Appl. Phys. A 71, 403–414 (2000) / Digital Object Identifier (DOI) 10.1007/s003390000553

Applied Physics A
Materials
Science & Processing

Small silicon memories: confinement, single-electron, and interface state considerations

S. Tiwari[1], **J.A. Wahl**[1], **H. Silva**[1], **F. Rana**[2], **J.J. Welser**[3]

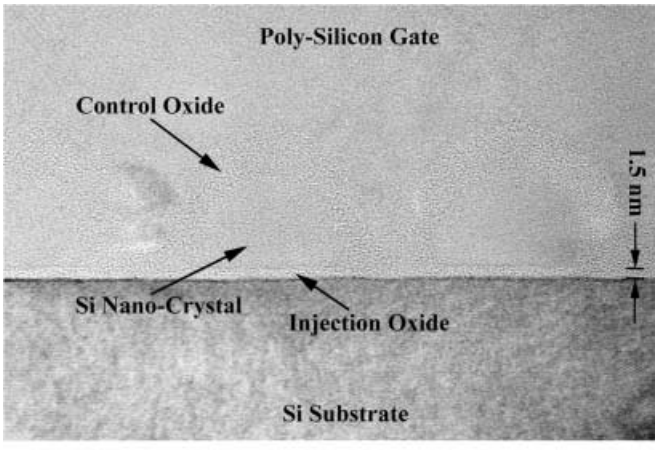
[1] School of Electrical Engineering, Cornell University, Ithaca, NY 14853, USA
[2] Department of Electrical Engineering & Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[3] IBM Research Division, T.J. Watson Research Center, Yorktown Heights, NY 10598, USA
(E-mail: st222@cornell.edu)

Received: 14 April 2000/Accepted: 17 April 2000/Published online: 6 September 2000 –

Abstract. Memories that utilize single-electron effects are an attempt at combining the discreteness observable in transport of electrons on to very small capacitances ($\sim 10^{-18}$ F) and into three-dimensionally quantum-confined states, with the reproducibility, architecture and integration of the field-effect devices. We discuss the role size plays in the operation and its variability for such memories. In particular, we discuss the implications of size effects through barriers on speed; through electrostatics on variability, acceptability and reproducibility of properties desired; through random variations and of tunneling on limits in the use of the field-effect, and through interface-states on the time-domain operation. For device properties and their variations, using silicon-on-insulator substrates, silicon and back-insulator thicknesses matter through the linear variations introduced in the electrostatic potential and quadratic variations introduced in the subband energies, the quantum-dots and nano-crystals matter secondarily through the electrostatics and the linear dependence of capacitance on size and the quadratic dependence of the allowed eigen-energies on size. We also discuss the implications of tunneling on time constants of charging of the confined states and in between the source and the drain for the ultimate structure size limit.

PACS: 85.42.+m; 85.40.Vb; 85.40.Tv

Nano-crystal and quantum-dot memories [1–3] (Figs. 1 and 2) are examples of flash memories that utilize quantum-dot(s) between the gate and the channel of a field-effect transistor to store electron(s), which screen the mobile charge in the channel, thus inducing a change in the threshold-voltage or conductivity of the underlying channel. These quantum-dots are transmissively coupled to the channel and are isolated from the gate, and their processing can be accomplished together with CMOS processing. Their reduced dimension and confinement brings forth two important features that are absent in the conventional floating-gate structures: a reduced density of states that restricts the states available for electrons and holes to tunnel, and the Coulomb blockade effect that arises from a larger electrostatic energy associated with placing a charged particle onto a smaller capacitance. The consequence of these two effects is a reduction of the number of charged carriers used in the operation of the device and results in low-power operation, increased reliability, faster speed due to use of smaller barrier thicknesses, and also a reduction in the "collective phenomena," that we have relied on to achieve reproducibility in microelectronics. Examples of such collective phenomena are the number of electrons flowing through the channel, the number of electrons transferred during a CMOS switching event, and the number of dopants used to control the threshold voltage. A smaller number of electrons flowing in the channel leads to larger fluctuations in the current, a smaller number of electrons transferred during switching leads to larger fluctuations in the switching voltage levels, and a smaller number of dopants leads to larger fluctuations in the threshold voltage. The scaling of device di-



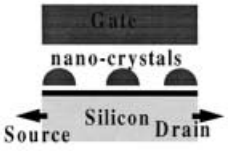


Fig. 1. A cross section of nano-crystal memory

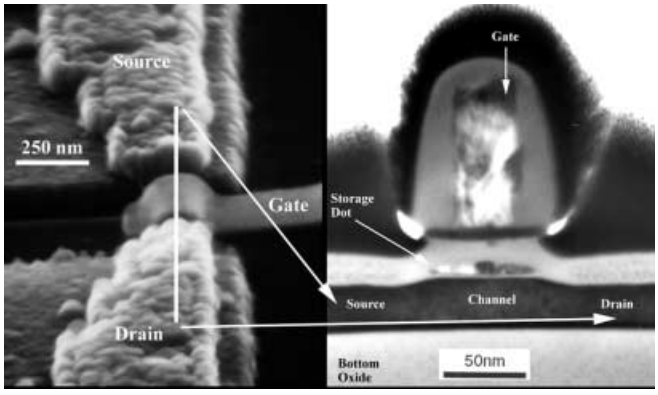


Fig. 2. A cross section of quantum-dot memory

mensions has been driven by higher function and lower cost gained from an increase of device density and performance, a lowering of power density, and mixing of logic and memory technologies. Logic and memory have to co-exist at such small dimensions, and the various forms of memories have to be capable of providing a range of performance from high speed to low power and non-volatility.

This paper discusses the consequences of size reduction on the characteristics and their variability for these devices. We will focus on three aspects. The first will be a description of the capacitances of such structures, while including the quantum-effects, and we will deduce their influence on the operation. The second will be life-time in the storage quantum-dot and the writing and erasing of such structures modeled using a rate equation derived from the equation of motion of the density matrix. This gives a reasonable description of the coupling between the three-dimensionally confined quantum-dot and one-dimensionally confined inversion layer (the writing process) or the unconfined system (the erasure process) and elucidates the role of barriers and the size of the quantum-dot. We will also derive from this the fluctuations in charge, threshold voltage shifts, and hence the noise to be expected during operation. The third will be the role of random defects at small dimensions and their influence on the variance in threshold voltage of the devices. For nano-crystal memories, we will also relate the storage of electrons to the percolation transport in the silicon channel. Finally, we will relate these to measurements on the device structures.

1 Background

In a floating-gate memory, when the gate energy is lowered with respect to that of the source and the drain, electrons transfer to the floating gate storage nodes. For nano-crystal memories, these are small single-crystal silicon islands (nano-crystals of an areal density in the $10^{11}\,\mathrm{cm}^{-2}$ range) that are deposited by chemical vapor deposition on the injection oxide. For the quantum-dot memory, this is a poly-crystalline island patterned at the intersection of the gate and the channel line.

Electrons stored on the island screen the charge in the channel and hence lead to less channel charge for the same applied gate-to-channel potential. This is effectively a change in the threshold voltage. The biggest implication of the scaling of the floating-gate region is that the number of electrons used in the device structure are discrete and small. But, this discreteness, or quantization, is not used directly in the device operation. Instead, it is coupled to the channel of the device. That is, these electrons trapped on the islands influence the conduction of a channel underneath them, and thus the conduction of the channel is a measure of the storage of the electrons. Barriers, used for storage of the electrons are thus important to the write, erase, and the refresh conditions. But, reading of the device, and the amount of signal delivered by the device, are related to the field-effect. The single electron, or the quantum effects, provide a perturbation to it that are detected through the influence of immobile charge on mobile charge. The device behaves as a gain cell and it is limited in size by field-effect.

For a nano-crystal density of ν_{nxt} of size t_{nxt}, a tunneling injection oxide of t_{inj}, a control oxide of t_{cntl}, and $\bar{\nu}$ average number of electrons per nano-crystal, the threshold-voltage is approximately given by:

$$\Delta V_{\mathrm{T}} = \frac{e\bar{\nu}\nu_{\mathrm{nxt}}}{\varepsilon_{\mathrm{ox}}}\left(t_{\mathrm{cntl}} + \frac{1}{2}\frac{\varepsilon_{\mathrm{ox}}}{\varepsilon_{\mathrm{nxt}}}t_{\mathrm{nxt}}\right).$$

The spacing between the nano-crystals should be less than the screening length in order to minimize percolative transport in the channel underneath. Figure 3 shows, at low currents, evidence of this percolation. The oxide in between the nano-crystals is kept large enough to suppress transport directly between the nano-crystals since leakage and subsequent loss to source and drain regions is one of the major methods for loss of charge in floating gate structures. The

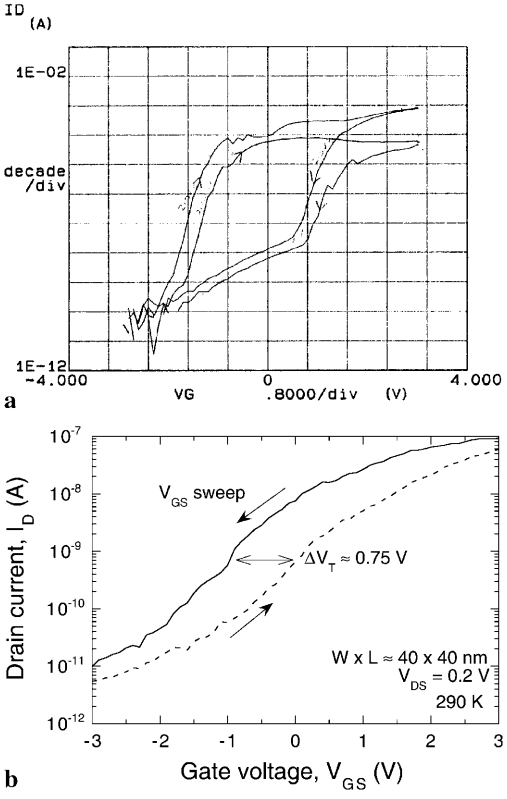


Fig. 3. The characteristics of a nano-crystal memory (*top*, injection oxide thickness $\sim 3\,\mathrm{nm}$) and quantum-dot memory (*bottom*) showing the hysteresis loop

Table 1. Spherical charge in SiO_2 gate stack with a 7-nm gate control oxide

Diameter /nm	C_Σ/aF	E_c/eV	C_{cntl}/aF	E_0/eV	Single electron V_T shift /V
30	6.68	0.011	5.27	~ 0.003	0.03
20	4.45	0.018	2.57	0.007	0.062
10	2.23	0.036	0.71	0.03	0.225
5	1.11	0.072	0.19	0.104	0.84
3	0.68	0.118	0.069	0.29	2.31
2	0.45	0.178	0.031	0.65	> 5
1	0.22	0.364	0.008	2.6	> 10

control oxide is designed to be thick enough (7–15 nm) so that the only path for electron transport to and from the nanocrystals is from the silicon underneath - inversion channel region during injection and depletion region during ejection. The barrier height of Si/SiO_2 interface is large (~ 3.15 eV). Oxide thicknesses in the 1–10 nm range controls the transmission efficiency over nearly 20 decades. This allows the memories to be made volatile and high speed using small injection oxide thickness, to non-volatile and slower speed using large injection oxide thickness.

Table 1 summarizes a number of characteristics for the poly-silicon/control oxide/silicon island/injection oxide/silicon gate stack. This is an approximate calculation meant to point out the main characteristics of the system. The capacitance (C_Σ) is the self-capacitance of the silicon dot used in the calculation of the charging energy (E_c). These vary linearly (for capacitance) and inverse linearly (for energy) with dimension. The quantization effect of confinement in energy (E_0) varies as the inverse square of the dimension. At dimensions below 10 nm, the capacitance is small enough that it requires energy of the order of room-temperature thermal energy to place an electron on the silicon island. As the dimension decreases further, the eigen-energy of the confined states allowed becomes larger than the single-electron electrostatic charging energy. When a single electron is stored on the island, it causes the channel threshold voltage to shift by a magnitude that is inversely controlled by the capacitance C_{cntl}. The shift should not exceed the operating voltages of the structure. Thus, dimensions of between 10 nm and 3 nm are usable and provide a substantial and observable effect. Yano et al. [4] and Nakazato et al. [5] describe other very interesting examples of the use of this large influence in a small dimension.

2 Small dimension effects

2.1 Single-electron effects

Neugebauer and Webb [6] recognized nearly four decades ago that, when a capacitance is reduced, the electrostatic energy required to charge the capacitance ($e^2/2C$) can be made to be of the order of magnitude of thermal voltage at dimensions in the 10-nm range which result in a capacitance in the aF (10^{-18} F) range. This implies that discrete single-electron transmission or storage events can be observed, and unless the electrostatic energy is available to the electron from the power supply, the transition is prohibited. This is known as Coulomb blockade. Fulton and Dolan [7], in 1987, demonstrated the first single-electron transistor, and in recent times, there has been tremendous interest in understanding of this mesoscopic system [8]. Here, we will summarize some the necessary conditions for observation of the single-electron events [9] in order to connect them to the properties of the small silicon memories. For the single-electron events to be clearly observable, a number of requirements must be met. The state that the electron occupies on the particle is confined, and for the event to be observable, the change in system energy upon transmission of an electron, is larger than thermal energy. The uncertainty principle tells us that the width of the eigenstate is $\Delta E \approx h/2\pi\tau$, where $\tau = 1/\Gamma$ is the lifetime (related inversely with the tunneling rate). The change in system energy ($\Delta U = QV = eIR_T$, where Q is the charge, V is the voltage, I is the current ($e\Gamma$), and R_T is the tunnel resistance), upon transition of an electron, is $\Delta U = e^2\Gamma R_T$. The energy width of the eigenstate is, therefore $\Delta E = h/2\pi\Gamma = h/2\pi\,\Delta U/(e^2R_T)$. A clear observation of Coulomb blockade requires $\Delta U \gg \Delta E$ which is the condition

$$R_T \gg \frac{(h/2\pi)^2}{e^2} = 4.1\,\mathrm{k\Omega}\,.$$

Note that this resistance is different from that of quantum resistance, usually alluded to in superconducting tunneling of Cooper pairs, of $R_q = h/4e^2$, which has the magnitude of 6.4 kΩ. So, the first condition is that the resistance of the barrier be larger than 4.1 kΩ. The second condition is that the energy ($e^2/2C$) be larger than kT. The overriding time constant for the transmission of this electron is an RC time constant, which is greater than 100 fs and can, in practical structures, be very large. This time constant dominates since it is larger than the time constant from uncertainty principle, i.e. of the certainty of observation, of $h/(e^2/2C)$, which is of the order of 10 fs, as well as of the transmission of a wave packet through the barrier, of $h/2\pi\,\mathrm{d}(\ln(T(E)))/\mathrm{d}E$ at $E = E_f$, which is very small. This large time constant also implies that a small current flows (1 electron/100 fs is $\sim 1.6\,\mu$A). In most experiments, this is typically a nA. Coupling the effect of stored electrons to field-effect of a transistor allows a larger current because carriers are more mobile in the barrier-free channel and do not have to be limited by the barrier impedence effects.

Now, let us consider a small particle which has these requisite properties. Figure 4 shows the transfer process of electron onto an island with a large density of states, such as a metal. For the moment, we assume that there is no spurious charge (electrically neutral; no trapped electron at interfaces or in the bulk) and hence electric field terminations occur only between the island and the electrodes, when a power supply is connected at the electrodes. Because only discrete tunneling events are allowed for the flow of charge and hence the change in electrochemical potential, the electrochemical potential of the particle – a nano-crystal – follows the inequality $\mu_{nxtl} \le e^2/(2C_\Sigma)$, where C_Σ is the total capacitance between the particle and its surroundings. This is equivalent to having polarization charge, or an offset charge Q_{ofs}, of

$$Q_{ofs} \approx C_\Sigma \frac{|\mu_{nxtl}|}{e} \le \frac{e}{2}\,.$$

Consider first the situation where the offset charge is zero, and C_1 and C_2 are the coupling capacitances. The Coulomb

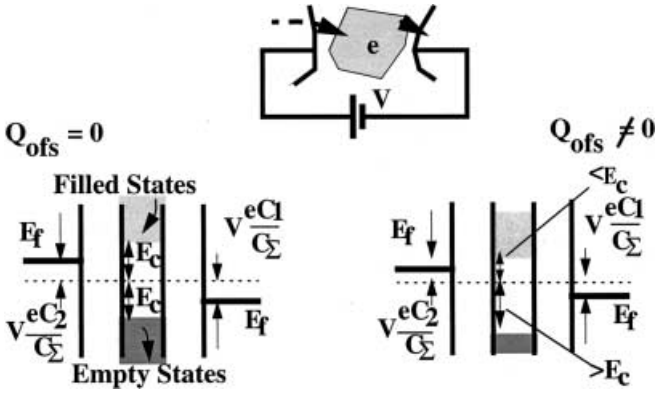


Fig. 4. Schematic of the transfer process of an electron upon application of a bias voltage V between two electrodes that the particle is confined in between. Both diagrams show Coulomb blockade condition, one in the absence of offset charge and another in the presence of offset charge. The capacitances of the two junctions are identified by the subscript 1 and 2

charging energy of E_c is accounted for in the energy diagram by raising the energy of the unoccupied states by E_c. Tunneling to the particle occurs when the energy of states in the lead align with the E_c-shifted unoccupied states of the particle. Tunneling from the occupied states of the particle also occurs with a change in energy of E_c; the energy diagram accounts for this by shifting the energy of occupied states by E_c. So, a barrier E_c exists for flow of an electron whether it is on to the particle or off the particle. For the moment we assume that there is no spurious charge (electrically neutral; no trapped charge at interfaces or in the bulk) and hence electric field terminations are only between the island and the electrodes, when a power supply is connected at the electrodes. Under favorable conditions, an electron tunnels from the left onto the particle with the expenditure of energy E_c by the power supply. The number of electrons in the occupied states has now increased by one and the electrochemical potential of the particle (aligned with the first electrode) is higher than the second electrode. Tunneling can now occur off the island to the second electrode because it is energetically favorable, and the system returns back to its initial state. Similar arguments hold when the first tunneling event is from the particle to the second electrode. In this case, the first tunneling event occurs upon alignment of the energies between the particle and the second electrode. This leads to the lowering of the energy of the particle, and now empty states are available for tunneling from the first electrode. In both cases the Coulomb blockade energy is still $e^2/2C_\Sigma$.

This offset charge represents the polarization of the particle due to electric fields terminated on the small quantum-dot. The electrochemical potential difference, at zero bias, is equivalent to a polarization charge, or an offset charge Q_{ofs}, of

$$Q_{\text{ofs}} \approx C_\Sigma \frac{|\mu_{\text{nxtl}}|}{e} \leq \frac{e}{2}\,.$$

This offset may be intentional, such as from an electric field due to a gate nearby whose potential can be varied as in a single-electron transistor, or it may be due to unintentional causes, such as an electron trapped on a defect or an interface state in the enclosing matrix of barrier. In the event of transfer of an electron from the reservoir on to the island under application of a bias V and in the presence of a polarization charge Q_{ofs}, the change in energy, for $C_1 \gg C_2$, is

$$eV = \frac{(e+Q_{\text{ofs}})^2}{2C_2} - \frac{Q_{\text{ofs}}^2}{2C_2} = \frac{e}{2C_2}\left(1+\frac{2Q_{\text{ofs}}^2}{e}\right),$$

i.e., if the offset charge is $-e/2$, conduction is allowed. And, the largest Coulomb blockade occurs for $e/2$, and is e/C_2. The offset can appear due to polarization induced by a gate. Thus, under gate modulation, it is possible to have a condition where there is no blockade, so the conduction can be modulated from off (blockade) to on condition. In the first single-electron transistor work of Fulton and Dolan [7], performed on aluminum junctions, both the existence of this offset charge and the tunability of this conduction was demonstrated. Likharev [10] provides a very complete description of the transport properties of the system. A more complete calculation of these capacitances using a self-consistent solution of the Poisson and the Schrodinger equation in the Hartree approximation and ignoring the exchange and correlation effects, can be performed using techniques such as that described in [11]. It thus includes the quantization effects. In the case of nano-crystal and quantum-dot memories, the coupling capacitance to the channel is made to be significantly larger, with additional coupling to the other nano-crystals in the vicinity and, in particular, for the end nano-crystals there is stronger coupling to the source and drain reservoirs. The latter is not surprising; it is one of the major mechanisms for leakage of charge in floating gate memories. While these results are secondarily geometry-specific (for example, box, sphere, and hemisphere shapes) because of the different degree of confinement, the estimates of capacitance described before (self-capacitance) are within 20% of the more sophisticated calculations. The coupling capacitances to the channel and source and drain regions, therefore, have a stronger influence on the characteristics.

2.2 Confinement and random effects in semiconductors

Single-charge tunneling, limited by the electrostatic energy argument of above, shows up best in metal systems, where the density of states is enormous and hence confinement does not place severe restrictions on the states occupied by the electron in the island. In practice, we work with semiconductor systems, where the density of states is many orders of magnitude lower. A consequence of this is an additional energy conservation term related to the energy of the confined state occupied by the electron. Hence the arguments of the required bias are modified by the subband energy term. It increases the energy requirement by E_0 for transit of one electron.

Interface states are a major source of the offset charge in these floating-gate structures. If we assume that surface states on the nano-crystals are the largest source of the offset charge, for a cubic quantum-dot, the mean threshold voltage and the standard deviation of the threshold-voltage are:

$$\overline{\Delta V_{\text{T}}} = \frac{e}{\varepsilon_{\text{ox}}}\left(t_{\text{cntl}} + \frac{1}{2}\frac{\varepsilon_{\text{ox}}}{\varepsilon_{\text{Si}}}t_{\text{dot}}\right)\frac{1}{4t_{\text{dot}}^2}\bar{\nu}\nu_{\text{nxt}}\,,$$

$$\sigma(V_{\text{T}}) = \frac{e}{\varepsilon_{\text{ox}}}\left(t_{\text{cntl}} + \frac{1}{2}\frac{\varepsilon_{\text{ox}}}{\varepsilon_{\text{Si}}}t_{\text{dot}}\right)\frac{1}{t_{\text{dot}}}(6N_{\square T})^{0.5}\,,$$

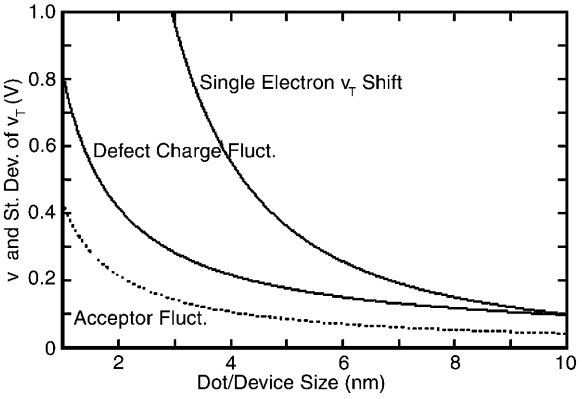


Fig. 5. The threshold voltage and its standard deviation as a function of quantum-dot size for a quantum-dot memory. For comparison the expectation from acceptor fluctuations is also included

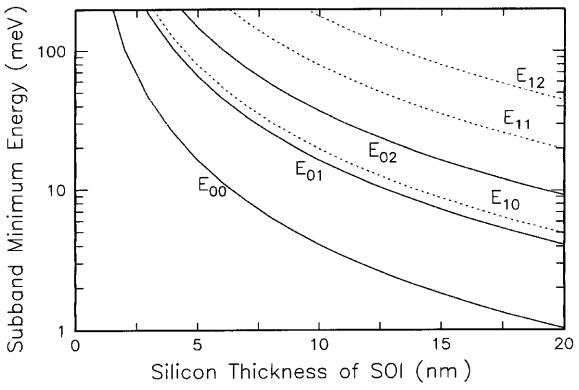


Fig. 6. Subband minimum energy as a function of the silicon thickness of an SOI structure

where $N_{\square T}$ is the interface state density. For oxide/silicon interfaces with thermally grown oxides the interface trap density is typically $\sim 5 \times 10^{10}\,\mathrm{cm}^{-2}\mathrm{eV}^{-1}$ or less. For comparison, the effect of random dopants is given by [12]:

$$\sigma(V_\mathrm{T}) = \frac{e}{C_\mathrm{gate}} \left(\frac{N_A z_\mathrm{d}}{3WL}\right)^{0.5} \left(1 - \frac{z_\mathrm{undop}}{z_\mathrm{d}}\right)^{1.5},$$

where z_d is the depletion region of the retrograde implant and z_undop is the lightly doped region width. Figure 5 compares the fluctuation effect due to defects, random dopants and the V_T shift to be obtained from a single electron. As dimensions decrease, the magnitude of the variations increases.

2.3 Comparison of magnitudes

Stray charge has already been seen to have a significant effect on the characteristics. Since the charging energy varies inversely with capacitance the charging energy varies linearly with dimensional variance, i.e.,

$$\frac{\Delta E_\mathrm{c}}{E_\mathrm{c}} = \frac{\Delta L}{L}.$$

Relative dimensional tolerances are similar to the relative voltage tolerances if the effect is an intrinsic part of device operation. Current, $(e\Gamma)$, is more severely affected by the dimensions and energy of the confining barrier due to the exponential dependence of quantum tunneling on the barrier height. Subband energies vary as inverse square of the dimensions. Thus the relative variation of the energy dependence is

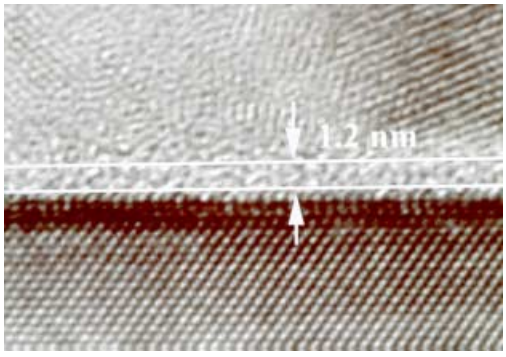
$$\frac{\Delta E_0}{E_0} = \frac{2\Delta L}{L},$$

a requirement nearly twice as strong as due to the Coulomb effects. This places a stronger constraint on devices structures when dimensions are below ~ 7 nm.

The variance in threshold voltage of bulk MOSFETs increases with decreasing dimensions partly because of the random (Poisson) distribution of dopants and the limited number of dopants used to achieve the threshold voltage. A plausible solution to this is elimination of dopants, such as in double-gate [13] and back-plane structures [14]. In the former, the electrostatics of the structure, through control of silicon channel thickness, allows for a normally off device whose threshold voltage is determined by the material parameters. This threshold voltage can not be made very high. In the latter, a back-gate is used and provides the back-barrier and controls the threshold voltage through an applied potential. In either case, the contribution of channel random doping is eliminated. But now, the variance in threshold voltage is determined by the lateral distribution of the dopants in the shallow contact regions, and can be interpreted as an effective channel length variation across the width of the device. Theoretically, such a variance is in the 1–5 mV range for 25-nm junction spacing, instead of the retrograde channel doping contribution of the order of 20–40 mV. Instead of channel doping, in the back-plane geometry (and with some modifications in the double-gate geometry), the thickness variations of silicon, through the linear electrostatic potential change ($\Psi \sim t$ instead of $\Psi \sim t^2$ for doped channels), now contributes to threshold voltage variation. Current SOI structures have a thickness variance of ~ 0.4 nm over 10-$\mu\mathrm{m}^2$ areas. Assuming that this can be improved to 0.3 nm, a 10-mV threshold voltage variance leads to a limit in usable silicon thickness of > 10 nm. Confinement introduces an inverse square dependence on the subband energy. This is worse than the linear electrostatic potential dependence. Figure 6 shows the variation of this energy as a function of the thickness of the silicon channel. Below 5 nm in thickness, the subband energies change very rapidly.

2.4 Tunneling

2.4.1 Tunneling in oxide. At small dimensions, two regions of tunneling are important: oxide tunneling (injection and control oxides) which determines the competing balance between injection and ejection, and between source and drain which determines the smallest limit of the device. Figure 7 shows a cross section of thin gate oxide and calculated tunneling current [15]. As oxide thickness decreases, a large current density can be obtained through thin barriers. Indeed, a 0.15-nm change in oxide thickness leads to a current density change by a factor of 10. Note that a larger roughness arises from the top poly-silicon gate/SiO_2 interface while the



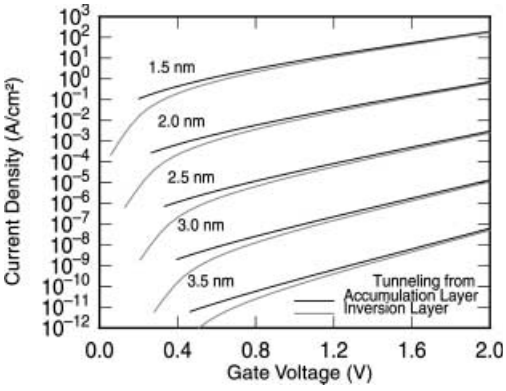


Fig. 7. A cross section of a thin gate oxide and the calculated gate tunneling current from accumulation regions and inversion region

bottom surface is nearly atomically smooth. Random effects should be expected from the random gate control capacitance and dopant activation in this poly-silicon/oxide region.

2.4.2 Tunneling in silicon. With a band gap of 1.1 eV and a low transverse effective mass ($0.19m_0$, Bohr radius ~ 3–4 nm), tunneling between the source/drain regions and the substrate, and from the source to the drain, becomes significant when the distance scales are ~ 10 nm. For bulk nMOSFET structures, inter-band tunneling appears when acceptor doping in channel or halo-doped regions approaches $2 \times 10^{19}\,\mathrm{cm}^{-3}$. This tunneling occurs at the source junction and at the reverse-biased drain-substrate junction, and is a stand-by power constraint similar in nature to that from oxide thickness. In thin silicon structures, inter-band tunneling (conduction to valence band and back) is avoided if the threshold voltage of the device allows $V_\mathrm{D} + (V_\mathrm{G} - V_\mathrm{T}) < E_\mathrm{g}/e$, a condition that prevents tunneling at the drain end, and is equivalent to the threshold voltage not exceeding ~ 0.55 eV for 0.5-V drain bias. This threshold voltage is consistent with requirements of low sub-threshold leakage currents for designs with good sub-threshold swing. So, intra-band tunneling between source and drain through the channel barrier is a fundamental constraint for adequate field-effect operation and needs to be satisfied in the quantum-dot memory. Figure 8 shows tunneling current between source and drain, for a 5-nm-thick sliver of silicon (junctions box doped $5 \times 10^{18}\,\mathrm{cm}^{-3}$ and 10 nm apart) for a quasi-two-dimensional self-consistent Schrodinger–Poisson calculation. The longitudinal masses of the doubly degenerate ellipsoids form the lowest energy ladder with the smaller

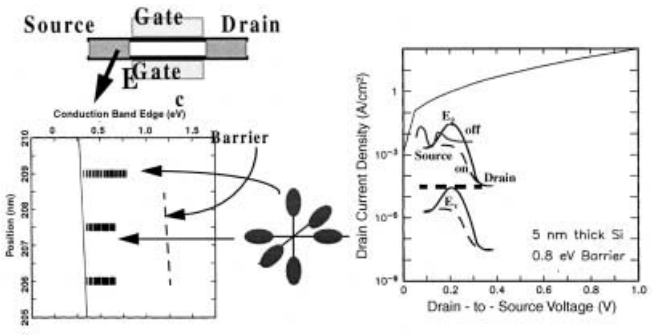


Fig. 8. Model tunneling current calculation between source and drain for a gate bias causing 0.8-eV barrier. Largest tunneling occurs from the low transverse mass valleys

transverse mass available for tunneling. The 4-fold degenerate in-plane ellipsoids cause tunneling through the longitudinal mass. Tunneling from the doubly-degenerate states dominates. This tunneling current, between source and drain, establishes a fundamental constraint of 10 nm for channel length, where the field-effect is not subsumed by tunnel effect, and is a practical constraint at which the stand-by current caused by tunneling leads to too high a stand-by power drain during chip operation. Geometries such as the straddle-gate structure [16] can satisfy this constraint while allowing for a small enough quantum-dot.

3 Quantum kinetic modeling

3.1 Quantum kinetic equation

Details of the modeling of the charging and discharging process are given elsewhere [17]; here we summarize only the salient steps of this modeling, and then discuss the implications. The Hamiltonian for the coupled system consisting of the channel region (in inversion: two-dimensional electron gas) coupled to the quantum-dot is:

$$H = H_{2\mathrm{deg}} + H_{\mathrm{qd}} + H_T \quad \text{or} \quad H = H_0 + H_T\,,$$
$$\text{with} \quad H_0 = H_{2\mathrm{deg}} + H_{\mathrm{qd}}$$

and

$$H_{2\mathrm{deg}} = \sum_n (\varepsilon_n + eV) a_n \dagger a_n\,,$$
$$H_{\mathrm{qd}} = \sum_m \varepsilon_m b_m \dagger b_m + E_\mathrm{s}(\nu)\,,$$

where $E_\mathrm{s}(\nu)$ is the electrostatic energy and m's and n's identify the quantum-dot and the channel states.

$$H_T = \sum_{n,m} T_{nm} a_n \dagger b_m + \mathrm{cc}$$

The state of the system is $|n_n, n_m\rangle$, where n_n and n_m represent the occupation number in the channel and the quantum-dot. It is useful to write time evolution of the density matrix in the Heisenberg representation:

$$\mathrm{i}\frac{h}{2\pi}\frac{\partial}{\partial t}\hat{P}_H(t) = [H, \hat{P}_H(t)]\,,$$

which yields the equation of motion in interaction representation:

$$\frac{\partial}{\partial t}\hat{P}_I(t) = -\mathrm{i}\frac{2\pi}{h}[H_T,\, \hat{P}_I(t_0)] + \left(\mathrm{i}\frac{2\pi}{h}\right)^2 \int_{t_0}^{t} [H_T(t), H_T(t'), \hat{P}_I(t')]\,\mathrm{d}t$$

and which can be formulated as the rate/master equation:

$$\frac{\partial}{\partial t}\boldsymbol{P}(t) = \boldsymbol{W} \bullet \boldsymbol{P}(t)\,.$$

To calculate the time-dependence of the charging and discharging, we first calculate self-consistently for all eigenstates in the channel for a given number of electrons stored in the quantum-dot for all gate voltages, repeat it for the different number of electrons allowed, determine the transition rates, and then determine the time-dependence from the rate equations. Probability of quantum-dot occupation number $\{n_m\}$ is:

$$p(\{n_m\})(t) = \sum_{\{n_n\}} \left\langle \{n_n\}, \{n_m\} \middle| \hat{P}_I(t) \middle| \{n_n\}, \{n_m\} \right\rangle$$

and probability of having ν electrons in the quantum-dot $\sum_m n_m = \nu$ with transition rates

$$p_\nu(t) = \sum_{\{n_m\}} \sum_{\{n_n\}} \left\langle \{n_n\}, \{n_m\} \middle| \hat{P}_I(t) \middle| \{n_n\}, \{n_m\} \right\rangle \sigma\left(\sum_m n_m, \nu\right)$$

determined using the coupling constants and the occupation statistics. The average and variance of electrons is:

$$\bar{\nu} \;= \langle \nu \rangle = \sum_{\nu=o}^{\nu_0} \nu p_\nu\,,$$

$$\sigma_\nu^2 = \langle \nu^2 \rangle - \langle \nu \rangle^2 = \sum_{\nu=0}^{\nu_0} \nu^2 p_\nu^2 - \left(\sum_{\nu=0}^{\nu_0} \nu p_\nu\right)^2 .$$

This now gives us the information from which many of the parameters of interest can be calculated.

3.2 Carrier statistics and charge fluctuations

From the master equation, the stationary solution follows from:

$$\boldsymbol{W} \bullet \boldsymbol{P}(t) = 0\,,$$

where the transition matrix $\boldsymbol{W}$ is of dimensions $\nu_{\max} \times \nu_{\max}$ where $\nu_{\max}$ is the number of electrons in the quantum-dot set for computational tractability, and the vector $\boldsymbol{P}(t) = [p_0(t), p_1(t), p_2(t), \ldots, p_{\nu_{\max}}(t)]$ is the probability of having 0, 1, 2, ... , $\nu_{\max}$ electrons in the quantum-dot.

This now allows determination of the time-dependence, relate to the classical expressions for current, threshold voltage, and also derive the spectrum for single quantum-dot. Figure 9 shows an evolution of electrons in a dot due to injection from inversion layer under three different bias conditions. At the 2-V bias condition, it takes nearly 100 ns before

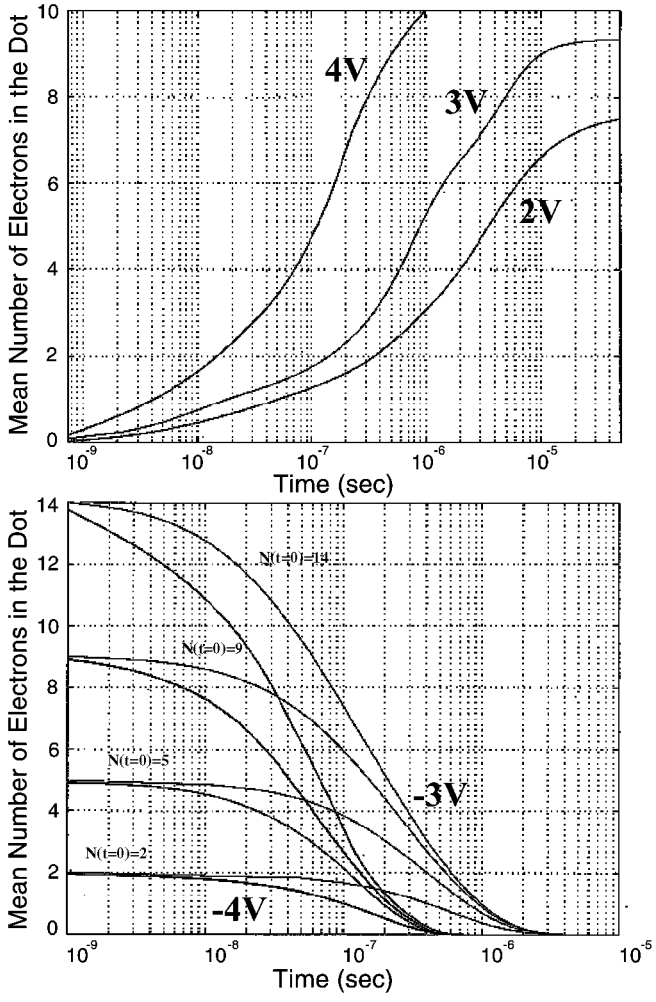


Fig. 9. The evolution of the mean number of electrons in a silicon quantum-dot (10 nm × 10 nm × 6 nm) from an inversion layer due to direct tunneling from an injection oxide of 1.5 nm at three different gate-to-inversion layer potentials. The evolution during ejection is shown on the right

the average reaches one electron. The transition rates are too low because of the large oxide barrier height and small overlap. But, it changes rapidly with bias so that less than 10 ns is needed at 4-V bias. The saturation in number of electrons between 100s of ns to 10s of μs for the differing bias voltages represents the effect of reduced dimensions. As the charging process nears flat-band conditions, the injection process begins to slow down for the same reasons that slow the process at low bias voltages. Now consider the same structure during erasure (Fig. 9) when a negative potential is applied at the gate to eject the electrons into the substrate. A number of starting electrons are considered for two differing voltages. The behavior does not have the detailed features of the injection process; the injection process reveals more of the details of the states being tunneled into. The time constants of ejection are however quite similar to that of injection. At 2 V, not shown, the process has very appreciably slowed down. The lifetime in the dot has become very large.

Figure 10 shows mean and variance in the number of electrons for a calculation in which a maximum of 3 electrons is allowed. The variance is 1/2 at gate voltages where the mean number of electrons is integer+1. The actual number

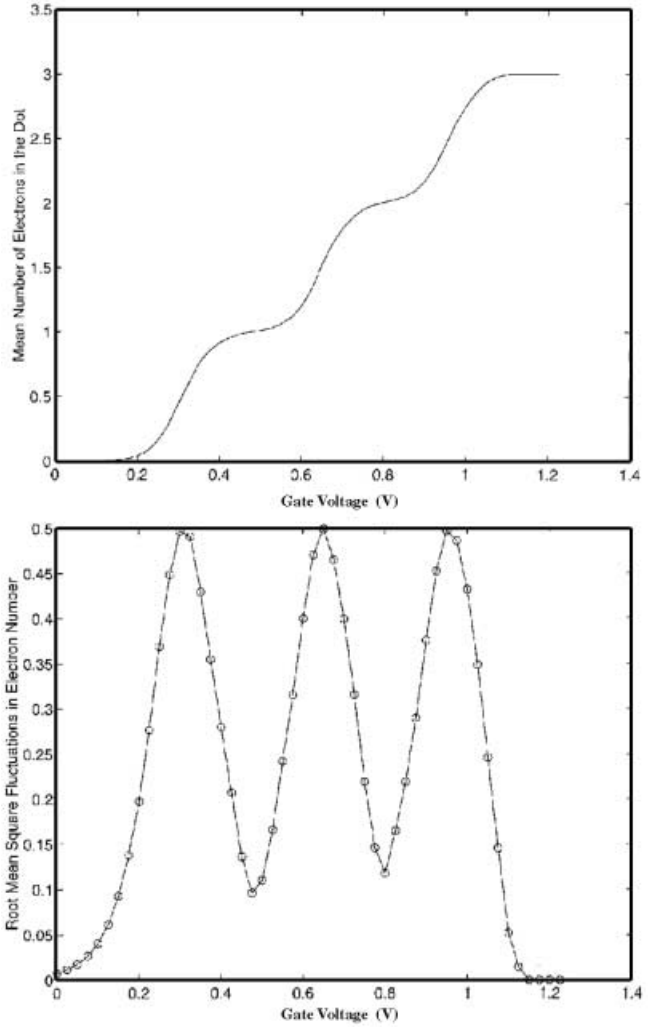


Fig. 10. The mean and variance as a function of gate voltage for occupation in the quantum-dot for a model example of quantum box coupled to the channel and controlled by the gate

of electrons in the quantum dot can take only integer values. A mean number of integer+1/2 implies that the actual number of electrons is fluctuating rapidly between integer and integer+1. The fluctuations in current, etc. follow from these calculations by coupling to the classical transport equations of the field-effect.

4 Experimental observations

Use of laterally uncoupled nano-crystals allows suppression of leakage mechanisms that limit the scaling of injection oxide in conventional floating-gate memory structures – due to leakage to other device regions. Thus, injection oxide can be reduced in thickness together with the use of small number of electrons. For nanocrystals 5 nm in dimension, the Coulombic charging energy is of ~ 0.07 eV and the subband energies ~ 0.1 eV, both larger than thermal energy at room temperature. For such nano-crystals that are 5 nm apart, i.e., a nano-crystal density of $1 \times 10^{12}\,\mathrm{cm}^{-2}$, with a control oxide thickness of 7 nm, the threshold shift is nearly 0.36 V for one electron per nano-crystal. Figure 11 shows

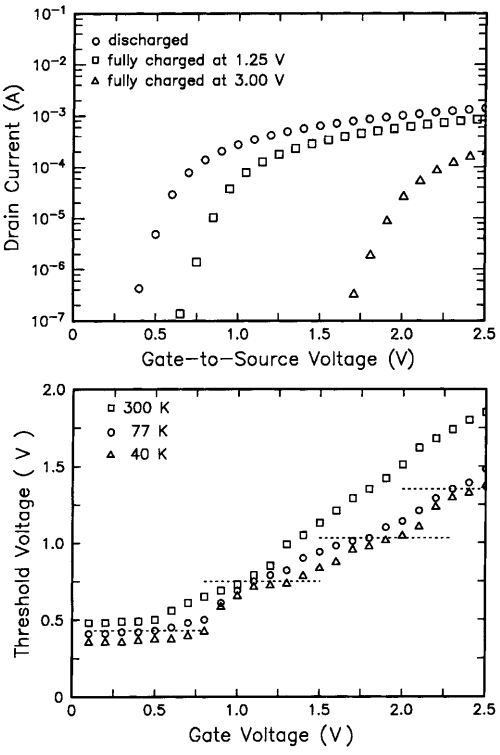


Fig. 11. Bistability in nano-crystal memories resulting from reduced charge storage in the floating-gate region. Injection oxide in this structure is 1.5 nm and it results in volatile operation. On the right is shown threshold voltage shift as a function of applied gate voltage for 300, 77 and 40 K

examples of operational characteristics, measured dynamically, that confirm these expectations. The devices exhibit convergence since, for any applied voltage, only a finite number of carriers can be accommodated. The figure also shows these threshold shifts at three different temperatures. At low temperatures, steps in the shifts, can be discerned corresponding to the storage, on an average, of $1, 2, \ldots$ electrons. While this effect is still present at room temperature, it is masked by the variability in nano-crystal size.

Similar behavior is observable in the single quantum-dot device structures. Figure 12 shows device characteristics emphasizing this behavior.

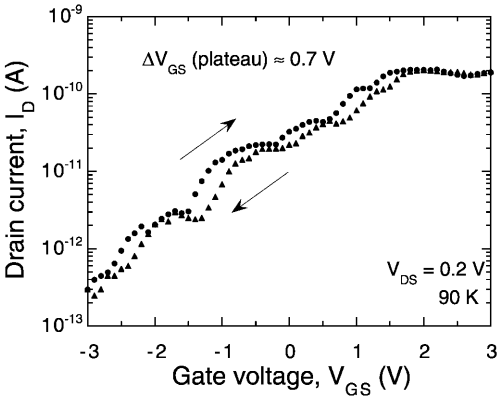


Fig. 12. Single-electron events lead to the discreteness observed in the current–voltage characteristics of the top figure for a quantum-dot memory

The response time – write and refresh – is summarized in Table 2 for approximately comparable nano-crystal density but varying injection oxide thickness in experimental structures. The write times are considerably better than standard flash memory structures, and the voltages are low. This tunability of operation, in power and speed, and operation at small voltages, show the desirability of use of small dimensions (also see [3–5]) in microelectronics because of their voltage, power, and compatibility with present practice of CMOS. We now focus on the experimental properties of the nano-crystal memories that are related to the use of small dimensions, particularly to understand the role of surfaces, smaller statistics, in the presence of confinement and finite charge effects.

The long-term quasi-stability of charge storage in the nano-crystal memory is summarized in Fig. 13 from a 1-MHz capacitance measurement of a large-area and short-gate-length structure. The discharged branch shows measurement of the capacitance at different gate biases over 10-min intervals following establishment of discharged state by application of a −4-V bias. Similarly, the charged branch shows measurements at different gate biases at 10 min intervals following establishment of charged state by application of a 4-V bias. A short and a long time constant are observed in these measurements, i.e., after an initial rapid change, such as the loss of the excess electrons of the charged branch, the electronic state of the structure changes slowly. The slow time-constant changes are most discernible in the voltage range of ±1 V around flat-band condition, a voltage range where the coupling between the nano-crystal and the channel (inverted or depleted) is the weakest. However, the existence of two time constants is indicative of a possible role for interface states. Note, however, that the preferred injection is still through an efficient transmission into the nano-crystal which has a large capture cross-section, and forms a path to possible localization at the interface defect.

The differences in the process of storage and erasure are best represented in the measurements of Fig. 14, which shows the change in capacitance as a function of time for low-voltage (1–2 V) charging and discharging, with the difference in time constants resulting from asymmetry of the conditions (dot-to-inversion region coupling versus dot-to-depletion region coupling).

Similar to the storage curves of Fig. 13, the discharging branch at the higher voltages appears to exhibit two time constants. The storage time data, following charging measurements of Fig. 14, are plotted in Fig. 15 as a function of applied gate-to-source and drain voltage, with the larger voltage results being extracted using smaller applied pulses. The charging follows an exponential relationship at the estimated tunnel oxide thickness of 1.8–2.0 nm, with a control oxide of 7 nm, of this structure. The exponential relationship is indicative that the coupling between the nano-crystals and the inversion region does follow a field-dependence that is linked to the total oxide thickness of the structure.

Table 2. Experimental write and refresh characteristics

Injection oxide thickness	Write condition	Write threshold voltage shift	Refresh time
1.6 nm	200 ns, 3 V	∼ 0.65 V	> 1 wk (RT) ∼ 1 h (85 °C)
2.1 nm	400 ns, 3 V	∼ 0.48 V	> 1 w (RT) ∼ 5 h (85 °C)
3.0 nm	1 μs, 3 V	∼ 0.55 V	large (RT) ≫ 1 h (85 °C)
3.6 nm	5 μs, 4 V	∼ 0.50 V	large (RT) large (85 °C)

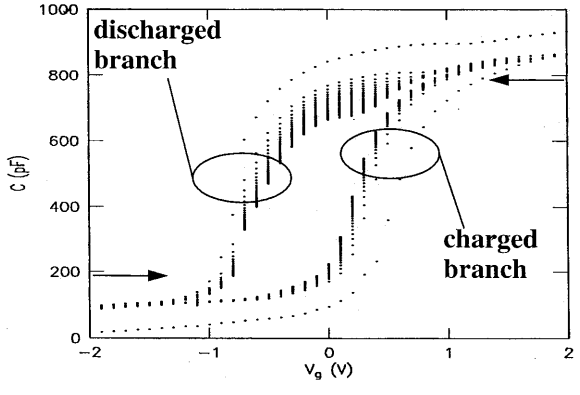


Fig. 13. *C–V* characteristics obtained for the charged and discharged branches for a large-area transistor structure

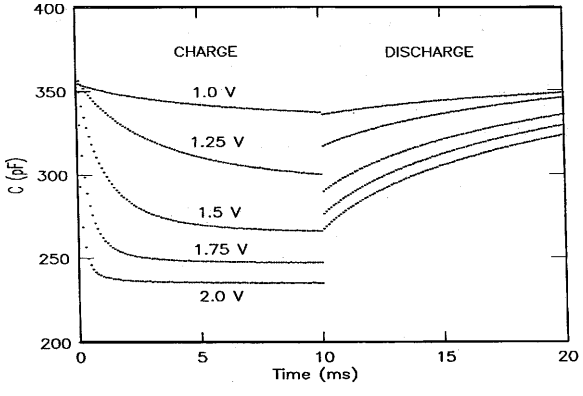


Fig. 14. Capacitance as a function of time during charging and discharging for a small-gate-length and large-area structure

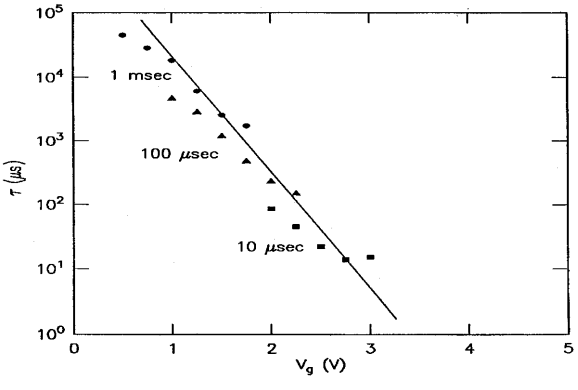


Fig. 15. Charging time constant as a function of gate voltage for the nano-crystal memory. The tunnel oxide thickness is estimated to be 1.8–2.0 nm and the control oxide is 7 nm

Logarithmic transfer characteristics during the charging and discharging are shown in Fig. 16 for a 3-nm tunnel oxide. The characteristics of the discharged branch are obtained by sweeping to a number of different gate voltages after charging at −4 V. The charged branch characteristics are similarly obtained following charging at different gate voltages and then sweeping the gate voltage to −4 V. In the discharged branch, injection begins to occur at the > ms time constant of the sweep at about 0.5 V leading to a dynamic change in the threshold voltage that leads to the decrease in observed drain current. At greater than 2.5 V, the threshold voltage shift saturates at ∼ 1.6 V. In the charged branch, the dynamic changes (ms time-constant) occur in the 0.5 to 2.5 V range. All the charging curves exhibit the presence of inversion layer at their initial applied bias and down to 0.5 to 1 V gate bias. The charging at 0.5 V shows a smaller threshold voltage shift and perhaps percolative transport due to insufficient charging of nanocrystals. These characteristics reinforce a feature of the characteristics of Fig. 13 – charge injection into the nanocrystals still allows the inversion region to persist, charge exchange takes place over ms time constants in the 1.5 to 2.5 V range, and below it is extremely slow.

Figure 17 plots the capacitance as a function of bias for a large area transistor structure. The measurement is made at a slow sweep rate with a bias sweep starting from a discharged condition. The dynamic threshold voltage shift that occurs within 0.5 V of flat-band again appears in these characteristics as a lowering of the net apparent capacitance. At high frequencies, both channel transmission-line and channel nano-crystal charging time effects reduce the measured capacitance. The two plateaus in capacitances near 2.0 V result from response of inversion region as well as of nanocrystals that are now efficiently transmissively coupled to the inversion region. At bias voltages of 2.0 V, there exists an inversion region in the structures as expected from Fig. 16 characteristics also, and as the bias voltage is increased above it, the charge density in the nano-crystals is sufficiently large and coupled efficiently enough to show as an additional increase in capacitance. As the bias is swept from 3 V to 4 V, this excess capacitance corresponds to the storage of mid-10^{11} cm^{-2} electron density in the nano-crystals, or 1–2 additional electrons stored per nano-crystal. The high-frequency behavior shows that the transmission-line effects dominate the nano-crystal coupling effects because of the large channel length.

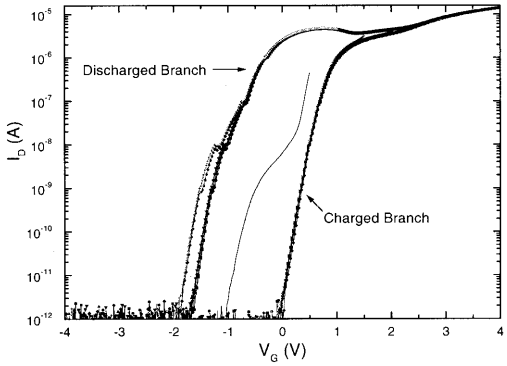


Fig. 16. Current, at slow sweep, as a function of gate voltages for charging and discharging at $V_{DS} = 0.1$ V. The discharged branch is obtained by sweeping from −4 V to 0.5, 1.0, 1.5, . . . , 4.0 V. The charged branch is obtained by charging for extended time at 0.5, 1.0, 1.5, . . . , 4.0 V and sweeping to −4 V. Charge movement can occur between the nano-crystals and the channel during the measurements for certain ranges of voltages and leads to the efficient effects observed at large voltages and hysteresis at low voltages. The extrapolated differences between the two branches allows calculation of the charge stored in the nano-crystals. In the logarithmic curves, at low voltages (0.5 V, middle curve), with less charging of nanocrystals, percolation transport can occur in the channel

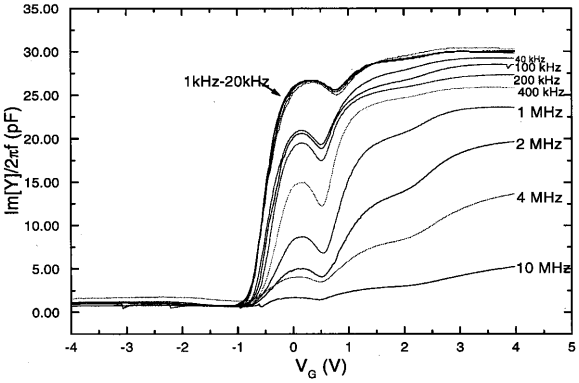


Fig. 17. *C–V* characteristics of a large-area transistor structure for a range of frequencies at a slow sweep rate starting from discharged condition

5 Percolation effects

The nano-crystal memory operates by screening of the channel from the gate by stored electrons in the nano-crystal. The nano-crystals have a disordered distribution on the oxide surface. The occupation of the nano-crystals by electrons is modeled to the first order by the rate equation derived before. When only a fraction of the nano-crystals are occupied, and this is a function of applied bias and time, transport takes place underneath through the unscreened areas. This is similar to the classic problem of bond percolation. The conductance of this area, for example measured at low drain-to-source voltages, should show percolative behavior with a criticality that is dependent on time and voltage. We model this system approximately by assuming a square lattice whose conductance we calculate between the edges. The occupation is determined using the rates, and for this calculation we do not consider any two-dimensional effects along the channel.

Figure 18 shows the application of the rates of charging to a square lattice of the quantum-dots as a snapshot in time for a voltage of 2.0 V between the gate and the channel. The occupation of an electron in a quantum-dot is represented by a filled dot. Its presence depletes the channel area underneath of electrons. A turn-off of the device requires blockage of the resistive conduction path between source and drain. In Fig. 18, this occurs between 10 ns and 100 ns.

Figure 19 shows the dependence of conductance on time and voltage with the other as parameter.

Percolation is clearly observable in these simulations, and a minimum necessary nano-crystal density and voltage and time are needed for reproducible operation. Note, however, that not all nano-crystals have to be filled with electrons for conduction to be shut off. Thus, time scales smaller than those required to fill all nano-crystals are sufficient

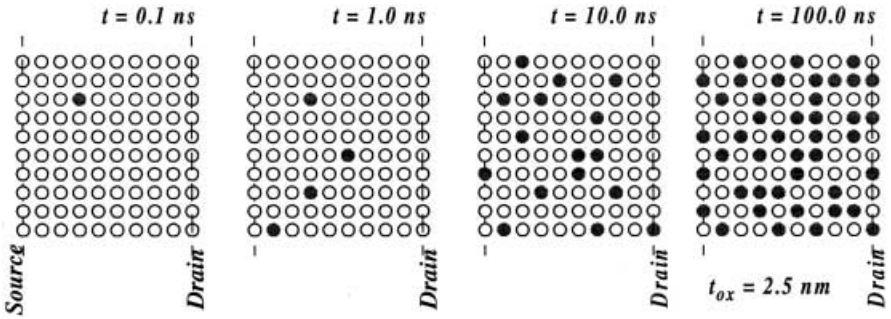


Fig. 18. Charging of a square lattice array of nano-crystals for tunneling injection oxide thickness of 2.5 nm and a gate bias voltage of 2.0 V as a function of time

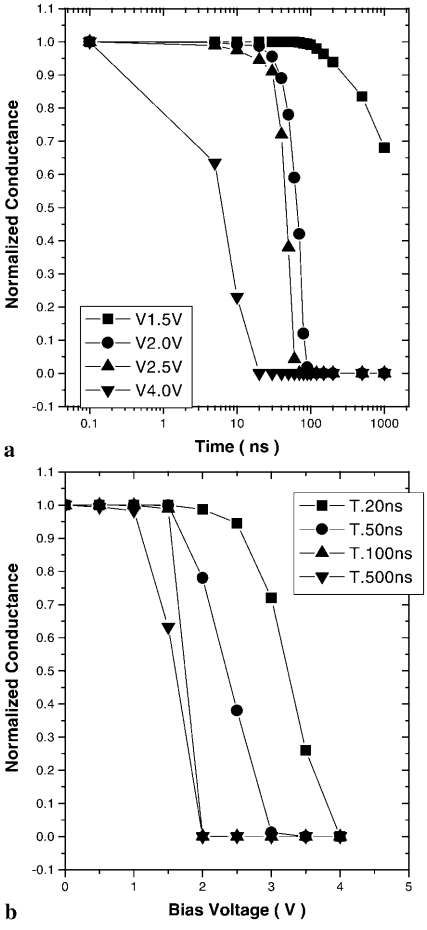


Fig. 19a,b. Normalized conductance as a function of time (**a**) and bias voltage (**b**) for the square lattice of the previous example

and probabilities significantly smaller than unity still allow operation.

6 Can we avoid use of collective phenomena?

Digital microelectronics optimizes device and interconnection technology with the circuit design to obtain the necessary functions at the characteristics desired (usually a combination of several attributes: speed, power dissipation, voltages of operation, density, reliability, noise immunity, cost, etc.). It requires a judicious blend, one of whose important components has been the reduction of dimensions. Smaller dimensions lead to faster devices, lesser area, and lower power – all desirable properties. High yields can be maintained only by improvements in technology and use of device designs that minimize variations arising from technology. CMOS, with the use of rail-to-rail circuits, restores levels and is less sensitive to variations so long as devices maintain reasonable active power gain. The use of collective phenomena has been implicit in this progress because the number of electrons used has always been large enough. It reduces the variances arising from random effects and usually also takes care of variations arising from systematic effects (dimensions, resistances, etc.) arising from the practice of technology. Reduction of dimensions, while reducing the power (usually), also increases the variance due to random effects[1] that appear in statistical fluctuations in voltages (threshold, threshold shifts, etc.) and in currents (magnitude, time and phase). So, implicit in this is a worsening of the noise-margin for operation of devices, in logic and in memory, and minimum voltages that can be tolerated and that can compensate for voltage noise margins. For a CMOS inverter gate driving another inverter gate, the number of electrons that are transferred during a low loaded switching event is nearly 100 electrons at dimensions of ~ 10 nm. Deviations in threshold-voltages change the amount of drive available approximately linearly with an average transconductance as the multiplication factor. A large consequence of any changes in drive current is a proportional change in switching times and hence problems in timing and clocking. For logic, therefore, the consequence of scaling size is serious; however, device design (through width) and circuit design (through careful timing analysis and design) can compensate and allow a functioning design.

For memories, the issues are more difficult. Static RAMs are flip-flops made using CMOS gates. The previous comments apply to it; SRAMs also use lower voltages and are much more sensitive to variances in threshold voltages due to the need to minimize imbalances in the flip-flop. Dynamic memories use a large number of electrons on the capacitors (~ 40 fC of charge, equivalent to $\sim 250\,000$ electrons). Fitting such a large number of electrons, as dimensions are scaled, is an increasingly difficult task because we are demanding that the capacitance not be scaled. An increase in the third dimension is needed and higher and higher aspect ratios are more and more difficult to achieve. An equally big problem is that of retention time. Transistors are designed to have fA of off-state current so that the electrons do not

[1] For a sample N, with Poisson distribution, approximatable by a normal distribution, the standard deviation is $N^{1/2}$

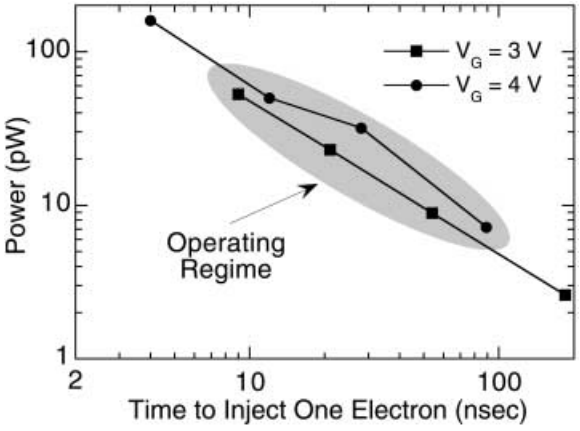


Fig. 20. Power-speed trade-off for single-electron operation in quantum-dot memory

leak easily and refresh cycles are slow. However, changes in threshold voltages change the off-state leakage current exponentially for barrier-modulated transport as is the case for electron diffusion. So, the leakage current changes by approximately a factor of 10 for every $\sim 60\,\mathrm{mV}$ change in V_T. This is very difficult to design around through a worst case design. Quantum-dot memories and the nano-crystal memories attempt to work around these issues by using the oxide barrier for storage of charge and using a gain cell (conversion of the change in electrostatic potential into the current carried by the device) and do it with low power by reducing the number of electrons needed to obtain the memory effect. Two problems arise: one is the need to work with low voltages and the second is the need to work with the smaller numbers and their consequent Poisson variance problem.

First consider the issue of voltages. Smaller size increases the eigen-energies and the electrostatic charging energy as seen in Table 1. The voltages needed for the charging are increased by the lever effect ($\sim t_\mathrm{cntl}/t_\mathrm{inj}$) approximately a factor of 3 to 5. To work with voltages of $\sim 3\,\mathrm{V}$, the charging energies should not exceed ~ 0.6 to $1\,\mathrm{eV}$. To compensate for random variations, in a quantum-dot memory one needs to work with ~ 5 electrons. So, this implies a size in between $10\,\mathrm{nm}$ and $3\,\mathrm{nm}$. These dimensions still maintain a charging energy larger than thermal voltages. For nano-crystal memories, the large number of quantum-dots help in minimizing the variance. A similar size range in dot size is still necessary. Figure 20 shows the power-speed trade-off that comes with the use of a single electron within the size constraints described. Using larger number of electrons scales it higher. However, it is still significantly smaller than that for alternative semiconductor memory structures. It is therefore quite likely that, if we are willing to trade speed for lower power, we will be able to work with smaller number of electrons and still achieve the control on electrical variations desired for microelectronics.

7 Conclusions

As field-effect devices reach their operational – fundamental and practical – limits, and assimilate semi-classical and quantum-mechanical effects in their operation, increased sensitivity in static and dynamic operational fluctuations are inevitable. Thickness and length control of barriers and channels are clearly a very essential requirement and they have an increasing variation due to quantum effects. And, since tunneling currents vary exponentially, the consequences of leakage can also be quite significant. Small silicon memories, such as the nano-crystal and quantum-dot, combine the field-effect with the discreteness that comes from use of small dimensions. For control of electrical variations, they have to solve similar issues as CMOS, and so long as dimensions, interface states, etc., can be controlled, and a few electrons are used, they can be practical and useful.

Acknowledgements. The work summarized in this paper has evolved over a number of years with contributions from many people, and discussions with many have helped clarify ideas. In particular, we express our gratitude to Kevin Chan, Arvind Kumar and the IBM silicon laboatory at Yorktown Heights. Support of IBM and of DARPA through grant N66001-7-1-8908 is also gratefully acknowledged.

References

1. S. Tiwari, F. Rana, K. Chan, H. Hanafi, W. Chan, D. Buchanan: Tech. Dig. IEDM 521, 1995
2. J.J. Welser, S. Tiwari, S. Rishton, K.Y. Lee, Y. Lee: IEEE Electron Dev. Lett. 278, 1997
3. S. Guo, E. Leobandung, S. Chou: Science **175**, 649 (1997)
4. K. Yano, T. Ishii, T. Hashimoto, T. Kobayashi, F. Murai, K. Seki: Tech. Dig. IEDM 541, 1993
5. K. Nakazato, R.J. Blaikie, H. Ahmed: J. Appl. Phys. **75**, 5123 (1994)
6. C.A. Neugebauer, M.B. Webb: J. Appl. Phys. **33**, 74 (1962)
7. T.A. Fulton, G.J. Dolan: Phys. Rev. Lett. **59**, 109 (1987)
8. H. Grabert, M.H. Devoret: *Single Charge Tunneling* (Plenum Press, New York 1992)
9. M.H. Devoret: In *Single Charge Tunneling*, ed. by H. Grabert, M.H. Devoret (Plenum Press, New York 1992)
10. K.K. Likharev: Phys. Rev. B **165**, 821 (1968)
11. A. Kumar, S.E. Laux, F. Stern: Phys. Rev. B **42**, 5166 (1990)
12. D. Burnett, S.W. Sun: Proc. SPIE **2636**, 83 (1995)
13. D.J. Frank, S.E. Laux, M.V. Fischetti: Tech. Dig. IEDM 553, 1992
14. I.C. Yang, C. Vieri, A. Chandrakasan, D.A. Antoniadis: Tech. Dig. IEDM 877, 1995
15. F. Rana, S. Tiwari, D.A. Buchanan: Appl. Phys. Lett. **69**, 1104 (1996)
16. S. Tiwari, J.J. Welser, P. Solomon: Tech. Dig. IEDM 737, (1998)
17. F. Rana, S. Tiwari, J.J. Welser: Superlattices Microstruct. **23**, 757 (1998)