 4(1)$$

where $V_{sc}(t)$ is the voltage signal from the oscilloscope and T_{sc} is the time duration of the waveform.

Since the minimum measurement frequency of the CT1 probe is of the order of 1MHz, the pre-pulse current level (which has zero frequency) cannot be considered accurate

and the final tenth of the waveform (where the current has decayed to zero) was used instead. The spurious d.c. signal prior to pulsing was eliminated by multiplying the signal by the Heaviside step function, ie.

$$I(t) = \left[I_{sc}(t) - \frac{10}{T_{sc}} \int_{0.9T_{sc}}^{T_{sc}} I_{sc}(\tau)\, d\tau \right] H(t - t_d) \tag{4(2)}$$

where $I_{sc}(t)$ is the current signal from the oscilloscope and t_d is the time-delay to the beginning of the pulse. Fig.4.4 shows the processed voltage and current signals obtained for a -50V ESD pulse applied to a 48.1Ω resistor. The close agreement between I(t) and V(t)/48.1 illustrates the accuracy of the measurement.

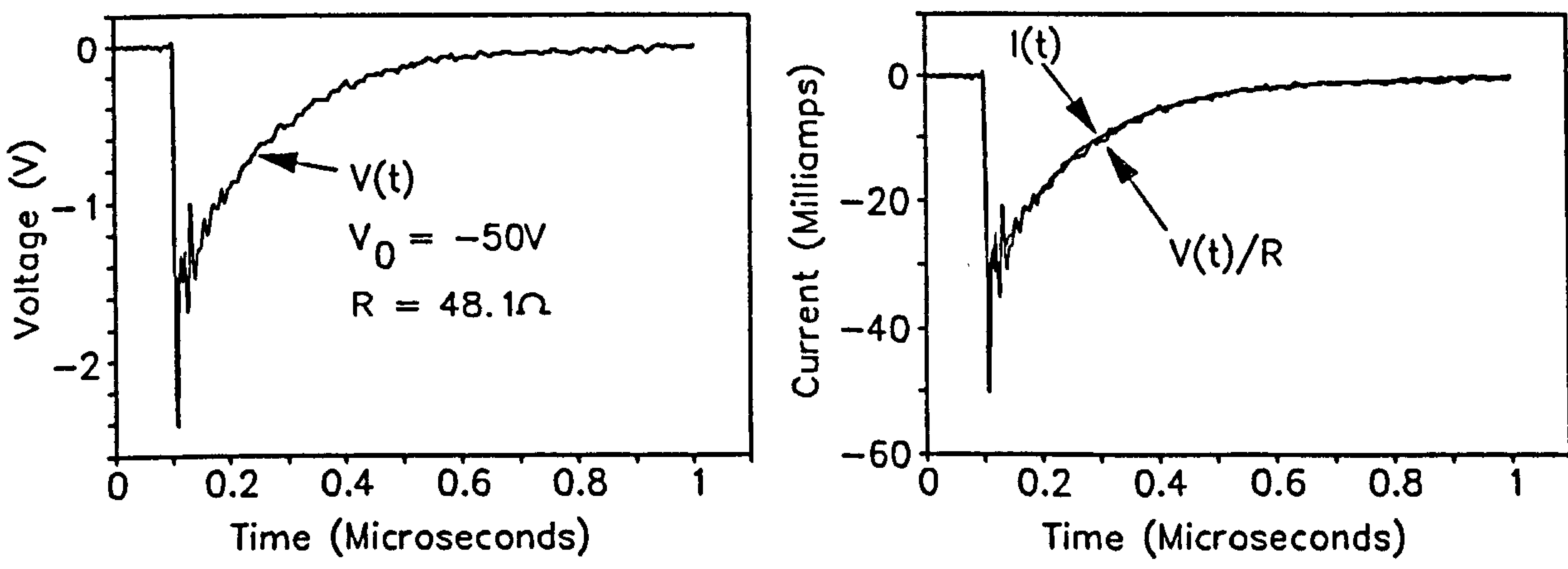


Figure 4.4: MathCad-processed waveforms of HBM electrostatic discharge pulse discharging into 48.1Ω, illustrating the accuracy of the waveform analysis technique.

4.2.3 Stress Generation

4.2.3.1 Human Body ESD Pulse Stress

Throughout the project, ESD was generated using the Hartley AutoZap ESD test system. This equipment produces 'Human Body Model' ESD pulses in accordance with the MIL-STD-883C specification (see Section 3.2.2). The system is shown schematically in Fig.4.5. Two high precision d.c. voltage sources were programmed to provide pulse voltages (V_0) between −4000V and +4000V in 20V steps. These voltages are used to charge the 'body' capacitance C_1 via the charging resistance R_1. C_1 is then discharged into the DUT via the 'body' resistance R_2. Charging and discharging is governed by a system of relays and the

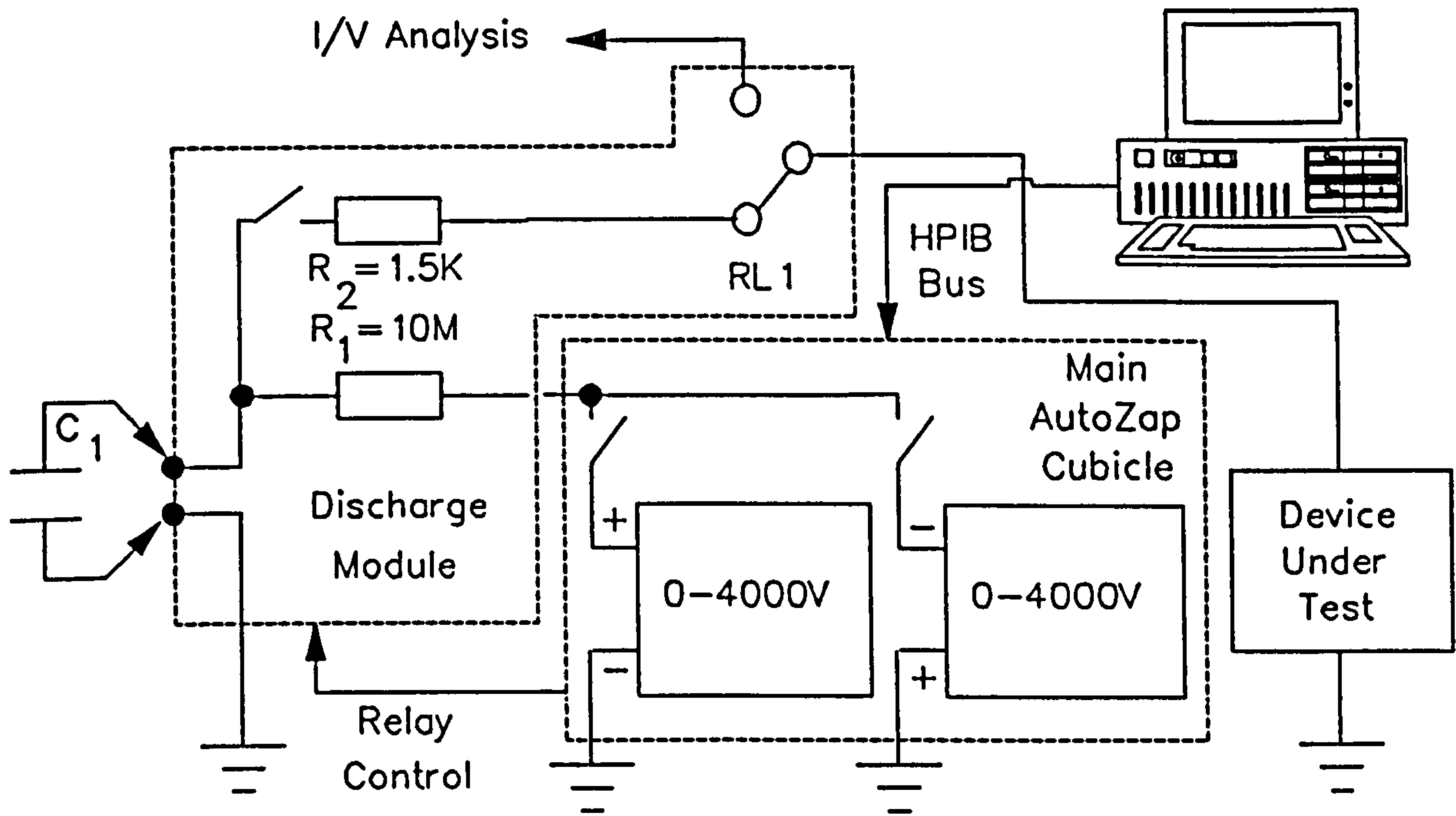


Figure 4.5: Schematic diagram of AutoZap ESD Simulation System.

entire operation is controlled by a Commodore 8296 desktop microcomputer.

For the purposes of these experiments, several alterations have been made to the standard hardware and software. The computer program was modified to allow the AutoZap to supply single pulses and incrementing pulse sequences according to user specifications. The negative voltage source was also modified to produce 0 to -800V in 4V steps. This source was used for both positive and negative polarity stressing by reversing the connections to the DUT.

The discharge module (containing the circuit elements C_1, R_1 and R_2 and the switch-over relay) and the device under test are usually mounted within the metallic AutoZap cabinet in order to protect the experiment from external electric fields. Since it is impractical to place the wafer microprober inside this enclosure, earlier studies used flying leads or co-axial cables running between the AutoZap and the microprober [1,2]. These cables presented large parasitic capacitances, introducing possible inaccuracies.

In the current study, the discharge module was removed from the AutoZap and mounted on the microscope stage beside the chuck. It was connected to the probes via short flying leads of negligible capacitance. The additional relay RL1 allows the DUT to be automatically switched between the AutoZap and other instruments.

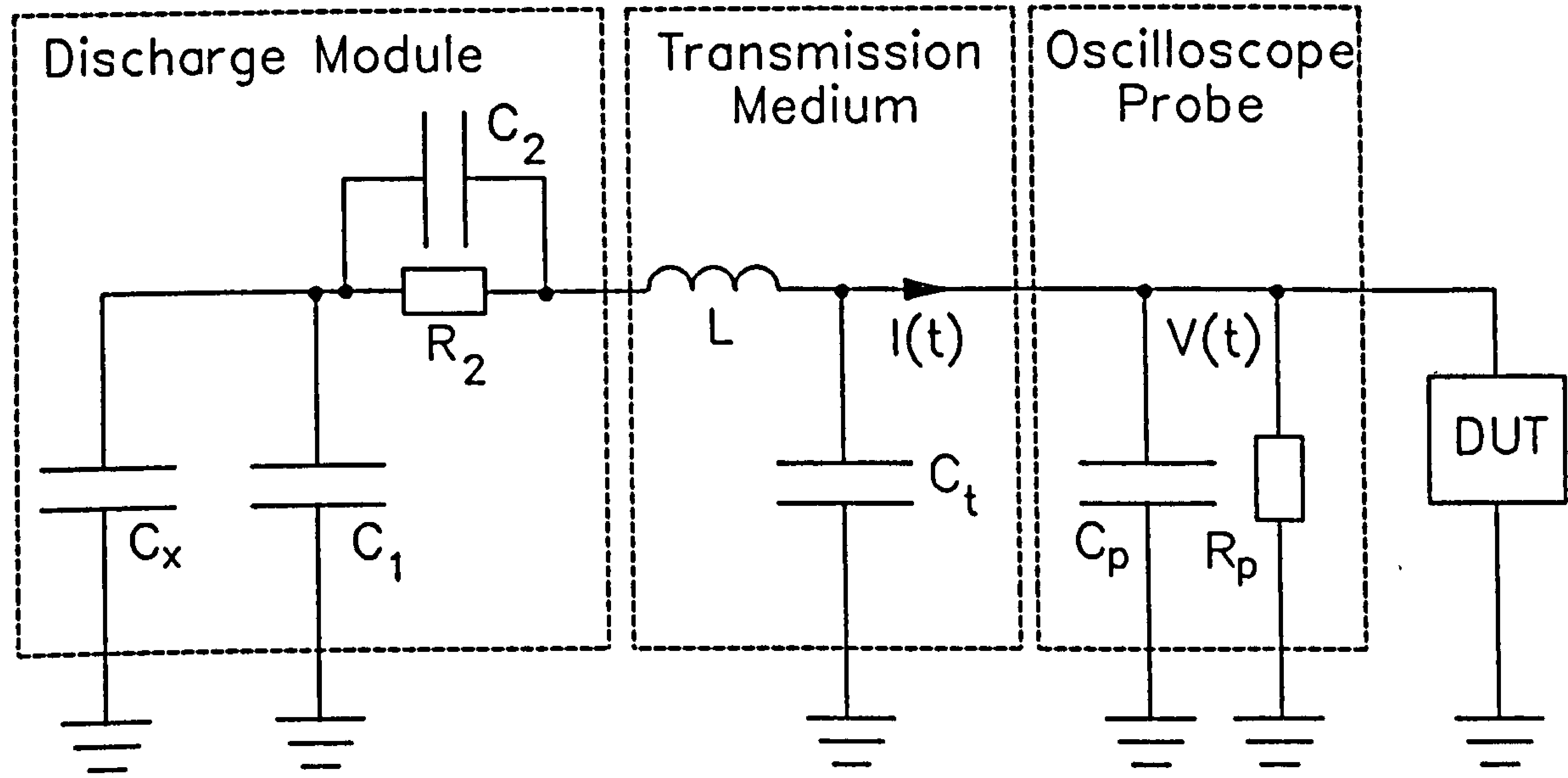


Figure 4.6: Equivalent circuit for ESD test system.

Fig.4.6 shows the equivalent circuit model used to characterise the AutoZap HBM test system, showing the following circuit elements:

C_1	'Human-body' discharge capacitor.
R_2	'Human-body' discharge resistor.
C_2	Parasitic capacitance parallel to R_2.
C_x	Parasitic capacitance parallel to C_1.
C_t	Total capacitance between discharge module and DUT, excluding capacitance of oscilloscope probe.
C_p	Oscilloscope probe capacitance.
R_p	Oscilloscope probe resistance.

Since the charge leakage via the probe resistance R_p can be ignored over the ESD rise-time constant, the total system charge can be assumed to be constant throughout t_{rise} [6], and the capacitances can be characterised using the following charge-balance technique:

With the wafer microprobe in the *up* position and a voltage probe connected, the effective pulse magnitude V_e (for an applied pulse voltage V_0=100V) was measured as a function of C_1 in the range 0-13pF. Assuming that the pulse charge is preserved, the graph of C_1 vs. $V_e/(V_0-V_e)$ should theoretically be linear, with a gradient of (C_t+C_p) and an intercept of $-C_x$. Hence (C_t+C_p) and C_x were calculated using the least-square method for probes A and B (individually and connected together in parallel) and the individual capacitances C_t, $C_p(A)$, $C_p(B)$ were determined by elimination (see Table 4.2). Since the C_1 vs. $V_e/(V_0-V_e)$ data shows no significant deviation from the linear model (Fig.4.7), the

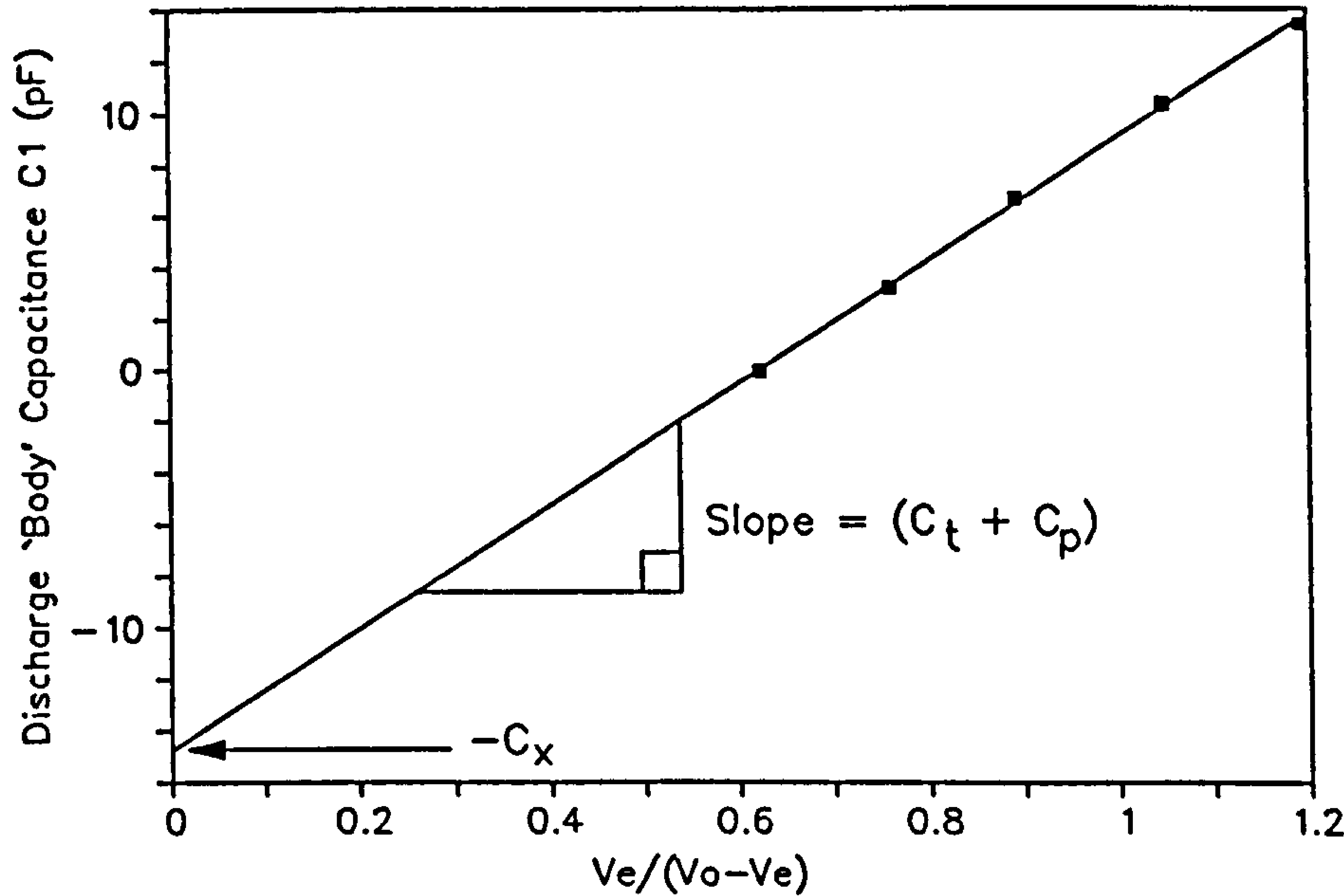


Figure 4.7: Graphical extrapolation of capacitance values in HBM circuit.

statistical errors in the experiment should be negligible. This was also confirmed by repeating the experiment and comparing results.

In order to determine C_2, R_2 was disconnected, a V_0=1000V pulse was applied and V_e was measured. Simple circuit theory allowed C_2 to be determined using the formula:

$$C_2 = \frac{(C_t + C_p)(C_x + C_1)V_e}{V_0(C_x + C_1) - V_e(C_t + C_p) - V_e(C_x + C_1)} \qquad 4(3)$$

The resulting value was C_2=1.029pF. The advantage of this technique (as opposed to measuring the components independently using an L-C-R meter) is that the system is arranged in exactly the same way as it is during a real ESD test.

Table 4.2: Capacitance Parameters

Arrangement	C_x (pF)	C_t (pF)	C_p(A) (pF)	C_p(B) (pF)
3" Chuck (-)	13.327	14.394	7.900	1.516
3" Chuck (+)	13.877	56.087	9.506	0.475
4" Chuck (-)	13.390	17.179	6.225	0.409
4" Chuck (+)	14.171	25.222	9.621	2.757

Although the inductance L produces a slight parasitic oscillation, the captured waveforms show that it has no noticeable effect on the ESD pulse magnitude V_c.

4.2.3.2 Constant Voltage Generation

Oxide breakdown under constant voltage stress conditions was examined using three separate techniques, which were suitable for long, medium and short time scales respectively. These are described below:

(i) Short Time-Scale Method (100ns-1ms) (Fig.4.8): The E-H 132L Pulse Generator provides 0-55V rectangular voltage pulses of either polarity. The pulse duration is controlled by the user and has a minimum value of 100ns. (Double pulse sequences can also be produced.) The output is diode 'clipped' at a value set by the Keithley 230 voltage source. Current and voltage waveforms are measured by the CT1 current probe and HP10431A voltage probe respectively. Fig.4.8 also shows typical current and voltage waveforms.

(ii) Long Time-Scale Method (>1s) (Fig.4.9): One SMU of the HP4145B parametric analyser was programmed to provide a constant voltage and monitor current as a function of time. Although the voltage and current are very precisely controlled and measured in this experiment, the time-response of the HP4145B (which is intended primarily for d.c. analysis) renders it inappropriate for time-scales below 0.5sec.

(iii) Medium Time-Scale Method (1ms-1s) (Fig.4.10): Constant voltage is produced by the Keithley 617 or 230 voltage source and current is sensed by a simple U.H.F. operational amplifier circuit. Since the op-amp input presents a virtual earth at low frequencies, it has minimal effect on the voltage across the oxide. However, this situation is only valid for frequencies below about 5MHz, rendering this technique unsuitable for very short time-scales. Furthermore, the Keithley voltage sources both have settling times of the order of 3ms.

4.2.3.3 Ramp Voltage Generation

Fig.4.11 shows the apparatus used to ramp the voltage across an MOS device. A purpose built function generator produces a variable speed voltage ramp which is applied to the DUT and a resistor R_s=1MΩ connected in series. R_s performs two functions: it limits the

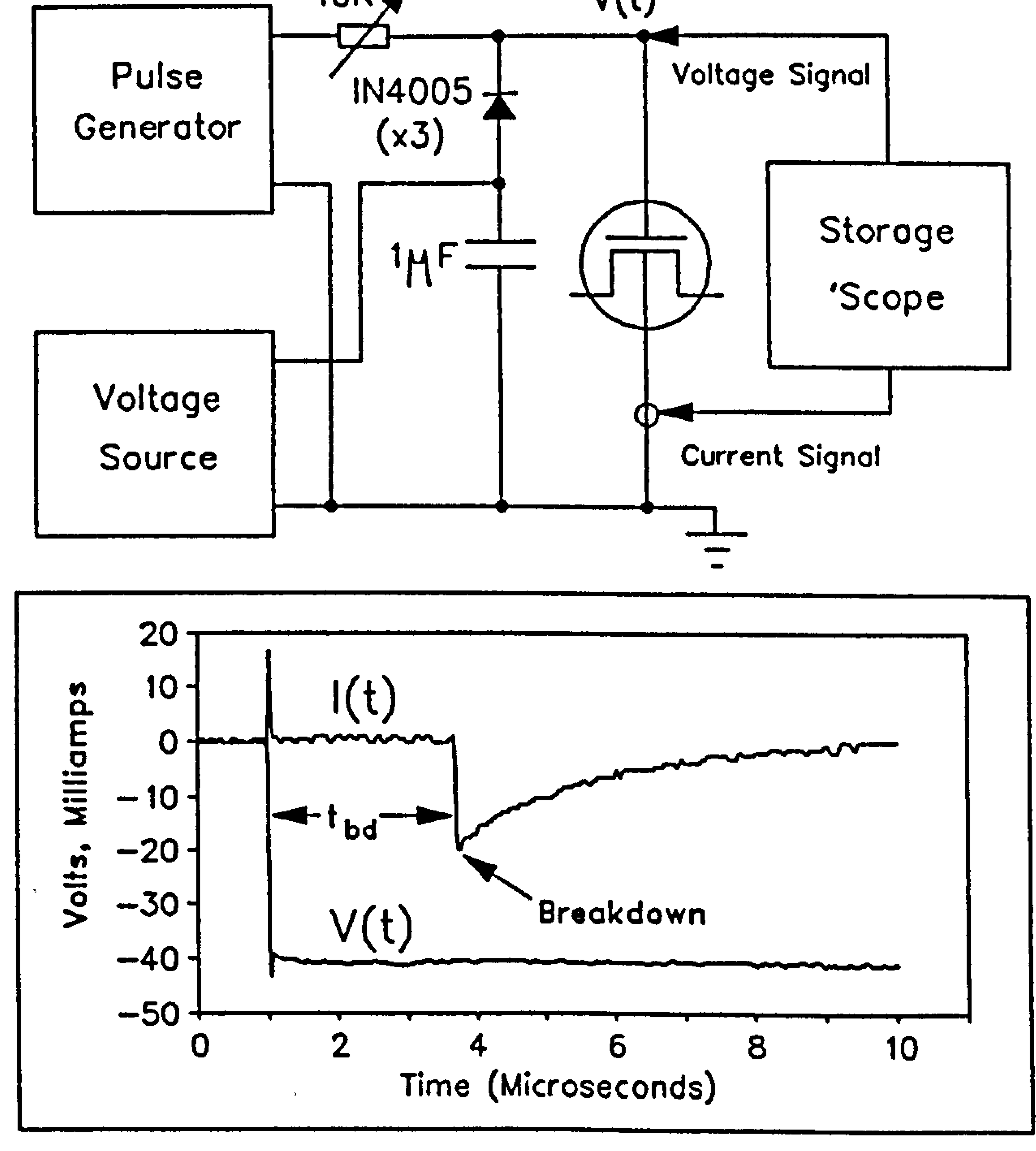


Figure 4.8: Apparatus Used For 'Short Time-Scale' Oxide Breakdown Measurements (inset shows typical voltage/current waveforms).

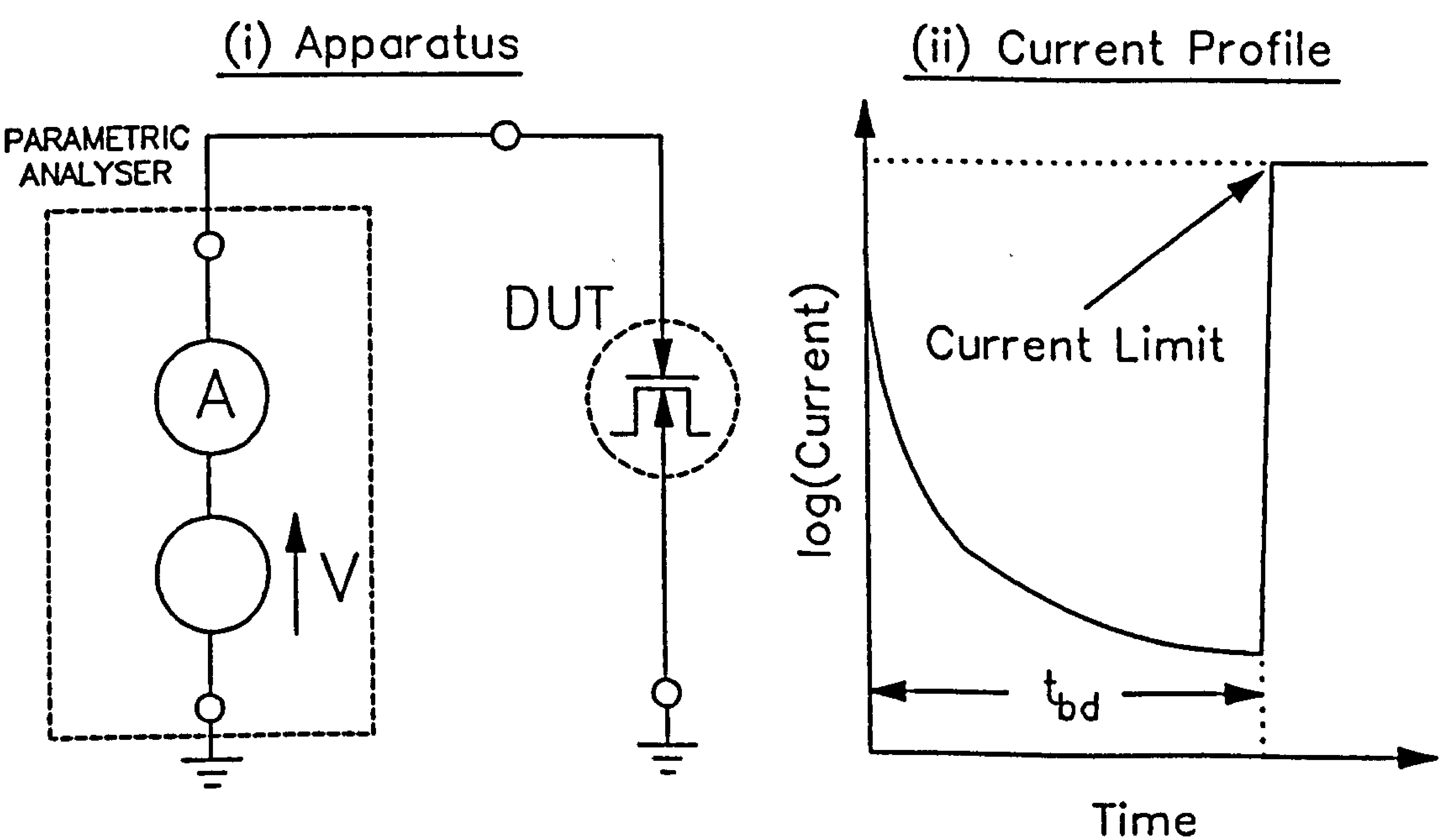


Figure 4.9: 'Long time-scale' apparatus and schematic current profile.

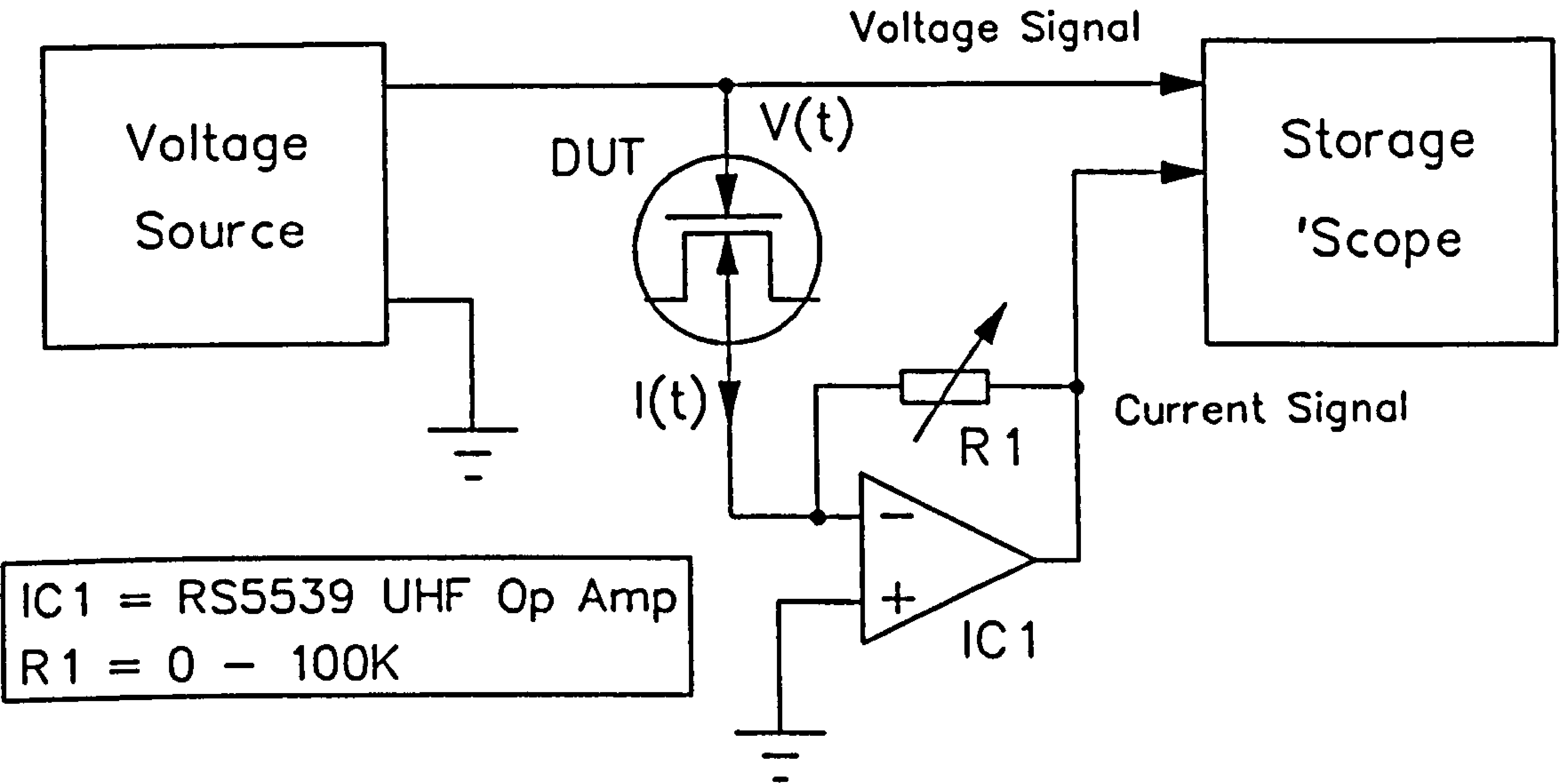


Figure 4.10: Apparatus Used for Medium Time-Scale Oxide-Breakdown Measurements.

current in the DUT and allows the DUT current to be monitored. The resistor voltage V_b and the overall voltage V_a were captured on the HP45111D oscilloscope and downloaded to the Walters 286 computer for analysis.

This apparatus is characterised in terms of the circuit model of Fig.4.12. Here C_{pc} and C_{cg} represent the total probe/chuck and chuck/ground capacitances respectively, while R_{cg} represents the parallel combination of R_s and the HP10440A oscilloscope probe resistance.

While R_s was measured using the Keithley 617 electrometer, C_{pc} and C_{cg} were measured in the following manner: With the microprobe positioned a fraction of a millimetre above the wafer surface, a voltage ramp was applied to the system and the V_a and V_b profiles were recorded. Simple circuit theory reveals that if $x(t)=V_b/V_a$ and $y(t)=\int_0^t V_b\,dt/V_a$ then

$$y(t) = -R_{cg}(C_{cg} + C_{pc})x(t) + R_{cg}C_{pc} \qquad 4(4)$$

Hence the x(t) and y(t) profiles were extracted from the experimental waveforms and the unknowns in Eqn.4(4) were determined using the method of least squares. The resulting parameter values (which were not significantly dependent upon the relative positions of chuck and prober) are listed in Table 4.3.

Table 4.3: Characterisation Parameters for Ramp Test Experiment

R_{cg} (KΩ)	C_{cg} (pF)	C_{pc} (pF)
916.7	24.69	1.491

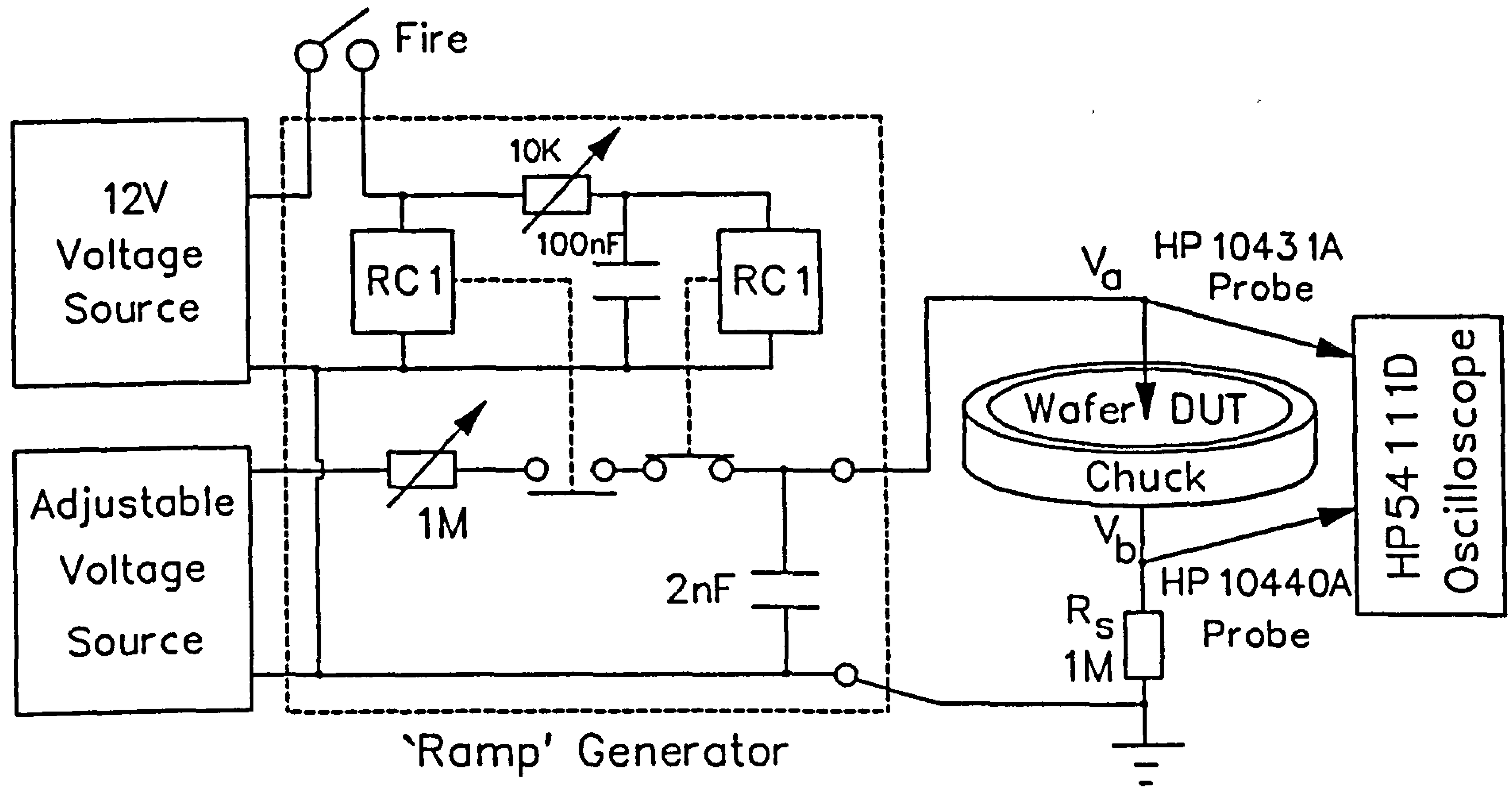


Figure 4.11: Apparatus for Voltage-Sweep Analysis.

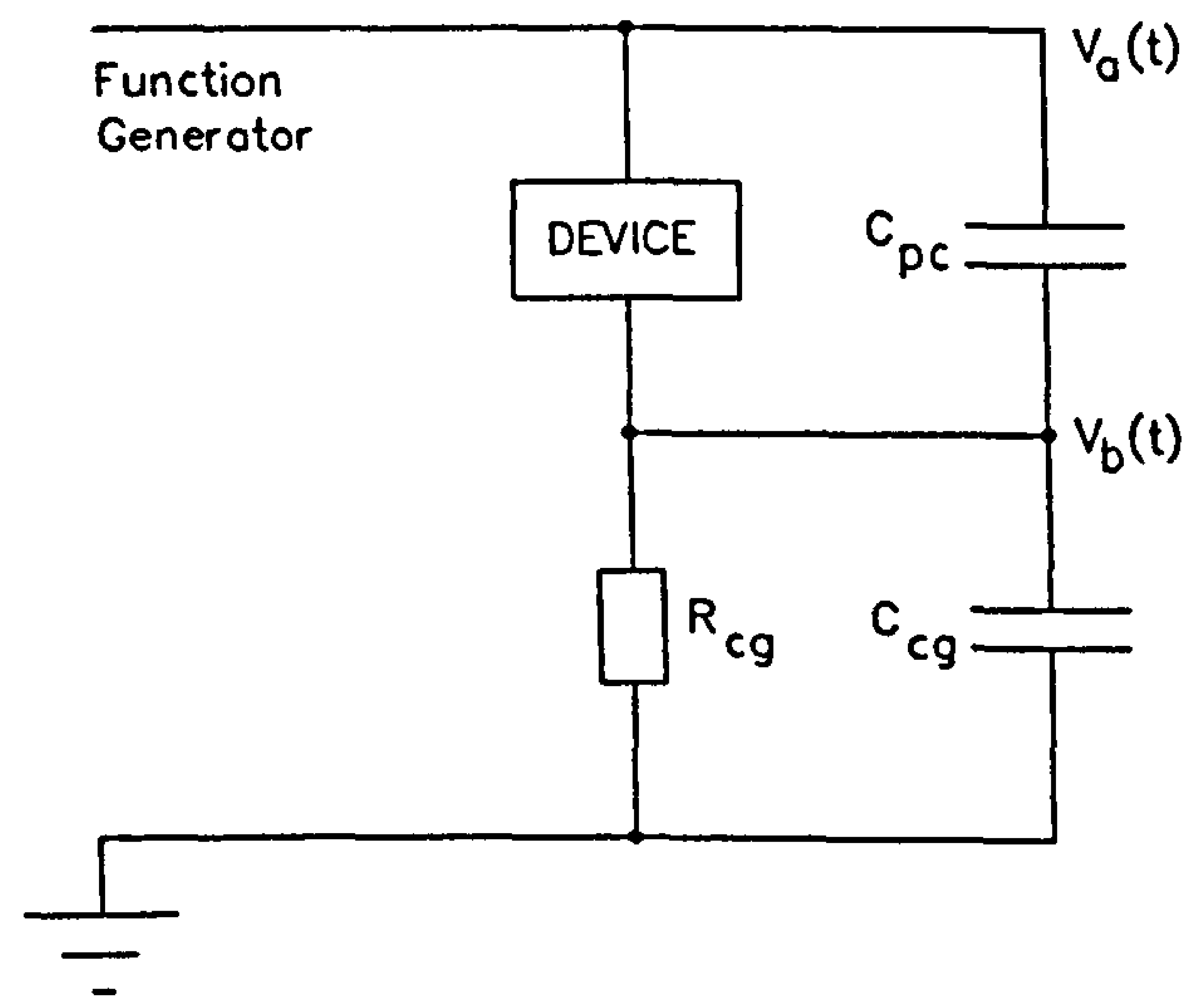


Figure 4.12: Characterisation of voltage ramp apparatus.

4.2.3.4 Constant Current Generation

The HP4145B parametric analyser was used to supply constant current. Since this instrument has a settling time of the order of 100ms, the technique was limited to relatively long time-scale measurements.

4.3 Test Samples

The sample structures were all fabricated on 3 or 4 inch wafers of bulk silicon or silicon-on-sapphire. The die on these wafers were not commercial circuits but process characterisation structures, especially designed to allow each device to be independently tested. However, the wafer quality was comparable to that encountered in commercial devices, so the experiments may be assumed to be realistic simulations of actual production situations. The wafers were supplied by four different manufacturers, who are referred to as suppliers A, B, C and D throughout this thesis. The samples are summarised below:

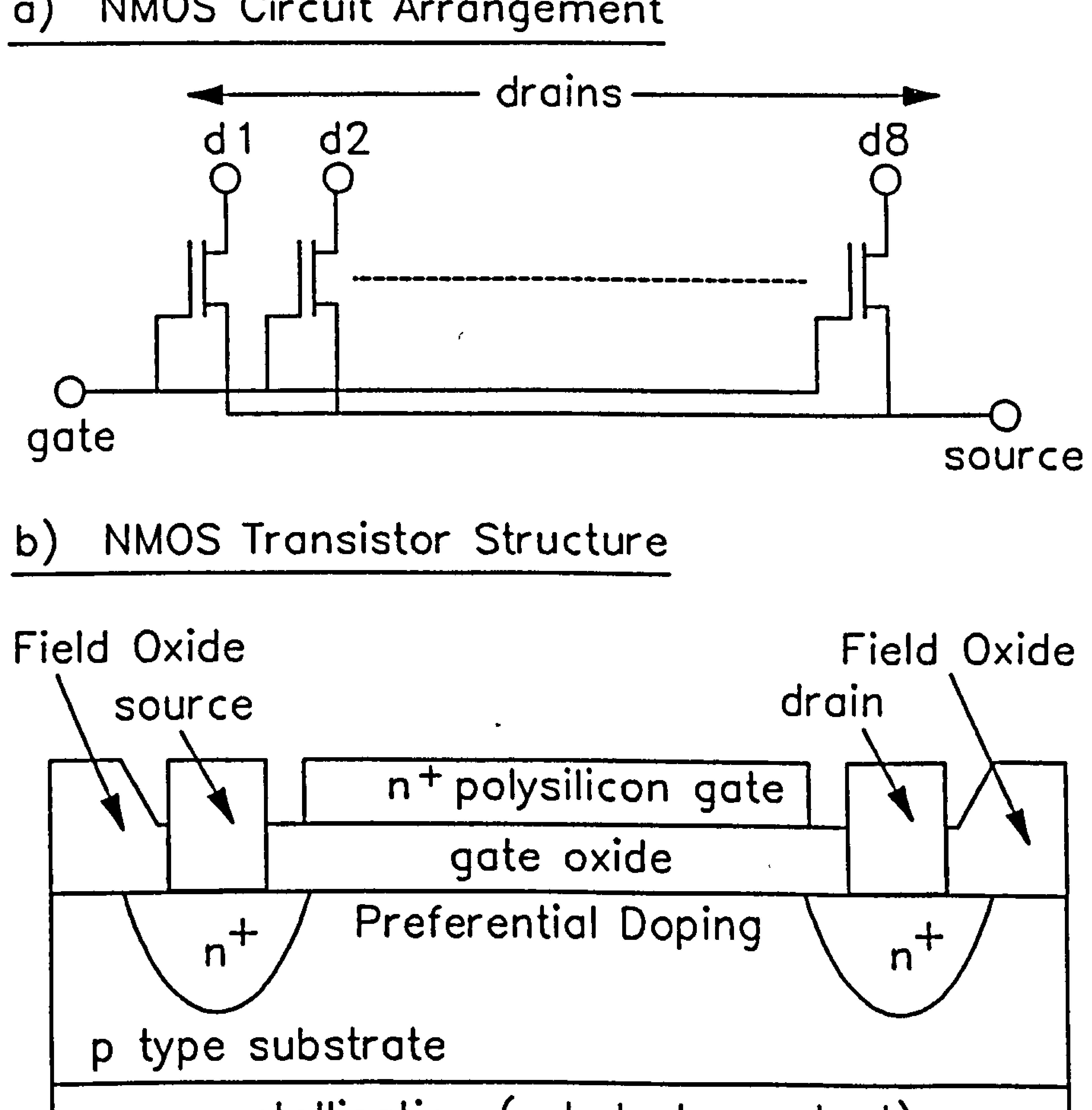


Figure 4.13: Schematic diagrams of NMOS structures.

4.3.1 Three Inch NMOS Bulk Silicon Wafers (Supplier A)

These wafers were composed of <100> 20Ωcm, p-type bulk silicon. The substrates received a preferential boron (p-type) implantation doping to $4.10^{15}cm^{-3}$ for the enhancement mode structures. Depletion mode structures had an additional phosphorus (n-type) implantation to $4.10^{15}cm^{-3}$. The wafers contained a selection of E and D-mode NMOS transistors, MOS capacitors and a various memory structures (which were not used in these experiments).

Gate oxide thicknesses were 400Å on all devices, with a Si-SiO_2 interface state density of $10^{11}cm^{-2}$ quoted by the manufacturers. All gate structures were composed of 4552Å vacuum deposited n^+-doped ($10^{21}cm^{-3}$) polysilicon. The field passivation oxide was 6000Å thick.

Channel lengths of both E-mode and D-mode transistors varied between approximately 1μm and 100μm. Sources and drains were produced by arsenic (n-type) implantation to $2.10^{20}cm^{-3}$. The MOSFETs were arranged in arrays of eight, with common sources and gates and individual drains (Fig.4.13a), allowing the characteristics of each device to be independently measured. Fig.4.13b shows a schematic cross section of an NMOS structure.

The wafers also contained rectangular wide area MOS capacitors with dimensions 175μm x 280μm. Some of these structures were fabricated upon the n-type $4.10^{15}cm^{-3}$ (E-mode) doping while others were fabricated on the p-type $4.10^{15}cm^{-3}$ (D-mode) doping. Some of the latter had further p-type implantation to $10^{20}cm^{-3}$. Electrical connection to these structures was made between the gate bond pads and the aluminium coated wafer base.

4.3.2 Three Inch CMOS Bulk Silicon Wafers (Supplier A)

These wafers were composed of <100> 25Ωcm, n-type bulk silicon and contained E-mode n-channel and E-mode p-channel MOSFETs. P-channel devices had a preferential arsenic (n-type) channel doping to $10^{16}cm^{-3}$, with boron (p-type) source and drain diffusions to $5.10^{19}cm^{-3}$. N-channel structures, fabricated in a $2.10^{15}cm^{-3}$ p-well, had preferential boron (p-type) channel doping to $10^{16}cm^{-3}$. Drain to source dimensions varied between 1 and 10μm. Circuit arrangement was identical to that of the NMOS structures described above.

Gate structures were composed of $10^{21}cm^{-3}$ doped polysilicon, evaporated onto 320Å SiO_2 gate oxides. The manufacturer's quoted Si-SiO_2 interface state density was $10^{11}cm^{-2}$. The field passivation oxide was 7200Å thick. The wafers also contained square 70μm x 70μm gate oxide capacitors.

4.3.3 Four Inch HMOS Bulk Silicon Wafers (Supplier B)

These wafers contained MOS transistor arrays similar to those described in Sections 4.3.1 and 4.3.2., with common gates and sources and individual drains. Channel lengths varied between 1 and 50μm. Additionally, rectangular gate oxide capacitors, 215 x 268μm were available. Gate electrodes were n^+ polysilicon and all gate oxides were 400Å thick.

4.3.4 Four Inch Bulk Silicon MOS Capacitor Wafers (Supplier C)

These wafers, made from p-type 17-23Ωcm bulk silicon, contained very wide area ($6.10^3 cm^2$) MOS capacitor structures. Gate oxides were 135Å thick (grown in 10% O_2 at 900°C), with n^+-type ($10^{20} cm^{-3}$) polysilicon gate electrodes. Electrical contact was made between a metallised spot on the centre of each gate and the metallised wafer base.

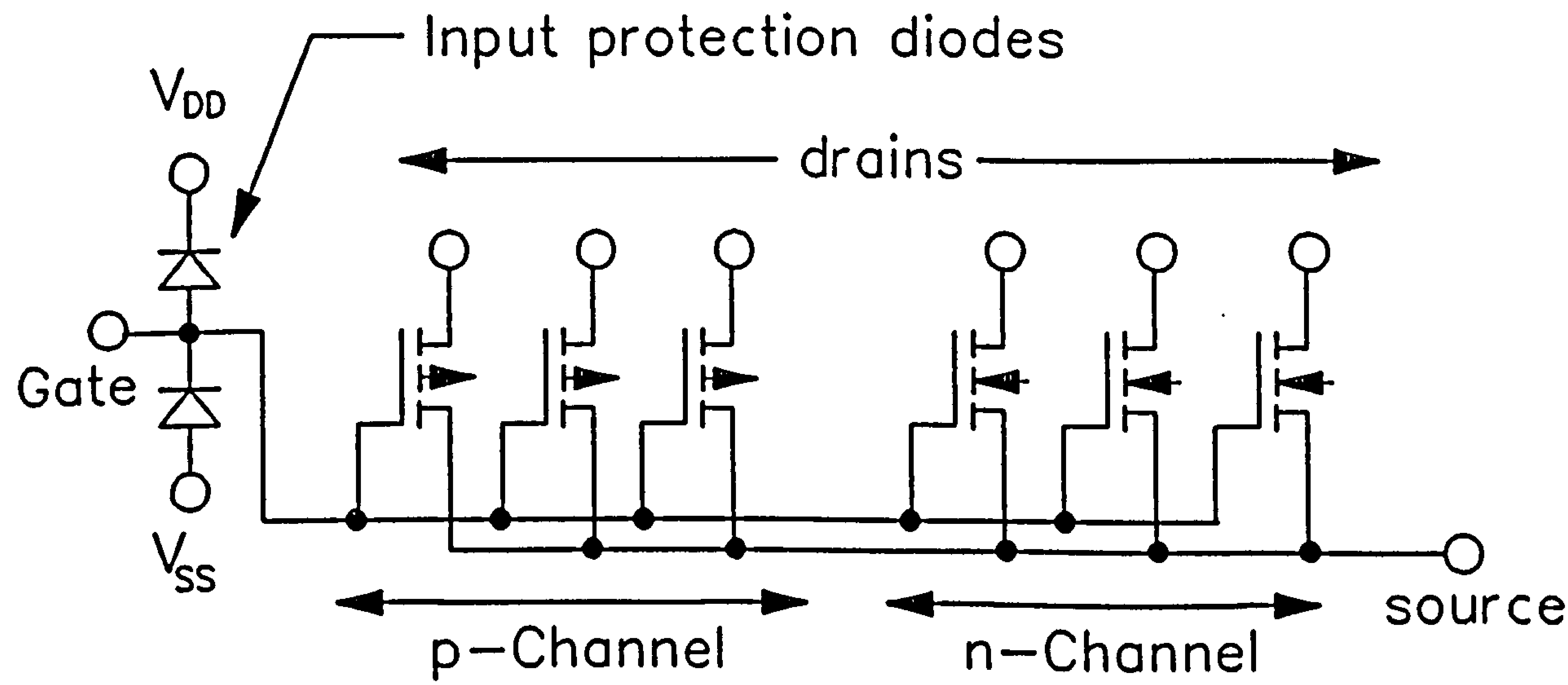


Figure 4.14: Organisation of Silicon-on-Sapphire (SOS) Test Structures.

4.3.5 Four Inch Silicon-on-Sapphire (SOS) Wafer (Supplier D)

These devices were identical to the SOS structure described in Chapter 2, consisting of E-Mode n and p-channel MOSFETS fabricated on individual Si islands on an insulating sapphire substrate. Like the MOS devices described above, the transistors were constructed in arrays (in this case of six), consisting of three n-channel and three p-channel devices

(Fig.4.14). While there was a common gate and source for all the devices on each array, the drains of individual devices could be probed independently. A protection circuit, consisting of two diodes was fabricated between the gate bond pad and the devices. Protection could easily be disabled, simply by leaving the V_{dd} and V_{ss} pads unconnected.

4.3.6 Past History of Samples

Some of the sample structures had previously been tested in order to characterise the fabrication processes. These tests measured such parameters as g_m and V_T, using voltages and currents of the magnitude expected during working life. These were non-destructive tests and, it may be assumed, did not stress the devices in any significant way.

4.4 Summary

1. This chapter has described the apparatus used for the practical work of the thesis. Techniques used to generate constant voltage, constant current, ramp voltage and ESD pulse stress have been outlined and characterised.

2. The test samples used in the experiments have also been described. These include NMOS, CMOS and HMOS bulk Si transistor structures, MOS capacitors and silicon-on-sapphire devices. The devices were obtained from four different suppliers.

4.5 References

1. Amerasekera,E.A., *"Failure Mechanisms in MOS Devices"*, Ph.D. Thesis, Loughborough University of Technology, pp.36-57, 1986.
2. Franklin,A.J., *"Electrical Overstress Failure in GaAs MESFET Structures"*, Ph.D. Thesis, Loughborough University of Technology, pp.84-98, 1990.
3. Oliver,M.A., Roberts,I, Walton,D.R., *"Temperature Control System to Heat Semiconductor Wafers"*, 2nd. Year Mini Project Report, Dept. of Electronic and Electrical Engineering, Loughborough University of Technology, May 1991.
4. Franklin,A.J., *"Electrical Overstress Failure in GaAs MESFET Structures"*, Ph.D. Thesis, Loughborough University of Technology, Appendix C (pp.C1-C5), 1990.
5. Tunnicliffe,M.J., Dwyer,V.M., Campbell,D.S., *"Slow Transient Voltage and ESD Breakdown in Unprotected MOS Gate Oxides"*, Proc. 1st. Components Engineering, Reliability and Test Conference (CERT), pp.78-87, 1990.

6. Amerasekera,E.A., Campbell,D.S., *"Charge-Limited Breakdown in MOS Capacitors"*, Proc. **17th.** European Solid State Device Research Conference (ESSDERC), pp.733-41, 1986.

Chapter 5

Experimental Procedures and Results

5.1 Introduction

This chapter presents a preliminary program of experiments, performed using the apparatus and test samples described in Chapter 4. MOS failure is examined over a wide range of experimental conditions, ranging from long-term oxide wearout to fast-pulse and ESD breakdown. The chapter does not cover all the experimental work of the project, but provides a foundation for the deeper diagnostic investigations which appear in Chapters Six, Seven and Eight.

5.2 Electrostatic Discharge Experiments

These experiments investigate the breakdown thresholds of the full range of MOS devices as functions of various parameters. The devices were stressed between their gate and substrate contacts using the AutoZap ESD apparatus described in Section 4.2.3.1.

Devices were subjected to ascending ESD pulse sequences, starting at $\pm$4V and ascending in steps of $\pm$4V until breakdown was observed. This technique is hereafter termed the *step test*. Breakdown was detected in terms of the oxide's I/V profile (which was measured after each pulse using the HP4145B parametric analyzer described in Section 4.2.2.1): The device was subjected to a 0-10V voltage ramp (in the same polarity as the stress pulse), and any leakage current detected above the noise level (approximately 100nA) was seen as an indication of breakdown.

All oxides were stressed and monitored under room temperature and moderate illumination (see Table 4.1) using standard HBM circuit parameters (discharge capacitance C_1=93.28pF and discharge resistance R_2=1.497KΩ) unless otherwise stated. The four-inch brass wafer chuck was used for all experiments except for those which required above-room-temperature conditions.

This section begins by reporting the experiments used to verify the inter- and intra-wafer uniformity of the test structures. It then proceeds to examine the sensitivity of ESD breakdown to variations in device size, temperature, illumination, C_1 and R_2.

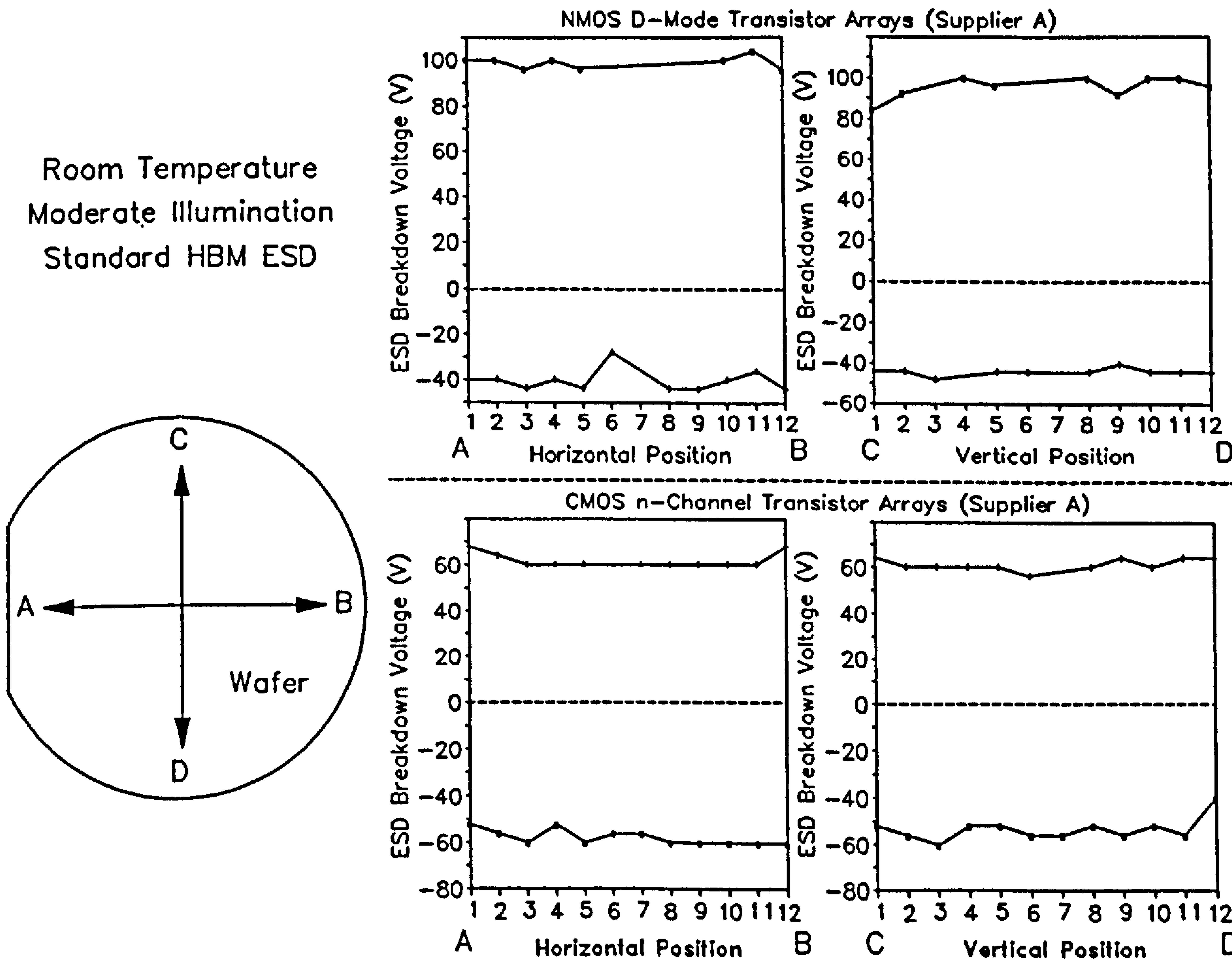


Figure 5.1: Positional Dependence of ESD Breakdown Thresholds in NMOS and CMOS Transistor Arrays (measured using Step Test).

5.2.1 Uniformity of Experimental Samples

The first stage of the research involved a characterisation of the wafers in terms of the uniformity of nominally identical test samples, selected from different geographical positions on different wafers. Fig.5.1 shows positive and negative polarity breakdown thresholds of NMOS and CMOS transistor arrays, selected by scanning a wafer horizontally (A to B) and vertically (C to D). Since no systematic variation exists and the experimental scatter is small (coefficients of variation: 4.8% [pos.], 10.2% [neg.]), the devices were considered to be roughly identical with minor statistical fluctuations. Fig.5.2 compares breakdown thresholds of devices selected from different wafers, showing that the samples are again uniform (mean values: +93.3/-43.3V [wafer 1], +92.7/-46V [wafer 2], +97.33/-44V [wafer 3].

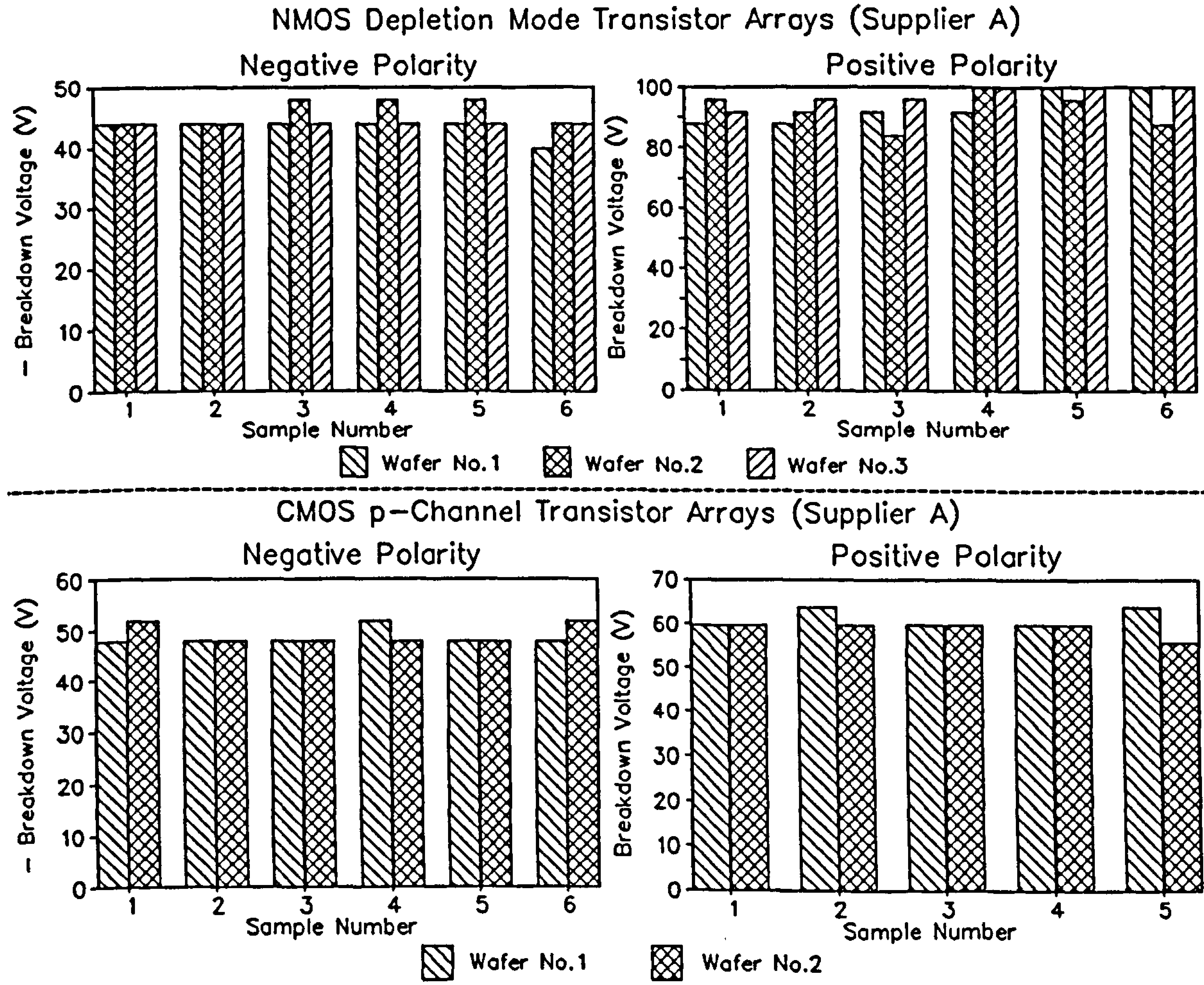


Figure 5.2: Comparison of ESD Breakdown Thresholds of NMOS and CMOS Devices Selected from Different Wafers (measured using Step Test).

Figs.5.3 and 5.4 show the positional variation and inter-wafer variation for HMOS devices. Although no positional dependence was detected, the breakdown voltages of HMOS Wafer No.2 were generally larger than those selected from Wafer No.1. For this reason, all HMOS experiments reported in this thesis were performed using Wafer No.1.

Fig.5.5 shows positional and inter-wafer variation in breakdown voltage for the MOS capacitor structures. The breakdown voltages clearly have a wide statistical variation, and a mean value which varies across wafer A. Breakdown voltages of Wafer B are also significantly lower than those of Wafer A, and Wafer B was also seen to contain a large number of pre-existing failures.

Fig.5.6 shows the positional variation of the silicon-on-sapphire breakdown voltage, which indicates approximate uniformity across the wafer. Since only one SOS wafer was available, no inter-wafer comparisons were possible.

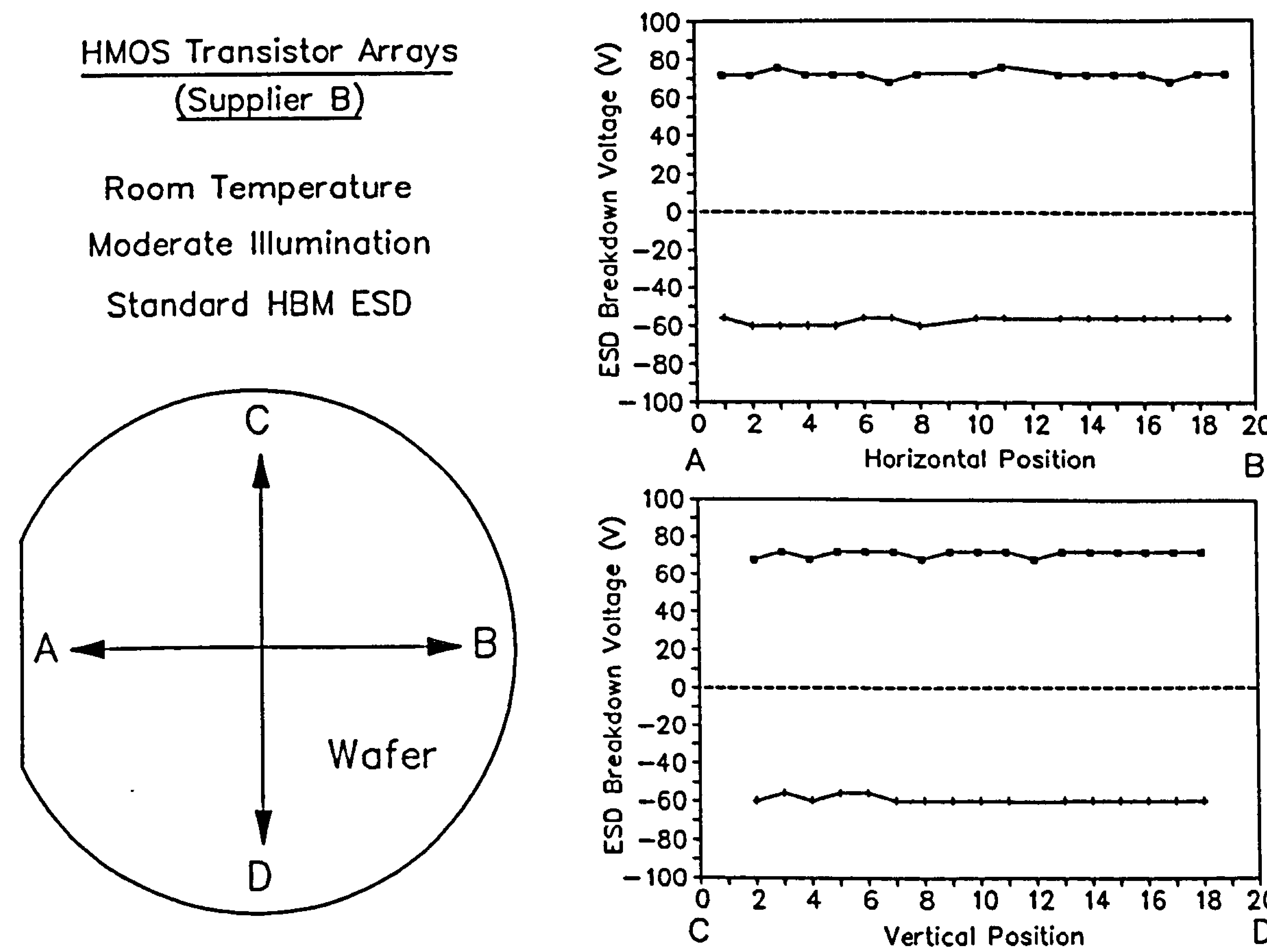


Figure 5.3: Positional Dependence of ESD Breakdown Voltage in HMOS Transistor Arrays (measured using Step Test).

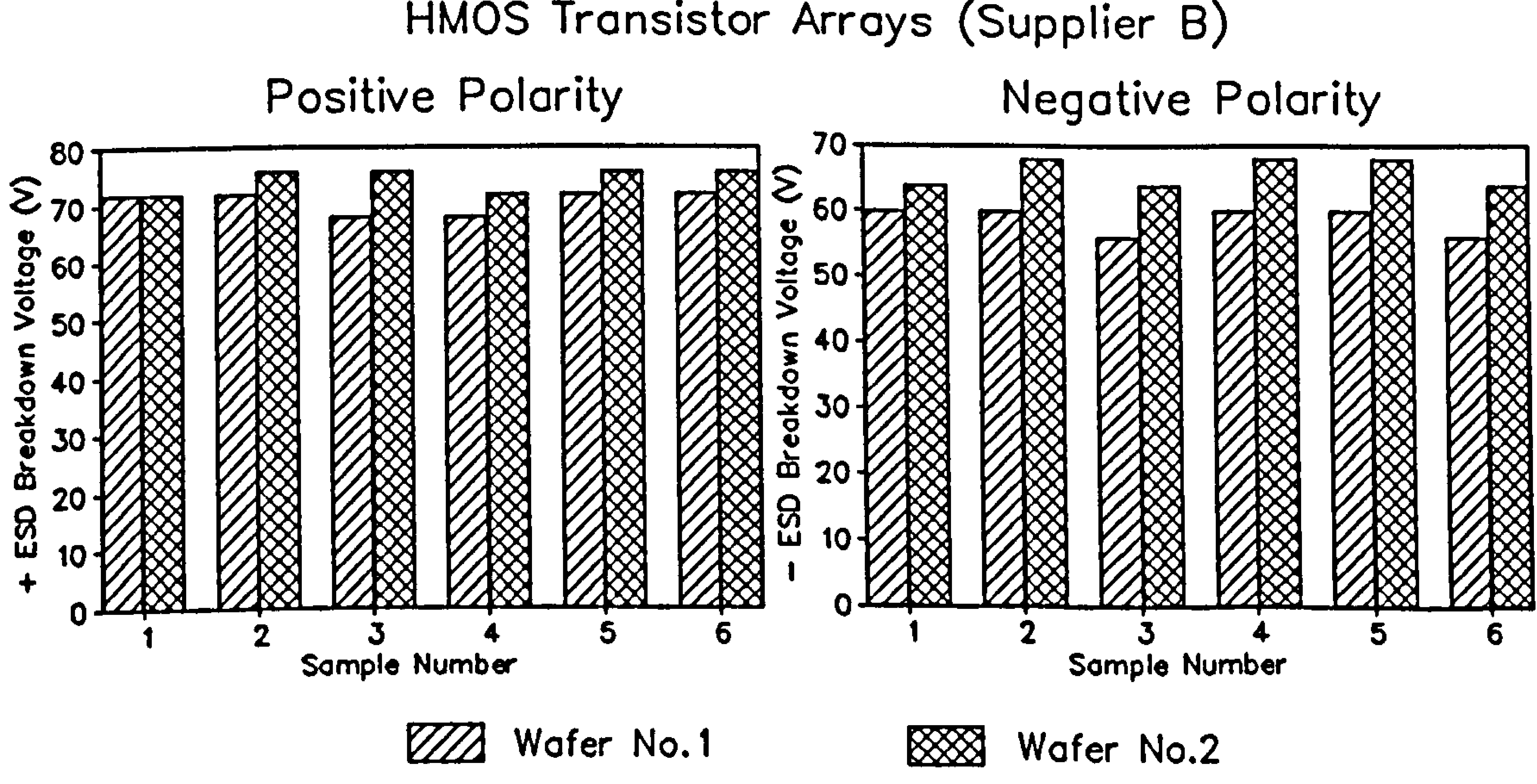


Figure 5.4: Comparison of ESD Breakdown Thresholds of HMOS Devices Selected from Different Wafers (measured using Step Test).

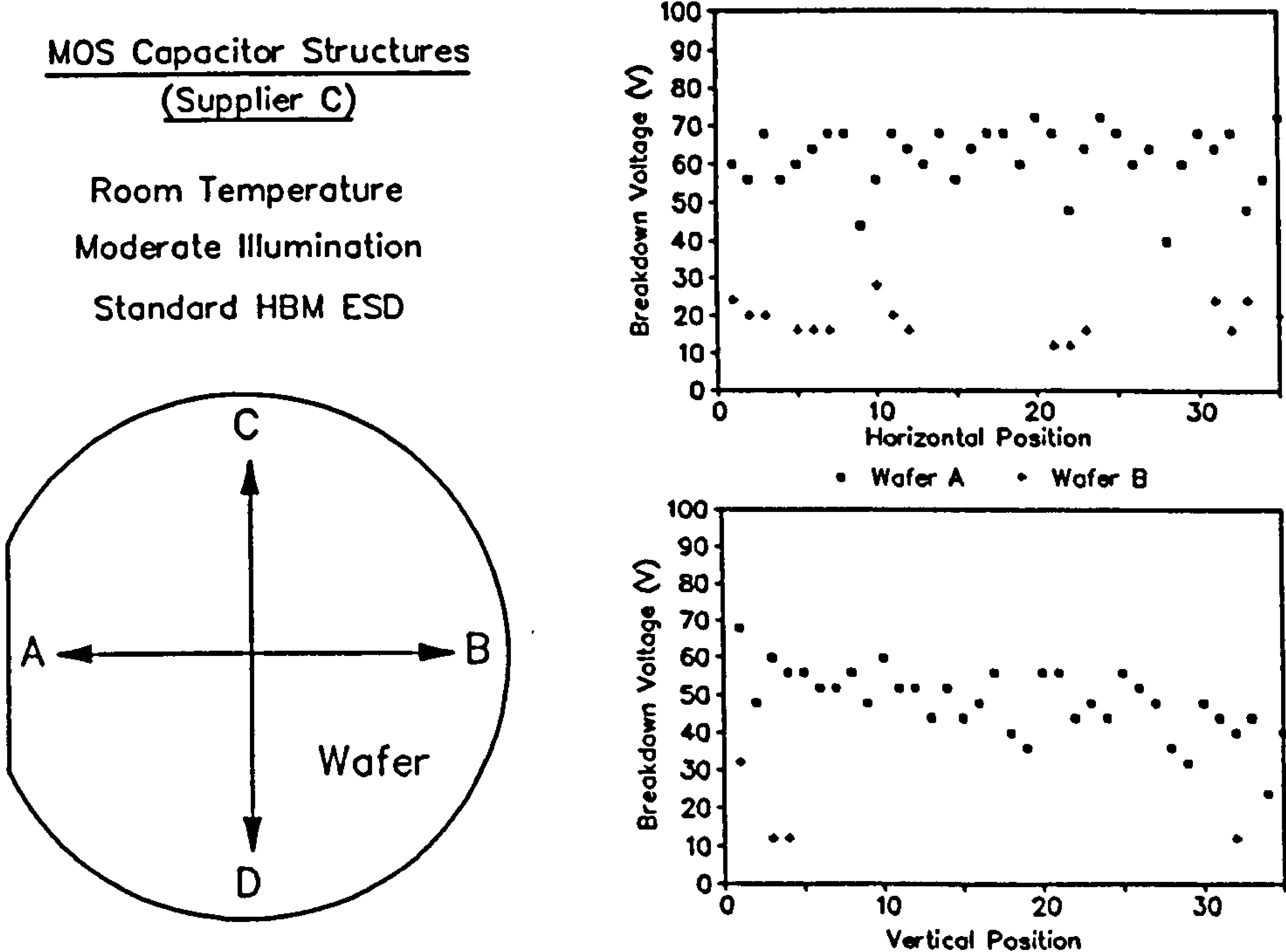


Figure 5.5: Positional and Inter-Wafer Dependence of ESD Breakdown Thresholds in MOS Capacitors (measured using Step Test).

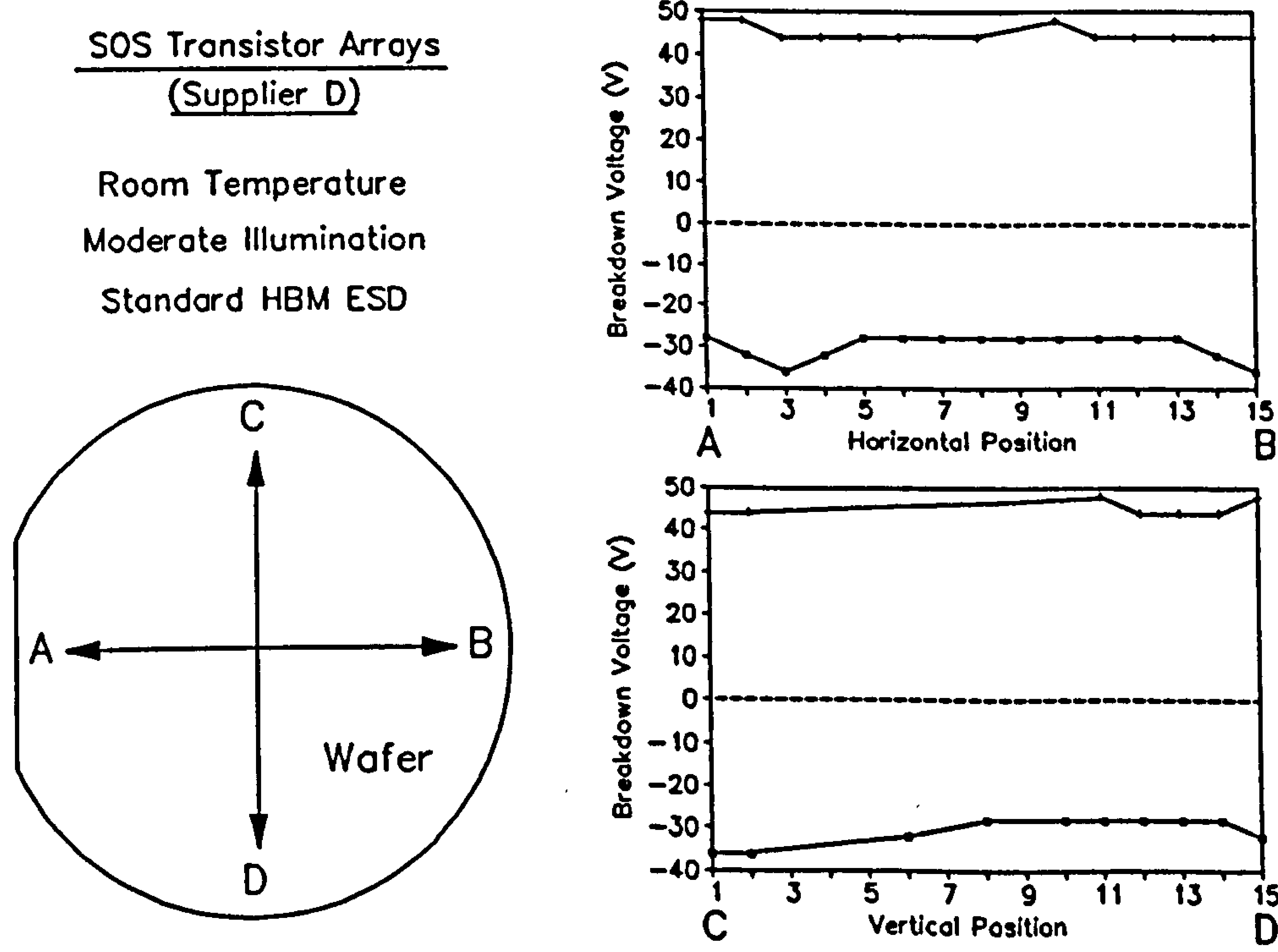


Figure 5.6: Positional Dependence of ESD Breakdown Threshold in SOS Transistor Arrays (measured using Step Test).

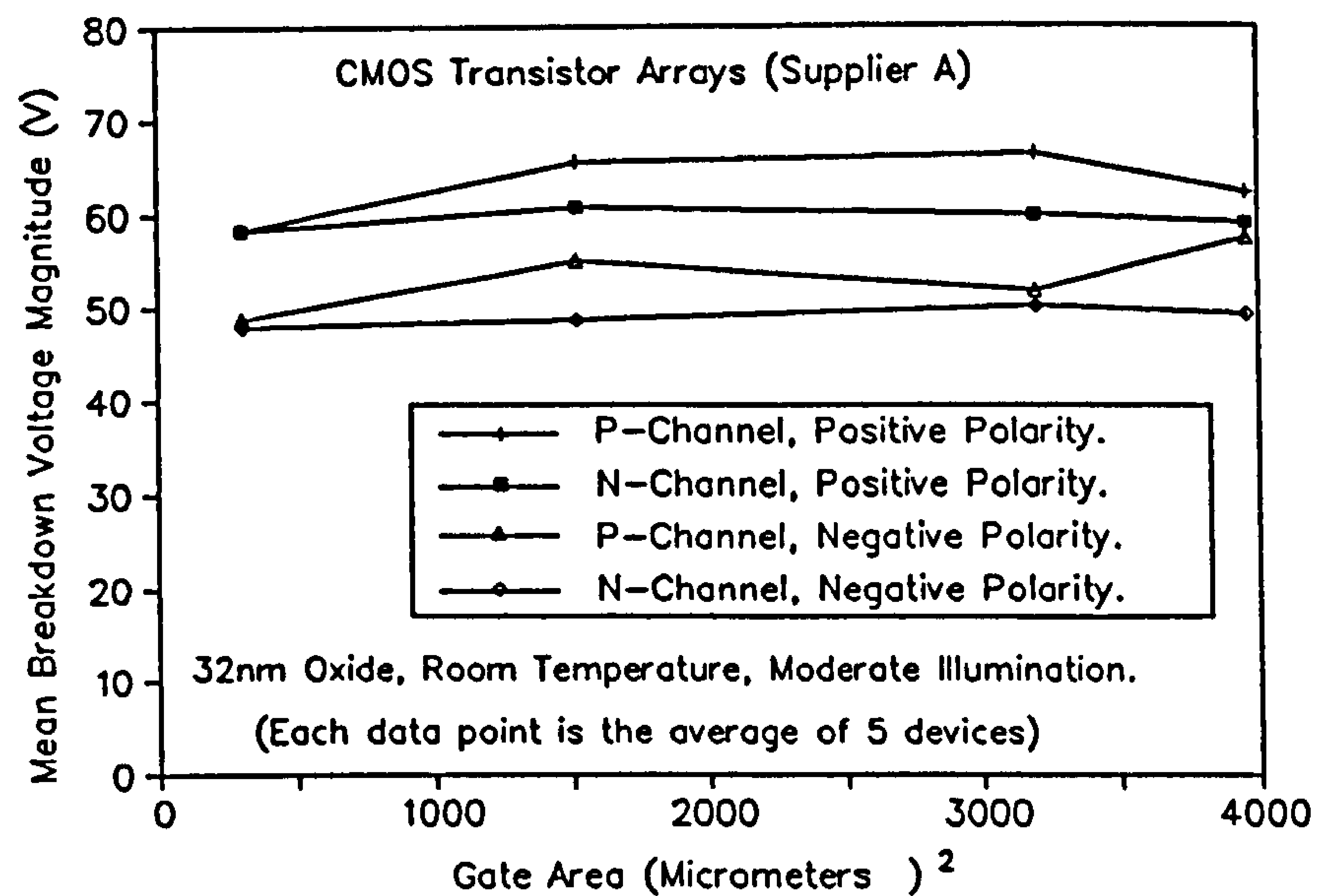


Figure 5.7: ESD Breakdown Thresholds of CMOS Devices as a Function of Gate Area.

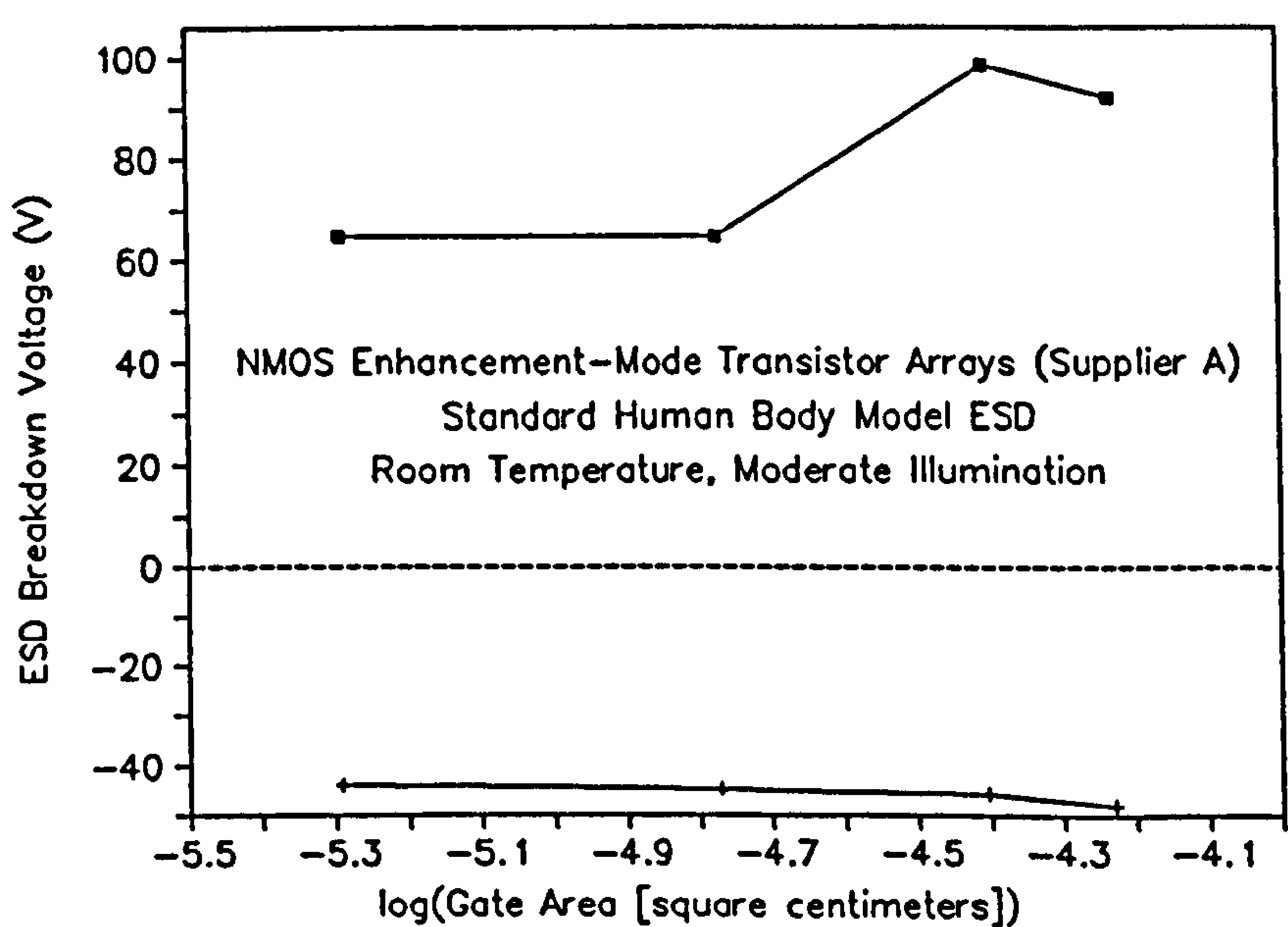


Figure 5.8: NMOS ESD Breakdown Threshold as a Function of Gate Area (each data point is the mean average of five devices).

5.2.2 Dimensional Dependence of ESD Breakdown

The relationship between the ESD breakdown voltage and the gate area was examined for CMOS and NMOS devices. Fig.5.7 shows positive and negative breakdown thresholds for both n and p-channel transistor arrays as a function of the total gate area. Each data point is the mean breakdown voltage of five identical structures. The experiment was repeated using NMOS E-Mode device structures and Fig.5.8 shows the results. (Note: Breakdown

voltages are plotted against log(area) merely in order to aid the presentation).

CMOS ESD breakdown appears to be largely insensitive to device size, although the positive thresholds are universally higher than their negative counterparts. The mean negative polarity threshold in NMOS devices appears to increase very slightly with area, although the total variation is less than the 4V measurement resolution. On the other hand, the positive polarity threshold varies by almost 35% across the measurement range. Although the variation is not monotonic, the smaller devices generally have smaller breakdown thresholds.

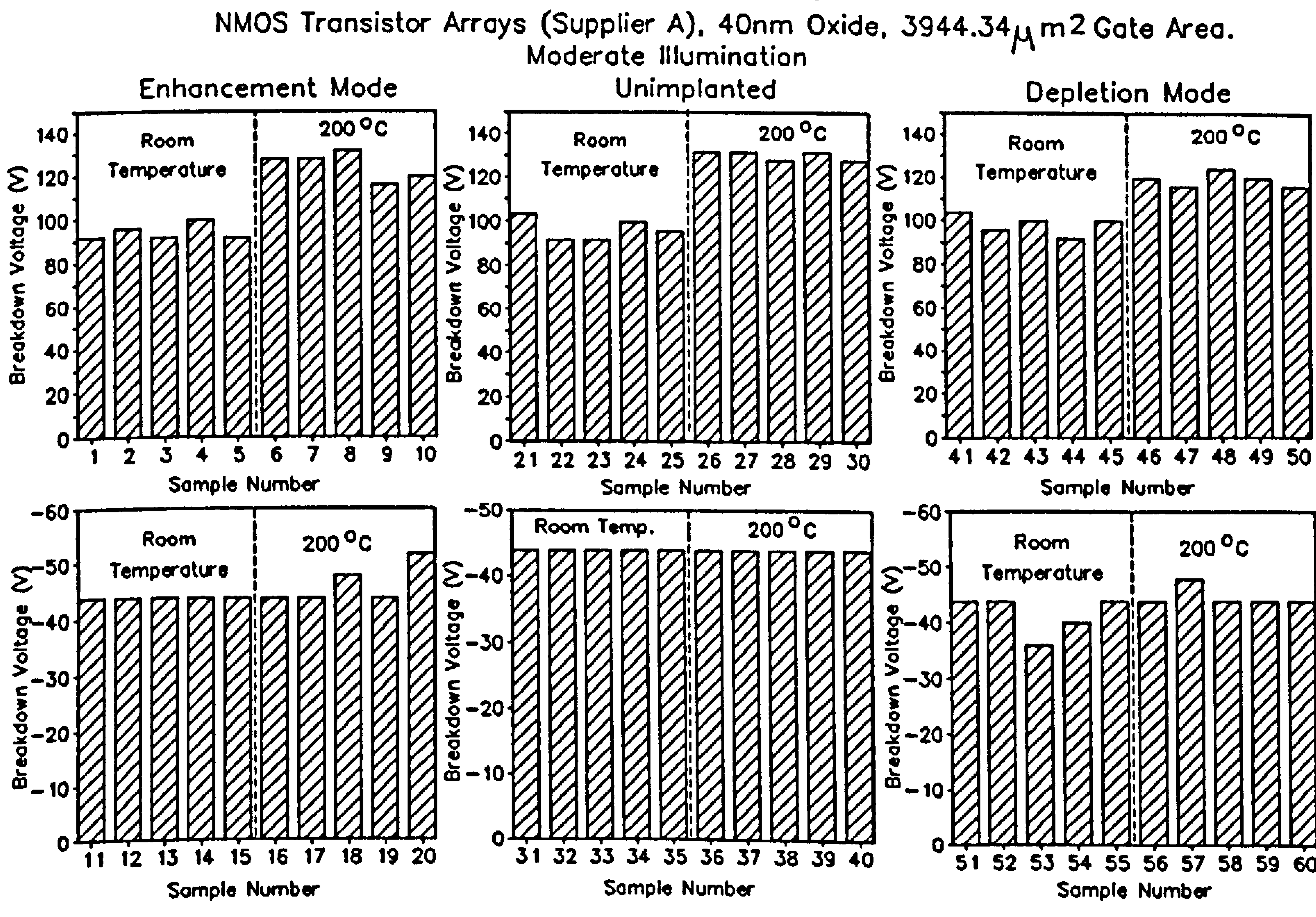


Figure 5.9: Temperature Dependence of ESD Breakdown in NMOS Transistor Arrays.

5.2.3 Temperature Dependence of ESD Breakdown

The ESD breakdown thresholds measured for NMOS transistor arrays at room temperature and 200°C are shown in Fig.5.9. The positive polarity thresholds are clearly higher at the elevated temperature, although the negative thresholds remain largely unchanged. (This temperature dependence of ESD breakdown is significant since it conflicts with the findings of earlier workers [1,2].)

The same experiment was performed using CMOS p and n-channel transistor arrays, and the results are shown in Figs.5.10 and 5.11 respectively. The positive polarity threshold is clearly temperature dependent, although to a lesser degree than in the NMOS structures. As with the NMOS devices, the negative polarity thresholds show no temperature sensitivity.

CMOS Transistor Arrays (Supplier A), 32nm Oxide, 305μm² Gate Area.
P–Channel Devices, Moderate Illumination

Negative Polarity Breakdown Voltage (V)

Room Temperature ← | → 200°C

Sample Number

Positive Polarity Breakdown Voltage (V)

Room Temperature ← | → 200°C

Sample Number

Figure 8.10: Effect of Temperature Upon the ESD Breakdown Thresholds of CMOS P-Channel Transistor Arrays.

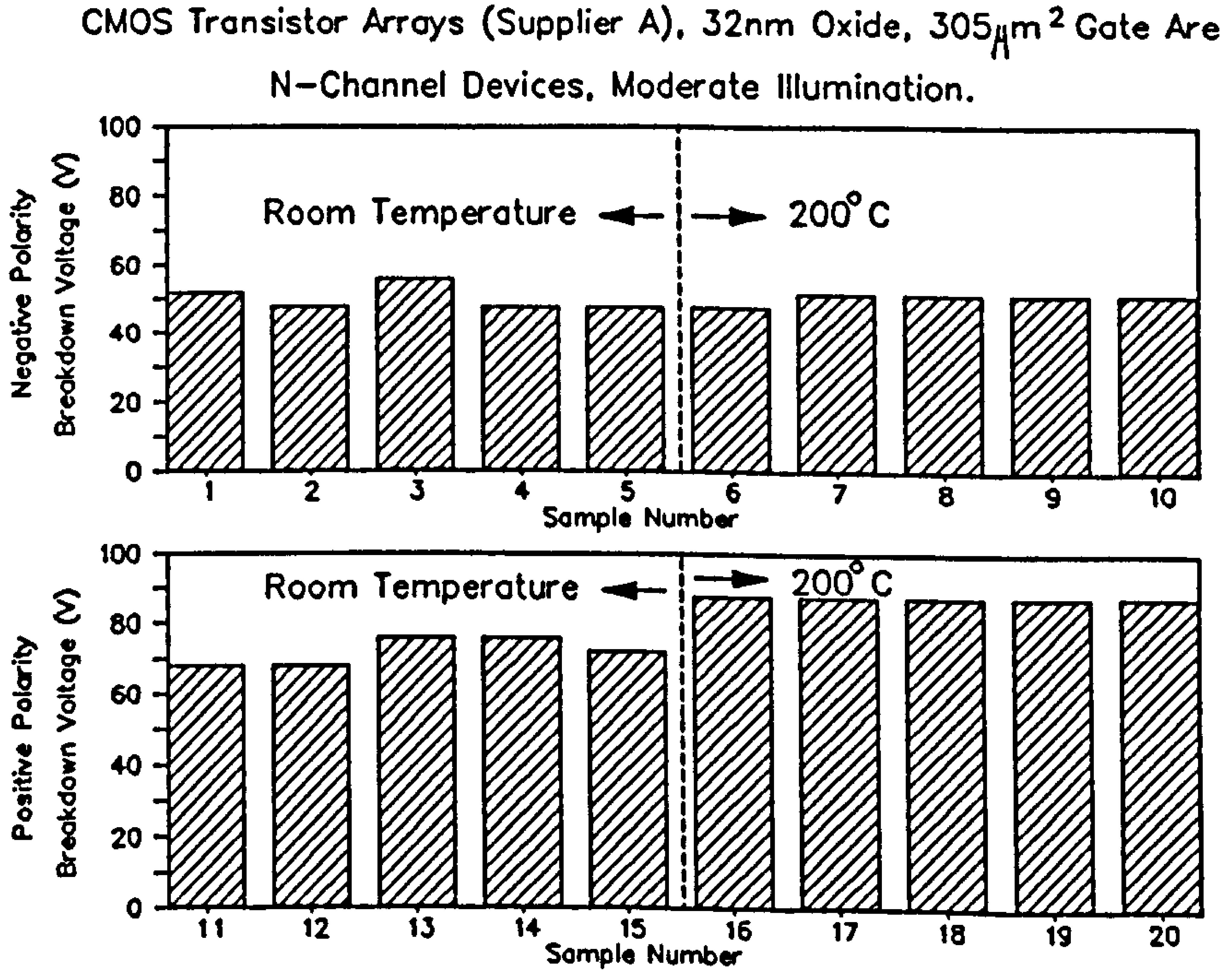


Figure 5.11: Temperature Dependence of ESD Breakdown in CMOS N-Channel Transistor Arrays.

5.2.4 ESD Breakdown as a Function of Luminous Intensity

The room-temperature ESD breakdown thresholds of NMOS transistor arrays were measured under low, moderate and high illumination. Fig.5.12 shows the results of these experiments. The negative polarity breakdown threshold is approximately constant and has a mean values of −44V under low illumination, −42.67V under moderate illumination and −44V under high illumination. However, the positive breakdown threshold was considerably greater under low illumination (mean value 372V) than under moderate illumination (mean value 73.33V) and high illumination (mean value 72.89V).

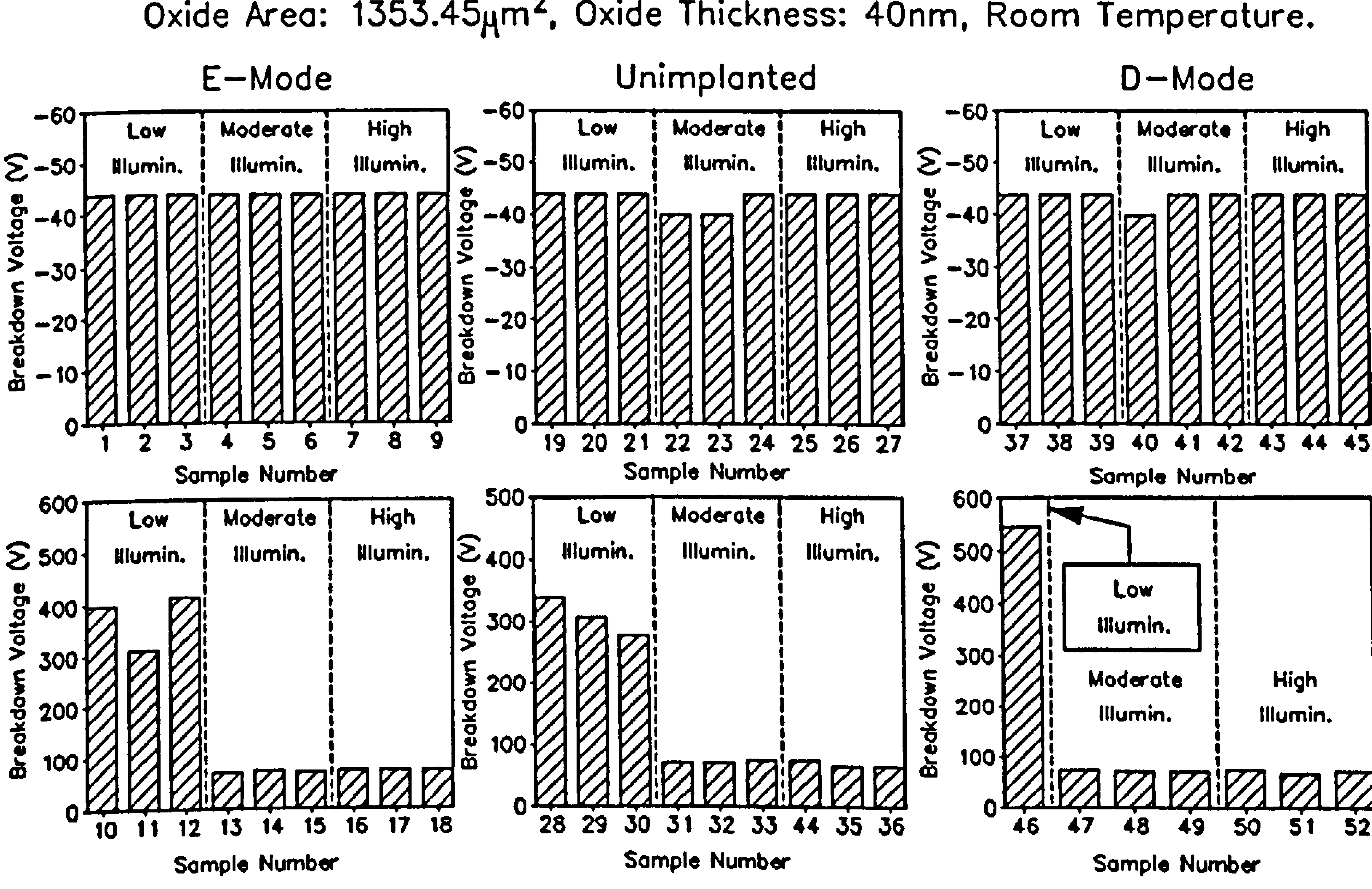


Figure 5.12: ESD Breakdown in NMOS Transistor Arrays as a Function of Luminosity.

Figures 5.13, 5.14 and 5.15 show the results of the same experiments performed using CMOS transistor arrays. These results show that the positive-polarity threshold is approximately constant and has a mean value of 59.6V at low illumination, 60V at moderate illumination and 58.4V under high illumination. The negative polarity thresholds of the n-channel structures also appear to be temperature insensitive, with mean values of −50.5V, −48.8V and −51.2V under low, medium and high illumination respectively. The only significant light-sensitivity appears in the p-channel structures under negative polarity conditions. The mean low-illumination threshold was −136V, while the mean moderate and high illumination thresholds were −50.4V and −48.8V respectively.

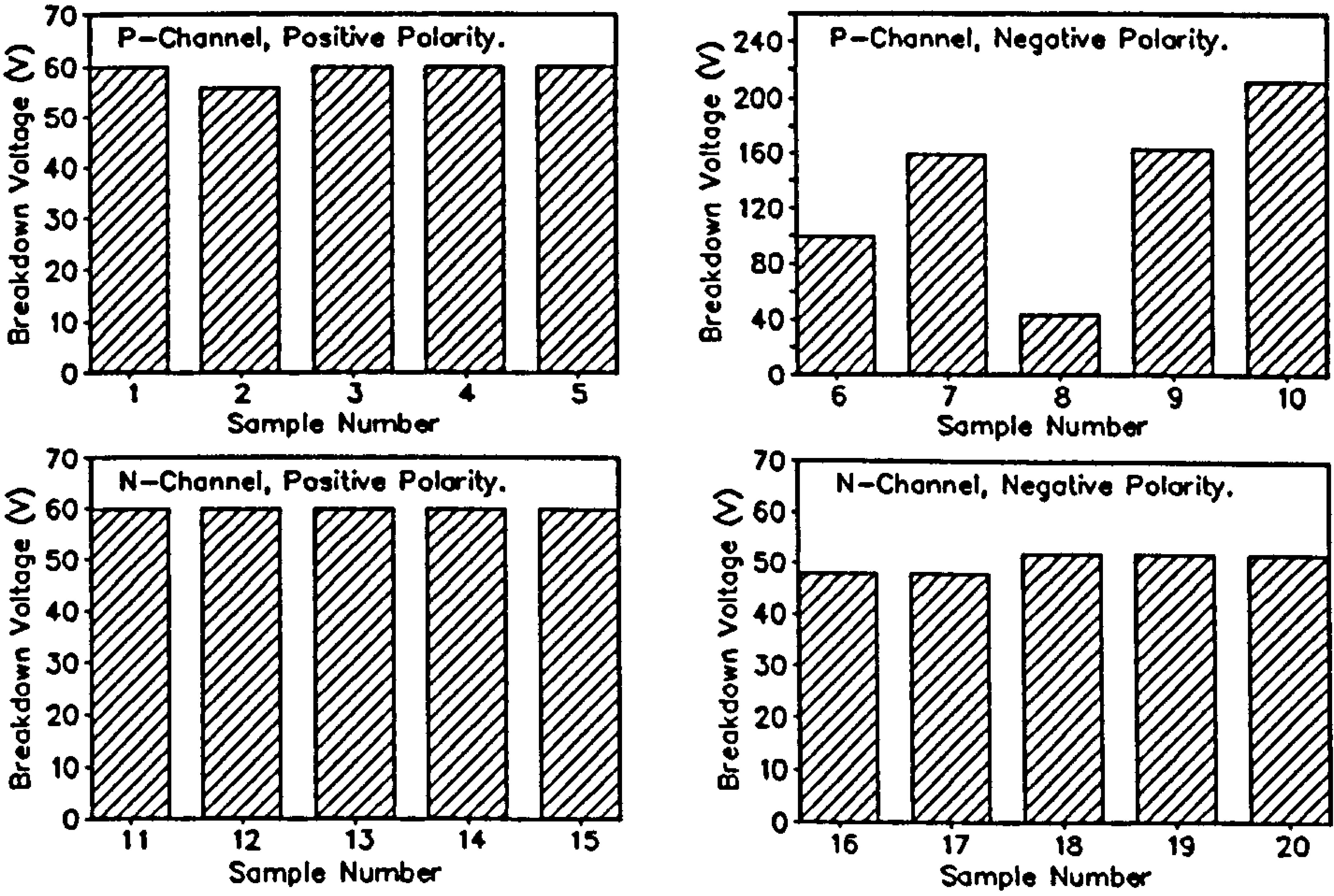


Figure 8.13: ESD Breakdown Thresholds of CMOS P-Channel Transistor Arrays under Low Illumination.

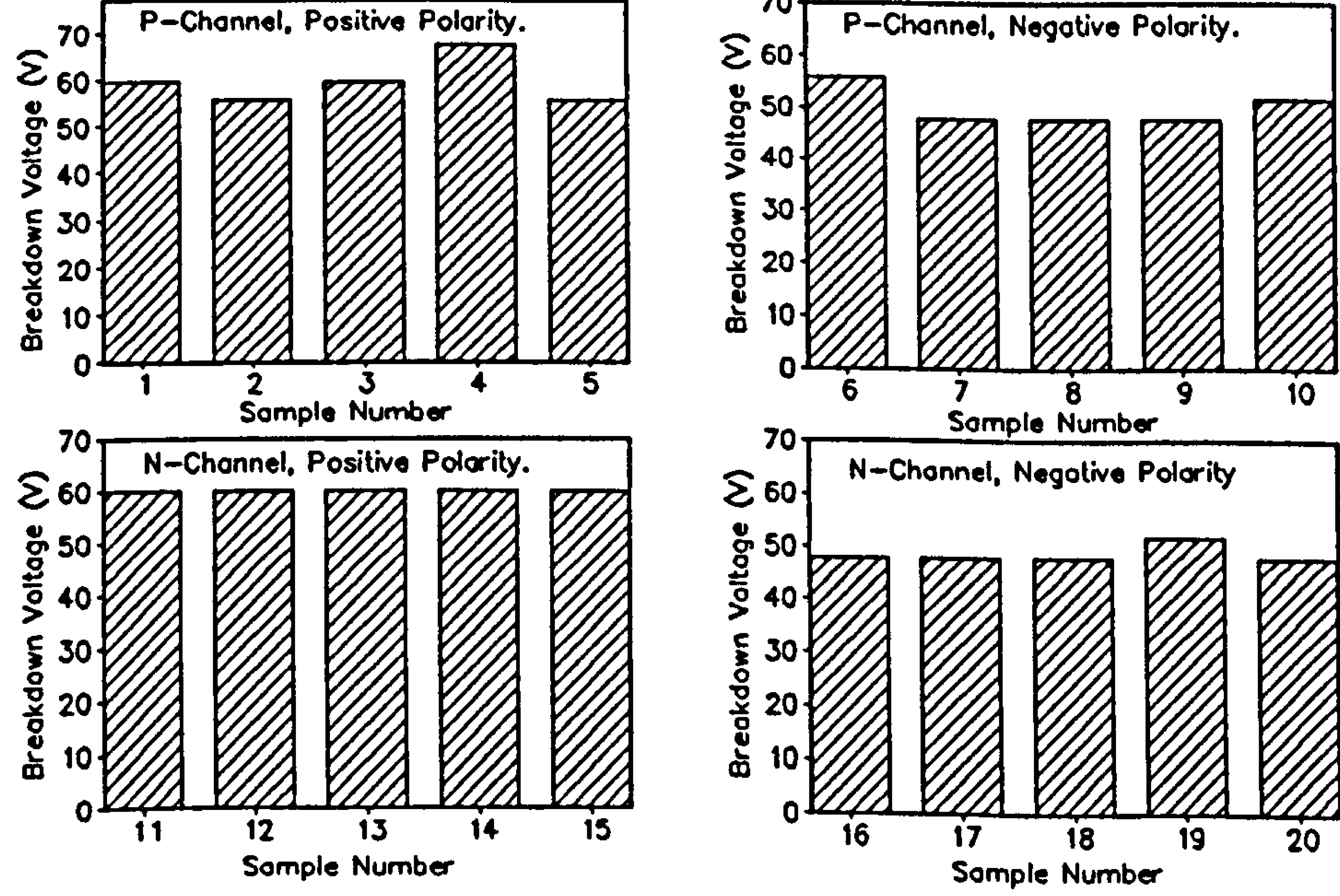


Figure 5.14: ESD Breakdown of CMOS N-Channel Transistor Arrays under Moderate Illumination.

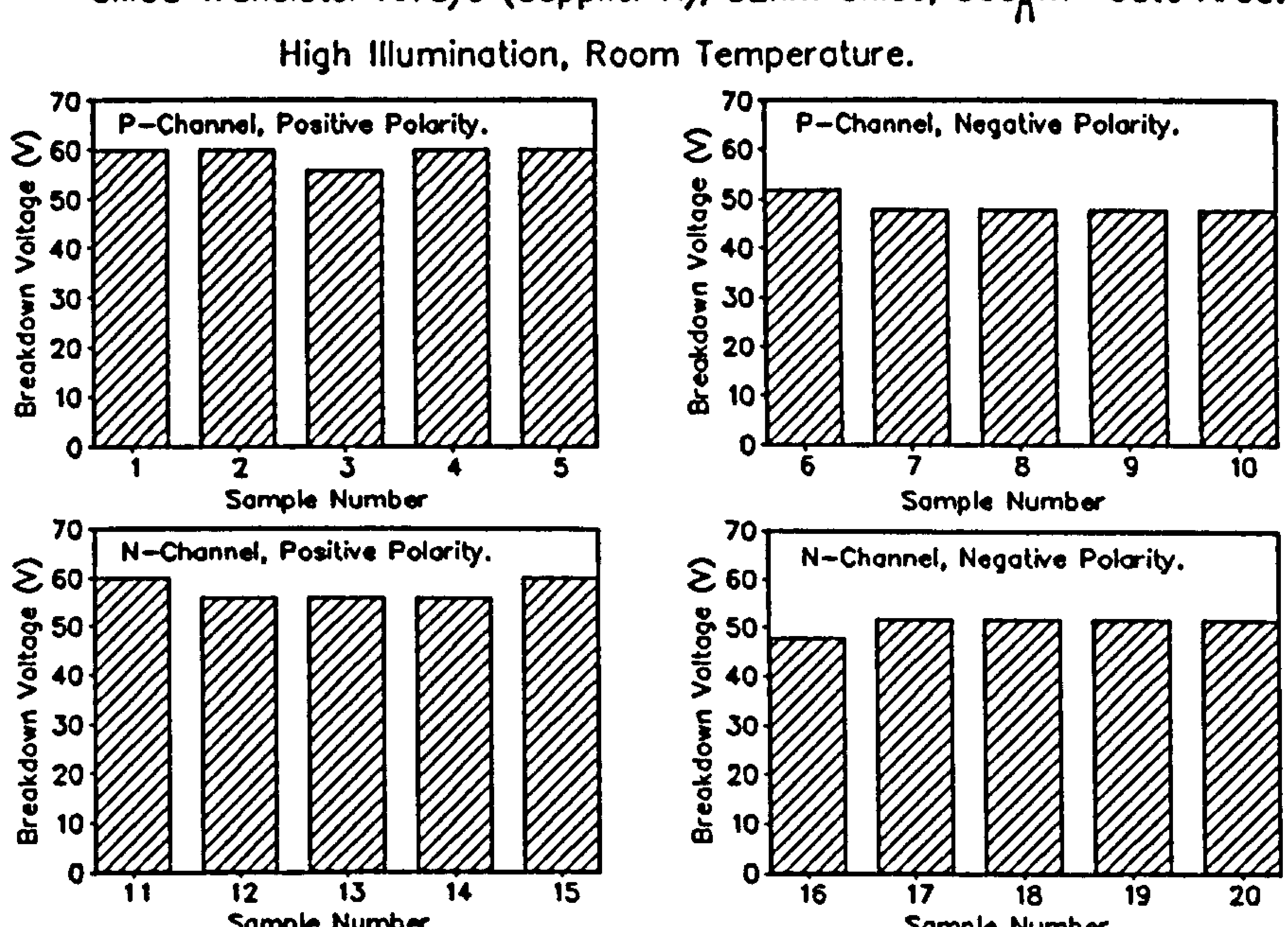


Figure 5.15: ESD Breakdown in CMOS N-Channel Transistor Arrays under High Illumination.

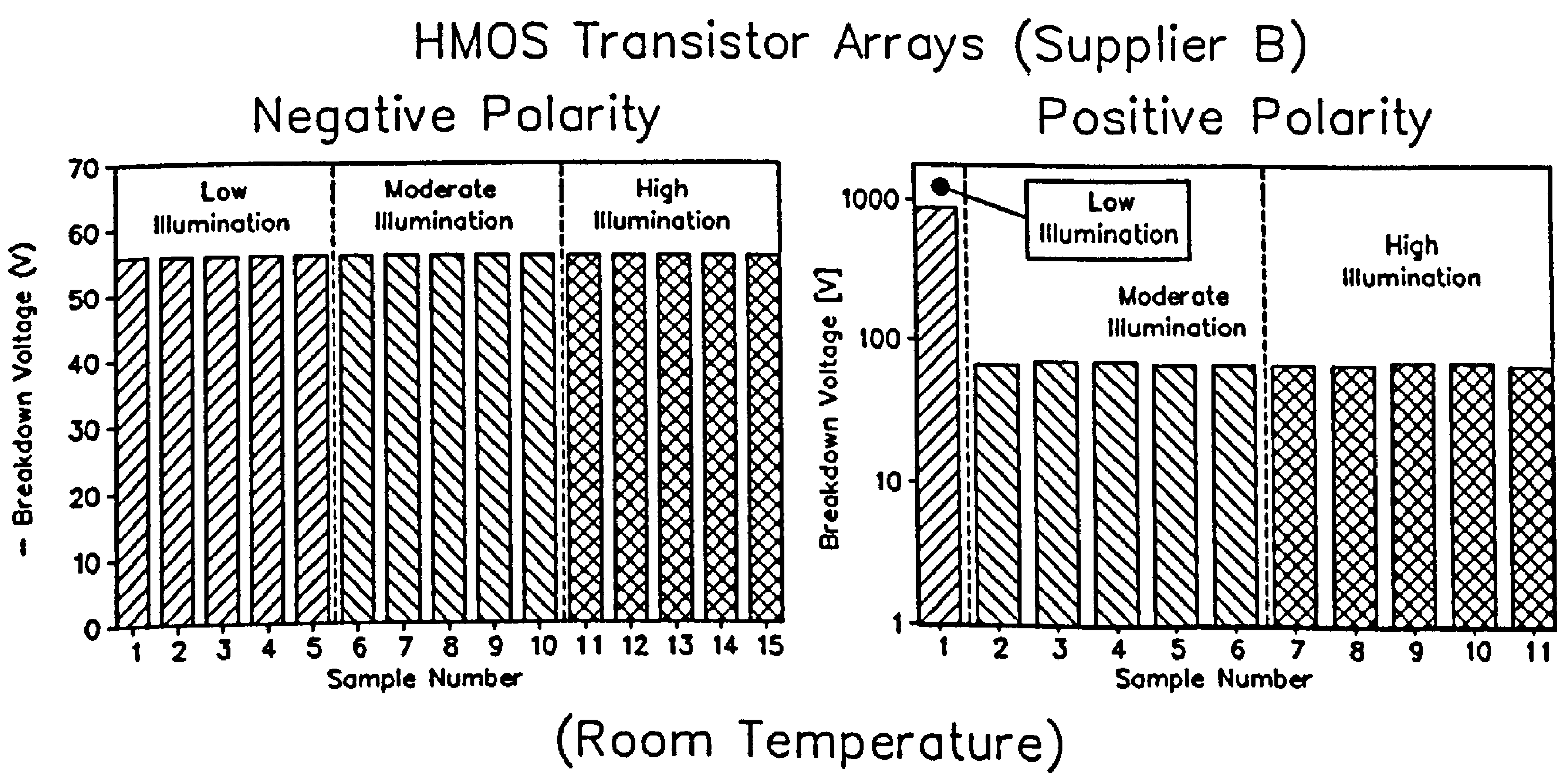


Figure 5.16: ESD breakdown thresholds in HMOS devices under low, medium and high illumination.

Fig.5.16 shows the ESD breakdown thresholds of HMOS transistor arrays measured at room temperature under low, moderate and high illumination conditions. The results are clearly in qualitative agreement with the NMOS results of Fig.5.12.

Fig.5.17(a) shows the positive and negative polarity breakdown thresholds of silicon-

on-sapphire transistor arrays measured under low, moderate and high illumination conditions. (Since no substrate contact existed on the SOS wafers, the devices were stressed between the gate and source terminals.) While the positive threshold is fairly consistent, the negative threshold appears to decrease with increasing illumination, and also shows a significant statistical scatter. Fig.5.17(b) shows the mean breakdown thresholds plotted against the illumination level, together with error bars to indicate ± 1 standard deviation. Since all the negative polarity error bars overlap, it is possible that the apparent trend in the negative data is merely an illusion caused by the statistical fluctuations.

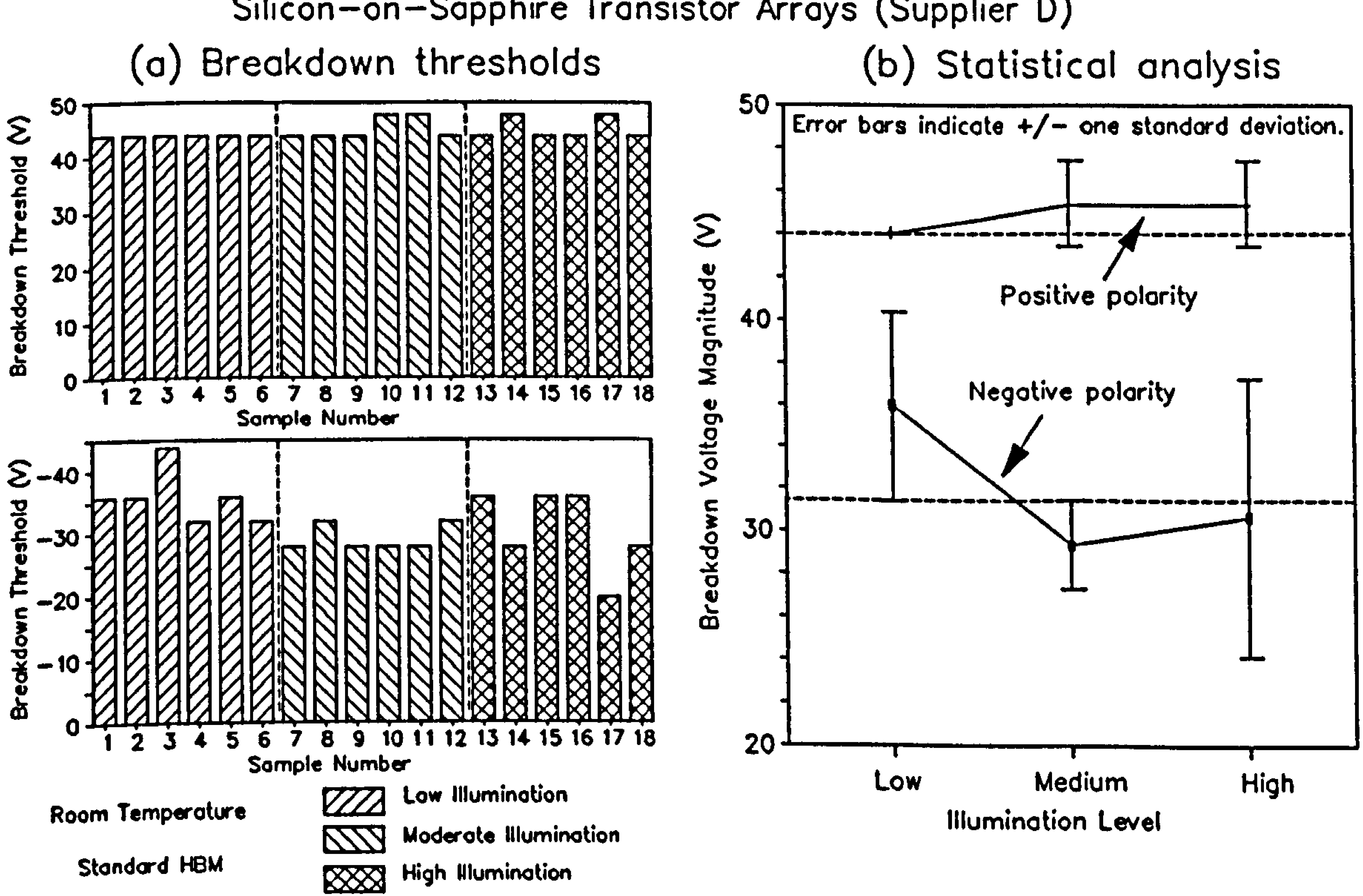


Figure 5.17: ESD breakdown thresholds in SOS devices under low, medium and high illumination.

5.2.5 ESD Breakdown as a Function of Discharge Capacitance

The room-temperature ESD breakdown thresholds of NMOS transistor arrays were measured using different values of discharge (or 'body') capacitance C_1 ranging between 3.24pF and 4.377nF. Fig.5.18 shows the resulting breakdown thresholds plotted against log(C_1). (Note: Breakdown thresholds are plotted against log(capacitance) rather than capacitance merely in order to aid the presentation). The data points indicate that the breakdown voltage magnitude decreases with increasing C_1, in accordance with the results of other workers [3,4]. The results also suggest saturation of V_{bd} for very large and very small capacitances.

The same experiment was performed using CMOS p-channel and n-channel transistor arrays and the results are shown in Figs.5.19 and 5.20. (The curves superimposed on these graphs indicate the data trends.) The graphs show similar features to the NMOS data, ie. V_{bd} decreasing with increasing C_1 and saturating at low and high capacitances. The saturation at low C_1 is more obvious in the positive polarity data (Fig.5.19) than in the negative polarity results (Fig.5.20).

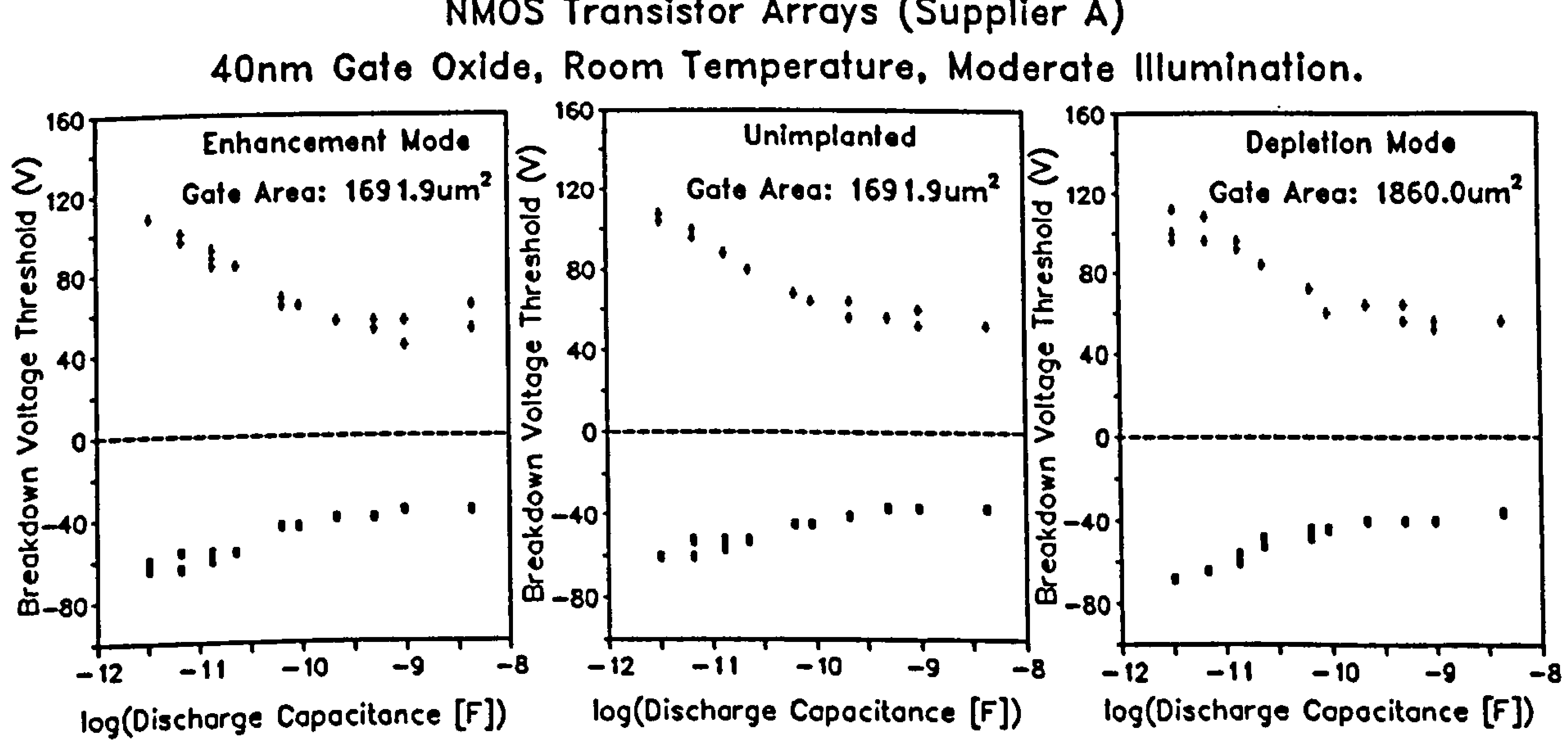


Figure 5.18: ESD Breakdown in NMOS Transistor Arrays as a Function of the AutoZap Discharge Capacitance.

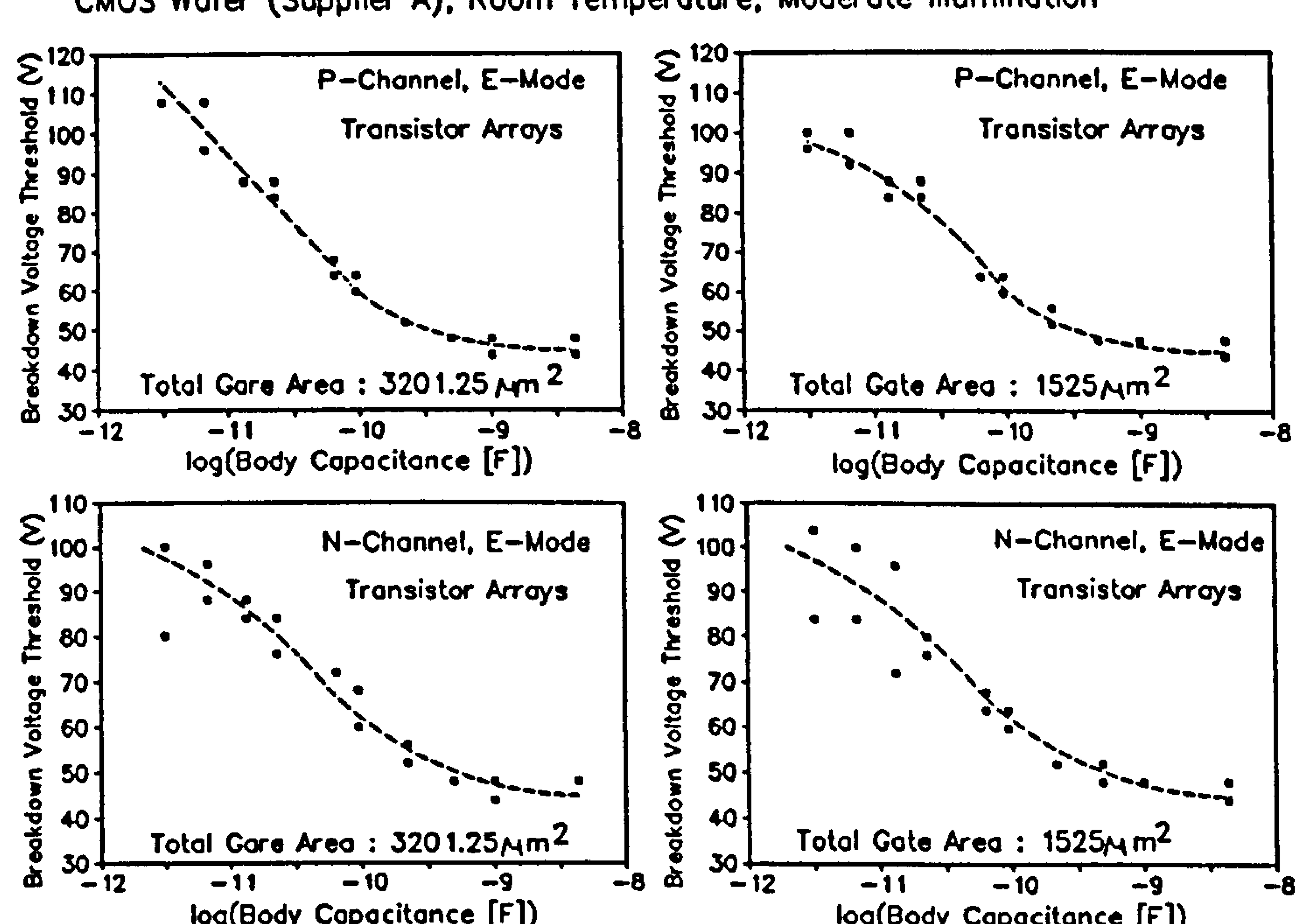


Figure 8.19: Positive Polarity ESD Breakdown in CMOS P-Channel Transistor Arrays as a Function of the Discharge Capacitance C_1.

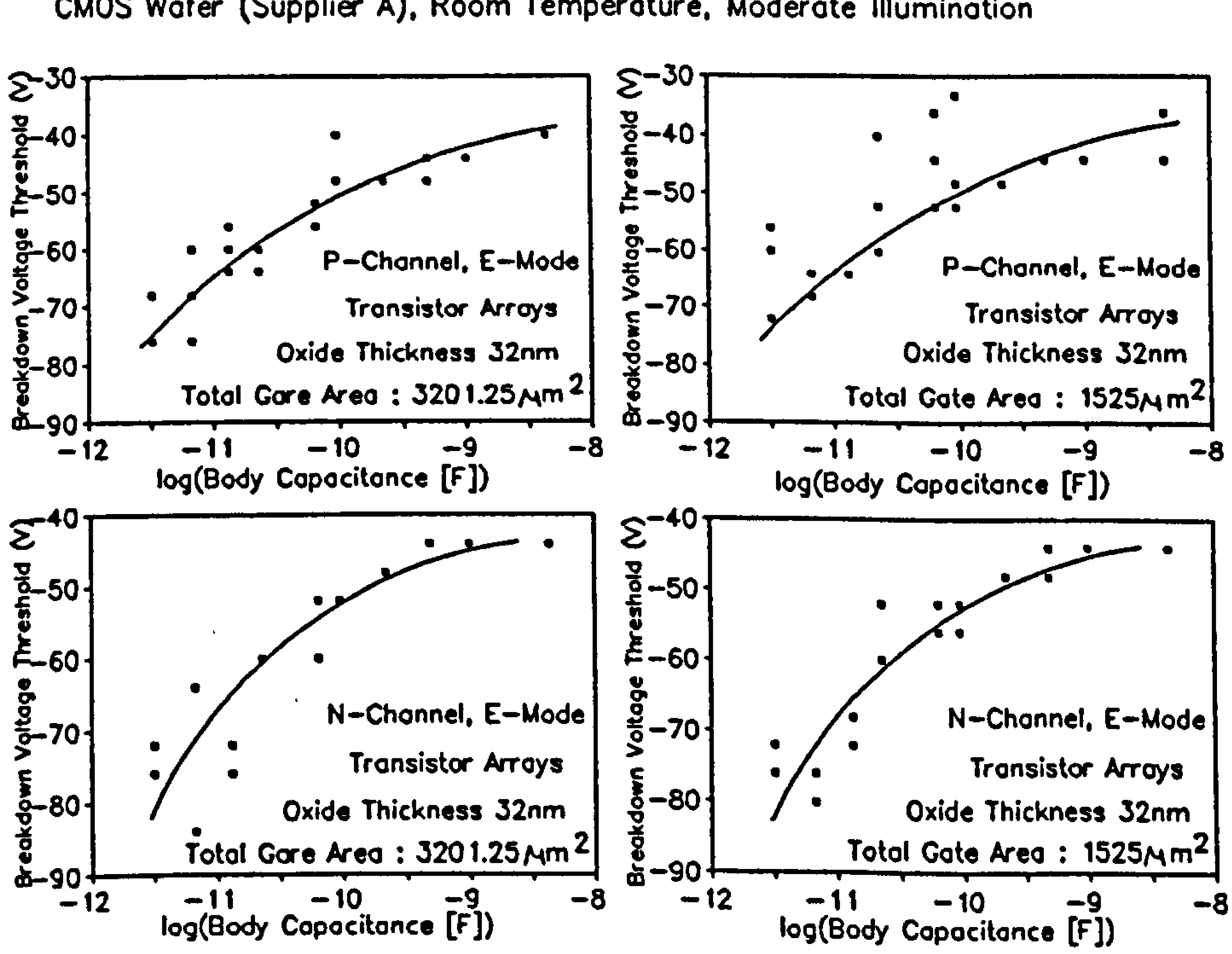


Figure 5.20: Negative Polarity ESD Breakdown of CMOS Transistor Arrays as a Function of Discharge Capacitance.

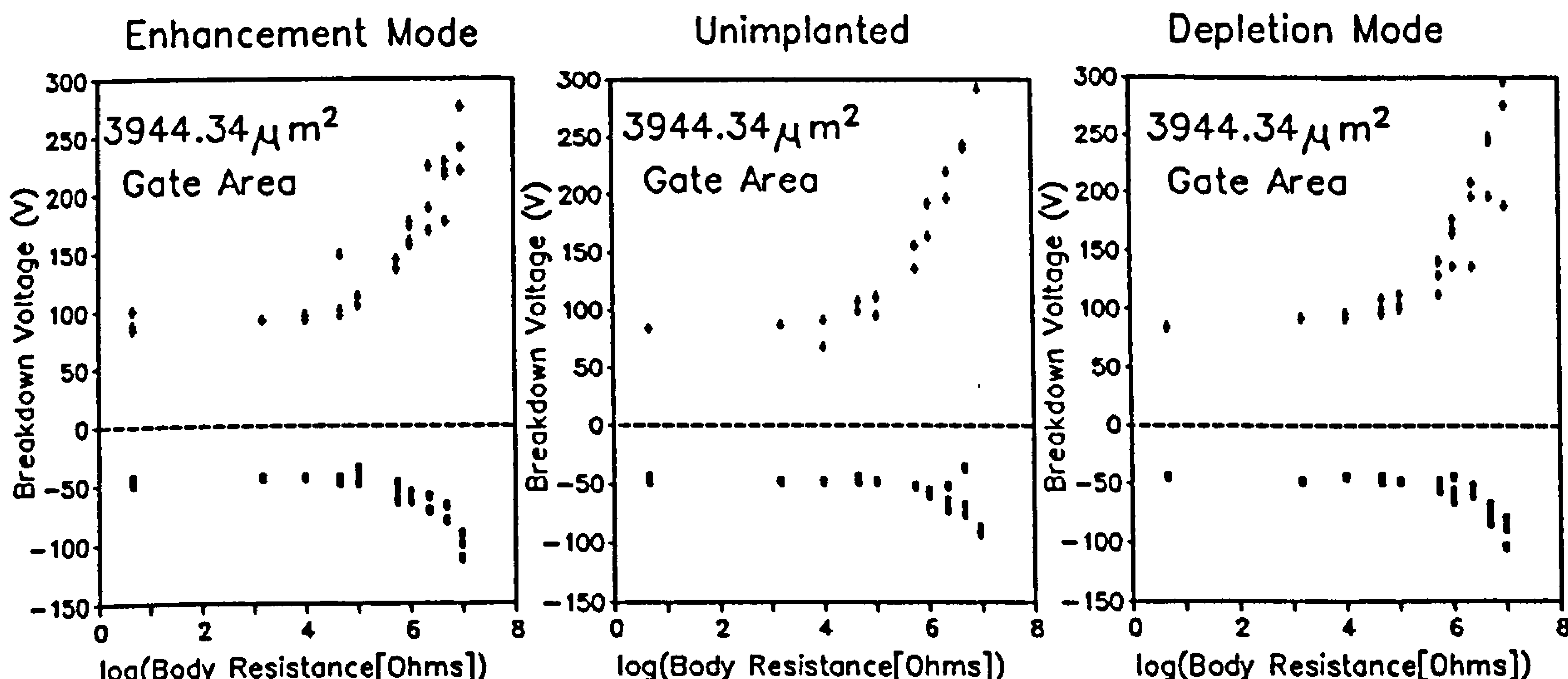


Figure 5.21: ESD Breakdown in NMOS Transistor Arrays as a Function of the Discharge Resistance R_2.

5.2.6 ESD Breakdown as a Function of Discharge Resistance

The room-temperature ESD breakdown thresholds of NMOS transistor arrays were measured using different values of discharge (or 'body') resistance R_2 ranging between 4.67Ω and 9.521MΩ. Fig.5.21 shows the resulting breakdown voltage thresholds plotted against $\log(R_2)$. (Note: Breakdown voltages are plotted against $\log(R_2)$ rather than R_2 merely in order to aid presentation.) For $R_2 < 100K\Omega$ the breakdown voltages are approximately constant, in agreement with earlier studies [4]. Above 100KΩ however, the breakdown voltage magnitude increases rapidly with $\log(R_2)$.

The same experiments were performed using CMOS p-channel and n-channel transistor arrays and the results are shown in Figs.5.22 and 5.23. The curves superimposed on the graphs indicate the data trends. The results are qualitatively similar to the NMOS data of Fig.5.21, although statistical scatter appears to be greater in the CMOS results.

CMOS Wafer (Supplier A), Room Temperature, Moderate Illumination

P–Channel, E–Mode
Transistor Arrays
Gate Area 3201.25μm²

P–Channel, E–Mode
Transistor Arrays
Gate Area 1525.00μm²

N–Channel, E–Mode
Transistor Arrays
Gate Area 3201.25μm²

N–Channel, E–Mode
Transistor Arrays
Gate Area 1525.00μm²

Breakdown Voltage Threshold (V)

log(Discharge Resistance [Ohms])

Figure 5.22: Positive Polarity ESD Breakdown in NMOS Transistor Arrays as a Function of the Discharge (Body) Resistance R_2.

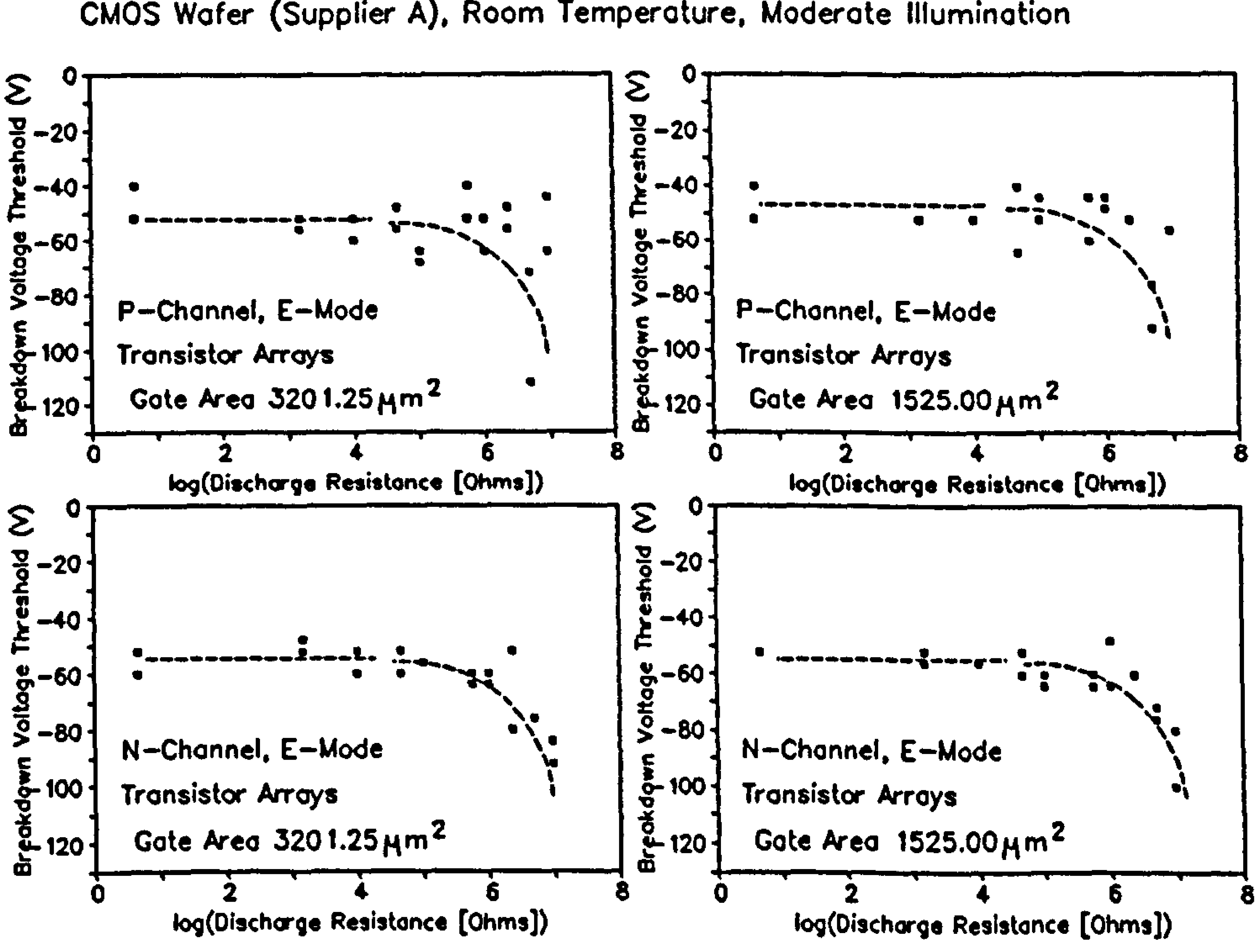


Figure 5.23: Negative Polarity ESD Breakdown in NMOS Transistor Arrays as a Function of the Discharge (Body) Resistance R_2.

4.2.7 Conclusions

The experiments described above represent a systematic investigation of ESD breakdown behaviour in MOS oxides. The study began by verifying the sample uniformity and proceeded to investigate the effects of varying each adjustable parameter. Although most of the devices showed good uniformity, the MOS capacitor structures from Supplier C showed a wide statistical scatter and were therefore excluded from the latter part of the study.

Many of the experiments, principally the temperature and illumination dependent tests, were not really intended to simulate real life conditions (devices are usually enclosed in light-proof packages and are unlikely to encounter human bodies at 200°C). These experiments were intended to investigate the mechanisms governing ESD breakdown, rather than to reproduce actual stress conditions. The significance of the results is examined in Chapter 6.

5.3 Ramp Voltage Experiments

The ramp voltage test is probably the oldest technique used to examine dielectric breakdown and has been used by numerous authors [e.g.8]. Although the voltage is usually increased linearly with time, the voltage 'ramps' used in the current work do in fact follow a $1\text{-}e^{-t/\tau}$ law. However, the techniques used to analyze the data (Section 6.5) are independent of the voltage waveshape. The following sections report the raw data only, the interpretation of the results being postponed to Chapter 6.

5.3.1 Negative Polarity

Negative polarity voltage-ramp experiments were performed using sixteen unimplanted NMOS capacitor structures (see Section 4.3.1), which were labelled Cap.1 to Cap.16 for identification. The apparatus described in Section 4.2.3.3 supplied negative-polarity voltage ramp waveforms, with ramp-rates (dV/dt) of -10, -100, -1000 and -10,000 kV/sec. The voltage and current signals (V_a and V_b) were captured on the HP54111D digital oscilloscope and downloaded to the Walters 286 computer for storage and analysis. Four capacitors were stressed at each ramp-rate, two at room temperature and two at 140°C. Figs.5.24 and 5.25 show the resulting waveforms.

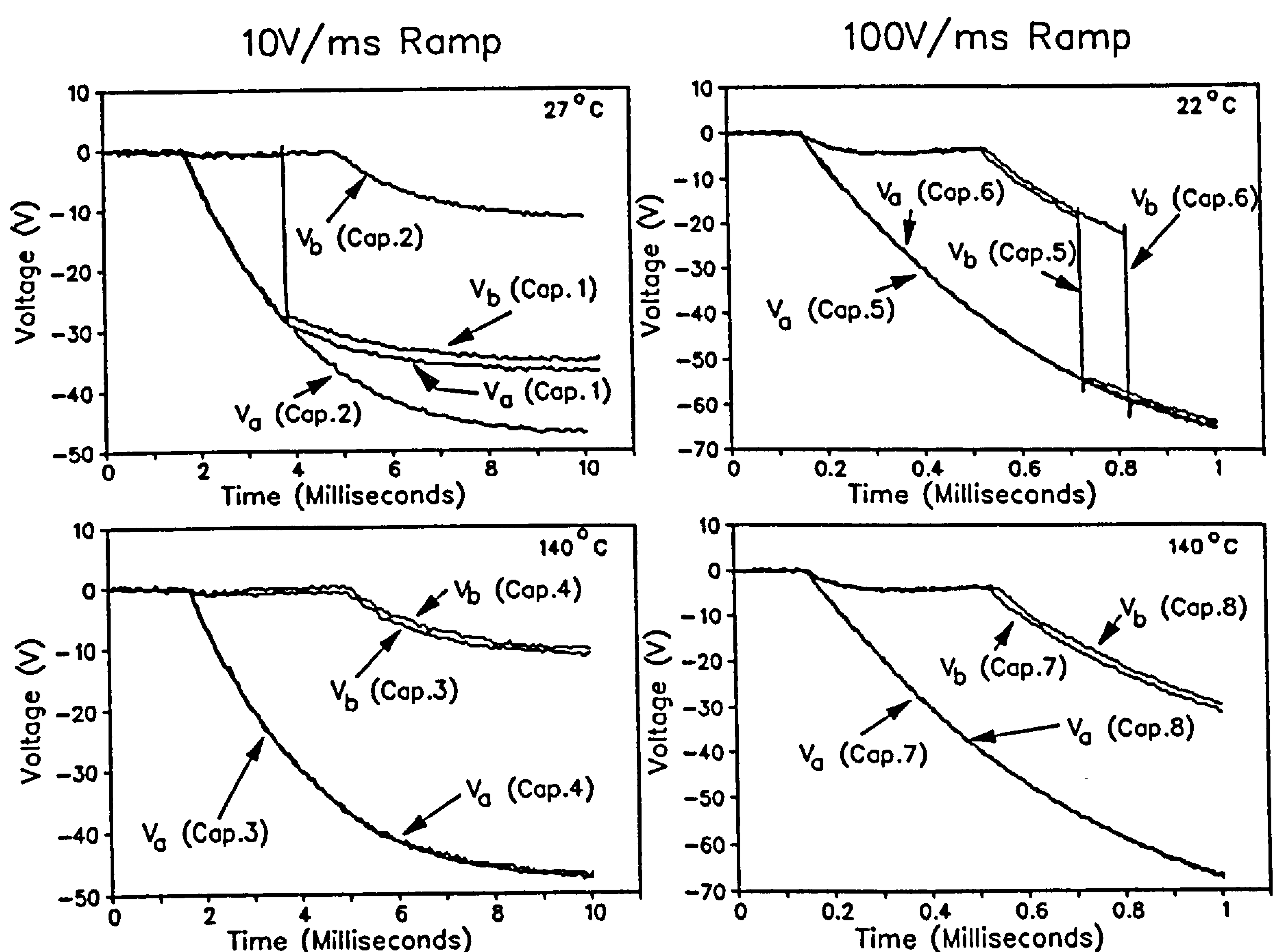


Figure 5.24: Waveforms for 10 and 100V/ms ramp experiments.

5.3.2 Positive Polarity

Positive polarity voltage-ramp experiments were performed using four NMOS capacitor structures, of which two were unimplanted (Caps.17 and 18) and two were E-Mode (Caps.19 and 20). All four structures were subjected to +100kV/sec voltage ramp waveforms. The resulting waveforms are shown in Fig.5.26.

5.4 Constant Voltage Experiments

These experiments investigate the time-dependent nature of breakdown in oxides subjected to constant voltage stress. The techniques used are described in Section 4.2.3.2. NMOS transistor arrays (Section 4.3.1) were stressed between gate and substrate, using a negative gate polarity in order to drive both oxide surfaces into accumulation (such that as much of the applied voltage as possible was dropped across the oxide). The experiments are summarised in the following sections:

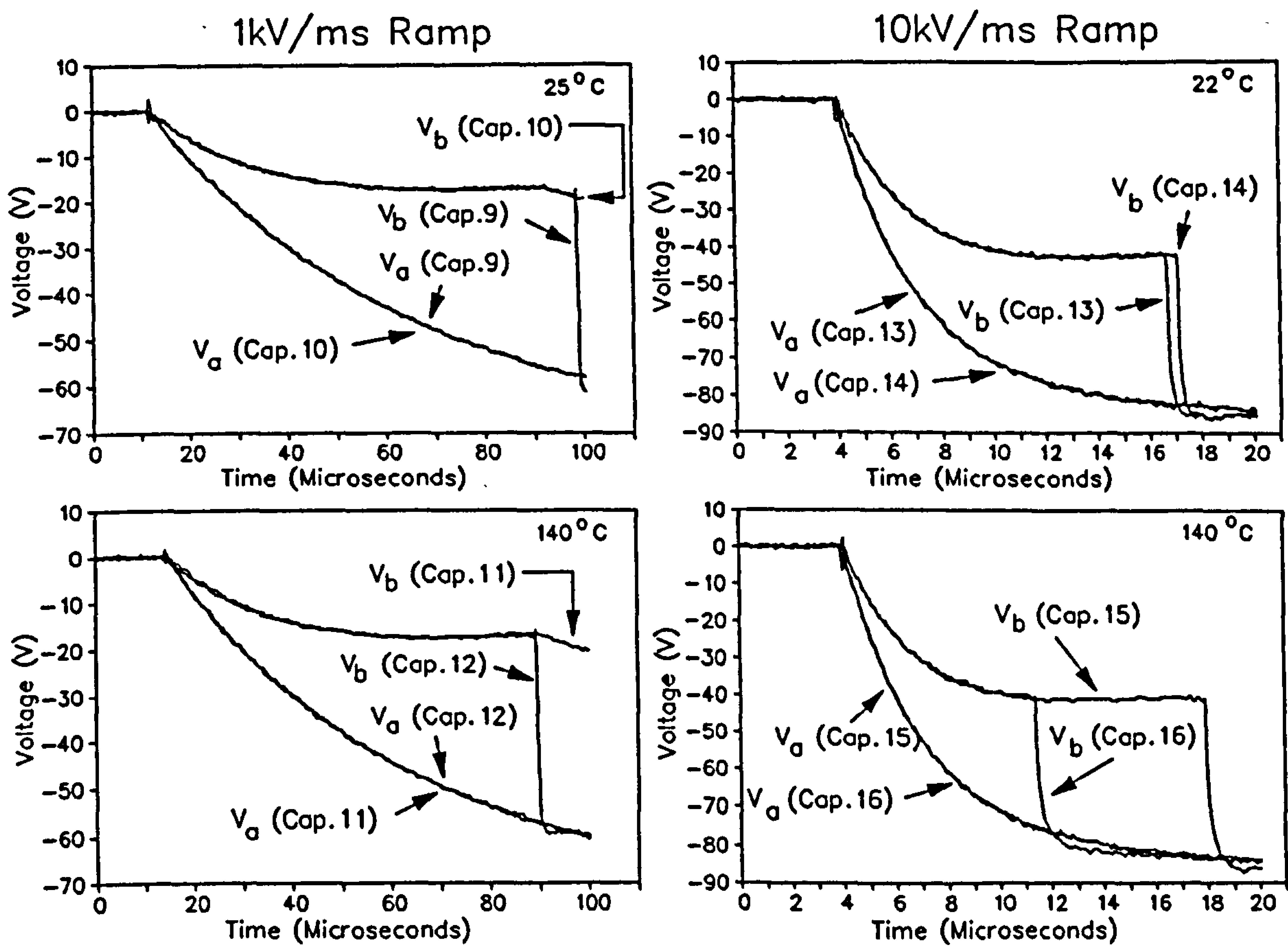


Figure 5.25: Waveforms for 1 and 10kV/ms voltage ramp experiments.

5.4.1 Temperature and Dimensional Dependencies

Constant voltage breakdown was studied as a function of device size and wafer temperature. The times-to-breakdown (t_{bd}) of one hundred and two NMOS transistor arrays under constant -37V gate potential were measured at temperatures between 30°C and 200°C (all under moderate illumination). The test structures were depletion-mode and had gate areas of 1353 μm^2, 3944 μm^2 and 5901 μm^2. The 'medium time-scale' technique (Section 4.2.3.2) was employed throughout the experiments and the test structures were selected from random positions across the wafer area.

Fig.5.27(a) shows all 102 data points plotted in the 'Arrhenius' mode, (ie. log(t_{bd}) vs. 1/T where T is in Kelvin) while Fig.5.27(b) shows the mean and mean$\pm 3\sigma$ values for log(t_{bd}) across the whole temperature range. Since Fig.5.27(b) shows no evidence of any dimensional dependence, the results were treated as a single data set. Although a wide statistical variation exists between devices, the general trend suggests an increasing breakdown time with increasing temperature. According to the Arrhenius equation (see Section 3.3.5), log(t_{bd}) should be a linear function of 1/T. Linear regression analysis [5] gives an optimum straight-line fit of $\log(t_{bd})=3.9159-1577.88/T$ (47.77% correlation coefficient, significant for 99%

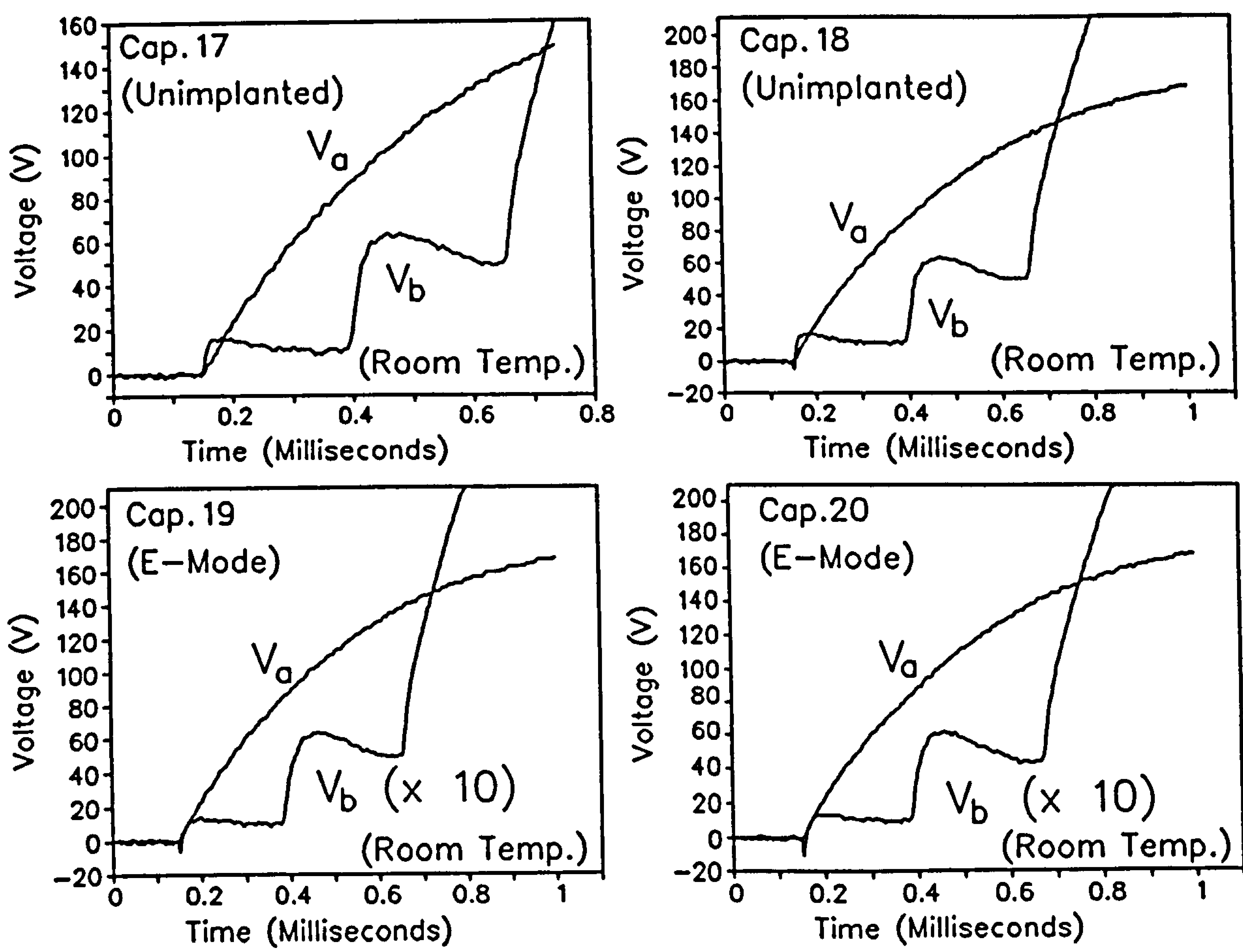


Figure 5.26: Results of positive polarity ramp voltage experiments.

confidence level), implying a *negative* activation energy E_a=-0.3133eV. This seriously conflicts with the results of other workers [e.g.6] and shall be discussed more fully in Chapter 6.

5.4.2 Illumination Dependence

The sensitivity of constant-voltage breakdown to illumination was experimentally investigated. The values of t_{bd} for identical NMOS D-mode transistor arrays under a constant -36V gate voltage were measured under low, medium and high illumination. Eight devices were tested at each illumination level. The 'medium time-scale' technique (Section 4.2.3.2) was employed and all measurements were performed at room temperature. The devices were selected from random positions across the wafer area. Fig.5.28 shows the data, together with the mean and mean$\pm 3\sigma$ values of $\log(t_{bd})$ for the three illumination levels. These results indicate that no significant illumination dependence exists.

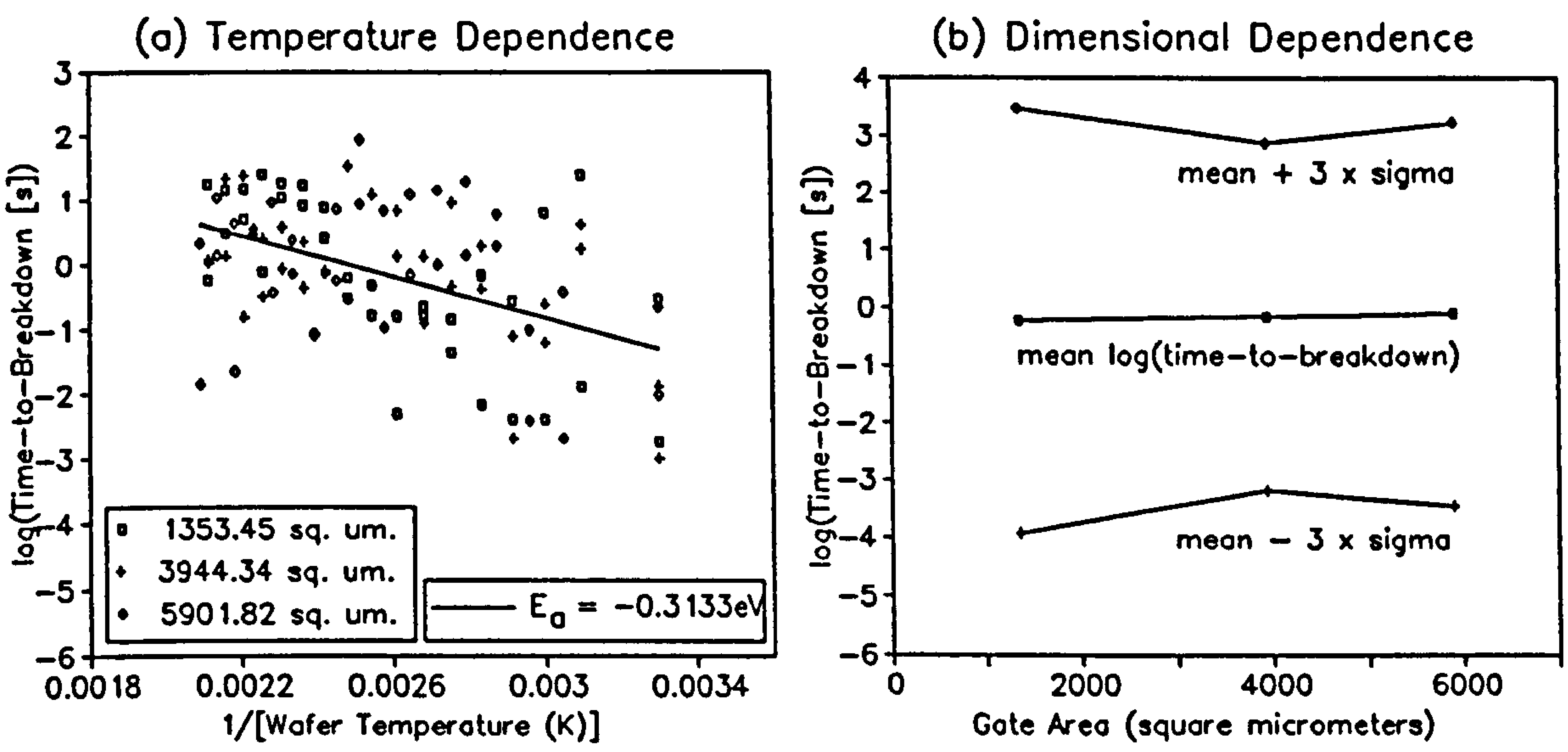


Figure 5.27: Results of temperature/dimensional dependent analysis.

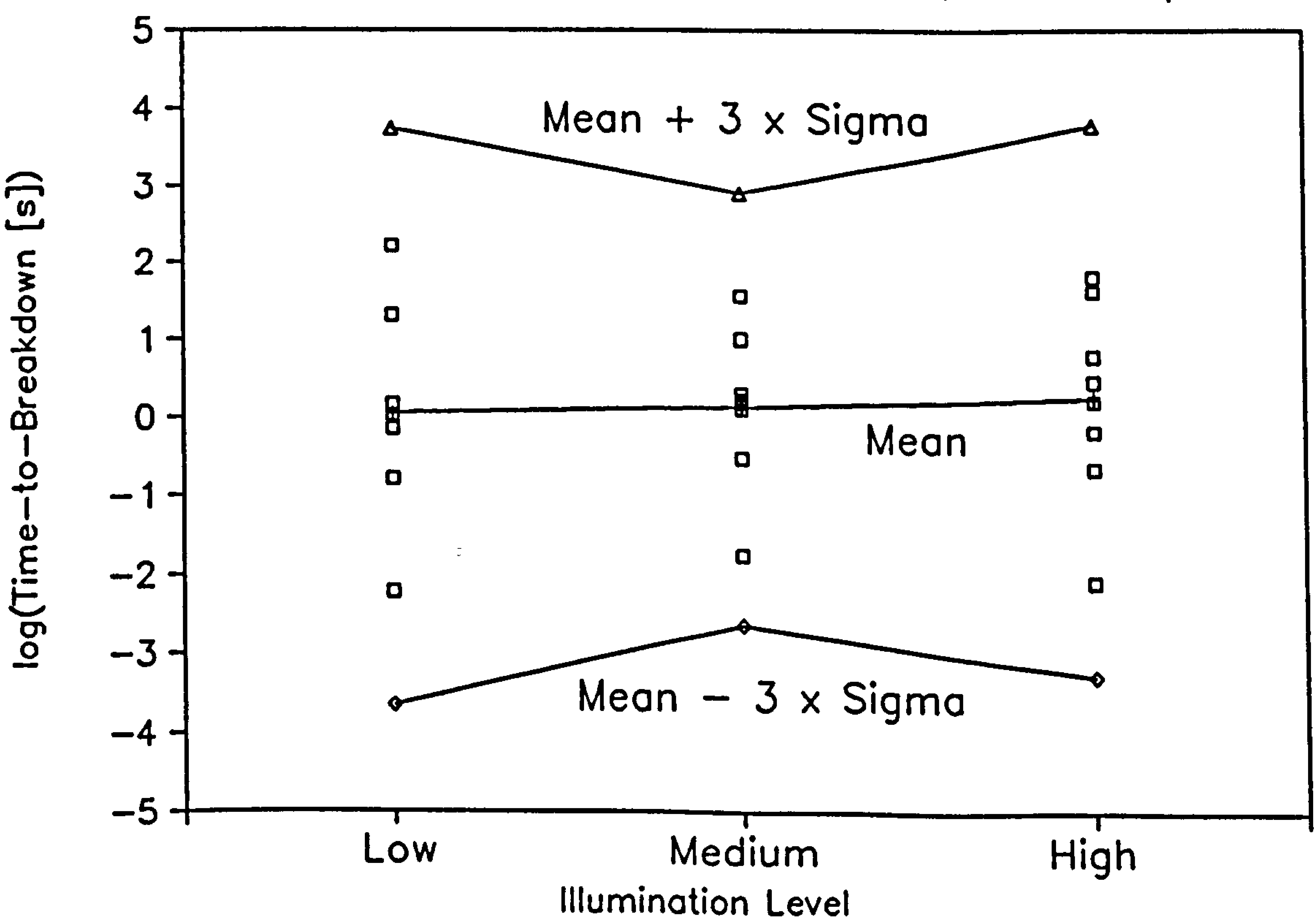


Figure 5.28: Time-to-breakdown as a function of illumination.

5.4.3 Doping Dependence

The relationship between constant voltage breakdown and silicon surface doping was experimentally examined. Values of t_{bd} were measured for NMOS E-mode (p^+-doped), unimplanted (p-doped) and D-mode (n-doped) transistor arrays of identical dimensions. The devices were subjected to -36V gate stress, under moderate illumination and room temperature, using the 'long time-scale' technique (see Section 4.2.3.2). Fig.5.29 shows the data, together with the mean and mean$\pm 3\sigma$ values of $\log(t_{bd})$ for the three doping conditions. These results indicate that the mean value of $\log(t_{bd})$ is not dependent upon the Si surface doping.

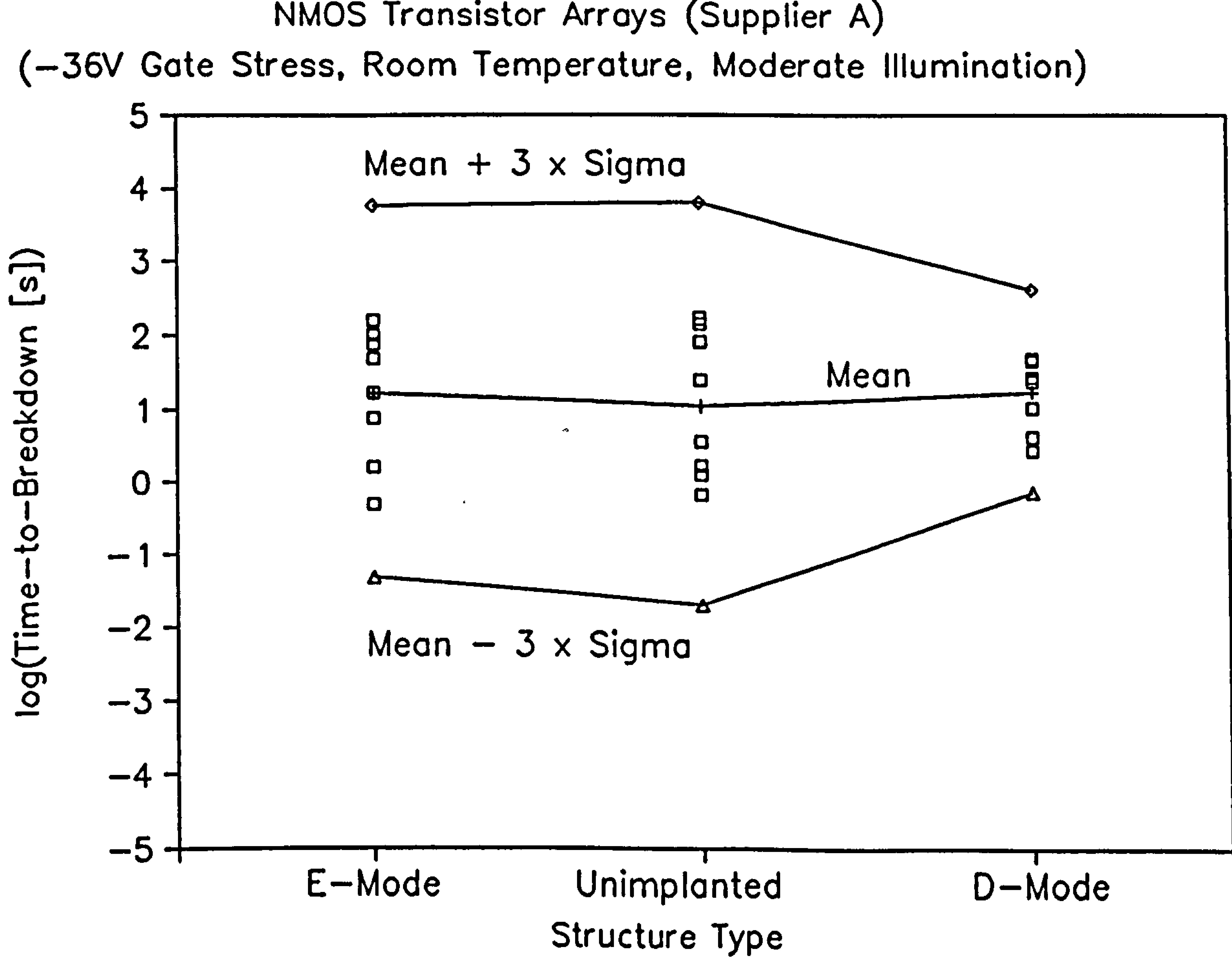


Figure 5.29: Comparative study of constant-voltage breakdown in dimensionally identical (gate area: 1353.45μm^2, gate oxide thickness: 40nm) E-mode, D-mode and unimplanted transistor arrays.

5.4.4 Electric Field Dependence

Constant voltage breakdown was examined as a function of the applied electric field. Values of t_{bd} were measured for NMOS E-mode transistor arrays under constant gate voltages between -34 and -46V. All three techniques of Section 4.2.3.2 were employed, allowing t_{bd} to be measured over more than ten orders of magnitude. Since the dimensional independence was established in Section 5.4.1, the devices were selected at random, irrespective of their gate areas. Fig.5.30 shows the results, with log(t_{bd}) plotted against the applied field.

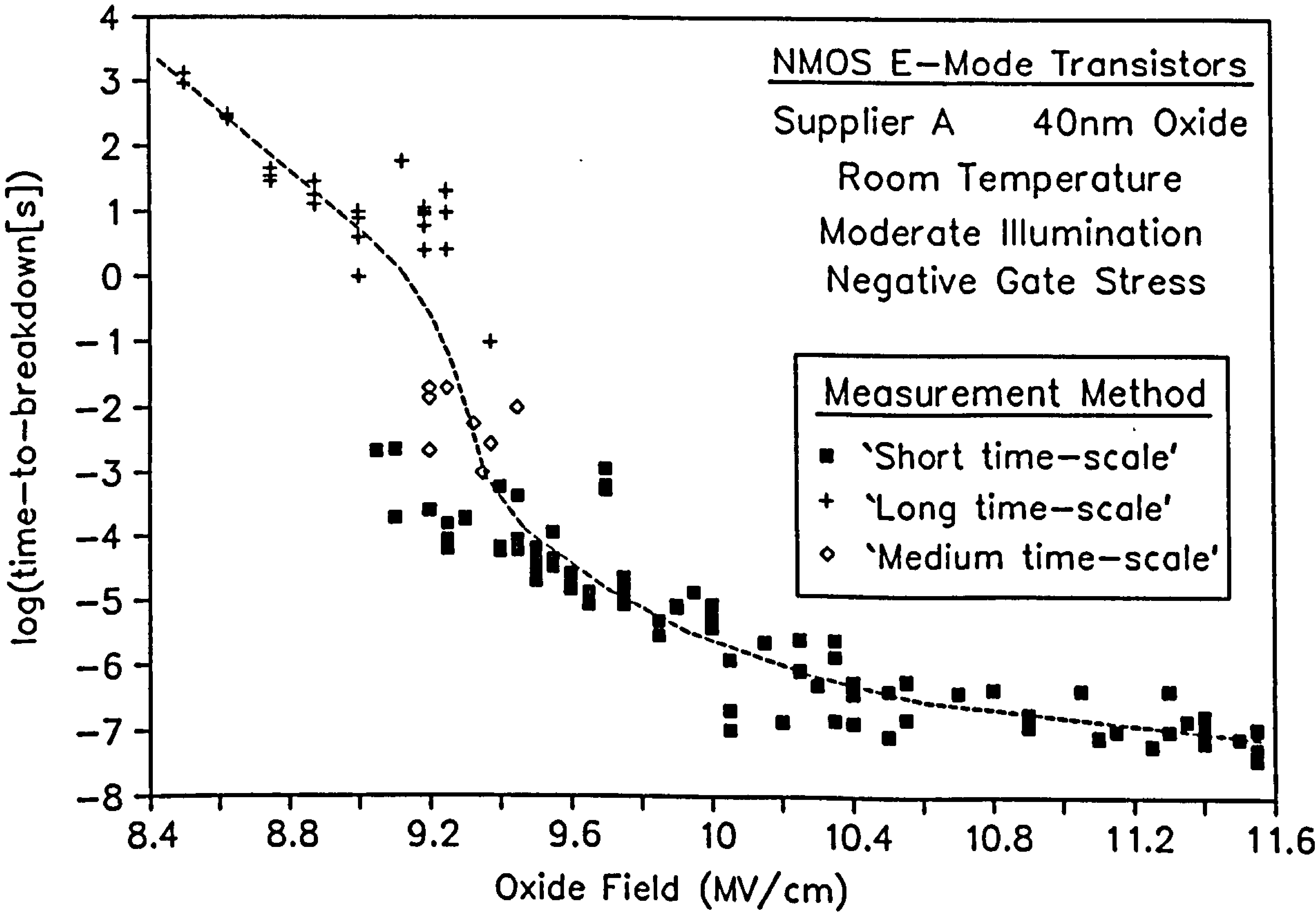


Figure.5.30: Time-to-breakdown vs. Field data obtained from NMOS transistor arrays [Field = Applied Voltage/Oxide Thickness (40nm)] .

5.5 Constant Current Injection Experiments

Constant current stress was generated using the HP4145B parametric analyzer in constant-current mode (see Section 4.2.3.4) and injected into the MOSFET gate oxides under the Fowler-Nordheim mechanism [7]. Constant current breakdown was characterised in terms of the injected charge-to-breakdown $Q_{bd} = I.t_{bd}$. As with the constant voltage experiments, the transistor gates were stressed negatively with respect to the substrate.

5.5.1 Dimensional Dependence

Constant current charge-to-breakdown was investigated as a function of the device gate area. A constant current of 0.5μA was injected into E-Mode NMOS transistor arrays with gate areas ranging between 1353.45 and 6831.3μm^2. Fig.5.31 shows the results plotted as log(Q_{bd}) against log(Area). The least-square linear fit with near-unity slope indicates that the areal injected-charge density Q_{bd}/Area (rather than the absolute charge Q_{bd}) is the critical parameter. This adds weight to the long-held belief that wearout is driven by the current, injected homogeneously across the oxide area [8].

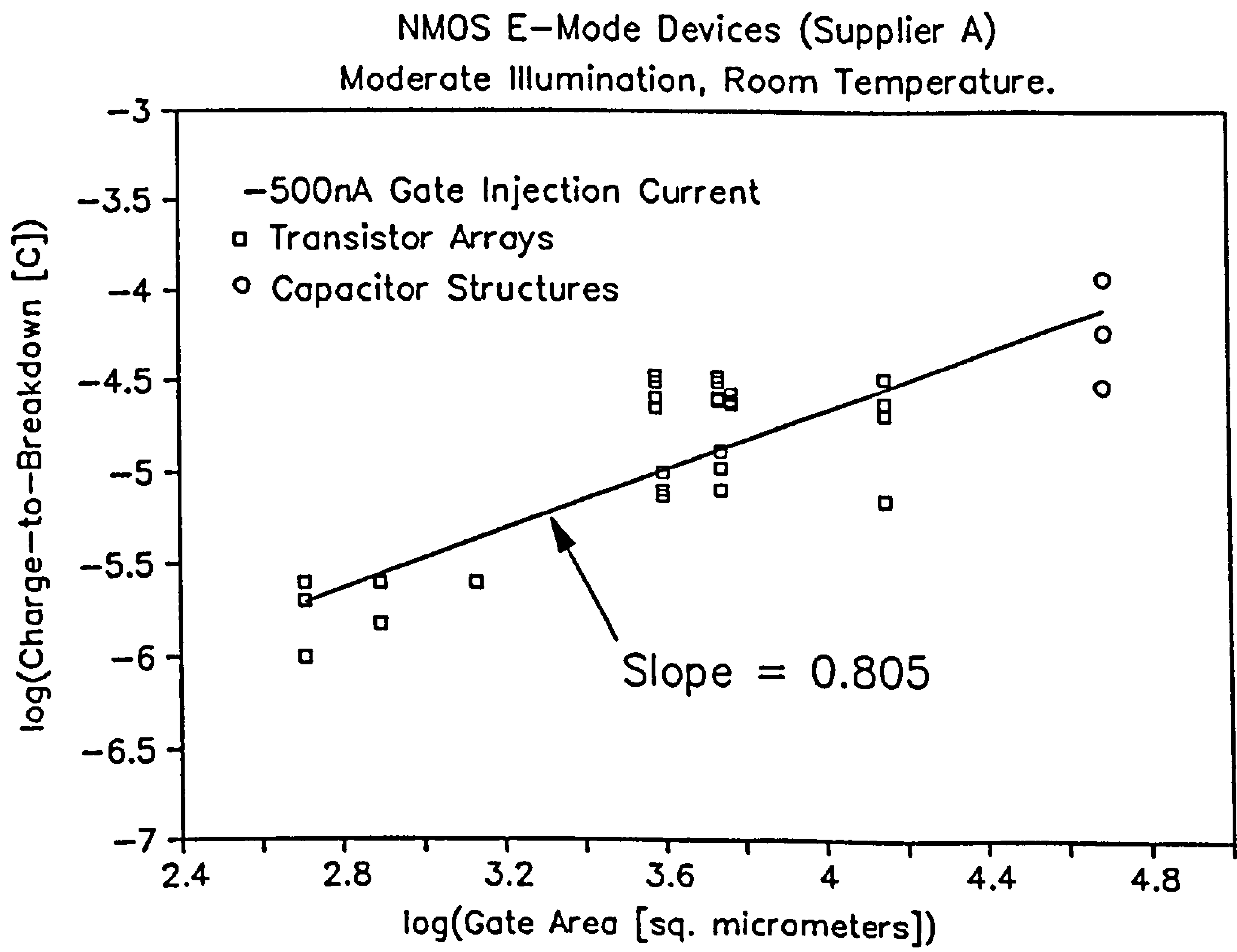


Figure 5.31: Constant-current charge-to-breakdown as a function of gate area.

5.5.2 Current Dependence

Constant current charge-to-breakdown was studied as a function of the injection current. Currents ranging between 1μA and 60μA were injected into identical NMOS E-Mode transistor arrays under moderate illumination and room temperature. Fig.5.32 shows the results plotted as log(Q_{bd}) vs. log(I). Below 25μA, Q_{bd} decreases slowly with increasing current while above 25μA it falls dramatically. These regions clearly correspond to the 'ductile' and 'brittle' failure modes identified by Wolters et al. [8], while 25μA represents the 'critical' current I_{cr} (see Fig.3.8 in Chapter Three).

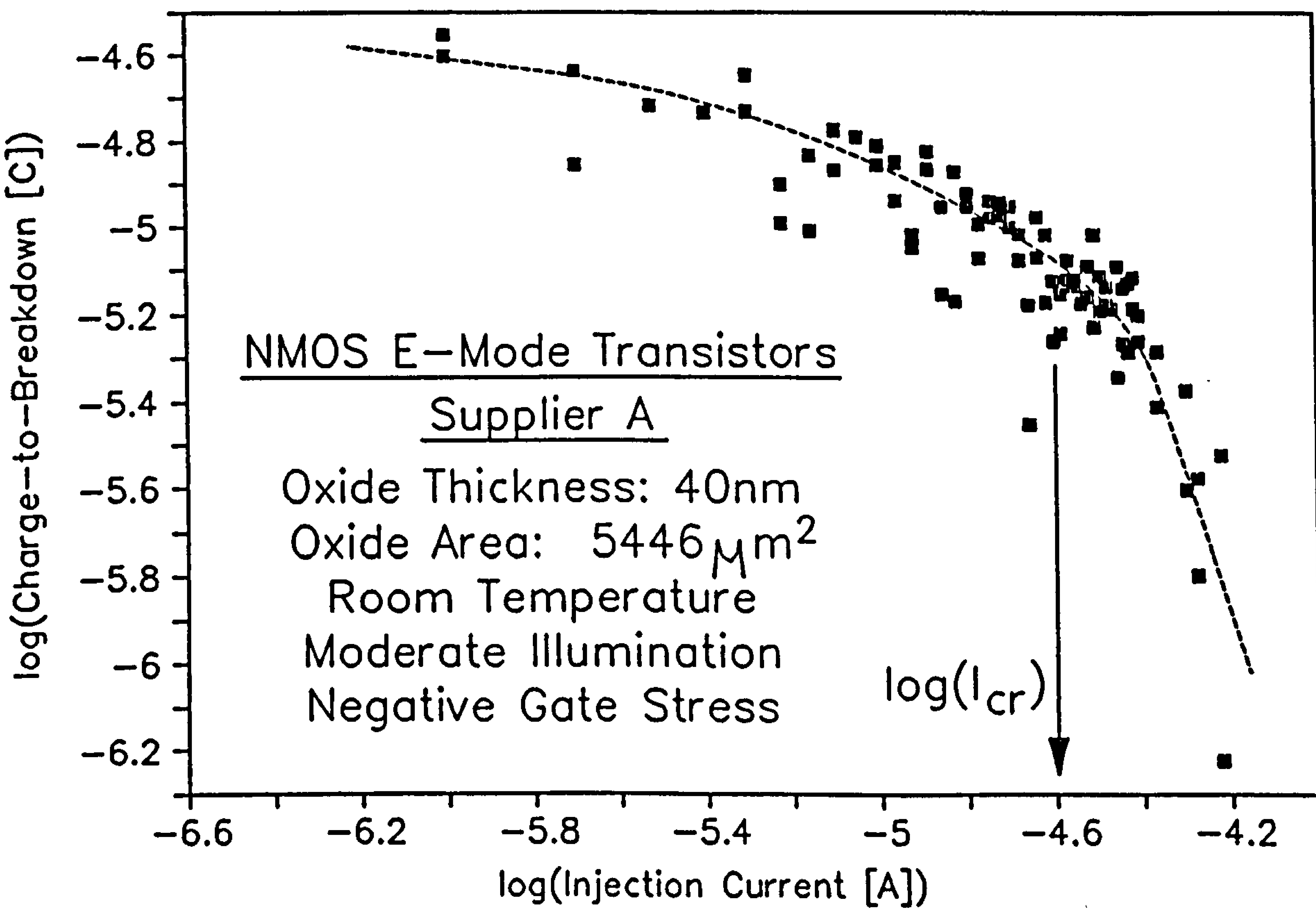


Figure 5.32: Constant current charge-to-breakdown as a function of current.

5.6 Summary

1. The uniformity of the experimental samples was studied in terms of their ESD breakdown thresholds. Tests were performed on the full range of experimental samples described in Chapter 4 (i.e. NMOS, CMOS, HMOS, MOS Capacitors & SOS). With the exception of the wide-area capacitor structures (Supplier C), the breakdown voltages were found to be reasonably consistent between wafers and between different positions on the wafers.

2. The relationship between ESD breakdown voltage and gate-oxide area in NMOS and CMOS devices was investigated. Although the negative polarity breakdown voltage remains constant, the NMOS positive polarity threshold is generally larger in the larger area devices.
3. ESD breakdown was investigated as a function of wafer temperature. Although the negative polarity breakdown voltages are temperature insensitive, the positive polarity thresholds increase with increasing temperature.
4. ESD breakdown was studied as a function of luminous intensity. In both NMOS and HMOS devices, the negative polarity breakdown voltages are light-insensitive. However, the positive polarity thresholds are considerably larger under low illumination than under moderate and high illumination. The reverse situation exists in CMOS devices, the positive threshold being light insensitive and the negative threshold increasing rapidly as illumination falls. ESD breakdown in SOS transistor arrays showed no light sensitivity whatsoever.
5. The sensitivity of ESD breakdown to the 'discharge' or 'body' capacitance C_1 was experimentally studied. The breakdown voltage magnitude generally decreased with increasing C_1, saturating towards constant values at low and high capacitance.
6. The sensitivity of ESD breakdown to the 'discharge' or 'body' resistance R_2 was experimentally studied. The breakdown voltage magnitude remained approximately constant for $R_2 < 100K\Omega$ and increased rapidly for $R_2 > 100K\Omega$.
7. Breakdown of NMOS capacitor structures under ramp-voltage stress was investigated. Negative polarity breakdown was studied as a function of the ramp rate *(dV/dt)* and wafer temperature in unimplanted capacitor structures. Positive-polarity ramp breakdown was examined in NMOS unimplanted and D-Mode devices. The resulting waveforms were recorded for later analysis.
8. Constant-voltage oxide breakdown was studied in NMOS transistor array structures. Time-to-Breakdown t_{bd} was found not to depend on device size, illumination or wafer temperature. However, t_{bd} was found to be a strong function of electric field, varying over ten orders of magnitude between 8.4 and 11.6MV/cm.
9. Constant-current oxide breakdown in NMOS transistor and capacitor structures was studied as a function of device size. The injected charge-to-breakdown Q_{bd} for a given current is approximately proportional to gate area, suggesting that the areal charge density and injection current density are the dominant parameters.
10. Constant-current oxide breakdown was examined as a function of the injection current. The resulting $\log(Q_{bd})$ vs. $\log(I)$ curve exhibits the same brittle/ductile transition identified by Wolters et al. [6].

5.7 References

1. Amerasekera,E.A., Campbell,D.S., *"Electrostatic Pulse Breakdown in NMOS Devices"*, Quality and Reliability Engineering International, 2, pp.107-16, 1986.
2. Amerasekera,E.A., Campbell,D.S., *"An Investigation of the Nature and Mechanisms of ESD Damage in NMOS Devices"*, Sol. St. Elect., 32(3), pp.199-206, 1989.
3. Amerasekera,E.A., Campbell,D.S., *"Charge Limited Breakdown in MOS Capacitors"*, Proc. 17th. Annual Solid State Device Research Conference, pp.733-41, 1987.
4. Maeda,N, Wada,T, *"The Dependence of Electrostatic Destruction Voltage on Device Structures of P-N Junctions and Insulated Films"*, Proc. 11th. EOS/ESD Symp. Proc., pp.157-61, 1989.
5. Lipson,C, Sheth,N.J., *"Statistical Design and Analysis of Engineering Experiments"*, McGraw-Hill Book Company, New York, pp.373-96, 1973.
6. Moazzami,R, Lee,J.C., Hu,C, *"Temperature Acceleration of Time-Dependent Dielectric Breakdown"*, IEEE Trans. Electron Dev., 36(11), pp.2462-5, 1989.
7. Lenzlinger,M, Snow,E.H., *"Fowler-Nordheim Tunneling into Thermally Grown SiO_2"*, J. Appl. Phys., 40(1), pp.278-83, 1969.
8. Wolters,D.R., van der Schoot,J.J., *"Dielectric Breakdown in MOS Devices: Part II: Conditions for the Intrinsic Breakdown"*, Phillips Journal of Research, 40(3), pp. 137-63, 1985.

Chapter 6

Discussions and Further Experiments

6.1 Introduction

This chapter analyses the initial experimental data presented in Chapter 5. The following discussion, together with further diagnostic investigations, leads to the development of a qualitative model of oxide dielectric breakdown. This model allows all the observed breakdown results (ESD, constant voltage, ramp voltage, constant current) to be explained in terms of a single set of physical mechanisms. This model is quantitatively developed in Chapter 7, where the resulting equations are compared with the experimental data.

6.2 Qualification of the ESD Data

This section presents a critical analysis of the experimental methods used for the ESD breakdown experiments of Chapter 5. The results of this analysis allow the data to be correctly interpreted later in this chapter. The results also permit the formulation of an 'ideal' ESD test system, with which future experiments may be performed.

6.2.1 AutoZap Voltage Resolution

In the step-test (see Section 5.2), the AutoZap ESD generator is programmed to produce a sequence of pulses, beginning at ±4V and ascending in magnitude in steps of 4V until breakdown is detected. The magnitude of the final pulse is then recorded as the device's breakdown strength. A device which breaks down under the twelfth pulse is therefore recorded as having a breakdown threshold of 12 x 4V = 48V, although its 'actual' breakdown threshold may lie anywhere between 44V and 48V. Hence a 4V measurement error exists throughout the ESD data of Chapter 5.

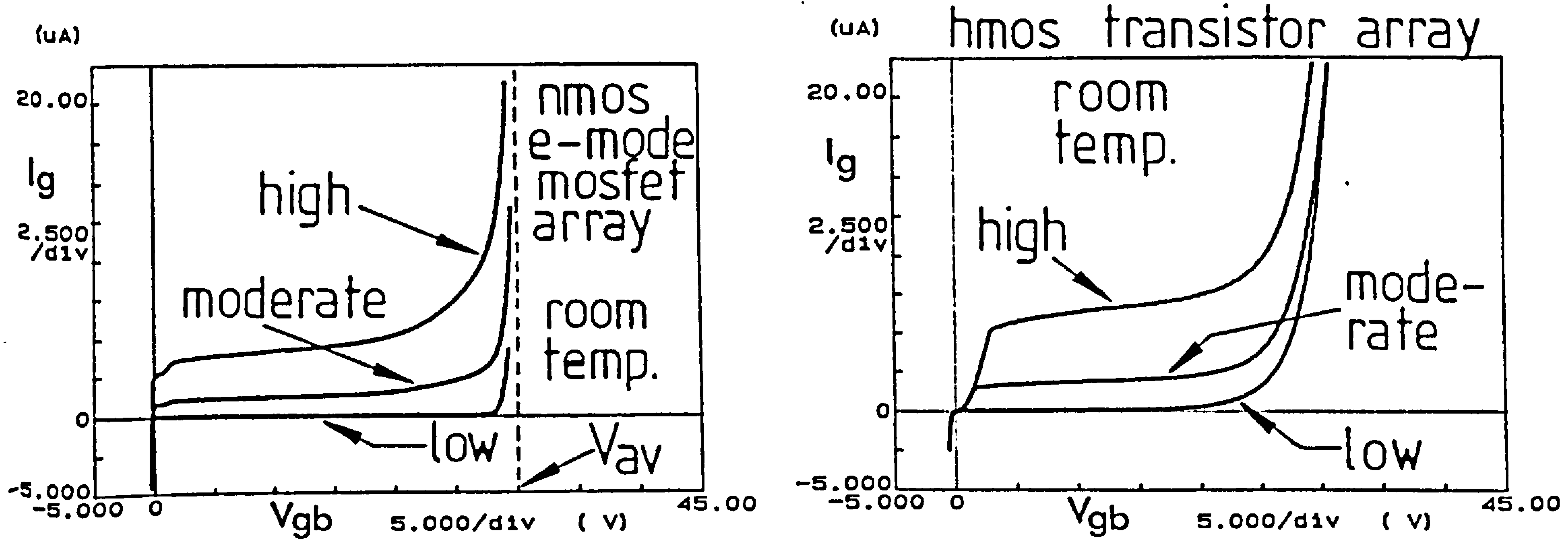


Figure 6.1: Post-Breakdown Characteristics of NMOS and HMOS Transistor Arrays ('low', 'moderate' and 'high' refer to the illumination levels defined in Table 4.1).

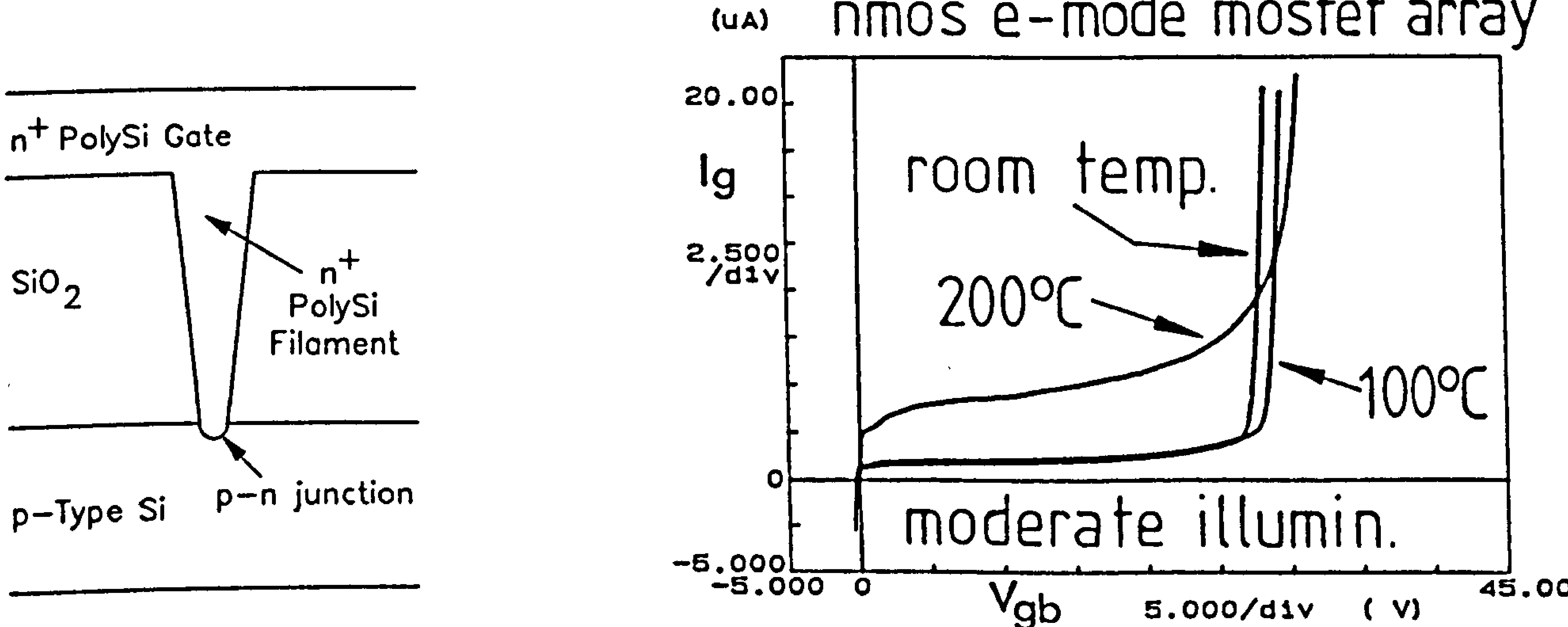


Figure 6.2: Oxide Breakdown Site Structure.

Figure 6.3: NMOS Post-Breakdown Characteristics at Room Temperature, 100°C and 200°C.

6.2.2 Analysis of the Breakdown Criterion

The breakdown criterion employed in Chapter 5 was that of Heimann [1], which defines a failed oxide as one which passes a high current at a low voltage, after being subjected to a high voltage stress. The validity of Heimann's criterion was examined by observing the post-breakdown characteristics of various MOS structures:

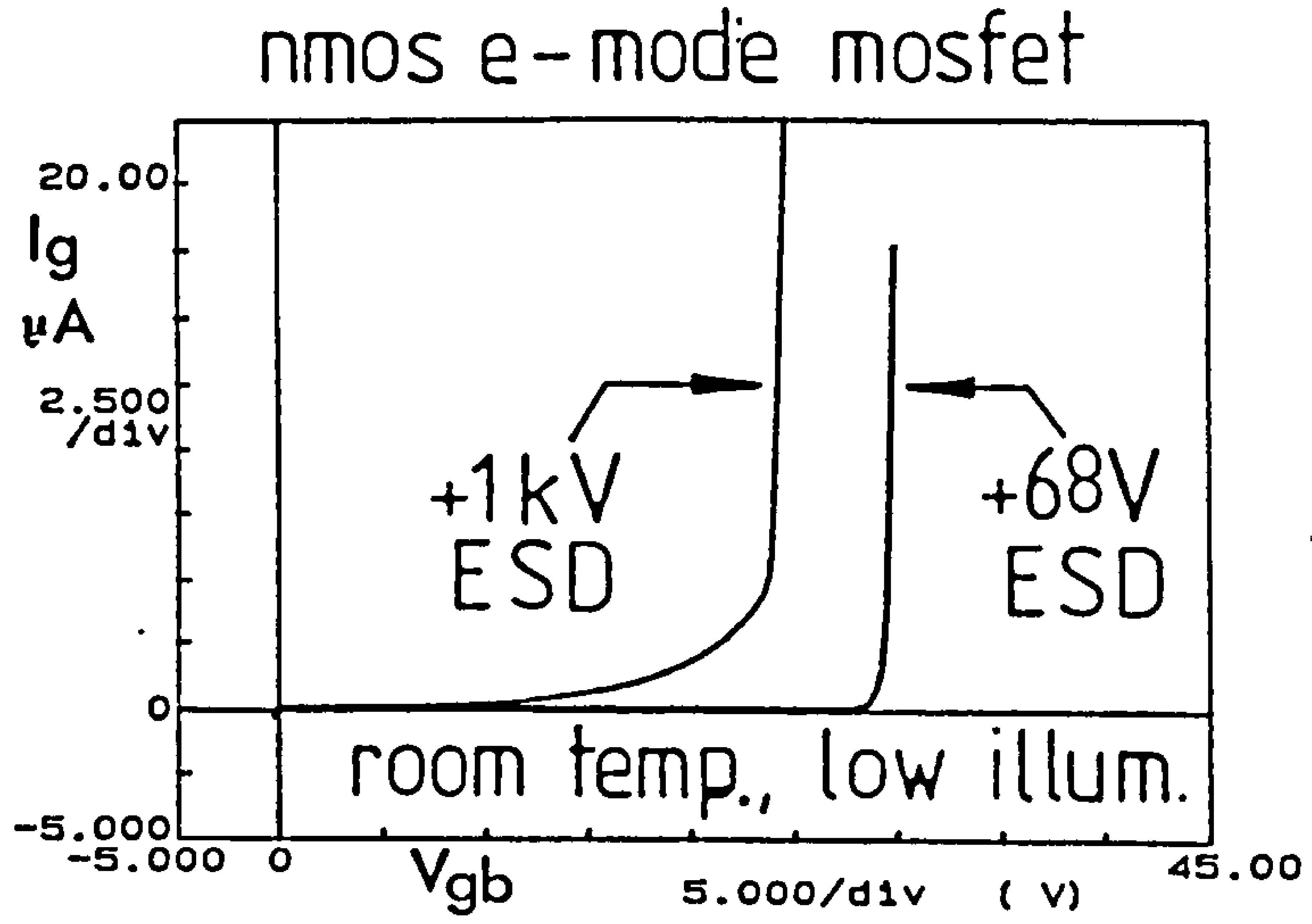


Figure 6.4: 'Softening' of Positive Post-Breakdown Characteristic.

6.2.2.1 NMOS and HMOS Devices

Fig.6.1 shows the post breakdown gate current I_g vs. gate/substrate voltage V_{gb} curves of NMOS and HMOS transistor arrays. (No significant variation was found to exist between the characteristics of E-mode, D-mode and unimplanted structures.) The characteristics may be understood by viewing the failed oxide as a short circuit, connecting the n^+-type gate structure to the p-type substrate and forming a p-n junction. This short-circuit is probably formed by a filament of gate material, diffusing through the ruptured oxide during breakdown (Fig.6.2). The p-n junction becomes forward biased under a negative gate voltage, causing the large negative currents observed in Fig.6.1. Under a positive voltage the junction is reverse-biased, allowing only a small leakage current to flow at voltages below the depletion-layer avalanche threshold V_{av}. This current depends upon the rate of electron-hole pair generation in the depletion layer and is therefore a strong function of luminous intensity (Fig.6.1) and temperature (Fig.6.3). Fig.6.1 shows that the current is practically undetectable under low illumination conditions, and hence the positive polarity breakdown fails to register unless the silicon undergoes the characteristic 'softening' described below.

'Softening' is characterised by a power-law current/voltage dependence ($I \propto V^n$ where n=constant) at voltages below the silicon avalanche threshold. This behaviour may be associated with foreign metal precipitates [2] and/or thermally-induced structural damage [3]. Fig.6.4 shows the softening of the positive-polarity curve induced by a 1kV ESD pulse, compared with the unsoftened characteristic associated with a 64V pulse. It seems probable that the dissipation of the high electrostatic pulse energy causes localised Joule heating in the junction region. This raises the silicon to a temperature close to its melting point (1415°C [4]), causing structural fatigue and the subsequent generation of defect centres. These defects

support Zener-type transitions in the space-charge region, creating a large electron-hole generation current which flows even under zero illumination [3]. (These centres may also affect the silicon impact ionisation coefficient, causing the reduction of V_{av} from 30V to 25V observed in Fig.6.4 [3].)

Hence the extremely large breakdown thresholds measured for NMOS and HMOS devices under positive polarity low-illumination conditions (Figs.5.12 and 5.16) may correspond to junction softening rather than to the oxide breakdown criteria, and must therefore be regarded as spurious.

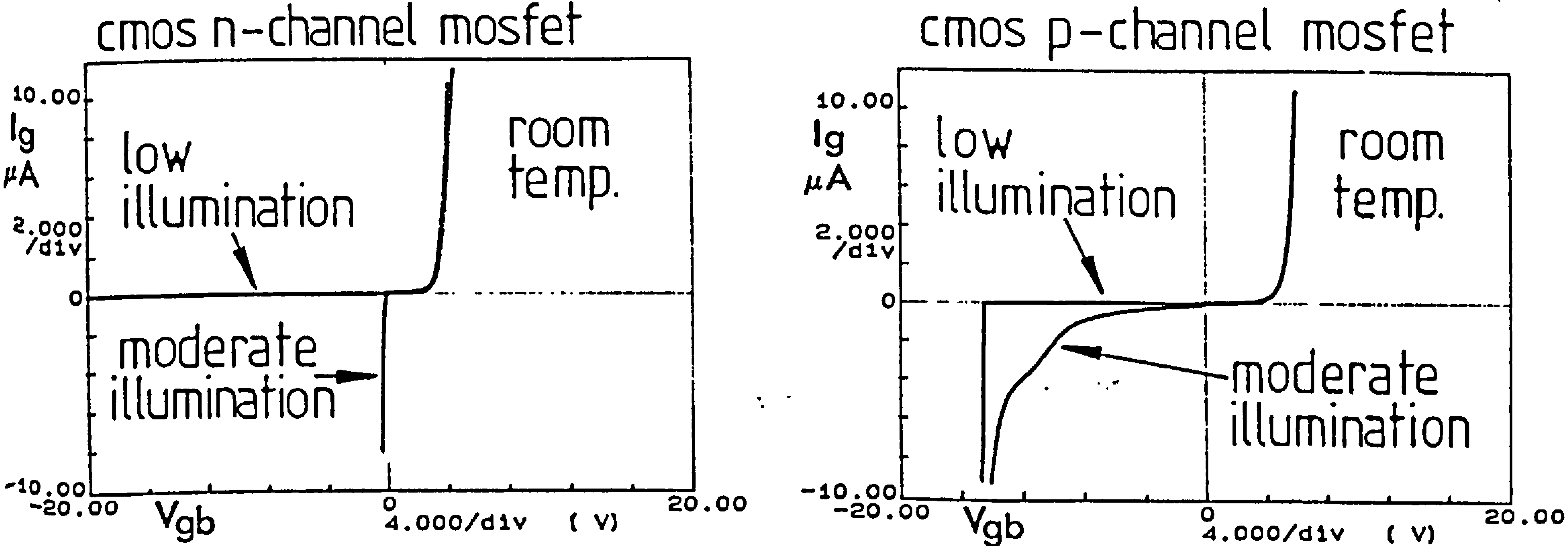


Figure 6.5: Post-breakdown characteristics of CMOS p-channel and n-channel transistor arrays.

6.2.2.2 CMOS Devices

Fig.6.5 shows the post-breakdown I/V characteristics of CMOS n- and p-channel transistor arrays. These curves are fundamentally different from those of Fig.6.1 and the following theory is proposed to explain their behaviour:

It is first necessary to consider the structure of the CMOS devices, which are fabricated on n-type wafers into which p-type 'tubs' or 'wells' are diffused (see Section 2.5.1.2). Since all the devices are enhancement-mode, the p-channel transistors are fabricated upon the n-substrate while the n-channel devices appear on the p-wells. Viewed between gate and substrate, the n-channel transistor appears as two p-n junctions, one consisting of the n^+-gate/p-well contact and the other consisting of the p-well/n-substrate junction. Since a positive gate potential forward biases the gate/well junction and reverse biases the well/substrate junction, the structure presents an open-circuit to low positive gate stress. Above approximately +4V, the depletion layers of the two junctions overlap causing

punchthrough, and the structure conducts a large forward current. If the polarity is reversed, the gate/well junction becomes forward biased and the well/substrate junction becomes reverse biased. The punch-through voltage of this combination is larger than in the positive polarity case and no conduction is detected below -20V. Hence a reverse current can only be caused by electron-hole generation, which may be supported thermally, optically (see the 'moderate Illumination' curve in Fig.6.5) or by Zener tunnelling in a 'softened' junction (Section 6.2.2.1).

In the p-channel devices, a damaged oxide creates a short-circuit between the n^+ gate and the n-type substrate. A positive gate potential therefore allows a large positive gate current to flow to the substrate. However, a negative gate voltage induces a p-type inversion layer beneath the oxide, which behaves in an identical manner to the p-well in the n-channel structure described above. (This behaviour has been previously observed and modelled by Soden et al. [5] and Syrzycki [6].) The damaged p-channel CMOS structure therefore displays a similar characteristic to its n-channel counterpart.

Thus the NMOS/HMOS situation is reversed in the CMOS structures, positive conduction occurring under all illumination conditions and negative conduction requiring optical stimulation or junction softening. The large breakdown thresholds measured under low-illumination negative-polarity conditions (see Fig.5.13) must therefore correspond to softening voltages and are therefore spurious in terms of oxide breakdown.

6.2.2.3 SOS Devices

Fig.6.6 shows the post-breakdown characteristics of an SOS transistor array, measured between gate and common-source. These curves were insensitive to illumination. Since the connections to the oxide are made via the heavily doped gate and source diffusions, no reverse-biased junctions appear in series with the oxide, and the device conducts in both polarities. Hence none of the SOS breakdown thresholds are spurious.

6.2.3 Cumulative Damage

Another possible source of inaccuracy in the ESD experiments is the sub-critical damage inflicted by the pulses prior to breakdown. It is possible that these pulses may weaken the oxide, causing it to fail at a voltage lower than its inherent breakdown strength. This hypothesis was tested by repeated ESD pulsing at a given voltage, recording the number of pulses n_{bd} required for breakdown. Fig.6.7 shows the results of this study which indicate

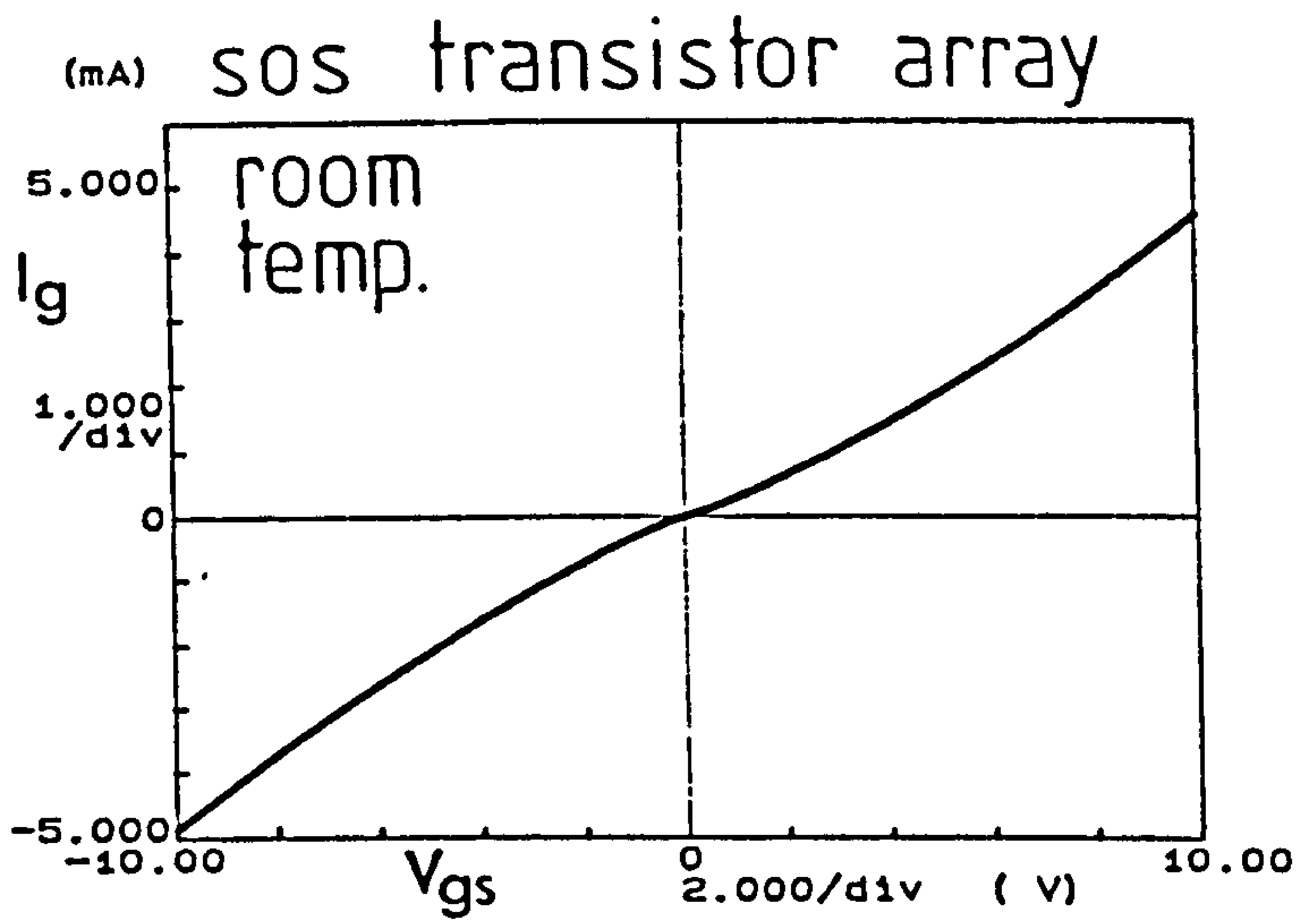


Figure 6.6: Post Breakdown Gate-Source Characteristic of SOS Transistor Array.

NMOS Unimplanted Transistor Arrays (Supplier A)
Gate Area = 1353.45μm^2, Oxide Thickness = 40nm,
Room Temperature, Moderate Illumination.

Negative Polarity

Average no. of Pulses to Breakdown.

50 40 30 20 10 0

20 40 60

Pulse Voltage Magnitude (V)
(Note: Each data point is the mean of 5 devices.)

Positive Polarity

Number of Pulses to Breakdown

40 35 30 25 20 15 10 5 0

40 50 60 70 80

Pulse Voltage Magnitude (V)
(Note: Each data point represents a single device.)

Figure 6.7: Results of Sequential ESD Pulse Experiment.

that n_{bd} decreases very rapidly with increasing pulse voltage. At voltages only 4V below the breakdown threshold, over 25 pulses are needed to cause failure, indicating that the cumulative sub-critical damage is only slight. The data of Chapter 5 can therefore be regarded as a valid representation of the inherent breakdown properties of the oxides, subject to the 4V resolution (Section 4.2.1).

6.2.4 Ideal Test System

These investigations allow the formulation of the following 'ideal' test strategy: The devices should be stressed in total darkness (simulating the conditions experienced by a packaged i.c.) and tested for breakdown under strong illumination. The switch from low to high illumination should be performed using a hand-operated shutter or an L.C.D. screen, covering and exposing the light source. This system is preferable to an electromechanical mechanism, which may introduce electromagnetic interference. Post-breakdown characteristics should also be measured in both polarities, using voltages in the range -20V to +20V (or a narrower range for oxides significantly thinner than 40nm).

6.3 Analysis of ESD Results

6.3.1 Breakdown as a Function of Illumination

In view of the errors introduced into the ESD threshold experiment under low illumination, further experiments were performed in order to determine whether or not a luminosity dependence really does exist for positive polarity NMOS breakdown. These experiments were conducted similarly to those of Chapter 5, using a 4V resolution pulse sequence and testing for breakdown after each pulse. However, breakdown was detected in terms of the *negative* post-breakdown characteristic, which Fig.6.1 shows to be open-circuit under all illumination conditions. Fig.6.8 shows the results, which indicate that the mean value of V_{bd} falls with increasing luminous intensity. Increased illumination also produces a less well defined breakdown voltage. This latter property (which is illustrated by the increase in the coefficient of variation with increasing illumination) is explained later in this chapter.

6.3.2 Breakdown as a Function of Temperature

While the NMOS and CMOS results in Chapter 5 indicate temperature insensitive negative polarity breakdown, the positive polarity thresholds increase perceptibly with temperature. Since the silicon avalanche threshold also increases with temperature [7], this observation adds credence to Amerasekera & Campbell's depletion-layer avalanche model for positive polarity breakdown [8]. (This theory is briefly outlined in Chapter 3.) The increase in V_{av} with temperature is also obvious from Fig.6.3.

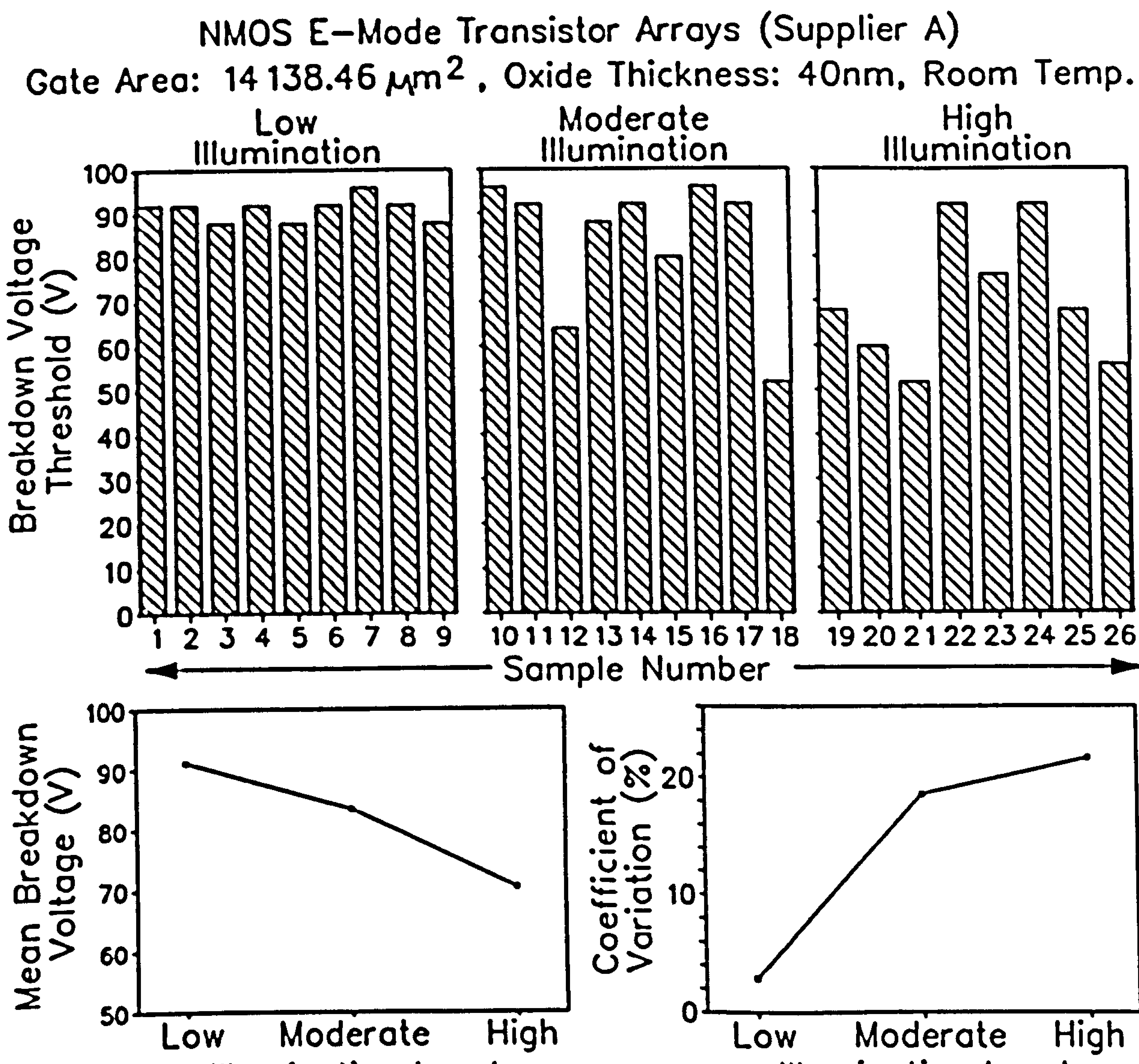


Figure 6.8: Positive polarity ESD breakdown in NMOS transistor arrays as a function of luminous intensity.

6.3.3 Breakdown as a Function of Device Size

Figs.5.7 and 5.8 show that while the CMOS positive and negative polarity breakdown thresholds have no significant dimensional dependence, the positive polarity NMOS threshold is highly sensitive to gate area. The breakdown voltage threshold is generally smaller for the smaller devices (although the variation is not entirely monotonic). This may be explained in terms of the depletion-layer avalanche theory [8] outlined in Section 3.3.7. According to several workers [9,10], the depletion layer in a wide area MOS capacitor with a lightly doped substrate ($N_a < 10^{17}cm^{-3}$) is flat in the centre of the device area and rounded at the device edges (Fig.6.9(i)). Avalanche breakdown therefore occurs at the rounded edges, where the electric fields are at their highest.

However, if the device size is reduced such that its dimensions are of the same order of magnitude as the depletion layer width W_d (~3μm), the flat area in the capacitor centre disappears and the entire depletion layer is forced into a dome-like structure (Fig.6.9(ii)). This greater degree of rounding might possibly increase the electric field per unit voltage, causing the avalanche voltage to fall. Thus the positive-polarity breakdown voltage decreases with decreasing dimensions. Although this explanation accounts for the observed data, it still needs to be verified by a numerical simulation of the depletion layer field system before it can be regarded as theoretically sound.

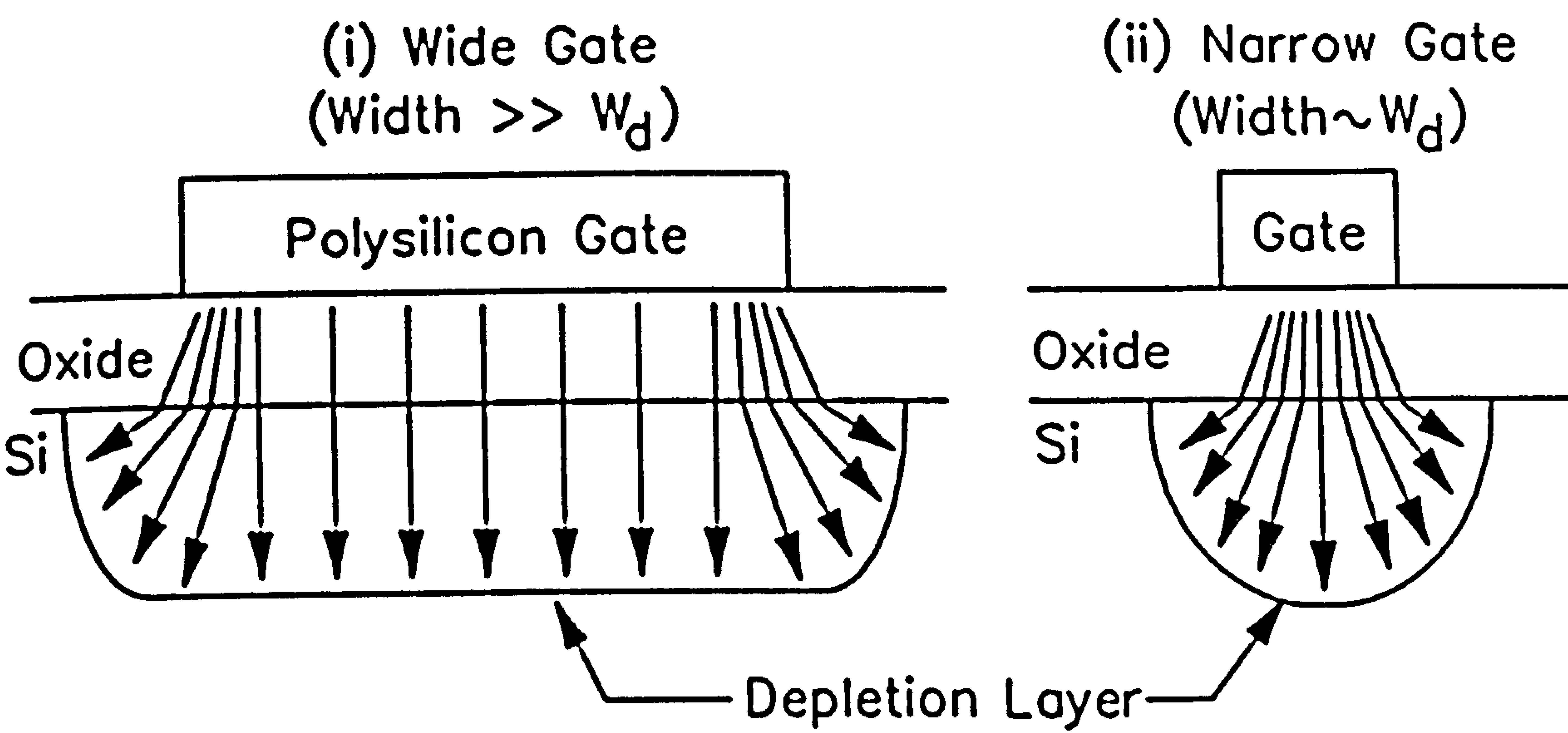


Figure 6.9: Schematic field distributions in wide and narrow depletion layers.

6.3.4 Breakdown as a Function of C_1 and R_2

The ESD experiments of Chapter 5 all employ variations of the human body model (or HBM) in which a charged capacitor C_1 is discharged through a resistance R_2 into the device under test. However, capacitance and resistance variations in this model allow other forms of ESD to be simulated. This is shown schematically in Fig.6.10.

Consider the voltage V(t) across discharge capacitance, which is given by

$$V(t) = V_0 - \frac{1}{C_1 + C_x}\int_0^t I(t)\, dt \qquad 6(1)$$

where I(t) is the displacement current flowing from (C_1+C_x). As $C_1 \rightarrow \infty$, the latter term in Eqn.6(1) tends to zero and the system takes on the qualities of a constant voltage source V_0.

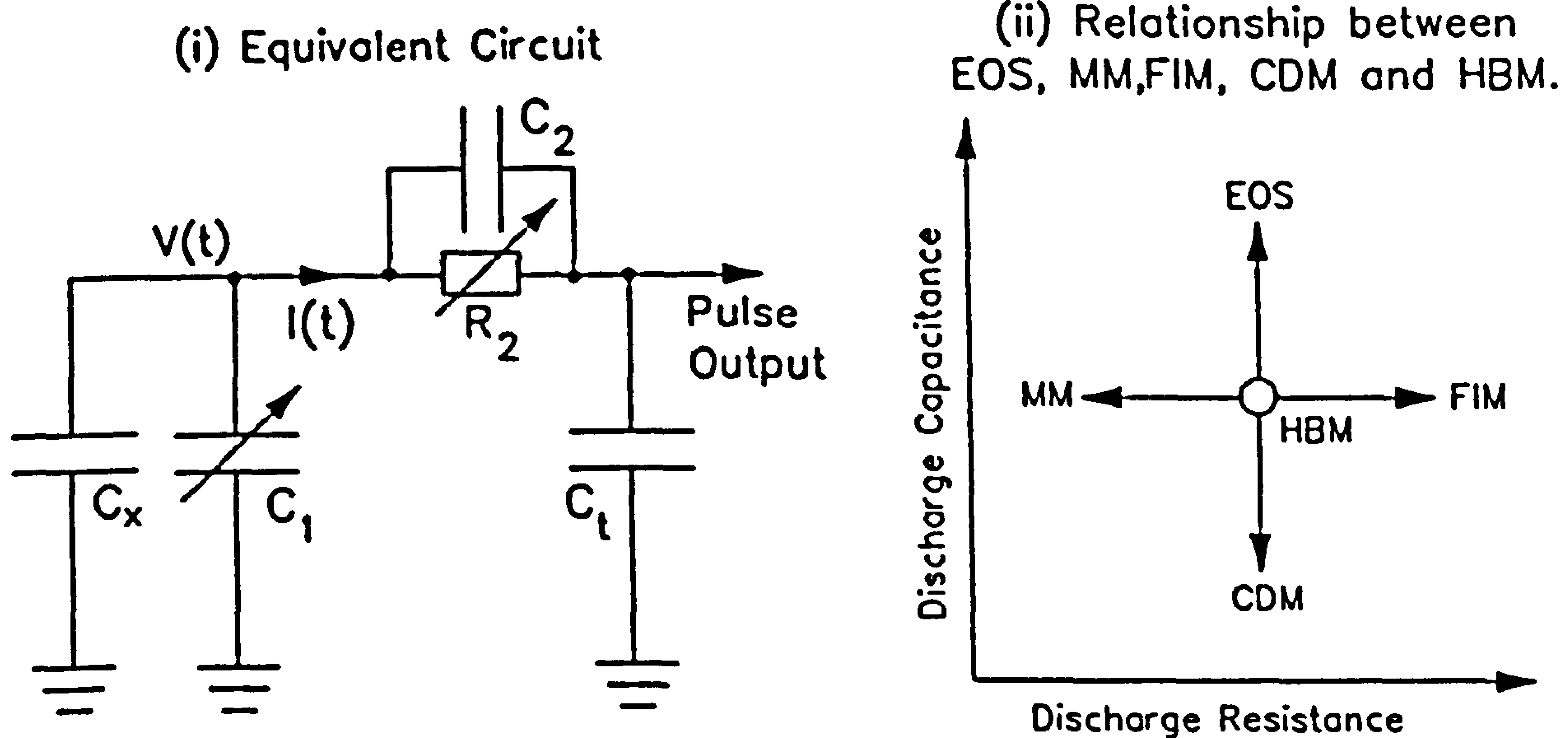


Figure 6.10: Schematic representation of the relationship between HBM and other ESD models.

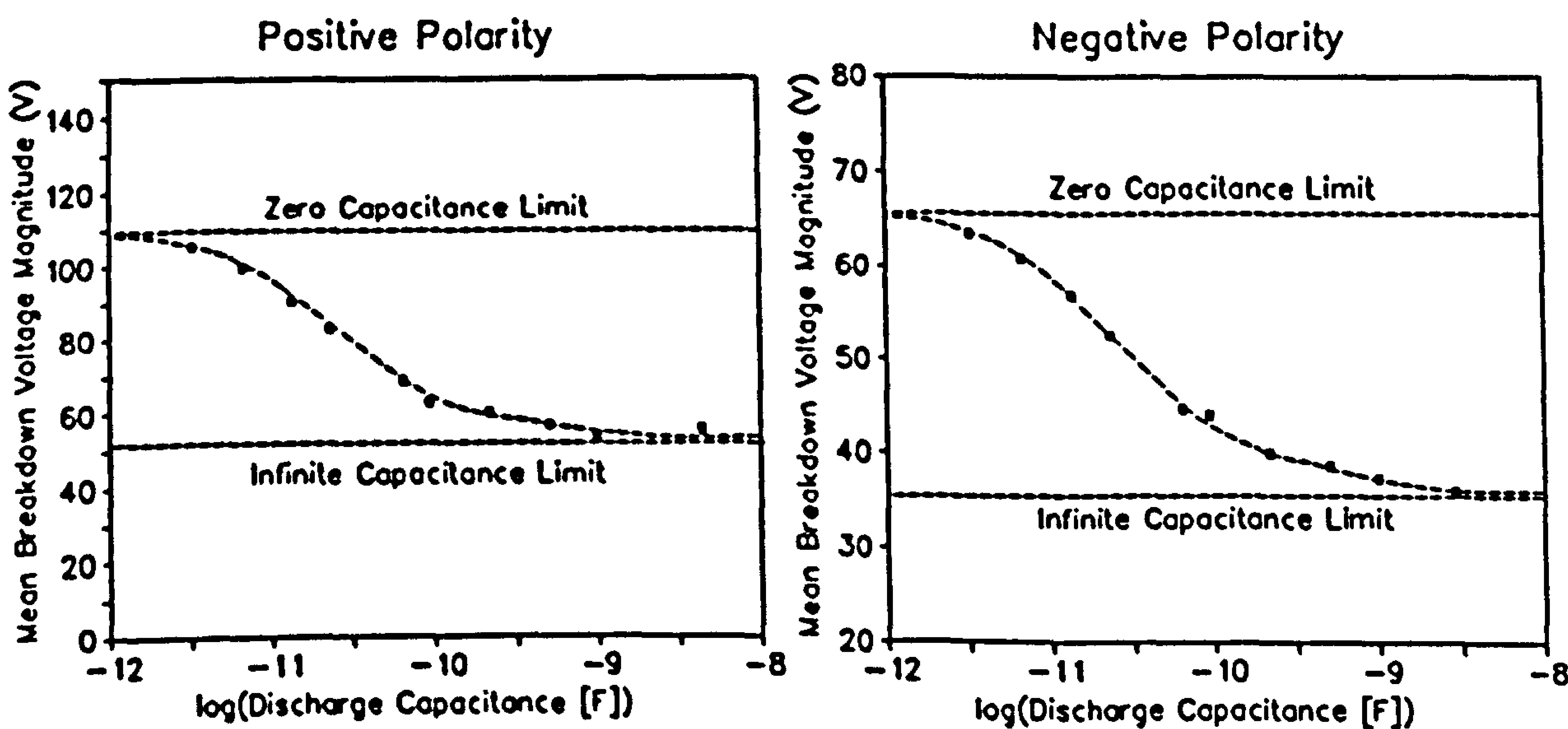


Figure 6.11: V_{bd} vs. $\log(C_1)$ curves showing saturation at high and low C_1.

This can be regarded as a simulation of continuous (EOS) voltage stress.

Conversely if $C_1 \rightarrow 0$ then the pulse charge is limited to the small stray capacitance C_x (~15pF) of the discharge module, simulating the low-capacitance conditions associated with the charged-device model (CDM). (In this latter case however, the correspondence is less direct since the finite pulse charge C_xV_0 is re-distributed between C_x and C_t during the pulse,

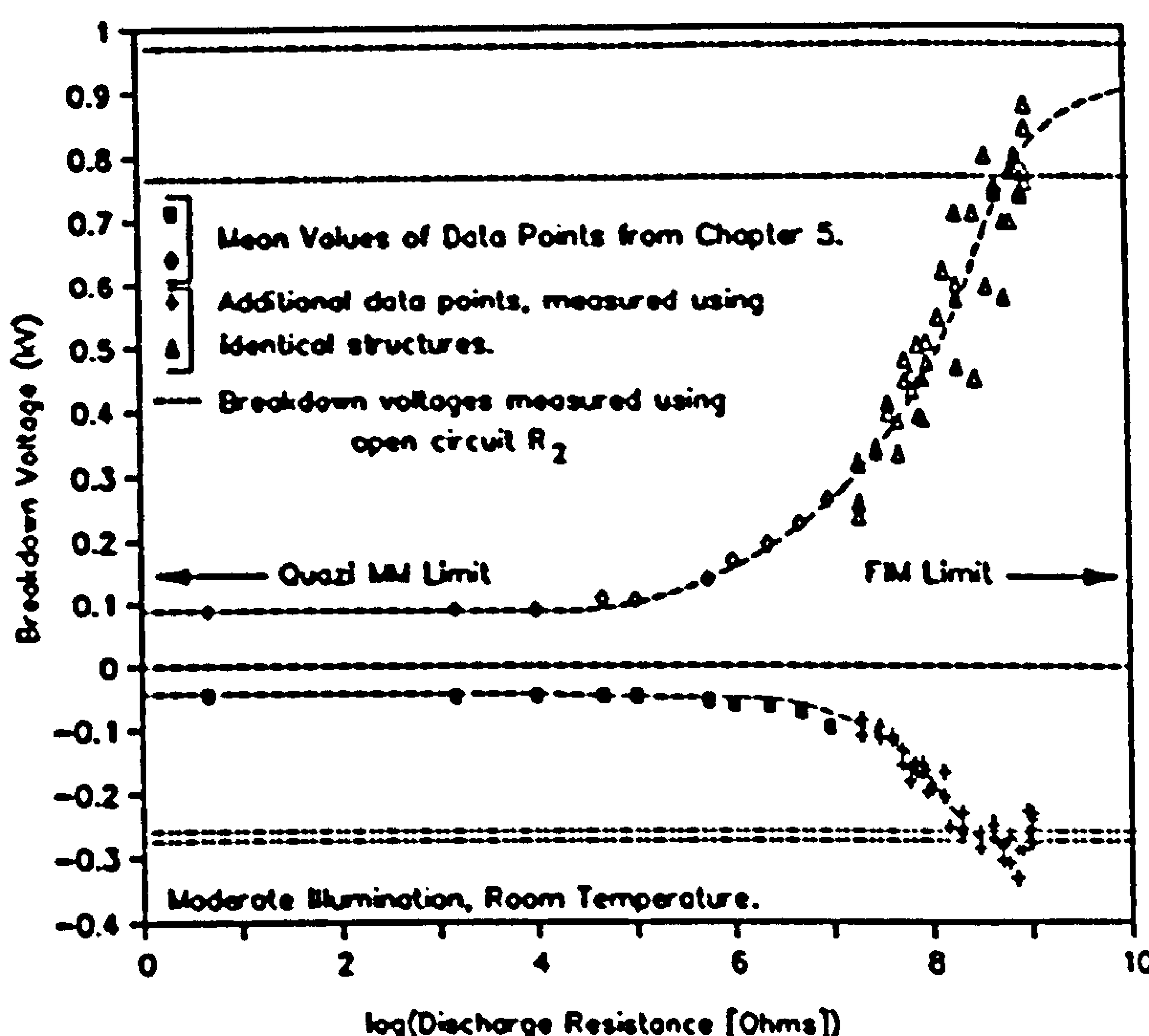


Figure 6.12: extension of V_{bd} vs. $\log(R_2)$ data up to 1GΩ, showing saturation towards 'FIM' limit.

reducing the effective pulse magnitude to a value well below V_0.)

Fig.6.11 shows the NMOS variable-capacitance data reproduced from Section 5.2.5, indicating the saturation of the breakdown voltages at low and high values of C_1, corresponding to the 'Quazi-CDM' and 'EOS' situations. These results bear a close correspondence with those of Maeda & Wada [11], which show the breakdown voltage saturating towards its 'intrinsic' value for $C_1 > 1000$pF.

Further forms of ESD are simulated by the resistance-dependent data (Section 5.2.6). A 'quazi-machine model' (MM) situation is clearly produced as R_2 tends toward zero (the 'true' machine model circuit contains C_1=200pF [see Section 3.2.2]). Conversely, if $R_2 \rightarrow \infty$ (open circuit) then the pulse can only be transmitted to the D.U.T. via the stray capacitance C_2. In this latter case, the experiment simulates the field-induced model (FIM) type ESD described in Section 3.2.2. Fig.6.12 shows V_{bd} vs. $\log(R_2)$ data for resistances up to 1GΩ, compared with the breakdown voltage thresholds measured with R_2 removed from the circuit (FIM simulation). Although there are significant statistical variations between devices, the curves show some evidence of saturation towards the 'FIM' limit as R_2 exceeds 100MΩ. The breakdown voltages in the 'machine model' limit ($R_2 \rightarrow 0$) are practically identical to those measured under HBM (C_2=100pF, R_2=1.5KΩ) conditions and the effects of R_2 upon V_{bd} are insignificant for values of R_2 below 100KΩ. These results therefore demonstrate the limits of the resistance-independent breakdown observed by Maeda & Wada [11] for discharge resistances in the range $0 < R_2 < 10$KΩ.

6.4 Analysis of Constant Voltage and Current Data

The constant voltage data of Section 5.4.4 (reproduced in Fig.6.13[a]) can clearly be divided into three domains, denoted I, II and III. These domains correspond to long, medium and short time-scales respectively, and have their respective boundaries at approximately 1s and 100μs. Fig.6.13(b) shows two data sets reproduced from other workers (Fong & Hu [12] and Chen & Hu [13]), showing the presence of same three domains. (Note: although the II/III boundary in Fig.6.13(b) is obvious, the I/II boundary is characterised by only a very slight change in slope). The following sections examine the physical mechanisms involved in each region.

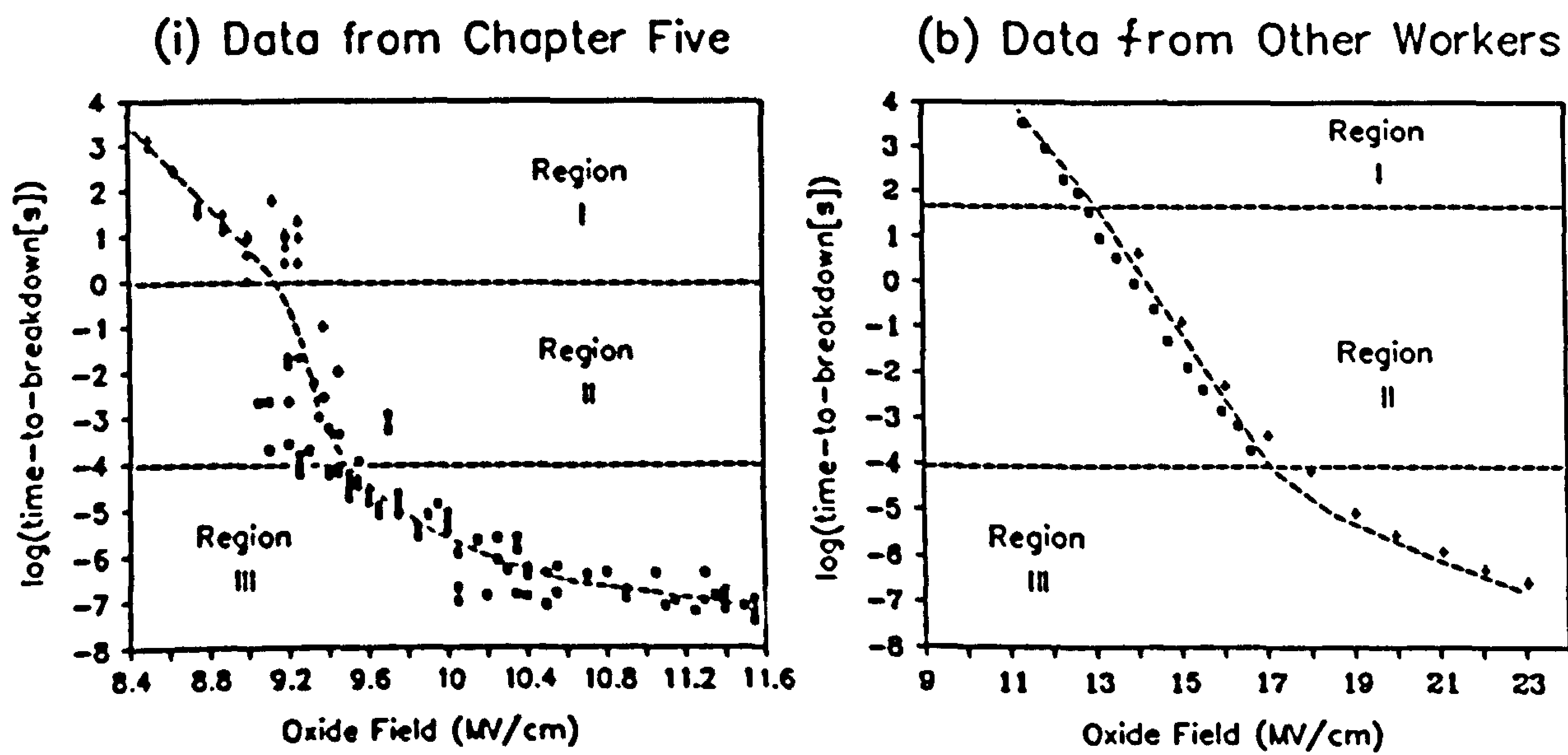


Figure 6.13: (a) Constant voltage t_{bd} vs. F data from Chapter 5, compared with (b) data reproduced from other workers. Data in graph (b) was produced using 10nm oxide MOS capacitors (■=Chen et al [13], +=Fong & Hu [12]).

6.4.1 Region III (Short Time-Scale)

While the breakdown time falls rapidly with increasing field in Region II, Region III is characterised by a slowly saturating t_{bd}. Since dielectric breakdown is believed to be linked to Fowler-Nordheim tunnelling [eg.14], it seems possible that the high-field saturation of t_{bd} is associated with some transient tunnelling phenomenon occurring over short time-scales. It is the purpose of this section to explore this hypothesis, both experimentally and theoretically.

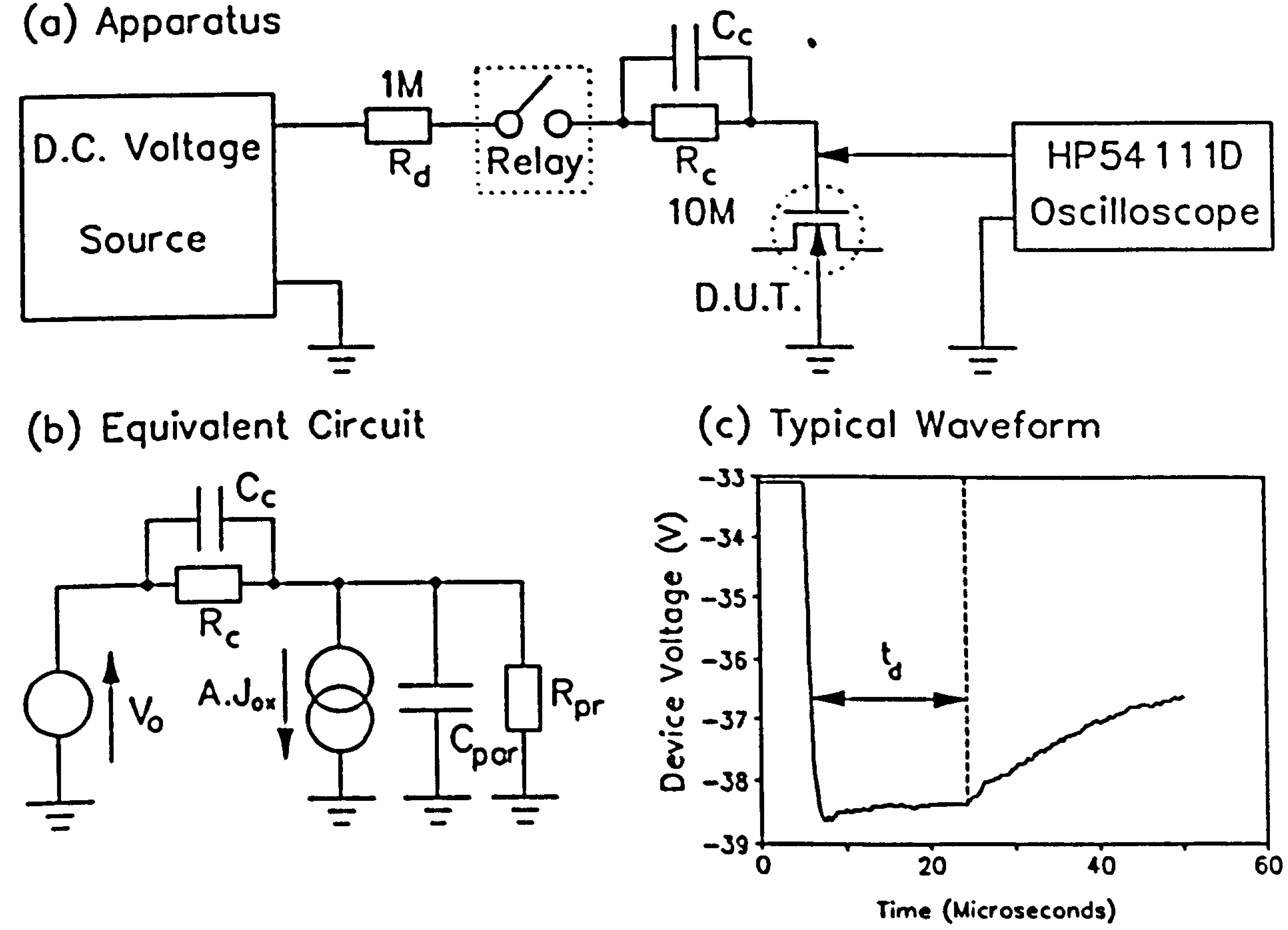


Figure 6.14: Transient Fowler-Nordheim tunnelling analysis: (a) Apparatus, (b) equivalent circuit and (c) typical voltage waveform.

6.4.1.1 Fast Transient Tunnelling Experiments

Under fast transient voltage stress, the tunnelling current J_{ox} may be several orders of magnitude smaller than the displacement current $\epsilon_0\epsilon_{ox}.dF/dt$. It is therefore very difficult to observe J_{ox} using standard techniques. For this reason, a specific experiment was devised which focuses on short time-scale tunnelling. Fig.6.14(a) shows the apparatus used in this experiment and Fig.6.14(b) shows the equivalent circuit for the apparatus. In this equivalent circuit, C_c and R_c are intentional circuit elements, R_{pr} is the resistance of the oscilloscope voltage probe and C_{par} represents the total capacitance appearing in parallel with the oxide (including the capacitances of the oxide and the oscilloscope probe, together with the parasitic capacitances of the connecting leads). The resistance R_d is included in order to damp down the inductive voltage spike which otherwise occurs at the leading edge of the waveform.

The operation of the circuit is as follows: As the relay is closed, a voltage transient is applied to the potential divider consisting of C_c and C_{par}. Hence (after a short rise-time

associated with R_d) a voltage $V_{ox}=V_0C_c/(C_c+C_{par})$ appears across the device-under-test. If the potential divider associated with R_c and R_{pr} is adjusted such that $R_c/R_{pr}= C_{par}/C_c$ then charge leakage from C_{par} via R_{pr} is exactly replenished by the current in R_c, and V_{ox} remains constant. However, if a tunnelling path opens across the oxide then this equilibrium is disturbed and V_{ox} decays with time. The $V_{ox}(t)$ profile is captured on the digital oscilloscope and downloaded to the Walters 286 computer.

The waveform of Fig.6.14(c) shows that the voltage decay (and hence the tunnelling current) begins only after a finite time delay t_d after the application of the voltage. Similar values of t_d were obtained by repeatedly pulsing the same device, showing that the processes active during t_d are reversed when the field is removed. Further experiments showed that t_d decreases with increasing oxide field and typical waveforms are shown in Fig.6.15. Fig.6.16 shows the t_d vs. field data from Fig.6.15 superimposed on the $\log(t_{bd})$ vs. field curve reproduced from Fig.6.13(a). The latter diagram shows that the t_d and t_{bd} data correlate closely in Region III, suggesting that Region III failure is largely governed by the tunnelling time-delay mechanism.

Additional results show that t_d is not significantly affected by variations in temperature (Fig.6.17) and luminous intensity (Fig.6.18). A theoretical analysis of the mechanisms which may be responsible for the tunnelling time delay is presented in the following section.

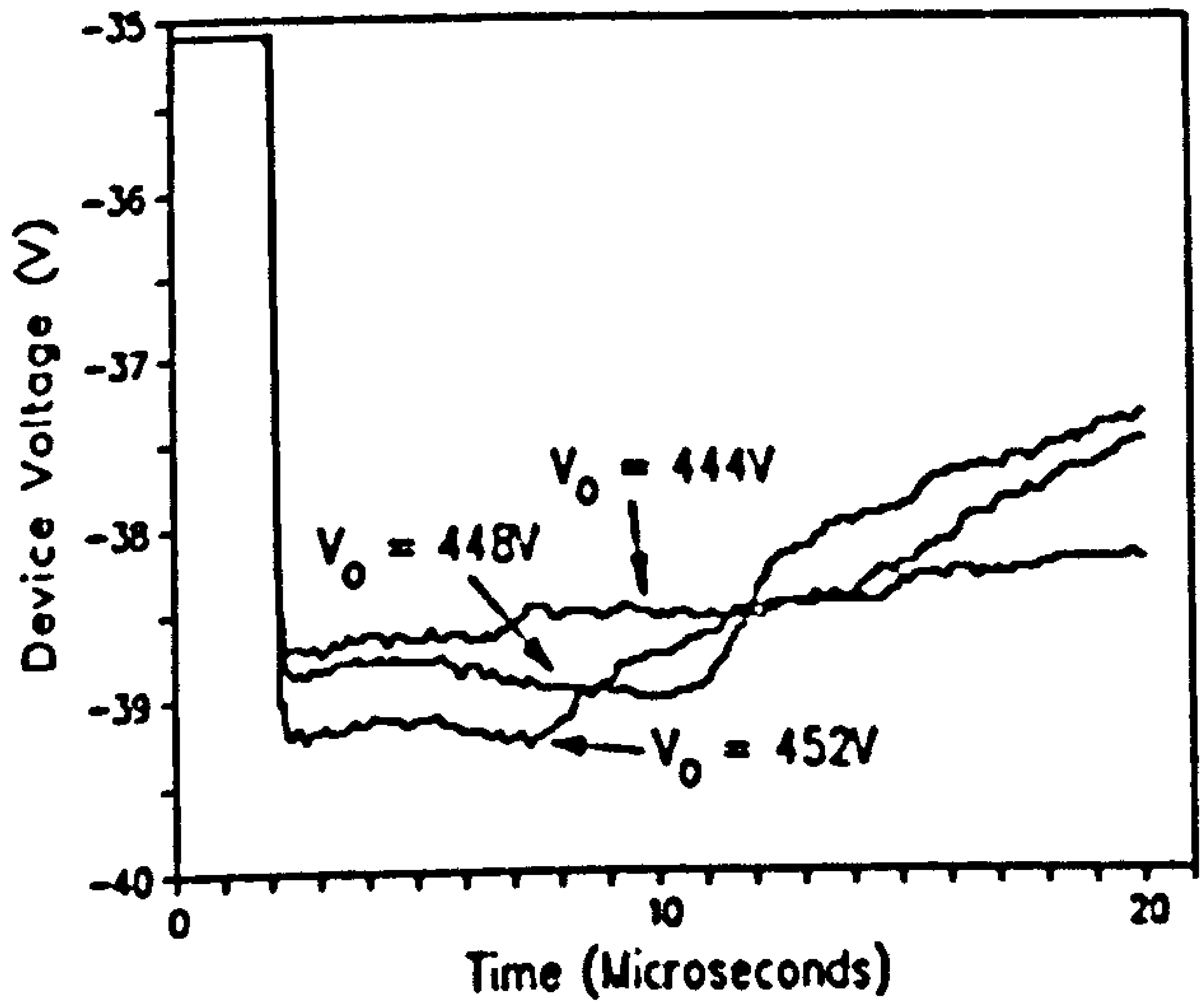


Figure 6.15: Tunnelling time-delay transients produced by three different stress voltages.

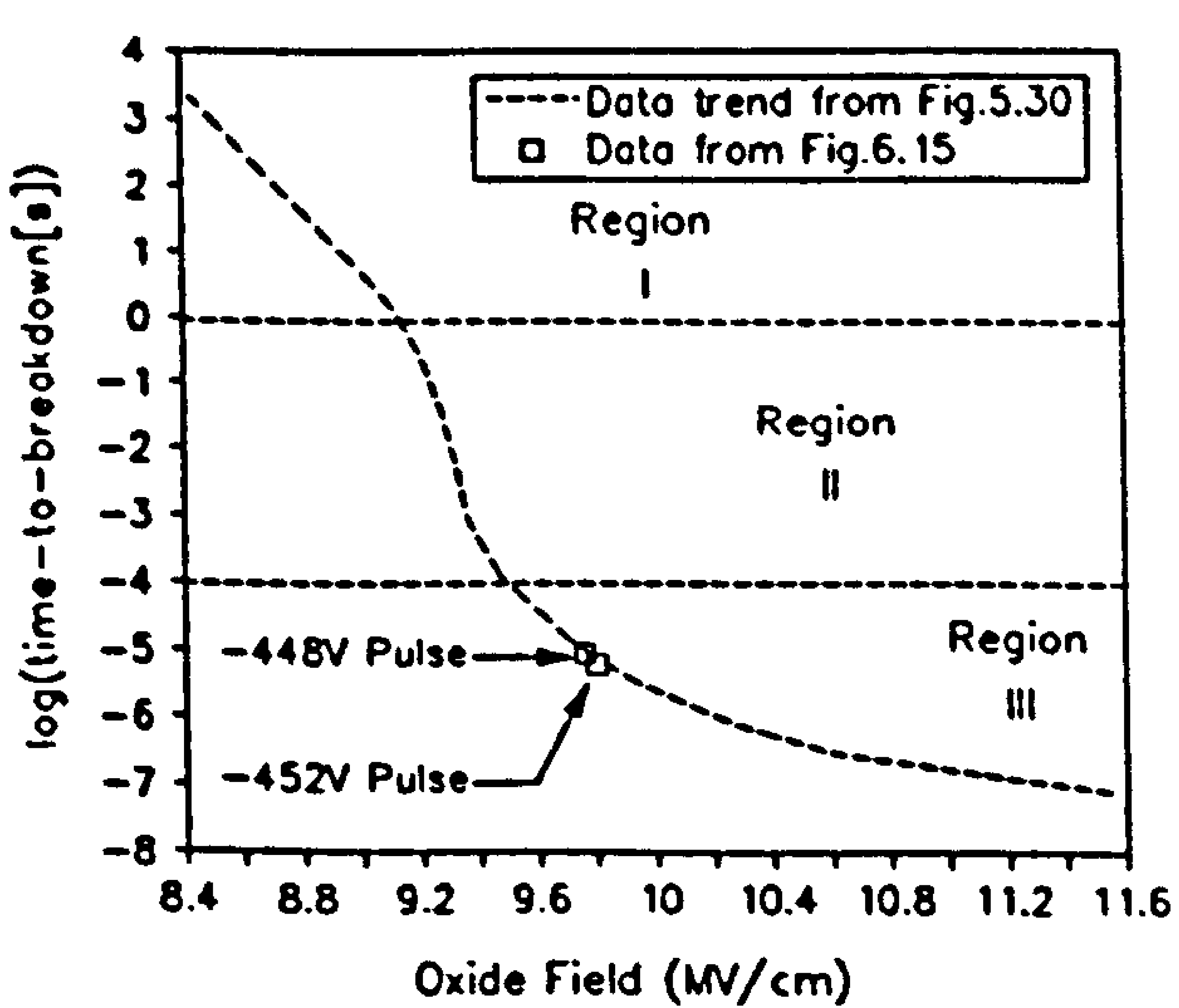


Figure 6.16: Tunnelling time-delays (from Fig.6.15), superimposed on time-to-breakdown vs. field curve from Fig.5.30. The data are in approximate agreement in Region III.

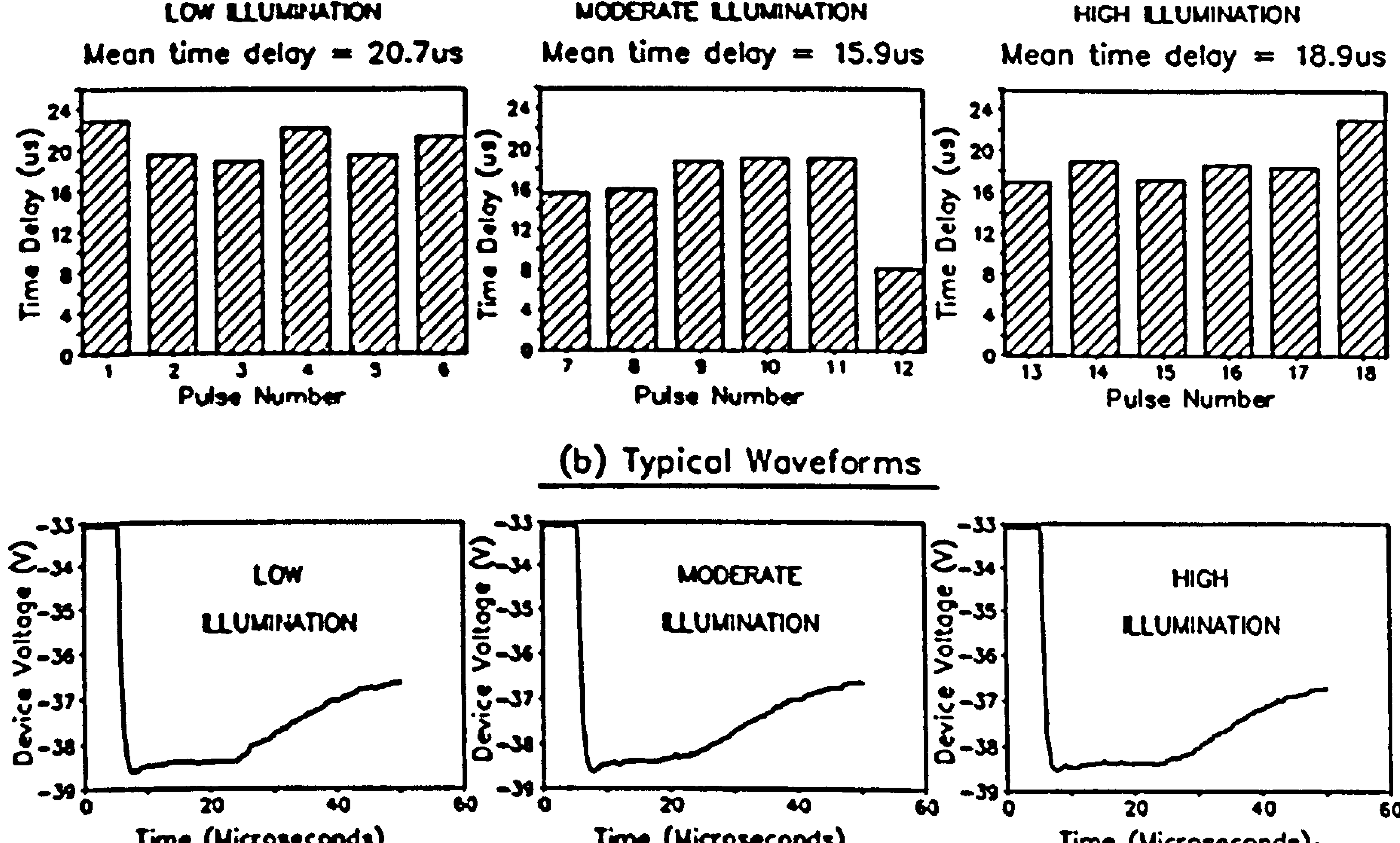


Figure 6.17: Tunnelling time delay as a function of luminous intensity.

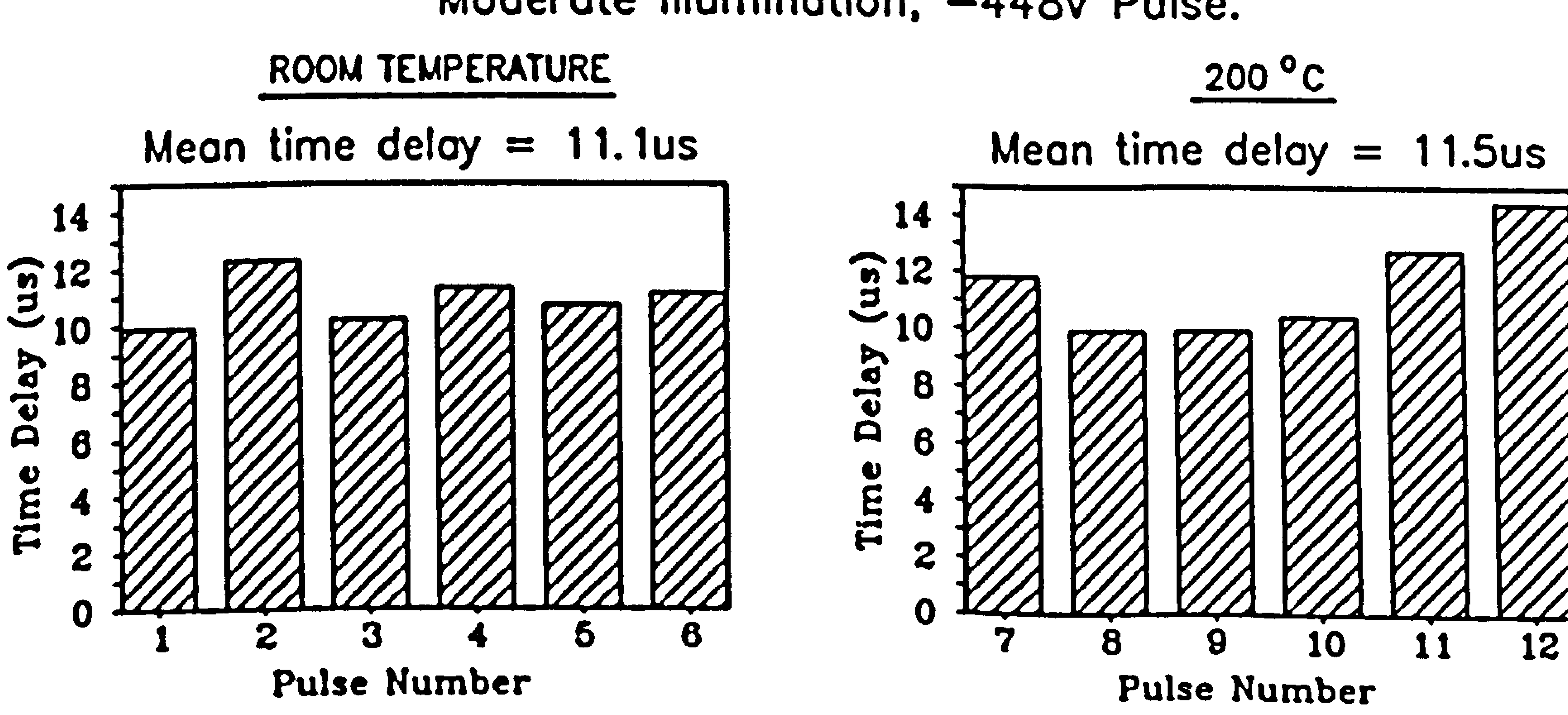


Figure 6.18: Tunnelling time delay as a function of wafer temperature.

6.4.1.2 Theoretical Analysis of Tunnelling Time Delay

Several theories were developed in order to explain the time delay t_d, all of which are described and analyzed below:

Theory 1: t_d is Governed by the Transit Time of Electrons in the Tunnelling Barrier.

According to de Moura et al. [15], the transit time of an electron in a barrier is given by the integral of the electron wavefunction divided by the tunnelling electron flux. For the case of a rectangular barrier, the tunnelling time ΔT_t is given by

$$\Delta T_t = \frac{m_{ox}^*}{2\hbar}\cdot\frac{K}{Q}\cdot\Delta x\cdot\left[\frac{Q}{K^2} - \frac{1}{Q}\right] + \frac{m_{ox}^*}{4\hbar}\cdot\frac{K}{Q}\cdot\left[\frac{1}{Q^2} + \frac{1}{K^2}\right]\sinh(2Q\Delta x) \qquad 6(2)$$

where

$$Q = \sqrt{\frac{2m_{ox}^*(V-E)}{\hbar^2}} \quad ; \quad K = \sqrt{\frac{2m_{ox}^*E}{\hbar^2}} \qquad 6(3)$$

and V is the barrier height, E is the energy of the tunnelling electron, Δx is the barrier thickness and m_{ox}^* is the effective mass of an electron in SiO_2. A triangular barrier, such as that experienced in an MOS structure, can be approximated by stacking a large number n of rectangular barriers in series (Fig.6.19), and summing the transit times, i.e.

$$T_t = \sum_{t=1}^{n} \Delta T_t \qquad 6(4)$$

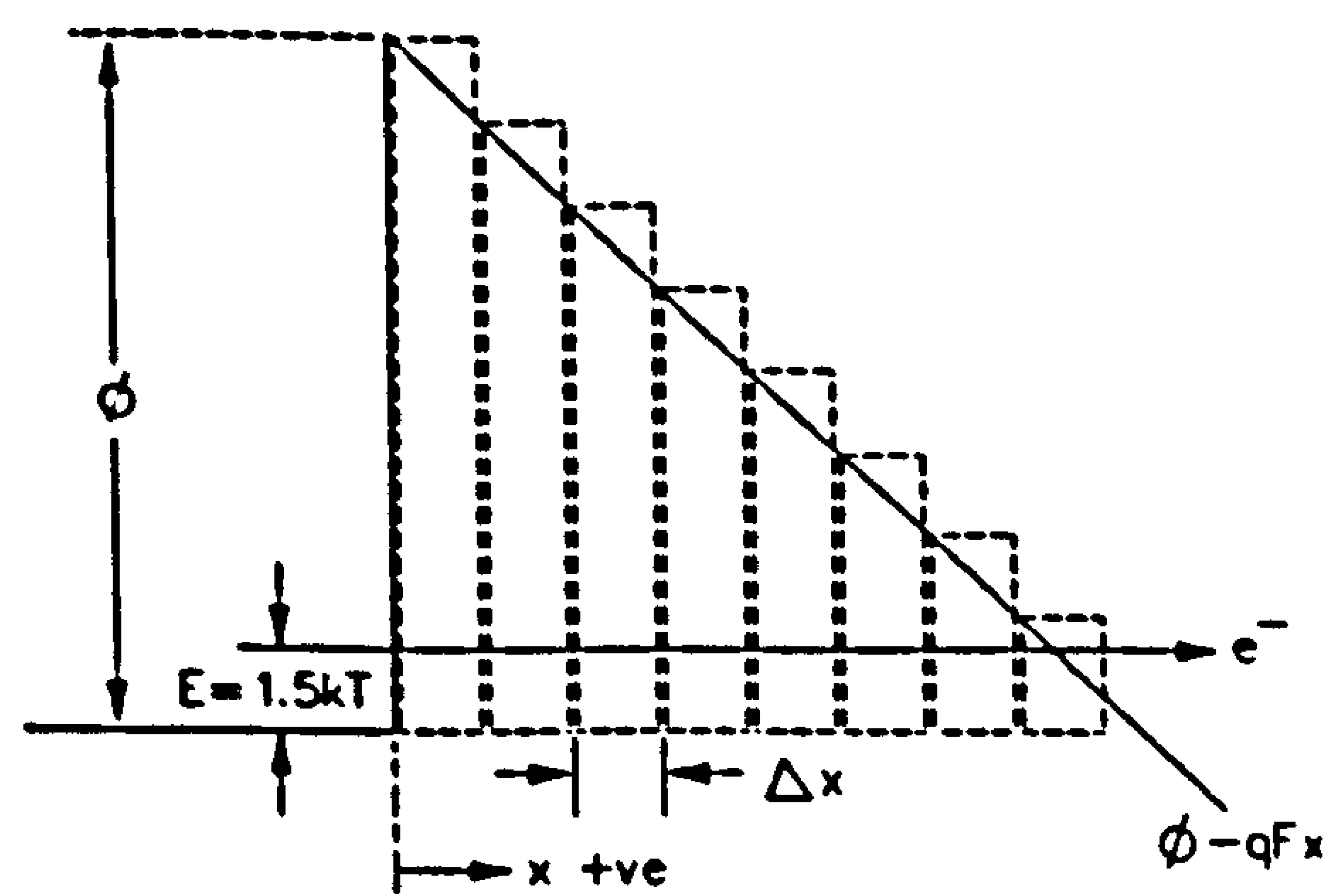


Figure 6.19: Resolution of triangular Fowler-Nordheim barrier into finite rectangular barriers.

Although this approximation is very crude (it assumes that $\Delta T_t \propto \Delta x$), it should be adequate for order-of-magnitude calculations. The calculation was performed for a Fowler-Nordheim barrier under 10MV/cm field. The barrier was divided into one hundred 0.0315nm thick rectangular slices, varying uniformly in height from 3.15eV to zero across 2.15nm. The effective mass m_{ox}^* was set at $0.5m_0$ [10] and the tunnelling electron energy was set at the mean thermal energy 1.5kT. The resulting value of T_t was found to be $1.921.10^{14}$ seconds, which is clearly many orders of magnitude smaller

than t_d. Thus a quantum-mechanical explanation of the tunnelling time delay can be eliminated.

Theory 2: t_d is Governed by the Response Rate of Surface Carriers.

A second possible explanation for the time delay is that t_d is associated with the response time of electrons at the PolySi-SiO_2 interface. Since in a p-substrate/n^+-gate structure under negative polarity these are majority carriers, their response is characterised by the *majority carrier response time* τ_{maj}. According to simple theory [17]

$$\tau_{maj} = \frac{\varepsilon_0 \varepsilon_{si}}{q \mu_n n} \qquad 6(5)$$

where μ_n is the electron mobility in Si and n is the cathode electron density. If $n=10^{21}cm^{-3}$ (the polysilicon doping level) then $\tau_{maj}=4.83.10^{-18}$ seconds, which is again far too small to explain the time delay.

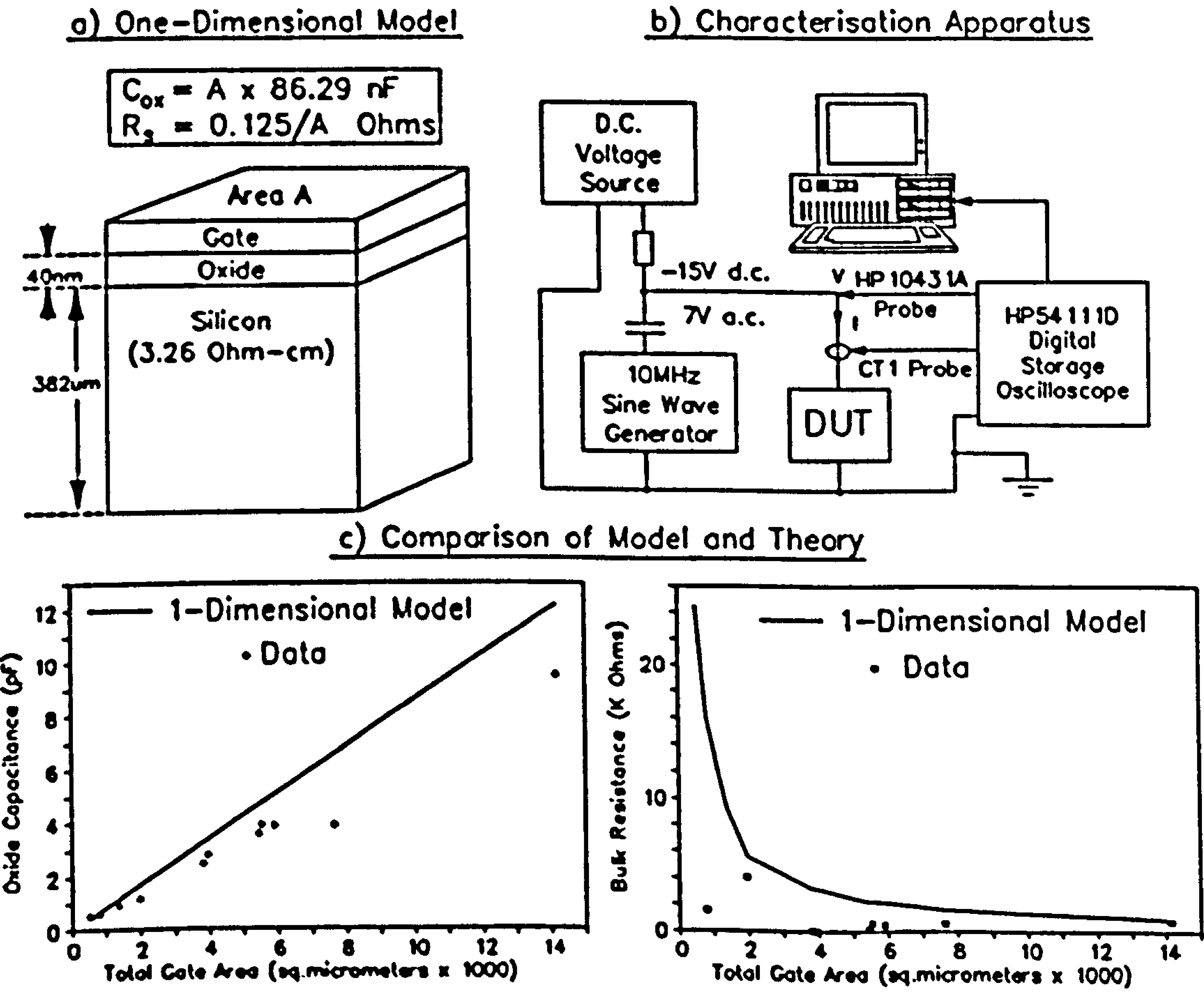


Figure 6.20: (a) One dimensional C-R model for MOS structure, (b) Apparatus for measuring C_{ox} and R_b, (c) Experimental results compared with model predictions.

Theory 3: t_d is Governed by the C-R Time Constant of the MOS System.

The time delay may be associated with the C-R time constant of the structure. Fig.6.20(a) shows a simple 'one-dimensional' model of an MOS device which assumes that oxide thickness is uniform, that all the current is limited to the region of Si directly below the oxide and that no fringing fields exist at the gate periphery. According to this model, if the oxide area is A and the oxide thickness is 40nm then C_{ox} is equal to 86.29 × A (nF) and the series bulk resistance R_s is equal to 0.125÷A (Ohms). The time-constant ($C_{ox}R_s$) is therefore 10.8ns, irrespective of device size, implying a five-time-constant relaxation time of 5 x 10.8ns = 53.9ns.

These predictions were compared with experimentally determined values of C_{ox} and R_s. Since the time-constant is too small for the 10kHz sampling frequency of the Wayne-Kerr 4210 L-C-R bridge (see Section 4.2.2.4), a technique based upon that of Wiley and Miller [18] was used. A 10MHz sine-wave-wave signal was applied to the device under test, together with a -10V d.c. bias to drive the Si surface into strong accumulation [Fig.6.20(b)]. Voltage and current profiles were captured by the HP54111D oscilloscope and downloaded to a Walters 286 computer, where the impedance Z_d of the structure was calculated. The capacitance and resistance were then extracted using the formulae

$$R_s = \Re(Z_d) \qquad C_{ox} = \frac{1}{2\pi f\,|\Im(Z_d)|} \qquad 6(6)$$

where f is the 10MHz sampling frequency. Fig.6.20(c) shows that the resulting C_{ox} and R_s are universally lower than the theoretical predictions. (This error [which is examined in Appendix A] is almost certainly due to the non-confinement of field and current to the one-dimensional section of Fig.6.20(a)). Hence the above-calculated 10.8ns must be regarded as the theoretical upper limit of the time-constant.

Although the 53.9ns relaxation time is consistent with the lower limit of the t_{bd} vs. F data (Fig.6.13[a]), it cannot explain the 20μs delay times illustrated in Figs.6.15 and 6.17. Hence the C-R time-constant theory can be eliminated from the analysis.

Theory 4: t_d is Caused by Oxide Charge Re-Distribution.

Having eliminated the three theories described above, it seems probable that the time delay is associated with an electric field perturbation, caused by the transient re-distribution of space-charge within the bulk oxide. The fact that current increases with time implies an increase of the electric field at the cathode. It can therefore be assumed that during the first few microseconds of stress, the oxide space-charge becomes *more positive* with time.

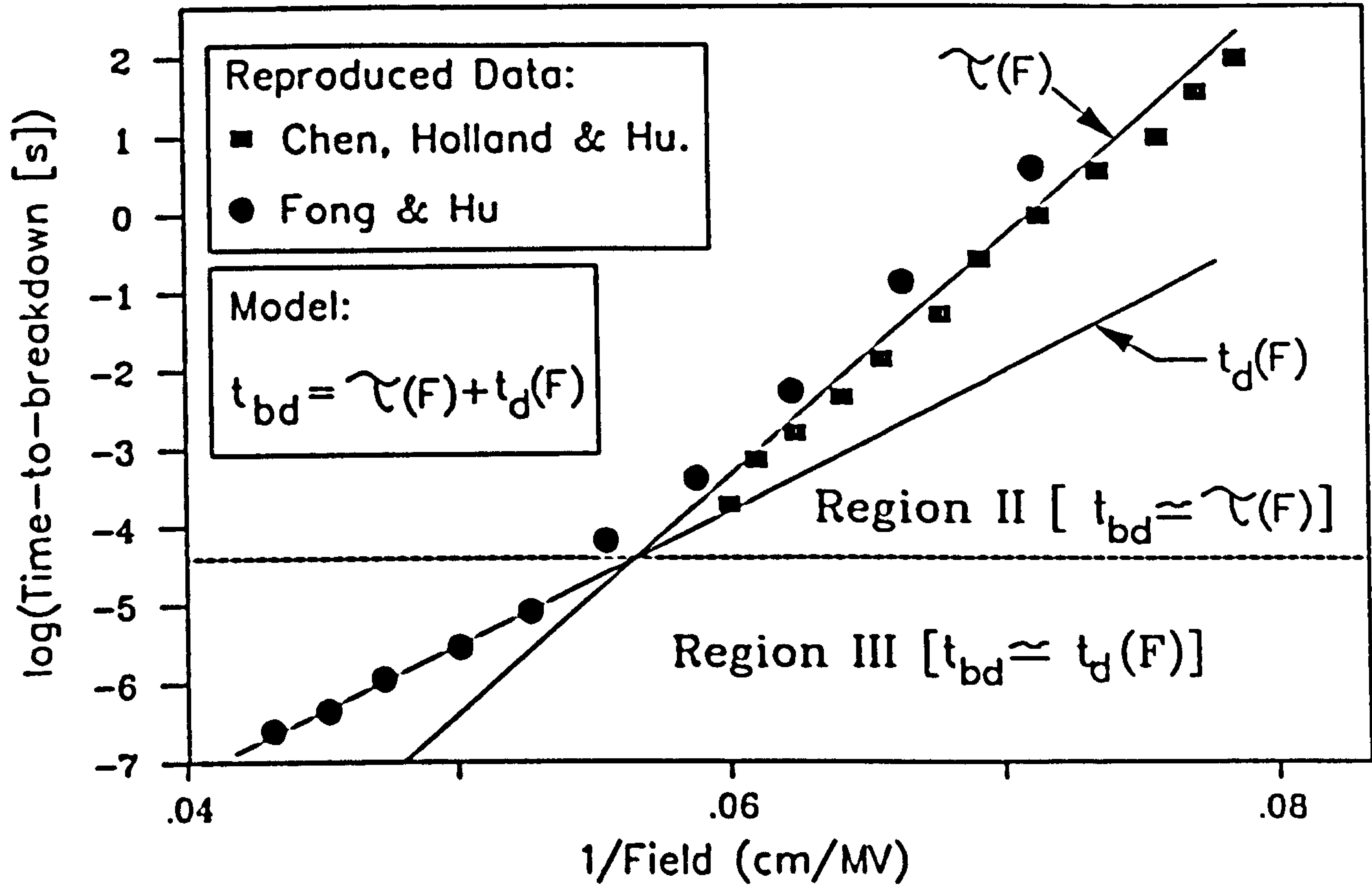


Figure 6.21: Theoretical dependencies of t_{bd} in Regions II and III.

6.4.2 Region II (Moderate Time-Scale)

Fig.6.13(a) shows that as t_{bd} exceeds approximately 100μs, the slope of the log(t_{bd}) vs. F graph increases considerably. However, since such wide statistical variations exist within this region, the corresponding data of Chen & Hu [12] and Fong & Hu [13] (Fig.6.13[b]) is used as the basis for the following discussion.

Fig.6.21 shows some of the data of Fig.6.13(b) reproduced on a log(t_{bd}) vs. 1/F format. In Region II, the graph appears to be linear, in accordance with the hole trapping theory [13,14], whilst the data points in Region III all lie above this straight-line model. This suggests that the time-to-breakdown function t_{bd}(F) is composed of two components, one of which is the TDDB causal wearout time τ(F) (see Section 3.3.5) and the other is the tunnelling time-delay t_d discussed above. The time-to-breakdown can therefore be expressed as

$$t_{bd}(F) = t_d(F) + \tau(F) \qquad 6(7)$$

Both of these components are shown on Fig.6.21. Throughout most of Region III, $t_d >> \tau$(F) and hence $t_{bd} \simeq t_d$. This produces the correlation between t_{bd} and t_d observed in Fig.6.16. In Region II however, $t_d << \tau$(F) and hence $t_{bd} \simeq \tau$(F).

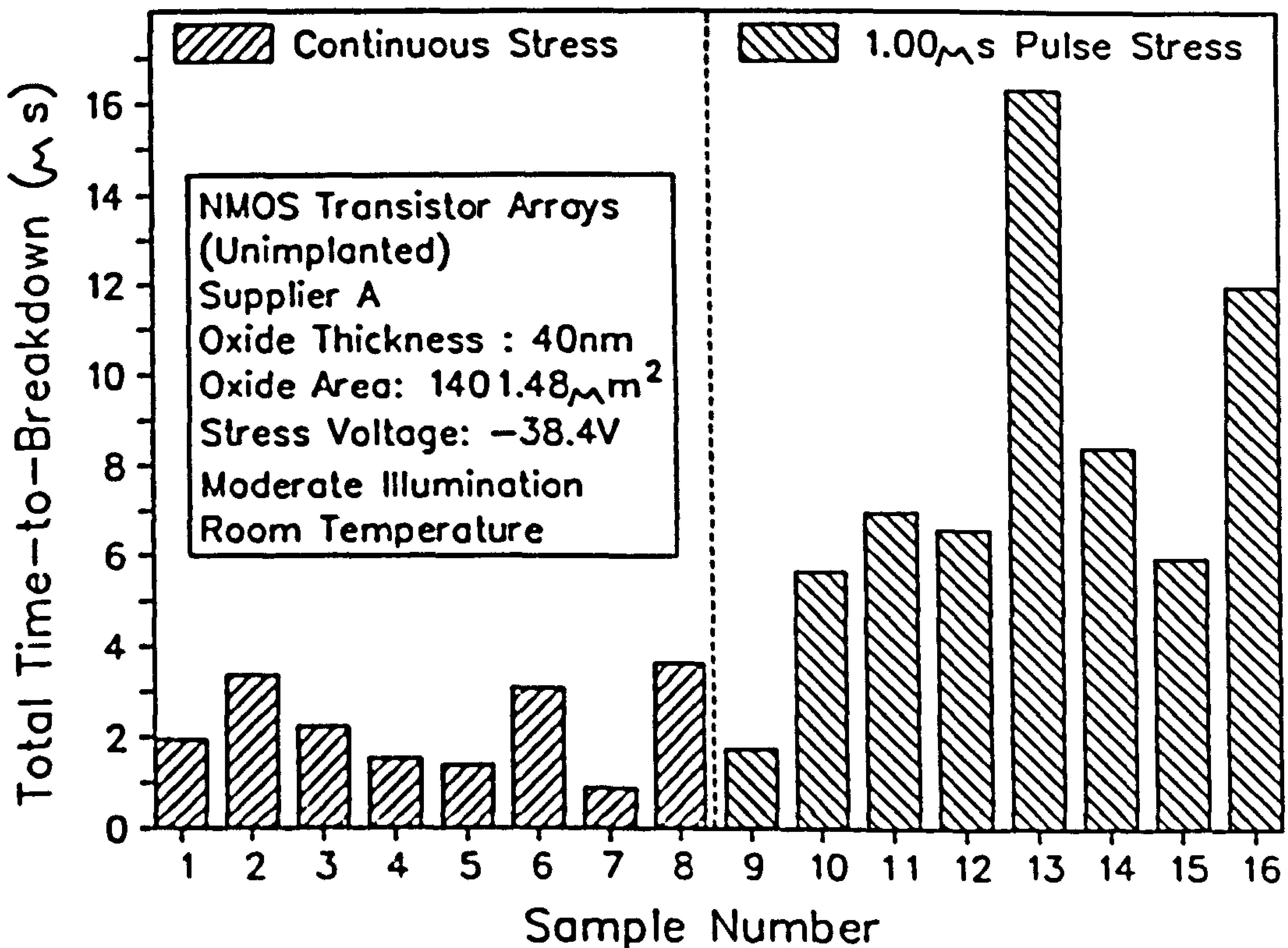


Figure 6.22: Comparison of total times-to-breakdown under continuous and 1μs pulse-sequence stress.

The above-stated theory, if correct, prohibits the direct application of causal wearout theory (Section 3.3.5) to short time-scale conditions, and thus invalidates the Fong & Hu model [12]. The theory has been tested by the following experiment: Eight identical NMOS transistor arrays were stressed at a continuous -38.4V and the times-to-breakdown were recorded. Eight more devices were subjected to sequences of -38.4V x 1μs rectangular voltage pulses and the total stress-time-to-breakdown (summed across each pulse in each sequence) was recorded for each device. The results, displayed in Fig.6.22, show that t_{bd} is generally much longer under pulsed stress than under continuous-voltage stress. This observation is consistent with the idea that t_d must elapse during each stress-period or pulse prior to wearout. Hence a continuous stress period requires only a single t_d (assuming F remains large enough to keep the oxide charge from relaxing), while a sequence of n pulses requires a total delay time of n.t_d. In order to bring the causal model into line with this observation, Eqn.3(1) must be re-written

$$\int_{t_d}^{t_{bd}} \frac{dt}{\tau(F)} = 1 \qquad 6(8)$$

such that the damage does not begin to accumulate until after t_d has elapsed.

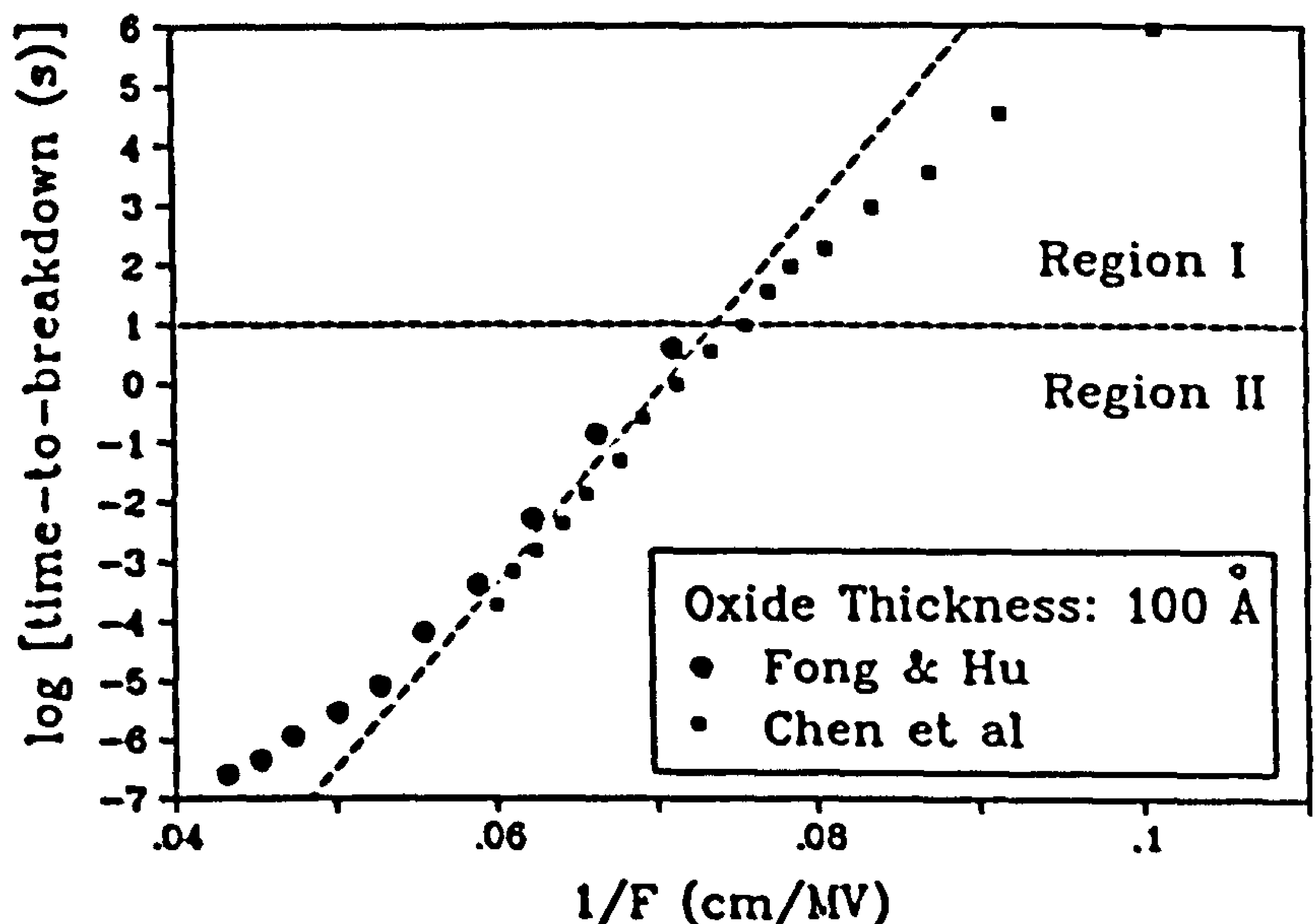


Figure 6.23: Data reproduced from Fig.6.13(b), plotted on log(t_{bd}) vs. 1/F format, showing the transition between regions I and II.

6.4.3 Region I (Long time-scale)

The long-time-scale portion of the log(t_{bd}) vs. F curve (Fig.6.13[a]) appears to have a slope which decreases with decreasing field. This trend is also apparent from the data of Fig.6.13(b), which is reproduced on a log(t_{bd}) vs. 1/F format in Fig.6.23. The I/II transition is more obvious on this format than on that of Fig.6.13(a), since all the Region I data points fall well below the $t_{bd}=\tau_0 e^{\gamma/F}$ model [13,14]. All these data suggest an acceleration of the oxide wearout mechanism in Region I.

Fig.6.24 shows a typical injection-current profile observed in Region I. The current decay has been observed by several earlier workers [eg.14] and has generally been attributed to the buildup of negative space-charge during stress. This space-charge is usually modelled in terms of trapped electrons in SiO_2 defect sites [19]. According to the Poisson equation, the trapped negative charge reduces the cathode electric field, causing the cathode/SiO_2 tunnelling current to decay with time. Thus the wearout process might be supposed to decelerate in Region I.

However, the negative space-charge also enhances the electric field at the anode of the oxide layer. Since many of the mechanisms governing wearout (eg. the creation and injection of hot holes [20]) are believed to occur at the anode, the negative charge may enhance wearout. Thus two opposing mechanisms are active in Region I, one opposing and the other enhancing breakdown. The shape of the curve in Fig.6.13(a) suggests that the latter mechanism predominates.

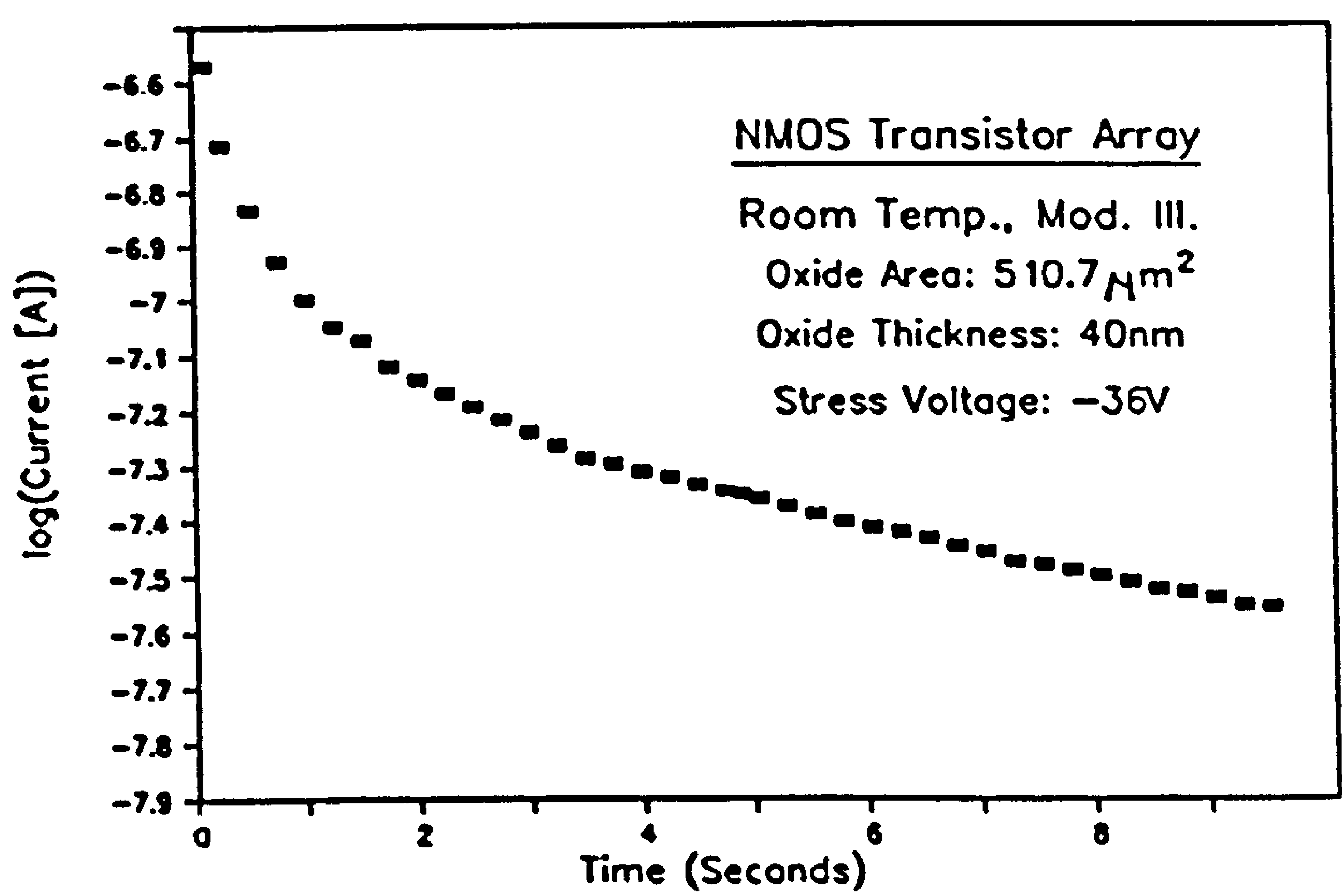


Figure 6.24: Typical injection-current transient for NMOS array stressed in Region I.

If the growth of negative charge is indeed caused by electron trapping, its growth must be governed by the electron injection rate and the time-period of stressing. This explains why the negative charge has no noticeable effect in Regions II and III, since the time-scales associated with these regions are too small to cause any noticeable charge buildup.

The three domains in the t_{bd} vs. F curve may therefore be characterised in terms of the dynamic charge states of the oxide, listed in Table 6.1.

Table 6.1

Dynamic Charge States of Oxides in Regions I, II and III.

Time Domain	Dynamic Charge-State of Oxide
Region I (>1s)	$dQ/dt < 0$
Region II (100μs to 1s)	$dQ/dt \simeq 0$
Region III (<100μs)	$dQ/dt > 0$

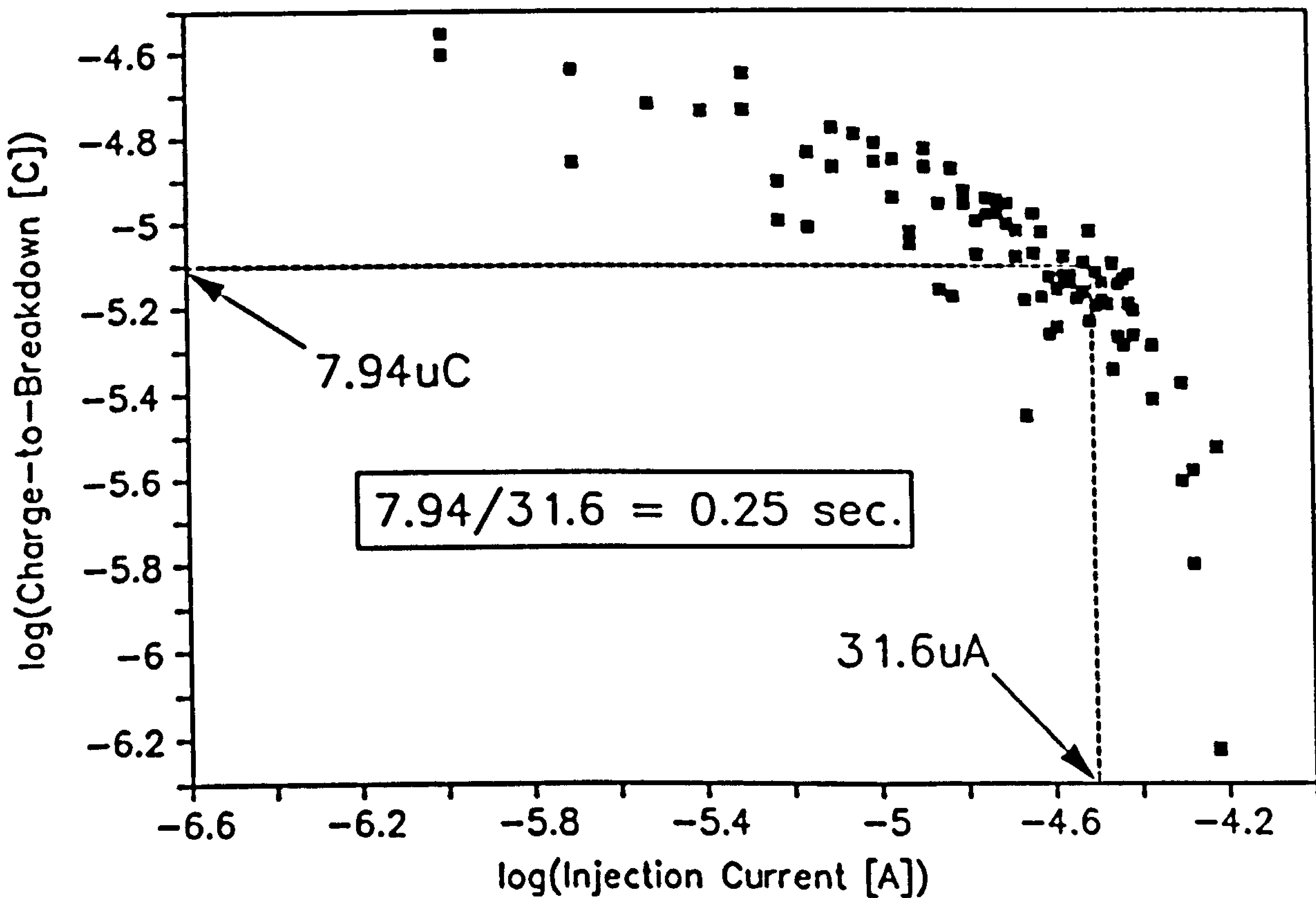


Figure 6.25: Constant current data reproduced from Chapter 5. Time-to-breakdown at the transition between 'brittle' and 'ductile' failure roughly corresponds to the boundary between Regions I and II of the constant-voltage data.

6.4.4 Analysis of Constant Current Data

Fig.6.25 reproduces the charge-to-breakdown vs. current data from Chapter 5. The 'critical' transition between the 'ductile' and 'brittle' failure modes (see Figure 3.8) occurs at approximately 31.6μA, at which point the charge-to-breakdown is 7.94μC. The corresponding time-to-breakdown is therefore equal to 7.94μC $\div$ 31.6μs = 0.25seconds. This correlates approximately to the transition between regions I and II in the constant-field data (Fig.6.13(a)). Brittle failure therefore appears to correspond to Regions II & III, while ductile failure occurs in Region I.

6.4.5 Temperature, Illumination, Doping and Size Dependencies of Constant Voltage Breakdown

The experiments reported in Section 5.4 showed that constant-voltage TDDB is independent of device size, illumination and substrate doping. The size-independence is consistent with the negative polarity ESD data (Fig.5.8), and suggests that the breakdown mechanism is intrinsic (ie. independent of the number of defect sites within the gate oxide area). The illumination and doping independencies are also in agreement with the ESD results (see Fig.5.12).

The results of the variable-temperature experiment (Fig.5.27) are important since although a high degree of scatter exists, the general trend indicates increasing t_{bd} with increasing temperature. This observation clearly conflicts with the results of previous workers who found an Arrhenius-type relationship with a positive activation energy [e.g.21]. The effectively negative E_a of Fig.5.27 presents further evidence of the inconsistency of SiO_2 properties (i.e. the fact that oxides produced by different manufacturers using different processing parameters can have totally different properties).

The increase in t_{bd} with T indicates that the wearout reaction is de-accelerated by temperature. This suggests the *low temperature breakdown* process (see Section 3.3.3), in which a dielectric becomes less susceptible to breakdown at elevated temperatures. An alternative possibility is that breakdown is triggered by the trapping of charge-carriers, who experience de-trapping under elevated temperatures.

6.5 Analysis of Ramp Voltage Data

The results of the ramp-voltage experiments of Section 5.3 require a certain amount of mathematical processing before their significance can be determined. The analysis is based upon the circuit model of Fig.4.12, the parameters of which were characterised in Section 4.2.3.3. The circuit parameters of the MOS structures were measured using the technique of Fig.6.20, which showed C_{ox}=41pF and R_b=400Ω. The following sections describe the processing methods and present the processed waveforms. The significance of the results is then discussed.

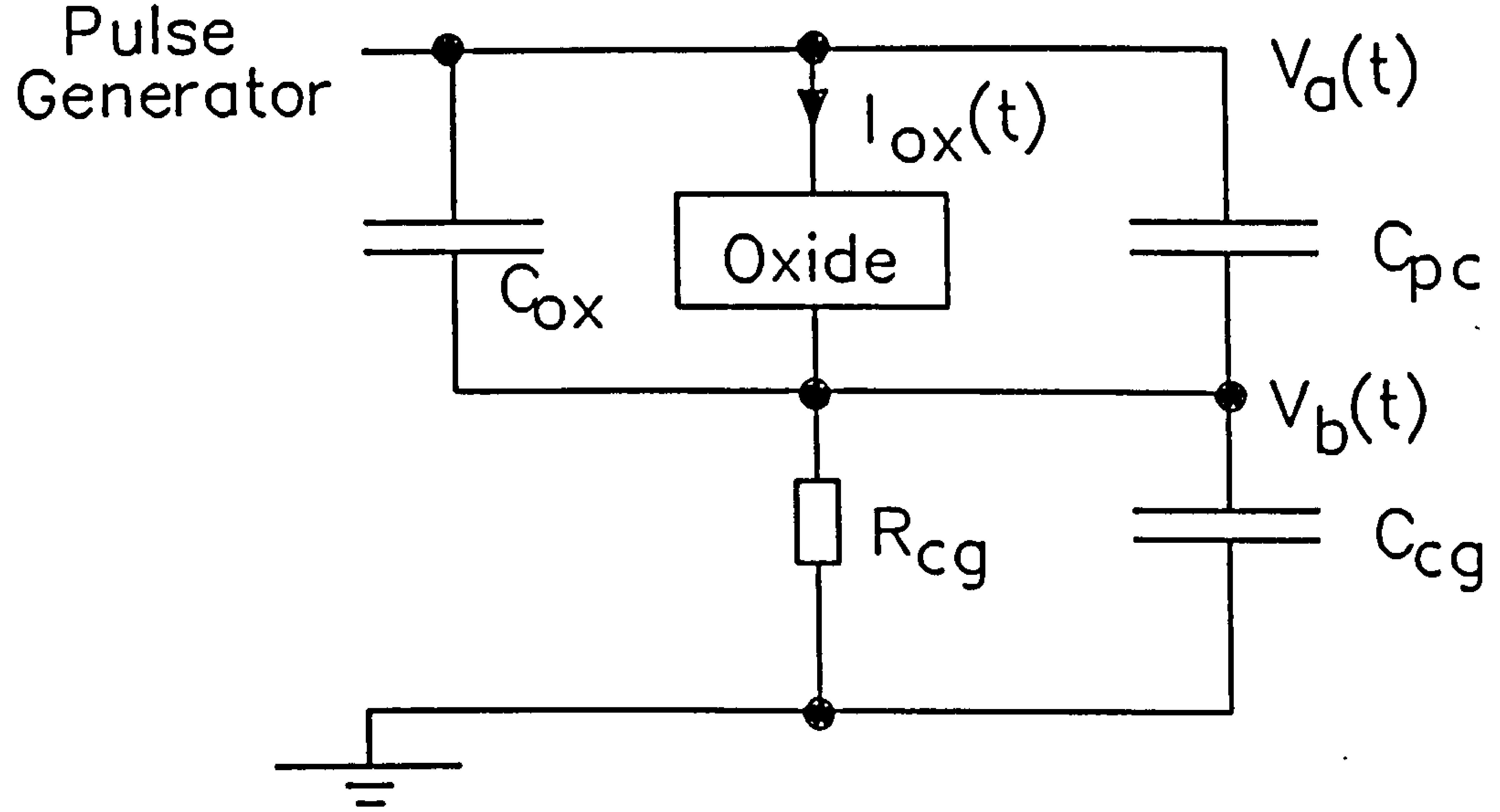


Figure 6.26: Equivalent circuit used for the analysis of negative polarity ramp data.

6.5.1 Negative Polarity

If a ramped negative gate stress is applied to a p-type MOS structure then accumulation layers appear at both oxide surfaces. Also, since the circuit current is only a few tens of μA[1], the voltage dropped across the bulk Si resistance (R_b=400Ω) may be neglected. Thus the entire device voltage may be assumed to appear across the oxide layer and the equivalent circuit of Fig.6.26 may be used to analyze the experiment.

If the oxide space-charge remains constant during the experiment (this may be assumed since the experimental time-domains lie mostly within the bounds of Region II [see Table 6.1]), then the oxide voltage profile $V_{ox}(t)$ and the injected charge profile $Q_{ox}(t)$ (=A. $\int J_{ox}.dt$) may be calculated from the formulae

$$V_{ox} = V_a - V_b \qquad 6(9)$$

$$Q_{ox} = \frac{1}{R_{cg}}\int_0^t V_b\,dt + C_{cg}V_b - (C_{ox}+C_{pc})(V_a - V_b) \qquad 6(10)$$

Figs.6.27 and 6.28 show the $Q_{bd}(t)$ profiles extracted from the 400Å oxide capacitor data.

Two of the samples (Nos.1 and 16) required much lower voltages and charges for breakdown than the others. These results probably indicate the presence of extrinsic oxide defects, and they are therefore excluded from the following analysis. The 10V/ms, 100V/ms

[1]. This approximation may be made by dividing V_b in Figs.5.24 and 5.25 by R_b.

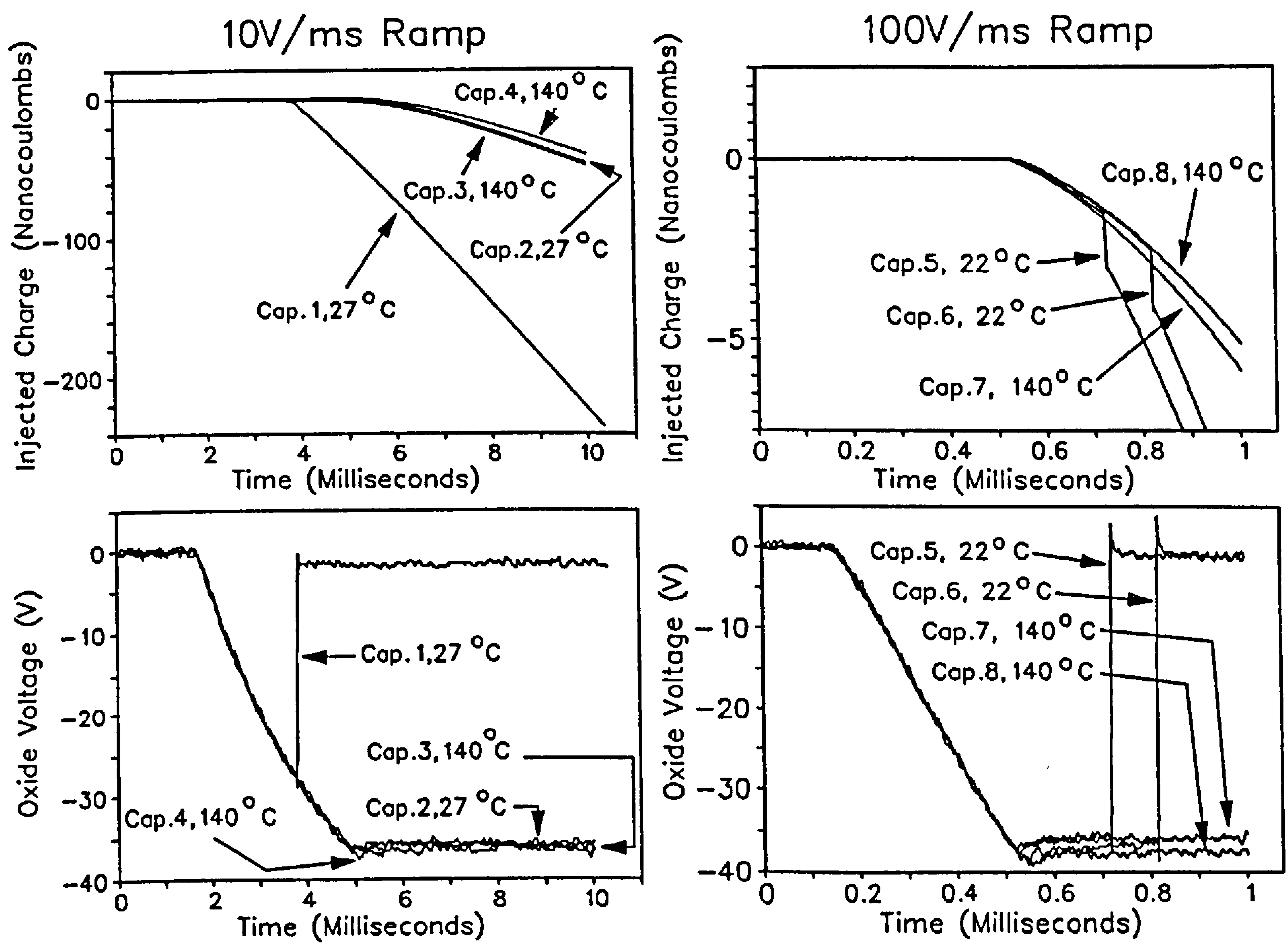


Figure 6.27: Reconstructed charge and voltage profiles for negative polarity ramp stress at 10 and 100V/ms.

and 1kV/ms data show oxide voltage pinning at approximately -37V, accompanied by a steady increase in Q_{ox}, indicating the onset of Fowler-Nordheim tunnelling. Pinning continues until breakdown, which is indicated by a rapid increase in Q_{ox} and a corresponding rapid decrease in V_{ox}. The value of Q_{ox} immediately prior to breakdown is equal to the breakdown charge Q_{bd}.

However, the 10kV/ms data show no evidence of voltage pinning and the $Q_{ox}(t)$ profiles are swamped beneath the system noise level. Furthermore, since the time domain associated with the 10kV/ms data (20μs) lies below the lower limit of Region II (see Table 6.1), the oxide space-charge might not be at a dynamic equilibrium and Eqn.6(10) may therefore be invalid. For this reason the 10kV/ms profiles may be unreliable and they are therefore excluded from the analysis.

The magnitude of the tunnelling current increases with increasing voltage transient speed (in accordance with the I=C.dV/dt law), causing a subsequent slight increase in $|V_{ox}|$. This causes Q_{bd} to fall in agreement with the constant current data of Fig.6.25.

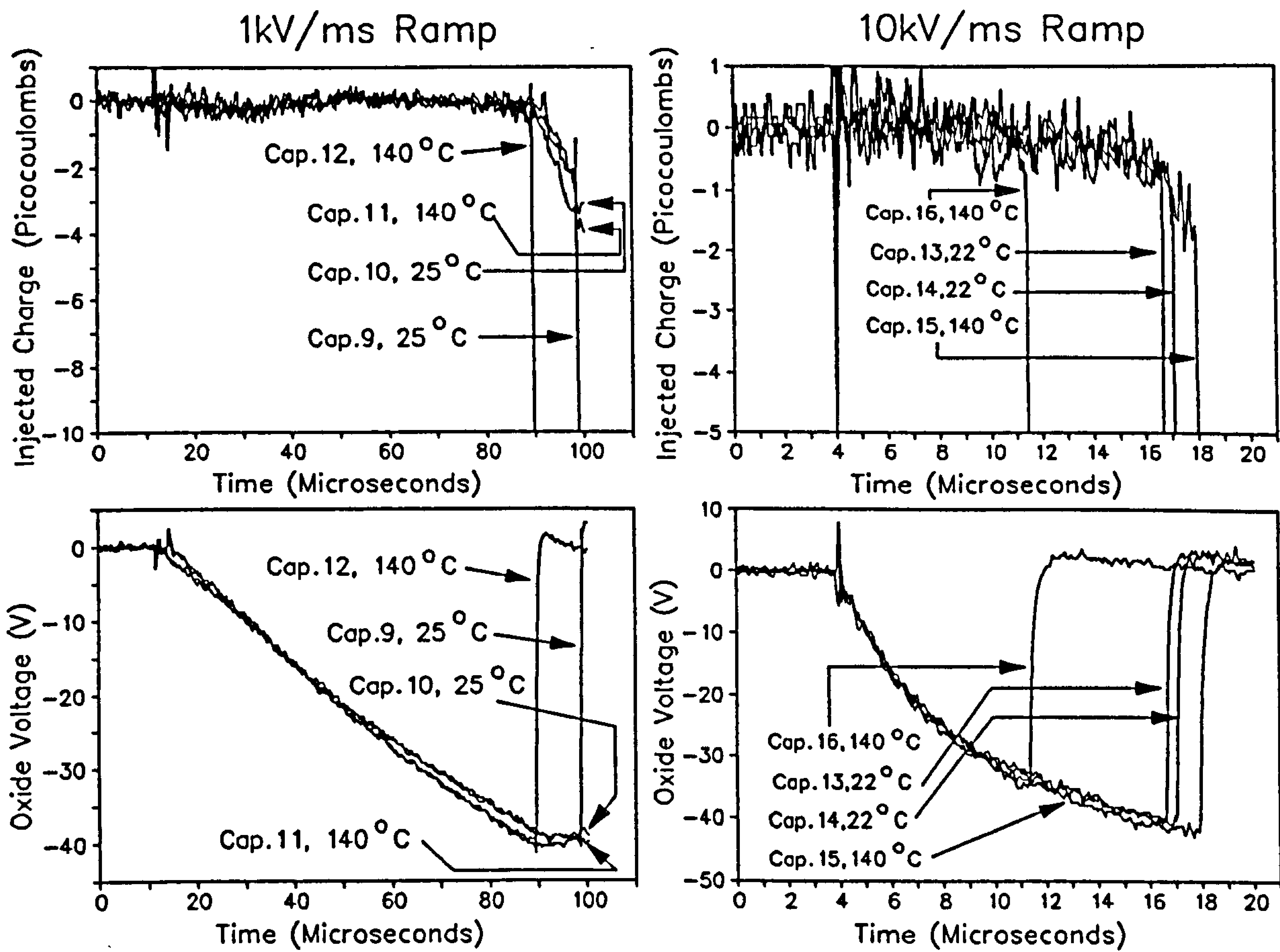


Figure 6.28: Reconstructed charge and voltage profiles for ramp voltage stress at 1 and 10kV/ms.

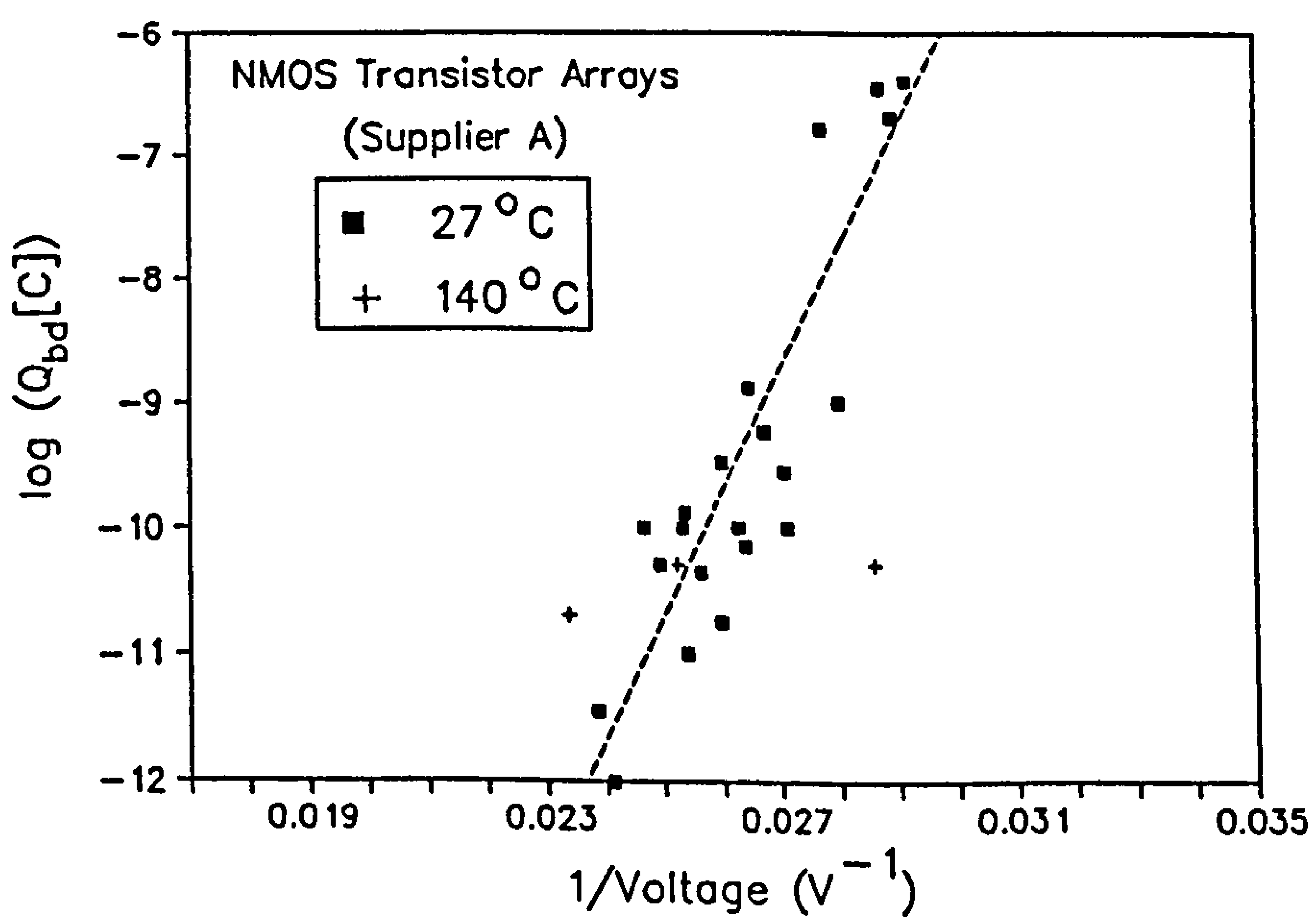


Figure 6.29: Q_{bd} vs. $1/V_{ox}$ data for NMOS transistor arrays, extracted from negative ramp-voltage profiles.

The ramp-voltage data may therefore be used to examine the Q_{bd} vs. V_{ox} relationship: When V_{ox}=-36V (Sample No.6), Q_{bd} is equal to -3nC. When V_{ox} increases in magnitude to −40V (Sample No.11), Q_{bd} falls in magnitude to -0.2nC. Although voltages and charges vary between samples, there is generally a rapid fall in Q_{ox} with increasing V_{ox}. This trend was investigated further by repeating the experiment with a large number of NMOS transistor arrays. Fig.6.29 shows the results plotted on a log(Q_{bd}) vs. 1/Field format. These results show that Q_{bd} can vary by six orders of magnitude over a voltage range of only 9V.

None of the above results show any evidence of temperature sensitivity of breakdown. The pinning voltages in Figs.6.27 and 6.28 are unaffected by temperature and the 140°C data points in Fig.6.29 are not significantly shifted from their room-temperature counterparts. However, the quantity of data is insufficient to draw any definite conclusions.

6.5.2 Positive Polarity

If positive gate stress is applied to a p-type capacitor, the p-type surface below the oxide becomes the cathode. A surface depletion layer initially forms in the cathode, which is later joined by an inversion layer along the silicon-oxide interface. Mobile electrons can then tunnel from the inversion layer into the oxide and cause breakdown.

The electrons in the inversion layer can be induced by thermal or optical carrier generation or by avalanche conduction in the depletion layer. The ramp time scales are too short to permit significant thermal/optical generation, whose time constants are of the order of 10ms-1s [23]. Inversion can therefore only result from avalanche conduction.

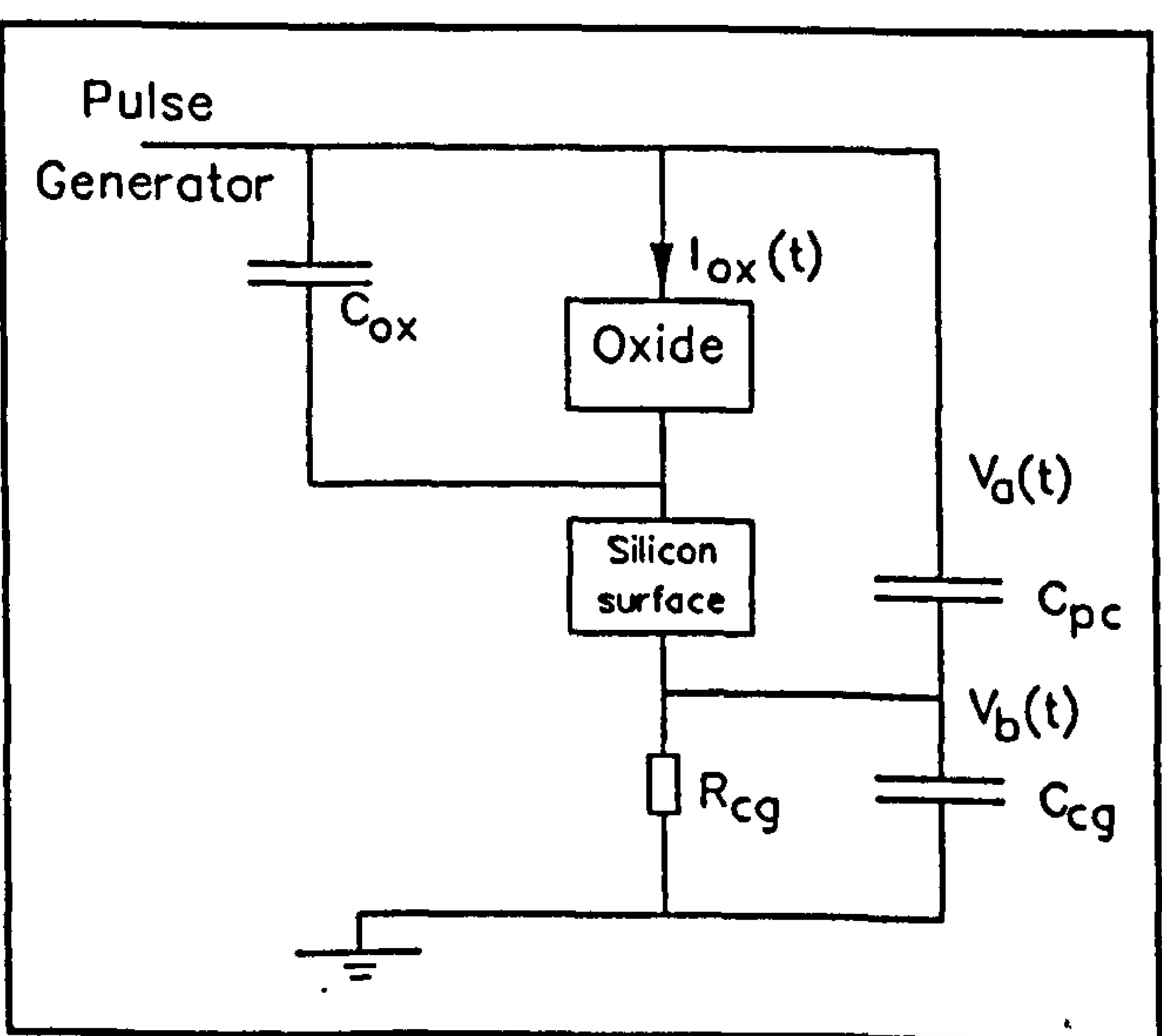


Figure 6.30: Positive polarity circuit model of slow transient system

Fig.5.26 shows that the current in the capacitor (represented by V_b) goes through three positive excursions. The first of these indicates the evacuation of holes from the cathode surface during depletion-layer formation. The second excursion indicates the onset of avalanche conduction and the third indicates tunnelling conduction in the oxide.

Fig.6.30 shows the circuit model used for the analysis of the positive polarity slow transient data. Throughout the experiment, the device voltage was shared between the depletion layer and the oxide (the inversion layer is practically 2-dimensional and makes no

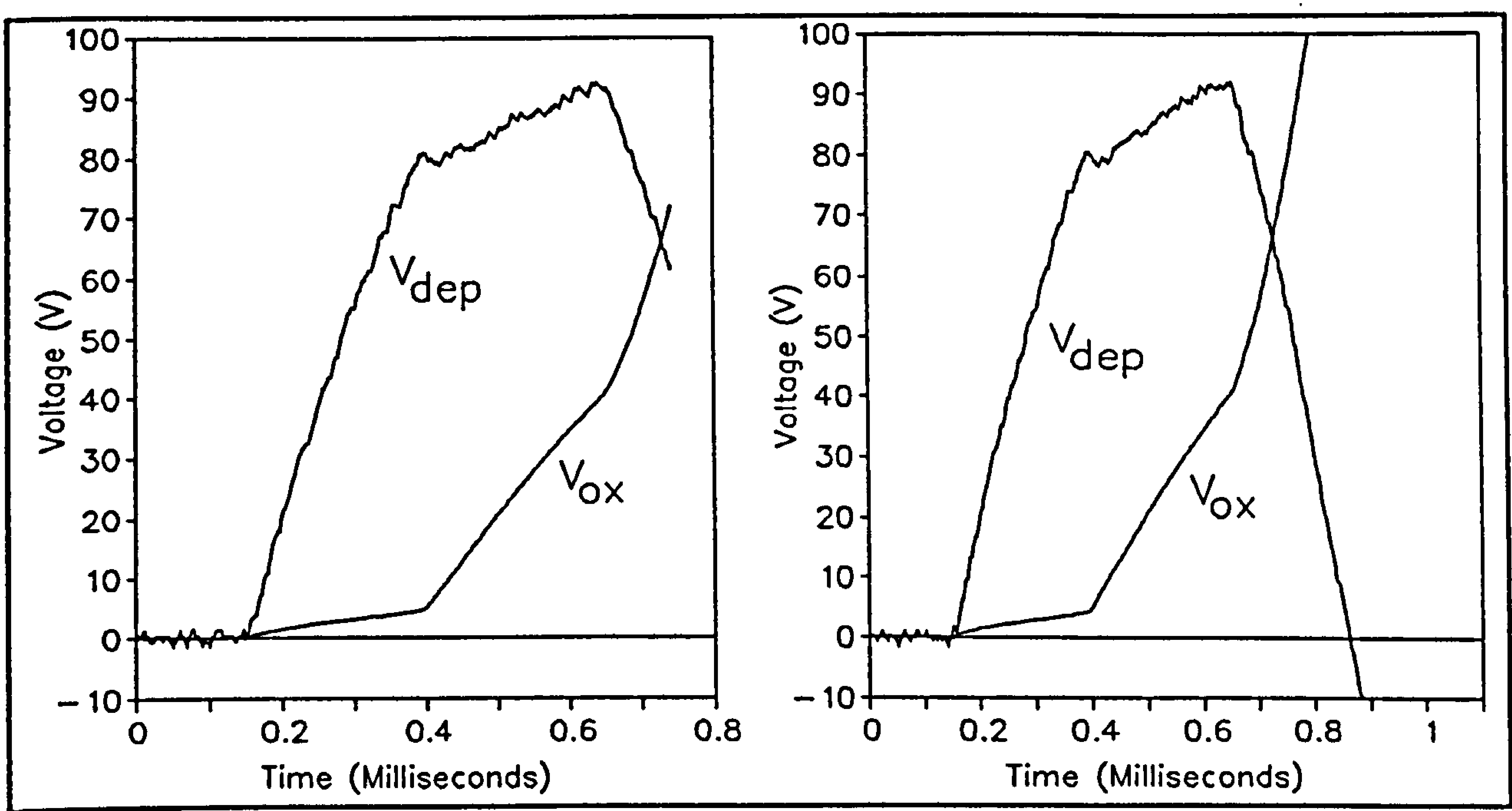


Figure 6.31: Reconstructed oxide and depletion layer voltages for unimplanted p-type MOS capacitors (Cap.17 on left, Cap.18 on right).

significant voltage contribution and the voltage drop across the bulk resistance is negligible).

If the magnitude of the oxide voltage is below the Fowler-Nordheim tunnelling threshold (i.e. the 37V 'pinning' voltage observed in Fig.6.27), the oxide behaves as an open circuit. Hence, by simple circuit theory, the oxide voltage V_{ox} is given by:

$$V_{ox} = \frac{V_b C_{cg}}{C_{ox}} + \frac{1}{C_{ox} R_{cg}} \int_0^t V_b \, dt - (V_a - V_b) \frac{C_{pc}}{C_{ox}} \qquad 6(11)$$

and the voltage V_{dep} across the surface depletion layer is given by

$$V_{dep} = V_a - V_b - V_{ox} \qquad 6(12)$$

Fig.6.31 shows the V_{ox} and V_{dep} curves re-constructed using this technique. The depletion layer voltage clearly pins close to a value V_{av} = 75V for all the devices tested and this value indicates the silicon avalanche threshold. However, the avalanche thresholds appear increase by about twenty per cent during the course of the waveform, rising from 75V to 91V over a period of 300μs. This is believed to be caused by Joule heating in the localised avalanche filaments or 'microplasmas' (which are probably located at the capacitor corners where the fields are at their maximum).

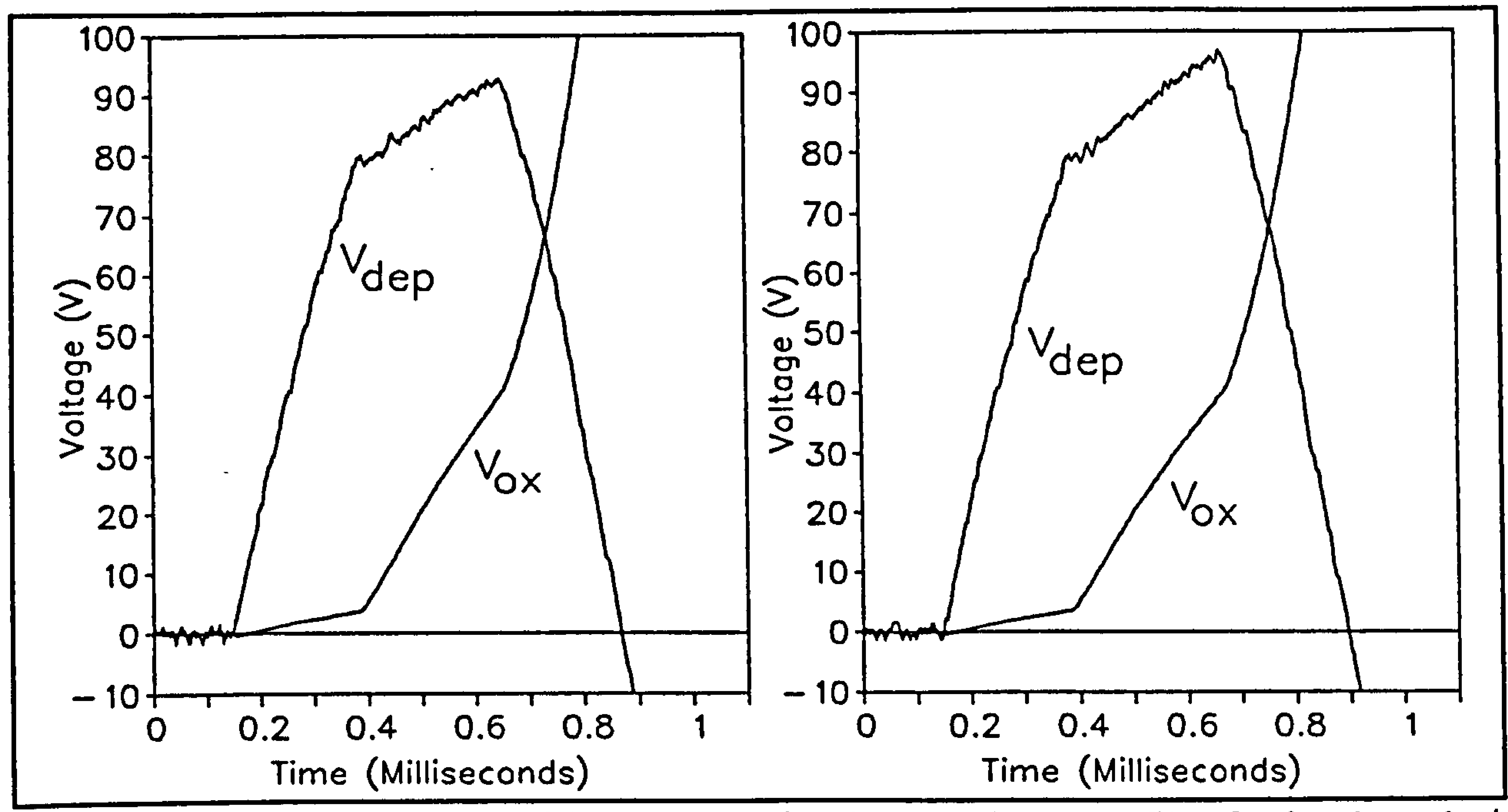


Figure 6.32: Reconstructed oxide and depletion layer voltage profiles for implanted p^+ type MOS capacitors (Cap.19 on left, Cap.20 on right).

The value of V_{ox} just prior to the beginning of the third (or tunnelling) current excursion indicates the voltage threshold for Fowler-Nordheim tunnelling. This threshold is clearly around 41V which is slightly higher than its negative polarity counterpart (Section 5.1.1). [Note: The sections of the V_{ox} and V_{dep} curves after the onset of tunnelling have no significance since Eqn.6(12) does not apply in the presence of tunnelling].

Earlier workers [24] have suggested that electron injection into the oxide directly from the hot avalanche plasma could be a contributory factor to breakdown. (This 'avalanche injection' mechanism is described in Section 3.4.3.) These waveforms fail to show any significant oxide current prior to tunnelling and therefore indicate that avalanche injection is not significant. (Avalanche injection would be exhibited as a decay in the apparent avalanche threshold with time). It therefore appears that for both positive and negative polarity cases, breakdown is the result of Fowler-Nordheim tunnelling.

6.5.3 Analysis of ESD Transient Data

After the apparently successful use of simple circuit theory for the analysis of the ramp voltage data, the same techniques were applied to ESD transient waveforms. Data were obtained from four unimplanted NMOS capacitors (Supplier A) which were subjected to standard HBM ESD pulse stress. Two of the samples (denoted 19 and 20) were pulsed at -70V, while the others (20 and 21) were stressed at +240V. The current and voltage waveforms were monitored by the Tectronix CT1 and HP10431A (probe A) probes and captured on the HP54111D digital oscilloscope. The waveforms were numerically processed using the techniques described in Section 4.2.2.5 and Fig.6.33 shows the resulting profiles. Breakdown in all four devices is clearly indicated by a rapid rise in current accompanied by a rapid fall in device voltage. The following sections describe the numerical analysis of these waveforms.

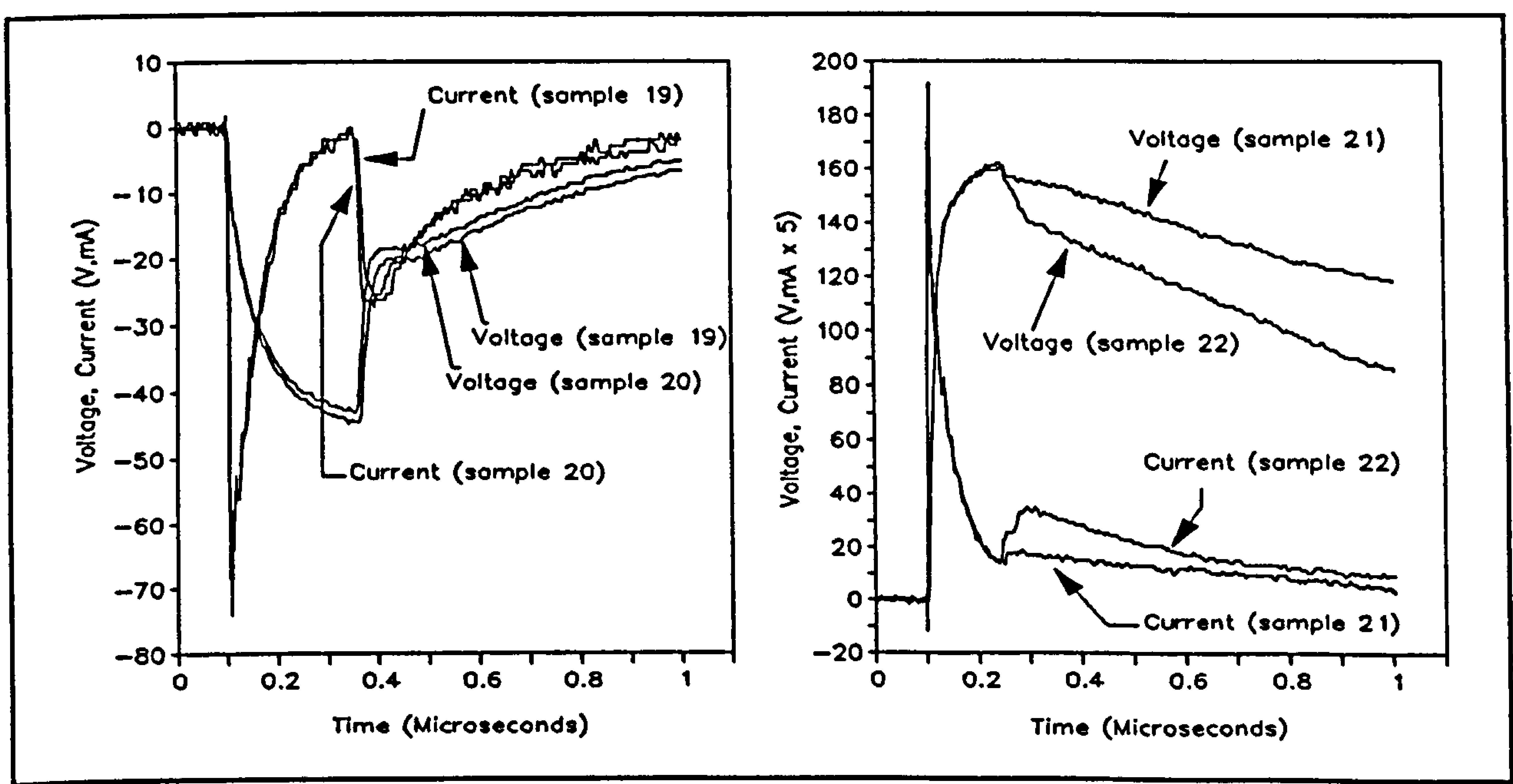


Figure 6.33: ESD waveforms obtained from unimplanted 400Å NMOS capacitors

6.5.3.1 Negative Polarity Analysis

Fig.6.34 shows the equivalent circuit used for the analysis of the negative ESD transient data. The device parameters (C_{ox} and R_b) were determined in Section 6.5 while the oscilloscope probe capacitance $C_p(A)$ was determined in Section 4.2.3.1. The probe-chuck capacitance C_{pc} (2pF) was measured using the Wayne-Kerr 4120 LCR bridge. Since the insertion impedance of the Tektronix CT1 probe was only of the order of 1Ω, it was excluded from the model. Also, since the total charge lost via the 1MΩ probe resistance during the pulse rise-time (about 1μs) is negligible, R_p was also excluded.

According to this circuit model, the following expressions may be used to determine

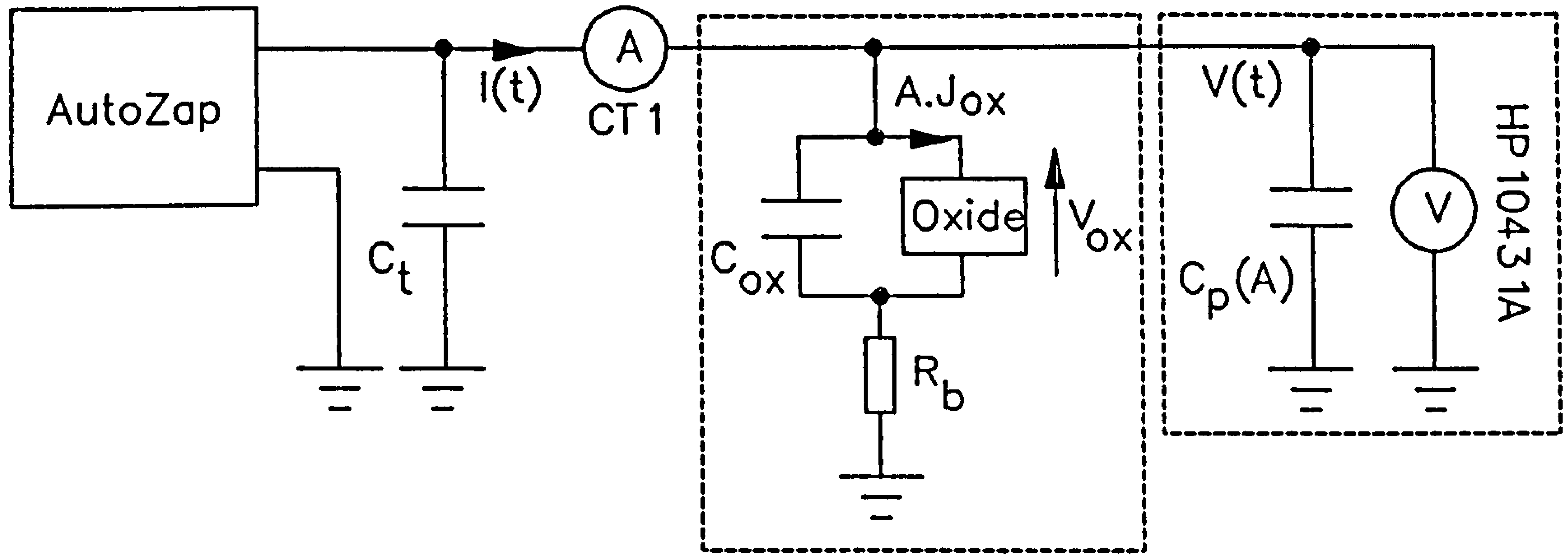


Figure 6.34: Negative-polarity circuit model used for the analysis of ESD transient data.

the oxide voltage V_{ox} and injected charge Q_{ox} profiles:

$$V_{ox}(t) = V(t) - R_b\left(I(t) - \left[C_{pc} + C_p(A)\right]\frac{dV}{dt}\right) \qquad 6(13)$$

$$Q_{ox}(t) = \int_0^t I(t)\,dt - \left(C_p(A) + C_{pc}\right)V(t) - C_{ox}V_{ox}(t) \qquad 6(14)$$

Fig.6.35 shows the resulting waveforms for Sample 19 (the waveforms obtained from sample 20 were practically identical to these). The injected charge Q_{bd} remains practically zero until the point of breakdown, at which point it rises rapidly. Also there is no evidence of pinning in the voltage profile, whose magnitude continues to rise until the point of breakdown. This is consistent with the theory of breakdown in Region III (see Section 6.4), where breakdown is governed by the tunnelling time-delay t_d. When t_d has elapsed and tunnelling begins, the oxide wears out over an infinitesimally short period of time and therefore appears to suffer instantaneous breakdown.

6.5.3.2 Positive Polarity Analysis

Since positive polarity ESD breakdown is believed to require depletion layer avalanche conduction, the device model in Fig.6.34 must be modified for positive polarity analysis. Fig.6.36 shows the modified circuit model, including a constant voltage source to

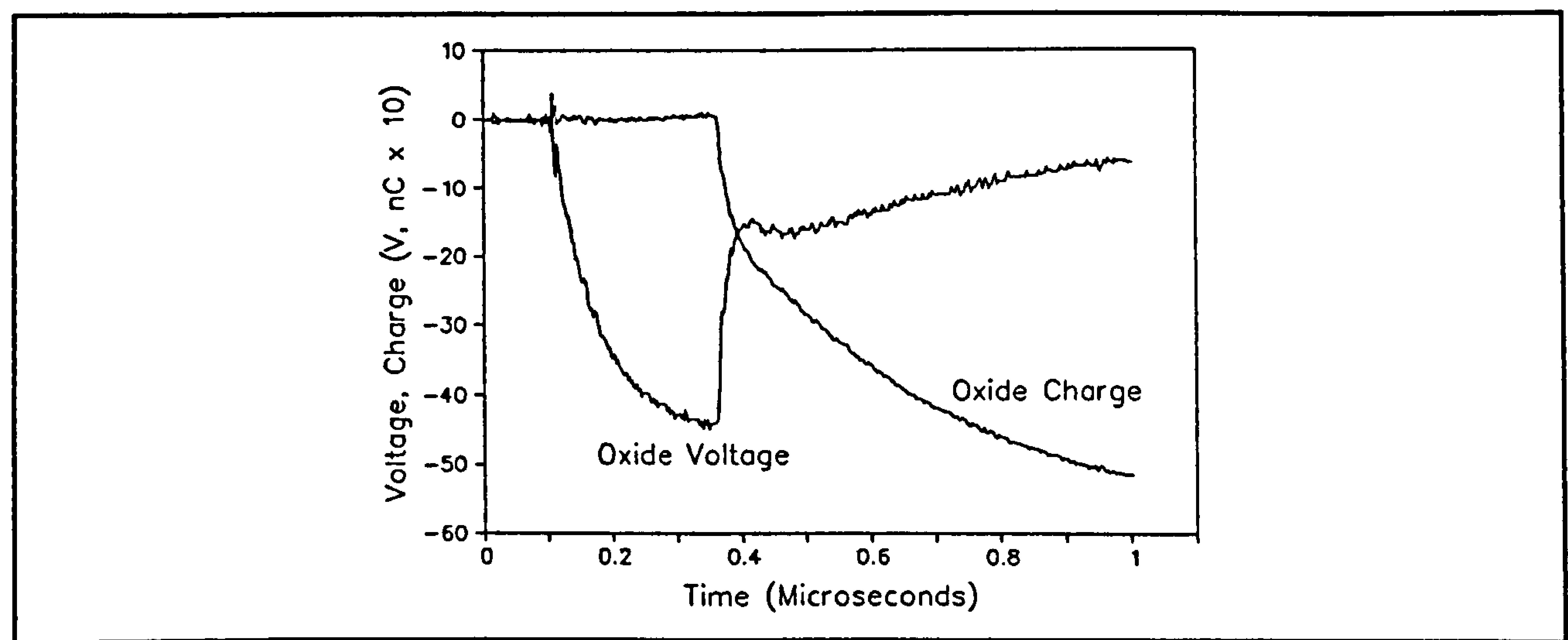


Figure 6.35: Q_{ox} and V_{ox} profiles extracted from ESD transient data

represent the pinning of the depletion-layer voltage at the avalanche threshold V_{av}. Prior to the onset of oxide tunnelling, the oxide voltage profile is given by

$$V_{ox} = \frac{1}{C_{ox}} \int_0^t I(t)\, dt - \frac{(C_p(A) + C_{pc})}{C_{ox}} V(t) \qquad 6(15)$$

and the total voltage V_{ser} in series with the oxide (avalanche voltage V_{av} plus bulk silicon voltage V_b) is given by

$$V_{ser} = V_{av} + V_b = V - V_{ox} \qquad 6(16)$$

Fig 6.37 shows the pre-breakdown V_{ox} and V_{ser} profiles extracted from the experimental data of sample 22 using these equations. (The profiles obtained from sample 21 were very similar to these). V_{ser} is clearly significantly larger than the values of V_{av} observed in Figs.6.31 and 6.32, and its magnitude decays with time (unlike Fig.6.31-2). This probably indicates a large voltage drop across the bulk resistance R_b, which decays as the displacement current falls. This hypothesis may be tested by calculating the product of the bulk resistance R_b and the device current at the point of maximum V_{ser} (about 0.15μs into the pulse). Since the current is equal to $C_{ox}.dV_{ox}/dt$=0.25mA, the result yields 400 x 43.10^{-3} = 17.3V, which is somewhat too small to account for the anomaly. However, the

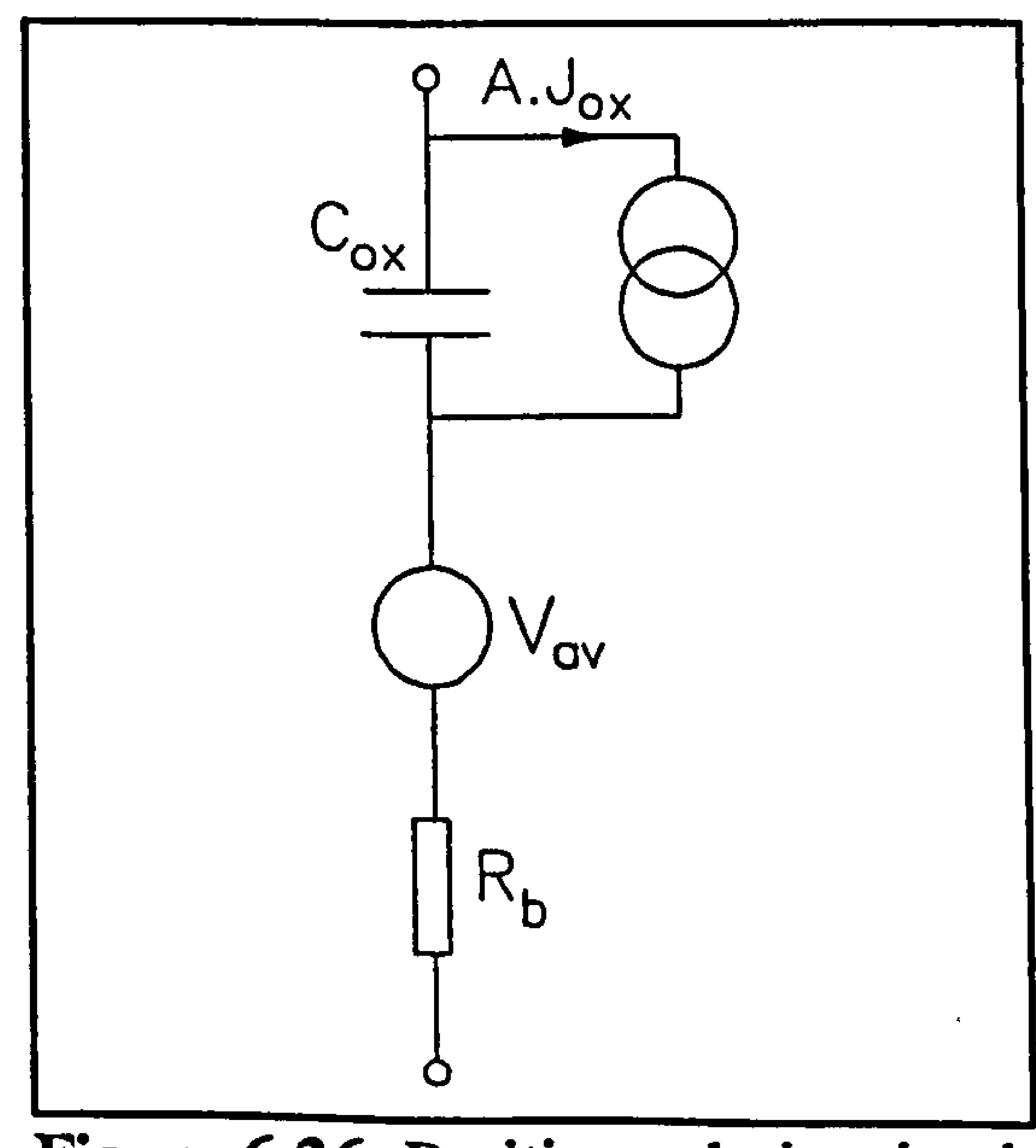


Figure 6.26: Positive polarity circuit model for MOS capacitor

localised nature of the avalanche filaments may give rise to an additional constriction resistance, increasing the effective value of R_b to a value considerably greater than the 400Ω measured under negative polarity conditions. Fig.6.37 shows that as the device current (dictated by dV_{ox}/dt) decays, V_{scr} also decays to a value of about 90V. This agrees approximately with the maximum value of V_{av} observed in Fig.6.31.

The oxide voltage profile $V_{ox}(t)$ of Fig.6.37 shows that the oxide voltage rises as high as 70V prior to the onset of tunnelling. The voltage exceeds the 41V tunnelling threshold (Fig.6.31) for a period of approximately 0.1μs prior to breakdown, and this time corresponds approximately to the lower limit of the t_{bd} vs. F data in Fig.5.30. The field corresponding to 70V in a 40nm thick oxide is equal to 17.5MV/cm, which is considerably larger than the maximum field in Fig.5.30 (11.6MV/cm). This shows that the high-field saturation of t_{bd} (Region III) continues far beyond the 11.6MV/cm limit of Fig.5.30.

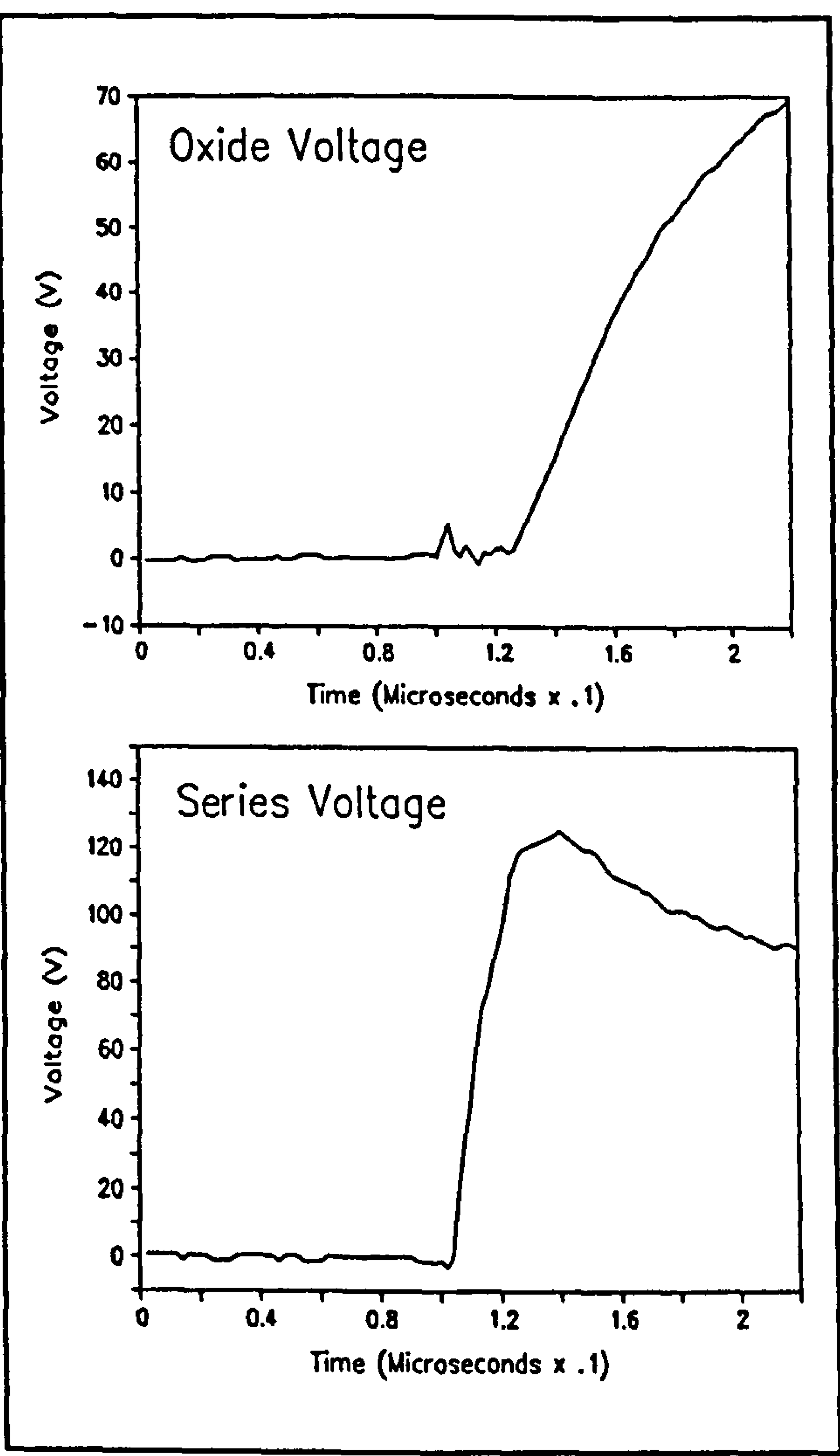


Figure 6.37: V_{ox} and V_{scr} profiles extracted from positive ESD transient data

6.6 Qualitative Model of Breakdown

This section develops a physical model of oxide dielectric breakdown. The model is based upon established solid-state theory and is supported by the experimental work of this thesis. The model attempts to explain all the observed breakdown data (ie. constant voltage, ramp voltage, constant current and ESD) in terms of a single set of physical mechanisms. Although equations are sometimes used in order to illustrate principles, the treatment is generally qualitative. The quantitative development of the model appears in Chapter 7.

6.6.1 Space Charge Evolution in Oxide

Section 6.4 showed how the three time domains in the t_{bd} vs. F curve could be explained in terms of dynamic space-charge perturbations in the SiO_2. Short time-scale behaviour (Region III) was associated with a positive dQ/dt while the long time-scale behaviour (Region I) was attributed to a negative dQ/dt. Over intermediate time-scales (100μs-1s), the oxide charge was proposed to be at a dynamic equilibrium (ie. $dQ/dt \simeq 0$).

Oxide space-charge evolution may be due to the drift of ionic contaminants and/or charge-carrier trapping in SiO_2 defect states. Although many early workers concentrated on the ionic drift model [25,26], their oxide samples had a far greater ionic impurity content than those produced by present-day processing. The following theory is therefore based upon the charge-trapping model.

The concept of carrier trapping has already been introduced in Section 2.2, where the properties of the major trap species were described. Thermally-grown silicon dioxide contains an inherent density of *pre-existing* (or *virgin*) trap states, to which additional *newly-generated* traps are added under current/voltage stress [19,27]. These traps, which exist in both the bulk oxide and the Si/oxide interfaces [28], can be either *donors*, (donating electrons to become positively charged) or *acceptors* (accepting electrons to become negatively charged) [29].

The mechanism by which the traps are produced has been disputed for some years. It has been argued that energies sufficient to damage the SiO_2 network (producing traps associated with broken Si-O bonds) can only be produced by electrons relaxing from the oxide conduction band [30]. Such relaxation may be caused by several mechanisms, including emission at the anode, capture in pre-existing traps [30] and recombination with trapped holes [31]. An alternative model was proposed by DiMaria and Stasiak [32] who suggested that hydrogen-related molecules, liberated from extrinsic defect sites near the anode, move toward the cathode where they interact with the SiO_2 network to form trapping sites.

A different philosophy maintains that at fields above 1.5MV/cm, the trap generation rate depends upon the number of 'hot' electrons in the oxide [32-34]. Under these conditions, electrons can reach energies as large as 4eV, while only 2.3eV are required for trap creation [32]. The trap generation rate is sometimes modelled as a constant [14,19] and sometimes as an exponential function of the electric field in the oxide [27].

The effect of these traps upon the Fowler-Nordheim tunnelling characteristics have been studied by many workers [eg.14,19], most of whom have identified bulk oxide traps (both virgin and newly-generated) as the main protagonists. Recent work by Changhua et al. [28] has detected the action of interface traps, although it is only the newly-generated states which have any significant effects. For the sake of simplicity however, the present work deals exclusively with bulk-oxide traps, and assumes that their density remains constant

during stress.

Many workers have postulated that two distinct bulk-trap species operate simultaneously in the oxide during injection [35,38]. The experimentally determined capture cross-sections of these traps identify them as neutral trap states (NTS) and coulombic repulsive traps (CRT) respectively (see Table 2.1). NTS are known to exhibit trapping/de-trapping (T-D) behaviour [27,35], allowing them to settle towards an equilibrium occupancy level which is a function of the injection current and the oxide electric field. This is sometimes called the *dynamic balance model* [27]. If N_1 is the NTS trap density, n_1 is the density of *filled* NTS, σ_1 is the NTS capture cross section and τ_e is the NTS de-trapping time constant then the trapping rate can be modelled:

$$\frac{dn_1}{dt} = \frac{J\sigma_1}{q}\left(N_1 - n_1\right) - \frac{n_1}{\tau_e} \qquad 6(17)$$

where J is the injection current density. (This equation is based upon Eqn.2(3) in Chapter 2 and assumes that the thermal and drift velocities of the electrons are equal.)

However, since the coulombic repulsive traps do not easily release their electrons at room temperature (ie. $\tau_e \rightarrow \infty$), they do not exhibit T-D behaviour [36] and the trapped space-charge continues to rise until the point of 100% occupancy. Hence if N_2 represents the CRT density, n_2 represents the *occupied* CRT density and σ_2 represents the CRT capture cross-section, then the CRT trapping rate can be modelled by the equation

$$\frac{dn_2}{dt} = \frac{J\sigma_2}{q}\left(N_2 - n_2\right) \qquad 6(18)$$

Both of the trap species appear as donors and acceptors. For the purposes of this simple discussion it will be assumed that N_1 and N_2 remain constant during stress. It will also be assumed that half of the total trap population are donors, while the other half are acceptors and that when the oxide is at equilibrium, all the donors are full, while all the acceptors are empty. Hence the conditions $n_1(0)=N_1/2$, $n_2(0)=N_2/2$ exist at the instant the stress is applied. These boundary conditions allow Eqns.6(17) and 6(18) to be solved, yielding the following equations for $n_1(t)$ and $n_2(t)$:

$$n_1(t) = \frac{\alpha}{\beta} - \left(\frac{\alpha}{\beta} - \frac{N_1}{2}\right)e^{-\beta t} \quad where \quad \alpha = N_1\frac{J\sigma_1}{q} \;;\; \beta = \frac{J\sigma_1}{q} + \frac{1}{\tau_e} \qquad 6(19)$$

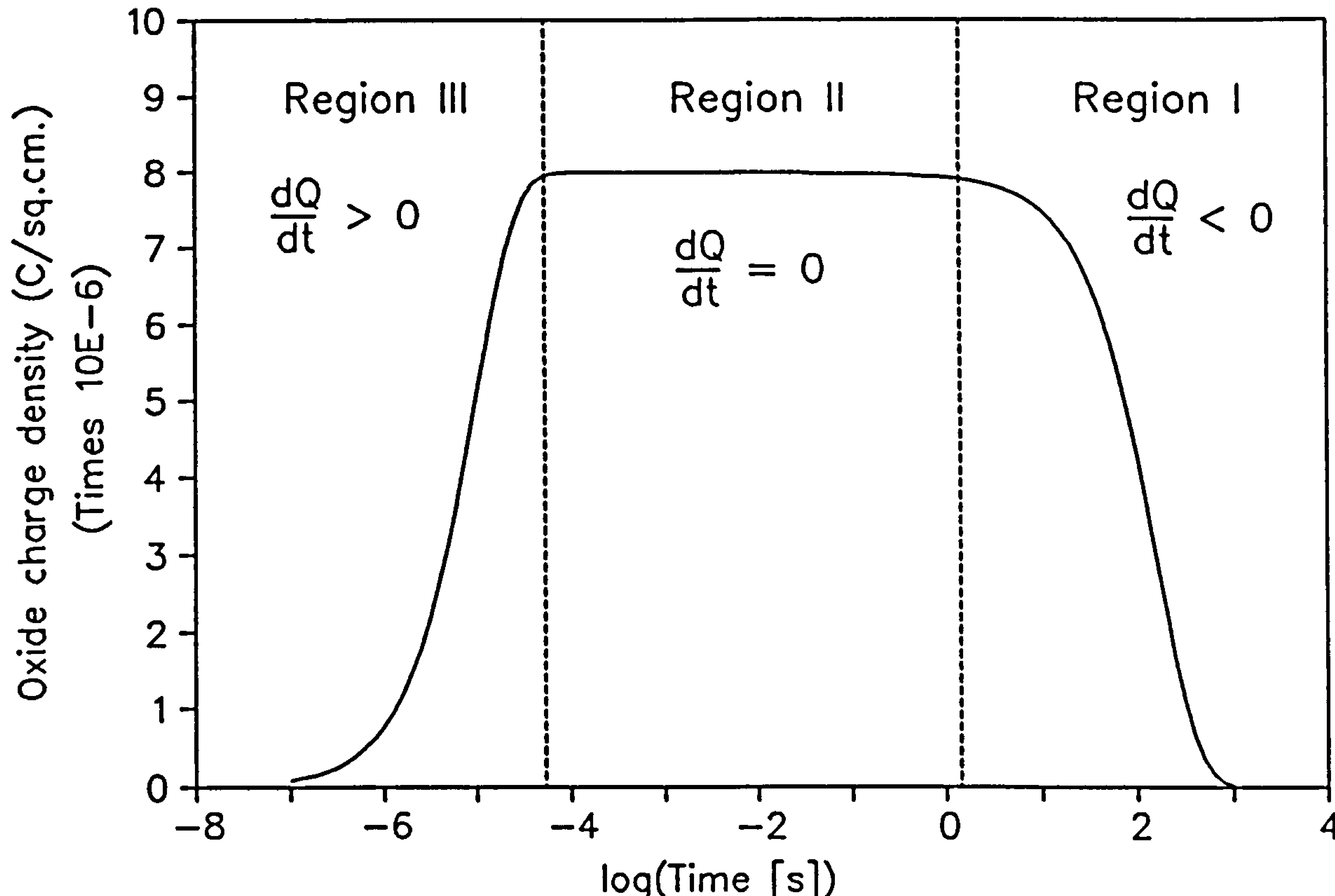


Figure 6.38: Q(t) vs. log(t) profile predicted using first-order trapping rate theory.

$$n_2(t) = N_2\left(1 - \frac{1}{2}\exp\left[-\frac{J\sigma_2}{q}t\right]\right) \qquad 6(20)$$

Assuming that the oxide is neutral at equilibrium trap-occupancy, the space-charge Q can be written

$$Q = -q\left[(n_1 + n_2) - \frac{(N_1 + N_2)}{2}\right] \qquad 6(21)$$

Thus the Q(t) profile was determined across the full time-scale of Fig.5.30, using the following typical parameter values: $N_1=10^{14}cm^{-2}$, $N_2=10^{14}cm^{-2}$, $\tau_e=10\mu s$ [37], $\sigma_1=10^{-16}cm^2$, $\sigma_2=10^{-19}cm^2$ (see Table 1.1) and $J=10mA/cm^2$. Fig.6.38 shows the resulting Q vs. log(t) profile.

It is clear from Fig.6.38 that three distinct time-domains exist, and that their boundaries correspond roughly to those of Regions I, II and III in Fig.6.13: In Region I, the NTS donor traps empty, causing the oxide to become more positive (dQ/dt>0). At the same time however, NTS fill up with injected electrons until eventually the trapping and de-trapping rates are equal. At this point, the oxide charge is at a dynamic equilibrium and

dQ/dt=0. However, as the stress continues, the acceptor CRT traps begin to fill, causing the oxide to become less positive (dQ/dt<0). This leads to Region I behaviour.

It must be stressed that the above model is greatly simplified and serves merely to show how two independent trapping mechanisms can give rise to the observed behaviour. In reality, the model parameters are likely to have complicated field/current dependencies and the total trap density increases with time due to the generation of defect sites.

6.6.2 Mechanisms of Oxide Wearout

Several theories of SiO_2 dielectric wearout have been reviewed in Section 3.3. The purpose of the present section is to determine which (if any) of these models adequately explains the experimental data of this thesis. The alternative models are summarised below:

(1) **Treeing Model:** Defect sites in the SiO_2 structure capture mobile electrons, whose excess energies are absorbed in the creation of further defects. Clusters of defects therefore grow during stress, forming 'tree' structures of high conductivity. When a 'tree' spans the oxide, it channels a large current through the dielectric, resulting in breakdown [30,39].

(2) **Electron Trapping Model:** Electrons tunnel into the oxide and are trapped in defect sites, causing internal field enhancement. When a critical field is reached, the Si-O bonds rupture, creating new defect states and permitting further trapping [40]. Alternatively, the enhanced field may be sufficient to support avalanche breakdown in the oxide [41].

(3) **BBII Hole Trapping Model:** Electrons tunnel into the oxide and are accelerated by the electric field. Some electrons gain sufficient energy to create electron-hole pairs by band-to-band impact ionisation (BBII). The holes drift back towards the cathode where some of them are trapped in surface defects. The subsequent positive space-charge enhances the cathode electric field, increasing the tunnelling current, causing the process to accelerate towards breakdown [14].

(4) **Surface Plasmon Model:** This model is identical to the BBII model, except that the oxide holes are generated by mobile electrons releasing their energy at the oxide anode. The emitted electrons transfer their energy to surface plasmons (SPs), which

quickly decay. The energy released by the SP decay results the creation of high energy electron-hole pairs. Some of the generated holes then tunnel back into the oxide [20,42].

(5) **Ion Activation Model:** Microscopic asperities along the cathode-oxide interface create localised filaments of tunnelling current. These filaments create high localised temperatures, which may be sufficient to free static impurity cations from their trapping sites. Once liberated, these ions drift toward the cathode where their positive space-charge enhances the tunnelling current and eventually causes breakdown [43].

(6) **Resonant Tunnelling Model:** If an oxide defect state in the Fowler-Nordheim barrier has an energy equal to a conduction-band state in the cathode, the two states can 'resonate'. Superposition of eigenfunctions shows that the barrier has a unity transmission coefficient and a short circuit therefore exists across the barrier. A large number of resonating states can therefore cause breakdown [44].

The electron trapping model (Model 1) fails to explain breakdown in Regions II and III, since the effects of negative charge trapping are only apparent in Region I (see Section 6.6.1). Model 5 can also be rejected on the grounds that Q_{bd} is approximately proportional to oxide area (Fig.5.31), suggesting that the injection current is homogeneously distributed rather than confined to a localised defect spot. The current profiles observed during stress showed no evidence of the random nondestructive breakdown events associated with resonant tunnelling breakdown [45], so this model was also rejected.

The most likely theories are therefore the positive charge related models (Models 3 and 4) and the treeing model (Model 1). For the purposes of this project, the positive charge model was used as a working assumption. This choice was made because:

(1) The charge-separation experiments of Chen et al. [46] shows that a constant generated hole-charge is required to support breakdown.

(2) MOS devices subjected to sub-catastrophic ESD stress display negative shifts in their C-V characteristics, indicating the presence of trapped positive charge [47]. The magnitude of this charge has been shown to increases until the point of breakdown [48].

(3) From the point of view of analytical modelling, hole generation and trapping are far simpler than tree propagation.

A choice must therefore be made between Models 3 and 4. The fact that breakdown occurs at voltages below the SiO_2 energy-gap potential [44] indicates that classical band-to-band impact ionisation is not the dominant mechanism. The theory that ionisation interactions take place between tail states [49] is also questionable, since these semi-localised levels may only extend a few meV into the oxide forbidden gap [50].

The fact that wearout is accelerated by the negative charge-trapping in Region I suggests that the anode (rather than the cathode) electric field is the critical parameter determining breakdown. This suggests Model 4, in which the rate of wearout is dictated by the anode hole injection rate, which depends in turn upon the anode electric field. According to the Fowler-Nordheim tunnelling theory, the electron injection rate, and hence the anode hole generation rate, is approximately proportional to $\exp(-\beta_1/F_{cat})$ where β_1 is a constant and F_{cat} is the cathode electric field. Similarly the probability of a hole entering the oxide at the anode is proportional to $\exp(-\beta_2/F_{an})$, where β_2 is a constant and F_{an} is the anode electric field. The positive charge generation rate, and hence the oxide wearout rate, is therefore proportional to $\exp(-\beta_1/F_{cat}-\beta_2/F_{an})$. It can therefore be seen that t_{bd} is proportional to $\exp(\beta_1/F_{cat}+\beta_2/F_{an})$. Since the charge-to-breakdown Q_{bd} is equal to the product of t_{bd} and the electron injection current ($\propto \exp(-\beta_1/F_{cat})$), Q_{bd} must be proportional to $\exp(\beta_2/F_{an})$. Hence the time and charge-to-breakdown decrease rapidly with increasing voltage, as observed in Figs.5.30 and 6.29.

In Region III of the t_{bd} vs. F curve, oxide wearout cannot begin until the oxide has become sufficiently positive to support cathode electron injection. However, since the wearout rate is so rapid at the fields associated with Region III (>9.5MV/cm), the oxide breaks down almost instantaneously once tunnelling begins. The time-to-breakdown t_{bd} is therefore dictated by the tunnelling time delay t_d (as shown in Figs.6.16 and 6.21).

While the wearout mechanism has no inherent dependence upon luminous intensity, elevated temperatures might possibly support cathode hole de-trapping, causing t_{bd} to decrease with increasing temperature (see Fig.5.27).

In view of its sound theoretical basis and its agreement with the observed data, this model is used as a working assumption for the remainder of this thesis.

6.6.3 Model of ESD Breakdown

In the following discussions, the theory developed above for constant voltage/current breakdown is extended to model failure under ESD conditions. The ESD breakdown data to be modelled are summarised below:

1. The oxide area dependent data for CMOS and NMOS devices (Section 5.2.2).
2. The temperature dependent data for CMOS and NMOS devices (Section 5.2.3).
3. The illumination dependent data for NMOS, CMOS, HMOS and SOS devices (Sections 5.2.4 and 6.3.1).
4. The body-capacitance dependent data for NMOS & CMOS devices (Section 5.2.5).
5. The body-resistance dependent data for NMOS & CMOS devices (Sections 5.2.5 and 6.3.4).

The discussion begins by considering ESD breakdown under negative polarity conditions. The model is then extended to embrace positive polarity breakdown.

6.6.3.1 Negative Polarity Model

Fig.6.34 shows the circuit model for an NMOS device under negative ESD gate stress. When a pulse of magnitude V_0 is applied, the charge $(C_x+C_1)V_0$ is distributed across the entire system capacitance $(C_x+C_1+C_t+C_{ox})$, attenuating the pulse magnitude by the 'capacitive loading' factor $L=(C_x+C_1)/(C_x+C_1+C_t+C_{ox})$ [8,11,51]. In order to support oxide breakdown, the pulse charge must clearly exceed the oxide's charge-to-breakdown Q_{bd}, which the above theory shows to decrease with increasing oxide field. Fig.6.39 superimposes the range of charges associated with ESD pulses between 35V to 65V (these values were taken from Fig.6.11) upon the Q_{bd} vs. $1/V_{ox}$ data of Fig.6.29. It is graphically shown that if all the ESD charge is injected into the oxide then the oxide voltage required to cause breakdown must be in the range 35.5V to 42.4V. However, the variation within this range is due largely to statistical fluctuation rather than to the variation in pulse charge. Furthermore, much of the variation may be due to the error of estimating the exact value of V_{ox}, which has a tendency to fluctuate during stress (see Fig.6.27). For this reason, calculations were based upon the mean voltage of the range 38.45V. This is denoted V_e^*, the 'effective' breakdown voltage of the oxide, and it corresponds approximately to the Fowler-Nordheim tunnelling threshold (denoted V_{FN}).

According to the 'capacitive loading' theory described above, the applied pulse

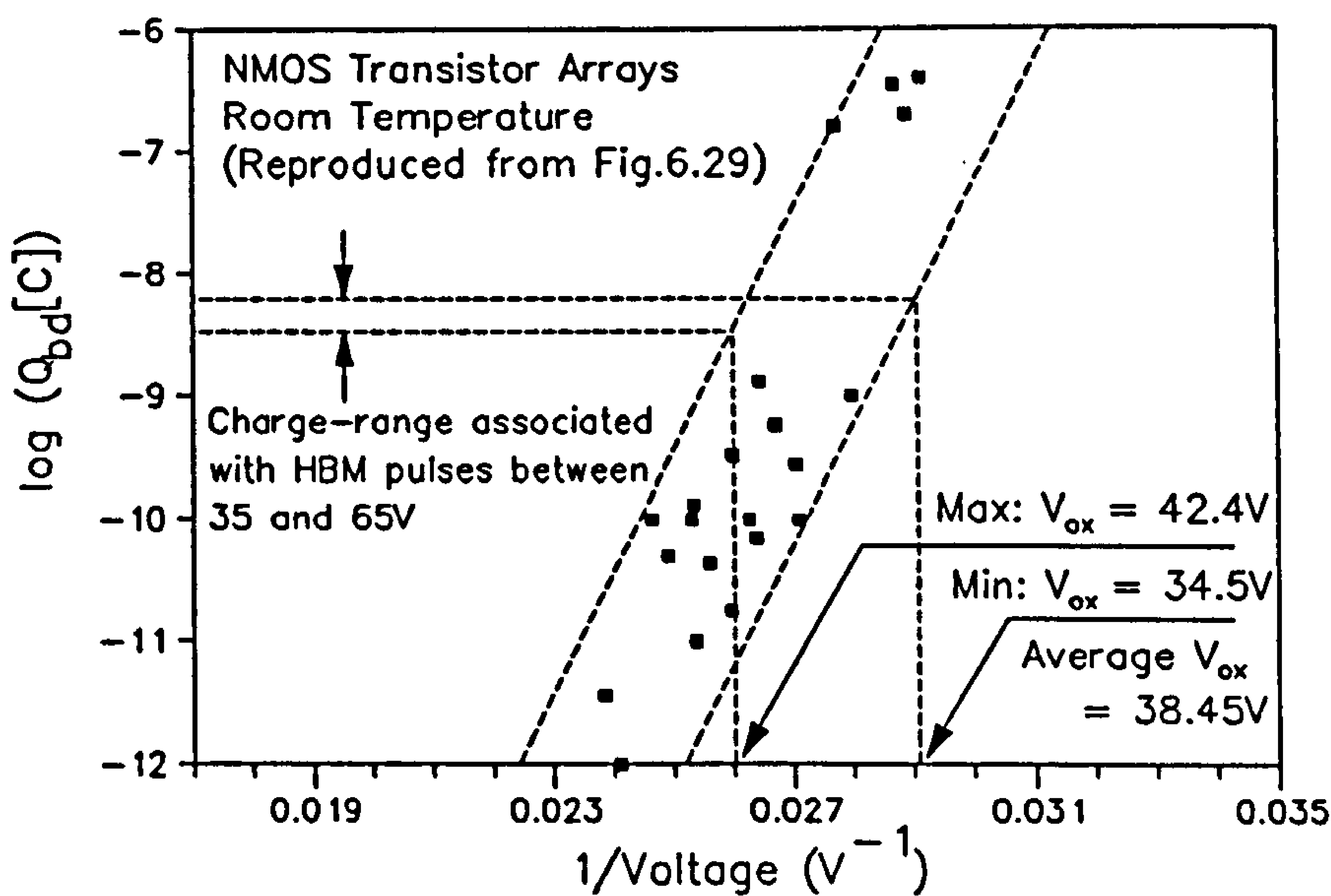


Figure 6.39: log(Q_{bd}) vs. 1/V_{ox} data reproduced from Fig.6.29, superimposed on the range of Q_{bd} values associated with ESD pulses in the range 35V to 65V.

voltage magnitude associated with V_e^* is equal to $V_e^*/L=V_e^*(C_x+C_1+C_t+C_{ox})/(C_x+C_1)$. Using the standard HBM body capacitance (C_1=93.28pF) together with the parameters in Table 4.2, the value V_{bd}=44.5V was calculated. This value agrees very closely with all the experimental values obtained from NMOS transistors throughout the thesis. It can therefore be seen that although the model is greatly simplified (it assumes that all the ESD charge is injected into the oxide and that V_{ox} remains constant throughout the pulse) it is surprisingly accurate.

Since oxide capacitances (C_{ox}) of the NMOS transistor arrays (see Fig.6.20) is generally far smaller than ($C_x+C_1+C_t$), variations in the oxide area A have no significant effect upon the capacitive loading factor L. Also, since Fig.5.31 shows only a linear variation of Q_{bd} with area, the oxide breakdown voltage is unlikely to be significantly affected by variations in A. Hence ESD breakdown should not be significantly dependent upon oxide area (see Figs.5.7 and 5.8).

As the value of C_1 varies between 3.24pF and 4.377nF, the oxide voltage required for breakdown can be shown to vary only slightly. However, since the capacitive loading factor increases with increasing C_1, tending towards $C_x/(C_x+C_t+C_{ox})$ as $C_1 \rightarrow 0$ and towards unity as $C_1 \rightarrow \infty$, it produces the 's'-shaped V_{bd} vs. C_1 curves observed in Figs.5.18, 5.19 and 6.11. In the negative-polarity CMOS data (Fig.5.20), the low-C_1 saturation of V_{bd} is probably swamped by the experimental scatter.

The V_{bd} vs. R_2 curves may be understood by considering the Q_{bd} vs. injection-current curve of Fig.6.25. Although smallest charge in this graph is two orders of magnitude greater

than the ESD charge range (Fig.6.37), the rapid slope of the graph beyond the critical current (ie. in Region II) suggests that it crosses the ESD charge $(C_x+C_1)V_0$ at a point corresponding to a current of 100μA. There is therefore not only a critical voltage V_c^*, but also a critical current I_{ox}^* required to support ESD breakdown. This current must be driven through the body resistance R_2, while the oxide voltage is pinned at the Fowler-Nordheim voltage V_{FN} ($\approx V_{ox}^*$). If R_2=1.5KΩ then the voltage required to support I_{ox}^* is 150mV, which is negligible. However, if R_2 increases to 100KΩ, the required voltage becomes 10V, which causes a significant increase in V_{bd}. Hence for low values of R_2 the breakdown voltage remains approximately constant, while for large values of R_2 it increases rapidly with R_2. This same trend was observed for both the CMOS and the NMOS structures (see Figs.5.21-23). The breakdown voltage increases with increasing R_2 until it reaches the limit imposed by the field-induced breakdown mechanism dictated by the capacitance C_2 (see Fig 6.12).

6.6.3.2 Positive Polarity Model

In Section 6.3, the depletion-layer avalanche model was used to explain the positive-polarity ESD results obtained from the NMOS devices. The original model [8] proposed that positive-polarity breakdown under a fast transient pulse can only occur if the oxide surface becomes inverted during the pulse period. Such a situation can only occur if the silicon surface depletion layer undergoes avalanche breakdown. The large hot-electron flux produced by the avalanche plasma supports rapid avalanche injection in the SiO_2, which causes immediate breakdown. However, the positive-polarity ramp-analysis of Section 6.5.2 showed that avalanche conduction could be sustained **without** the occurrence of oxide breakdown. Indeed, oxide conduction was only detected at voltages exceeding the Fowler-Nordheim tunnelling threshold. It seems therefore that avalanche conduction merely provides a supply of electrons to the Fowler-Nordheim junction, and the subsequent tunnelling supports breakdown by the same mechanisms as those which apply under negative polarity stress. The depletion layer can therefore be modelled as a reverse-biased avalanche diode (breakdown voltage V_{av}) connected in series with the oxide (see Fig.6.40[i]), such that the device needs to receive an effective voltage of magnitude $(V_c^*+V_{av})$ in order to suffer oxide breakdown.

However, Section 3.3.7 raised the following objection to the application of this avalanche model to ESD breakdown in unprotected devices: If no conduction path exists in parallel with the oxide then the pulse decay time must be dictated by the rate of charge injection, which cannot begin until the oxide surface is inverted. The work of this chapter allows this problem to be solved. To begin with, the effects of sub-avalanche inversion can be included in the model by means of a resistor R_{gen} placed in series with the oxide (see

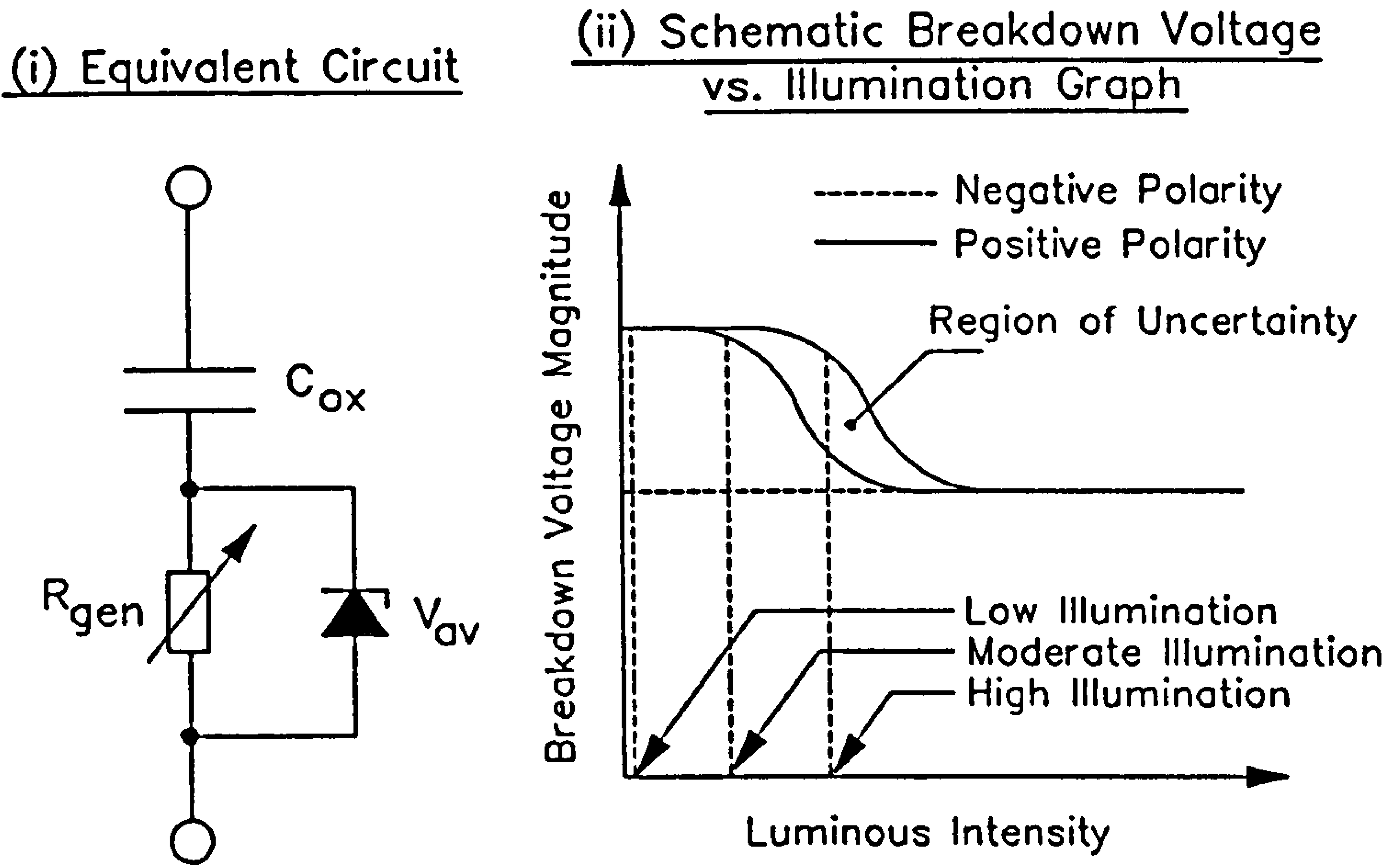


Figure 6.40: Model of positive polarity ESD breakdown: (i) Equivalent circuit, (ii) Schematic V_{bd} vs. Illumination curves.

Fig.6.40). The growth of the inversion layer under sub-avalanche conditions is clearly governed by the time-constant $C_{ox}R_{gen}$, which is equal to the minority carrier response time τ_{min}. In an unilluminated device, this is typically in the range 10-1000ms [23]. If τ_{min}=10ms and C_{ox}=4pF then R_{gen} must equal 2.5GΩ. Section 6.6.3.1 showed that breakdown requires an injection current of about 100μA which, in the absence of depletion-layer avalanche conduction, would require a voltage of 250kV across R_{gen}. It is therefore obvious that breakdown at voltages of the order of +100V must be caused by avalanche breakdown. Since avalanche breakdown is most likely to occur at the end of the pulse risetime (when the depletion-layer field is at its maximum) Amerasekera and Campbell's [8] assumption is justified.

In order to test this hypothesis, an NMOS capacitor was stressed at +50V using the apparatus shown in Fig.6.41(a). The value 50V was chosen since it exceeds V_e^* but is smaller than V_e^*+V_{av} (Fig.6.31 shows that V_{av} lies between 80 and 90V). The voltage V_k across the capacitor C_k was used to monitor the total charge injected as a function of time. Fig.6.41(b) displays the resulting V_k(t) profile, which indicates that 535μC are required to cause breakdown. Since this charge corresponds to an HBM pulse magnitude of over 5MV, this adds experimental evidence to the assertion that positive polarity ESD breakdown in p-substrate devices requires depletion-layer avalanche conduction.

Sections 6.3.2 and 6.3.3 showed how the depletion layer avalanche model can explain the temperature and size dependencies of positive-polarity ESD breakdown in NMOS devices. The model can also be used to explain the illumination-dependence observed in Fig.6.8. It

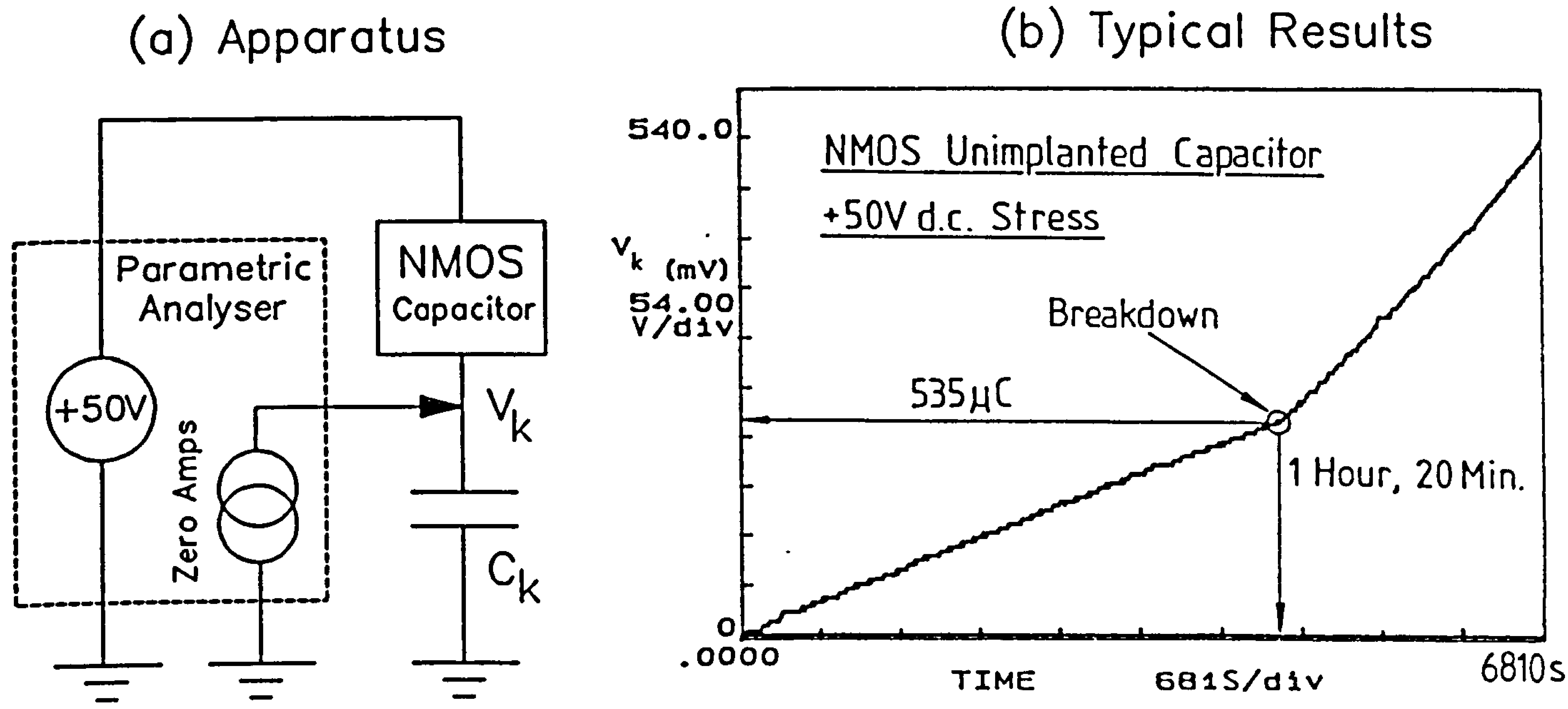


Figure 6.41: Positive polarity sub-avalanche experiment.

is first necessary to consider some of the mechanisms governing τ_{min}. At relatively low temperatures (<140°C), the dominant mechanism is electron-hole-pair (e.h.p.) generation/recombination in the depletion layer [23]. The generation may be direct (i.e. band-to-band) or indirect (i.e. via trapping levels in the Si forbidden gap). If the device is bombarded with photons then the e.h.p. generation rate must increase, causing τ_{min} and R_{gen} to fall. The generation rate must also depend upon the density of generation/recombination centres, and τ_{min} and R_{gen} therefore vary between devices according to the local silicon quality. The 10-1000ms range for τ_{min} [23] corresponds to zero-light conditions, and its value may be far smaller under illumination. If τ_{min} were to fall to 400ns under strong illumination, the corresponding value of R_{gen} would decrease to 400KΩ, which would require only 10V in order to carry the 100μA breakdown current. Since $V_{av}>10V$ (see Fig.6.1), breakdown would not require avalanche conduction and it would therefore occur at a lower voltage than it would under low illumination conditions.

The situation is shown schematically in Fig.6.40(ii): As illumination increases, R_{gen} decreases, causing V_{bd} to decrease also. However, random fluctuations in trap density cause the variation in R_{gen} to be greater in some devices than in others. Hence the decrease in the mean value of V_{bd} is accompanied by an increase in its coefficient of variation. Both these phenomena are observed in Fig.6.8.

The NMOS and HMOS gate-substrate characteristics (I_g. vs. V_{gb}) shown in Fig.6.42 add experimental evidence to the above theory. The negative polarity curves show a large tunnelling current above the Fowler-Nordheim threshold V_{FN} (about 35V), which is unaffected by variations in temperature and illumination. However, the positive polarity

Figure 6.42: I_g vs. V_{gb} characteristics for undamaged NMOS and HMOS transistors.

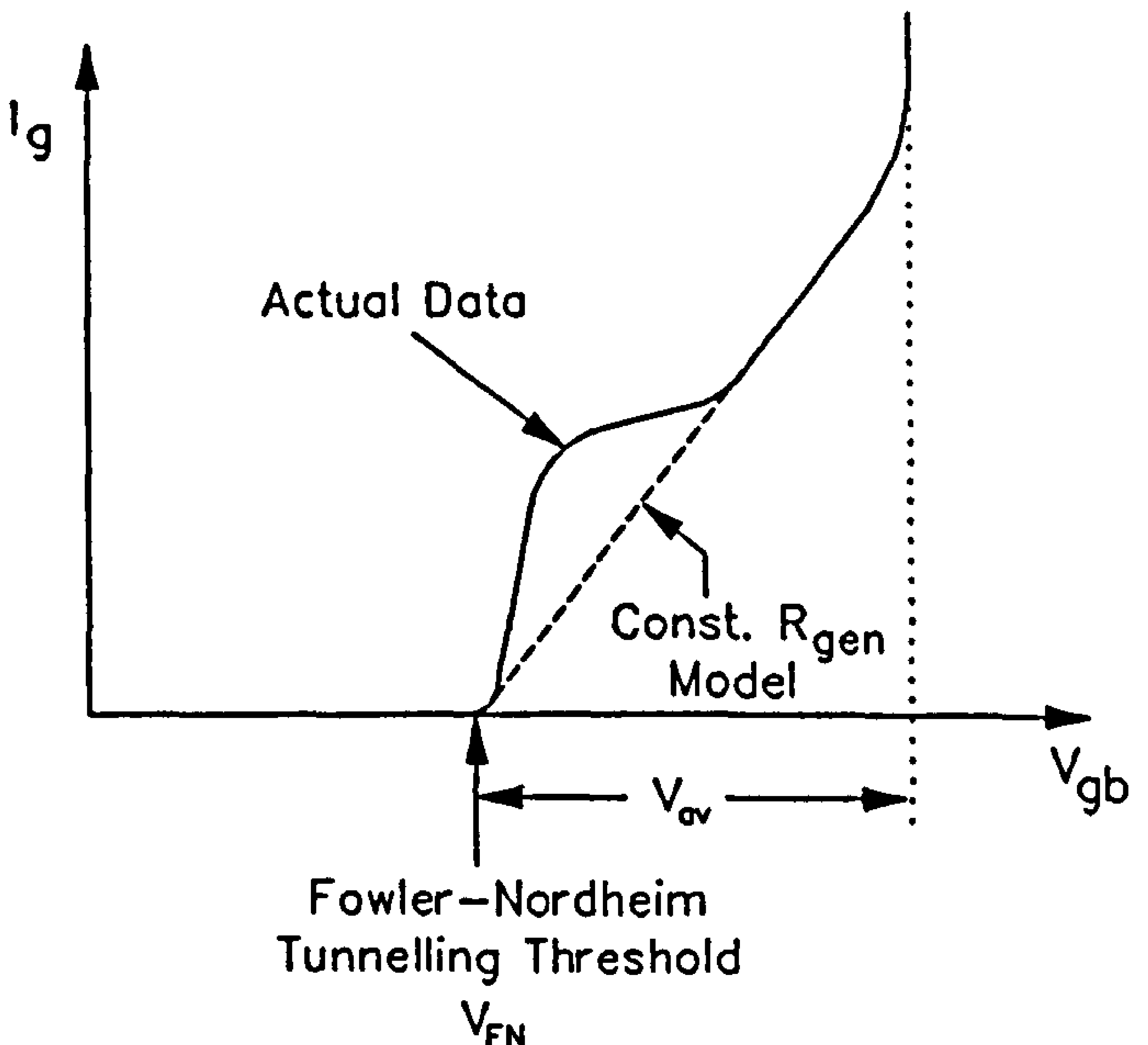


Figure 6.43: Schematic comparison of real and theoretical I_d vs. V_{gb} curves.

curves show that the tunnelling current is limited to a value dependent upon the luminous intensity. In this region, the oxide voltage 'pins' at V_{FN}, and the additional device voltage is dropped across the depletion layer. Although the curves are not linear as predicted by the 'constant R_{gcn}' model above (see Fig.6.43), it is clear that the model is qualitatively correct. 'Effective' values of R_{gcn} dV_{gb}/dI_g calculated from the NMOS characteristic show that R_{gcn} decreases from the order of GΩ, to about 2MΩ as the illumination increases.

Having explained the polarity dependencies in NMOS and HMOS transistors, it is now necessary explain the similar effects observed in the CMOS devices. Since the CMOS wafers have n-type substrates and n^+-type gates, the depletion-layer avalanche model is not applicable. In order to shed further light on this problem, the CMOS I_g vs. V_{gb} characteristics were measured at room temperature and at 200°C. The results, displayed in Fig.6.44, show that the characteristics are practically identical, irrespective of device type (p or n-channel) and stress polarity. The current increases rapidly as the voltage magnitude exceeds about 40V, indicating the Fowler-Nordheim voltage V_{FN}. This may be assumed to be the effective breakdown voltage V_e^* of the structures (see Section 6.6.3.1), which is polarity insensitive but increases slightly with increasing temperature. However, since for both wafer chucks the value of C_t is greatest under positive polarity conditions (see Table 3.2), the loading factor L and hence the breakdown voltage $V_{bd}=V_e^*/L$ is greater under positive polarity stress than under negative polarity stress. Since V_e^* increases with temperature, both positive and negative-polarity breakdown should, in principle, be temperature sensitive. However, the results of Section 5.2.3 show that while the positive breakdown threshold increases with temperature, the negative threshold remains constant. This may be explained by considering the voltage resolution of the ESD step-test (see Section 6.2.1). In the positive polarity case, the temperature variation is greater than the 4V voltage resolution and is therefore detected by the experiment. In the negative polarity case the temperature variation in V_e^* is attenuated by the capacitive loading factor to such an extent that its variation becomes swamped by the 4V resolution.

Finally, the silicon-on-sapphire (SOS) data must be considered (see Fig.5.17). Fig.6.45 shows the I_g vs. V_{gs} curves of the SOS transistor arrays, showing that the magnitude of V_e^* is approximately 26V, irrespective of stress polarity. Thus the apparent polarity dependence observed in Fig.5.17 can be explained in the same manner as that of the CMOS results above. The curves of Fig.6.45 were unaffected by variations in luminous intensity. Hence the apparent illumination dependence of negative polarity breakdown in Fig.5.17 must be an illusion caused by statistical fluctuations in breakdown voltage between devices. (Fig.5.17(b) showed that this is consistent with the degree of statistical scatter.)

Figure 6.44: I_g vs. V_{gb} characteristics for undamaged CMOS transistors.

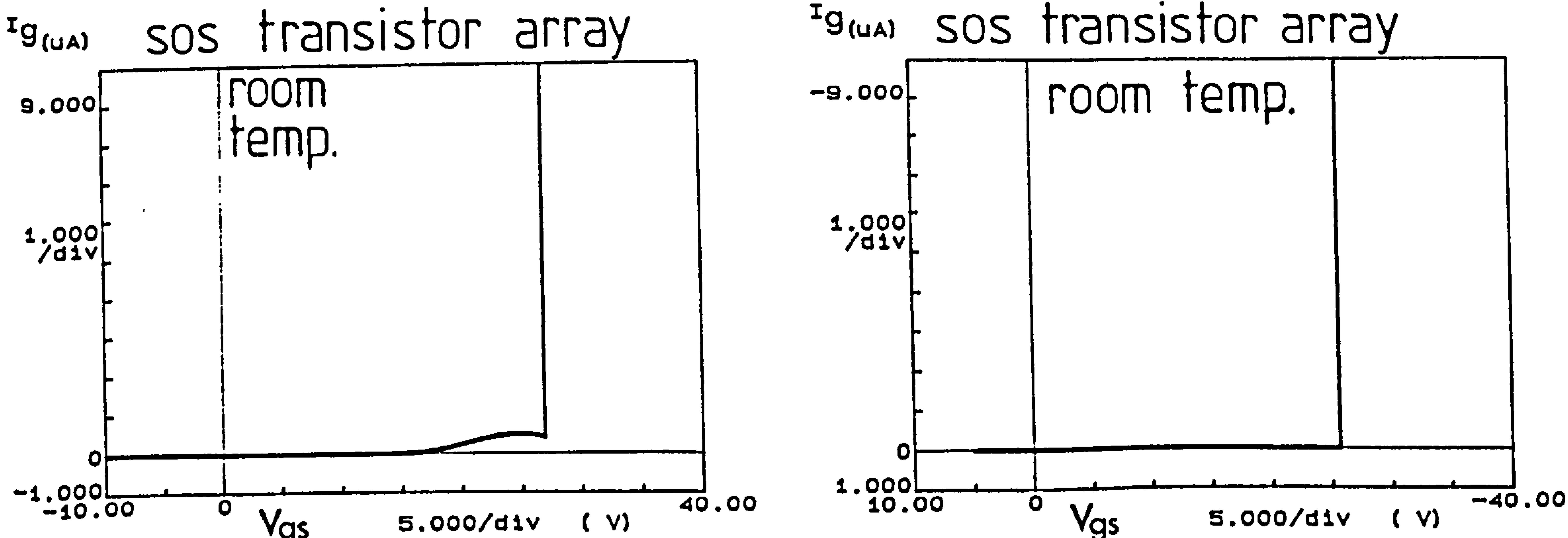


Figure 6.45: I_g vs. V_{gs} characteristics for SOS transistor arrays.

6.7 Summary

1. The ESD data of Chapter 5 was qualified by means of a critical analysis of the experimental techniques used in its acquisition. It was shown that the low-illumination data for NMOS and HMOS under positive polarity was spurious, together with the low-illumination CMOS data under negative stress. The results of this analysis led to the formulation of an ideal test strategy for studying ESD breakdown.
2. Some of the data which the above-mentioned analysis revealed as spurious was re-measured using the improved test techniques. The mechanisms responsible for the observed illumination, temperature and size dependencies of ESD breakdown were then discussed.
3. The sensitivity of ESD breakdown to discharge capacitance C_1 and discharge resistance R_2 was examined. It was shown that data obtained using extreme values of the HBM parameters C_1 and R_2 could be regarded as simulations of other forms of stress such as EOS, MM, CDM and FIM.
4. The constant voltage t_{bd} vs. F curve can be seen to consist of three distinct time domains, which were denoted I, II and III. Additional experiments showed that Region III (i.e. ultra short time-scale) breakdown is governed by a delay time t_d between the application of stress and the onset of Fowler-Nordheim tunnelling. It was concluded that t_d was probably due to charge re-distribution in the oxide.
5. Breakdown in Regions I, II and III were characterised in terms of different dynamic charge-conditions in the oxide. While Region III is caused by the oxide becoming more positive and Region I with the oxide becoming more negative, Region II behaviour is associated with the intermediate condition of stable oxide charge.
6. The 'critical current' in the Q_{bd} vs. injection-current curve (first observed by Wolters et al.[37]) was shown to correspond to the transition point between Regions I and II.
7. The negative-polarity voltage ramp data was mathematically processed in order to extract the oxide voltage and charge injection profiles. The charge-to-breakdown Q_{bd} was shown to decrease very rapidly with increasing V_{ox}. The results were found to have no significant temperature dependencies.
8. The positive-polarity voltage ramp data was mathematically processed in order to extract the V_{ox} and V_{dep} profiles. Although V_{dep} was seen to pin at the avalanche voltage V_{av}, the latter was observed to increase gradually during the pulse. This was attributed to an increase in the impact-ionisation coefficient caused by Joule heating in the silicon.
9. ESD transient data were analyzed in the same manner as the voltage ramp waveforms. The oxides were shown to undergo instantaneous breakdown at the

beginning of Fowler-Nordheim tunnelling. The bulk resistivity was shown to be an important parameter in this analysis.

10. A model of charge evolution in the bulk oxide was developed. This model showed how the simultaneous action of neutral and coulombic-repulsive traps could produce the Regions I, II and III observed in the constant-voltage data.
11. The various models of oxide breakdown (first introduced in Chapter 3) were reviewed in the light of the experimental data. It was decided that Fischetti's surface plasmon model provided the best explanation of the observed results.
12. A qualitative model of negative polarity ESD breakdown was developed and compared with the experimental data. The model (which was based upon the Q_{bd} vs. V_{ox} data obtained earlier in the chapter) showed how capacitive loading could explain the capacitance-dependent breakdown data. The observation that a critical current is required for charge-limited breakdown allowed the resistance-dependent data to be explained.
13. The depletion-layer avalanche model [8] was modified in the light of the experimental data, and used to explain the positive-polarity ESD data. An analysis of sub-avalanche conditions in the silicon allowed the requirement for depletion-layer avalanche to be justified. The polarity-dependent breakdown in CMOS and SOS devices was explained in terms of the difference in capacitive-loading conditions which Chapter 4 showed to prevail under positive and negative conditions.

6.8 References

1. Heimann,P.A., *"An Operational Definition for Breakdown of Thin Thermal Oxides on Silicon"*, IEEE Trans. Electron Dev., ED30(10), pp.1366-8, 1983.
2. Goetzberger,A, Shockley,W, *"Metal Precipitates in Silicon p-n Junctions"*, . Appl. Phys., 31(10), pp.1821-4, 1960.
3. Plummer,A.R., *"The Effect of Heat Treatment on the Breakdown Characteristics of Silicon pn Junctions"*, J.Electronics & Control, 5(5), pp.405-16, 1958.
4. Streetman,B.G., *"Solid State Electronic Devices"* (Second Edition), Prentice/Hall International, London, p.443, 1980.
5. Soden,J, Hawkins,C.F., *"Test Considerations for Gate Oxide Shorts in CMOS ICs"*, IEEE Design and Test of Computers, pp.56-64, August 1986.
6. Syrzycki,M, *"Modelling of Gate Oxide Shorts in MOS Transistors"*, IEEE Trans. Computer-Aided Design, 8(3), pp.193-202,1989.
7. Sze,S.M., *"Physics of Semiconductor Devices"* (Second Edition), John Wiley and Sons, New York, p.107, 1981.
8. Amerasekera,E.A., Campbell,D.S., *"Charge-Limited Breakdown in MOS Capacitors"*, Proc. **17th.** Annual Euro. Sol. State Device Research Conf.

(ESSDERC), pp.733-41, 1987.

9. Goetzberger,A, Nicollian,E.H., *"MOS Avalanche and Tunnelling Effects in Silicon Surfaces"*, J. Appl. Phys., 38(12), pp.4582-8, 1967.
10. Rusu,A, Bulucea,C, *"Deep-Depletion Breakdown Voltage of Silicon-Dioxide/Silicon MOS Capacitors"*, IEEE Trans. Electron Dev., ED26(3), pp.201-5, 1979.
11. Maeda,N, Wada,T, *"The Dependence of Electrostatic Destruction Voltage on Device Structures of P-N Junctions and Insulated Films"*, Proc. 11th. EOS/ESD Symp. Proc., pp.157-61, 1989.
12. Fong,Y, Hu,C, *"The Effects of High Electric Field Transients on Thin Gate Oxide MOSFETs"*, Proc. 9th. EOS/ESD Symp., pp.252-7, 1987.
13. Chen,I-C, Hu,C, *"Accelerated Testing of Time-Dependent Breakdown of SiO_2"*, IEEE Electron Dev. Lett., EDL8(4), pp.140-2, 1987.
14. Chen,I-C, Holland,S.E., Hu,C, *"Electrical Breakdown in Thin Gate and Tunneling Oxides"*, IEEE Trans. Electron Dev., ED32(2), pp.413-22, 1985.
15. de Moura,M.A, de Albuquerque,D.F., *"Remarks on the Traversal Time in a Tunnelling Process"*, Sol. State Comm., 74(5), pp.353-4, 1990.
16. Weinberg,Z.A., *"On Tunnelling in Metal-Oxide-Silicon Structures"*, J.Appl.Phys., 53(7), pp.5052-6, 1982.
17. Nicollian,E.H., Brews,J.R., *"MOS (Metal Oxide Semiconductor) Physics and Technology"*, John Wiley & Sons, New York, p.101, 1982.
18. Wiley,J.D., Miller,G.L., *"Series Resistance Effects in Semiconductor CV Profiling"*, IEEE Trans. Electron Dev., ED22(5), pp.265-72, 1975.
19. Liang,M-S, Hu,C, *"Electron Trapping in Very Thin Thermal Silicon Dioxides"*, Proc. IEDM '81, pp.396-9, 1981.
20. Fischetti,M.V., *"High Field Electron Transport in SiO_2 and Generation of Positive Charge at the Si-SiO_2 Interface"*, Insulating Films on Semiconductors, pp.181-9, 1986.
21. Moazzami,R, Lee,J.C., Hu,C, *"Temperature Acceleration of Time-Dependent Dielectric Breakdown"*, IEEE Trans. Electron Dev., ED36(11), pp.2462-5, 1989.
22. Wolters,D.R., van der Schoot,J.J., *"Dielectric Breakdown in MOS Devices, Part I: Defect-Related and Intrinsic Breakdown"*, Phillips Journal of Research, 40, pp.115-36, 1985.
23. Nicollian,E.H., Brews,J.R., *"MOS (Metal Oxide Semiconductor) Physics and Technology"*, John Wiley & Sons, New York, p.105-56, 1982.
24. Amerasekera,E.A., *"Failure Mechanisms in Semiconductor Devices: Progress Report No.6"*, Contract No.(LUT)ELJB7/(MOD)A5a/1412, Loughborough University of Technology, October 1987.
25. Raider,S.I., *"Time-Dependent Breakdown of Silicon Dioxide Films"*, Appl. Phys. Lett., 23(1), pp.34-6, 1973.
26. Osburn,C.M., Weitzman,E.J., *"Electrical Conduction and Dielectric Breakdown in Silicon Dioxide Films on Silicon"*, J. Electrochem. Soc., 119(5), pp.603-9, 1972.
27. Avni,E, Shappir,J, *"Trap Generation and Occupation in Stressed Gate Oxides under Spatially Variable Oxide Electric Field"*, Appl. Phys. Lett., 51(22), pp.1857-9, 1987.
28. Changhua,T, Mingzhen,X, Yangyuan,W, *"Effect of Newly Generated Interface Traps on Fowler-Nordheim Voltage Shift"*, Chinese Physics, 10(3), pp.799-806, 1990.
29. Davis,J.R., *"Instabilities in MOS Devices"*, Gordon and Breach, New York, 1981.
30. Wolters,D.R., van der Schoot,J.J., *"Dielectric Breakdown in MOS Devices, Part III:*

The Damage Leading to Breakdown", Phillips Journal of Research, 40, pp.164-92, 1985.

31. Chen,I.C., Holland,S, Hu,C, *"Electron-Trap Generation by Recombination of Electrons and Holes in SiO_2"*, J. Appl. Phys., 61(9), pp.4544-8, 1987.
32. DiMaria,D.J., Stasiak,W.J., *"Trap Creation in Silicon Dioxide Produced by Hot Electrons"*, J. Appl. Phys., 65(6), pp.2342-6, 1989.
33. Buchanan,D.A., DiMaria,D.J., *"Interface and Bulk Trap Generation in Metal-Oxide-Semiconductor Capacitors"*, J. Appl. Phys., 67(12), pp.7439-52, 1990.
34. DiMaria,D.J., *"Correlation of Trap Creation with Electron Heating in Silicon Dioxide"*, Appl. Phys. Lett., 51(9), pp.655-7, 1987.
35. Avni,E, Shappir,J, *"A Model for Silicon Dioxide Breakdown under High Field and Current Stress"*, J.App.Phys., 64(2), pp.743-8, 1988.
36. Avni,E, Shappir,J, *"Effects of Temperature Annealing on Charge-Injection-Induced Trapping in Gate Oxides of Metal-Oxide-Silicon Transistors"*, J. Appl. Phys., 63(5), pp.1563-8, 1988.
37. Wolters,D.R., van der Schoot,J.J., *"Dielectric Breakdown in MOS Devices, Part II: Conditions for the Intrinsic Breakdown"*, Phillips Journal of Research, 40, pp.137-63, 1985.
38. Zhang,J.F., Watkinson,P, Taylor,S, Eccleston,W, Young,N.D., *"Electron Conduction and Charge Trapping Behaviour in SiO_2 Prepared by Plasma Anodisation"*, Proc. 20th. Annual Euro. Sol. State Device Research Conf. (ESSDERC), pp.265-8, 1987.
39. Verwey,J.F., Amerasekera,E.A., Bisschop,J, *"The Physics of SiO_2 Layers"*, Rep. Prog. Phys., 53, pp.1297-331, 1990.
40. Harari,E, *"Dielectric Breakdown in Electrically Stressed Thin Films of Thermal SiO_2"*, J.Appl.Phys., 49(4), pp.2478-89, 1978.
41. Meyer,W.K., Crook,D.L., *"Model for Oxide Wearout due to Charge Trapping"*, Proc. 21st. IEEE Reliability Physics Symp., pp.242-7, 1983.
42. Fischetti,M.V., *"Model for the Generation of Positive Charge at the Si-SiO_2 Interface based on Hot-Hole Injection from the Anode"*, Phys. Rev. B, 31(4), pp.2099-2113, 1985.
43. Ridley,B.K., *"Mechanism of Electrical Breakdown in SiO_2 Films"*, J.Appl.Phys., 46(3), pp.998-1007, 1975.
44. Ricco,B, et al., *"Novel Mechanism for Tunnelling and Breakdown of Thin SiO_2 Films"*, Phys.Rev.Lett., 51(9), pp.1795-8, 1983.
45. Suñé,J, Farrés,E, Placencia,I, Barniol,N, Martín,F, Aymerich,X, *"Nondestructive Multiple Breakdown Events in Very Thin SiO_2 Films"*, Appl. Phys. Lett., 55(2), pp.128-30, 1989.
46. Chen,I-C, Holland,S, Young,K.K., Chang,C, Hu,C, *"Substrate Hole Current and Oxide Breakdown"*, Appl. Phys. Lett., 49(11), pp.669-71, 1986.
47. Amerasekera,E.A., Campbell,D.S., *"ESD Pulse and Continuous Voltage Breakdown in MOS Capacitor Structures"*, Proc. 8th. EOS/ESD Symp., pp.208-13, 1986.
48. Tunnicliffe,M.J., Dwyer,V.M., Campbell,D.S., *"Parametric Drift in Electrostatically Damaged MOS Transistors"*, Proc. 14th. EOS/ESD Symp., 1992.
49. Chen,I-C, Holland,S, Hu,C, *"Hole Trapping and Breakdown in Thin SiO_2"*, IEEE Electron Dev. Lett., EDL7(3), pp.164-7, 1986.
50. Mott,N, *"Properties of Non-Crystalline Materials"*, Oxford University Press, Oxford,

1987.

51. Tunnicliffe,M.J., Dwyer,V.M., Campbell,D.S., *"Slow Transient Voltage and ESD Breakdown in Unprotected MOS Gate Oxides"*, Proc. 1st. Components Engineering, Reliability and Test Conference (CERT), pp.78-87, 1990.2

Chapter 7

Mathematical Model of Oxide Wearout

7.1 Introduction

The experimental investigations of Chapters 5 and 6 have shed some light upon the qualitative nature of oxide breakdown over a wide spectrum of experimental conditions. The principal conclusions can be summarised as follows:

1. When an oxide is abruptly exposed to a large electric field (~10MV/cm), a time delay t_d elapses prior to the beginning of Fowler-Nordheim tunnelling injection. The delay time decreases with increasing electric field, but is unaffected by variations in temperature and illumination. The delay is probably caused by transient electron trapping/de-trapping in neutral trap states (NTS).
2. When exposed to a prolonged electric field stress, the tunnelling current decays with time. This is probably caused by electron trapping in the much slower coulombic-repulsive trap (CRT) states.
3. The two charge-trapping mechanisms divide the t_{bd} vs. F curve into three distinct domains, which have been denoted Regions I, II and III. Region III breakdown is dominated by the tunnelling time delay t_d, while Region I failure is dictated by CRT electron trapping. Breakdown in Region II is associated with an oxide space-charge in dynamic equilibrium.
4. The boundary between Regions I and II appears to coincide with the critical current J_{cr} between 'ductile' and 'brittle' failure [1].
5. Since oxide damage accumulation cannot begin until $t > t_d$, the standard 'causal' equation is inapplicable under short time-scale stress. A modified causal equation, which only integrates damage for $t > t_d$, was therefore proposed.
6. The ESD breakdown criterion can be described in terms of a critical oxide voltage V_{FN} and a critical tunnelling current I_{ox}^{*}. Only if these criteria are met can the limited ESD pulse charge support breakdown.
7. Positive polarity breakdown in an unilluminated p-substrate device can only result from avalanche conduction in the Si depletion layer.

The present chapter extends this understanding on a quantitative level and develops a mathematical model of the oxide wearout process. For the sake of simplicity, the analysis is limited as much as possible to the following data sets obtained from NMOS devices:

(i) NMOS ESD data for variable C_1 (Fig.6.11)
(ii) NMOS ESD data for variable R_2 (Fig.6.12)
(iii) NMOS constant-voltage data (Fig.6.13(i))
(iv) NMOS constant-current data (Fig.6.25).

The chapter begins by assuming a simple working model of dielectric wearout, based upon the results of Chapter 6. This theory is used to develop analytical models of breakdown under constant current, constant voltage and ESD conditions. The model predictions are then compared with the various data sets.

7.2 Analytical Model of Oxide Wearout

The earlier parts of this chapter are based upon the 'causal' model of oxide wearout, modified in order to account for the tunnelling time-delay t_d (see Section 6.4.2). Chapter 6 showed that if damage begins to accumulate at $t=t_d$ then the causal equation must be re-written

$$\int_{t_d}^{t_{bd}} \frac{dt}{\tau(F)} = 1 \qquad 7(1)$$

The time-to-breakdown function $\tau(F)$ shall be given the trial form [2]:

$$\tau(F) = \tau_0 e^{\frac{\gamma}{F}} \qquad 7(2)$$

where τ_0 and γ are constants. Finally, the oxide injection current density J (for $t > t_d$) will be assumed to follow the Fowler-Nordheim equation

$$J = kF^2 e^{-\frac{B}{F}} \approx J_0 e^{-\frac{B}{F}} \qquad 7(3)$$

where k and B are constants, J_0 is approximately constant over the limited field-range applicable to breakdown and F is the average oxide field given by V_{ox}/T_{ox} (Appendix A gives the derivation of this latter formula). It should be noted that this theory only applies to the

short time-scale wearout behaviour (Regions II and III), since it contains no model of long time-scale current decay. The modelling of Region I breakdown is considered later in the chapter.

7.3 Analytical Model of Constant Voltage/Constant Current Breakdown

Constant voltage breakdown can be modelled by a simple corollary of Eqns.7(1-2): If F (and hence $\tau(F)$) remain constant with time then Eqn.7(1) can be re-written

$$t_{bd}(F) = \tau_0 e^{\frac{\gamma}{F}} + t_d(F) \tag{7(4)}$$

Assuming that the field-dependence of $t_d(F)$ is known, t_{bd} can be modelled as a function of F. If $t_{bd} >> t_d$ (as in Region II) then Eqn.7(4) becomes $t_{bd}=\tau_0.\exp(\gamma/F)$.

Current-induced breakdown can be modelled by transforming Eqn.7(1) from the time-domain to the charge-domain. This is achieved by changing the variable of integration from t to Q_{ox} (=J × t), i.e.

$$\int_{t_d}^{t_{bd}} \frac{dt}{\tau(F)} = \int_{t=t_d}^{t=t_{bd}} \frac{(J\,dt)}{J\tau(F)} = \int_{0}^{Q_{bd}} \frac{dQ_{ox}}{\Theta(J)} = 1 \tag{7(5)}$$

where the *charge-to-breakdown function* $\Theta(J) = J \times \tau(F(J))$ can be obtained by combining Eqns.7(2-3), i.e.

$$\Theta(J) = \tau_0 J_0^{\frac{\gamma}{B}} J^{1-\frac{\gamma}{B}} \tag{7(6)}$$

Hence under a constant current J, Eqn.7(5) can be re-written

$$Q_{bd}(J) = \Theta(J) = \tau_0 J_0^{\frac{\gamma}{B}} J^{1-\frac{\gamma}{B}} \tag{7(7)}$$

Having obtained analytical expressions for t_{bd} and Q_{bd} as functions of F and J, the next stage is to fit these equations to the measured data. For this purpose, the peak (i.e. Region II) negative polarity injection current was measured in NMOS transistors as a function of field. This was performed using the HP4145B parametric analyzer (Section 4.2.2.1) in time-domain mode (Fig.7.1(a)) and the values of k, J_0 and B were determined by fitting Eqn.7(3)

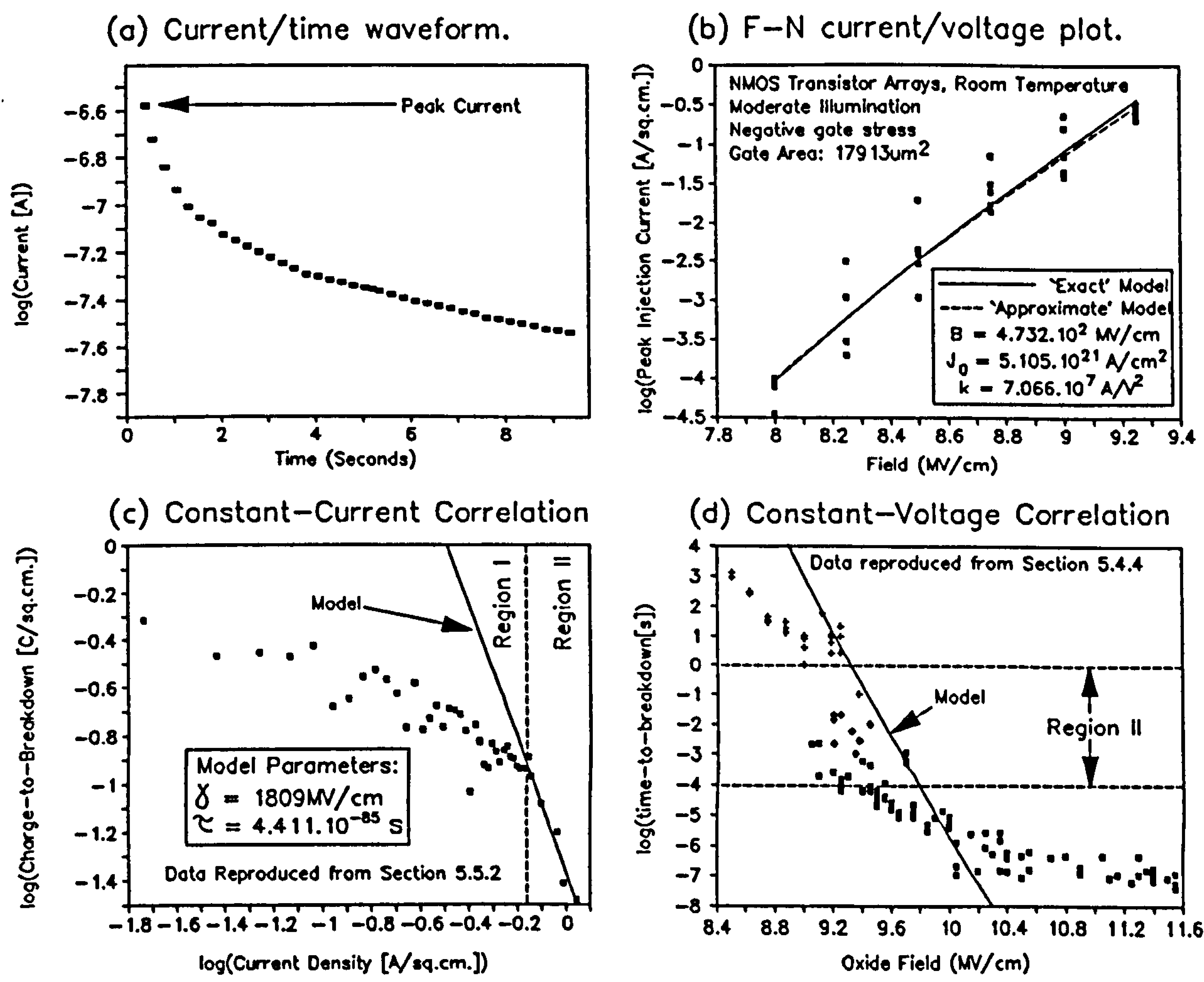


Figure.7.1: Correlation between analytical breakdown model and Region II experimental data.

to the data (see Fig.7.1(b))[1]. Secondly, the slope and intercept of the $\log(Q_{bd})$ vs. $\log(J)$ curve in Region II yield $(1-\gamma/B)$ and $\tau_0 J_0^{\gamma/B}$, allowing the computation of γ and τ_0 (Fig.7.1(c))[2]. The overall consistency of the model was then verified by comparing the experimental t_{bd} vs. F data in Region II with the model equation $t_{bd}=\tau_0 e^{\gamma/F}$. The results are plotted in Fig.7.1(d), which show that the model curve barely skims the upper limit of the data distribution. This error may possibly be associated with the fact that much of the data used to calculate J_0 and B (Fig.7.1(b)) was measured using fields in Region I. However, the predicted $\log(t_{bd})$ vs. F curve has the correct approximate slope, and is shifted by little more than 300kV/cm from the centre of the distribution (a field which corresponds to only 1.2V

[1]. B and k were obtained from the slope and intercept of the $\log(J/F^2)$ vs. 1/F plot. J_0 was then adjusted to give optimum correlation of $J=J_0e^{-B/F}$ within the data field.

[2]. The Q_{bd} vs. J data was reproduced from Fig.5.32. The data points in Fig.7.1(c) represent the mean values of Q_{bd} obtained for each injection-current level.

across a 40nm oxide).

Fig 7.1 contains all the parameters obtained from the above analysis. These parameters should not be regarded as a hard and fast characterisation of the data, particularly in view of the sparseness of the Region II constant current data and the wide experimental scatter. They should merely be regarded as *plausible* parameters, which provide a tolerable fit with the experimental data (allowing for the existence of large statistical variations in device properties).

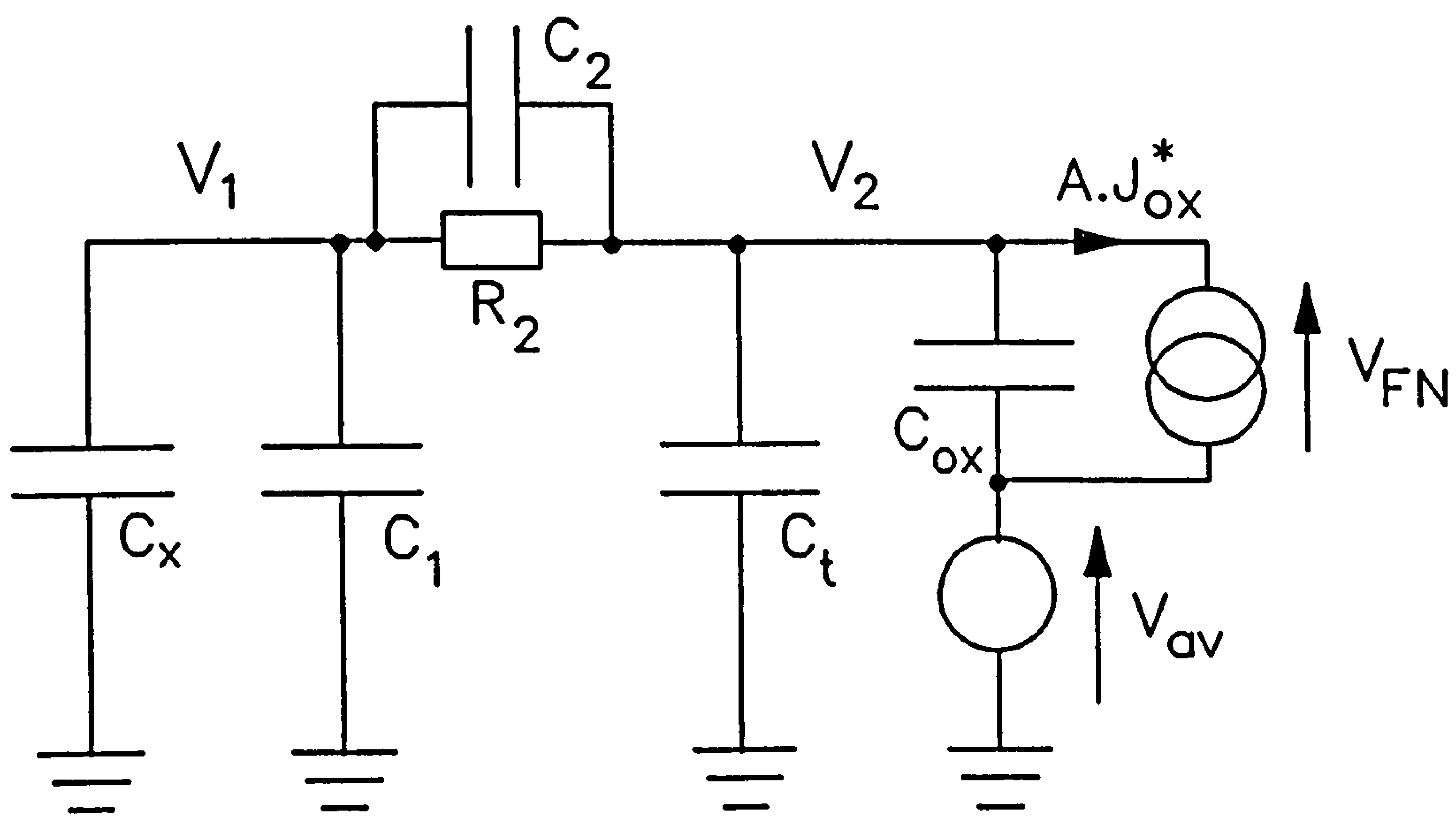


Figure 7.2: Equivalent circuit model for ESD System.

7.4 Analytical Model of ESD Breakdown

Chapter 6 concluded that during an ESD event (in which the charge is limited to the order of 180nC), breakdown can only be supported if the following criteria are met:

a. The oxide voltage V_{ox} must be raised to the threshold V_{FN} for Fowler-Nordheim tunnelling.

b. A critical threshold current $I_{ox}^* = A.J_{ox}^*$ must be forced into the oxide under the Fowler-Nordheim tunnelling mechanism.

The following section quantitatively extends this understanding and develops an analytical model of ESD breakdown. This model is based upon the equations presented earlier in this chapter. Fig.7.2 shows a general equivalent circuit model for the ESD system.

The voltage source V_{av} represents the silicon surface depletion layer in its avalanche condition. If no surface depletion layer exists (as under negative polarity stress), V_{av} is simply set to zero.

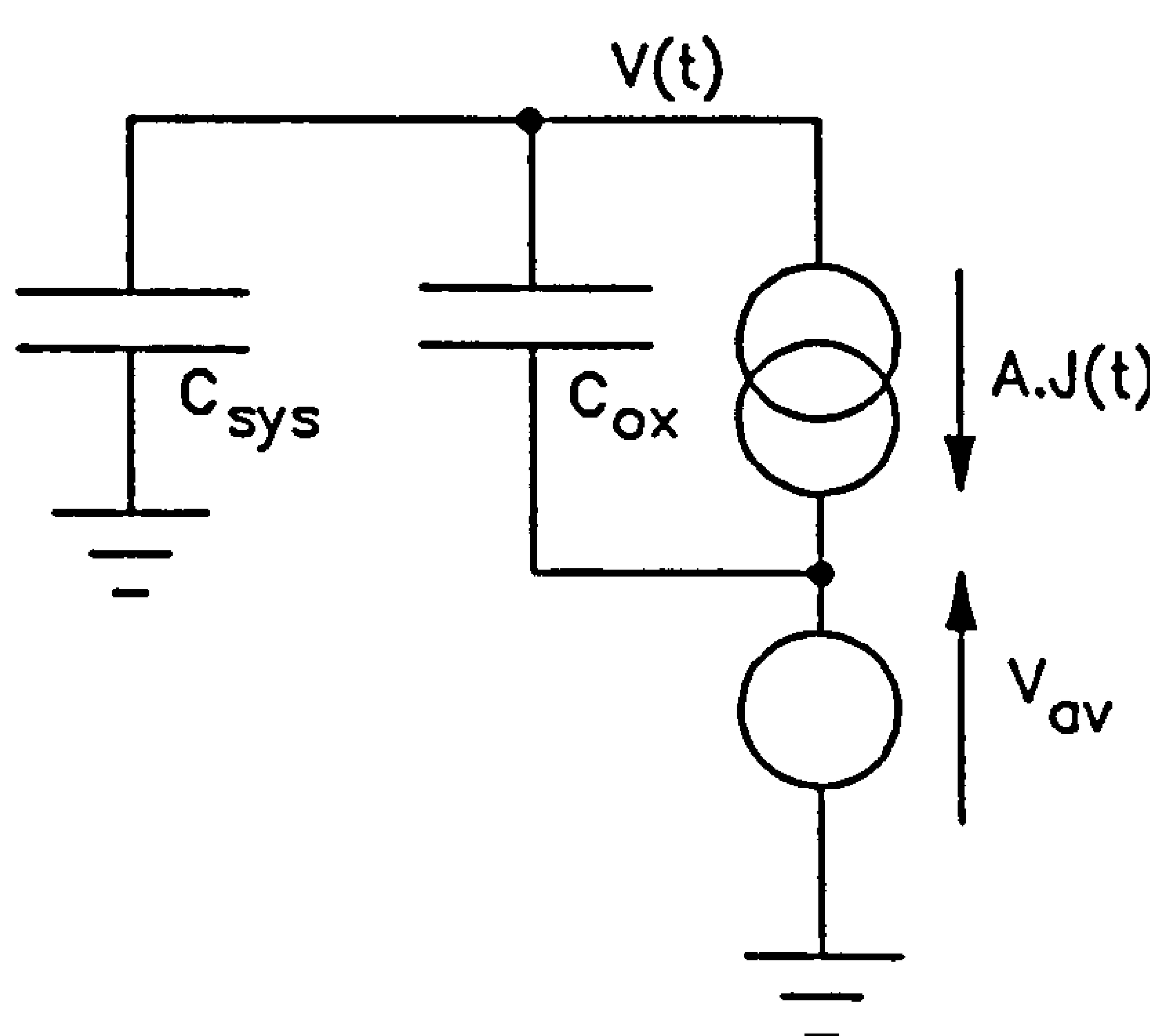


Figure 7.3: Equivalent circuit model for low-R_2 ESD breakdown.

7.4.1 Low-Resistance Breakdown

7.4.1.1 Breakdown Voltage Threshold

The ESD analysis begins by considering the low-resistance situation ($R_2 \sim k\Omega$), for which the breakdown voltage is largely independent of R_2 (see Fig.6.12). Since the pulse rise-time constant ($\tau_{rise} \sim R_2C_t$) is substantially smaller than the tunnelling time-delay t_d, the voltage across R_2 decays to zero long before the start of tunnelling. Additionally, the voltage dropped across R_2 by I_{ox}^*($\sim$0.1mA) is only about 150mV, and can therefore be neglected. Hence V_1 and V_2 may therefore be considered equal throughout the pulse decay, and the pulse-system capacitances C_x, C_1 and C_t may be replaced by a single equivalent capacitance $C_{sys}=(C_x+C_1+C_t)$.

The resulting equivalent circuit model is shown in Fig.7.3. According to this model, the pulse voltage decay profile V(t) is governed by the differential equation:

$$C_{sys}\frac{dV}{dt} + C_{ox}\frac{d(V-V_{av})}{dt} = -A\,k\left(\frac{V-V_{av}}{T_{ox}}\right)^2 \exp\left(\frac{-B\,T_{ox}}{V-V_{av}}\right) \qquad 7(8)$$

and since V_{av} is constant, the L.H.S. of Eqn.7(8) can be re-written $(C_{sys}+C_{ox})dV/dt$. Solving this equation together with the boundary condition $V(t_d)=V_e$ (where V_e is the *effective* pulse

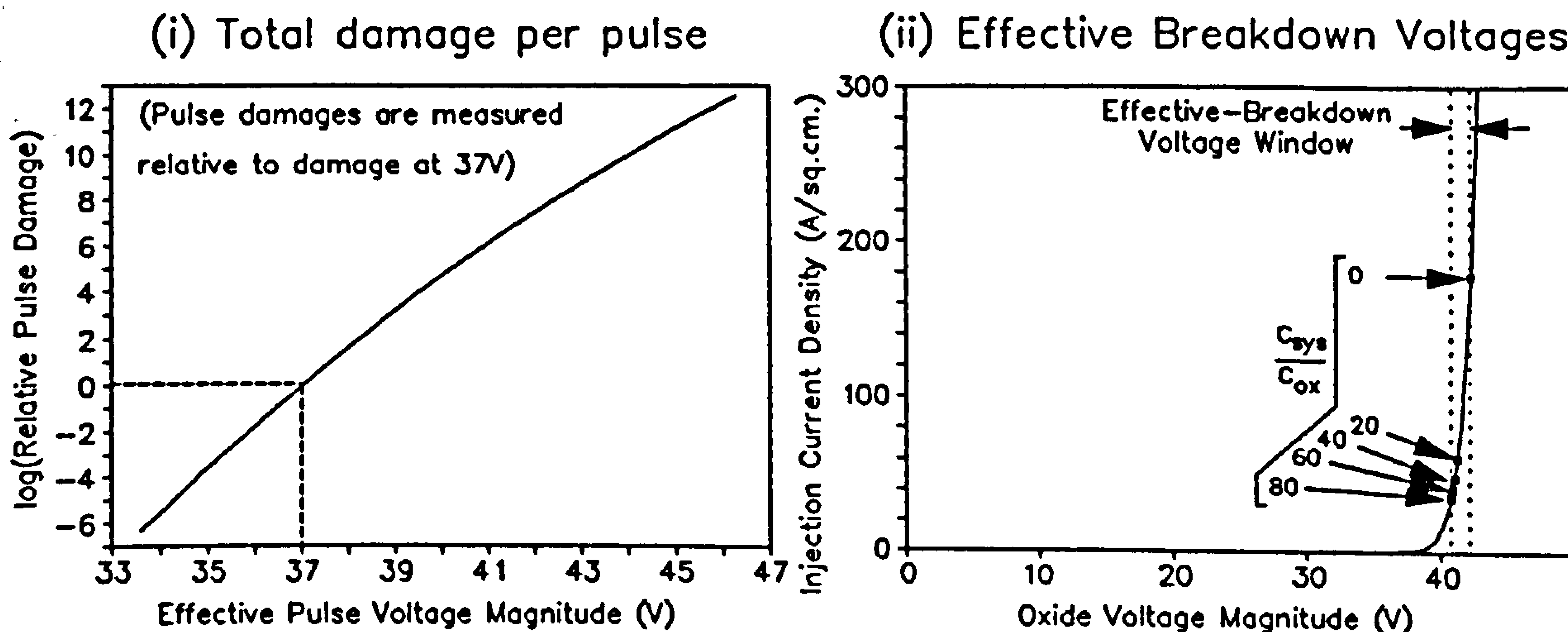


Figure 7.4: ESD oxide wearout behaviour predicted by low-R_2 model.

voltage magnitude) yields the following expression for V(t)

$$V(t) = V_{av} + \frac{B\,T_{ox}}{\log_e\left[e^{\frac{BT_{ox}}{V_e - V_{av}}} + \frac{AkB(t-t_d)}{(C_{sys}+C_{ox})\,T_{ox}}\right]} \qquad 7(9)$$

This allows the computation of the oxide field profile $F(t)=[V(t)-V_{av}]/T_{ox}$, which may be combined with Eqns.7(1-2) in order to determine the breakdown criterion. For this purpose, it is convenient to define the *damage function* s(t):

$$s(t) = \int_{t_d}^{t} \frac{dt}{\tau[F(t)]} = \frac{1}{\tau_0}\int_{t_d}^{t} e^{-\frac{\gamma}{F(t)}}\,dt \qquad 7(10)$$

such that $s(t_{bd})=1$. Inserting the ESD field profile into Eqn.7(10) yields

$$s(t) = \frac{(C_{sys}+C_{ox})T_{ox}}{Ak\tau_0(B-\gamma)}\left[\left(e^{\frac{BT_{ox}}{V_e-V_{av}}} + \frac{AkB(t-t_d)}{(C_{sys}+C_{ox})T_{ox}}\right)^{1-\frac{\gamma}{B}} - e^{\frac{(B-\gamma)T_{ox}}{(V_e-V_{av})}}\right] \qquad 7(11)$$

and hence the *total* damage $S_t(V_e)$ inflicted by a complete ESD pulse of effective magnitude V_e is given by

$$S_t(V_e) = s(t\to\infty) = \frac{(C_{sys}+C_{ox})}{Ak\tau_0(\gamma-B)}\,e^{-\frac{(\gamma-B)T_{ox}}{V_e-V_{av}}} \qquad 7(12)$$

Fig.7.4(i) shows the S_t vs. V_e curve predicted from the data in Fig.7.1 using Eqn.7(12)).

Since a 1V increment raises S_t by two orders of magnitude, pulses below the breakdown voltage inflict negligible damage upon the oxide (see also Section 6.2.3). It is therefore possible to define the effective breakdown voltage V_e^* such that $S_t(V_e^*)=1$, yielding

$$V_e^* = V_{av} + \frac{(\gamma - B)\,T_{ox}}{\log_e\left[\dfrac{(C_{sys}+C_{ox})T_{ox}}{Ak\tau_0(\gamma - B)}\right]} = V_{av} + \frac{H\,T_{ox}}{\log_e\left[\dfrac{\varepsilon_0\,\varepsilon_{ox}}{k\,\tau_0 H}\left(1+\dfrac{C_{sys}}{C_{ox}}\right)\right]} \qquad 7(13)$$

where $H=\gamma-B$ and $C_{ox}=\epsilon_0\epsilon_{ox}A/T_{ox}$. Hence the *oxide* breakdown voltage $V_{ox}^*=V_e^*-V_{av}$ can be written

$$V_{ox}^* = \frac{H\,T_{ox}}{\log_e\left[\dfrac{\varepsilon_0\,\varepsilon_{ox}}{k\,\tau_0 H}\left(1+\dfrac{C_{sys}}{C_{ox}}\right)\right]} \qquad 7(14)$$

Thus the breakdown voltage of a given oxide of given thickness T_{ox} depends solely upon the ratio of the oxide capacitance to the pulse-system capacitance. Fig.7.4(ii) shows that even this dependence is not particularly significant. The predicted values of V_{ox}^* do however agree roughly with the V_{ox}^* values predicted in Fig.6.39.

Since avalanche breakdown in the NMOS structures only occurs under positive stress, the positive and negative breakdown thresholds $V_e^{*(pos)}$ and $V_e^{*(neg)}$ are related by the formula

$$V_e^{*(pos)} = V_e^{*(neg)} + V_{av} \qquad 7(15)$$

(assuming that the value of V_{ox}^* is polarity-independent).

The circuit model of Fig.7.2 can be used to determine the applied ESD pulse voltage V_{bd} associated with a peak oxide voltage of V_{ox}^*. Since during the pulse risetime, the total charge $V_{bd}(C_x+C_1)$ must be shared out between the capacitances C_x, C_1, C_t and C_{ox}, the following charge-balance equation can be written

$$(C_x + C_1)\,V_{bd} = (C_x + C_1 + C_t)\left(V_{ox}^* + V_{av}\right) + C_{ox}\,V_{ox}^* \qquad 7(16)$$

which may be re-arranged to yield the following expression for V_{bd}:

$$V_{bd} = V_{ox}^*\left(1+\frac{C_{ox}+C_t}{C_x+C_1}\right) + V_{av}\left(1+\frac{C_t}{C_x+C_1}\right) \qquad 7(17)$$

Fig.7.5 compares the experimental V_{bd} vs. C_1 data with the predictions of Eqns.7(17). The solid curves were calculated using the capacitance-values measured in Section 4.2.3.1 (adjusting V_{ox}^* and V_{av} for optimum correlation). Although the prediction is reasonably

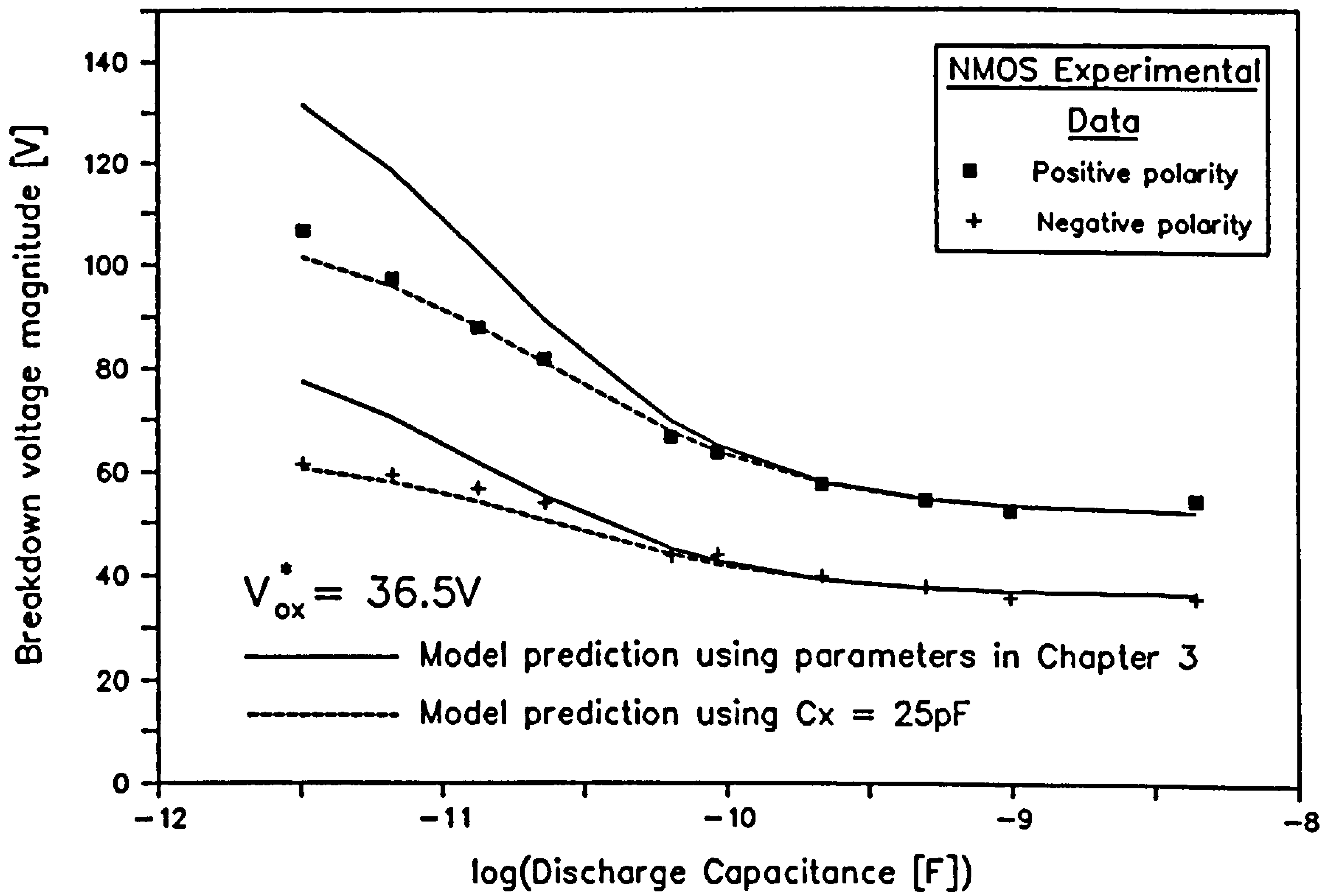


Figure 7.5: Comparison of low-R_2 ESD model with experimental V_{bd} vs. C_1 data.

accurate for large values of C_1, the accuracy is lost for $C_1 < 100pF$.

The dashed curves represent the model predictions obtained by elevating C_x to 25pF. The accuracy of these latter curves suggest that the source of the error lies in the characterization technique used to determine C_x. The fact that the optimized value of V_e^* is somewhat lower than its predicted value (see Fig.7.4) may be related to the similar error in the constant voltage model (Fig.7.1(d)). However, the fact that a single value of V_{ox}^* is sufficient to model breakdown across the whole C_1 range shows that the breakdown threshold is capacitance-independent (see Fig.7.4(ii)), suggesting that the model is basically correct.

A curious feature of this model is that it can create the illusion that breakdown is energy-dependent: If the stray capacitance C_x was not known to exist, the total pulse energy for breakdown (E_{bd}) might be supposed to equal $\frac{1}{2}C_1V_{bd}^2$, yielding the following expression for V_{bd}:

$$V_{bd} = \sqrt{\frac{2 \cdot E_{bd}}{C_1}} \qquad 7(18)$$

Fig.7.6 shows how the V_{bd} vs. C_1 curves predicted by Eqns.7(17) and 7(18) can correspond closely to each other across a wide range of C_1. It is probably this coincidence which led earlier workers to the false conclusion that ESD breakdown is thermally activated [3].

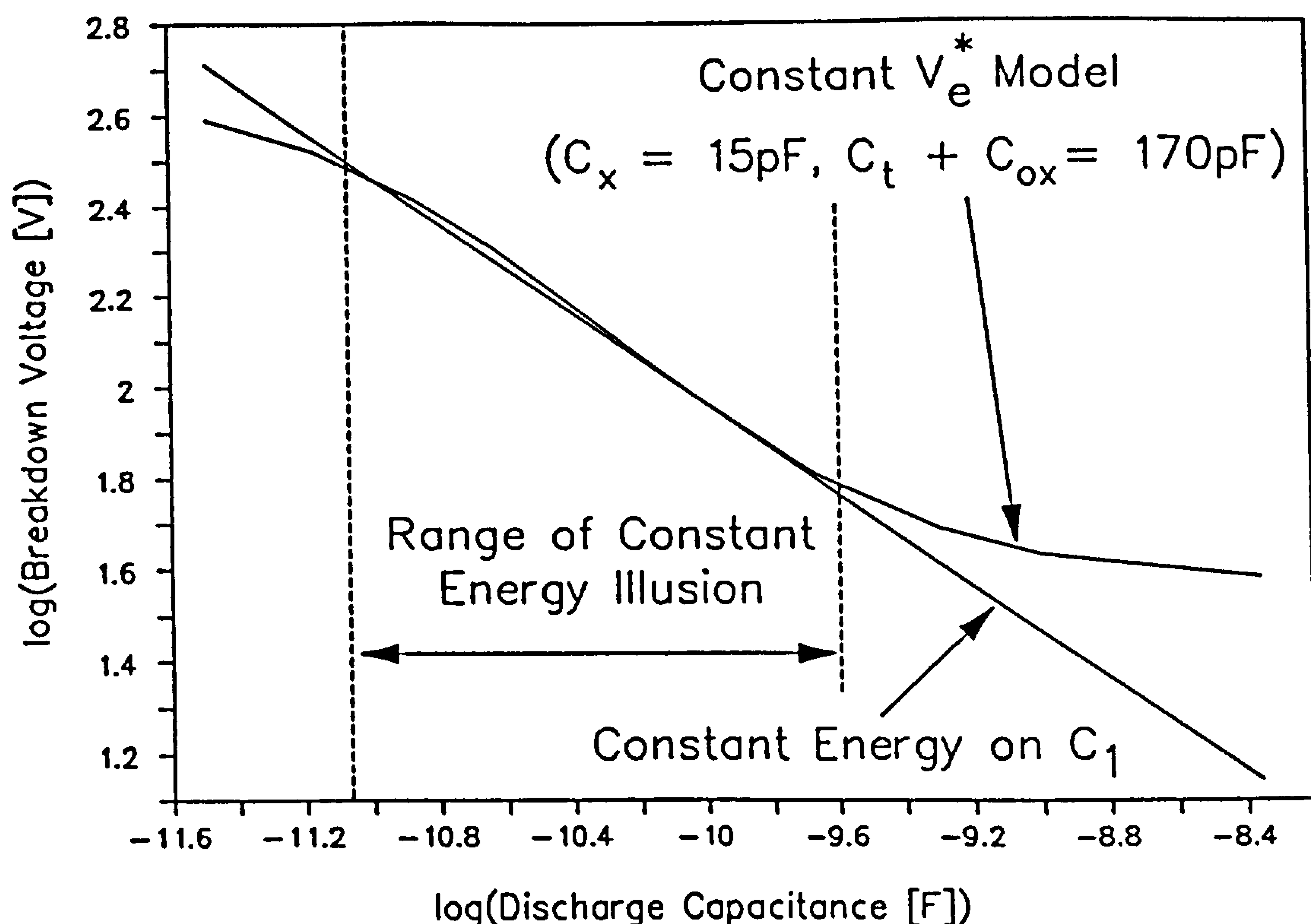


Figure 7.6: Illustration of how the constant-V_{ox}^* model can create the illusion of constant pulse energy.

7.4.1.2 Time-to-Breakdown

Eqn.7(11) allows the time-to-breakdown t_{bd} to be expressed as a function of the ESD pulse voltage. Setting $s(t_{bd})=1$ and solving for t_{bd} produces the expression

$$t_{bd} = t_d + \frac{(C_{sys}+C_{ox})T_{ox}}{AkB}\left(\left[e^{-\frac{(\gamma-B)T_{ox}}{V_e-V_{av}}} - \frac{Ak\tau_0(\gamma-B)}{(C_{sys}+C_{ox})T_{ox}}\right]^{-\frac{B}{\gamma-B}} - e^{\frac{BT_{ox}}{V_e-V_{av}}}\right) \qquad 7(19)$$

which may be simplified to yield

$$t_{bd} = t_d + \frac{\tau_0 H}{B} e^{\frac{HT_{ox}}{V_{ox}}}\left[\left(e^{-\frac{HT_{ox}}{V_{ox}}} - e^{-\frac{HT_{ox}}{V_{ox}^*}}\right)^{-\frac{B}{H}} - e^{\frac{BT_{ox}}{V_{ox}}}\right] \qquad 7(20)$$

where V_{ox} is the peak oxide voltage. When $V_{ox} \gg V_{ox}^*$, Eqn.7(20) can be simplified further to give

$$t_{bd} \approx t_d + \tau_0 e^{\frac{\gamma T_{ox}}{V_{ox}}} = t_d + \tau\left(\frac{V_{ox}}{T_{ox}}\right) \qquad 7(21)$$

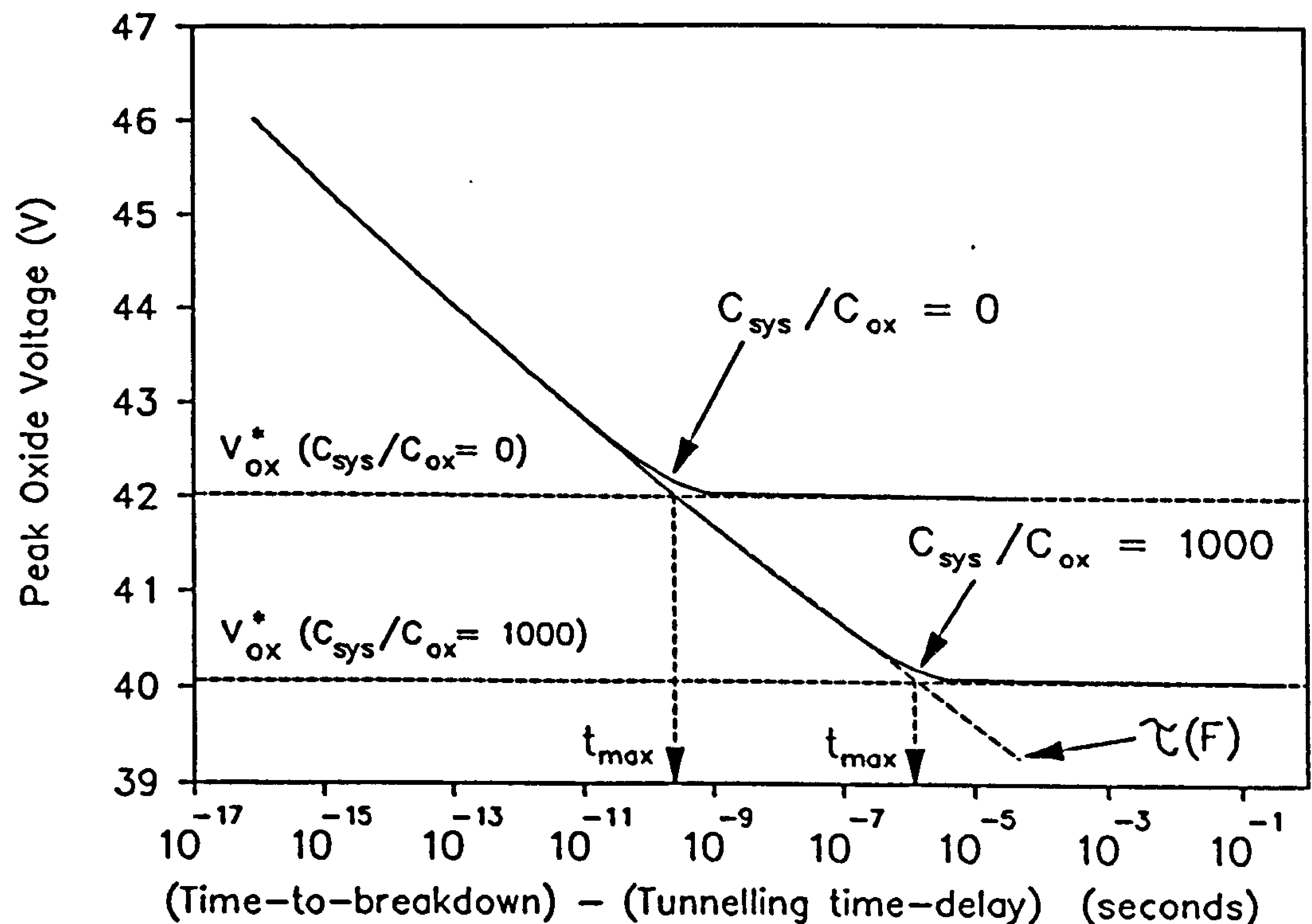


Figure 7.7: Time-to-breakdown vs. oxide voltage under ESD conditions.

Fig.7.7 shows (t_{bd}-t_d) vs. the oxide pulse voltage V_{ox} for two values of C_{sys}/C_{ox}. The value of (t_{bd}-t_d) clearly follows the τ-function until a fraction of a volt below V_{ox}^*, at which point it rises rapidly to infinity (at $V_{ox}=V_{ox}^*$). Hence the longest measurable value of t_{bd} (denoted t_{max}) is approximately equal to the sum of t_d and the value of τ at V_{ox}. The discontinuous rise in t_{bd} below this point creates the illusion that breakdown is not causal at all, but entirely voltage-dependent and time-independent.

7.4.2 High-Resistance Breakdown

7.4.2.1 Voltage Pinning Model

As R_2 exceeds the order of about 100kΩ, $AJ_{ox}^*R_2$ becomes significant, causing V_1 and V_2 to diverge during the pulse-decay. This undermines the assumption upon which Eqn.7(8) is founded, invalidating the above model. The sharpness of the Fowler-Nordheim tunnelling curve (Fig.7.4(ii)) suggests that it can be represented by an ideal voltage-reference 'diode', the pinning voltage (V_{FN}) of which can be conveniently defined as any typical value of V_{ox}^*. Hence in order to support breakdown, the oxide voltage must be increased to the pinning threshold V_{FN} and a critical current density J_{ox}^*, dependent upon C_{sys}/C_{ox} (see Fig.7.4(ii))

must be injected.

Consider the equivalent circuit of Fig.7.2. As the pulse is applied, charge is transferred from (C_1+C_x) to C_t and C_{ox}, causing the voltage V_1 to fall (from its initial value V_0), and V_2 to rise. When the oxide voltage reaches the Fowler-Nordheim pinning threshold,

$$V_2 = V_{av} + V_{FN} \qquad 7(22)$$

and

$$V_1^{(0)}(C_x + C_1) = V_0(C_x + C_1) - V_{FN}C_{ox} - C_t(V_{FN} + V_{av}) \qquad 7(23)$$

where $V_1^{(0)}$ is the value of V_1 at the exact moment the pinning threshold is reached. Since $dV_2/dt=d(V_{FN}+V_{av})/dt=0$, the oxide injection current must equal the sum of the currents in R_2 and C_2, i.e.

$$AJ = -(C_x + C_1)\frac{dV_1}{dt} = \frac{(V_1 - V_2)}{R_2} + C_2\frac{dV_1}{dt} \qquad 7(24)$$

which may be re-arranged to yield the differential equation

$$(C_x + C_1 + C_2)\frac{dV_1}{dt} = \frac{V_2 - V_1}{R_2} \qquad 7(25)$$

For the purpose of solving this equation, t=0 will be defined as the instant at which V_1 reaches $V_1^{(0)}$. Under this boundary condition, the solution to Eqn.7(25) becomes

$$V_1(t) = V_2 + \left(V_1^{(0)} - V_2\right) e^{-\frac{t}{\tau_c}} \qquad 7(26)$$

where $\tau_c=R_2(C_x+C_1+C_2)$. Hence according to Eqn.7(24), the injection current profile may be written

$$A \cdot J(t) = I_{max}\, e^{-\frac{t}{\tau_c}} \qquad 7(27)$$

where

$$I_{max} = \frac{(C_x + C_1)\left(V_1^{(0)} - V_2\right)}{R_2(C_x + C_1 + C_2)} \qquad 7(28)$$

At the breakdown threshold, I_{max} must clearly equal $A.J_{ox}^*$. If this criterion is inserted into Eqn.7(28) and $V_1^{(0)}$ and V_2 are eliminated using Eqns.7(22) and 7(23), the following

expression for $V_0=V_{bd}$ may be derived:

$$V_{bd} = V_{FN}\left(1 + \frac{C_{ox}+C_t}{C_x+C_1}\right) + V_{av}\left(1 + \frac{C_t}{C_x+C_1}\right) + A\,J_{ox}^{*}R_2\left(1 + \frac{C_2}{C_x+C_1}\right) \quad 7(29)$$

In order to evaluate Eqn.7(29), it is necessary to determine the value of J_{ox}^{*}. This can be achieved by considering the shape of the injection current profile J(t) during voltage pinning. According to Eqns.7(27-28), J(t) is given by

$$J(t) = J_{max}\,e^{-\frac{t}{\tau_c}} \quad 7(30)$$

where $\tau_c=R_2(C_x+C_1+C_2)$ and $J_{max}=(C_x+C_1)(V_1^{(0)}-V_2)/A\tau_c$. Hence the charge-injection profile Q(t) can be determined:

$$Q(t) = \int_0^t J(t)\,dt = J_{max}\tau_c\left(1-\exp\left[-\frac{t}{\tau_c}\right]\right) = \tau_c\left(J_{max} - J\right) \quad 7(31)$$

The total *injectable* charge is therefore given by $Q(\infty) = J_{max}\tau_c$, which must equal Q_{bd} in order to support oxide breakdown. Solving Eqn.7(31) for J and inserting it into Eqn.7(6) yields the following expression for the *charge-to-breakdown* function

$$\Theta(Q) = \tau_0 J_0^{\frac{\gamma}{B}}\left(J_{max} - \frac{Q}{\tau_c}\right)^{1-\frac{\gamma}{B}} \quad 7(32)$$

which when inserted into Eqn.7(5) yields

$$S_t = \int_0^{J_{ox}^{*}\tau_c} \frac{dQ}{\Theta(Q)} = \frac{\tau_c B}{\tau_0\gamma}\left(\frac{J_{ox}^{*}}{J_0}\right)^{\frac{\gamma}{B}} = 1 \;\rightarrow\; J_{ox}^{*} = J_0\left(\frac{\tau_0\gamma}{\tau_c B}\right)^{\frac{B}{\gamma}} \quad 7(33)$$

Substituting this expression into Eqn.7(29) together with $\tau_c=R_2(C_x+C_1+C_2)$ yields

$$V_{bd} = V_{FN}\left(1+\frac{C_{ox}+C_t}{C_x+C_1}\right) + V_{av}\left(1+\frac{C_t}{C_x+C_1}\right) + A\,\frac{(C_c+C_1+C_2)^{m_1}}{C_x+C_1}\,G_1 R_2^{m_1} \quad 7(34)$$

where

$$m_1 = 1 - \frac{B}{\gamma} \quad ; \quad G_1 = J_0\left(\frac{\tau_0\gamma}{B}\right)^{\frac{B}{\gamma}} \quad 7(35)$$

It is clear that as $R_2 \to 0$, Eqn.7(34) resembles Eqn.7(17), with V_{ox}^* replaced by V_{FN}. Hence the high-resistance model presents a more general version the model of Section 7.4.1.

7.4.2.2 Field-Dependent (FIM) Breakdown

The above expressions relates to breakdown driven by current in R_2. If breakdown is field-induced via the capacitance C_2, an alternative model is required. Assume that when V_0 is equal to the field-induced breakdown threshold V_{bd}^f, a voltage of $V_{FN}+V_{av}$ is induced across the DUT at the moment the pulse is applied. If a constant charge-displacement exists around the circuit then:

$$V_{FN}+V_{av}+\frac{C_{ox}V_{FN}+C_t(V_{FN}+V_{av})}{C_2}=\frac{V_{bd}^f(C_x+C_1)-[C_{ox}V_{FN}+C_t(V_{FN}+V_{av})]}{C_x+C_1} \qquad 7(36)$$

Since the circuit resistance is limited to that of the bulk silicon (approx. 500Ω), the series voltage dropped by $A.J_{ox}^*$ is negligible. Eqn.7(36) can now be re-arranged to yield

$$V_{bd}^f = V_{FN}\left(1+\frac{1}{\alpha_c}\frac{C_{ox}+C_t}{C_x+C_1}\right)+V_{av}\left(1+\frac{1}{\alpha_c}\frac{C_t}{C_x+C_1}\right) \qquad 7(37)$$

where the coupling coefficient α_c is given by

$$\alpha_c = \frac{C_2}{C_x+C_1+C_2} \qquad 7(38)$$

(Note that as $C_2 \to \infty$, the capacitive coupling becomes an open-circuit and Eqn.7(37) becomes identical to Eqn.7(17).) The actual ESD breakdown data follows either Eqn.7(34) or 7(37), depending upon which of the two predicts the lowest breakdown voltage.

7.4.3 Comparison with Experimental Data

7.4.3.1 Empirical Fitting of Models to ESD Data

Fig.7.8 shows the V_{bd} vs. R_2 curves predicted by Eqns.7(29) and 7(32), fitted to the NMOS ESD data (reproduced from Fig.6.12, using the average breakdown voltage for each value of C_1 or R_2). The models clearly provide an accurate description of the data over a wide range of conditions. Particularly interesting is the sharp 'switch' between the two

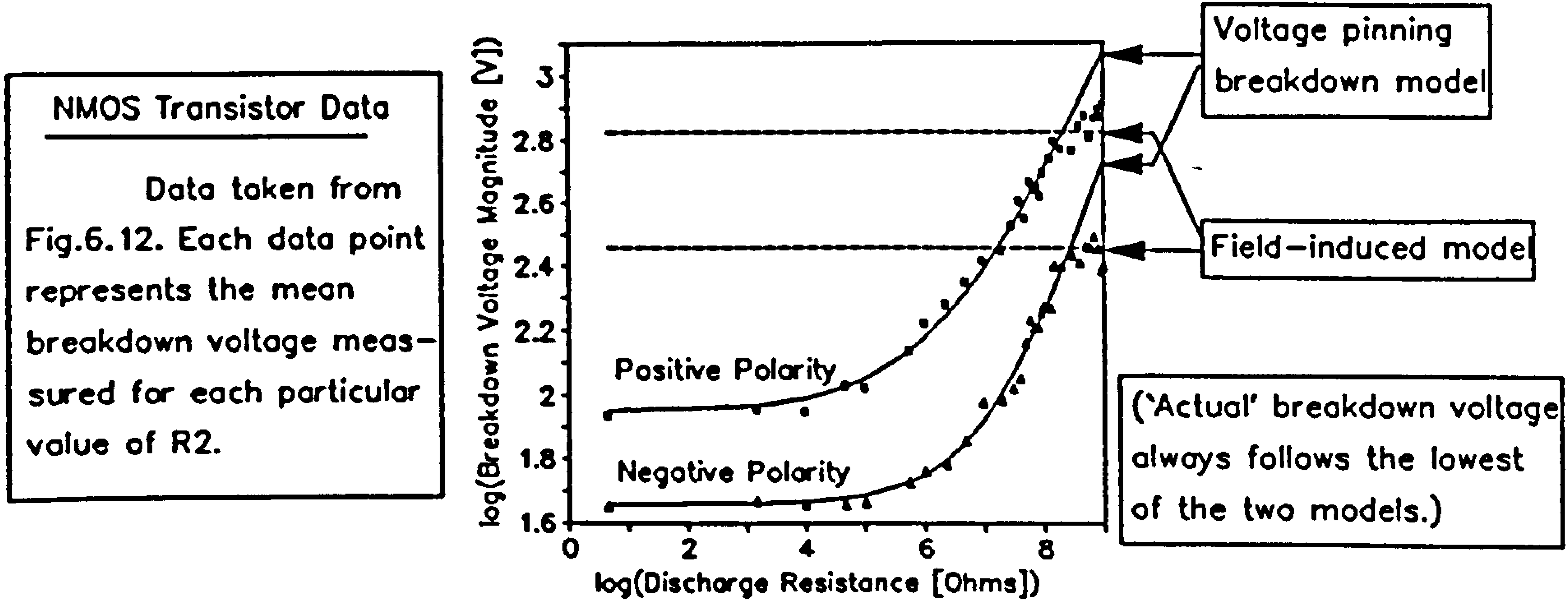


Figure 7.8: Empirical fitting of ESD data to analytical model.

breakdown models at R_2=100MΩ.

The independently measured ESD circuit parameters (see Section 4.2.3.1) were used except that an anomalously large value of C_2 (3.34pF rather than 1.029pF) was needed in order to fit the field-induced breakdown model to the high-R_2 tail of the data. This discrepancy (together with the C_x discrepancy in Section 7.4.1.1) is believed to be due to inaccuracy in the characterization technique employed in Chapter 4.

In fitting the model to the data, the values of V_{FN}, V_{av}, G_1 and m_1 were adjusted to give optimum correlation. Table 7.1 shows all parameter values used.

Table 7.1: Optimized Parameter Values

Parameter	Negative Polarity	Positive Polarity
V_{FN}	38.952V	38.952V
V_{av}	0V	32V
G_1	0.005	0.007
m_1	0.555	0.419

7.4.3.2 Comparison With Constant Current/Voltage Data

Eqn.7(35) allows the parameters m_1 and G_1 to be computed from the values of J_0, τ_0, γ and B determined in Fig.7.1. Fig.7.9 shows the negative-polarity V_{bd} vs. R_2 curve calculated from these parameters, compared with the experimental data. It is clear that the predicted breakdown thresholds are universally too large.

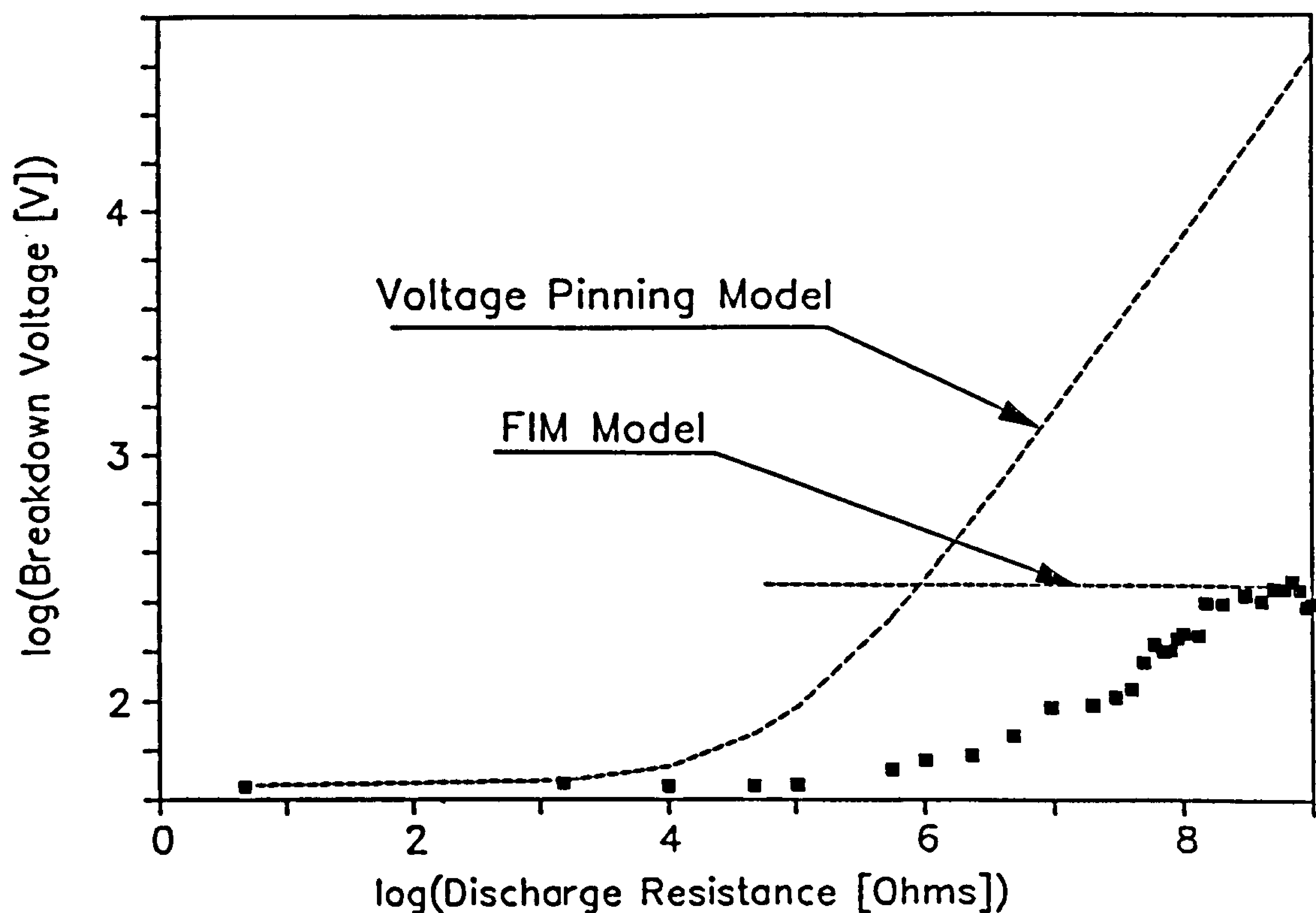


Figure 7.9: Comparison of Single-Pulse HBM voltage pinning model and experimental data.

A possible explanation for this discrepancy is that Eqn.7(29) predicts the criterion for breakdown under a *single* ESD pulse, rather than under an incrementing pulse sequence (i.e. a 'step-test'). Although the two situations are identical in the low-R_2 domain (see Section 7.4.1.1), they are not necessarily the same under high-resistance conditions.

If the current increment is ΔJ, then the total damage function $S_{seq}(n)$ for a sequence of n pulses is given by

$$S_{seq} = \sum_{r=0}^{n} \frac{\tau_c B}{\tau_0 \gamma} \left(\frac{n \cdot \Delta J}{J_0} \right)^{\frac{\gamma}{B}} \qquad 7(39)$$

If a large number of pulses are required for breakdown then the summation in Eqn.7(34) may be replaced by an integral, i.e.

$$S_{seq} = \int_0^{J_{max}} \frac{1}{\Delta J} \frac{\tau_c B}{\left(\tau_0 \gamma J_0^{\frac{\gamma}{B}} \right)} J^{\frac{\gamma}{B}} dJ = \left(\frac{\tau_c B^2}{\Delta J \cdot \tau_0 \gamma J_0^{\frac{\gamma}{B}} (\gamma + B)} \right) J_{max}^{\frac{\gamma + B}{B}} \qquad 7(40)$$

where J_{max} is the maximum current of the pulse sequence. According to Eqns.7(18) and

7(23), the current increment ΔJ corresponds to an applied voltage increment ΔV of

$$\Delta V = A \cdot \Delta J \cdot \frac{R_2(C_x + C_1 + C_2)}{C_x + C_1} \qquad 7(41)$$

The breakdown criterion clearly requires $S_{seq}=1$ at $J_{max}=J_{ox}^*$, Eqn.7(40) may be re-arranged to yield an expression for J_{ox}^*. This may be combined with the appropriate expressions for ΔJ (from Eqn.7(41)) and τ_c $[=R_2(C_x+C_1+C_2)]$ to yield

$$J_{ox}^* = \left[\Delta V \cdot \frac{C_x + C_1}{R_2^2 (C_x + C_1 + C_2)^2 A} \cdot \frac{\tau_0 \gamma J_0^{\frac{\gamma}{B}} (\gamma + B)}{B^2}\right]^{\frac{B}{\gamma + B}} \qquad 7(42)$$

This may be inserted into Eqn.7(29), to yield the following breakdown voltage expression

$$V_{bd} = V_{FN}\left(1+\frac{C_{ox}+C_t}{C_x+C_1}\right) + V_{av}\left(1+\frac{C_t}{C_x+C_1}\right) + A^{n_2}(\Delta V)^{p_2}\frac{(C_c + C_1 + C_2)^{m_2}}{(C_x + C_1)^{n_2}} G_2 R_2^{m_2} \qquad 7(43)$$

where

$$m_2 = \frac{\gamma - B}{\gamma + B} \; ; \; n_2 = \frac{\gamma}{\gamma + B} \; ; \; p_2 = \frac{B}{\gamma + B} \qquad 7(44)$$

and

$$G_2 = \left[\frac{\tau_0 \gamma J_0^{\frac{\gamma}{B}} (\gamma + B)}{B^2}\right]^{\frac{B}{\gamma + B}} \qquad 7(45)$$

However, Fig.7.10 shows that although this new development to the model does significantly reduce the predicted values of V_{bd} for large R_2s, the reduction is not sufficient to explain the experimental data.

Fig.7.11(a) provides a clue as to a possible reason for this inconsistency: The left-hand graph shows total ESD charge (Q_{tot}) vs. peak injection-current (J_{max}), superimposed upon the Q_{bd} vs. J data from Fig.7.1(d) (for $V_{bd} < V_{bd}^f$). For the former curve, the total ESD charge was determined using the expression:

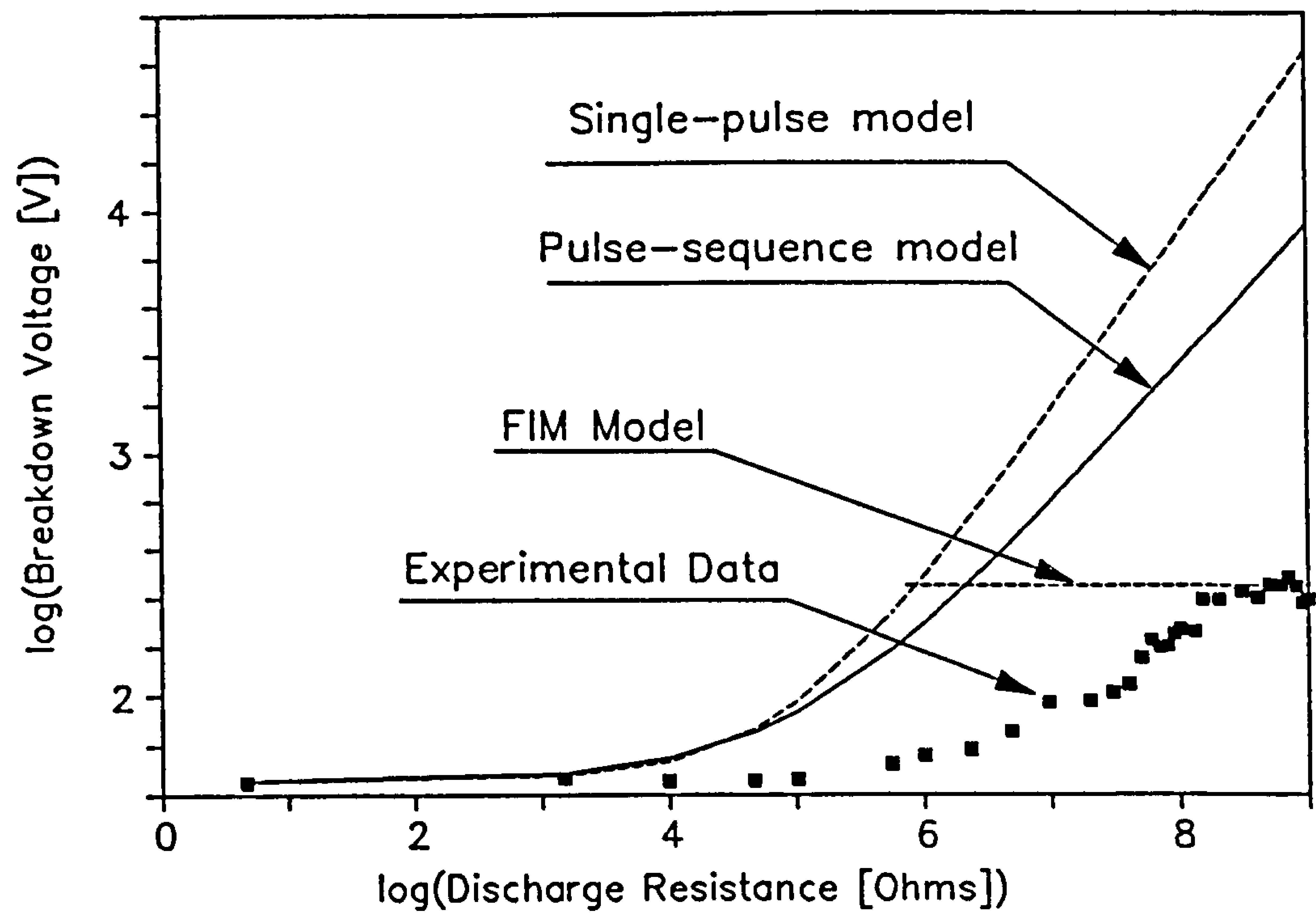


Figure 7.10: Comparison of pulse-sequence HBM voltage pinning model and experimental data.

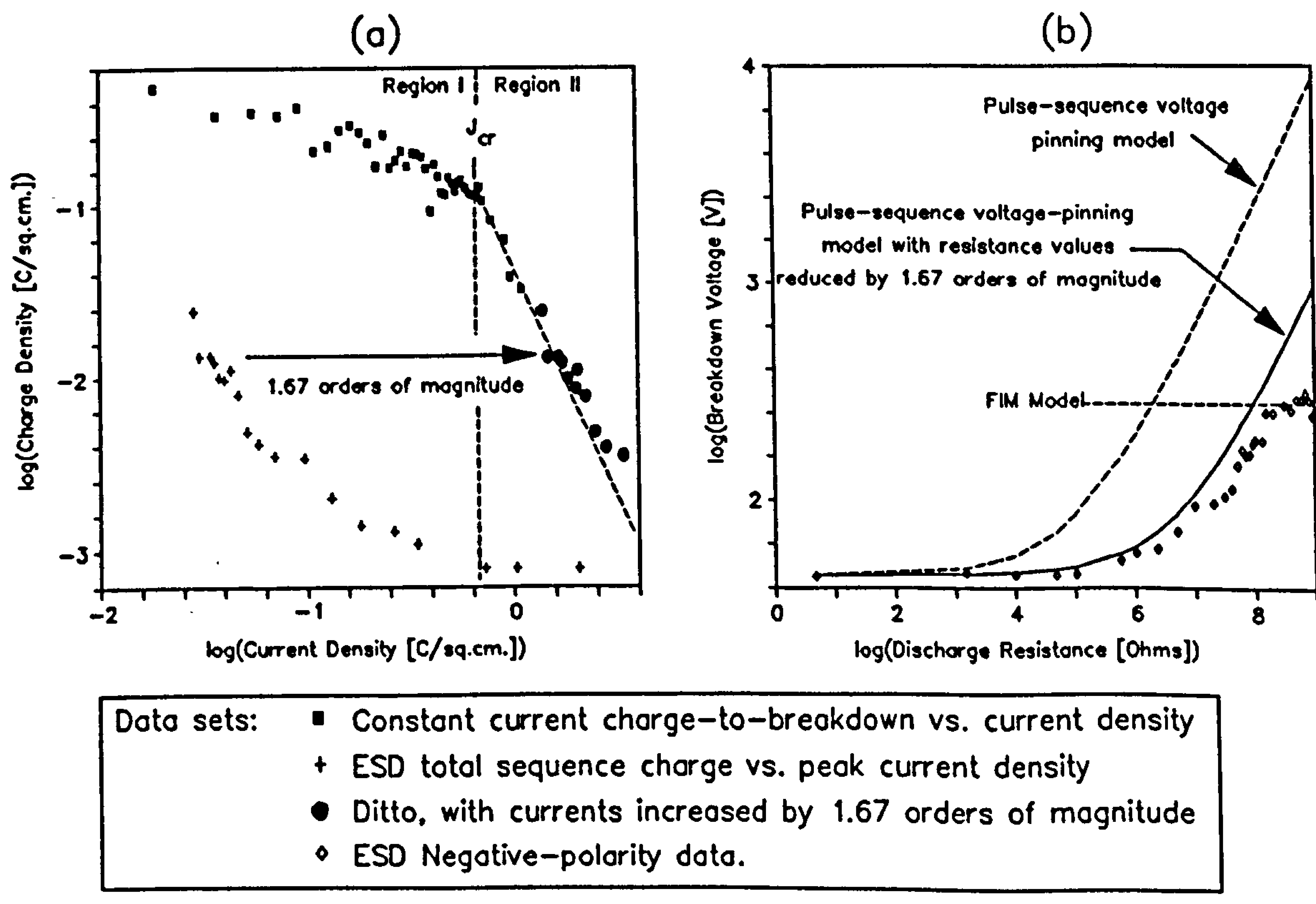


Figure 7.11: Comparison of Total ESD charge vs. estimated peak current data with the constant-current Q_{bd} vs. J results.

$$Q_{tot} = \frac{1}{A}\sum_{r=1}^{V_{bd}/\Delta V}\left(C_c + C_1\right) r\cdot\Delta V - \frac{1}{A}\frac{\left(C_x + C_1\right)}{2\cdot\Delta V} V_{bd}^2 \qquad 7(46)$$

while the peak current was given by Eqn.7(28). The resulting graphs show that the constant-current Q_{bd} is more than an order of magnitude larger than its ESD counterpart. This eliminates the possibility that the discrepancy arises from minor errors in the estimation of parameters. Since there is no reason to doubt the methods used to determine Q_{bd} and Q_{tot}, the discrepancy may be due to an underestimation of the peak ESD by a factor of $10^{1.67}$ (see Fig.7.11(a)). According to Eqn.7(28), a times $10^{1.67}$ increase in J_{max} may be modelled by scaling R_2 by a factor of $10^{-1.67}$. Fig.7.11(b) shows how such scaling eliminates nearly all of the discrepancy between the model and the data. Hence a physical justification for the enhancement of J_{max} would lead to a fully consistent model of breakdown. This problem is addressed in the following section.

7.4.4 'Enhanced Injection' Model of High-Resistance Breakdown

The expression (i.e. Eqn.7(28)), from which the peak ESD current was determined, assumes that the Fowler-Nordheim current imposes an ideal voltage clamp. However, the finite tunnelling time-delay t_d may cause the oxide voltage to overshoot its pinning value, creating a temporarily enhanced injection current (Fig.7.12(a)). The large voltage sensitivity of the tunnelling current (see Fig.7.1(b)) may result in an enhancement of more than an order of magnitude, providing the current magnification required for correlation (see Fig.7.11). The oxide voltage overshoot observed in the ramp-voltage profiles of Fig.6.27 adds credence to this theory.

The overshoot mechanism is illustrated in Fig.7.12(a), in which the time delay between the oxide voltage reaching V_{FN} and the onset of tunnelling is assumed to be equal to the tunnelling time-delay t_d. If the 'ramp-rate' r (=dV_{ox}/dt at $V_{ox}=V_{FN}$) remains approximately constant throughout t_d, then the peak oxide voltage V_{ox}^{max} (at the end of t_d) is given by the expression

$$V_{ox}^{max} = V_{FN} + r\cdot t_d \qquad 7(47)$$

Simple circuit theory shows that the ramp rate r and the applied pulse voltage V_0 are related by the equation

Figure 7.12: Illustration of t_d-induced overshoot model.

$$V_0 = V_{FN}\left(1 + \frac{C_{ox} + C_t}{C_x + C_1}\right) + V_{av}\left(1 + \frac{C_t}{C_x + C_1}\right) + R_2\left[(C_t + C_{ox})\left(1 + \frac{C_2}{C_x + C_1}\right) + C_2\right] r \quad 7(48)$$

If a critical value of V_{ox}^{max} is required to support breakdown, then the breakdown criterion may be expressed in terms of a critical ramp rate $r=r^*$, which corresponds to $V_{ox}^{max}=V_{ox}^*$. Hence the breakdown voltage may be written

$$V_{bd} = V_{FN}\left(1 + \frac{C_{ox} + C_t}{C_x + C_1}\right) + V_{av}\left(1 + \frac{C_t}{C_x + C_1}\right) + R_2\left[(C_t + C_{ox})\left(1 + \frac{C_2}{C_x + C_1}\right) + C_2\right] r^* \quad 7(49)$$

In order to obtain an analytical expression for V_{ox}^*, the section of the circuit model (Fig.7.2) to the right of R_2 must be assumed to act in isolation. The oxide breakdown voltage for such an isolated circuit can be determined from Eqn.7(14), setting $C_{sys}=C_t$, i.e.

$$V_{ox}^{*} = \frac{(\gamma - B)\,T_{ox}}{\log_e\left[\frac{\varepsilon_0\,\varepsilon_{ox}}{k\,\tau_0\,(\gamma - B)}\left(1 + \frac{C_t}{C_{ox}}\right)\right]} \qquad 7(50)$$

Hence Eqns.7(47-50) may be combined to give a complete model for the single-pulse breakdown threshold.

However, if the oxides are subjected to a step-test, the damage inflicted by sub-critical pulses needs to be included in the model. As in Section 7.4.3.2, the total sequence damage S_{seq} is given by summing all the pulse damages throughout the sequence, i.e.

$$S_{seq} = \sum_{r=1}^{V_{ox}^{P}/\Delta V_{ox}} \frac{\varepsilon_0\,\varepsilon_{ox}}{k\,\tau_0\,(\gamma - B)}\, e^{-\frac{H\,T_{ox}}{r\cdot\Delta V_{ox}}} \qquad 7(51)$$

where V_{ox}^{P} and ΔV_{ox} are the peak voltage in the sequence and the voltage resolution (as they appear across the oxide, see Fig.7.12(b)). For a large number of pulses, the summation may be replaced by an integral:

$$S_{seq} \approx \frac{1}{\Delta V}\,\frac{\varepsilon_0\,\varepsilon_{ox}}{k\,\tau_0\,(\gamma - B)}\left(1 + \frac{C_t}{C_{ox}}\right)\int_0^{V_{ox}^{P}} e^{-\frac{(\gamma - B)T_{ox}}{V_{ox}}}\, dV_{ox} \qquad 7(52)$$

Assuming that most of the oxide damage occurs when V_{ox} is in the vicinity of V_{FN}, the integral in Eqn.7(52) can be approximated:

$$\int_0^{V_{ox}^{P}} e^{-\frac{(\gamma - B)T_{ox}}{V_{ox}}}\, dV_{ox} \approx V_{FN}^{2}\int_0^{V_{ox}^{P}} \frac{1}{V_{ox}^{2}}\, e^{-\frac{(\gamma - B)T_{ox}}{V_{ox}}}\, dV_{ox} = \frac{V_{FN}^{2}}{(\gamma - B)\,T_{ox}}\, e^{-\frac{(\gamma - B)T_{ox}}{V_{ox}^{P}}} \qquad 7(53)$$

When V_{ox}^{P} is equal to the *sequence* breakdown voltage $V_{ox}^{*(seq)}$, S_{seq} must be equal to 1. Substituting these values into Eqn.7(53) and combining the result with Eqn.7(50) yields the following expression for $V_{ox}^{*(seq)}$:

$$V_{ox}^{*(seq)} = \frac{V_{ox}^{*}\cdot V_{seq}}{V_{ox}^{*} + V_{seq}} \qquad 7(54)$$

where V_{ox}^{*} is the single-pulse breakdown voltage (Eqn.7(50)) and V_{seq} is given by

$$V_{seq} = \frac{(\gamma - B)T_{ox}}{\log_e\left[\frac{V_{FN}^2}{(\gamma - B)\,T_{ox}\,\Delta V_{ox}}\right]} \qquad 7(55)$$

The relationship between V_{ox}^{max} and V_0 can be determined from Eqns.7(47-48), and hence ΔV_{ox} can be shown to be given by given by

$$\Delta V_{ox} = \frac{t_d \cdot \Delta V}{R_2\left((C_t + C_{ox})\left(1 + \frac{C_t}{C_x + C_1}\right) + C_2\right)} \qquad 7(56)$$

where ΔV is the applied pulse voltage resolution. All these equations may now be combined in order to produce an expression for V_{bd}.

Fig.7(13) shows the curves predicted by the single-pulse and pulse-sequence models, assuming t_d=20μs (see Fig.6.14(c)). The multiple-pulse model agrees reasonably well with the experimental data, although its slope is generally too large. This error may be due to the fact that the model assumes that t_d is constant, while in practice it is known to vary slightly with experimental conditions.

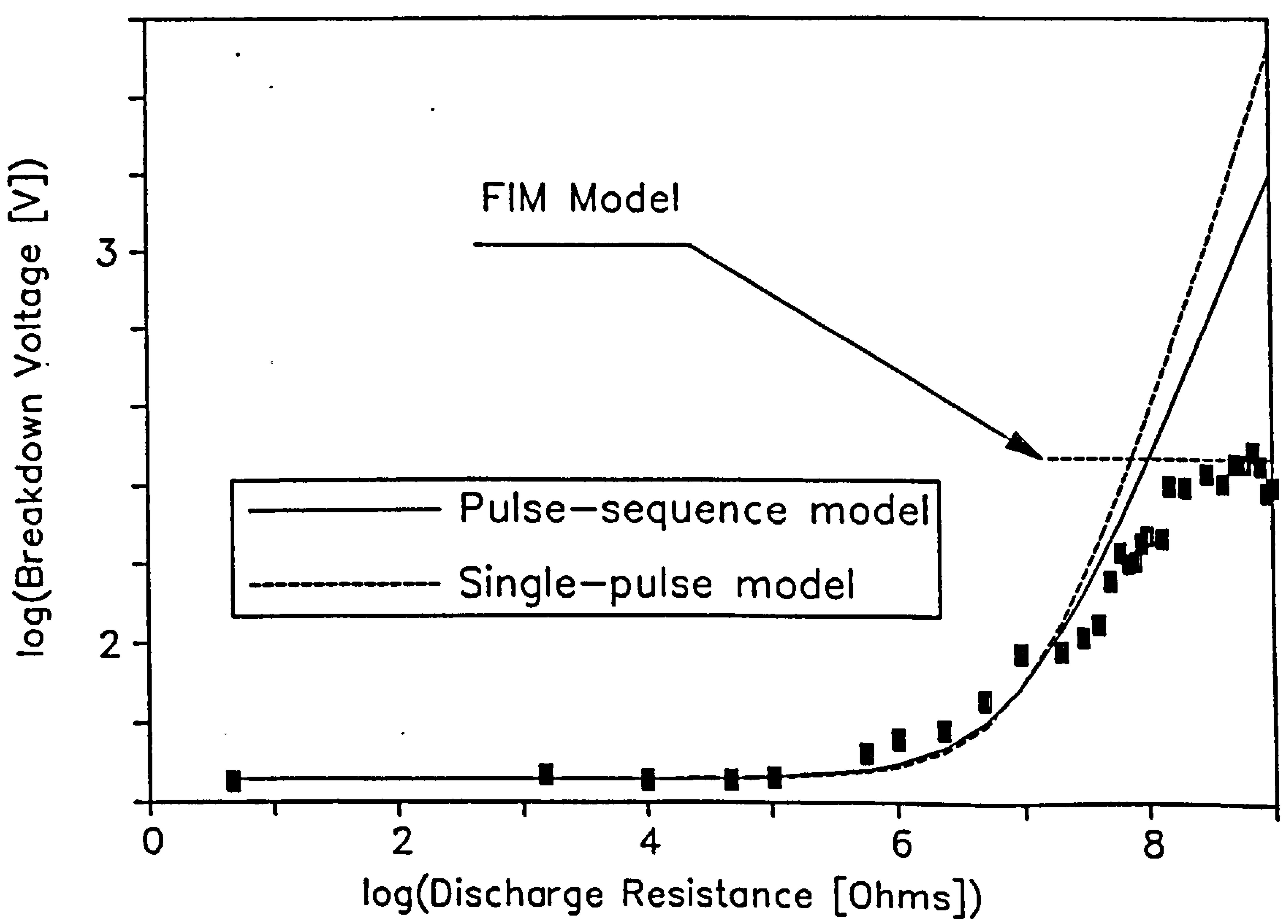


Figure 7.13: Comparison between 'critical-overshoot' model and experimental data.

7.5 Summary and Conclusions

The analysis in Section 7.4 has shown how low-resistance (i.e. HBM or MM) ESD can be modelled in terms of the discharge of a single capacitance into a Fowler-Nordheim junction. The results are consistent with the experimental data, indicating a sharply defined breakdown voltage (largely independent of stress conditions), below which the oxide remains robust and above which the oxide wears out very rapidly. However, the model assumes that the oxide displacement current is independent of any perturbation of the oxide space-charge during stress. Simple theory shows that if the electron traps are concentrated in a charge sheet positioned x_t from the cathode, then the displacement current is given by $C_{ox}.dV_{ox}/dt + q.(T_{ox}/x_t-1).dn/dt$ where n is the trapped electron density. The neglecting of the latter term and the subsequent simplification to $C_{ox}.dV_{ox}/dt$ appears to be justified by its successful application in Section 6.5.3 to the calculation of ESD charge injection (i.e. the extrapolated injection charge in Fig.6.35 remains zero prior to breakdown).

Section 7.4.2 developed a model of high-resistance ($100k\Omega < R_2 < 100M\Omega$) breakdown, based upon the assumption that the Fowler-Nordheim current pins the oxide voltage throughout the pulse decay. Parameter optimization produced excellent correlation with the experimental data under both positive and negative polarity conditions. The extremely high resistance data ($R_2 > 100M\Omega$) were also successfully modelled in terms of a field-induced ESD pulse propagated via the parasitic capacitance C_2. The optimized parameters were, however, wildly inconsistent with their corresponding values measured from the constant current/voltage stress data. This anomaly was solved by including the current-enhancement effects of the oxide-voltage overshoot associated with the tunnelling time-delay t_d.

The optimized values of the parasitic capacitances C_2 and C_x were also significantly greater than their corresponding values measured during circuit characterization (Table 4.2). It seems most likely that this anomaly is due to an inaccuracy in the characterization techniques employed in Chapter 4. For example, it may be that the parallel C/R model of the oscilloscope probe is not fully adequate under high-frequency conditions.

The next logical stage in this work should involve the development of a more detailed model of oxide breakdown based upon physical principles, and the comparison of the results with the analytical model of Section 7.3. This model must firstly include the effects of CRT trapping/de-trapping upon the oxide charge profile. A first attempt at such an analysis has already been performed by the author [3] and the paper is reproduced in Appendix C. Although this model produced some excellent results, it was based upon the assumption that field-assisted thermionic emission (i.e. the Poole-Frenkel effect) governs de-trapping. The observed temperature dependence of t_d (Fig.6.18) presents a strong argument against this theory, suggesting that trap-to-band tunnelling and/or trap-to-band impact ionization (TBII)

are the genuine mechanisms. Although TBII has already been used by Avni and Shappir [4] to model steady-state trap occupancy, no time-domain analysis has (to the author's knowledge) been published.

Finally, the dielectric wearout mechanism itself must be modelled. Chapter 6 concluded that wearout was probably the result of hole-charge generation and trapping in the cathode. These holes are probably generated at the anode by the emission and decay of surface plasmons [4], and tunnel back into the oxide via the Si-SiO_2 potential barrier.

7.6 References

1. Wolters,D.R., van der Schoot,J.J., *"Dielectric Breakdown in MOS Devices - Part II: Conditions for the Intrinsic Breakdown"*, Phillips J. of Res., 40(3), pp.137-63, 1985.
2. Moazzami,R, Lee,J.C., Hu,C, *"Temperature Acceleration of Dielectric Breakdown"*, IEEE Trans. Electron Dev., 36(11), pp.2463-5, 1989.
3. Tunnicliffe,M.J., Dwyer,V.M., Campbell,D.S., *"A Model of ESD Oxide Breakdown in Unprotected MOS Transistors"*, Proc. European ESD Symposium, Eindhoven, Neth., 1991.
4. Avni,E, Shappir,J, *"Trap Generation and Occupation in Stressed Gate Oxides under Spatially Variable Oxide Electric Field"*, Appl. Phys. Lett., 51(22), pp.1857-9, 1987.
5. Fischetti,M.V., *"High Field Electron Transport in SiO_2 and Generation of Positive Charge at the Si-SiO_2 Interface"*, Insulating Films on Semiconductors, pp.181-9, 1986.

Chapter 8

Latent Failure and Parametric Drift in Damaged MOS Devices

8.1 Introduction

So far, this thesis has concentrated on the causes of dielectric failure in MOS structures, rather than the effects of the subsequent damage. This chapter is therefore devoted to studying the impact of ESD damage upon device and circuit performance, with particular reference to *Latent failure* and *parametric drift*.

The concept of 'latent' or 'walking-wounded' failure has already been introduced in Chapter 3. To recap, a device may receive an ESD pulse during fabrication, which causes it to sustain sub-catastrophic damage. Such a device may pass a functional quality inspection, enter the field, and fail prematurely during working life. The possible consequences of such failure have prompted considerable experimental work over the past decade: In the 1980s, McAteer et al. [1,2] confirmed that sub-catastrophic ESD can reduce the reliability of electronic components. Crockett's experiments on packaged CMOS [3] showed that low-voltage (1kV) ESD pulses rendered devices more sensitive to subsequent high-voltage (2.5kV) stress. Later work by Aur et al [4] showed a reduced hot-electron reliability in MOSFETs subjected to ESD pulses.

Another consequence of sub-catastrophic ESD is parametric drift, i.e. the variation of the operational parameters of a damaged device. Holmes [5], for example, showed that an ESD pulse below the breakdown threshold produced a negative shift in the strong-inversion threshold voltage V_T. This was often sufficient to drive enhancement mode MOSFETs into depletion mode, causing logical malfunctions. Several other workers [6-8] have found that ESD-induced oxide breakdown distorts a MOSFETs I/V characteristics, causing a range of degradation modes, ranging from reduced-g_m characteristics to a total loss of transistor action. Fig.8.1 shows these degradation modes as observed by Amerasekera & Campbell [6]. Soden and Hawkins [8] found that g_m reduction in CMOS circuits is accompanied by an increase in the supply current, introducing a battery-failure hazard in portable equipment. These phenomena were also predicted by Syrzycki's theoretical analysis [9].

The present chapter examines the relationships between latent damage and characteristic variation. Attempts are made to correlate parametric drift and oxide wearout in MOSFETs subjected to ESD. The degradation of walking-wounded devices to catastrophic failures under *working* voltage conditions is also studied.

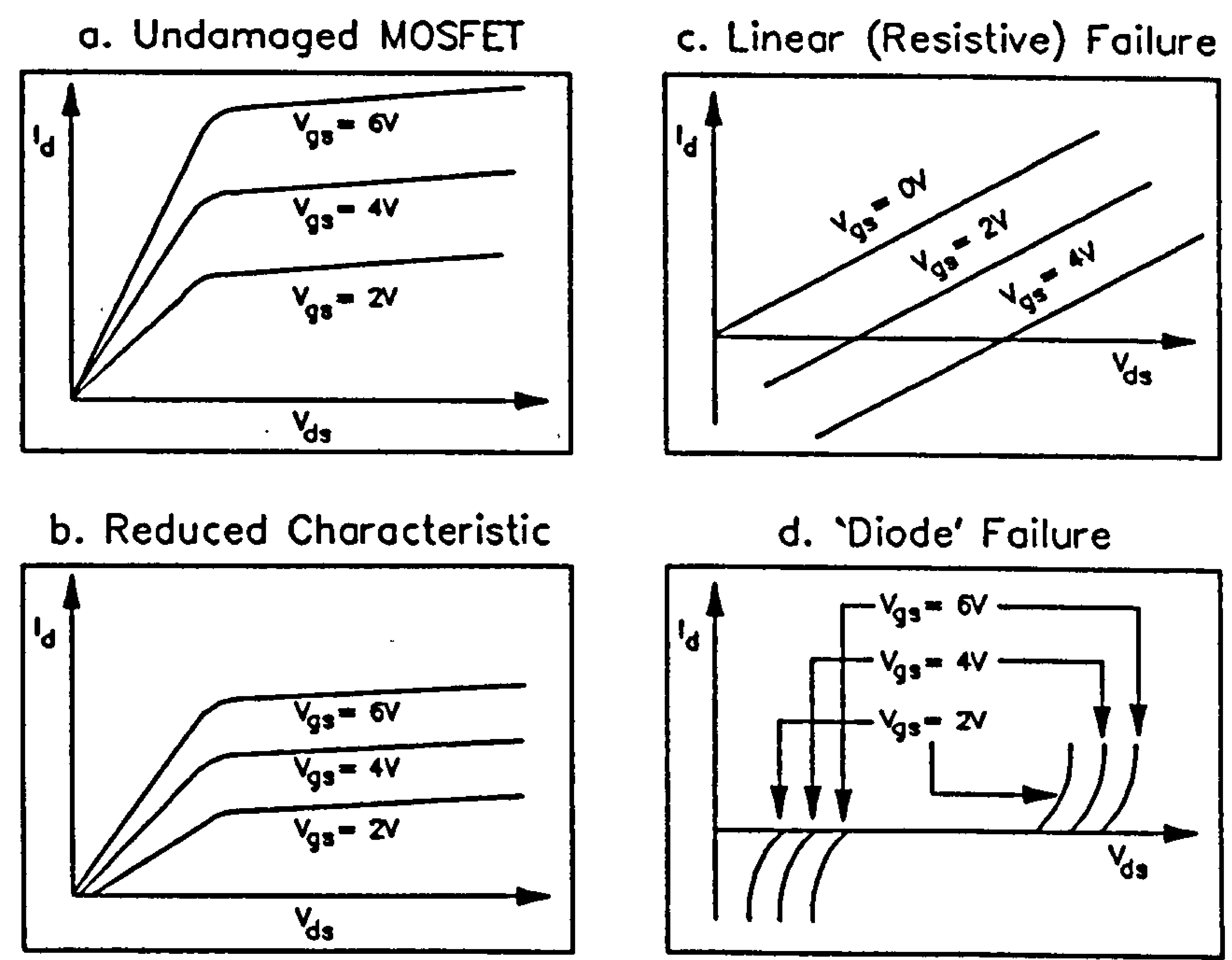


Figure 8.1: MOSFET Faulure Modes identified by Amerasekera & Campbell.

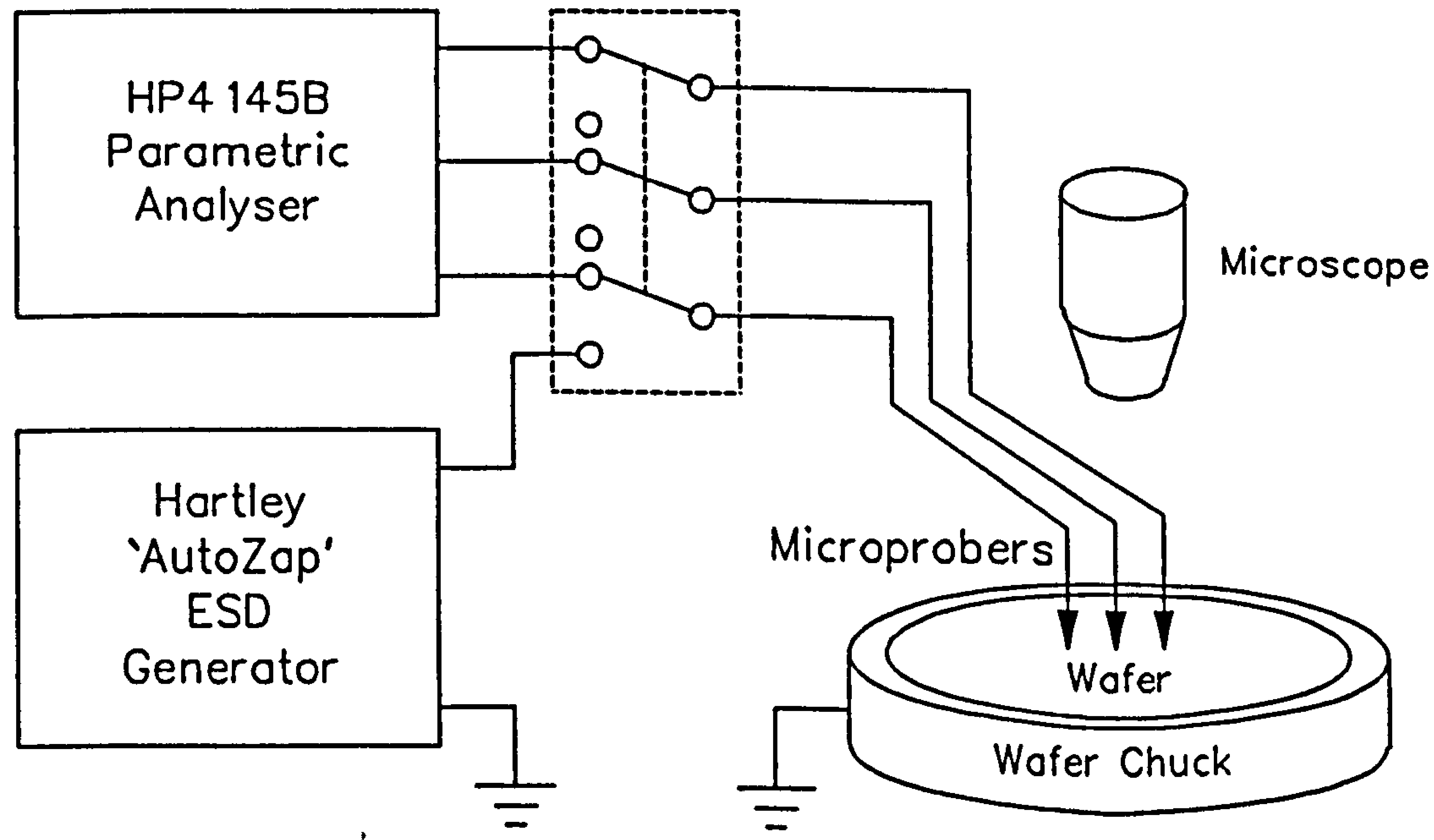
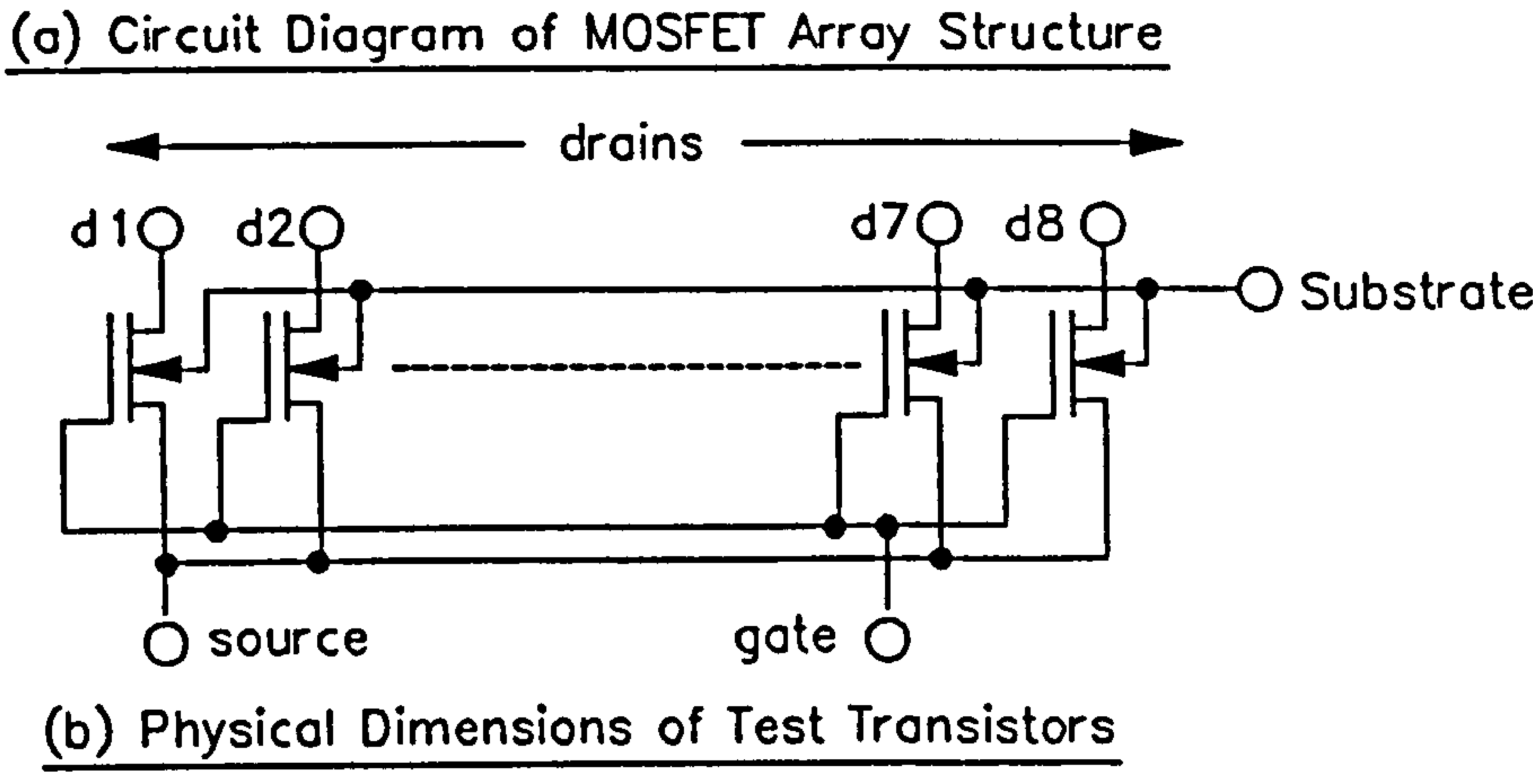


Figure 8.2: Block diagram of the experimental apparatus. The device may be switched between the 'AutoZap' ESD source and the HP4145B parametric analyser.

8.2 Apparatus

Fig.8.2 shows the apparatus used in these studies. The devices were tested at wafer-level using the chuck/microprober system described in Chapter 4. The Hartley 'AutoZap' (see Section 4.2.3.1) supplied standard HBM pulses, which were applied to the device under test (DUT) between the gate and the wafer substrate contacts. The effect of ESD upon the DUT's d.c. characteristics was observed using the Hewlett Packard HP4145B parametric anayser. Room temperature (approx. 25°C) was maintained throughout the experiments and the DUTs were illuminated by the microscope illumination system during testing.



(b) Physical Dimensions of Test Transistors

Type	d1	d2	d3	d4	d5	d6	d7	d8
NMOS(E)	W=25 L=3.5	W=32 L=3.5	W=35 L=3.5	W=42 L=3.5	W=46 L=3.5	W=63 L=3.5	W=40.8 L=3.5	W=102.9 L=3.5
NMOS(D)	W=25 L=3.5	W=32 L=3.5	W=35 L=3.5	W=42 L=3.5	W=46 L=3.5	W=63 L=3.5	W=40.8 L=3.5	W=102.9 L=3.5

(W = channel width, L = channel length, all dimensions in micrometers)

Figure 8.3: (a) Interconnection of devices in array structure, (b) Device dimensions.

8.3 Test Samples

The bulk-silicon NMOS wafers are discussed in Chapter 4. The transistors, which were both D and E-Mode, were fabricated in arrays of eight (denoted d1-d8) with common gate and source terminals and individual drains. Since the ESD pulse charge (C_1V_p) may not be evenly distributed between the device gates, each array was analysed as a single unit rather than a set of independent structures. Although this adds an unfortunate complication, it simulates an actual i.c. input in which many gate structures are connected to a single pin. Fig.8.3 shows a circuit diagram of the interconnections, together with the dimensions of the various structures.

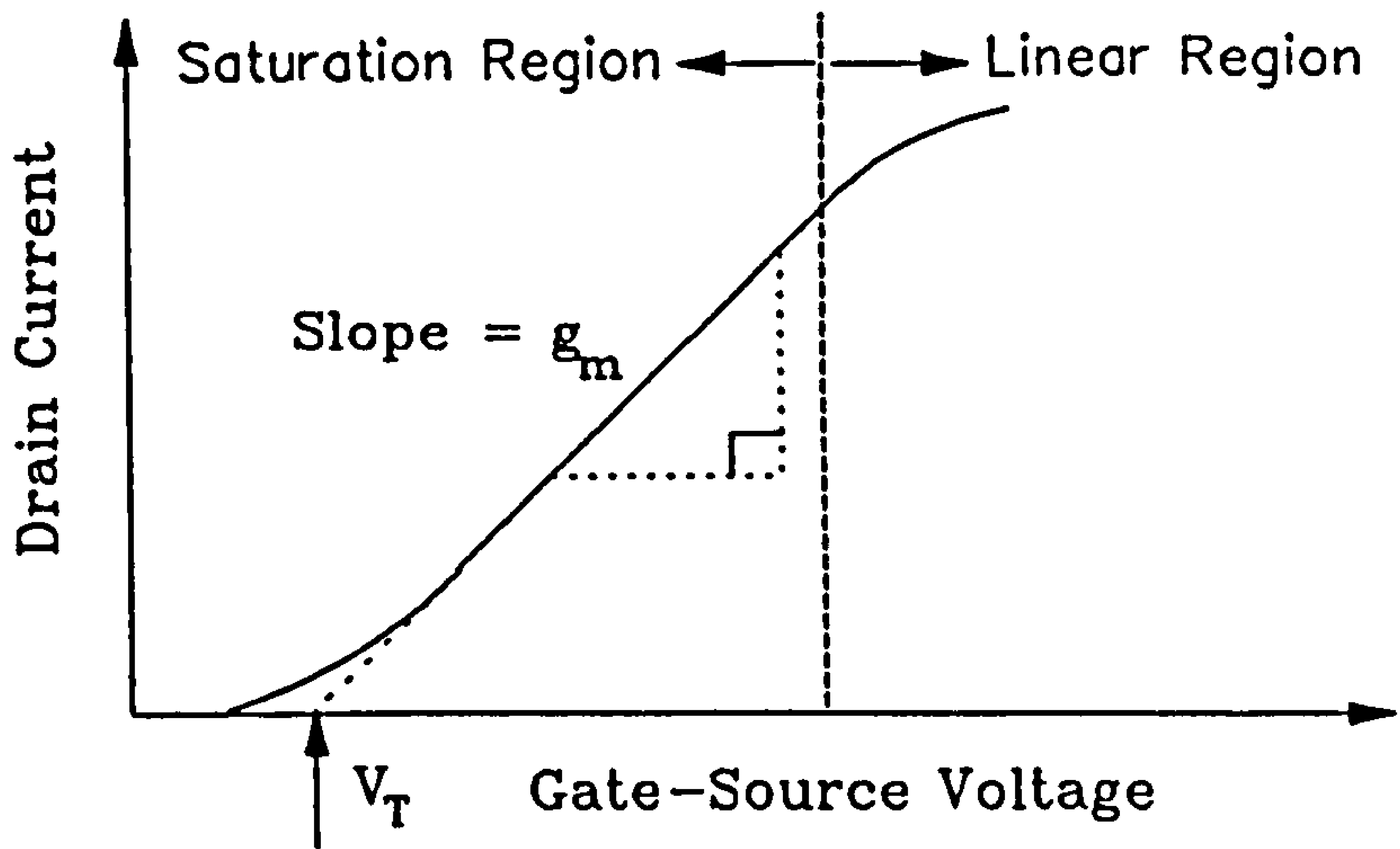


Figure 8.4: Extraction of V_T and g_m parameters from MOSFET transfer characteristic.

8.4 Sub-Breakdown Latency and Parametric Drift

8.4.1 Experiment

The initial phase of this study concerned sub-breakdown ESD damage, i.e. latant damage produced by ESD pulses *below* the oxide breakdown voltage threshold V_{bd}. Parametric drift was characterised in terms of the transconductance g_m and the strong-inversion threshold voltage V_T. These parameters were defined as the gradient and intercept of the I_d vs. V_{gs} curve in the saturation region (Fig.8.4). V_{ds} was held at 7V during this experiment.

The experiments were performed using NMOS E-mode transistor arrays. Each device in each array was characterised in terms of its initial transconductance $g_m(0)$ and initial threshold voltage $V_T(0)$. Negative polarity ESD pulse sequences were applied to the gate structures and g_m and V_T were measured for each device after each pulse. The process was continued until oxide breakdown was observed in terms of a finite gate current I_{gs}.

The data was then processed in the following manner: The values of g_m and V_T for each device after each pulse were expressed as a percentages of $g_m(0)$ and $V_T(0)$, yielding the percentage values $g_m(\%)$ and $V_T(\%)$. The average values of $g_m(\%)$ and $V_T(\%)$ for all the devices in the array was plotted against the number of pulses n.

Fig.8.5 shows the resulting graph for -48.5V ESD pulses. While g_m remains approximately constant prior the breakdown, it suddenly drops after breakdown (in agreement with earlier studies [6-8]). The average value of V_T decreases during the pulse sequence, suggesting the occurrence of positive charge trapping. Although this agrees qualitatively with Holmes' results [5], the variation does not exceed 10% and the latent damage may remain undetected until breakdown.

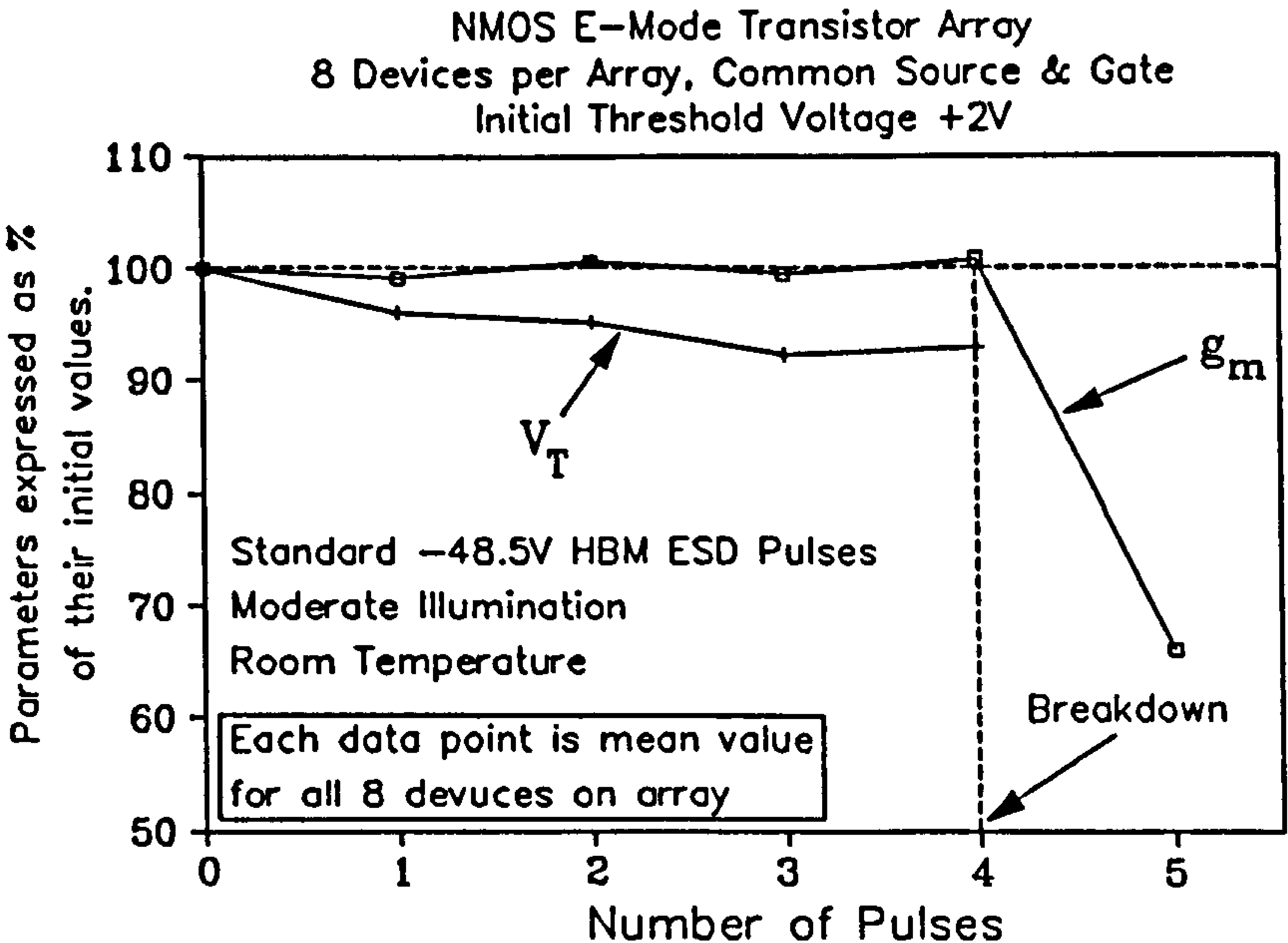


Figure 8.5: Sub-breakdown parametric drift in NMOS enhancement transistor array subjected to -48.5V ESD pulse sequence.

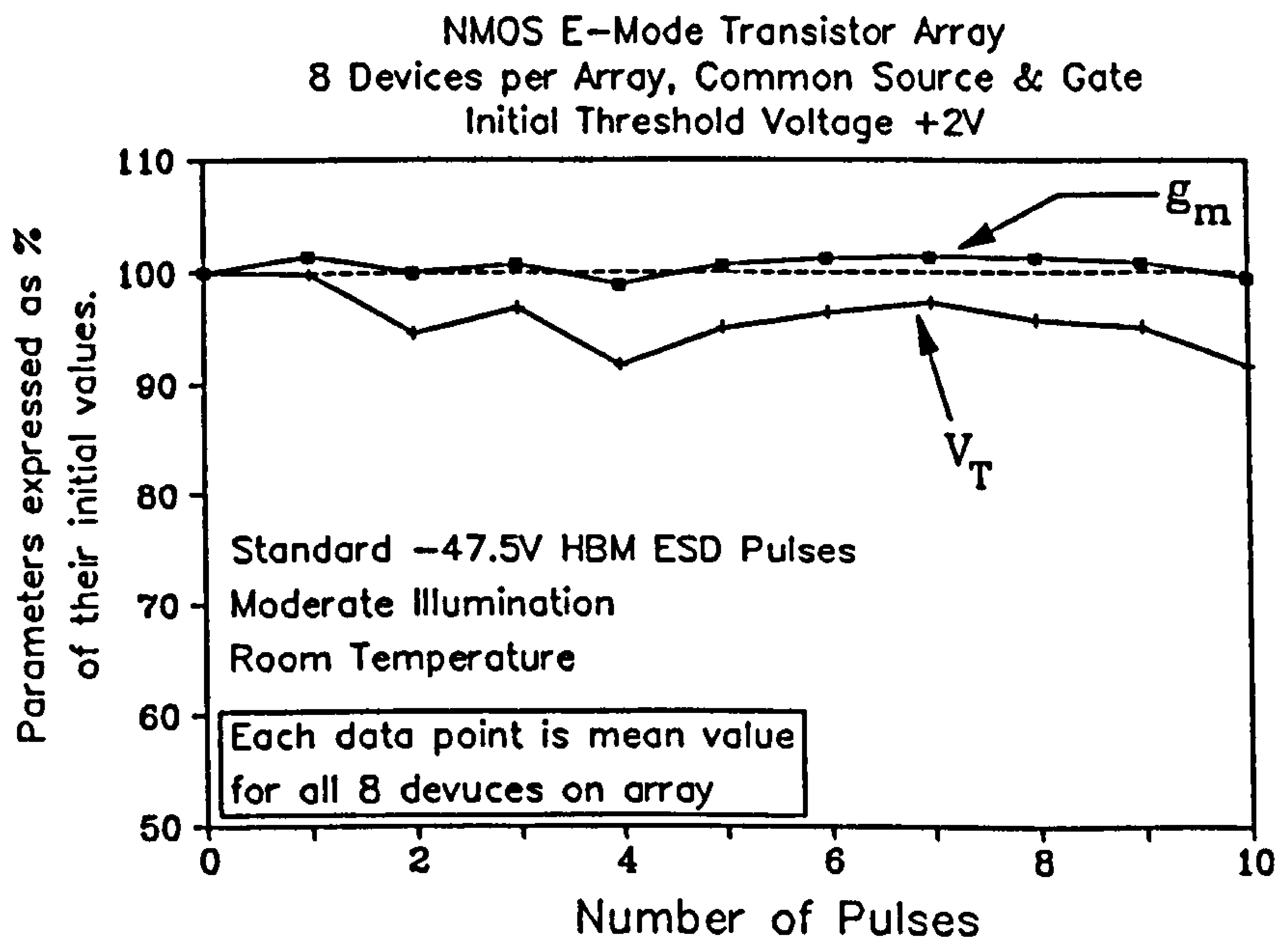


Figure 8.6: Sub-breakdown parametric drift in NMOS enhancement transistor array subjected to -37.5V ESD pulse sequence.

Fig.8.6 shows the results of an identical experiment performed using -47.5V ESD pulses. At this lower voltage, the latent damage per pulse is smaller than in the previous experiment and the array withstands a far greater number of pulses prior to breakdown (a further 20 pulses were required to cause failure). As with the -48.5V experiment, g_m remains approximately constant while V_T decreases. However, the decrease in V_T is not monotonic and exhibits noticable fluctuations, suggesting that positive and negative charge trapping may

predominate alternately throughout the pulse sequence. Again the shift does not exceed 10% and is too slight to cause functional errors.

The voltage dependence of latent failure was further examined by repeatedly pulsing NMOS array structures and noting the number of pulses n required to cause breakdown. Fig.8.7 shows n plotted as a function of the ESD pulse magnitude V_p, each data point representing the average value obtained from five structures. (These same data have already been presented in Fig.6.7.) The voltage 'window' ΔV_{lat} associated with significant latent failure can be conveniently defined as the voltage range between n=10 and n=1, which is clearly in the region of 2V.

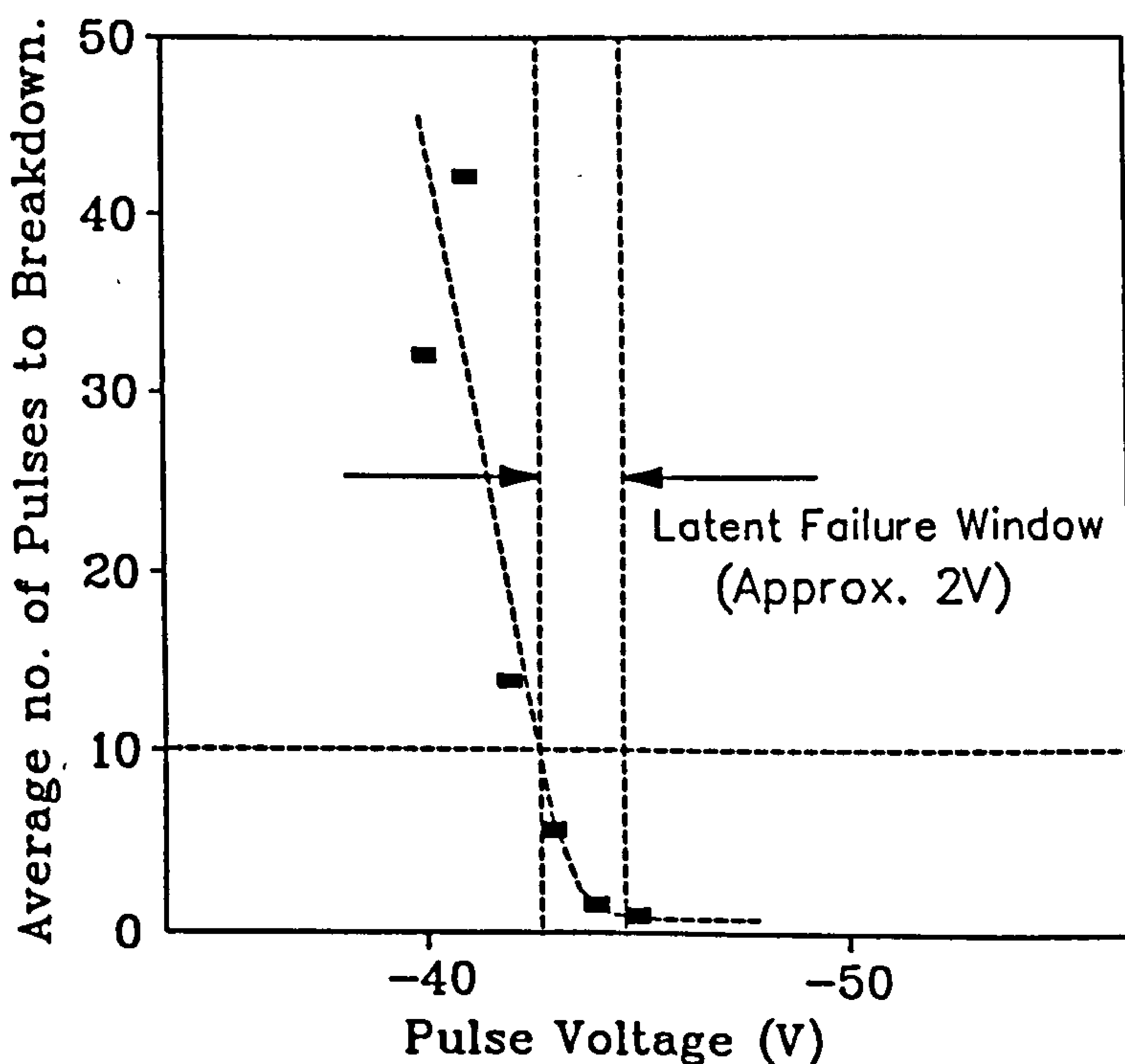


Figure 8.7: Pulses-to-breakdown n vs. pulse voltage V_p in NMOS E-Mode transistor arrays. Latent window ΔV_{lat} is defined as voltage range between n=10 and n=1.

8.4.2 Theoretical Analysis

The value of ΔV_{lat} can also be computed using oxide wearout theory of Chapter 7. According to the model, the damage fraction S_t produced by a pulse of magnitude V_0 is given by

$$S_t(V_0) = \frac{C\,T_{ox}}{\tau_0 A k H} e^{-\frac{H T_{ox}}{L V_0}} \qquad 8(1)$$

where C is the total capacitance parallel to the oxide, A is the oxide area (i.e. $C_x+C_l+C_t+C_x$), T_{ox} is the oxide thickness, L is the pulse attenuation factor due to capacitive

loading (given by $[C_1+C_x]/C$) and τ_0, k and H are constants defined in Chapter 7. S_t is defined such that $S_t=0$ and $S_t=1$ represent virgin and failed devices respectively. S_t is therefore approximately equal to 1/n, and the latent failure window can be re-defined as

$$\Delta V_{lat} = V_0|_{S_t=1} - V_0|_{S_t=0.1} \qquad 8(2)$$

Combining Eqns.8(1) and 8(2), together with the approximation $e^{-2}\approx 0.1$ yields

$$\Delta V_{lat} = V_{bd}\left(1 - \frac{H\,T_{ox}}{H\,T_{ox} + 2\,L\,V_{bd}}\right) \qquad 8(3)$$

where V_{bd} is the breakdown voltage magnitude (i.e. the value of V_p for $S_t=1$). Since $C_1+C_x=106.607$pF (see Chapter 4), $T_{ox}=40$nm, H=1335.8MV/cm (see Chapter 7), C=117.23pF (independently measured) and $V_{bd}=45$V (Fig.8.7), ΔV_{lat} is approximately 0.68V. Since the measured value was 2V (see Fig.8.7), the prediction is of the same order of magnitude as the experimental result.

8.5. Latent Damage and Parametric Drift in Damaged-Oxide Devices

8.5.1 Experimental Procedure

Results are reported for 64 NMOS transistors, of which 32 were E-mode and 32 were D-Mode. The devices, all of which conformed to their nominal specifications, had received no prior stress. The procedure was as follows: The devices were initially characterised in terms of their I_d vs. V_{ds} characteristics and were then subjected to single ESD pulses of −200V, -100V, +100V, and +200V. One specemin of each array type was stressed at each voltage. The I_d vs. V_{ds} curves were then re-measured and compared with the original characteristics.

Since many of the devices were in a delicate walking-wounded condition immediately after stress, a characterisation method was required which subjected them to minimum measurement stress. The magnitudes of V_{ds} and V_{gs} were therefore limited to the range −1V to 1V during characterisation. Fig.8.8 shows typical characteristics of undamaged devices.

The characteristics of some of the degraded devices were then re-measured using V_{ds} and V_{gs} ranges of 0-10V, in order to examine the effects of working-voltage stress. Since V_{gs} is common to all the devices in an array, only one device per array was subjected to such measurement.

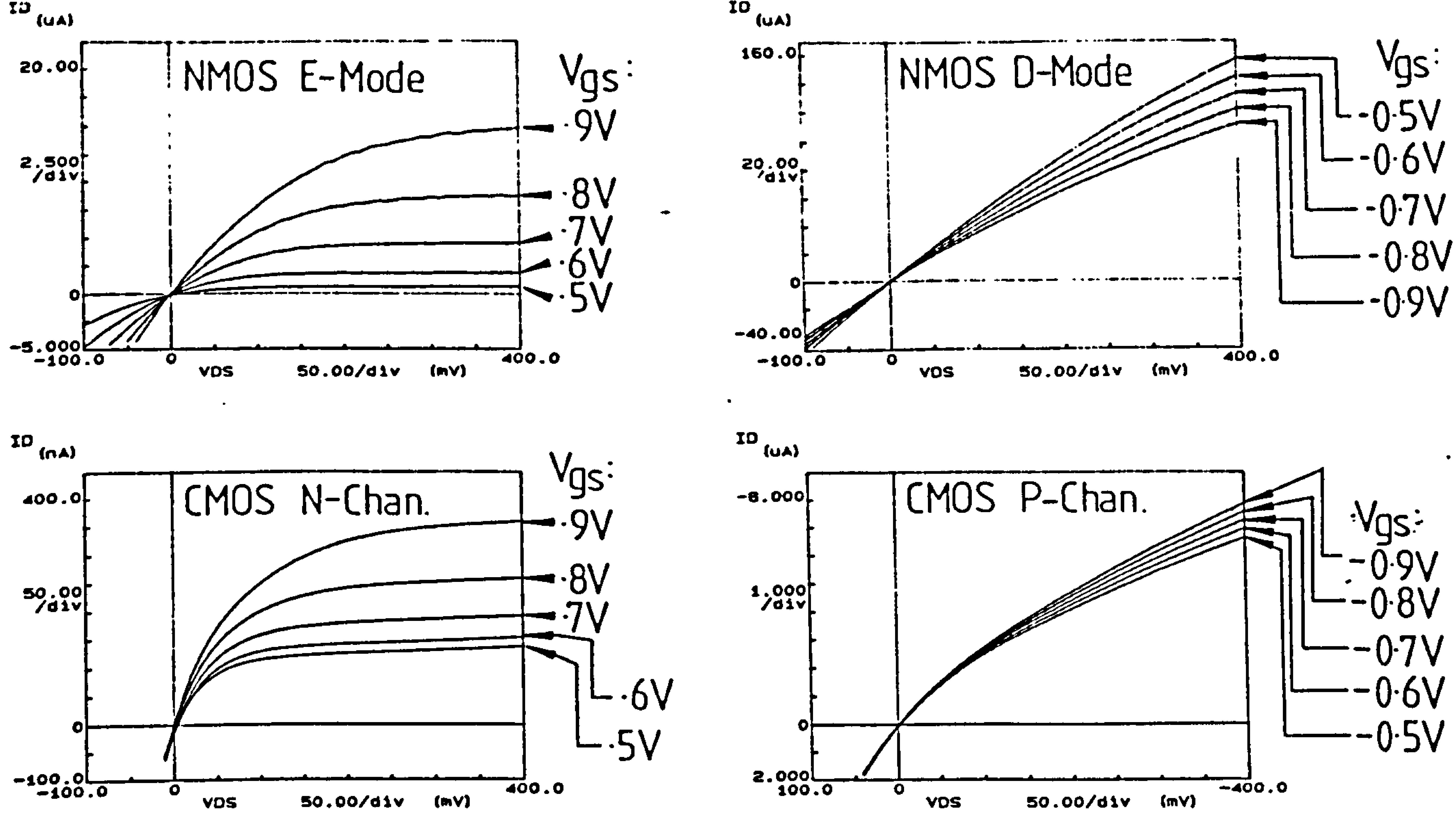


Figure 8.8: Typical I_d vs. V_{ds} characteristics of virgin NMOS transistor structures.

8.5.2 Results

The post-ESD characteristics were classified into the following categories:

Group 1.	Unchanged characteristics (cf.Fig.8.1[a]).
Group 2.	Characteristics with negatively-shifted V_T.
Group 3.	Characteristics with reduced g_m, converging at origin (cf.Fig.8.1[b]).
Group 4.	Characteristics with reduced g_m *not* converging at origin (cf.Fig.8.1[b]).
Group 5.	Linear resistive characteristics (cf.Fig.8.1[c]).
Group 6.	Characteristics with zero g_m (cf. Fig.8.1[d]).

Typical characteristics from each category are shown in Fig.8.9. Groups 2-4 clearly belong to the 'walking-wounded' category, while Groups 5 and 6 are catastrophic failures. The threshold-voltage variation in Group 2 was sometimes of the order of several volts, causing E-mode devices to become functionally D-mode. The transconductance reduction in Groups 3 and 4 varied between about 5% and 95% (anything beyond a 95% reduction was classed as a Group 6 failure). Fig.8.10 shows which devices fell into each category.

Many of the devices whose characteristics were re-measured using $0<V_{ds}<10V$ retained their stable walking-wounded characteristics. Others, however, underwent further

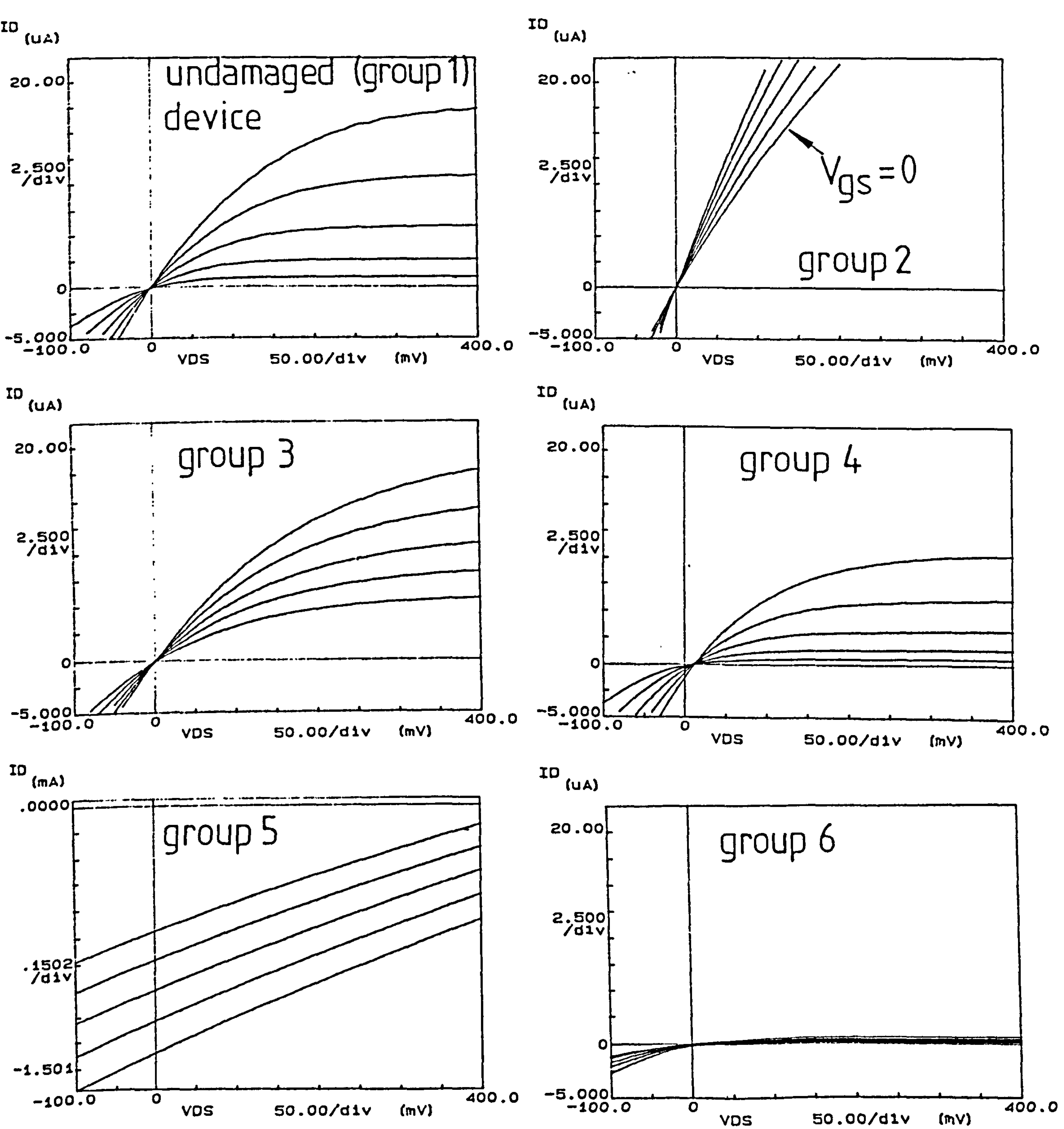


Figure 8.9: Post-breakdown characteristics of NMOS E-Mode transistors.

NMOS D-Mode Transistors

Pulse Voltage	d1	d2	d3	d4	d5	d6	d7	d8
-200V	1	1	1	1	1	2	2	2
-100V	3	3	1	1	1	1	1	1
+100V	2	2	2	2	2	5	2	2
+200V	2	2	2	2	2	2	2	5

NMOS E-Mode Transistors

Pulse Voltage	d1	d2	d3	d4	d5	d6	d7	d8
-200V	6	6	6	6	6	6	4	4
-100V	6	6	6	3	3	3	3	1
+100V	3	5	2	3	3	3	3	5
+200V	2	2	2	2	2	2	5	2

Figure 8.10: Distribution of failure categories 1 to 5 in NMOS arrays.

degradation during characterisation. Fig.8.11 shows the characteristics of two walking-wounded devices (originally of Groups 2 and 4) after they had received three characterisations. These devices are clearly unstable under working voltage stress and therefore constitute a latent hazard.

Further experiments were performed in order to observe this degradation phenomenon in action. Fig.8.12 shows the transfer characteristics (I_d vs. V_{gs}) for an NMOS E-Mode device during the degradation cycle. Characteristics A and B were measured immediately before and after the application of a -80V ESD pulse, while characteristics C and D show the results of subsequent characterisations. Since characteristics B and C possess positive transconductances ($\partial I_d/\partial V_{ds}$), they must be classified as walking-wounded states. However, characteristic D is clearly a catastrophic failure of Group 5 (see Fig.8.9). Degradation is displayed by the spasmodic jumps of characteristics B and C, which cause the device to drift towards catastrophic failure under working-voltage stress.

In a limited number of cases, repeated characteristic re-measurement caused an eventual return to transistor action (positive g_m), introducing the possibility of no-fault-found (NFF) behaviour. In this recovered state, the transistors were extremely robust and could often withstand ESD pulses up to several hundred volts. However, these cases were rare and most catastrophically failed devices retained in their fault characteristics indefinitely.

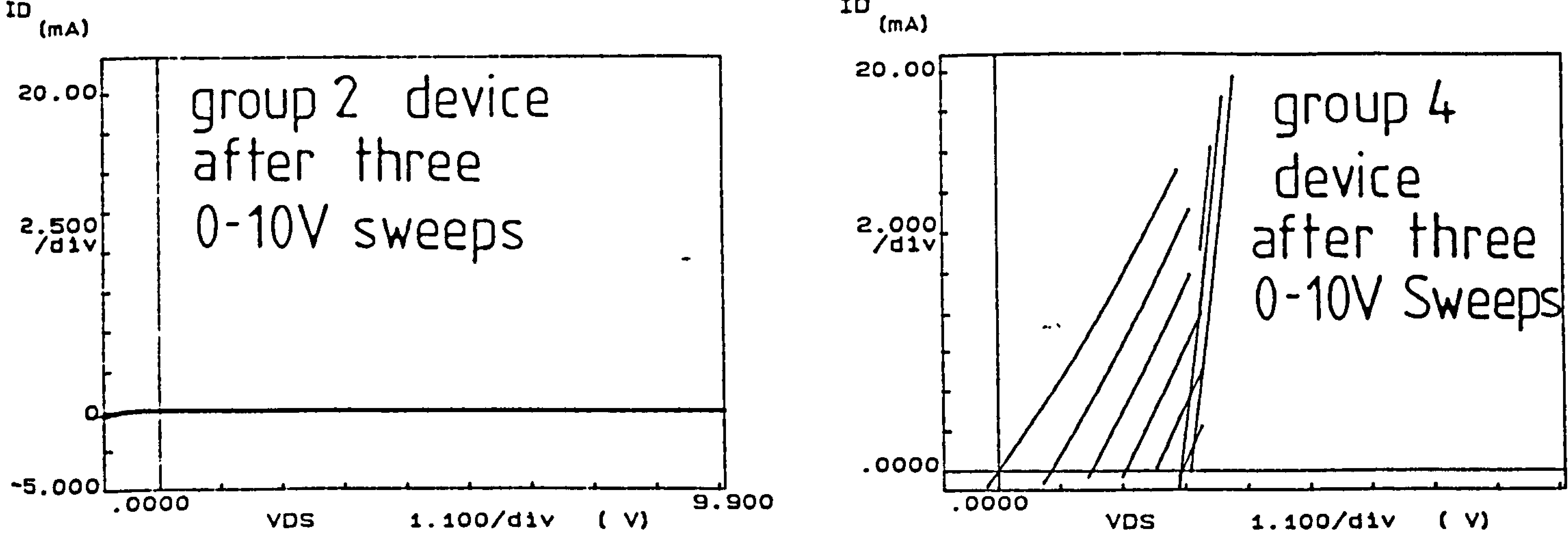


Figure 8.11: Characteristics of two walking-wounded devices after three 0-10V characterisation sweeps.

A
B
C
D
11.00
I_d
1.100 /div
.0000
.0000
2.000/div (V) V_{gs}

Figure 8.12: I_g vs. V_{gs} characteristics during degradation.

8.5.3 Discussion and Modelling

Dielectric breakdown in MOS oxides is believed to be accompanied by the intrusion of one or more filaments of polysilicon gate material into the SiO_2 [8] (see also Section 6.2.2). These filaments can form conductive paths between the source, drain and gate terminals, or between the gate electrode and the channel. The geometry and position of the filaments dictate the nature of the failure characteristics [9].

8.5.3.1 Qualitative Analysis of Failure Characteristics

Group 2 characteristics are clearly extreme cases of the negative V_T drift encountered in Section 3. In these cases the positive-charge trapping *was* sufficient to drive E-mode devices into D-mode (or D-mode devices even further into depletion).Group 3 failures probably result from filaments connecting the gate to the source terminal. Since the filament resistance R_f forms a potential divider with the parasitic gate resistance R_g, the effective gate-source voltage is reduced by a factor $R_f/(R_f+R_g)$, thereby reducing the effective transconductance by the same factor. The characteristic is uniformly scaled-down by the process and the characteristics continue to converge at the origin.

Group 4 failures are most probably caused by gate/drain or gate/channel filaments. When V_{gs} is high and V_{ds} is low, current flows in through the gate and out through the source, creating a negative I_d for positive V_{ds}. Hence the characteristics fail to converge at the origin. The g_m reduction is caused by inversion-layer charge being sucked out of the channel region into the gate via the damaged oxide.

Group 5 represents an extreme case of the gate-drain breakdown, when the current in the filament resistance R_f dominates the transistor drain current. Similarly Group 6 represents an extreme case of gate-channel breakdown, where practically all inversion charge is removed via R_f.

The spasmodic nature of characteristic shifts in Fig.8.12 suggests thermal expansion of defects due to joule heating. As the power dissipation in a defect increases during a voltage sweep, the temperature increases quazi-statically until a critical temperature is reached. At this point the defect expands rapidly, reducing R_f, until a new equilibrium is reached. Since this transition is too fast for the parametric analyzer to measure, the characteristics appear to consist of discrete curves, punctuated by discontinuities.

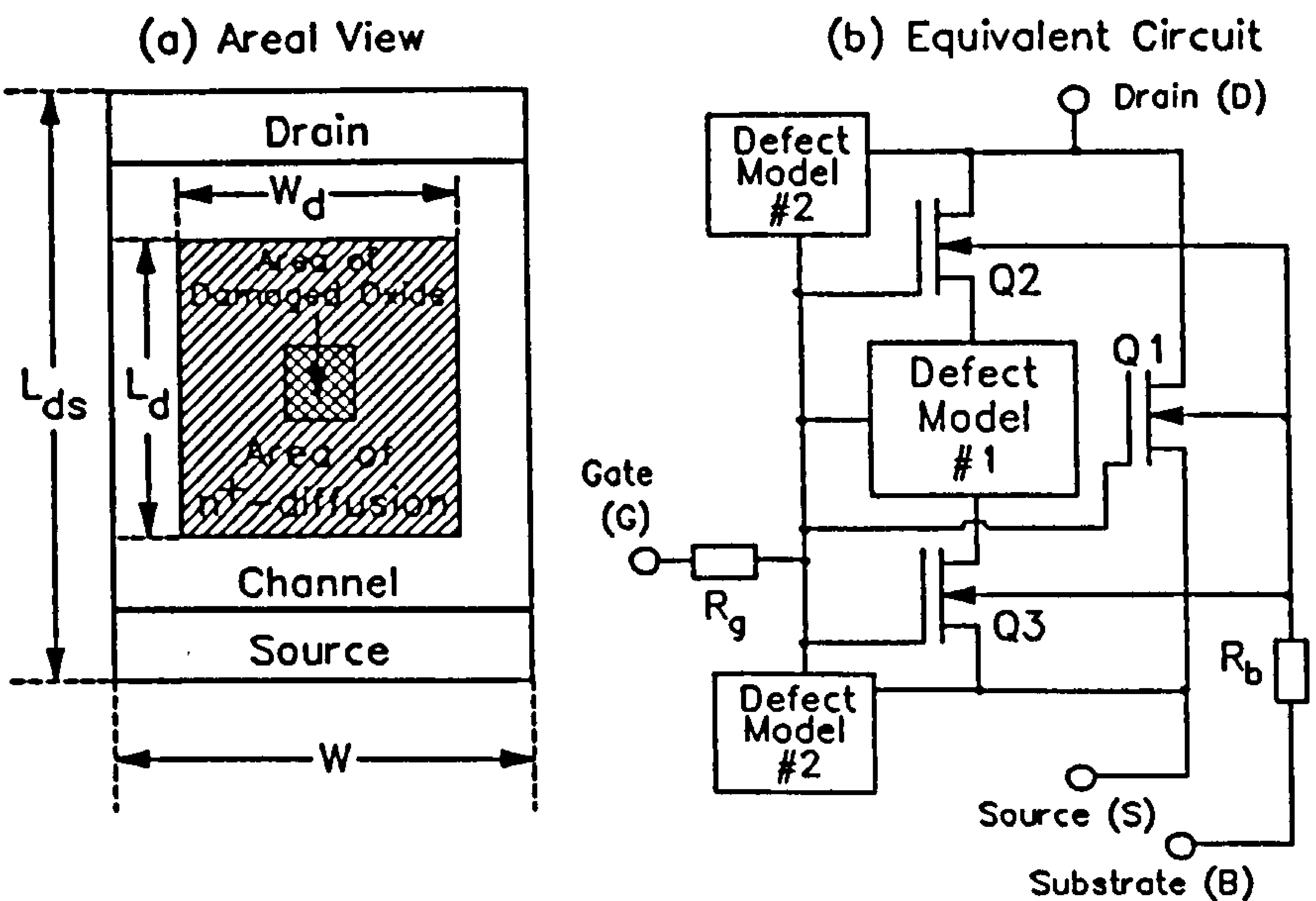


Figure 8.13: Schematic areal view and equivalent circuit for damaged MOSFET.

8.5.3.2 PSpice Model of Failed Device

A lumped-element model of a failed transistor was developed using the MicroSim *PSpice* software. The methodology (which was based upon that of Syrzycki [9]) is illustrated in Fig.8.13. Defect models #2 represents the gate/source and gate/drain filament structures described above. Since gate-channel breakdown (defect model #1) is accompanied by solid-state diffusion of n^+ doping into the channel region [9], a rectangular equipotential of dimensions L_d x W_d, was placed within the channel, dividing the structure into three different transistors. The PSpice LEVEL 1 MOSFET model (i.e. first-order Shichman-Hodges model, see Section 2.3.2) was selected, together with V_T=0. (Although it assumes constant mobility, it easily applicable to the characteristics of Fig.8.9, in which the fields are insufficient to cause mobility modulation.) Fig.8.14 show failure characteristics simulated using this model, together with the associated parameter values. (Note: the parameters R_f and R_g were chosen arbitarily for these simulations and are not corroborated by any independent measurements.)

8.5.3.3 Effect of Degradation upon CMOS Operation

Since the impact of device degradation upon logic operation is a complex subject, the present discussion is initially limited to the simple CMOS invertor (or NOT) gate. The left portion of Fig.8.15 shows a graphical representation of the operation of such a circuit. The two transistor characteristics are superimposed and the operating points for

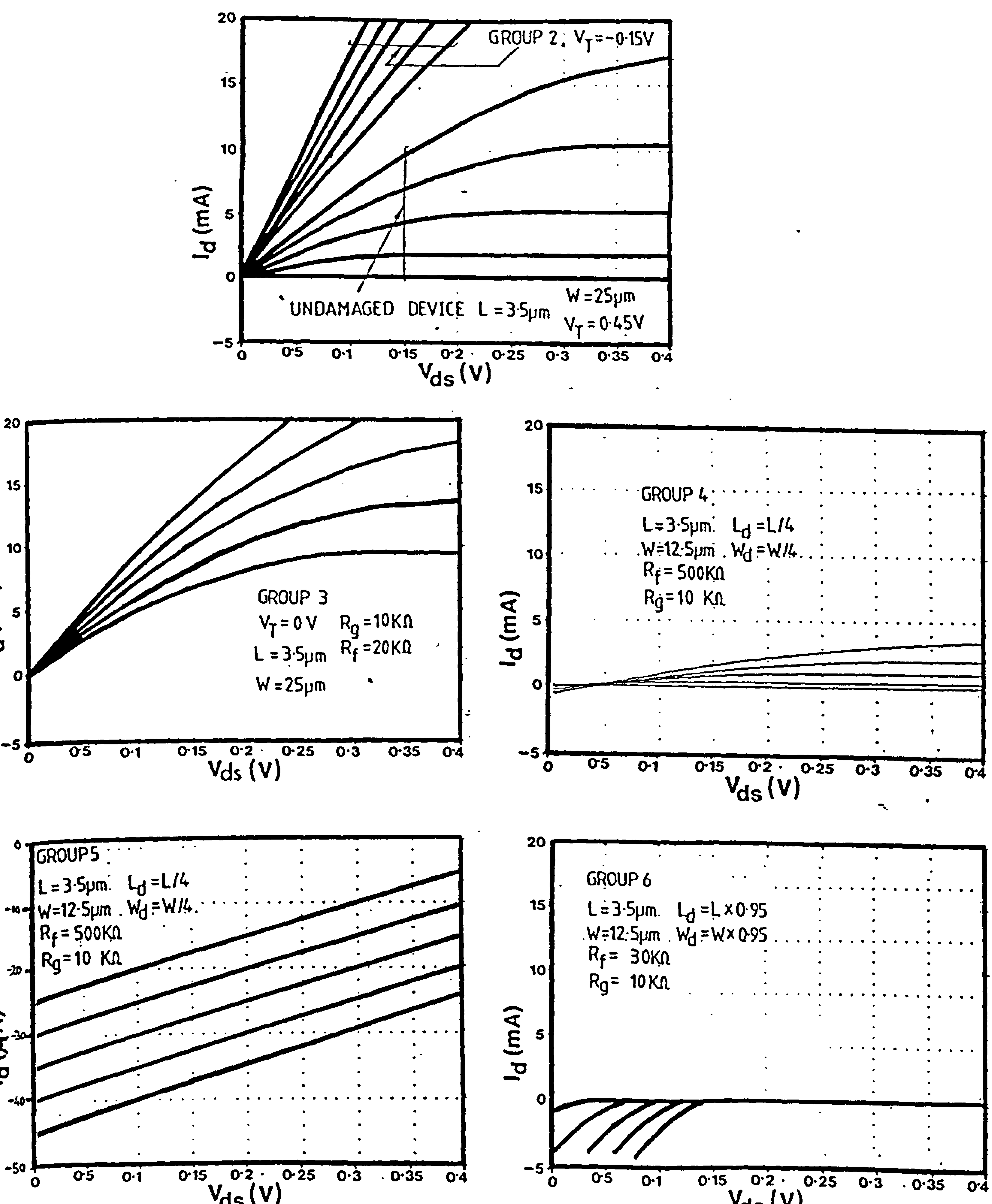


Figure 8.14: Failure characteristics predicted using PSpice model.

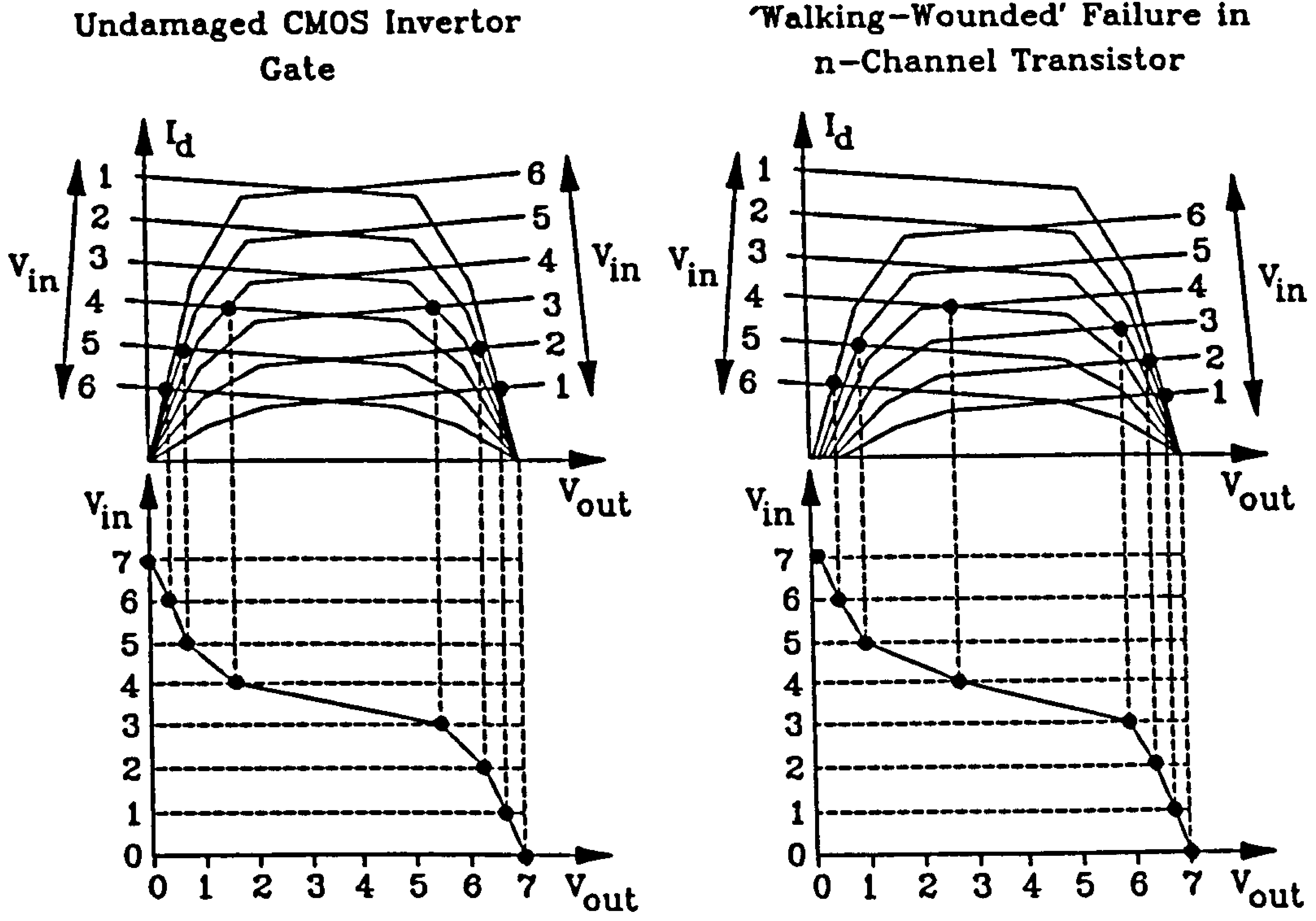


Figure 8.15: Schematic representation of the transfer characteristics of a healthy CMOS invertor (left) and CMOS invertor with 'walking-wounded' n-channel transistor.

a selection of input voltages are transferred to an input vs. output graph. The output clearly exhibits a logical 1 while the input is held at 0 and vice versa.

The right portion of Fig.8.15 shows the impact of a degraded (group 4) n-channel transistor on the gate operation. Although the output voltage is positively shifted, the logical inversion function is maintained. This illustrates how walking-wounded failure is not necessarily detectable in terms of the circuit logic performance [8]. However, continued g_m degradation causes the curve to shift steadily to the right until the circuit enters a 'stuck-at-1' condition. Fig.8.16 illustrates the effects of linear catastrophic failure (Group 5) on a CMOS invertor. This type of failure clearly introduces a voltage-follower behaviour, irrespective of which device (n-channel or p-channel) has failed.

The above findings help to illustrate the limits of digital fault analysis and location techniques. Several algorithms are avaliable, including the 'M/y-box' technique of Chen, Lee and Shen [10]. The M/y box is a conceptual logic element with three terminals, denoted P, N and Y (see Fig.8.17). It is connected to the output of a CMOS device with P connected to the p-channel and N to the n-channel gate blocks. The logic states of P and N (logic 1 or 0) indicate open and closed circuit paths to the supply rails V_{dd} and V_{ss}. If P and N are both 1 or both 0 then the circuit is healthy and the output is 0 or 1 respectively. The n and p-channel transistor blocks can therefore be represented by gate-level circuits (according to rules laid down by Chen et al. [10]) and Fig.8.18 shows the resulting representation of a

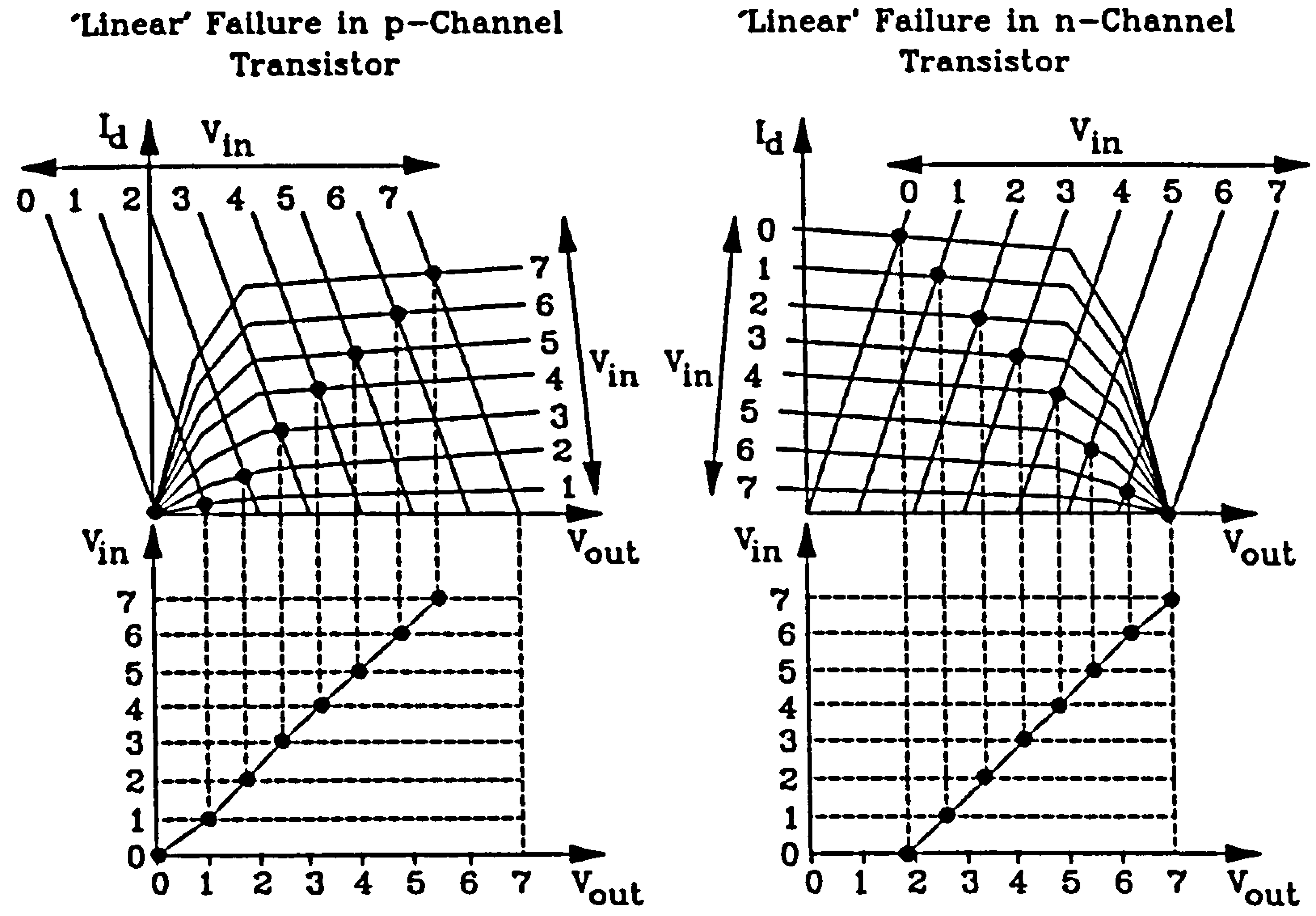


Figure 8.16: Schematic representation of the transfer characteristics of CMOS invertor gate with linear failure in p-channel transistor (left) and CMOS invertor gate with linear failure in n-channel transistor (right).

CMOS NAND gate.

Faulty logic operation can be represented as follows: If P=0 and N=1 then the gate output is simultaneously connected to V_{dd} and V_{ss} producing an output of 1 or 0 depending upon which supply rail predominates. This condition is expressed by an M/y box output Y=y. Conversely, if P=1 and N=0 then the output floats, recalling its previous logic state. This is expressed by an M/y box output Y=M.

Consider a CMOS NAND gate with a faulty n-channel transistor. Since a group 2 fault creates a short circuit, it can be represented by setting one of the inputs of the lower AND gate in Fig.8.18(c) to 0. Similarly a Group 6 (open circuit) failure can be simulated by setting the same gate input equal to 1. However, a Group 5 (linear resistive) failure presents greater problems, since it implies a direct connection between the transistor gate and drain. Since the M/y box algorithm only caters for short- and open-circuits between drain and source terminals, the Group 5 failure lies beyond its scope. The improvement of logic-fault analysis techniques is therefore an open area for new research.

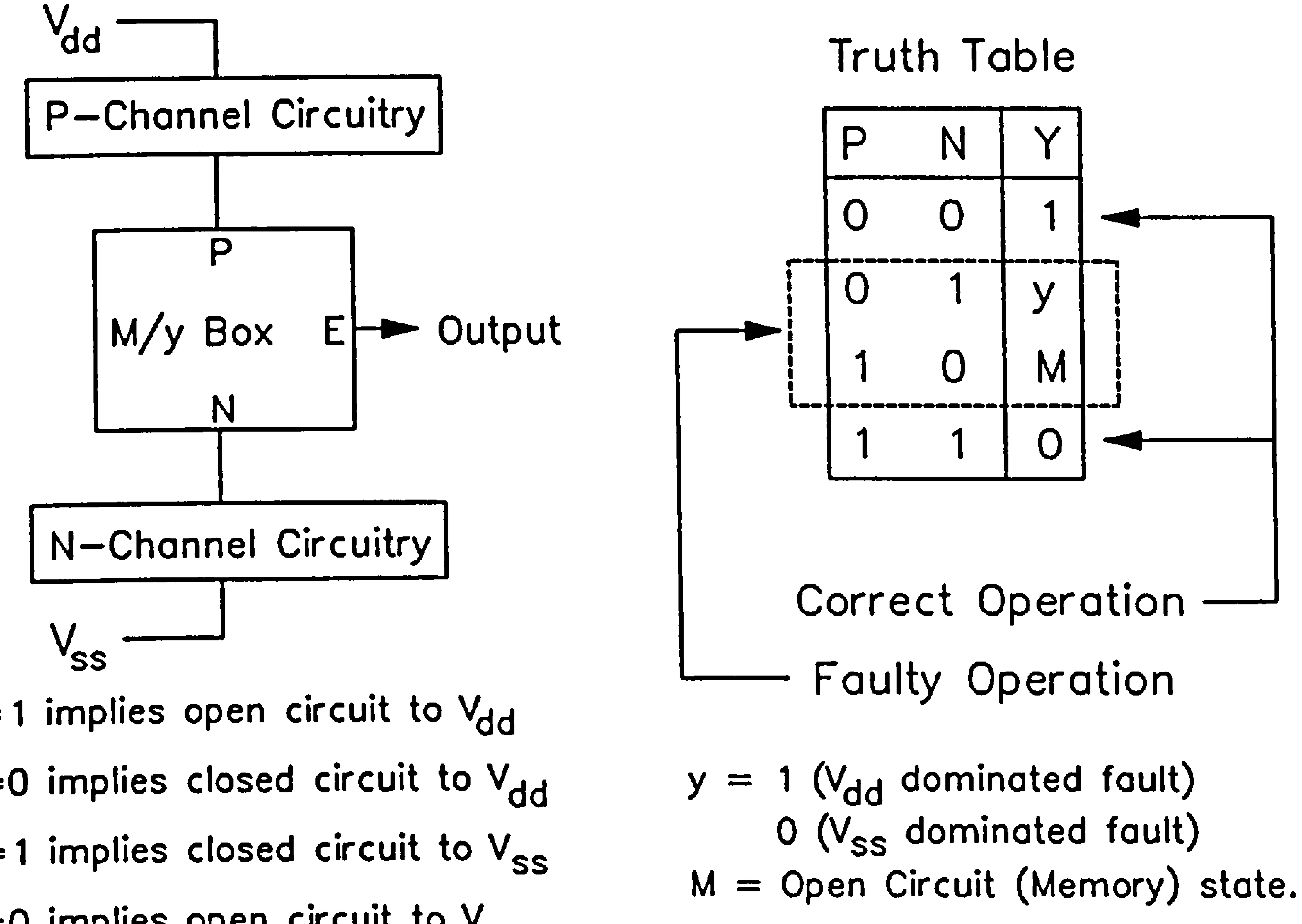


Figure 8.17: M/y box logic element: Circuit diagram and truth table.

8.6. Summary and Conclusions

The relationships between latent damage and d.c. parametric drift in ESD damaged MOSFETs have been studied both experimentally and theoretically. The following conclusions can be drawn from the results:

1. Sub-breakdown ESD stress causes a negative drift in the strong-inversion threshold voltage. This phenomonen has been observed by many earlier workers [eg.11] (although Greason [12] recently observed a positive V_T shift in negatively stressed n-substrate MOS structures). However, since the V_T shift did not exceed 10%, the latent oxide damage was not readily apparent in terms of the device's d.c. characteristics.

2. Although sub-breakdown latency is difficult to detect, the extreme narrowness of the latent failure window (determined by theory and experiment) suggests that it is

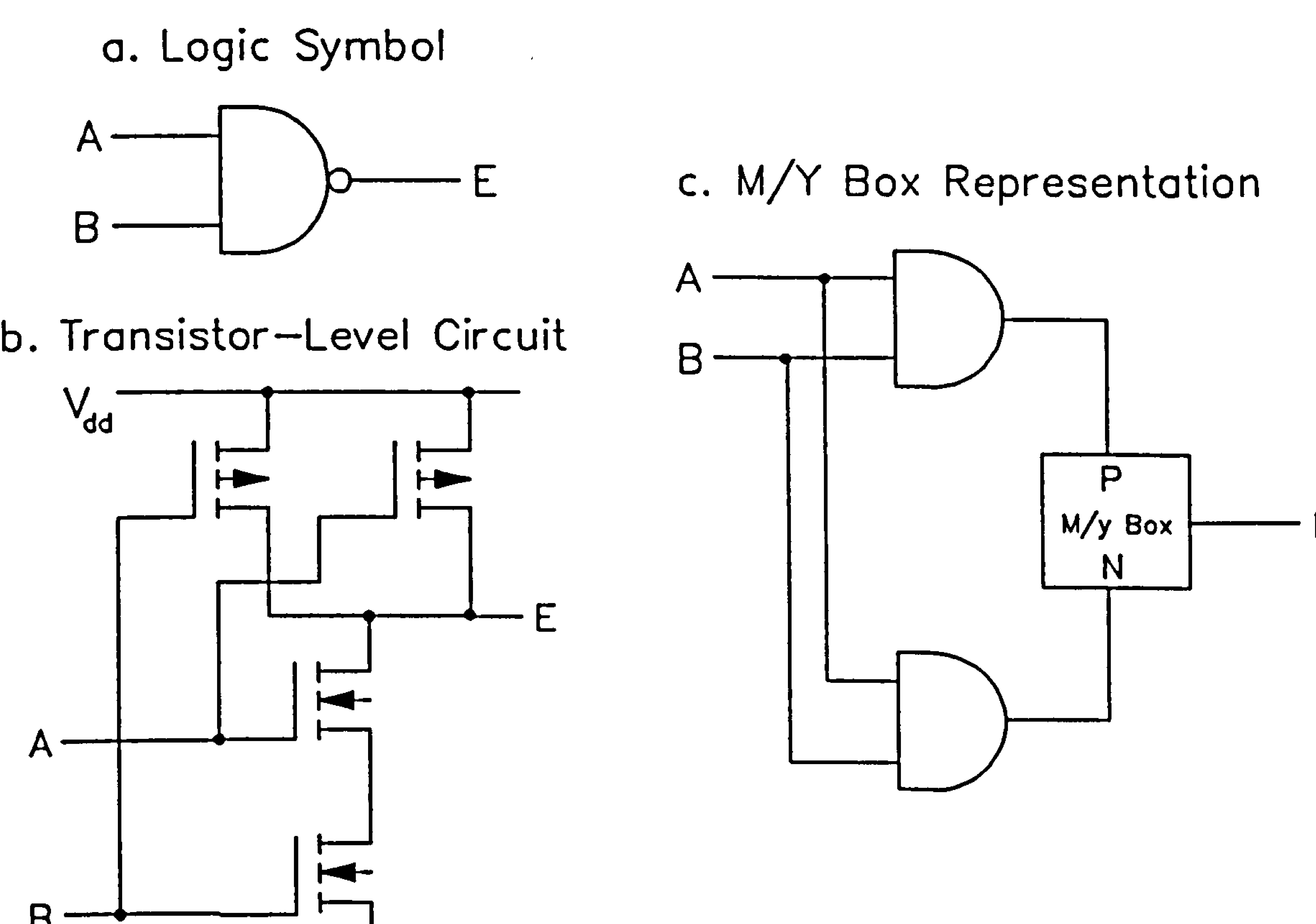


Figure 8.18: Application of M/y box in CMOS NAND gate.

unlikely to be a major reliability hazard. Devices subjected to random-magnitude ESD are most likely to receive pulses greater than V_{bd}, causing immediate failure, or below V_{bd}-ΔV_{lat}, leaving them virtually undamaged.

3. MOS devices with damaged oxides were found to exhibit either walking-wounded or catastrophic-failure characteristics. The former were shown to degrade into the latter under working-voltage stress, providing a latent failure mechanism. The 'jumpy' nature of the characteristic degradation suggested spasmodic defect expansion under Joule heating.

4. The Syrzycki methodology [9] was employed in order to model the transistors at their various stages of degradation. The effect of degradation on CMOS operation was also examined.

5. The results of this chapter were used to illustrate some limitations of the M/y-box technique for digital fault analysis [10].

However, the reported experimental data is not sufficient to provide a complete picture of the degradation processes. This work merely demonstrates that damaged devices *do* degrade under working-voltage stress and suggests some explanations of this phenomenon.

8.7. References

1. McAteer,O.J., Twist,O.J., Walker,R.C., *"Identification of Latent ESD Failures"*, Proc. 2nd. EOS/ESD Symp., pp.55-63, 1980.
2. McAteer,O.J., Twist,O.J., Walker,R.C., *"ESD Latent Failures"*, Proc. 4th. EOS/ESD Symp., pp.41-8, 1982.
3. Crockett,R.G.M., Smith,J.C., Hughes,J.F., *"ESD Sensitivity and Latency Effects of some HCMOS Integrated Circuits"*, Proc. 6th. EOS/ESD Symp., pp.196-201, 1984.
4. Aur,S, Chatterjee,A, Polgreen,T, *"Hot Electron Reliability and ESD Latent Damage"*, Proc. 26th. IEEE Reliability Physics Symp., pp.15-18, 1988.
5. Holmes,G.C., *"An Investigation into the Effects of Low Voltage ESD Transients on MOSFETs"*, Proc. 7th. EOS/ESD Symp., pp.170-4, 1985.
6. Amerasekera,E.A., Campbell,D.S., *"Electrostatic Pulse Breakdown in NMOS Devices"*, Qual. & Rel. Eng. Int., 2, pp.107-116, 1986.
7. Amerasekera,E.A., Campbell,D.S., *"An Investigation of the Nature and Mechanisms of ESD Damage in NMOS Devices"*, Sol. St. Elect., 32(3), pp.199-206, 1989.
8. Soden,J, Hawkins,C.F., *"Test Considerations for Gate Oxide Shorts in CMOS ICs"*, IEEE Design and Test of Computers, pp.56-64, August 1986.
9. Syrzycki,M, *"Modelling of Gate Oxide Shorts in MOS Transistors"*, IEEE Trans. Computer-Aided Design, 8(3), pp.193-202, 1989.
10. Chen,J.-E., Lee,C.L., Shen,W.-Z., *"My-box Representation for Faulty CMOS Circuits"*, Proc. IEE, 137G(3), pp.225-32, 1990.
11. Amerasekera,E.A., Campbell,D.S., *"ESD Pulse and Continuous Voltage Breakdown in MOS Capacitor Structures"*, Proc. 8th. EOS/ESD Symposium, Las Vegas, Nevada, Sept.23-5, 1986. pp.208-13, 1986.
12. Greason,W.D., Chum,K., *"The Integrity of Gate Oxide Latent Failures Under EOS/ESD Conditions"*, Proc. 14th. EOS/ESD Symp., pp.106-11, 1992.

Chapter 9

Discussion and Conclusions

9.1 Introduction

This final chapter summarises the investigations recorded earlier in this thesis, discusses their scientific significance, and comments upon their relevance to the general advancement of semiconductor device technology.

The chapter begins by analyzing the overall objectives of the research and then proceeds to evaluate the extent to which each area of investigation has contributed to the furthering of these aims. It ends by discussing possible areas for future research.

9.2 General Aims of This Study

Chapter 1 outlined of the history of microelectronics technology and showed how recent device miniaturization has led to a dramatic increase in the number of components per chip. Since a VLSI chip malfunction can be caused by the failure of any one of its 1,000,000 transistors, there is an increasing demand for high-quality, high-reliability devices. To complicate matters, the continuous reduction of device dimensions introduces novel failure mechanisms, the physics of which are required to be understood. Without such a fundamental understanding, VLSI reliability optimization becomes something of a 'black art', relying upon empirical design rules and feedback from lengthy and expensive quality assurance tests. However, a sound theoretical model of the relevant failure mechanisms should (in principle) allow reliability to be 'built in' at the design stage, thereby eliminating the need for extensive performance feedback and process adjustment.

The general aim of this thesis is to contribute towards the development of such a model. Since MOS is commonly used throughout present-day microelectronics, the research was limited to this technology. Research was also confined to *reliability* rather than *yield* problems, the latter term relating to *time zero* failures (i.e. devices which are catastrophically degraded as they leave the fabrication stage [1]). All devices showing such failure were excluded from the experiments.

9.3 Literature Survey Chapters

The thesis began by reviewing the current state-of-the-art in MOS reliability theory. Chapters 1 to 3 examined MOS device technology and its associated failure mechanisms. Gate oxide breakdown under electrostatic discharge (ESD) was finally identified as a fruitful area for the original research of the thesis. A better understanding of this mechanism may help optimise the design of protection-circuitry (whose ESD-blocking properties are not always perfect [2]), thereby assisting the development of a 'built-in' reliability strategy (see Section 9.2).

Chapters 1 and 2 reviewed the theory and practice of MOS technology. Techniques for the growth of thermal SiO_2 were described, together with its applications as a passivation layer and as a gate dielectric. The atomic and electronic structures were also briefly described and the nature of oxide defect sites was discussed.

The theory of the MOS transistor was then examined. Since the range of available MOS-related devices is so extensive, the treatment given was somewhat limited. Attempts were made to include all the generic MOS-related structures, with the exception of the E^2PROM memory devices, whose ultra-thin tunnelling oxides place them in a somewhat different category. The basic 'first order' theory of the MOSFET was presented, together with the 'second order' effects of miniaturisation upon this model.

Chapter 3 presented the results of an extensive literature survey on MOS failure mechanisms. This included both time-dependent and event-dependent failure due to intrinsic and extrinsic (i.e. defect-related) mechanisms. The sources of event-related stress were examined with a particular emphasis on ESD, which is known to account for a very large proportion of electronic failures. The physics of ESD and the various models developed to simulate it were examined. This study was augmented by a section on reliability improvement, examining such issues as static elimination, protection circuit design and the removal of extrinsic defects by screening.

Although latchup, spiking, electromigration and hot electron injection were all covered, most attention was given to gate oxide dielectric breakdown, which seems to the most important and least understood MOS failure mechanism. The main generic theories of dielectric breakdown were described and discussed. The intrinsic and extrinsic breakdown modes were described, and the latter was explained using the statistical *effective thickness* model.

The limited body of literature on short time-scale and ESD oxide breakdown physics was reviewed. However, the conflicting claims of the various workers, and the ambiguity of the experimental data led to the conclusion that future research was required. This area was therefore chosen as the main topic of the thesis.

9.4 Experimental Studies of Oxide Breakdown

The aim of the preliminary experimental research was to characterise the full spectrum of breakdown phenomena between short-pulse/ESD time-scales and long time-scale dielectric wearout. Chapter 4 described and characterised the apparatus and test structures used for the experiments. The apparatus included a wafer-level microprober station, a d.c. parametric analysis system, a Hartley AutoZap ESD-pulse simulator and a 1GHz digital oscilloscope for monitoring short-pulse/ESD transients. The test samples were supplied on unscribed silicon wafers by four independent semiconductor manufacturers. (To ensure anonymity, these were denoted A,B,C and D). The structures included NMOS, CMOS, HMOS and silicon-on-sapphire (SOS) transistors, together with wide-area MOS-capacitors.

The preliminary experimental program was reported in Chapter 5. The study began by examining the uniformity of ESD oxide breakdown behaviour in the various structures. Positive and negative polarity breakdown thresholds in NMOS, CMOS, HMOS and SOS were found to be reasonably consistent between devices selected from different wafers and different geographical wafer positions. The positive thresholds were, however, significantly greater than their negative counterparts. The MOS Capacitor samples from Supplier C showed a large degree of experimental scatter and were therefore excluded from any further experiments.

Further experiments investigated the sensitivity of the ESD breakdown thresholds to device area, wafer temperature, luminous intensity, channel doping (i.e. n-channel, p-channel, E-Mode or D-Mode), discharge capacitance C_1 and discharge resistance R_2. Positive polarity breakdown thresholds in CMOS and NMOS structures were found to increase with increasing temperature while the negative thresholds remained unchanged. Breakdown voltages decreased with increasing C_1, and increased rapidly with increasing R_2 for $R_2 > 100k\Omega$. Some of the samples showed increased breakdown thresholds under low illumination, although Chapter 6 showed this effect to be illusory.

The negative-polarity response of NMOS capacitors to ramp-voltage stress was examined as a function of ramp speed and wafer temperature. The resulting data (which were mathematically processed in Section 6.5.1), showed that the oxide voltage was 'pinned' at approximately -40V (probably by Fowler-Nordheim tunnelling) prior to breakdown. An increase in the ramp speed caused an increased pinning voltage and a subsequent decrease in the injected charge required to support breakdown (Fig.6.29). No significant temperature dependence was detected. The corresponding positive-polarity responses were also measured. The analysis of these waveforms showed that oxide charge injection from a deeply-depleted cathode requires a device-voltage equal to the sum of the Fowler-Nordheim threshold V_{FN} and the depletion layer avalanche voltage V_{av} (see Figs.6.6.31 and 6.32).

Constant-voltage oxide wearout in NMOS devices was observed as a function of device size, wafer temperature, luminous intensity, structure type (i.e. D-Mode, E-Mode etc.) and electric field. Time-to-breakdown t_{bd} was found to increase with increasing temperature (an observation which contradicted earlier results) and to decrease with increasing field F. It was convenient to divide the t_{bd} vs. F curve into three domains, denoted I, II and III (see Fig.6.13). In the high-field domain (Region III), t_{bd} was found to be dominated by the time-delay t_d between stress-application and the onset of the Fowler-Nordheim injection current. However, in the low-field domain (Region I) the injection current was seen to decrease with time (Fig.6.24).

Constant-current wearout was examined as a function of device size and injection-current. The charge-to-breakdown Q_{bd} was found to be approximately proportional to gate area, confirming that the areal injected charge density is the critical parameter determining breakdown. For a given device area, the value of Q_{bd} was found to fall with increasing current. Above a certain critical current, Q_{bd} falls extremely rapidly, a situation corresponding to the brittle/ductile transition identified by Wolters et al. [3]. This transition was shown to correspond to the boundary between regions I and II in the constant-voltage breakdown data.

9.5 Theoretical Analysis of Oxide Breakdown

A qualitative framework for the modelling of SiO_2 dielectric breakdown was presented at the end of Chapter 6. According to this model, the division of the t_{bd} vs. F curve into Regions I, II and III is due to the simultaneous action of coulombic-repulsive and neutral trap sites and the subsequent oxide space-charge evolution. The first-order rate equation was used to model the capture and emission of electrons which affect the cathode field and hence the tunnelling current. The various possible mechanisms of oxide wearout were reviewed, and the *Surface Plasmon* model was found to be the most consistent with the observed phenomena. The model was then extended to ESD breakdown, which was analyzed not so much as a 'fast' voltage pulse but as a finite charge package available for injection into the gate oxide. This led to a theoretical confirmation of Amerasekera & Campbell's depletion-layer avalanche theory for positive-polarity ESD breakdown [4]. The resulting model provided a qualitative explanation for all of the ESD data reported in this thesis.

The quantitative development of the breakdown model was reported in Chapter 7. The first stage of the analysis assumed a version of the causal oxide wearout model, slightly modified in order to account for the tunnelling time-delay t_d identified in Section 6.4.1.1). This model was combined with the equivalent circuit of the ESD system in order to yield

expressions for the breakdown threshold as a function of discharge capacitance C_1 and discharge resistance R_2. Initially, low resistance ESD ($R_2 < 100k\Omega$) situation was modelled in terms of a single capacitance discharging into a Fowler-Nordheim tunnel junction. This analysis showed that the oxide breakdown potential is approximately constant and independent of the circuit capacitances.

Several models of high-resistance ($100k\Omega < R_2 < 100M\Omega$) ESD breakdown were developed and compared with the experimental data. However, the most successful was found to be a model based upon the oxide-voltage overshoot associated with the time-delay t_d. The limit of the high-resistance model as $R_2 \rightarrow 0$ was shown to be equivalent to the low-resistance model. Ultra-high resistance ($R_2 > 100M\Omega$) breakdown was modelled in terms of a field-induced pulse propagated via the parasitic capacitance C_2 in parallel with R_2.

The models were successfully fitted to the experimental data (although, strangely, the optimized values of the parasitic capacitances C_x and C_2 were about twice their expected values). This model showed that the apparent 'constant energy contour' (upon which some earlier workers have based their claims for a thermal breakdown mechanism [4]) is a mathematical 'quirk' produced by the parasitic circuit elements.

9.6 Studies of Parametric Degradation

Chapter 8 examined the effects of oxide dielectric damage upon device and circuit performance, with particular emphasis on latent failure and parametric drift. The results led to two important conclusions. Firstly, they showed how HBM ESD-induced latent failure is far more likely to result from a punctured gate-oxide than an incomplete oxide wearout. They also show how a 'walking-wounded' device can degrade to become a catastrophic failure under working-voltage conditions (0-20V).

The PSpice computer modelling showed how the various characteristic degradation modes (including those identified by previous workers [6]) can be modelled in terms of gate-oxide resistance paths and the formation of equipotentials in the channel region associated with n^+ gate-dopant intrusion.

Finally the M/y-box fault-analysis technique was discussed. It was shown that the linear-resistive failure mode (Group 5) cannot be easily modelled using the M/y-box algorithm, since it involves a direct connection between the input and output terminals of a transistor. More complex algorithms may therefore be required in order to accommodate this type of failure.

The investigations in this chapter are largely tentative and are intended merely as a springboard for a much deeper investigation to be performed in the near future.

9.7 Final Comments

Although the aims of this project were primarily practical (i.e. the development of predictive failure models for 'building-in reliability'), some of the results may contribute to the general advancement of dielectric wearout theory. Firstly the detection of a finite time delay t_d between stress application and tunnelling is a significant result. Although it has not (to the author's knowledge) been previously documented, t_d fits in with the results of several other workers (see Fig.6.21). A second important result is the correspondance between the I/II boundary in the constant-voltage data and the critical current J_{cr} in the constant-current results. This result allows the so-called 'brittle-ductile' transition of Wolters et al. [3] to be explained in terms of oxide space-charge evolution.

Some of the novel techniques developed and employed during this project may also be of general scientific interest. For example, although the concept of 'charge-limited breakdown' had previously been suggested [4], the visualization of an ESD event as a finite charge-package (rather than as a time-limited pulse) has never before been seriously examined.

The ultimate objective of this research is the improvement of device reliability. Section 9.1 mentioned the concept of 'built-in' reliability, which involves the use of educated design and processing techniques to manufacture devices 'right-first-time'. Since the recent increase in i.c. complexity has introduced difficulties in quality assessment (i.e. 'test-in' quality), built-in reliability has become increasingly desirable. However, 100% built-in reliability would need extremely accurate modelling, impeccably high material quality and highly repeatable processing.

A middle ground between 'built-in' and 'test-in' reliability is the philosophy of 'design-for-testability', in which circuits are designed to be easily tested. Such techniques may include the fabrication of a test-structure on each chip, whose properties may be assumed to characterize the quality of all the nearby circuitry. 'Wafer-level-reliability' (WLR) techniques may be performed upon these structures in order to assess their susceptibility to the relevant failure mechanisms (i.e. oxide wearout, electromigration, hot-electron injection, etc.). Theoretical modelling will then be needed in order to extrapolate the behaviour of the active circuitry from the test-structure data. Hence the application of the theory of Chapter 7 to WLR data modelling is a further area for exploration.

9.8 Suggestions for Future Research

The most obvious flaws in the analytical ESD model of Chapter 7 are the contradictory values of the parasitic capacitances C_x and C_2, as measured by the system characterisation (Table 4.2) and by the empirical model-fit (Section 7.4.3.1). Although this may be explained in terms of the inadequacy of the parallel C/R probe model at high frequencies, this theory must somehow be verified. Additionally, the 'enhanced-injection' model of Section 7.4.4 assumes that t_d is constant, while Fig.6.15 shows it to have a noticeable field dependence. Since the inclusion of field-dependence may possibly improve the correlation between theory and experiment in Fig.7.13, this might provide scope for the immediate advancement of the work. Characterization of constant current/voltage breakdown under positive polarity conditions may allow the model to be applied to the positive polarity ESD results.

Longer-term plans for the extension of the theoretical models have already been outlined at the end of Chapter 7. Further development may also involve the inclusion of an element of statistical uncertainty in order to model the wide experimental 'scatter' observed in many of the data (particularly the MOS Capacitor data of Fig.5.5). This might be achieved by combining the model equations with those of the effective thickness theory (see Section 3.3.6), although statistically-significant quantities of experimental data will then be required in order to verify the resulting model.

The resulting equations could finally be incorporated in a marketable software package designed for reliability modelling applications. Although such packages have already been developed [e.g.7], they assume continuous oxide wearout, a principle which this thesis has shown to be non-universal.

The parametric-drift analysis of Chapter 8 is an area wide open for further research. Further studies could be aimed at producing a reliable damaged-transistor model, with a facility for simulating the walking-wounded drift towards catastrophic failure. This could possibly be achieved by the application of thermal modelling to the heating and subsequent expansion of oxide shorting sites [8]. Improved fault analysis techniques could then be devised in order to model the effects of all the observed failure modes on logic operation.

9.9 Summary

1. The aims of the thesis, (i.e. the investigation of MOS reliability and the development of a sound theoretical basis for 'built-in' reliability techniques) were outlined.

2. The literature-survey chapters were described and their conclusions outlined.
3. The experimental studies on SiO_2 dielectric breakdown were outlined and the results were summarised.
4. The analysis of the experimental data was summarised, together with the theoretical models used to explain the results.
5. The conclusions of the parametric drift studies (Chapter 8) were presented.
6. The chapter ended by discussing of the overall significance of the results of the thesis, and suggesting areas of future research.

9.10 References

1. Dimirrijev,S., Stojadinocic,N, Stamenkovik,Z, *"Yield Model for In-Line Integrated Circuit Production Control"*, Sol. State Electron., 31(5), pp.975-9, 1988.
2. Duvvury,C, Rountree,R.N., Adams,O, *"Internal Chip ESD Phenomena Beyond the Protection Circuit"*, IEEE Trans. Electron Dev., 35(12), pp.2133-9, 1988.
3. Wolters,D.R., van der Schoot,J.J., *"Dielectric Breakdown in MOS Devices - Part II: Conditions for the Intrinsic Breakdown"*, Phil. J. of Res., 40, pp.137-63, 1985.
4. Amerasekera,E.A., Campbell,D.S., *"Charge Limited Breakdown in MOS Capacitors"*, Proc. 17th. Annual Sol. St. Dev. Res. Conference, pp.733-41, 1987.
5. Tunnicliffe,M.J., Dwyer,V.M., Campbell,D.S., *"A Model for ESD Oxide Breakdown in MOS Transistors"*, European ESD Symposium, Eindhoven, Netherlands, 3-5 Sept. 1991.
6. Amerasekera,E.A., Campbell,D.S., *"An Investigation of the Nature and Mechanisms of ESD Damage in NMOS Devices"*, Sol. St. Elect., 32(3), pp.199-206, 1989.
7. Rosenbaum,E, Lee,P.M., Moazzami,R, Ko,P.K., Hu,C, *"Circuit Reliability Simulator - Oxide Breakdown Module"*, 1989 International Electronic Device Meeting (IEDM) Technical Digest, pp.331-4, 1989.
8. Dwyer,V.M., Franklin,A.J., Campbell,D.S., *"Thermal Failure in Semiconductor Devices"*, Sol. State Elect., 33(5), pp.553-60, 1990.

Appendix A

Mathematical Analysis of the MOS Structure

A.1 Introduction

This appendix presents a collection of mathematical proofs and investigations relating to the work of this thesis. It begins with a discussion of oxide capacitance in small dimension MOS structures. This discussion explains the discrepancies observed in Fig.6.20.

The oxide charge-injection mechanism is then considered. The Fowler-Nordheim theory for a metallic or degenerate-semiconductor cathode is developed and extended in order to model injection from a nondegenerate cathode.

The appendix ends with a demonstration of how high-field carrier accumulation and inversion permit the standard Fowler-Nordheim equation to be employed in all practical cathode doping situations.

A.2 Oxide Capacitance in Small Dimension Devices

In a wide-area gate-oxide capacitor, the oxide capacitance C_{ox} is given by

$$C_{ox} = \frac{\varepsilon_0 \varepsilon_{ox} A}{T_{ox}} \qquad \text{A(1)}$$

where A is the gate area and T_{ox} is the oxide thickness. As A decreases, the fringing fields around the capacitor periphery become more significant and Eqn.A(1) ceases to apply. Two alternative situations can arise, depending upon the geometry of the gate edges.

Consider a simple square capacitor structure with area A and periphery $4A^{½}$. If the gate edges are sharp [Fig.A.1(i)] then the total capacitance can be modelled using a simple corollary of Cetner, Iniewski and Jakubowski's two-dimensional interconnection-line capacitance formula [1], i.e.

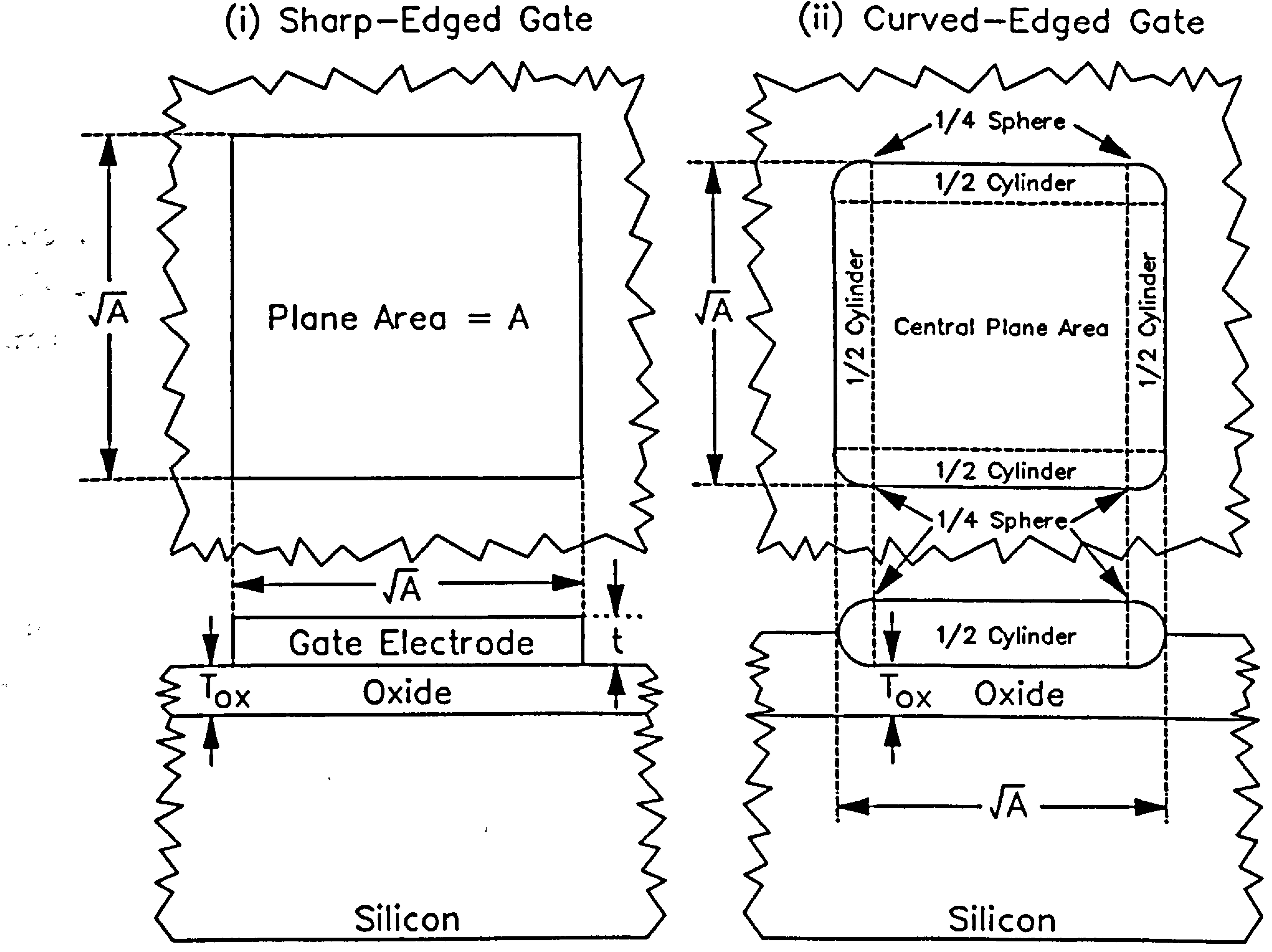


Figure A.1: Plan-views and cross sections of small-dimension capacitor structures.

$$C_{ox} = \frac{\varepsilon_0 \varepsilon_{ox} A}{T_{ox}} \left[1 + \frac{4\,T_{ox}}{\sqrt{A}} \log_e\left(1 + \frac{t}{T_{ox}}\right) + \frac{4\,T_{ox}}{\sqrt{A}} \log_e\left(1 + \frac{\sqrt{A}/2}{T_{ox} + t}\right)\right] \qquad A(2)$$

where t is the gate electrode thickness. This formula clearly predicts a capacitance greater than Eqn.A(1). However, if the gate edges are rounded (as with a etched gate-well) then this model cannot be used. Instead, the structure can be modelled using the system of half-cylinders and quarter-spheres shown in Fig.A.1(ii). The capacitance of a cylinder of radius r and length L, parallel to an infinite equipotential surface is given by

$$C = \frac{\pi\, \varepsilon_0\, \varepsilon_r\, L}{\log_e\left[1 + \frac{2h}{r}\right]} \qquad A(3)$$

where h is the minimum distance between the plain and the cylinder. (Note: This formula assumes $h << L$.) Similarly, the capacitance between a sphere of radius r and an

equipotential plain is given by

$$C = \frac{2\pi \varepsilon_0 \varepsilon_r}{\frac{1}{r} + \frac{1}{2h + r}} \qquad \text{A(4)}$$

The structure of Fig.A.1(ii) can be modelled by using a half of Eqn.A(3) for each half cylinder, a quarter of Eqn.6(4) for each quarter sphere, and Eqn.A(1) for the central plane area. Summing all these components gives the following expression for C_{ox}

$$C_{ox} = \frac{\pi \varepsilon_0 \varepsilon_{ox}\left(2\sqrt{A} - 4r\right)}{2\log_e\left[1 + \frac{2T_{ox}}{r}\right]} + \frac{2\pi \varepsilon_0 \varepsilon_{ox}}{\frac{1}{r} + \frac{1}{2T_{ox} + r}} + \frac{\varepsilon_0 \varepsilon_{ox}\left(\sqrt{A} - 2r\right)^2}{T_{ox}} \qquad \text{A(5)}$$

where r is the radius of curvature of the gate edges. Fig.A.2 shows the log(C_{ox}) vs. log(A) curves predicted by Eqns.A(1), A(2) and A(5). Note that Eqn.A(5) predicts values of C_{ox} below those of Eqn.A(1), a result which agrees with the experimental data also shown in Fig.A.2. Eqn.A(2) on the other hand predicts a slightly larger value of C_{ox} than Eqn.A(1). Although this theory is highly simplified (it assumes a square gate structure), it shows that deviation of the capacitance from the one-dimensional theory is probably due to the gate-edge rounding associated with etching.

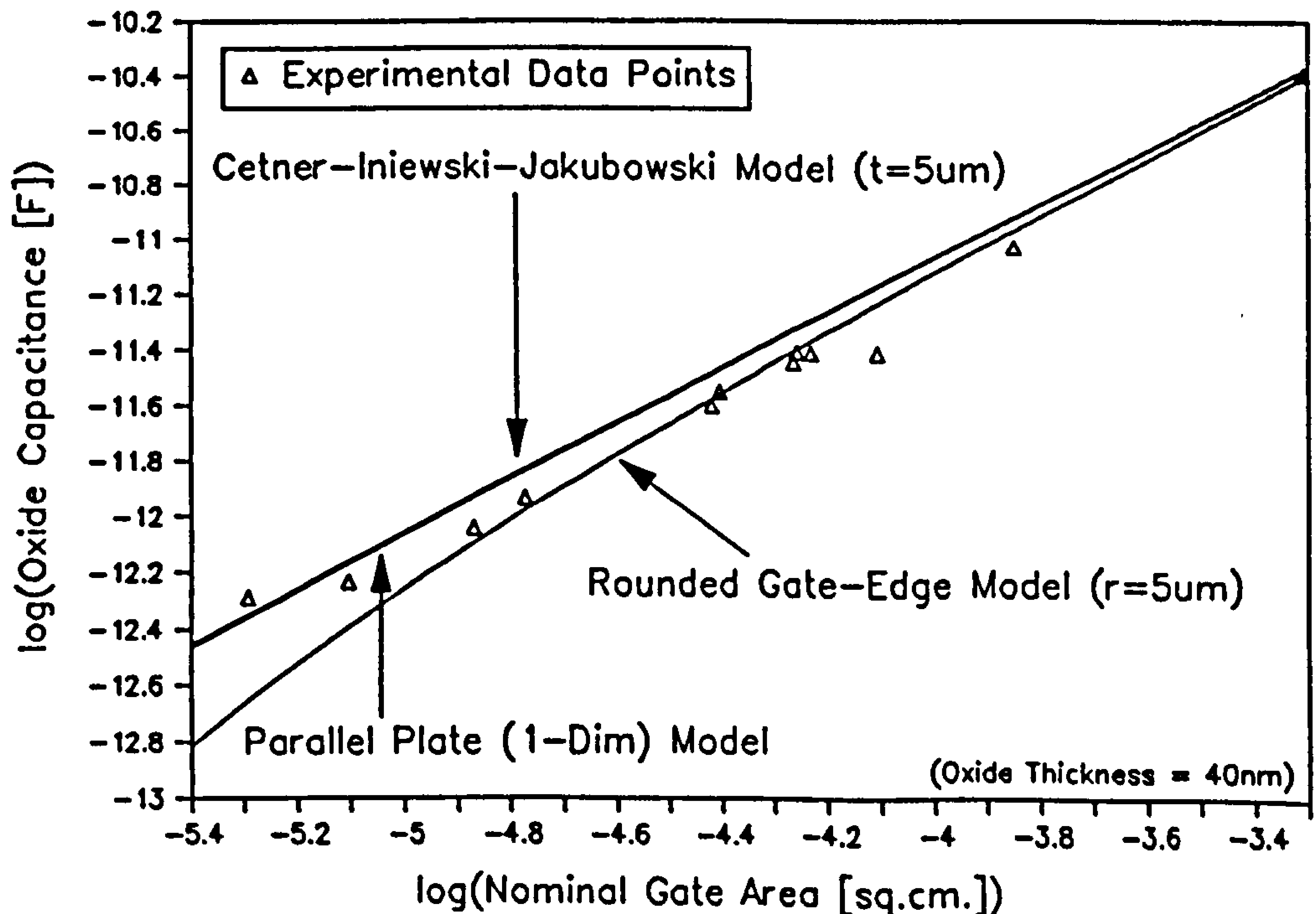


Figure A.2: Small-dimension capacitances - comparison of theory and experiment.

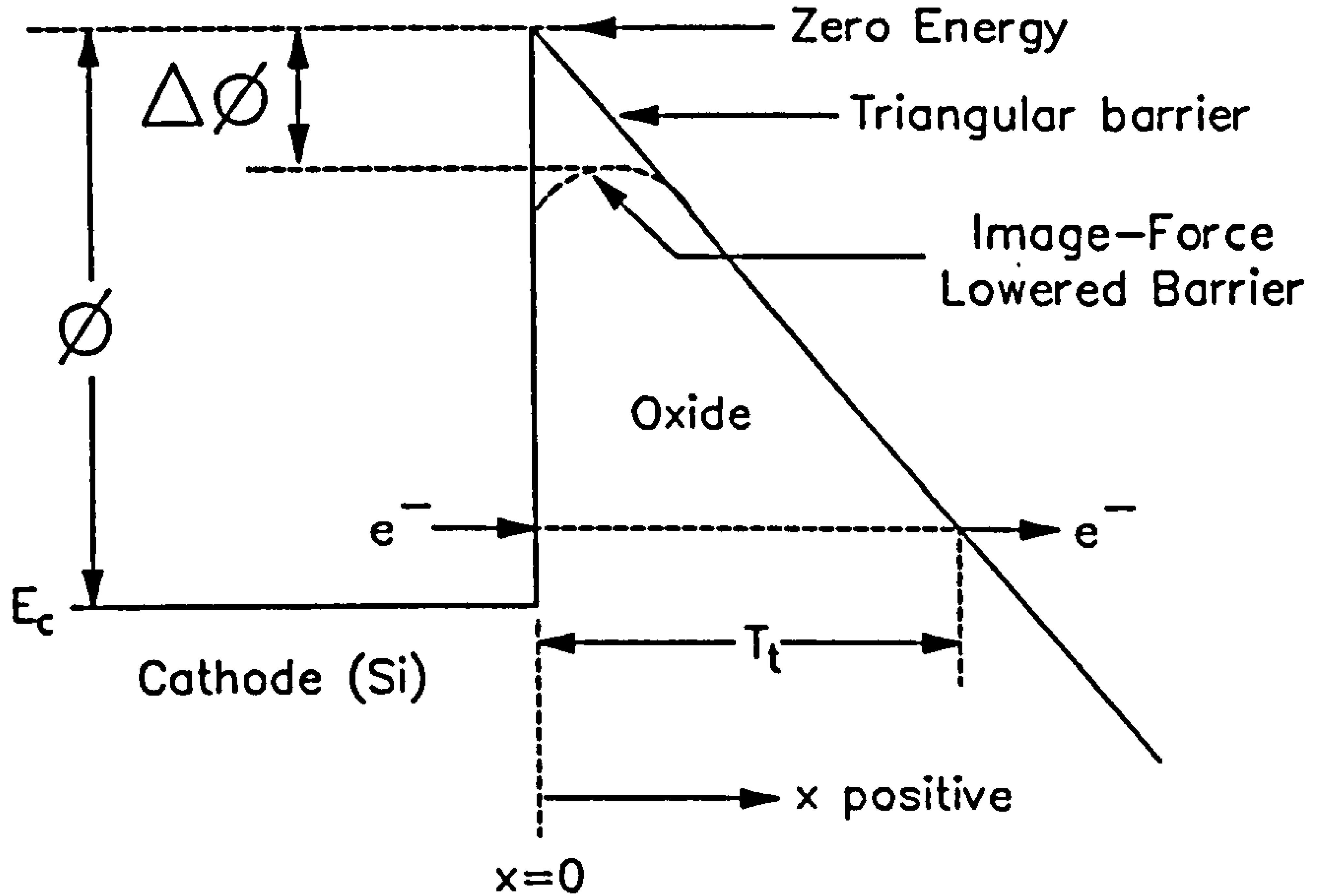


Figure A.3: Schematic Representation of Fowler-Nordheim Tunnelling Barrier

A.3 Fowler-Nordheim Tunnelling Theory

In 1928, Fowler and Nordheim published an historic theoretical paper concerning field-assisted tunnelling of electrons from a metal cathode into a vacuum [2]. Many subsequent workers [e.g.3] have shown that this same theory is applicable to electron tunnelling into wide bandgap insulators such as SiO_2. Fig.A.3 shows a schematic representation of the tunnelling process. The tunnelling current depends upon the barrier transmission coefficient and the rate of electron supply to the cathode/SiO_2 interface. If the electrons at the cathode surface behave as a free Fermi gas, the tunnelling current into the dielectric is given by [4]

$$J = q \int_{p_x=0}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{p_x}{m_{si}^*} \cdot \frac{2}{h^3} \cdot D(E_x) f(E) \cdot dp_x dp_y dp_z \qquad \text{A(6)}$$

where (p_x, p_y, p_z) is the electron momentum, p_x is the momentum normal to the interface, E_x is the energy associated with p_x, E is the total electron energy, $D(E_x)$ is the barrier transmission probability and f(E) is the Fermi-Dirac function, given by

$$f(E) = \frac{1}{1 + \exp\left[\frac{E - E_f}{kT}\right]} \qquad \text{A(7)}$$

where E_f is the electron Fermi-level at the cathode surface. If a parabolic SiO_2 conduction

band is assumed then

$$E = E_c + \frac{p_x^2 + p_y^2 + p_z^2}{2m_{si}^*} \quad ; \quad E_x = E_c + \frac{p_x^2}{2m_{si}^*} \qquad A(8)$$

where E_c is the energy of the conduction-band edge at the interface. The relative values of E_c and E_f determine the mobile electron density, ie.

$$n = \int_{E_c}^{\infty} N(E) f(E)\, dE \qquad A(9)$$

where N(E) is the density-of-states function for the silicon conduction band, which is given by

$$N(E) = \frac{1}{2\pi^2}\left(\frac{2m_{si}^*}{\hbar}\right)^{\frac{3}{2}} \left(E - E_c\right)^{\frac{1}{2}} \qquad A(10)$$

Since the numerical evaluation of Eqn.A(6) is extremely lengthy, it can be combined with Eqn.A(8) to yield a single integral

$$J = q \int_{-\infty}^{\infty} D(E_x) N(E_x)\, dE_x \qquad A(11)$$

where $N(E_x).dE_x$ is the electron *supply function* (ie. the number of electrons in the energy range E_x to E_x+dE_x which interact with the interface per unit time), given by

$$N(E_x)\, dE_x = \frac{4\pi m_{si}^* kT}{h^3} \log_e \left[1 + \exp\left(-\frac{(E_x - E_f)}{kT}\right)\right] dE_x \qquad A(12)$$

The transmission coefficient $D(E_x)$ is determined by solving the Schrödinger equation across the cathode-SiO_2 potential barrier [2]. If the oxide band structure is assumed to be parabolic then

$$D(E_x) \approx \exp\left[-2 \int_0^{T_t} \frac{2\pi}{h} \sqrt{2m_{ox}^* \left(q\Phi(x) - E_x\right)}\, dx\right] \qquad A(13)$$

where $\Phi(x)$ is the barrier potential energy profile, T_t is the thickness of the tunnelling barrier (Fig.A.3) and m_{ox}^* is the effective electron mass in the SiO_2 forbidden gap. Although the

problem has been solved using the Franz E-k relationship [5] (see Section 2.2.3), the parabolic band-structure model has been found to suffice [6]. The effective mass m_{ox}^* has been assigned numerous values throughout the literature (Table A.1), most of which are in the region of $0.5m_0$ where m_0 is the electron rest mass. (In the conduction band, m_{ox}^* is approximately equal to m_0 [7].)

Table A.1: Examples of m_{ox}^* Values Used by Different Workers

Authors	Value of m_{ox}^* Used
Lenzlinger and Snow [3]	$0.42m_0$
Weinberg [6]	$0.5m_0$
Kreiger and Swanson [8]	$0.362m_0$

The simplest model of the junction assumes a triangular potential barrier (Fig.A.3), where $\Phi(x)=E_c$ for $x<0$ and $\Phi(x)=-qFx$ for $x>0$ (F is the cathode electric field). Integrating Eqn.A(13) for such a barrier yields

$$D(E_x) = \exp\left[-\frac{4\sqrt{2m_{ox}^*}\,|E_x|^{3/2}}{3\hbar qF}\right] \qquad A(14)$$

where the oxide conduction band edge is defined as the zero-energy point. For a metallic or degenerate-semiconductor cathode, the zero-temperature tunnelling current J(0) can be found by integrating Eqn.7(6) between $-\infty$ and E_f. Using a linear approximation for $D(E_x)$ about $E_x=E_f$, the integration yields [4]:

$$J(0) = \lambda F^2 e^{-\frac{\beta}{F}} \qquad A(15)$$

where

$$\lambda = \frac{q^3 m_{si}^*}{8\pi h \phi m_{ox}^*} \quad ; \quad \beta = \frac{8\pi\sqrt{2m_{ox}^*}\,\phi^{3/2}}{3hq} \qquad A(16)$$

and ϕ is the effective work function of the Si-SiO_2 potential barrier and m_{si}^* is the electron effective mass in silicon (approximately $1.1m_0$ [9]). Eqn.A(15) (often called the Fowler-Nordheim or F-N equation) produces a linear graph of $\log(J/F^2)$ vs. $1/F$. The tunnelling

current J(T) for a finite temperature T has been computed as follows [4]

$$J(T) = \frac{\pi kT/d}{\sin(\pi kT/d)} J(0) \qquad \text{A(17)}$$

where

$$d = \frac{hqF}{4\pi\sqrt{2m_{ox}^{*}\phi}} \qquad \text{A(18)}$$

Fig.A.4 shows the $\log(J/F^2)$ vs. $1/F$ curve predicted by numerical solution of Eqn.A(11) compared with that given by Eqns.A(15-18) for the same parameters. The analytical approximation is clearly adequate for fields in excess of 3MV/cm. However, Lenzlinger and Snow [3] discovered that a larger m_{ox}^{*} was needed in Eqn.A(18) than in Eqn.7(16) in order to make the model fit their experimental data. This paradox was explained by introducing a temperature dependence to the parameter ϕ.

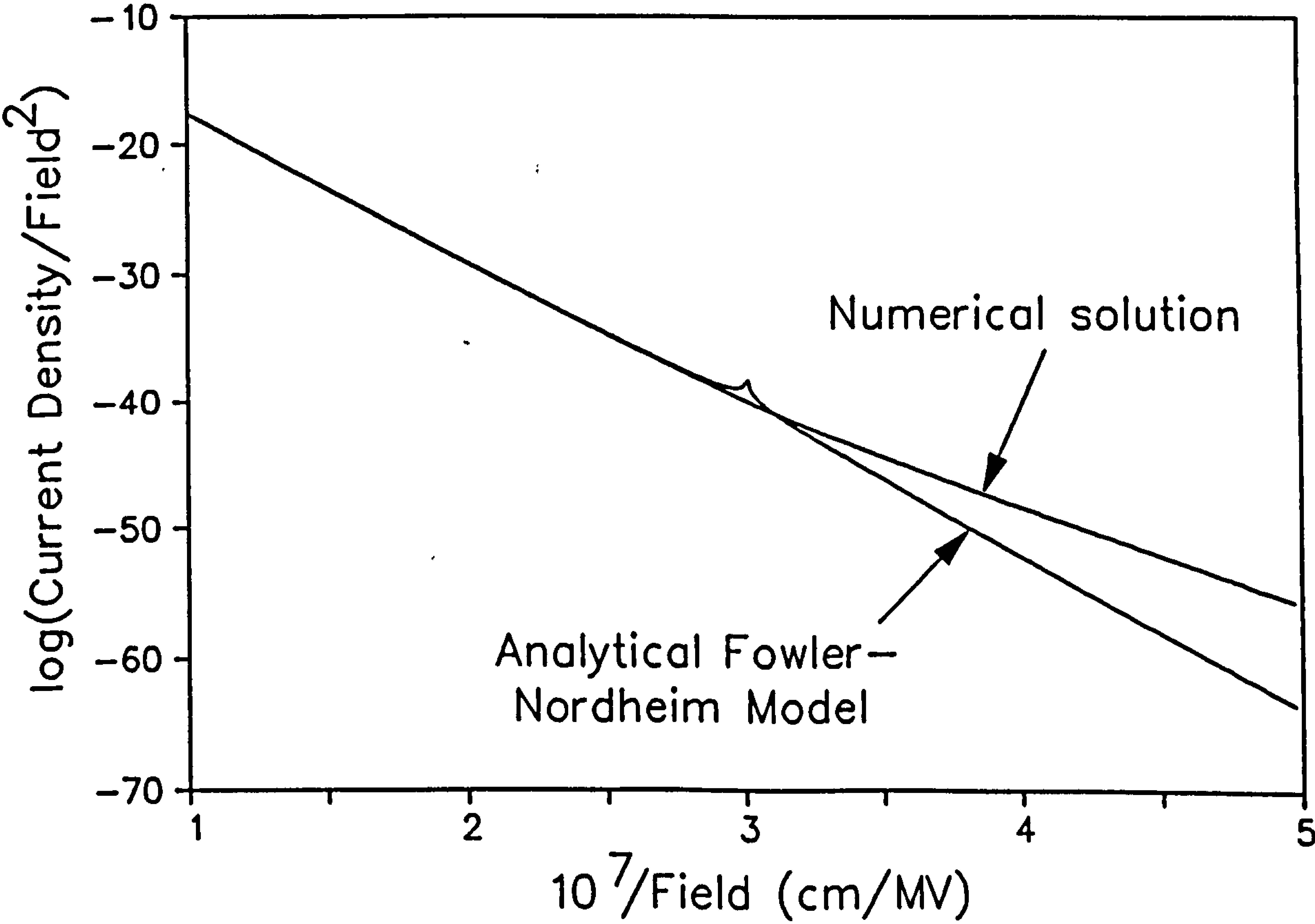


Figure A.4: Comparison of analytical and numerical tunnelling models (T=300K, $m_{ox}^{*}=0.5m_0$).

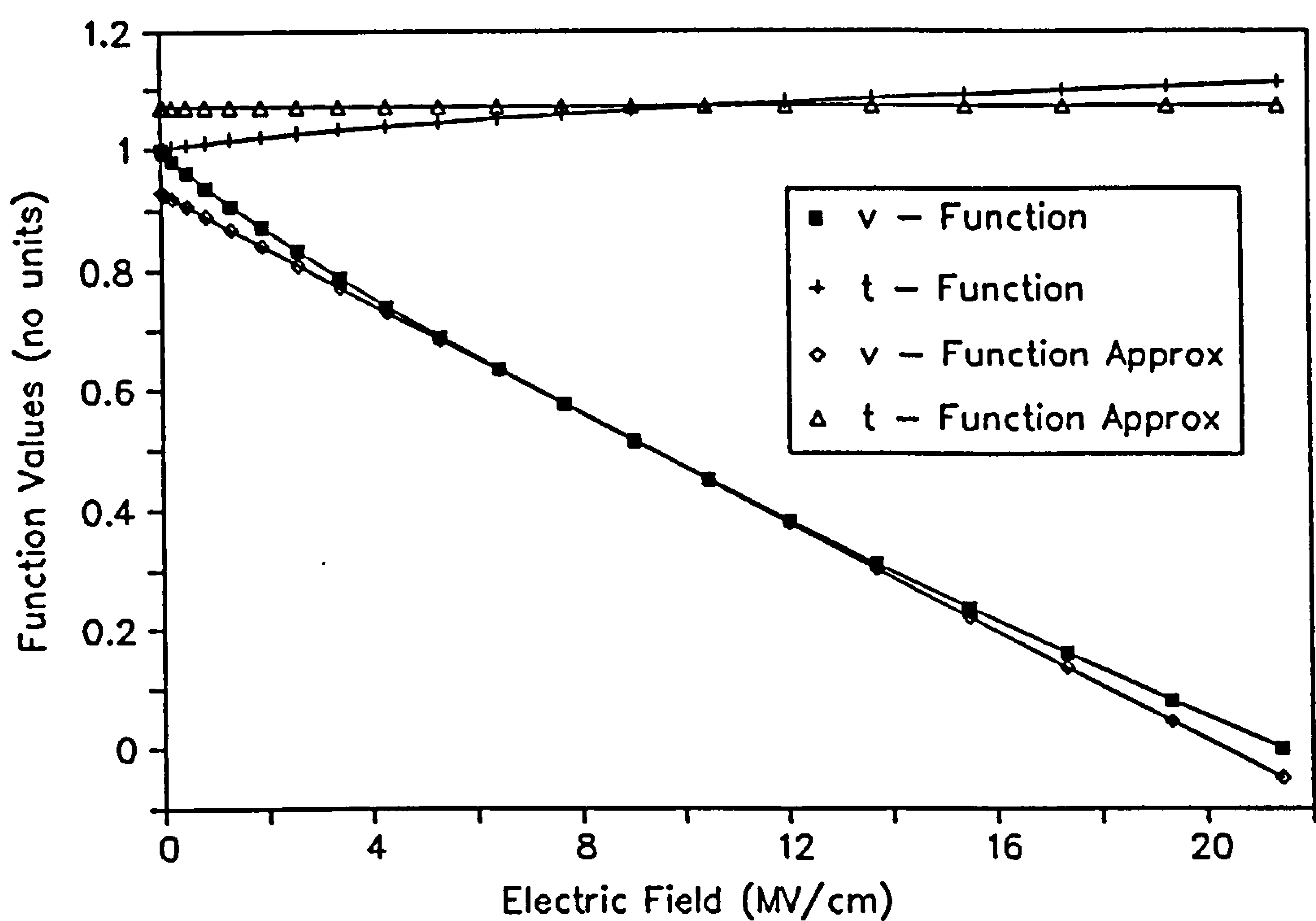


Figure A.5: Classical barrier-lowering functions and linear approximations.

A.4 Image Force Barrier Lowering

According to classical theory, an otherwise triangular Fowler-Nordheim barrier can be distorted due to the 'image-force' experienced by carriers in the oxide [10] (Fig.A.1). The charge -q of every injected electron induces an image-charge +q in the cathode. The proximity of this positive charge causes the electrons to lose potential energy. The height of the Si-SiO_2 barrier is therefore lowered. The following modified forms of β and λ (in Eqn.A(16)) have been developed to account for this effect

$$\lambda = \frac{q^3 m_{si}^*}{8\pi h \phi m_{ox}^* t^2(\Delta\phi/\phi)} \quad ; \quad \beta = \frac{8\pi\sqrt{2 m_{ox}^*}\,\phi^{3/2}}{3hq} v\left(\frac{\Delta\phi}{\phi}\right) \qquad \text{A(19)}$$

where t and v are tabulated functions of $\Delta\phi/\phi$, $\Delta\phi$ being the magnitude of barrier lowering (Fig.A.3) [4,11]. A similar modification is applied to the parameter d (in Eqn.A(18)), ie.

$$d = \frac{hqF}{4\pi\sqrt{2 m_{ox}^*\phi}\, t(\Delta\phi/\phi)} \qquad \text{A(20)}$$

Since $\Delta\phi = (q^3F/4\pi\epsilon_0\epsilon_{ox})^{1/2}$ (see Chapter 3), t and v can both be expressed as functions of field, which are plotted in Fig.A.5. This figure shows that for a limited field range, the following

linear approximations can be used:

$$v(F) = a - bF \quad ; \quad t(F) = a \qquad \text{A(21)}$$

where a=0.929, b=4.573.10^{-8}cm/V and c=1.07 [12]. Hence the Fowler-Nordheim equation can be re-written

$$J(0) = kF^2 e^{-\frac{B}{F}} \qquad \text{A(22)}$$

where k=(λ/c)e$^{\beta b}$ and B=βa. The log(J/F^2) vs. 1/F linearity is therefore preserved with the inclusion of the classical image force barrier lowering.

In 1976, Weinberg and Hartstein [13] discovered that experimental data was more consistent with the simple triangular-barrier model of Eqns.A(15) and A(16) than with the above image-force theory. The fact that the image force definitely exists for hot carrier and photo-injection [14] produces a paradox.

The following solution was offered by Hartstein et al. [15,16], who considered the electron as a wave function ψ_1 incident upon a barrier with a transmission coefficient of D. On impact, the function becomes $\psi_2=\psi_t+\psi_r$, where ψ_t and ψ_r represent the transmitted and reflected components respectively. Since the transmitted component is given by $|\psi_t|^2=D.|\psi_1|^2$, the charge presented by the tunnelling electron is reduced to $D.q$ and the barrier lowering is similarly attenuated. Hence a F-N tunnelling electron, for which D<<1, experiences negligible barrier lowering, while a photo-emitted or 'hot' electron, for which D=1, experiences the full image-force.

Although the above argument is now generally accepted, it has sometimes been criticized on the basis of experimental results. For example, Krieger and Swanson's [8] Si-SiO_2 tunnelling data are equally explainable with and without the image force (assuming different values of m_{ox}^*). This chapter proceeds upon the assumption that the image force *is* negligible, although this choice is arbitrary and has as yet no sound theoretical or experimental justification.

A.5 The Effect of Cathode Electron Concentration

The theory so far has assumed a metallic or degenerate n$^+$ semiconductor cathode. The effects of cathode surface doping on tunnelling and breakdown have been studied by Amerasekera and Campbell [18], who suggested that the dependence of N(E_x) upon the cathode electron density introduces a polarity dependence of the tunnelling characteristics.

They also suggested that increased trap occupation in an n-type surface may increase $D(E_x)$ (although Krieger and Swanson [8] claim that this effect should be negligible). This section presents a theoretical study of tunnelling as a function of cathode doping.

The first step is to examine the effect of surface electron density on the tunnelling current. For the general case, the value of J for a given n and F can be computed by numerical solution of Eqns.A(7-14), replacing the lower limit of the integral in Eqn.A(11) with E_c. However, in a non-degenerate semiconductor, the following analytical approximations can be used:

$$f(E) \approx \exp\left[\frac{E_f - E}{kT}\right] \quad ; \quad n \approx N_c \exp\left[-\frac{E_c - E_f}{kT}\right] \qquad A(23)$$

where N_c is the effective density of states for the silicon conduction band, given by

$$N_c = 2\left(\frac{2\pi m_{si}^* kT}{h^2}\right)^{\frac{3}{2}} \qquad A(24)$$

Under these approximations, Eqn.A(6) can be re-written:

$$J = \frac{4q\pi kT}{h^3}\cdot\frac{n}{N_c}\int_0^\infty p_x D(E_x)\, e^{-\frac{p_x^2}{2m_{si}^* kT}}\, dp_x \qquad A(25)$$

The exponential term in the integrand peaks so sharply at an energy 1.5kT above the conduction band edge, that the corresponding value of $D(E_x)$ can be assumed throughout the energy range. Since the oxide conduction band edge is defined as zero energy, $E_f=-\phi$ and Eqn.A(23) can be rearranged to yield

$$E_c = kT\cdot\log_e\left[\frac{N_c}{n}\right] - \phi \qquad A(26)$$

where ϕ is assumed to be 3.15eV irrespective of the value of n. Under all these assumptions, Eqn.A(25) can be re-written

$$J = R^* T^2 \frac{n}{N_c}\cdot D\left(kT\left(\log_e\left[\frac{N_c}{n}\right] + \frac{3}{2}\right) - \phi\right) \qquad A(27)$$

where R^* is the effective Richardson constant (equal to $4m_{si}^*\pi qk^2/h^3$). According to Eqn.A(14), the transmission coefficient in Eqn.A(27) can be written

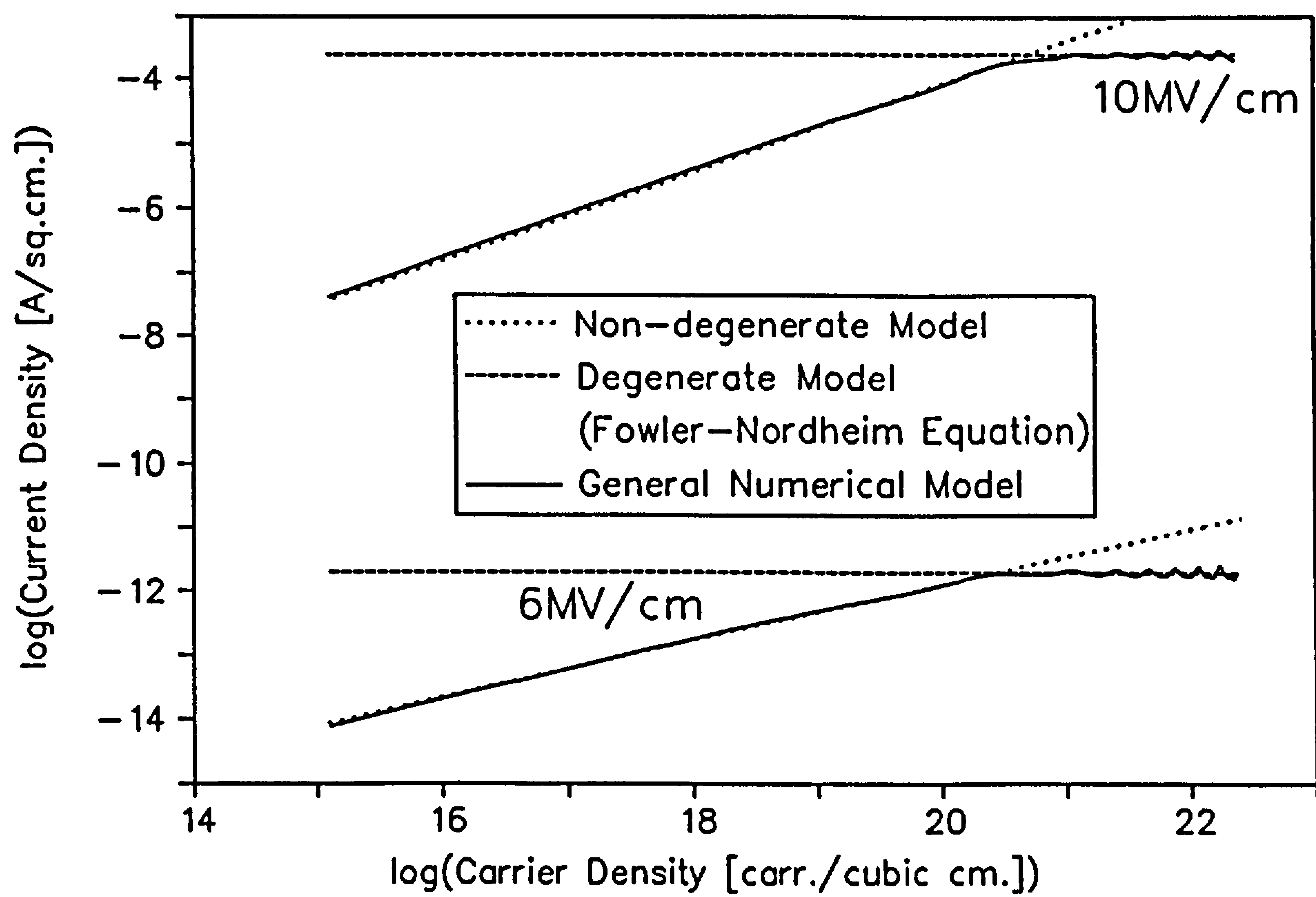


Figure A.6: Tunnelling current as a function of cathode electron density.

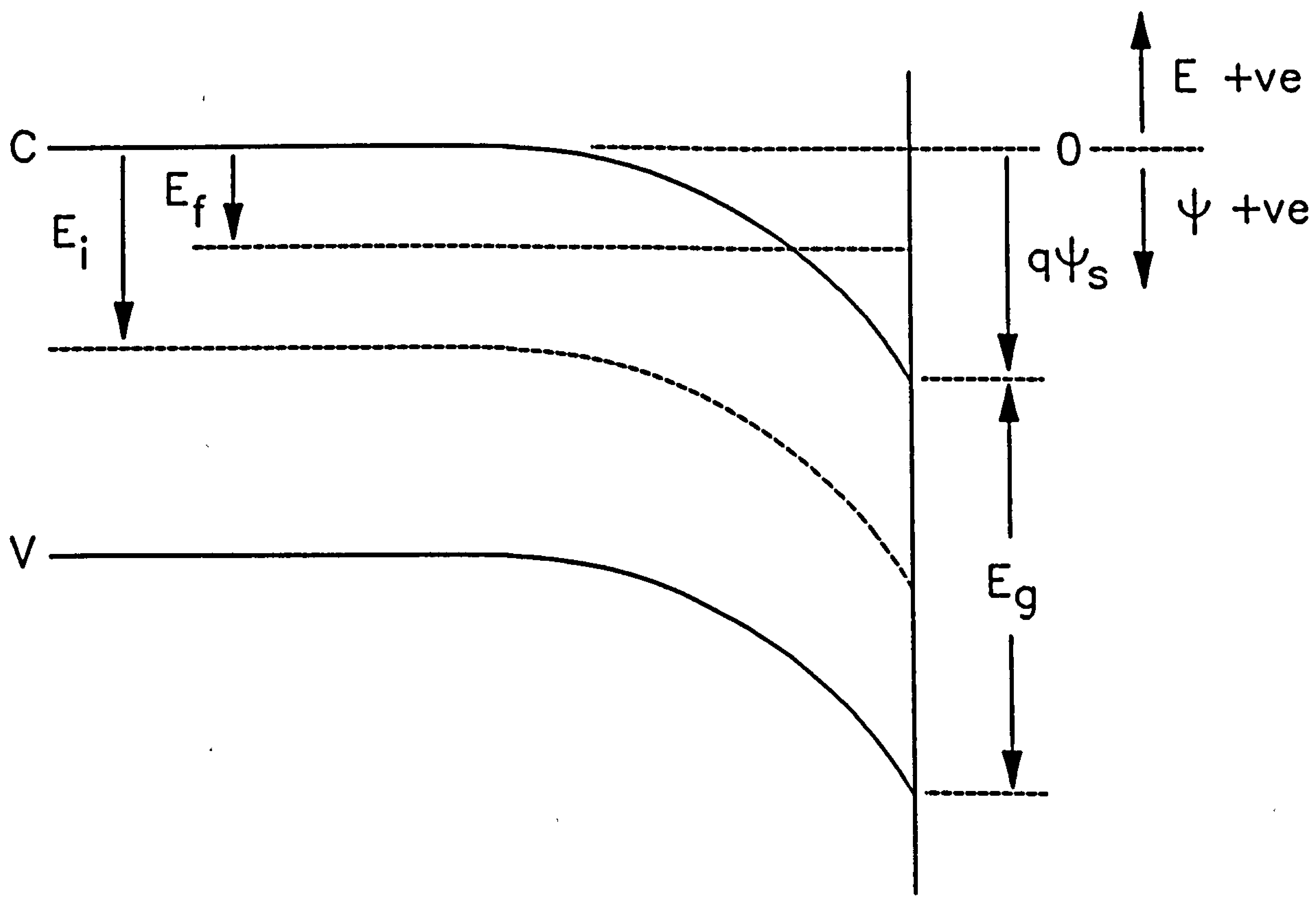


Figure A.7: Model of Band Bending and Electron Accumulation.

$$D = \exp\left[-\frac{4\sqrt{2m_{ox}^{*}}}{3\hbar qF}\left|kT\cdot\log_e\left(\frac{N_c}{n}\right) - \phi + \frac{3}{2}kT\right|^{\frac{3}{2}}\right] \qquad A(28)$$

Applying the binominal expansion to Eqn.A(28), yields

$$D \simeq \exp\left[-\frac{4\sqrt{2m_{ox}^{*}}}{3\hbar qF}\left[\left(\phi - \frac{3}{2}kT\right)^{\frac{3}{2}} + \frac{3}{2}kT\sqrt{\phi - \frac{3}{2}kT}\,\log_e\left(\frac{n}{N_c}\right)\right]\right] \qquad A(29)$$

which allows Eqn.A(27) to be re-written as

$$J \simeq R^{*}T^{2}\left(\frac{n}{N_c}\right)^{Y(F)} D\left(\frac{3}{2}kT - \phi\right) \qquad A(30)$$

where

$$Y(F) = 1 - \frac{2kT\sqrt{2m_{ox}^{*}}}{\hbar qF}\sqrt{\frac{3}{2}kT - \phi} \qquad A(31)$$

Fig.A.6 shows the room temperature J vs. n curves predicted by numerical solution of Eqns.A(7-14) compared with those predicted by the standard Fowler-Nordheim theory of Eqns.A(15-16) and the nondegenerate model of Eqn.A(30-31) for two different electric fields. The two analytical models clearly provide regional approximations appropriate for degenerate and non-degenerate domains. The switch-over between the two regions takes place fairly abruptly.

A.6 The Effect of Cathode Doping

Since the degenerate model is accurate for $n \geq 10^{21}cm^{-3}$, it is suitable for injection from a $10^{21}cm^{-3}$ doped n^{+}-type polysilicon gate. If the cathode is more lightly n-doped (or even p-doped), it still becomes degenerate n-type as an accumulation or inversion layer forms under a high electric field (Fig.A.7). The relationship between surface charge density, field and potential was explored for non-degenerate surfaces by Kingston et al. [20] and extended to degenerate surfaces by Seiwatz and Green [21]. According to this latter model, surface field F_s and surface potential ψ_s are related by the equation

$$F_s = \frac{1}{L_D}\frac{kT}{q}\left(\frac{N_D - N_A}{n_i}\frac{q\psi_s}{kT} + K_1 - K_2\right) \tag{7(32)}$$

where

$$K_1 = \frac{2}{3}\frac{\left[F_{\frac{3}{2}}\left(\frac{E_f + q\psi_s}{kT}\right) - F_{\frac{3}{2}}\left(\frac{E_f}{kT}\right)\right]}{F_{\frac{1}{2}}\left(\frac{E_i}{kT}\right)} \tag{7(33)}$$

$$K_2 = \frac{2}{3}\frac{\left[F_{\frac{3}{2}}\left[-\frac{q\psi_s + E_g + E_f}{kT}\right] - F_{\frac{3}{2}}\left[-\frac{E_g - E_f}{kT}\right]\right]}{F_{\frac{1}{2}}\left[-\frac{E_g + E_i}{kT}\right]} \tag{7(34)}$$

$$L_D = \sqrt{\frac{\varepsilon_0 \varepsilon_{si} kT}{2q^2 n_i}} \tag{7(35)}$$

$$F_j(\eta) = \int_0^{\infty} \frac{x^j}{1 + \exp(x - \eta)} dx \tag{7(36)}$$

The surface potential ψ_s and the surface carrier density n(0) are related by

$$n(0) = 4\pi\left[\frac{2 m_p^* kT}{h^2}\right]^{\frac{3}{2}} F_{\frac{1}{2}}\left(\frac{E_f + q\psi_s}{kT}\right) \tag{7(37)}$$

Fig.A.8 shows the n(0) vs. F curve predicted for $4.10^{15}cm^{-3}$ n-doped silicon. The surface carrier density is clearly driven above $10^{21}cm^{-3}$ by fields below 1MV/cm, showing that the degenerate Fowler-Nordheim equation should be adequate for modelling breakdown at fields in the 10MV/cm region.

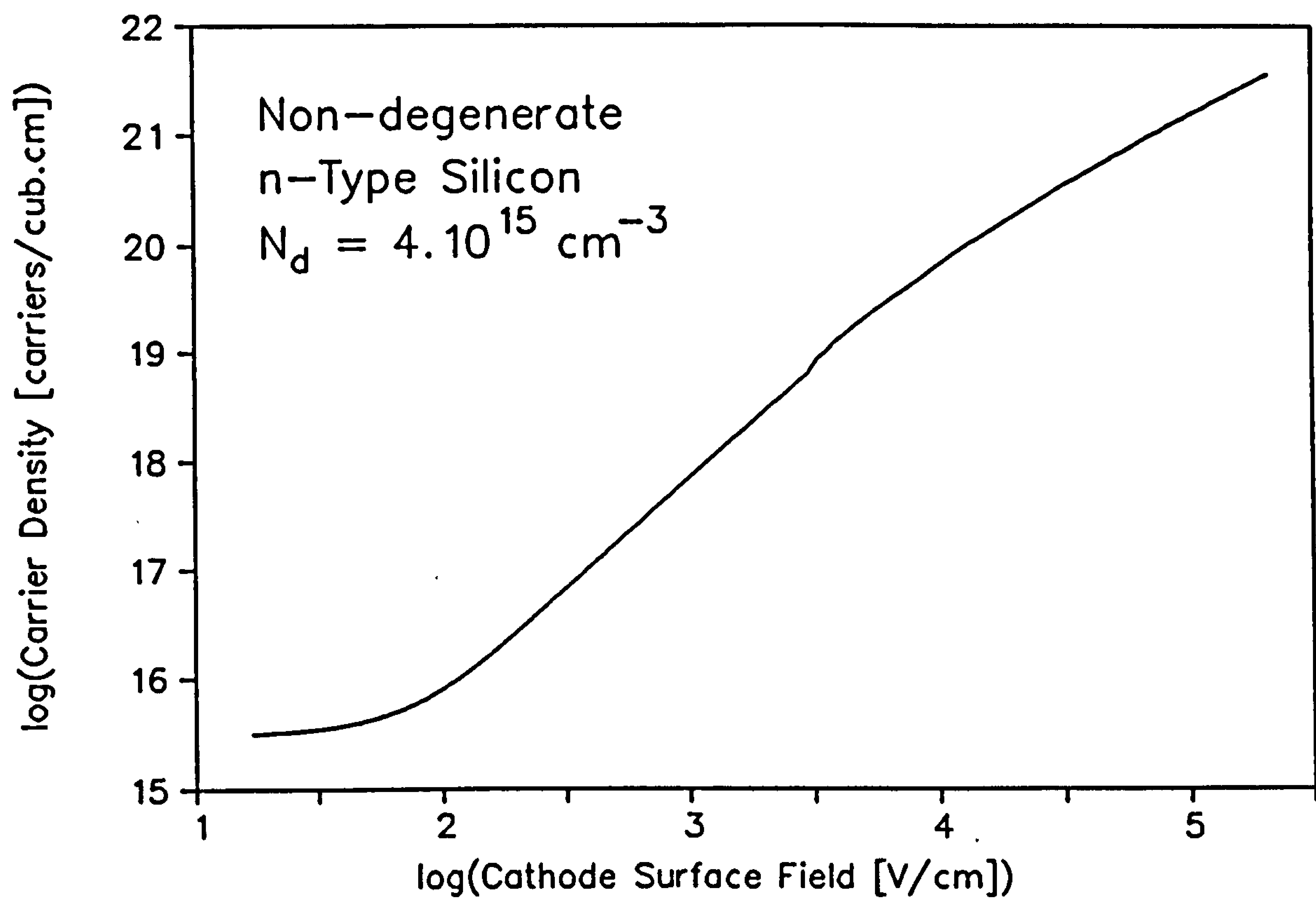


Figure A.8: Surface carrier density vs. field relationship predicted by Seiwatz model for $4.10^{15}cm^{-3}$ n-doped Si.

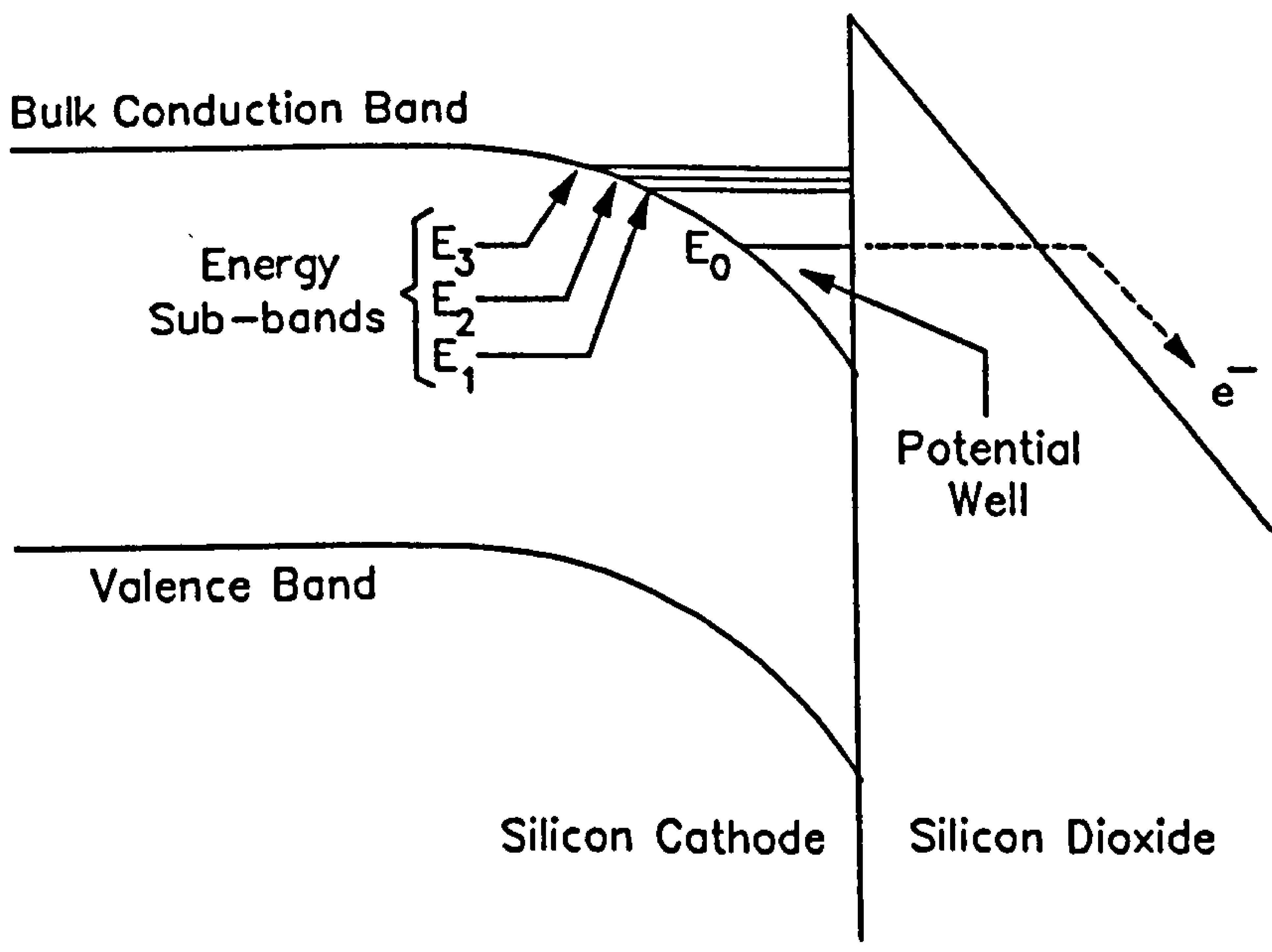


Figure A.9: The formation of surface sub-bands at the p-type Si/SiO_2 interface.

A.7 Effect of Cathode Surface Quantization

However, complications are introduced if the silicon is heavily p-doped. The subsequent surface band bending introduces a potential well in the cathode, quantizing the x-momentum of the inversion electrons and affecting the supply function and the transmission coefficient [6]. The surface electron wavefunctions are diffracted into a number of two-dimensional sub-bands (Fig.A.9), the lowest of which (E_0) dominates tunnelling [21]. (An excellent review of two-dimensional theory is given by Mott [22].) Although the analysis of tunnelling in such conditions is extremely complex, the current has been shown to be approximately 2.5 times the current predicted by the Fowler-Nordheim equation [21], a deviation which is insignificant when viewed in the field domain.

A.7 References

1. Cetner,A, Iniewski,K, Jakubowski,A, *"A Simple Formula for Two-Dimensional Capacitance"*, Solid State Electronics, 31(5), pp.973-4, 1988.
2. Fowler,R.H., Nordheim,L, *"Electron Emission in High Electric Fields"*, Proc. Roy. Soc., A119, p.173-81, 1928.
3. Lenzlinger,M, Snow,E.H., *"Fowler-Nordheim Tunnelling into Thermally Grown SiO_2"*, J.Appl.Phys., 40(1), pp.278-83, 1969.
4. O'Dwyer,J.J., *"The Theory of Electrical Conduction and Breakdown in Solid Dielectrics"*, Clarendon Press, Oxford, pp.77-94, 1973.
5. Av-Ron,M, Shatzkes,M, DiStefano,T.H., Gdula,R.A., *"Electron Tunneling at Al-SiO_2 Interfaces"*, J.Appl.Phys., 52(4), pp.2897-908, 1981.
6. Weinberg,Z.A., *"On Tunnelling in Metal-Oxide-Silicon Structures"*, J.Appl.Phys., 53(7), pp.5052-6, 1982.
7. Verwey,J.F., Amerasekera,E.A., Bisschop,J, *"The Physics of SiO_2 Layers"*, Rep. Prog. Phys., 53, pp.1297-331, 1990.
8. Krieger,G, Swanson,R.M., *"Fowler-Nordheim Electron Tunneling in Thin Si-SiO_2-Al Structures"*, J.Appl.Phys., 52(9), pp.5710-7, 1981.
9. Streetman,B.G., *"Solid State Electronic Devices"*, Prentice/Hall International, N.J., p.65, 1980.
10. Nordheim,L.W., *"The Effect of the Image Force on the Emission and Reflexion of Electrons by Metals"*, Proc. Roy. Soc., 121, pp.626-39, 1928.
11. Burgess,R.E., Kroemer,H, *"Corrected Values of Fowler-Nordheim Field Emission Functions v(y) and s(y)"*, Phys. Rev., 90(4), p.515, 1953.
12. Tunnicliffe,M.J., Dwyer,V.M., Campbell,D.S., *"Experimental and Theoretical Studies of EOS/ESD Oxide Breakdown in Unprotected MOS Structures"*, Proc. 12th. EOS/ESD Symp. pp.162-8, 1990.
13. Weinberg,Z.A., Hartstein,A, *"Photon Assisted Tunneling from Aluminium into Silicon Dioxide"*, Solid State Comm., 20, pp.179-82, 1976.
14. Garrigues,M, Balland,B, *"Hot Carrier Injection into SiO_2"*, Chapter 9 in *"Instabilities*

in Silicon Devices, Vol.1: Silicon Passivation and Related Instabilities", Editors: G.Barbottin & A.Vapaille, North-Holland (Amsterdam), 1986.

15. Hartstein,A, Weinberg,Z.A., *"On the Nature of the Image Force in Quantum Mechanics with Application to Phonon-Assisted Tunnelling and Photoemission"*, J.Phys.C: Solid State Physics, **11**, pp.L469-73, 1978.
16. Hartstein,A, Weinberg,Z.A., *"Unified Theory of Internal Photoemission and Photon-Assisted Tunneling"*, Physical Rev. B, **20**(4), pp.1335-8, 1979.
17. Appelbaum,J.A., Hamann,D.R., *"Variational Calculation of the Image Potential near a Metal Surface"*, Physical Rev. B, **6**(4), pp.1122-30, 1972.
18. Amerasekera,E.A., Campbell,D.S., *"ESD Pulse and Continuous Voltage Breakdown in MOS Capacitor Structures"*, Proc. **8th.** EOS/ESD Symposium, pp.208-13, 1986.
19. Kingston,R.H., Neustadter,S.F., *"Calculation of the Space Charge, Electric Field, and Free Carrier Concentration at the Surface of a Semiconductor"*, J. Appl. Phys., **26**(6), pp.718-20, 1955.
20. Seiwatz,R, Green,M, *"Space Charge Calculations for Semiconductors"*, J. Appl. Phys., **29**(7), pp.1034-40, 1958.
21. Weinberg,Z.A., *"Tunnelling of Electrons from Si into Thermally Grown SiO_2"*, Solid State Elect., **20**, pp.11-18, 1977.
22. Mott,N, *"Conduction in Non-Crystalline Materials"*, Clarendon Press, Oxford, pp.107-13,1987.

Appendix B

Abstracts of Progress Reports

B.1 Progress Report No.6[1]

Tunnicliffe,M.J., Dwyer,V.M., *"Failure Mechanisms in Semiconductor Devices Progress Report No.6"*, Contract No.(LUT)ELJB7/(MOD)A5a/1412, Loughborough University of Technology, October 1988.

This Report covers the work performed between November 1987 and October 1988 on the project "Failure Mechanisms in Semiconductor Devices".

Previous research is summarised and the existing study position is defined for the benefit of the new staff working on the project.

The new experimental work is presented. This focuses on the transient response of MOS structures to constant voltage stress.

Several modifications to the existing breakdown models are proposed. These modifications are supported by the results of earlier reports and by the new experimental data.

B.2 Progress Report No.7

Tunnicliffe,M.J., Dwyer,V.M., *"Failure Mechanisms in Semiconductor Devices Progress Report No.7"*, Contract No.(LUT)ELJB7/(MOD)A5a/1412, Loughborough University of Technology, August 1990.

This report surveys work undertaken on the project "Failure Mechanisms in Semiconductor Devices" (Contract No. (LUT)ELJB7/(MOD)A5a/1412) between October 1988 and August 1990.

Experimental studies concentrated on breakdown and pre-breakdown behaviour in

[1]. Progress reports 1 to 5 were written by Dr.E.A. Amerasekera, and describe research performed before the author joined the project.

MOS capacitor and transistor structures under 'slow' voltage ramp (10-1000V/ms), 'rectangular' pulse and ESD pulse conditions. The Hewlett Packard 54111D oscilloscope, linked to the Walters 286 computer, proved a valuable tool and provided much interesting information on these phenomena. The experimental data was numerically analyzed and a mathematical model of ESD oxide breakdown under both positive and negative polarity stress was developed.

These studies were supplemented by an investigation of the ESD sensitivity of silicon-on-sapphire devices. The structures (p and n-channel transistor arrays) were seen to degrade in much the same manner as the bulk silicon structures investigated earlier in the course of this contract.

B.3 Progress Report No.8(a)

Tunnicliffe,M.J., Dwyer,V.M., *"Failure Mechanisms in Semiconductor Devices Progress Report No.8(a) :- Failure Mechanisms in MOS Technology: a Literature Survey"*, Contract No.(LUT)ELJB7/(MOD)A5a/1412, Loughborough University of Technology, December 1991.

This document presents an up-to-date review of research into the failure physics of MOS devices, compiled as part of the *'Failure Mechanisms in Semiconductor Devices'* contract at the International Electronics Reliability Institute, Loughborough University of Technology. The document begins with a brief summary of the experimental and theoretical work conducted on this contract. The remainder of the work is devoted to a global review of MOS breakdown, based on more than two hundred specially selected references.

The review begins with a description of the MOS structure and its properties. This is followed by a discussion of the MOS transistor, its theory of operation and the various integrated-circuit technologies based upon it.

The causes of failure (ie. intrinsic, extrinsic, defect and time related) are then described, with particular emphasis upon electrostatic discharge (ESD) stress. This is followed by a discussion of the various failure mechanisms such dielectric breakdown, hot-carrier injection and electromigration.

B.4 Progress Report No.8

Tunnicliffe,M.J., *"Failure Mechanisms in Semiconductor Devices Progress Report No.8 (Final Report)"*, Contract No.(LUT)ELJB7/(MOD)A5a/1412, Loughborough University of Technology, November 1992.

This report presents an experimental and theoretical investigation of electrical failure in MOS structures, with a particular emphasis on short-pulse and ESD failure. It begins with an extensive survey of MOS technology, its failure mechanisms and protection schemes. A program of experimental research on MOS breakdown is then reported, the results of which are used to develop a model of breakdown across a wide spectrum of time scales. This model, in which bulk-oxide electron trapping/emission plays a major role, prohibits the direct use of *causal* theory over short time-scales, invalidating earlier theories on the subject.

The work is extended to ESD stress of both polarities. Negative polarity ESD breakdown is found to be primarily oxide-voltage activated, with no significant dependence on temperature of luminosity. Positive polarity breakdown depends on the rate of surface inversion, dictated by the Si avalanche threshold and/or the generation speed of light-induced carriers. A unified analytical model, based upon the above-mentioned oxide breakdown model is developed to predict ESD breakdown under any given conditions.

The project ends with an experimental and theoretical investigation of the effects of ESD breakdown on device and circuit performance. Breakdown sites are modelled as resistive paths in the oxide, and their distorting effects upon transistor performance are studied. The degradation of a damaged transistor under working stress is observed, giving a deeper insight into the latent hazards of ESD damage.

Appendix C

Reproductions of Published Papers

This appendix reproduces a series of papers, written by the author, which relate to the work of this thesis. The papers were originally published in fully refereed specialist conferences. Much of the content of these papers also appears in the main body of the thesis.

Paper I

Tunnicliffe,M.J., Dwyer,V.M., Campbell,D.S.,

"Slow Voltage Transient and ESD Breakdown in Unprotected MOS Gate Oxides",

Proc. 1st. Components Engineering Reliability and Test Conference (CERT)
Gatwick Airport, London, U.K., 9th-11th. May, 1990,

pp.78-87, 1990.

PAGE

NUMBERING

AS ORIGINAL

SLOW TRANSIENT VOLTAGE AND ESD BREAKDOWN IN UNPROTECTED MOS GATE OXIDES

by

M.J. Tunnicliffe, V.M. Dwyer and D.S. Campbell

Electronic Component Technology Group,
Department of Electronic and Electrical Engineering,
Loughborough University of Technology,
Loughborough, Leicestershire, LE11 3TU, U.K.
Tel.(0509) 222855 Fax. (0509) 222854

ABSTRACT

The breakdown of unprotected MOS oxides under transient voltage and ESD stress has been studied both experimentally and theoretically. It is shown that under relatively slow voltage transient conditions (~10V/ms), when the series resistance of the voltage source is high (1MΩ), the tunnelling injection at the oxide cathode limits the voltage across the oxide, allowing it to withstand a considerable injected charge prior to breakdown. Effectively the oxide acts as its own protection diode.

As the transient speed is increased, the oxide voltage is forced slightly upwards, causing a dramatic decrease in the charge required for breakdown. Under the very fast voltage transients associated with ESD stress, the charge needed for breakdown becomes almost negligible. Hence, when the charge available is limited to only a few nanocoulombs, breakdown depends soley upon the magnitude of the applied stress voltage. This observation is used to derive a general expression for the ESD breakdown voltage threshold of an unprotected oxide.

The effect of temperature upon transient breakdown has also been studied but no distinct trends were observed.

1. Introduction

Failure of silicon MOS devices is often the result of ESD (electrostatic discharge) events. An ESD event can be caused by the transfer of static electricity from a charged human body to the gate terminal of a device. This induces a fast electric field transient in the gate oxide, which can lead to dielectric breakdown.

Numerous attempts have been made to identify the mechanisms responsible for breakdown in thin SiO_2 films. Most studies have concentrated on breakdown under continuous or relatively slow transient voltage stress. It has been shown that, under these conditions, breakdown is linked to the injection of charge from the cathode (the negatively biased oxide surface) into the oxide, and only occurs after a certain quantity of charge has been injected [1-3,7]. It has been found that the voltage threshold required to induce breakdown, under very slow voltage transient conditions, (~1V/min) decreases with increasing temperature [4].

Only a few workers have examined breakdown under the very fast voltage transients which appear under ESD conditions. Fong & Hu [5] have developed a simple model to predict the lifetime of an oxide under a fast field transient of arbitary shape. This model is based on the same basic assumptions used to describe breakdown under slow transient stress. However, Amerasekera and Campbell [6] have found that ESD breakdown in unprotected NMOS structures is independent of temperature, suggesting that ESD and continuous (or slow transient) voltage breakdown are due to separate mechanisms [4]. They went on to show that the energy required to support ESD breakdown appears to be approximately constant, implying a thermal criterion for breakdown [7].

The present work examines the behaviour of the oxide under both slow voltage transient and ESD conditions. An accurate model of ESD breakdown in unprotected MOS oxides is developed.

2. Theoretical Analysis

2.1 Oxide Breakdown Theory

Fig.1 shows a simplified energy band diagram an MOS structure. Breakdown is related to the tunnelling injection of electrons from the cathode into the oxide conduction band. The magnitude of the injection current density J_{ox} is given by the

Fowler-Nordheim equation

$$J_{ox} = kF^2 e^{-\frac{\beta}{F}} = J_o e^{-\frac{\beta}{F}} \qquad (1)$$

where F is the electric field magnitude at the cathode, and k and β are constants. In addition, the field F varies sufficiently slowly that J_o can be regarded as constant. The total injected charge Q_{bd} required to support breakdown is given by

$$Q_{ox} = A \int_0^{t_{bd}} J_{ox}\, dt \qquad (2)$$

where A is the oxide area and t_{bd} is the time delay to breakdown.

Several models have been developed to predict Q_{bd} from the stressing conditions. Throughout this work, the 'hole trapping' oxide breakdown model [2] will be assumed. Although there have recently been some objections to this model [3], it is simple and predicts most of the observed breakdown phenomena and is therefore adequate as a working assumption.

A simple version of the model is developed below. The breakdown process consists of two distinct phases: pre-breakdown and breakdown. These are discussed below.

(i) Pre-breakdown. [Fig.1(i)]:

Electrons tunnelling from the cathode into the oxide conduction band are accelerated by the electric field and generate electron-hole pairs by collision ionisation. The newly created holes drift toward the cathode where some of them are trapped. (This hole trapping occurs primarily in localised 'weak' oxide regions). The positive space-charge density $Q_p(t)$, associated with these trapped holes at time t, is given by [2]

$$Q_p(t) = \eta \int_0^{t} J_{ox}\, \alpha_0 (T_{ox}-T_t)\, e^{-\frac{H}{F}}\, dt \qquad (3)$$

where η is the trapping efficiency, J_{ox} is the tunnelling current, T_{ox} is the oxide thickness, T_t is the tunnelling length [see Fig.1(i)], α_0 and H are constants. [Note: Eqn.(3) assumes that electric field F is homogeneous throughout the dielectric.]

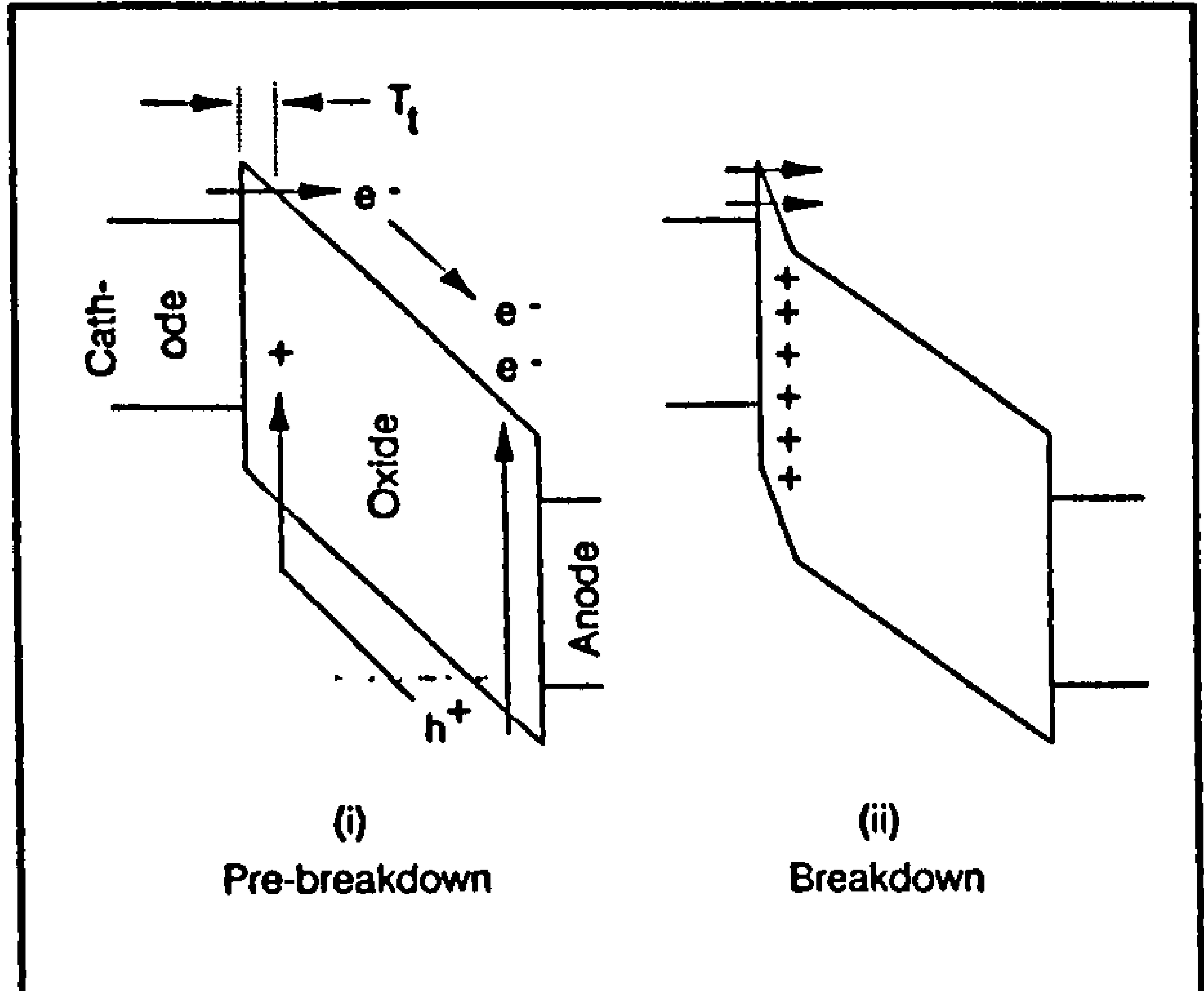


Figure 1: Energy band diagrams for stressed MOS structures before and during breakdown.

(ii) Breakdown. [Fig.1(ii)]:

When $Q_p(t)$ exceeds a certain critical value Q_p^*, the electric field at the cathode becomes noticably enhanced and the tunnelling current increases. This causes the above processes to accelerate very rapidly, leading almost immediately to breakdown. Inserting $Q_p = Q_p^*$ into Eqn.(3) and re-arranging yields

$$\int_0^{t_{bd}} J_{ox}\, e^{-\frac{H}{F}}\, dt = \frac{K}{T_{ox}-T_t} \qquad (4)$$

where t_{bd} is the time to breakdown and K ($=Q_p^*/\eta\alpha_0$) is a constant whose value represents the resilience of the oxide to stress. If the field F remains constant with time then Eqn.(2) may be combined with Eqn.(4) to yield

$$Q_{bd} = \frac{AK}{T_{ox}-T_t}\, e^{\frac{H}{F}} = \frac{AK}{T_{ox}-T_t}\, e^{\frac{HT_{ox}}{V_{ox}}} \qquad (5)$$

where V_{ox} is the voltage across the oxide.

Eqn.(5) shows how an oxide under a relatively low stress voltage can withstand considerable charge injection prior to breakdown. As the voltage is increased the injected charge required to support breakdown rapidly falls.

2.2 Oxide Breakdown under ESD Stress

Fig.2 shows a simplified equivalent circuit for an ESD 'human body model' or HBM test system (MIL-STD-883C) connected to an unprotected MOS device. The discharge or 'body' capacitance C_1 is charged to the applied stress voltage V_o and is then allowed to discharge via the 'body resistance' R into the DUT (device under test).

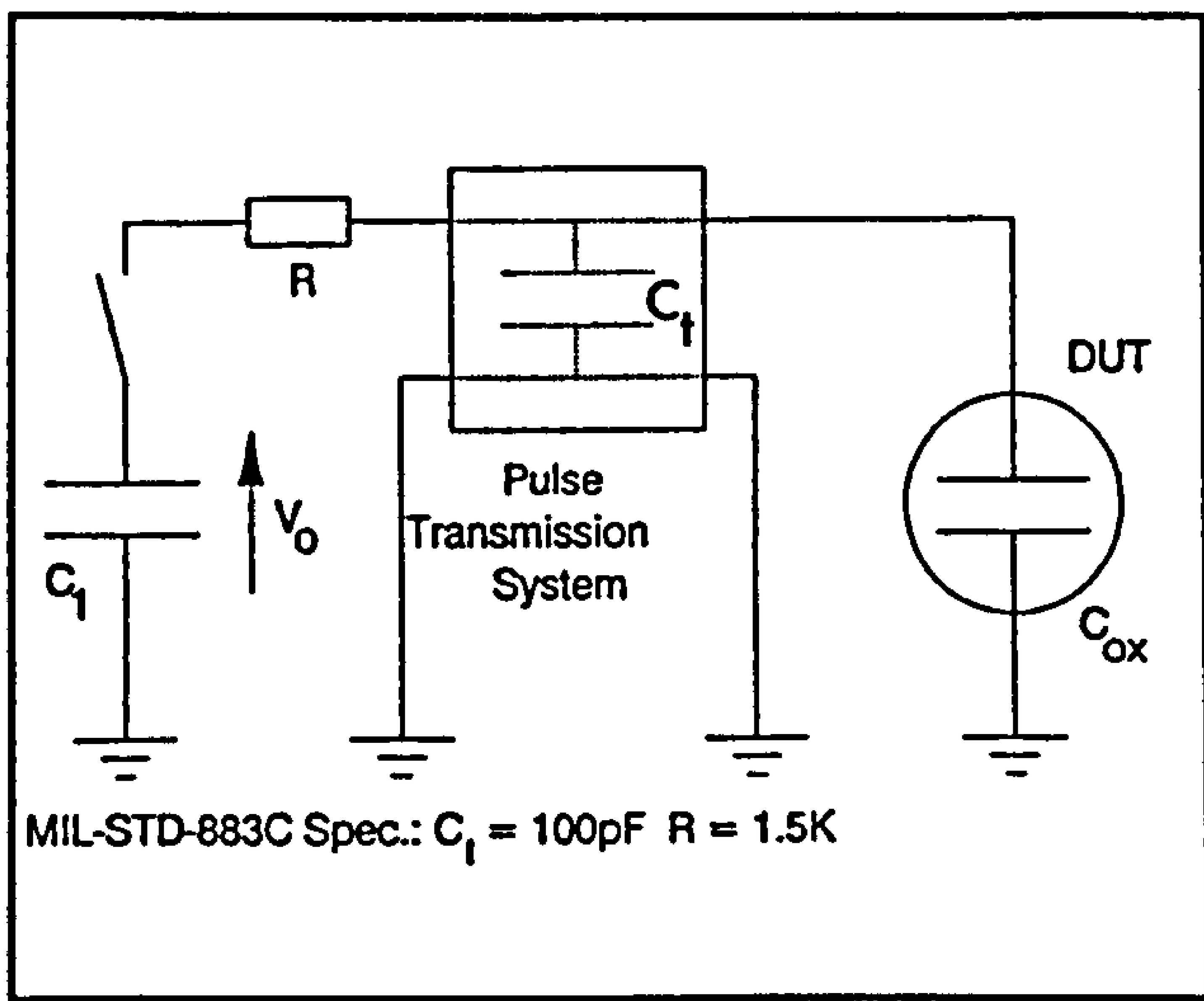


Figure 2: 'Human Body Model' ESD test system (MIL-STD-883C).

The voltage and current transients recieved by the DUT are extremely fast, with time constants of the order of 150ns. The charge available for injection into the DUT is limited to $C_1.V_0$ and the energy is limited to $\frac{1}{2}C_1.V_0^2$. Both the DUT and the pulse transmission system present capacitive loads to the test system, causing the 'effective' voltage V_{eff} appearing across the DUT to be somewhat lower in magnitude than V_0 [7]. Simple circuit theory gives the relationship between V_0 and V_{eff} to be

$$V_{eff} = \frac{V_0\,C_1}{C_1 + C_{ox} + C_t} \quad (6)$$

where C_{ox} is the capacitance of the oxide and C_t is the capacitance of the pulse transmission system.

It will be assumed that an ESD pulse of effective voltage V_{eff}, applied to an unprotected MOS oxide, is equivalent to a constant voltage pulse of θV_{eff} where θ is a constant. Under this assumption, Eqns.(5) and (6) may be combined to yield the following relationship between Q_{bd} and V_0:

$$Q_{bd} = \frac{AK}{T_{ox}-T_t}\exp\frac{HT_{ox}(C_1+C_{ox}+C_t)}{\theta C_1 V_0} \quad (7)$$

This relationship is plotted in Fig.3, using reasonable parameter values (some parameters were taken from references [2] & [7], θ was set to 1 and K was chosen to give optimum correlation with the results presented later in this paper).

Breakdown can only occur if the available charge $C_1.V_0$ exceeds Q_{bd}. Fig.3 shows $C_1.V_0$ plotted against V_{eff}, for several values of C_1. The value of V_{eff} at which the $C_1.V_0$ curve crosses the Q_{bd} curve represents the effective ESD breakdown voltage threshold V_{eff}^*. Note that the value of V_{eff}^* varies only slightly when C_1 is varied in the range 100-200pF. This indicates that the breakdown threshold should depend only slightly upon stressing conditions.

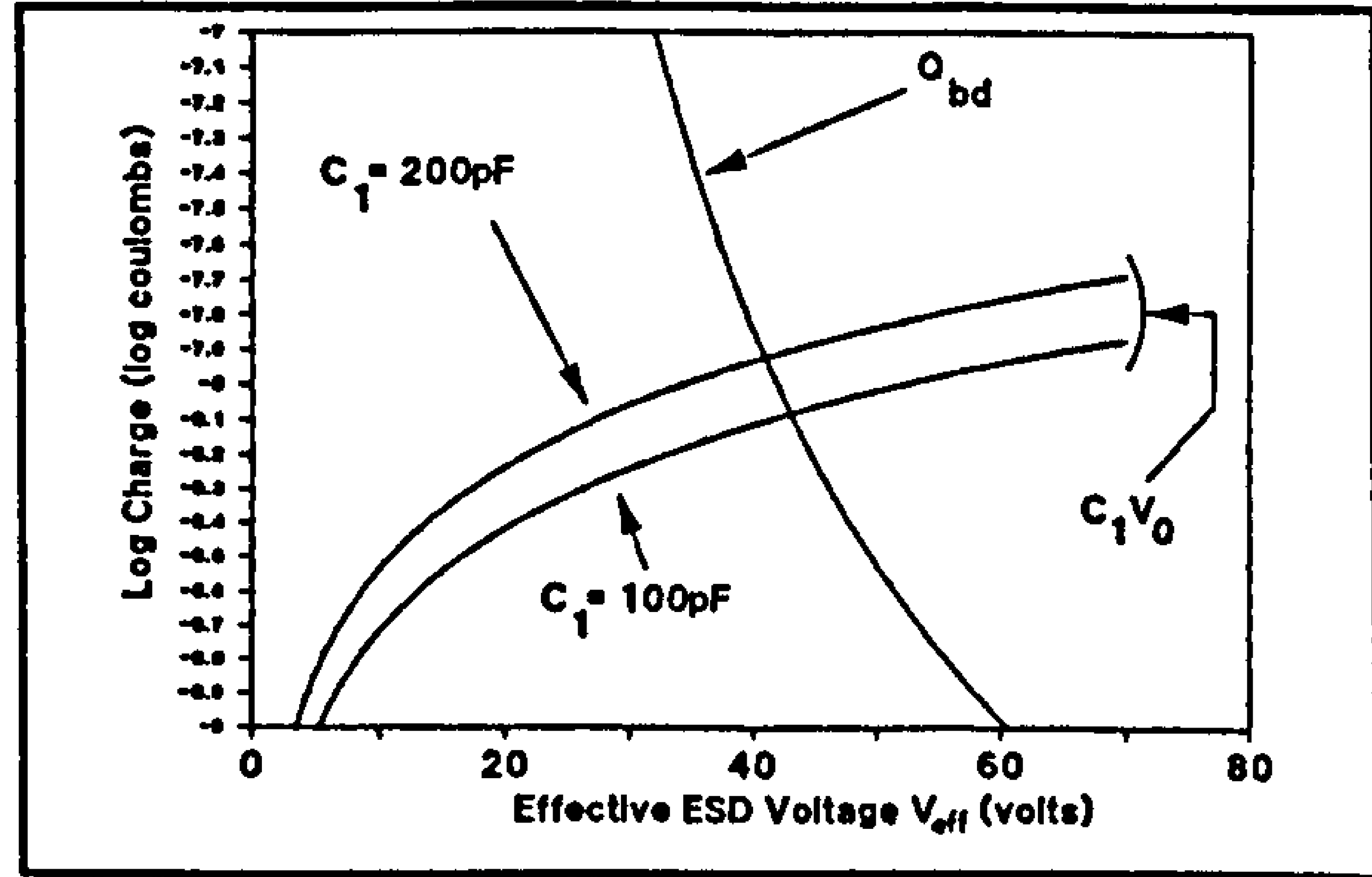


Figure 3: Illustration of ESD breakdown criterion for MOS devices. The point at which the C_1V_0 curve crosses the Q_{bd} curve indicates the breakdown threshold.

3. Experimental

3.1 Description of Test Samples

The test structures used were wide area gate oxide capacitors with n^+-type polysilicon gates and p type silicon substrates, fabricated on bulk silicon wafers (Fig.4). The properties of these capacitors are listed below:

Gate Area (A): 49×10^{-5} cm^2
Oxide Thickness (T_{ox}): 400 Å
Gate Doping (N_D): 1×10^{21} cm^{-3}
Substrate Doping (N_A): 4×10^{15} cm^{-3}.

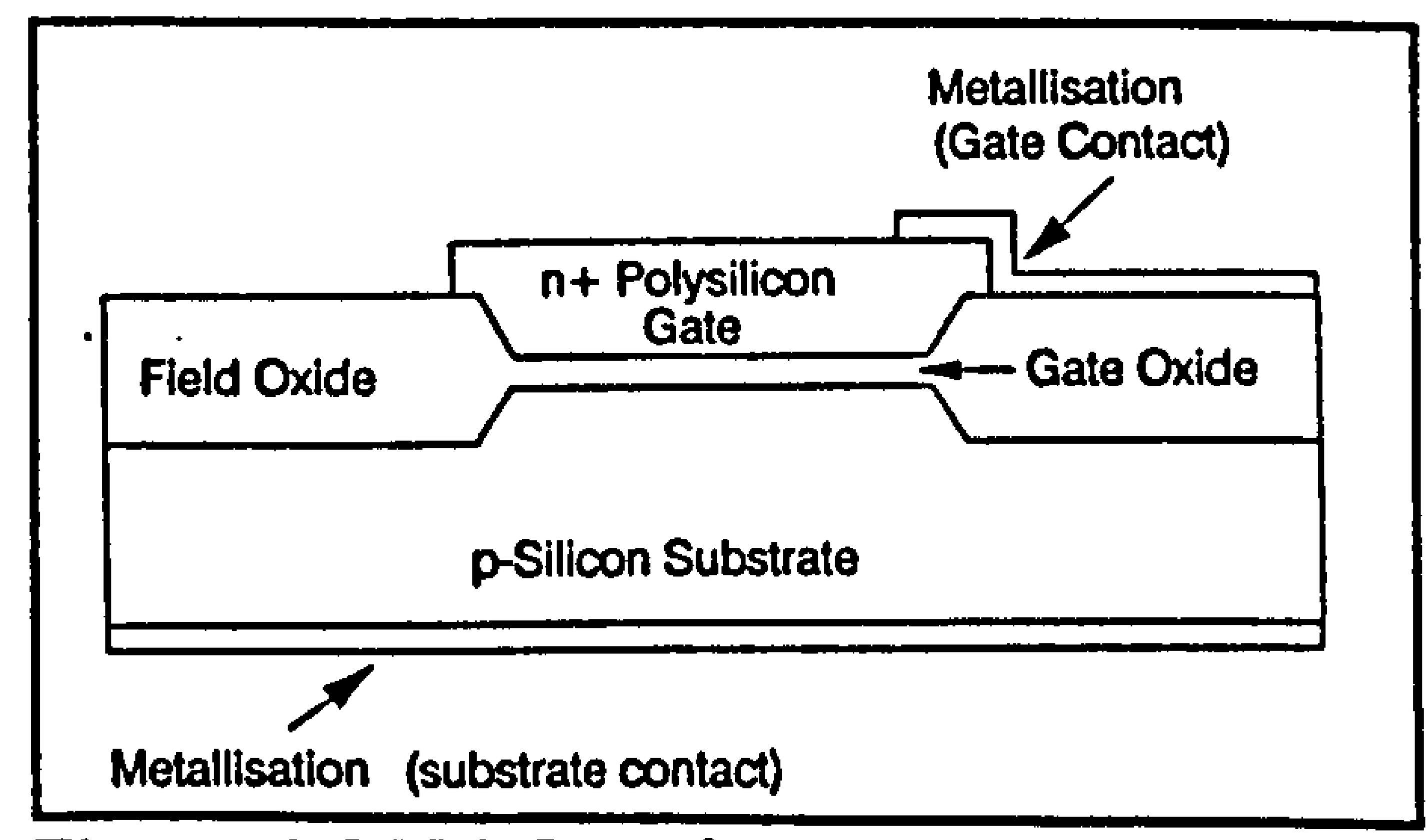


Figure 4: MOS Capacitor Test Structure (cross section).

In all the experiments, the polysilicon gates were negatively stressed such that majority carriers accumulated at both oxide surfaces and hence most of the applied voltage appeared across the oxide.

Results obtained from 21 capacitors are presented. Each capacitor was allocated a serial number between 1 and 21 by which it is identified throughout this paper.

3.2 Wafer Chuck and Microprober

The wafers were mounted on a chuck and the individual capacitor gates were accessed by a microprober (Fig.5). The ambient temperature was thermostatically controlled between room temperature and 140°C. Electrical contact was made to the capacitors via both the microprober and the chuck (which was in contact with the metallised wafer base).

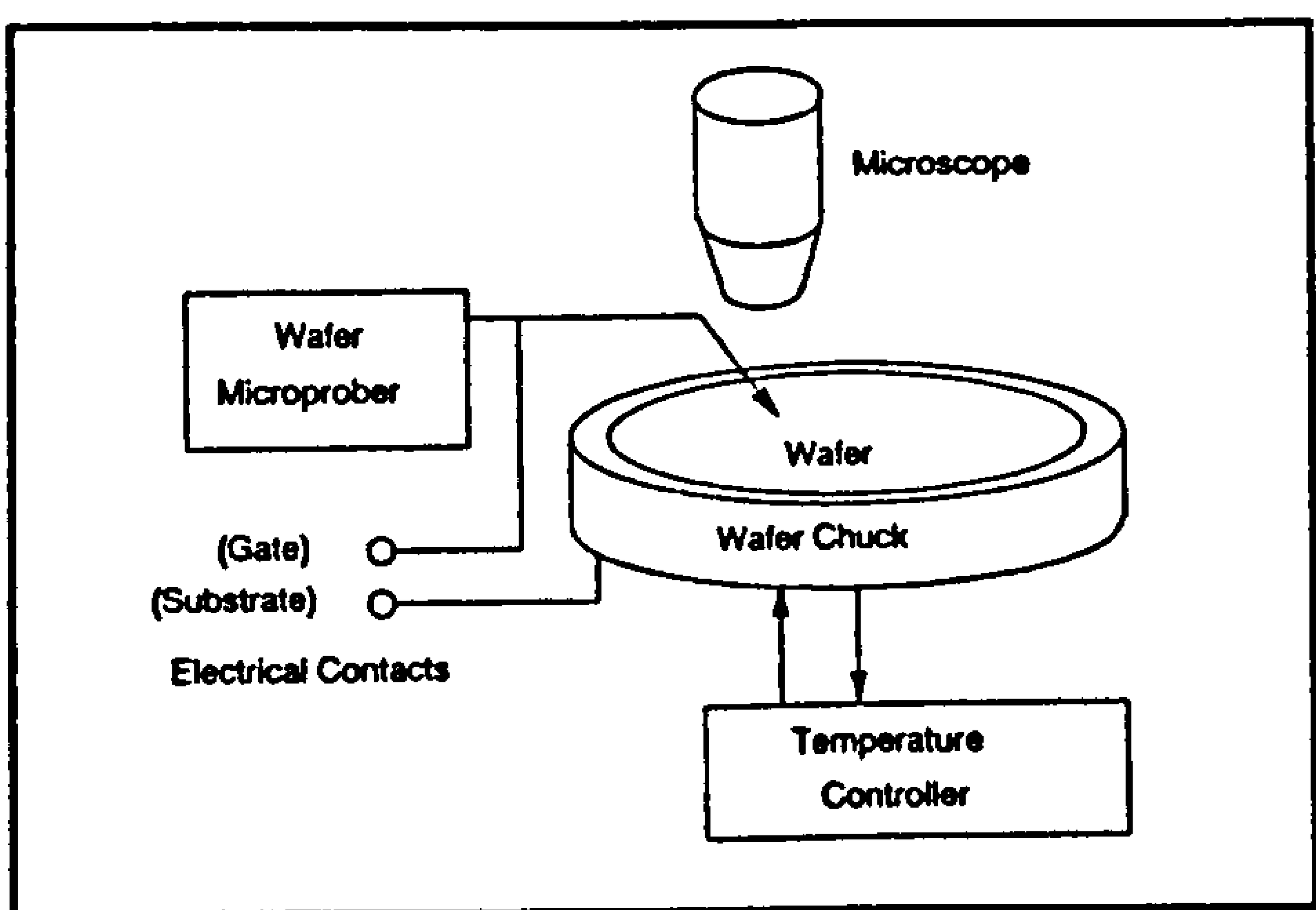


Figure 5: Wafer/Chuck Assembly

3.3 Characterisation of Test Samples

The test structures were characterised in terms of their oxide capacitances C_{ox} and bulk resistances R_b. These parameters were measured using the apparatus shown in Fig.6. A 7V, 2MHz sinusoidal voltage was applied to the structure, together with a -15V d.c. bias (to drive the silicon surface into accumulation). The voltage was monitored by an HP 10440A probe and the current by a Tektronix CT1 current probe. Voltage and current signals were captured on an HP 54111D oscilloscope and downloaded to a Walters 286 personal computer for analysis. The a.c. impedence of the structure was subsequently calculated and C_{ox} and R_b were extracted using simple a.c. theory. This technique yielded approximate values of C_{ox} = 41pF and R_b = 400Ω, which were reasonably consistent between samples. The measured value of C_{ox} was in close agreement with its theoretical value of 42.3pF ($=\epsilon_0\epsilon_r A/T_{ox}$).

3.4 'Slow' Voltage Transient Experiments

Fig.7 shows the apparatus used for the slow transient experiments. A purpose built pulse generator produced voltage transients with varying speeds which were applied to the device under test via short flying leads. The current in the system was monitored in terms of the voltage V_b across a resistor R_s (=1MΩ) connected in series with the wafer chuck. R_s also served to limit the current in

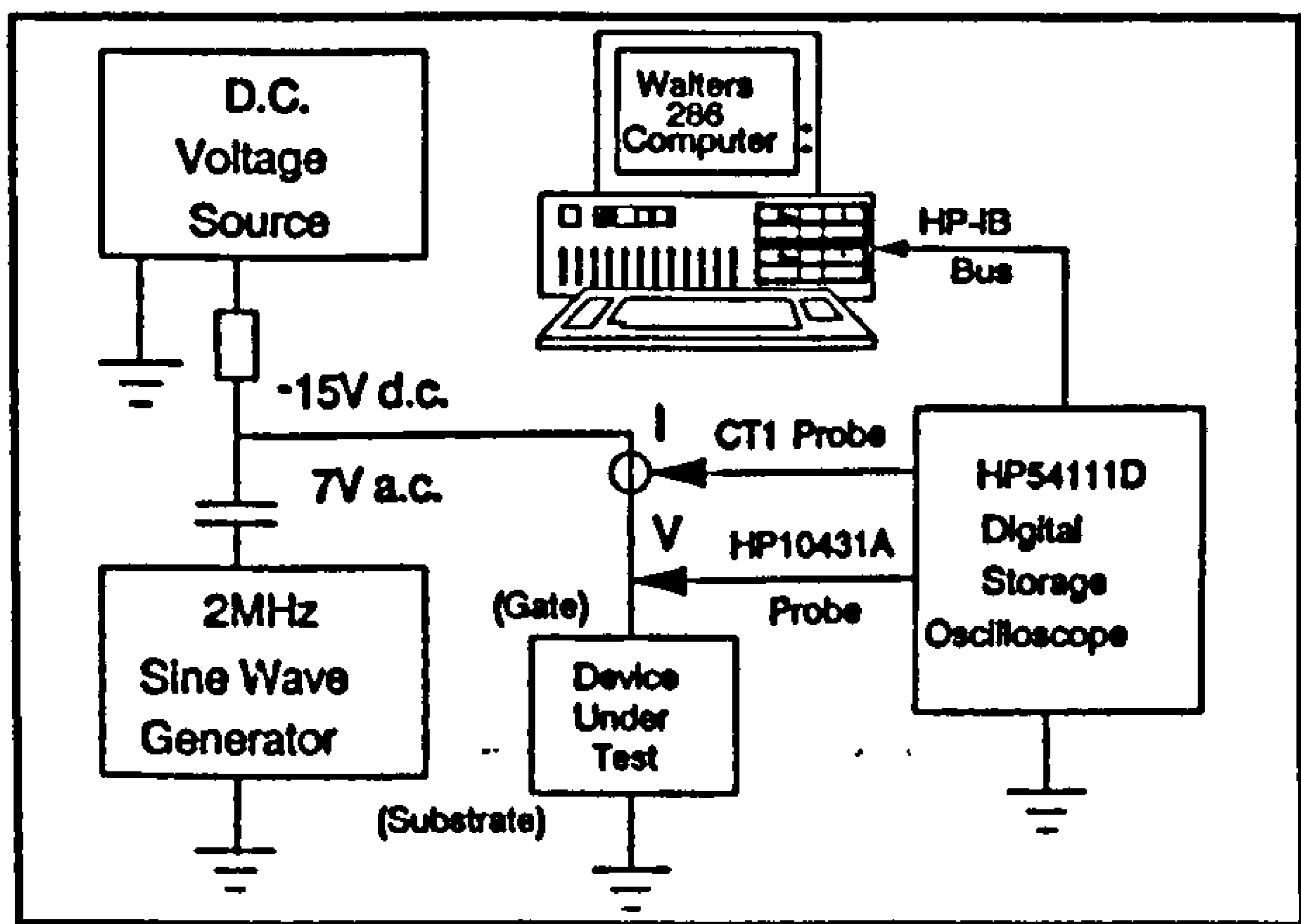


Figure 6: Sample Characterisation System

the oxide. V_b was again monitored by the HP 10440A voltage probe while V_a, the voltage across the whole system, was monitored by an HP 10431A voltage probe. The waveforms of V_a and V_b were captured on the oscilloscope and downloaded to Walters P.C.

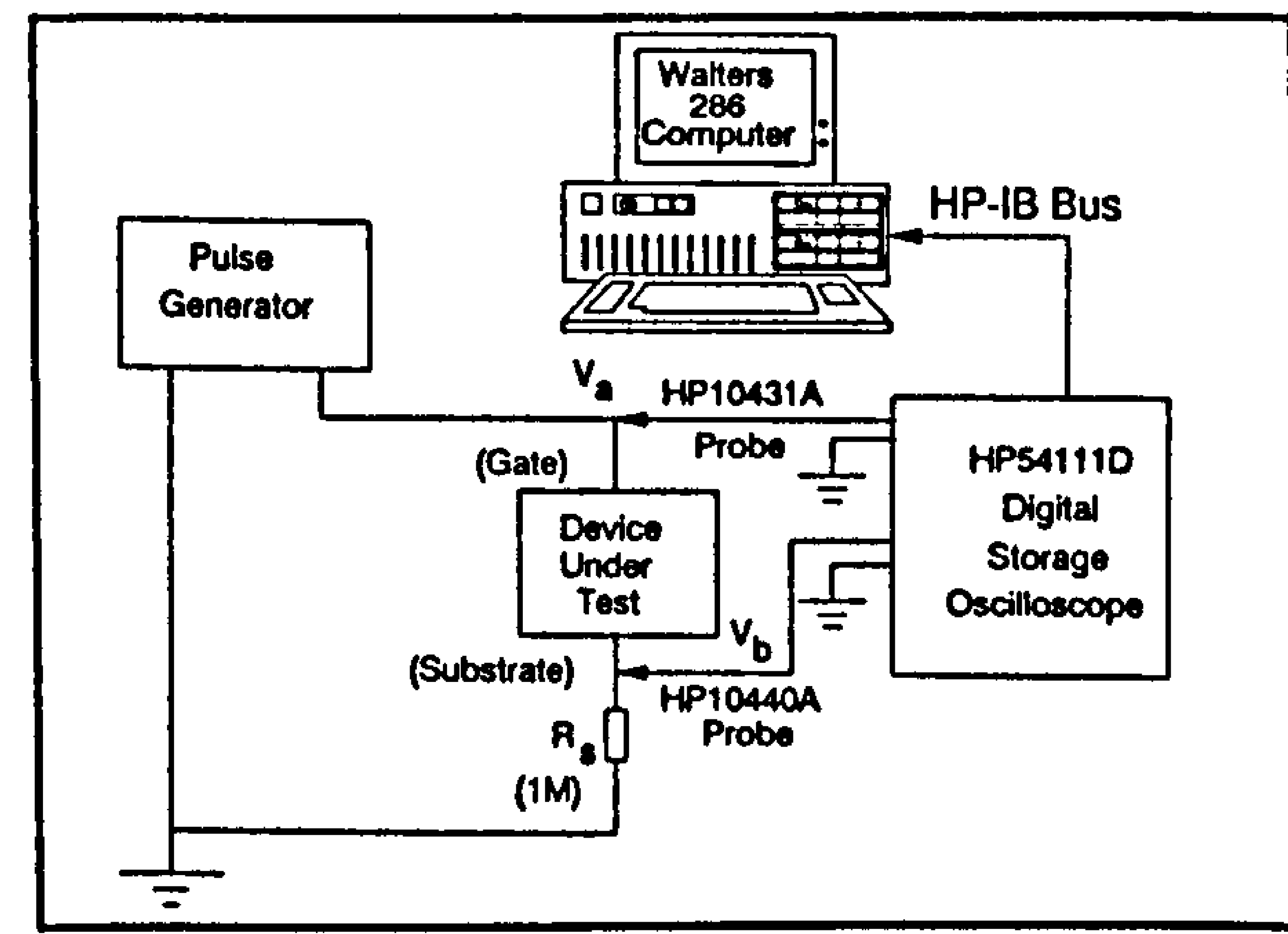


Figure 7: Slow Voltage Transient Apparatus

The following tests were performed:

Transient Speed (V/ms)	Temperature (°C)	Serial Nos.
10	27	Cap.1 & 2
10	140	Cap.3 & 4
100	22	Cap.5 & 6
100	140	Cap.7 & 8
1,000	25	Cap.9 & 10
1,000	140	Cap.11 & 12
10,000	22	Cap.13 & 14
10,000	140	Cap.15 & 16.

The resulting waveforms were captured and recorded.

3.5 ESD Experiments

Fig.8 shows the apparatus used for the ESD experiments. A Hartley AutoZap ESD generator produced a negative HBM ESD pulse which was applied to the device under test via a 1m length of 50Ω co-axial cable. The voltage V_{ox} across the capacitor under test was monitored by the HP10431A voltage probe and the current I was monitored by the CT1 probe. The current and voltage waveforms were again captured on the oscilloscope and downloaded on to the P.C. for numerical interpretation.

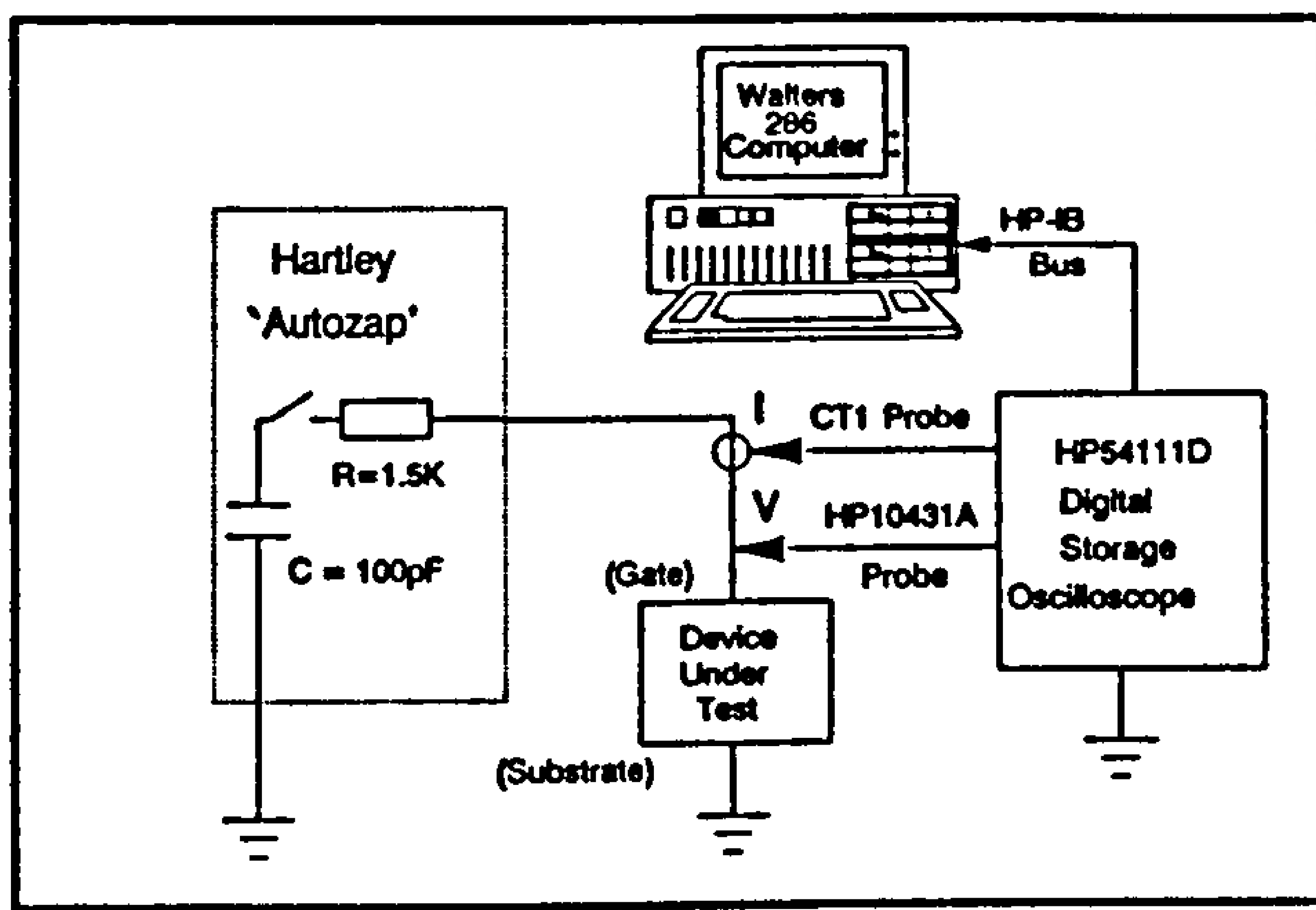


Figure 8: ESD Apparatus

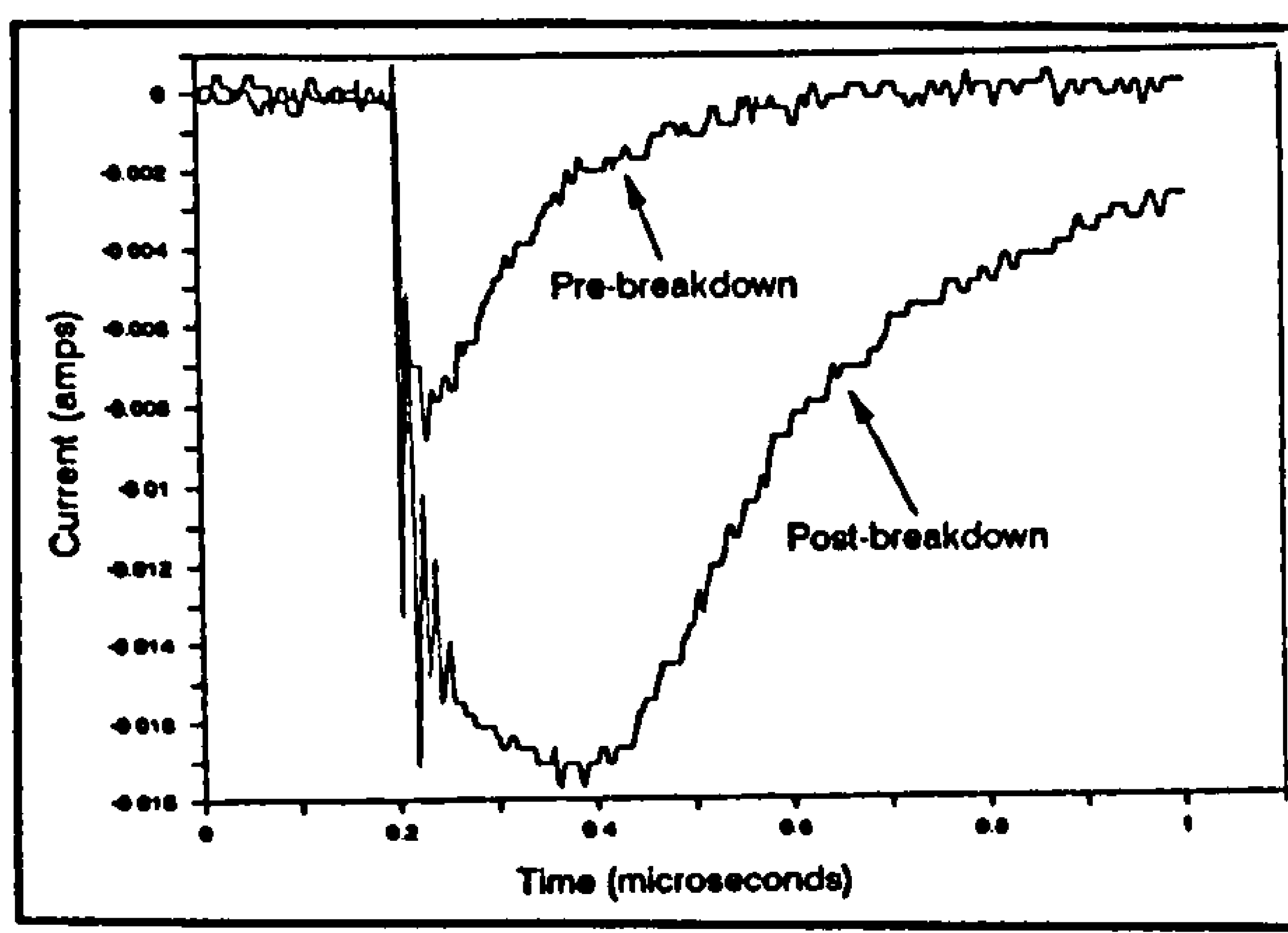


Figure 9: ESD current waveforms before and after breakdown.

Two types of ESD experiment were preformed:

(i) ESD Transient Experiment:
With the chuck at room temperature, capacitor 17 was subjected to a -104V ESD pulse and capacitor 18 to a -108V ESD pulse. The resulting waveforms were captured and recorded.

(ii) ESD Breakdown Threshold Experiment:
The ESD breakdown voltage threshold for capacitors 19, 20 and 21 were determined in the following manner. The samples were subjected to sequences of ESD pulses, starting at V_o = -4V and increasing in magnitude in steps of 4V until breakdown was observed. Breakdown was indicated by a distinctive change in the shape of the current waveform (see Fig.9). To prevent any disturbance to the experiment due to long-term charge leakage via the 1MΩ voltage probe resistance, the HP10431A probe was disconnected during these experiments.

4. Results

4.1 'Slow' Transient Results

The waveforms obtained from the experiments described in Section 3.3 are shown in Figs.10 to 13, together with the corresponding capacitor serial numbers and wafer temperatures.

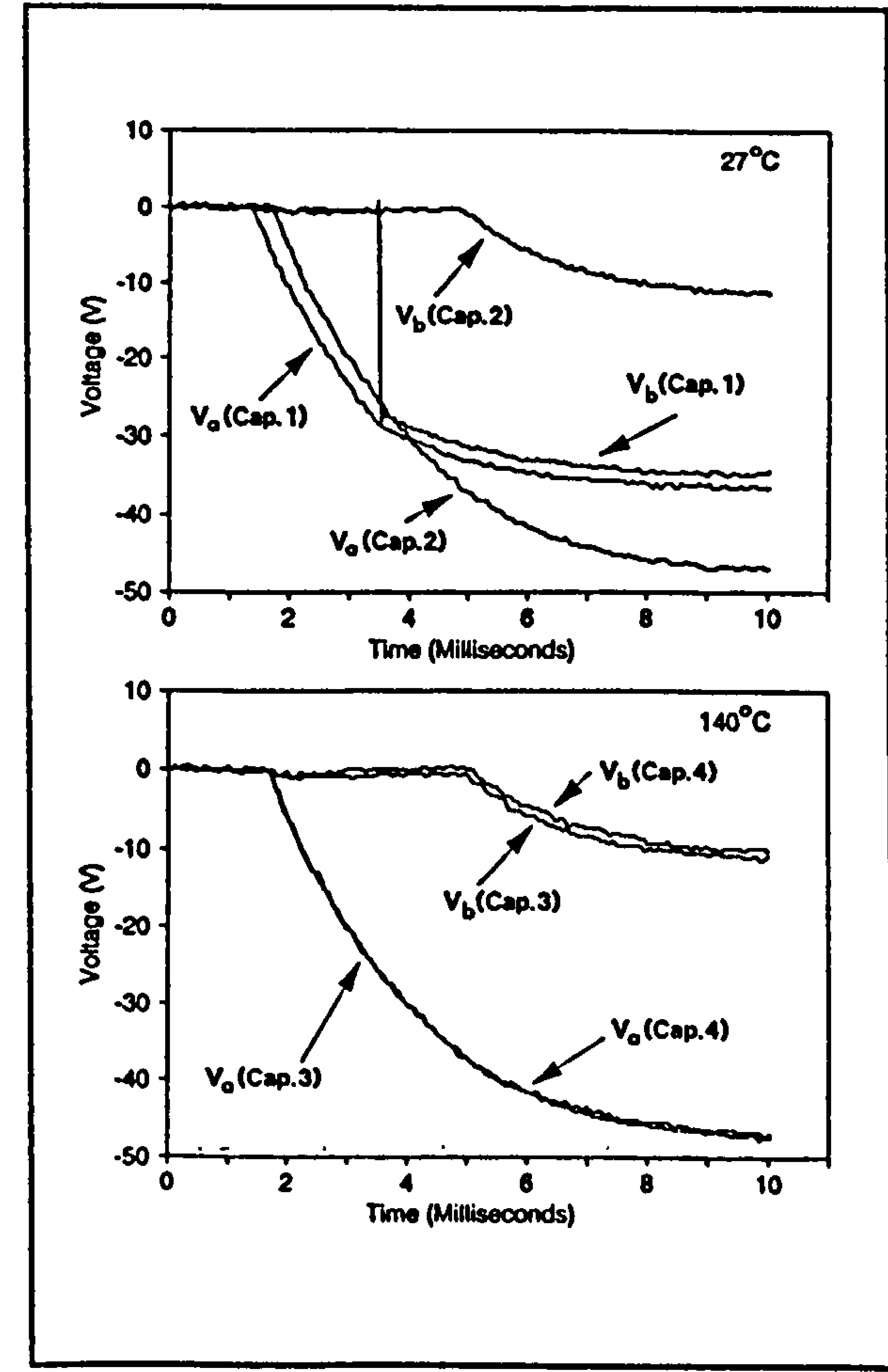


Figure 10: 10V/ms waveforms

4.2 ESD Transient Results

Fig. 14 shows the waveforms obtained from the ESD transient experiments, together with the corresponding capacitor serial numbers.

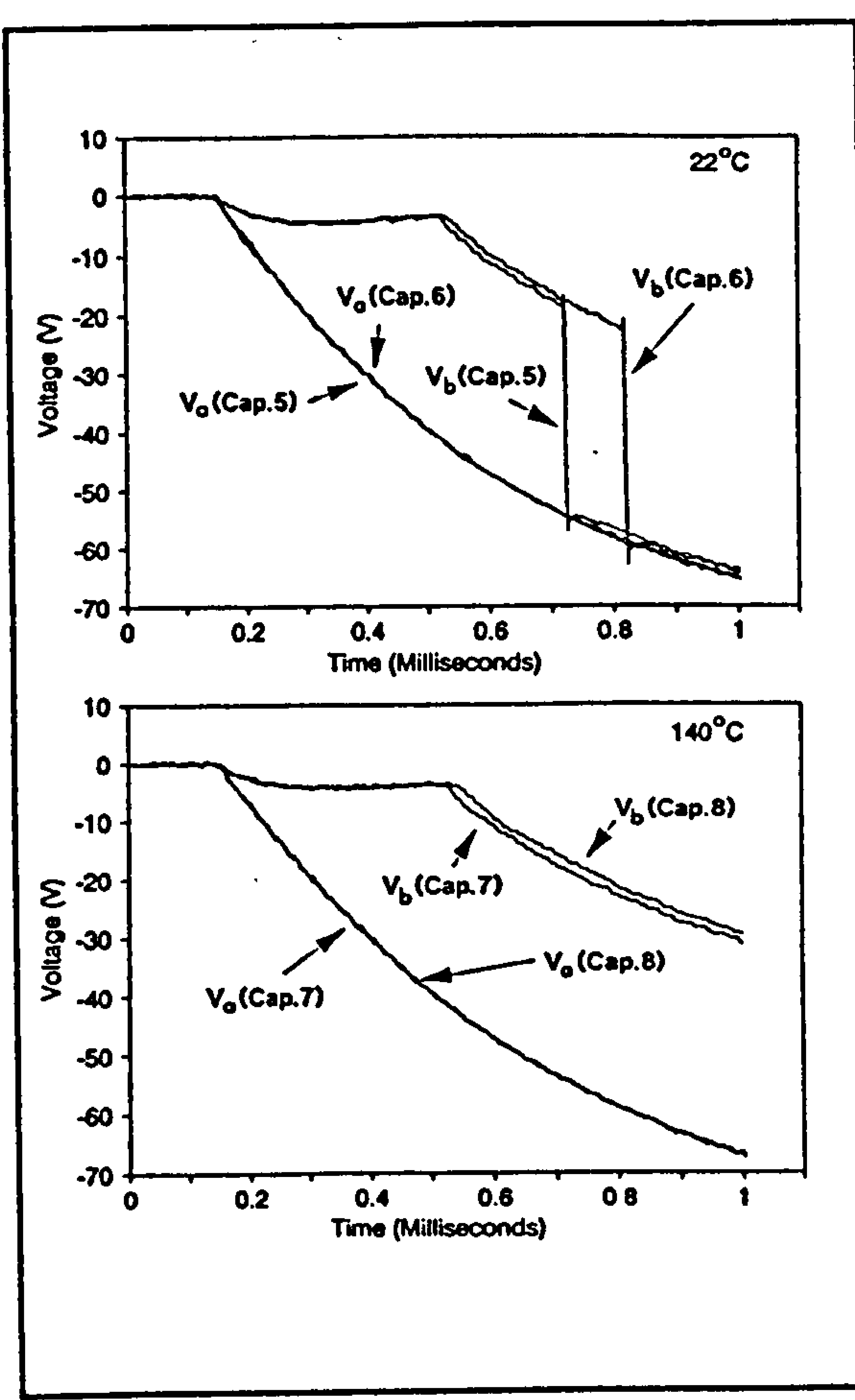


Figure 11: 100V/ms waveforms

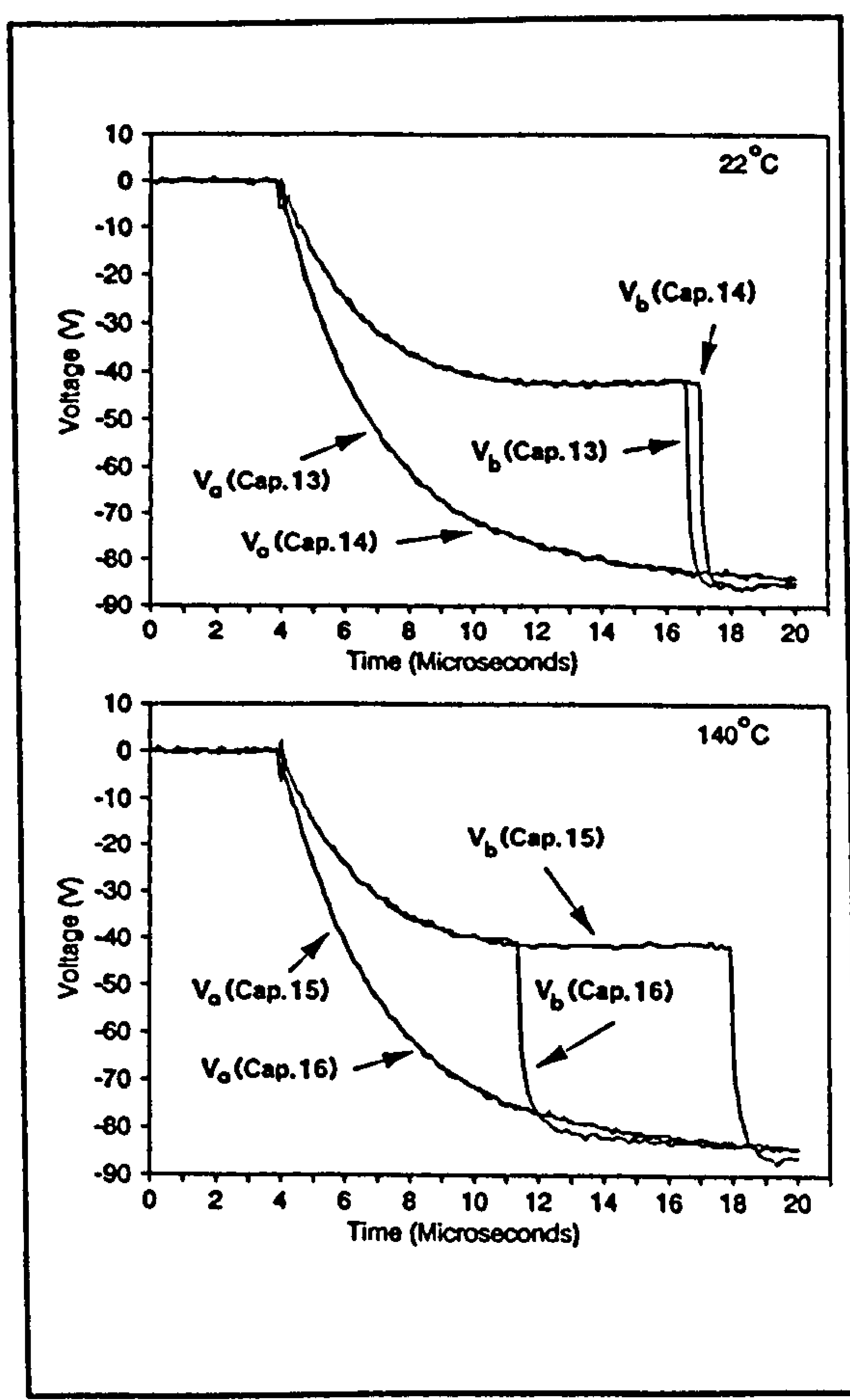


Figure 13: 10,000V/ms waveforms

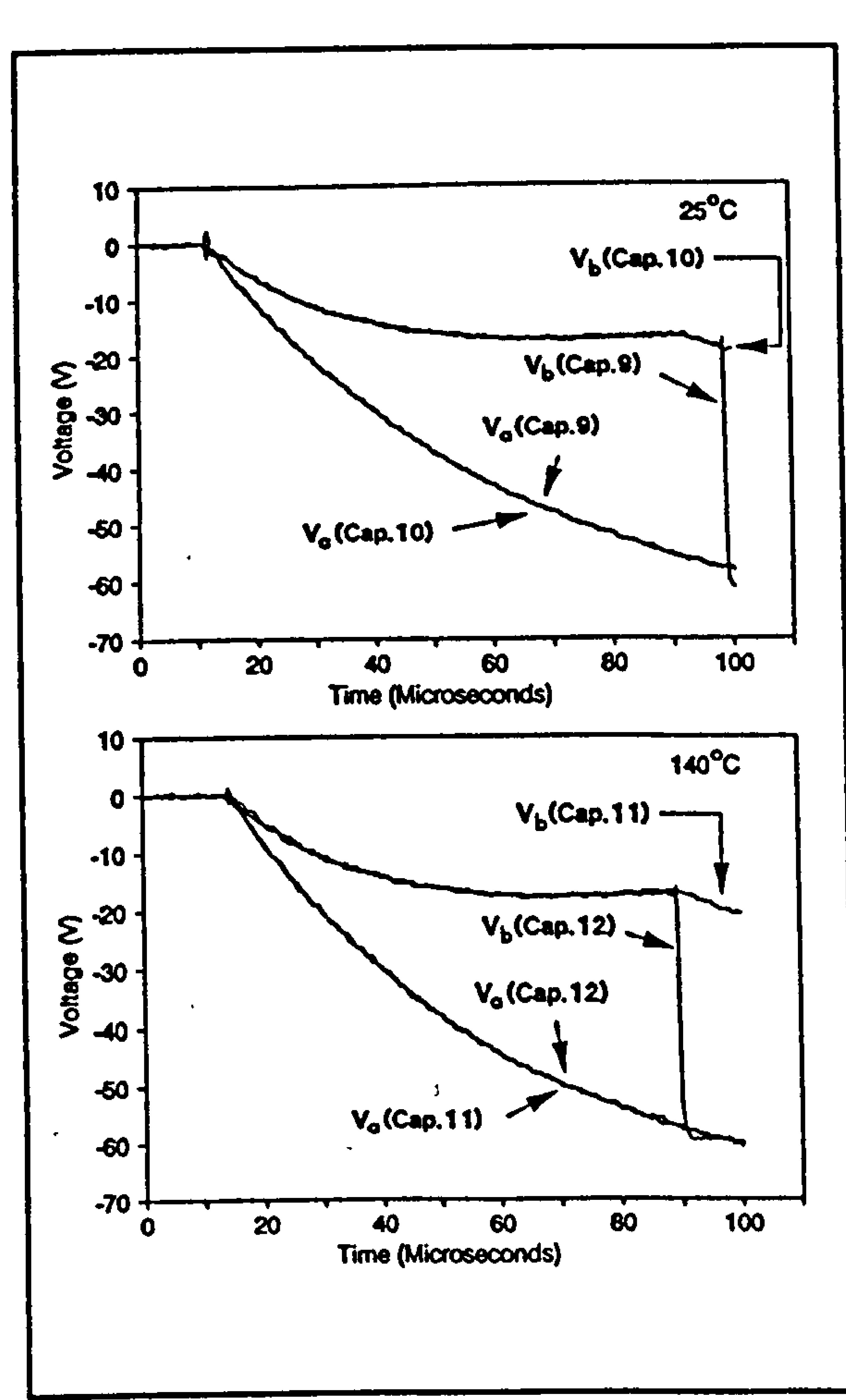


Figure 12: 1,000V/ms waveforms

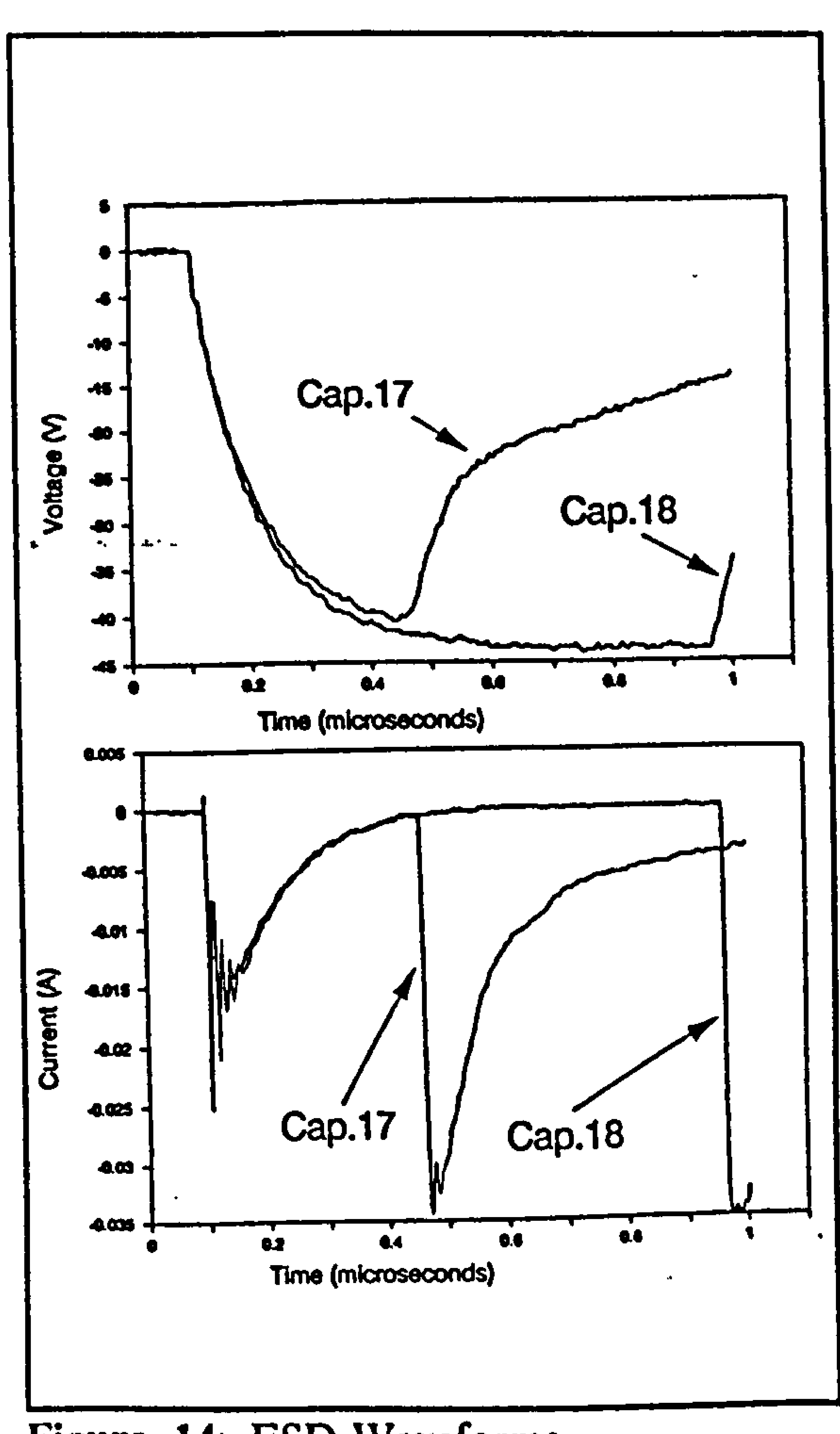


Figure 14: ESD Waveforms

4.3 ESD Breakdown Threshold Results

Capacitors 19, 20 and 21 broke down at V_o =-92V.

5. Interpretation of the Results

The results in Section 4 require some numerical processing before they can be compared with the theory in Section 2. This section describes the processing techniques and shows how the relevant data was extracted.

5.1 Slow Transients

The circuit model of Fig.15 was used in the interpretation of the slow transient data. R_{cg} represents the combined effect of R_s in parallel with the HP10440A probe resistance. The combination of C_{cg} and C_{pc} replace the capacitances of the voltage probe, the flying leads and the chuck/microprober system, Fig. 15. When the oxide is in-circuit, C_{ox} swamps C_{pc} and C_{pc} is therefore neglected in the analysis. The bulk resistance R_b has little effect upon these experiments and is therefore not included in the model.

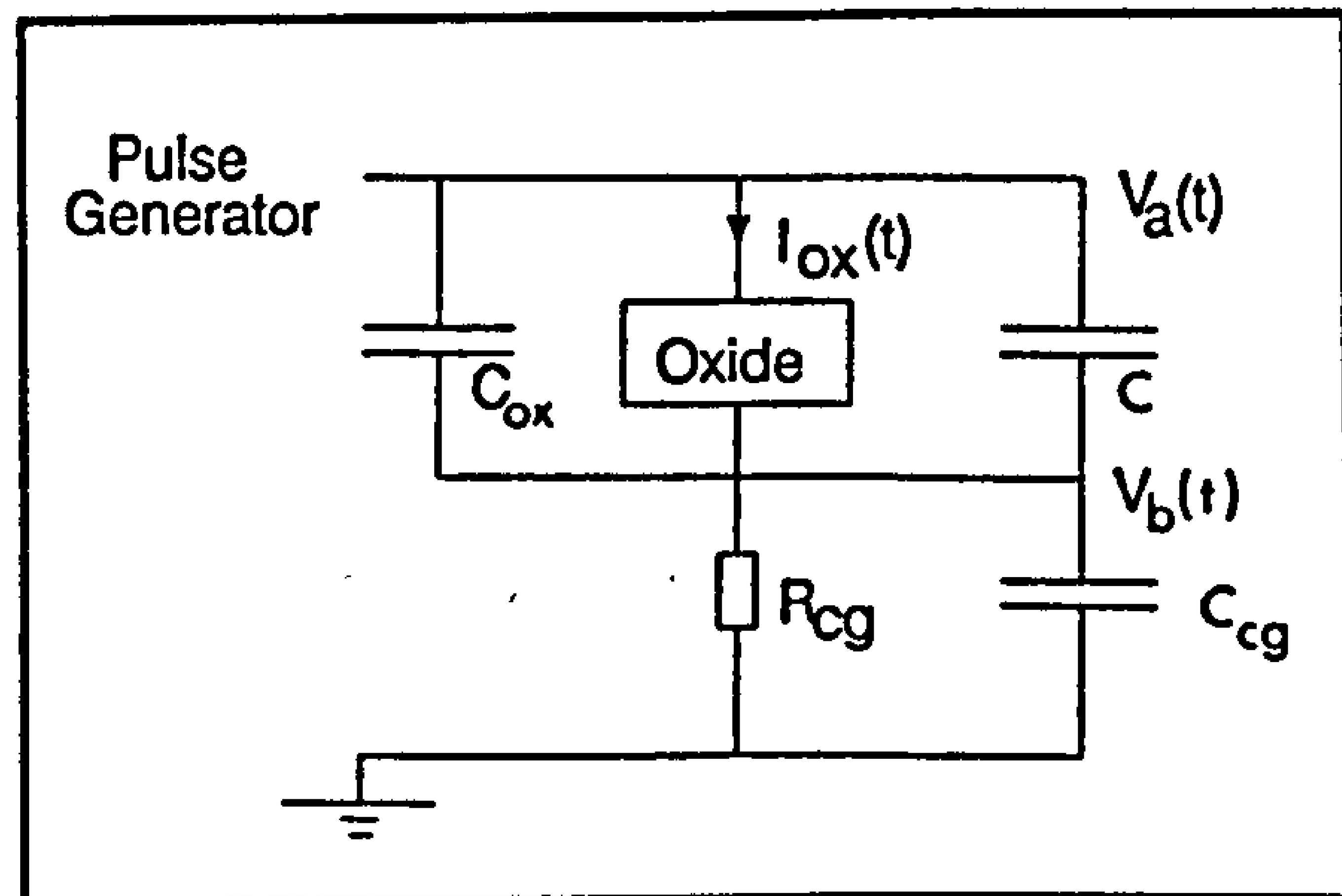


Figure 15: Circuit model of slow transient apparatus

The oxide voltage V_{ox} is clearly given by

$$V_{ox} = V_a - V_b \tag{8}$$

and the injected charge Q_{ox} ($=A.\int J_{ox}.dt$) is given by

$$Q_{ox} = \frac{1}{R_{cg}}\int_0^t V_b\, dt + C_{cg}V_b - C_{ox}(V_a - V_b) \tag{9}$$

The values of the circuit elements were determined using the portion of the waveforms for which V_{ox} < 35V, since in this region the oxide behaves as an open circuit (ie. Q_{ox}=0).

Setting Q_{ox}=0 into Eqn.(9) and rearranging yields

$$y = -R_{cg}(C_{ox} + C_{cg})\, x + R_{cg} C_{ox} \tag{10}$$

where $x=V_b/V_a$ and $y=\int V_b.dt/V_a$. Values of R_{cg} and C_{cg} were obtained from equation (10) by linear regression and the known value of C_{ox} (= 42.26pF Section 3.3). Figs.16 to 19 show the $Q_{bd}(t)$ profiles
extracted from the experimental data using Eqn.(7).

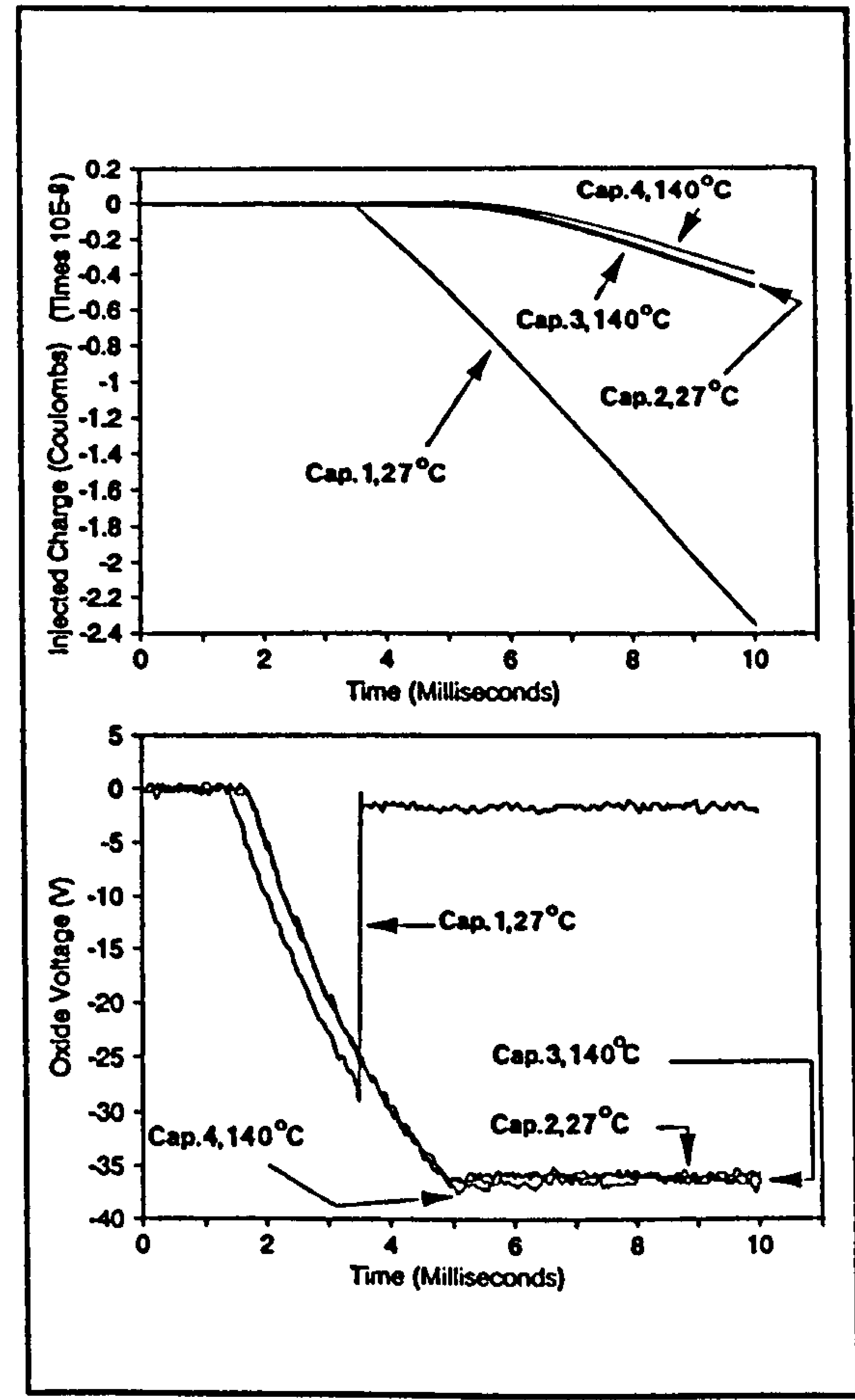


Figure 16: $Q_{ox}(t)$ & $V_{ox}(t)$ for 10V/ms data

5.2 ESD

The circuit model of Fig.20 was used in the analysis of the ESD data. C_x is the internal stray capacitance of the AutoZap discharge module, C_t is the capacitance of the co-axial cable and C_y represents the combined capacitances of the HP10431A probe, the leads and the

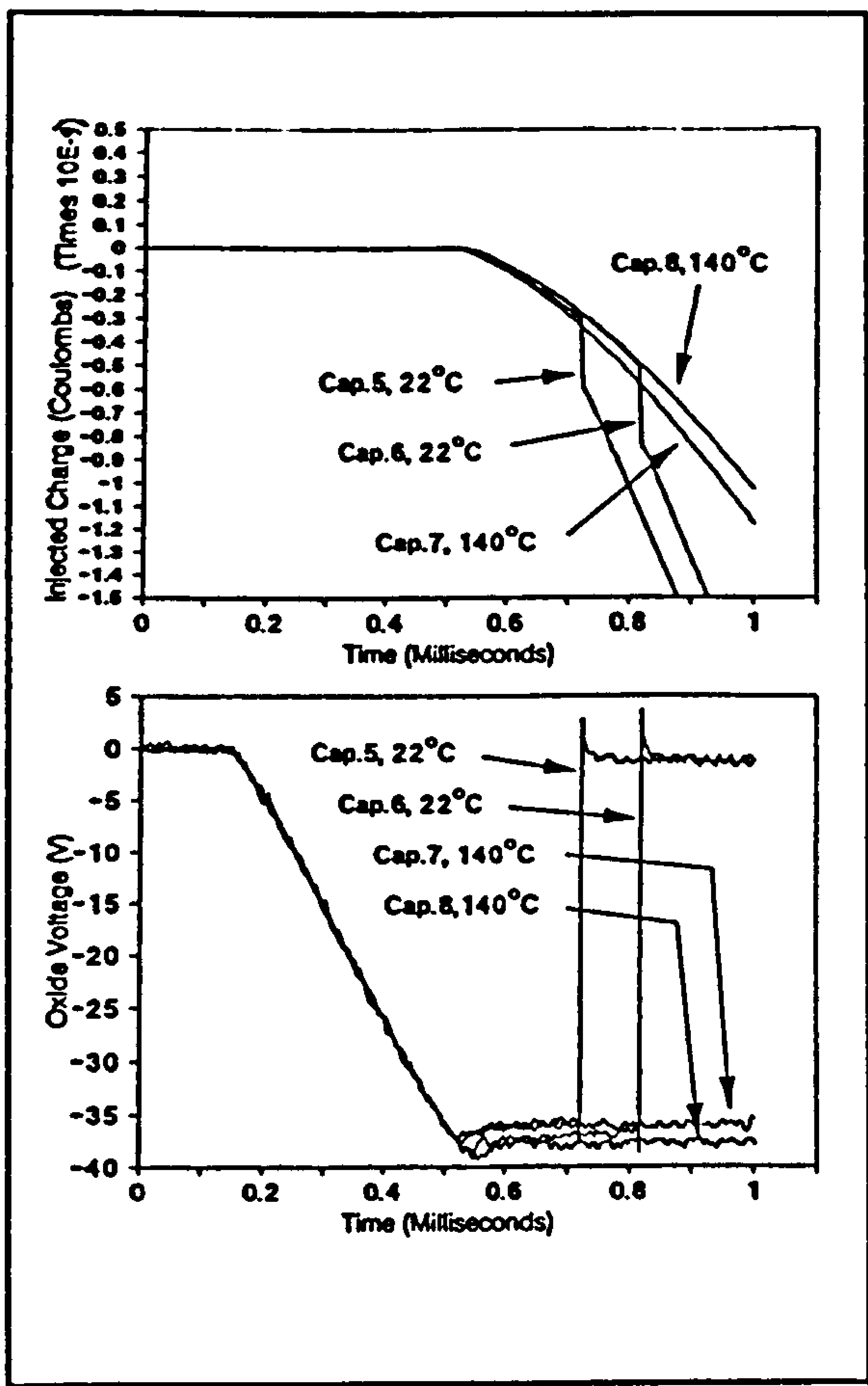


Figure 17: $Q_{ox}(t)$ & $V_{ox}(t)$ for 100V/ms data

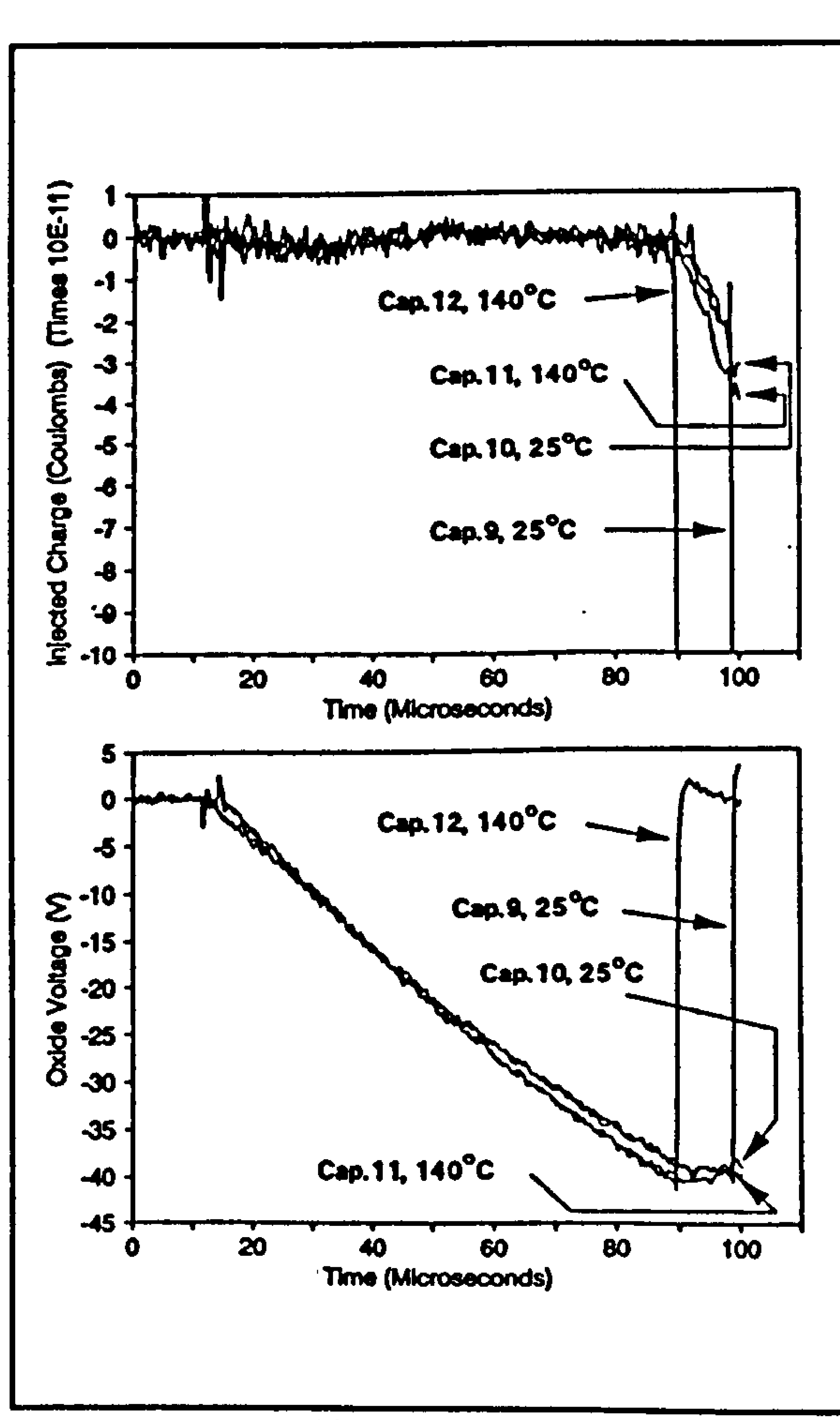


Figure 18: $Q_{ox}(t)$ and $V_{ox}(t)$ for 1,000V/ms data

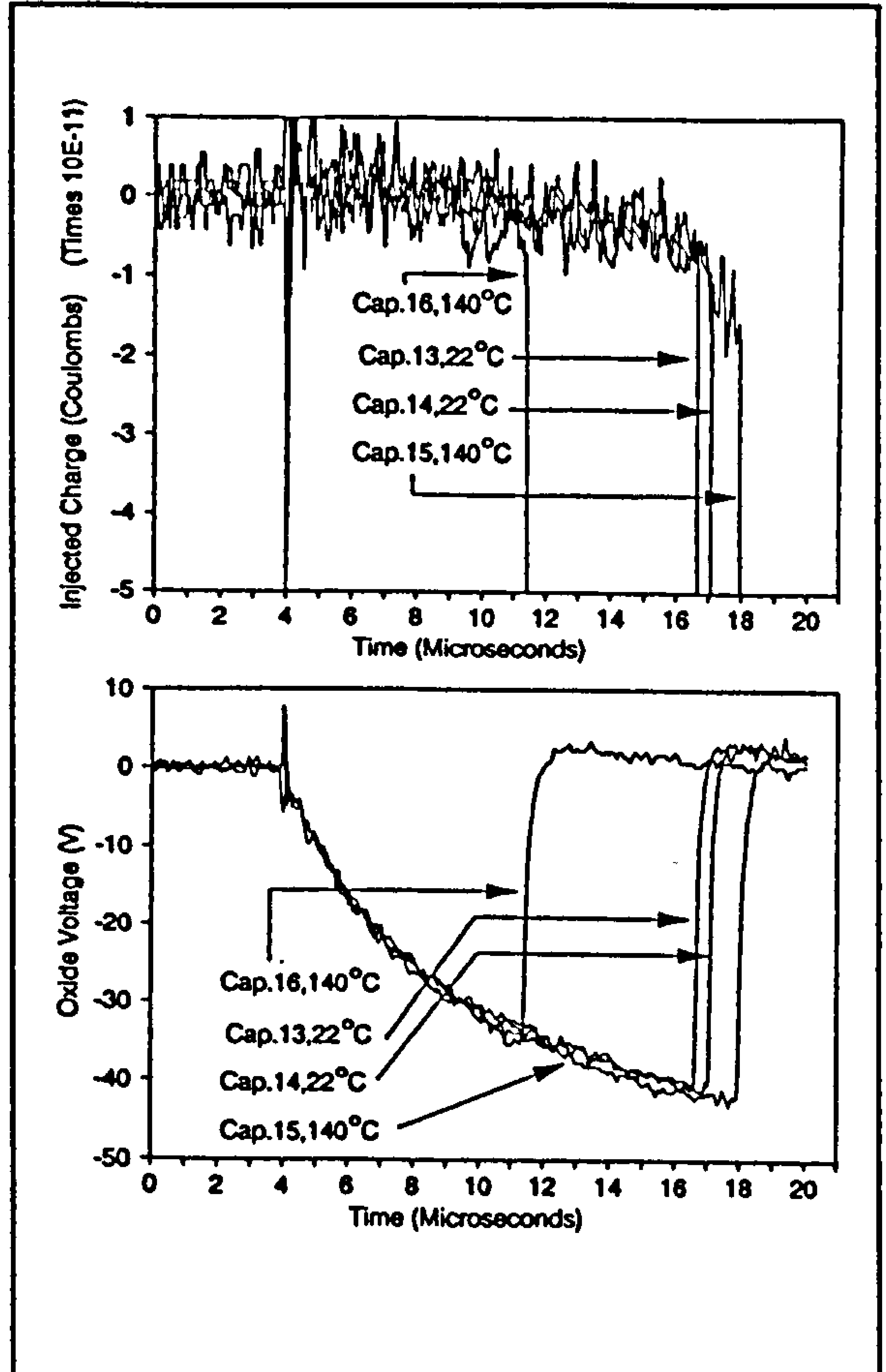


Figure 19: $Q_{ox}(t)$ and $V_{ox}(t)$ for 10,000 V/ms data

microprober/chuck system. The current probe and the voltage probe resistance were not included in the model since they have negligible effect upon the waveforms.

Simple circuit theory shows that the oxide voltage V_{ox} is given by

$$V_{ox} = V - R_b(I - C_y \frac{dV}{dt}) \qquad (11)$$

and the injected charge Q_{ox} is given by

$$Q_{ox} = \int_0^t I\,dt - C_y V - C_{ox} V_{ox} . \qquad (12)$$

The circuit elements were measured using a Wayne-Kerr 4210 LCR bridge and were found to have the following values:

C_x = 1pF C_y = 10pF C_t = 87pF C_1 = 94.9pF

Fig.21 shows the V_{ox} and Q_{ox} profiles reconstructed from the experimental data using Eqns.(11) and (12).

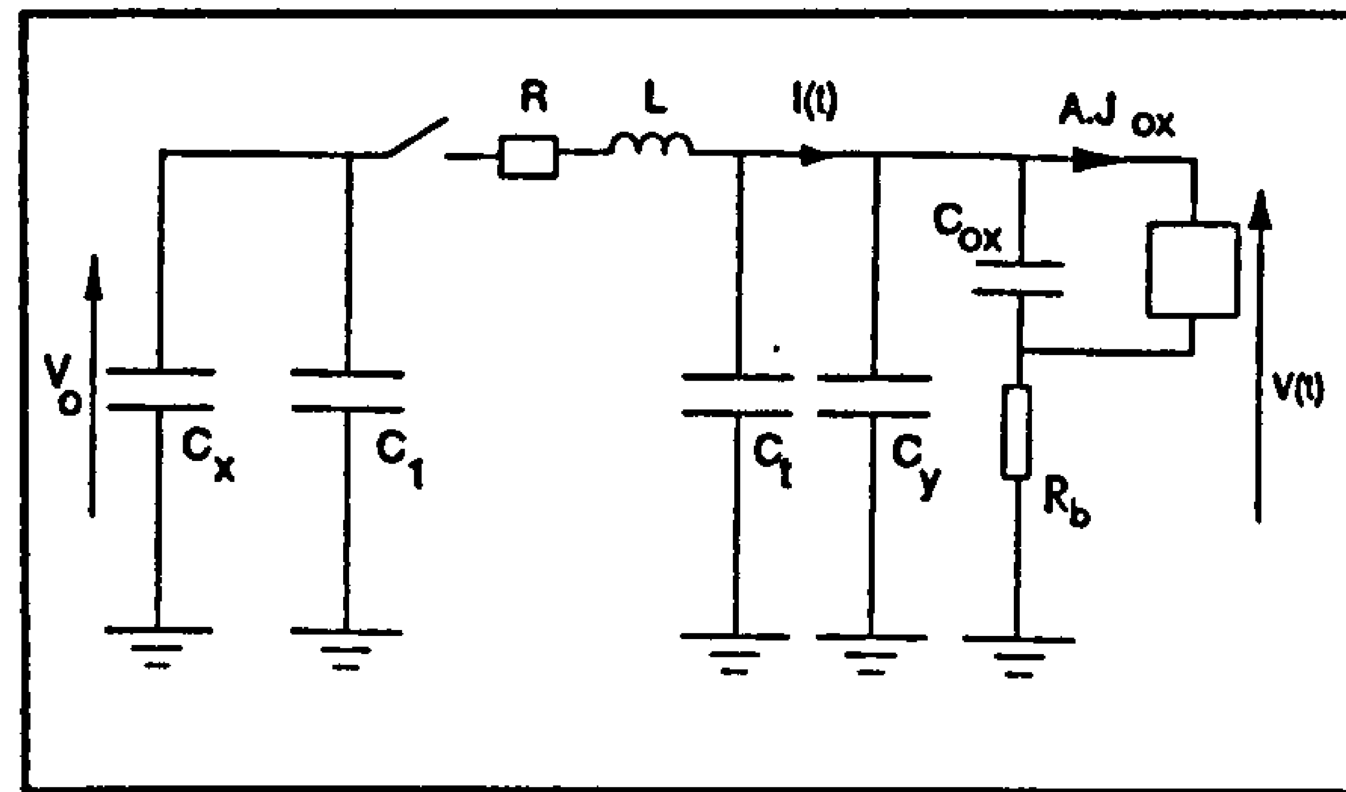


Figure 20: Circuit Model of ESD Apparatus

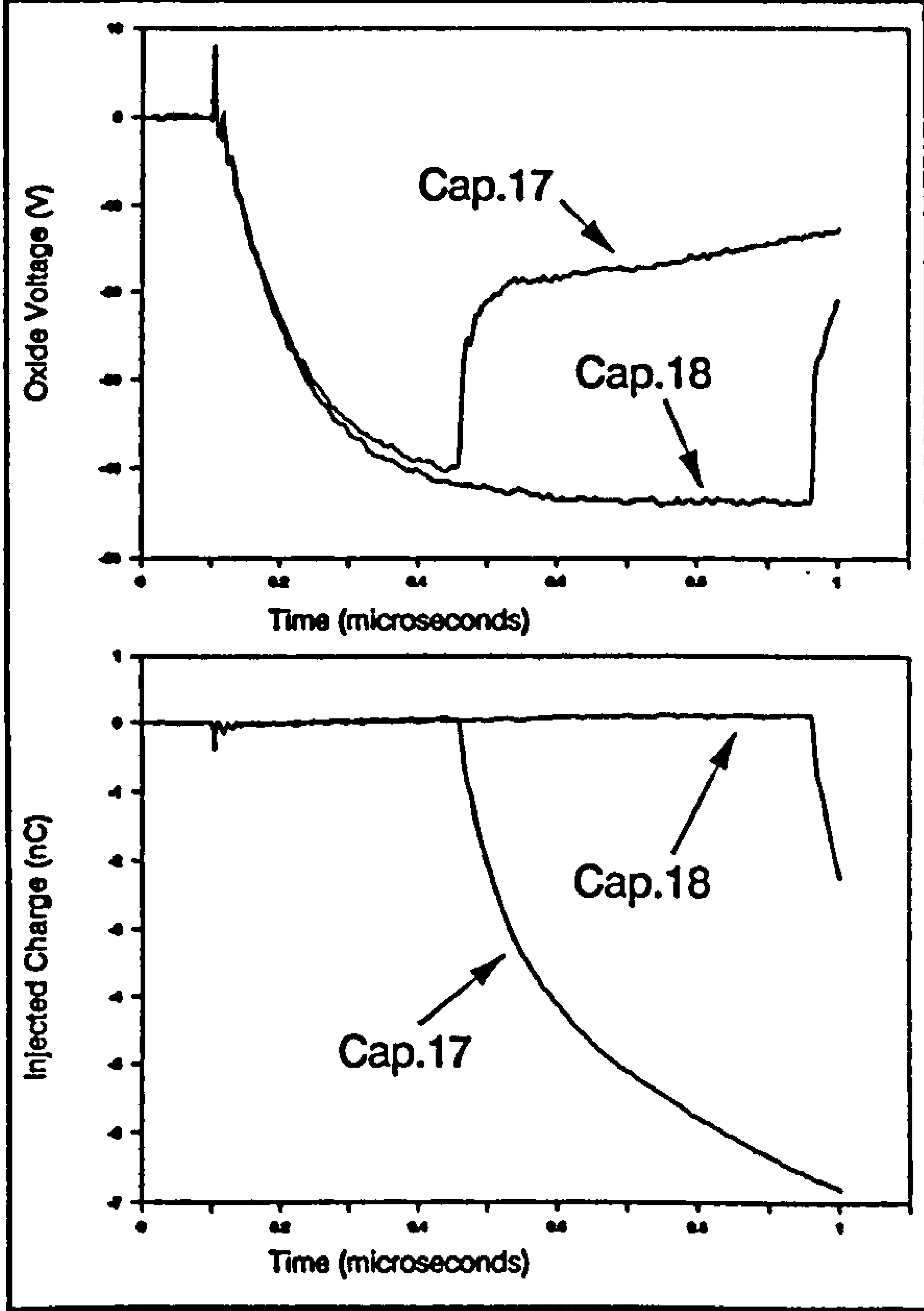


Figure 21: $V_{ox}(t)$ & $Q_{ox}(t)$ profiles for ESD data

6. Discussion

6.1 Slow Transient Results

The waveforms in Figs.16,17 and 18 clearly indicate 'pinning' of the oxide voltage due to tunnelling injection at the oxide cathode. Pinning continues until the onset of dielectric breakdown, indicated by a rapid increase in Q_{ox} and a rapid decrease in V_{ox}. The value of Q_{ox} just prior to breakdown defines Q_{bd}.

As the speed of the voltage transient is increased (Figs.16-19), the voltage magnitude at which V_{ox} pins increases, causing a variation in Q_{bd}. It is therefore possible to examine the Q_{bd} vs. V_{ox} relationship for the oxide. For example, when $|V_{ox}|$ = 36V (Capacitor 6, Fig.17), $|Q_{bd}|$ = 3 nC. While, when $|V_{ox}|$ is increased to 40V (Capacitor 11, Fig.18), $|Q_{bd}|$ falls to 2nC. Although Q_{bd} and V_{ox} vary between samples, the general trend indicates a steeper dQ_{bd}/dV_{ox} than is predicted by the theory in Section 2 (see Fig.3). Two of the samples tested (Capacitors 1 and 16) seem to require substantially less voltage and charge for breakdown than the others. These samples probably had major defects in their oxides.

It is worth noting that the voltage pinning, and hence the limitation of Q_{bd} is caused by the same mechanism which supports breakdown and hence the oxide can be visualised as acting as its own protection diode.

6.2 ESD Results

The fast transient speed of an ESD pulse ($\sim$1,000V/μs), together with the low resistance of the ESD source (1.5KΩ) cause V_{ox} to be forced higher than in the case of a slow transient pulse. Fig.21 shows that V_{ox} is forced up to about 43V, causing breakdown with a Q_{bd}<<1nC. This is consistent with the steepness of the Q_{bd} vs. V_{ox} curve discussed in Section 6.1 above. These results suggest that under ESD conditions, the single criterion for breakdown is that V_{ox} should exceed a certain critical value V^*. Hence, according to the circuit model of Fig.20 [see also Eqn.(6)], the breakdown voltage threshold V_{bd} should be given by

$$V_{bd} = \frac{V^*(C_x + C_1 + C_t + C_{ox})}{C_x + C_1} \quad (13)$$

The value of V_{bd} predicted by Eqn.(13) using V^*=-40V is -93.9V which correlates well with the experimental value of -92V. There is clearly an excellent correlation between theory and experiment.

6.3 The Effects of Temperature

There is very little difference between the voltage/current relationships measured at 27 and 140°C. This is consistent with the Fowler-Nordheim equation which predicts very little variation of tunnelling current/field relationship with temperature [8]. This indicates that the temperature dependence observed by Amerasekera and Campbell [4] occurs only for very long time scale stressing (>10ms) and that the breakdown observed in the slow transient tests is due to the same temperature independent mechanism as ESD breakdown [6].

7. Conclusions

Experimental studies have been conducted into dielectric breakdown in MOS oxides under both slow voltage transient and ESD stress. The experimental techniques have been described and the results have been presented.

The results indicate that the Q_{bd} vs. V_{ox} curve is extremely steep in the 0-30nC region, causing the oxide voltage required for ESD breakdown to be approximately constant. This observation was used to derive an accurate expression for the ESD breakdown voltage threshold of an MOS device.

The effects of varying temperature upon breakdown behaviour were also studied, but no trends were observed.

Acknowledgements

This work was being carried out with the support of the Procurement Executive, Ministry of Defence, and we are grateful to them for permission to publish this paper. The authors would like to thank Mr. D. Shore for many valuable discussions, and other members of the Component Technology group for help with typing the manuscript.

8. References

1. Wolters,D.R., van der Schoot,J.J., "Dielectric Breakdown in MOS Devices (Parts I,II & III)", Phillips Journal of Research, 40, pp.115-92, 1985.

2. Chen,I-C, Holland,S, Hu,C, "Electrical Breakdown in Thin Gate and Tunneling Oxides", IEEE Transactions on Electron Devices, ED-32, pp.413-22, 1985.

3. Avni,E, Shappir,J, "A Model for Silicon Dioxide Breakdown under High Field and Current Stress", Journal of Applied Physics, 64, pp.743-8, 1988.

4. Amerasekera,E.A., Campbell,D.S., "Oxide Breakdown in MOS Structures under ESD and Continuous Voltage Stress Conditions", Proc. European Conference on Reliability, pp.325-330, Copenhagen, 1986.

5. Fong,Y, Hu,C, "The Effects of High Electric Field Transients on Thin Gate Oxide MOSFETs", 1987 EOS/ESD Symposium Proceedings, pp.252-7, 1987.

6. Amerasekera,E.A.,Campbell,D.S.,"Electrostatic Pulse Breakdown in NMOS Devices", Quality and Reliability Engineering International, 2, pp.107-116, 1986.

7. Amerasekera,E.A., Campbell,D.S., "Charge Limited Breakdown in MOS Capacitors", Proc. 17th. Annual European Solid State Device Research Conference, pp.733-41, 1987.

8. O'Dwyer,J.J., "The Theory of Electrical Conduction and Breakdown in Solid Dielectrics", Clarendon Press, Oxford, pp.77-94, 1973.

Paper II

Tunnicliffe,M.J., Dwyer,V.M., Campbell,D.S.,

"Experimental and Theoretical Studies of EOS/ESD Oxide Breakdown in Unprotected MOS Structures",

Proc. 12th. Electrical Overstress/Electrostatic Discharge Symposium Lake Buena Vista, 11th.-13th. September, 1990,

pp.162-8, 1990.

Experimental and Theoretical Studies of EOS/ESD Oxide Breakdown in Unprotected MOS Structures

by

M.J.Tunnicliffe, V.M.Dwyer and D.S.Campbell

Electronic Component Group,
Dept. of Electronic and Electrical Engineering,
Loughborough University of Technology,
Loughborough, Leicestershire, LE11 3TU, U.K.
Tel. Inside U.K.: 0509 222855
Outside U.K.: 44 509 222855

Abstract

A study of oxide breakdown under EOS (continuous, ramped and pulsed voltage) and ESD in a variety of MOS devices is presented. The results are used to develop a more accurate version of the 'causality' theory for oxide breakdown. This theory is used in the development of a model for negative polarity ESD breakdown. The major conclusion of this model is that ESD breakdown is almost entirely dependent upon the oxide field. It is then shown (both experimentally and theoretically) that positive polarity breakdown in a p-type structure can only be supported by avalanche conduction in the deeply depleted silicon surface. This fact is used to extend the negative polarity ESD model to predict positive polarity breakdown.

1. Introduction

The thin gate oxides of MOS devices are highly susceptible to dielectric breakdown under EOS and ESD stress conditions. Over recent years, numerous attempts have been made to identify the causes of electrical breakdown in thin SiO_2 layers. Experiments indicate that ESD breakdown is caused by the injection of charge from the cathode surface into the oxide under Fowler-Nordheim tunnelling [eg.1,2,3]. The voltage threshold for breakdown under very slow voltage ramp conditions has been shown to decrease with increasing temperature [4], whilst under faster transient conditions (including ESD) it appears to be temperature independent [5,6]. Breakdown behaviour also seems to be affected by the type and density of the doping in the cathode material [7].

At present there is no simple model which explains all the observed EOS and ESD breakdown behaviour. The purpose of this paper is to present the results of a range of EOS and ESD experiments on MOS devices and develop a theoretical framework to link all observed phenomena to a common mechanism.

2. Fowler-Nordheim Tunnelling in MOS structures

Under high electric field F, the electron tunnelling current density J_{ox} from the cathode surface into the oxide is governed by the Fowler-Nordheim equation [8]

$$J_{ox} = \frac{q^3 m_{si}^* F^2}{8 \pi h \phi m_{ox}^* t^2} \exp\left[-\frac{8 \pi \sqrt{2 m_{ox}^*}\, \phi^{3/2} v}{3 h q F}\right] \qquad (1)$$

where q is electron charge, h is Plank's constant, ϕ is the zero-field Si-SiO_2 barrier height m_{ox}^* and m_{si}^* are the effective electron masses in SiO_2 and silicon respectively and v and t are tabulated functions which can be approximated as

$$t(F) = 1.07 \qquad (2)$$

$$v(F) = 0.929 - 4.573.10^{-8} F \qquad (3)$$

in the region of F=10MV/cm. Using these approximations, Eqn.(1) can be written

$$J_{ox} = k F^2 e^{-\frac{B}{F}} \qquad (4)$$

where $B = 0.929[8\pi(2m_{ox}^*)^{½}\phi^{3/2}/3hq]$ and
$k = [q^3 m_{si}^*/8\pi h m_{ox}^*]\exp[-8\pi(2m_{ox}^*)^{½}\phi^{3/2}(4.523.10^{-8})/3hq]/(1.07)^2$. If an electron's effective mass in SiO_2 is equal to its rest mass then k=27.269A/V^2 and B=364.4MV/cm.

Since the exponential factor in Eqn.(4) varies with F much faster than the F^2, a further approximation can be made:

$$J_{ox} = J_0 e^{-\frac{B}{F}} \qquad (5)$$

where J_0 is a constant, approximately equal to 2.461.10^{15}A for fields in the 9.5MV/cm region.

3. Theory of Oxide Breakdown

Numerous physical models have been developed to explain dielectric breakdown in SiO_2. Some theories [eg.1,3,9] maintain that electron traps in the bulk oxide are responsible for breakdown. Others [eg.2] claim that breakdown is supported by trapped holes close to the cathode/oxide surface.

It is, however, generally agreed that the process is not instantaneous, but involves a cumulative 'wearout' of the dielectric during tunnelling injection. This has been described as the 'causality' theory.

Breakdown can be characterised in two ways:

a. Time-to-breakdown (t_{bd})

According to the general causality model, t_{bd}, the time delay till breakdown is related to the stress field profile F(t) by the equation [12]

$$\int_0^{t_{bd}} \frac{dt}{\tau(F(t))} = 1 \qquad (6)$$

where τ is a function of F(t). If F remains constant with time then Eqn.(1) can be written

$$t_{bd} = \tau(F) \qquad (7)$$

Thus $\tau(F)$ may be seen as the time-to-breakdown for a constant field F. Several forms have been given to the function $\tau(F)$. One of the most accurate and theoretically plausible is [10,11]

$$\tau(F) = \tau_0 e^{\frac{\gamma}{F}} \qquad (8)$$

where τ_0 and γ are constants. Fig.1 shows the above model fitted to the constant field experimental data from Refs.10 and 11.

Over a narrow field range (approx. 13-18 MV/cm), the model appears to be accurate whilst at low fields (to the right of Fig.(1)), t_{bd} is shorter than expected. This may be due to electronic charge, trapped during the lengthly stressing period, enhancing the internal oxide field and accelerating the breakdown process. At high fields, t_{bd} is longer than expected. Possible reasons for this will be discussed later.

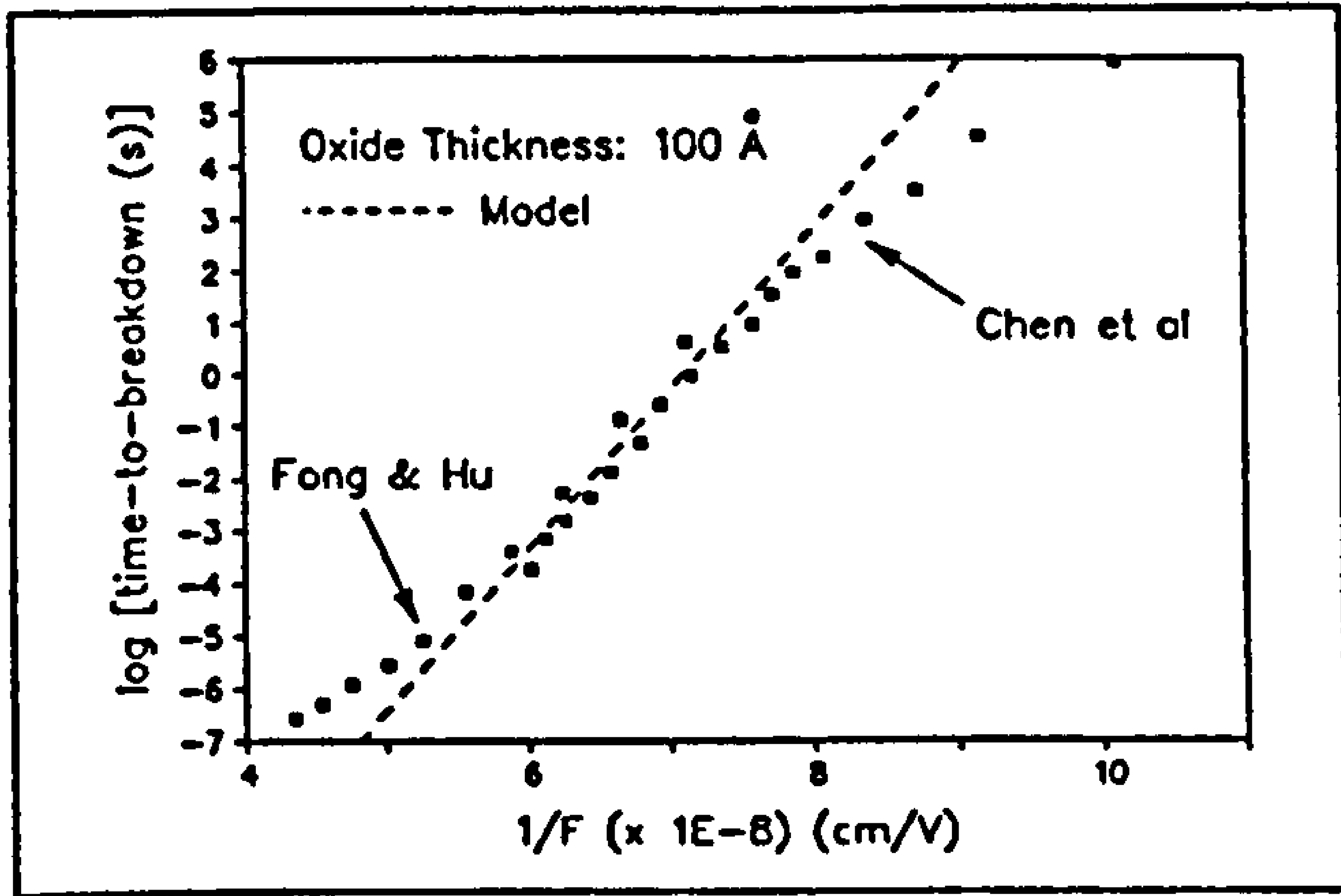


Figure 1: log t_{bd} vs. 1/Field data (after Chen et al.[10] & Fong & Hu [11]).

b. Charge-to-breakdown (Q_{bd})

Breakdown can also be characterised in terms of the 'charge-to-breakdown' or Q_{bd}, ie. the total charge density injected into the oxide prior to breakdown given by

$$Q_{bd} = \int_0^{t_{bd}} J_{ox}\, dt \tag{9}$$

If the field F remains constant with time, Eqns.(9) and (5) can be combined to yield

$$Q_{bd} = J_{ox} t_{bd} = Q_0 e^{\frac{H}{F}} \tag{10}$$

where $Q_0 = J_0 \tau_0$ and $H = \gamma - B$. According to the model of Chen, Holland and Hu [2], H represents the exponential factor governing impact ionisation in the SiO_2 and is therefore independent of B. Hence a plot of Q_{bd} vs. field should provide a characterisation of breakdown independent of any space-charge induced variations in B [10]. Q_{bd} is also an important parameter in the analysis of ESD breakdown, where the charge available to support breakdown is limited to the charge stored in the 'human body' capacitance [6].

Oxide breakdown under ESD stress exhibits none of the 'causal' nature discussed above but appears to be field dependent and time independent. This led earlier workers to believe that ESD breakdown is due to a different mechanism from continuous voltage/current (EOS) breakdown [4]. This paper presents the alternative view and demonstrates that the causality theory is consistent with the observed ESD phenomena.

4. Test Samples

The following test samples were used in the experiments detailed below:

p-type MOS Capacitors,

1 device/array (see Fig.2).

Type 1: gate area: $49.10^{-5}cm^2$, polysilicon gate doping: $10^{21}cm^{-3}$ (n-type), substrate doping: $5.10^{14}cm^{-3}$ (p-type), oxide thickness: 400Å.

NMOS Transistor Arrays,

10 devices/array, common gate & source, polysilicon gate doping 10^{21} cm^{-3} (n-type), channel doping $4.10^{15}cm^{-3}$ (p-type), source/drain implant doping: 2.10^{20} cm^{-3}, oxide thickness 400Å (see Fig.3).

Type 2:	Total gate area 51.07μm^2
Type 3:	Total gate area 380.95μm^2
Type 4:	Total gate area 78.6μm^2
Type 5:	Total gate area 544.6μm^2
Type 6:	Total gate area 589.42μm^2

Devices were stressed between gate and substrate, all other terminals being left floating.

5. Ramped Field Experiments

Fig.4(a) shows the apparatus used for voltage ramp experiments on MOS devices. A purpose built pulse generator produced a negative polarity, variable speed voltage transient $V_a(t)$, applied to the device under test via a pair of short flying leads. The current was monitored in terms of the voltage transient $V_b(t)$ across a resistor R_s (1MΩ) connected in series with the chuck. R_s also served to limit the tunnelling current in the

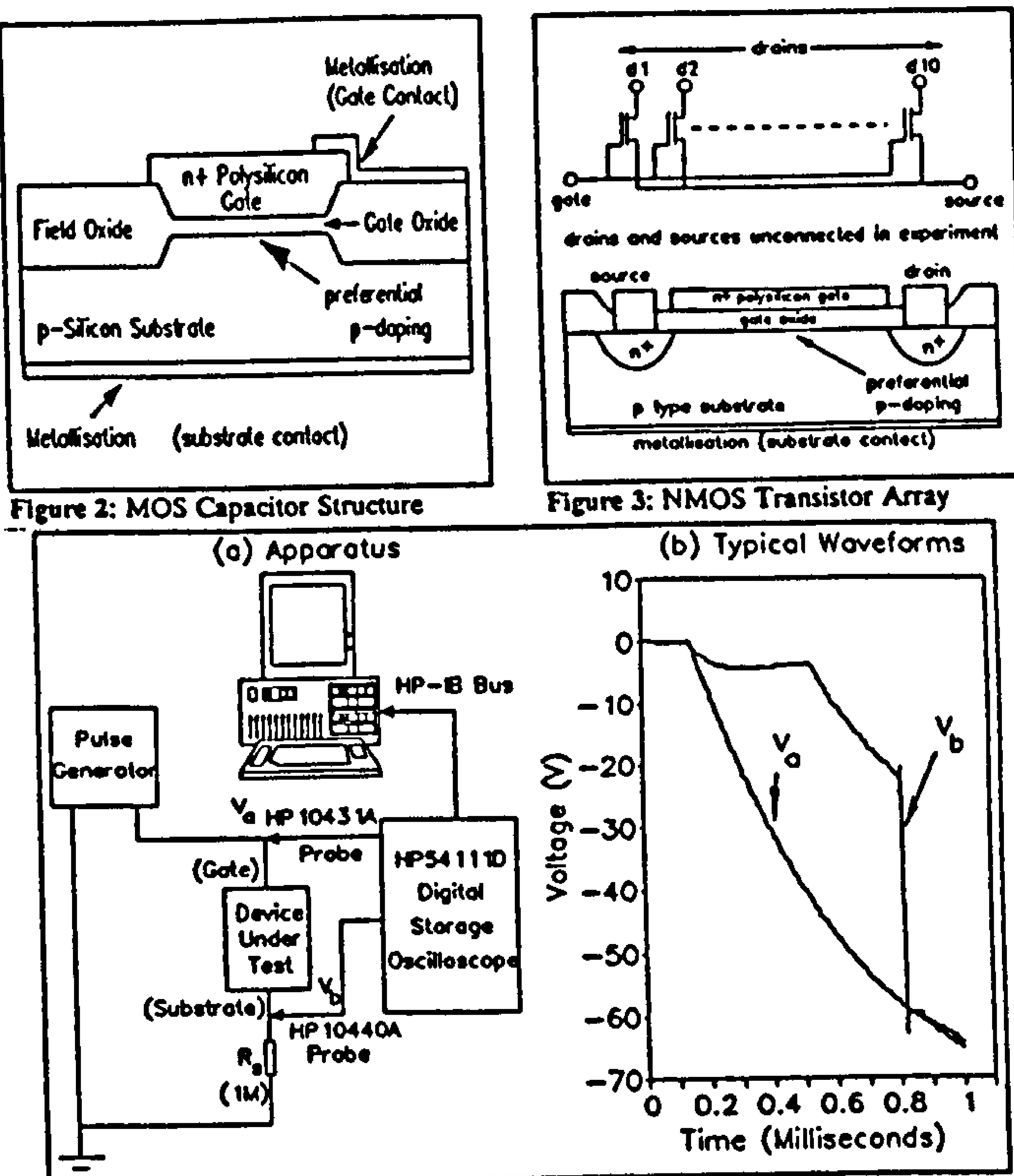


Figure 2: MOS Capacitor Structure

Figure 3: NMOS Transistor Array

Figure 4: Field ramp experiments: (a) Apparatus, (b) Typical waveforms

oxide. V_a and V_b were monitored by HP10431A and HP10440A voltage probes respectively. The waveforms were captured on an HP54111D digital storage oscilloscope and downloaded to a desktop computer for analysis.

Typical V_a and V_b waveforms obtained using negative polarity transients on NMOS transistor structures are shown in Fig.4(b). Fig.5(a) shows the equivalent circuit used to analyze the waveforms. The oxide voltage V_{ox} and injected charge Q_{ox} profiles were extracted using the relations

$$V_{ox} = V_a - V_b \tag{11}$$

$$Q_{ox} = \frac{1}{R_{cg}} \int_0^t V_b\, dt + C_{cg} V_b - (C_{ox} + C_{pc})(V_a - V_b) \tag{12}$$

where C_{pc} is the total stray capacitance in parallel with the device under test, C_{cg} is the total capacitance in parallel with R_s and R_{cg} is represents the combined resistances of R_s and the HP10440A voltage probe.

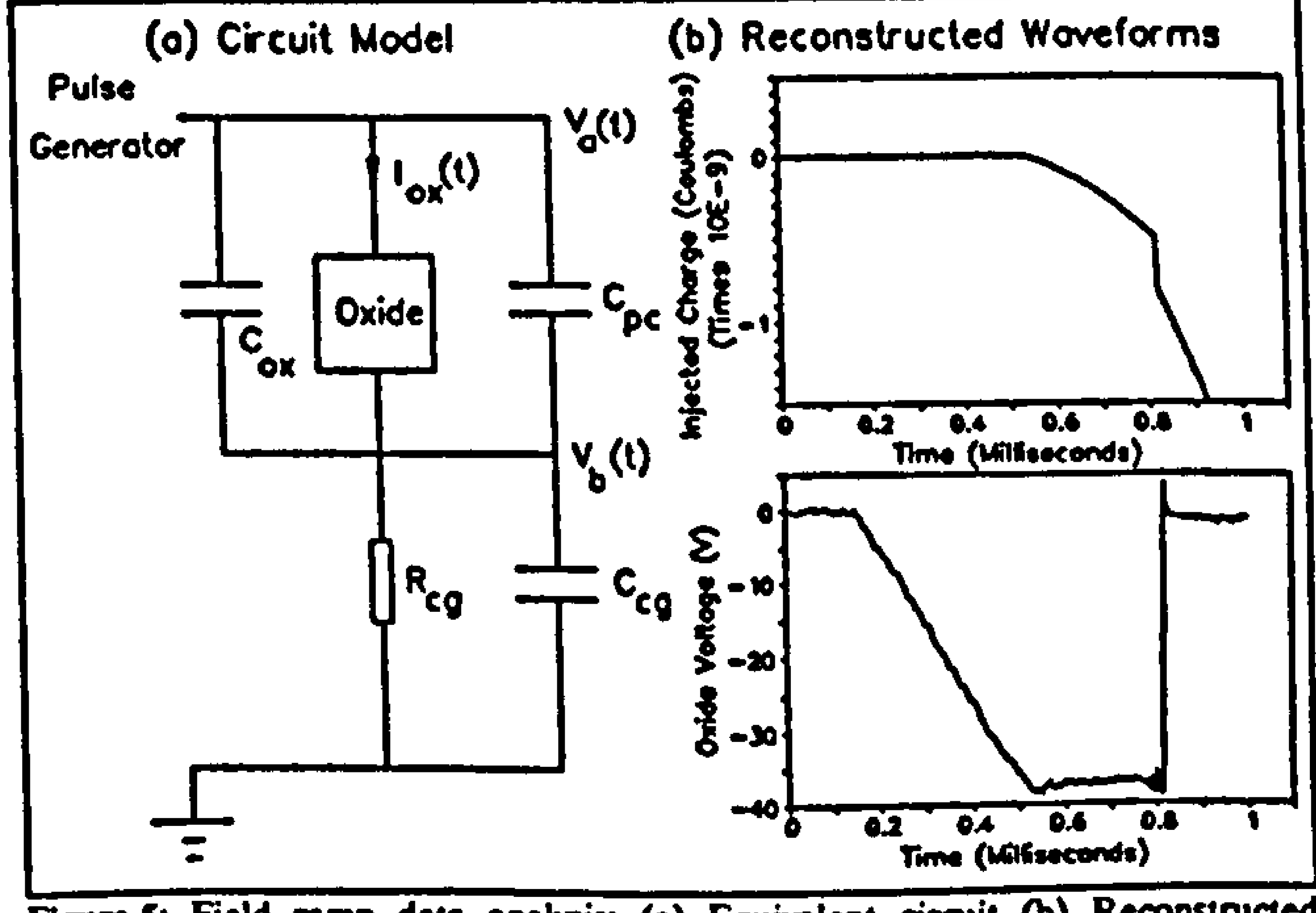


Figure 5: Field ramp data analysis: (a) Equivalent circuit (b) Reconstructed voltage/charge waveforms

Typical $V_{ox}(t)$ and $Q_{ox}(t)$ profiles are shown in Fig.5(b). V_{ox} clearly pins at a value V_{ox}^{p} corresponding to the threshold for 'significant' tunnelling conduction (approx. -37V for a 400Å oxide) and drops rapidly when Q_{ox} reaches the charge-to-breakdown, ie. $A.Q_{bd}$ (see Eqn.(10)) where A is the oxide area over which tunnelling occurs.

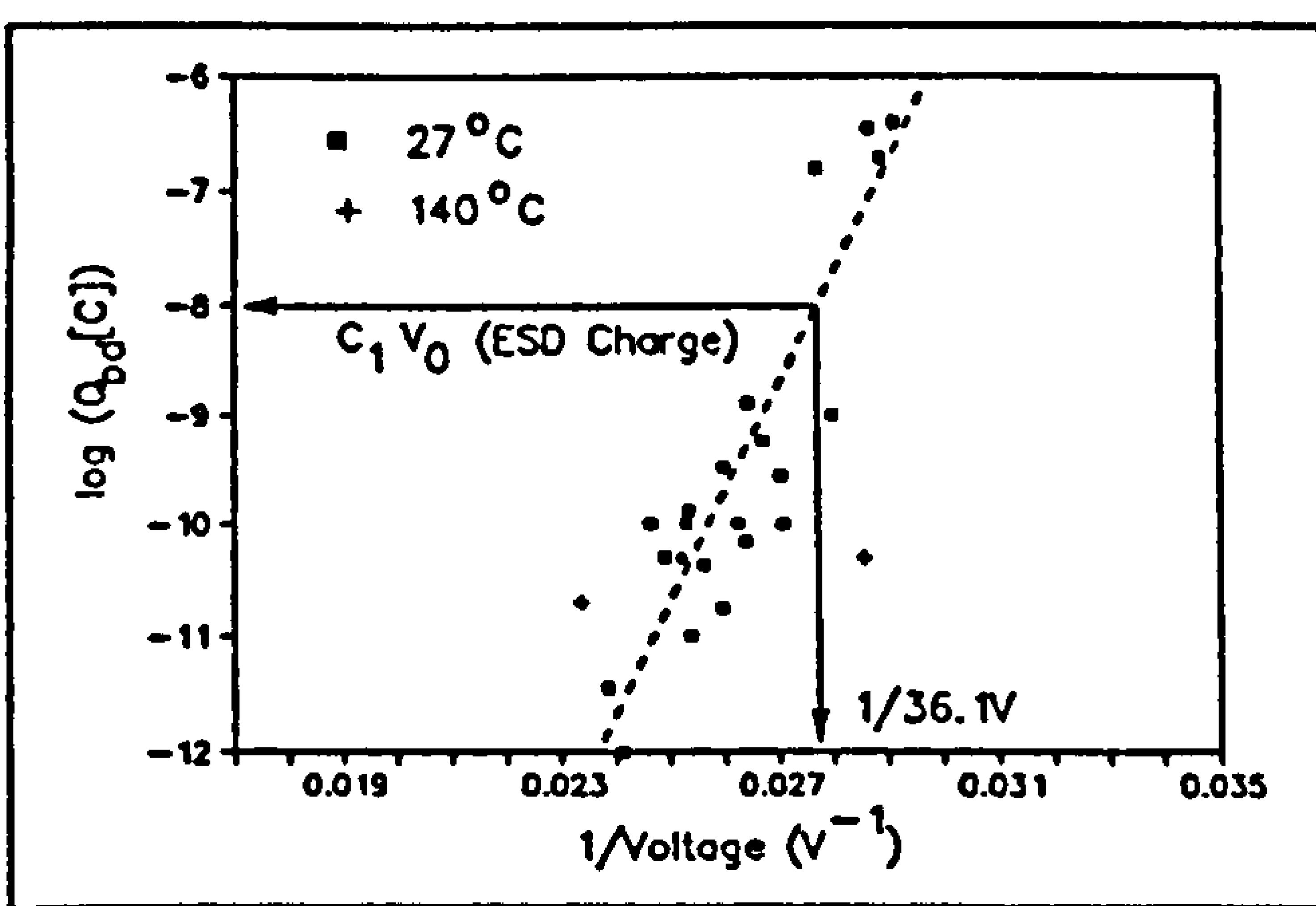


Figure 6: log(Breakdown Charge) vs. $1/V_{ox}{}^{p}$ for Field Ramp Data

The relationship between the charge-to-breakdown and $V_{ox}{}^{p}$ is shown in Fig.6. The curve is approximately linear, in agreement with Eqn.(10). The point which lies well below the curve probably indicates an 'extrinsic' failures, akin to those observed, for example, by Wolters and van der Schoot [1].

It is interesting to compare the measured values of breakdown charge with the levels of charge available during an ESD event. For a 100V ESD pulse supplied by a MIL-STD-883C 'human body model' system ('body' capacitance C_1 = 100pF), the upper limit of the charge available for tunnelling is equal to 10nC. Referring to Fig.6, it appears that this level of charge is attained when the oxide voltage is about -36V.

6. Constant Field Pulse Experiments

The following experiment was performed to determine t_{bd} (and hence τ) as a function of field F. The gates of NMOS E-mode transistor arrays with 400Å oxides were subjected to negative polarity constant voltage pulses of varying magnitudes. The times-to-breakdown were measured and recorded.

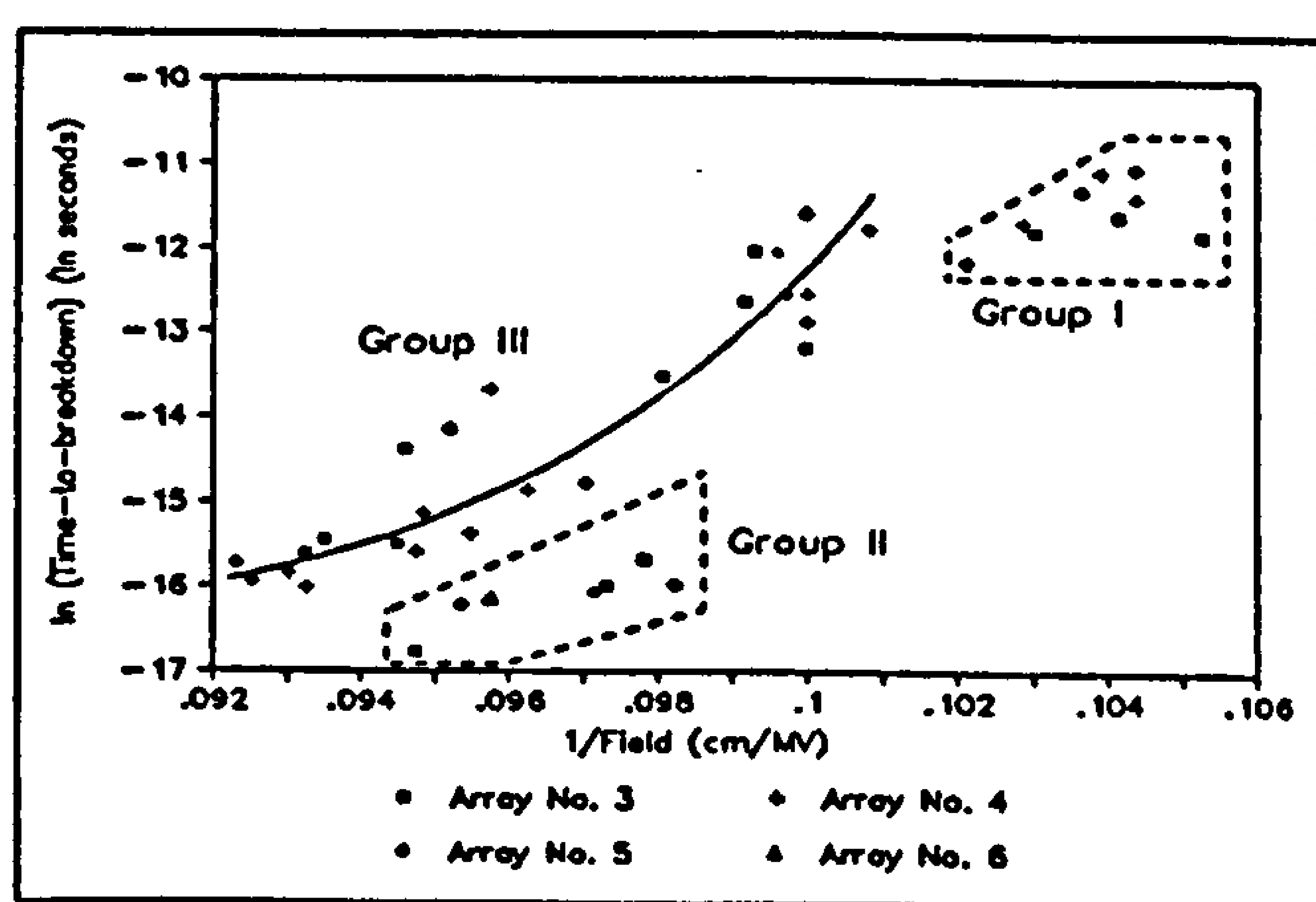


Figure 7: $\log_e t_{bd}$ vs. 1/F data reproduced from the results of other workers

The resulting $\log_e t_{bd}$ vs. 1/F curve is shown in Fig.7. The results were grouped into the following three categories:

Group I: These results are believed to form the 'tail end' of a failure distribution, the whole of which could not be observed using the equipment available, ie. many oxides stressed in this field region failed to break down within the measurement time range of the oscilloscope. These results were therefore excluded from the analysis.

Group II: These devices failed at much shorter times than expected. They were therefore assumed to contain the same 'extrinsic' defects observed by earlier workers [eg.1].

Group III: These results form a curve not unlike those measured by Fong and Hu [11] and Chen et al [10] (see Fig.1), linear for high 1/F but becoming concave-up as the field increases.

7. Short-time-scale Fowler-Nordheim Tunnelling

The nonlinearity of the $\log_e(t_{bd})$ vs. 1/F curve at high fields indicates an imperfection in the causality breakdown model as stated in Section 3. The following experiment shed some light on the reasons for this inconsistency:

A fast voltage pulse was applied to a small dimension MOS transistor array via a 10MΩ resistor (see Fig.8(a)). Charge passing through the oxide was not readily replaced due to the high source resistance and hence tunnelling was indicated by a decay in the device voltage with time. A small capacitance in parallel with the 10MΩ resistor sharpened the risetime and hence produced a constant voltage pulse.

Fig.8(b) shows a typical voltage waveform obtained from this experiment. It is clear that tunnelling begins only after a time delay Δt after the application of the voltage. The value of Δt varies very slowly with changing voltage (far slower than the exponential factor in the causality equation) and shall, for the purposes of this paper, be regarded as having a constant value of 200ns.

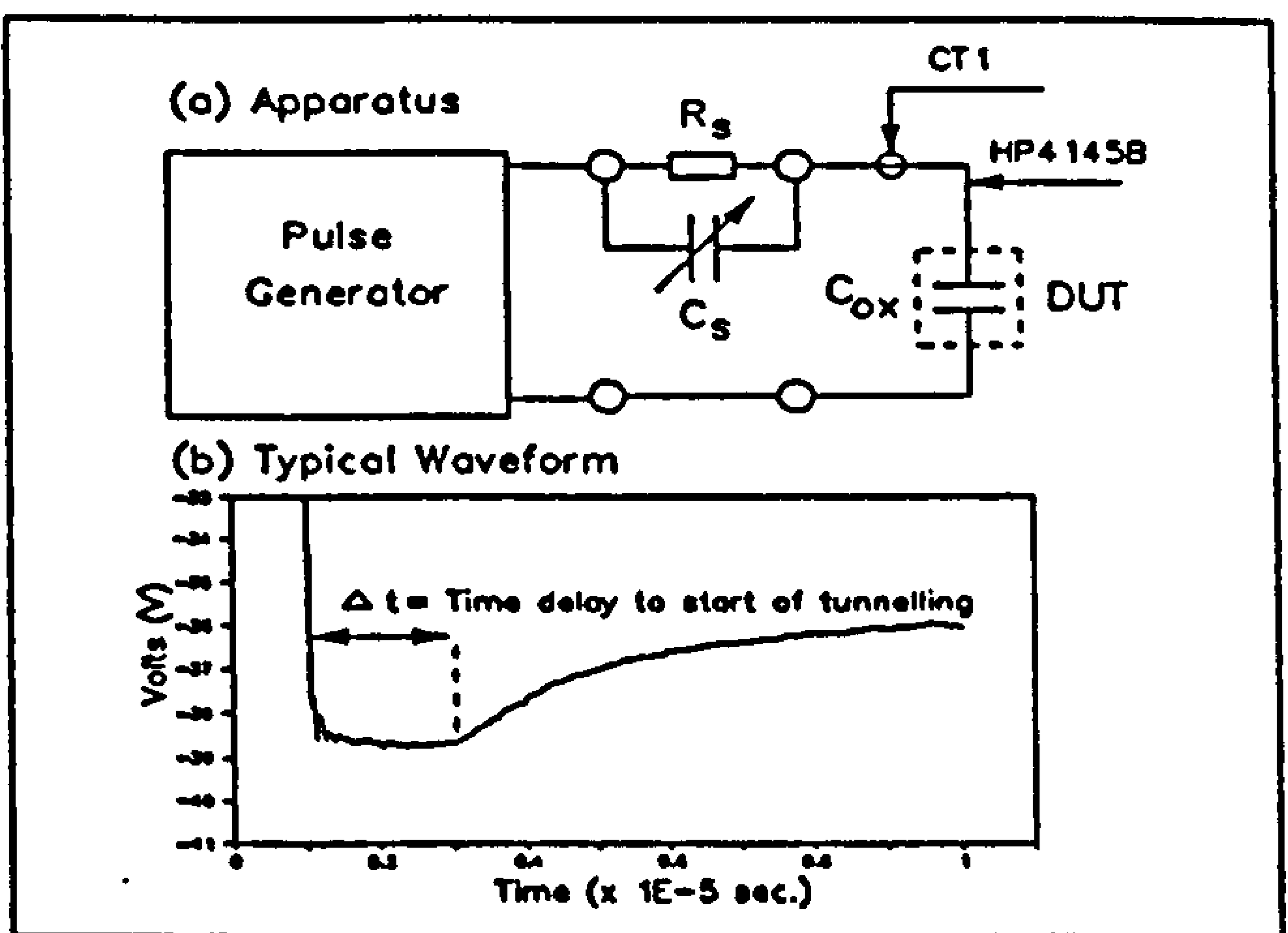


Figure 8: Short time scale tunnelling (a) Apparatus (b) Typical Waveform

8. Modified Causality Model

Since oxide breakdown is believed to be caused by tunnelling [1,2,3,9,10], it is reasonable to assume that the damage leading to breakdown accumulates only during the period over which tunnelling occurs. Hence Eqn.(6) can be re-written

$$\int_{\Delta t}^{t_{bd}} \frac{dt}{\tau(F)} = 1 \qquad (13)$$

or for a constant field

$$t_{bd} = \tau_0 e^{\frac{\gamma}{F}} + \Delta t \qquad (14)$$

Fig.9 shows the $\log_e(t_{bd})$ vs. 1/F curve predicted by Eqn.(14), using Δt = 200ns, γ = 1000MV/cm and $\tau_0 = 2.286.10^{-49}$s, compared to the 'intrinsic' (Group III) experimental data.

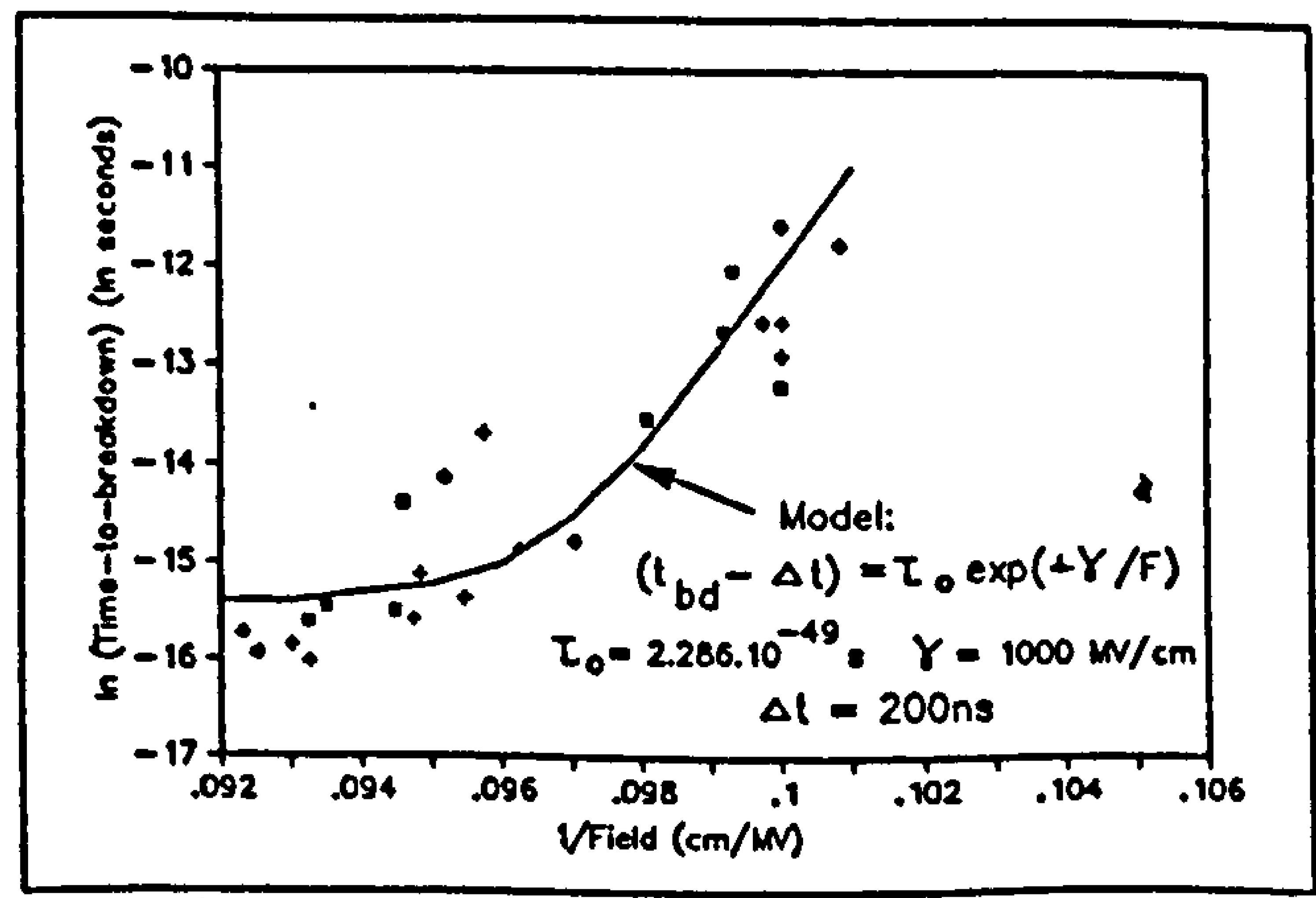


Figure 9: Comparison between 'intrinsic' pulse breakdown data and modified causality model

9. Model of Negative Polarity ESD Breakdown

In this section a model of oxide breakdown under negative polarity ESD (HBM) stress is developed, based upon the modified causality theory presented in Section 8. The predictions of the ESD model are experimentally tested in Section 10.

When negative polarity stress is applied to a p-type MOS structure, both electrode surfaces are driven into accumulation and most of the device voltage appears across the oxide. Hence to a fair approximation, the situation can be modelled using the simple circuit of Fig.10(a) where C_1 and R represent the human body resistance and capacitance and C_{ox} represents the oxide capacitance.

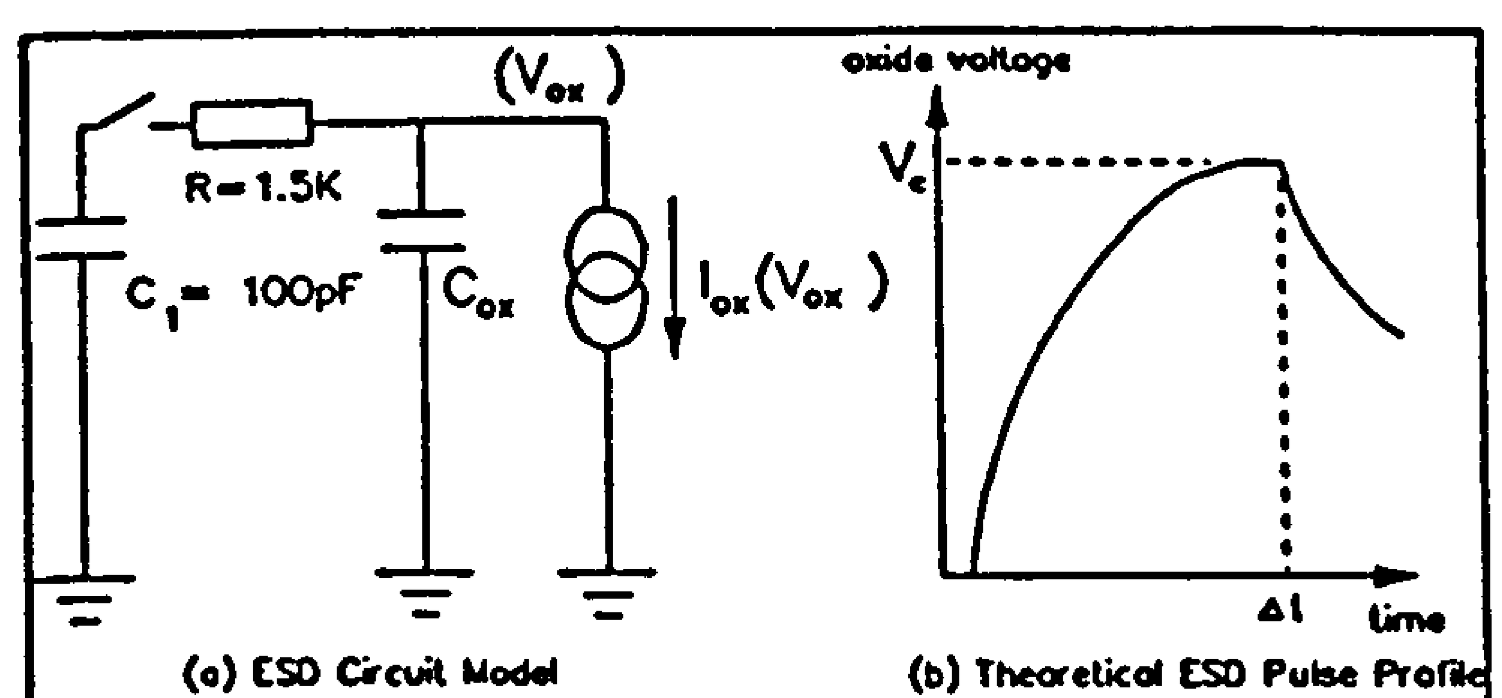


Figure 10: Negative polarity ESD model: (a) Equivalent circuit, (b) Hypothetical voltage waveshape (not to scale).

9.1 Modelling the Voltage Pulse Profile

The current source in Fig.10(a) represents the tunnelling injection at the oxide cathode. As in the previous section, an abrupt commencement of tunnelling will be assumed after a time delay Δt from the beginning of the pulse [see Fig.10(b)]. Hence the tunnelling current I_{ox} is given by

$$I_{ox} = \frac{Ak}{T_{ox}^2} V_{ox}^2 \, e^{-\frac{BT_{ox}}{|V_{ox}|}} H(t-\Delta t) \qquad (15)$$

where A is the oxide area over which tunnelling occurs (assumed to be equal to the total oxide area), T_{ox} is the oxide thickness and H is the Heaviside step function.

(i) **Pre-Tunnelling (Pulse rise and dwell) [0 < t < Δt]:**

Neglecting any oscillatory effects caused by stray inductance, the pulse profile is given by

$$V_{ox}(t) = V_e\left[1 - e^{-\frac{t}{\tau_r}}\right] \qquad (16)$$

where τ_r is the time constant of the pulse rise and V_e is the 'effective' stress voltage received by the oxide (ie. the maximum value to which V_{ox} can rise in the absence of oxide conduction. V_e is generally lower than the 'applied' ESD voltage V_a since the oxide and the transmission medium present a capacitive load to the ESD source [13].

(ii) **Tunnelling (Pulse decay) [Δt < t < ∞]:**

If tunnelling commences after five rise-time constants ($5\tau_r > \Delta t$) then $V_{ox}(\Delta t)$ is equal to V_e. If the tunnelling current is of the order of μA then the voltage across the 1.5KΩ body resistance is negligible and all the capacitances parallel to the oxide (C_1, C_{ox}, transmission line capacitances etc.) act as a single capacitor C. Under this assumption, the decay profile can be modelled by the differential equation

$$C\frac{dV_{ox}}{dt} = -I_{ox}(V_{ox}) \qquad (17)$$

subject to the boundary condition $V_{ox}(\Delta t) = V_e$. Solving Eqn.(17) yields

$$|V_{ox}(t)| = \frac{BT_{ox}}{\log_e\left[\frac{AkB(t-\Delta t)}{CT_{ox}} + e^{\frac{BT_{ox}}{|V_e|}}\right]} \qquad (18)$$

9.2 Modelling Oxide Damage and Failure

The voltage profile of Eqn.(18) can be inserted into the modified causality model (Section 8). For this purpose let

$$s(t) = \int_{\Delta t}^{t} \frac{dt}{\tau(F)} \qquad (19)$$

such that $s(t_{bd}) = 1$. s(t) can be seen as a measure of the oxide damage inflicted by a field profile F(t) over a time t. Inserting the field profile $F(t) = V_{ox}(t)/T_{ox}$ into Eqn.(19) yields

$$s(t, V_e) = \frac{CT_{ox}}{\tau_0 Ak(B-\gamma)}\left[\left[\frac{AkB(t-\Delta t)}{CT_{ox}} + e^{\frac{BT_{ox}}{|V_e|}}\right]^{\frac{B-\gamma}{B}} - e^{\frac{(B-\gamma)T_{ox}}{|V_e|}}\right] \qquad (20)$$

Setting $s(t_{bd}, V_e) = 1$, Eqn.(20) can be re-arranged to give t_{bd} as a function of V_e, ie.

$$t_{bd} = \Delta t + \frac{CT_{ox}}{AkB}\left[\frac{Ak\tau_0(\gamma-B)}{CT_{ox}} + e^{-(\gamma-B)\frac{T_{ox}}{|V_e|}}\right]^{-\frac{B}{\gamma-B}} - \frac{CT_{ox}}{AkB} e^{\frac{BT_{ox}}{|V_e|}} \qquad (21)$$

The $(t_{bd}-\Delta t)$ vs. V_e curve predicted by Eqn.(21) is plotted in Fig.11 using the values of k, B, γ and τ_0 calculated previously and the values of T_{ox} and A appropriate to a p-type capacitor Type 1 (see Section 3). It is clear that the transition from a very large t_{bd} to a very small t_{bd} occurs over a narrow voltage range. Hence the oxide breaks down either very quickly or not at all.

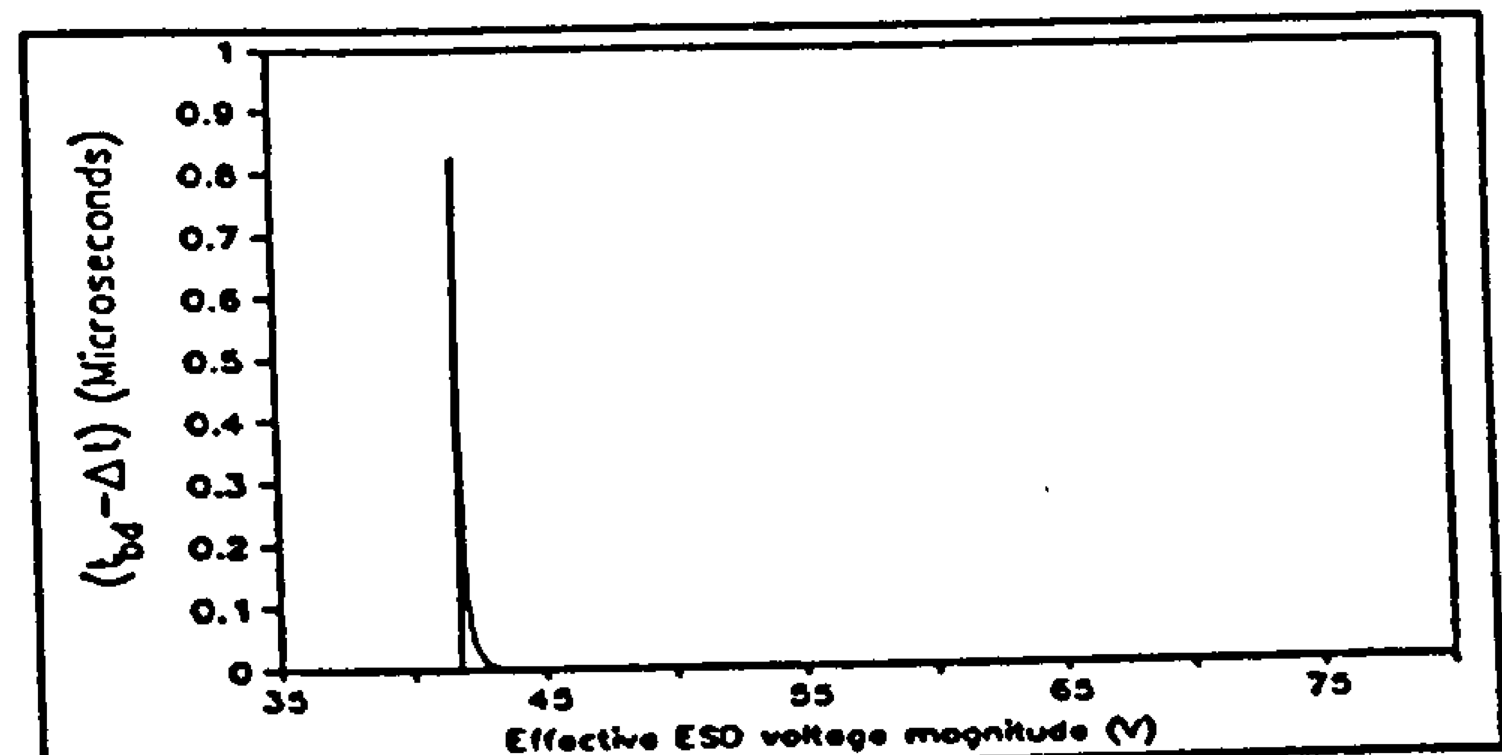


Figure 11: Theoretical $(t_{bd}-\Delta t)$ vs. V_e curve

Let $S(V_e)$ represent the total oxide damage inflicted by a complete ESD pulse of effective magnitude V_e, ie.

$$S(V_e) = s(t=\infty, V_e) = \frac{CT_{ox}}{\tau_0 Ak(\gamma-B)} e^{\frac{(\gamma-B)T_{ox}}{|V_e|}} \qquad (22)$$

ESD breakdown thresholds are normally determined by a 'step' test in which a sequence of ESD pulses of ascending magnitude are applied until breakdown occurs at the n'th pulse

$$|V_e(r)| = |r.\Delta V| \qquad (23)$$

where r=1,2,3,...n and ΔV is the voltage resolution. Hence the breakdown voltage threshold V_e^* is defined as $V_e(n) = n.\Delta V$. The total oxide damage $S_t(r.\Delta V)$ after pulse number r is equal to the total damage produced by pulses 1 to r, ie.

$$S_t(r.\Delta V) = \sum_{i=1}^{r} \frac{CT_{ox}}{\tau_0 Ak(\gamma-B)} e^{\frac{(\gamma-B)\,T_{ox}}{|i.\Delta V|}} \qquad (24)$$

Fig.12 shows the additional damage S(r.ΔV) inflicted by each successive pulse as a percentage of the total damage S_t(r.ΔV), plotted against r for ΔV=3V. Clearly the vast majority of damage is inflicted by the final pulse in the sequence [ie. S(r.ΔV) < S((r+1)ΔV). Hence for 'sensible' magnitude of ΔV (3V+), the criterion for ESD breakdown in a pulse sequence test can be expressed as

$$S(V_e^*) = \frac{CT_{ox}}{\tau_0 Ak(\gamma-B)} \exp\left[\frac{(\gamma-B)T_{ox}}{|V_e^*|}\right] = 1 \qquad (25)$$

where V_e^* is the effective ESD breakdown voltage threshold. This expression can be re-arranged to yield

$$|V_e^*| = \frac{(\gamma-B)T_{ox}}{\log_e\left[\frac{CT_{ox}}{Ak\tau_0(\gamma-B)}\right]} = \frac{(\gamma-B)T_{ox}}{\log_e\left[\frac{C}{A}\right] + \log_e\left[\frac{T_{ox}}{k\tau_0(\gamma-B)}\right]} \qquad (26)$$

Variations in the $\log_e$[C/A] term are small compared to the other term in the denominator and can therefore be neglected (this is demonstrated by the calculations in Table 1). Hence the effective breakdown voltage threshold is a function of the quality and thickness of the oxide and largely independent of the oxide area and stress conditions. Its value is approximately -38V for a 400Å oxide which is reasonably consistant with the -36.1V estimated by the charge/voltage data in Fig 6.

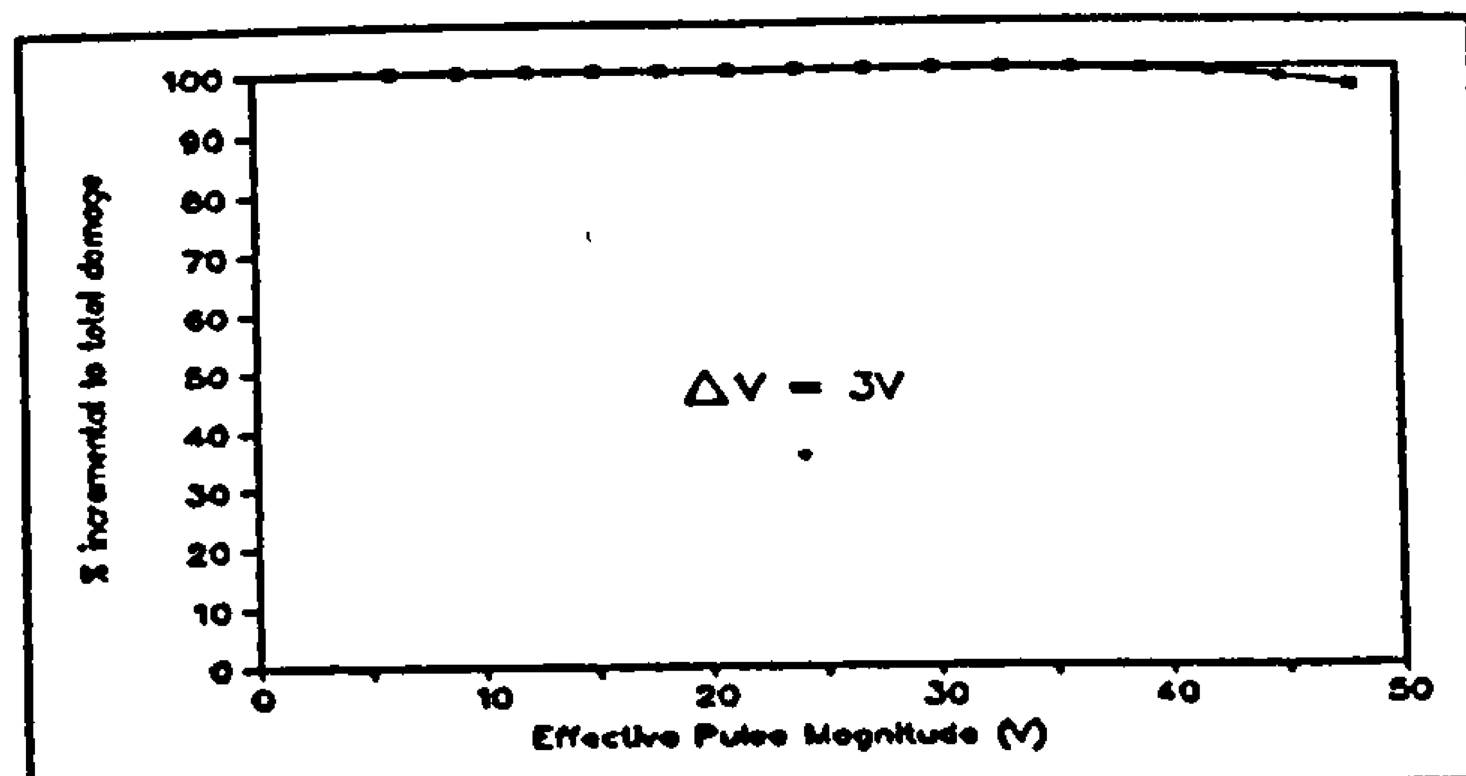


Figure 12: Incremental damage S_i plotted as percentage of total damage S_t for an ESD pulse sequence

10. Experiments on Negative Polarity ESD Breakdown

Fig.13(a) shows the apparatus used to measure negative polarity ESD breakdown voltage thresholds. The Hartley AutoZap ESD source supplied ESD pulses to the devices under test via a co-axial cable, and the current profiles were monitored by a Tektronix CT1 current probe. MOS capacitors and transistor arrays were subjected to negative polarity ESD pulse sequences, beginning at -60V and increasing in magnitude in steps of 4V until breakdown was observed. Breakdown was indicated by a distinctive change in the shape of the current waveform [6].

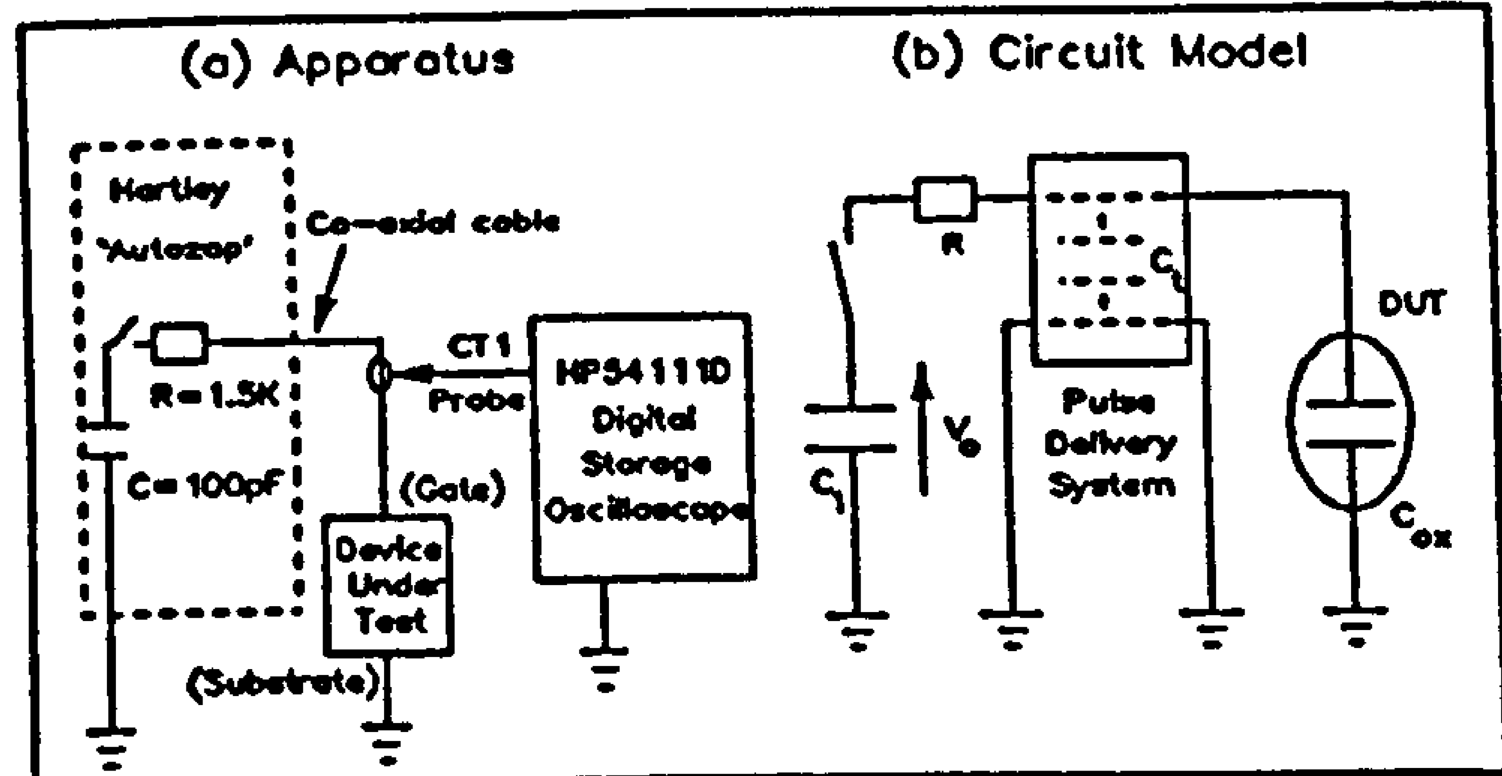


Figure 13: ESD Apparatus and Equivalent Circuit

Fig.13(b) shows a circuit model for the ESD apparatus, where C_t is the capacitance of the co-axial transmission cable. If an oxide voltage $V_e^* = -40V$ is required to cause breakdown (Section 6) then the 'applied' ESD breakdown voltage V_{bd} is given by

$$V_{bd} = \frac{V_e^* (C_1 + C_t + C_{ox})}{C_1} \tag{27}$$

C_t, C_1 and C_{ox} were measured using a Wayne-Kerr 4210 LCR bridge and V_e^* was calculated for each structure using Eqn.(26). C_1 was equal to 95.5pF for all tests. Theoretical values of V_{bd} were thereby deduced and are compared with their experimental values in Table 1.

Table 1

Structure Type	Experimental V_{bd}	(C_t+C_{ox})	V_e^*	Theoretical V_{bd}
Type 1	-96V	158.6pF	-41.32V	-109.94V
Type 2	-80V	118.4pF	-37.269V	-83.47V
Type 3	-80V	120.2pF	-38.396V	-86.00V
Type 4	-80V	118.5pF	-37.506V	-84.04V
Type 5	-84V	121.0pF	-38.602V	-87.51V
Type 6	-84V	121.1pF	-38.648V	-87.66V

There is a fairly close correlation between theoretical and experimental V_{bd} for the transistor array structures (Types 2-6). A small inaccuracy (approx.15%) exists for the capacitor structures (Type 1). This is possibly due to an inaccuracy in the calculated V_e^* caused by the tunnelling surface area not being exactly equal to the gate area. (A decrease in A in Eqn.(26) produces a decrease in the magnitude of V_e^*.)

11. Model of Positive Polarity ESD Breakdown

Under positive polarity stress, the p-type silicon is driven first into deep depletion and then into inversion. The inversion layer can be formed either by thermally induced minority electrons or by avalanche conduction in the depletion layer. The rate of inversion layer growth under thermal carrier generation is governed by the minority carrier response time' τ_{min}, between 10ms and 1s [14]. For the purposes of this model, τ_{min} will be taken to be approximately 10ms. This situation can be conveniently modelled for a 41pF MOS capacitor by including a R_s=250MΩ resistor in series with the oxide capacitance, such that the oxide voltage (and hence the inversion layer charge) responds with a time constant of 10.25ms.

Fig.14(a) shows the approximate equivalent circuit for a p-type MOS capacitor under positive ESD stress. The behaviour of the circuit can be modelled by numerical solution of the differential equations:

$$\frac{V_2 - V_1}{R_s} = C_1 \frac{dV_1}{dt} \tag{28}$$

$$\frac{V_1 - V_2}{R_s} = C_{ox} \frac{dV_2}{dt} + I_{ox} \tag{29}$$

subject to the boundary conditions $V_1(0) = V_0$ and $V_2(0) = 0$. Fig.14(b) shows the numerically predicted $V_2(t)$ profiles corresponding to +1KV and +100KV ESD pulses. It is clear from these waveforms that the slow minority carrier generation rate causes the tunnelling current to 'pin' the oxide voltage. It was shown in Sections 6 and 7 that an oxide voltage V_2 of about 40V is required to support ESD breakdown. Hence, referring to Fig.14 it appears that if thermal carrier generation were the only mechanism of inversion, even with the smallest value of τ_{min} (10ms) a device could withstand positive ESD pulses well in excess of 1KV.

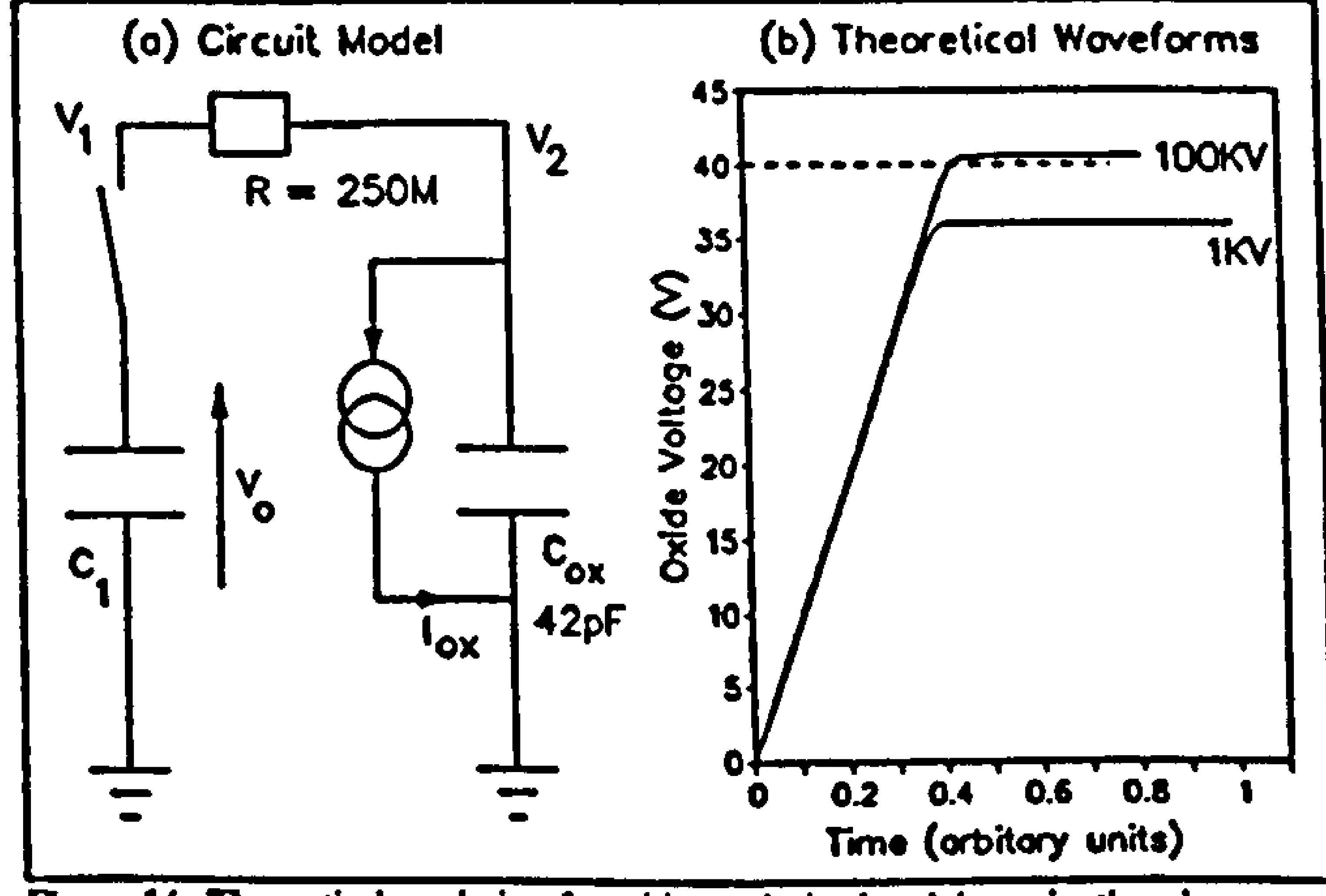


Figure 14: Theoretical analysis of positive polarity breakdown in the absence of avalanche conduction

This shows that the positive polarity ESD breakdown at voltages below 1KV [7,13] must be supported by depletion layer avalanche conduction. Avalanche conduction occurs during the initial 'deep depletion' phase, before thermal carrier generation has begun and the depletion layer voltage is at its peak. The device voltage V_e^* required for breakdown should therefore be given by

$$V_e^* = V_{av} + V_{ox}^* \tag{30}$$

where V_{ox}^* is the oxide breakdown voltage (given by Eqn.26 as approximately 40V) and V_{av} is the depletion layer avalanche threshold.

12. Experiments on Positive Polarity Breakdown

These experiments were designed to test the theory developed in Section 8.

12.1 Determination of Avalanche Threshold

The depletion layer avalanche threshold was determined using the voltage ramp apparatus described in Section 5. Positive polarity voltage ramps were applied to unimplanted MOS capacitors (Type 1).

Fig.15 shows the equivalent circuit used to analyze the data. The oxide voltage V_{ox} was reconstructed using the relation

$$V_{ox} = \frac{V_b C_{cg}}{C_{ox}} + \frac{1}{C_{ox} R_{cg}} \int_0^t V_b \, dt - (V_a - V_b) \frac{C_{pc}}{C_{ox}} \tag{31}$$

while the voltage V_{dep} across the surface depletion layer was reconstructed using

$$V_{dep} = V_a - V_b - V_{ox} \tag{32}$$

Fig.16 shows the V_{ox} and V_{dep} curves re-constructed using this technique. The depletion layer voltage saturates close to a value V_{av} = 80V for all the devices tested, which indicates the silicon avalanche threshold. The avalanche threshold appears to increase with time. This increase is believed to be caused by Joule heating in the localised avalanche filaments (or 'microplasmas') located at the capacitor corners.

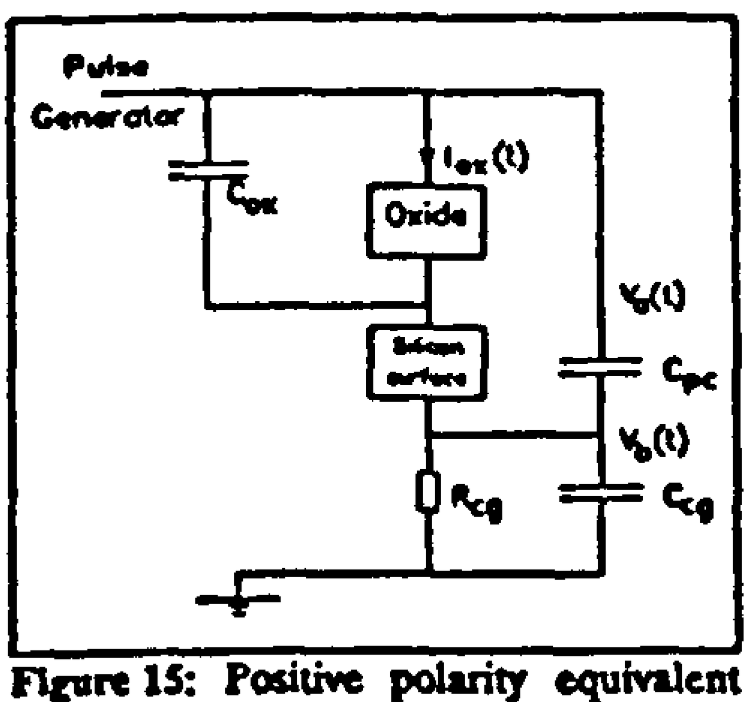


Figure 15: Positive polarity equivalent circuit for voltage ramp apparatus

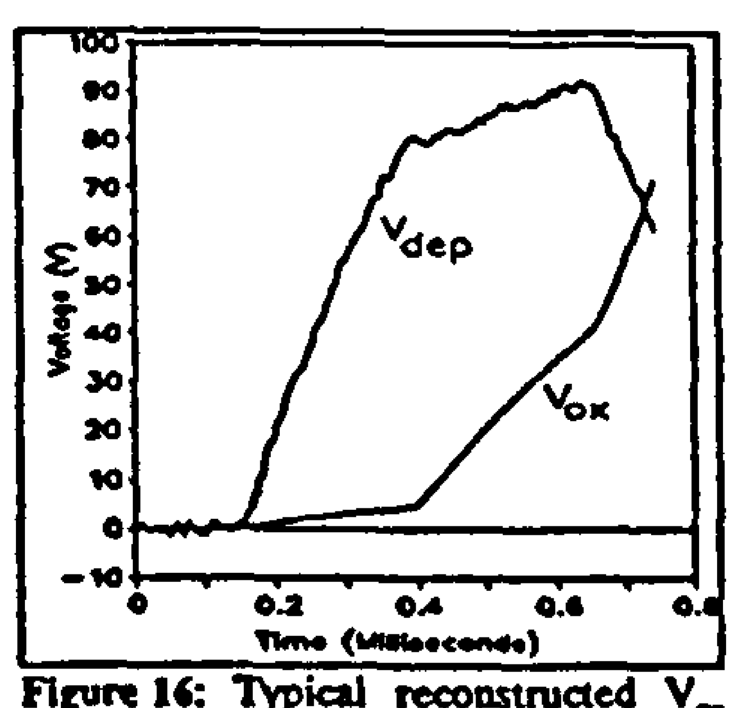


Figure 16: Typical reconstructed V_{ox} and V_{dep} profiles

The apparent decrease in V_{dep} towards the end of the reconstructed waveforms indicates the onset of significant Fowler-Nordheim tunnelling. Tunnelling begins when V_{ox} reaches a threshold of around 41V (which is approximately consistent with the negative polarity value, see Section 5). The sections of the V_{ox} and V_{dep} curves after this point have no physical significance since Eqn.(31) ceases to hold in the presence of tunnelling.

An earlier study [13] suggested instantaneous oxide breakdown when depletion layer underwent avalanche. These results show that this is not the case. Avalanche conduction merely provides a continuous supply of electrons for injection into the oxide while breakdown itself depends upon the oxide field, in accordance with Eqn.(26).

12.2 Sub-avalanche Breakdown

The apparatus of Fig.17(a) was used to examine positive polarity oxide breakdown in the absence of depletion layer avalanche conduction. A continuous +50V D.C. voltage was applied to a p-type MOS capacitor (Type 1). The oxide tunnelling charge profile was determined by monitoring the voltage V_k across a large capacitor C_k=2.2mF connected in series with the device under test to collect injected charge. Fig.17(b) shows a typical $Q_{ox}(t)$.[$=C_kV_k(t)$] profile. The sharp increase in dV_k/dt after 1 hour 15 minutes indicates oxide breakdown. Q_{bd} (the value of Q_{ox} at breakdown) is equal to 510μC, which is more than 50,000 times the charge available in a 100V ESD pulse.

This result adds experimental evidence to the theory of Section 8. It shows that even if the device voltage exceeds the oxide breakdown potential [The applied 50V well exceeds the 40V oxide breakdown voltage given by Eqn.(26)], the slow electron generation rate causes the oxide voltage to remain low and hence t_{bd} and Q_{bd} to be very high [Eqns.(10) and (14)].

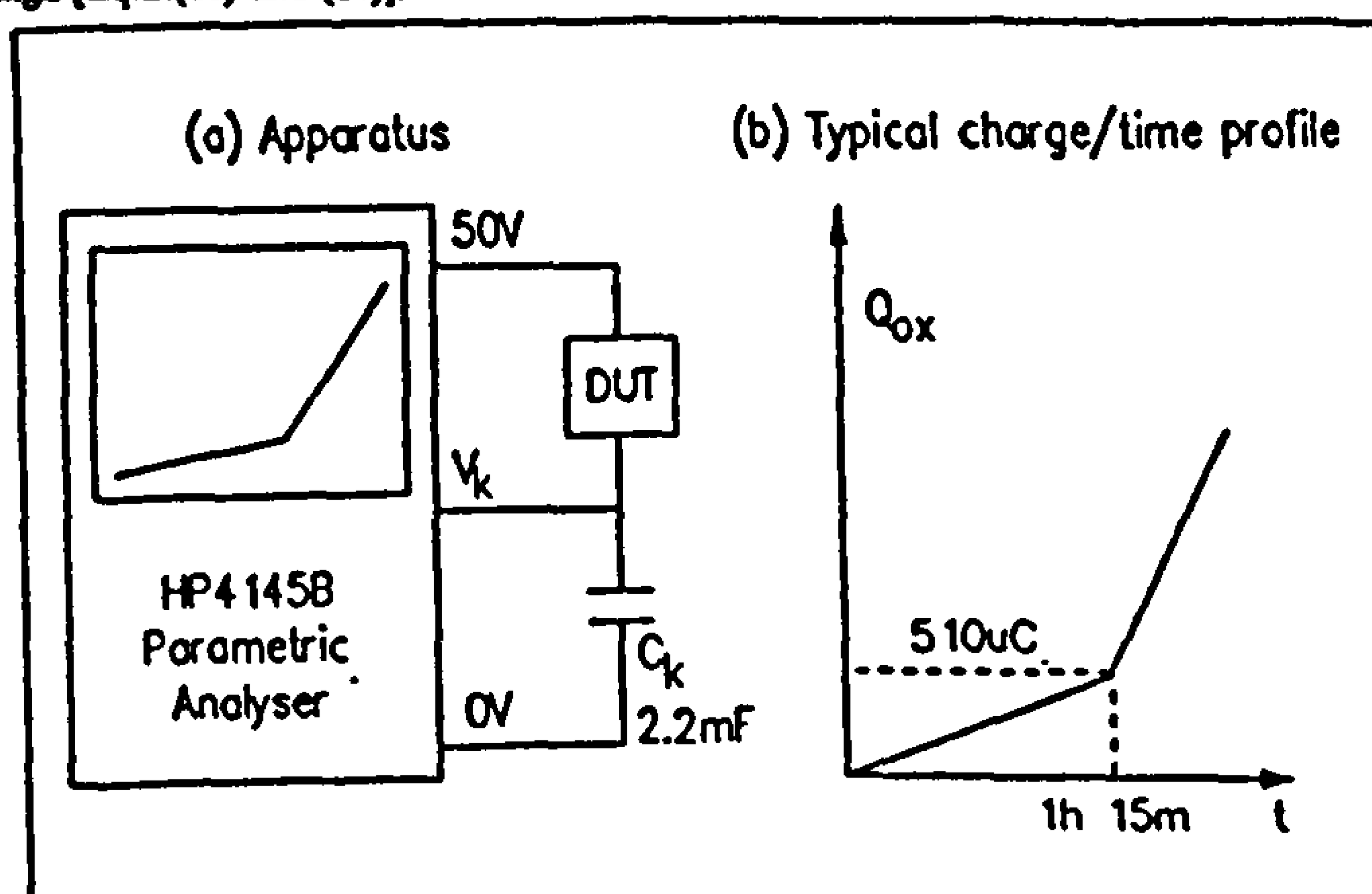


Figure 17: Sub avalanche positive polarity experiment

12.3 Positive Polarity ESD Breakdown Thresholds

The experiments described in Section 7 were repeated using positive polarity ESD pulses. Fig.18 shows the equivalent circuit for the positive polarity ESD system.

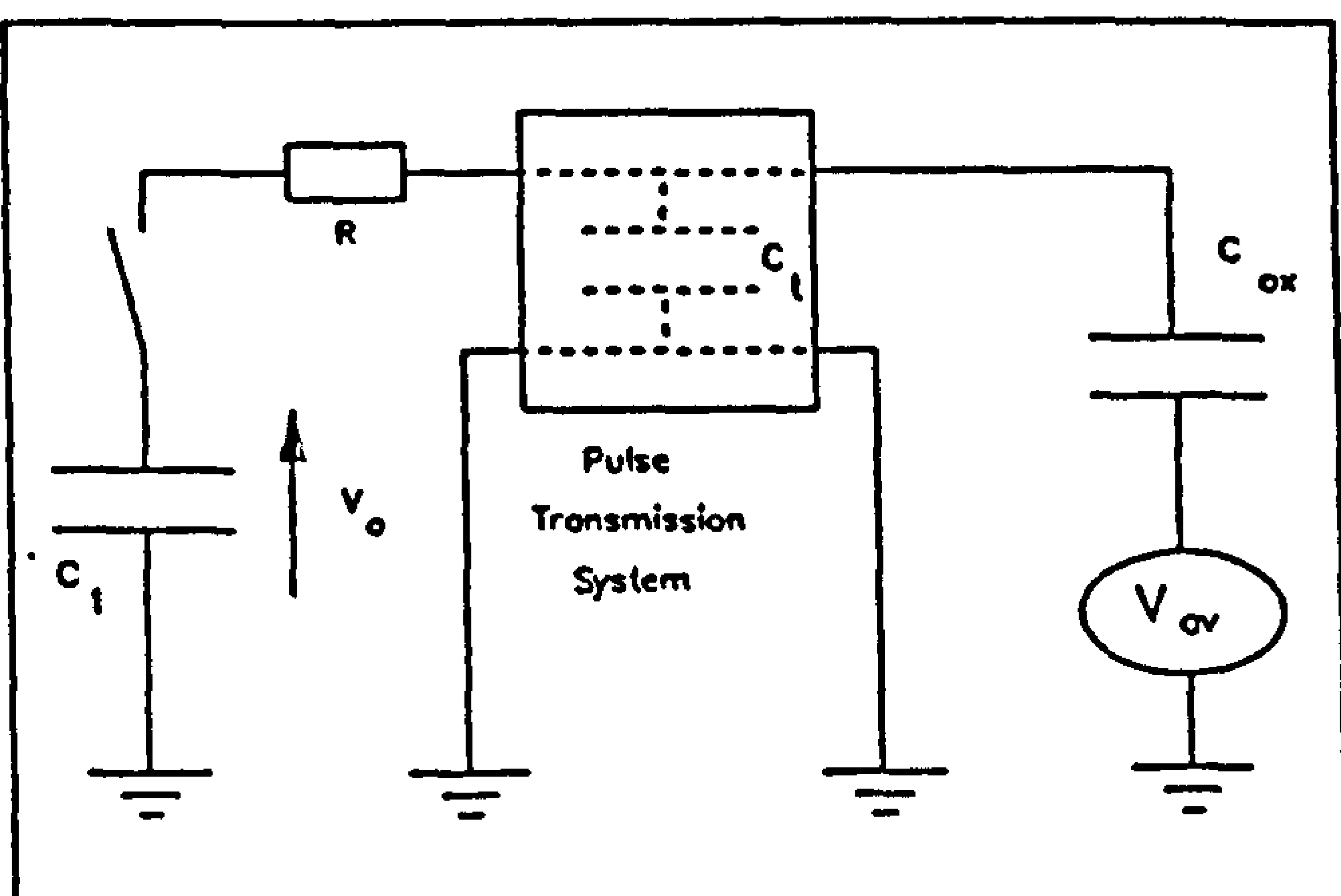


Figure 18: Positive polarity ESD circuit model

According to this equivalent circuit, the applied breakdown voltage threshold V_{bd} is given by

$$V_{bd} = \frac{(C_1 + C_t)(V_{av} + V_{ox}^*) + C_{ox} V_{ox}^*}{C_1} \tag{33}$$

Theoretical values of V_{bd} were calculated, assuming that the values for V_{ox}^* are equal in magnitude to the values of V_c^* used in Section 7 and taking V_{av}=80V. C_1 and C_t were equal to 95.5pF and 117.6pF for all the experiments. Experimental and theoretical V_{bd} values are compared in Table 2.

Table 2

Structure	Experimental V_{bd}	V_{ox}^*	C_{ox}	Theoretical V_{bd}
Type 1	324V	41.32V	41pF	288.45V
Type 2	116V	37.269V	.8pF	261.99V
Type 3	132V	38.396V	2.6pF	264.62V
Type 4	192V	37.506V	.9pF	262.56V
Type 5	220V	38.602V	3.4pF	266.02V
Type 6	204V	38.648V	3.5pF	266.17V

For the case of the MOS capacitor (Type 1) the experimental V_{bd} is about 11% higher than the theoretical value. This discrepancy may be due to a resistive voltage drop in series with V_{av} introduced by current constriction around the avalanche filaments. Hence Eqn.(33) provides a slightly pessimistic estimate of a device's ESD sensitivity.

Correlation for the transistor results is, in general, poor. The major source of inaccuracy is probably V_{av}, which was only measured for an MOS capacitor structure.

13. Conclusions

Oxide Breakdown in MOS structures under a variety of EOS and ESD conditions has been studied both experimentally and theoretically. A fairly close correlation has in general been observed between the experimental data and the theoretical predictions.

The following specific conclusions can be drawn from this work:

1. The non-linearity of the log(t_{bd}) vs. 1/Field curve at high electric fields may be due to the finite time response of the Fowler-Nordheim tunnelling mechanism. This finite response is possibly due to the finite time constants of the traps at the cathode-oxide interface.
2. Analytical modelling has shown that the apparently time-independent characteristics of ESD breakdown (namely the near-instantaneous failure and the absence of latent oxide damage) are consistent with the causality breakdown model.

3. Experiment and theory have shown that positive polarity breakdown in a p-type structure under sub-kilovolt ESD conditions can only result from avalanche conduction in the deeply depleted cathode. The total device voltage required for breakdown is equal to the oxide breakdown potential plus the depletion layer avalanche threshold.

Hence a wide range of EOS and ESD phenomena have been explained in terms of a single theory, namely a modified version of the causality breakdown model.

14. Acknowledgements

The work contained in this paper was carried out with the support of the Procurement Executive, Ministry of Defence, to whom we are grateful for permission to publish. The authors would like to thank other members of the Electronic Component Group at Loughborough University, and Mr.D.Shore of the M.O.D. for many valuable discussions.

15. References

1. Wolters,D.R., van der Schoot,J.J., "Dielectric Breakdown in MOS Devices", Phillips Journal of Research, 40, pp.115-92, 1985.

2. Chen,I.C., Holland,S, Hu,C, "Electrical Breakdown in Thin Gate and Tunnelling Oxides", IEEE Trans. Electron Dev., ED-32, pp.413-22, 1985.

3. Harari,E., "Dielectric Breakdown in Electrically Stressed Thin Films of Thermal SiO_2", J.Appl.P., 49, pp.2478-89, 1978.

4. Amerasekera,E.A., Campbell,D.S., "Oxide Breakdown in MOS Structures under ESD and Continuous Voltage Stress Conditions", Proc. European Conference on Reliability, Copenhagen, pp.325-330, 1986.

5. Amerasekera,E.A., Campbell,D.S., "An Investigation of the Nature and Mechanisms of ESD Damage in NMOS Transistors", Solid State Electronics, 32, pp.199-206, 1989.

6. Tunnicliffe,M.J., Dwyer,V.M., Campbell,D.S., "Slow Transient Voltage and ESD Breakdown in Unprotected MOS Gate Oxides", Proc. Components Engineering, Reliability and Test Conference, Gatwick Airport, London, pp.78-87, 1990.

7. Amerasekera,E.A., Campbell,D.S., "ESD Pulse and Continuous Voltage Breakdown in MOS Capacitor Structures", 1986 EOS/ESD Symp. Proc., pp.208-213, 1986.

8. O'Dwyer,J.J., "The Theory of Electrical Conduction and Breakdown in Solid Dielectrics", Clarendon Press, Oxford, pp.77-94, 1973.

9. Avni,E, Shappir,J, "A Model for Silicon Dioxide Breakdown under High Field and Current Stress", J.Appl.P., 64, pp.743-8, 1988.

10. Chen,I.C., Hu,C, "Accelerated Testing of Time-Dependent Breakdown of SiO_2", IEEE Electron Dev. Lett., EDL-8, pp.140-2, 1987.

11. Fong,Y, Hu,C, "The Effects of High Electric Field Transients on Thin Gate Oxide MOSFETs", 1987 EOS/ESD Symp. Proc., pp.252-7, 1987.

12. Coleman,D.J., Hunter,W.R., Brown,G.A., Chen,I.C., "Extensions of the Effective Thickness Theory of Oxide Breakdown", 27th Annual Proceedings, Reliability Physics 1989, pp.39-42, 1989.

13. Amerasekera,E.A., Campbell,D.S., "Charge Limited Breakdown in MOS Capacitors", Proc. 17th Ann. Eur. Sol. St. Dev. Res. Conf. (ESSDERC), pp.733-41, 1987.

14. Nicollian,E.H., Brews,J.R., "MOS Physics and Technology", John Wiley & Sons (New York), 1982.

Paper III

Tunnicliffe,M.J., Dwyer,V.M., Campbell,D.S.,

"A Model for ESD Oxide Breakdown
in Unprotected MOS Transistors",

European ESD Symposium, Eindhoven, Netherlands, 3rd.-5th. Sept.,
1991.

A Model for ESD Oxide Breakdown in Unprotected MOS Transistors

M.J.Tunnicliffe, V.M.Dwyer, D.S.Campbell.

International Electronics Reliability Institute,
Department of Electronic and Electrical Engineering,
Loughborough University of Technology,
Loughborough, Leicestershire, LE11 3TU, United Kingdom.
Tel. 0509 222855, Fax. 0509 222854.

Abstract

Experimental and theoretical studies on the ESD sensitivity of unprotected MOS transistors are presented. An experimental examination of short time-scale Fowler-Nordheim injection shows an apparent time delay between the application of stress voltage and the onset of tunnelling. A model based on transient thermal emission of trapped electrons is developed to explain this phenomenon. This model, combined with the 'causal' theory of time dependent dielectric breakdown (TBBD) is used to develop a model of fast transient/ESD oxide breakdown under negative polarity conditions. This model can be extended to embrace positive polarity ESD breakdown, as well as EMP and other spurious pulse phenomena.

Symbols not defined in text

h	Plank's constant ($6.63.10^{-34}$Js).
k	Boltzmann's constant ($1.38.10^{-23}$J/K).
m_0	Electron rest mass ($9.11.10^{-31}$kg).
q	Electronic charge ($1.6.10^{-19}$C).
ϵ_0	Permittivity of free space ($8.85.10^{-14}$F/cm).
ϵ_{ox}	Relative permittivity of SiO_2 (3.9).
ϕ	Si-SiO_2 potential barrier (3.2eV).
$\hbar$	$h/2\pi$ ($1.06.10^{-34}$Js).

1. Introduction

The miniaturisation of metal oxide semiconductor (MOS) transistors requires the production of thinner and thinner gate oxides. Ultra thin oxides are highly susceptible to electrostatic discharge (ESD) damage and can fail catastrophically at static potentials as low as 100V. For this reason, a considerable amount of research has recently been performed on protection circuit optimisation [eg.1]. However, only a few workers have investigated the basic physical processes of fast-pulse/ESD oxide breakdown [eg.2,3], which are still poorly understood. An improved understanding of these mechanisms may contribute to the design of better protection circuits.

Several models of oxide breakdown under long time-scale stress have been developed [eg.4,5], and extended to fast-pulse and ESD conditions [eg.2,6,7]. However, several observed phenomena remain unexplained by these theories. This paper continues the work on a deeper, more quantitative level and develops a model better suited to the short time-scale conditions associated with ESD and electromagnetic pulse (EMP) stress.

2. Theory

Several conflicting models of oxide dielectric breakdown have been proposed. It is generally agreed, however, that electrons tunnelling from the cathode continuously erode the dielectric over a period of time. This leads eventually to unstable current acceleration, causing thermal dissociation of the dielectric. This is called 'time dependent dielectric breakdown' or TDDB.

The erosion may involve a 'tree' of high trap density, growing from the anode toward the cathode, causing breakdown when it spans the dielectric [5,8]. Alternatively, the continuous trapping of electrons may distort the internal oxide field profile, producing high localised fields which rupture the SiO_2 network [9]. The injected electrons may also induce holes in the dielectric by band-to-band impact ionisation (BBII) [4] or by surface plasmon interactions at the anode [10]. Some of these holes are trapped in surface states near the SiO_2/cathode interface, presenting an areal charge density Q_p. When Q_p reaches a critical value Q_p^*, the local electric field is enhanced, increasing the electron injection rate and accelerating the process toward breakdown [4].

For the purposes of this paper the latter model is

assumed (although any model of current-activated erosion would suffice), giving the following condition for breakdown

$$\eta \int_0^{t_{bd}} J \cdot \alpha \, dt = Q_p^* \tag{1}$$

where t_{bd} is the time delay to breakdown after the application of stress, α is the hole generation efficiency (the number of generated holes per injected electron), η is the hole trapping efficiency (the number of trapped holes per generated hole) and J is the tunnelling injection current. If the holes are generated by impact ionisation then α is related to the electric field F by the equation

$$\alpha(F) = \alpha_0 e^{-\frac{H}{F}} \tag{2}$$

where α_0 and H are constants. α is also known to increase with increasing oxide thickness [2].

At this stage, a constant electric field $F=V_{ox}/T_{ox}$ (V_{ox}=oxide voltage) is assumed throughout the dielectric until the trapped charge density reaches Q_p^*. The tunnelling current density J is therefore given by the Fowler-Nordheim equation [11]

$$J = \lambda F^2 e^{-\frac{B}{F}} \approx J_0 e^{-\frac{B}{F}} \tag{3}$$

where

$$\lambda = \frac{q^3 m_{si}^*}{8\pi h \phi m_{ox}^*} \quad ; \quad B = \frac{4\sqrt{2m_{ox}^*}\phi^{\frac{3}{2}}}{3\hbar q} \tag{4}$$

m_{si}^* is the effective electron mass in Si ($1.1m_0$), m_{ox}^* is the effective electron mass in the SiO_2 *forbidden gap* and J_0 is approximately constant over a limited field range. Fitting Eqn.(4) to an experimental J vs. F characteristic requires an m_{ox}^* of approximately $0.362m_0$ [12] (various values between $0.3m_0$ and $0.5m_0$ are reported throughout the literature). This is considerably lower than the $m_{ox}^*=1.0m_0$ of the SiO_2 *conduction band* [13] or the $m_{ox}^*=0.96m_0$ measured from Lenzlinger and Snow's temperature dependent tunnelling data [14]. Although the classical image-force accounts for the discrepancy [12], this explanation is discarded in the present paper in accordance with Hartstein & Weinberg's quantum-mechanical argument [15].

If the Fowler-Nordheim equation is combined with Eqns.(1) and (2), the breakdown condition can be written

$$\int_0^{t_{bd}} \frac{dt}{\tau(F)} = 1 \tag{5}$$

where $\tau(F)$ is the value of t_{bd} for a constant field F, given by

$$\tau(F) = \tau_0 e^{-\frac{\gamma}{F}} \tag{6}$$

where τ_0 and γ are constants.

Two distinct modes of breakdown have been identified, ie. *intrinsic* and *extrinsic*. Intrinsic breakdown is governed by the inherent properties of the MOS structure while extrinsic breakdown is associated with structural defects [16]. Intrinsic breakdown in any given structure has a constant t_{bd} for a given F, allowing it to be characterised by a constant τ_0 and γ. Extrinsic breakdown has a shorter t_{bd}, which may be characterised by an effective oxide thickness $T_{eff}<T_{ox}$ [17]. Extrinsic breakdown varies randomly between devices, generally following a Poisson distribution [16].

Under fast transient conditions, the J vs. F characteristic departs from the simple Fowler-Nordheim model [7] and Eqn.(6) cannot be directly applied. The examination of this departure and the development of a modified tunnelling model suitable for fast transient conditions are amongst the main objectives of this paper.

3. Experimental Samples

The samples used in these experiments were NMOS transistor and capacitor structures, fabricated specifically for process characterisation. All devices had 40nm oxide thickness, $10^{21}cm^{-3}$ n^+-type gate doping and $4.10^{15}cm^{-3}$ p-type substrate doping. No implant doping was present. Several different gate areas were used and the relevant dimensions are quoted where necessary throughout the paper. The devices were tested at wafer-level using a microprober system at room temperature. All electrical connections were made between the gate electrodes and the metallised wafer base, leaving source and drain contacts floating.

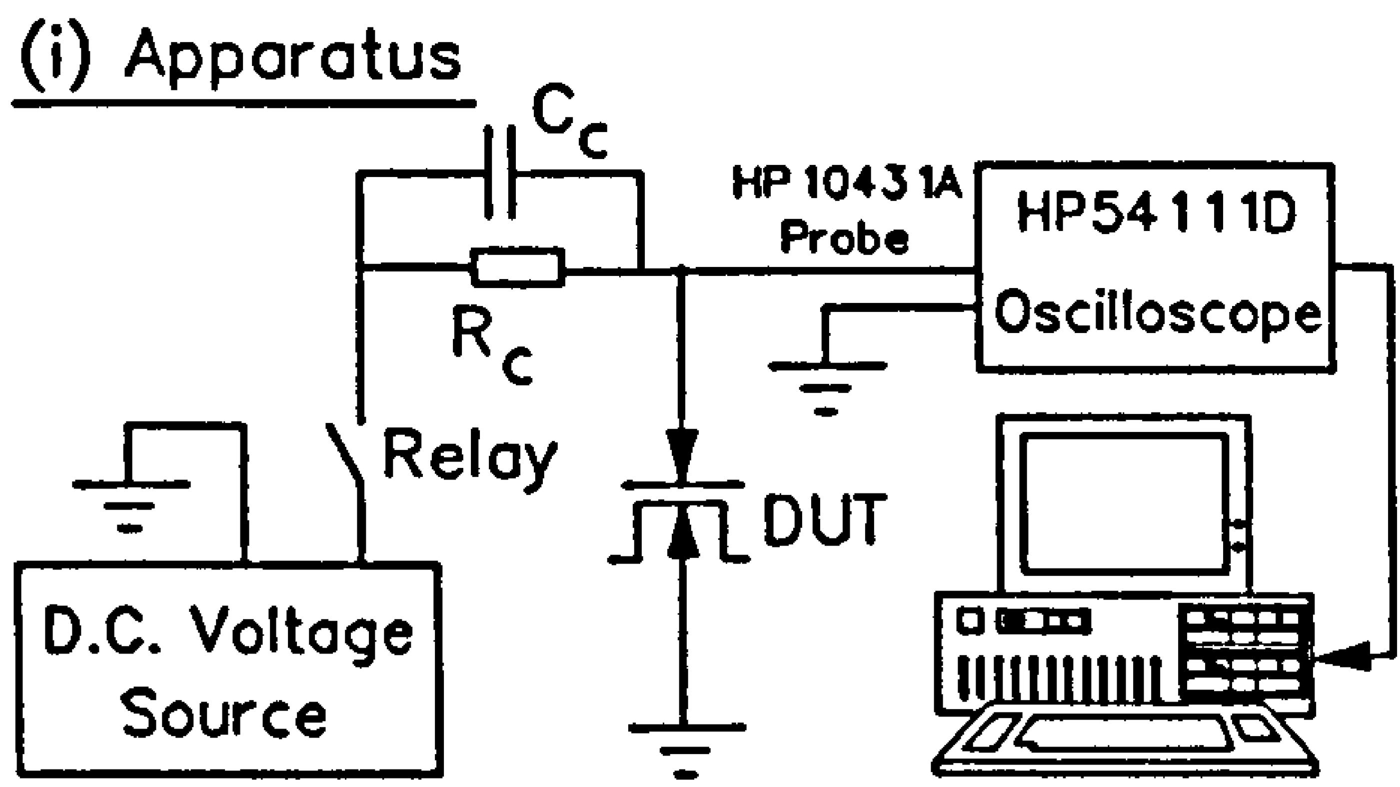
4. Short Time-Scale Fowler-Nordheim Tunnelling

4.1 Experiment

Under very fast transient voltage stress, the tunnelling current in an MOS oxide may be several

orders of magnitude smaller than the displacement current $\epsilon_0\epsilon_{ox}.dF/dt$. It is therefore very difficult to observe using standard techniques.

Fig.1(i) shows an experiment devised to study tunnelling under fast transient conditions. (This technique was first introduced in an earlier paper by the authors [7]). Fig.1(ii) shows a simple equivalent circuit model for the apparatus. While C_c and R_c are intentional circuit elements, R_{pr} is the resistance of the oscilloscope voltage probe. C_{par} represents the total capacitance appearing directly in parallel with the oxide, consisting of the oxide capacitance, the capacitance of the oscilloscope probe and any other stray capacitances.



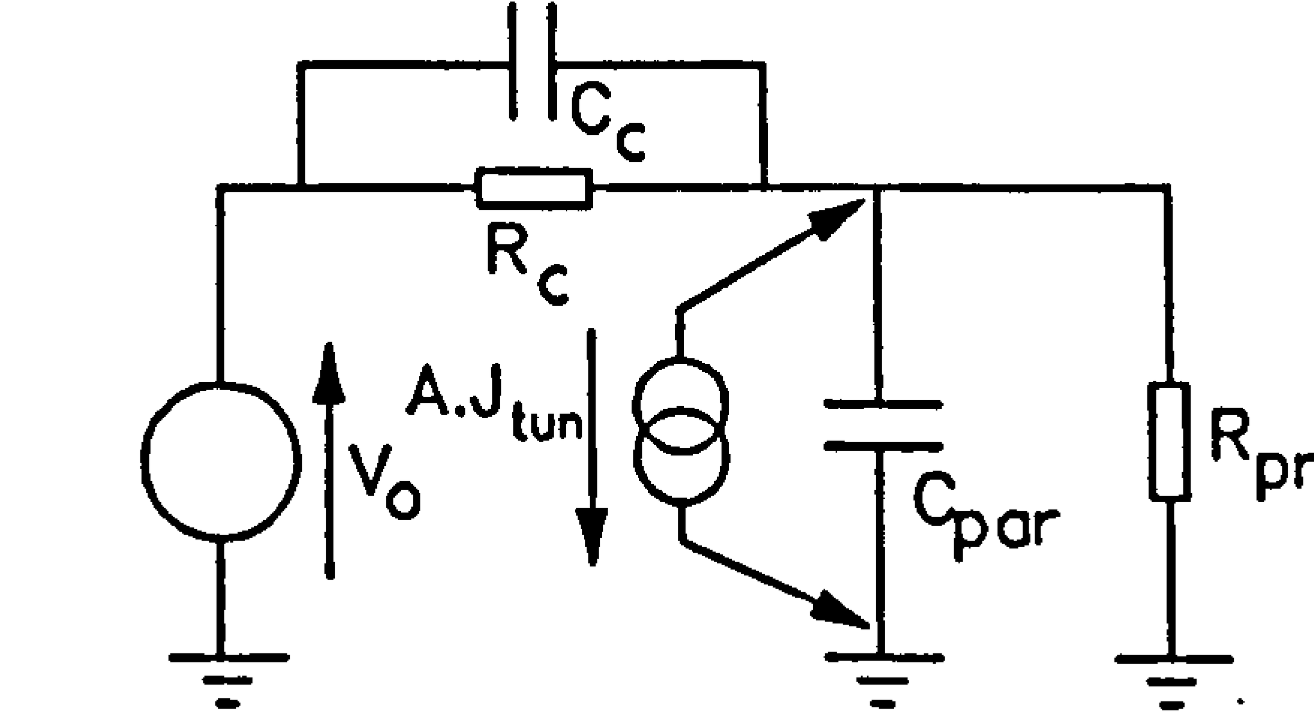


Fig.1: Apparatus and Equivalent Circuit Model for Short Time-Scale Tunnelling Experiment.

The operation of the circuit is as follows: As the relay is closed, the voltage V_0 is applied to the potential divider consisting of C_c and C_{par}. This causes a voltage $V_{ox} = V_0C_c/(C_c+C_{par})$ to appear across the device under test (DUT). If the potential divider formed by R_c and R_{pr} is adjusted such that $R_{pr}/(R_c+R_{pr}) = C_c/(C_c+C_{par})$ then charge leakage from C_{par} via R_{pr} is replenished by the current in R_c and V_{ox} remains constant. If, however, a tunnelling current path opens across C_{par} then the equilibrium is disturbed and V_{ox} decays with time. The $V_{ox}(t)$ profile is captured on a digital oscilloscope and downloaded to a desktop computer for storage and analysis.

The success of this experiment depends upon obtaining a sufficiently low value of C_{par}. If C_{par} is too high then the decay of $V_{ox}(t)$ is too slow to be detected. Additionally, the maintenance of a high field in the presence of tunnelling leads to rapid breakdown [see Eqn.(1)], preventing the sample from being re-tested. For this reason, only small dimension MOSFETs are suitable for this experiment.

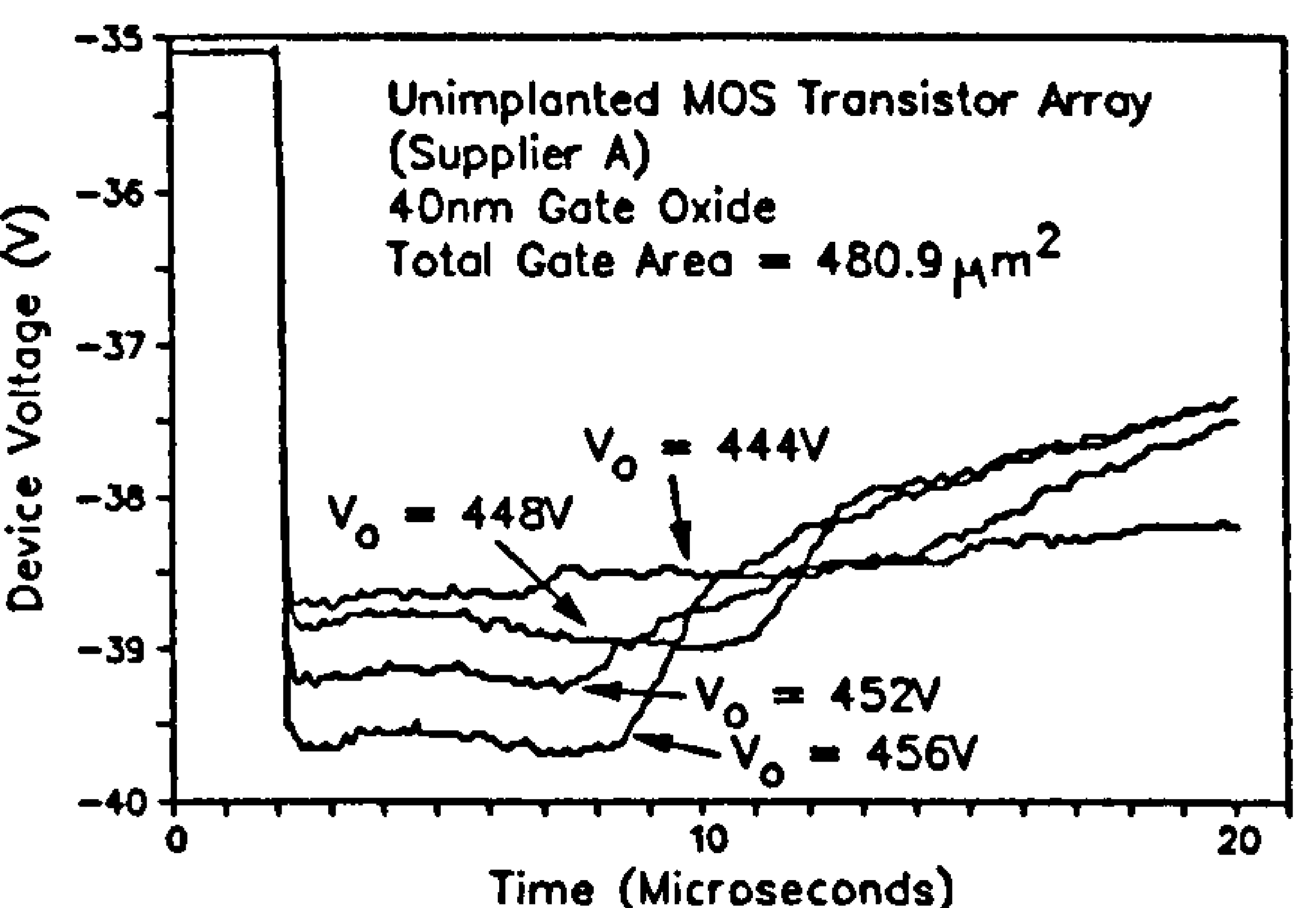


Fig.2: Experimental $V_{ox}(t)$ Profiles.

4.2 Results and Discussions

Fig.2 shows a typical set of waveforms obtained from the above experiment using small dimension transistor structures. The n^+-type polysilicon gate was biased negatively with respect to the p-type substrate, allowing rapid majority carrier accumulation at both oxide surfaces and eliminating any depletion layers in series with the oxide. The waveforms suggest a rapid rise of the tunnelling current after a time delay Δt.

The experimental Δt is many orders of magnitude greater than the traversal time of electrons in the tunnelling barrier [18], so a quantum-mechanical cause is unlikely. Alternatively the time constant of the oxide capacitance C_{ox} and the substrate resistance R_s may be responsible for the delay. This hypothesis is difficult to test, since the circuit elements are too small for accurate measurement by standard impedance analysis equipment. However, measurement of the gate-substrate resistance of a broken-down transistor (which can be used as an upper limit for R_s) and calculation of $C_{ox} = \epsilon_0\epsilon_{ox}A/T_{ox}$ shows that $C_{ox}R_s$ is far too small to account for the results.

Earlier work assumed that the oxide current changed abruptly from zero to the value predicted by the Fowler-Nordheim equation after a time delay Δt [7]. Examination of Fig.2 however, shows that this is unlikely to be the case. The final value to which V_{ox} settles seems to decrease as the maximum value of V_{ox} increases. This suggests that the mechanism is history dependent. Thus during the decay, the oxide 'remembers' the value of V_{ox} during the initial dwell period Δt and adjusts the current accordingly. Hence the tunnelling current lags behind the oxide field, as if the oxide were behaving as a kind of inductance.

4.3 Framework for Modelling

According to the circuit model of Fig.1(ii), the apparatus can be modelled using the differential equation

$$\frac{1}{R_c}(V_0-V_{ox})-(C_c+C_{par})\frac{dV_{ox}}{dt} = A\cdot J(t)+\frac{V_{ox}}{R_{pr}} \tag{7}$$

where A is the oxide area. It will be assumed that J is dictated by an 'effective' oxide field $F_c(t)$, related to J(t) by the standard Fowler-Nordheim equation, ie.

$$J = \lambda F_c^2 e^{-\frac{B}{F_c}} \tag{8}$$

The time-lag hypothesis of Section 4.2 suggests a first-order dynamical relationship between $F_c(t)$ and the overall field $F = V_{ox}(t)/T_{ox}$.

Numerical solution of Eqn.(7) using a trial form for the F/F_c relationship yields a theoretical $V_{ox}(t)$ profile which can be compared with the experimental results. Some features of the F/F_c relationship may be deduced by direct inspection of Fig.1. Firstly the monotonic decay of $|V_{ox}(t)|$ suggests that the process of Eqn.(7) is not quickly reversible. Secondly, the rapid decrease of Δt with increasing V_{ox} indicates strong field dependence.

These features suggest a transient distortion of the oxide field system due to de-trapping. Theoretical studies by Avni and Shappir [5] show that an electrically neutral oxide can become positively charged under high fields by impact-induced emission of trapped electrons (ie. trap-to-band impact ionisation or TBII). Since very few mobile electrons are available during Δt, a model based on spontaneous de-trapping is suggested. Trapped electrons can be spontaneously emitted by trap-band tunnelling or by the Poole-Frenkel effect, ie. thermal excitation over a field-lowered trap-band potential barrier [19]. For the purposes of this paper the latter mechanism is assumed.

4.4 Physical Model

For simplicity, a homogeneous two-dimensional sheet of traps is assumed. This sheet contains N_0 traps/cm^2 and is situated at a distance x_n from the anode. Although prolonged charge injection generates further traps in the oxide [4,5,8,9], N_0 can be assumed to be constant over the short time periods involved in this experiment.

At zero field, the equilibrium trapped electron density n_0 is compensated by a fixed positive charge density $q.p_0$, leaving the oxide electrically neutral. The trap occupancy can be modelled by the first order rate equation, governing the opposing effects of trapping and de-trapping

$$\frac{dn}{dt} = \left(\frac{J}{q}\sigma + \frac{1}{\tau_f}\right)(N_0 - n) - \frac{n}{\tau_e} \tag{9}$$

where n is the density of filled traps, σ is the capture cross section, τ_e and is the de-trapping time constant and τ_f is the zero-current trap filling time constant. The apparent non-reversibility of the process implies that the last term in Eqn.(9) should predominate, making dn/dt always negative.

The value of τ_e is assumed to be governed by the overall oxide field F. If the Poole-Frenkel effect is responsible for de-trapping then Hartke's three-dimensional emission model [19] can be used, ie.

$$\frac{\tau_e(0)}{\tau_e(F)} = \frac{1}{2} + \left[\frac{kT}{\beta\sqrt{F}}\right]^2\left[1+\left(\frac{\beta\sqrt{F}}{kT}-1\right)\exp\left(\frac{\beta\sqrt{F}}{kT}\right)\right] \tag{10}$$

where $\beta=(q^3/\pi\epsilon_0\epsilon_{ox})^{1/2}$ and $\tau_e(0)$ is the zero-field de-trapping time-constant. If $n=p_0$ at zero field and current (see above) then τ_f must be equal to $\tau_e(0)(N_0/p_0-1)$. The effective field F_c governing the tunnelling injection is defined as the field in the barrier region, which, according the Poisson equation, is given by

$$F_c = F + \frac{q\,x_n}{\epsilon_0\,\epsilon_{ox}\,T_{ox}}(p_0 - n) \tag{11}$$

Strictly speaking, the presence of charge $q(p_0-n)$ in the oxide invalidates Eqn.(7). However, simple calculations show that Eqn.(7) remains an adequate approximation.

The static J vs. F characteristic can be computed by setting dn/dt in Eqn.(9) to zero and performing a simultaneous numerical solution of Eqns.(8) to (11). Fig.3 shows the results obtained from sensible parameter values ($x_n=0.3T_{ox}$, $p_0=10^{14}$cm^{-2}, $N_0=10^{15}$cm^{-2}, $\sigma=10^{-15}$cm^2), using $m_e^*=m_0$ and adjusting $\tau_e(0)$ such that the relaxation time $5\tau_e$ is of the same order as Δt (10μs) at 10MV/cm (see Fig.3). The computed data clearly provide a plausible approximation to the Fowler-Nordheim model with $m_e^*=0.318m_0$. This is approximately compatible with the results of Kreiger and Swanson [12], which require $m_e^*=0.362m_0$.

Fig.4 shows the $V_{ox}(t)$ computed by simultaneous numerical solution of Eqns.(7) to (11). The profile clearly exhibits an approximately constant V_{ox} followed by an abrupt decay, in good qualitative agreement Fig.2.

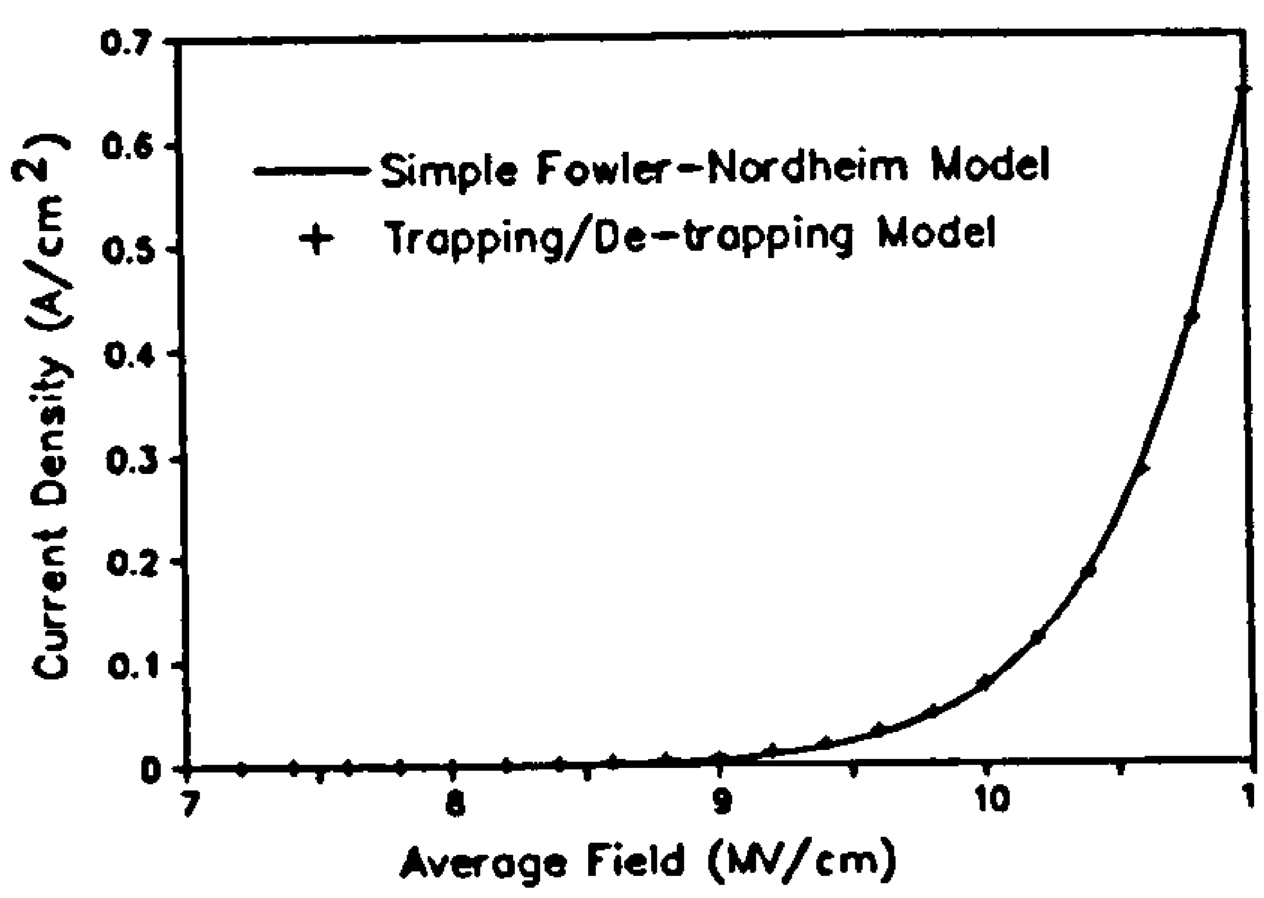


Fig.3: J vs. F data obtained from Eqns.(8) to (11) [$m_e^*=m_0$], compared with standard Fowler-Nordheim model [$m_e^*=0.318m_0$].

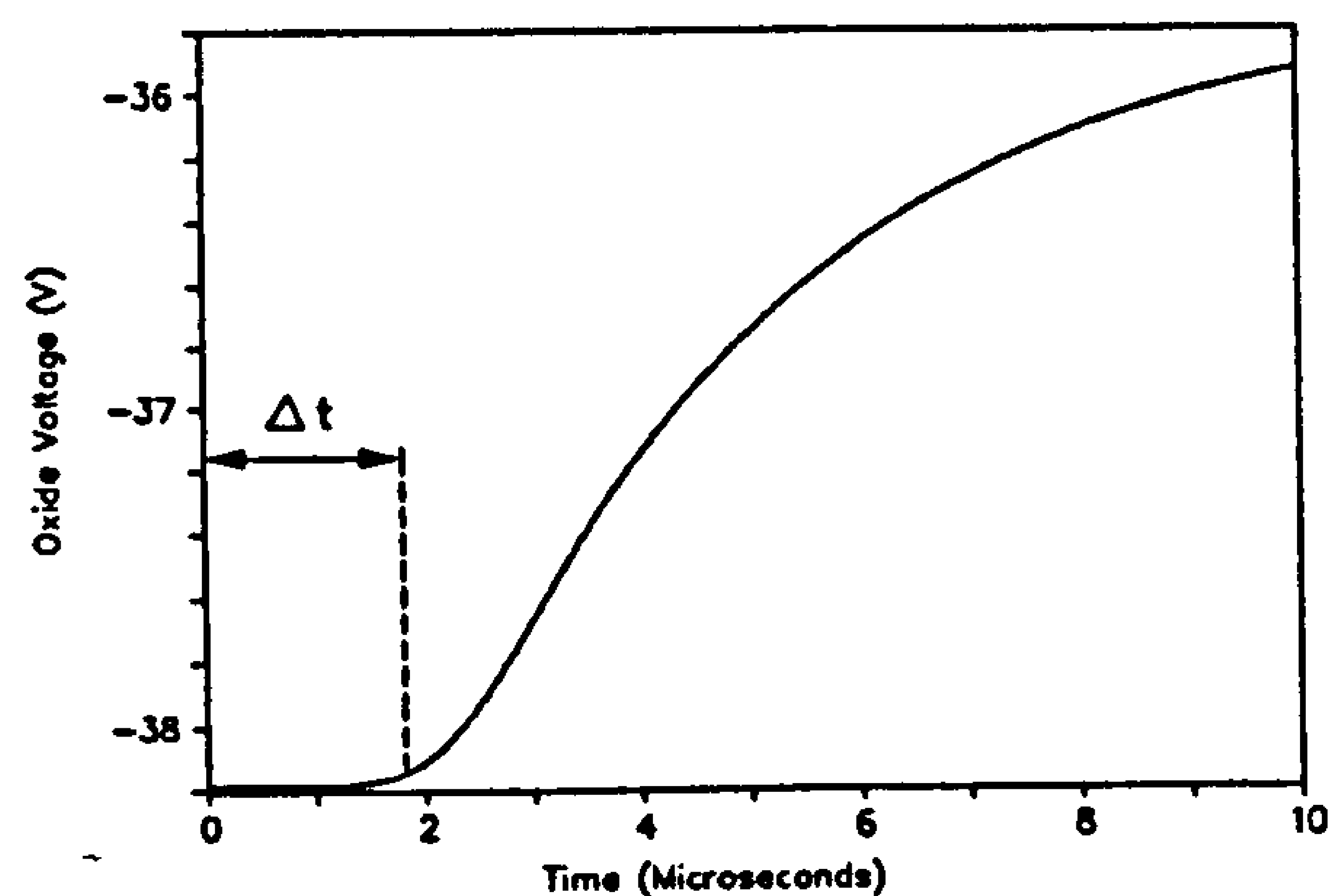


Fig.4: Transient oxide voltage profile obtained from numerical solution of trapping/de-trapping equations. (Compare with Fig.2.)

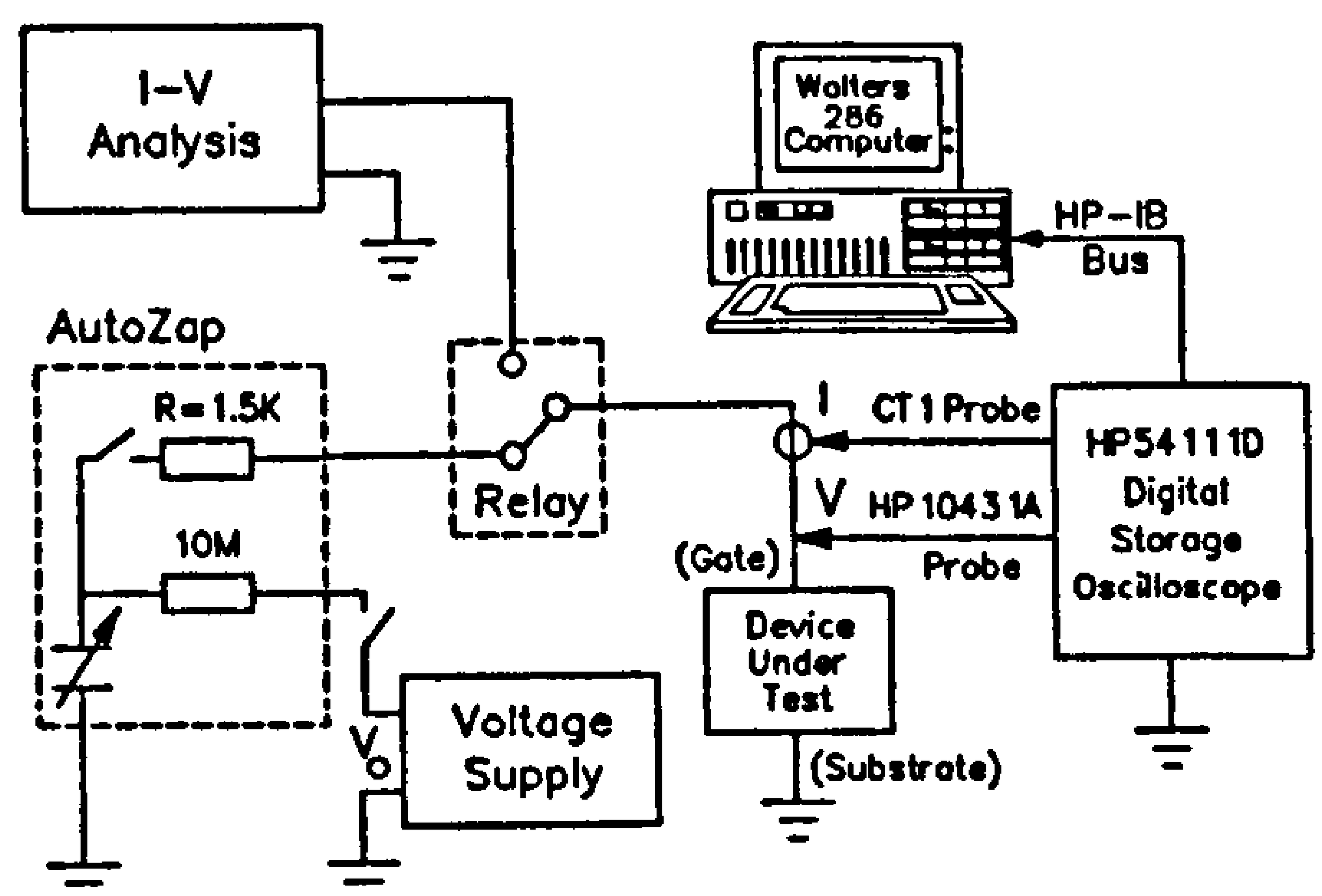


Fig.5: Apparatus used in ESD experiments.

5. ESD Breakdown

5.1 Experiment

Fig.5 shows the ESD apparatus used. Human body model (HBM) ESD pulses were supplied by a Hartley 'AutoZap' ESD simulator with a variable discharge capacitance C_l. C_l was charged to a given voltage V_0 and discharged via the resistance R (1.5KΩ) into the device under test. The discharge module was mounted close to the wafer microprober in order to minimise parasitic elements.

Voltage/current waveforms could be captured by an HP54111D digital storage oscilloscope and downloaded to a desktop computer. Analysis of these waveforms allowed the ESD system to be characterised in terms of a circuit model (see below).

Oxide breakdown was detected by measuring the I/V characteristics of each oxide before and after pulsing. Devices were stressed between gate and substrate, with all other contacts floating.

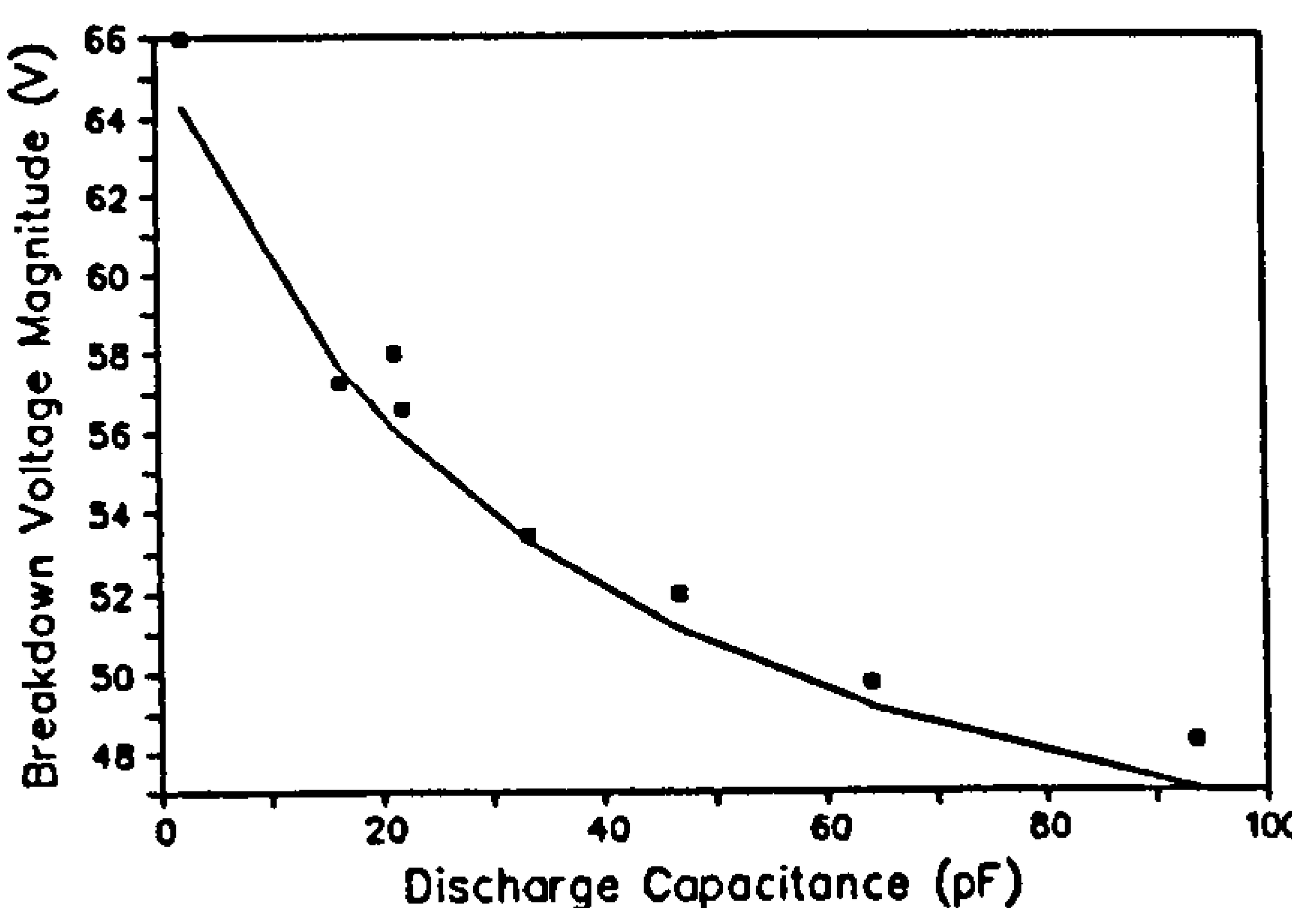


Fig.6: Negative polarity ESD threshold data.

5.2 Negative Polarity ESD Breakdown

Negative polarity breakdown thresholds V_{bd} were measured by applying pulses of increasing magnitude (2V between pulses), until oxide breakdown was observed. Although the voltage probe was necessary for circuit characterisation, it was removed in these experiments in order to minimise unnecessary circuit elements. Fig.6 shows the negative polarity breakdown threshold as a function of the discharge capacitance C_l. The results are clearly consistent with earlier investigations, showing a decreasing breakdown voltage threshold with increasing C_l [20].

5.3 ESD Circuit Model

In a p-substrate MOS device under negative polarity stress, accumulation occurs at both oxide surfaces and the vast majority of the device voltage appears across the oxide capacitance C_{ox}. Fig.7 shows an equivalent circuit for the ESD system under negative polarity conditions, showing the following parameters:

C_x	Stray capacitance of discharge module.	35pF.
C_t	Capacitance of transmission system.	20pF.
R	AutoZap 'body' resistor.	1.5KΩ.
C_p	Probe capacitance.	8pF.

R_p Probe resistance. 1MΩ.

The last two elements do not apply to the ESD threshold experiments since the voltage probe was unconnected.

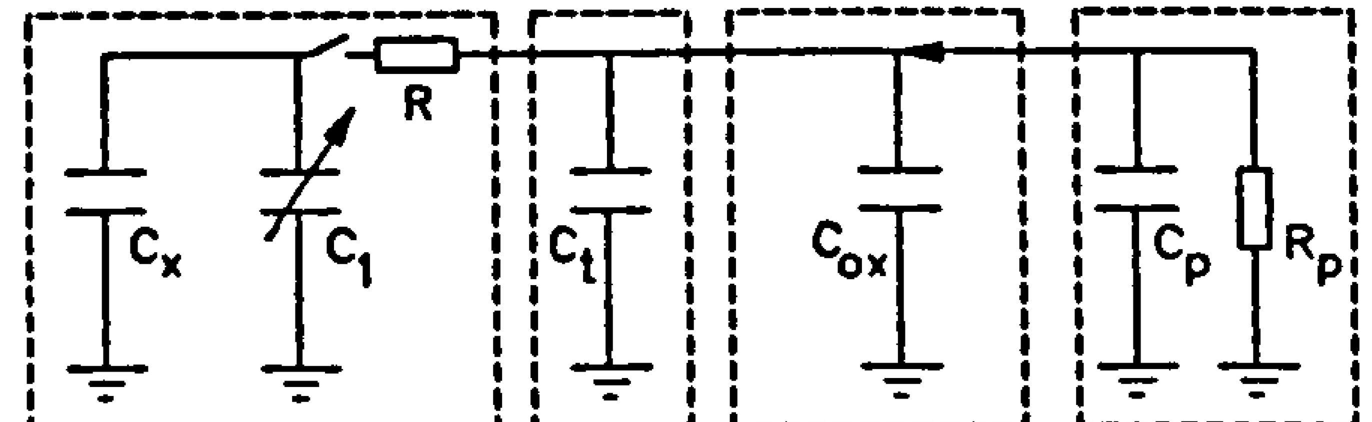


Fig.7: Negative polarity equivalent circuit model for negative polarity ESD experiment.

The *effective* voltage V_e of an ESD pulse is the voltage which actually appears across the device under test. This differs from the *applied* voltage V_0 because of the capacitive loading effect of the oxide [20] and the packaging and pulse transmission media [6,7]. For the equivalent circuit of Fig.7 (without C_p and R_p), V_e is given by

$$V_e = V_0\left[\frac{C_x + C_1}{C_x + C_1 + C_t + C_{ox}}\right] \quad (12)$$

When the pulse is applied, the device voltage rises to V_e with a characteristic rise-time constant τ_r given by

$$\tau_r = \frac{R}{\dfrac{1}{C_x+C_1} + \dfrac{1}{C_t+C_{ox}}} \quad (13)$$

In the present experimental conditions, τ_r cannot be more than a small fraction of a microsecond. Since this is far smaller than the emission time constant τ_e (Section 4.4), the oxide voltage can always be assumed to have reached V_e before the onset of significant tunnelling (this has also been shown experimentally [6]). For tunnelling currents of the order of μA, the voltage across R can be neglected and the situation resolves into a single capacitor $C=(C_x+C_1+C_t+C_{ox}+C_p)$, charged initially to V_e and discharging into the tunnel junction.

5.4 Analytical ESD Model

An earlier publication assumed an abrupt increase in tunnelling current from zero to the value predicted by Eqn.(3) at time Δt after the application of stress [7]. This yields an analytical field profile F(t) which can be combined with the theory of Section 2 to predict the breakdown condition. For this purpose it is convenient to define a transient *oxide wearout factor* $s(t)=Q_p/Q_p^*$ predicting breakdown when $s(t_{bd})=1$. According to the analytical F(t) profile, s(t) is given by [7]

$$s(t) = \frac{-CT_{ox}}{\tau_0 A\lambda H}\left[\left(\frac{A\lambda B(t-\Delta t)}{CT_{ox}} + e^{\frac{BT_{ox}}{V_e}}\right)^{-\frac{H}{B}} - e^{\frac{-HT_{ox}}{V_e}}\right] \quad (14)$$

where $H=\gamma$-B (Note: V_e denotes the effective ESD voltage *magnitude* and is positive irrespective of stress polarity). The value of s(t) increases monotonically, saturating towards a value $S(V_e)$ given by

$$S(V_e) = \frac{CT_{ox}}{\tau_0 A\lambda H}\exp\left[-\frac{HT_{ox}}{V_e}\right] \quad (15)$$

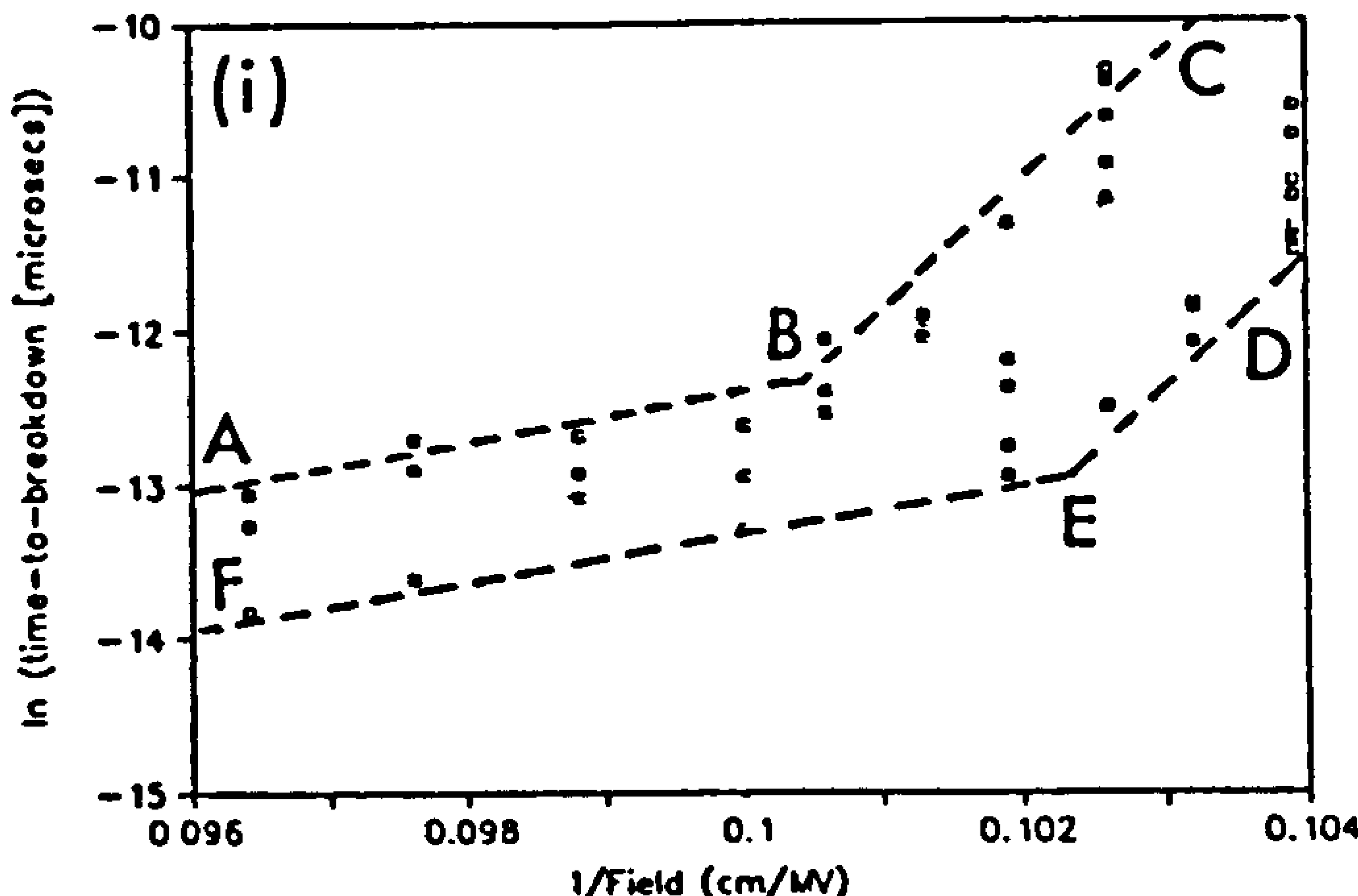
The effective ESD breakdown voltage V_e^* can therefore be defined such that $S(V_e^*)=1$, yielding the following expression

$$V_e^* = \frac{-HT_{ox}}{\log_e\left(\dfrac{A\lambda\tau_0 H}{CT_{ox}}\right)} \quad (16)$$

Fig.8(i) shows some constant-field time-to-breakdown data, plotted ln(time) vs. 1/Field. The data points form a distribution reflecting random oxide inhomogeneities, which lies roughly within the area ABCDEF. The graph has two characteristic regions: BCDE, in which breakdown is governed by causal wearout and ABEF ($t_{bd}<6\mu s$), which is dominated by the tunnelling time Δt. γ and τ_0 must be calculated from the former region [7]. Assuming that statistical variations between samples are exhibited in τ_0 rather than γ, γ=825.62MV/cm and $4.879.10^{-43}s<\tau_0<4.502.10^{-42}s$. These values, together with λ and B appropriate for $m_{ox}^*=0.318m_0$, allow V_e^* to be calculated as a function of C_{ox} [Fig.8(ii)]. The variation of V_e^* over the full range of C_1 is clearly small, showing that ESD breakdown can be accurately characterised in terms of a single effective breakdown voltage V_e^*. V_e^* is also relatively insensitive to the statistical variations between devices, varying as little as 3.5% over the full range of τ_0.

5.5 Numerical ESD Model

The analytical model rests upon the assumption that no tunnelling and no oxide degradation occur in the interval 0<t<Δt. According to Section 4.4 however, a small current *does* flow during this time. Since this current is too small to produce a significant oxide charge depletion it does not register



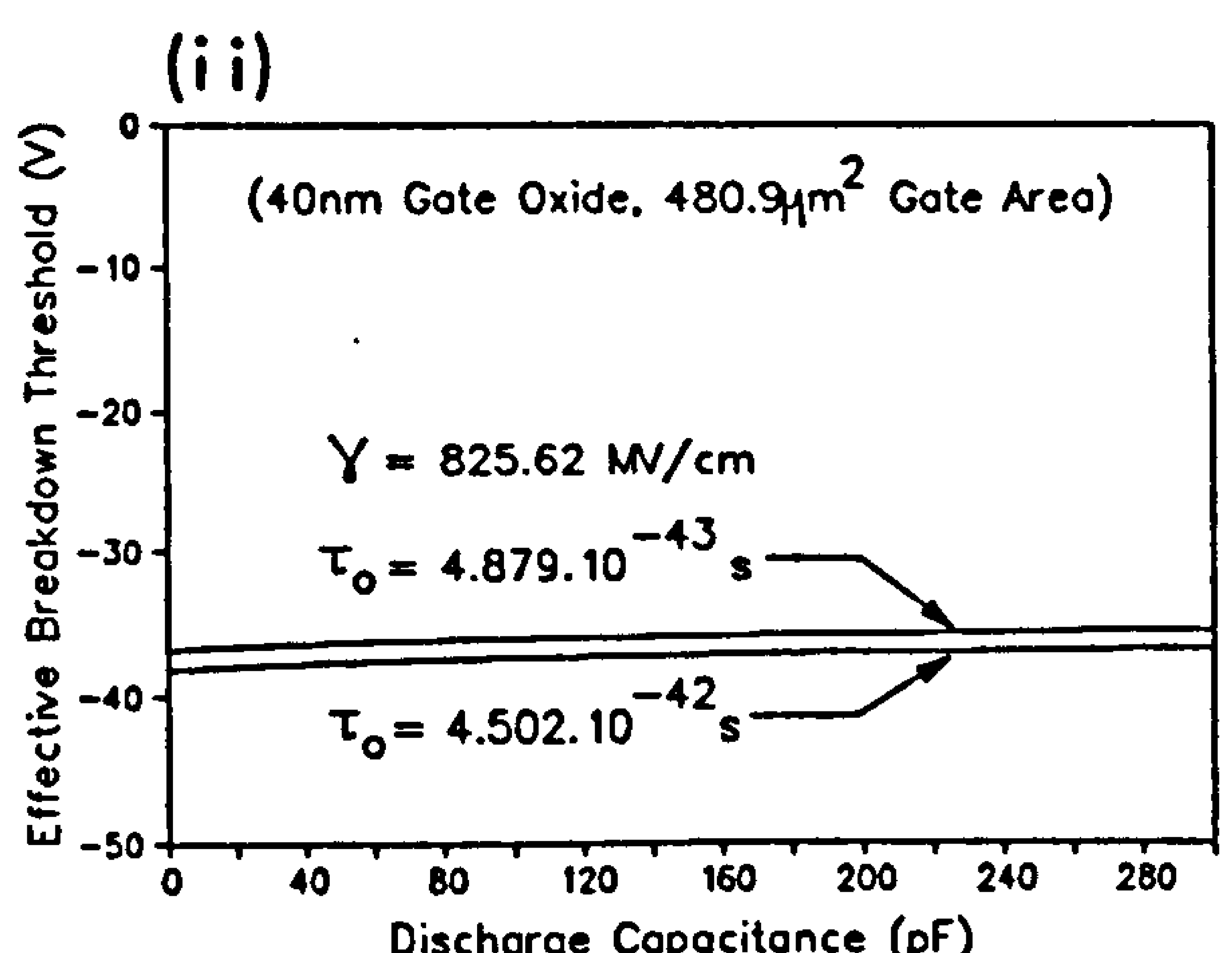


Fig.8: (i) Constant-field breakdown-time data (D-mode NMOS), (ii) Theoretical effective breakdown threshold as a function of C_1.

in the experiment of Section 4.1. However, it does not necessarily follow that the current is unable to degrade the oxide during this time. The assumptions used in the determination of Eqns.(14-16) may therefore be unsound.

An alternative approach is to solve Eqns.(8) to (11) simultaneously with the equation

$$C\frac{dF}{dt} = -J(t) \qquad (17)$$

to obtain the F(t) and J(t) profiles. These can then be combined with Eqn.(1) in order to obtain the $s(t)=Q_p/Q_p^*$ profile which is compared with the analytical version in Fig.8. This clearly illustrates that degradation during $0<t<\Delta t$ is negligible compared to subsequent degradation and that the analytical model is adequate.

The analytical and numerical models both predict an almost constant value of V_e^* irrespective of the system capacitance C.

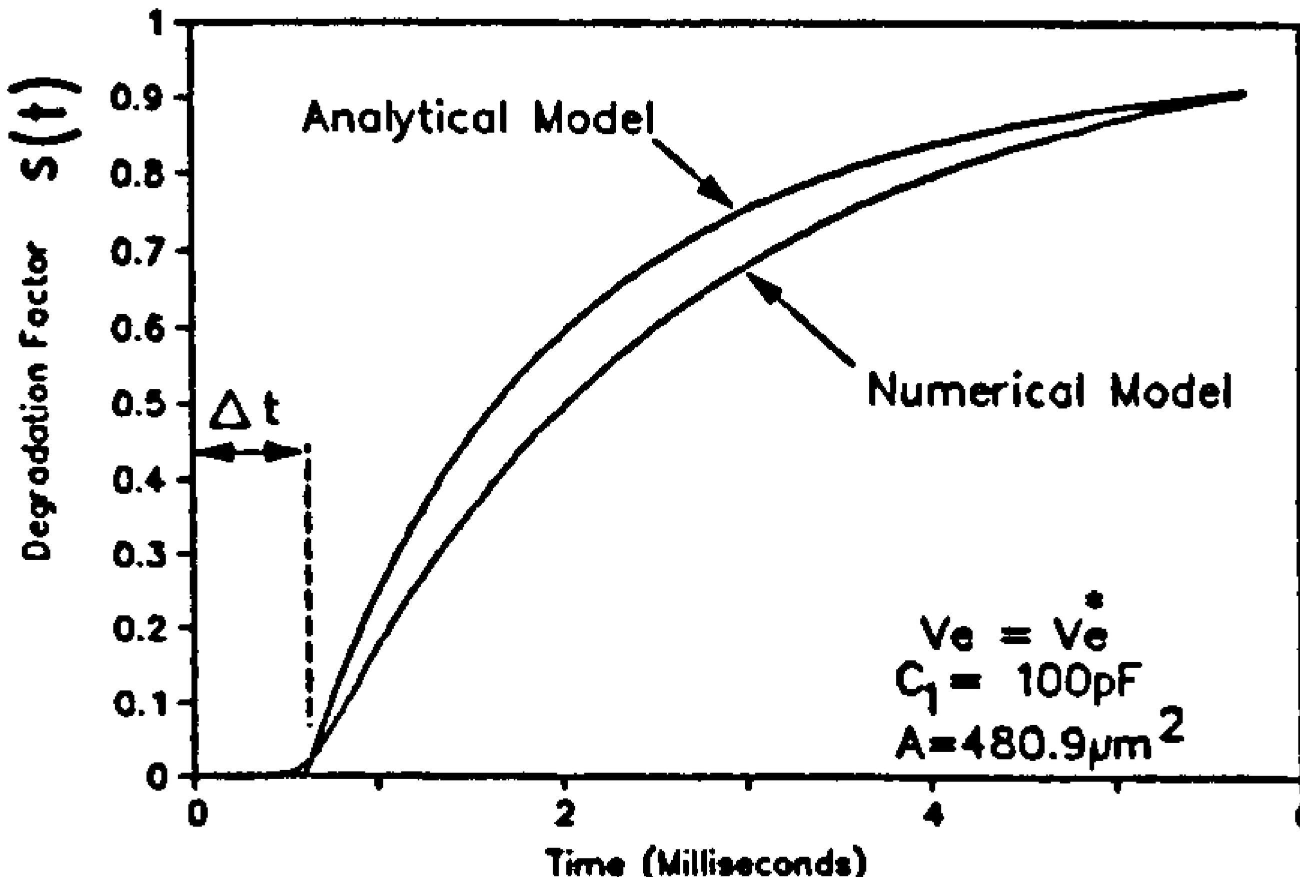


Fig.9: Transient degradation curves predicted by analytical and numerical models.

5.6 Constant Energy Effect

A curious feature of this model is that it produces, under some conditions, the illusion of a constant energy to breakdown. If the capacitance C_x was not known to exist, then the total electrostatic energy E might be supposed to be equal to $\frac{1}{2}C_1V_0^2$. If remains constant, then the V_{bd} vs. C_1 curve would have the form

$$V_{bdn} = \frac{const.}{C_1^{1/2}} \qquad (18)$$

Fig.10 shows this curve superimposed on the curve of Eqn.(12) for C_{ox}=40pF, V_e^*=-38V. The curves are practically coincident for a very wide range of C_1. It is probably this coincidence which led earlier workers to conclude that ESD breakdown is thermal rather than charge/field activated as proposed in this paper [20].

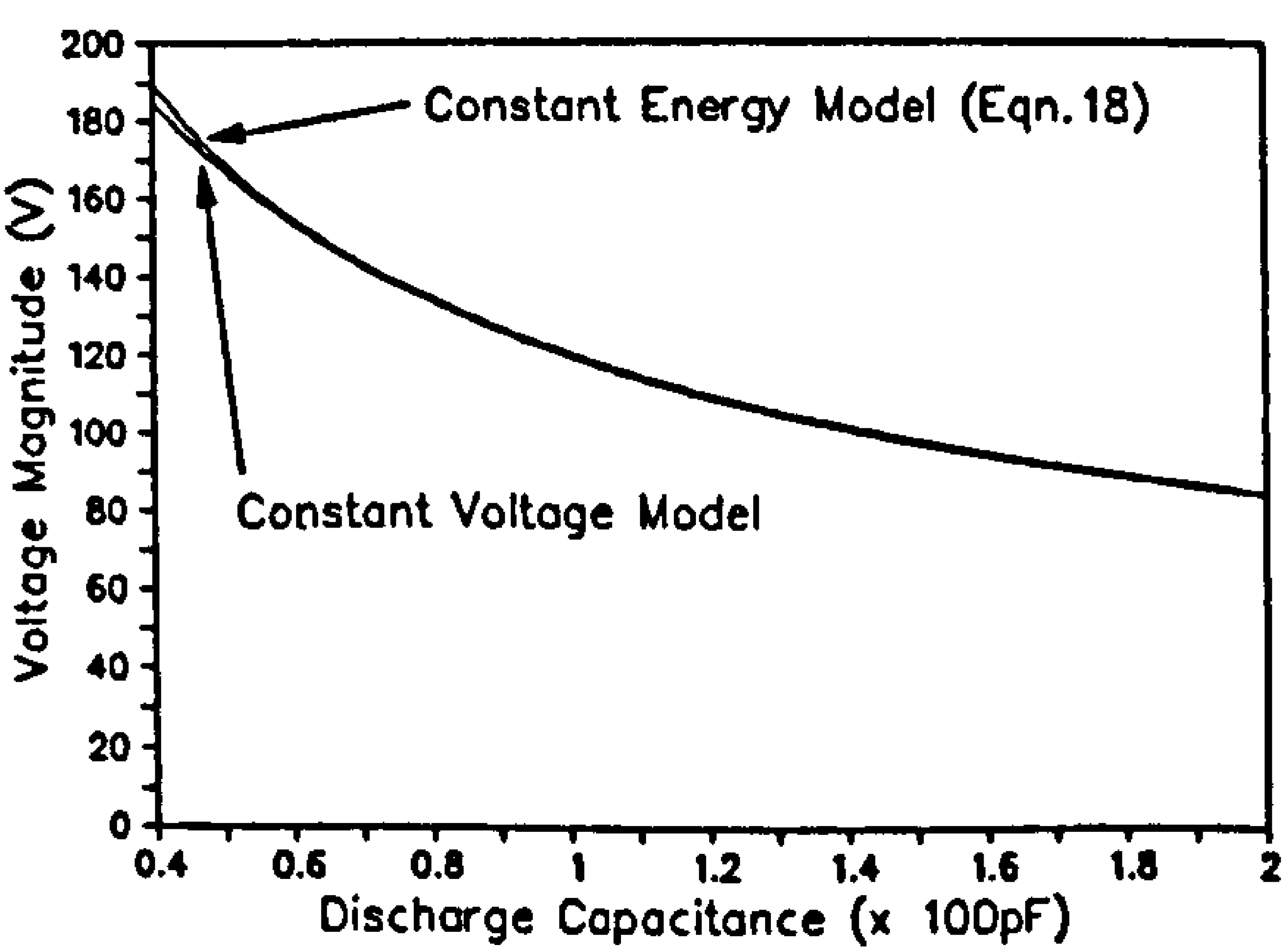


Fig.10: Comparison of constant energy and constant V_e^* profiles.

5.7 Positive Polarity ESD Breakdown

Early experiments showed that an oxide's ESD breakdown threshold is affected by the doping of the cathode material [3]. It was later suggested that positive polarity ESD breakdown in a p-substrate device can only result from depletion layer avalanche breakdown [20]. Subsequent experiments showed this theory to be correct [7].

The extension of the above model to positive polarity ESD requires not only the inclusion of the depletion layer avalanche voltage V_{av} in series with the device, but also a nonlinear resistance introduced by current construction around the avalanche filaments. This is a complex situation and will be addressed in a separate paper.

6. Conclusions

The ESD sensitivity of a variety of unprotected MOS structures has been studied theoretically and experimentally. The theoretical models, based upon the causal wearout of the oxide under Fowler-Nordheim tunnelling, provided accurate predictions of the observed breakdown phenomena.

The following important conclusions can be drawn from the work:

1. The anomalous 'tunnelling time delay' observed in earlier studies [7] can be modelled by field-dependent spontaneous emission of trapped electrons, causing space-charge evolution within the oxide. This model also permits equal effective electron masses to be assigned to the oxide conduction band and forbidden gap.

2. The above model of tunnelling combined with the causal theory of oxide breakdown predict an effective ESD breakdown threshold which is almost constant with changing capacitance. This fact, when combined with a circuit model of the ESD system, can create the false impression that breakdown is energy limited. This led earlier investigators to the false conclusion that ESD breakdown is thermally activated [20].

7. Acknowledgements

This work was conducted with the support of the Procurement Executive, Ministry of Defence, to whom the authors are grateful for permission to publish. The authors would also like to thank the other members of the International Electronics Reliability Institute at Loughborough University and Mr. David Shore of the MOD for many valuable discussions.

References

[1] Rountree,R.N., Hutchins,C.L., *"NMOS Protection Circuitry"*, IEEE Trans. Electron Dev., 32(5), pp.910-7, 1985.

[2] Fong,Y, Hu,C, *"The Effects of High Electric Field Transients on Thin Gate Oxide MOSFET's"*, Proc. 9th. EOS/ESD Symp., pp.252-7, 1987.

[3] Amerasekera,E.A., Campbell,D.S., *"ESD Pulse and Continuous Voltage Breakdown in MOS Capacitor Structures"*, Proc. 8th. EOS/ESD Symp., pp.208-13, 1986.

[4] Chen,I.C., Holland,S.E., Hu,C, *"Electrical Breakdown in Thin Gate and Tunnelling Oxides"*, IEEE Trans. Electron Dev., 32(5), pp.413-22, 1985.

[5] Avni,E, Shappir,J, *"A Model for Silicon Dioxide Breakdown under High Field and Current Stress"*, J.Appl.P., 64(2), pp.743-8, 1988.

[6] Tunnicliffe,M.J., Dwyer,V.M., Campbell,D.S., *"Slow Transient Voltage and ESD Breakdown in Unprotected MOS Gate Oxides"*, Proc. 1st. Components Engineering, Reliability and Test Conference (CERT), pp.78-87, 1990.

[7] Tunnicliffe,M.J., Dwyer,V.M., Campbell,D.S., *"Experimental and Theoretical Studies of EOS/ESD Oxide Breakdown in Unprotected MOS Structures"*, Proc. 12th. EOS/ESD Symp., pp.162-8, 1990.

[8] Wolters,D.R., van der Schoot,J.J., *"Dielectric Breakdown in MOS Devices: Part III: The Damage Leading to Breakdown"*, Phillips Journal of Research, 40, pp.164-92, 1985.

[9] Harari,E, *"Dielectric Breakdown in Electrically Stressed Thin Films of Thermal SiO_2"*, J.Appl.P., 49(4), pp.2478-89, 1978.

[10] Fischetti, *"High Field Electron Transport in SiO_2 and Generation of Positive Charge at the Si-SiO_2 Interface"*, Insulating Films on Semiconductors, pp.181-9, 1986.

[11] O'Dwyer,J.J., *"The Theory of Electrical Conduction and Breakdown in Solid Dielectrics"*, Clarendon Press, Oxford, pp.77-94, 1973.

[12] Kreiger,G, Swanson,R.M., *"Fowler-Nordheim*

Tunnelling in Thin Si-SiO$_2$-Al Structures", J.Appl.Phys., 52(9), pp.5710-7, 1981.

[13] Verwey,J.F., Amerasekera,E.A., Bisschop,J, *"The Physics of SiO$_2$ Layers"*, Rep. Prog. Phys., 53, pp.1297-331, 1990.

[14] Lenzlinger,M, Snow,E.H., *"Fowler-Nordheim Tunnelling in Thermally Grown SiO$_2$"*, J.Appl.Phys., 40(1), pp.278-83, 1969.

[15] Hartstein,A, Weinberg,Z.A., *"Unified Theory of Internal Photoemission and Phonon Assisted Tunneling"*, Physics Rev. B, 26(4), pp.1335-8, 1979.

[16] Wolters,D.R., van der Schoot,J.J., *"Dielectric Breakdown in MOS Devices: Part I: Defect Related and Intrinsic Breakdown"*, Phillips Journal of Research, 40, pp.115-36, 1985.

[17] Coleman,D.J., et al, *"Extensions of Effective Thickness Theory of Oxide Breakdown"*, Proc. 27th. IEEE Reliability Physics Symposium, pp.34-8, 1989.

[18] de Moura, M.A., de Albuquerque, D.F., *"Remarks on the Traversal Time in a Tunneling Process"*, Sol. St. Comm., 74(5), pp.353-4, 1990.

[19] Hartke,J.L, *"The Three-Dimensional Poole-Frenkel Effect"*, J.Appl.Phys., 39, pp.4871-3, 1968.

[20] Amerasekera,E.A., Campbell,D.S., *"Charge Limited Breakdown in MOS Capacitors"*, Proc. 17th. Annual European Solid State Device Research Conference (ESSDRC), pp.733-41, 1987.

Martin J.Tunnicliffe graduated from Bradford University, Bradford, West Yorkshire, U.K. in 1987, with a B.Eng. degree in Electrical and Electronic Engineering. He is currently a Research Assistant in the International Electronics Reliability Institute (IERI), Loughborough University of Technology, Loughborough, Leicestershire, U.K., where he is studying the effects of electrical overstress and ESD on MOSFET structures, and working toward a Ph.D. in 1992.

Vincent M. Dwyer graduated from Cambridge University, Cambridge, U.K. in 1981. He has conducted postgraduate research in Theoretical Surface Physics at York University, York, U.K. (leading to a D.Phil. degree), Trinity College, Dublin, Ireland and Warwick University, Warwick, U.K. He is currently lecturing at Loughborough University of Technology, Loughborough, Leicestershire, U.K., where he is also a member of the International Electronics Reliability Institute (IERI). His research interests include optoelectronics, thermal breakdown of diode junctions and EOS/ESD effects in semiconductor devices.

David S. Campbell is Professor of Component Technology and Director of the International Electronics Reliability Institute (IERI), Loughborough University of Technology, Loughborough, Leicestershire, U.K. Throughout his eighteen years with the university he has worked on ferro-electrics, thin film, thick film and hybrid circuits, interconnection technology and semiconductor device reliability. He has published over a hundred articles and contributed to eight books.

Paper IV

Tunnicliffe,M.J., Dwyer,V.M., Campbell,D.S.,

"Parametric Drift in Electrostatically Damaged MOS Transistors",

Proc. 14th. Electrical Overstress/Electrostatic Discharge Symposium
Dallas, Texas, 16th.-18th. Sept., 1992,

pp.112-20, 1992.

PARAMETRIC DRIFT IN ELECTROSTATICALLY DAMAGED MOS TRANSISTORS

by

M.J. TUNNICLIFFE, V.M. DWYER and D.S. CAMPBELL

International Electronics Reliability Institute,
Department of Electronic and Electrical Engineering,
University of Technology, Loughborough, Leicestershire, LE11 3TU, U.K.
Telephone Inside U.K.: 0509 222855, Outside U.K.: 44 509 222855.

Abstract

The relationship between parametric drift and latent damage in ESD-stressed MOSFETs is studied. Sub-breakdown damage causes minor characteristic distortion and may remain undetected until failure. However, such damage is only produced within a narrow stress-voltage window.

Oxide breakdown may cause straightforward malfunction (*catastrophic failure*) or degraded transistor action (*walking-wounded* behaviour). Walking wounded devices can degrade further under working voltages (0-10V), providing a latent failure mechanism. Prolonged stress may cause partial recovery in catastrophically failed devices. These phenomena are attributed to polysilicon intrusion into the gate oxide and channel regions.

Catastrophic and walking-wounded failure are modelled using the *PSpice* circuit simulation system. The effects of degradation upon CMOS logic are also examined.

1. Introduction

The importance of electrostatically-induced latent damage was established in the early 1980s by McAteer et al [1,2], who showed that sub-catastrophic ESD can reduce the reliability of electronic components. Such damage can be inflicted in the factory during almost any stage of manufacture. Unless the affected devices are eliminated, they can enter the field and fail prematurely during working life.

The fear of latent damage has prompted a considerable research effort over the past decade. For example, Crockett's experiments on packaged CMOS circuits [3] showed that low-voltage (1kV) ESD pulses rendered devices more sensitive to subsequent high-voltage (2.5kV) stress. Later work by Aur et al [4] showed a reduced hot-electron reliability in MOSFETs subjected to ESD pulses.

Sub-catastrophic ESD can also cause characteristic variation in components. Holmes [5], for example, showed that an ESD pulse below the breakdown threshold produced a negative shift in the strong-inversion threshold voltage V_T. This was often sufficient to drive enhancement (E) mode MOSFETs into depletion (D) mode, causing logical malfunctions. Several other workers [6-8] have found that ESD-induced oxide breakdown distorts a MOSFETs I/V characteristics, causing a reduction of the transconductance g_m or a total loss of transistor action. The former condition is described as 'walking wounded' while the latter is termed 'catastrophic failure' [6]. Soden and Hawkins [8] found that g_m reduction in CMOS circuits is accompanied by an increase in the supply current, introducing a battery-failure hazard in portable equipment. These phenomena were also predicted by Syrzycki's theoretical analysis [9].

The present work examines the relationships between latent damage and characteristic variation in MOS transistors. Attempts are made to correlate parametric drift and oxide wearout in MOSFETs subjected to ESD. The degradation of walking-wounded devices towards catastrophic failure under working voltage conditions is also studied. Tentative attempts are made to model this degradation process using the MicroSim *PSpice* software.

2. Apparatus and Test Structures

2.1 Apparatus

Fig.1 shows the apparatus used in these studies. The devices were tested at wafer-level using a chuck/microprober system. The Hartley 'AutoZap' supplied human-body-model (HBM) ESD pulses, which were applied to the device under test (DUT) between the gate and the wafer substrate contacts. Fig.2(1) shows the equivalent circuit of the AutoZap. The 2-way relay allowed the discharge capacitance C_1 to be charged to the required pulse voltage V_p and subsequently discharged into the device-under-test (DUT) via the discharge resistor R_2. Fig.2(2) shows a typical ESD pulse voltage waveform for a 1.5KΩ resistive load. The effect of ESD upon the DUT's d.c. characteristics was observed using the Hewlett Packard HP4145B parametric analyzer. Room temperature (approx. 25°C) was maintained throughout the experiments and the DUTs were illuminated by the microscope illumination system during testing.

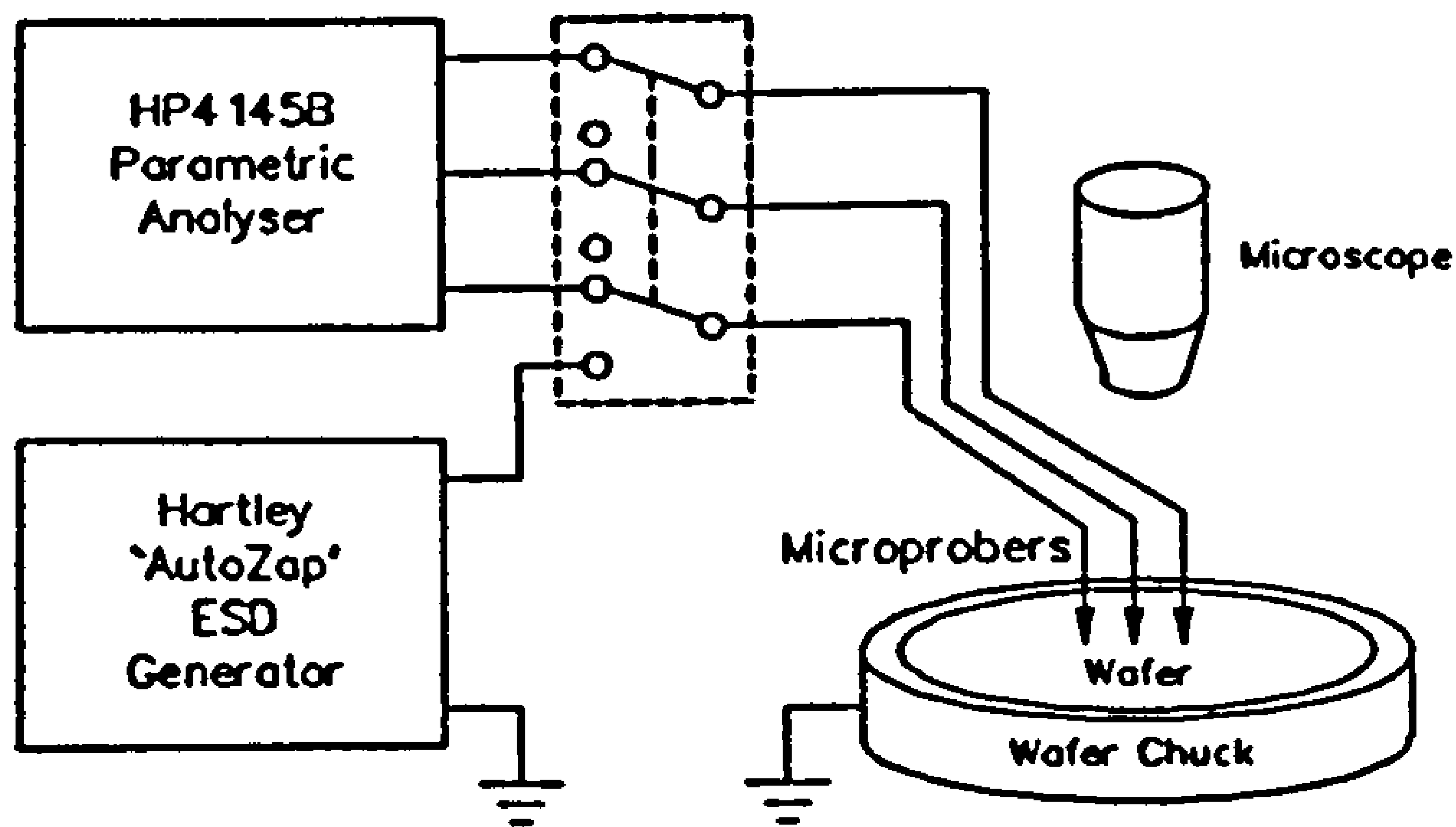


Figure 1: Block diagram of the experimental apparatus. The device may be switched between the 'AutoZap' ESD source and the HP4145B parametric analyzer.

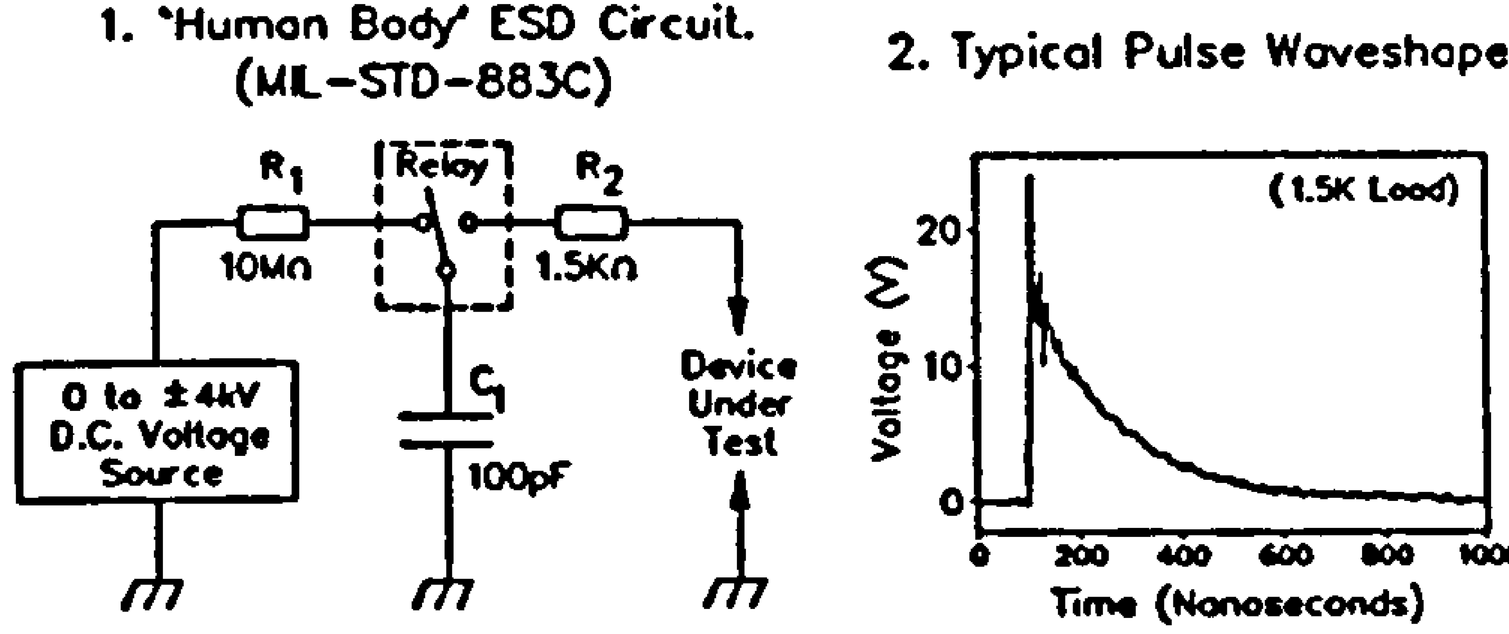
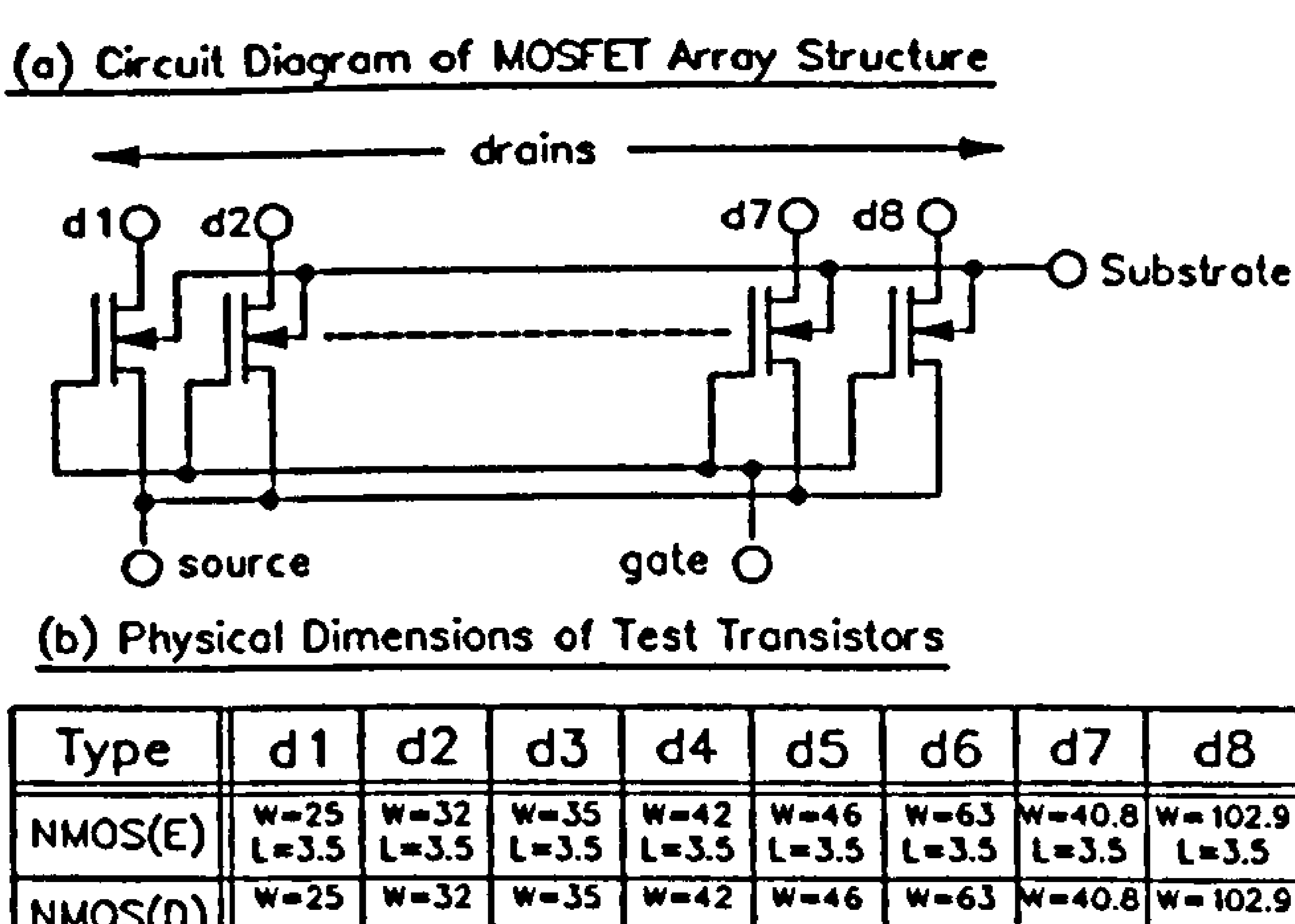


Figure 2: (1) AutoZap equivalent circuit [MIL-STD-883C human body model (HBM) specification] and (2) Typical output voltage waveform [into 1.5KΩ resistive load].

2.2 Test Samples

NMOS transistor structures were fabricated on 3" diameter wafers of <100> bulk silicon, p-doped to $6.6.10^{14}cm^{-3}$. The n^+ source and drain regions were formed by 100keV arsenic ion-implantation to a density of $2.10^{20}cm^{-3}$. Channels were preferentially implant-doped to a net concentration of $4.10^{15}cm^{-3}$ using boron (p) in the E-mode devices and phosphorus (n) in the D-mode devices. Gate oxides were grown in O_2 under 950°C to a thickness of 40nm, upon which the 455nm polysilicon gate electrodes were CVD deposited. The gate polysilicon was As (n^+) doped to a density of $10^{21}cm^{-3}$. The field-oxide was 0.6μm thick, and was supplemented by a 1.5μm Si glass passivation layer.

The transistor structures were fabricated in arrays of eight devices (denoted d1-d8), with common gate and source terminals and individual drains. Since the ESD pulse charge (C_1V_p) may not be evenly distributed between the device gates, each array was analyzed as a single unit rather than a set of independent structures. Although this adds an unfortunate complication, it simulates an actual i.c. input in which many gate structures are connected to a single pin. Fig.3 shows a circuit diagram of the interconnections, together with the dimensions of the various structures.



(b) Physical Dimensions of Test Transistors

Type	d1	d2	d3	d4	d5	d6	d7	d8
NMOS(E)	W=25 L=3.5	W=32 L=3.5	W=35 L=3.5	W=42 L=3.5	W=46 L=3.5	W=63 L=3.5	W=40.8 L=3.5	W=102.9 L=3.5
NMOS(D)	W=25 L=3.5	W=32 L=3.5	W=35 L=3.5	W=42 L=3.5	W=46 L=3.5	W=63 L=3.5	W=40.8 L=3.5	W=102.9 L=3.5

(W = channel width, L = channel length, all dimensions in micrometers)

Figure 3: (a) Interconnection of devices in MOS transistor array, (b) Dimensions.

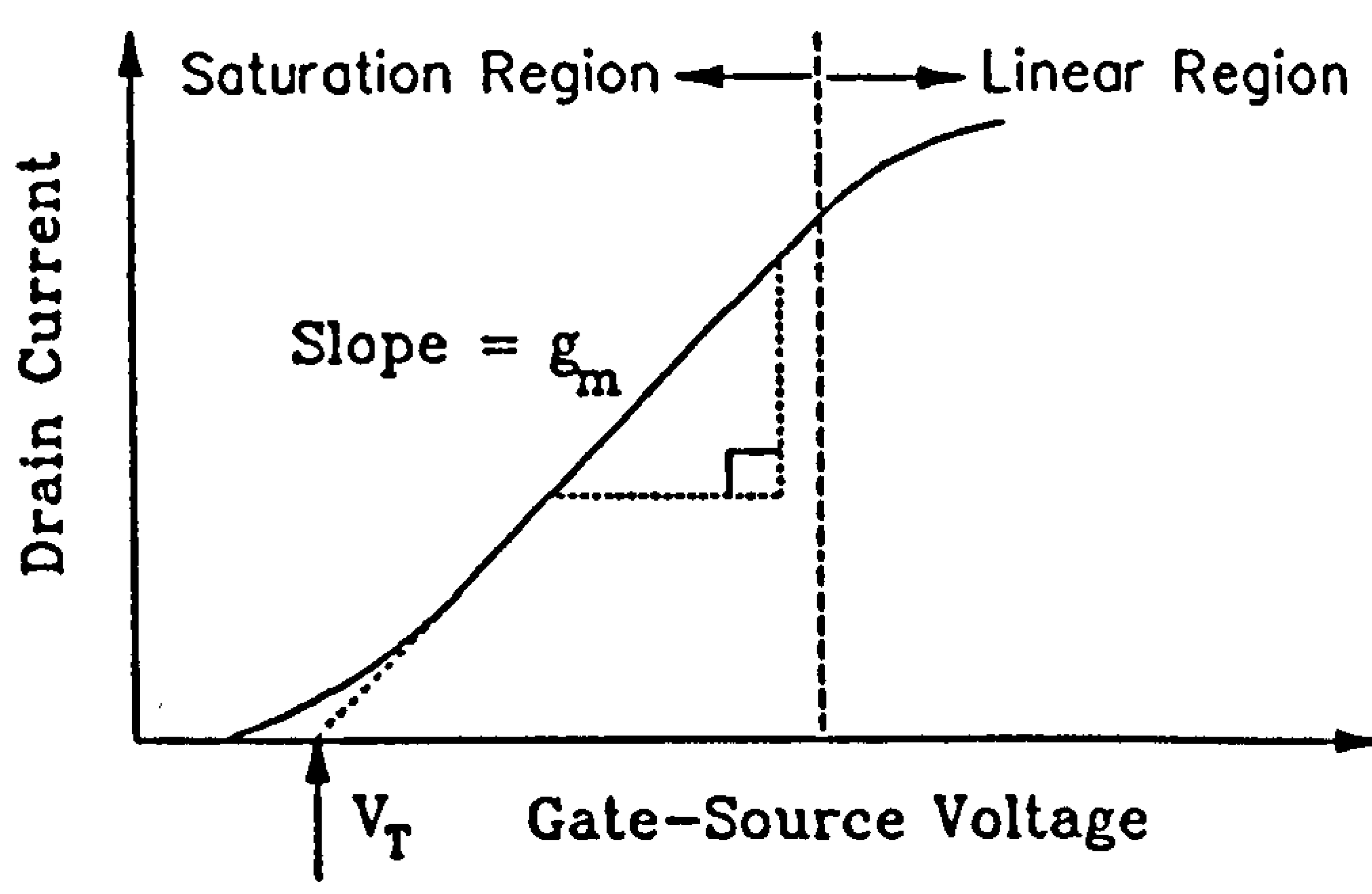


Figure 4: Extraction of V_T and g_m parameters from MOSFET transfer characteristic.

3. Sub-Breakdown Latency and Parametric Drift

3.1 Experiment

The initial phase of this study concerned sub-breakdown ESD damage, ie. latent damage produced by ESD pulses *below* the oxide breakdown voltage threshold V_{bd}. Parametric drift was characterised in terms of the transconductance g_m and the strong-inversion threshold voltage V_T. These parameters were defined as the gradient and intercept of the I_d vs. V_{gs} curve in the saturation region (Fig.4). V_{ds} was held at 7V during this experiment.

The experiments were performed using NMOS E-mode transistor arrays. Each device in each array was characterised in terms of its initial transconductance $g_m(0)$ and initial threshold voltage $V_T(0)$. Negative polarity ESD pulse sequences were applied to the gate structures and g_m and V_T were measured for each device after each pulse.

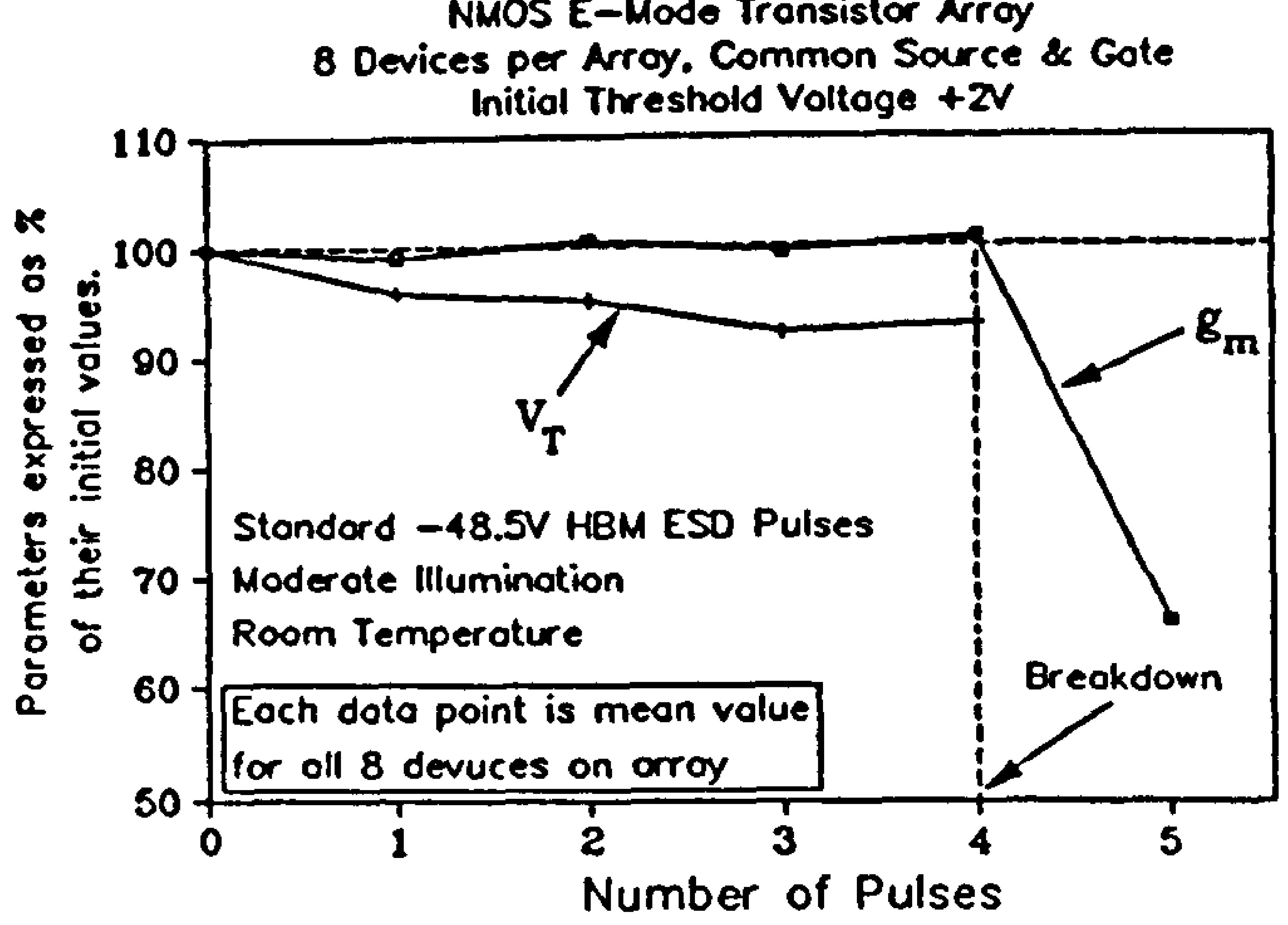


Figure 5: Sub-breakdown parametric drift in NMOS enhancement transistor array subjected to -48.5V ESD pulse sequence.

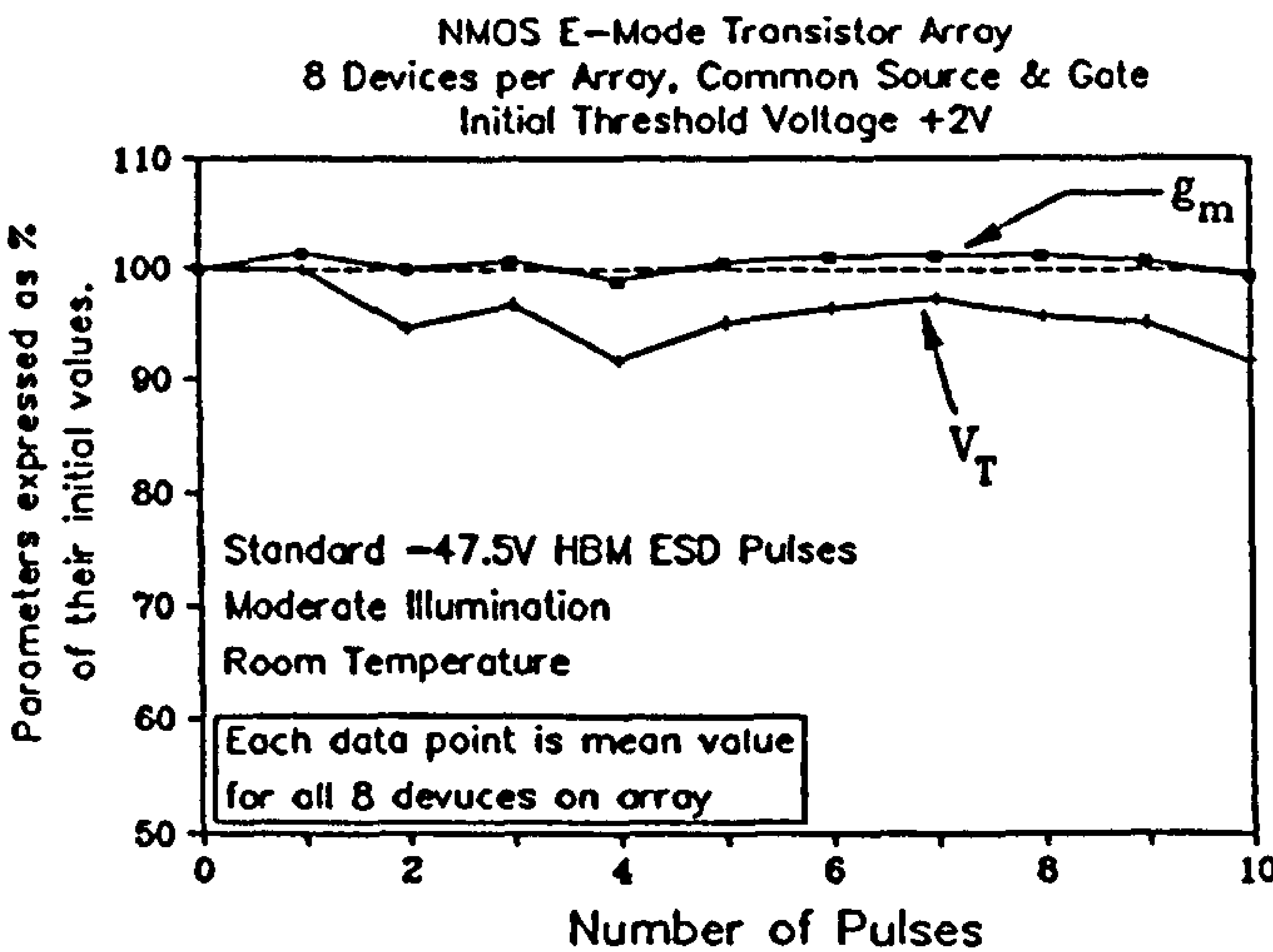


Figure 6: Sub-breakdown parametric drift in NMOS enhancement transistor array subjected to -37.5V ESD pulse sequence.

The process was continued until oxide breakdown was observed in terms of a finite gate current I_{gs}.

The data was then processed in the following manner: The values of g_m and V_T for each device after each pulse were expressed as a percentages of $g_m(0)$ and $V_T(0)$, yielding the percentage values $g_m(\%)$ and $V_T(\%)$. The average value of $g_m(\%)$ and $V_T(\%)$ for all the devices in the array was plotted against the number of pulses n.

Fig.5 shows the resulting graph for -48.5V ESD pulses. While g_m remains approximately constant prior the breakdown, it suddenly drops after breakdown (in agreement with earlier studies [6-8]). The average value of V_T decreases during the pulse sequence, suggesting the occurrence of positive charge trapping. Although this agrees qualitatively with Holmes' results [5], the variation does not exceed 10% and the latent damage may remain undetected until breakdown.

Fig.6 shows the results of an identical experiment performed using -47.5V ESD pulses. At this lower voltage, the latent damage per pulse is smaller than in the previous experiment and the array withstands a far greater number of pulses prior to breakdown (a further 20 pulses were required to cause failure). As with the -48.5V experiment, g_m remains approximately constant while V_T decreases. However, the V_T decrease is subject to fluctuations, suggesting that positive and negative charge trapping may predominate alternately throughout the pulse sequence. Again the shift does not exceed 10% and is too slight to cause functional errors.

The voltage dependence of latent failure was further examined by repeatedly pulsing NMOS array structures and noting the number of pulses n required to cause breakdown. Fig.7 shows n plotted as a function of the ESD pulse magnitude V_p (each data point represents the average value obtained from five structures). The voltage 'window' ΔV_{lat} associated with significant latent failure can be conveniently defined as the voltage range between $n=10$ and $n=1$, which is clearly in the region of 2V.

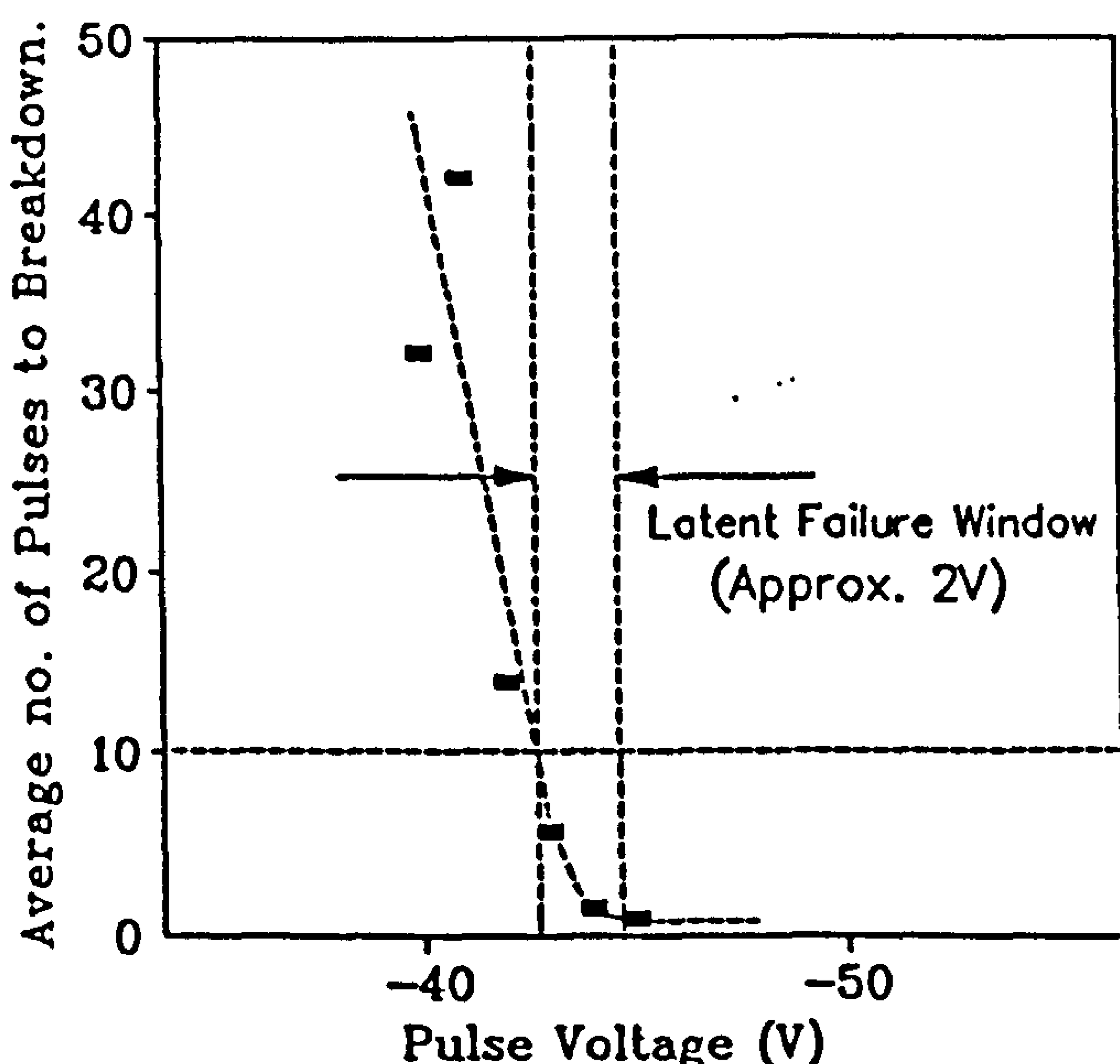


Figure 7: Pulses-to-breakdown n vs. pulse voltage V_p in NMOS E-Mode transistor arrays. Latent window ΔV_{lat} is defined as voltage range between $n=10$ and $n=1$.

3.2 Theoretical Analysis

The value of ΔV_{lat} can also be computed using oxide wearout theory. According to the Chen-Holland-Hu model [10], field-assisted tunnelling supports electron-hole pair generation and subsequent hole-trapping in the stressed oxide. The trapped hole charge (which may be responsible for the observed V_T drift) distorts the field profile, enhancing the tunnelling current and accelerating the process to towards dielectric breakdown. Application of this model to ESD oxide breakdown [11,12] shows that the damage factor S_t produced by a pulse of magnitude V_p can be expressed as

$$S(V_p) = \frac{C T_{ox}}{\tau_0 A k H} \exp\left(-\frac{H T_{ox}}{L V_p}\right) \qquad (1)$$

where C is the total capacitance parallel to the oxide, A is the oxide area, T_{ox} is the oxide thickness, L is the pulse attenuation factor due to capacitive loading (given by $L=C_1/C$) and τ_0, k and H are constants. S_t is defined such that $S_t=0$ and $S_t=1$ represent virgin and failed devices respectively. S_t is therefore approximately equal to $1/n$, and the latent failure window can be re-defined as

$$\Delta V_{lat} = V_p(S_t=1) - V_p(S_t=0.1) \qquad (2)$$

Combining Eqns.(1) and (2), together with the approximation $e^{-2} \approx 0.1$ yields

$$\Delta V_{lat} = V_{bd}\left(1 - \frac{H T_{ox}}{H T_{ox} + 2 L V_{bd}}\right) \qquad (3)$$

where the breakdown voltage magnitude V_{bd} (ie. the value of V_p for $S_t=1$) is given by

$$V_{bd} = \frac{H T_{ox}}{L \cdot \log_e\left(\frac{C T_{ox}}{\tau_0 A k H}\right)} \qquad (4)$$

Since $C_1=100$pF, $T_{ox}=40$nm, $H=550.7$MV/cm [measured for identical structures in ref.12], $C=117.23$pF [independently measured] and $V_{bd}=45$V, ΔV_{lat} can be shown to be 1.52V. Since the measured value was 2V (see Fig.7), the theory agrees roughly with the experiment.

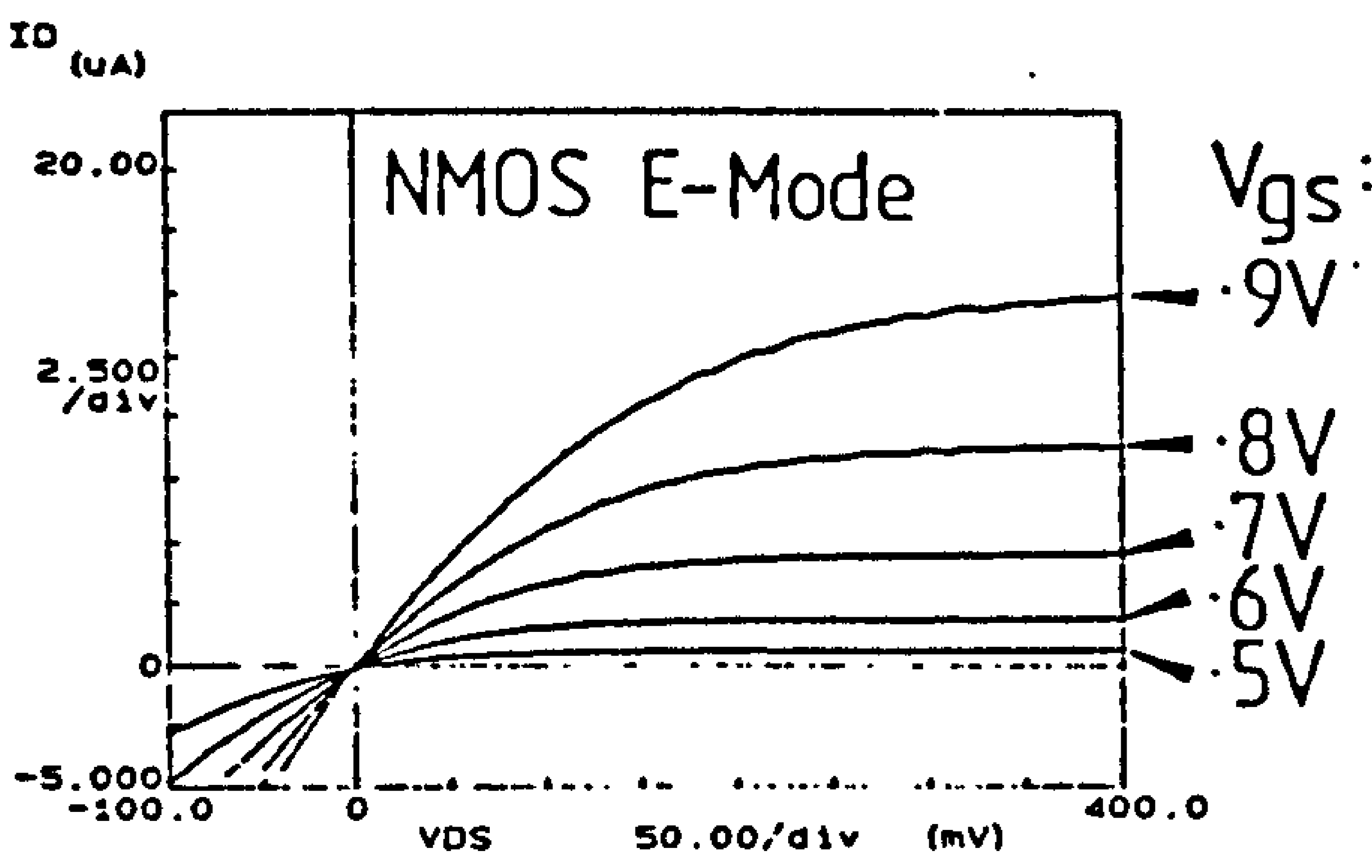
4. Latent Damage and Parametric Drift in Damaged-Oxide Devices

4.1 Experimental Procedure

Results are reported for 64 NMOS transistors, of which 32 were E-mode and 32 were D-Mode. The devices, all of which conformed to their nominal specifications, had received no prior stress. The procedure was as follows: The devices were initially characterised in terms of their I_d vs. V_{ds} and characteristics and were then subjected to single ESD pulses of −200V, -100V, +100V, and +200V. One specimen of each array type was stressed at each voltage. The I_d vs. V_{ds} curves were then re-measured and compared with the original characteristics.

Since many of the devices were in a delicate walking-wounded condition immediately after stress, a characterisation method was required which subjected them to minimum measurement stress. The magnitudes of V_{ds} and V_{gs} were therefore limited to the range −1V to 1V during characterisation. Fig.8 shows typical characteristics of undamaged devices.

The characteristics of some of the degraded devices were then re-measured using V_{ds} and V_{gs} ranges of 0-10V, in order to examine the effects of working-voltage stress upon the characteristics. Since V_{gs} is common to all the devices in an array, only one device per array was subjected to such measurement.



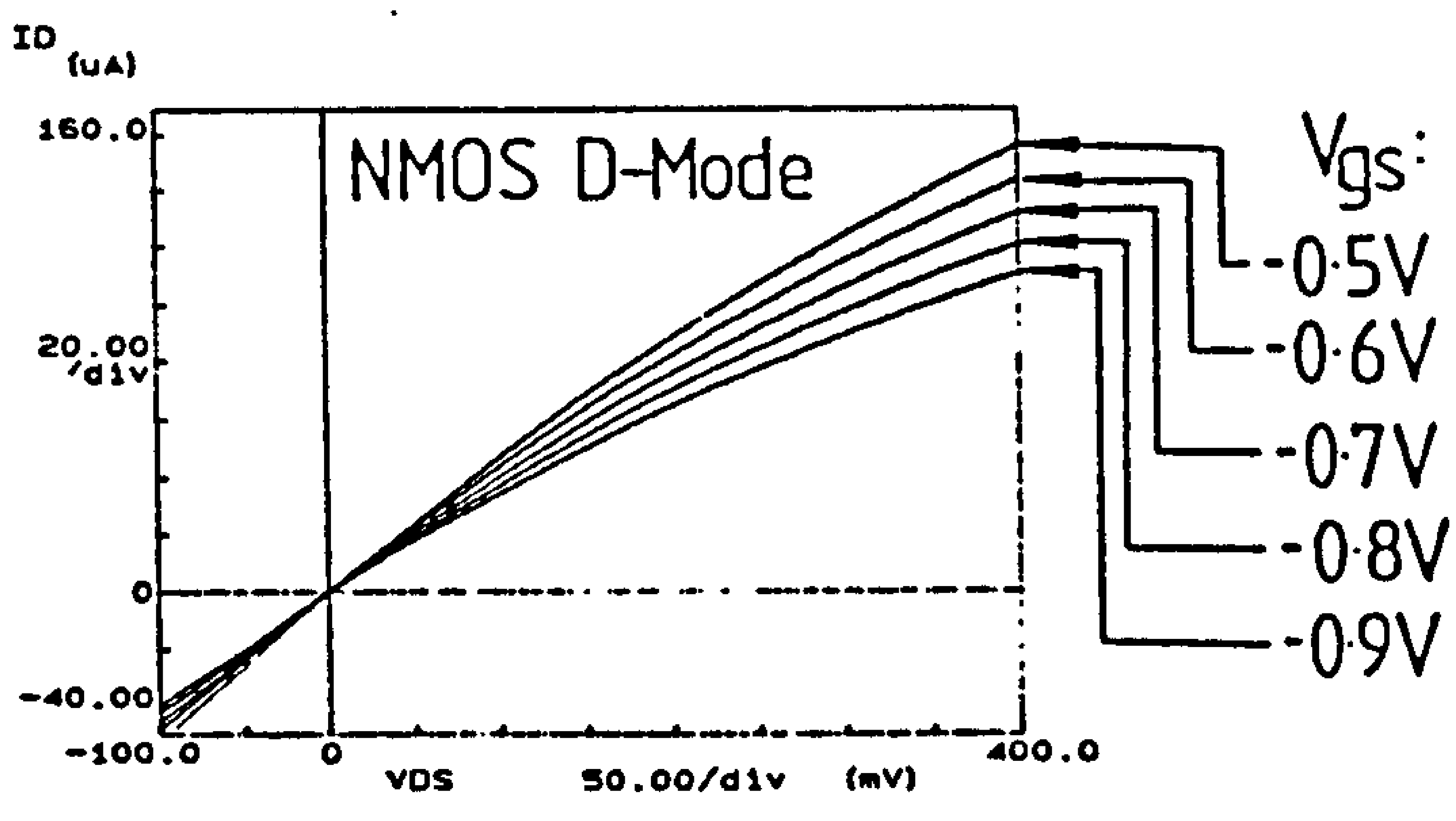


Figure 8: Typical I_d vs. V_{ds} characteristics of virgin transistor structures.

4.2 Results

The post-ESD characteristics were classified into the following categories:

Group 1. Unchanged characteristics.
Group 2. Characteristics with negatively-shifted V_T.
Group 3. Characteristics with reduced g_m, converging at origin.
Group 4. Characteristics with reduced g_m *not* converging at origin.
Group 5. Linear resistive characteristics.
Group 6. Characteristics with zero g_m.

Typical characteristics from each category are shown in Figs.9 and 10. Groups 2-4 clearly belong to the 'walking-wounded' category, while Groups 5 and 6 are catastrophic failures. The threshold-voltage variation in Group 2 was sometimes of the order of several volts, causing E-mode

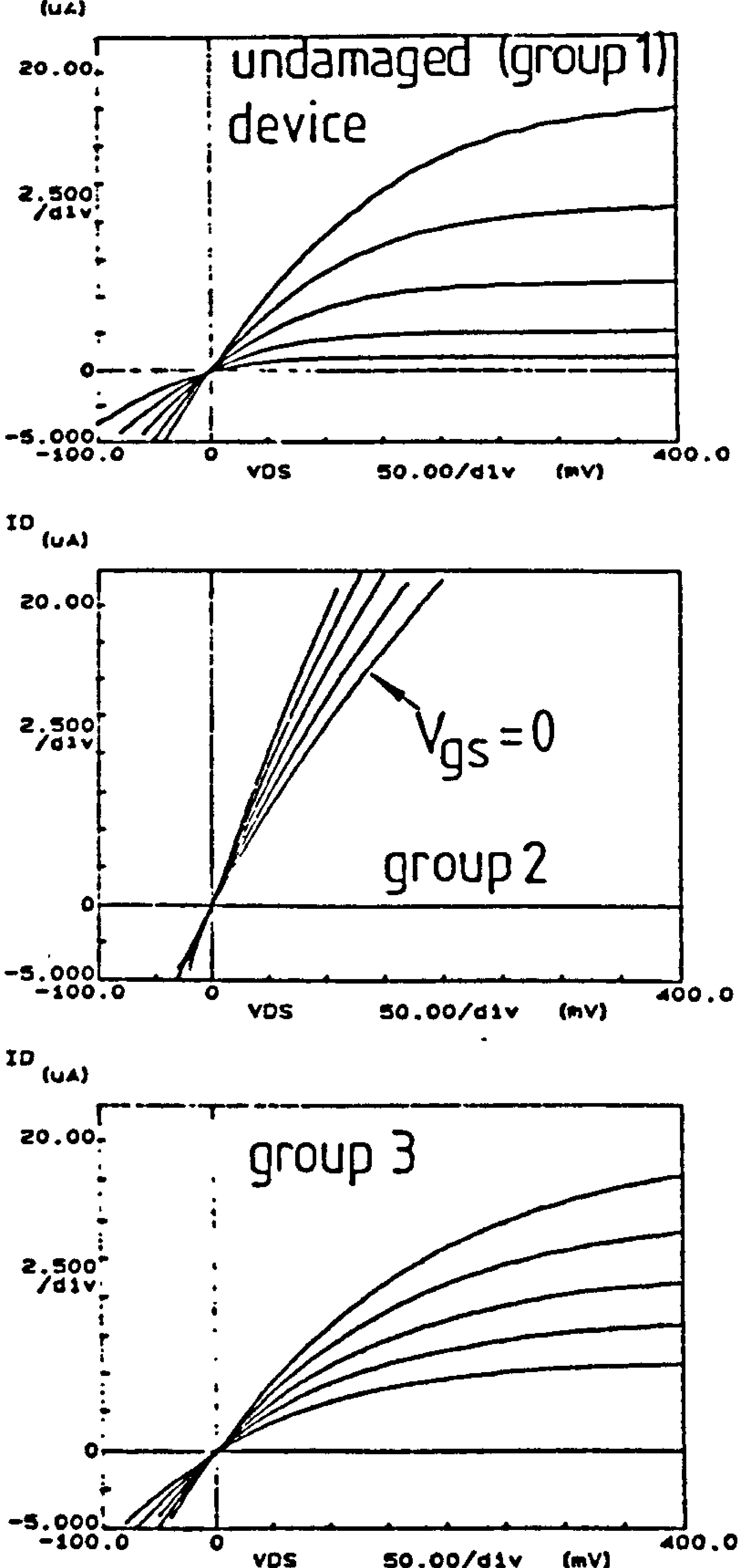


Figure 9: Post-breakdown characteristics of NMOS E-Mode transistors (Groups 1-3).

devices to effectively become D-mode. The transconductance reduction in Groups 3 and 4 varied between about 5% and 95% (anything beyond a 95% reduction was classed as a Group 6 failure). Fig.11 shows which of the devices on each array fell into each category.

Many of the devices whose characteristics were re-measured using $0<V_{ds}<10V$ retained their stable walking-wounded characteristics. Others, however, underwent further degradation during characterisation. Fig.12 shows the characteristics of two walking-wounded devices (originally of Groups 2 and 4) after they had received three characterisations. (A typical undamaged device characteristic is shown for comparison.) These devices are clearly unstable under working voltage stress and therefore constitute a latent hazard.

Further experiments were performed in order to observe this degradation phenomenon in action. Fig.13 shows the transfer characteristics (I_d vs. V_{gs}) for an NMOS E-Mode device during the degradation cycle. Characteristics A and B were measured immediately before and after the application of a -80V ESD pulse,

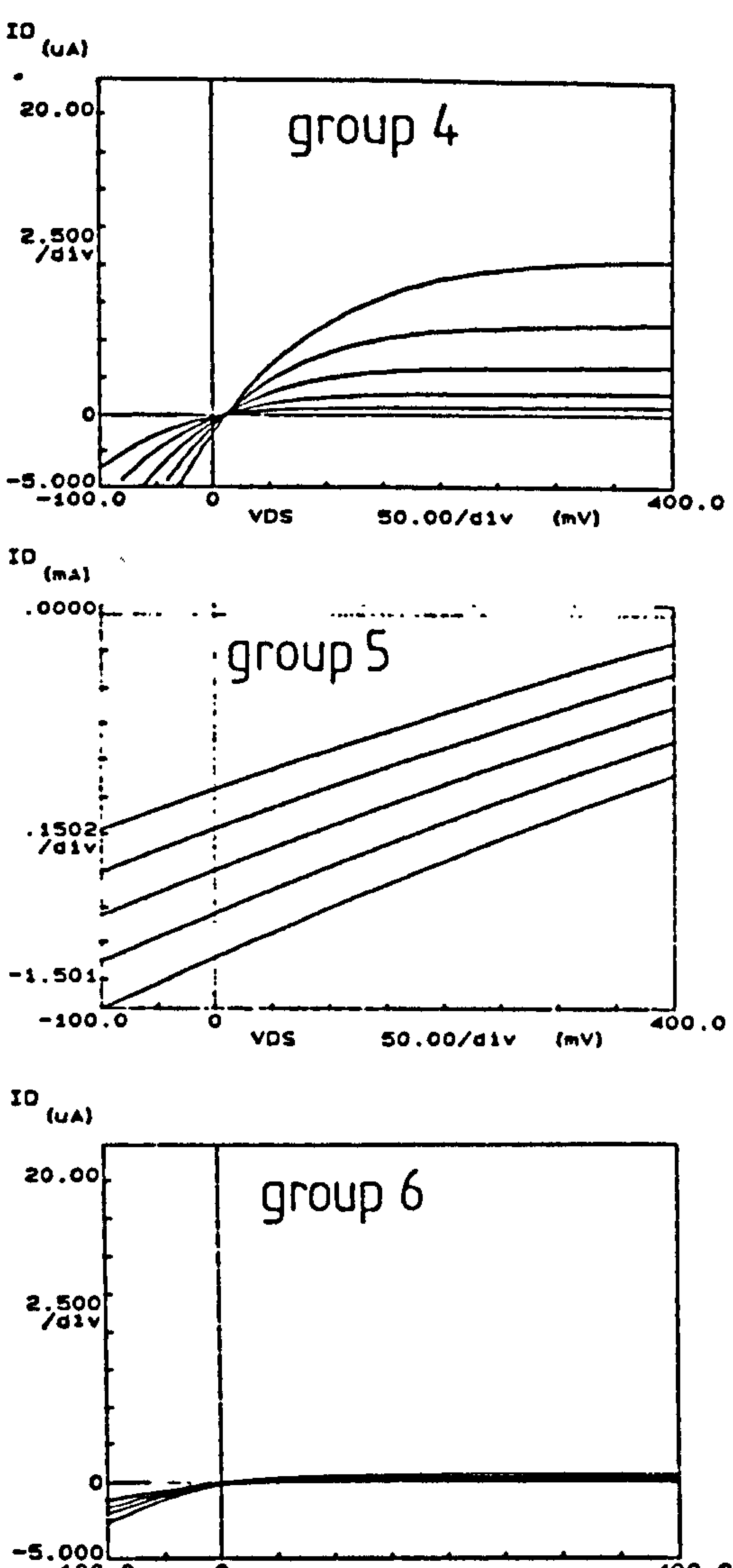


Figure 10: Post-breakdown characteristics of NMOS E-Mode transistors (Groups 4-6).

NMOS D-Mode Transistors

Pulse Voltage	d1	d2	d3	d4	d5	d6	d7	d8
-200V	1	1	1	1	1	2	2	2
-100V	3	3	1	1	1	1	1	1
+100V	2	2	2	2	2	5	2	2
+200V	2	2	2	2	2	2	2	5

NMOS E-Mode Transistors

Pulse Voltage	d1	d2	d3	d4	d5	d6	d7	d8
-200V	6	6	6	6	6	6	4	4
-100V	6	6	6	3	3	3	3	1
+100V	3	5	2	3	3	3	3	5
+200V	2	2	2	2	2	2	5	2

Figure 11: Distribution of failure catagories 1 to 5 in NMOS arrays.

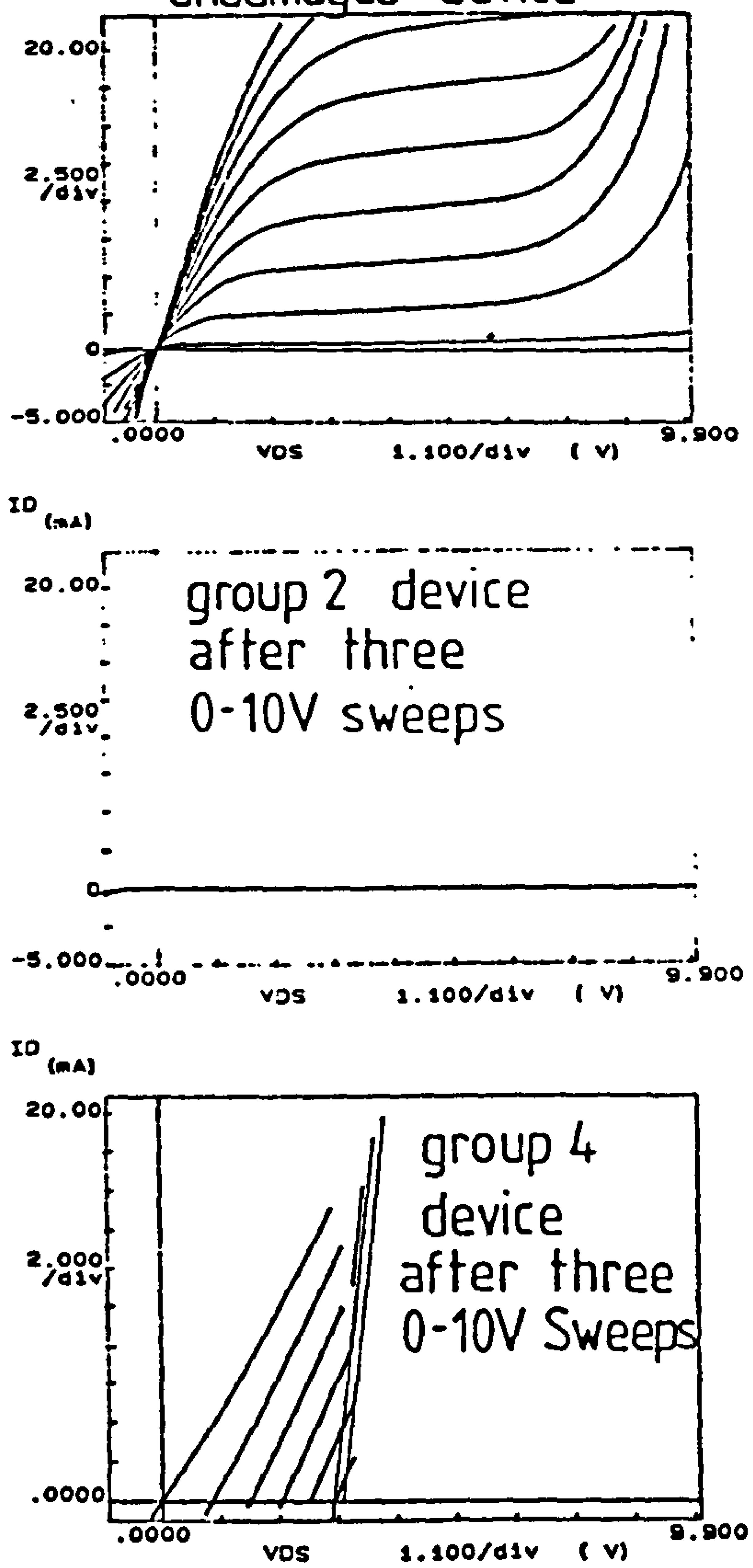


Figure 12: Characteristics of walking-wounded devices (originally in Groups 2 and 4) after three 0-10V characterisation sweeps. Typical undamaged characteristic is shown for comparison.

while characteristics C and D show the results of subsequent characterisations. Since characteristics B and C possess positive transconductances ($\partial I_d/\partial V_{gs}$), they must be classified as walking-wounded states. However, characteristic D is clearly a catastrophic failure of Group 5 (see Fig.10). Degradation is displayed by the spasmodic jumps of characteristics B and C, which cause the device to drift towards catastrophic failure under working-voltage stress.

In a limited number of cases, repeated characteristic re-measurement caused an eventual return to transistor action (positive g_m). In this recovered state, the transistors were extremely robust and could often withstand ESD pulses up to several hundred volts. However, these cases were rare, and most catastrophically failed devices retained in their characteristics indefinitely.

Figure 13: Characteristics during degradation.

4.3 Discussion and Modelling

Dielectric breakdown in MOS oxides is believed to be accompanied by the intrusion of one or more filaments of polsilicon gate material into the SiO_2 [8]. These filaments can form conductive paths between the source, drain and gate terminals, or between the gate electrode and the channel. The geometry and position of the filaments dictate the nature of the failure characteristics [9].

4.3.1 Qualitative Analysis of Failure Characteristics

Group 2 characteristics are clearly extreme cases of the negative V_T drift encountered in Section 3. In these cases the positive-charge trapping *was* sufficient to drive E-mode devices into D-mode (or D-mode devices even further into depletion).

Group 3 failures probably result from filaments connecting the gate to the source terminal. Since the filament resistance R_f forms a potential divider with the parasitic gate resistance R_g, the effective gate-source voltage is reduced by a factor $R_f/(R_f+R_g)$, thereby reducing the effective transconductance by the same factor. The characteristic is uniformly scaled-down by the process and the characteristics continue to converge at the origin.

Group 4 failures are caused by gate/drain and gate/channel filaments. When V_{gs} is high and V_{ds} is low, current flows in through the gate and out through the source, creating a negative I_d for positive V_{ds}. Hence the characteristics fail to converge at the origin. The g_m reduction is caused by inversion-layer charge being sucked out of the channel region into the gate via the damaged oxide.

Group 5 represents an extreme case of the gate-drain breakdown, when the current in the filament resistance R_f dominates the transistor drain current. Similarly Group 6 represents an extreme case of gate-channel breakdown, where practically all inversion charge is depleted via R_f.

The spasmodic nature of characteristic shifts in Fig.13 suggests thermal expansion of defects due to joule heating. As the power dissipation in a defect increases during a voltage sweep, the temperature increases quazi-statically until a critical temperature is reached. At this point the defect expands rapidly, reducing R_f, until a new equilibrium is reached. Since this transition is fast for the parametric analyzer to measure, the characteristics appear to consist of discrete curves, punctuated by discontinuities.

4.3.2 PSpice Model of Failed Device

A lumped-element model of a failed transistor was developed using the MicroSim *PSpice* software. The methodology (which was based upon that of Syrzycki [9]) is illustrated in Fig.14. Defect models #2 represents the gate/source and gate/drain filament structures described above. Since gate-channel breakdown (defect model #1) is accompanied by solid-state diffusion of n^+ doping into the channel region [9], a rectangular equipotential of dimensions L_d x W_d, was placed within the channel, dividing the structure into three different transistors. The PSpice LEVEL 1 MOSFET model (Shichman-Hodges) was selected, together with V_T=0. (Although it assumes constant mobility, it easily applicable to the failure characteristics of Figs. 9 and 10, in which the fields are insufficient to cause mobility modulation.) Figs.15-19 show failure characteristics simulated using this model.

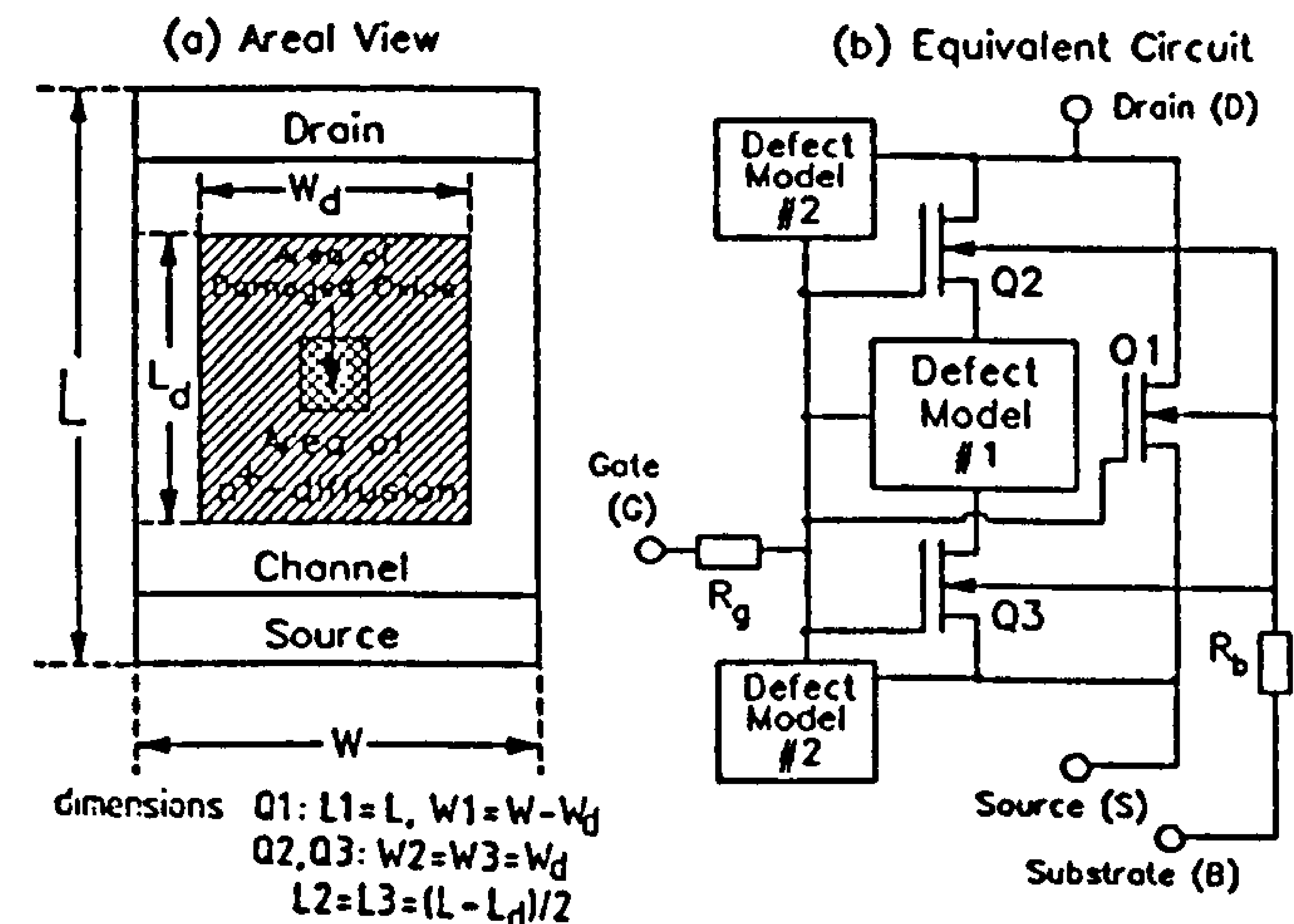


Figure 14: Schematic areal view and equivalent circuit for damaged MOSFET.

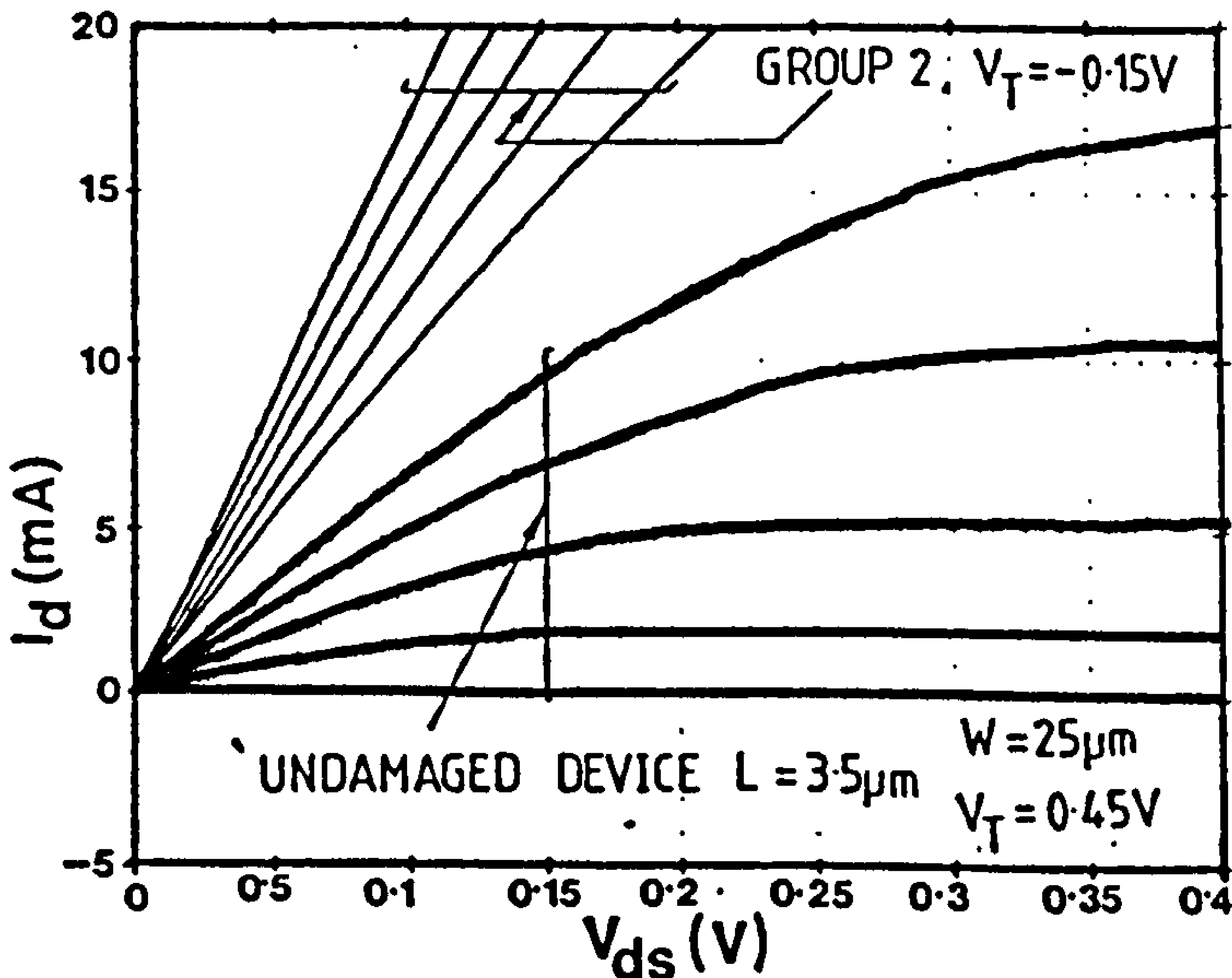


Figure 15: Group 2 failure characteristic predicted using PSpice model.

4.3.3 Effect of Degradation on CMOS Operation

The impact of device degradation upon logic operation is a complex subject which shall be dealt with more fully in a later publication [13]. The present discussion is confined to the simple CMOS invertor (NOT) gate. The left portion of Fig.20shows a graphical representation of the operation of such a circuit. The two transistor characteristics are superimposed and the operating points for a selection of input voltages are transferred to an input vs. output

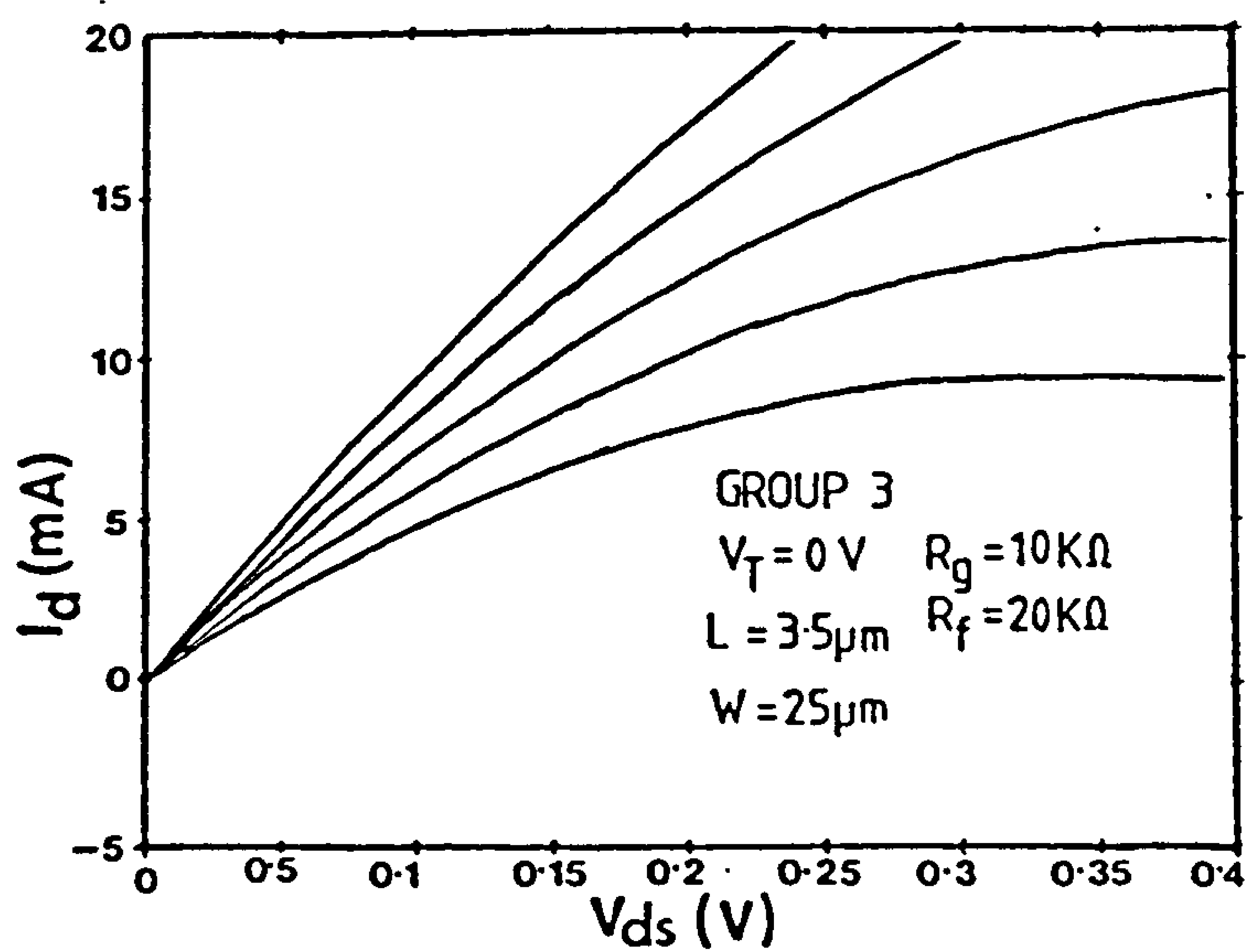


Figure 16: Group 3 failure characteristic predicted using PSpice model.

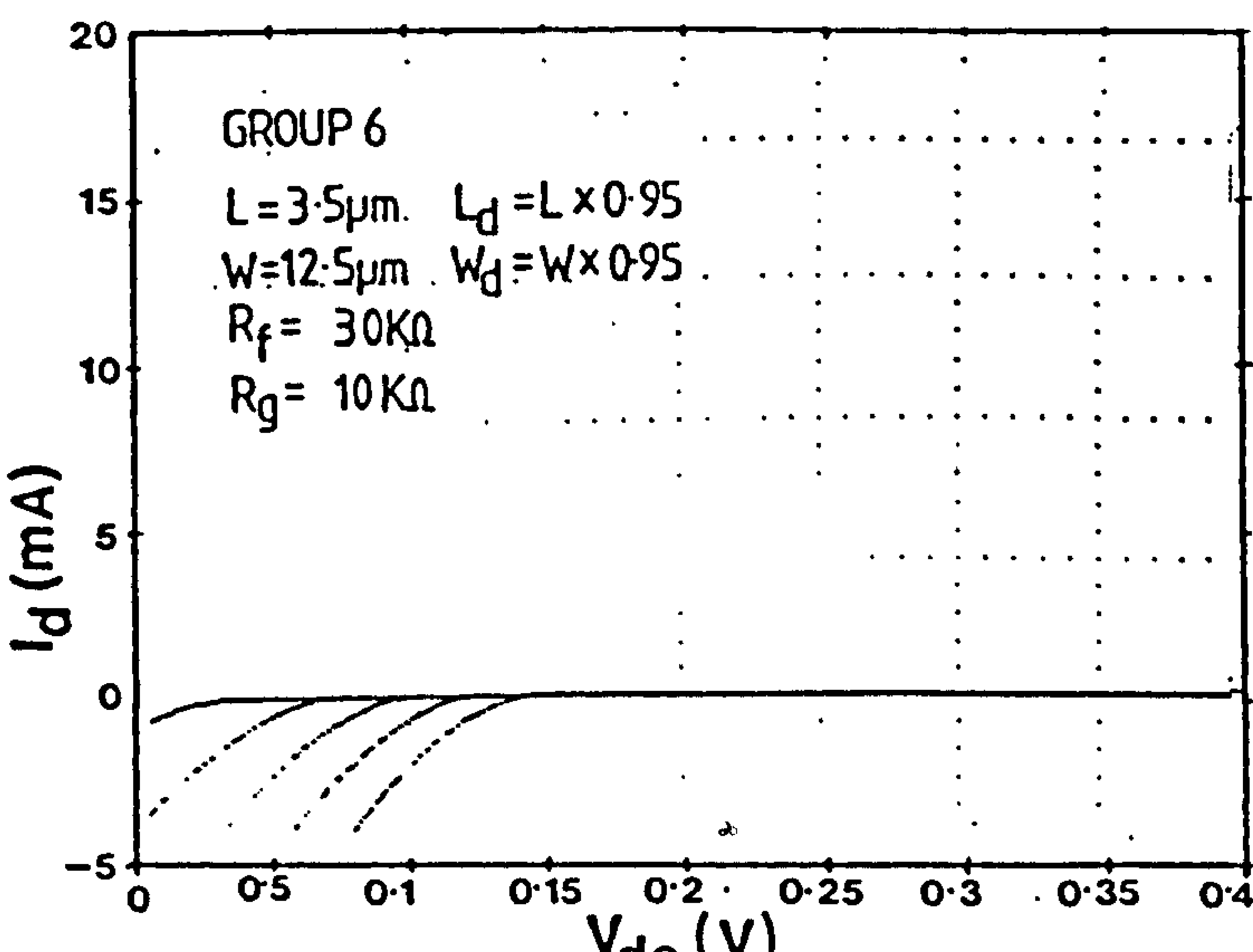


Figure 19: Group 6 failure characteristic predicted using PSpice model.

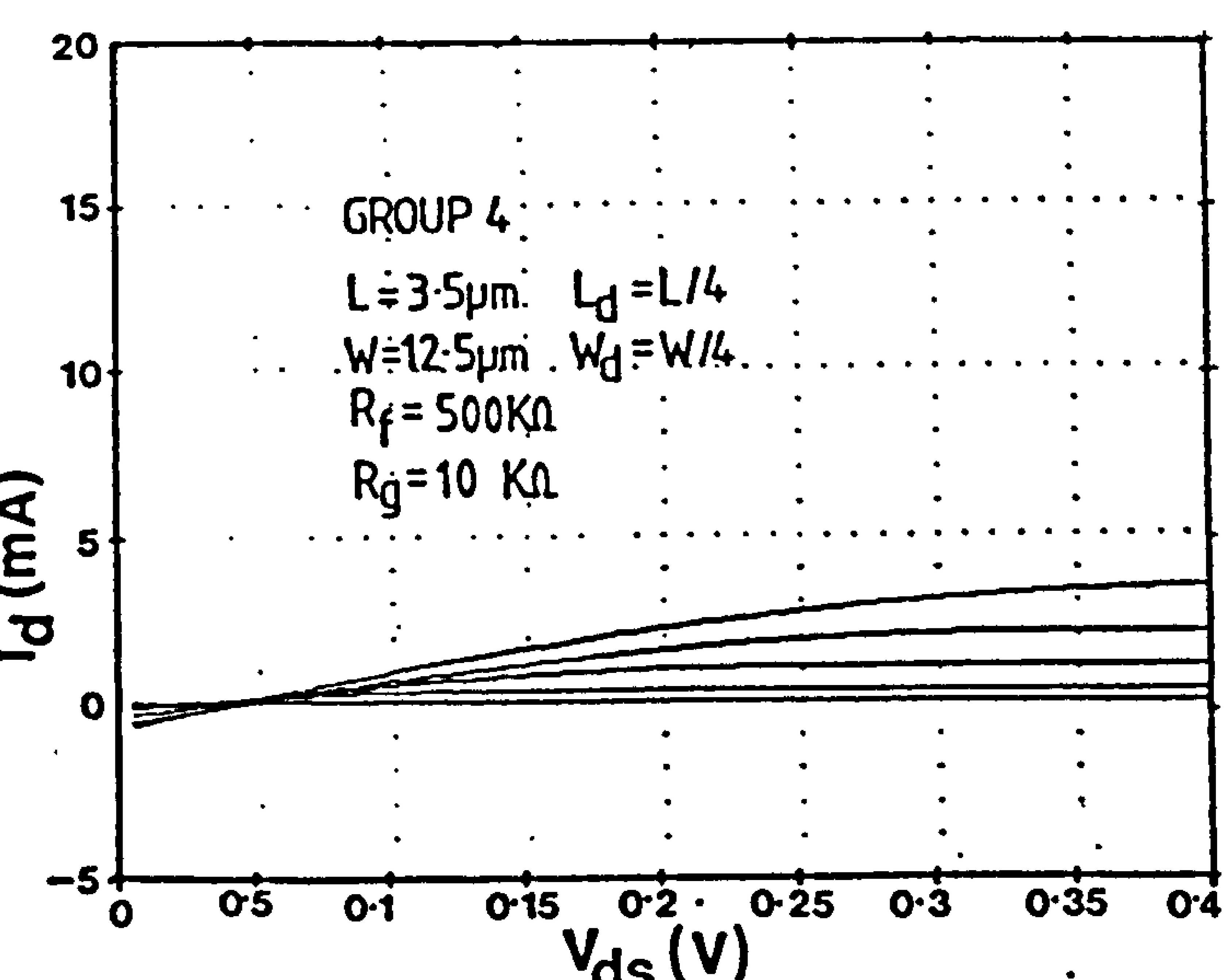


Figure 17: Group 4 failure characteristic predicted using PSpice model.

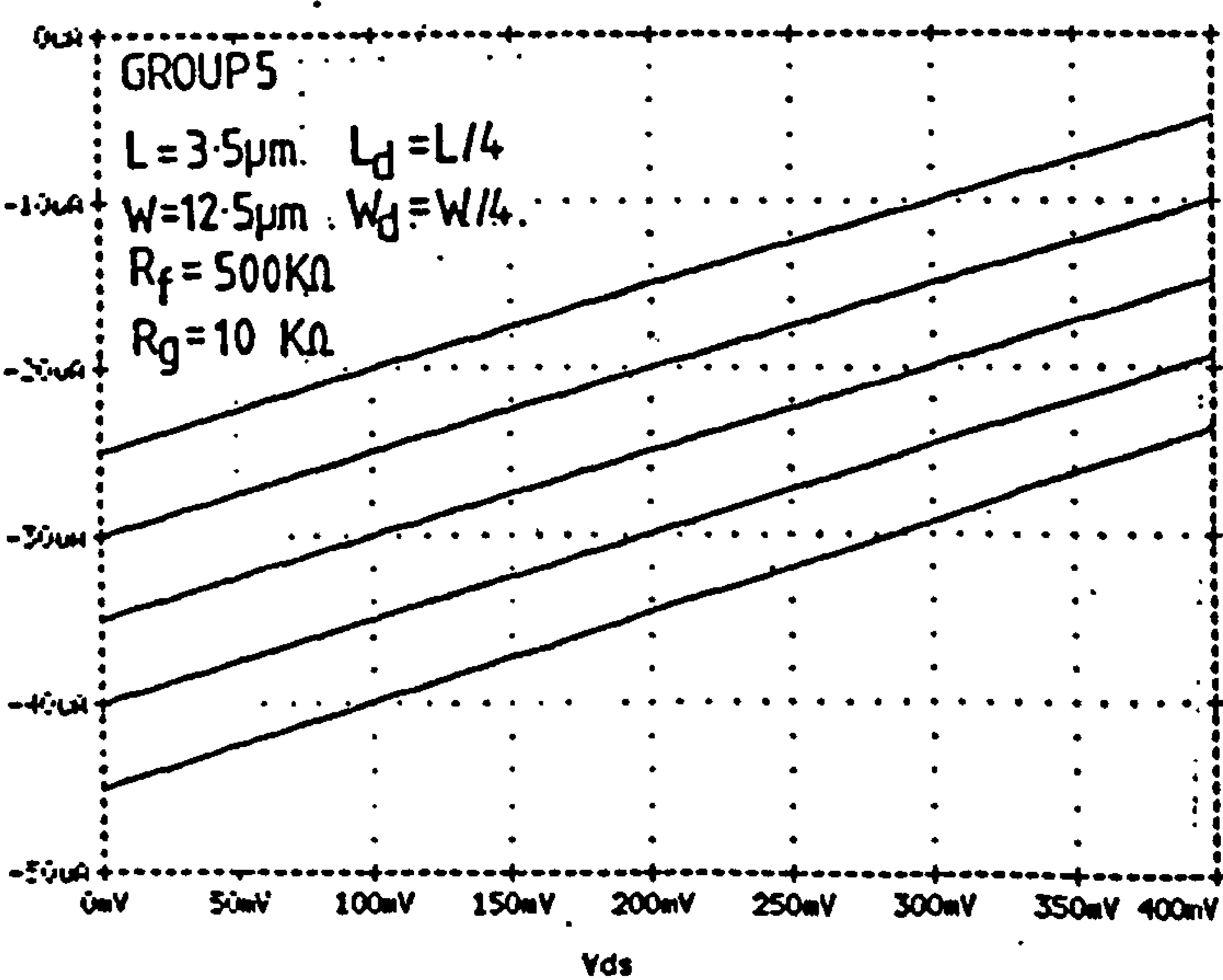


Figure 18: Group 5 failure characteristic predicted using PSpice model.

graph. The output clearly exhibits a logical 1 while the input is held at 0 and vice versa.

The right portion of Fig.20 shows the impact of a degraded (group 4) n-channel transistor on the gate operation. Although the output voltage is positively shifted, the logical inversion function is maintained, illustrating how walking-wounded failure is not necessarily detectable in terms of logic performance [8]. However, continued g_m degradation causes the curve to shift steadily to the right until the circuit enters a 'stuck-at-1' condition.

Fig.21 illustrates the effects of linear catastrophic failure (Group 5) on a CMOS invertor. This type of failure clearly introduces a voltage-follower behaviour, irrespective of which device (n-channel or p-channel) has failed.

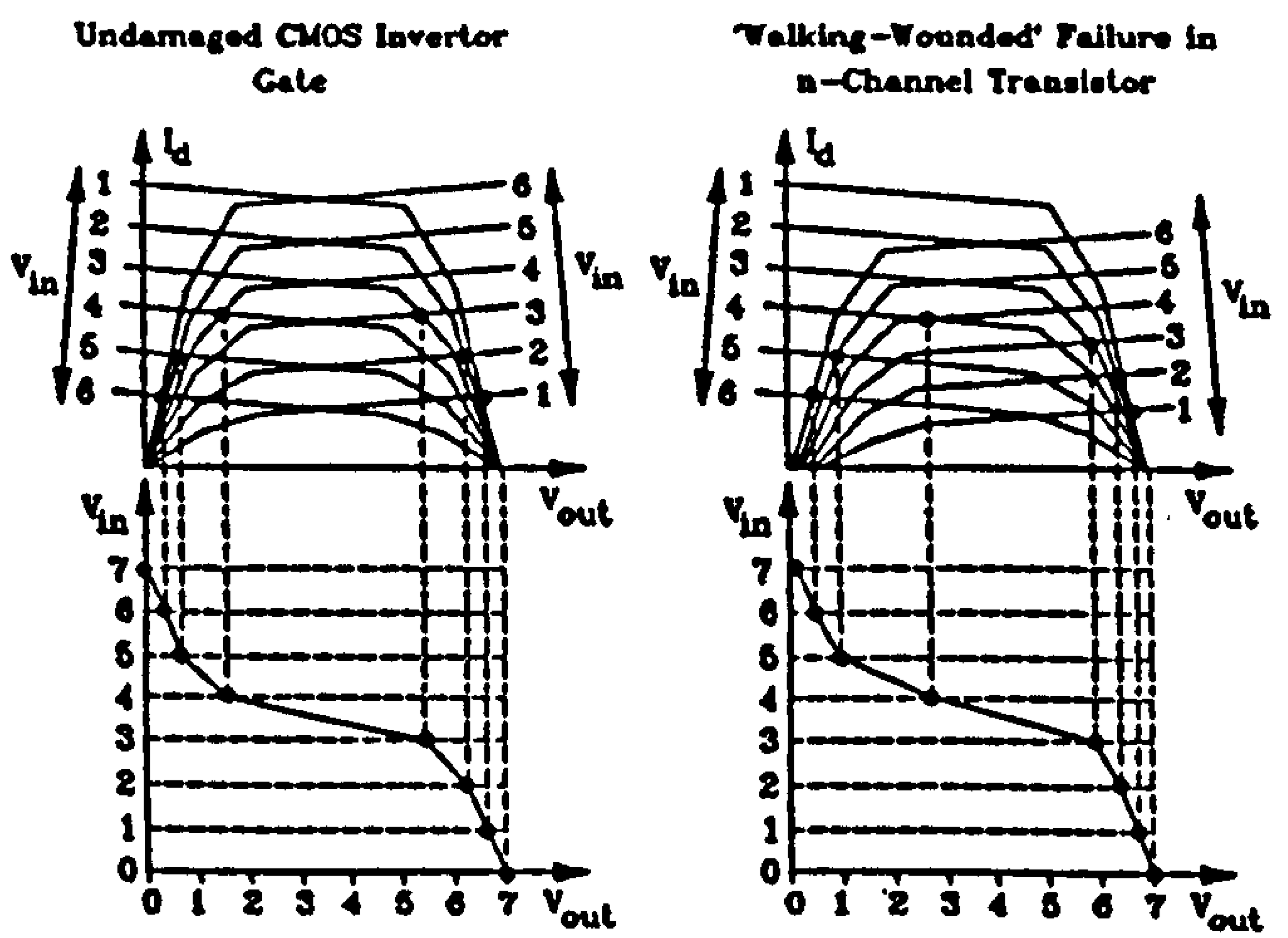


Figure 20 : Schematic representation of the transfer characteristics of a healthy CMOS invertor (left) and CMOS invertor with 'walking-wounded' n-channel transistor.

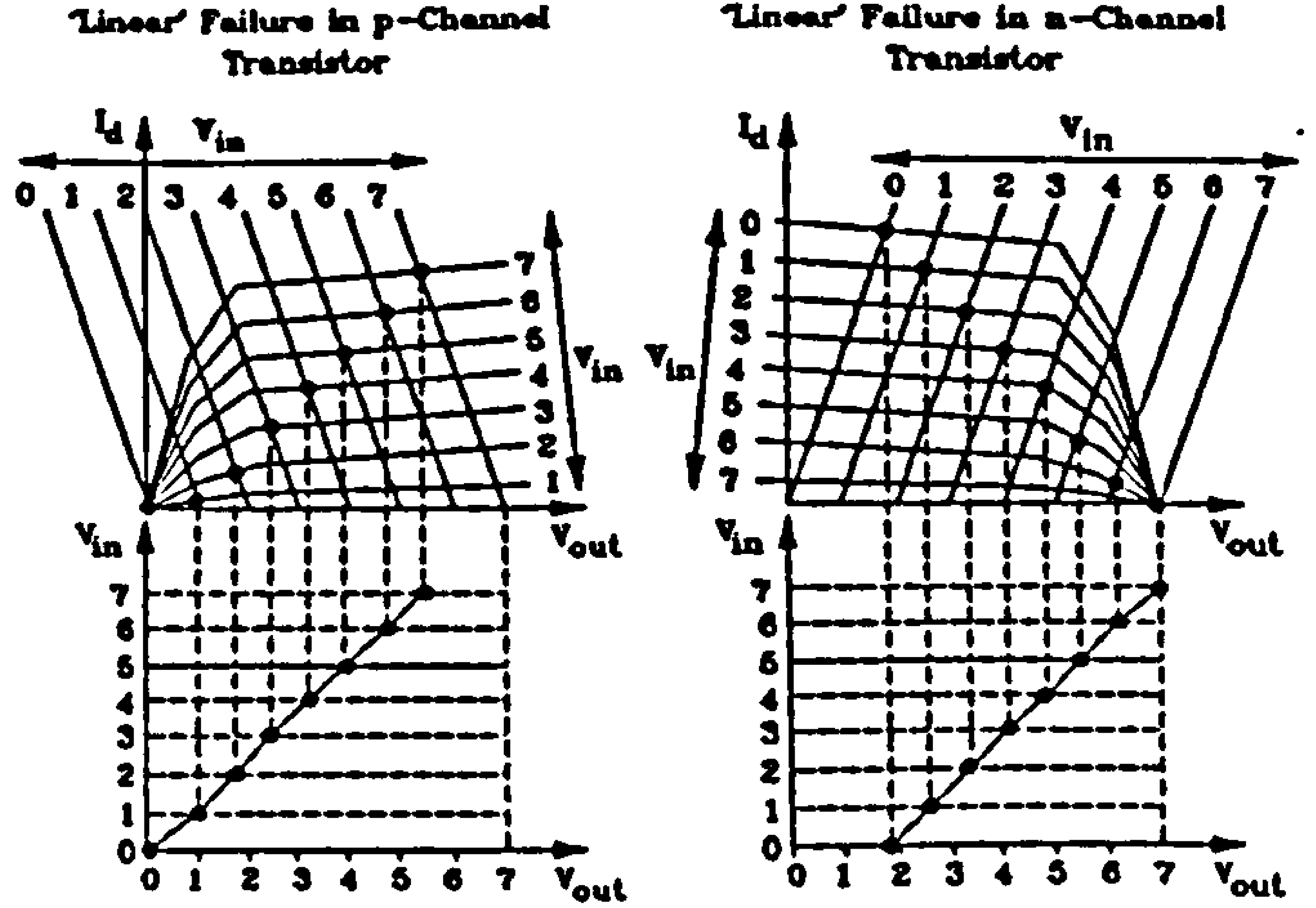


Figure 21: Schematic representation of the transfer characteristics of CMOS invertor gate with linear failure in p-channel transistor (left) and CMOS invertor gate with linear failure in n-channel transistor (right).

5. Conclusions

The relationships between latent damage and d.c. parametric drift in ESD damaged MOSFETs have been studied both experimentally and theoretically. The following conclusions can be drawn from the results:

1. Although sub-breakdown latency is difficult to detect in terms of d.c. parametric drift, the extreme narrowness of the latent failure window (determined by theory and experiment), suggests that it is unlikely to be a major reliability hazard. Devices subjected to random-magnitude ESD are most likely to receive pulses greater than V_{bd}, causing immediate failure, or below V_{bd}-ΔV_{lat}, leaving them virtually undamaged.

2. MOS devices with damaged oxides were found to exhibit either walking-wounded or catastrophic-failure characteristics. The former were shown to degrade into the latter under working-voltage stress, providing a latent failure mechanism. The 'jumpy' nature of the characteristic degradation suggested spasmodic defect expansion under Joule heating.

3. The Syrzycki methodology [8] was employed in order to model the transistors at their various stages of degradation. The effect of degradation on CMOS operation was also examined.

However, the reported experimental data is not sufficient to provide a complete picture of the degradation processes. This work merely demonstrates that damaged devices *do* degrade under working-voltage stress and suggests some explanations as to why this occurs. It also provides a springboard for a deeper study, the results of which are to be published later in the year.

6. References

[1] McAteer,O.J., Twist,O.J., Walker,R.C., *"Identification of Latent ESD Failures"*, Proc. 2nd. EOS/ESD Symp., pp.55-63, 1980.

[2] McAteer,O.J., Twist,O.J., Walker,R.C., *"ESD Latent Failures"*, Proc. 4th. EOS/ESD Symp., pp.41-8, 1982.

[3] Crockett,R.G.M., Smith,J.C., Hughes,J.F., *"ESD Sensitivity and Latency Effects of some HCMOS Integrated Circuits"*, Proc. 6th. EOS/ESD Symp., p.196-201, 1984.

[4] Aur,S, Chatterjee,A, Polgreen,T, *"Hot Electron Reliability and ESD Latent Damage"*, Proc. 26th. IEEE Reliability Physics Symp., pp.15-18, 1988.

[5] Holmes,G.C., *"An Investigation into the Effects of Low Voltage ESD Transients on MOSFETs"*, Proc. 7th. EOS/ESD Symp., pp.170-4, 1985.

[6] Amerasekera,E.A., Campbell,D.S., *"Electrostatic Pulse Breakdown in NMOS Devices"*, Qual. & Rel. Eng. Int., 2, pp.107-116, 1986.

[7] Amerasekera,E.A., Campbell,D.S., *"An Investigation of the Nature and Mechanisms of ESD Damage in NMOS Devices"*, Sol. St. Elect., 32(3), pp.199-206, 1989.

[8] Soden,J, Hawkins,C.F., *"Test Considerations for Gate Oxide Shorts in CMOS ICs"*, IEEE Design and Test of Computers, pp.56-64, August 1986.

[9] Syrzycki,M, *"Modelling Gate Oxide Shorts in MOS Transistors"*, IEEE Trans. Computer-Aided Design, 8(3), pp.193-202, 1989.

[10] Chen,I-C, Holland,S.E., Hu,C, *"Electrical Breakdown in Thin Gate and Tunneling Oxides"*, IEEE Trans. Electron Dev., ED32(2), pp.413-22, 1985.

[11] Tunnicliffe,M.J., Dwyer,V.M., Campbell,D.S., *"Experimental and Theoretical Studies of EOS/ESD Oxide Breakdown in Unprotected MOS Structures"*, Proc. 12th. EOS/ESD Symposium, pp.162-8, 1990.

[12] Tunnicliffe,M.J., Dwyer,V.M., Campbell,D.S., *"A Model for ESD Oxide Breakdown in MOS Transistors"*, European ESD Symposium, Eindhoven, Netherlands, 3-5 Sept. 1991.

[13] Tunnicliffe,M.J., Dwyer,V.M., To be published, 1992.