Application and Integration of Quantum-Effect Devices for Cellular VLSI

Thesis by

Harold Levy

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1995

(Defended December 12, 1994)

© 1995

Harold Levy

All rights reserved

Acknowledgements

A mole of thanks go to my advisor **Tom McGill** for providing the resources and political exposure that few graduate students get to experience; his pursuits of research excellence and laboratory safety were greatly appreciated, and our debates on technological issues were always enlightening and amusing.

I would also like to thank **Tom McGill, Carver Mead, Demitri Psaltis, Axel Scherer**, and **Amnon Yariv** for serving on my thesis committee.

My friends and colleagues in the McGill group have contributed extensively to this project and my happiness over these past years; I am particularly indebted to **Doug Collins** for his insights on experimental science, his sense of humor, and his zeal for scuba diving. **David Ting** was an especially good teacher of things theoretical, and I had many fruitful discussions with **Ron Marquardt, Mark Phillips, Johanes Swenberg**, and **Mike Wang**, as well as **Ed Croke, Yixin Liu, Rob Miles, Peo Pettersson, Chris Springfield**, and **Ed Yu**. I must also thank **Doug Collins, David Chow**, and **Jan Söderström** for growing the wafers used in this project and for helping me with my research at the beginning. Of course without **Marcia Hudson**, the McGill group secret weapon, I would have struggled daily with the bureaucracy; her cheer always made the workday enjoyable.

Tobi Delbrück, Chuck Neugebauer, Buster Boahen, Bhusan Gupta, John Harris, John Lazzaro, and **Ogden Marsh** were my connections to silicon reality, and they deserve a double high-five for their willingness to listen to my ramblings about alien devices and material systems and to provide valuable feedback. I also thank **Axel Scherer** for his expert advice on etching and deposition, and I enjoyed stimulating discussions with **John Hopfield** about neural networks and with **Jim McCaldin** about biological applications of solid-state devices. My friends **Greg Willette, Hong Jiao**, and **Steve Winters** taught me much at Caltech as well.

I am sure my sanity over these years was maintained by the interaction I had with my friends

outside Caltech, particularly **Tom Mucciaro, Richard and Cecilia Levin**, and **Russell Wolf**.

I would like to thank my family for all of the love and support they have eagerly given me all these years. **Sandy and Fran Levy** have been optimal parental units, and **David Levy** has always been an awesome dude and a brother.

Finally, I thank my wife **Donna Levy** for making it all worthwhile.

Abstract

Cellular VLSI is that subclass of electronic systems for which small perturbations in a repeated cell design can dramatically influence the cost and performance of the entire system. This thesis presents examples of how the room-temperature quantum effects of tunneling and resonance may be used to condense the functionality of many conventional VLSI devices into a smaller and more efficient subunit, thus yielding tremendous benefits for the system as a whole. In particular, two and three-terminal applications of a complimentary pair of quantum-effect devices, the resonant-tunneling diode and the tunneling-switch diode, are presented.

The first example is an image-segmentation network for machine vision, implemented by using resonant-tunneling diodes in one and two-dimensional networks to extract boundaries between regions of constant spatial texture. In this case a single quantum-effect device may replace up to thirty-three CMOS transistors per pixel.

The second example is an artificial neural-network processor based on multistate resistors for synaptic conductances. These programmable resistors were produced by combining a vertically-integrated stack of resonant-tunneling diodes with a resistive load and a single MOSFET driven in its ohmic region. This macrostructure has the potential to provide synaptic changes on the picosecond time scale at length scales well below one micron.

The third example is a current-mode transistorless memory array based on a two-dimensional network of cells containing only a single tunneling-switch diode and a resistive load. The resulting system has the potential for reaching more than an order-of-magnitude more cell density than state-of-the-art DRAM arrays, while operating at state-of-the-art SRAM speeds and reasonable power consumption.

Contents

Acknowledgements iii

Abstract v

Contents vii

List of Figures xi

List of Tables xv

1 INTRODUCTION **1**

MOS EVOLUTION 1

QUANTUM EFFECTS 3

THESIS ROAD-MAP 4

REFERENCES 5

I DEVICES **7**

2 THE RESONANT-TUNNELING DIODE **9**

DEVICE PHYSICS 9

Intraband Devices 10

Interband Devices 14

The Tunnel Diode 16

CIRCUIT PRINCIPLES 21

Load-Line Analysis 21

Multistability 23

Oscillations 26

REFERENCES 31

3 **THE TUNNELING-SWITCH DIODE** **33**

DEVICE PHYSICS 33

2-Terminal Properties 34

3-Terminal Properties 43

CIRCUIT PRINCIPLES 43

Load-Line Analysis 46

Circuit Duality 46

REFERENCES 49

II **SYSTEMS** **51**

4 **IMAGE SEGMENTATION** **53**

BIOLOGY 53

CIRCUIT PRINCIPLES 57

Resistive Networks 57

Feature Extraction 59

Nonlinear Networks 61

1D DEMONSTRATION 63

TECHNOLOGY COMPARISON 66

2D SIMULATIONS 67

REFERENCES 73

5 **NEURAL NETWORKS** **75**

SYSTEM PRINCIPLES 76

Neurons & Synapses 76

Multilayer Networks 77

CIRCUIT PRINCIPLES 78

Vector-Matrix Multipliers 78

RTD Synapses 78

HARDWARE DEMONSTRATION 82

REFERENCES 87

6 COMPUTER MEMORY **89**

SYSTEM PRINCIPLES 89

Conventional Voltage-Mode Memory 89

TSD Transistorless Current-Mode Memory 91

FABRICATION ISSUES 98

PERFORMANCE ESTIMATES 100

SCALING ISSUES 102

OTHER ISSUES 104

REFERENCES 105

List of Figures

1.1 Evolutionary cycle for MOS technology 2

2.1 Intraband RTD quantum-mechanics 10
2.2 Intraband RTD current-voltage characteristic 11
2.3 Intraband RTD energy functionals 13
2.4 Interband RTD quantum-mechanics 14
2.5 Interband RTD current-voltage characteristic 15
2.6 Interband RTD energy functionals 16
2.7 Tunnel diode band-energy profile 17
2.8 Tunnel diode current-voltage characteristic 18
2.9 Tunnel diode energy functionals 19
2.10 RTD load-line analysis 22
2.11 RTD stack current-voltage characteristic 24
2.12 RTD stack load-line analysis 25
2.13 NDR biasing scheme for obtaining AC gain 26
2.14 RTD generalized bias circuit 27
2.15 Frequency response of AC gain 28
2.16 Effect of unstable NDR oscillations 29

3.1 TSD device structure 34
3.2 TSD *I-V* characteristic 35
3.3 TSD band-energy diagrams ($V = 0$, $V < 0$) 37
3.4 TSD band-energy diagrams ($V > 0$) 39
3.5 TSD *C-V* characteristic 42

3.6 Charge-injection by light 44
3.7 Charge-injection by current 45
3.8 TSD load-line analysis 46
3.9 RTD/TSD duality comparison 47

4.1 Illustration of the vertebrate retina 54
4.2 Response functions in the retina 55
4.3 Retina response to a luminance step 56
4.4 One-dimensional resistive network 57
4.5 1D resistive network simulation 60
4.6 Nonlinear network devices 62
4.7 1D RTD segmentation network 64
4.8 1D segmentation comparison 65
4.9 Resistive fuse schematic 66
4.10 2D RTD segmentation network 68
4.11 Modified Newton-Raphson simulation 69
4.12 Mandril segmentation 70
4.13 Lena segmentation 71

5.1 Neuron activation function 76
5.2 Multilayer neural network 77
5.3 Vector-matrix multiplier schematic 79
5.4 Tristable load-line circuit 80
5.5 RTD synapse circuit 81
5.6 RTD multiplier stage 82
5.7 RTD *I-V* overlay 83
5.8 Binary activation function 83
5.9 Primary logic problems 84
5.10 XOR neural network test 85
5.11 AND, OR, and NOT neural network tests 86

6.1 Conventional voltage-mode RAM 90
6.2 Ideal TSD memory load-line behavior 91

6.3 TSD memories under static bias 92
6.4 TSD memory read scheme 94
6.5 Alternative TSD memory read scheme 95
6.6 TSD memory write scheme 96
6.7 2x2 demonstration circuit 97
6.8 TSD memory cell layout 98
6.9 TSD memory fabrication scheme 99
6.10 Donor concentration vs. epilayer thickness 103
6.11 Avalanche breakdown voltage vs. donor concentration 103

List of Tables

2.1 Comparison between the RTD and tunnel diode 20

4.1 Comparison between RTD and CMOS segmentation elements 67

6.1 TSD memory performance estimation 101

Chapter 1

INTRODUCTION

This thesis is an exploration of ways to harness, at the system level, the quantum-mechanical phenomena that are beginning to impede the evolution of silicon-based MOS technology. In particular, this thesis shows that the *cellular* class of systems stands most to gain by this idea, and that the examples contained herein could be just the beginning if a few material-system breakthroughs occur.

MOS EVOLUTION

The VLSI era has been a virtual supernova of progress since Jack Kilby and Robert Noyce filed their respective patents on the concept of integrated circuits in 1959 [1]. From watches to cars, tools to games, hardly any modern convenience does not exist in some form with a *chip* in it. Furthering the performance of these chips, whether it be in speed, density, or energy efficiency, is directly linked to the reduction of transistor dimensions [2].

Recently some researchers have sought to predict limitations to this reduction for silicon MOS technology by analyzing its evolutionary cycle [3, 4]. A flow diagram of this cycle based on the ideas presented in [3] is shown in Figure 1.1.

The entry point to the cycle is motivated by the desire to increase the system density, speed, or energy efficiency as denoted by the "START" label. To accomplish this, fabrication process designers develop the lithographic capability to produce transistors with shorter channel lengths. This length reduction will lower the punch-through voltage for the channel which in turn reduces the operating

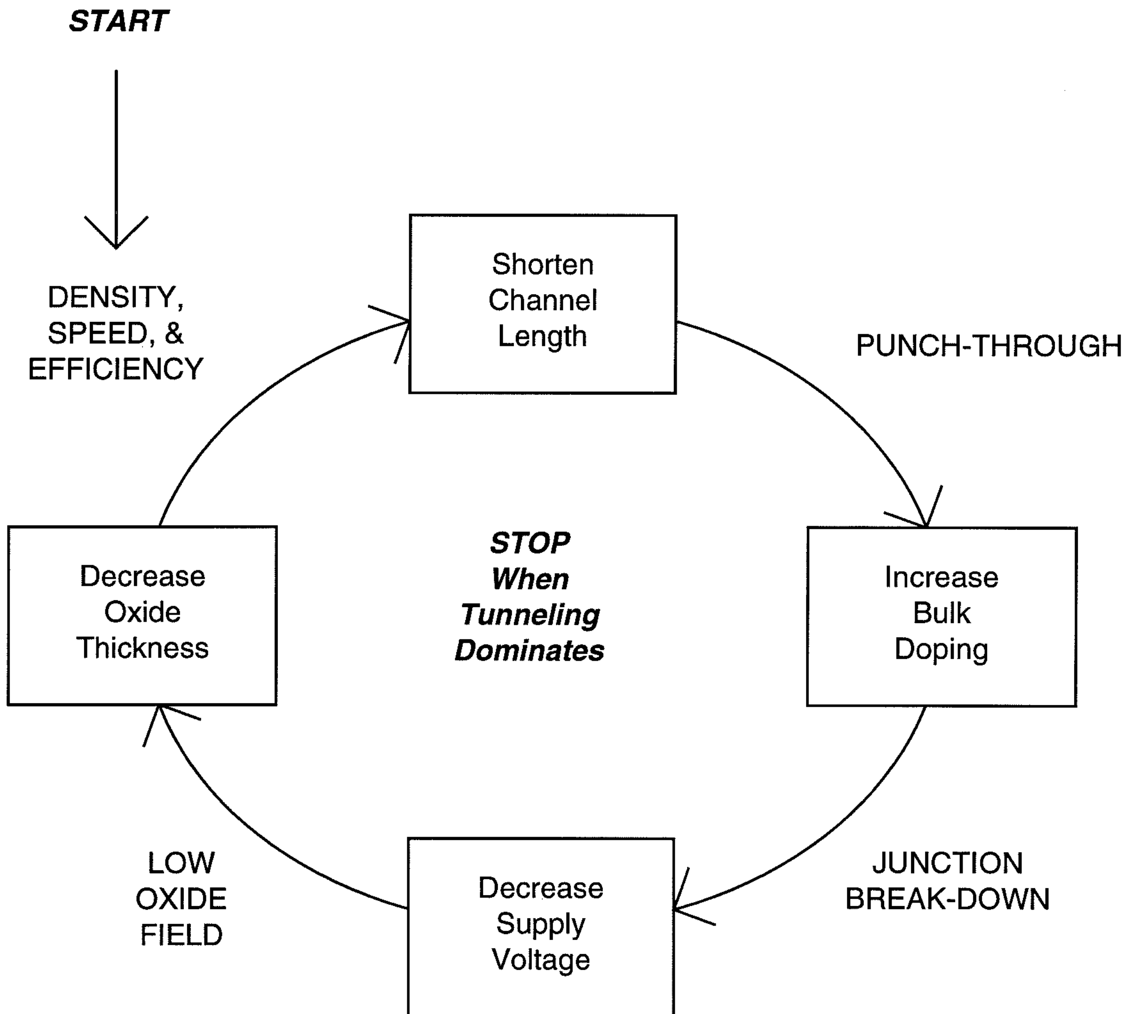


Figure 1.1: The evolutionary cycle for silicon-based MOS technology (adapted from [3]).

range over which the gate can influence the channel current.

To alleviate the punch-through problem, the doping concentration of the bulk semiconductor underneath the gate may be increased to provide more screening of the electric field at the drain, thus making the channel more difficult to deplete. Unfortunately this will then lower the breakdown voltage of the *pn*-junction formed between the drain and the channel, again reducing the operating range over which the gate can influence the channel current.

If the power-supply voltage for the transistors is reduced to below this breakdown voltage, then the electric field from the gate will be lowered and cause a reduced ability to control the semiconductor surface potential through the oxide. This field may be restored by using a thinner oxide, which would then complete one evolutionary cycle.

QUANTUM EFFECTS

If the evolutionary cycle for MOS technology is repeated enough, channel lengths and oxide thicknesses will soon approach a distance comparable to the probability wavelength of a charge carrier. When this happens, the tunneling current through the channel or gate will begin to become larger than the diffusion or drift current influenced by the gate potential, eventually preventing the device from performing useful computation [3].

Before this point, however, there will be an intermediate regime where tunneling exists but does not dominate the transistor physics. It is here that the opportunity to integrate classical and quantum effect devices together can be used to a possible system advantage, and this is the central idea behind this thesis.

Note that since silicon MOS fabrication does not yet routinely employ feature sizes in the quantum regime, quantum effect devices have been developed in various heterojunction test-bed material systems such as GaAs, GaSb, InAs, InP, etc. using epitaxial growth methods that can modulate material composition with monolayer control [5, 6]. If some of the device concepts developed with these materials can be transferred to silicon, then perhaps there will be a revolution in the way computation is performed with VLSI systems.

THESIS ROAD-MAP

This thesis is divided into two parts, one on devices and one on systems. The part on devices presents a review of the device physics and circuit principles for two quantum effect devices that work exceptionally well at room temperature: the resonant-tunneling diode and the tunneling-switch diode. The part on systems presents examples of how these devices can be used to greatly impact the performance of cellular VLSI systems, and these examples are the main contribution of this thesis.

The first system example is an image-segmentation network for machine vision applications, implemented by using resonant-tunneling diodes in one and two-dimensional networks to extract boundaries between regions of constant spatial texture [7]. In this case a single quantum-effect device may replace up to thirty-three CMOS transistors per pixel.

The second example is an artificial neural-network processor based on multistate resistors for synaptic conductances [8]. These programmable resistors were produced by combining a vertically-integrated stack of resonant-tunneling diodes with a resistive load and a single MOSFET driven in its ohmic region. This macrostructure has the potential to provide synaptic changes on the picosecond time scale at length scales well below one micron.

The third example is a current-mode transistorless memory array based on a two-dimensional network of cells containing only a single tunneling-switch diode and a resistive load [9]. The resulting system has the potential for reaching more than an order-of-magnitude more cell density than state-of-the-art DRAM arrays, while operating at state-of-the-art SRAM speeds and reasonable power consumption.

REFERENCES

[1] W. R. Runyan and K. E. Bean, *Semiconductor Integrated Circuit Processing Technology*. Reading, MA: Addison-Wesley, 1990.

[2] C. Mead and L. Conway, *Introduction to VLSI Systems*. Reading, MA: Addison-Wesley, 1980.

[3] C. A. Mead, "Scaling of MOS technology to submicrometer feature sizes," *Analog Integrated Circuits and Signal Processing*, vol. 6, no. 1, pp. 9–25, 1994.

[4] M. Nagata, "Limitations, innovations, and challenges of circuits and devices into a half micrometer and beyond," *IEEE J. Solid-State Circuits*, vol. 27, pp. 465–472, Apr. 1992.

[5] S. M. Sze, Ed., *High-Speed Semiconductor Devices*. New York, NY: Wiley, 1990.

[6] R. K. Watts, Ed., *Submicron Integrated Circuits*. New York, NY: Wiley, 1989.

[7] H. J. Levy, D. A. Collins, and T. C. McGill, "Extracting discontinuities in early vision with networks of resonant tunneling diodes," in *Proceedings of the 1992 IEEE International Symposium on Circuits and Systems*, (San Diego, California), pp. 2041–2044, May 10-13 1992.

[8] H. J. Levy and T. C. McGill, "A feedforward artificial neural network based on quantum effect vector matrix multipliers," *IEEE Trans. Neural Networks*, vol. 4, pp. 427–433, May 1993.

[9] H. J. Levy and T. C. McGill, "Transistorless, multistable current-mode memory cells and memory arrays and methods of reading and writing to the same," *Patent Disclosure CIT-2238*, filed August 1993.

Part I

DEVICES

Chapter 2

THE RESONANT-TUNNELING DIODE

The resonant-tunneling diode (RTD) is an experimental realization of the canonical *particle-in-a-box* construct in quantum-mechanics. The potential-energy box is created using epitaxial layers of semiconductors with different band-gaps; the band-gaps are arranged to form a modulated potential energy profile over a distance comparable to the probability wavelength of a charge carrier. This chapter briefly reviews RTD device physics and basic circuit principles.

DEVICE PHYSICS

The RTD came about in the early 1970's as a result of developing *molecular-beam epitaxy* (MBE), a fabrication technique that allows modulation of device composition with atomic monolayer control [1]. MBE can consequently produce semiconductor sandwiches with layers thin enough along the growth direction to permit quantum-mechanical effects (e.g. tunneling, resonance, and confinement) to dominate the functionality of the device. The layers are astonishingly uniform across at least 2-inch diameter wafers, so that the RTD device can be made with macroscopic or VLSI feature sizes and handle current-densities high enough for practical application [2]. Currently, other epitaxial techniques such as *chemical vapor deposition* (CVD) are now used in addition to MBE to fabricate a wide variety of epitaxial devices for both electronic and optical applications [3, 4].

There are two classes of RTD devices distinguished by where the charge carriers tunnel from and where they tunnel into. The first is the *intraband* class, those devices where the majority carriers in the electrodes are in the same band (conduction or valence) as the quantum-well state. The second

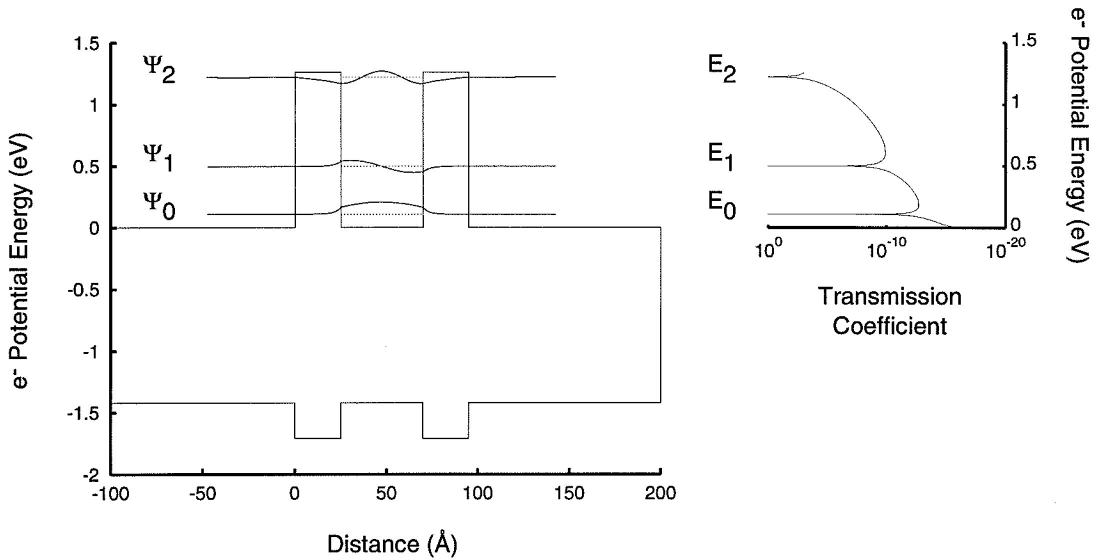


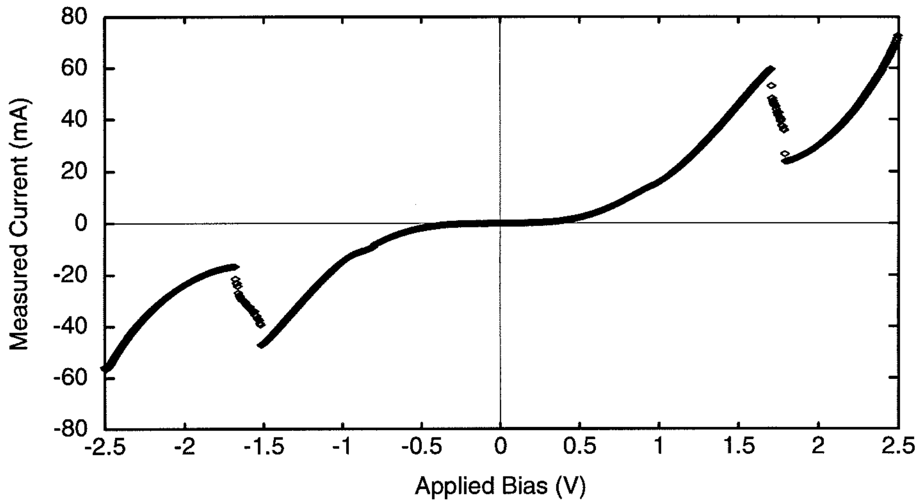
Figure 2.1: Unbiased potential-energy profile for an intraband RTD and its associated quantum-mechanical transmission.

is the *interband* class, distinguished by having the quantum-well state in the opposing band. As will be seen from the next two subsections, these two classes of RTD's exhibit different current-voltage (*I*-*V*) characteristics.

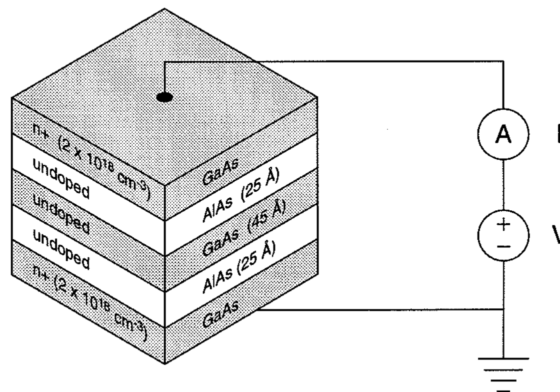
Intraband Devices

Figure 2.1 shows an approximation[1] of the unbiased potential-energy profile for a typical intraband RTD implemented with heterojunction MBE growth technology. For this specific example, the electrode and well materials are GaAs and the barrier material is AlAs. If the charge carriers are modeled as probability waves, only a discrete number of quasibound states Ψ_n exist in the well that satisfy the conservation of energy and momentum (i.e. Schrödinger's equation). These states are not bounded by the well as in the particle-in-a-box construct because of the finite barrier heights and widths. Furthermore, the transmission coefficient (the ratio of transmitted to incident probability densities) for this device is sharply peaked at only the energy levels E_n of the quasibound states. This suggests that quantum-mechanical transport through the device will be enhanced whenever an

[1]Throughout this section many high-order device physics issues will be ignored for the sake of brevity. Figure 2.1, for example, ignores band-bending, barrier-well state mixing, notch states, etc. Also note that in this and future plots the potential-energy reference is arbitrary and moved around at will to make the calculations simpler.



(a)



(b)

Figure 2.2: Measured *I-V* characteristic for an intraband RTD.

incident charge carrier *resonates* with a specific state within the quantum well. This is very similar to the Fabry-Perot effect used in optics to design thin-film optical filters [5].

Figure 2.2(a) shows the measured room-temperature *I-V* characteristic for MBE sample III-101 grown by Jan Söderström in the McGill-group laboratory. This device has the band-energy profile depicted in Figure 2.1, and the material composition shown in Figure 2.2(b). To explain either the intraband or interband RTD *I-V* characteristic, it is important to consider all of the charge transport modes. The direction of intended charge transport is along the growth direction (i.e. voltage is applied across the epitaxial layers), and there are two dominating mechanisms of transport, diffusion and resonant-tunneling. The drift contribution is negligible since the number of charge carriers in

the undoped barrier and well regions is typically many orders of magnitude less than that of the surrounding electrode layers. There is also a small amount of *scattering* transport of electrode charge carriers moving incidently non-parallel to the growth direction (more on this later).

Figure 2.3(a) shows the calculated energy-state population probability $f(E)$ for the device shown in Figure 2.2 at room-temperature. Note that the probability dies off exponentially, and since diffusion depends on how many charge carriers are above a net barrier energy, the diffusion current grows exponentially with a linear increase in potential energy (i.e. applied voltage). Hence for large applied voltages (e.g. $|V| \gtrsim 2$ volts in Figure 2.2(a)), most of the electrons are just diffusing across the device without the need for resonant-tunneling. This precludes energy levels $E_{n>0}$ from contributing significantly to the I-V characteristic at room temperature.

Figure 2.3(b) shows the density of energy states $g(E)$ above the conduction band edge E_C; the concentration of available free electrons at a particular energy outside the quantum well is thus the product $g(E)f(E)$, plotted in Figure 2.3(c). The energies of the free electrons are dispersed across momenta parallel and perpendicular to the wafer plane, as in $E = E_C + (\hbar k_\perp)^2/(2m^*) + (\hbar k_{||})^2/(2m^*)$, which is plotted as the dark parabola in Figure 2.3(d). Note there is a continuum of these parabolas for $E > E_C$ and that they are populated according to Figure 2.3(c). For the quasi-bound states the perpendicular momentum is fixed according to the quantum-mechanically allowed values given by $E = E_n + (\hbar k_{||})^2/(2m^*)$. The energy dispersion for E_0 is plotted as the gray parabola in Figure 2.3(d), and since it is entirely contained within the electrode dispersion, all of it is available for tunneling from energy levels populated according to Figure 2.3(c) for $E \geq E_0$. This picture clearly shows that there are few electrons available for tunneling at zero bias, explaining the apparent threshold voltage in the intraband I-V characteristic (e.g. $|V| \lesssim 0.5$ volt in Figure 2.2(a)).[2]

When bias is applied, the voltage sets the potential energy difference between the two electrodes and E_0 is brought closer to E_C (and eventually below it) for the electrode at higher potential energy. The concentration of free electrons that can tunnel into E_0 is the integral of $f(E)g(E)$ along the E_0 dispersion; note that this number is a maximum when $E_0 = E_C$. When $E_0 < E_C$ the E_0 dispersion is no longer contained *at all* in the electrode dispersion so that no free electrons are available for tunneling. This explains the abrupt resonance drop-off in the intraband RTD I-V characteristic (e.g. $|V| \sim 1.5$ volts in Figure 2.2(a)). The current does not drop completely to zero because of the presence

[2]This threshold voltage can be minimized by using InGaAs in the well which has a lower E_C; however, there is always a kink at the origin in the I-V characteristic [2].

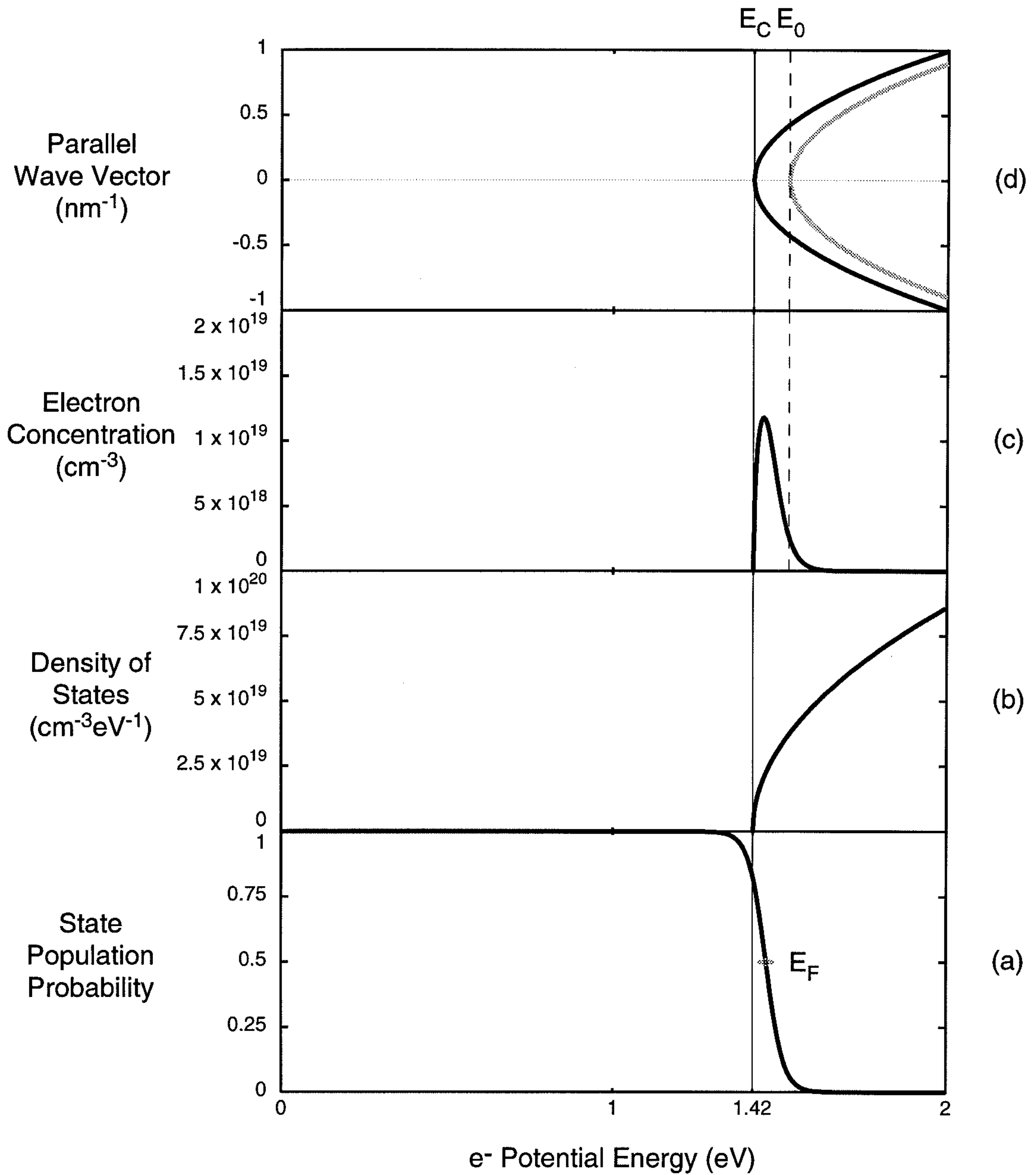


Figure 2.3: Various energy functionals for the intraband device shown in Figure 2.2.

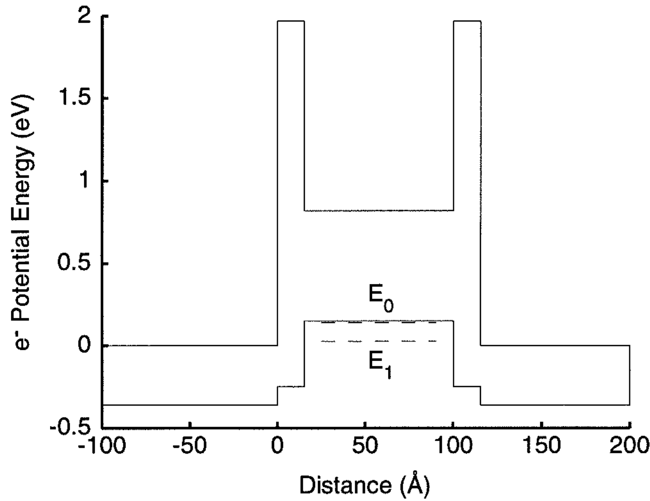


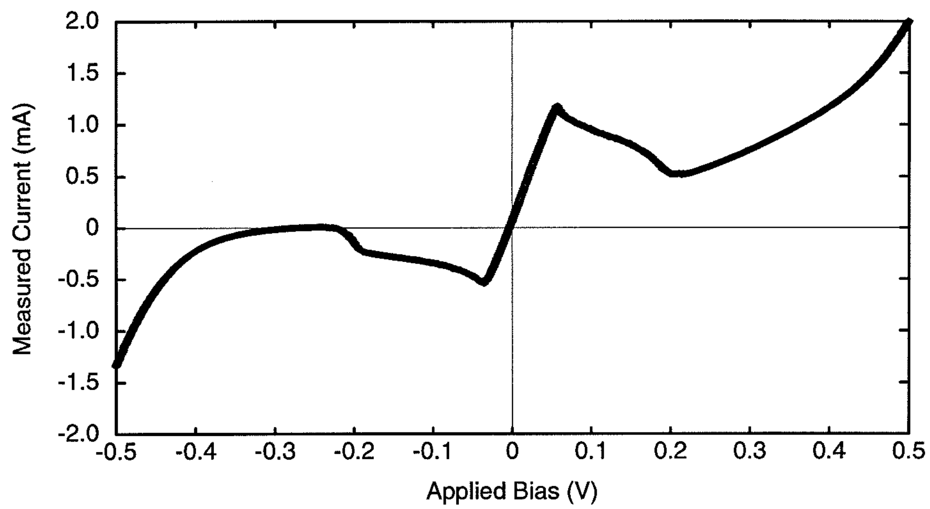
Figure 2.4: Unbiased potential-energy profile for an interband RTD and its associated quantum well energy levels.

of diffusion as well as inelastic scattering transport.

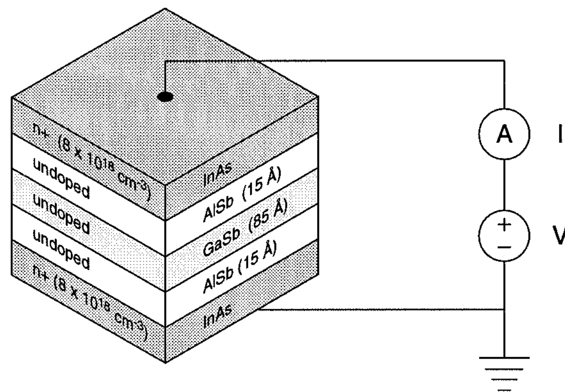
Interband Devices

In 1989 the first resonant-interband-tunelling diode (RIT) was fabricated using the InAs/GaSb/-AlSb material system [6]. To first-order, the RIT embodies all of the same device physics of the intraband RTD with the exception that the quantum-well states the electrode carriers tunnel into are in the opposing band, as shown in Figure 2.4. This makes for some interesting changes in the RTD *I-V* characteristic, as shown in Figure 2.5(a) for room-temperature MBE sample III-286 grown by Doug Collins in the McGill-group laboratory. This device has the band-energy profile depicted in Figure 2.4, and the material composition shown in Figure 2.5(b). The major points to note are that there is no voltage threshold to the resonance or kink at the origin, and that the resonance falls off much more gradually than that for the intraband RTD.

To explain these differences the corresponding $f(E)g(E)$ electron concentration is plotted in Figure 2.6(a) and energy dispersion in Figure 2.6(b). The E_0 dispersion points the opposite way from Figure 2.3(d) because the effective masses of electrons and holes are of opposite sign (they move in opposite directions in an applied electric field). Note that only *part* of the E_0 dispersion is contained within the electrode dispersion, so that the concentration of free electrons available for tunneling is the integral of $f(E)g(E)$ along this segment only. Since at zero bias this integral is sizeable, there is



(a)



(b)

Figure 2.5: Measured *I-V* characteristic for an interband RTD.

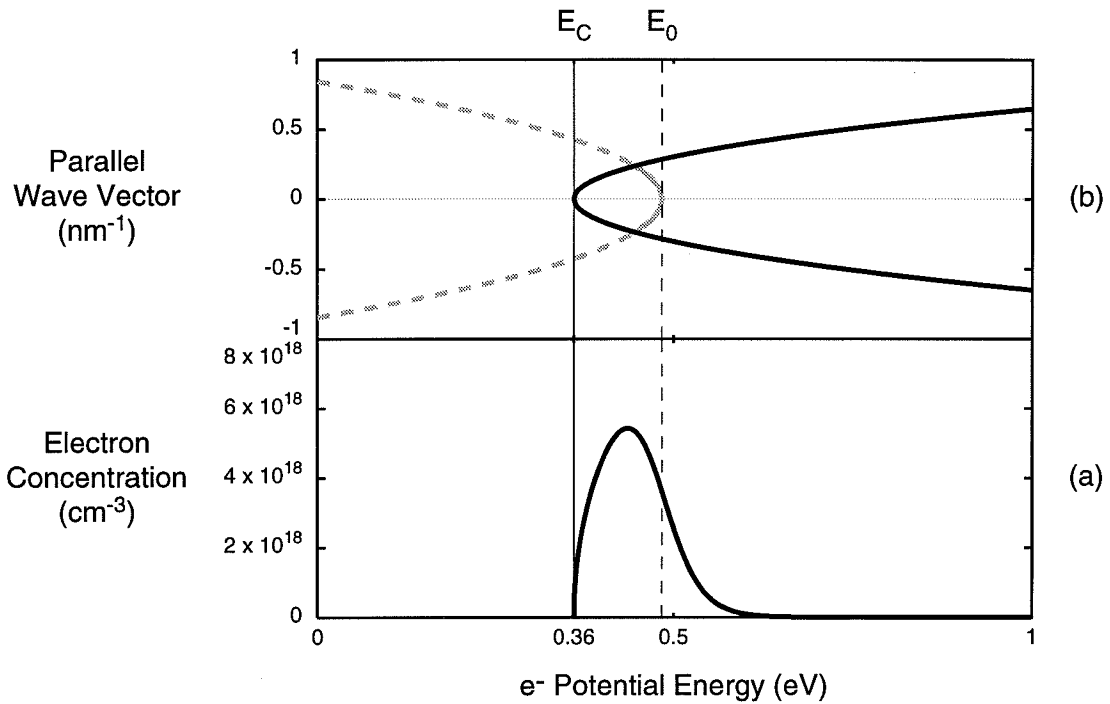


Figure 2.6: Various energy functionals for the interband device shown in Figure 2.5.

no apparent threshold voltage in the interband *I-V* characteristic. Furthermore, the resonance maximum occurs when the E_0 dispersion segment is positioned at the $f(E)g(E)$ peak with applied bias, instead of at E_C, so that there is a gradual resonance drop-off beyond this point until $E_0 < E_C$ and the resonance is extinguished.

The Tunnel Diode

Many of the RTD circuit principles are inherited from those of the first semiconductor tunneling device, the *tunnel diode,* discovered and explained by Leo Esaki in 1958 (who later won the Nobel prize for the work in 1973) [7, 8]. Figure 2.7 shows an approximation of the band-energy profile for a Ge tunnel diode studied in [8]. The tunnel diode is simply a *pn*-junction with adequate dopant concentrations to make the depletion region thin enough to permit tunneling. Since the interface between the *p* and *n* regions is a point of tremendous concentration gradient, the free majority carriers on both sides of the junction diffuse across until the resulting electric field caused by the remaining dopant ions causes a drift current of minority carriers that is equal and opposite to the diffusion current. The larger the diffusion gradient, the larger the electric field and thinner the depletion region.

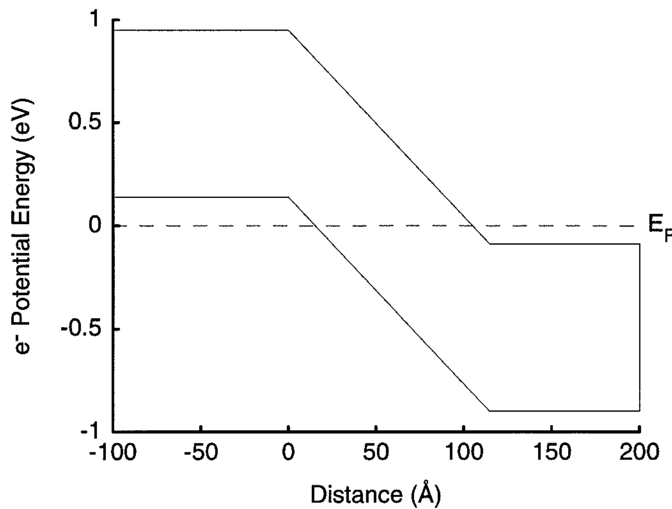
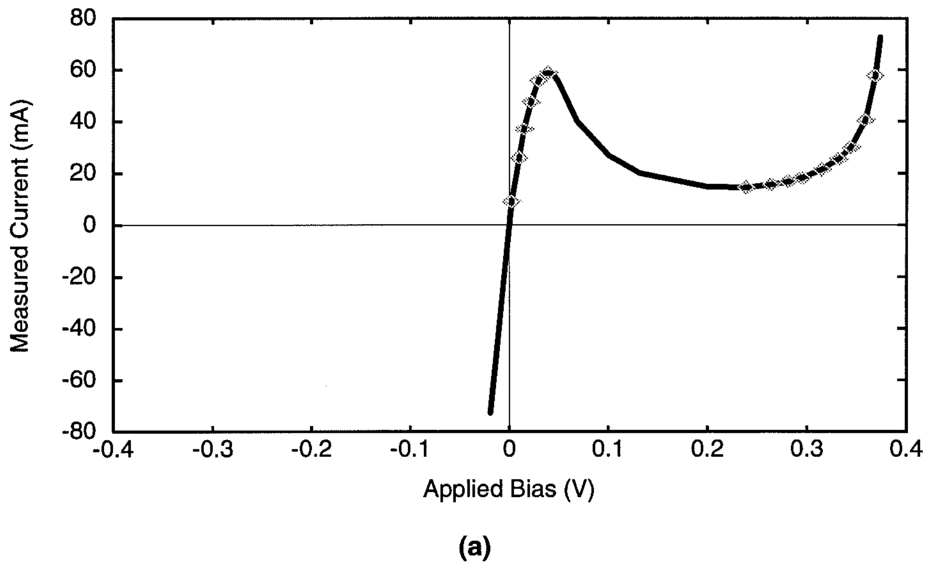


Figure 2.7: Tunnel diode band-energy profile showing ~100Å tunneling barrier formed by the depletion region.

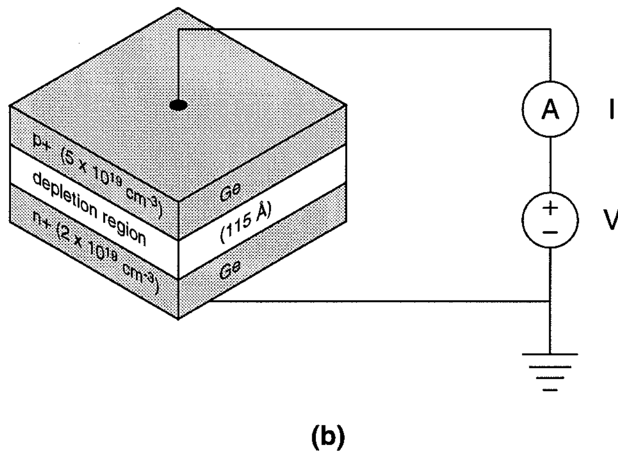
Figure 2.8(a) shows the measured *I-V* characteristic of the device displayed in Figure 2.7, and its material composition is shown in Figure 2.8(b) [8]. First note that the device behaves like a Zener diode in the reverse-bias direction, though the breakdown voltage is at the origin because the depletion region is thin enough for tunneling to begin with. To explain the *I-V* characteristic further, various calculated energy functionals for the device at zero-bias are plotted in Figure 2.9. The solid black lines are those attached to the conduction-band edge for the *n*+-Ge, and the solid gray lines are those attached to the valence band edge for the *p*+-Ge; reverse bias will move these two band edges further apart and forward bias will push them together.

Figure 2.9(a) shows the majority carrier concentrations for the two Ge layers as a function of energy. These are the electrons in the conduction band for *n*+-Ge and holes in the valence band for *p*+-Ge. Figure 2.9(b) shows the *compliment* of these concentrations; that is, holes in the conduction band for *n*+-Ge and electrons in the valence band for *p*+-Ge (note that these are not the minority carrier concentrations, i.e. holes in the valence band for *n*+-Ge and electrons in the conduction band for *p*+-Ge). Lastly, Figure 2.9(c) shows the energy dispersions for electrons and holes at the band edges surrounding the depletion region. Now consider the forward and reverse bias parts of the *I-V* characteristic separately.

In reverse bias, the overlap between majority carrier concentrations on either side of the depletion barrier is reduced while the overlap between the compliment concentrations is enhanced. This



(a)



(b)

Figure 2.8: Measured *I-V* characteristic for a tunnel diode (adapted from [8]).

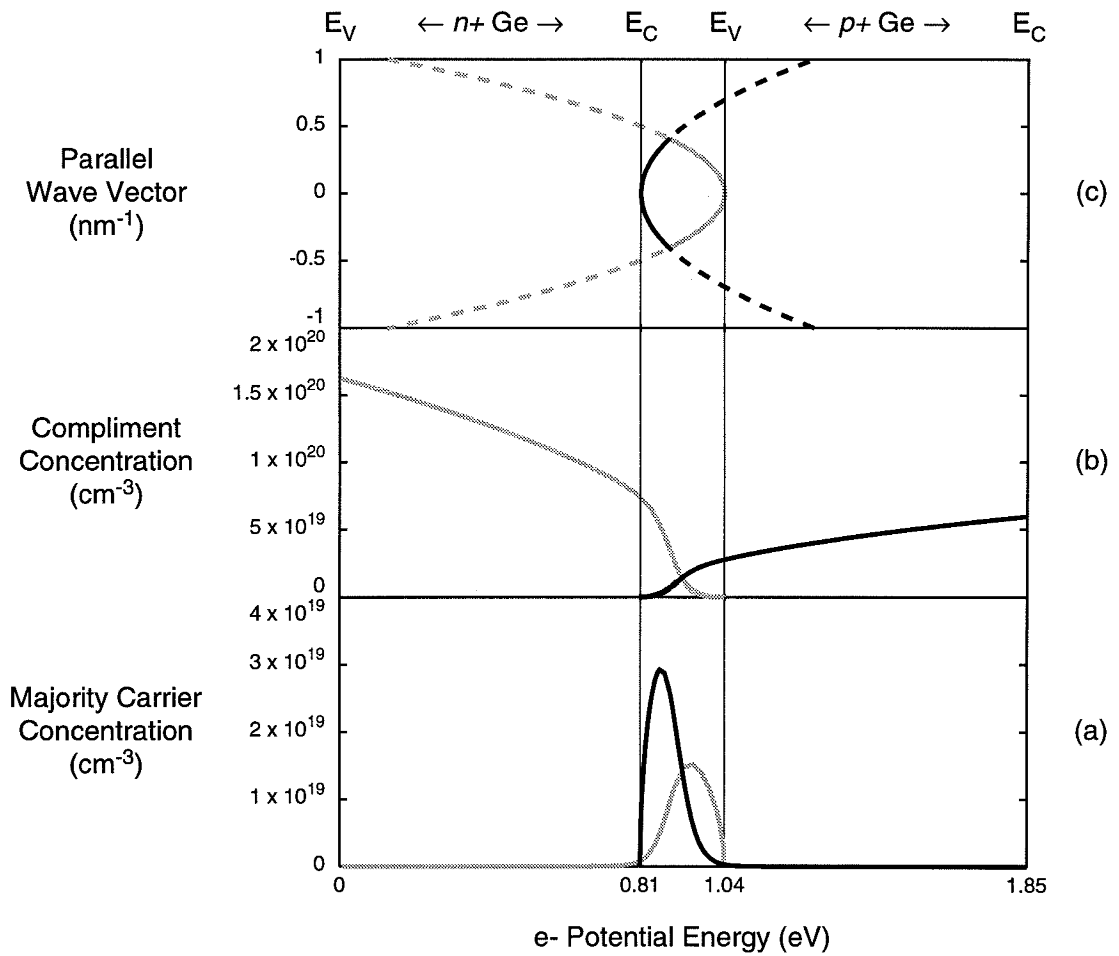


Figure 2.9: Various energy functionals for the tunnel diode shown in Figure 2.8.

Table 2.1: Brief comparison between the RTD and tunnel diode.

Attribute	Resonant Tunneling Diode	Tunnel Diode
I-V Characteristic	symmetric or asymmetric	asymmetric only
Typical PVR (300K)	4–70	< 20
I-V onset	linear or exponential	exponential only
Peak-voltage and peak-current	easily controllable and reproducible	difficult to control and reproduce

means that electrons will be tunneling from the valence band of the $p+$-Ge into hole states in the conduction band of the $n+$-Ge. In forward bias, the overlap between compliment concentrations is reduced and the overlap between majority carrier concentrations is enhanced. Hence electrons will be tunneling from the conduction band of the $n+$-Ge into hole states in the valence band of the $p+$-Ge. Eventually when the applied forward bias is large enough and the $n+$-Ge conduction band edge is pushed above the $p+$-Ge valence band edge, the tunneling current is shut-off and diffusion takes over. In all of these situations the valid overlaps are over the energy region spanned by the intersecting dispersion relationships shown in Figure 2.9(c). Note that this is the *direct* tunneling picture, which is valid in indirect band-gap materials like Ge for large potential differences between electrodes (e.g. $V \sim 1$ volt in Figure 2.7). For these materials there is also *indirect* tunneling between dispersions centered at different momenta, but this is small compared to direct tunneling because of the need for scattering to conserve momentum [9].

To conclude this section on device physics, note that all of these devices have a peak in current in one or more quadrants of their *I-V* characteristic, and as a result have regions of *negative differential resistance* (NDR). A figure of merit with NDR-type devices is the *peak-to-valley* current ratio (PVR). A comparison between RTD and tunnel diode attributes is given in Table 2.1 [1, 9–11]. The RTD differs from the tunnel diode in many respects, but perhaps most advantageously by having fabrication

design parameters (e.g. barrier thicknesses and well widths) apart from doping concentrations that allow the *I-V* characteristic to be tailored to specific requirements. In addition, fabrication of RTDs has proven to be both reproducible and uniform, and they have been implemented with a wide variety of semiconductor material systems [1, 12, 13].

CIRCUIT PRINCIPLES

The RTD is a two-terminal device that behaves like a nonlinear resistor (the capacitance across the structure is typically less than 10 $fF/\mu m^2$). Note that the RTD is *passive* like a resistor is; the current flowing through the device varies only with differential changes across the terminals (the *normal* input mode) and not when both terminals change together (the *common* mode). Hence an RTD may be useful in various signal-processing circuits that require large common-mode input ranges or large normal-mode to common-mode rejection ratios (CMRR). However, this can be a disadvantage in other types of circuits that require decoupling between the input and output currents for fan-out purposes (e.g. computer logic). There is neither *isolation* nor *restoration* of signals like there is in transistor circuitry, and hence RTD's can only augment transistor technology not replace it.

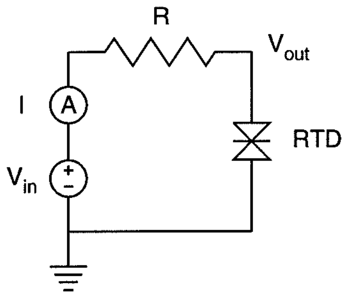
This augmenting comes from two key features the RTD offers in a single package: multistability and oscillation. These features are selected by various loading schemes that bias the RTD in either the positive differential resistance (PDR) or NDR regions of the *I-V* characteristic. Before describing these schemes in detail, a technique will be required to help visualize *I-V* fixed points in circuits with nonlinear devices.

Load-Line Analysis

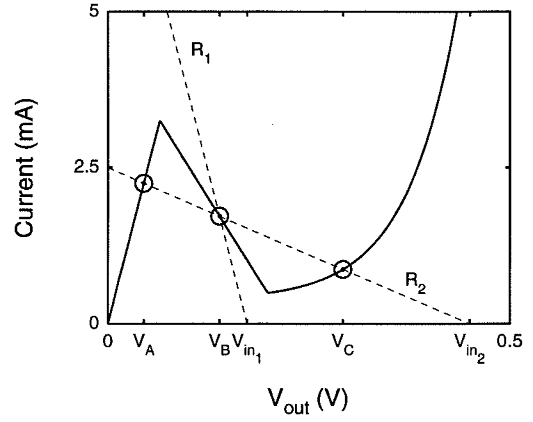
Load-line analysis is just a graphical visualization of the solution to Kirchoff's current law for a two element circuit. Consider Figure 2.10(a) for example; if the RTD was replaced by an ordinary resistor the circuit would just be a voltage divider with a V_{out} given by the solution to

$$I = \frac{(V_{in} - V_{out})}{R} \quad \Longleftrightarrow \quad I = \frac{(V_{out} - V_{in})}{R_{\text{RTD}}}$$

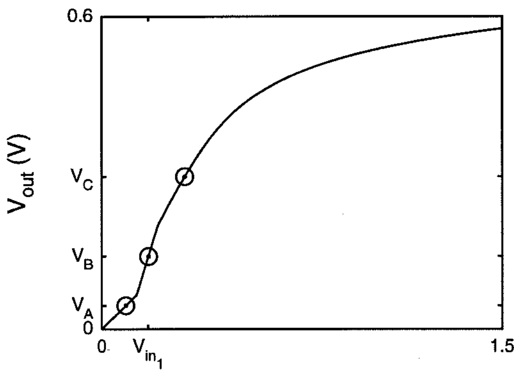
where R_{RTD} is the value of the resistor where the RTD is. The solution to this equation is of course just the intersection of the two *I-V* relationships, and the graphical presentation of this is especially enlightening when used with nonlinear *I-V* devices like the RTD.



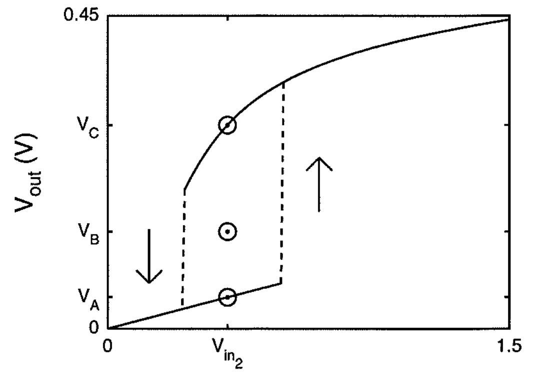
(a)



(b)



(c)



(d)

Figure 2.10: Load-line analysis for the RTD.

Figure 2.10(b) shows the RTD load-line analysis for two values of the resistor labelled R in Figure 2.10(a). At first glance it appears that there is one possible V_{out} solution for R_1 and three possible V_{out} solutions for R_2 for the specific values of V_{in} labelled V_{in_1} and V_{in_2}, respectively. Note that this is only a static picture of the circuit; to determine which of these states are *accessible* the dynamic picture for varying V_{in} must be examined.

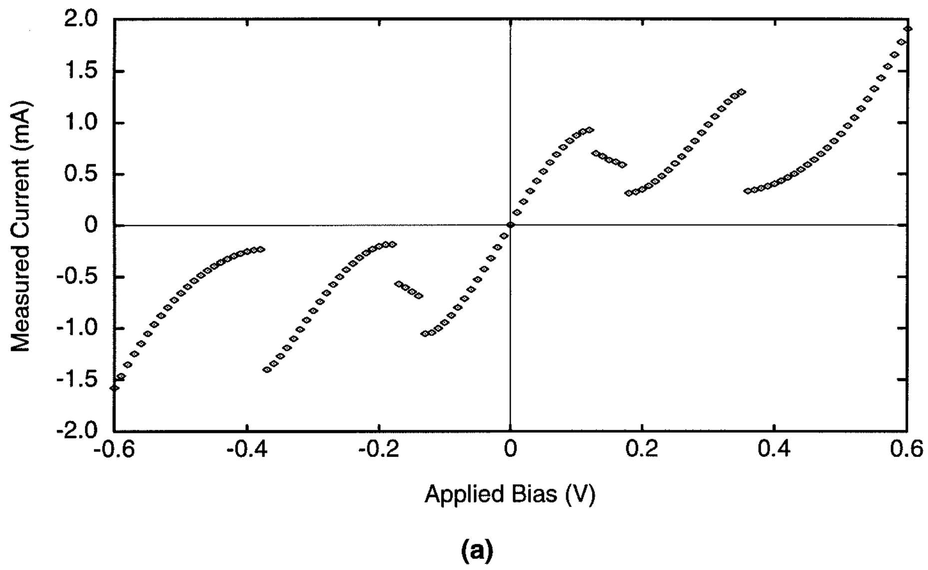
The dynamic picture for R_1 is shown in Figure 2.10(c) and for R_2 in Figure 2.10(d). The key points are that R_1 only intersects the RTD *I-V* characteristic at one point for all V_{in}, and that R_2 intersects the RTD *I-V* at two points for some range of V_{in}. The fixed point at V_B is *inaccessable* for R_2 at *any* value of V_{in}. Also notice that which $V_{out} \in \{V_A, V_C\}$ actually occurs depends upon the history of V_{in} shown by the arrows in Figure 2.10(d). This means the R_2 circuit is *hysteretic* and *bistable.*

Multistability

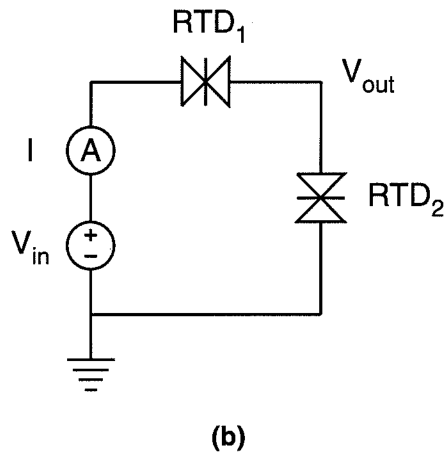
The previous section presented a bistable circuit based on the RTD *I-V* characteristic with one current peak in it. Various biasing schemes have been concocted to obtain *I-V* characteristics with multiple peaks in them to yield subcircuits with many more than two stable states. The simplest conceptually are those that use voltage-dividers to bias each RTD in an emitter-coupled array differently from the others so that they switch-off (i.e. they lose quantum transport) at different values of V_{in} [14, 15]. Unfortunately, the required voltage-divider circuitry is difficult to implement satisfactorily with VLSI technology because of its inefficient use of real-estate and reliance upon precise doping concentrations.

Another technique for constructing multiple-peak *I-V* characteristics is to use RTD devices in series [16, 17]. Figure 2.11(a) shows the resulting *I-V* characteristic for MBE sample III-370 grown by David Chow in the McGill-group laboratory, and Figure 2.11(b) shows the associated circuit schematic. This sample contains two RIT-type RTD's grown in series and separated by 4000 Å of InAs. The relatively thick intermediate layer is important for breaking any coherence that might develop between the wave-functions of the two quantum wells which would cause the stack to act as just a single device [18]. This means that this sample is an integrated *macrostructure*; the same *I-V* characteristic would result from wiring up two separate devices in series, and at least nine current peaks have been produced by such an integrated stack of RTD devices [19].

The explanation of the multiple *I-V* peaks for the serially-connected RTD's is subtle compared to the parallel case. If the devices are exactly identical then the voltage drop across each device is



(a)



(b)

Figure 2.11: Measured *I-V* characteristic for a double-RTD stack.

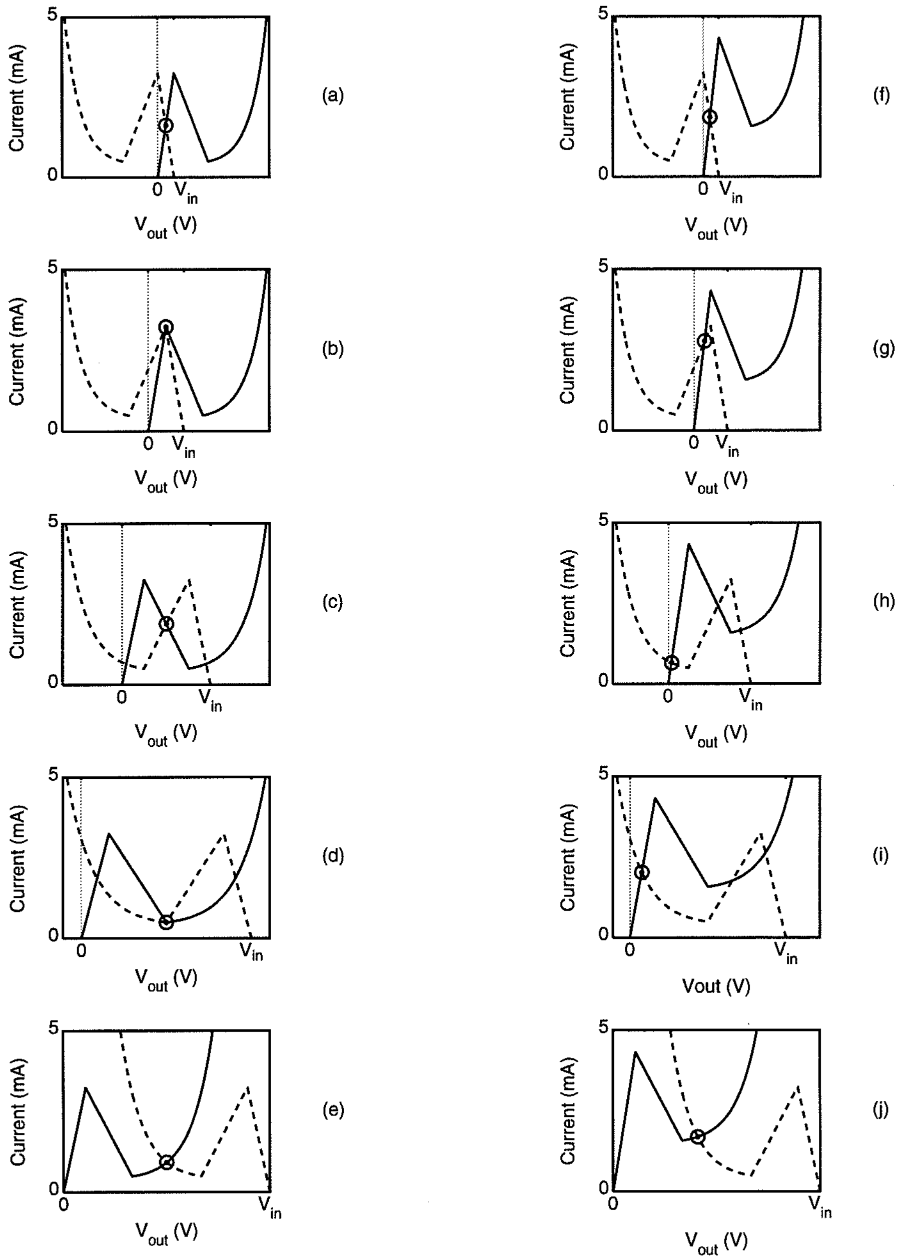


Figure 2.12: Load-line analysis for the RTD stack.

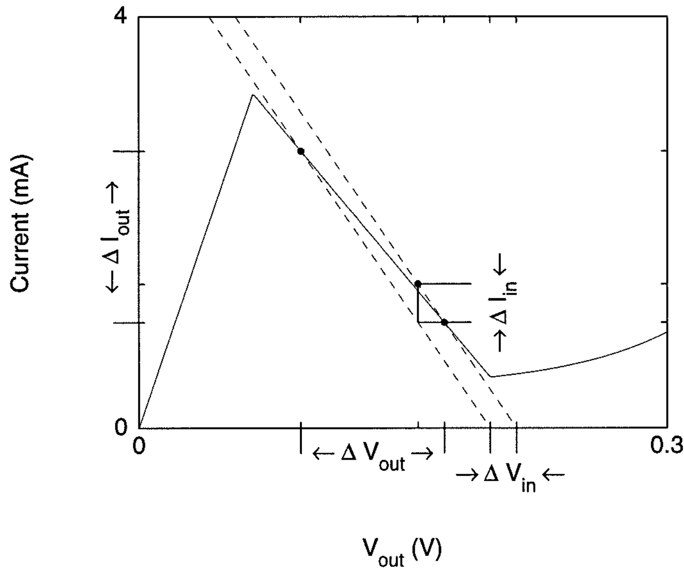


Figure 2.13: Biasing scheme for achieving AC gain from an NDR region.

exactly the same and the devices would switch-off at the same value of V_{in}, resulting in only *one I-V* peak. Figure 2.12(a) through Figure 2.12(e) show the load-line analysis for just this scenario. The solid *I-V* line is for RTD_2 and the dashed is for RTD_1. The value of V_{out} is indicated by the circle for various values of increasing V_{in}. In reality the devices are not exactly the same and hence the load-line analysis shown in Figure 2.12(f) through Figure 2.12(j) is more appropriate. Note that the device with *lower peak current* switches-off first.

Oscillations

Figure 2.13 shows the load-line analysis for biasing an RTD in the NDR region of its *I-V* characteristic. Note that when the load-line resistance is less than that of the NDR region it is possible to achieve AC voltage gain; for some given input voltage change $|\Delta V_{in}|$ there is an even greater output voltage change $|\Delta V_{out}|$ which is shown geometrically in the figure. The situation is similar for current. If $|\Delta I_{in}|$ is the current change through just the load-line for a given ΔV_{in}, then the current through both elements is amplified to a greater $|\Delta I_{out}|$. The additional power is coming from the bias source that is offsetting ΔV_{in} into the NDR region.

It is useful to explore the frequency dependence of this gain to determine the stability and response of NDR-based circuits. Figure 2.14 shows the generalized bias circuit for an RTD, including

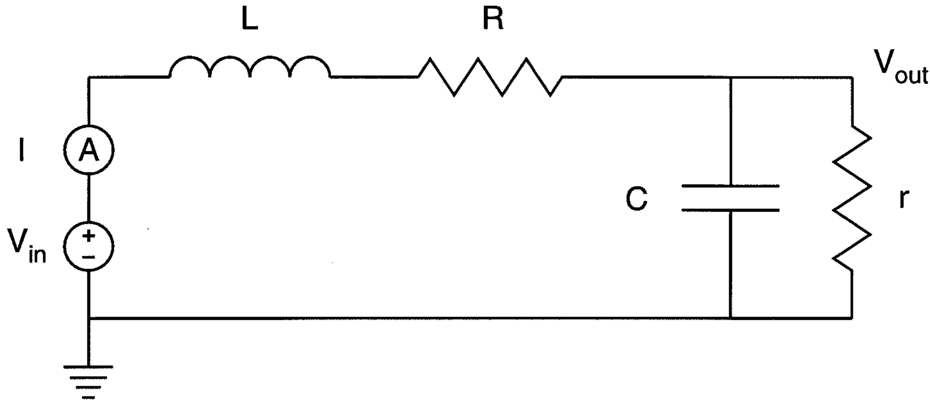


Figure 2.14: RTD generalized bias circuit.

extrinsic or parasitic inductance L, series resistance R, extrinsic or parasitic capacitance C, and the resistance of the RTD r which is negative in the NDR region we are confining V_{in} to. For sinusoidal inputs of the form $V_{in} = e^{st}$, where $s = \sigma + j\omega$, the impedances of the individual components are

$$\begin{aligned} Z_L &= Ls \\ Z_R &= R \\ Z_C &= \frac{1}{Cs} \\ Z_r &= r \end{aligned}$$

so that the total impedance of the circuit is given by

$$Z = Ls + R + \frac{r}{1 + rCs}.$$

The output signal V_{out} is measured across C and r which have a combined parallel impedance of $r/(1 + rCs)$. Kirchoff's current law for the circuit yields

$$\begin{aligned} \frac{V_{out}}{V_{in}} &= \frac{r/(1 + rCs)}{Z} \\ &= \frac{r}{R + r - \omega^2 rLC + j\omega(L + RrC)} \end{aligned}$$

where $s = j\omega$ for the purposes of extracting the frequency response. The voltage gain is just the magnitude of this ratio, and it is plotted in Figure 2.15(a) for $L = 0.1$pH, $R = 40\Omega$, $C = 1$pF, and $r = -49\Omega$ (typical values for a laboratory measurement of a typical RTD). Similarly for current we have

$$\frac{I_{out}}{I_{in}} = \frac{V_{in}/Z_{\text{with}}}{V_{in}/Z_{\text{without}}}$$

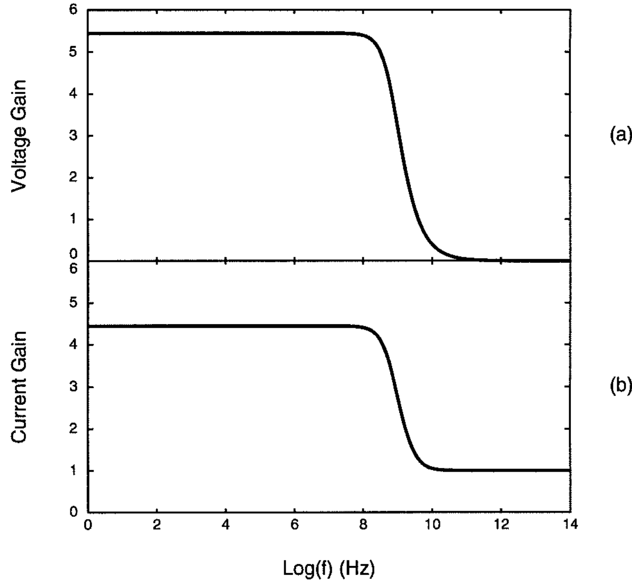


Figure 2.15: Frequency response of AC gain.

$$
\begin{aligned}
&= \frac{Ls + R}{Ls + R + r/(1 + rCs)} \\
&= \frac{Lj\omega + R}{Lj\omega + R - \frac{(jr)/(C\omega)}{r - j/(C\omega)}}
\end{aligned}
$$

where Z_{with} and Z_{without} are the total circuit impedances with and without the capacitor/RTD pair, respectively. The frequency response of the current gain is plotted in Figure 2.15(b). Note that the circuit can amplify microwave frequencies.

The voltage gain of the circuit goes to infinity at the zeros of the circuit impedance. Since the impedance is second-order in s there are two zeros for $Z(s) = 0$:

$$
\begin{aligned}
s &= -\frac{1}{2}s_* \pm \sqrt{\frac{1}{4}s_*^2 - \frac{1}{LC}\left[1 + \frac{R}{r}\right]} \\
s_* &= \frac{R}{L} + \frac{1}{rC}.
\end{aligned}
$$

Circuit stability implies signal dampening with $\text{Re}[s] = \sigma < 0$. To ensure this for the above s values note that there are two conditions, $s_* > 0$ and $1 + R/r > 0$. These may be condensed into the single expression $L/(r^2C) < -R/r < 1$ [20]. If the real and imaginary parts of the impedance are examined separately with $s = j\omega$, each part vanishes at a particular frequency. For the real part the *cut-off* frequency f_r is that frequency at which the RTD effectively shorts, and for the imaginary part the *resonant* frequency f_i is that frequency for which the circuit has unrestricted oscillation. These

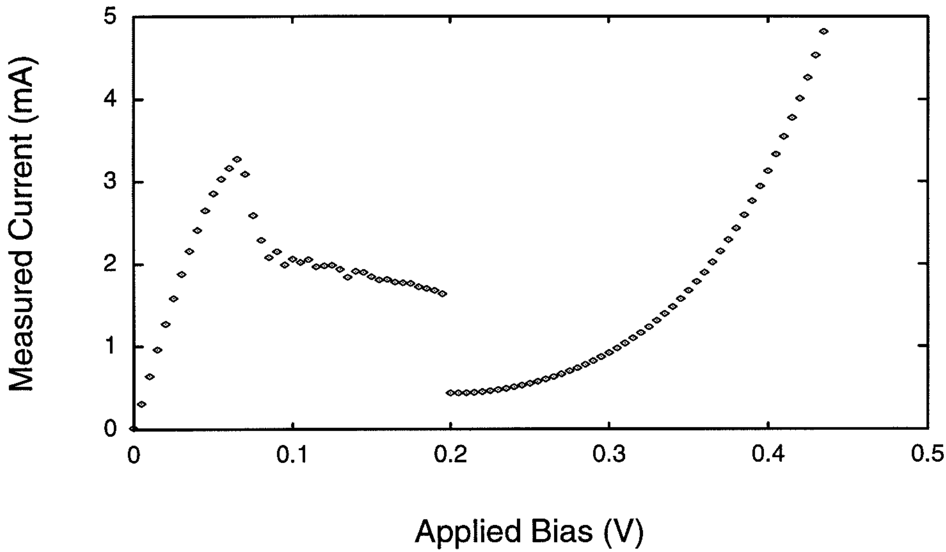


Figure 2.16: Effect of unstable NDR oscillations on the measured *I-V* characteristic.

two frequencies are given by

$$f_r = \frac{1}{2\pi(-r)C}\sqrt{\frac{-r}{R} - 1}$$

$$f_i = \frac{1}{2\pi}\sqrt{\frac{1}{LC} - \frac{1}{(rC)^2}}.$$

Note that if $f_r < f_i$ the circuit will short before it has a chance to oscillate uncontrollably; this is a stable operating regime and it satisfies $L/(r^2C) < -R/r < 1$. If $f_r \geq f_i$ any energy present at the resonant frequency will destabilize the circuit. When measuring the I-V characteristic of an NDR device it is common for the setup to act as an antenna for energy at the resonant frequency under $f_r \geq f_i$ conditions, resulting in an NDR plateau as shown in Figure 2.16.

The occurrence of plateaus may be circumvented by using very small devices (e.g. $< 25\mu m$ in diameter), since the reduced conductance will raise the magnitude of $-r$, thus decreasing f_r and increasing f_i so as to enhance the stability of the device in the presence of ambient microwave-frequency excitation. The f_r and f_i frequencies are typically in the 10-100 GHz range for devices the size of a wire-bond pad (e.g. $150\mu m \times 50\mu m$), and there are observations of RTD oscillations all the way up to 2.5 THz for $5\mu m$ square devices mounted in special apparatus and pumped with far-infrared wavelength lasers [21].

REFERENCES

[1] L. L. Chang, L. Esaki, and R. Tsu, "Resonant tunneling in semiconductor double barriers," *Applied Physics Letters*, vol. 24, pp. 593–595, June 1974.

[2] H. Riechert, D. Bernklau, J.-P. Reithmaier, and R. D. Schnell, "MBE growth of high performance GaAs/GaAlAs and InGaAs/GaAlAs double barrier quantum well structures for resonant tunneling devices," in *Resonant Tunneling in Semiconductors: Physics and Applications* (L. L. Chang, E. E. Mendez, and C. Tejedor, Eds.), New York, NY: Plenum Press, 1991.

[3] S. M. Sze, Ed., *High-Speed Semiconductor Devices.* New York, NY: Wiley, 1990.

[4] R. K. Watts, Ed., *Submicron Integrated Circuits.* New York, NY: Wiley, 1989.

[5] E. Hecht and A. Zajac, *Optics.* Reading, MA: Addison-Wesley, 1979.

[6] J. R. Söderström, D. H. Chow, and T. C. McGill, "New negative differential resistance device based on resonant interband tunneling," *Applied Physics Letters*, vol. 55, no. 11, pp. 1094–1096, 1989.

[7] L. Esaki, "New phenomenon in narrow Ge p-n junctions," *Physical Review*, vol. 109, p. 418, 1958.

[8] L. Esaki, "Discovery of the tunnel diode," *IEEE Trans. Electron Devices*, vol. ED-23, pp. 644–647, July 1976.

[9] S. M. Sze, *Physics of Semiconductor Devices.* New York, NY: Wiley, 2nd ed., 1981.

[10] D. J. Day, Y. Chung, C. Webb, J. N. Eckstein, J. M. Xu, and M. Sweeny, "Double quantum well resonant tunnel diodes," *Applied Physics Letters*, vol. 57, pp. 1260–1261, Sep. 1990.

[11] P. Cheng and J. S. Harris, Jr., "Improved design of AlAs/GaAs resonant tunneling diodes," *Applied Physics Letters*, vol. 56, pp. 1676–1678, Apr. 1990.

[12] K. Ismail, B. S. Meyerson, and P. J. Wang, "Electron resonant tunneling in Si/SiGe double barrier diodes," *Applied Physics Letters*, vol. 59, pp. 973–975, Aug. 1991.

[13] D. H. Chow, J. R. Söderstöm, D. A. Collins, D. Z.-Y. Ting, E. T. Yu, and T. C. McGill, "Novel InAs/GaSb/AlSb tunnel structures," *SPIE Proceedings on Quantum-Well and Superlattice Physics III*, vol. 1283, Mar. 1990.

[14] J. Söderström and T. G. Andersson, "A multiple-state memory cell based on the resonant tunneling diode," *IEEE Electron Device Letters*, vol. 9, pp. 200–202, May 1988.

[15] F. Capasso, S. Sen, A. Y. Cho, and D. Sivco, "Resonant tunneling devices with multiple negative differential resistance and demonstration of a three-state memory cell for multiple-valued logic applications," *IEEE Electron Device Letters*, vol. 8, pp. 297–299, July 1987.

[16] A. A. Lakhani and R. C. Potter, "Combining resonant tunneling diodes for signal processing and multilevel logic," *Applied Physics Letters*, vol. 52, pp. 1684–1685, May 1988.

[17] S. Sen, F. Capasso, D. Sivco, and A. Y. Cho, "New resonant tunneling devices with multiple negative resistance regions and high room temperature peak-to-valley ratio," *IEEE Electron Device Letters*, vol. 9, pp. 402–404, Aug. 1988.

[18] E. Wolak, B. G. Park, K. L. Lear, and J. S. Harris, Jr., "Variation of the spacer layer between two resonant tunneling diodes," *Applied Physics Letters*, vol. 55, pp. 1871–1873, Oct. 1989.

[19] A. C. Seabaugh, Y.-C. Kao, and H.-T. Yuan, "Nine-state resonant tunneling diode memory," *IEEE Electron Device Letters*, vol. 13, pp. 479–481, Sep. 1992.

[20] C. Kidner, I. Mehdi, J. R. East, and G. I. Haddad, "Power and stability limitations of resonant tunneling diodes," *IEEE Trans. Microwave Theory and Techniques*, vol. 38, pp. 864–872, July 1990.

[21] T. Sollner, W. Goodhue, P. Tannenwald, C. Parker, and D. Peck, "Resonant tunneling through quantum wells at frequencies up to 2.5 THz," *Applied Physics Letters*, vol. 43, pp. 588–590, Sep. 1983.

Chapter 3

THE TUNNELING-SWITCH DIODE

The tunneling-switch diode (TSD) is a quantum-effect member of a family of binary switching devices known as thyristors. Thyristors use a conductive potential-energy barrier in series with an asymmetrical *pn*- or *np*-junction to switch their device physics between a high-impedance depletion mode and a low-impedance inversion mode. The TSD uses a tunneling barrier to accomplish this and consequently exhibits extremely fast switching speeds. This chapter briefly reviews TSD device physics and basic circuit principles.

DEVICE PHYSICS

Yamamoto and Morimoto discovered the TSD[1] in 1972 in the course of their work on negative differential resistance (NDR) in metal-*np*-junction devices [1]. Their search for a more stable and reproducible device led them to the TSD, first made by growing a thin thermal oxide (< 100 Å) on top of an asymmetrical *pn*- or *np*-junction. Shortly thereafter Kroger and Wegener explored maximizing the yield of devices by using various other insulator materials and deposition methods; perhaps the most promising is SiO_xN_y deposited by chemical vapor deposition (CVD) [2, 3].

Researchers have shown that a tunneling barrier is sufficient but not necessary for producing the thyristor switching phenomenon. An incredibly wide variety of conductive potential-energy

[1] Unfortunately, this device has no canonical name; it is usually referred to by its layers, as in a "metal-insulator-semiconductor switch." This thesis uses "tunneling switch diode" to help distinguish the device from other thyristors as well as from MOS transistors.

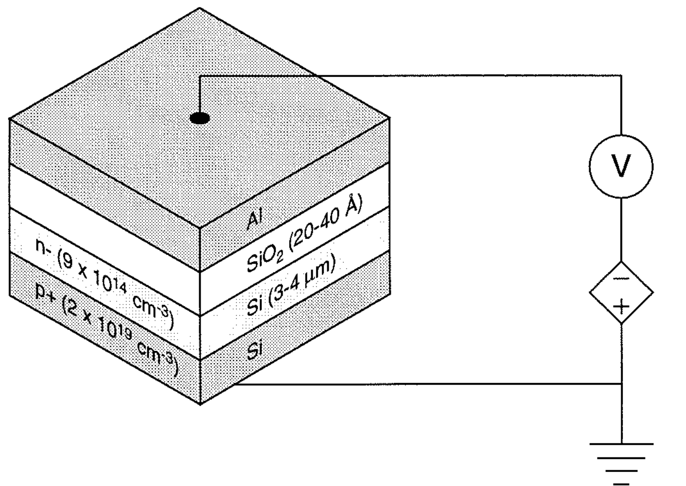


Figure 3.1: Typical device structure for a TSD.

barriers has been shown to work, including a 10^3-10^4 Å layer of intrinsic polysilicon [3] and even a 600-800 Å layer of tin oxide for high-speed gas sensing applications [4]. Of course a *pn*-junction may also be used, and the resulting *pnpn* structure is the famous Shockley diode widely used in power electronics [5]. As with these classical-physics thyristors, the TSD exhibits a variety of interesting two and three terminal device physics, and the next two subsections explore these sets of properties separately.

2-Terminal Properties

Figure 3.1 shows the device structure for a typical test TSD made in our lab. After obtaining the *n-/p+* wafer from an industrial CVD fab facility, a sample is cleaned by the RCA procedure [6, 7] and soaked in buffered HF until the solution beads on the semiconductor surfaces. The sample is then heated at 700 $^\circ C$ under O_2 gas for 15 minutes to produce a 20-30 Å SiO_2 tunneling barrier (this thickness is inferred from *C-V* measurements). We use a rapid thermal annealer for the heating to maximize reproducibility [8]. Next, electrodes are formed on top of the oxide by sputtering aluminum through a shadow mask (usually a scanning-electron microscopy grid). Finally an ohmic contact is made by removing the oxide from the *p+* layer with buffered HF and sintering a sputtered aluminum layer at 450 $^\circ C$ for 4 minutes under argon gas.

The measured *I-V* characteristic for a device made in this way is shown in Figure 3.2. The compliment of this device with a *pn*-junction instead of an *np*-junction behaves similarly for the opposite

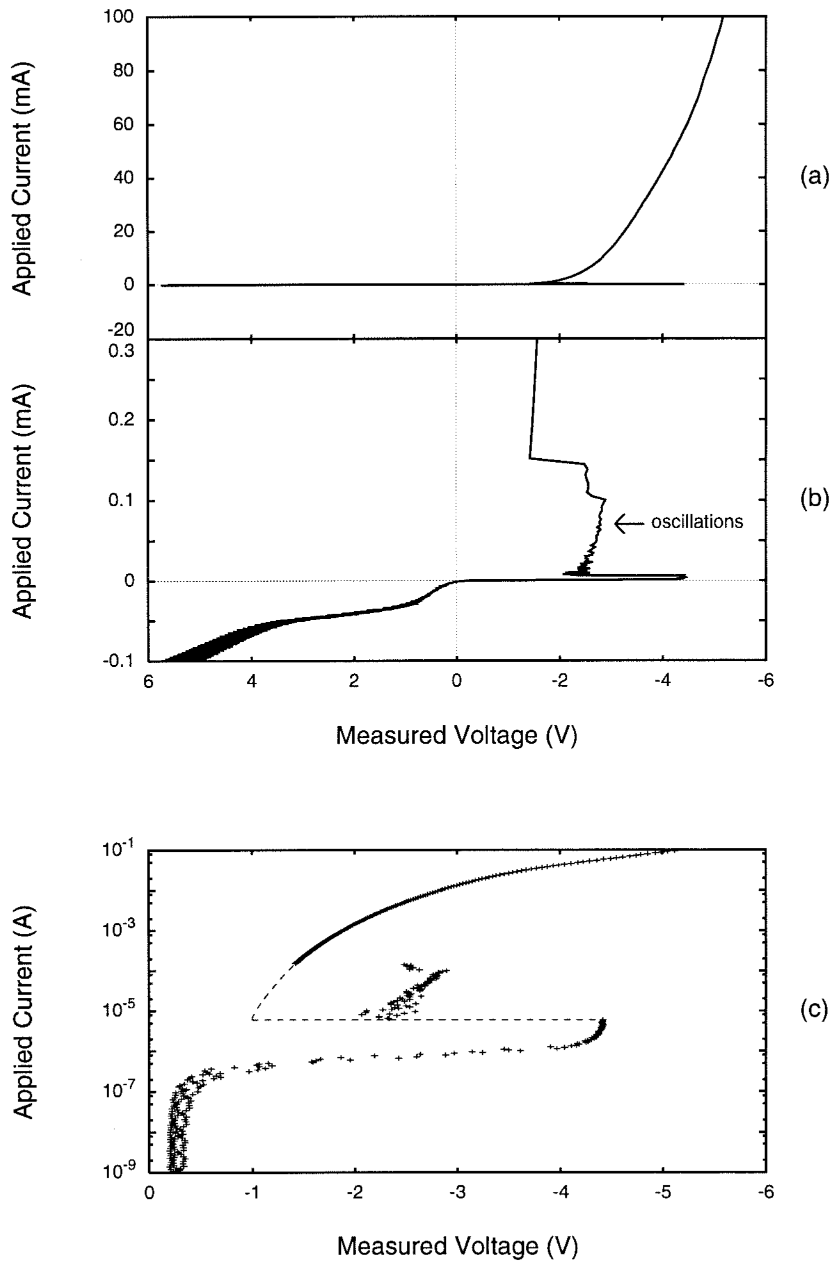


Figure 3.2: Measured *I-V* characteristic for a TSD shown at three different scales.

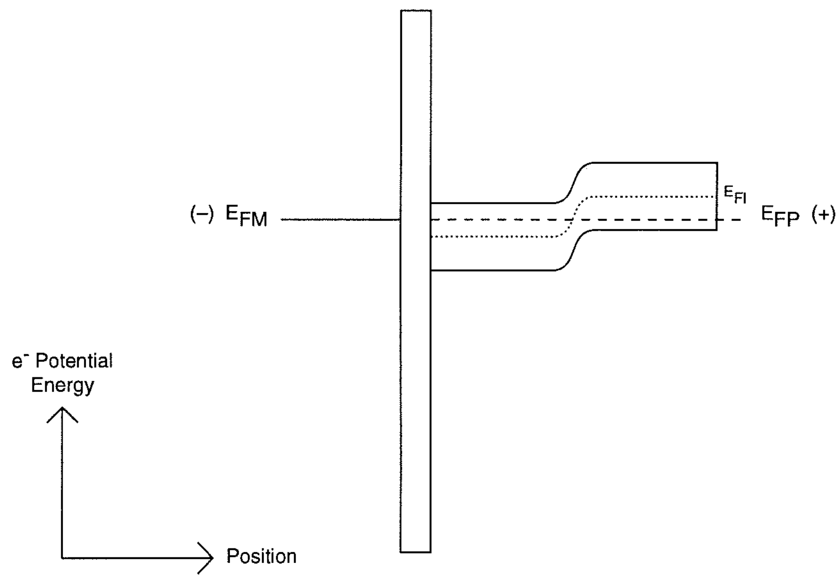
biasing scheme. The first thing to note is that the data was taken by injecting a variable current into the device and measuring the resulting voltage drop. This avoids hysteresis and protects the device in case it cannot handle the current density just beyond the peak-voltage in the low-impedance portion of the *I-V* characteristic (the device used for Figure 3.2 can handle much more current than that). Also note there are oscillations in the NDR region, which are typical for a big device with low NDR and hence a low resonant frequency in the presence of parasitic inductance. If the resonant frequency could be greatly increased, as it would for much smaller devices (this device was a 100 μm square), the *I-V* characteristic in the NDR region would likely follow that indicated by the dashed line in the figure. Other extrinsic features of Figure 3.2 include the roll-off in the high-current region due to parasitic series resistance and the bend in the low-current region due to the absence of the origin in a logarithmic plot.

To explain how this *I-V* characteristic comes about, consider the idealized band-energy diagrams in Figure 3.3 and Figure 3.4. Figure 3.3(a) depicts the zero-bias state of the device at thermal equilibrium, where the fermi-level of the metal E_{FM} is equal to the fermi-level of the p+ layer E_{FP}. The intrinsic fermi-level E_{FI} is shown for convenience in deducing depleted and inverted regions in later figures. These diagrams are idealized in the sense that they ignore the difference between the work-functions of the metal and the semiconductor, the surface states at the oxide-semiconductor interface, and any fixed oxide charge and/or oxide traps. Also note that the oxide potential-energy barrier is not based upon the band structure of crystalline SiO_2, but instead represents some average barrier for the amorphous state of the actual layer.

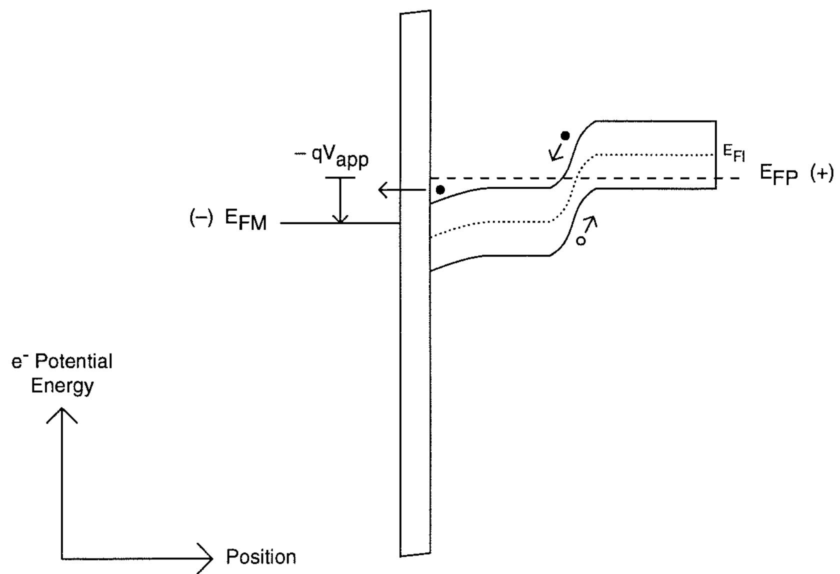
The conductivity of the barrier layer is paramount in producing the switching characteristic under applied-bias conditions. For insulators like SiO_2, the layer thickness must allow *direct tunneling* to prevent inversion at the surface of the n- layer in the high-impedance state[2]. This constrains the thickness of SiO_2 layers to 20-40 Å [11, 12].

Figure 3.3(b) illustrates what happens in reverse-bias, where because the barrier is more conductive than the np-junction, the junction bears most of the voltage drop. The dominant source of current in this case is thermal emission of electrons and holes from mid-gap traps in the depletion region of the np-junction, where the emitted electrons get swept to the barrier and tunnel through it

[2]Note that this is a fundamentally different regime than EEPROM structures that use Fowler-Nordheim tunneling and hot-electron injection to *selectively* control the conductivity of a SiO_2 barrier layer that is never thinner than ~40 Å [9–11]. In TSD devices the np-junction provides the high impedance, not the oxide.

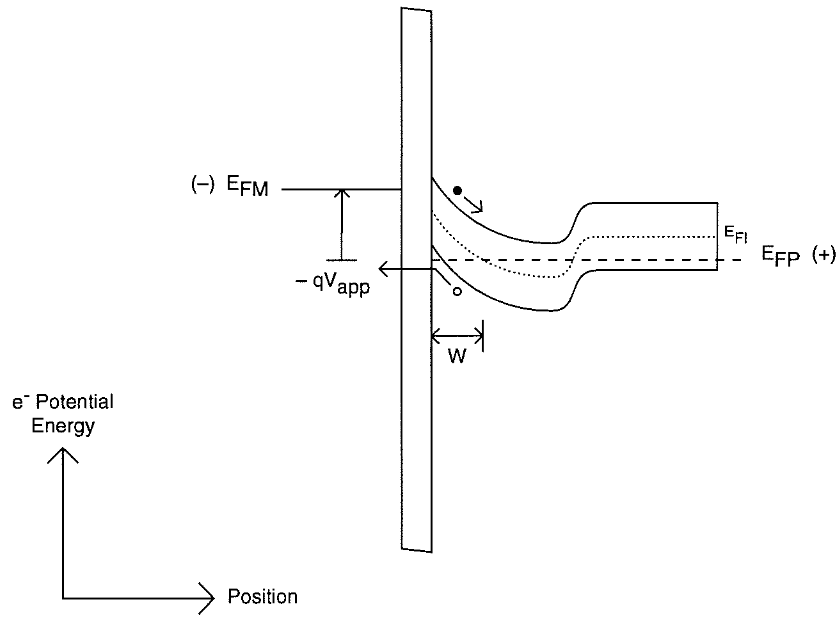


(a)

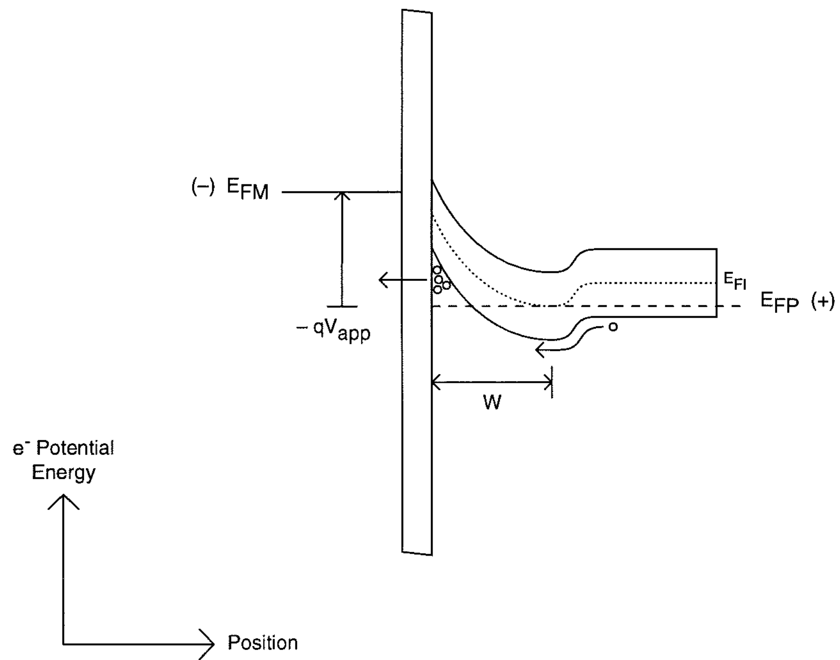


(b)

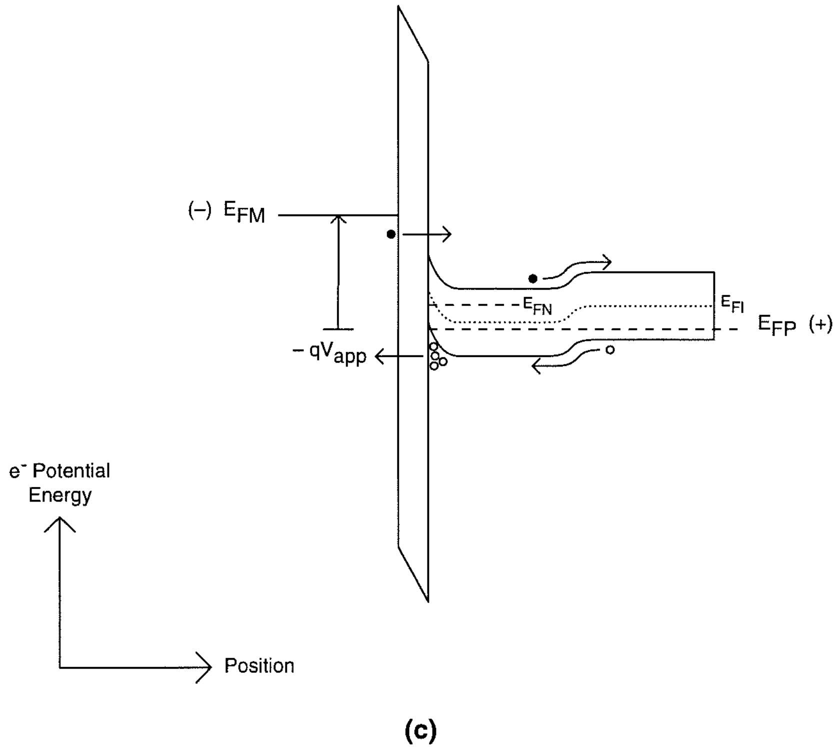
Figure 3.3: Idealized zero-bias (a) and reverse-bias (b) band-energy diagrams for a TSD.



(a)



(b)



(c)

Figure 3.4: Idealized forward-bias band-energy diagrams for a TSD before switching (a), during switching (b), and after switching (c) into the low-impedance state.

while the emitted holes get swept to the $p+$ layer. Consequently, this portion of the I-V characteristic is essentially the same as that of a reverse-biased diode.

The forward-bias picture is much more subtle. The easiest way to think about what is going on is to consider the screening of the electric field generated at the oxide electrode by charge density accumulated at the surface of the n- semiconductor layer. For example, if the oxide layer is too thick then there is no way for holes accumulated at the semiconductor surface to escape; these holes will then screen the electric field from the oxide electrode and prevent the charge density from changing deeper in the semiconductor, preventing switching from occurring. If the oxide layer is too thin then there is no way for holes to accumulate at all, and so the potential energy of the n- layer up to the np-junction depletion region effectively follows the potential energy of the electrode. This means the I-V characteristic will just be that of the np-junction and no switching will occur in this case either.

For intermediate oxide layer thicknesses, the tunneling barrier will act as a slight carrier bottleneck and some accumulation will occur, but not enough to completely screen-out the electric field from the oxide electrode, thus changing the charge density deeper in the semiconductor (i.e. deplet-

ing it) as the electric field gets stronger. Note that this means the potential drop is predominantly across the *n*- layer, and so the *np*-junction does not get forward-biased. This is why the peak voltage extends way beyond the canonical 0.7 V diode threshold; until the *n*- depletion extends all the way to the *np*-junction, it is the primary source of charge carriers via field-separated electrons and holes thermally emitted from mid-gap traps, as shown in Figure 3.4(a). This generation current is small-enough that the electrons recombine at the *np*-junction instead of appreciably lowering the junction barrier, and the same is true for any electrons tunneling-in from the conduction band of the oxide electrode.

The switching point can actually occur in response to one of three possibilities. The first possibility is that the depletion of the *n*- layer extends all the way to the *np*-junction, as shown in Figure 3.4(b), thus allowing it to be forward-biased by additional field from the oxide electrode. The second possibility, which would be more likely for heavier-doped *n* layers, is that the field in the depletion region becomes large enough for avalanche to occur, and hence the resulting electron current would raise the potential of the *n* layer and forward-bias the *np*-junction. The third possibility is that E_{FM} goes above the conduction-band edge of the *n*- surface allowing electrons to tunnel from the valence band of the metal into the conduction band of the *n*- layer and then onward to forward-bias the *np*-junction.

Once the *np*-junction is forward-biased, holes from the *p*+ layer can rush across and begin to accumulate dramatically at the surface of the *n*- layer. This accumulation screens the electric field from the oxide electrode and simultaneously lowers the voltage-drop across the *n*- layer while increasing the voltage-drop across the oxide. The increased oxide field enhances any electron tunneling present to turn-on the *np*-junction even more, which of course then enhances the hole current providing the accumulation. This feedback mechanism continues until inversion occurs at the surface of the *n*- layer, as shown in Figure 3.4(c).

The resulting low-impedance state is retained as long as there is enough hole current to maintain inversion and keep the majority of the applied voltage-drop across the oxide instead of the semiconductor, or as long as there is enough electron tunneling from the oxide electrode to keep the *np*-junction forward-biased. If these currents become low-enough to be consumed by recombination at the *np*-junction, the device switches back to the high-impedance state.

The short time it takes the TSD to switch between these states is what distinguishes it from all other thyristors. Since the tunneling barrier is so conductive, any accumulated internal charge

rapidly leaks away when the bias is reduced to switch the device off. When the bias is increased to switch the device on, the device is similarly limited only by the turn-on time of the *np*-junction. Researchers have accordingly observed ~1 ns switching times in both directions [13, 14].

The 1 MHz / 0.5 VAC *C-V* characteristic of a TSD is shown in Figure 3.5, along with the *I-V* characteristic to point out the voltage where the device switches on. A high frequency was used to prevent depletion-region generation currents from increasing the measured capacitance. At any given bias there are three possible sources of intrinsic device capacitance, that of the insulator C_i, the *n*- depletion region C_d, and the *np*-junction depletion region C_j. Since these capacitances are in series with each other, the total capacitance is given by

$$C_{total} = \frac{C_j C_i C_d}{C_j C_i + C_i C_d + C_j C_d}. \tag{3.1}$$

For a large negative bias applied to the device, the oxide electrode is positive relative to the *p*+ layer and hence both accumulation of electrons at the *n*- layer surface and depletion-layer widening for the *np*-junction occur, as shown in Figure 3.3(b). Since the *np*-junction depletion layer is quite wide ($C_j \ll C_i$) and since there is no *n*- layer depletion ($C_d = \infty$), Equation 3.1 reduces to $C_{total} \sim C_j$ in this bias range. Note that eventually the accumulated charge becomes so great that it screens the electric field from the oxide electrode and prevents further widening of *np*-junction depletion layer, making the capacitance eventually independent of applied bias.

For a large positive bias applied to the device, the oxide electrode is negative relative to the *p*+ layer and hence depletion of the *n*- layer occurs, as shown in Figure 3.4(a). Since the width of this depletion region becomes very wide ($C_d \ll C_i$ and $C_d \ll C_j$) Equation 3.1 reduces to $C_{total} \sim C_d$ in this bias range. Note that since there is no inversion, the depletion region continues to widen with increasing bias, thus lowering the capacitance until the TSD eventually switches on, at which point the capacitance abruptly increases as the impedance drops.

Lastly, at some slightly negative applied bias (positive oxide electrode potential) the *n*- layer is in accumulation with $C_d = \infty$, and the *np*-junction is not yet reverse-biased so that $C_j \lesssim C_i$. This reduces Equation 3.1 to $C_{total} \sim (C_j C_i)/(C_j + C_i)$ and allows the use of the peak capacitance as a lower bound on C_i, thus providing an upper bound on the oxide thickness d via $C_i = \epsilon A/d$. For the 60 μm diameter device used for Figure 3.5, this indicates an oxide thickness of no more than ~ 20.6 Å.

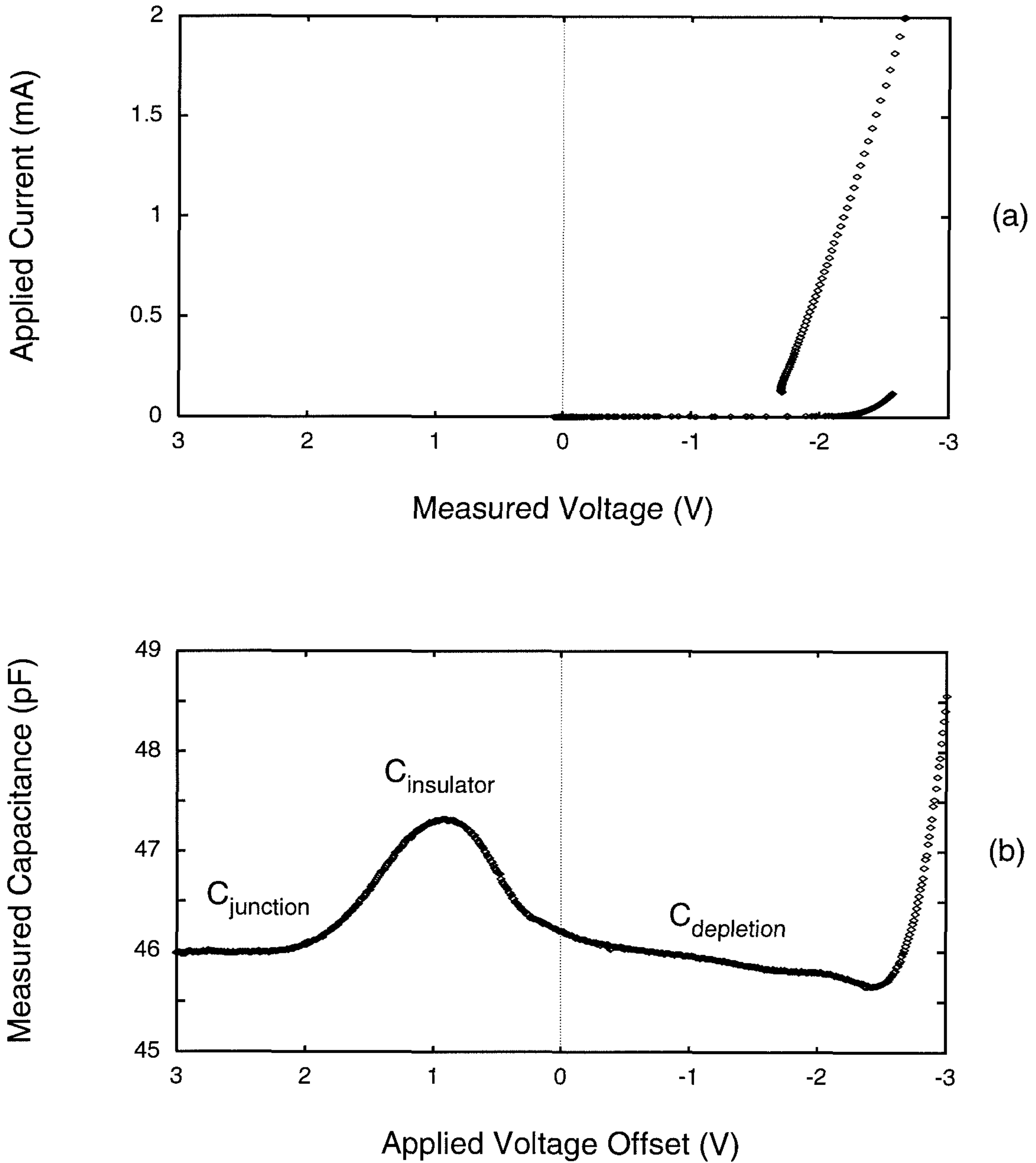


Figure 3.5: Measured *C-V* characteristic for a TSD at 1 MHz / 0.5 VAC.

3-Terminal Properties

All thyristors are easily adapted to three-terminal configurations because of the ease in which minority carriers may be externally injected into the depletion/avalanche layer, thus reducing the amount of forward bias required to switch the device on [5, 13, 15]. For example, Figure 3.6 illustrates what happens when light is shined on the *n*- layer surrounding the TSD used in Figure 3.2. As the light intensity is increased, more electron-hole pairs are generated in the surrounding *n*- layer and diffuse beneath the oxide electrode of the TSD, flowing through the depletion region as shown in Figure 3.4(a). The injected holes of course tunnel through the oxide into the metal valence band, but what is special is that the injected electrons raise the potential of the *n*- layer deeper in the device (i.e. the depletion region separates the injected holes and electrons in opposite directions), extending the depletion region and reducing the required electric field from the oxide electrode to switch the device on. The *I-V* peak-voltage, therefore, will decrease with increasing light intensity as shown in the figure.

The same situation occurs when either an explicit ohmic contact is made to the surrounding *n*-layer or charge is injected from a nearby *coplanar* device [13, 16]. Figure 3.7, for example, illustrates what happens when an adjacent coplanar TSD to the one used in Figure 3.2 is switched on and then further forward-biased. Note that the peak-voltage immediately drops from ~4.4 V to ~1.8 V when the adjacent device is turned on, and then gradually further decreases as the current-density of the adjacent device is increased. This phenomenon has been used to build a charge-coupled shift register out of TSD devices [13].

CIRCUIT PRINCIPLES

The TSD shares many of the same circuit principles with the RTD, and so this section will highlight only some of the differences. Note that the TSD is the *I-V* dual to the RTD by exhibiting a peak in voltage instead of a peak in current. Although the TSD has an asymmetric *I-V* characteristic, a symmetric one may be produced by using two devices wired-up in an antiparallel (*diac*) configuration. It may also be possible to produce a TSD-based diac by integrating two TSD devices back-to-back [17].

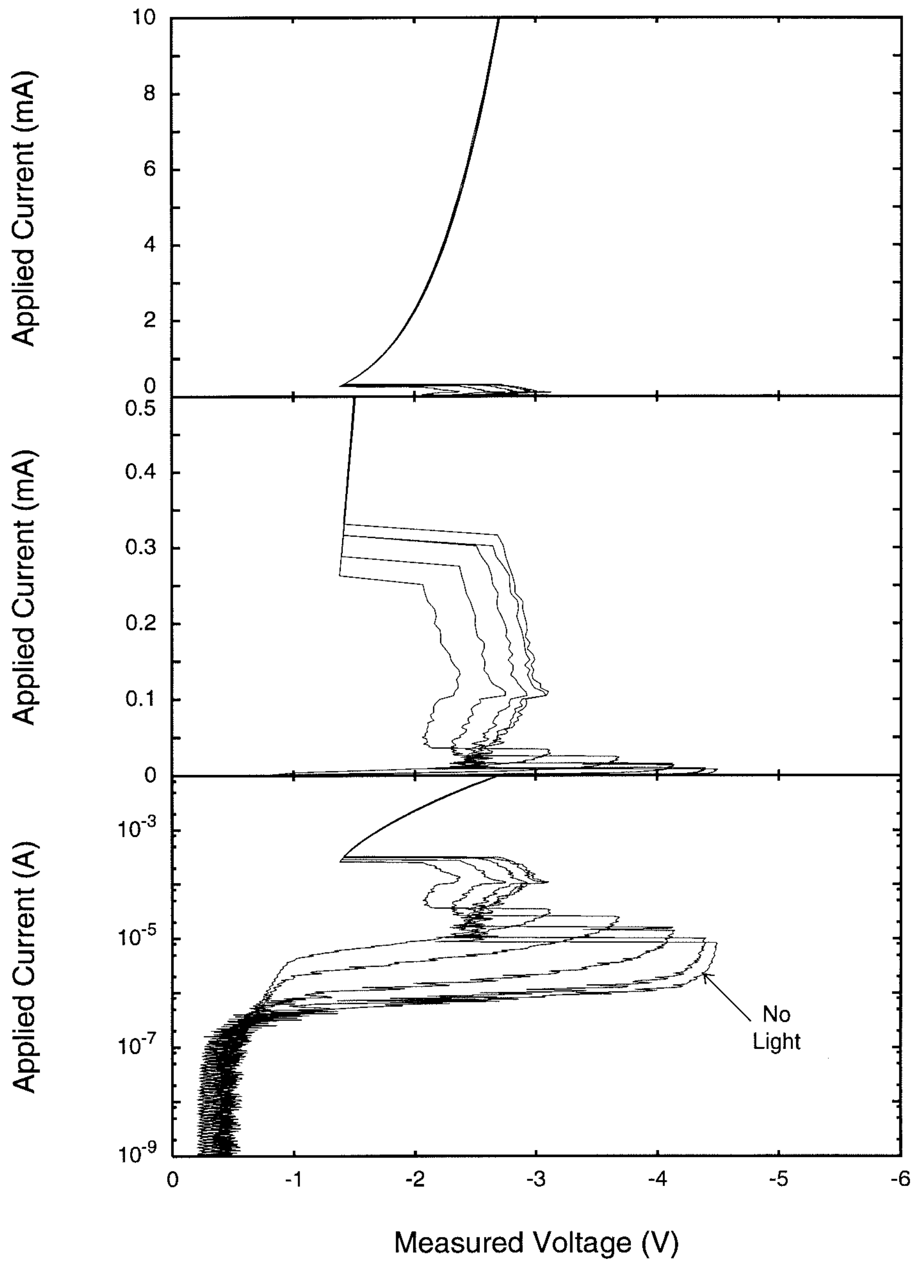


Figure 3.6: Reduction of the TSD peak voltage with exposure to increasing charge-injection from a microscope lamp (shown at three different scales).

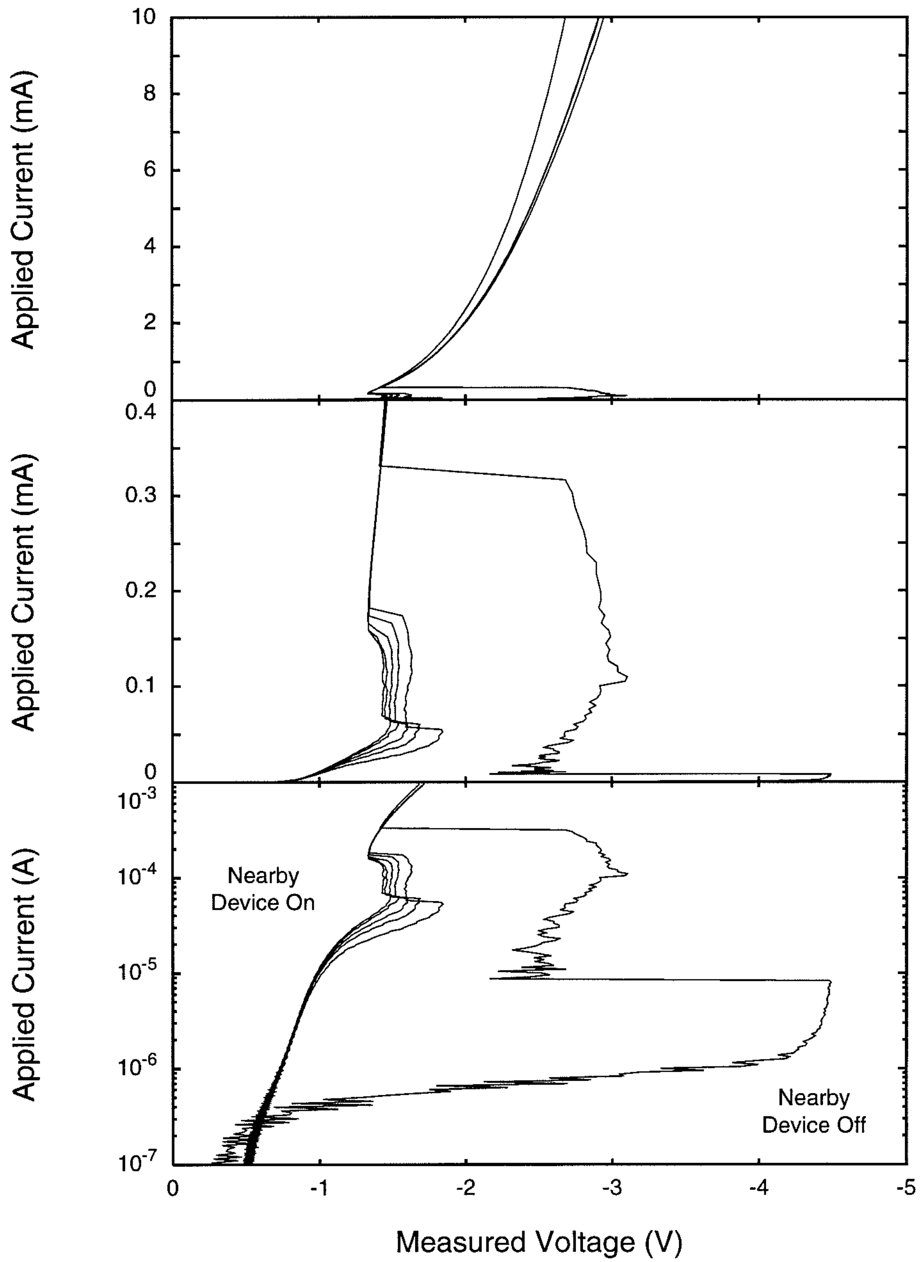


Figure 3.7: Reduction of the TSD peak voltage with exposure to increasing charge-injection from a nearby coplanar device (shown at three different scales).

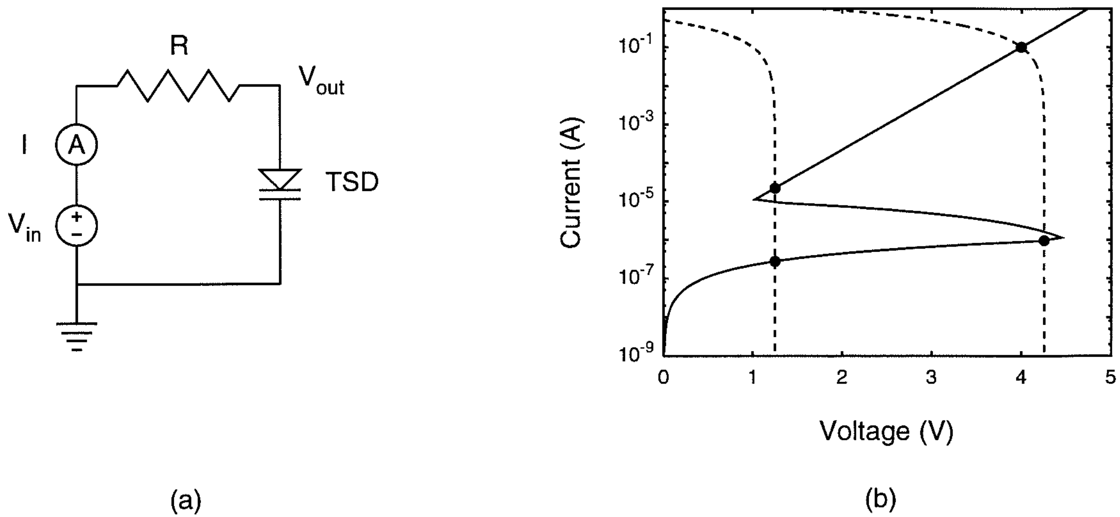


Figure 3.8: Load-line analysis for the TSD.

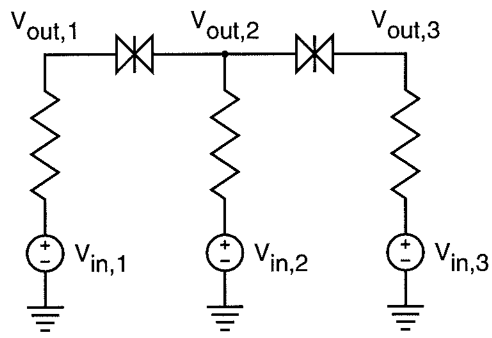
Load-Line Analysis

Figure 3.8(a) shows the schematic diagram for a simple TSD load-line circuit, and Figure 3.8(b) shows the corresponding load-line analysis. The resistive load (dashed line) is shown for two values of V_{in}, and the *I*-*V* fixed points are shown as filled circles. Note that a linear perturbation in V_{in} results in a nominal change in the high-impedance fixed points and an exponential change in the low-impedance ones. This is distinctively different from the RTD, where all of the fixed points are clustered rather close together in *I*-*V* space as shown in Figure 2.10. This feature will be exploited in a later chapter that discusses minimizing the power dissipation of a large array of bistable circuits.

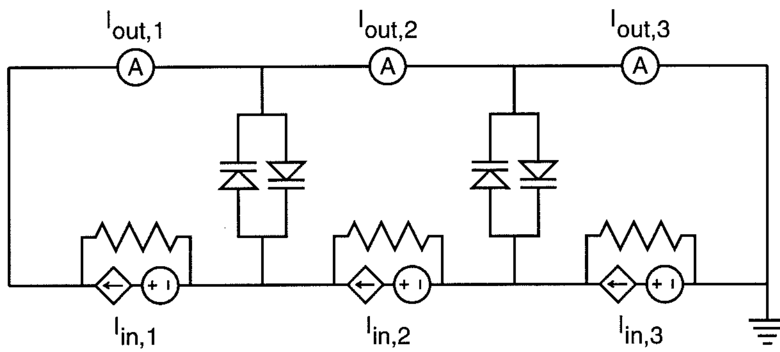
Circuit Duality

Since the TSD and RTD are duals of each other, we can always construct the TSD equivalent of an RTD circuit function[3] and vice versa. We can do this, for example, by replacing all resistance values with their reciprocals, connecting all series elements in parallel (and vice versa), and switching current for voltage (and vice versa) in all input and output signals. In the absence of any wiring or interface efficiency for one implementation over the other, there may be some motivation to prefer the TSD version since it can be implemented in silicon. However, as Figure 3.9 shows, there can be a

[3]We may have difficulty in fabricating the exact duals of RTD's and TSD's, but the approximation will usually yield functionally equivalent circuits.



(a)



(b)

Figure 3.9: An example of circuit duality for RTD and TSD devices.

considerable wiring/interface problem for the TSD version of an RTD voltage-mode circuit. Notice that the input and output signals for the circuit (all currents) are serially connected and floating, a difficult configuration to efficiently implement with conventional VLSI technology. The RTD circuit is the basis for an image-segmentation network discussed in a later chapter.

REFERENCES

[1] T. Yamamoto and M. Morimoto, "Thin-MIS-structure Si negative-resistance diode," *Applied Physics Letters*, vol. 20, pp. 269–270, Apr. 1972.

[2] H. Kroger and H. A. R. Wegener, "Bistable impedance states in MIS structures through controlled inversion," *Applied Physics Letters*, vol. 23, pp. 397–399, Oct. 1973.

[3] H. Kroger and H. A. R. Wegener, "Steady-state characteristics of two terminal inversion-controlled switches," *Solid-State Electronics*, vol. 21, pp. 643–654, 1978.

[4] Y. K. Fang, F. Y. Chen, J. D. Hwang, and B. C. Fang, "Tin oxide gated metal-insulator-semiconductor switch diode for room-temperature high speed gas sensing applications," *Applied Physics Letters*, vol. 62, pp. 490–492, Feb. 1993.

[5] F. E. Gentry, F. W. Gutzwiller, N. Holonyak, Jr., and E. E. V. Zastrow, *Semiconductor Controlled Rectifiers: Principles and Applications of p-n-p-n Devices.* Englewood Cliffs, NJ: Prentice-Hall, 1964.

[6] W. R. Runyan and K. E. Bean, *Semiconductor Integrated Circuit Processing Technology*. Reading, MA: Addison-Wesley, 1990.

[7] B. Garrido, J. Samitier, J. R. Morante, L. Fonseca, and F. Campabadal, "Influence of the silicon wafer cleaning treatment on the Si/SiO_2 interfaces analyzed by infrared spectroscopy," *Applied Surface Science*, vol. 56-58, pp. 861–865, 1992.

[8] R. Singh, "Rapid isothermal processing," *J. Applied Physics*, vol. 63, pp. 59–114, Apr. 1988.

[9] S. Keeney, R. Bez, D. Cantarelli, F. Piccinini, A. Mathewson, L. Ravazzi, and C. Lombardi, "Complete transient simulation of flash EEPROM devices," *IEEE Trans. Electron Devices*, vol. 39, pp. 2750–2757, Dec. 1992.

[10] J. S. né, P. Olivo, and B. Riccò, "Quantum-mechanical modeling of accumulation layers in MOS structure," *IEEE Trans. Electron Devices*, vol. 39, pp. 1732–1739, July 1992.

[11] S. Kar and W. E. Dahlke, "Interface states in MOS structures with 20-40 å thick SiO_2 films on nondegenerate Si," *Solid-State Electronics*, vol. 15, pp. 221–237, 1972.

[12] E. H. Nicollian and J. R. Brews, *MOS (Metal Oxide Semiconductor) Physics and Technology.* New York, NY: Wiley, 1982.

[13] T. Yamamoto, K. Kawamura, and H. Shimizu, "Silicon p-n insulator-metal (p-n-I-M) devices," *Solid-State Electronics*, vol. 19, pp. 701–706, 1976.

[14] H. Kroger and H. A. R. Wegener, "Steady-state characteristics of three terminal inversion-controlled switches," *Solid-State Electronics*, vol. 21, pp. 655–661, 1978.

[15] S. M. Sze, *Physics of Semiconductor Devices.* New York, NY: Wiley, 2nd ed., 1981.

[16] H. Kroger and H. A. R. Wegener, "Controlled-inversion transistors," *Applied Physics Letters*, vol. 27, pp. 303–304, Sep. 1975.

[17] K.-F. Yarn, Y.-H. Wang, and C.-Y. Chang, "GaAs bidirectional bistability switch using double triangular barrier structures," *J. Applied Physics*, vol. 75, pp. 2695–2698, Mar. 1994.

Part II

SYSTEMS

Chapter 4

IMAGE SEGMENTATION

Any type of image-processing system is usually an excellent candidate for improvement at the cellular level because it is comprised of a large two-dimensional array of either pixels (e.g. as in a CCD camera), connections between pixels (e.g. as in a recognition system), or both. Usually the quality and/or performance of these systems increases dramatically with the dimensions and density of the array, so that any slight improvement made at the cellular level can have a tremendous effect on the entire system.

This chapter examines the task of *image segmentation* that is sometimes required in the early stage of machine vision systems. Since biology has already solved this problem, a brief introduction to edge-detection mechanisms in the retina will be given to motivate the circuit built from quantum-effect devices. The circuit is derived from earlier work by others using Si CMOS, and a comparison will show that a single resonant-tunneling diode (RTD) stack may adequately perform the function of as many as 33 transistors, thus offering the potential for extremely high pixel densities with rapid settling speed.

BIOLOGY

Figure 4.1 shows an illustration of the cells and organization found in the retina of a vertebrate organism as gleaned from scanning electron microscopy [1]. The key things to note are that there are photoreceptors (*rods* and *cones*) interconnected by orthogonal signal paths. The longitudinal signal paths are effected by *horizontal* and *amacrine* cells, and the transverse paths are effected by *bipolar*

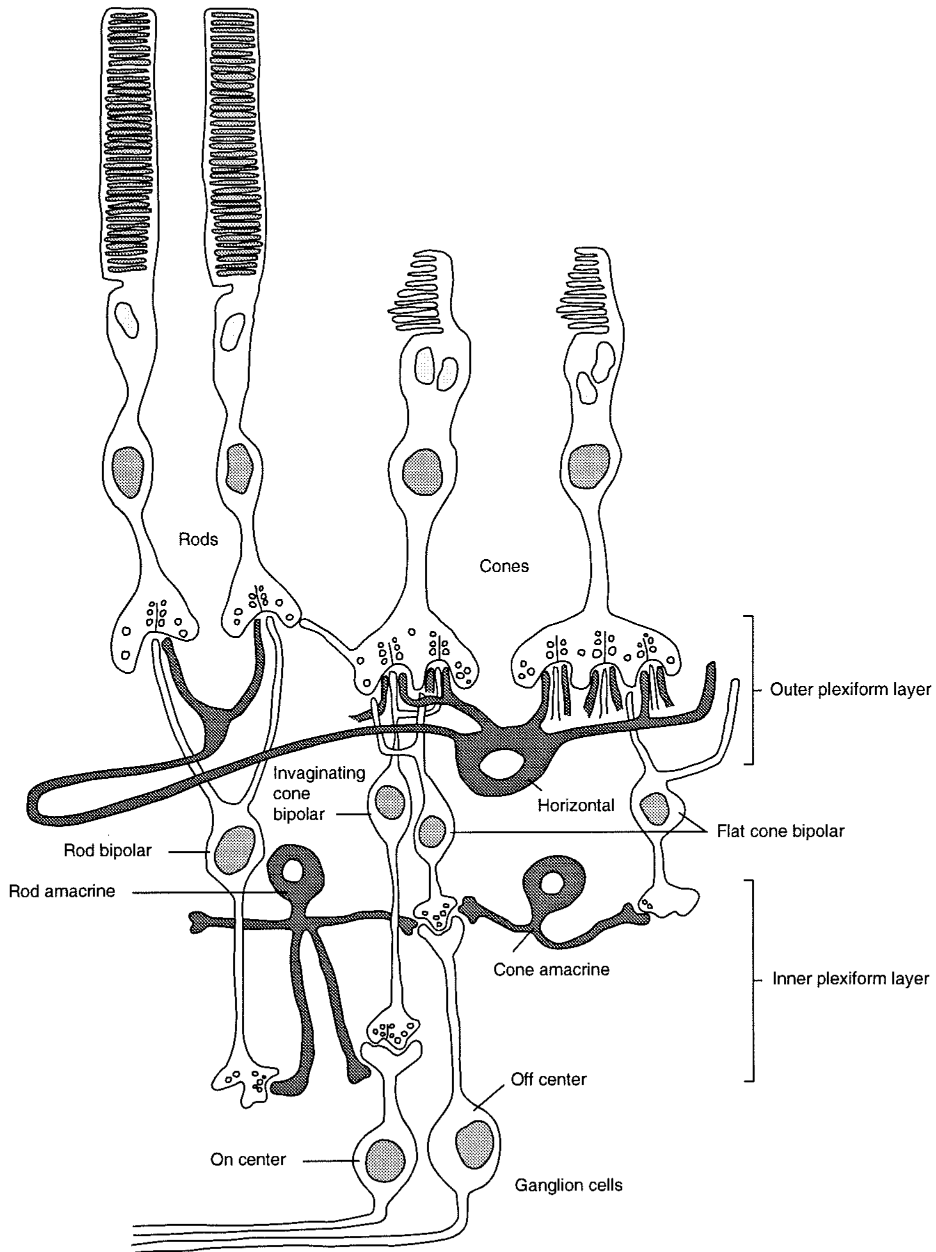
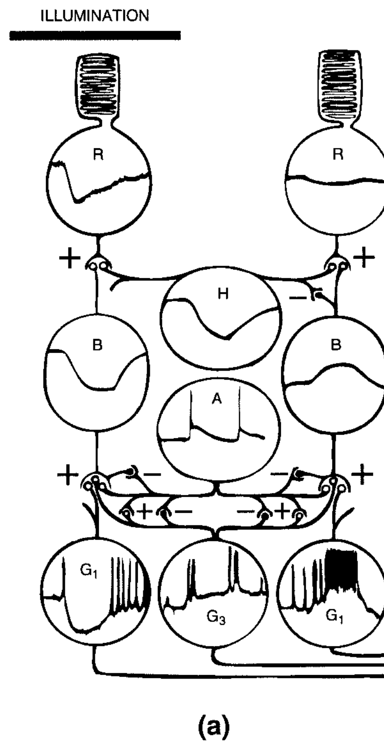
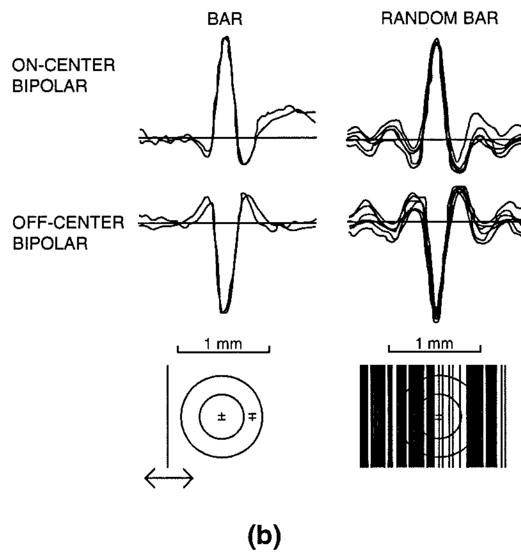


Figure 4.1: Illustration of the cells in the vertebrate retina (adapted from [1]).



(a)



(b)

Figure 4.2: Response functions in the retina (adapted from [2] and [3]).

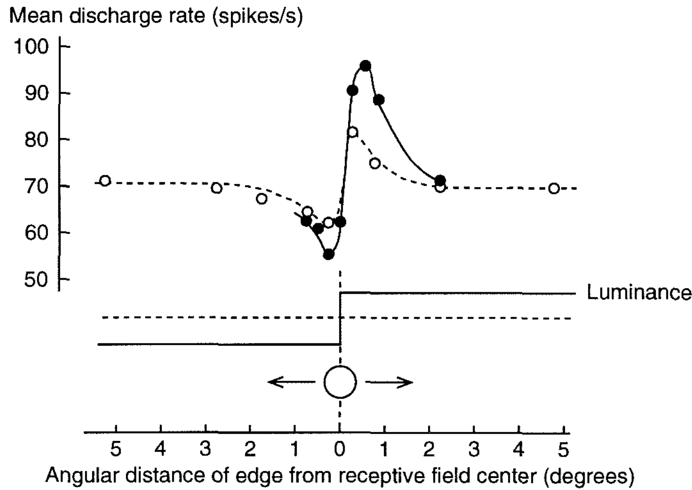


Figure 4.3: Retina response to a luminance step (adapted from [4]).

and *ganglion* cells. The outputs of the ganglion cells form the optic nerve which sends the output of the retina on up to the higher-processing sections of the brain.

Figure 4.2(a) shows the measured response function in time for each of these cell types in the mudpuppy retina when the left photoreceptor is stimulated *temporarily* with illumination [2]. Note that the horizontal cell H conveys what is going on spatially in the neighborhood of the two receptors R, and likewise the amacrine cell A expresses temporal information. Since this chapter is concerned with segmentation of static images, the horizontal cells are the natural target of curiosity.

The horizontal cells form what biologists call *receptive fields*. These are localized groups of photoreceptors and transverse cells that operate on a function of the light activity in their neighborhood, and there are two types of spatial receptive fields in the retina: *on-center* and *off-center*. The type shown in Figure 4.2(a) is an off-center receptive field as shown by the activation of the ganglion cells G_1.

Figure 4.2(b) shows the response of bipolar cells when a luminocity bar is presented (either in a sweep or as random flashes) to these two types of receptive fields in a catfish retina [3]. When the bar is aligned with the center of a receptive field a maximum response is elicited, and as the bar is moved away the response goes inhibitory for a bit and then dies off. The key point is that bipolar cells output the difference between the receptive field center and the *local* neighborhood about the field.

As a last piece of biological data, Figure 4.3 shows the response from cat ganglion cells in an

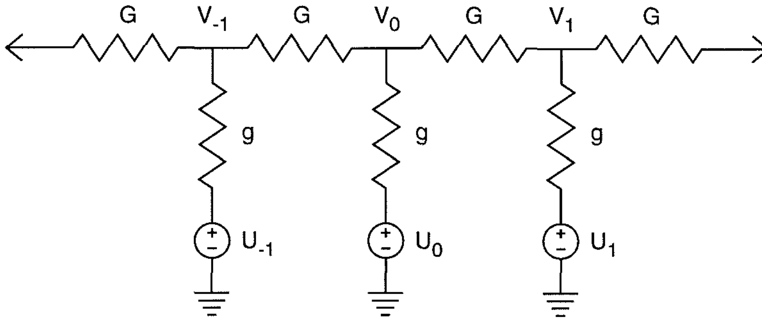


Figure 4.4: A one-dimensional resistive network.

on-center receptive field when exposed to a luminance step centered at various positions in and about the field [4]. Note that there is a peaked response when the *edge* is aligned with the center of the receptive field, and that the response peak is spread out over a some angular distance and hence *many* cells. This is not a problem for biological systems with as many as 10^9 photoreceptors per retina [1], but it can be expensive to implement in VLSI technology where each pixel and/or connection between pixels often consumes a non-trivial amount of wafer real-estate. Nevertheless, the essential aspects of the retina and other biological sensory systems have been emulated with analog VLSI technology by Carver Mead and his group since the 1980's [5]. The next section will quickly review the circuit principles behind the RTD circuit which are based on some of that work.

CIRCUIT PRINCIPLES

Resistive Networks

The functionality of the horizontal cells may be modeled by a *resistive network* [5]. Figure 4.4 shows a small portion of an infinite one-dimensional resistive network. The voltage sources U_i are used to model the photoreceptor output under the presence of some pattern of illumination. In general the signals may be the U_i voltages or currents injected directly into the G network. To understand what this network does computationally, it is helpful to examine how current and voltage change per unit cell of the network.

Start with the simplest case of $U_i = 0$. If we process the unit cell from left to right, then we get the difference equations

$$I_{in} = G(V_{x-1} - V_x) \tag{4.1}$$

$$I_{out} = I_{in} - gV_x \tag{4.2}$$

where *in* and *out* refer to lateral current flow at a specific node x. If increasing Δx is from left to right, then these difference equations may be approximated with the differential equations

$$\frac{dV}{dx} = -\frac{I}{G} \tag{4.3}$$

$$\frac{dI}{dx} = -gV. \tag{4.4}$$

For $\sqrt{G/g} \gtrsim 1$ this approximation is excellent [5]. These equations can help us see how a particular voltage spreads through the network. For example, if we combine Equations 4.3 and 4.4 we obtain

$$\frac{d^2V}{dx^2} - \frac{g}{G}V = 0. \tag{4.5}$$

This is a diffusion equation with the solution

$$V(x) = V_0 e^{-|x|/L} \tag{4.6}$$

$$L = \sqrt{G/g} \tag{4.7}$$

where L is the characteristic length-constant for the exponential decay. This solution means that if we fix a single output node (e.g. V_0), its value will tail off exponentially throughout the network. Consequently, a resistive network forms exponentially-weighted neighborhoods of interaction. If we now compute how much current is required to fix a node like V_0, we will be able to analyze the effect a given photoreceptor output (e.g. U_0) has on the entire network.

Since Equation 4.6 is the voltage solution for the whole network, we can plug it into Equation 4.3 to get an expression for the current required to fix any individual node

$$I_x = \sqrt{Gg}V_x. \tag{4.8}$$

Note that Equation 4.1 and Equation 4.2 defined the unit cell from left to right only, so that this is the current flowing in a semi-infinite network with V_x at the origin. To get the total current for an infinite network we must join two semi-infinite networks together and use

$$I_x = (2\sqrt{Gg} + g)V_x = G_* V_x \tag{4.9}$$

where $G_* = 2\sqrt{Gg} + g$ is the conductance to ground of the entire infinite network at a given node V_x. Since we can also express this conductance as $G_* = g(2L+1)$, then for large L Equation 4.9 is just twice that of Equation 4.8, which we expect when the continuous approximation holds.

At last we are prepared to write down the expression for all V_x given some pattern of I_x or U_x inputs (photoreceptor outputs). Since each injected current contributes a little to each node of the network, and since the network is made entirely from linear elements, we can use the superposition principle to write

$$\begin{aligned} V(x) &= \sum_y V_y e^{-|y-x|/L} \\ &= \sum_y \frac{I_y}{G_*} e^{-|y-x|/L} \\ &= \sum_y \frac{U_y}{G_* G_{in}} e^{-|y-x|/L} \end{aligned} \tag{4.10}$$

where $1/G_{in} = 1/g + 1/(2\sqrt{Gg})$ is the reciprocal of the conductance to ground for each U_i input. This equation explicitly shows that the output at any node is the exponentially-weighted average of the inputs to the network. With this in mind we can now explore how a network like this can be used to perform spatial feature-extraction on an image.

Feature Extraction

Figure 4.5(a) shows a circuit simulation of a one-dimensional resistive network presented with a textured or noisy luminance step similar to that used in Figure 4.3. The U_i values are shown as discrete points, and the V_i values are shown as lines for different values of L ($L = 3$ is solid, $L = 2$ is dashed, and $L = 1$ is gray). Notice that the V_i outputs reflect the exponentially-weighted average output of U_i activity – this provides an effect analogous to biological receptive fields.

The V_i outputs reflect the extracted *features* from the U_i inputs. The small-signal texture and noise information is averaged-out while the large-signal discontinuities at the boundaries are preserved, although in a linear resistive network the boundaries are smeared over a distance comparable to the length-constant of the network, as shown in the figure.

Figure 4.5(b) shows the currents required to maintain the U_i inputs (in practice the U_i inputs are fixed and the V_i outputs are left free to reflect the network computation). Note that these currents are proportional to the difference between the U_i inputs and the V_i outputs, thus providing on-center receptive field computation like that shown in Figure 4.3.

In comparison with this difference computation, Figure 4.5(c) shows the spatial derivative of the V_i outputs. This quantity can sometimes be more useful in edge-detection applications since the response to edges is smoother, unipolar, and single-peaked.

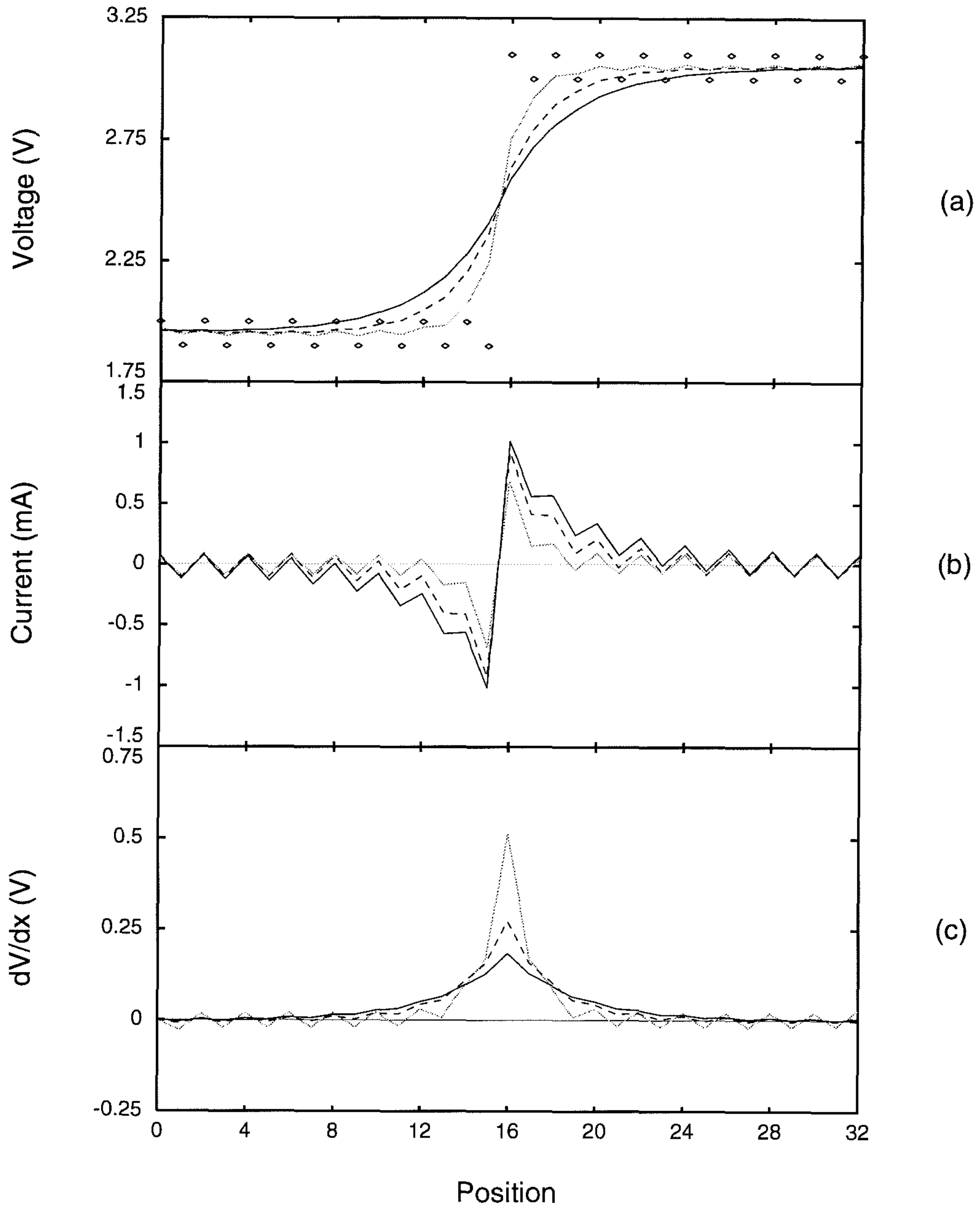


Figure 4.5: Simulation of a 1D resistive network.

In each of the computations presented in Figure 4.5, the response is spread-out over many unit cells, taking poor advantage of the limited number of cells in expensive artificial hardware. If devices with nonlinear *I-V* characteristics are used instead of the G resistors in the network, the circuit may sharpen its response to discontinuities by greatly reducing the lateral conductance for larger signals between adjacent V_i, thus effectively *segmenting* the network into smaller smoothing units for each region of uniform texture or noise. This means the region boundaries may be preserved with much less smearing than what occurs with linear resistive networks (i.e. there is less mixing between regions).

Nonlinear Networks

Figure 4.6 shows several devices with nonlinear *I-V* characteristics for potential use as the lateral (G) conductors in image segmentation networks. Note that the length-constant is derived from the chord conductance I/V and not the differential conductance dI/dV because it is the *amount* not the derivative of current that flows between adjacent pixels which participates in the local averaging. This is a result of using a device with a first and third quadrant *I-V* characteristic; even though dI/dV may be negative a positive current still flows for that value of applied bias.

To compare the way the effective length-constants of these devices vary with applied bias, we can plot *relative* length-constants assuming the same linear g conductance for each network, given by

$$L_{rel} = L\sqrt{g} = \sqrt{\frac{I}{V}}. \tag{4.11}$$

Figure 4.6(a) shows a simulated *I-V* characteristic for a saturating resistor (solid line) like that of the HRES circuit presented in [5]. The dashed line shows the *I-V* function of a linear resistor with roughly the same small-signal conductance as the saturating resistor. Figure 4.6(b) shows how the resulting L_{rel} depends on the voltage between adjacent V_i outputs. Note that while there is always some averaging occurring (the current and hence L_{rel} never drops to zero), the length-constant drops reasonably fast below that of the linear resistor as the size of the voltage (the extracted discontinuity) grows. Figure 4.6(c) through Figure 4.6(h) show devices that result in L_{rel} dropping faster than that for the saturating resistor.

Figure 4.6(c) shows a simulated *I-V* characteristic for the resistive fuse circuit developed by John Harris et al. [6]. The ideal resistive fuse can perform perfect segmentation since no current flows

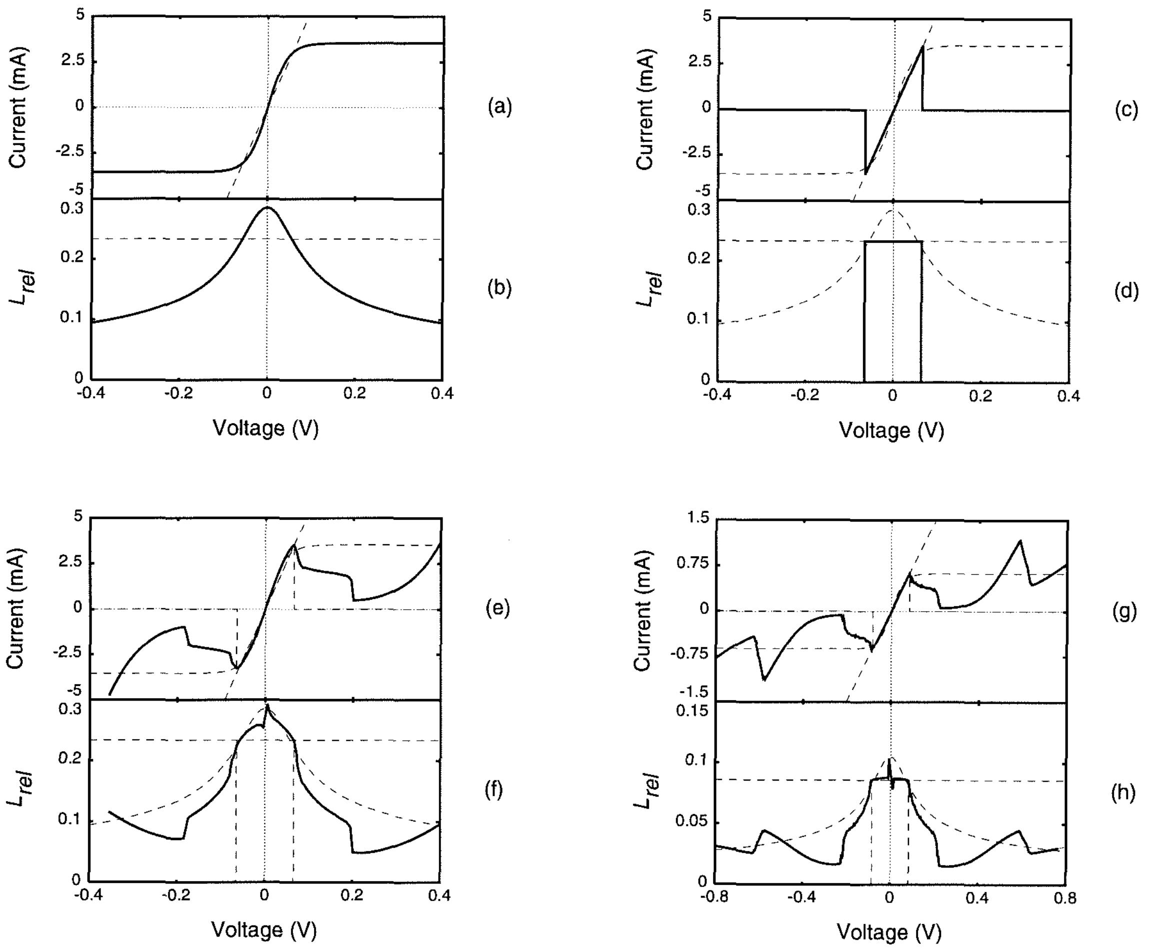


Figure 4.6: Nonlinear network devices and their relative length-constants.

between regions whose average value differ by more than the peak voltage, thus preserving the discontinuity without *any* smearing. This is reflected in the L_{rel} plot in Figure 4.6(d) plot which also shows the values for comparable linear and saturating resistors.

Figure 4.6(e) through Figure 4.6(h) show the measured *I-V* characteristic and the resulting L_{rel} for single and double RTD elements. The single RTD is from MBE sample III-419, and the double RTD (two singles wired in series) is from MBE sample III-286, both grown by Doug Collins in the McGill-group laboratory. Note that the segmentation effectiveness is in between the saturating resistor and the resistive fuse, and that the fine *I-V* structure of the RTD does not adversely affect this for voltages less than where the tail current is approximately the peak current. If a wider range of voltages needs to be handled, the *I-V* tail can easily be pushed to higher voltage by using more RTD's in series, as shown in Figure 4.6(h). In this case it is important to keep the series resistance low so that the initial peak voltage does not move up too much with each additional RTD.

For the nonlinear network elements with peaked *I-V* characteristics, the dynamic range of the input signals must be scaled appropriately to ensure that the typical image discontinuities are greater than the peak voltage. This may be accomplished either by using differential transconductance amplifiers for the g conductances [6] or by using specialized receptor circuits that adjust a narrow intensity response within a wider range through adaptation [7, 8]. In either case, the required level-shifting and gain-control may be easily performed at a nominal cost of wafer real-estate. It has been shown that for a given image discontinuity in one-dimension, V_{step}, the peak voltage should be set below $V_{step}(\sqrt{Gg} + g)/(2G + \sqrt{Gg} + g)$ to extract it [9].

Another issue concerning these type of network elements is that they can impart hysteresis to the circuits they are in. For CMOS circuits with adjustable parameters (e.g. peak voltage, NDR, etc.), a continuation method may be used whereby the *I-V* characteristic is continuously adjusted from that of a saturating resistor to that of a resistive fuse [6]. This ensures that all devices start-out in their smoothing (i.e. resistive) state for each sampling of the network inputs. An equivalent method that is available to passive devices like the RTD is just to temporarily switch all network inputs to the same value.

1D DEMONSTRATION

This section will present an explicit demonstration of how well RTD devices segment one-dimen-

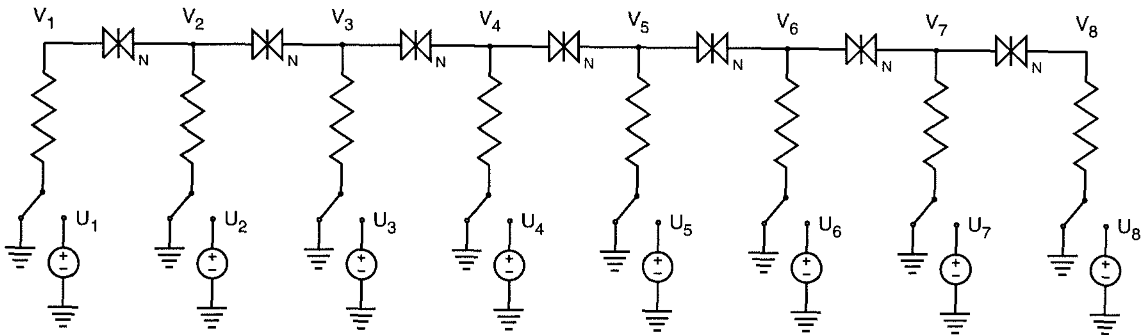


Figure 4.7: Schematic diagram for the one-dimensional RTD segmentation network.

sional line data. Figure 4.7 shows the schematic diagram for the one-dimensional RTD segmentation test-circuit [10]. The circuit was tested with one RTD per lateral connection and additionally simulated with a piecewise-linear interpolation of the measured *I-V* characteristic for the double-RTD element (sampled at 16mV intervals). The general use of an arbitrary number of RTD's in series per lateral connection is denoted by the N subscript near the RTD symbol.

The *I-V* characteristics for the single and double RTD elements are shown in Figure 4.6(e) and Figure 4.6(g), respectively. Since each RTD can add a degree of hysteresis to the circuit, some mechanism for resetting the devices is necessary, and this was accomplished by using CMOS analog switches to select either ground or a U_i input voltage.[1] The network length-constant can be programmed by putting resistors in series with the CMOS switches, but the resistance of the switches alone ($\sim$137Ω) was high enough to keep $L\sim2.7$ for the single-RTD network. For the double-RTD network the length-constant was set to $\sim$1.6 to maximize the output performance.

In order to make a comparison between the RTD networks and those using saturating resistors and resistive fuses, the same set of U_i inputs used to test these other networks in [6] was used. These inputs are shown as points in Figure 4.8. Figure 4.8(a) shows the segmentation performance for the single-RTD network (solid line) in comparison to the HRES-type saturating resistor network tested in [6] (dashed line). Figure 4.8(b) similarly shows the performance for the double-RTD network, and Figure 4.8(c) shows the performance for the resistive fuse network tested in [6].

The key thing to note is that while the single-RTD element allows considerable amount of mixing across the discontinuity, the double-RTD element yields performance almost equivalent to the

[1] In a two-dimensional real-time video implementation of this circuit, the resetting would naturally occur between each frame (using the vertical sync signal for example).

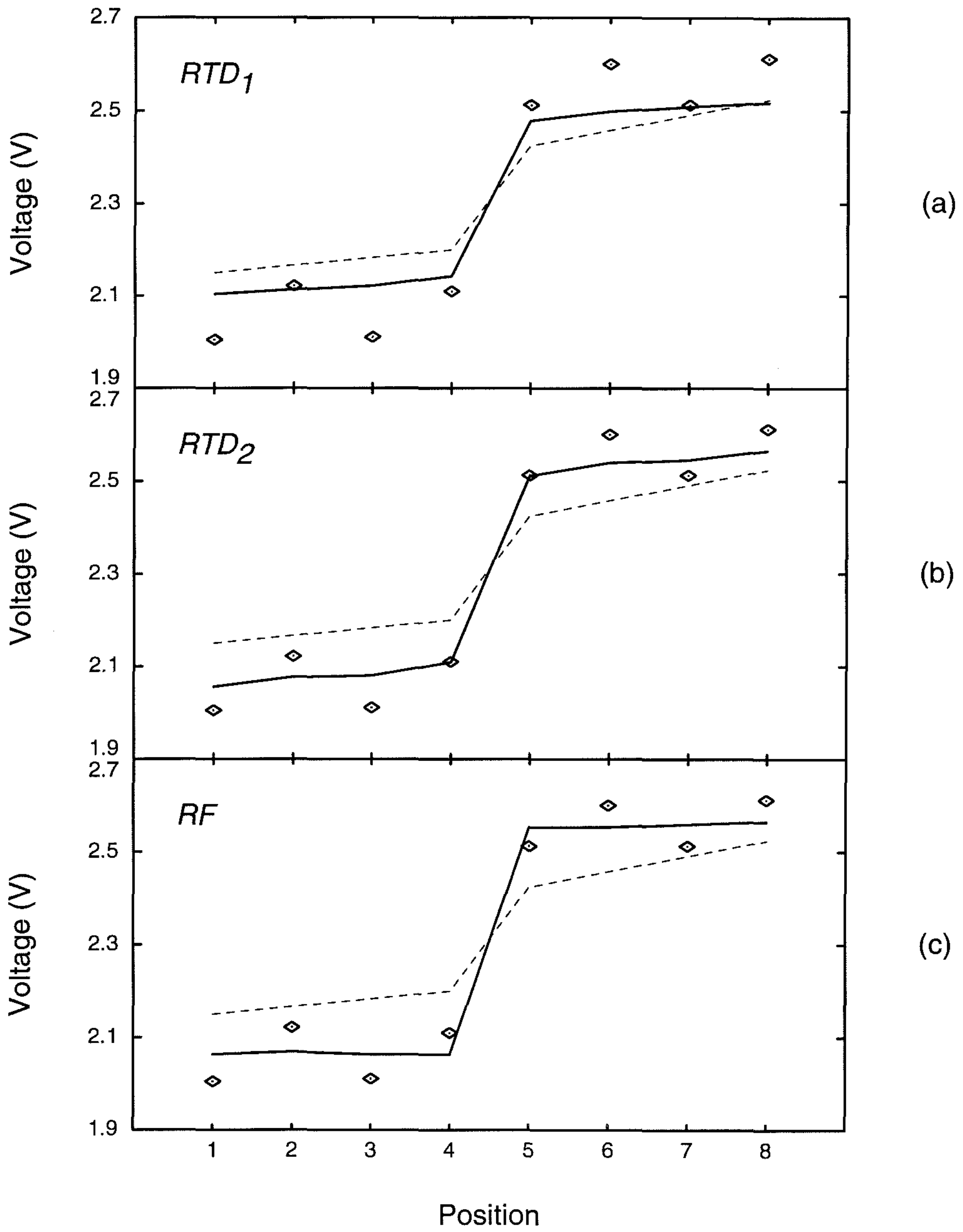


Figure 4.8: One-dimensional segmentation comparison for various nonlinear devices.

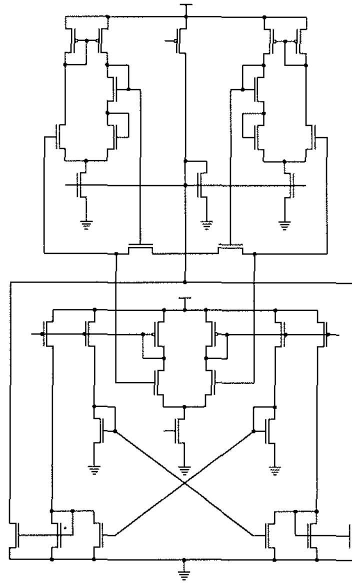


Figure 4.9: Schematic diagram for the resistive fuse circuit used in [6].

resistive fuse. Also note that the saturating resistor performs much better than the linear network of Figure 4.5, but that it does not do as well as either RTD network or the resistive fuse network. This is just because the saturating resistor is the worst-case fuse (its off-current is equal to its peak current) and the RTD elements are better fuses before they reach far up the tail-part of their *I-V* characteristics.

TECHNOLOGY COMPARISON

The use of RTD devices was shown in the previous section to provide the equivalent functionality of CMOS subcircuits with many transistors. For example, the schematic diagram of the resistive fuse circuit used to generate the data in Figure 4.8(c) from [6] has 33 transistors, as shown in Figure 4.9. While recent progress in the development of these types of CMOS circuits has brought the transistor count down [11], an 11-transistor resistive fuse still requires 7500 μm^2 and has a useable common-mode range of only 1.5 V. A 4-transistor resistive fuse has been fabricated as well [11], but its peak voltage was ~2 volts; this would likely necessitate scaling the network inputs to 10 volts or more to put salient image features into the fuse region of the *I-V* characteristic. Table 4.1 shows a comparison between RTD and CMOS segmentation elements for various attributes. An interesting sidelight is that RTD devices can be fabricated on the same wafer as long-wavelength infrared de-

Table 4.1: Brief comparison between RTD and CMOS segmentation elements

Attribute	RTD	Saturating Resistor & Resistive Fuse
Area	minimal ($\lesssim 1\mu m^2$)	large ($\gtrsim 10^4 \mu m^2$)
Speed	extremely fast (GHz)	fast (MHz)
Common-mode range	infinite	narrow
Parameters	static	dynamic
Fabrication	unconventional	conventional
Stability	bistable	saturating resistor is monostable, resistive fuse is bistable

tectors using III-V materials (e.g. GaAs/AlAs, GaSb/AlSb/InAs, etc.), offering the potential for an integrated infrared imager and feature-extractor.

Also, some mention of the dual to the RTD segmentation network should be mentioned. This would make use of a peaked *voltage-current* (*V-I*) characteristic that is analogous to the RTD *I-V* characteristic. Unfortunately, the input and output currents are series-connected and hence floating; it is difficult to implement this in a design that readily interfaces with the off-chip world. In addition, devices such as thyristors and their bipolar-transistor equivalents either switch too slowly or do not have the required symmetric properties.

2D SIMULATIONS

The natural extension to the demonstration circuit of Figure 4.7 is the two-dimensional array of interconnected pixels as shown in Figure 4.10. Simulations on various test images were conducted by scaling 0–255 gray scale values into 0–1 volt for use with the *I-V* characteristic in Figure 4.6(g). These values then comprised the U_{ij} input array, and the circuit was simulated from an initial $V_{ij} = 0.5$ volt condition (the 0.5 volt value was chosen to shorten the average simulation time).

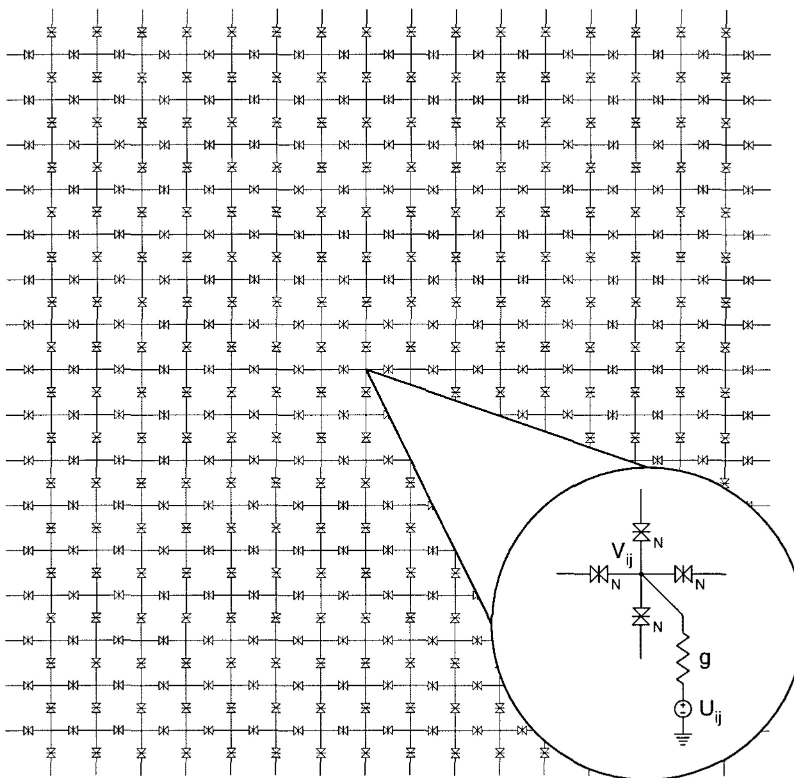


Figure 4.10: Schematic diagram for the two-dimensional RTD segmentation network.

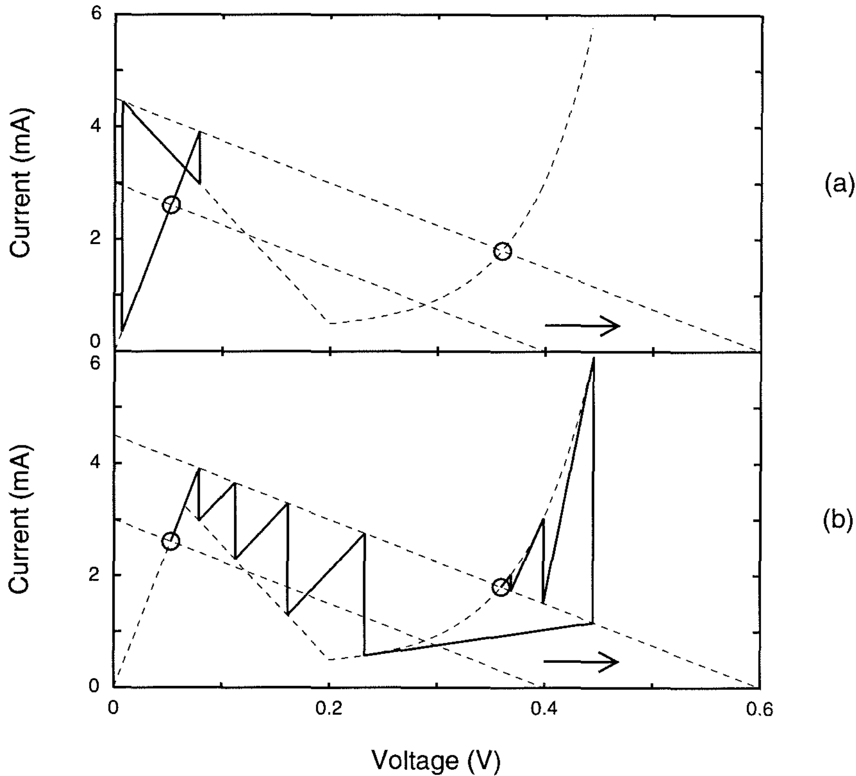


Figure 4.11: Simulation convergence with the modified Newton-Raphson technique.

The simulation technique is described as follows.

First a piecewise-functional description of the measured *I-V* characteristic was determined by fitting various parametric line-segments and exponentials to it. For the general asymmetric double-RTD element it takes typically fourteen of these functions to emulate all of the fine structure in the *I-V* characteristic. Next the derivatives dI/dV were established for all of these functions and assembled into a stand-alone device simulation routine.

The network solution was then found by using a modified form of the Newton-Raphson technique [12] for solving Kirchoff's current law. In the unmodified form, the technique starts with an initial V_{ij} condition and computes the net current ΔI_{ij} flowing into each node as well as the net conductance $G_{ij} = dI_{ij}/dV_{ij}$ for each node. A temporary output array V'_{ij} is then set according to $V'_{ij} = V_{ij} + \Delta I_{ij}/G_{ij}$. After all of the V'_{ij} have been established, the array is updated with $V_{ij} = V'_{ij}$. The procedure iterates until the maximum $\Delta I_{ij}/G_{ij}$ for the whole array is less than some pre-specified voltage tolerance.

There is a problem with this technique when discontinuous switching is present, which is the

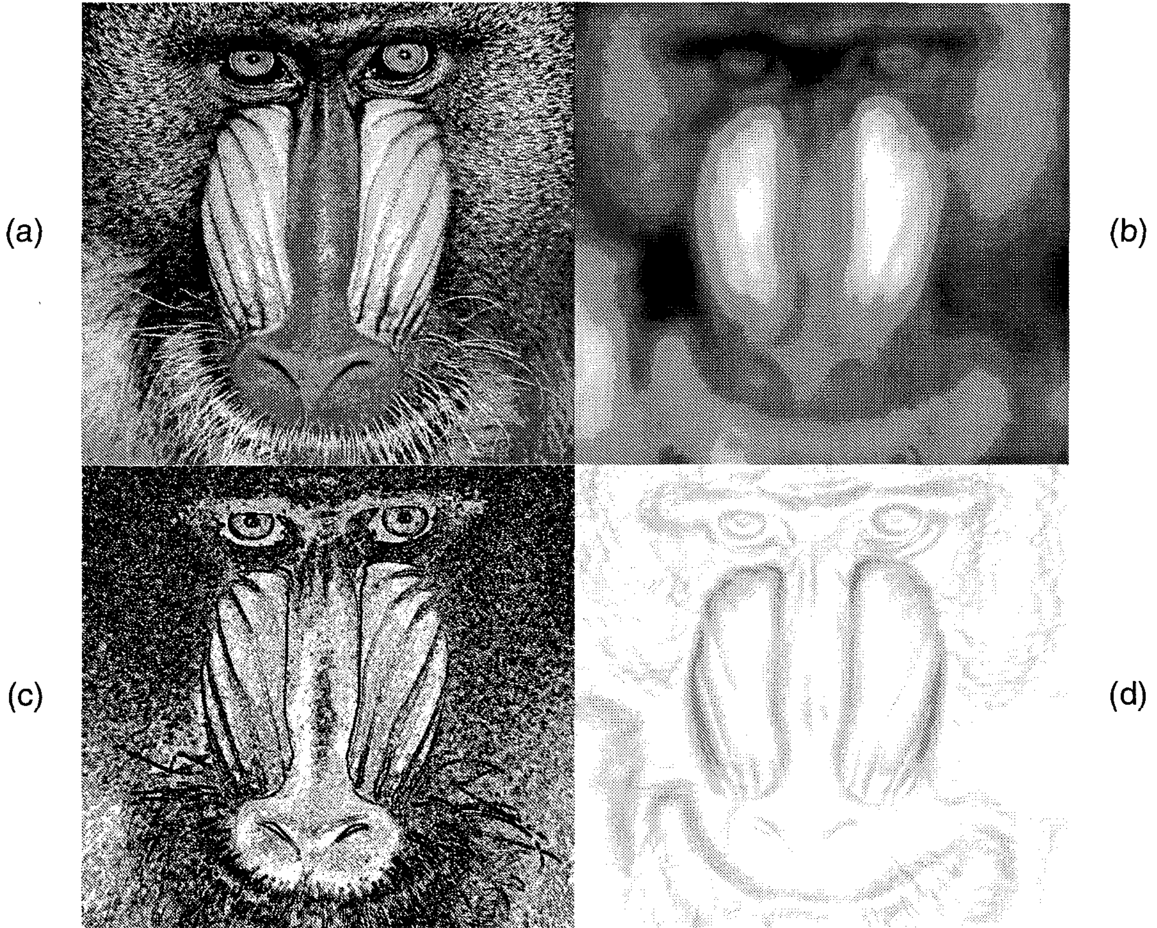


Figure 4.12: Two-dimensional RTD segmentation simulation on the mandril image.

case when the U_{ij} are initially applied. An illustration of the problem is shown in Figure 4.11(a). This figure shows what happens in the simple load-line circuit when the input voltage is discontinuously increased. Note that because the output voltage is always changed by an amount proportional to the conductance at the floating node (in this case one of the RTD terminals), the algorithm is launched into an endless loop, as shown in the figure, and never finds the solution state. The solution to the problem is to use the absolute value quantity $|\Delta I_{ij}/G_{ij}|$ as the voltage increment instead, as shown in Figure 4.11(b).

The circuit simulation has been run on a variety of standard test images. Figure 4.12(a) shows the classic mandril test image digitized into a 512 x 480 array of pixels, and Figure 4.12(b) shows the simulated output of an RTD segmentation network with $L = 10.0$ and voltage tolerance set to 1 μV. Notice the regions of averaged texture separated by sharp boundaries. Figure 4.12(c) and Figure

Figure 4.13: Two-dimensional RTD segmentation simulation on the lena image.

4.12(d) are the spatial derivatives of Figure 4.12(a) and Figure 4.12(b), respectively, as computed by the Sobel operator

$$\begin{aligned}
\Delta V_x &\equiv V_{x+1,y-1} + 2V_{x+1,y} + V_{x+1,y+1} - V_{x-1,y-1} - 2V_{x-1,y} - V_{x-1,y+1} \\
\Delta V_y &\equiv V_{x-1,y+1} + 2V_{x,y+1} + V_{x+1,y+1} - V_{x-1,y-1} - 2V_{x,y-1} - V_{x+1,y-1} \\
V_{x,y} &= \sqrt{\Delta V_x{}^2 + \Delta V_y{}^2}.
\end{aligned}$$

Since texture and noise have virtually been removed in Figure 4.12(b), its spatial derivative is much cleaner and portrays a more salient representation of the object in the original image. Figure 4.13(a) through Figure 4.13(d) show the analogous effect for the lena test image with the same network length constant.

REFERENCES

[1] E. R. Kandel and J. H. Schwartz, *Principles of Neural Science*. New York, NY: Elsevier, 2nd ed., 1985.

[2] J. E. Dowling, *The Retina: An Approachable Part of the Brain*. Cambridge, MA: Belknap Press of Harvard University Press, 1987.

[3] A. Gallego and P. Gouras, Eds., *Neurocircuitry of the Retina*. New York, NY: Elsevier, 1985.

[4] P. Buser and M. Imbert (translated by R. H. Kay), *Vision*. Cambridge, MA: MIT Press, 1992.

[5] C. A. Mead, *Analog VLSI and Neural Systems*. Reading, MA: Addison-Wesley, 1989.

[6] J. Harris, C. Koch, J. Luo, and J. Wyatt, "Resistive fuses: analog hardware for detecting discontinuities in early vision," in *Analog VLSI Implementation of Neural Systems* (C. Mead and M. Ismail, Eds.), Norwell, MA: Kluwer, 1989.

[7] T. Delbrück, *Investigations of Analog VLSI Visual Transduction and Motion Processing*. Ph.D. thesis, Caltech, Pasadena, CA, Nov. 1992.

[8] J. A. Anderson, "General introduction," in *Neurocomputing: Foundations of Research* (J. A. Anderson and E. Rosenfeld, Eds.), Cambridge, MA: MIT Press, 1988.

[9] J. Harris, *Analog Models for Early Vision*. Ph.D. thesis, Caltech, May 1991.

[10] H. J. Levy, D. A. Collins, and T. C. McGill, "Extracting discontinuities in early vision with networks of resonant tunneling diodes," in *Proceedings of the 1992 IEEE International Symposium on Circuits and Systems*, (San Diego, California), pp. 2041–2044, May 10-13 1992.

[11] P. C. Yu, S. J. Decker, H.-S. Lee, C. G. Sodini, and J. L. Wyatt, Jr., "CMOS resistive fuses for image smoothing and segmentation," *IEEE J. Solid-State Circuits*, vol. 27, pp. 545–553, Apr. 1992.

[12] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*. New York, NY: Cambridge University Press, 2nd ed., 1992.

Chapter 5

NEURAL NETWORKS

There is the idea that if the nervous system of a biological organism is emulated at a low-enough level by an artificial system, then the two systems will have equivalent intelligence [1]. Nobody knows yet how low "low-enough" needs to be, but the study of neural networks typically assumes that only a functional emulation of processing units (*neurons*) and their interconnections (*synapses*) is sufficient to capture the essence of what the nervous system can do. The hope is to build artificial computing systems with biological capabilities such as pattern recognition, generalization, adaptation, and autonomy.

Neural networks usually have many more connections between neurons than neurons themselves. In the cerebellum, for example, there may be as many as 10^5 synapses per neuron [2]. Researchers have shown that the computational capabilities of these networks are directly related to the number and variability of the synapses, and that the architecture of the connectivity determines the suitability of a network for a given problem or task [3, 4].

The synaptic circuit analogue is thus a prime candidate for an application of quantum-effect devices that dramatically impacts the capabilities of an entire cellular system. This chapter will present how RTD devices may be used to provide a compact and fast multistate conductance that implements a rudimentary but useful form of synaptic function in VLSI technology. The first section will quickly review some neural network principles that motivate the RTD-based subcircuit.

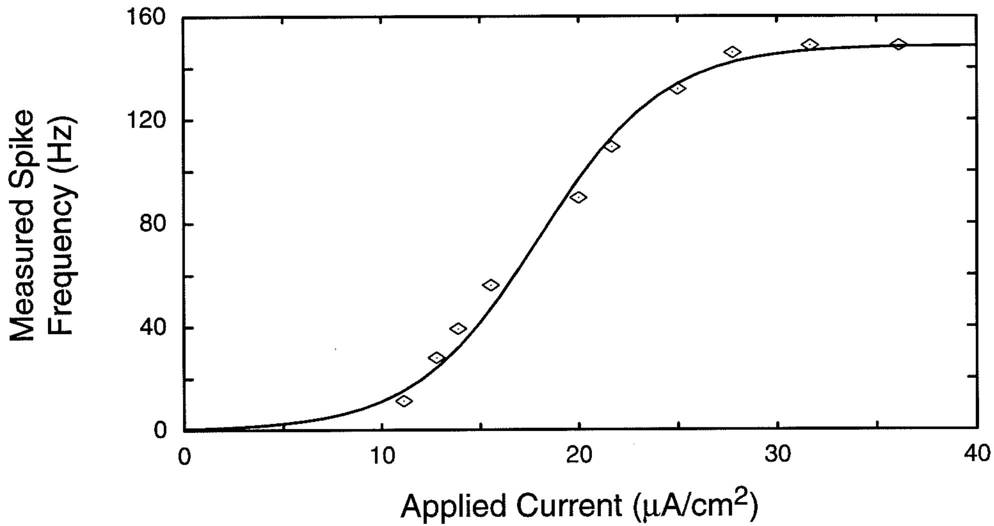


Figure 5.1: Activation function for a neuron (adapted from [5]).

SYSTEM PRINCIPLES

Neurons & Synapses

Figure 5.1 shows the measured input-output *activation* function for a crab neuron adapted from [5]. Most neurons output a voltage spike-train at a frequency that follows this sigmoidal transfer function. The input signal is normally the sum of all the currents received at the synapses, and these synaptic currents are injected by spike-trains from other neurons. Notice that the activation serves to provide both thresholding and saturation; these nonlinear characteristics allow neural networks to perform tasks that linear programming methods cannot [4].

A common artificial abstraction of the real neuron activation function is a transimpedance relationship of a form like

$$V_{out} = V_0 \tanh(\beta I_{in}) \tag{5.1}$$

where V_{out} is a scalar voltage representation of the neuron output spike-train frequency, I_{in} is the net input current summed over all positive (excitatory) and negative (inhibitory) synaptic inputs, β is the gain (i.e. slope of the linear part of the activation function), and V_0 is a scaling factor.

The synapses are variable conductances through which neurons receive their inputs. Both neurons and synapses have complicated time-dependent properties that are only just beginning to be understood, but it is clear that stored information is represented in the nervous system by the synap-

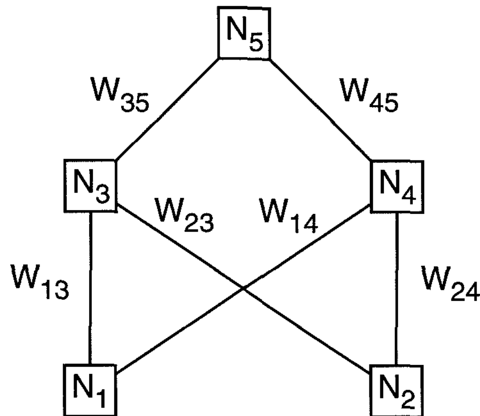


Figure 5.2: The arbor diagram for a multilayer neural network.

tic conductances [3].

A common abstraction of synaptic function is a linear conductor, so that the net input current for a given neuron j is given by

$$I_{in,j} = \sum_i G_{ij} V_{out,i} \tag{5.2}$$

where G_{ij} is the synaptic conductance (either positive or negative) from neuron i to neuron j, and $V_{out,i}$ is the output of neuron i.

Multilayer Networks

Neural networks come in many different architectures and sizes, but a common theme in neural networks designed for use with artificial learning algorithms is multilayer organization. Figure 5.2 shows the connectivity schematic or *arbor diagram* for a simple three-layer *feedforward* neural network. The neurons are labeled N_{1-5} and the synaptic conductances are represented as mathematical input weights W_{ij}. Neurons N_1 and N_2 form the input layer, N_3 and N_4 form an intermediate or hidden layer, and N_5 forms the output layer. Note that the computation each stage (i.e. pair of adjacent layers) performs may be expressed in vector-matrix notation; the first stage, for example, may be expressed as

$$\begin{pmatrix} N_{in,3} \\ N_{in,4} \end{pmatrix} = \begin{pmatrix} W_{13} & W_{23} \\ W_{14} & W_{24} \end{pmatrix} \begin{pmatrix} N_{out,1} \\ N_{out,2} \end{pmatrix}. \tag{5.3}$$

The inputs to the next stage are just computed from $N_{out,i} = F(N_{in,i})$, where F is the sigmoidal neuron activation function.

In general the network may have feedback between the layers, and a learning algorithm may be used to set the W_{ij} values to perform a given task [4]. Usually the learning algorithms require some level of regularity in the flow of information, and hence the frequent use of multilayer networks as opposed to a more nonuniform (but realistic) connection schemes. The next section considers some circuit principles behind multilayer network implementations.

CIRCUIT PRINCIPLES

Vector-Matrix Multipliers

From Equation 5.3 we can see that a vector-matrix multiplier can be a useful computational engine in a multilayer neural network, but the need for both positive and negative weighting elements makes building four-quadrant multipliers out of positive-weighting components like resistors and transistors non-trivial.

Figure 5.3 shows a scheme for implementing the four-quadrant vector-matrix multiplication with only positive-weighting elements. This multiplication may be expressed as

$$V_j = \sum_i W_{ij} U_i = \sum_i (P_{ij} - Z) U_i \tag{5.4}$$

where the possibly negative weight W_{ij} has been split into a differential pair of positive weights P_{ij} and Z, shown as small boxes in the figure.

There are some important issues concerning the memory involved in maintaining the states of a set of W_{ij} weighting elements. Systems have been built where the weight control value is set by the voltage from a capacitor [6, 7] and/or the voltage from a digital-to-analog converter fed by an external RAM chip [6–8]. These types of systems are large and cumbersome, and are relatively slow because of capacitive time delays and/or non-local memory access. At the other extreme there have been attempts at taking explicit advantage of memory effects in materials [9, 10], but these have had a little impact because of their nonlinearity and limited variability. The next subsection presents RTD-based weighting elements as an intermediate solution between these approaches.

RTD Synapses

The primary function of a synapse circuit is to provide variability. Integrated stacks of RTD devices can provide a discrete set of stable I-V operating points by using them in simple load-line cir-

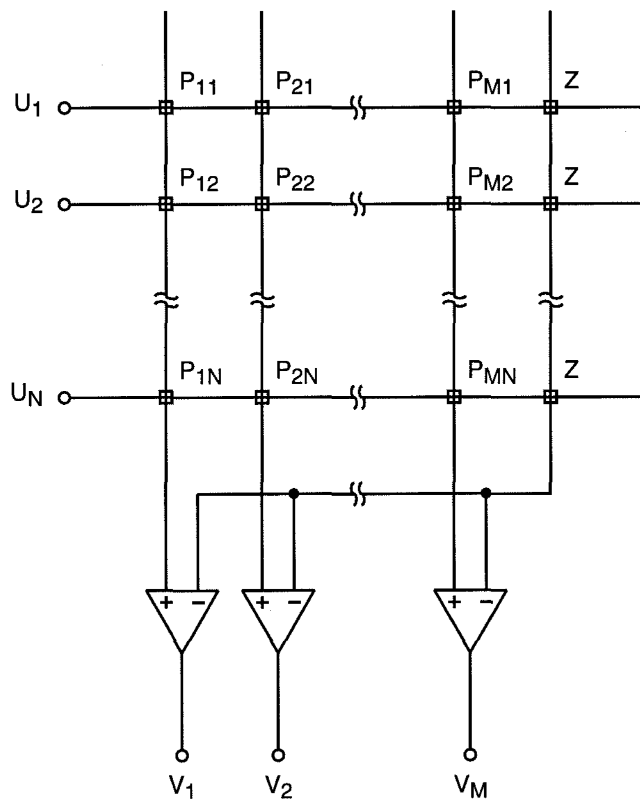
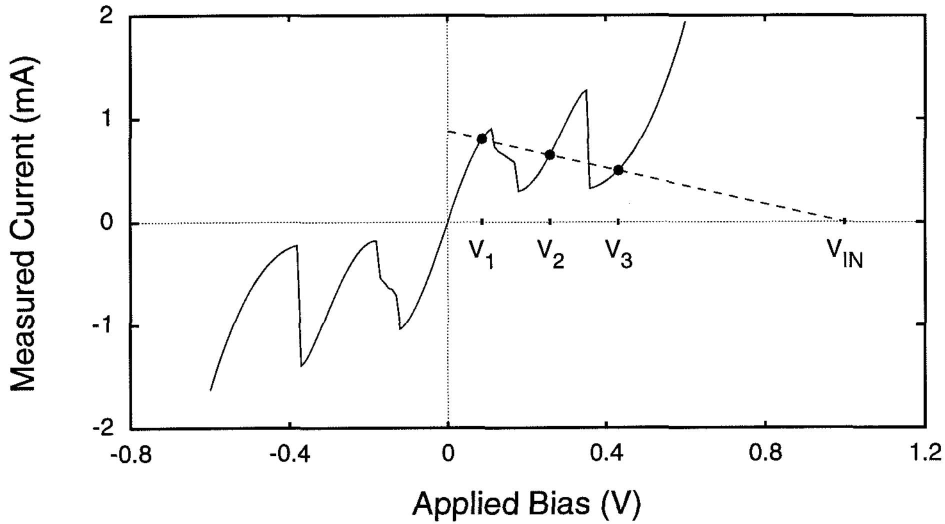
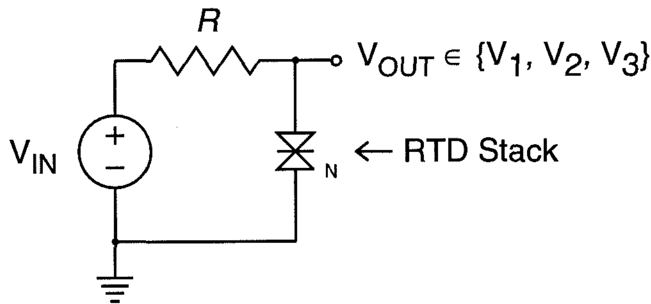


Figure 5.3: Schematic diagram of a four-quadrant vector-matrix multiplier.



(a)



(b)

Figure 5.4: Tristable load-line circuit.

cuits, as shown in Figure 5.4. Figure 5.4(a) shows the measured *I-V* characteristic of MBE sample III-370 grown by David Chow in the McGill-group laboratory. It is a two-device stack, and it is being biased with a 1300 Ω serial load with $V_{in} = 1.0V$ and $V_{out} \in \{V_1, V_2, V_3\}$. Switching among these states is accomplished either by pulsing V_{in} or by adding another pulsable current source to the V_{out} node.

Note that for N peaks in the *I-V* characteristic there will be $N + 1$ stable output voltages for the load-line circuit. While as many as nine have been integrated into a single stack [11], multiple stacks may be combined together to produce the sum of their individual state counts minus one [12]. The general use of N RTD devices in series is denoted by the N subscript on the RTD symbol in Figure 5.4(b). Also note that a transistor load may be used instead of a resistor in order to ensure all *I-V*

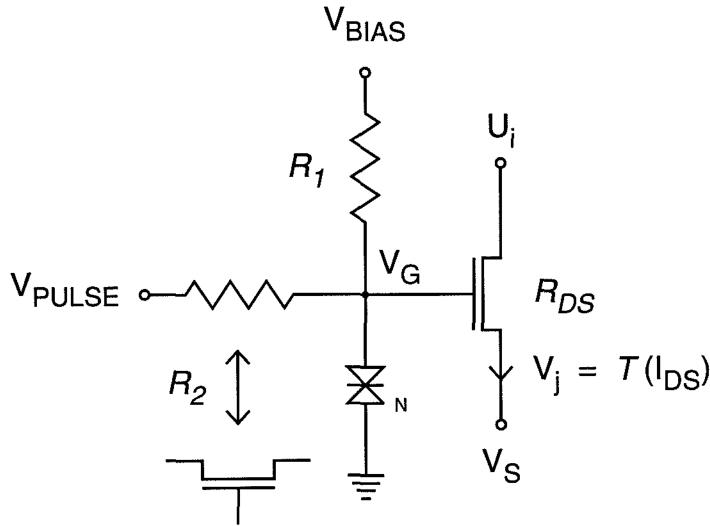


Figure 5.5: RTD synapse circuit.

peaks are intersected by the load-line [13].

The advantage to using RTD stacks to provide multistability is that the stacks themselves are perpendicular to the lithography plane and may be made well below 1 μm^2, thus providing tremendous numbers of states per unit area while being capable of state-to-state transitions at frequencies as high as 110 GHz [14].

In order to harness the variability that the RTD load-line circuit provides, an adjustable conductor must be added to the synaptic subcircuit. As an example, a single MOSFET operated in its ohmic region has been used [15]. Figure 5.5 shows this example subcircuit. The static global bias voltage V_{BIAS} maintains the MOSFET gate voltage V_G at one of the multistable operating points induced by the RTD stack in series with the resistive load R_1. The current injected by V_{PULSE} through the resistance R_2 is used to select the desired V_G operating point and hence the conductance of the MOSFET channel R_{DS}. In a VLSI array of these subcircuits some method of addressing which synapse to change is necessary, and this is provided by using a transistor in place of R_2 as shown in the figure.

For a given input voltage U_i a current $I_{DS} = (U_i - V_S)/R_{DS}$ will flow. Note that it is imperative to keep the source voltage V_S of the MOSFET constant so that I_{DS} depends only on U_i. In addition, some transimpedance function T is required to convert the current flowing into V_S from many synapses into an output voltage V_j.

Figure 5.6 shows an example of how this can be accomplished for a complete vector-matrix multiplier stage with one input and one output. Transimpedance amplifiers marked as T output a volt-

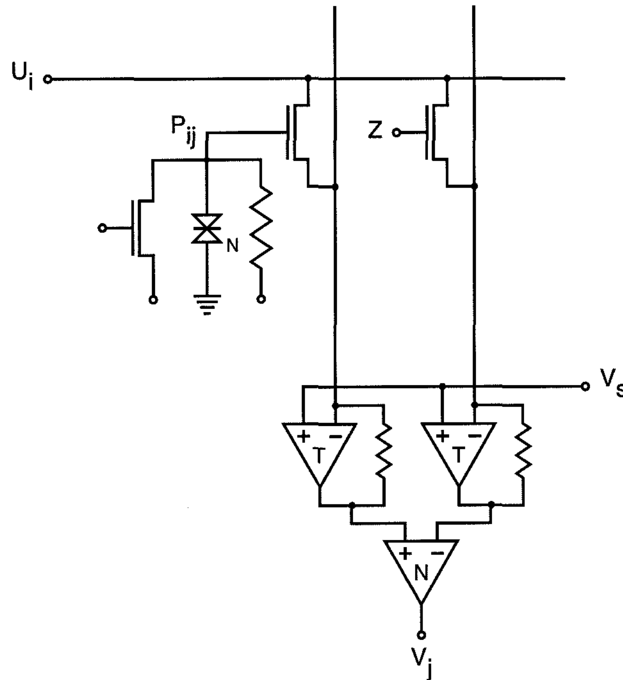


Figure 5.6: RTD vector-matrix multiplier stage.

age that keeps both of their inputs equal to V_S. Hence the more current that flows into them the larger their output. The amplifier marked as N is used to effect both the four-quadrant weighting and the neuron activation function.

Note that the voltage Z effectively maps the *physical* values of P_{ij} into an arbitrary set of *logical* W_{ij} values. For example, if Z is chosen to be at the middle of the range of available P_{ij}, then the effective $[W_{ij}]$ range is $\pm[P_{ij}]/2$. For the case of a tristable synapse circuit, setting $Z = V_2$ yields possible logical weights of $\{-1, 0, 1\}$. The next section presents a hardware demonstration of just this scenario.

HARDWARE DEMONSTRATION

Figure 5.7 shows an overlay of the I-V characteristics for six double-RTD stacks used to make the synapses in a network with the architecture shown in Figure 5.2. The uniformity is important when inputs through different synapses must sum to zero for a given W_{ij} solution set to work for a particular task.

Amplifiers like that marked N in Figure 5.6 were used in a high-gain configuration to produce a

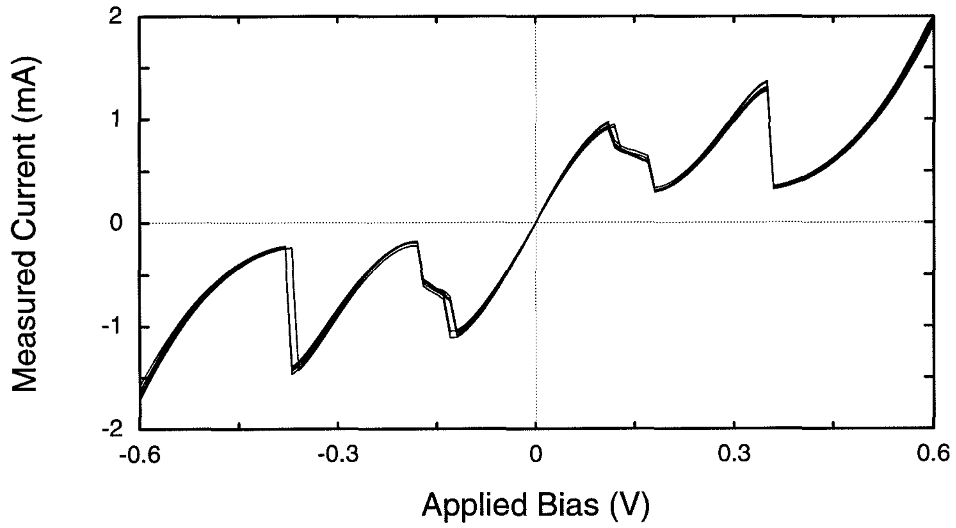


Figure 5.7: Overlay of RTD *I-V* characteristics.

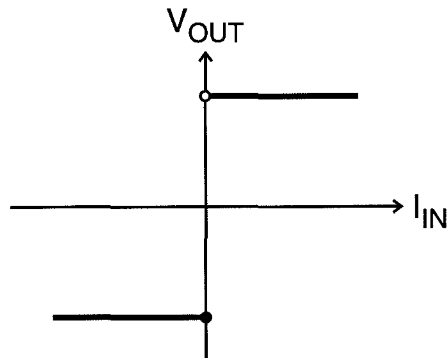


Figure 5.8: A binary approximation to a neuron activation function.

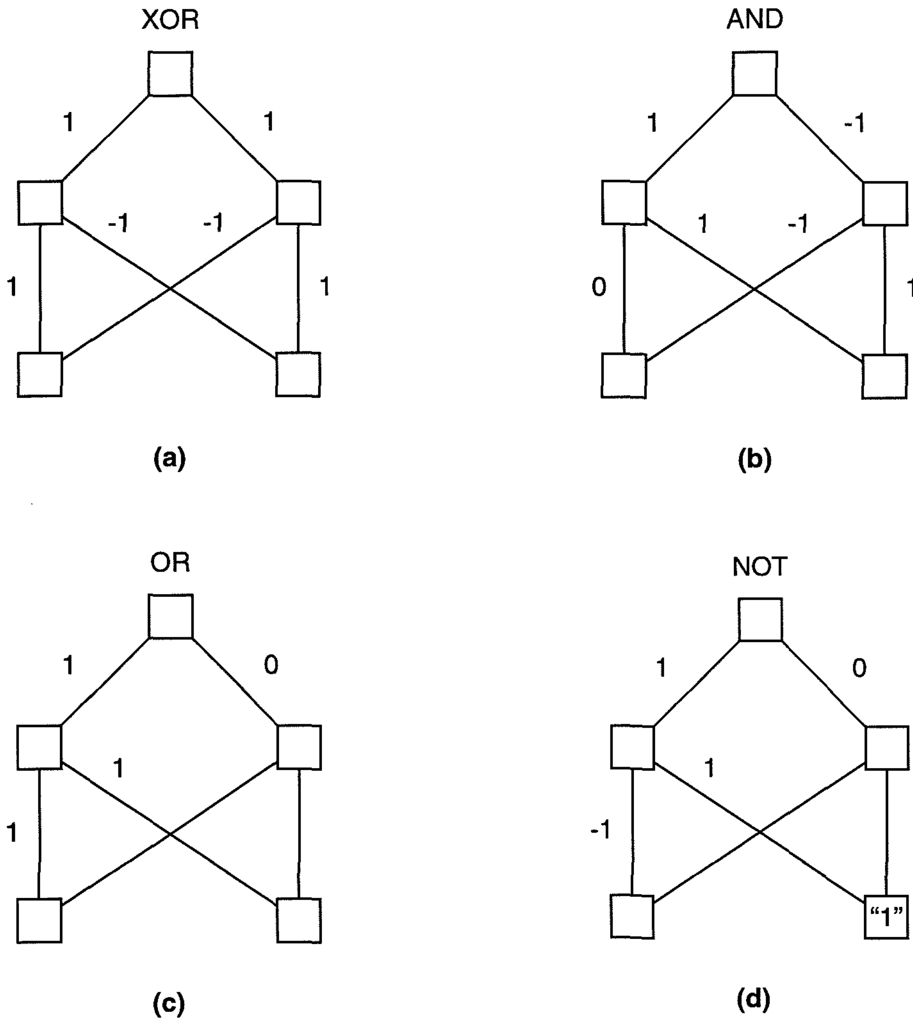
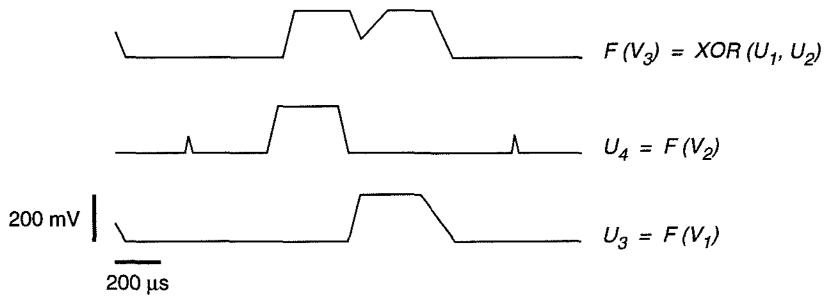


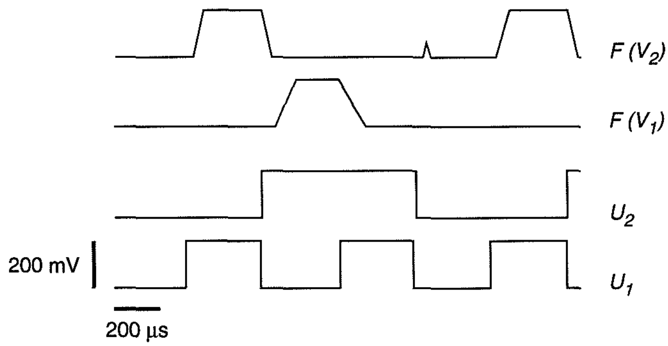
Figure 5.9: Synaptic weight solutions to the primary logic problems.

binary approximation to the neuron activation function as shown in Figure 5.8. In order to test the synapses at the extrema of their input range, the primary logic problems were used as tests for the network. Note that the system is an analog circuit, and the purpose of these tests is not to suggest replacement of digital logic with this neural network. The W_{ij} states that solve each of the logic problems are shown in Figure 5.9.

Simple clocking circuits were built to run through the truth tables for each of the problems. Oscilloscope traces for the separate vector-matrix multiplier stages required to implement the architecture in Figure 5.2 are shown in Figure 5.10 for the XOR problem. Figure 5.11 shows the input-output performance of the network on the other logic problems.

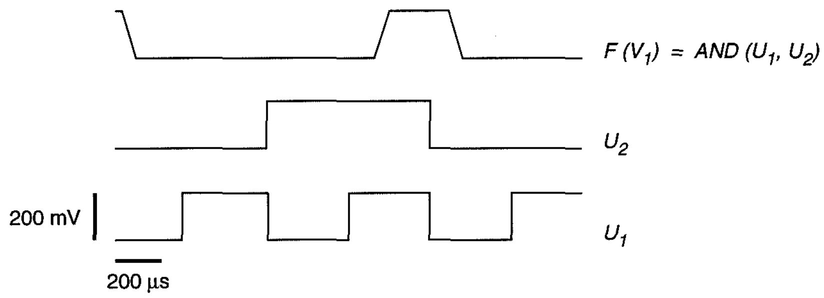


(a)

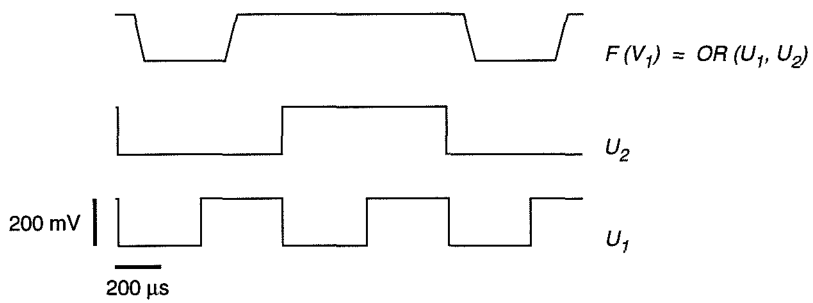


(b)

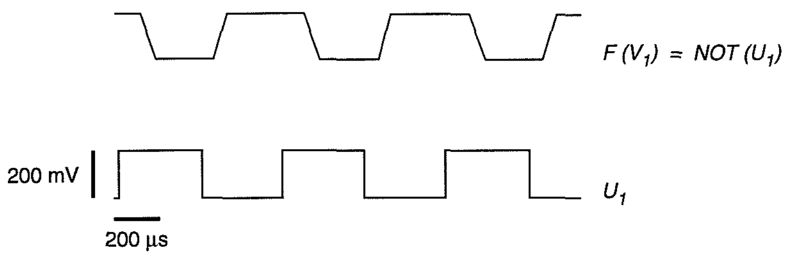
Figure 5.10: XOR test.



(a)



(b)



(c)

Figure 5.11: AND, OR, and NOT tests.

REFERENCES

[1] P. M. Churchland and P. S. Churchland, "Could a machine think?," *Scientific American*, vol. 262, no. 1, pp. 32–37, 1990.

[2] E. R. Kandel and J. H. Schwartz, *Principles of Neural Science.* New York, NY: Elsevier, 2nd ed., 1985.

[3] P. S. Churchland and T. J. Sejnowski, *The Computational Brain.* Cambridge, MA: MIT Press, 1992.

[4] D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, *Parallel Distributed Processing*, vol. 1. Cambridge, MA: MIT Press, 1986.

[5] J. A. Anderson, "General introduction," in *Neurocomputing: Foundations of Research* (J. A. Anderson and E. Rosenfeld, Eds.), Cambridge, MA: MIT Press, 1988.

[6] F. J. Kub, K. K. Moon, I. A. Mack, and F. M. Long, "Programmable analog vector-matrix multipliers," *IEEE J. Solid-State Circuits*, vol. 25, pp. 214–221, Feb. 1990.

[7] Y. Tsividis and S. Satyanarayana, "Analogue circuits for variable-synapse electronic neural networks," *Electronic Letters*, vol. 23, pp. 1313–1314, Nov. 1987.

[8] A. M. Chiang, "A CCD programmable signal processor," *IEEE J. Solid-State Circuits*, vol. 25, pp. 1510–1517, Dec. 1990.

[9] E. G. Spencer, "Programmable bistable switches and resistors for neural networks," in *AIP Conference Proceedings 151: Neural Networks for Computing* (J. S. Denker, Ed.), (New York, NY), pp. 414–419, American Institute of Physics, AIP, 1986.

[10] A. P. Thakoor, J. L. Lamb, A. Moopenn, and J. Lambe, "Binary synaptic connections based on memory switching in a-Si:H," in *AIP Conference Proceedings 151: Neural Networks for Computing* (J. S. Denker, Ed.), (New York, NY), pp. 426–431, American Institute of Physics, AIP, 1986.

[11] A. C. Seabaugh, Y.-C. Kao, and H.-T. Yuan, "Nine-state resonant tunneling diode memory," *IEEE Electron Device Letters*, vol. 13, pp. 479–481, Sep. 1992.

[12] S.-J. Wei and H. C. Lin, "Multivalued SRAM cell using resonant tunneling diodes," *IEEE J. Solid-State Circuits*, vol. 27, pp. 212–216, Feb. 1992.

[13] Z. X. Yan and M. J. Deen, "A new resonant-tunnel diode-based multivalued memory circuit using a MESFET depletion load," *IEEE J. Solid-State Circuits*, vol. 27, pp. 1198–1202, Aug. 1992.

[14] E. Özbay and D. M. Bloom, "110-GHz monolithic resonant-tunneling-diode trigger circuit," *IEEE Electron Device Letters*, vol. 12, pp. 480–482, Sep. 1991.

[15] H. J. Levy and T. C. McGill, "A feedforward artificial neural network based on quantum effect vector matrix multipliers," *IEEE Trans. Neural Networks*, vol. 4, pp. 427–433, May 1993.

Chapter 6

COMPUTER MEMORY

Perhaps the systems likely to benefit the most from advances in cellular VLSI are computer memories, currently containing as many as 64 million identical cells arranged into simple arrays on a single chip. Since the presence of so many cells places a severe limitation on the power consumption per cell, today's digital memories store information as *voltages* isolated by transistors. This chapter presents a design for a *transistorless current-mode* digital memory based on the tunneling-switch diode (TSD) that could exhibit nominal power consumption, rapid read/write speeds, and incredibly high cell densities.

SYSTEM PRINCIPLES

Designing a computer memory is a tedious exercise in constraint solving, and progress has largely been evolutionary instead of revolutionary [1]. Computer memory must, of course, be fast, small, energetically frugal, inexpensive to build, and hopefully integratable with other VLSI technology (e.g. the same as what is used to build processor chips). The memory design proposed in this chapter is based upon the likelihood that evolution in conventional VLSI technology will soon approach the parameter space where quantum effect devices like the TSD can enter the mainstream.

Conventional Voltage-Mode Memory

To begin the presentation, first consider the design of a conventional *voltage-mode* random-access memory (RAM) like that shown in Figure 6.1. Information is represented as a two-dimensional ar-

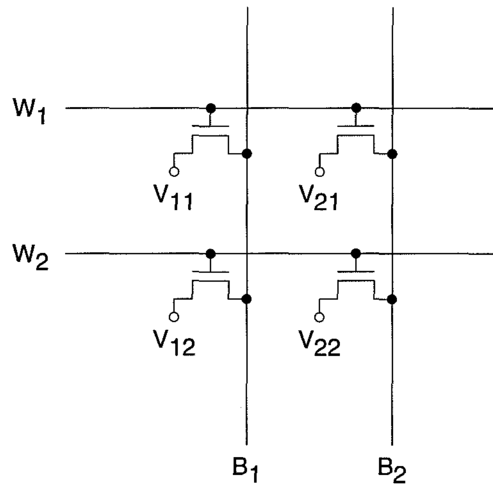


Figure 6.1: A conventional voltage-mode RAM with isolation transistors.

ray of voltages V_{xy} in memory cells that store charge either on a leaky capacitor, as in the case of a dynamic RAM (DRAM), or on the gate of a transistor in a small subcircuit with feedback, as in the case of a static RAM (SRAM). In either case, isolation transistors are required to keep the memory cells from shorting together and losing information.

Separate from how the information is *stored* is the issue of how the information is *sensed*. In the case of a DRAM the stored charge must be directly used to determine the information content of a memory cell because that is all there is obtainable from the cell. In a SRAM, however, the stored charge is used to control an internal transistor that can source a current for sensing purposes. Note that in both of these memory designs charge must be transferred from the memory cell to the gate of a transistor somewhere in the sensing circuitry; this charge is used to discriminate between discrete memory cell states.

The amount of time it takes to make a sensing decision is dependent upon the amount of memory cell sensing charge Q_{sense} available and its mechanism of transfer to the sensing circuitry. Since a capacitor discharges with exponential decay, DRAM states typically take longer to discriminate between than SRAM states which are conveyed by a large current. A typical value of required Q_{sense} for an advanced DRAM or SRAM design is 100 fC, which yields sense times of ~50 ns for minimum-sized DRAM capacitors and ~1 ns for 150 μA SRAM cells [1–4].

Note that these voltage-mode memory cells exhibit extremely low power dissipation when they are not being read or written to because hardly any current is flowing. The ability of the SRAM

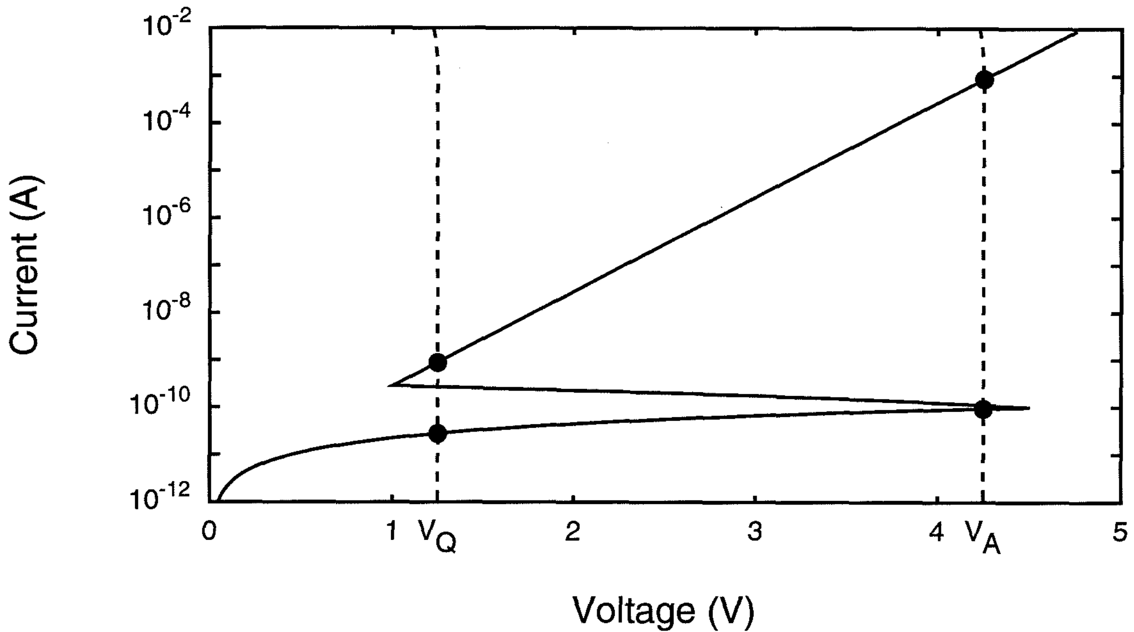


Figure 6.2: Ideal load-line behavior for the TSD memory.

cell to provide a large sensing current only when it is addressed comes at the expense of additional lithography real-estate used for at least four transistors. What would be optimal is a device technology that provides similar dichotomous behavior without any transistors, and this is the goal of the TSD-based design.

TSD Transistorless Current-Mode Memory

Figure 6.2 shows a conception of an ideal load-line behavior goal. It is ideal in that a linear perturbation in bias voltage from V_Q to V_A can exponentially separate the I-V fixed-points over six orders of magnitude. The essence of the TSD memory is to have 10^9 static memory cells drawing $\lesssim 1$ nA with their quiescent V_Q bias while having a small number of addressed cells providing ~ 1 mA with V_A bias (if they are in the low-impedance state) to the sensing circuitry. Since the information is stored as a current, there is no need for isolation transistors to prevent shorting.

To clarify this point, Figure 6.3 shows a 2x2 array of memory cells (simple load-line circuits) with a static applied bias of $V_s = V_{s+} - V_{s-}$ across them. This would correspond to the V_Q state in Figure 6.2. Note that the total power dissipation is kept very low regardless of which impedance state the cells are in.

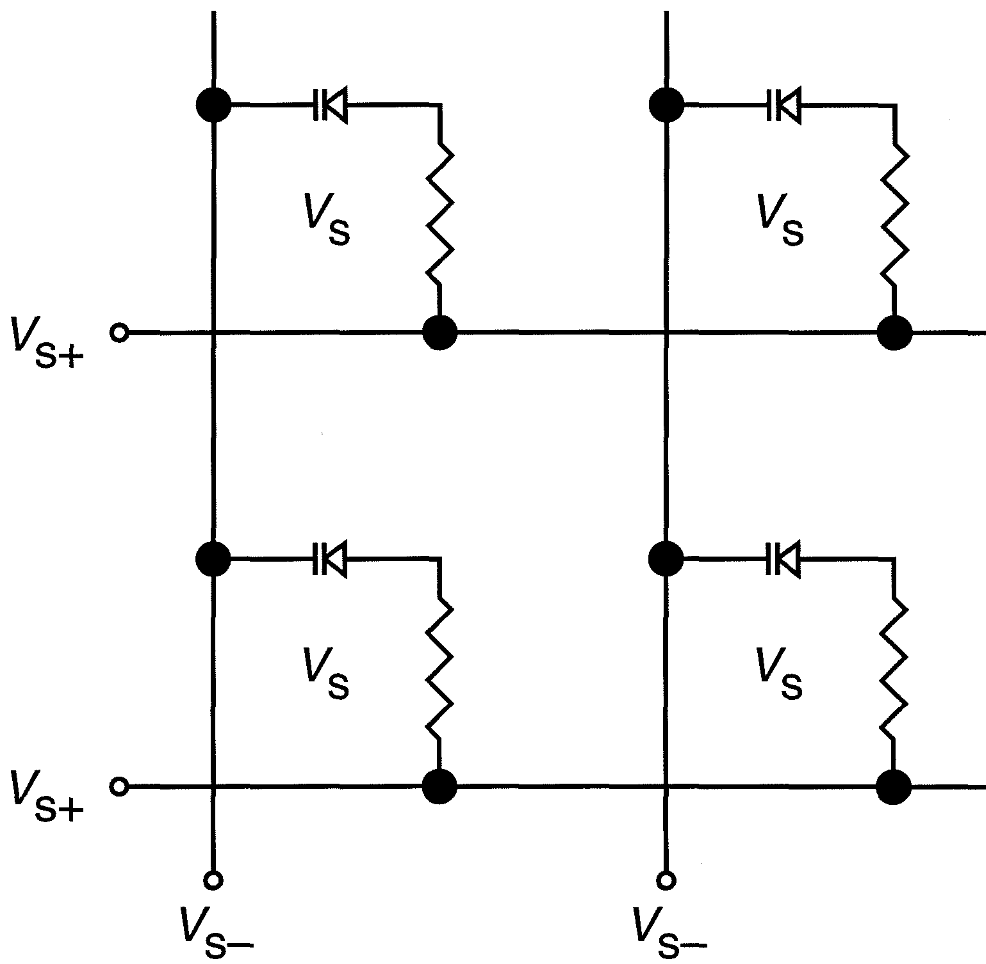


Figure 6.3: TSD memories under static bias.

When a particular cell is to be sensed, a reading scheme like that shown in Figure 6.4 can be used. In this scheme the word line voltage is increased by an amount δ to obtain a load-line condition like V_A in Figure 6.2 for the entire row of cells. If any of these cells are in the low-impedance state, then they will source a current that is six orders of magnitude greater than any other cell on their corresponding bit line and allow this state to be rapidly discerned by the $\sim$1 mA current. If these cells are in the high-impedance state, then this current will not be present. A specific cell connected to the word line may be examined by choosing the appropriate bit line to sense. Note that because the increased bias $V_s + \delta$ does not exceed the peak-voltage of the TSD I-V characteristic, the impedance states of the memory cells do not change. Another possible reading scheme with lower power consumption for exponential-I-V memory cells shown in Figure 6.5. This scheme takes advantage of the fact that $e^{\delta} > 2e^{\delta/2}$ for $\delta > 2\ln 2$, so that only the addressed cell is biased to the V_A point of Figure 6.2. Of course many other similar distributions of δ across the word and bit lines are possible, and the optimal choice will depend on the dimensions of the memory cell array.

To change the impedance state of a specific cell, a write scheme like that shown in Figure 6.6 can be used. In this case a large-enough or small-enough voltage is applied to a specific cell to force it to go into the desired impedance state. To ensure only the cell connected to a specific word-line/bit-line address pair is changed, the voltage difference from V_s required to change the impedance state can be split across the word line and bit line. This is shown by changing the word line to $V_{s+} + \Delta$ and the bit line to $V_{s-} - \Delta$ so that the cell to be changed sees a total change of 2Δ, while all of the other cells on the pertaining word line and bit line see only a change of Δ. The value of Δ is thus chosen so that only 2Δ is enough to change the impedance state. Note that implicit in this scheme is the specification that there will be a different Δ for writing the low-impedance state ($\Delta > 0$) than for writing the high-impedance state ($\Delta < 0$). From Figure 6.2 it is apparent that changing to the high impedance state from the V_Q bias point will require a much lower $|\Delta|$ than for changing to the low-impedance state.

The schematic diagram for a working 2x2 memory test circuit we have constructed from discrete devices is shown in Figure 6.7; it contains steering logic for applying the various reading and writing voltages to the word and bit lines. We now look forward to testing the circuit in a monolithic form.

An important issue in implementing these addressing schemes with a VLSI technology is the presence of large-enough transistors in the addressing circuitry to be able to source and sink the $\sim$1 mA current (the word line addressing transistors will have to potentially source this much cur-

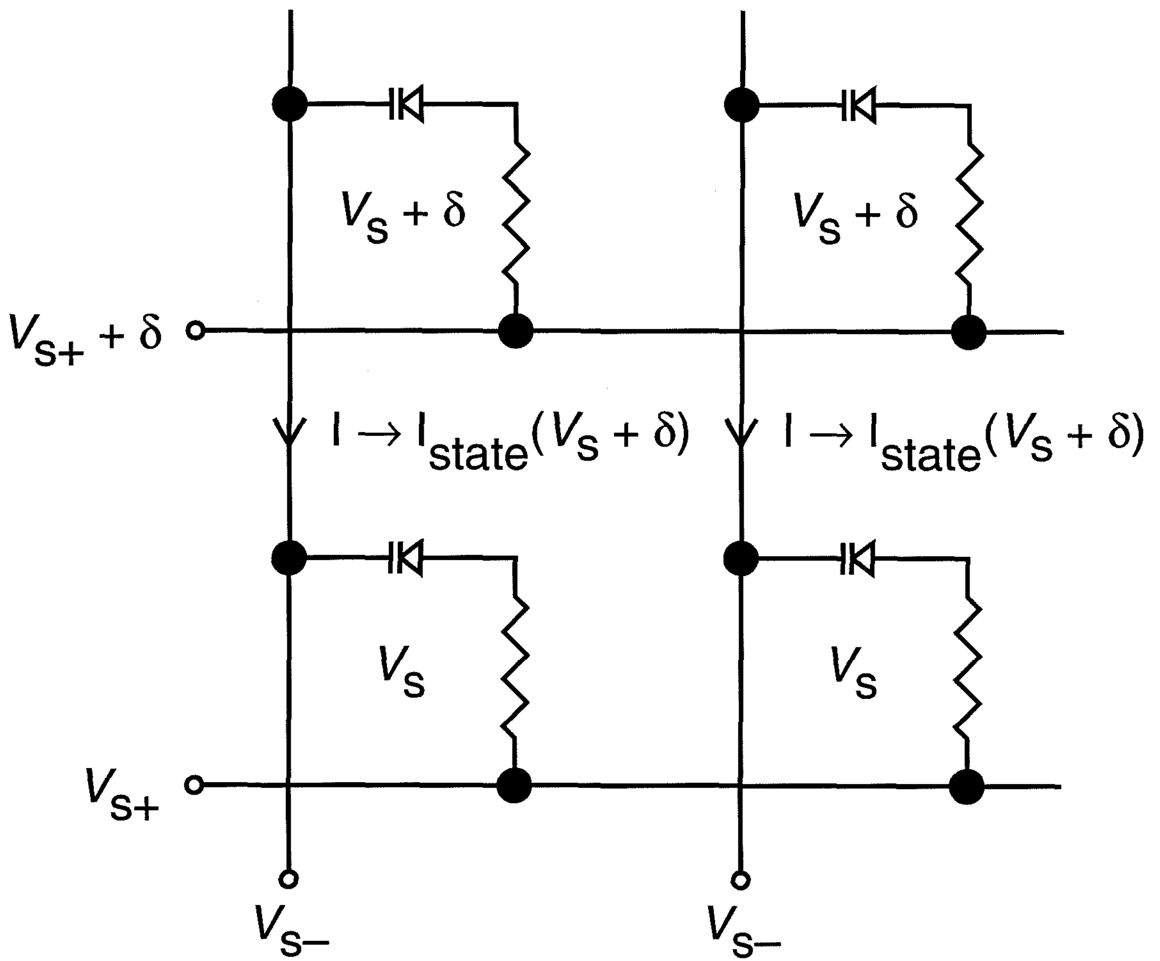


Figure 6.4: TSD memory read scheme.

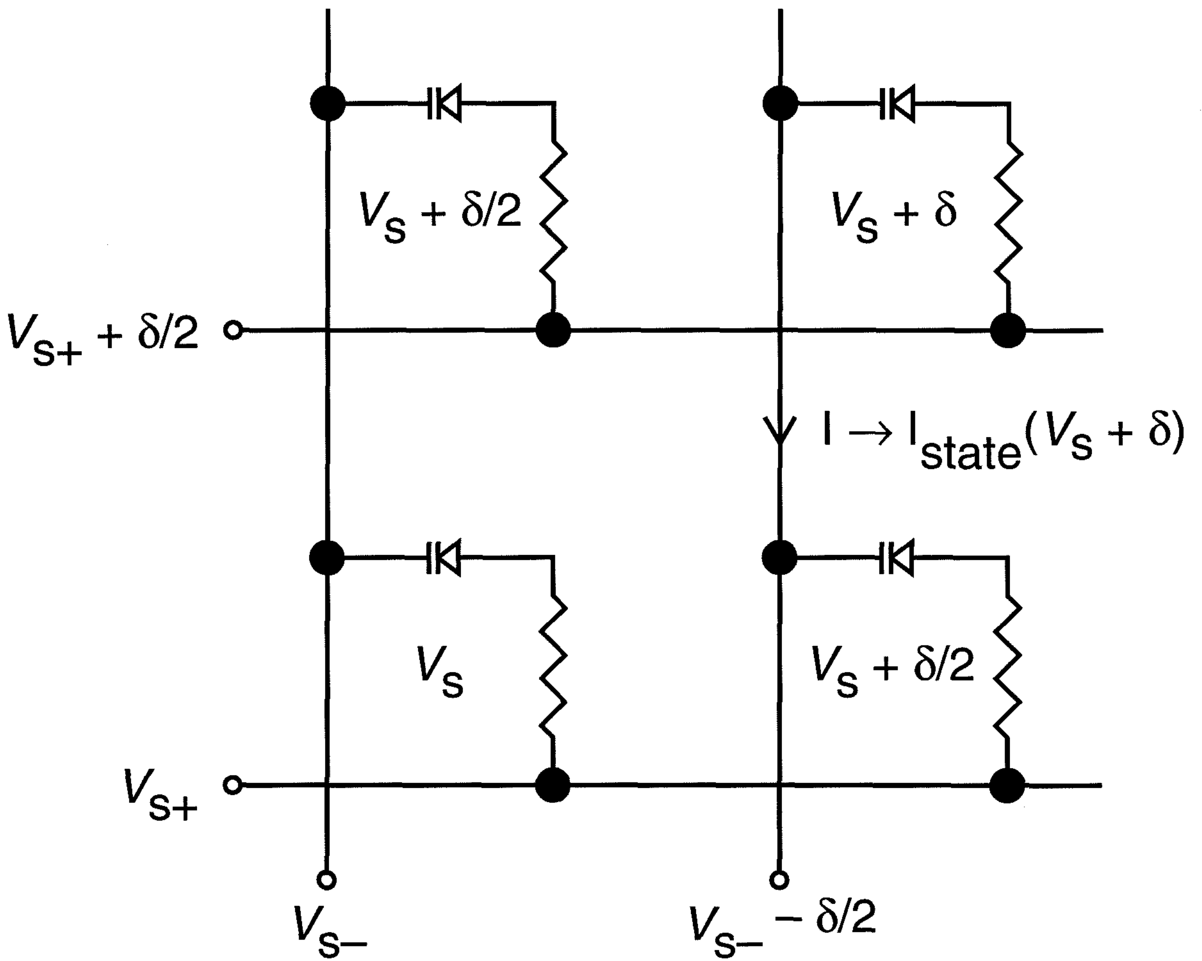


Figure 6.5: Alternative TSD memory read scheme that consumes less power for exponential *I-V* devices.

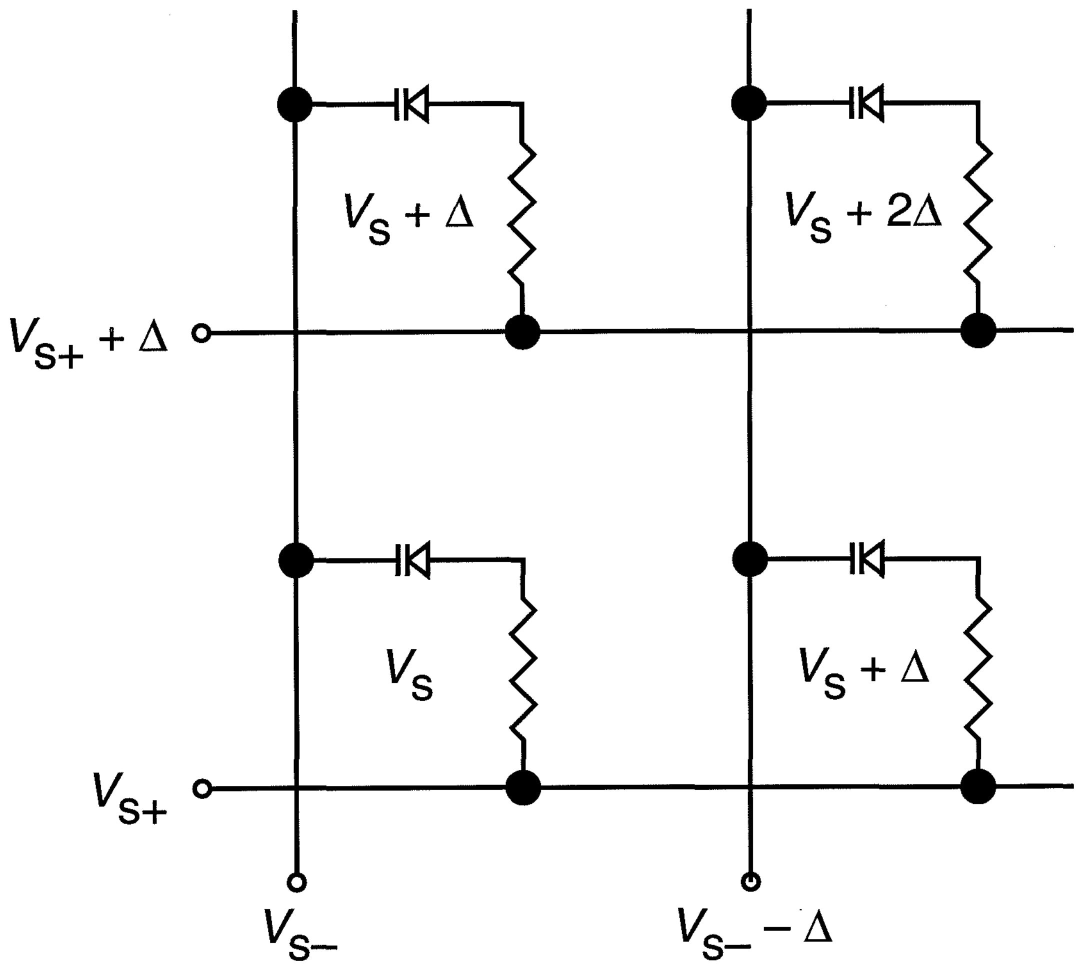


Figure 6.6: TSD memory write scheme.

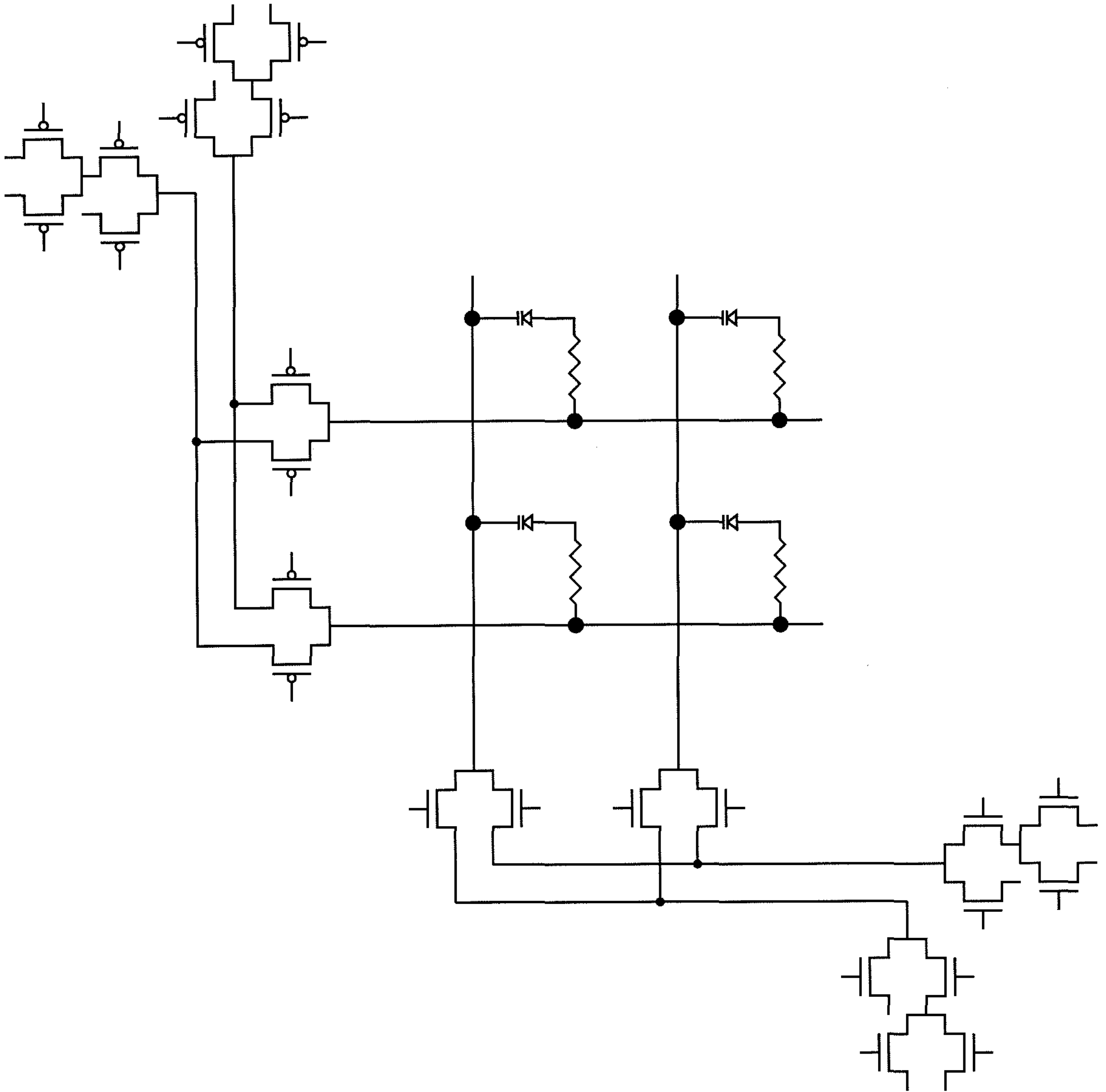


Figure 6.7: A 2x2 demonstration circuit with read/write steering logic (built from discrete devices).

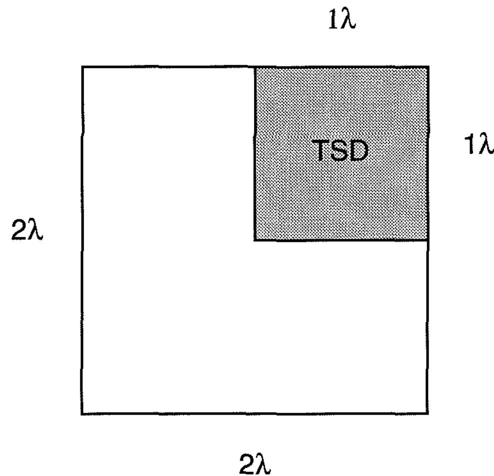


Figure 6.8: TSD memory cell layout.

rent for each of the memory cells on the word line if they are in the high-impedance state). These transistors will likely not be minimum sized, and consequently there will be the design headache of aligning the physical layout pitch of the memory array with that of the larger addressing transistors. This is a common and surmountable problem for any kind of addressable high-density array such as what is used in memory or imaging chips.

FABRICATION ISSUES

A possible physical layout design for the TSD-based memory cell is shown in Figure 6.8. If 1 λ is the minimum feature size attainable on a given fabrication line, then a 1 λ x 1 λ device would be used with a surrounding 1 λ space between devices, thus yielding a 4 λ^2 cell size. Cells could be isolated from each other either by etching or by ion-implantation.

A possible *self-aligned* fabrication scheme for integrating the devices together with word lines and bit lines is shown in Figure 6.9. Figure 6.9(a) shows a wafer produced by an epitaxial method like CVD or MBE to provide an insulating substrate, a conducting layer, and a TSD layer. The insulating substrate could be provided by sapphire or a reverse-biased *pn*-junction. The conducting layer could be provided by either a highly doped semiconductor layer or perhaps by a metal silicide. Note that the load-line resistance does not need to be explicitly fabricated because the intrinsic series resistance of the TSD layer is more than adequate.

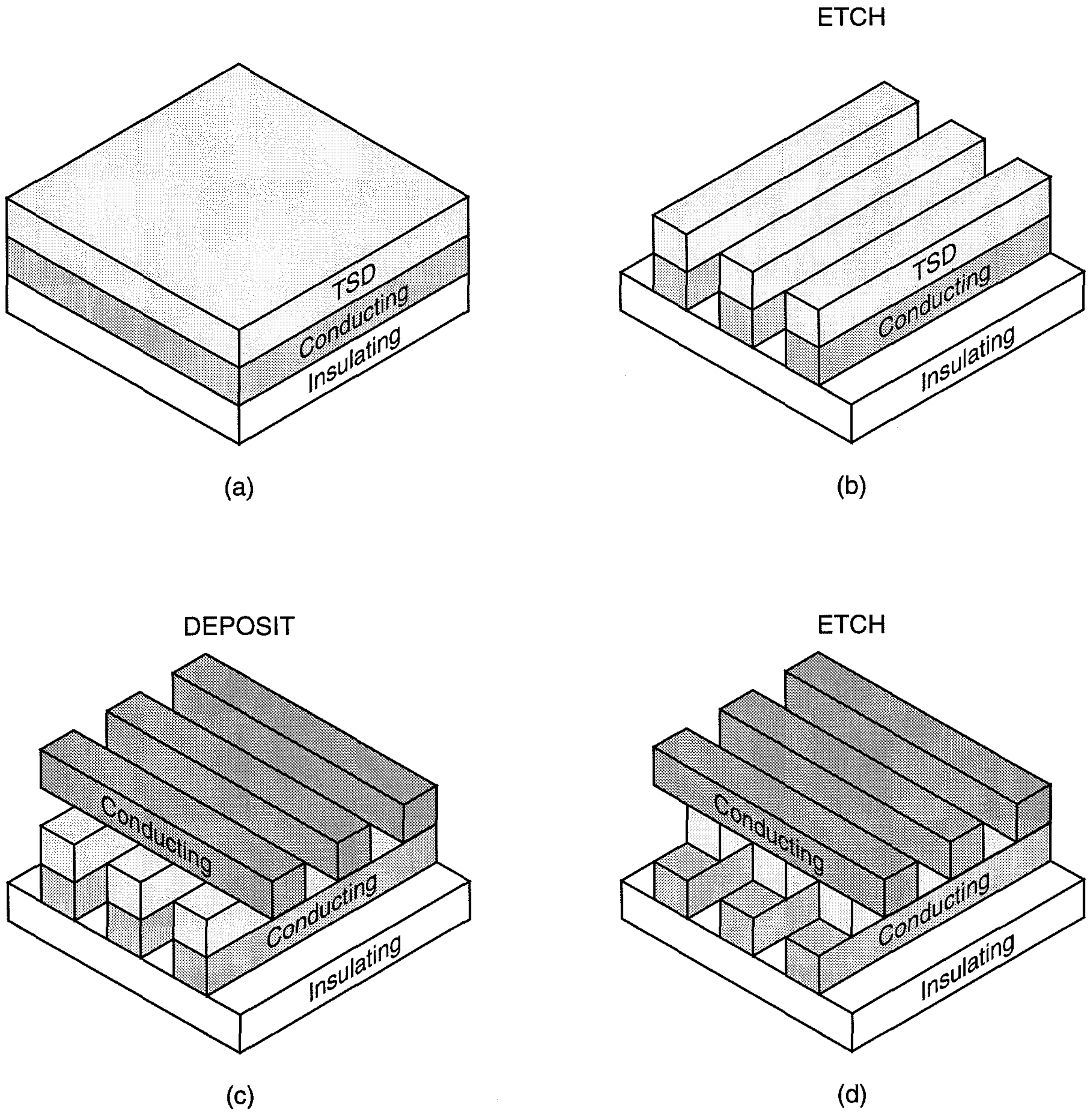


Figure 6.9: A possible TSD memory fabrication scheme.

After the tunneling barrier is formed on the wafer to complete the formation of the TSD layer, one set of addressing lines can be formed by etching the device layers into columns as shown in Figure 6.9(b). Then the other and orthogonal set of addressing lines can then be deposited onto the columns (the interstitial spaces may first be filled with dielectric or may be left as air), as shown in Figure 6.9(c).

Lastly, an etch step may be performed to define memory cells only where the orthogonal addressing lines intersect, as shown in Figure 6.9(d), thus yielding the memory cell cross-section portrayed in Figure 6.8. The self-aligned nature of this fabrication procedure should greatly improve the yield of memory cells in the array.

An important fabrication issue in a scheme like this is ensuring a low-enough resistance for the addressing lines. If there is a considerable voltage drop between cells, then when one of the cells in the low-impedance state is addressed with increased applied bias, it will pull-down the bias across all of the other cells on the same address lines, possibly switching those cells into the the high-impedance state. To make the system function with a considerable voltage drop between cells, the quiescent bias value V_s must be increased to provide a larger amount of state-protecting hysteresis; unfortunately this will also exponentially increase the power dissipation of those cells in the low impedance state. If this is a problem, however, laterally-displaced addressing lines can be used so that they both may be fabricated from the same highly conductive and arbitrary material; in this case there is a slight loss in cell density instead of an increase in power dissipation.

PERFORMANCE ESTIMATES

An estimate of performance can be obtained by examining the memory cell size required to source enough current to provide Q_{sense} in a specified amount of time. This sensing current I_{sense} is simply specified by the amount of time τ_{sense} we are willing to wait (or have to wait) for the Q_{sense} to come out of the memory cell:

$$I_{sense} = \frac{Q_{sense}}{\tau_{sense}}. \tag{6.1}$$

If J_{TSD} is the maximum current-density the TSD-based memory cell can handle, then the minimum area A_{TSD} the TSD must be made with is given by

$$A_{TSD} = \frac{I_{sense}}{J_{TSD}}. \tag{6.2}$$

Table 6.1: Performance estimation for the TSD memory.

Maximum Transient Current Density, J_{TSD} (A/cm^2)	**Required β_{low} For $1A/cm^2$**	**TSD Size (μm) and Cell Density (cm^{-2}) For $Q_{sense} = 200fC$**		
		$200\mu A \times 1ns$	$20\mu A \times 10ns$	$2\mu A \times 100ns$
10^5	$25,000$	$.45 \times .45$ 125MB	$.14 \times .14$ 1.25GB	$.04 \times .04$ 12.5GB
10^4	$2,500$	1.4×1.4 12.5MB	$.45 \times .45$ 125MB	$.14 \times .14$ 1.25GB
10^3	250	4.5×4.5 1.25MB	1.4×1.4 12.5MB	$.45 \times .45$ 125MB

MB = Megabits, GB = Gigabits

Note that A_{TSD} is equivalent to 1 λ^2 from Figure 6.8, so that the total memory cell area A_{cell} is four times A_{TSD} and the fabrication line must be able to provide feature sizes as small as $\sqrt{A_{TSD}}$.

We can estimate the static current draw per cell by considering the ratio β_{low} of the low-impedance state currents for the quiescent and addressed bias points, shown as 10^6 in Figure 6.2:

$$\beta_{low} \equiv \frac{I_{A,low}}{I_{Q,low}} \tag{6.3}$$

$$I_{static} = \frac{I_{sense}}{\beta_{low}}. \tag{6.4}$$

Table 6.1 presents some performance estimates for a conservative Q_{sense} of 200 fC. To read the table, start with a value of J_{TSD}; this will determine the minimum TSD size for a specified $I_{sense} \times \tau_{sense}$. Once the TSD size is known, it is multiplied by four and used to compute how many cells could fit within 1 cm^2. If this number of cells is not to draw more than 1 A of quiescent current, then the TSD devices must provide at least the β_{low} displayed in Table 6.1.

For a state-of-the-art memory product like the NEC μPD42S64xx 64 megabit DRAM with λ =

$0.35\mu m$ and $\tau_{sense} = 50ns$ (the cell size is $0.85\mu m \times 1.69\mu m$) [5], Table 6.1 indicates a cell density increase of a factor of ~3 for the same feature size and a factor of ~9 for the same sense time and a conservative $J_{TSD} = 10^4 A/cm^2$. This factor would approach two orders of magnitude if J_{TSD} could be raised to $10^5 A/cm^2$!

To date, TSD devices have been produced in silicon with switching times of 1 ns in both directions [6, 7] and with current densities of $\sim 10^4 A/cm^2$ [8]. We have so far fabricated TSD devices with β_{low} parameters of $\sim 10^4$. We fully expect to come close to the goal of 10^6 by lowering the series resistance of the TSD semiconductor layers and by perhaps using a tunneling barrier like SiO_xN_y which provides greater tunneling transmission than SiO_2 for the same thickness.

SCALING ISSUES

This section considers how the performance estimates of Table 6.1 may be affected by parameter constraints imposed by fabrication. For example, fabrication techniques such as etching are usually capable of some maximum aspect ratio; as the lateral dimensions of the device are scaled down, similar reductions will have to be made in the vertical dimension as well. Changing this vertical dimension, however, will change the *I-V* characteristic if no other device parameter is used for compensation. Thus it will be helpful to have an estimate of what *n*- epilayer thickness t_d and donor concentration N_d yield the same peak-voltage V_{peak}.

Figure 6.10 plots N_d as a function of t_d for two values of V_{peak} assuming a punch-through switching mechanism. This is calculated as follows [9]. First note that the distance to be punched-through is the epilayer thickness up to the beginning of the depletion region of the *np*-junction:

$$x_d = t_d - W_j \tag{6.5}$$

where W_j is the depletion width for the one-sided abrupt *np*-junction. From Poisson's equation we know that the voltage V_d required to deplete this region is given by

$$V_d = \frac{qN_d}{2\epsilon}x_d^2 \tag{6.6}$$

where q is the electronic charge and ϵ is the dielectric constant [10]. Before punch-through most of the voltage drop is across the depletion region, so we use the approximation $V_{peak} \sim V_d$. Note that

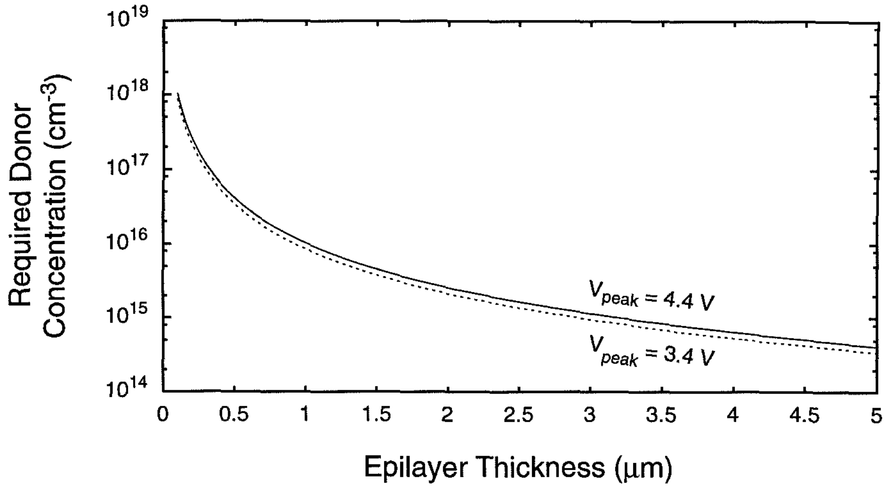


Figure 6.10: Required donor concentration as a function of epilayer thickness for various *I-V* peak-voltages (punch-through mode).

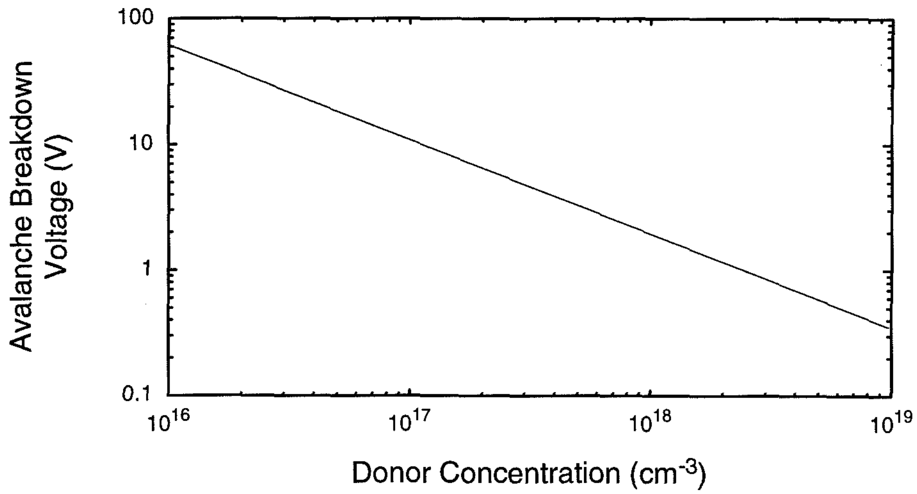


Figure 6.11: Avalanche breakdown voltage as a function of donor concentration.

the value of W_j also depends on N_d as

$$W_j = \sqrt{\frac{2\epsilon}{qN_d}(V_{bi} - 2kT/q)} \tag{6.7}$$

where V_{bi} is the built-in junction potential and kT is the Boltzmann thermal energy [10]. Figure 6.10 is obtained by solving Equation 6.6 and Equation 6.7 for N_d given t_d and V_{peak}.

When very high donor concentrations are used, avalanche breakdown in the epilayer will occur before punch-through. Figure 6.11 plots avalanche breakdown voltage as a function of donor concentration using the approximation

$$V_a = 60\left(\frac{E_g}{1.1}\right)^{3/2}\left(\frac{N_d}{10^{16}}\right)^{-3/4} \tag{6.8}$$

where V_a is the breakdown voltage in volts and E_g is the room-temperature bandgap in eV [10].

Now when Figure 6.10 and Figure 6.11 are analyzed together, we can see that V_{peak} can be kept constant when t_d is decreased by concurrently increasing N_d, but only until the avalanche breakdown voltage drops below the punch-through voltage, somewhere above $N_d \sim 10^{17} cm^{-3}$. This indicates that t_d can be made below 0.5 μm, which allows lateral dimensions below 0.1 μm for a reasonable aspect ratio of 5:1.

OTHER ISSUES

This chapter concludes with some comments on other issues concerning the feasibility of the TSD memory circuit. For example, the temperature stability of the *I-V* characteristic is important for picking the quiescent and addressed bias levels that ensure the retention of the memory cell states. Some experimental studies have been conducted [8, 9] with results that certain structures can be made with exceptional temperature stability [8]. Other issues, such as the uniformity of *I-V* characteristics throughout the memory cell array, and the lateral isolation of memory cells from each other, are presently unresolved. Fortunately these issues are also pertinent to the MOSFET evolutionary cycle, indicating that much effort will be spent in these areas in the years to come.

REFERENCES

[1] K. Itoh, "Trends in megabit DRAM circuit design," *IEEE J. Solid-State Circuits*, vol. 25, pp. 778–789, June 1990.

[2] T. Nagai, K. Numata, M. Ogihara, M. Shimizu, K. Imai, T. Hara, M. Yoshida, Y. Saito, Y. Asao, S. Sawada, and S. Fujii, "A 17-ns 4-Mb CMOS DRAM," *IEEE J. Solid-State Circuits*, vol. 26, pp. 1538–1543, Nov. 1991.

[3] E. Seevinck, P. J. van Beers, and H. Ontrop, "Current-mode techniques for high-speed VLSI circuits with application to current sense amplifier for CMOS SRAM's," *IEEE J. Solid-State Circuits*, vol. 26, pp. 525–536, Apr. 1991.

[4] H. Nambu, K. Kanetani, Y. Idei, N. Homma, K. Yamaguchi, T. Hiramoto, N. Tamba, M. Odaka, K. Watanabe, T. Ikeda, K. Ohhata, and Y. Sakurai, "High-speed sensing techniques for ultrahigh-speed SRAM's," *IEEE J. Solid-State Circuits*, vol. 27, pp. 632–640, Apr. 1992.

[5] "NEC μPD42S64xx DRAM," *Semiconductor International*, p. 46, June 1993.

[6] T. Yamamoto, K. Kawamura, and H. Shimizu, "Silicon p-n insulator-metal (p-n-I-M) devices," *Solid-State Electronics*, vol. 19, pp. 701–706, 1976.

[7] H. Kroger and H. A. R. Wegener, "Steady-state characteristics of three terminal inversion-controlled switches," *Solid-State Electronics*, vol. 21, pp. 655–661, 1978.

[8] H. Kroger and H. A. R. Wegener, "Steady-state characteristics of two terminal inversion-controlled switches," *Solid-State Electronics*, vol. 21, pp. 643–654, 1978.

[9] J. G. Simmons and A. A. El-Badry, "Switching phenomena in metal-insulator-n/p+ structures: theory, experiment and applications," *Radio and Electronic Engineer*, vol. 48, pp. 215–226, May 1978.

[10] S. M. Sze, *Physics of Semiconductor Devices*. New York, NY: Wiley, 2nd ed., 1981.