IEEE Circuits and Devices, vol. 16, No. 4, pp. 16-21, July 2000

NOVORAM: A New Concept for Fast, Bit-Addressable Nonvolatile Memories

Konstantin K. Likharev

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

I. Introduction

Field-induced ("Fowler-Nordheim") tunneling is widely used for writing and/or erasing data in nonvolatile floating gate memory cells - see, e.g., Refs. 1, 2. Unfortunately, if standard silicon dioxide barriers are used, the write/erase process is rather long, typically on the order of a microsecond. This drawback restricts the application of floating gate memories to niches were data is written in large blocks (e.g., flash memories [1]) and does not allow them to compete for the much larger market of random access memories [2] for which fast (few-nanosecond) write/erase operations and high cell endurance are necessary.

The reason for the slowness of the write/erase using Fowler-Nordheim tunneling through SiO_2 or any other uniform layer is that such a barrier cannot combine the low transparency necessary for long retention time at low voltages V across the barrier with the high transparency necessary for fast read/erase at acceptable voltages (say, below 10 MV/cm). This statement is illustrated by Fig. 1 which shows the electric field dependence of the tunnel current density j and the floating gate recharging time scale

$$\tau(E) \equiv CV/j(E), \qquad (1)$$

where C is the specific capacitance of the barrier and $V = Ed$ is the applied voltage, for two values of SiO_2 barrier thickness d. One can see that the usual value d = 9 nm the tunnel conductance at E < 4.5 MV/cm is so low that the floating gate charge creating this field may be stored for more than 10 years (the *de-facto* industrial standard for nonvolatile memories). However, a recharging time faster than 1 microsecond requires high fields (E > 14 MV/cm) which are unacceptable because they lead to Frenkel-Pool conductance and hence to low barrier endurance. One might think the situation would be improved by using thinner barriers with higher tunnel transparency. Figure 1 shows, however, that this is not the case: thinner barriers have virtually the same transparency at higher fields (because Fowler-Nordheim tunneling at fixed electric field is essentially independent on barrier thickness), while at lower fields these barriers have additional transparency due to tunneling through the barrier as a whole, so that the retention time degrades.

There are strong (though so far theoretical) indications that the situation may be improved dramatically using the recently suggested [4, 5] "crested" tunnel barriers. The goal of this chapter is to give a brief review of this exciting opportunity. In the next section, basic properties of the crested barriers will be reviewed. In Sec. III, I will discuss possible structure and properties of NOVORAM (nonvolatile random access memory), the simplest memory using the crested barriers. Finally, in Conclusion (Sec. IV) the scaling properties of NOVORAM will be compared with those of DRAM and of hypothetical room-temperature single-electron memories.

II. Crested Tunnel Barriers

In order to suggest a better alternative to the usual, uniform tunnel barriers, let us first analyze the reason why the transparency of these barriers changes so slowly in an applied electric field (Fig. 1). This reason in almost obvious from Fig. 2a which shows the potential energy profile of the usual (rectangular) tunnel barrier: an applied electric field tilts the profile and hence decreases the barrier thickness, but its highest point is close to the electron source, and remains virtually unaffected.

The situation changes immediately if we consider a barrier with a "crested" potential profile peaking in the middle, for example the triangular barrier shown in Fig. 2b [6]. In this case, the maximum height U_m of the barrier becomes quite sensitive to the applied voltage V, decreasing as

$$U_m (V) = U_m(0) - eV/2. \qquad (2)$$

Because of this enhanced barrier suppression, the tunnel current changes much faster [9].

The implementation of crested barriers is straightforward using composite semiconductors such as GaAs/AlGaAs, where the barrier shaping may be achieved with either a gradual change of the layer composition during its epitaxial growth [13, 14] or by modulation doping [8]. 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-bandgap materials (SiO_2, Si_3N_4, etc.) both of these approaches run into fabrication problems; for example suitable dopants with shallow levels, necessary for modulation doping, have not yet been found, to the best of my

knowledge. (One possible exception, which still has to be explored, is $Al_xGa_{1-x}N$ – see, e.g. Ref. 17).

Fortunately, there is another possible solution to this problem, which seems much more practical [4, 5]. The triangular barriers shown in Fig. 2b may be reasonably well approximated by the "staircase" potential pattern formed in layered barriers, for example in a trilayer barrier (Fig. 2c). Figure 3 shows the current density j and gate recharging time scale τ for such a barrier with the following parameters (for definitions, see Fig. 2c): U' = 2.0 eV, m' = 0.2 m_0, ε' = 7.5, d' = 4 nm; U = 3.6 eV, m = 0.48 m_0, ε = 8.5, d = 5 nm. (This particular set of effective masses, conduction band offsets, and dielectric constants has been selected because it corresponds to the published data [15-20] for the $n^+Si/Si_3N_4/AlN/Si_3N_4/n^+Si$ system. These materials seems like good candidates for trilayer barrier implementation, especially because these materials may be grown on silicon and on each other with acceptable speed, using mostly a variety of metal organic vapor deposition techniques – for a review, see Ref. 17.) The current has been calculated using the standard quasiclassical approximation, taking into account the quantization of 1D electron energy in the triangular quantum well at the level interface (shown schematically with bold lines in Fig. 2c) – for details, see Ref. 21 (which describes application of this technique for the calculation of a slightly different phenomenon of resonant electron emission. The quantization results in resonant tunneling of electrons via the corresponding energy subbands and electric charging of the quantum wells by tunneling electrons. In order to obtain the results shown in Fig. 3, this charging has been taken into account in a self-consistent way [22].

Figure 3 shows that this particular barrier may combine an acceptable 10-year retention time for a considerable specific floating gate charge $Q_s = CV_s < 5$ mF/m$^2 \times 5$V $\approx$ 25 mC/m^2 with a few-nanosecond write/erase time at relatively low electric field $F \approx V_w/d$ ~ 5 MV/cm.

III. NOVORAM

If confirmed experimentally, the acceleration of Fowler-Nordheim tunneling in layered barriers may have several important applications including ultra-dense electrostatic data storage [4, 23], acceleration of write/erase in single-electron memories [24, 25], and electronic cooling using resonant Fowler-Nordheim emission [21]. However, I believe that this effect may find its most straightforward application in fast, bit-addressable, floating gate non-volatile random-access memories (NOVORAM) [4, 5].

Figure 4 shows a possible structure of the NOVORAM cell. A floating gate is separated from the readout *n*-type MOSFET by a thick (~12 nm) SiO_2 gate oxide which keeps good channel insulation throughout the device operation, so that both charging and discharging of the gate is carried out through the layered, crested tunnel barrier with a potential profile similar to that shown in Fig. 2c. Data bits are stored as charges of either $+Q_s$ (for binary 1) or $-Q_s$ (for binary 0) of the floating gate, which create at the tunnel barrier relatively low voltages $\pm V_s$ ensuring sufficient retention time. (For the example shown in Fig. 3, V_s should be somewhat below 5 V.)

In order to write 1, the word line voltage is raised to a value $+V_w > V_s$ (for the example shown in Fig. 3, V_w should be close to 10 volts) while the potential of source and

drain of the transistor are lowered to about $-V_w$. Since the capacitances of the crested barrier and the gate oxide are close to each other (the former layer should be somewhat thicker than the latter, but its materials are likely to have higher dielectric constants), the initial voltage drop applied to the barrier is close to V_w, ensuring its rapid (few-nanosecond) charging to $Q \approx +Q_0$. The erase (or rather write 0) operation is completely symmetric and may be carried out with the same high speed.

In order to read the stored information, the selected word line and the source are kept at zero potential (while the deselected word lines are biased by $-V_w$) and the drain is biased by a drain voltage $+V_r$. (For the current 0.18-0.25 μm fabrication technologies, V_r should be between 1 and 2 volts, but in future it may be scaled down in proportion with the MOSFET channel length). At this bias, in cells storing binary 1 and connected to the selected word line the floating-gate-to-source voltage V_g is only slightly above $+V_s$ and hence is reliably above the transistor threshold voltage V_t (on the order of +1 volt [26]). As a result, the readout transistors in these cells are open. At the same time, in the selected cells storing binary 0, $V_g \approx - V_s$ is well below V_t and their transistors remain closed. In cells of a deselected word line, the voltage $-V_w$ applied to the line brings the net potential of the floating gate to (V_s - V_w), the value which is below V_t even for the cells storing charge $+Q_s$ (even more so for the cells with charge -Q_s), so that their transistors remain closed regardless of the cell contents.

In order to understand why the operation described above is sufficient for NOVORAM operation, let us examine the possible memory matrix organization (Fig. 5) which is essentially the generic NOR architecture for floating gate memories [1]. It is

straightforward to check that if the ratio V_w/V_s is below 2 (as it is for the case presented in Fig. 3) the write operations described above do not cause harmful disturb effects in semi-selected cells, while the read operation does not degrade the retention time considerably. Hence, no disturb-compensation schemes are necessary.

This simple NOVORAM architecture should allow very compact cell layout, like that shown in Fig. 6 for SOI-based MOSFET technology. (This example has been selected because it allows scaling of NOVORAM deep into the 10-nm region – see Sec. IV below.) Even this relatively conservative layout allows the cell area to be kept within $8F^2$, where F is the minimum feature (half-pitch) size. As a result, NOVORAM may be more dense than dynamic random access memory (DRAM) even at the current technological level, being in addition non-volatile and potentially faster than DRAM.

IV. Discussion

Figure 7 shows the projected scaling of NOVORAM using the $8F^2$ rule, together with projections for DRAM [27] and single-electron (more exactly, hybrid SET/FET) memories [24]. Since NOVORAM cells should not have large storage capacitors, this memory is inherently scalable (while DRAM is not, due to the need for a nearly-fixed value of its storage capacitor). On the other hand, in contrast to room-temperature single-electron memories, NOVORAM can be implemented with the current level of patterning technology, and hence provide a convenient evolutionary way to electronic circuits of extremely high integration scale.

The upper limit of NOVORAM density will probably be determined by the following factors:

(i) As the floating gate size approaches ~10 nm, the number $N = Q_s/e$ of electrons corresponding to the stored charge, approaches ~10. At this state the r.m.s. fluctuation of N may become so large that the rate of soft errors exceeds the level which may be compensated by acceptable redundancy techniques. (Using Coulomb blockade effects for storing just one or a few electrons is possible as a matter of principle, but is hampered by the background charge randomness – for details, see Ref. 25.)

(ii) If readout MOSFETs are implemented using currently available materials, they can hardly be scaled down below ~10 nm gate length [28, 29].

If no ways to circumvent these problems are found, they will limit the NOVORAM density at a level about 10^{11} bits/cm^2 [Fig. 7], somewhat lower than that for the SET/FET memories. However, even this conservative estimate implies memory chips with integration scale well above 1 terabit per die.

To summarize, I believe that NOVORAM is a convenient new paradigm which may enable the Moore-law-type progress of semiconductor memory technology to be extended well into the nanoscale, terabit range.

Acknowledgments

Important contributions by Alexander Korotkov, and useful discussions with J. Brewer, 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 DARPA (via ONR) and SRC.

References

[1] *Nonvolatile Semiconductor Memory Technology*, ed. by W. D. Brown and J. E. Brewer (IEEE Press, New York, 1998).

[2] A. K. Sharma, *Semiconductor Memories* (IEEE Press, New York, 1997).

[3] J. Maserjian, in: *The Physics and Chemistry of SiO_2 and the Si/SiO_2 Interface*, ed. by C.H. Helms and B.E. Deal (Plenum, New York, 1988).

[4] K. K. Likharev, "Layered tunnel barriers for nonvolatile memory devices", *Appl. Phys. Lett.*, vol. 73, pp. 2137-2139, Oct. 1998.

[5] A. N. Korotkov and K. K. Likharev, "Resonant Fowler-Nordheim tunneling and its possible applications", in: *1999 IEDM Tech. Digest* (IEEE Press, Piscataway, NY, 1999), pp. 223-226.

[6] A positive effect of "graded" tunnel barriers on the speed of field-induced electron injection was noticed long ago [7, 8]. However, the asymmetrical barriers studied in those works could not provide short erase time and hence the bit-addressable memory operation as a whole. 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.

[7] D. J. DiMaria, "Graded or stepped energy band-gap-insulator MIS structures (GI-MIS or SI-MIS)", *J. Appl. Phys.*, vol. 50, pp. 5826-5829, Sep. 1979.

[8] F. Capasso, F. Beltram, R. J. Malik, and J. F. Walker, "New floating-gate AlGaAs/GaAs memory devices with graded-gap electron injector and long retention times", *IEEE Electron Device Lett.*, vol. 9, pp. 377-379, Aug. 1988.

[9] A somewhat similar effect may be achieved using grain-enriched interfaces [10], electrode surface texturing [11], or granular floating gates [12], 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.

[10] D. J. DiMaria and D.W. Dong, "High-current Injection into SiO2 from Si Rich SiO2 Films and Experimental Applications", *J. Appl. Phys*., vol. 51, pp. 2722-2735, 1980.

[11] H. J. Buhlmann, M. Olcer, and M. Ilegems, "Enhanced Field-Emission from Plasma-textured Si-SiO_2 Interfaces", *Electron. Lett*., vol. 22, pp. 212-214, Feb. 1986.

[12] H. I. Hanafi, S. Tiwari, and I. Khan, "Fast and long retention-time nano-crystal memory", *IEEE Trans. on Electron. Dev*. vol. 43, pp. 1553-1558, Sep. 1996.

[13] C. L. Allyn, A. C. Gossard, and W. Weigmann, "New rectifying semiconductor structure by molecular-beam epitaxy", *Appl. Phys. Lett*., vol. 36, pp. 373-376, 1980.

[14] R. J. Malik, T. R. AuCoin, R. L. Ross, K. Board, C. E. C. Wood, and L. F. Eastman, "Planar-doped barriers in GaAs by molecular-beam epitaxy", *Electronics Lett*., vol. 16, pp. 836-837, 1980.

[15] *Silicon Nitride and Silicon Dioxide Thin Insulating Films*, ed. by V. J. Kapoor and K. T. Hankins (The Electrochemical Society, Pennington, NJ, 1987), pp. 7, 23.

[16] V. I. Belyi, L. L. Vasilyeva, A. S. Ginovker, V. A. Gritsenko, S. M. Repinsky, S. P. Sinitsa, T. P. Smirnova, and F. L. Edelman, *Silicon Nitride in Electronics* (Elsevier, Amsterdam, 1987), pp. 148, 162.

[17] S. Strite and H. Morkoç, "GaN, AlN, and InN: A review" *J. Vac. Sci. Technol.* B, vol. 10, pp. 1237-1266, Jul/Aug. 1992.

[18]. J. T. Wallmark and J. H. Scott, *RCA Review*, vol. 30, pp. 335-340, 1969.

[19] V. W. L. Chin, T. L. Lancey, and T. Osotchan, “Electron mobilities in gallium, indium, and aluminum nitrides”, *J. Appl. Phys*., vol. 75, pp. 7365- 7372, June 1994.

[20] V. M. Bermudez, T. M. Jung, K. Doverspike, and A. E. Wickenden, “The growth and properties of Al and AlN films on GaN(0001)-(1×1)”, *J. Appl. Phys*., vol. 79, pp. 110-119, Jan. 1996.

[21] A. N. Korotkov and K. K. Likharev, “Possible Cooling by Resonant Fowler-Nordheim Emission”, *Appl. Phys. Lett.*, vol. 75, pp. 2491-2493, Oct. 1999.

[22] Resonant tunneling is important only if the subband broadening due to scattering of electrons in subbands is lower than the inter-subband spacing. In the opposite limit, electron transfer through the triangular potential well at the interface (Fig. 2c) may be treated as sequential tunneling. Calculations carried out with this (implicit) assumption [4] have shown that the layered barrier performance may be similar, if the outer layer thickness d' is re-adjusted (e.g., to about 2 nm instead of 4 nm for the example shown in Fig. 3).

[23] A. N. Korotkov and K. K. Likharev, “New Prospects for Electrostatic Data Storage Systems”, in: *Proc. of 8th Goddard Conference and 7th IEEE Symposium on Mass Storage Systems* (NASA, Greenbelt, MD, 2000), pp. 197-202.

[24] K. K. Likharev and A. N. Korotkov, “Ultradense Hybrid SET/FET Dynamic RAM: Feasibility of Background-Charge-Independent Room-Temperature Single-Electron Digital Circuits”, in: *Proc. of 1995 ISDRS* (U. Virginia, Charlottesville, 1995), pp. 355-358.

[25] K. Likharev, "Single-Electron Devices and Their Applications", *Proc. IEEE*, vol. 87, pp. 606-632, Apr. 1999.

[26] The threshold voltage is relatively high because of the large gate oxide thickness.

[27] *International Technology Roadmap for Semiconductors, 1998 Update/1997 Edition* (Semiconductor Industry Association, San Diego, CA, 1999), available on the Web at www.semichips.org.

[28] F.G. Pikus and K. K. Likharev, "Nanoscale Field-Effect Transistors: An Ultimate Size Analysis", *Appl. Phys. Lett*., vol. 71, pp. 3661-3663, Dec. 1997.

[29] Y. Naveh and K. Likharev, “Modeling of 10-nm-scale Ballistic MOSFETs”, *IEEE Electron. Dev. Lett*., vol. 21, pp. 242-244, May 2000.

Figure Captions

Fig. 1. Tunneling current density j (in A/m^2, dashed lines) and the floating gate recharging time scale τ as defined by Eq. (1) (in seconds, solid lines) for two SiO_2 barriers with d = 5 nm and d = 9 nm, as functions of the electric field E in the barrier (in MV/cm), calculated using the standard quasiclassical approximation. The effective electron mass in the SiO_2 conduction band was accepted to be isotropic and equal to $0.4m_0$, the effective barrier height is 3.2 eV. (These parameters give good agreement with experiment - see, e.g., Ref. 3.)

Fig. 2. Conduction band edge profiles of various tunnel barriers (solid lines) and their deformation at high applied voltage (dashed lines): (a) usual, uniform tunnel barrier; (b) ideal crested barrier; (c) realistic trilayer crested barrier. In the lower panel (c), thick horizontal lines show the position of electron subbands formed in the triangular quantum at the interface between the first and the second layer (schematically).

Fig. 3. Tunneling current density j (dashed lines) and the floating gate recharging time scale τ (solid lines) for a trilayer crested barrier with parameters corresponding to the $n^+Si/Si_3N_4/AlN/Si_3N_4/n^+Si$ system (see the text) and total barrier thickness $d' + d + d' = 4 + 5 + 4 = 13$ nm. Results for two SiO_2 barriers are also shown for comparison.

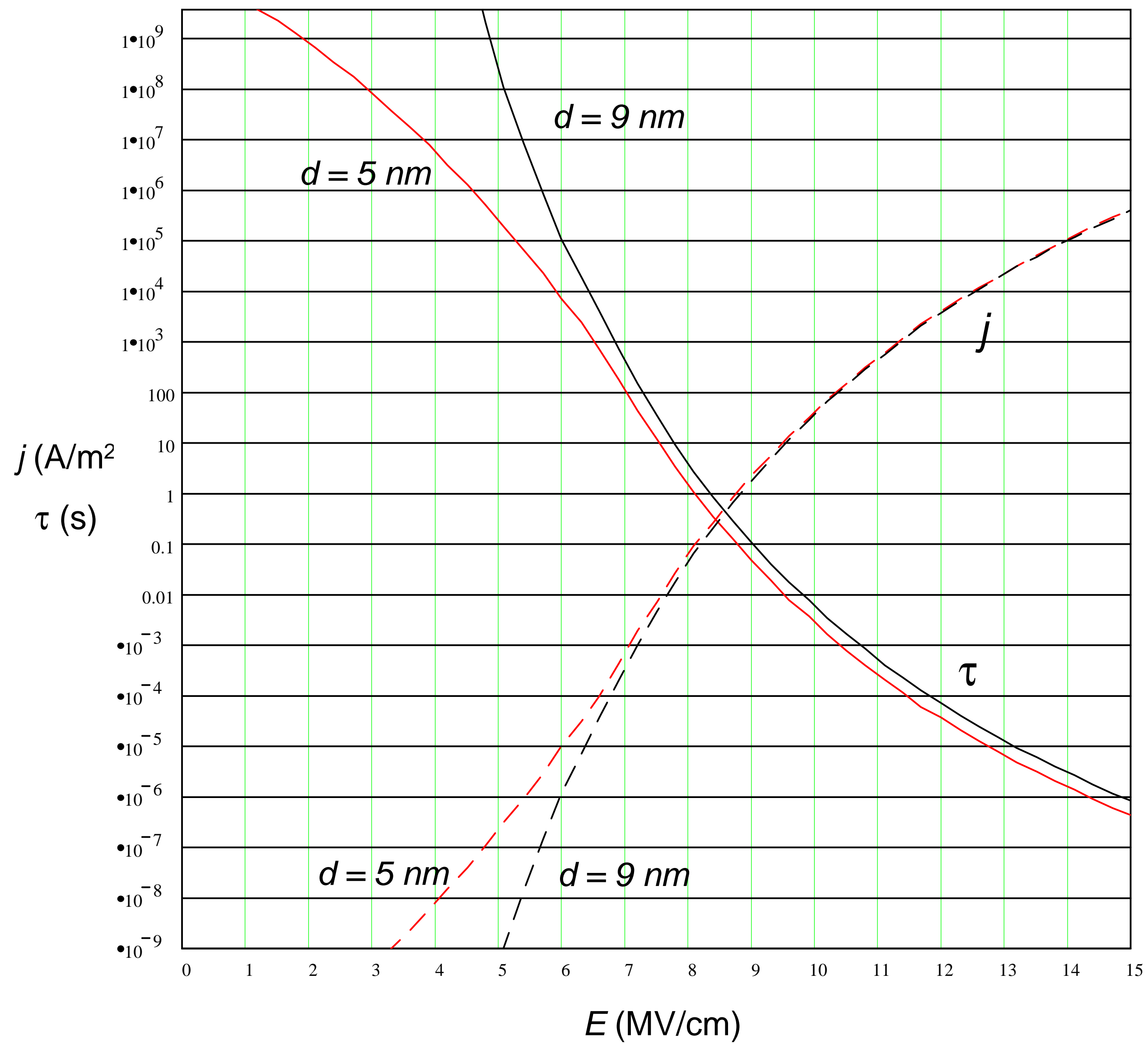
Fig. 4. NOVORAM cell: structure and basic operations.

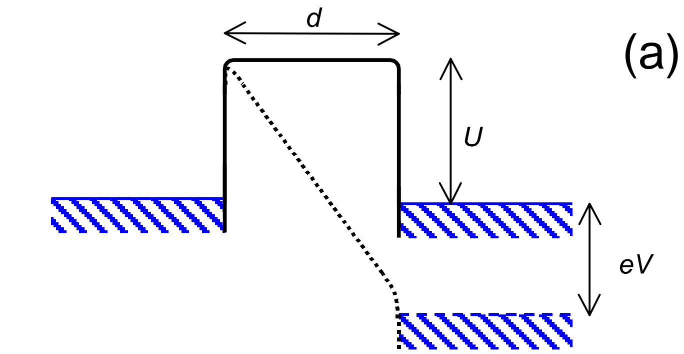
Fig. 5. Possible NOR architecture of the NOVORAM matrix.

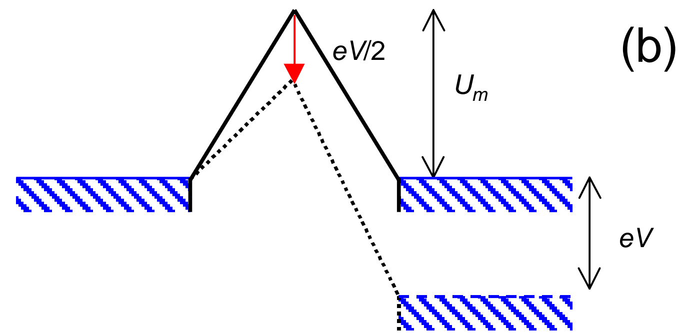
Fig. 6. Possible conservative layout of the NOVORAM cell using SOI (e.g., SIMOX) technology.

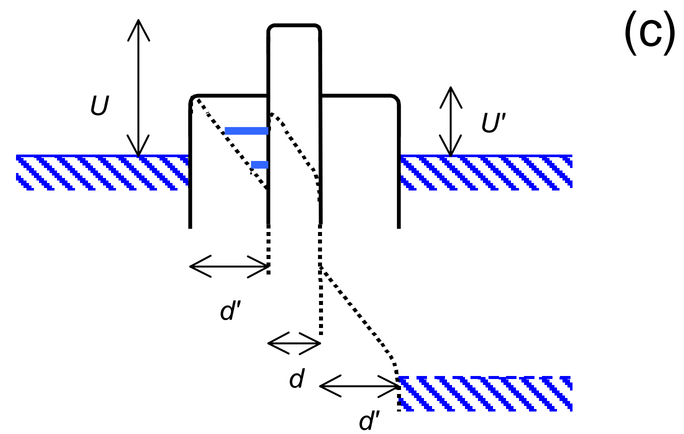
Fig. 7. Scaling prospects for several bit-addressable memories. Solid lines show the estimated relation between the minimum feature size and density. Dotted line and points characterize low-temperature operation of single-electron memories (their room

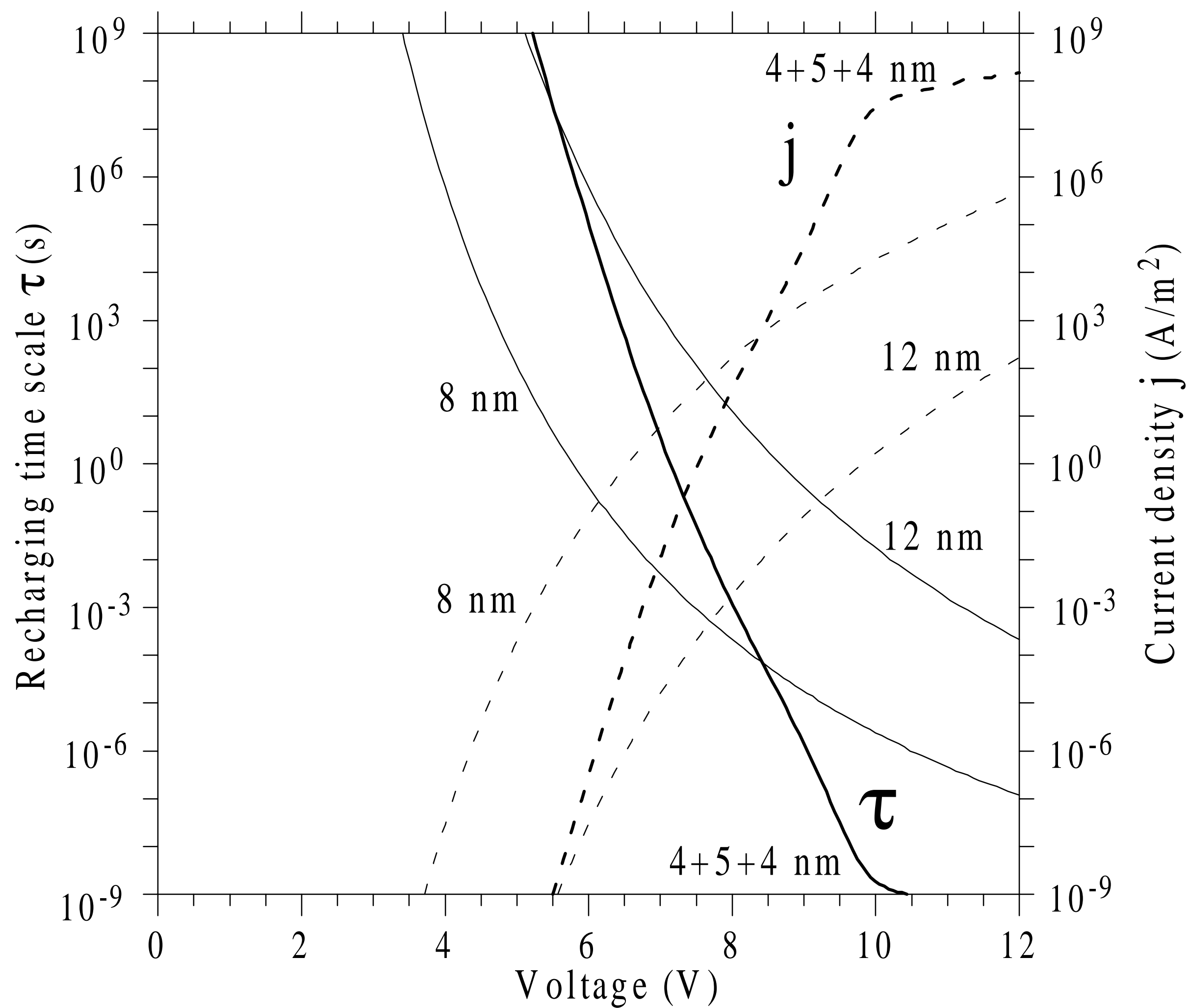
temperature operation requires ~3 nm technology.) Dashed lines and open points indicate the regions where major physical problems are anticipated (in addition to fabrication challenges). The DRAM projections are borrowed from the recent industrial forecast [27], while those for single-electron memories are adopted from the recent review [25].


d = 9 nm
d = 5 nm
j
τ
d = 5 nm
d = 9 nm
j (A/m2
τ (s)
E (MV/cm)
1•10^9
1•10^8
1•10^7
1•10^6
1•10^5
1•10^4
1•10^3
100
10
1
0.1
0.01
•10^−3
•10^−4
•10^−5
•10^−6
•10^−7
•10^−8
•10^−9
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

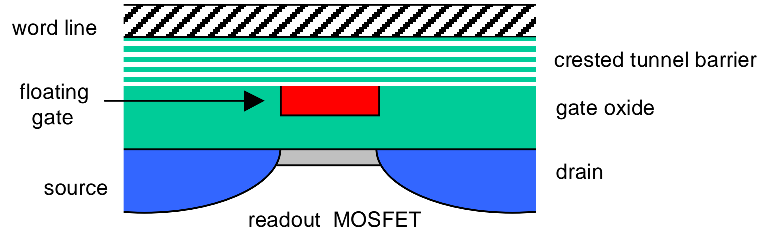

d
(a)
U
eV


(b)
eV/2
U_m
eV

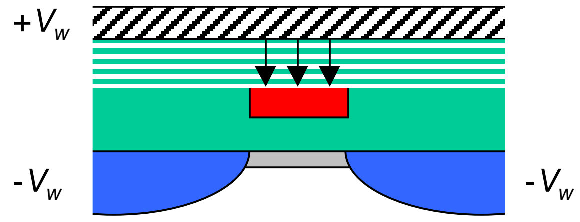

(c)
U
U'
d'
d
d'



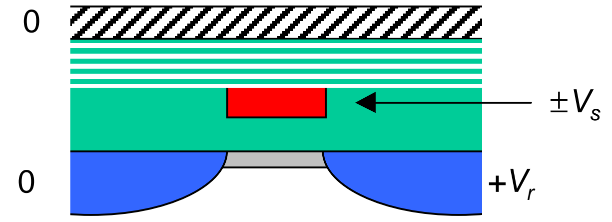
4+5+4 nm
j
12 nm
8 nm
12 nm
8 nm
τ
4+5+4 nm
Recharging time scale τ (s)
Current density j (A/m2)
Voltage (V)



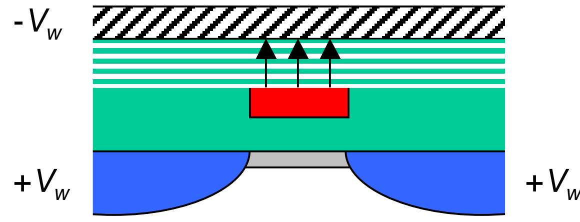
WRITE 1



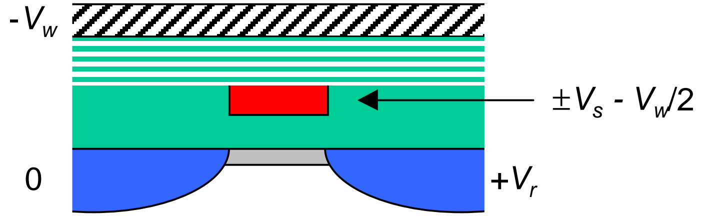
READ SELECT

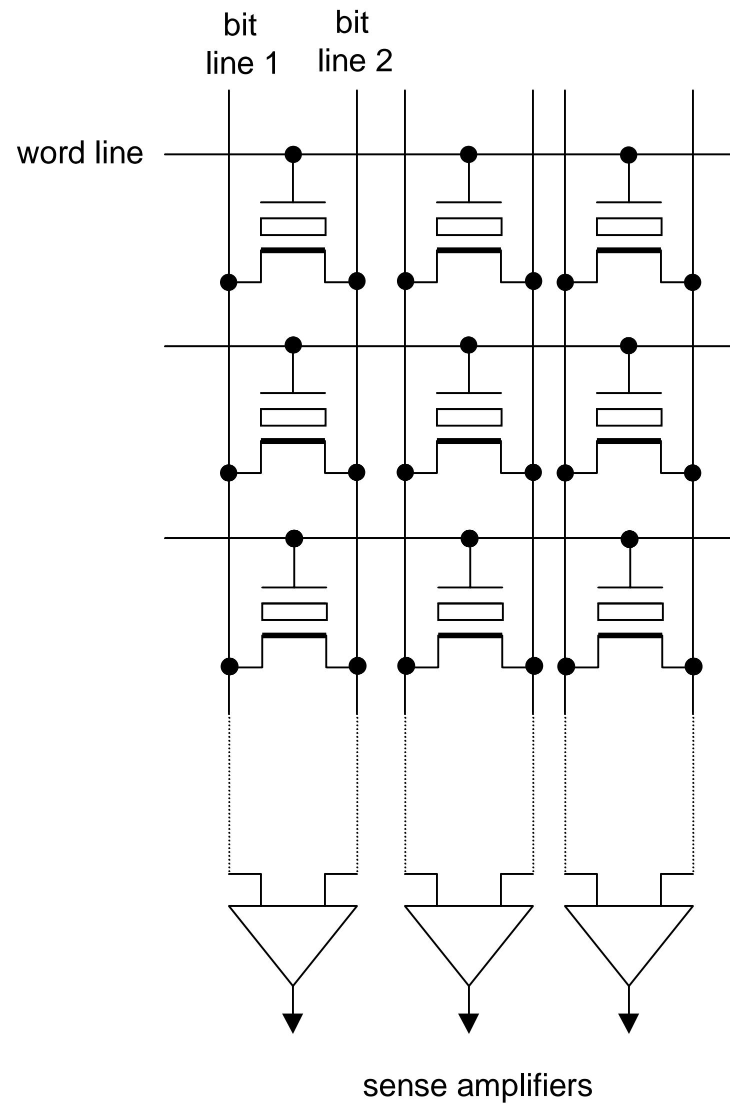


WRITE 0

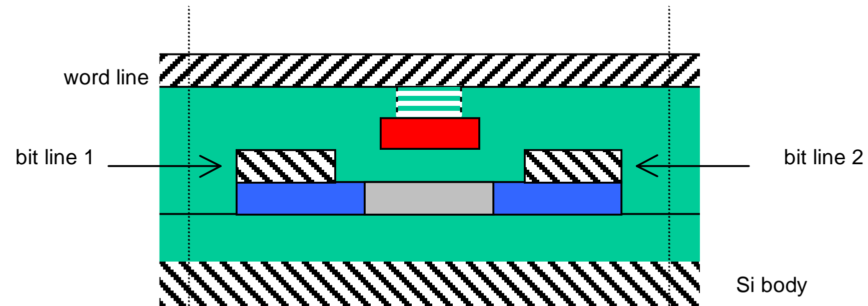


READ DESELECT

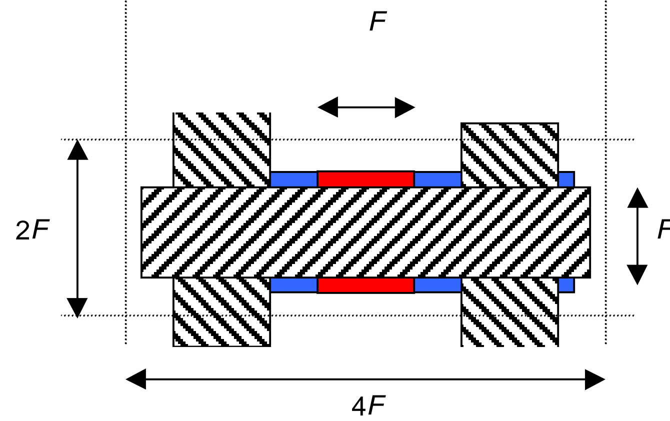


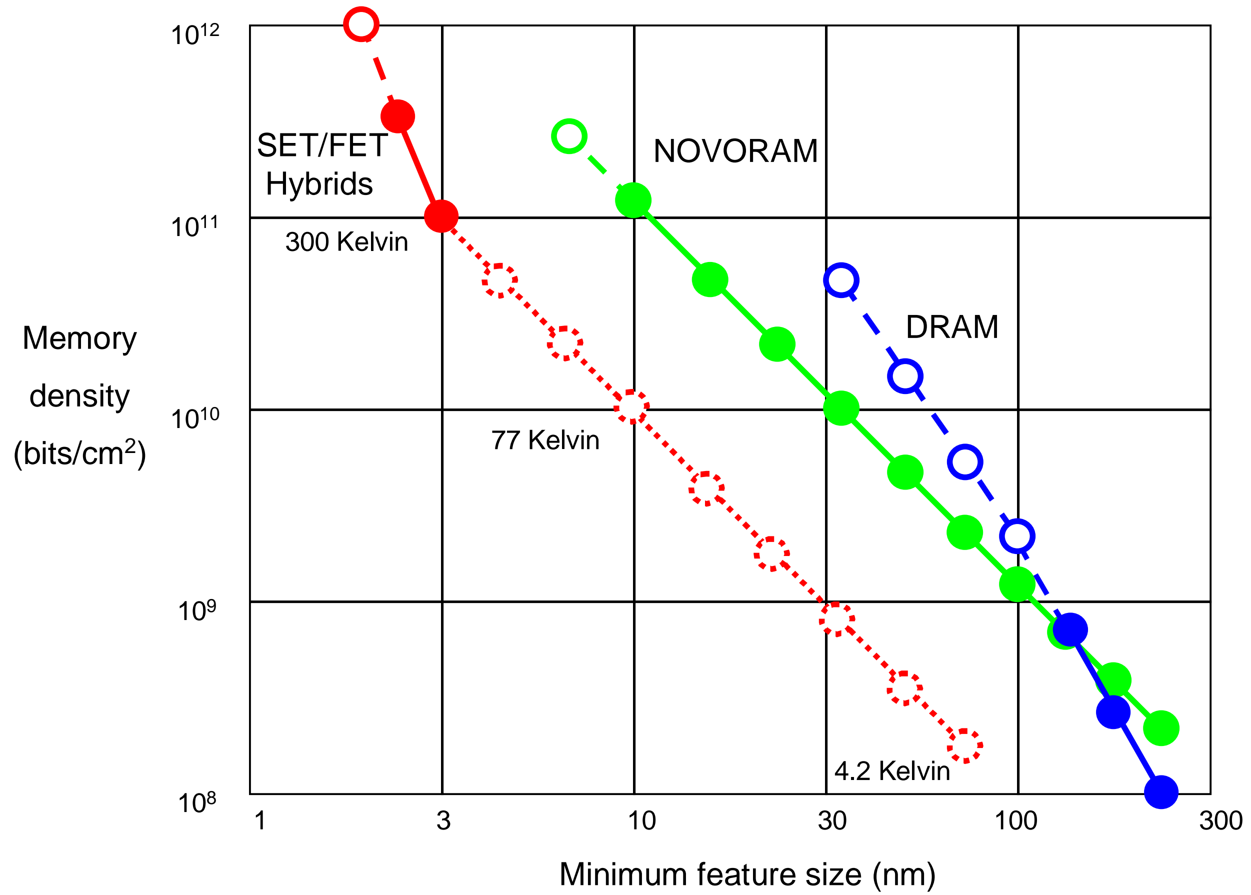

bit
line 1
bit
line 2
word line
sense amplifiers

Cross - section



Top view




Memory
density
(bits/cm2)
10^12
10^11
10^10
10^9
10^8
SET/FET
Hybrids
300 Kelvin
77 Kelvin
4.2 Kelvin
NOVORAM
DRAM
1
3
10
30
100
300
Minimum feature size (nm)