Process Development for a Silicon Planar Resonant-Tunneling Field-Effect Transistor

by

Mike Chou

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 1994

Author .....
Department of Electrical Engineering and Computer Science
January 14, 1994

Certified by .....
Henry I. Smith
Professor
Thesis Supervisor

Certified by .....
Dimitri A. Antoniadis
Professor
Thesis Supervisor

Accepted by .....
Frederic R. Morgenthaler
Chairman, Departmental Committee on Graduate Students

Process Development for a Silicon Planar Resonant-Tunneling Field-Effect Transistor

by

Mike Chou

Submitted to the Department of Electrical Engineering and Computer Science
on January 14, 1994, in partial fulfillment of the
requirements for the degree of
Master of Science in Electrical Engineering

Abstract

Field-induced resonant tunneling in a 2-dimensional electron gas (2-DEG) offers exciting possibilities for quantum-based device applications as well as a fertile ground for fundamental studies of electrical transport in the mesoscopic regime. The Planar RESonant-Tunneling Field-Effect Transistor (PRESTFET), designed for maximum flexibility, can achieve resonant tunneling (RT) under a variety of bias conditions. Although such a device has been successfully fabricated in the GaAs/AlGaAs system and shown to exhibit RT at low temperatures, the inherent design and material parameters limit the minimum dimensions of the device to about 600 Å, below which fringing fields destroy the structure of the quantum well. This thesis investigates the feasibility of using the silicon/silicon-dioxide system as the platform for a PRESTFET which can accomodate finer gate dimensions. Fabrication techniques for a two-level gate, two-level dielectric MOS transistor have been developed which allow very close (50-100 Å) coupling between the tunneling gates and the 2-DEG. Computer simulations and theroretical modeling were also done to assess device performance at various lithographic dimensions. It was concluded that because of the inherently low electron mobility in silicon, strong resonant tunneling effects are not likely to be observed until minimum device dimensions are reduced to around 100 Å.

Thesis Supervisor: Henry I. Smith
Title: Professor

Thesis Supervisor: Dimitri A. Antoniadis
Title: Professor

Acknowledgments

I am extremely grateful to my thesis supervisors, Prof. Henry Smith and Prof. Dimitri Antoniadis, for their enthusiastic support and guidance in this research. Their valuable advise and constant encouragement were most important when I encountered difficulties in fabrication. Prof. Smith has taught me how to tackle real problems in a no-nonsense, step-by-step approach, and his help was critical in making the all-important lift-off process work. Prof. Antoniadis was instrumental in providing me with a fundamental understanding of the silicon MOSFET system and the associated processing technologies. I believe that because of their patient teachings, I have developed the skills and the confidence of a better scientist, and these are the qualities that I will carry with me in all my future endeavours, whether or not they are in the field of science and engineering.

It has been a great pleasure to work in the MIT Integrated Circuits Laborotory (ICL), where most of the fabrication was carried out. I am deeply indebted to the entire staff of the ICL. It's the people behind the machines that make technological breakthroughs come true. Their enormous dedication and enthusiasm are especially endearing, given that their jobs are always exhausting and often thankless. I would like to thank P. Tierney and T. Tyson for their patience and help with my photomask preparation, stepper training, and nitride-etch characterizations. I'd also like to thank J. D. Maria and J. Walsh for their putting up with my sloppy, and sometimes dangerous, e-beam evaporation techniques. J. Walsh also helped me a great deal on the Prometrix measurements. O. Hurtado was also extremely helpful in suggesting important ideas for implementing the wafer-holder and the heat-shielding fixtures for the e-beam. B. Foley and R. Cuikay are greatly appreciated for always being able to do quick implants on short notices. B. Foley also helped me a great deal in my dry-etch characterizations and metal dummy preparations. I'd like to thank J. Bishop for his help in diffusion and LPCVD; I hope is heart condition is better now. Special thanks to N. Polce for figuring out various diffusion and LPCVD recipes, as well as troubleshooting the rapid thermal annealer in the TRL; his interest in the project and

willingness to discuss various processing possibilities are greatly appreciated. I am grateful to P. Burkhart, who has always tried to solve my problems and accomodate my requests with a sense of urgency. I'd like to also thanks Dr. Linus Cordes for providing useful information and approving new processes quickly and painlessly.

I am also grateful to J. Carter, the manager of the NanoStructures Lab, for his patient explanations of the art of e-beam evaporation, and to M. Mondol for his technical support in the NSL. I'd like to also express thanks to T. McClure and R. Perilli in the Microlab for their constant and enthusiastic help. Also, it has been a great pleasure to work with J. Martin on the SIMS analysis; his active interest in the project and lively discussions are deeply appreciated. I'd also like to thank Prof. J. Chung and Prof. R. Reif for their useful advise.

My fellow graduate students have been a wonderful source of both technical and moral support. R. Ghanbari, G. Rittenhouse, M. Burkhardt, J. Jacobs, and C. Eugster have all taught me valuable physics. V. Wong, H. Hu, S. Hector, C. Hsu, D. Sobek, N. Gupta, and W. Chu have all helped me a great deal in my experiments. I want to especially thank my office-mates, M. Burkhardt and N. Gupta, for their moral support.

Finally, this work would not have been possible without the loving support of my parents, who gave up almost everything they had to ensure that I would get a better chance at success in life. Here, I shall dedicate this thesis to my beloved mother and father.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The critical dimensions of electronic devices have undergone enormous down-scaling over the past two decades in a quest for high-speed performance. As device dimensions become comparable to the electron mean-free-path, it is necessary to investigate the effects of quantum interference and diffraction, which dictate the fundamental transport mechanisms. A thorough understanding of quantum-mechanical transport, combined with novel computation architectures and advanced lithographic technologies may some day lead to a revolutionary class of electronics.

The double-barrier resonant-tunneling (DBRT) diode has perhaps received the most attention in the field of quantum-effect electronics in the past decade because of its pronounced negative differential conductance (NDC) and its potential for very high-speed operation. Impressive figures for the state-of-the-art DBRT diodes include peak-to-valley ratios as large as 30 at room temperature [9] and oscillation frequencies as high as 400 GHz [10]. This type of conventional DBRT is based on the vertical tunneling of electrons through a layered, sandwich structure of III-V semiconductor material, grown by molecular-beam epitaxy (MBE). Sharp interfaces and extremely thin layers (down to 15 Å[14]) are mainly responsible for the high performance. However, the inherent two-terminal design prevents it from performing transistor-like operations, which require a third terminal to modulate the barrier heights or the well depth independently of the source-drain bias. Another drawback of the vertical diode is its difficulty in circuit integration, which is most natural on a flat, two-dimensional

plane.

An alternative device, the Planar RESonant Tunneling Field Effect Transistor (PRESTFET) confines the electrons in a plane and induces the tunnel barriers electrostatically via two thin metal gates. This device had been fabricated in the GaAs/AlGaAs MODFET (MODulation-doped Field-Effect Transistor) structure and was shown to exhibit resonant tunneling in the 2-DEG at 4.2 K [18]. However, because of the relatively wide well and barriers ($\sim$ 600 Å), the quantum effects were weak and observable only at low temperatures. In order to further reduce the critical dimensions of the PRESTFET, a new material system, in addition to finer lithography, is needed because the effects of fringing fields in the MODFET system place a practical limit on the gate electrode separation at the thickness of the n-AlGaAs layer, which can be no smaller than 400-500 Å[18].

The aim of this research is to develop a new process, based on the silicon MOSFET, which can be combined with very fine x-ray or e-beam lithography to implement a PRESTFET with critical dimensions much smaller than 600 Å. The basic design is a dual-dielectric, two- level gate structure, in which a large, upper metal gate induces a 2-DEG at the Si-SiO_2 interface and a pair of very fine lower gates creates the double barriers by modulating the potential at the 2-DEG. The lower tunnel gates are separated from the 2-DEG by a very thin, high-quality thermal oxide on the order of 40-100 Å thick. By reducing the effects of fringing, this will allow gate fingers as thin as 40 Å to still have close electrostatic control over the 2-DEG. The second-level dielectric will be 100 Å of silicon nitride, which is used to isolate the upper and lower gates. This layer has to be as thin as possible to let the upper gate overcome the fringing effects of the lower gates.

The most important drawback of the silicon MOSFET system is the low electron mobility relative to the GaAs MODFET. Low mobility implies a high degree of scattering, both elastic and inelastic. Inelastic scattering is detrimental to quantum-effect devices because it destroys the coherence, and hence the wave nature, of the electrons. Elastic scattering due to impurities and lattice disorder is also undesirable, though less detrimental, because it also tends to 'smear out' the energy distribution of

electrons. The special processing techniques required to produce fine structures, such as electron-beam evaporation and x-ray lithography, may damage the 2-D interface and further reduce mobility. Various annealing techniques were performed to restore the mobility, and its low-temperature value were used as a measure of the degree of coherence in the silicon system. Computer simulations were performed to analyze the effects of barrier shape, width, and height, as well as the amount of scattering, on the tunneling current. We shall see that at presently realizable lithographic linewidths ($\sim$ 500 Å), planar resonant tunneling will be difficult to observe in silicon. Resonant tunneling is predicted to occur in silicon if critical dimensions are reduced to about 100 Å.

Chapter 2

Theory of Resonant Tunneling

Tunneling, in which electrons with an incident energy lower than that of a potential barrier actually penetrate the barrier, is a purely quantum- mechanical phenomenon. The wave nature of the electrons manifests itself by producing many counter-intuitive phonomena in the sub-micron regime. Resonant tunneling is perhaps the most obvious and dramatic demonstration of electron interference because it has a clear optical analogy in the Fabry-Pérot resonator. However, the case of electrons tunneling in a solid is much more complicated than that of photons through layers of glass. First of all, electrons experience forces by virtue of their charge: an electrical current can exist only under a bias. Because they also interact with one another, the shape of the potential distribution at resonance, when the quantum well is occupied by many electrons, differs greatly from that off resonance, when the well is occupied by few electrons. Thus the potential will be a function of bias, and can be modeled correctly only by solving *both* Poisson's and Schrödinger's equations self-consistently, a formidable task. For the optical case, since photons do not mutually interact, the 'potential', or dielectric permitivity, is independent of the state of the system (i.e. on or off resonance). Secondly, in a real solid at finite temperatures, electrons suffer many collisions, both elastic and inelastic. Elastic scattering, due to lattice disorder and impurities, can be taken into account by broadening the density-of-states distribution [2], which effectively 'smears' the structure in the transmission function, T(E). Inelastic scattering, due to phonons and electron-electron interactions, breaks

the phase coherence, and hence the wave nature, of the electrons; this makes it difficult to treat electrons in the coherent Fabry-Pérot formalism. In the optical case, since photons generated by a laser typically have a coherence length on the order of 10 centimeters (and up to 100 kilometers [22] if special care is taken), and no phase-breaking occurs in typical materials, the photons can be treated as infinitely coherent for cavity widths much below 10 cm. We see that charged interactions and finite scattering can complicate the dynamics of electrons enormously. Many aspects of resonant tunneling, such as the exact nature of the scattering mechanisms and tunneling times, are still controvertial topics and await experimental study.

2.1 Coherent Resonant Tunneling

2.1.1 Formal Equivalence to Optics

The Fabry-Pérot effect in optics is a well-known phenomenon [22] [33] in which light at select frequencies can penetrate through a slab of material (or air) with highly reflective coatings on both sides. This is an interference effect produced by multiple reflections, and manifests the wave property of light. In standard text-book quantum mechanics, electrons are almost always treated as a completely coherent disturbance whenever *wavefunctions* are used. In this treatment, resonant tunneling is *exactly* analogous to the Fabry-Pérot effect. This formal equivalence can be established by writing out the time-independent equations that describe electrons and electromagnetic waves [11]:

$$\text{Schrödinger(QM): } \frac{\partial^2}{\partial x^2}\Psi(x) + \frac{2m}{\hbar^2}[E - V(x)]\Psi(x) = 0$$
$$\text{Maxwell(EM): } \frac{\partial^2}{\partial x^2}A(x) + \frac{\omega^2}{c^2}\epsilon_r(x)A(x) = 0$$

$\Psi(x)$ is the electron wave function and $A(x)$ can represent a 1D electric field. $\epsilon_r(x)$ is the relative permittivity as a function of space, and is sufficient to describe the cavity and the environment. Energy, E, in the electron case is analogous to frequency, ω, in the optics case; both describe a single time-harmonic eigenfunction. From

the above equations, we see a striking parallel between the two cases. A quantum-mechanical potential barrier can be modeled as a highly reflective surface in optics because $(E - V(x)) < 0$ for a QM barrier corresponds to $\epsilon_r(x) < 0$, which exists in a plasma medium with $\omega < \omega_p$ [23] (e.g. a silvered mirror at optical frequencies). One point worth mentioning is that although the *dispersion relationships* , i.e. ω as a function of k, corresponding to the full, *time-dependent* Schrödinger's and Maxwell's equations are quite different (because they're different orders in time), the single-frequency, time-independent behaviours are very similar.

Given that the two time-independent behaviors are similar, we summarize the Fabry-Pérot results and then apply them to resonant tunneling in the coherent limit. In optics, the cavity is fully described by the individual transmission coefficient,T_1, and the reflection coefficient, R_1, of the mirrors, together with d, the spacing between the mirrors. For simplicity, T_1 and R_1 are taken to be real; i.e. no phase delays caused by the walls. By summing the transmission *amplitudes* of waves which have been reflected n number of times inside the cavity over the index n, the steady-state transmission coefficient T can be easily obtained [33]:

$$T = \frac{1}{1 + F \sin^2(\phi/2)}, \tag{2.1}$$

$\phi = 2kd, F = 4R_1/(1 - R_1)^2$, where ϕ is the phase accumulated in one round-trip traversal in the cavity, and F is a parameter which describes the sharpness of the resonances. The closer R_1 is to 1, the larger F is, and the more *selective* the filter becomes. Equation 2.1 predicts complete transmission, $T = 1$, whenever $kd = n\pi$, or $d = n(\lambda/2)$, for integer n. Resonant transmission corresponds to setting up *standing waves* inside the cavity. At resonance, the incident waves are completely transmitted, and none reflected. This occurs because the large intensity of standing waves built-up inside the cavity 'leaks' out through both ends of the cavity in such a way as to destructively cancel the large reflected component of *incident* waves, and at the same time propogates constructively in the forward direction. In the limit of highly reflective mirrors, $T_1 \ll 1$, the shape of the resonant peak $T(\omega)$ is approximated by a

Lorentzian with

$$\Gamma = FWHM = \Delta\omega \simeq \frac{c}{d}T_1, \tag{2.2}$$

$$\tau = \frac{1}{\Delta\omega} = \frac{d}{c}\frac{1}{T_1}, \tag{2.3}$$

τ is the average lifetime of a photon, equal to the inverse of the escape frequency, $\frac{c}{d}T_1$. We see that the width of the resonance is related to the lifetime by the uncertainty relation $\tau\Delta\omega = 1$.

The exact same analysis applies to resonant tunneling of electrons through a symmetric, double-barrier. However, complications arise because the phase ϕ accumulated by the electron in the well depends on the exact shape of the potential as well as the incident energy; the *magnitudes* of the individual transmission and reflection coefficients (T_1, T_2) also depend strongly on energy. Actual transmission coefficients are obtained numerically, but the Fabry-Pérot analogy is still extremely useful for grasping the fundamental physics. The next section deals with electrons only, emphasizing the differences from optics.

2.1.2 Electron RT

In an ideal quantum device, no inelastic scattering takes place, and electrical transport is completely phase-coherent. To calculate electrical current, one first calculates the transmission coefficient of the structure as a function of carrier energy at a given bias, and apply its Landauer's conductance formula [20] over the entire energy spectrum of incident carriers, which is a function of temperature, voltage, and free transverse momenta [3] [4]. We see that even in the fully-coherent limit, electrical transport is more complicated than light propagation.

The transmission coefficient of a double-barrier structure as a function of energy was treated by Bohm [7] in the WKB approximation and by Kane [5] in the wave-matching formalism. For electron energies below the barrier height, the transmission is very small, except at those discrete energy values which correspond to constructive

interference inside the well. At these energy values, electrons can "tunnel" through the barrier structure with little or no reflection. The maximum transmission is unity for symmetric barriers and $4T_1T_2/(T_1+T_2)^2$ for asymmetric structures [26] , where T_1 and T_2 are the individual barrier transmission probabilities. For simple rectangular barriers under zero bias, the resonant energies satisfy a simple relationship [5]

$$2k_3d = \pi(2n+1) + \tan^{-1}(\kappa_2/k_3) + \tan^{-1}(\kappa_4/k_3) \tag{2.4}$$

where $k_i = \sqrt{\frac{2m^*}{\hbar}(E - V_i)}$ and $\kappa_i = \sqrt{\frac{2m^*}{\hbar}(V_i - E)}$ are the wave vectors in the propagating and decaying regions, respectively. See Figure 2-1.

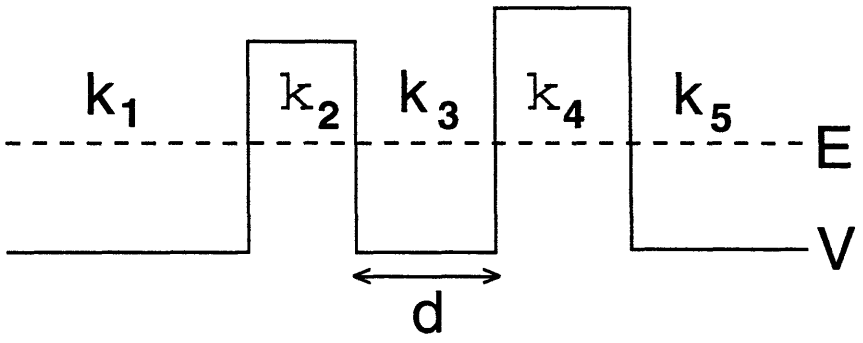

Figure 2-1: Double Rectangular Barriers

This is exactly analogous to the Fabry-Pérot cavity in optics, where light waves at specific wavelengths can travel unimpeded through two highly reflective mirrors. In both cases, at resonance, the waves that have bounced back and forth in the well are in phase with one another and also with those just coming in through the first barrier. Equation 2.4 differs from the simple standing-wave relationship $kd = n\pi$ by an energy-dependent phase factor which accounts for the extra phases picked up from bouncing off each wall once. A more general form of Equation 2.4 for high barriers is [26]:

$$\frac{1}{\hbar}\int 2m[E - V(x)]^{1/2}dx = (n - \gamma)\pi, \tag{2.5}$$

where the integral is calculated throughout the well, n is an integer, and

$\gamma = 0$ for infinite walls on both sides,

$\gamma = \frac{1}{4}$ for one wall and one gradual, continuous edge,

$\gamma = \frac{1}{2}$ for two continuous edges.

From steady-state calculations of $T(E)$, one can actually infer something about the dynamics of tunneling. The transmission probability near a resonant energy can be approximated by a Lorentzian [13]:

$$T = T_{res}\frac{\frac{1}{4}\Gamma_e^2}{(E - E_r)^2 + \frac{1}{4}\Gamma_e^2} \tag{2.6}$$

where E_r is the resonant energy, Γ_e is the full-width at half-maximum (FWHM) of the Lorentzian, and $T_{res} = 4T_1T_2/(T_1 + T_2)^2$ is the peak value. This Lorentzian shape can be obtained by Taylor expansions near E_r of Equation 2.1, and is valid for T_1, T_2 both small. Similar to Equations 2.2 and 2.3 for light, we have the following expressions for electrons:

$$\Gamma_e = \hbar\frac{v(T_1 + T_2)}{2d} \tag{2.7}$$

$$\tau_e = \frac{2d}{v(T_1 + T_2)} \tag{2.8}$$

these equations are exactly analogous to the Fabry-Pérot expressions, except that the speed of light is substituted for the electron velocity. The Heisenberg Uncertainty Relationship,

$$\tau_e\Gamma_e = \Delta t\Delta E = \hbar, \tag{2.9}$$

also holds. The time constant τ_e is the life-time of the metastable state [7] and is well known in the field of nuclear physics [6]; it can also be interpreted as the transient time required to built up to the steady-state transmission behaviour [14]. The subscript e denotes an elastic process. In the Fabry-Pérot example, the transmission coefficient

is a sum of infinitely many wavefronts that were first incident upon the cavity at different times; implicit is the assumption that the light was turned on a 'long' time ago (much larger than τ). From Equation 2.8, we see that the time required to achieve resonance can be extremely long for two thick barriers ($T_1, T_2 \ll 1$), even though the steady-state transmission can be of order one. This places severe requirements on the coherence time (or coherence length) of the incident electron waves in a solid. The fundamental limit on the switching speed of any resonant-tunneling device is determined by τ_e, but usually, circuit parasitics [14] impose lower practical limits.

In contrast with optics, where no forcing is needed for photons to propagate, electrical conduction requires a finite voltage bias, which also distorts the potential profile. See Figure 2-2a. The transmission coefficient as a function of energy differs from the equilibrium case and needs to be recalculated. When the bias is such that one of the resonant energies of the well lines up with the conduction band edge E_c on the left side, the current goes through a maximum becasue electrons at the Fermi level can tunnel into the quasi-eigenstate and then through the structure. See Figure 2-2b. As the bias voltage is increased further, the resonant level drops below the left-hand side E_c and current drops because carriers are prohibited by energy-conservation to tunnel into the quasi-eigenstate. See Figure 2-2c. This is the mechanism of negative differential conductance (NDC).

2.2 Sequential Resonant Tunneling

As emphasized by Capasso [14] , the observation of NDC does not necessarily imply a Fabry-Pérot mechanism. The presence of inelastic scattering, in which electrons lose their phase memory, give another physical picture for RT. M. Büttiker [13] [12] developed a very elegant way of incorporating the effect of a single inelastic scatterer quantum-mechanically by introducing a fictitious junction in the middle of the well, which is capable of transferring an electron inside the well, with probability ϵ, into a fictitious side reservoir which is held at a chemical potential, μ_{side}, such that there is no net current supplied by this side reservoir to the well. The value of μ_{side} is

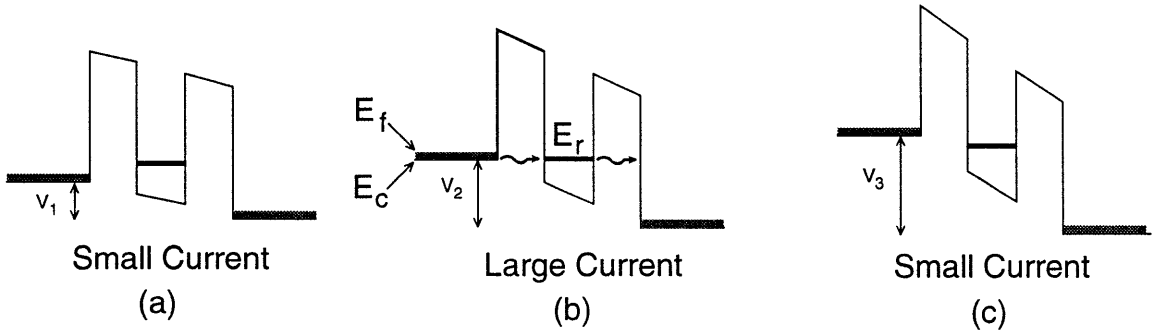

Figure 2-2: Mechanism for Negative Differential Conductance.$V_1 < V_2 < V_3$

a function of the potential difference between the actual contacts on either side of the diode, and the device can be considered two-terminal. Since the current that reemerges from the side reservoir back into the well adds *incoherently* to the current due to electrons that did not suffer scattering, the phase-randomizing function of the side reservoir is performed seamlessly. In this manner, Büttiker separated the total transmission into a coherent part and a sequential part,

$$T_{tot} = T_c + T_i, \tag{2.10}$$

T_c is the probability for an electron to traverse the double barrier without being scattered, whereas T_i is the probability that an electron will traverse the double barrier by going into and out of the side reservoir (inelastic scatterer) at least once. A characteristic energy width is used to describe each physical process. Γ_e, the transmission peak FWHM, is used to describe the elastic process of resonant tunneling; Γ_i, related to the scattering time τ_i by $\Gamma_i = \hbar/\tau_i$, is used to describe the inelastic process of phase-randomization. According to Breit and Wigner [8], the total width Γ_{tot} describing the overall process is a sum of the partial widths due to both elastic and inelastic processes,

$$\Gamma_{tot} = \sum \Gamma_e + \sum \Gamma_i \tag{2.11}$$

This states that after inelastic scattering is introduced, T_{tot} still exibits a Lorentzian-lineshaped resonance, but with the peak transmission *reduced* by the ratio Γ_e/Γ_{tot} and the FWHM *broadened* by the same factor. Breit and Wigner also showed that T_c and T_i, the coherent and sequential components of T_{tot}, are *both* characterized by the same width Γ_{tot}. This result makes a highly non-intuitive statement that the *purely coherent* transmission width is also affected by the degree of inelastic scattering. Büttiker [13] further showed that that ratio of each component of transmission to the total transmission is equal to the ratio of the specific partial width to the total width: $T_c/T_{tot} = \Gamma_e/\Gamma_{tot}$,and $T_i/T_{tot} = \Gamma_i/\Gamma_{tot}$. Given these results, we can summarize the effect of increasing the degree of inelastic scattering (as $\epsilon \rightarrow 1$):

1. The total transmission, $T(E)$, is flattened and broadened gradually until it becomes flat and independent of energy.

2. The sequential component of the tunneling current becomes a larger fraction of the total current until *all* current is accounted for by sequential tunneling.

In the completely incoherent limit, i.e. as the scattering probability inside the well goes to one, the transmission through the double barrier structure is insensitive to special geometrical arrangements , such as the width of the well, between the two barriers. This is because the electron loses its phase memory in the well and is thus incapable of 'sensing' the structure of the well. In this limit, $T_{tot} = T_1T_2/(T_1 + T_2)$, corresponding to adding the individual quantum resistances ($\sim 1/T$) of the two barriers [12], and no interesting current-voltage structures remain.

Since the relative magnitudes of the partial widths Γ_e and Γ_i determine the quantum nature of the system, we must try to maximize Γ_e and minimize Γ_i, according to the results given above. Maximizing Γ_e implies increasing the individual barrier transmission T_1, T_2. This may be done by either reducing barrier width or lowering barrier height; it will be shown later that the latter is unfavorable. Minimizing Γ_i

implies maximizing the inelastic scattering time τ_i, and this means that the material quality must be high; this is the reason why mobility is a critical issue for RT devices. For our purposes, the momentum relaxation time [30],

$$\tau_p = \frac{m^* \mu_n}{q} \tag{2.12}$$

is used to give us the ball-park figure for the inelastic scattering time τ_i, and the Heisenberg Relation (Eq. 2.9) is used to find the inelastic width Γ_i. In Eq. 2.12, τ_p is the momentum relaxation time, m^* the electron effective-mass, μ_n the electron mobility (usually measured), and q the electronic charge.

2.3 Distinction between Coherent and Sequential Resonant Tunneling

It is important to understand which tunneling mechanism, whether coherent or sequential, dominates the physical system at hand in order to predict whether interesting non-linearities, such as resonant tunneling, will be observed. By comparing the magnitudes of three time scales of the system, Büttiker classified the current-conduction mechanism into three regimes: Coherent RT, Sequential RT, and Incoherent tunneling.

The three time scales are : the round-trip travel time in the well, the intrinsic transient response time of the resonance, and the inelastic scattering time in the well. Let w be the width of the well and v the group velocity of the carriers in the well, then $\nu = v/2w$ is the attempt frequency of tunneling, and $\tau_\nu = 1/\nu$ is the time required for a single round-trip travel in the well. If T_1, T_2 are the individual tunneling probabilities of each barrier, then $\tau_e = \frac{1}{\nu(T_1+T_2)}$ is the time required to build up the resonant electron density inside the well; $\Gamma_e = \hbar/\tau_e$ is the intrinsic resonant width . The inelastic events are characterized by a scattering time τ_i.

Coherent RT is dominant in the regime $\tau_i \gg \tau_e \gg \tau_\nu$ since the electron density can achieve its resonant, steady-state value before inelastic scatering occurs. This case is

rarely observed experimentally, as evidenced by the much lower peak-to-valley ratios in actual devices than coherent theory would predict. Sequential RT dominates in the regime $\tau_e \gg \tau_i \gg \tau_\nu$, where carriers are scattered after many round-trip traversals in the well, but before complete constructive interference is achieved. Most of the experimental RT observed point to this regime, where the peak-to-valley ratio is of the order 1-30. Complete incoherent transport occurs for the third regime, $\tau_e \gg \tau_\nu \gg \tau_i$, where carriers lose phase memory before traversing the well. The transmission in this case is $T_1T_2/(T_1 + T_2)$ for all energies, and no resonant enhancement is seen at the discrete levels. The quasi- eigenstates do not exhibit themselves at all because all information about the geometric arrangements of the well is destroyed by the inelastic scattering [13] . In the design of a RT device, we need to be sure that at least the sequential RT criteria are satisfied.

2.4 Other Broadening Mechanisms

Inelastic scattering is perhaps the most detrimental mechanism in smearing out the quantum effects of RT because it destroys the phase memory of the electrons. Other non-idealities, such as elastic scattering due to random disorder in the solid, finite temperature and source-to-drain voltage, and free motion in transverse directions, further weaken RT by effectively "broadening" the carrier energy spectrum, which in the ideal case would be peaked at the Fermi level, E_f. Bagwell [2] has developed an elegant method to account for these effects by convolving the transmission coefficient function with the energy spectra of the various broadening mechanisms.

Finite temperature enables carriers within an energy range of $3.5k_BT$, centered at E_f, to contribute to conduction. A finite source-drain voltage, V_{DS}, allows electrons within a energy range of eV_{DS} on the source side to contribute to conduction. Elastic disorder scattering broadens the resonant levels by $\hbar/\tau$, where τ is the elastic scattering, or momentum relaxation, time. The inclusion of free motion in each transverse direction can be done by convolution with the 1D density of states which has a $1/\sqrt{E}$ dependence.

Chapter 3

Previous Work

3.1 Vertical RT diodes

Because of the ability of molecular-beam epitaxy (MBE) to grow very thin(down to several monolayers), high-quality materials and achieve abrupt interfaces, almost all the exprimental RT work has been based on the vertical structure in the AlGaAs system, shown in Figure 3-1. The barriers are formed by the layers containing high concentrations of aluminum, which corresponds to a wider band-gap.

In a typical vertical RT diode, with 3.0 nm thick barriers, a 4.5 nm wide well, and a barrier height of 300 meV, caculations [14] for the zero-bias case gives a single transmission peak at ~100 meV with a FWHM of ~4 meV. For coherent RT to be present, the broadening due to inelastic scattering must be small compared to the transmission FWHM. For the state-of-the-art AlGaAs/GaAs heterojunctions, the electron mobility is ~7000 cm^2/Vs at 300 K. This implies a scattering time $\tau_i \approx 0.3$ ps, which corresponds to a broadening of ~2 meV. Since this is of the order of the intrinsic resonant width, we expect the tunnel current to contain both coherent and sequential contributions. The experimental peak-to-valley ratio of 1.3 for this structure [14] seems to suggest that the sequential RT is dominant.

More recently, improved peak-to-valley current ratios were obtained in the InGaAs / InAlAs materials system. By replacing the InAlAs barriers with strained layer AlAs and incorporating InAs into the well structure, Broekaert [9] has achieved a P/V ratio

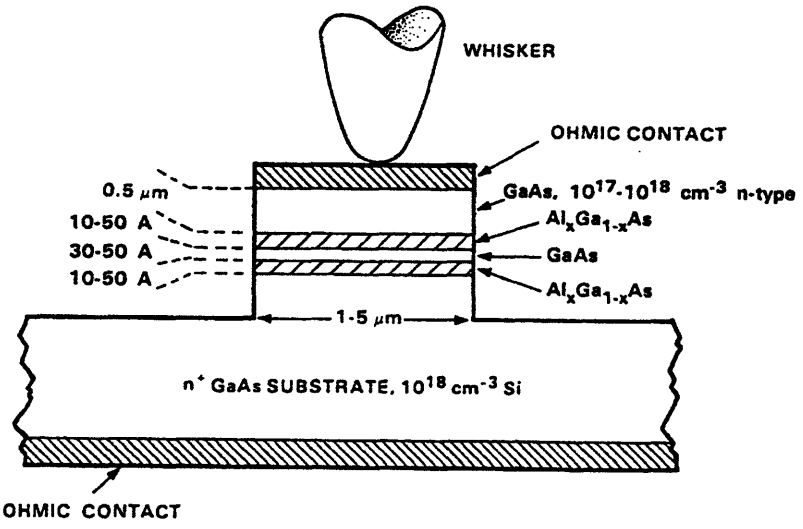

Figure 3-1: The Vertical DBRT Diode

of 30 at room temperature. The use of a binary compound (InAs) instead of a ternary compound (InGaAs) in the well eliminates alloy scattering and is likely to account for the enhanced quantum effect. This demonstrates the importance of scattering in the RT phenomenon.

3.2 The GaAs PRESTFET

Ismail [18] has fabricated a planar RT field-effect transistor (PRESTFET) based on the GaAs MODFET structure, as shown in cross-section in Figure 3-2.

The GaAs/AlGaAs layers were grown by MBE, and the tunnel gates were defined by direct-write electron-beam lithography and then lifted-off. The highest resolution device had gate fingers of width $\sim$ 600 Å separated by $\sim$ 600 Å (pitch=1200 Å). The advantages of this structure compared to the vertical diode is three-fold. First, the height of each barrier can be independently controlled by varying the gate bias, making "transistor" action possible. Secondly, the degrees of freedom in the direction perpendicular to transport is reduced to 1, compared to 2 in vertical diodes; this

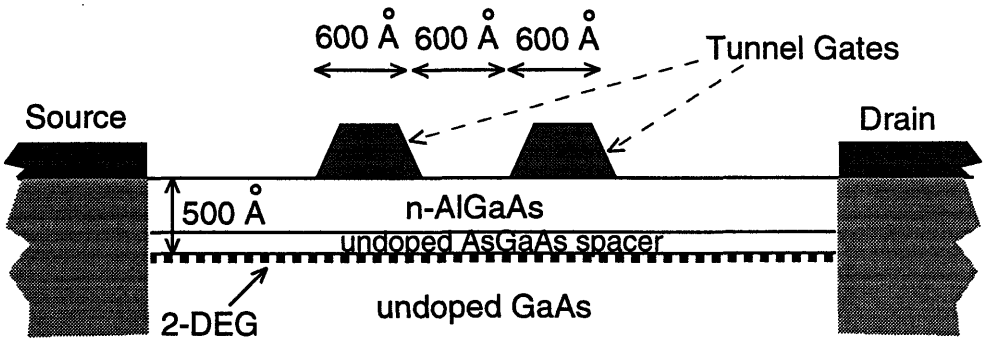

Figure 3-2: The GaAs PRESTFET

reduces dimensional broadening. Finally, the mean-free-path and coherence length can be much greater than in vertical diodes.

The major limitation of planar RT devices is the size of the tunnel gates, which are defined lithographically and are at least an order of magnitude larger than the critical dimensions in vertical diodes. Since the intrinsic resonant width, Γ_e, is directly proportional the tunnel probability through the individual barriers, and since tunnel probability decreases exponentially with increasing barrier thickness, wide barriers imply extremely sharp and narrow transmission peaks, which are easily washed out by small amounts of inelastic scattering. This dictates low-temperature operation and high mobilities. In addition, low-temperature (<10 K) operation also serves to prevent thermal smearing among levels since the wider well reduces the energy-level separations to 2-5 meV. Ismail's GaAs samples had a maximum mobiliity of 250,000-400,000 cm^2/Vs at 4.2K, which corresponds to a scattering width of 0.4-0.7 meV. The intrinsic transmission widths of the lowest resonant energies for the 1200 Å-pitch device are likely to be orders of magnitude smaller than the scattering width, and thus no RT is expected at those energies. However, it is possible that one of the quasi- eigenstates resides very close to the top of the potential barrier, where transmission is large due to the reduced barrier thicknesses and the high incident carrier energy. This is probably the explanation for Ismail's observation of negative

differential transconductance (NDT) for a GaAs PRESTFET biased below threshold [18]. See Figure 3-3

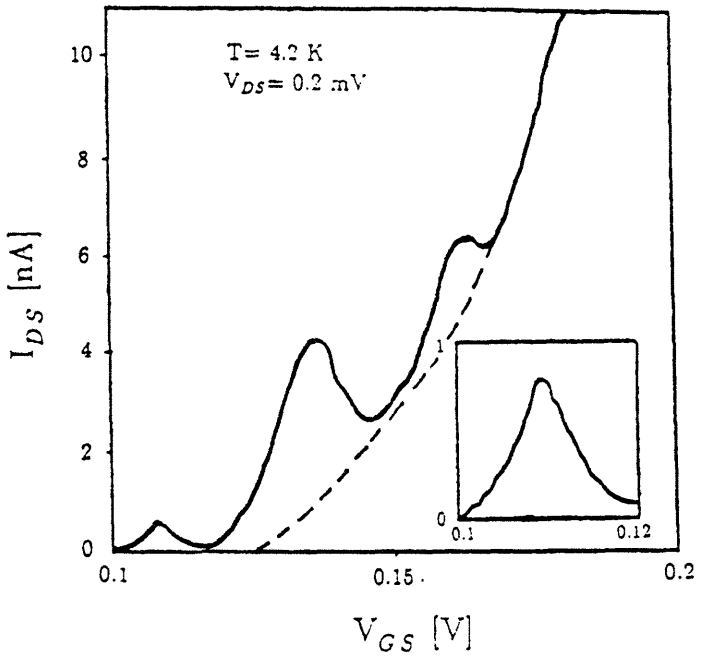

Figure 3-3: The first peak shows Negative Differential Transconductance below threshold

Ismail's device showed RT directly in the current measurements when the source-drain voltage (V_{DS}) is held small and constant, and the gates (V_{GS}) were swept together. On the other hand, when V_{GS} is fixed (below threshold) and V_{DS} is swept, RT can be seen only in the differential output conductance (g_d). This is because for large V_{DS}, the width of the carrier energy distribution becomes comparable to the level spacing, and this smears out the RT. Another possible, parallel mechanism is that a large V_{DS} breaks the symmetry between the barriers and lowers the peak transmission probability at resonance.

Chapter 4

The Proposed Silicon PRESTFET

It is clear from the previous work on planar RT devices that the major thrust has to be in the reduction of the barrier and the well widths in order to broaden the intrinsic resonant widths beyond the scattering rate and to increase the level spacing of the quasi- eigenstates. Ismail's work on the GaAs PRESTFET is close to the practical limit of the MODFET structure. The need of a high-mobility 2-DEG with an electron concentration of $1.5-5\times10^{11}$ cm^{-2} at the heterointerface places stringent requirements on the doping and thickness of the n-AlGaAs layer and the undoped AlGaAs spacer layer. The optimum combination was found [18] to be 420 Å of n^+AlGaAs, silicon-doped to 1×10^{18} cm^{-3}, on top of a 75 Å-thick spacer layer. Thus, the tunnel gates were spaced about 500 Å away from the actual 2-DEG. Because of fringing-field effects, the minimum separation between the tunnel gates is about this value, 500 Å.

4.1 Device Structure

With improved lithography, a new materials system needs to be considered which can minimize the effects of fringing. Here, we propose a planar structure based on the silicon MOSFET. See Figure 4-1.

This new device incorporates two levels of thin dielectric insulators as well as two levels of metal gates. A large top gate is biased to populate a 2-DEG at the Si-SiO_2

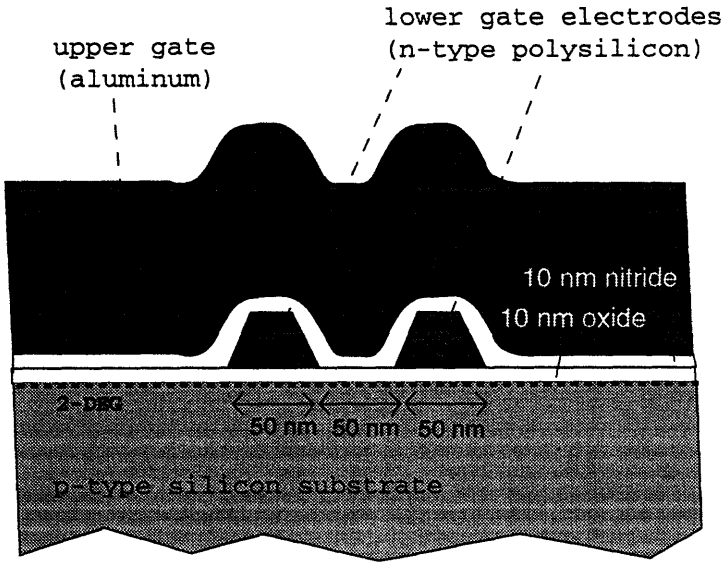

Figure 4-1: The Silicon PRESTFET

interface, while a pair of thin lower gates are biased to raise the potential locally to form the barriers. The potential variation can be tightly controlled because both dielectrics (SiO_2 and Si_3N_4) can be extremely thin (40-100 Å). This makes practical the lithographic formation of fine gate structures with critical dimensions as small as the oxide and nitride thicknesses. The single most important advantage of this device is our ability to grow a very thin, defect-free oxide on the silicon, which allows close coupling of tunnel electrodes to the 2-DEG.

4.2 Observability of RT in silicon: A Comparison with GaAs

The main disadvantage of making quantum effect devices on silicon is the low electron mobility in the 2-DEG, which is typically 600-1000 cm^2 /Vs at 300K and can be 10,000 cm^2/Vs at 4.2K. This is more than an order of magnitude lower than that in high-

quality GaAs MODFET structures, as seen in Ismail's previous work. Mobility is a critical concern at the 600 Å linewidth, achieved by Ismail. This is evidenced by the fact that the GaAs PRESTFET's exposed by focused-ion-beam (FIB) lithography failed to display even negative differential transconductance because the mobility was lowered seven times, relative to the e-beam exposed devices, by ion damage [18]. We shall make some crude comparisons of the silicon and GaAs devices, using the three time scales proposed by Büttiker, to estimate the observability of RT in the proposed silicon device.

Many approximations are made since the exact barrier height and the energy level that corresponds to the observed tunneling peak are not known for Ismail's GaAs PRESTFET. See Figure 3-3. However, it is fairly certain that the resonance observed below threshold occured with E_f very close to the top of the barrier since the transmission would otherwise be vanishingly small due to the thick barriers. See Figure 4-2.

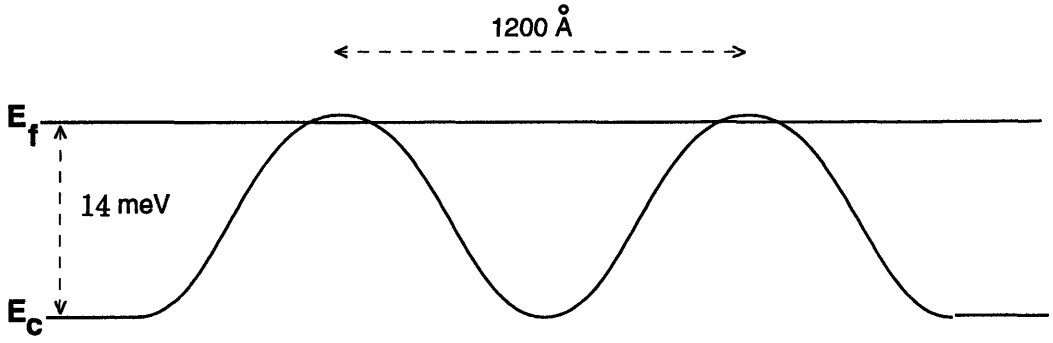

Figure 4-2: Hypothetical Potential Profile and E_f at Resonance

The height of the barriers at resonance is approximately 14 meV, since this is the difference $(E_f - E_c)$ corresponding to a carrier density of $\sim 2 \times 10^{11}$ cm^{-}2 at the heterointerface. By making further assumptions that a silicon PRESTFET with a similar tunnel gate structure (pitch=1200 Å) achieves resonance with the same potential profile and E_f (though the ideal E_f was found to be $\sim$5 meV due to mobility

considerations), we can calculate an approximate traversal time, τ_ν, for both materials. Also, from typical mobility figures, we can estimate the scattering times, τ_i. The value of τ_e is difficult to estimate because it depends exponentially on the *exact* position of E_f relative to the barrier maximum. However, it is reasonable to expect $\tau_e \gg \tau_i$ for both materials, since we are dealing with wide barriers. See Table 4.1.

	μ_n (cm^2/Vs)	τ_i (psec)	τ_ν (psec)	τ_e (psec)
GaAs	400,000	15	0.87	$\gg \tau_i, \tau_\nu$
Si	10,000	1.1	1.5	$\gg \tau_i, \tau_\nu$

Table 4.1: Estimated Time Scales for the GaAS and Si PRESTFET's with 50 meV High Barriers and 1200 Å pitch, @ 4.2 K.

By comparing τ_i with τ_ν, we see that the GaAs PRESTFET operates in the sequential RT regime ($\tau_e \gg \tau_i \gg \tau_\nu$). The situation is less obvious for silicon, in which $\tau_i \sim \tau_\nu$. For this case, the magnitude of τ_e will determine whether RT is observable. But in comparison with the GaAs case, RT in the silicon device is expected to be much weaker. Significant reduction in linewidth below 600 Å may be required to achieve strong RT.

The main experimental aspects of this research is to develop the materials system in the silicon environment, which can be built upon with ultra-fine x-ray or e-beam lithography to fabricate the silicon PRESTFET. The resulting technology may also be applied to fabricate other silicon devices requiring small critical dimensions, such as the short-channel MOSFET.

Chapter 5

Technology Development and Test Results

The bulk of this thesis involves the development of a robust fabrication sequence and the characterization of materials which are suitable for the fabrication of the proposed 500 Å minumum linewidth silicon PRESTFET. The experimental work focused on the deposition and activation of the all-important lower gate material – amorphous silicon, and on the control of the radiation damage introduced to the 2-DEG by high-energy electrons / photons during fabrication. Because of time constraints, actual tunnel gate structures and nanolithography were not implemented. Ordinary MOS transistors with large gate dimensions (20 μm wide $\times$ 16 μm long) were fabricated to test the fabrication sequence.

By carefully controlling the fabrication environment and deposition parameters, we were able to successfully e-beam evaporate, implant, lift-off and activate a 1500 Å-thick layer of amorphous silicon, a process which allows the definition of very fine tunnel gate structures with a refractive material (thinner layers required for finer structures). It was found that diffusion and activation of As dopant atoms in the e-beam evaporated Si films were not possible at moderate temperatures (900-950 °C). Rapid thermal annealing above 1050 ° C was necessary to redistribute and activate the implanted dopants. We were also able to control the radiation damage imparted onto the test transistors during e-beam evaporation with an additional 900 °C furnace

anneal in N_2. The resulting transistor characteristics showed a full restoration of room temperature mobility to $\sim$ 700 cm^2/Vs and also displayed a maximum low temperature differential mobility of 10,000 - 12,000 cm^2/Vs at 4 Kelvin. The low temperature mobility obtained here is similar to the values reported by J. S. Thomas [28] on quasi-1D MOSFETS which allowed the observation of universal conductance fluctuations at 300 mK. This suggests that the fabrication sequence developed is suitable for high-mobility quantum device research.

5.1 Test Device Design and Fabrication Sequence

The test process is a two-level-gate, two-level dielectric process similar to the actual PRESFET design, the only difference being that only large, conventional transistors were laid out and that the lower gate material is 3 times as thick as the proposed PRESTFET. The actual thickness of the lower, amorphous silicon (a-Si) gate deposited was 1500 Å instead of the proposed 500 Å because the low energy($\sim$ 10 KeV), high-mass Antimony was not available on-site as an implantation source; instead, 20 KeV Arsenic was used, which has a range of 300 Å and straggle of 300 Å in amorphous silicon, according to SUPREM simulations. The substrates used were lightly doped (10-20 Ω-cm) p-type silicon wafers, corresponding to a doping of $10^{15}/cm^3$. The lower gate dielectric was 100 Å of thermal SiO_2 grown at 900 °C, which is high enough to form a high-quality Si-SiO_2 interface. E-beam evaporation and lift-off were used to deposit the a-Si lower gate on the gate oxide. The second, upper dielectric film used was 100 Å of LPCVD silicon nitride (Si_3N_4), deposited at a modest 800 °C. The top gate was formed by thermal evaporation and wet-etching of aluminum; sputtering and plasma-etching were avoided in order to prevent further radiation damage in the plasma environment.

Optical masks for photolithographic patterning were laid out, which contained conventional, single-gate, long channel transistors with gate dimensions 20 μm (width) $\times$ 16 μm (length). Two distinct types of transistors were patterned: one which is gated only by the a-Si sitting on gate-oxide (henceforth called polysilicon-gate tran-

sistors), and the other type gated only by the aluminum metal sitting on top of gate nitride and gate oxide, without the lower gate in between the dielectrics (henceforth called aluminum-gate transistors). Capacitors utilizing both gate levels separately, and Van der Pauw test structures were also included as test structures. The electron mobility was measured as the transport (not Hall) mobility of the transistors. The polysilicon (after anneal of a-Si) resistivity, and the oxide, nitride film thicknesses and capacitances were measured from the test structures.

The basic fabrication sequence is very similar to the standard NMOS technology developed at Integrated Circuits Laborotory here at MIT [32]. Notable variations from it include:

1. For the lift-off of the lower gates, photoresist is directly spun and patterned onto the gate oxide before the e-beam evaporation of silicon. This leaves the critical gate oxide exposed to the environment, and extreme care must be taken so that organic and metallic contaminants do not adhere to it.

2.Also, because pure aluminum (thermally evaporated) was used as contact metal instead of $AlSi_{1\%}$, spiking in the source/drain contact areas is a problem that needs special attention. The solution was to do a shallow arsenic implant which define the actual channel length and a deep phosphorous implant which just surrounds the contact hole regions. Since phosphrous diffuses much more rapidly than arsenic at high temperatures, we were able to get a junction depth of ~ 0.4 μm for phosphorous implants, enough to prevent spiking, while maintaining a shallow, 0.25 μm junction for arsenic, which is important in minimizing short-channel effects.

3. Another complication arises because the plasma environment, SF_6, generally used to etch nitride films, aslo attack the underlying polysilicon layer. The solution was to deposit a LTO (low temperature oxide) hard mask, pattern it and wet-etch the nitride in hot phosphoric acid.

The fabrication took place in four facilities here at MIT: the Integrated Circuits Lab (ICL), the Technology Research Lab (TRL), the NanoStructures Lab (NSL), and the Microlab. The final, optimized process is summarized below:

- Selective P+ field-implantation and field-oxidation using LOCOS.
- Dummy gate oxide growth, $\sim$ 100 Å.
- Photolithography and shallow source/drain N+ implants (Arsenic, 90 KeV).
- Photolithography and deep source/drain N+ implants (Phosphorous, 90 KeV).
- Dummy Gate wet etch.
- 100 Å Gate Oxide growth, 900 °C dry O_2.
- Lower gate pattern on 1 μm-thick photoresist using contact lithography.
- UV-Ozone clean to remove organic contaminants on the gate oxide.
- Electron-beam evaporation of amorphous silicon, $\sim$1500 Å. Chamber pressure $< 10^{-7}$ torr. Heat shield fixture used.
- Low-energy, $\sim$20 KeV, Arsenic implant into amorphous silicon, dose 3-5$\times 10^{15}/cm^2$.
- Acetone lift-off of resist.
- RCA clean, 100 Å LPCVD Si_3N_4 deposition at 800 °C.
- Rapid thermal anneal at 1050-1080 °C to activate gate implants, $\sim$ 20 seconds.
- Regular 900 °C N_2 furnace anneal to reduce structural damage in the 2-DEG interface, $\sim$ 1 hour.
- 1000 Å Low-Temperature-Oxide (LTO) Deposition, 400 °C.
- Photolithography of Contact Cuts.
- Buffered-Oxide-Etch (BOE) contact holes in LTO to form hard mask.
- Remove resist, wet etch the nitride in hot phosphoric acid with LTO as hard mask.
- BOE wet etch of LTO hard mask and source/drain contact area gate oxide.
- Piranha Clean. Thermal evaporation of aluminum. (Upper gate metal)
- Pattern and wet etch aluminum in PAN acid at 40 °C.
- Sinter in N_2 tube at $\sim$ 425 °C, $\sim$ 15 minutes.

5.2 Lower Gate Metallization

The critical step in fabricating the silicon PRESTFET is the definition of the lower tunnel gates, for which extremely narrow linewidths, on the order of a few hundred

Angstroms, are required. This section presents in more detail the development and characterization of the amorphous silicon process.

5.2.1 Choice of Gate Material

The choice of a suitable material to be used for the lower gates is critical because it must withstand subsequent high-temperature processing as well as provide etch resistance to the gas or chemical used to etch contact holes to the lower gates. In a Si-MOSFET type quantum-effect device fabricated by Thomas [28], which also employed a two-level gate system, tungsten (specifically, 50 Å Cr - 150 Å W - 75 Å Cr) was chosen because of its high melting point and reasonable conductivity. However, this material was very difficult to work with for several reasons: (1) Tungsten tends to oxidize when heated to a high temperature, such as during annealing. (2) Tungsten layers thicker than ~200 Å are destroyed by stress build-up from the anneal. (3) Thin layers of tungsten have pinholes and may fail to act as an etch stop to the hydrofluoric acid (HF) used to etch through the upper-level oxide to the tungsten lower gate.

To overcome the above problems, we chose silicon as a candidate material for the tunnel gates. After deposition, it can be implanted with low-energy arsenic or antimony and then annealed to form a conducting film. The resulting polysilicon could withstand a wide range of thermal cycling and also was etch-resistant to a variety of chemical etchants.

5.2.2 Evaporation and Lift-Off

Lift-off [31] [16] is the technique chosen for lower gate metallization because of its ability to define narrow metal lines in an additive manner. In this process, high-resolution resist, such as PMMA, is coated directly on the gate oxide, then exposed either with x-ray or e-beam lithography and developed. The metal is then e-beam evaporated onto the sample; when the resist is finally removed, the metal remains where the resist had been developed away earlier. In an alternative, subtractive process, where the metal is first deposited over the entire oxide surface and then

etched away selectively after resist has been spun on top and patterned, the lateral over-etching (undercut) severely limits the resolution [31]. Also, in a reactive-ion-etcher (RIE), it is difficult to stop the etch at the underlying thin oxide because of inadequate selectivity.

In our experiment, amorphous silicon (melting point = 1420 °C) was evaporated from a carbon (graphite) crucible onto the substrate via electron beam heating in a high-vacuum environment ($\leq 10^{-7}$ torr); the requirement on high vacuum is not very stringent, but the better the vacuum is, the lower the final resistivity is for the polysilicon after annealing. Contaminants, notably oxygen (see Figures 5-2, 5-3, 5-4), are introduced from the chamber ambient to the film and may form complexes with implanted ions, preventing the ions from being activated during anneal. The beam spot was focused to its minimum area, roughly $(5\ \text{mm})^2$, in order to best approximate a point source for easy lift-off. A special stainless arm-like fixture was built to hold a 4-inch wafer directly above the crucible. The distance from the source to the target wafer was roughly 18 inches, making the point-source approximation reasonable.

In order to minimize resist heating and edge-rounding during the initial melting of the silicon source, a second stainless-steel shutter was built and affixed to the original shutter such that it is located about 1 inch below the substrate and would swing in and out along with the original shutter. Resist heating is detrimental to lift-off because a rounded resist profile would create a continous coverage of the substrate by the evaporant, making lift-off impossible [31].

It is also important for the lithography to generate sharp resist profiles to begin with. Contact photolithography was used with a 1-to-1 chrome mask and 400 nm UV light to expose the resist. With 1 μm-thick resist and the special low resist-heating arrangement, a 1500 Å thick film of amorphous silicon could be easily lifted-off, after evaporation and ion-implantation, in hot (60 °C) acetone without ultrasound agitation.

5.2.3 Dopant Activation and Contamination Control

The implanted dopants (As) were activated and redistributed only by a high temperature rapid thermal anneal at 1050-1080 °C (temperatures above 1050 °C are required to RTA-activate dopants [19]). An ordinary 920 °C furnace anneal for 1 hour produced a very high-resistivity film and also failed to redistribute the dopants.

The distribution of the implanted arsenic atoms and of the vaious contaminants were measured via the SIMS technique, using either rare isotopes (in the case of carbon, arsenic, and silicon) or molecular species (in the case of oxygen) to identify each element. In the following SIMS plots, the count rates of various species are given as a function of depth from the surface of the sample, the 1500 Å-thick amorphous silicon film. These rates can be related to relative abundances once the detector sensitivity of the instrument to those species is calibrated and the natural abundances of the isotopic/molecular forms are known. But for our present purpose of tracking the impurity distribution, this is not necessary.

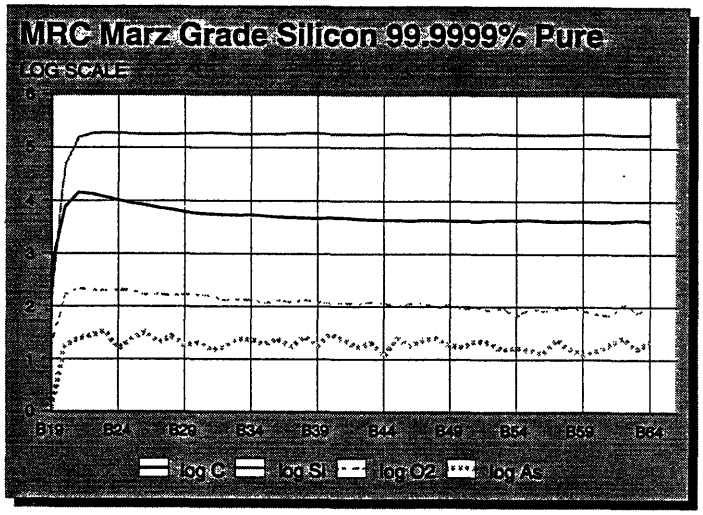

Figure 5-1: SIMS data for target material

Figure 5-1 shows the SIMS count rates for the pure silicon taget (MRC MARZ-grade, 99.9999% pure) *before* evaporation and implant. The matrix atoms (Si) as well as three types of impurities (O,C,As) are monitored. Once again, we emphasize that even though the carbon count is higher than the oxygen count, it does not

necessarily imply that there are more carbon atoms than oxygen atoms in the material because the count rates are for selected isotopic/molecular species and that machine sensitivity was not calibrated. It merely serves as a template against which subsequent measurements are compared.

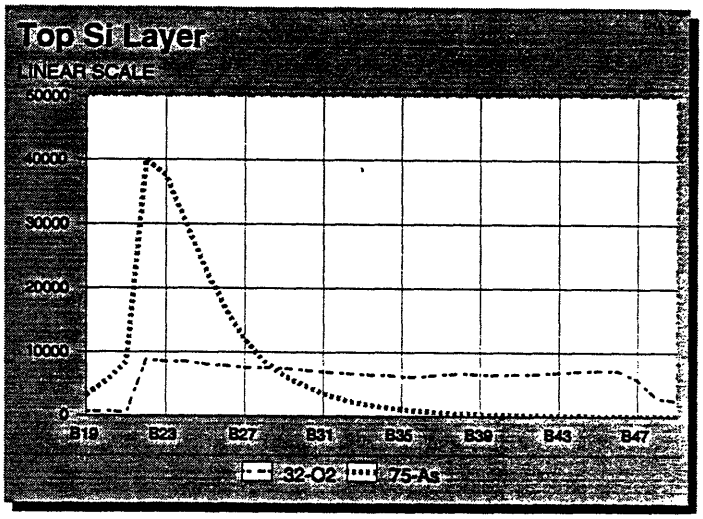

Figure 5-2: SIMS data for post-implant, pre-activation film

Figure 5-2 shows the SIMS data for a 1500 Å-thick, evaporated a-Si film after a 20 KeV, $3^{15}/cm^2$ arsenic implant. The arsenic distribution peaks at about 300 Å from the surface and resembles the Gaussian shape typically expected for implant profiles. The e-beam evaporation process, carried out at $\sim 10^{-6}$ torr, was seen to introduce significant contamination; the carbon content is 10 times higher, and the oxygen content 50-100 times higher than the pure, starting material.

Figure 5-3 shows the same data for the implanted film after a one-hour, 920 °C furnace anneal in N_2. The film was capped by a 100 Å-thick layer of LPCVD nitride to prevent the oxidation of silicon and out-diffusion of arsenic. We see that no significant redistribution of the implanted arsenic took place. Sheet resistivity measurements gave results above the megohms range. This suggests that the arsenic atoms may have formed large complexes with the impurity atoms (probably oxygen) which prevented their movement and incorporation into the substitutional sites in the poly-silicon matrix.

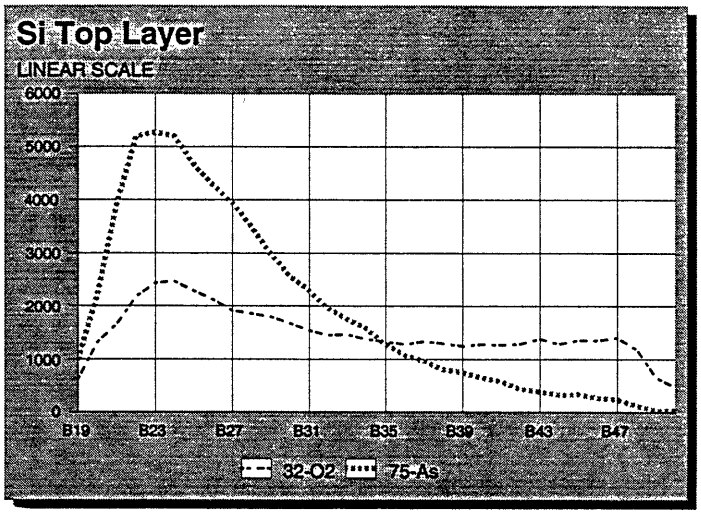

Figure 5-3: SIMS data for 920 °C, 1-hour activation

Figure 5-4 shows the SIMS data for the implanted film after a high-temperature rapid thermal anneal, with nitride capping. The temperature was held between 1050 and 1080 °C for ~20 seconds and stayed near the peak 1080 °C for ~10 seconds. We see that the implanted arsenic atoms are now completely redistributed throughout the thickness of the film. Resistivity measurements yield 8 kΩ/□, which is a reasonably low value for use as a gate electrode.

The fact that a 900 °C, 1 hour furnace anneal is adequate for activating the implanted arsenic in a LPCVD deposited polysilicon film [19] [16] but not in the e-beam deposited film suggests that contamination, probably by oxygen, introduced during the evaporation process is the key problem. It's possible that the high-temperature RTA anneal may have enabled the diffusion and incorporation of the dopants in the silicon matrix by providing enough thermal activation energy to break up the arsenic-contaminant complexes. By lowering the chamber pressure to 10^{-7} torr during evaporation, the sheet resistance of the final, RTA-annealed film was lowered to ~800 Ω/□ (for an As implant dose of $5 \times 10^{15}/\text{cm}^2$).

For the RTA anneal to be successful, the evaporated a-Si film must be capped by a dielectric film (100 Å nitride used in our case) to prevent oxidation and degradation of the a-Si by the ambient, since extremely high temperaturs are reached in a chamber

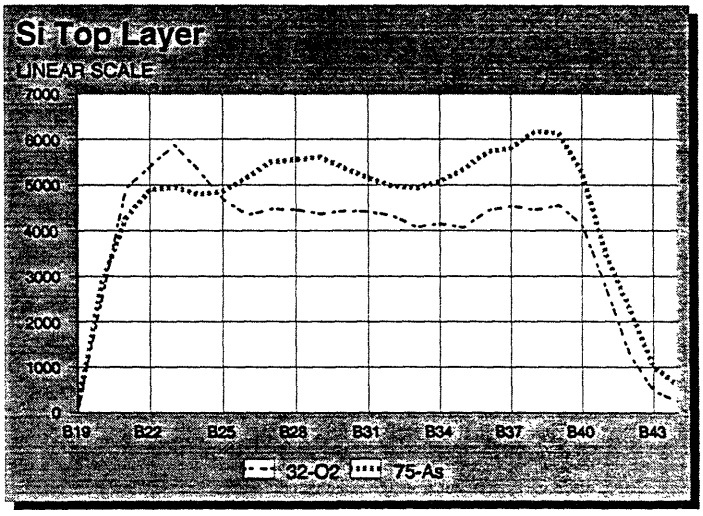

Figure 5-4: SIMS data for RTA activation, 1080 °C peak temperature.

which is not pumped down.

If the above procedures are carried out carefully, the resulting silicon film is polycrystralline, has a low sheet resistance, and is immune to wet etchants such as BOE and hot phosphoric acid.

5.3 Damage and Mobility Concerns

Various processing steps involving ionizing radiation cause damage to the Si-SiO_2 interface by creating oxide charges and interface states [21]. The electron-beam evaporation [17] of the amorphous silicon, the x-ray or e-beam lithography [24], and the various plasma sputtering or etching [21] of films are such examples. These processes significantly lower the electron mobility, the most critical parameter that determines the observability of quantum effects.

In order to maximize the mobility of the devices, we tried to avoid ionizing environments where possible, and to anneal out the damage caused by the unavoidable fabrication processes. We expect the short, high-temperature RTA activation of the lower gate implants to also serve the additional function of partially annealing out the damage introduced during e-beam evaporation and x-ray (or e-beam) lithography.

However, we found that this was inadequate, and that a low-temperature, dedicated damage anneal was necessary to further improve the mobility of the test transistors. To avoid possible damage introduced by the final deposition and etching of the metal contacts, the final aluminum metallization for the upper gate was done by thermal evaporation instead of sputtering, and the aluminum was etched in a wet process instead of in a plasma.

In our particular experiments, since fine lithography was not done, the damage is entirely due to the electron-beam evaporation, during which electrons accelerated to 10 KeV strike a target material, producing back-scattered electrons and X-rays which impinge on the substrate and produce damage in the form of lattice dislocations and oxide charges. In an attempt to anneal out this damage, a 900 °C, 1-hour furnace anneal in N_2 was done immediately after the RTA dopant activation. The resulting transistor characteristics are compared to a batch which did not receive this extra damage anneal. Device data provided in the next section shows that this additional damage anneal can improve the electron room-temperature mobility by ~10% and low-temperature mobility by ~25%.

5.4 Device Test Results

5.4.1 Room Temperature Data

The final MOSFET and capacitor characteristics for both the damage-annealed and the un-annealed batches are quite well-behaved. For both the poly-gate and the aluminum-gate transistors, the gate-substrate and the subthreshold leakage currents are in the picoamps range. The subthreshold behavior is linear and drops off at ~65 mA/decade for all transistors. The threshold voltages (V_T) for both types of transistors are in the range 0-150 mV, close to the theoretical values of roughly 0 V. The only deviation is that the poly-gate capacitors yield capacitance values that are 10-15% lower than expected. This suggests that there may be incomplete activation of dopants near the poly-Si / gate-oxide interface. Also, the mobility, measured at

V_{DS}=50 mV, of the poly-gate transistors tended to be lower than aluminum-gate transistors by ~10%. This may be due to the fact that the gate oxide for the poly-gate transistors was directly exposed to high-energy particlers while a thick ($\sim$ 1μm) layer of photoresist protected the gate areas of the aluminum-gate transistors during e-beam evaporation. The room-temperature data show only a slight improvement in the mobility for devices which received the extra damage anneal. The results are summarized below:

Extra Anneal	Gate Type	V_T(mV)	C'(F/cm^2)	μ_n(cm^2/Vs)
No	aluminum	147	2.70×10^{-7}	660
	polysilicon	36	2.65×10^{-7}	580
Yes	aluminum	7.3	2.40×10^{-7}	670
	polysilicon	33	2.64×10^{-7}	620

Table 5.1: HP 4145 room-temperature data.

Table 5.1 gives maximum differential mobility from the maximum differential transconductance (G_M) measured at V_{DS}=50 mV, via the formula [30]:

$$\mu_n = \frac{G_M}{(\frac{W}{L})C'_{ox}V_{DS}}. \tag{5.1}$$

5.4.2 Low-temperature Data

Measurements were also made with a lock-in amplifier with a very small signal (V_{DS}=100 μV) both at room temperature and at 4 degrees Kelvin. At 4 K, the threshold voltages for both types of transistors shifted up to ~0.5 V due to the temperature dependence of the semiconductor work-function ϕ_F. The differential transconductance at 4 K is strongly dependent on the gate voltage, V_{GS}, and peaks at roughly 1 volt for aluminum-gate transistors and 0.75 volts for poly-gate transistors. This is directly related to the strong mobility dependence on the transverse electric field at low temperatures, where phonon scattering is negligible. At low fields (small V_{GS}), ionized impurity scattering dominates the mobility because the low concentration of electrons in the 2-DEG cannot effectively shield the impurity potential [1] ; at high fields (large V_{GS}), surface roughness scattering limits the mobility [15] because the

channel electrons are pulled very close to the Si/SiO_2 interface. Maximum mobility (or G_M) occurs at a moderate gate voltage, where neither ionized impurity or surface roughness scattering dominates. Table 5.2 summarizes the results. For completeness, room-temperature measurements by the same instrument are included.

Extra Anneal	Gate Type	μ_n(cm^2/Vs), T=300K	μ_n(cm^2/Vs), T=4K
No	aluminum	681	9,200
	polysilicon	652	10,400
Yes	aluminum	733	11,700
	polysilicon	697	12,800

Table 5.2: AC lock-in measurements at room temperature and low temperature.

All mobility values are calculated using experimental gate capacitances measured at room-temperature exept for the poly-gate transistors at 4 K, in which cases the above method yields mobilities in the 14,000-17,000 cm^2/Vs range, which was believed to be too high. Instead, the theoretical capacitance of 3.45×10^{-7} F/cm^2 calculated from 100 Å of gate oxide was used to deduce the poly-gate transistor mobilities at 4 K. It appears that the dopants in the polysilicon gate become active throughout the thickness of the film at low temperatures whereas as an insulating layer exists within the gate near the interface at room temperature. The reason for this is unclear to the author, and low-temperature capacitance measurements may provide more clues.

The AC lock-in measurements done at $V_{DS} = 100$ μV show, by averaging the poly-gate and aluminum-gate results, that the extra 900 °C damage anneal improved the room-temperature mobility by roughly 10% and the low-temperature mobility by roughly 25%. This clearly indicates the importance of the additional damage anneal. For the poly-gate transistors, the after-anneal room-temperature mobility of ~700 cm^2/Vs shows an almost complete restoration compared to n-MOSFET's which do not go through ionizing environments during processing. The low temperature values of ~10,000 cm^2/Vs are similar to the results obtained by Thomas [28] in his quasi-1D MOSFET experiments; this mobility value was high enough for Thomas to have observed universal conductance fluctuations in his devices. In conclusion, a robust fabrication process has been developed which allow the patterning of very fine gate

electrodes and still maintain high quality device characteristics.

Chapter 6

Simulations and Calculations

In order to optimize the design of the Si PRESTEFT device structure and to determine the best bias for resonant tunneling, a good understanding of the transmission coefficient behavior as a function of electron energy, barrier width, height, and shape is necessary. If our experimental environment were an ideal quantum system, i.e. completely coherent, such optimization would not be necessary because quantum mechanical interference would always manifest itself perfectly. Unfortunately, it is clear that in our actual proposed silicon PRESTFET, electron phase-randomization due to various kinds of scattering makes quantum effects unobservable on scales larger than the mean-free-path. To place more exact requirements on device dimensions and bias conditions, numerical simulations are necessary. Given the degree of coherence we have achieved in the silicon 2-DEG, which we relate to the experimental low-temperature mobility (μ ~10,000 cm^2/Vs @4K), we can get an upper limit on the critical device dimensions which will allow resonant tunneling to be observed, as well as the optimum design parameters for the actual transistor, such as the gate oxide thickness.

In the following sections, some computer modeling and calculations are shown. First, two kinds of potentials shapes, the rectangular and the rounded, are compared at various length scales and energies. It was found that at device dimensions presently realizable (e.g. by e-beam lithography. See Khalid Ismail's work [18]), the rounded potential profile is more tolerant than the rectangular one to scattering effects. Some

2D semiconductor Poisson solutions were also done using the PISCES software to determine the actual potential profile for the proposed Si PRESTFET. It was found that with this design (i.e. 500 Å lines and spaces, 100 Å thick gate oxide and 100 Å thick gate nitride), the potential barriers and well take on a rather rectangular shape. This suggests that the gate oxide thickness might be increased to round off the barriers to increase the observable tunneling current, although at the cost of reducing the level spacing and an increase in inter-level smearing. Of course, the above trade offs were made because of limitations on the lithographic technology. The optimum solution is to reduce the tunneling gate linewidth and separation well below the presently achievable level. It is estimated that at 100-200 Å linewidth, we will observe a strong resonant tunneling effect in silicon.

6.1 Method of Transmission Coefficient Calculation

The calculations of T(E) were done by discretizing the potential profile into many intervals in space; within each interval the potential is approximated as having a constant value [27]. A 2-by-2 connection matrix is used to relate the forward and reverse wave amplitudes on either side of each interval. Then these matrices are multiplied to yield a final 2-by-2 matrix which represents the *entire* potential structure, whether it contains one or multiple barriers and wells. The transmission coeffient is extracted from this final matrix, and of course, the values of the matrix elements are energy-dependent.

6.2 Transmission Behavior for Rectangular and Rounded Double-Barriers

In an MBE-grown vertical double-barrier resonant-tunneling (DBRT) diode, the potential profile is rectangular because of the sharp interfaces between the barrier and

the well material and the resulting jump in the conduction band energy. Because very thin (down to $\sim$ 15 Å, or 3 monolayers) layers of high-purity material can be deposited, the level separation of the quasi-eigenstates of the well is large (typically 1-10 eV) and the tunneling current is large (typically 1-100 mA for a $10 \times 10\mu m^2$ diode) and observable at room temperature. See, for example, the work by Broekaert *et al*[9]. The temperature tolerance and large peak-to-valley ratio is a direct consequence of the tight quantum confinement. In fact, nearly all vertical DBRT devices built up to date are based on rectangular barriers, with one notable exception, the work by Sen *et al* [29], in which a parabolic quantum well is sandwiched between two thin barriers. The 500 Å parabolic well (relatively wide by MBE standards) was formed by short-period ($\sim$ 15 Å), variable duty-cycle, $GaAs/Al_{0.3}Ga_{0.7}As$ supperlattices in which the Al content within each layer produced the correct amount of band-bending for a parabolic well. Resonant tunneling was observable on this device only up to a temperature of 100 K, because the quantum nature of the device was compromised by enlarging the well. The I-V characteristics show strong resonances at 7.1 K, and the energy levels were shown to be indeed spaced evenly, about $\sim$ 90 meV apart. It is an impressive demonstration of the reliability of MBE technology for producing high-quality interfaces because the presence of 30 such interfaces in the well did not produce enough scattering to reduce the coherence of the electrons.

For our proposed planar device, the PRESTFET, it is much harder to have precise control over the exact shape of the potential because it is induced electrostatically by metal gates with minimum lateral dimensions on the order of or greater than their distance from the 2-DEG. For the GaAs PRESTFET built by Ismail [18], coincidentally, both the critical lateral gate dimension and the gate-to-2DEG distance were limited to 600 Å. A further reduction in gate dimensions will not reduce the potential dimensions at the 2DEG because of electrostatic fringing. For the silicon PRESTFET, though, we have the luxury of varying the gate-to-2DEG distance by varying the gate oxide and nitride thicknesses. The thinnest gate oxide that can be reliably grown in our lab is approximately 50 Å, although thinner oxides can be grown at the cost of poorer interface quality. Thus, the silicon system allows us the possibility to

form quantum wells with dimensions on the same order as the vertical DBRT devices! This is very exciting indeed.

Given the present state-of-the-art in electron-beam direct-write technology and also in x-ray proximity printing in the immediate future, we can easily fabricate two gate fingers with linewidths and separation on the order of 500 Å. So we compare the transmission of rectangular barriers with rounded barriers at this linewidth. The rectangular barrier system has a 500 Å wide well in between two 500 Å thick barriers. The rounded barrier system is modeled by two periods of a cosine function with $\lambda/2$ = 500 Å. The barrier height for both was set to 30 meV. Typical values for $(E_f - E_c)$ in MOS systems at low temperatures are in the range 5-20 meV.

In figure 6-1, we plot the single- and double-barrier transmission as a function of energy for both the rectangular and the rounded barrier shapes, on a log scale. For both cases, the double-barrier transmission shows distinct peaks corresponding to the quasi-eigenstates of the well. The transmissions are exactly unity at those energies because the barriers are symmetric, but they appear smaller because a finite, discrete set of energy values were used to make the plots. Several important features should be noticed. First, the level spacing increases quadratically with energy for the rectangular system ,and is roughly constant for the rounded system, which approximates a parabola near the well bottom. The level spacing is about 2-3 meV near the barrier maximum for both cases (more obvious in Figure 6-2). Secondly, for both systems, the minimum (or off-resonance) double-barrier transmission is roughly the square of the single-barrier tramsmission at the same energy. This is the fully coherent behavior when no specific phase-matching conditions are satisfied. Thirdly, and probably most importantly, we notice that the single- barrier T(E) drops much more rapidly below 30 meV (barrier maximum) for the rectangular case than the rounded case. Since we know from semiclassical arguments that

$$\Gamma_e \sim \frac{\hbar}{2w} v(T_1 + T_2) \tag{6.1}$$

,which states that the width of a resonant peak, or its 'leakiness', is proportional to

the single-barrier transmission probability. The incident electrons are not monochromatic (i.e. they are distributed within a range of energies); thus a wider resonant peak allows more current to flow since the double-barrier structure serves as a more 'tolerant' energy filter in this case. Hence, the rounded barriers are expected to produce larger tunneling currents in general. Another consequence of Eq. 6.1 is that since single-barrier T(E) drops exponentially with energy for both systems, tunneling through lower levels is impossible because the vanishingly small Γ_e implies undetectable currents as well as complete smearing due to the large ratio Γ_i/Γ_e.

The transmission behavior near the energies where tunneling is likely to occur is plotted in figure 6-2 for the same structures described above. For the rectangular case, both the single- and double-barrier transmission display many irregularly-spaced peaks for E > 30 meV. These are called Ramsauer resonances and do not correspond to true tunneling, since the incident energy is above the barrier height. They are also the result of constructive interference of waves reflected by the sharp discontinuities of the potential, and since four such sharp 'edges' are seen by electrons with energies above 3o meV, the condition for constructive interference is more complicated. Below the 30 meV mark, the rectangular double-barrier transmission shows a tiny peak at $\sim$ 29.5 meV, which will not give rise to RT because its FWHM, or corresponding Γ_e, is much smaller compared to the experimental Γ_i of 0.6 meV. The Ramsauer resonances, on the other hand, have FWHM on the order of 1-3 meV, which is larger than Γ_i. We thus expect to observe some non-linear structure in I-V for a rectangular double-barrier potential due to the Ramsauer effect , but it's likely to be weak because the Ramsauer level spacing is not much larger than the inelastic smearing of 0.6 meV.

Now we turn to the rounded potential structure. Peaks in T(E) occur only for double-barriers, whereas the single-barrier T(E) is smooth throughout. This is expected because the rounded single-barrier has no sharp discontinuities. The double-barrier transmission shows a significant tunneling peak at 29 meV, with a FWHM of 0.25 meV. The level spacing is about 2-3 meV. Given an inelastic smearing of 0.6 meV, RT will probably not be observable for this quasi-eigenstate.

Since our goal is to observe resonant tunneling, we shall focus on the tunneling

phenomenon, in which the particle energy is below the barrier height. Ramsauer resonances, although interesting, will not be studied further here.

6.3 Linewidth Dependence of RT

As the barrier widths are reduced, the quantum well becomes more leaky; more tunneling current flows and inelastic smearing becomes less detrimental. Thus we next investigate the effect of reducing the critical dimension of the gates, or the linewidth. Figure 6-3 plots the single-barrier $T_1(E)$ for potentials characterized by various linewidths, for the rectangular system and the rounded system. The barrier height is kept constant at 30 meV. Since from Eq. 6.1 we know that the tunneling current and resistance against inelastic smearing is related to the single-barrier transmission probablity, figure 6-3 helps us pinpoint the efffect of linewidth reduction.

Note that the rectangular case is plotted on a log scale, while the rounded case is plotted on a linear scale; this is because the former has a strong energy dependence, while the latter has a weaker energy dependence. The rectangular barrier results in T(E) which is drastically dependent on linewidth : the rate at which T(E) decreases below 30 meV and the value of T(E) at E=30 meV are both strong functions of linewidth. For the rounded barrier, T(E) is a much more gradual function of linewidth. A very interesting result is that T(E)=0.5 for E=30 meV *independent* of the linewidth. This may be a direct consequence of better impedence matching of the smooth- varying structure to the reservoir for the rounded-barrier case. Comparing the values of $T(E_{max})$ at E_{max}=30 meV, we see that for linewidths $>$ 10 nm, rectangular barriers are not favorable because small transmission probabilities imply huge transient build-up times (electron life-time), which means an infinitely coherent system is required for RT. Also, even if infinite coherence is obtained, the RT current may fall below the detection limit of practical instruments. Thus we conclude that for linewidth greater than 10 nm, rounded potential profiles are preferable, given that the silicon PRESTFET is coherence-limited.

In order to be relatively immune from inelastic smearing and to achieve a sizable current, we saw that rounded barriers are preferable, given current lithographic limits. However, this choice comes with a penalty, in that RT level spacing (near the top of the barrier) is smaller than that of a rectangular barrier system. See figure 6-4. It is shown that at a given linewidth, the level spacing is smaller for a rounded system than a rectangular system. Roughly, this can be understood by noticing that in a parabolic well, the discrete energy levels are equally spaced, whereas in a square well, the energy levels go up quadratically with quantum number. Another possible way to see this is that structures with sharp discontinuities (such as a square barrier) tend to exhibit stronger non-linearities than those with gradual change (such as a rounded barrier). Figure 6-4 assumes a fixed barrier height of 30 meV.

6.4 Barrier Height Dependence of RT

The optimum barrier parameters include the linewidth and the barrier height. The linewidth is mostly determined by the lithographic process, and the barrier height is mostly determined by the voltage bias applied on the gate electrodes. We introduce here a *tunneling coefficient*, which we define as the single-barrier transmission coefficient *evaluated* at an incident energy equal to the height of the potential barrier. This is a useful figure since for our thick, 500 Å wide barriers, resonant tunneling is likely to occur only when the barriers are reduced to match the Fermi energy of the 2-DEG. Figure 6-5 plots this tunneling coefficient as a function of barrier height for both rectangular and rounded barriers, with the linewidth held constant at 500 Å. We see that, surprisingly, the transmission for the rounded barrier is *always* 1/2 independent of barrier height, whereas the transmission for the rectangular barrier decreases rapidly as height is raised, and is always lower than the rounded transmission. Clearly, inelastic smearing and tunneling current arguments makes the rounded system more favorable.

Again, we see that while rounded barriers are less susceptible to inelastic effects (shorter build-up times) and allow more current to flow (wider transmission peaks),

rectangular barriers result in less inter-level smearing because of larger RT level spacings near the top. This is seen in figure 6-6. Level spacing in general grows with increasing barrier height. This might suggest that we can work with the rounded barrier at a relatively coarse linewidth, say 500 Å, and simply raise the barrier potential indefinitely to reduce inter- level smearing. This does not work because in order to achieve RT through the top level, the Fermi level must be at the same height as the potential barrier. In order to raise the Fermi level, and thus support a high electron density, a large top gate bias must be applied, resulting in very large electric fields in the direction perpendicular to transport. This produces severe scattering and destroys coherence. At low temperatures, where the mobility is highly dependent on the field, a very narrow window exists where the mobility is maximum, and this defines where we must bias the device. From actual transistor measurements, the maximum mobility of $\sim$ 10,000 cm^2/Vs for aluminum-gated transistors peak at $V_{GS} \sim 1$ V. This corresponds to $(E_f - E_c) = 5$ meV. At this value, the level spacing is only 1.5 meV for rounded barriers, from Figure 6-6. Since the total transmission width, Γ_{tot} is roughly 0.85 meV ($\Gamma_i = 0.6$ meV, $\Gamma_e = 0.25$ meV) for a 500 Å linewidth, parabolic potential structure under optimum biasing , we expect significant inter-level smearing in addition to inelastic smearing.

6.5 Sensitivity of RT to Inelastic Scattering

It has been shown in previous chapters that for a purely coherent system characterized by an elastic width Γ_e, the net effect of adding an inelastic mechanism characterized by an inelastic width Γ_i, after summing both coherent and sequential contributions to the transmission, is to smear out the sharp resonances in the transmission coefficient. The original structure is reduced *and* broadened by the ratio Γ_{tot}/Γ_e, where Γ_{tot} is the FWHM of the new transmission function and is the sum of Γ_e and Γ_i. Figure 6-7 displays the value of the *peak* transmission coefficient for a symmetric double-barrier system as a function of the degree of coherence, characterized by a single parameter *epsilon* (ϵ); ϵ is the probablity for an electron to be scattered inelastically into a side

reservoir upon each incidence of the scattering junction. The 'leakiness' of the system, characterized by the single-barrier transmission T_1, is used as a parameter such that devices with different linewidths (or barrier heights) may be compared. The plot is based on the simple yet elegant single-scatterer/side-reservoir model developed by M. Büttiker [13]. The basic equations are:

$T = T_c + T_i$,

$T_c = (1-\epsilon)T_1T_2/|Z|^2$,

$T_i = S_bS_f/(S_b + S_f)$,

$S_b = \epsilon T_1[(1 + (1-\epsilon)R_2]/|Z|^2$,

$S_f = \epsilon T_2[(1 + (1-\epsilon)R_1]/|Z|^2$,

$|Z|^2 = 1 + (1-\epsilon)^2R_1R_2 + 2(1-\epsilon)R_1^{1/2}R_2^{1/2}cos(\phi)$, where T_c and T_i are the coherent and sequential components of the transmission probability, respectively. S_f and S_b are the forward and backward scattering probabilities from the side reservoir. $|Z|^2$ describes the Lorenzian energy dependence common to both the coherent and sequential components, and ϕ is the phase due to a round-trip traversal of the well.

Figure 6-7 clearly shows the importance of a 'leaky' well (corresponding to large individual transmission coefficients) when inelastic effects are important. For a hypothetical silicon PRESTFET with a 500 Å well and barriers, the highest mobility estimates (μ_n ~10,000 cm^2/Vs) yield a scattering probability per incidence, ϵ, equal to 0.3, using the relation $\Gamma_i = -2\hbar\nu\ln(1-\epsilon)$ [13]. It is important to note that for $\epsilon = 0$ (i.e. no scattering), symmetric double barriers yield T=1 at resonance independent of the 'leakiness', or individual barrier transmissions. At $\epsilon = 0.3$, the resonant transmission is still large for leaky barriers ($T_1 = T_2 = 0.5$), but is reduced dramatically for the other three cases. This explains why for thick barriers, resonant tunneling can only occur for carrier energies near the top of the barriers, at which the well becomes more 'leaky'. The large value, 0.3, of the scattering probability comes from both the large size of the well and the low mobility inside silicon. This clearly shows why the proposed silicon PRESTFET is coherence-limited.

6.6 PISCES simulations

Using the software PISCES-2B [25], we obtained two-dimensional Poisson solutions in silicon for the proposed 500 Å minimum-linewidth PRESTFET. The simulations were done at a temperature of 80 K, the lowest temperature that allowed convergence for the software used. It is expected that the electrostatic parameters at 4 Kelvin will not be significantly different from the 80 K results. The device simulated has 100 Å of gate oxide and 100 Å of gate nitride as the dielectrics. N+ polysilicon is assumed to be the tunnel gate material, and aluminum is assumed the top gate material. The background doping in the bulk silicon is n-type, 10^{15} cm^{-3}. All physical parameters are consistent with the fabrication experiments actually carried out.

Figure 6-8 shows the conduction band energy, E_c, at the interface, along the direction of electron transport, for various top gate voltages. The tunnel gates and back gate were held constant at 0 volts. As the top gate bias becomes more positive, the electron concentration, away from the tunnel gates, of the 2-DEG increases. This corresponds to an increase of $(E_f - E_c)$, as expected. An important result here is that at the optimum top gate biasing of 1 volt, where the mobility is largest, $(E_f - E_c)$ is only about 5 meV. For RT to occur, the barriers must be lowered to match the Fermi energy of the 2-DEG approximately (otherwise the transmission peak widths are too small, resulting in huge resonance build-up times as well as a tunneling current too small to be measured). Lowering the barriers causes a reduction in the inter-level spacing (as seen in Figure 6-6), a unfavorable effect.

The lower (tunnel) gates must be biased such that the potential barriers line up with E_f for RT to occur. Figure 6-9 displays the actions of lower-gate biasing on the potential profile. The top gate voltage is fixed at 2 volts. For ease of simulation, only the left lower-gate voltage is swept. We see that about half a volt is required to 'push' the barrier down to the Fermi level. The most striking feature here is that the 'deformed' barrier takes on a rather square profile, with the barrier top relatively flat. This is because the 500 Å -wide gate is only 100 Å away from the 2-DEG, greatly reducing the elestrostatic fringing, which was a problem for the GaAs

PRESTFET. The fact that the lower gates can come so close to the 2-DEG is not necessarily an advantage at coarse linewidths, since this results in a more rectangular barrier. Because rounded, parabolic potentials are favored at currently achievable linewidths, it may be necessary to increase the gate oxide thickness, which should produce more gradual, rounded barriers as a result of increased fringing. The nitride thickness should be kept at a minimum so that the top gate can still wield control on the well potential; otherwise, fringing fields from the tunnel gates will 'wash out' the quantum well. Whether fine-tuning the gate oxide and nitride thicknesses will significantly improve our chances of seeing RT still needs to be examined carefully.

The conclusion from the above calculations and discussion is that for a low-mobility system, such as silicon, extreme confinement is required for quantum effects to become observable. Simulations show that at a linewidth of 100 Å, a rounded double-barrier system yields an elastic width, Γ_e, of 3.4 meV, about five times larger than the inelastic width, Γ_i, of 0.6 meV. This implies that resonant tunneling will be much less affected by inelastic scattering. Intuitively, this is because large elastic width corresponds to small electron lifetimes and thus the RT transient build-up time will also be small, making scattering less likely to occur. Reducing *both* the barrier thickness and the well width are necessary: The former reduces effects of inelastic smearing by reducing transient build-up times while the latter reduces effects of inter-level smearing by enlarging level spacing.

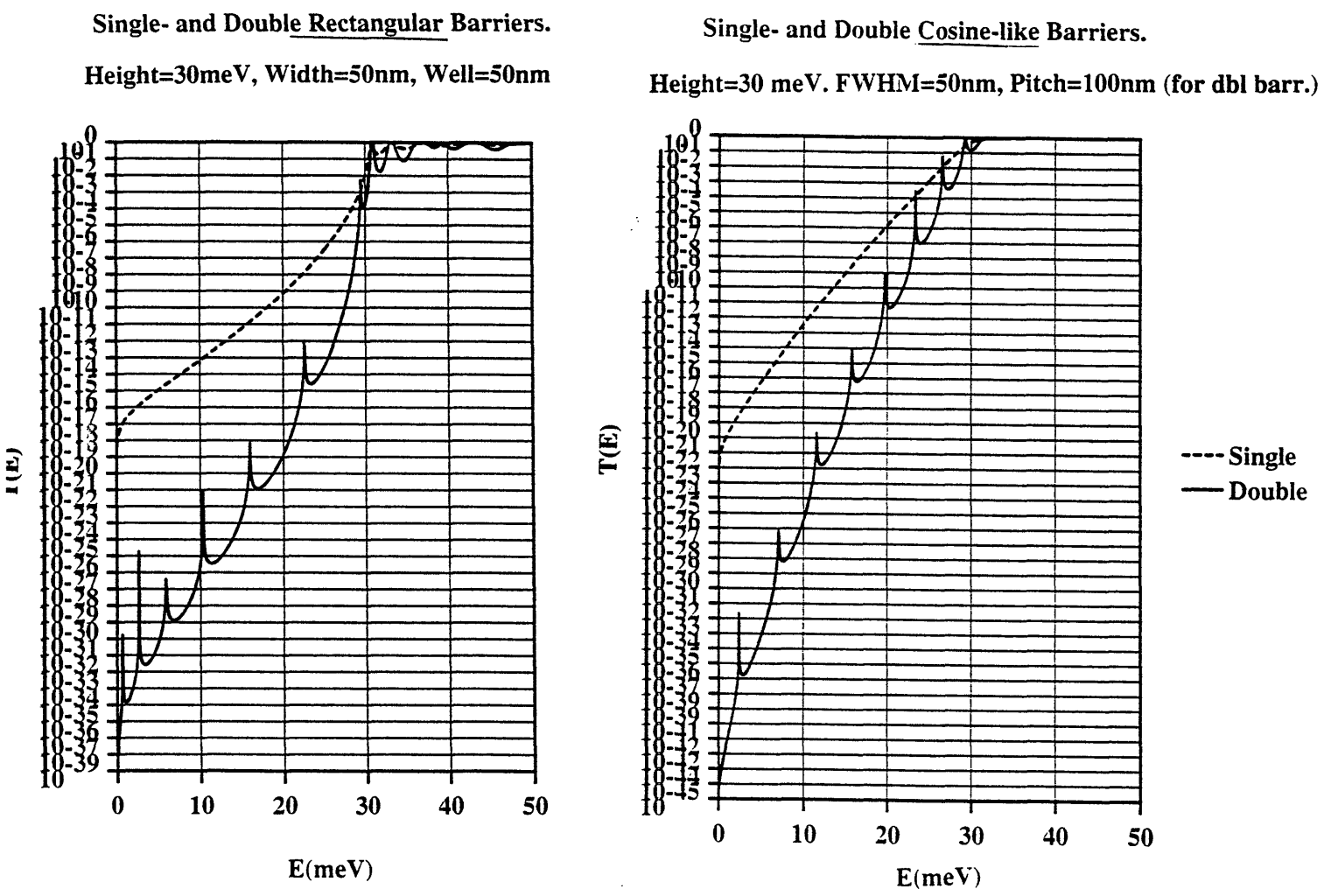

Figure 6-1: Rectangular (left) and Rounded (right) Barrier Transmission

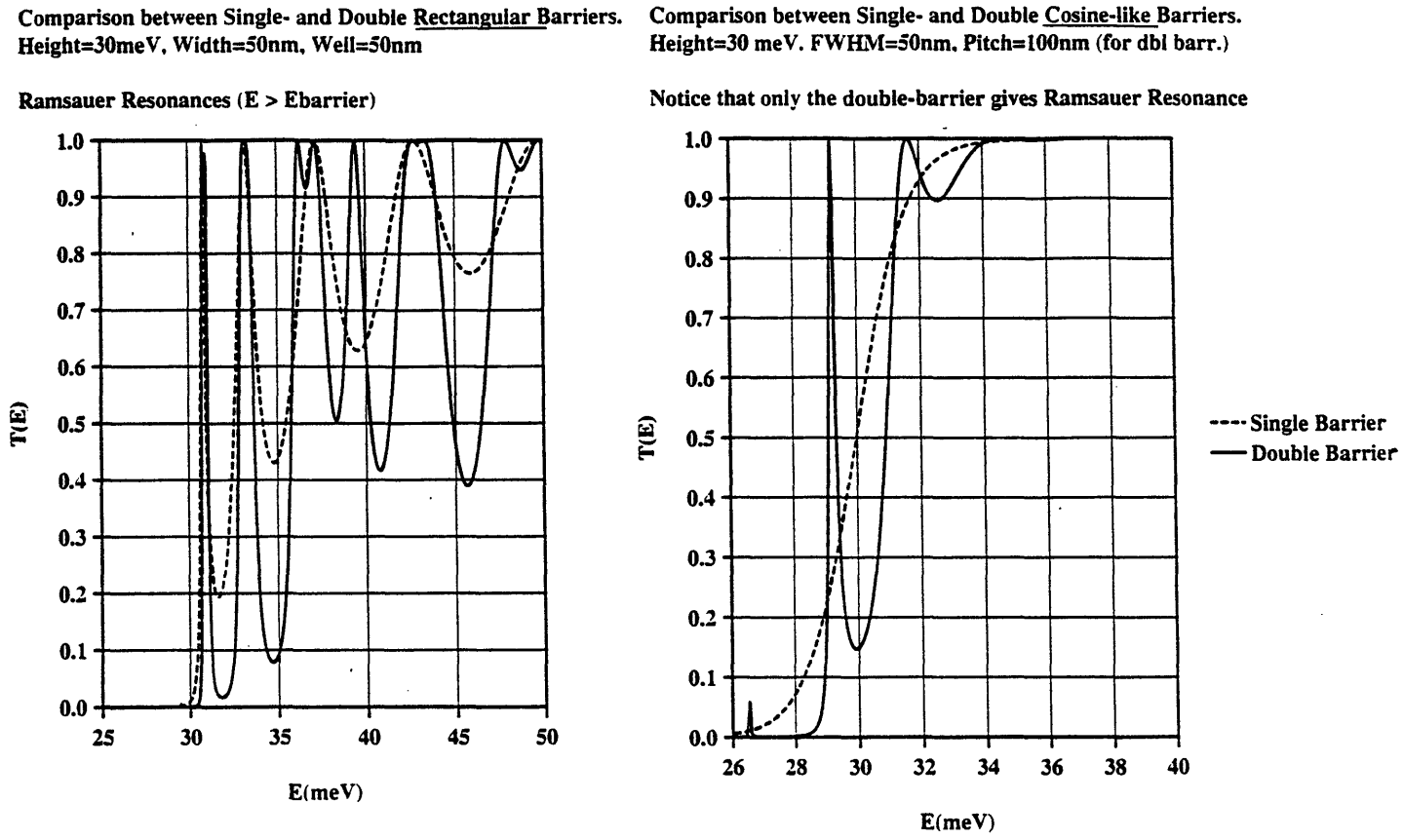

Figure 6-2: Transmission for Energies near the Barrier Top

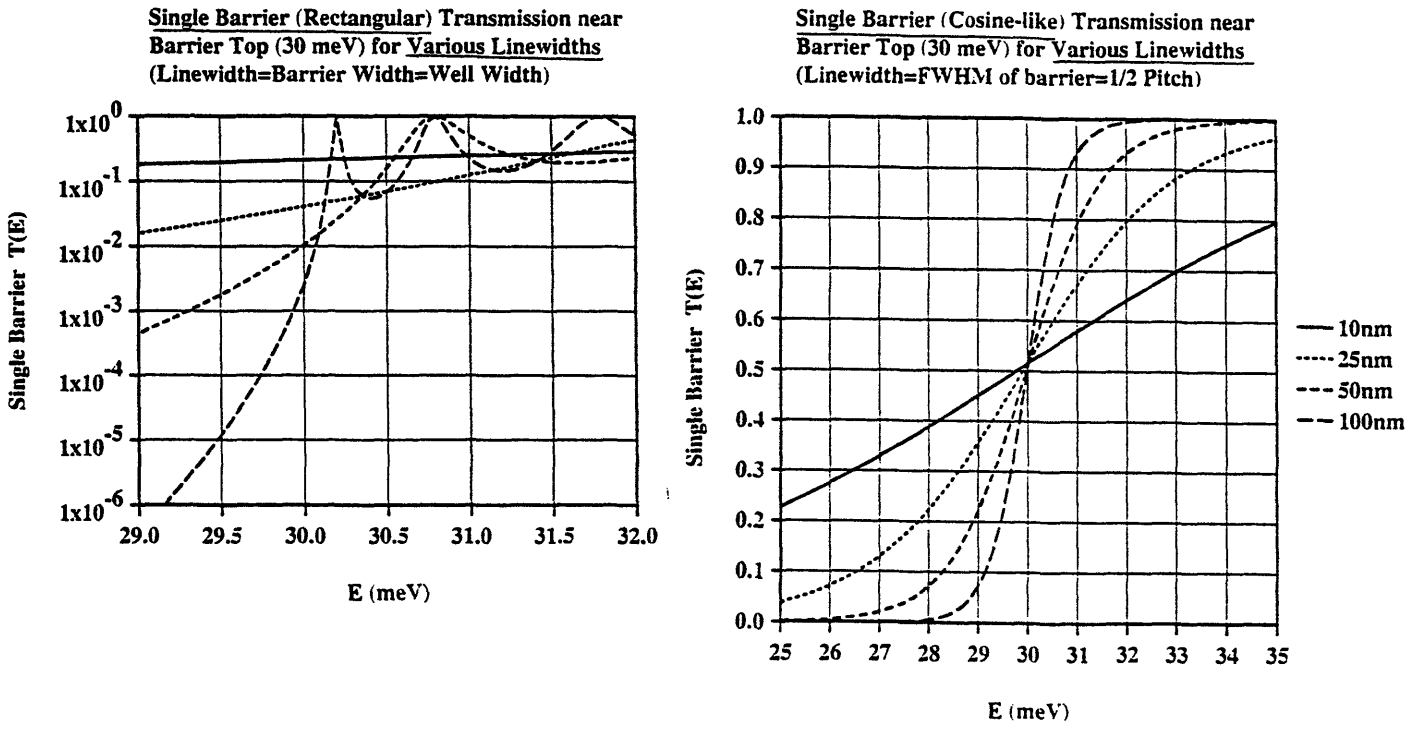

Figure 6-3: Single-Barrier Transmission with Linewidth as Parameter. Note the different abscissa/ordinate scaling in the two plots.

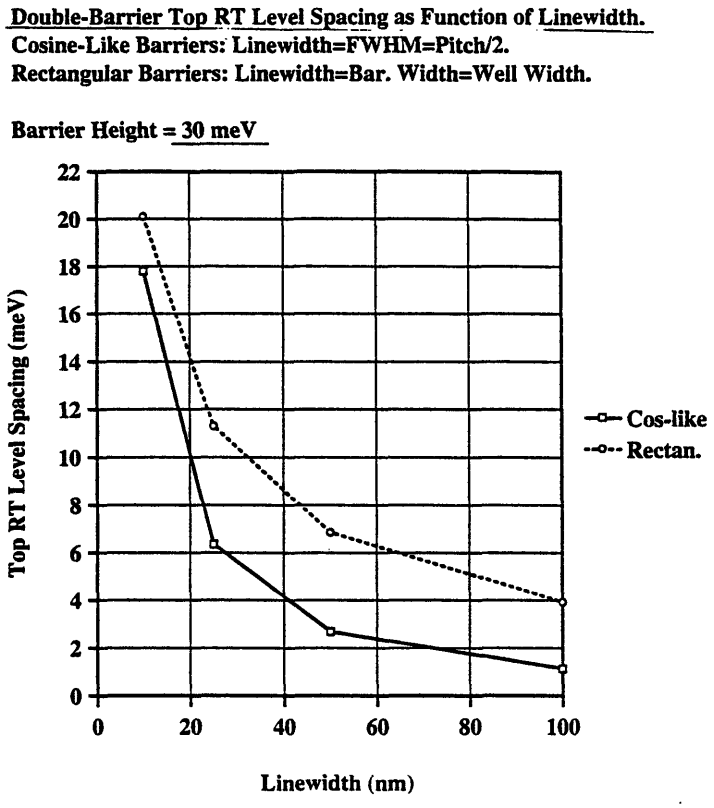

Figure 6-4: Level Spacing as Function of Linewidth

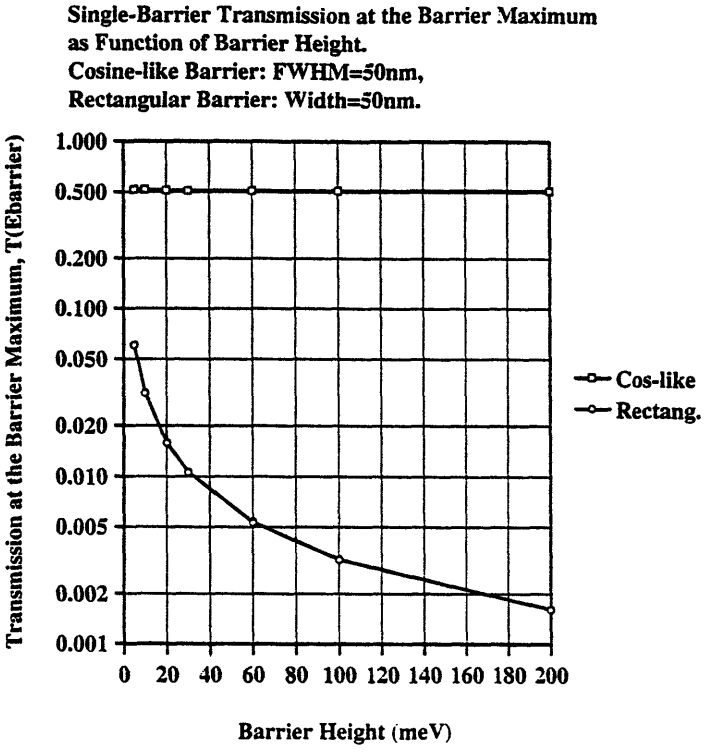

Figure 6-5: Single-Barrier Transmission as Function of Barrier Height

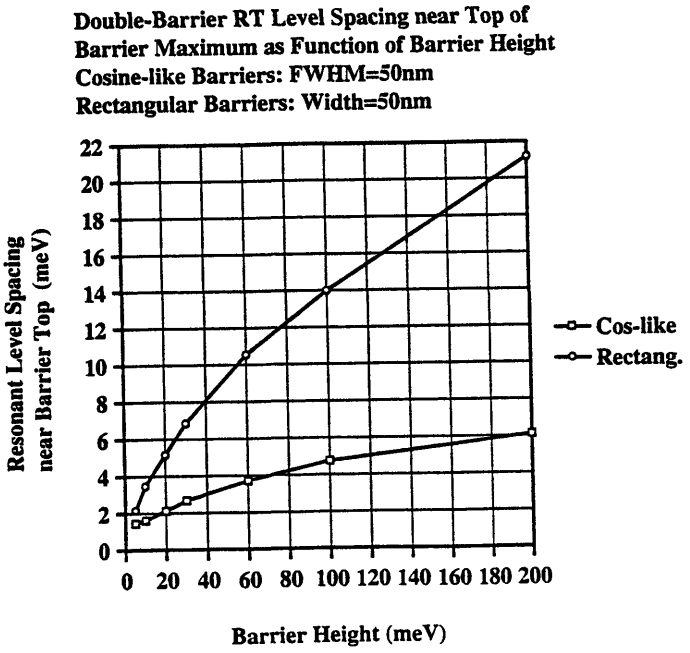

Figure 6-6: RT Level Spacing as Function of Barrier Height

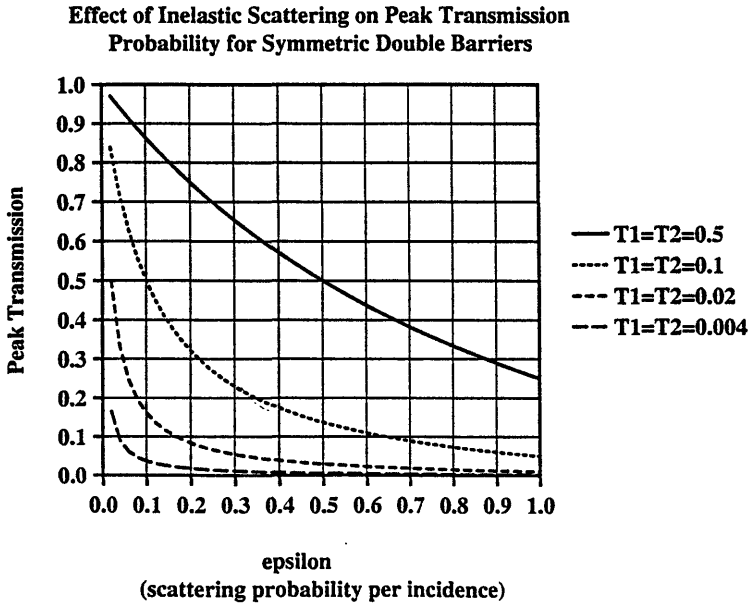

Figure 6-7: Peak RT Transmission as Function of Scattering

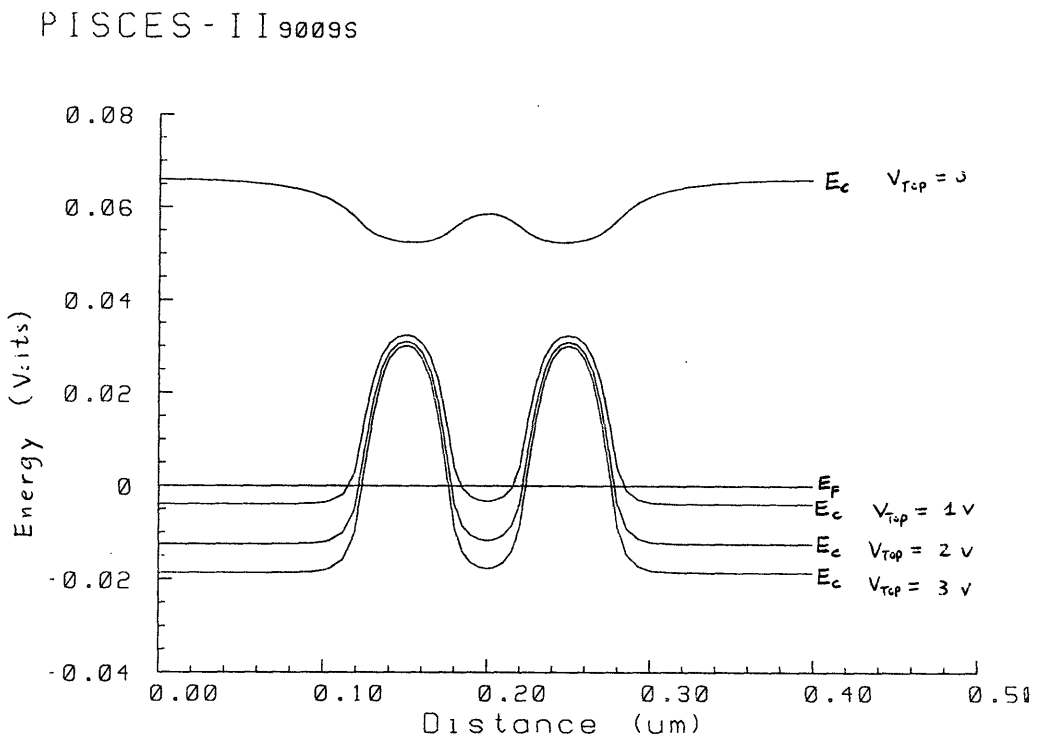

Figure 6-8: Conduction Band Energy, Top Gate Biasing. Lower Gates and Back Gate at 0 V. T = 80 K.

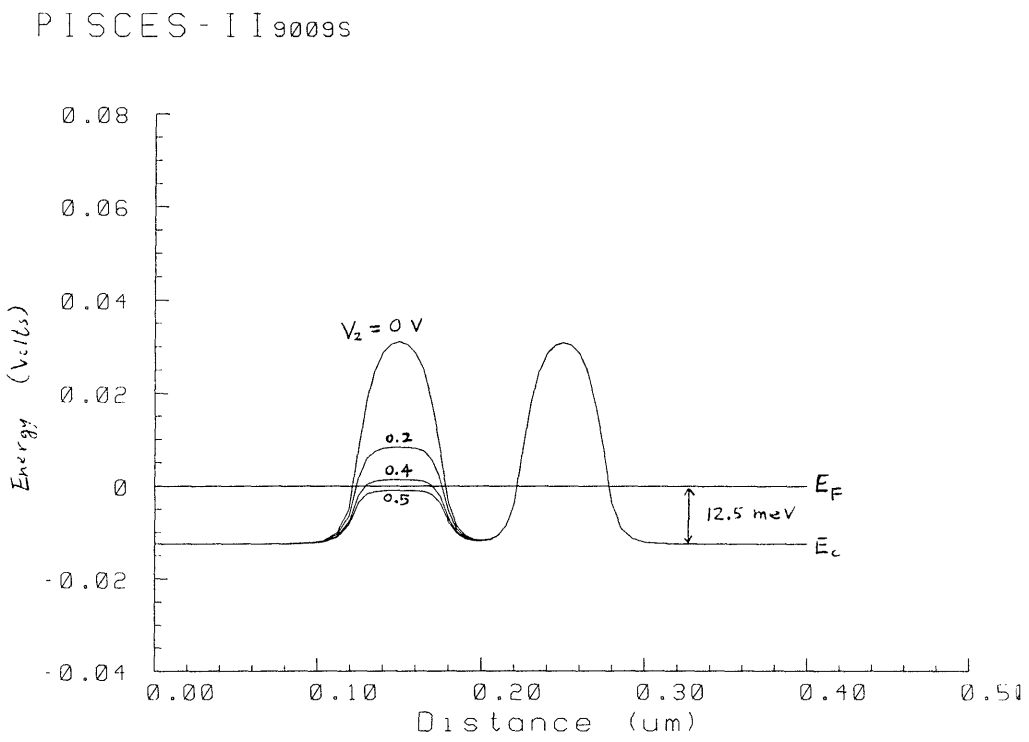

Figure 6-9: Conduction Band Energy, Lower Gate Biasing. Top Gate at 2 V. Back Gate at 0 V. T = 80 K.

Chapter 7

Conclusions

We have developed a process suitable for the fabrication of a silicon PRESTFET with a 500 Å minimum linewidth. The fine gates can be formed by lift-off of e-beam evaporated amorphous silicon after implantation of low-energy antimony(Sb). Either x-ray or e-beam lithography can be used to pattern these fine gates at this dimension. The diffusion and activation of implanted arsenic(As) ions in the evaporated silicon were found to require very high temperatures. A 1050-1080 °C, 20 second rapid thermal anneal successfully activated the dopants, whereas a 900 °C, 1-hour furnace anneal was not successful. The phenomenon was attributed to oxygen contamination introduced into the film during e-beam evaporation; higher vacuums during evaporation were found to result in lower resistivities, and hence better film quality. The damage introduced in the test devices by energetic particles during e-beam evaporation was significantly reduced with an additional 900 °C furnace anneal. The room temperature and low temperature mobilities measured from ordinary test transistors indicate that the device quality is high enough for quantum-effect research. If the minimum gate dimensions are to be reduced below 500 Å, a new process is needed since doping of the gate via implantation is no longer feasible; the dopants may be evaporated or spun onto the silicon gate, or a gate material other than silicon may be considered.

The intrinsic, low electron mobility in the Si/SiO_2 2-DEG implies a high degree of scattering; hence quantum interference effects are difficult to observe unless extremely tight confinement is produced. In the case of resonant tunneling, the most critical

observability criterion is that the electron inelastic scattering time must be much longer than the transient time required to build up the full, resonant wavefunction inside the well. The inelastic scattering time is limited by the material chosen and the temperature of device operation, whereas the build-up time is a function of the barrier structure and the incident carrier energy. To minimize this resonant build-up time, the barriers need to be made thinner, effectively making the well 'more leaky'. At resonance, the barrier height must be close to the value of $(E_f - E_c)$ which corresponds to maximum carrier mobility; thus barrier heights cannot be varied at will to shorten the build-up transient. A second observability criterion for RT is that the energy levels of the quasi-eigenstates inside the well must be far enough apart from one another to prevent inter-level smearing and allow a finite source-drain voltage to be applied. Specifically, the spacing between the resonant peaks in the total transmission function, T(E), after incorporating inelastic scattering and other broadening mechanisms, must be larger than the characteristic widths (e.g. FWHM) of the individual peaks. This is achieved by shrinking the width of the quantum well.

It is estimated in this work that a silicon PRESTFET with a 100 Å minimum linewidth (two 100 Å wide gate fingers, separated by 100 Å) should exhibit resonant tunneling at 4 K, given a maximum electron moblility of ~10,000 cm^2/Vs measured at this temperature. The intrinsic MOS design of this device will allow even finer gates to function effectively since the gate oxide thickness can be reduced to 20-30 Å , although at the cost of degrading the Si/SiO_2 interface quality. Future work on the silicon PRESTFET should focus on the definition of the two parallel, ultra-fine gate electrodes as well as maintaining high device mobility during fabrication. The prospect of a lithographically defined quantum-effect device in silicon is indeed very exciting because it opens up venues for a new class of ULSI electronics based on the most reliable and well-characterized semiconductor material known to man.

Bibliography

[1] V.M. Agostinelli, H.S. Shin, and A.F. Tasch. A comprehensive model for inversion layer hole mobility for simulation of submicrometer MOSFET's. *IEEE Trans. Electron Devices*, 38(1):151, 1991.

[2] P. F. Bagwell. *Quantum Mechanical Transport in Submicron Electronic Devices.* PhD thesis, Massachusetts Institute of Technology, 1990.

[3] P.F. Bagwell, T. Broekaert, T.P. Orlando, and C.G. Fonstad. Resonant tunneling diodes and transistors with a one-,two-, or three-dimensional electron emitter. *J. Appl. Phys.*, 68(9):4634, 1990.

[4] Philip Bagwell and Terry P. Orlando. Landauer's conductance formula and its generalization to finite voltages. *Phys. Rev. B.*, 40(3):1456, 1989.

[5] E. Bernstein and S. Lundqvist, editors. *Tunneling Phenomena in Solids*, chapter 1. Plenum, New York, 1969. Chapter written by E. O. Kane.

[6] J. Blatt and V. F. Weisskopf. *Theoretical Nuclear Physics.* Springer, Berlin, 1979.

[7] D. Bohm, editor. *Quantum Theory.* Prentice-Hall, Englewood Cliffs, N.J., 1951.

[8] G. Breit and E. Wigner. title unknown. *Phys. Rev. Lett.*, 49:519, 1936.

[9] T. P. Broekaert, W. Lee, and C. Fonstad. Pseudomorphic $In_{0.53}Ga_{0.47}As$ / AlAs / InAs resonant tunneling diodes with peak-to-valley current ratios of 30 at room temperature. *Appl. Phys. Lett.*, 53(16):1545, 1988.

[10] E. R. Brown. title unknown. In *Advanced Heterostructure Transistor Conference*, Kona, Hawaii, 1988.

[11] Martin Burkhardt. Private Communication.

[12] M. Büttiker. Role of quantum coherence in series resistors. *Phys. Rev. B.*, 33:3020, 1986.

[13] M. Büttiker. Coherent and sequential tunneling in series barriers. *IBM J. Res. Develop.*, 32(1):63, January 1988.

[14] F. Capasso, editor. *Physics of Quantum Electron Devices.* Springer-Verlag, Berlin Heidelberg, 1990.

[15] Y.C. Cheng and E.A. Sullivan. On the role of scattering by surface roughness in silicon inversion layers. *Surface Science*, 34:717, 1973.

[16] S.K. Ghandhi. *VLSI Fabricarion Principles, Silicon ans Gallium Arsenide.* John Wiley & Sons, 1983.

[17] M. Hamasaki. Radiation effects on thin-oxide MOS capacitors caused by electron beam evaporation of aluminum. *Solid-State Electronics*, 26(4):299, 1983.

[18] K. Ismail. *The Study of Electron Transport in Field-Effect-Induced Quantum Wells on GaAs/AlGaAs.* PhD thesis, Massachusetts Institute of Technology, 1989.

[19] T. Kamins. *Polycrystalline Silicon for Integrated Circuit Applications.* Kluwer Academic Publishers, Boston/Dordrecht/Lancaster, 1988.

[20] R. Landauer. Electrical transport in open and closed systems. *Z. Phys. B - Condensed Matter*, 68:217, 1987.

[21] T.P. Ma and P. V. Dressendorfer, editors. *Ionizing Radiation Effects in MOS Devices and Circuits.* John Wiley & Sons, 1989.

[22] K.D. Möller. *Optics.* University Science Books, Mill Valley, CA, 1988.

[23] T.P. Orlando, P.F. Bagwell, H.I. Smith, and S.D. Senturia. *Physics for Solid State Applications.* MIT 6.730 Course Notes, 1985.

[24] M. Peckerar, R. Fulton, and P. Blaise. Radiation effects in MOS devices caused by x-ray and e-beam lithography. *J. Vac. Sci. Technol.*, 16(6):1658, 1979.

[25] M.R. Pinto, C.S. Rafferty, H.R. Yeager, and R.W. Dutton. *"PISCES-2B," Supplemetary Report.* Stanford Electronics Laborotories, Department of Electrical Engineering, Stanford University, Stanford, CA, 94305, 1985.

[26] B. Ricco and M. Y. Azbel. Physics of resonant tunneling. The one-dimensional double-barrier case. *Phys. Rev. B.*, 29(4):1970, 1984.

[27] George Rittenhouse. Private Communication.

[28] J. Scott-Thomas. *Conductance Oscillations Periodic in the Charge Density of One Dimensional MOSFET Structures.* PhD thesis, Massachusetts Institute of Technology, 1990.

[29] S. Sen et al. Observation of resonant tunneling through a compositionally graded parabolic quantum well. *Appl. Phys. Lett.*, 51:1428, 1987.

[30] M. Shur. *Physics for Semiconductor Devices.* Prentice Hall, Englewood Cliffs, N.J., 1990.

[31] H.I. Smith. *Submicrometer Structures Technology.* MIT 6.781 Course Notes, 1986.

[32] P.K. Tedrow and C.G. Sodini. Twin-well CMOS Process, Version 1.3. Technical report, Massachusetts Institute of Technology, Room 39-521, Cambridge, MA 02139, July 1993.

[33] M. Young. *Optics and Lasers.* Springer-Verlag, Berlin Heidelberg, 1986.