

Layered tunnel barriers for nonvolatile memory devices

Konstantin K. Likharev^{a)}

State University of New York at Stony Brook, Stony Brook, New York 11794-3800

(Received 23 June 1998; accepted for publication 9 August 1998)

Fowler–Nordheim tunneling of electrons through “crested” energy barriers (with the height peak in the middle) is much more sensitive to applied voltage than that through barriers of uniform height. Calculations for trilayer barriers, with layer parameters typical for wide-band-gap semiconductors, have shown that by merely doubling the voltage, the tunnel current may be changed by more than 16 orders of magnitude. It is argued that this effect may be used for the implementation of nonvolatile random-access memories combining a few ns cycle time with a few years retention time and for ultradense electrostatic data storage. © 1998 American Institute of Physics. [S0003-6951(98)02341-9]

Field-induced (Fowler–Nordheim) tunneling¹ is the basic process used for writing and erasing data in electrically alterable floating gate memory cells—see, for example, Refs. 1 and 2. It is also responsible for the main disadvantage of these structures, a long write/erase time, typically in the microsecond range. The goal of this work has been to analyze whether the charge injection process may be sped up significantly by using profiled (“crested”) tunnel barriers with a potential maximum in the middle.

Floating gate memory applications require the tunnel barrier to have negligible tunneling (corresponding to a gate charge retention time of at least 1 year) for relatively low voltages applied to the barrier, $V < V_1$. Parameter V_1 characterizes the maximum voltage during data storage, including that created by the stored charge and external voltages applied to write/erase data in other cells of the same row or column (“half-select crosstalk”). On the other hand, in full-select mode the applied voltage (V_2) should suppress the barrier to such an extent that tunneling current recharges the gate quickly. In order to compete with dynamic random access memory (DRAM) technology for bit-addressable applications, the gate recharging time should be below 10 ns, while in order to keep the memory architecture simple, the ratio V_2/V_1 should be as low as possible (ideally, below 2, allowing us to use just one transistor per cell).

The usual uniform barriers [Fig. 1(a)] cannot satisfy these two conditions simultaneously. Figure 2(a) shows the current density j and the gate recharging time scale

$$\tau(V) \equiv C_0 V / j(V), \quad (1)$$

as functions of voltage V for a typical barrier. [In Eq. (1), C_0 is the capacitance per unit area of the tunnel barrier]. The current has been calculated using the standard quasiclassical approximation, in the assumption of the isotropic and parabolic dispersion law for electrons both in the source conduction band and under the barrier, and taking into account the image charge effect.¹ The results indicate that, for example, a 5-nm-thick barrier of height $U = 3.6$ eV may provide a 3 year retention time ($\sim 10^8$ s) for voltages below $V_1 \approx 3.3$ V, while the write time at $V_2 = 2V_1 \approx 6.6$ V is about 3 ms, far too long for bit-addressable applications. A change in the barrier thickness d to either side only makes the situation

worse [see the results for $d = 10$ nm in Fig. 2(a)]. A change in the height of the barrier (say, to $U = 3.2$ eV typical for SiO_2) also does not change the situation much. This relatively slow dependence of the barrier transparency on the electric field is due to the fact that the highest part of the barrier, closest to the electron source, is only weakly affected by the applied voltage: $U_{\text{max}}(V) \approx U_{\text{max}}(0)$ —see the dashed line in Fig. 1(a).

Now consider a “crested” barrier with the potential barrier height peaking in the middle and gradually decreasing toward the conducting electrodes [Fig. 1(b)]. Figure 2(a) shows that the current through such a barrier changes much faster, so that a voltage change from $V_1 \approx 3.2$ V to $V_2 \approx 5.95$ V $< 2V_1$ decreases the recharging time from 10^{+8} to 10^{-8} s. The reason for this dramatic improvement is that in the crested barrier the highest part (in the middle) is pulled

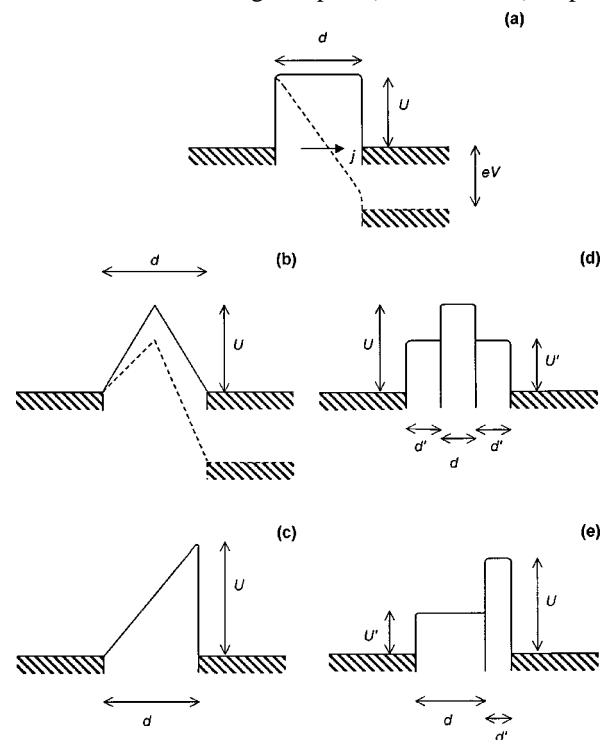


FIG. 1. Conduction band edge diagrams of various tunnel barriers: (a) a typical uniform barrier; (b) idealized crested symmetric barrier; (c) idealized asymmetric barrier; (d) crested, symmetric layered barrier; and (e) asymmetric layered barrier. Dashed lines in panels (a) and (b) show the barrier tilting caused by applied voltage V .

^{a)}Electronic mail: klicharev@ccmail.sunysb.edu

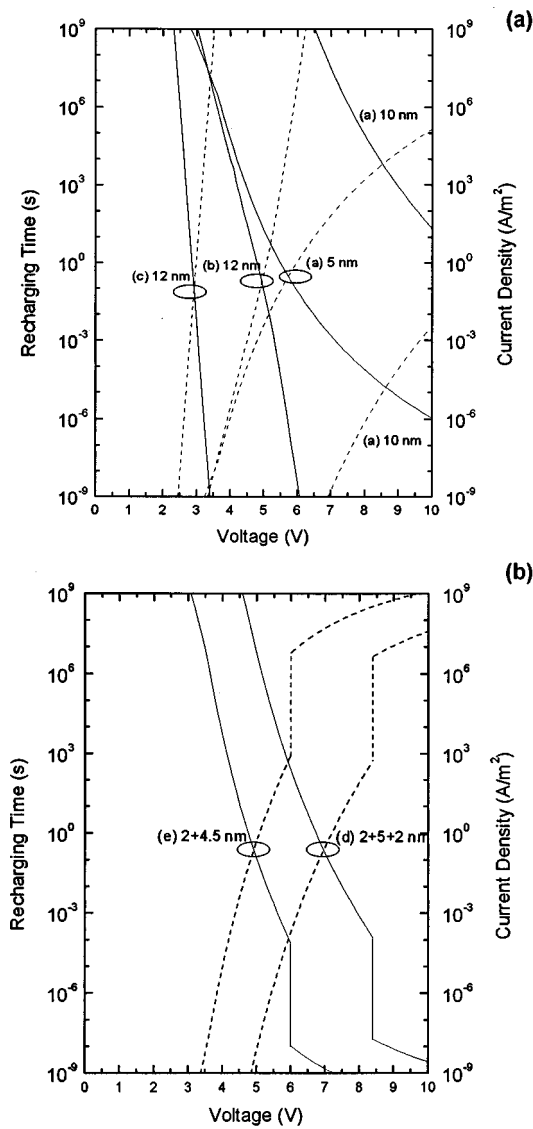


FIG. 2. Tunnel current j (in A/m^2 , solid lines) and time scale τ of floating gate recharging (in s, dashed lines) for various barriers, as functions of applied voltage, calculated using the quasiclassical approximation. The curve labeling corresponds to Fig. 1. The effective carrier mass in the electrodes has been assumed to be isotropic and to equal $0.2m_0$ (modeling n^+ -Si). Other parameters are as follows: for the material with higher barrier $U=3.6$ eV, $m=0.48m_0$, and $\epsilon=8.5$ (the parameters correspond to AlN); for the material with lower barrier, $U'=2.0$ eV, $m'=0.2m_0$, and $\epsilon'=7.5$ (Si_3N_4).

down by the electric field very quickly: $U_{\max}(V)=U_{\max}(0)-eV/2$.

A similar positive effect on the sensitivity to electric field of thermionic emission (which dominates for barriers comparable with $k_B T$) was noticed earlier.³⁻⁵ Moreover, the use of this effect for floating gate memories has been suggested.⁵ However, the authors of that work considered asymmetric triangular barriers [Fig. 1(c)]. Although the injection characteristics of such barriers may be even better than those of symmetric crested barriers [see curves (c) in Fig. 2(a)], this is only true for one current direction (say "write"). The speed of the reciprocal process ("erase") is low, thus excluding the possibility of bit addressable applications. Of course, this opportunity may be restored by connecting two barriers with opposite barrier slopes in parallel, but this option may be too complex for practical applications.⁶

The implementation of crested barriers is straightforward in composite semiconductors, where the barrier shaping may be achieved with either a gradual change of the layer composition during its epitaxial growth^{3,5} or by modulation doping.⁴ However, the maximum barrier height (conduction band offset) available in these materials is too small to provide sufficient retention time at room temperature. For most prospective wide band materials (SiO_2 , Si_3N_4 , AlN, etc.) both these approaches run into fabrication problems; for example for these materials suitable dopants with shallow levels, necessary for modulation doping, have not yet been found.

Fortunately, there is another possible solution to this problem, which seems much more practical. Both the symmetric and asymmetric barriers shown in Figs. 1(b) and 1(c), respectively, may be reasonably well approximated by "staircase" potential patterns formed in layered barriers [Figs. 1(d) and (e)]. I have performed calculations of the function $j(V)$ for the following systems: n^+ -Si/ Si_3N_4 /AlN/ Si_3N_4 / n^+ -Si [trilayer, symmetric barrier, Fig. 1(d)] and n^+ -Si/ Si_3N_4 /AlN/ n^+ -Si (bilayer, asymmetric barrier) within a broad range of layer thicknesses d and d' . This particular set of materials has been selected since both silicon nitride and aluminum nitride had been successfully deposited on silicon substrates, mostly using a variety of chemical vapor deposition techniques—see, for example, Ref. 7. Also, for these materials the relevant data, including the conduction band offsets, effective masses, and dielectric constants, have been published.^{7,8}

Figure 2(b) shows the $j(V)$ and $\tau(V)$ dependences for the sets $\{d, d'\}$ providing the lowest ratio V_2/V_1 for the retention time of 3 years and write/erase time of 10 ns. The sharp current step in each plot is due to the beginning of charge accumulation in a potential dip which is formed at the interface between the first and second layers as a result of potential tilting by the applied electric field. Beyond the step, direct tunneling through the barrier as a whole is replaced with sequential tunneling via the accumulated free electron layer at the interface. In the current version of the theory these steps are vertical, while in reality they would be spread over the voltage range on the order of $\Delta V=(\hbar/d)\times(U/me)^{1/2}$, where m is the effective mass of the electron under the barrier. For the accepted parameters ΔV is small—on the order of 0.03 V.

The calculation results show that with the appropriate choice of layer thicknesses, the ratio V_2/V_1 may be below 2 for barriers of both types, indicating that the time performance specified above may be reached even with the simplest memory organization shown in Fig. 3(a).

One more encouraging result is that the low V_2/V_1 ratio may be combined with a low absolute value of field necessary for fast write/erase: for both cases shown in Fig. 2(b) the field is below 10 MV/cm; for the symmetric, trilayer barrier it is as low as 6.5 MV/cm. At so low an electric field, the hopping (Frenkel-Poole) conductance of the nitrides via deep localized states¹ should not be essential, ensuring very high endurance of the barriers under electric stress.⁹ Finally, it is quite possible that other combinations of materials may have even better performance.

If confirmed experimentally, the acceleration of Fowler-

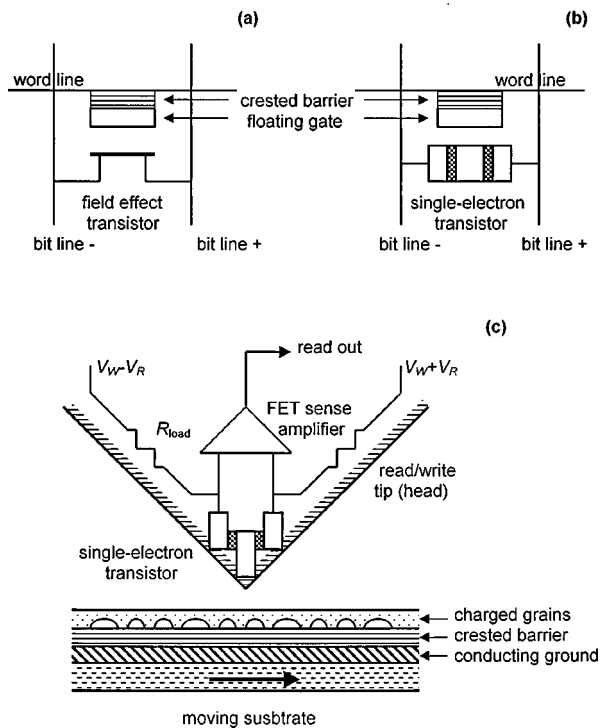


FIG. 3. Possible applications of crested barriers: (a) a cell of the nonvolatile random access memory (NOVORAM), (b) a hybrid SET/FET memory cell, and (c) a system for ultradense electrostatic data storage.

Nordheim tunneling in layered barriers may have several important applications, first of all in bit-addressable memories such as the one shown in Fig. 3(a). Apparently such nonvolatile random-access memory (NOVORAM) may be more dense than DRAM even at the current technological level. Moreover, since NOVORAM cells should not have large storage capacitors, such memory would be more scalable than DRAM. Simple estimates indicate that by using nanoscale ballistic field-effect transistors (FETs),¹⁰ NOVORAM cells may be eventually scaled down to ~ 5 nm minimum feature size and have a density of $\sim 10^{11}$ bits/cm², enabling terabit scale integration.

In the future, crested barriers may also be employed in the recently proposed hybrid single-electron transistor (SET/FET) memories with dynamic readout using a SET as a sense preamplifier/modulator^{11,12} [see Fig. 3(b)]. This type of readout allows the problem of random background charge (which plagues other digital single-electron device concepts) to be circumvented. Recently, a low-temperature prototype of such a memory cell was demonstrated experimentally.¹³ Room-temperature operation of this memory requires a ~ 4 nm technology,¹¹ but if a further reduction in the minimum size becomes available, this memory may be scaled down deeper than NOVORAM, to a density on the order of 10^{12} bits/cm² (at ~ 1 nm minimum feature size).

Finally, the use of SETs may also allow crested tunnel barriers to be used for ultradense electrostatic data storage [Fig. 3(c), Ref. 12]. In this system, a read/write head is flown over a substrate with the crested barrier separating a conducting (ground) layer and a layer of nanometer size metallic grains, not necessarily of similar size or shape. The binary unity is coded by the few-electron charging of a small group of grains. Write 1 is achieved by the application of a sufficiently high voltage V_w to both inputs on the tip. The voltage

suppresses the tunnel barrier and pulls electrons from the ground electrode into the grains. The recorded data may be read out by the application of the opposite voltages $\pm V_r$ to the inputs. This voltage biases the SET which is extremely sensitive to an electric field, in this case created by the grain charge. The SET output signal is further amplified by a closely located FET and then sent out. Recent experiments¹⁴ may be considered as the first step toward application of such readout.

Preliminary estimates show that the electrostatic recording may provide data storage density up to $\sim 10^{11}$ bits/cm², i.e., about two orders of magnitude higher than the presently demonstrated magnetic recording density, provided that the read/write head can be flown at a comparable height (~ 50 nm) above the substrate surface. In contrast to earlier approaches to electrostatic data recording (see, e.g., Ref. 15), the use of crested tunnel barriers may make possible a write/read speed above 300 Mbytes/s per channel, which seems adequate even for the mentioned unparalleled bit density.

Useful discussions with H. Goronkin, T. Ishii, S. Luryi, A. Seabaugh, M. Shur, M. Spencer, S. Sze, and S. Tiwari are gratefully acknowledged. This work was supported in part by ONR/DARPA.

- ¹S. M. Sze, *Physics of Semiconductor Devices*, 2nd ed. (Wiley, New York, 1981).
- ²B. Prince, *Semiconductor Memories*, 2nd ed. (Wiley, Chichester, UK, 1991); in *Nonvolatile Semiconductor Memory Technology*, edited by W. D. Brown and J. E. Brewer (IEEE, New York, 1998).
- ³C. L. Allyn, A. C. Gossard, and W. Weigmann, *Appl. Phys. Lett.* **36**, 373 (1980).
- ⁴R. J. Malik, T. R. AuCoin, R. L. Ross, K. Board, C. E. C. Wood, and L. F. Eastman, *Electron. Lett.* **16**, 837 (1980).
- ⁵F. Capasso, F. Beltram, R. J. Malik, and J. F. Walker, *IEEE Electron Device Lett.* **9**, 377 (1988).
- ⁶A somewhat similar effect may be achieved using grain-enriched interfaces [D. J. DiMaria and D. W. Dong, *J. Appl. Phys.* **51**, 2722 (1980)], electrode surface texturing [H. J. Buhmann, M. Olcer, and M. Ilegems, *Electron. Lett.* **22**, 212 (1986)], or granular floating gates [H. I. Hanafi, S. Tiwari, and I. Khan, *IEEE Trans. Electron Devices* **43**, 1553 (1996)], due to the electric field concentration on convex parts of the conductor surfaces. All these methods, however, are inherently irreproducible due to the randomness of the exact shape of the surfaces.
- ⁷*Silicon Nitride and Silicon Dioxide Thin Insulating Films*, edited by V. J. Kapoor and K. T. Hankins (The Electrochemical Society, Pennington, NJ, 1987), pp. 7, 23; V. I. Belyi, L. L. Vasilyeva, A. S. Ginovker, V. A. Gritsenko, S. M. Repinsky, S. P. Sinita, T. P. Smirnova, and F. L. Edelman, *Silicon Nitride in Electronics* (Elsevier, Amsterdam, 1987), pp. 148, 162; S. Strite and H. Morkoç, *J. Vac. Sci. Technol. B* **10**, 1237 (1992).
- ⁸J. T. Wallmark and J. H. Scott, *RCA Rev.* **30**, 335 (1969); V. W. L. Chin, T. L. Lancey, and T. Osotchan, *J. Appl. Phys.* **75**, 7365 (1994); V. M. Bermudez *et al. ibid.* **79**, 110 (1996).
- ⁹Interface trap charging should be also unimportant for bit-addressable memories with their low duty cycle. For example, for the case presented in Fig. 1(d) the traps on the would discharge through the 2 nm Si₃N₄ layers in less than a ns after the write/erase operation has been completed, making the cell ready for a new cycle.
- ¹⁰L. Guo, E. Leobandung, and S. Chou, *Appl. Phys. Lett.* **70**, 850 (1997); F. G. Pikus and K. K. Likharev, *ibid.* **71**, 3661 (1997).
- ¹¹A. N. Korotkov and K. K. Likharev, in: *Proceedings of the 1995 ISDRS* (University of Virginia Press, Charlottesville, VA, 1995), p. 355.
- ¹²K. K. Likharev and A. N. Korotkov, *VLSI Design* (Amsterdam) **3**, 341 (1997).
- ¹³C. D. Chen, Y. Nakamura, and J. S. Tsai, *Appl. Phys. Lett.* **71**, 2038 (1997).
- ¹⁴M. J. Yoo, T. A. Fulton, H. F. Hess, R. L. Willett, L. N. Dunkleberger, R. J. Chichester, L. N. Pfeiffer, and K. W. West, *Science* **276**, 579 (1997).
- ¹⁵R. C. Barrett and C. F. Quate, *J. Appl. Phys.* **70**, 2725 (1991).