Oxide-trap-enhanced Coulomb energy in a metal-oxide-semiconductor system

Ming-Pei Lu and Ming-Jer Chen*
Department of Electronics Engineering, National Chiao-Tung University, Hsin-Chu, Taiwan
(Received 15 April 2005; revised manuscript received 12 October 2005; published 16 December 2005)

Coulomb energy is essential to the charging of a nanometer-scale trap in the oxide of a metal-oxide-semiconductor system. Traditionally the Coulomb energy calculation was performed on the basis of an interfacelike trap. In this paper, we present experimental evidence from a 1.7-nm oxide: Substantial enhancements in Coulomb energy due to the existence of a deeper trap in the oxide. Other corroborating evidence is achieved on a multiphonon theory, which can adequately elucidate the measured capture and emission kinetics. The corresponding configuration coordinate diagrams are established. We further elaborate on the clarification of the Coulomb energy and differentiate it from that in memories containing nanocrystals or quantum dots in the oxide. Some critical issues encountered in the work are addressed as well.

DOI: 10.1103/PhysRevB.72.235417 PACS number(s): 72.20.Jv, 73.40.Qv, 73.50.Td, 73.61.Ng

I. INTRODUCTION

In a metal-oxide-semiconductor (MOS) system, a Coulomb barrier arises during the charging of a nanometer-scale trap in the oxide. Thus, a critical energy to overcome the barrier, namely, Coulomb energy, plays a vital role in the capture kinetics.[1,2] Traditionally the Coulomb energy was calculated on the basis of an interfacelike trap. This treatment essentially remains valid if the oxide used is much thicker. However, with the currently aggressive downscaling of the oxide thickness, the oxide trap is likely situated deeper into the oxide from the SiO_2/Si interface and therefore, the Coulomb energy is expected to be affected due to enhanced image charge. However, little work has been done in this direction since the introduction of the Coulomb energy concept.[1,2] On the other hand, it is noteworthy that the definition of the Coulomb energy in the case of the oxide trap[1,2] is significantly different from that in memories containing nanocrystals or quantum dots in the oxide.[3–6] However, such a confusing issue has not yet been clarified.

In this paper, we exhibit experimental evidence for the Coulomb energy enhancement in the presence of a deeper oxide trap. The other corroborating evidence is achieved based on a multiphonon theory with the configuration coordinate diagrams taken into account. We further elaborate on the clarification of the Coulomb energy in a MOS system containing a nanometer-scale trap in the oxide and differentiate it from that in a MOS memory containing a nanocrystal or dot in the oxide, followed by a concrete discussion on the critical issues encountered in the work.

II. EXPERIMENT

The *n*-channel metal-oxide-semiconductor field-effect transistors (MOSFETs) with varying channel lengths and widths (60 nm to 600 nm) were fabricated in a state-of-the-art manufacturing process. The key process parameters as obtained by capacitance-voltage ($C-V$) fitting were n^+ polysilicon doping concentration $=1.3\times10^{20}$ cm^{-3}, gate oxide thickness$=1.7$ nm, and channel doping concentration $=8\times10^{17}$ cm^{-3}. To detect a potential oxide trap with fluctuating occupancy, the random telegraph signals (RTS) measurement is a good means.[1,2,7–9] The RTS measurement-equipment and method used were the same as that described elsewhere.[10] The operating conditions at room temperature were $V_D=10$ mV and with V_G ranging from 0.2 to 0.4 V. The purpose of the low voltage operation is twofold: (i) it can ensure no extra trap created during the long-term RTS measurement; and (ii) the devices under study can readily reduce to a near-equilibrium one-dimensional(1D)MOS system. We conducted extensive RTS measurement across the whole wafer and found that as expected, the occurrence probability of RTS events in underlying devices is extremely low. For those devices identified with RTS, it was found that (i) the same abrupt transitions between two distinct states in drain current also simultaneously occur in source current; and (ii) no such noticeable changes can be observed in gate or bulk current, opposed to the recent literature[11] with a smaller oxide thickness ($\sim$1.3 nm). Therefore, the RTS events encountered in our work are due to the transfer of a single electron between a certain process-induced defect in the oxide and the underlying conductive channel layer. The capture time associated with the upper level of RTS current and the emission time associated with the lower current level both were exponentially distributed. The mean of the capture time distribution, designated τ_c, divided by the mean of the emission time distribution τ_e is given in Fig. 1 against gate voltage for two devices labeled Traps *A* and *B*. The inset of Fig. 1 shows the corresponding time evolutions of RTS drain current at a certain gate voltage. Figure 1 reveals that while initially the τ_c/τ_e ratio is comparable between Traps *A* and *B*, with gate voltage increasing further, the Trap *B*'s τ_c/τ_e drops with a faster rate than Trap *A*.

III. ANALYSIS AND PHYSICAL INTERPRETATIONS

The size of the trap under study must be significantly less than the oxide thickness used (1.7 nm) since no noticeable change in the gate current was observed. Hence, the trap responsible for the measured RTS in drain current is a nanometer-scale trap. To explore the measured τ_c/τ_e, it is necessary to know in advance the amount of the image or

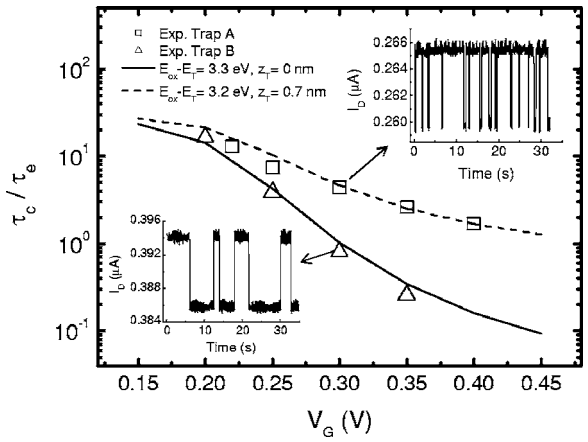

FIG. 1. Measured mean capture time to mean emission time ratio versus gate voltage for two devices labeled Traps *A* and *B*. The inset shows the time records of RTS drain current at a fixed gate voltage of 0.3 V. The fitting lines from Eq. (2) are also shown.

induced charge on the gate as a single electron is inserted into the oxide trap. First of all, it is well recognized that once a single electron is inserted into the oxide, the Debye screening length of a single electron (~70 nm) (Refs. 4 and 6) develops laterally around a *negatively charged* nanometer-scale trap in the oxide. Here, the Debye screening length is the effective size of the "cloud" of the induced charges on the electrodes. Thus, only within the Debye screening length can the plate capacitor approximation readily apply, leading to a capacitive coupling equivalent circuit as shown in Fig. 2. The capacitance model accounts for the effect of the trap depth and the charge sharing between gate, inversion layer, and silicon depletion region. Owing to the insertion of one electron into a depth z_T from the SiO_2/Si interface, the gate oxide capacitance per unit area C_{ox} associated with the oxide thickness t_{ox} can be separated into two distinct components: the trap to anode (near the gate) capacitance per unit area $C_g=C_{ox}t_{ox}/(t_{ox}-z_T)$ and the trap to cathode (near the channel) capacitance per unit area $C_c=C_{ox}t_{ox}/z_T$. The other capacitances such as the inversion-layer capacitance per unit area C_{inv} and the silicon depletion capacitance per unit area C_{dep} can be quantified using a self-consistent Schrödinger-Poisson equations solver with the process parameters mentioned above as input. Figure 3 shows the simulated results

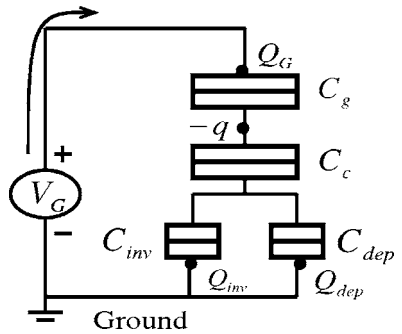

FIG. 2. Capacitive coupling equivalent circuit, accounting for the effect of the trap depth and the charge sharing between gate, inversion layer, and silicon depletion region.

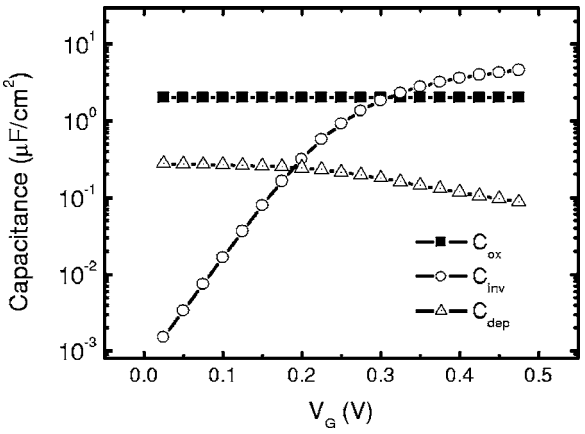

FIG. 3. Simulated results of the key capacitance components versus gate voltage.

of the key capacitance components versus gate voltage. The proposed capacitance model exactly reduces to that by Schulz[1] for the case of $z_T=0$. Indeed, the calculated results on a 17-nm oxide are consistent with those in the literature.[2]

While a single electron is inserted into the trap, the potential change ΔV in the trap reads as $\Delta V=q/(A_{DB}\times C_{eff})$ where A_{DB} is the effective Debye screening area and C_{eff}, the equivalent capacitance per unit area seen from the trap to the ground, can be derived from the model. Then the image charge (positive) Q_G developed on the gate electrode can be expressed as $Q_G=\Delta V\times(A_{DB}\times C_g)$. Combining both equations while eliminating the common factor (i.e., Debye screening area), one achieves $Q_G(=qC_g/C_{eff})$

$$Q_G = q \times \frac{z_T \times (C_{inv} + C_{dep}) + C_{ox}t_{ox}}{t_{ox}C_{ox} + t_{ox}(C_{inv} + C_{dep})}. \quad (1)$$

The calculated gate image charge as depicted in Fig. 4 remains constant until a 2DEG (2D electron gas) layer critically appears (at $V_G\approx 0.1$ V), and then due to increasing screening by the inversion-layer charge, the gate image charge decreases with increasing gate voltage. Specifically, the figure reveals that an increase in the trap depth can substantially increase the gate image charge. In the presence of a 2DEG layer, the source and drain are electrically tied to-

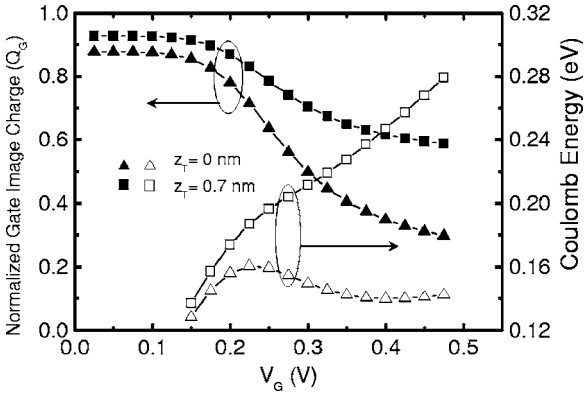

FIG. 4. Calculated gate image charge and Coulomb energy versus gate voltage for two trap depths in the oxide.

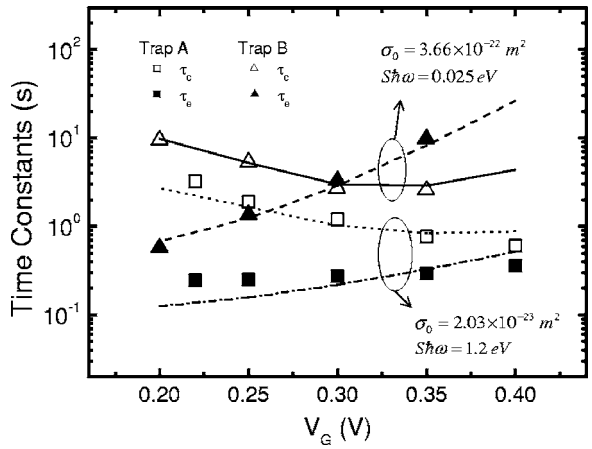

FIG. 5. Comparison of the measured and calculated capture time constants and emission time constants versus gate voltage.

gether and thereby the Coulomb energy can readily be written as $\Delta E \approx Q_G V_G$.[1,2] The calculated Coulomb energy is together plotted in Fig. 4, showing that the Coulomb energy associated with the interface trap increases with gate voltage until encountering a certain peak. However, such a peak point disappears in the case of nonzero trap depth and the Coulomb energy instead piles up over the conventional value.

According to the principle of detailed balance with the Coulomb energy included, the τ_c/τ_e ratio can read as[1]

$$\frac{\tau_c}{\tau_e} = e^{(E_T - E_F + \Delta E)/k_B T}. \tag{2}$$

In Eq. (2), the trap level E_T relative to the quasi-Fermi level E_F is a function of gate voltage and can readily be quantified using the Schrödinger-Poisson solver. The best fitting results achieved using Eq. (2), with $z_T = 0.7$ nm and $E_{OX} - E_T = 3.2$ eV for Trap A and $z_T = 0$ nm and $E_{OX} - E_T = 3.3$ eV for Trap *B*, are shown in Fig. 1. Here E_{OX} denotes the oxide conduction band edge. Evidently, the fitting quality is fairly good. The extracted $E_{OX} - E_T$ values are close to the SiO_2/Si interface barrier height, as expected due to the low voltage operation. It is hence argued that an interface trap exists in the Trap *B* device while a 0.7-nm deep trap in the oxide prevails in Trap *A*. In other words, the conventional Coulomb energy appears to work well for the Trap *B* device but leads to poor quality in fitting the Trap *A* data. Such a remarkable difference in τ_c/τ_e between Traps *A* and *B* can therefore serve as experimental evidence of the Coulomb energy enhancement.

Other corroborating evidence can be obtained through the fitting of the measured mean capture time versus gate voltage as shown in Fig. 5. Since the capture kinetics involve the thermal activation process at room temperature of operation, a multiphonon emission theory was utilized to calculate the capture time

$$\frac{1}{\tau_c} = \sigma \nu_{th} \frac{n_s}{z_{qm}} e^{-\Delta E/k_B T} \tag{3}$$

where ν_{th} is the carrier thermal velocity ($\approx 1.23\times10^5$ m/s), n_s is the inversion-layer electron density per unit area, and

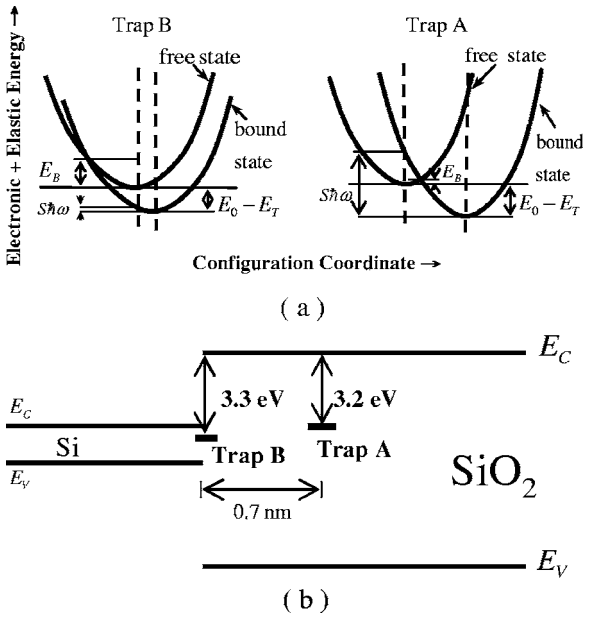

FIG. 6. Schematic configuration coordinate diagrams used for a phenomenological description of the capture and emission kinetics encountered in Traps *A* and *B*. The corresponding energy band diagrams in flatband conditions are also given, schematically showing the trap depth and its energetic level in the oxide.

z_{qm} is the average thickness of the inversion layer. σ is the multiphonon capture cross section and can be written as

$$\sigma = \sigma_0 e^{-E_B/(k_B \times T)}. \tag{4}$$

The prefactor σ_0 involves the interaction between the trap state and free electron wave function.E_B is the thermal activation barrier height and according to multiphonon emission theory the thermal activation barrier height at high temperature ($k_B T > \hbar\omega/2$) can reduce to[12,13]

$$E_B = \frac{(E_0 - E_T - S\hbar\omega)^2}{4S\hbar\omega}, \tag{5}$$

where E_0 is the energy level of the lowest subband for unprimed valley and $S\hbar\omega$ is the lattice relaxation energy (S is the Huang-Rhys factor). Fitting the τ_c data in Fig. 5 to Eq. (3) yielded the lattice relaxation energy $S\hbar\omega$ of 1.2 and 0.025 eV for Traps *A* and *B*, respectively; and σ_0 of 2.03×10^{-23} and 3.66×10^{-22} m^2 for Traps *A* and *B*, respectively. The fitting quality is again good and the same parameters readily reproduced the τ_e data as depicted in Fig. 5. Specifically, the extracted σ_0 values are physically reasonable from the viewpoint of the penetration of the wave function into the oxide: the capture cross section decreases with increasing trap depth from the SiO_2/Si interface. The extracted values of the lattice relaxation energy also correctly reflect the status of the trap: A deeper trap (i.e., Trap A in our work) is accompanied with a higher lattice relaxation energy.[14,15] Using the above extracted results, we constructed a configuration coordinate diagram of the underlying electron-lattice system as schematically shown in Fig. 6 for both devices. Also plotted in Fig. 6 are the MOS energy band diagrams (removing the polysilicon part) in the flatband case, showing the spatial

distance and energetic level of the trap. The calculation results show that the thermal activation barrier E_B of Trap A is substantially smaller than Trap B, as is clearly indicated in Fig. 6.

IV. FURTHER CONSIDERATIONS

A. On the definition of Coulomb energy in trap case

Good reproduction of the measured time constants over gate voltage range, such as those in Figs. 1 and 5, is essential and crucial in the areas of MOSFET RTS. This means that the Coulomb energy involved must quantitatively follow that in Fig. 4. The corresponding Coulomb energy lies between 120 and 280 meV, comparable with that (250 meV) in the similar RTS measurements by Schulz.[1]

As a single electron is inserted into the oxide trap, the total energy of the MOS system will change. The change in energy of the system can be divided into two parts: one is the storage energy and the other is the work done by the voltage source. The change in the storage energy term is

$$\Delta E_S = \frac{q^2}{2 \times A_{DB} \times C_{eff}}. \quad (6)$$

ΔE_S was calculated to have a value of around 1 meV for the Debye screening length of 70 nm, which is negligibly small in magnitude. This means that the Coulomb energy in terms of the work ($\approx Q_G V_G$) done by an external voltage source dominates. Therefore, the definition of $\Delta E \approx Q_G V_G$ as adopted in the areas of MOSFET RTS[1,2] is valid.

B. On the nanocrystals case

There are several fundamental differences between a MOS system with a nanometer-scale trap in the oxide and a MOS system with a nanocrystal or dot in the oxide. First, the self-capacitance of a nanocrystal dot in the oxide can be well linked to the actual dot diameter (this promises applications as a nanoscale floating gate) whereas from the MOS electrostatics point of view, it is the Debye screening length prevailing in the trap case. Second, in our RTS measurement the gate voltage was fixed such as to ensure a quasiequilibrium MOS system; and different gate voltages under such quasiequilibrium conditions produced different RTS data. However, during typical Coulomb blockade experiments on nanocrystalline memories, the gate voltage must continuously change in order to produce a series of Coulomb blockade events. Third, once captured, the electrons essentially remain in the dots (unless a potential leakage is present or the retention time is exceeded); however, this is not the case for the oxide trap, as evidenced by the fluctuating occupancy.

The experimentally determined Coulomb energy in the nanocrystalline dots memories[3–6] ranged from 46 to 168 meV. However, the definition of the Coulomb energy is significantly different from that in Refs. 1 and 2. Instead, an alternative treatment on the basis of the Coulomb blockade theory was widely adopted in the areas of nanocrystalline dots memories. For example, the product of the gate voltage shift between two subsequent Coulomb blockade events and the gate-to-dot coupling coefficient can be directly connected to the critical energy required to overcome the barrier due to the single electron storage energy and the quantum confinement induced energy separation. The single electron storage energy is defined as the Coulomb energy $\Delta E \approx q^2/2C_{dot}$ where C_{dot} is the self-capacitance of the dot. Obviously, different situations encountered can lead to different definitions on the Coulomb energy.

V. CRITICAL ISSUES

A. Screening length

Due to the usage of a heavily doped n^+ polysilicon gate, one may consider the Thomas-Fermi screening length instead as employed in the metal case. However, a self-consistent Schrödinger-Poisson solving over the range of gate voltage under study reveals a band bending across a polydepletion region near the oxide. The corresponding electron density at the interface is found to be about one order of magnitude less than the immobile positively charged impurity concentration. Hence, in the presence of the polydepletion in our work, the Debye-Hueckel screening length considerably applies, which should be much larger than the Thomas-Fermi screening length (of the order of 1 nm) in the metal gate case. To further support this argument, from the measured RTS relative amplitude at $Vg=0.2$ V, we estimate the amount of the affected area to be *at least* 28 and 35 nm across the charged trap for $z_T=0.7$ nm and $z_T=0$ nm, respectively. Thus, the cited 70 nm for the Debye screening length remains reasonable. Even the replacement with a lower value of 28 or 35 nm causes little error.

B. Silicon depletion charge

The Coulomb energy also includes the contribution by the charge induced at the edge of the semiconductor depletion region. The corresponding amount of energy is the product of the induced charge at the edge of the semiconductor depletion region times the difference ($\sim$0.07 eV) between Fermi level and valence band edge at the bulk part of the substrate. The depletion image charge at $Vg=0.2$ V is found to be $0.1e$ and $0.07e$, respectively, for $z_T=0$ and 0.7 nm, and each decreases with increasing gate voltage. As a result, the Coulomb energy due to the depletion image charge becomes of the order of a few milli-electron-volts and drops with increasing gate voltage. Obviously, the role of the charge induced at the edge of the semiconductor depletion region is so insignificant that the depletion image charge can be neglected in the present work.

C. Electron tunneling

First of all, a deeper oxide trap may not always dictate a longer time. According to the configuration coordinate diagrams that describe the electron-lattice coupling, our data point to the opposite case: A deeper oxide trap produces a smaller time constant. This is reasonable since all the extracted parameters can find their physical origins as detailed above. If the electron tunneling were involved only, then the capture time would be the sum of the tunneling time from the

channel conduction band edge to certain oxide depth z_T plus the subsequent multiphonon emission time such as to lower the energy of the tunneling electrons to the same level as the trap. One can estimate the tunneling time of around 10^{-9} sec across z_T of 0.7 nm (Ref. 16) and can reasonably hypothesize that the multiphonon emission time is a spontaneous event (as can be easily understood from the configuration coordinate diagrams in Fig. 6; the hypothesis also works well for the areas of the trap assisted tunneling), leading to a capture time of the order of 10^{-9} sec. Obviously, the possibility of the electron tunneling must in principle be removed since the measured capture times fall within 0.5 to 6 sec. On the other hand, once trapped the electrons may instantly tunnel to the gate electrode, contributing to the gate current. In other words, under such situations, no RTS in drain or gate current can be detected due to the extremely slow detection process in measurement setup. Moreover, in our work the gate current was found to be several orders of magnitude less than the drain current, indicating the absence of the electron tunneling in determining the experimental RTS drain current.

Note that the high and low levels of RTS current represent the different stable states as denoted the free and bound state in the configuration coordinate diagrams in Fig. 6. The detailed balance essentially applies only to two such states, rather than the abrupt transitions between the two. The capture and emission time constants represent the critical times required to overcome the barrier height and reach the crossing point, then instantly entering into the other stable state.

Eventually, the measured discrete switching RTS drain current indicates that the transit time between the high and low levels is substantially less than the integration time in measurement setup.[10] In other words, the abrupt transition between two stable states represents a spontaneous event with respect to the measurement setup. Hence, the corresponding transient displacement current through the gate electrode may escape detection. This explains why we saw only a flat gate current level (with typical thermal or shot fluctuations around it) over the whole observation time.

VI. CONCLUSION

We have presented experimental evidence concerning the Coulomb energy enhancement in a MOS system with a nanometer-scale oxide trap. Other corroborating evidence based on a multiphonon theory has elucidated the measured capture and emission kinetics. The corresponding configuration coordinate diagrams have been established. We have further elaborated on the clarification of the Coulomb energy and have differentiated it from that in memories containing nanocrystals as a floating gate. Some critical issues encountered in the work have been addressed as well.

ACKNOWLEDGMENT

The authors would like to thank Professor Supriyo Datta and Professor Mark Lundstrom of Purdue University for access to the self-consistent Schrödinger-Poisson simulation program. This work was supported by the National Science Council of Taiwan under Contract No. NSC93-2215-E-009-002.

*Electronic address: chenmj@faculty.nctu.edu.tw

[1] M. Schulz, J. Appl. Phys. **74**, 2649 (1993).
[2] H. H. Mueller, D. Wörle, and M. Schulz, J. Appl. Phys. **75**, 2970 (1994).
[3] S. Tiwari, F. Rana, H. Hanafi, A. Hartstein, E. F. Crabbe, and K. Chan, Appl. Phys. Lett. **68**, 1377 (1996).
[4] L. Guo, E. Leobandung, and S. Y. Chou, Science **275**, 649 (1997).
[5] M. Saitoh, N. Takahashi, H. Ishikuro, and T. Hiramoto, Jpn. J. Appl. Phys., Part 1 **40**, 2010 (2001).
[6] S. Huang, S. Banerjee, R. T. Tung, and S. Oda, J. Appl. Phys. **93**, 576 (2003).
[7] K. S. Ralls, W. J. Skocpol, L. D. Jackel, R. E. Howard, L. A. Fetter, R. W. Epworth, and D. M. Tennant, Phys. Rev. Lett. **52**, 228 (1984).
[8] K. R. Farmer, C. T. Rogers, and R. A. Buhrman, Phys. Rev. Lett. **58**, 2255 (1987).
[9] M. J. Kirton and M. J. Uren, Adv. Phys. **38**, 367(1989).
[10] M. J. Chen and M. P. Lu, Appl. Phys. Lett. **81**, 3488 (2002).
[11] A. Avellan, W. Krautschneider, and S. Schwantes, Appl. Phys. Lett. **78**, 2790 (2001).
[12] D. V. Langin, in *Deep Centers in Semiconductors*, edited by S. T. Pantelides (Gordon and Breach, Yverdon, 1992).
[13] B. K. Ridley, *Quantum Processes in Semiconductors*, 3rd ed. (Clarendon, Oxford, 1993).
[14] A. Palma, A. Godoy, J. A. Jiménez-Tejada, J. E. Carceller, and J. A. López-Villanueva, Phys. Rev. B **56**, 9565 (1997).
[15] W. B. Fowler, J. K. Rudra, M. E. Zvanut, and F. J. Feigl, Phys. Rev. B **41**, 8313 (1990).
[16] I. Lundström and C. Svensson, J. Appl. Phys. **43**, 5045 (1972).