Effects of interface roughness and conducting filaments in metal–oxide–semiconductor tunnel structures

D. Z.-Y. Ting
Thomas J. Watson, Sr. Laboratory of Applied Physics, California Institute of Technology, Pasadena, California 91125 and Department of Physics, National Tsing Hua University, Hsinchu 30043, Taiwan

T. C. McGill
Thomas J. Watson, Sr. Laboratory of Applied Physics, California Institute of Technology, Pasadena, California 91125

(Received 21 January 1998; accepted 21 May 1998)

The current–voltage characteristics of n^+ poly-Si/SiO_2/p-Si tunnel structures containing nonuniform ultrathin oxide layers are studied using three-dimensional quantum mechanical scattering calculations. We find that, in general, roughness at the Si/SiO_2 interface renders the oxide layer more permeable. In the direct-tunneling regime, interface roughness induces lateral localization of wave functions, which leads to preferential current paths. But in the Fowler–Nordheim tunneling regime it affects transport primarily through scattering. These two distinct mechanisms lead to opposite current density dependencies on island size. We have also examined oxide-embedded conducting filaments, and found that they act as highly efficient localized conduction paths and lead to dramatic increases in current densities. Depending on the filament length, our model can mimic experimental current voltage for ultrathin oxides having undergone either quasibreakdown or breakdown. We also found that the lower bias current densities in the structure with long filaments are greatly enhanced by resonant tunneling through states identified as quantum dots, and that this current enhancement is highly temperature dependent. We also report on the dependence of current–voltage characteristics on filament diameter size and filament density. [S0734-211X(98)07804-X]

I. INTRODUCTION

The continued scaling of metal–oxide–semiconductor (MOS) device structures has brought much attention to ultrathin oxides. Normal operation of MOS field-effect transistor with 1.5 nm direct-tunneling gate oxide has been reported.[1] Tunneling through oxide barriers, as a mechanism for leakage currents, is of particular interest. Typical theoretical analysis models the oxide layer as a one-dimensional barrier with an effective barrier height and an effective mass. The barrier height may be obtained experimentally or treated as a fitting parameter, while the effective mass is normally used as a parameter for fitting measured current–voltage (I–V) characteristics. Tunneling coefficients can be calculated using the well-known WKB approximation. Approximate integration of tunneling coefficient curves, with the appropriate Fermi factors describing carrier statistics, then yields an analytical I–V curve formula for the direct-[2] and Fowler–Nordheim[3] tunneling regimes, which can be used conveniently for comparison with experimental data. A somewhat similar treatment uses multiple scattering theory instead of the WKB approximation to compute tunneling coefficients to provide clarification of mechanisms for leakage currents through ultrathin oxides.[4] A still more advanced treatment solves Poisson and Schrödinger equations self-consistently for accumulated layers in MOS devices to calculate tunneling currents.[5]

Nonuniformity in oxides are, typically, not treated theoretically due to the much increased complexity and computational demands. Yet, they can have dramatic effects on the current–voltage characteristics of MOS tunnel structures with ultrathin oxide barriers. One example is interface roughness. If we view interface roughness as local fluctuations in oxide thickness, then this fluctuation as a percentage of total oxide thickness can be quite large in ultrathin oxides. In addition, recently Cundiff and co-workers[6] found experimental evidence that, in typical industrial oxides, roughness at the Si/SiO_2 interface increases with decreasing oxide layer thickness; this further enhances the importance of interface roughness in ultrathin oxides. Another type of nonuniformity is conducting filaments embedded in oxides. It has been shown that constant current stressing of MOS structures in the Fowler–Nordheim tunneling regime can lead to quasibreakdown or breakdown in ultrathin oxides, which are characterized by dramatic increases in leakage currents. Based on experimental observations, several groups have proposed the formation of oxide-embedded conducting filaments as a model for breakdown. Hirose and co-workers[7] proposed that the onset of dielectric degradation takes place rather homogeneously close to the SiO_2/Si interface where the Si–O bonds are heavily strained. Based on their data, they postulated the existence of localized conducting filaments approximately 50 nm in diameter, and extending for no more than 3 nm from the SiO_2/Si interface into the oxide layer. Apte and Sarawat[8] proposed a physical-damage model of dielectric breakdown where the damages in the form of broken bonds in the strained SiO_2 layer near the anode links up with islands of bulk damages to create filamentary paths, which enables excessive conduction. Halimaoui and co-workers[9]

envisioned that the oxide layer contains narrow paths (defects) running from anodes to cathodes. Under currents stressing, they merge to form larger conducting paths, resulting in quasibreakdown; further stressing leads to the merging of quasibreakdown paths and causes breakdown. In this article we use a three-dimensional (3D) model which allows us to analyze the current–voltage characteristics of MOS tunnel structures containing nonuniform oxide layers. Specifically, we examine the cases of interface roughness and conducting filaments.

II. METHOD

Standard treatment uses a one-dimensional potential to describe the oxide barrier. With interfacial nonuniformity, we need to use a three-dimensional description. In principle, variations in the nonuniform potential extend indefinitely in the directions along the interface. In practice, we do not perform computation on an infinite domain, but use instead a quasi-3D supercell geometry to approximate the physical structure. We treat the problem of tunneling through a nonuniform barrier using the open-boundary planar supercell stack method (OPSSM).[10] The device structure treated by OPSSM consists of an active layer sandwiched between two semi-infinite flatband electrode regions. Let the z axis be the direction perpendicular to the interfaces. Then, the active region is composed of a stack of N_z layers perpendicular to the z direction, with each layer containing a periodic array of rectangular planar supercells of $N_x \times N_y$ sites. A one-band nearest-neighbor tight-binding Hamiltonian is used to describe the potential and effective-mass variations over this volume of interest. Our model is formally equivalent to the one-band effective-mass equation[11]

$$-\frac{\hbar^2}{2}\nabla\cdot\frac{1}{m^*(\mathbf{x})}\nabla\psi+V(\mathbf{x})\psi=E\psi, \tag{1}$$

discretized over a Cartesian grid, and subject to periodic boundary conditions (with supercell periodicity) in the x and y directions, and open-boundary conditions in the z direction. Since we are free to choose the values of $V(\mathbf{x})$ and $m^*(\mathbf{x})$ at each of the $N_x \times N_y \times N_z$ sites in our computational domain, we have tremendous flexibility in dictating the geometry of the device structure we simulate. OPSSM solves the quantum mechanical scattering problem exactly for the 3D geometry described by the planar supercell stack, and allows us to compute transmission coefficients with a high degree of numerical accuracy and efficiency. Note that even though the supercell geometry imposes an artificial periodicity to make computations tractable, the use of sufficiently large supercells can minimize supercell artifacts and yield excellent descriptions of the physical problem. Once transmission coefficients are obtained, current densities can be obtained using the standard formula[12]

$$J=\frac{em^*kT}{2\pi^2\hbar^3}\int_0^\infty T(E,V)\ln\left[\frac{1+e^{(E_F-E)/kT}}{1+e^{(E_F-E-eV)/kT}}\right]dE. \tag{2}$$

III. RESULTS AND DISCUSSION

Our model of the MOS tunnel structure consists of an n^+ poly-Si electrode, followed by the oxide layer, and finally, a p-type silicon region. The conduction-band edge is chosen to be at $E_c^M=0$, and the poly-Si Fermi level at $E_F^M=0.1$ eV. The tunneling barrier height at the n^+ poly-Si/SiO_2 interface is taken to be $\Phi^B=3.25$ eV,[7] and the SiO_2/p-Si conduction-band offset is taken as 3.29 eV.[13] The p-Si Fermi level is chosen to be 0.88 eV below the Si conduction-band edge. Thus, at zero gate bias, the p-Si conduction-band edge is 0.98 eV higher than the n^+ poly-Si conduction-band edge, and a bias of $V_{FB}=-0.92$ V is required to bring the oxide into flatband (FB) condition. The effective masses of the n^+ poly-Si, SiO_2, and p-Si are taken to be 1.0, 0.35, and 0.9 m_0, respectively. For convenience, we also assume flatband conditions in the electrodes, and let all the voltage drop occur in the oxide. This should be valid for the high doping levels considered for this structure. We use a cubic mesh with discretization distance of 0.135 75 nm, and 32×32 or 64×64 planar supercells in our simulations. We will consider the characteristics of these structures under negative gate biases (i.e., p-Si lowered relative to poly-Si).

A. Interface roughness

We consider a MOS tunnel structure with a 0.27 nm rough interfacial layer sandwiched in between a 1.36 nm pure oxide layer and the p-Si region. We assume that the rough interfacial layer consists of a 50%–50% mixture of oxide and Si in random configurations. The Si sites, and the oxide sites, for that matter, may aggregate and form patches. We will call the silicon patches islands, and characterize them by their lateral extent (average island size, λ) and the thickness of the interfacial layer (island height, h).

Figure 1 shows the calculated J–V curves for three MOS tunnel structures with rough Si/SiO_2 interfaces characterized by average island sizes of $\lambda=0.33$, 0.97, and 2.95 nm. For comparison, we also construct a reference structure with a smooth interface by replacing the rough interfacial layer with a pure oxide layer of the same thickness (resulting in a total oxide thickness of 1.63 nm). We note that in the direct tunneling regime ($|V_G|<4$ V), current density increases with island size. For instance, at $|V_G|=2$ V, the $\lambda=0.33$, 0.97, and 2.95 nm structures show current densities at 2.7, 3.4, and 4.6 times higher than the reference structure, respectively. This is the result of lateral localization of tunneling electrons, and can be understood by analyzing transmitting state wave functions in the rough interfacial layer. Let the silicon island transmission fraction be defined as the ratio of the sum of silicon site probability densities in the interfacial layer divided by the total probability density in the same. Since in this case the interfacial layer consists of 50% silicon sites and 50% oxide sites, a fraction greater than 0.5 would indicate a preference for transmission through the silicon islands. Figure 2 shows the silicon island transmission fractions as functions of electron incident energy for the three structures at $|V_G|=2$ V and $|V_G|=5$ V, representing direct and Fowler–Nordheim tunneling cases, respectively. At $|V_G|$

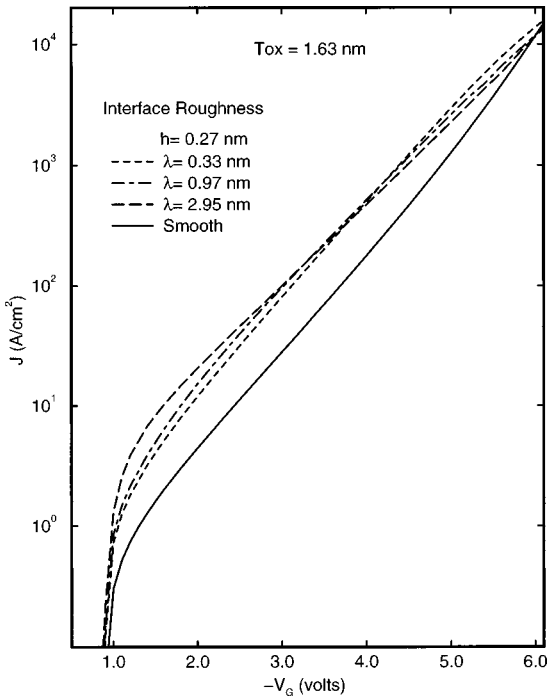

FIG. 1. Supercell calculation of current density–voltage curves for a set of n^+ poly-Si/SiO_2/p-Si tunnel structures with varying degrees of interface roughness. The oxide thickness is 1.63 nm.

$=2$ V, the electron deBroglie wavelength in the Si portion of the rough interfacial layer is approximately $\lambda_e \approx 1$nm. In this instance the Si island transmission fractions are fairly constant in the energy range shown, yielding values of 0.54, 0.65, and 0.87 for the $\lambda = 0.33$, 0.97, and 2.95 nm structures, respectively. Evidently, in the structure with average island size λ much smaller than the electron deBroglie wavelength λ_e, there is only a very slight preference for transmission through the silicon sites. But in the structure with λ considerably larger than λ_e, an electron can readily distinguish the oxide energy barriers form the silicon open pathways, and preferentially traverses the silicon sites to which it is laterally localized. The localization in the more conducting portion of the rough interfacial layer leads to higher current densities for structures with larger islands.

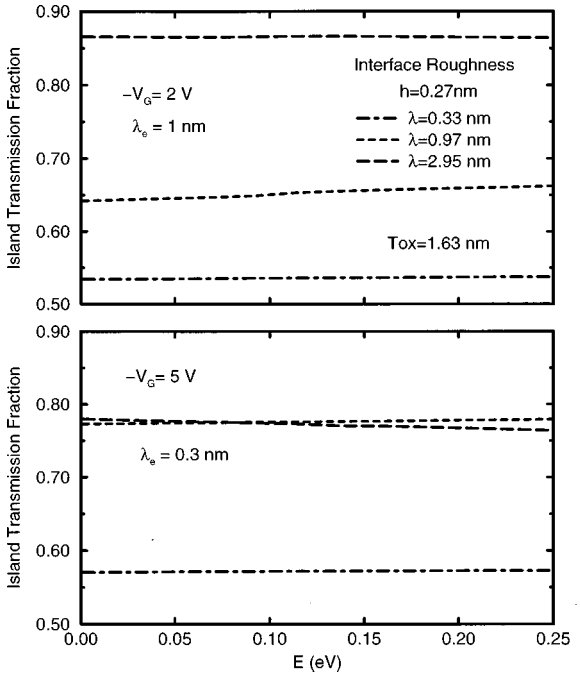

FIG. 2. Silicon island transmission fractions as functions of incident electron energy at two different gate biases for a set of MOS tunnel structures with varying degrees of interface roughness.

The bottom portion of Fig. 2 shows the silicon island transmission fractions for $|V_G| = 5$ V. Comparing to the $|V_G| = 2$ V case, the trailing interface is biased lower with respect to the poly-Si electrode. Thus, upon reaching the rough interface, an electron would be relatively more energetic (we assume it does not suffer inelastic scattering), with correspondingly shorter deBroglie wavelength of $\lambda_e \approx 0.3$ nm. For the structures with $\lambda = 0.97$ and 2.95 nm (both considerably larger than λ_e), the silicon island transmission fractions have comparable values of ≈ 0.77. And for the smaller island structure ($\lambda = 0.33$ mm), the fraction is only ≈ 0.57, which still does not much exceed 50%. Judging by these results, at $|V_G| = 5$ V, we might expect comparable current densities for the $\lambda = 0.97$ and 2.95 nm structures, and a smaller current density for the $\lambda = 0.33$ nm structure. But this is not the case. Figure 1 shows that in the Fowler–Nordheim tunneling regime ($|V_G| > 4$ V), current density decreases with island size; this trend is the opposite of that for the direct tunneling case. Clearly, a physical mechanism other than localization must be invoked in order to explain this behavior.

In the Fowler–Nordheim tunneling regime ($|V_G| > 4$ V), the conduction-band edge at the trailing interface of the barrier (in our case, the rough interfacial layer) is biased below incoming electron energies. Unlike in the direct-tunneling regime, where an electron traverses the oxide portions of the interface with evanescent characteristics, in the Fowler–Nordheim tunneling regime, an electron transmits through both the oxide and silicon portions of the interfacial layer with propagating characteristics. Therefore, the rough interface affects transport primarily through scattering, rather than lateral localization. Here, the larger islands produce more scattering, and thereby reduce the transmission amplitude in the forward direction. Figure 1 shows that, indeed, contrary to the direct-tunneling case, in the Fowler–Nordheim tunneling current densities decrease with increasing island size. Note also that at higher biases the size of the tunneling current is primarily determined by the leading edge of the tunnel barrier. Since the four structures differ only at the trailing edge, the differences in their current densities tend to be less pronounced than in the direct-tunneling regime.

B. Conducting filaments

It has been conjectured that quasibreakdown and breakdown in MOS tunnel structures are the results of current-stressing-induced conducting filaments in the oxide layer.[7–9]

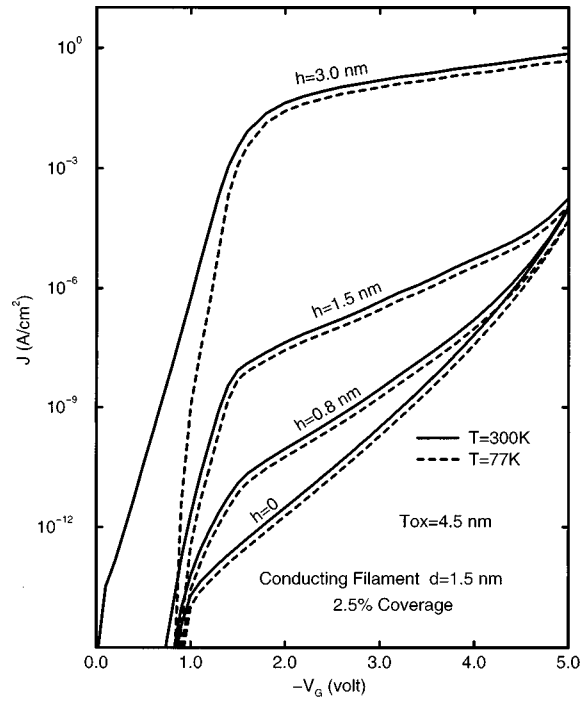

FIG. 3. Calculated current density–voltage curves at 300 and 77 K for a set of n^+ poly-Si/SiO_2/p-Si tunnel structures with oxide-embedded cylindrical conducting filaments with various cylinder heights. Oxide thickness is 4.5 nm. Conducting filaments have a diameter of 1.55 nm, and cover approximately 2.5% of the cross-sectional area.

In our simulation we consider a set of structures with 4.5 nm thick oxides, embedded with cylindrical conducting filaments 1.55 nm in diameter. 64×64 planar supercells were used in our simulations. The filaments account for approximately 2.5% of our computational domain in cross-sectional area, and extend from the SiO_2/Si interface into the oxide layer with cylinder heights of $h=0.8$, 1.5, and 3.0 nm. A fourth, ''undamaged'' ($h=0$) structure is also included for comparison. Because the nature of the filamentary conducting material is not well known, we choose to fill the cylinders with silicon for simplicity. Figure 3 shows the J–V curves for these structures calculated at 300 and 77 K. We note that, in general, current densities increase dramatically with filament length h. This again is due to lateral localization of transmitting state wave functions. We similarly define the filament transmission fraction for a transmitting state as the sum of probability densities over the filament sites, divided by the total probability densities in the filament-containing layers. Table I shows filament transmission fractions for tunneling states with incoming energy equal to E_F^M at gate biases of $|V_G|=2$ V (direct tunneling) and $|V_G|=5$ V (Fowler–Nordheim tunneling). In all cases (with perhaps the exception of $h=0.8$ nm at $|V_G|=5$ V), the filament transmission fraction greatly exceeds 2.5%, the fraction of cross-sectional area occupied by the filaments. This clearly indicates that conduction is strongly localized to the filaments. Table I also shows that the confinement is weaker in the Fowler–Nordheim regime, especially for the shorter filaments. This is because the trailing edge of the oxide does not act as confining barriers when its band edge is biased below the incoming electron energy. Again, since tunneling properties in this regime are primarily determined by the leading edge of the oxide barrier, the $h=0.8$ and 1.5 nm structures show current densities which converge with that of the undamaged ($h=0$) structure at high bias. The filament in the $h=3.0$ nm structure extends sufficiently close to the leading edge of the tunnel barrier so that the large current density increases persist even at higher biases. Our $h=1.5$ nm and $h=3.0$ nm curves bear strong qualitative resemblance to experimental I–V curves for the quasibreakdown[7,9] and breakdown[9] cases, respectively. It is worth noting that within our model we have reproduced both the quasibreakdown and the breakdown behaviors using the same mechanism, with the only difference being whether the filaments extend from the trailing interface far enough into the oxide layer towards the leading interface to have a substantial impact on the Fowler–Nordheim tunneling characteristics. We note that, in particular, we can reproduce the breakdown behavior without using oxides which run through the entire length of the oxide layer; we have kept the filament height to under 3 nm, as suggested by Hirose and co-workers.[7]

TABLE I. Filament transmission fraction calculated using transmitting state wave functions for electrons with incident energy $E=E_F^M$.

$\lvert V_G\rvert$ (V)	$h=0.8$ nm	$h=1.5$ nm	$h=3.0$ nm
2.0	0.43	0.58	0.66
5.0	0.06	0.16	0.64

The $h=3.0$ nm curve differs from the others in its low-bias temperature dependence. In the $h=0$, 0.8, and 1.5 nm cases, the 77 and 300 K results appear essentially the same. In the $h=3.0$ nm curve, however, current densities at low biases ($|V_G|<|V_{FB}|$) increase significantly with temperature. This turns out to be due to resonant tunneling through quantum dots states. We describe their properties in detail elsewhere.[14] Suffice it to say here that the quantum dots are laterally localized in the cylinders, and electrostatically confined along the third direction. When the bias is low ($|V_G|<|V_{FB}|$), the p-Si band edge is actually higher than that of the incoming electrode. Therefore, current contributing resonance levels must be above the p-Si band edge to allow electrons to scatter elastically into the p-Si conduction band through resonant tunneling. This, typically, places these resonance levels far above the Fermi level or the incoming electrode, therefore, their contributions to current densities are highly sensitive to temperature.

In Fig. 4 we examine the dependence of current densities on the fraction of gate area occupied by the filaments. We have performed calculations where we kept the filament diameter constant, but reduced the supercell size from 64×64 to 32×32, thereby increasing filament density by a factor of 4. We find that the corresponding current density increase is almost exactly fourfold in the direct-tunneling regime. The fact that the current density scales linearly with filament density indicates strongly that filament conduction

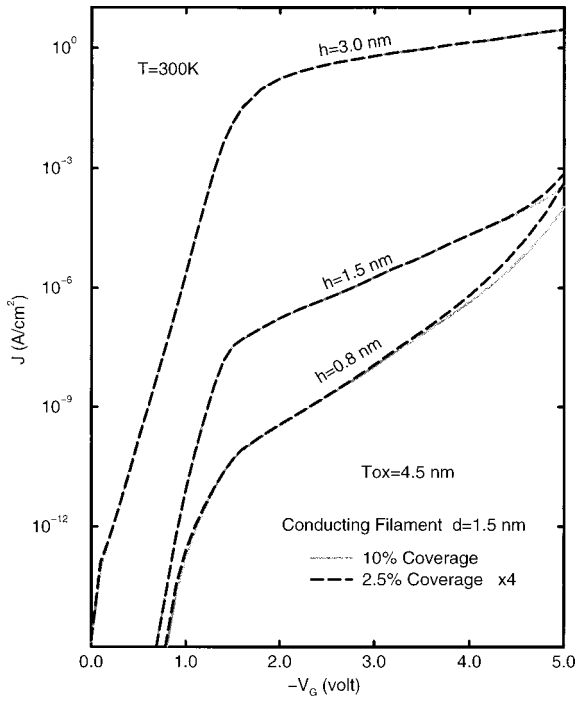

FIG. 4. Calculated current density–voltage curves at 300 K for two sets of MOS tunnel structures with oxide-embedded cylindrical conducting filaments. Both sets of structures have an oxide thickness of 4.5 nm, filament diameter of 1.55 nm, and cylinder heights of 0.8, 1.5, and 3.0 nm. The filaments in the first and second sets, respectively, cover 10% and 2.5% of the cross-sectional area. We have multiplied the second set of current densities by a factor of 4 for ease of comparison.

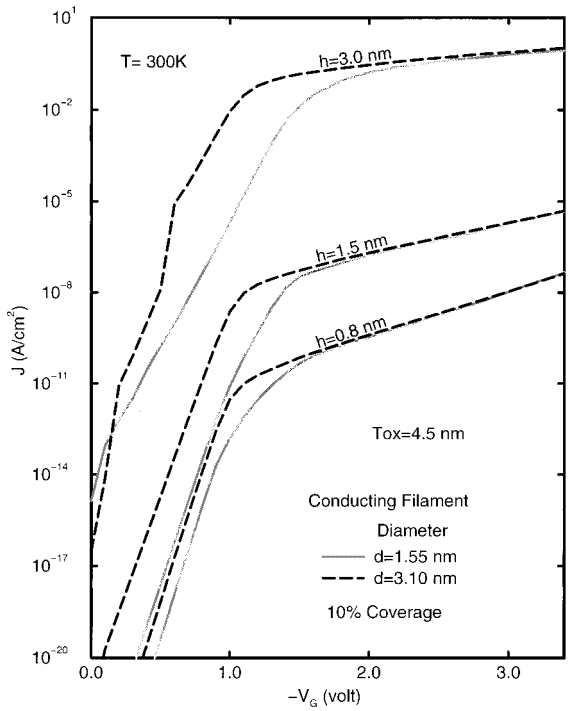

FIG. 5. Calculated current density–voltage curves at 300 K for two sets of MOS tunnel structures with oxide-embedded cylindrical conducting filaments. Oxide thickness is 4.5 nm, and cylinder heights are 0.8, 1.5, and 3.0 nm. In all cases, the cylinders cover 10% of the cross-sectional area, but have different diameters: 1.55 nm for the first set, and 3.10 nm for the second set.

provides the dominant conducting mechanism. In the Fowler–Nordheim regime, the scaling is dependent on filament length. In the case of long filaments ($h=3.0$ nm), the current density scales linearly with filament density well into the Fowler–Nordheim regime. But, in the cases of shorter filaments, the current densities do not scale up as rapidly. In fact, for the $h=0.8$ nm structure, at $|V_G|=5$ V, current densities are essentially independent of filament density, since, again, high-bias currents are primarily controlled by the leading edge of the barrier, which is essentially unaffected by short filaments.

In Fig. 5 we examine the dependence of current densities on filament diameter. We simulated structures with a filament diameter of $d=1.55$ nm using 32×32 planar supercells, and structures with $d=3.10$ nm using 64×64 supercells. In both cases, the filaments cover 10% of the cross-sectional area. In general, the larger diameter filaments result in higher current densities at low bias. But the difference between the current densities in the larger and smaller diameter structures diminishes with increasing gate bias. This is best understood by considering the electron deBroglie wavelength in the filaments. As the gate bias ($|V_G|$) increases, the energy of a typical injected electron increases relative to the band edge of the filament, and the deBroglie wavelength λ_e of the electron decreases correspondingly. At lower biases, λ_e might be comparable to the filament diameters. The transmission properties of the filaments then depend on the number of available modes in the cylindrical filaments, and this generally favors the larger diameter filaments. But at higher biases, λ_e is much shorter than the diameter sizes considered here, and the transmission properties then scale linearly with the cross-sectional area of the filaments.

IV. SUMMARY

We performed 3D quantum mechanical calculations to analyze the current–voltage characteristics of n^+ poly-$Si/SiO_2/p$-Si tunnel structures containing nonuniform ultrathin oxide layers. We find that, in general, roughness at the Si/SiO_2 interface renders the oxide layer more permeable. In the direct-tunneling regime interface roughness induces lateral localization of wave functions, which leads to preferential current paths, and is characterized by current densities which increase with island size. In the Fowler–Nordheim tunneling regime, however, interface roughness affects transport primarily through scattering, which increases with island size, manifesting in current densities which decrease with island size. We have also examined oxide-embedded conducting filaments, and found that they act as localized conduction paths and lead to dramatic increases in current densities. Depending on the filament length, our model can produce current–voltage characteristics reminiscent of those observed experimentally for ultrathin oxides having undergone either quasibreakdown or breakdown. We also found that the lower bias ($|V_G|<|V_{FB}|$) current densities in structures with long filaments are greatly enhanced by resonant tunneling through states identified as quantum dots, and that

this current enhancement is highly temperature dependent. We also report on the dependence of current–voltage characteristics on filament diameter size and filament density.

ACKNOWLEDGMENTS

The authors would like to thank O. J. Marsh, E. S. Daniel, and Z. Q. Zhang for helpful discussions, and M. A. Barton for technical assistance. This work was supported by the U.S. Office of Naval Research (ONR) under Grant No. N00014-89-J-1141, the U.S. Air Force Office of Scientific Research (AFOSR) under Grant No. F49620-96-1-0021, and by the ROC National Science Council under Grant No. NSC 87-2112-M-007-005.

[1]H. S. Momose, M. Ono, T. Yoshotomi, T. Ohguro, S. Nakamura, M. Saito, and H. Iwai, IEEE Trans. Electron Devices **43**, 1233 (1996).
[2]See J. G. Simmons, J. Appl. Phys. **34**, 1793 (1963), and references therein.
[3]See M. Lenzlinger and E. H. Snow, J. Appl. Phys. **40**, 278 (1969), and references therein.
[4]S. Nagano, M. Tsukiji, K. Ando, E. Hasegawa, and A. Ishitani, J. Appl. Phys. **75**, 3530 (1994).
[5]F. Rana, S. Tiwari, and D. A. Buchanan, Appl. Phys. Lett. **69**, 1104 (1996).
[6]S. T. Cundiff, W. H. Knox, F. H. Baumann, K. W. Evans-Lutterodt, M.-T. Tang, M. L. Green, and H. M. van Driel, Appl. Phys. Lett. **70**, 1414 (1997).
[7]M. Hirose, J. L. Alay, T. Yoshida, and S. Miyazaki, in *The Physics and Chemistry of SiO_2 and the Si–SiO_2 Interface—3*, edited by H. Z. Massoud, E. H. Poindexter, and C. R. Helms (The Electrochemical Society, Pennington, NJ, 1996), Vol. 96-1, p. 485.
[8]P. P. Apte and K. C. Saraswat, IEEE Trans. Electron Devices **49**, 1595 (1994).
[9]A. Halimaoui, O. Brière, and G. Ghibaudo, Microelectron. Eng. **36**, 157 (1997).
[10]D. Z.-Y. Ting, S. K. Kirby, and T. C. McGill, J. Vac. Sci. Technol. B **11**, 1738 (1993).
[11]D. J. Ben Daniel and C. B. Duke, Phys. Rev. **152**, 683 (1966).
[12]R. Tsu and L. Esaki, Appl. Phys. Lett. **22**, 562 (1973).
[13]J. L. Alay and M. Hirose, J. Appl. Phys. **81**, 1606 (1997).
[14]D. Z.-Y. Ting and T. C. McGill (unpublished).