Flash Memory Scaling

Al Fazio

Abstract

In order to meet technology scaling in the field of solid-state memory and data storage, the mainstream transistor-based flash technologies will start evolving to incorporate material and structural innovations. Dielectric scaling in nonvolatile memories is approaching the point where new approaches will be required to meet the scaling requirements while simultaneously meeting the reliability and performance requirements of future products. High-dielectric-constant materials are being explored as possible candidates to replace the traditional SiO_2 and ONO (oxide/nitride/oxide) films used today in memory cells. Likewise, planar-based memory cell scaling is approaching the point where scaling constraints force exploration of new materials and nonplanar, three-dimensional scaling alternatives. This article will review the current status and discuss the approaches being explored to provide scaling solutions for future transistor floating-gate-based nonvolatile memory products. Based on the introduction of material innovations, it is expected that the planar transistor-based flash memory cells can scale through at least the end of the decade (2010) using techniques that are available today or projected to be available in the near future. More complex, structural innovations will be required to achieve further scaling.

Keywords: flash memory, floating gates, nonvolatile memory, scaling.

Introduction

Floating gate flash memory[1,2] is now the fastest-growing memory segment, driven by the rapid growth of portable devices such as digital cameras and cellular phones. The technology allows for data stored in multiple memory cells to be erased in a single action (a "flash") by means of an applied voltage. Flash memory is categorized into two basic approaches: NOR flash, characterized by a fast initial access time for high read performance; and NAND flash, characterized by a slow initial access time and high write performance. This article will focus on the scaling of NOR flash memory; however, many of the scaling considerations described in this article are common to both NOR and NAND floating-gate flash memories.

As flash memory devices begin to scale into the sub-100-nm lithography regime, scaling is becoming a greater challenge due to the high electric fields required for the programming and erase operations and the stringent leakage requirements for long-term charge storage. These requirements are imposing fundamental scaling limitations on the cell operating voltages and on the physical thickness of the tunneling dielectric. Overcoming these limitations will require innovations in cell structure and device materials. This article will outline the major scaling challenges for flash memory and identify some of the potential solutions to enable further scaling to occur.

Flash Cell Basic Operation

A floating gate nonvolatile flash memory cell is shown schematically in Figure 1. It is a metal oxide semiconductor (MOS) transistor with two gates, a floating gate and a control gate. The memory cell consists of an n-channel transistor with the addition of an electrically isolated polysilicon floating gate. Electrical access to the floating gate is only through a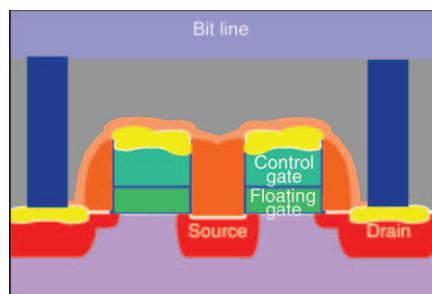 capacitor network of surrounding SiO_2 layers and source, drain, transistor channel, and polysilicon control gate terminals. Any charge present on the floating gate is retained due to the inherent $Si-SiO_2$ energy barrier height, leading to the nonvolatile nature of the memory cell. Characteristic of the structure is a thin tunneling oxide (~100 Å), an oxide/nitride/oxide (ONO) interpoly dielectric (IPD) that resides between the two polysilicon gates, and a short electrical channel length. Because the only electrical connection to the floating gate is through capacitors, the flash cell can be thought of as a linear capacitor network with an n-channel transistor attached. The threshold voltage of the device can be changed by modifying the charge on the floating gate, which can retain this charge for many years. Data can be stored in the memory by adding or removing charge. In the simplest form of a memory cell, two threshold levels (high and low) can store one bit of information in each cell. This concept can be extended to store more levels, commonly called an MLC (multilevel cell).[3] Four levels would allow two bits of data per cell to be stored. Adding and removing charge from the floating gate is normally achieved by Fowler–Nordheim tunneling or hot carrier injection.

Flash Cell Programming

Programming a flash cell means that charge, or electrons, are added to the floating gate. Figure 2 shows the cell bias conditions during program operation. A high drain-to-source bias voltage is applied, along with a high control gate voltage. The gate voltage inverts the channel (turns it to the ON state), while the drain bias accelerates electrons toward the drain. Programming a flash cell by channeling electrons with high kinetic energy—so-called hot electrons—can be understood by use of the "lucky" electron model,[4] as illustrated by the energy-band diagram in Figure 3. In the model, an electron crosses the channel without collision—a process that requires some luck—thereby gaining sufficient kinetic energy to surmount the 3.2 eV $Si-SiO_2$ energy barrier. Prior to entering the drain and being swept away, this lucky electron experiences a collision with the

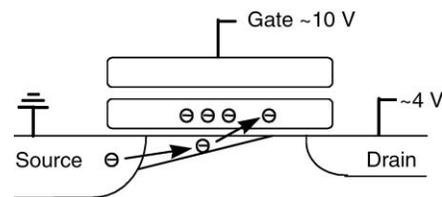*Figure 1. Flash memory cell: cross section along the channel.*

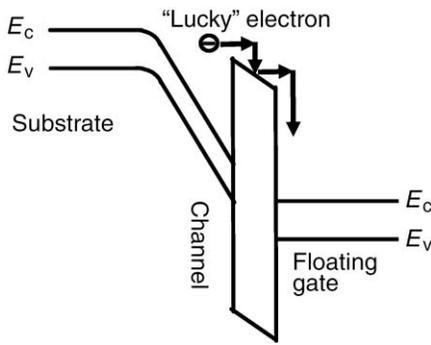*Figure 2. Memory cell bias conditions for programming.*

Figure 3. Energy-band diagram for channel hot-electron programming. E_c is the energy of the conducting band and E_v is the energy of the valence band.

silicon lattice and is redirected toward the Si-SiO$_2$ interface with the aid of the gate field. The electron is subsequently captured on the floating gate and retained as stored charge.

Erasing a Flash Cell

As mentioned earlier, in a flash cell, the contents of the memory, or charge, are removed by means of applying an electrical voltage, erasing them in a "flash." The electrical erasure of flash memory is achieved by Fowler–Nordheim tunneling,[5] for which the bias conditions are shown in Figure 4. Under these conditions, a high field (8–10 MV/cm) is present between the floating gate and the channel. As evidenced by the energy-band diagram of Figure 5, electrons tunneling through the first ~30 Å of the SiO$_2$ are then swept into the silicon. When the erase operation has been completed, electrons have been removed from the floating gate, reducing the cell threshold. While programming is selective to each individual cell, erasing is not, with many cells (typically, 64 kbytes) being erased simultaneously.
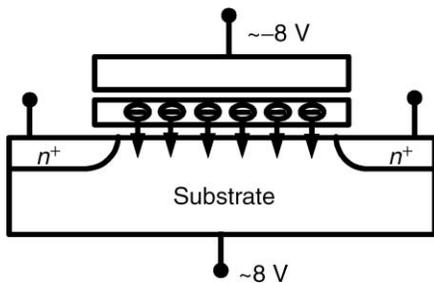

Figure 4. Memory cell bias conditions for erase operation. The floating gate is the rectangle in the middle, and the channel is the region between the n$^+$ areas at the top of the substrate layer.

Flash Memory Scaling Challenges and Alternatives

Figure 6 depicts the layout of a flash memory cell and some of the major scaling constraints. Throughout the history of the development of flash, the cell size has been reduced through a combination of lithography and self-alignment techniques.[6] Today, it is common for many of the cell components to be fully or partially self-aligned, eliminating the need for lithography registration between layers. Figure 7 illustrates various self-alignment techniques that have been deployed. Once the cell becomes fully self-aligned, further scaling is largely determined by the electrical limiters of the cell, where scaling the flash cell requires scaling the cell's capacitive network and basic transistor features. The following sections explore how the introduction of new materials and structures will enable flash memory to continue scaling beyond simple linear extrapolations.

Channel-Length Scaling: Challenges and Alternatives

To achieve hot carrier programming, a voltage of about 4 V is required from the drain to source to produce electrons of high enough energy to overcome the 3.2 eV Si-SiO$_2$ barrier height. One can extrapolate the historical scaling trend for channel length and extrapolate potential limits. This extrapolation is based on the fact that the basic planar cell structure is the same for all of the generations, and scaling is achieved by reducing specific cell dimensions. The difference between the gate length L_{gate} and the effective electrical gate length L_{eff} is the lateral diffusion of the source and drain underneath the gate. The convergence point of the trend represents a projection of a scaling rate limiter of the current planar cell structure. As shown in Figure 8, scaling the gate length will be limited below the 70 nm lithography node due to the inability of the shorter channel length to withstand the required programming voltage. To continue scaling at the same rate, more revolutionary ideas will be needed to scale the L_{eff} more aggressively.

Scaling the channel length could be achieved through engineering of the Si-SiO$_2$ barrier. By choosing dielectric alternatives to SiO$_2$, the barrier can be tailored to allow hot-electron injection to occur at lower voltages. Some promising alternatives have been proposed.[7] Three-dimensional structures[8] are an alternative way to address the gate-length scaling constraint. Both "fin" (protruding above the silicon like a fish's fin) and "U" (a U-shaped trough etched below the silicon surface) structures show promise to allow further scaling while maintaining the channel length required for pro-
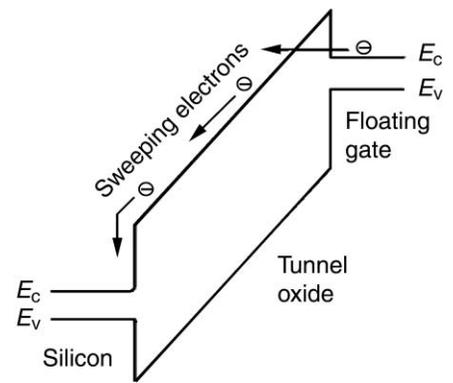

Figure 5. Energy-band diagram for Fowler–Nordheim tunneling erase operation.

gramming. Figure 9 shows an example of a "finfet" (a MOS field-effect transistor made in a fin structure) type of flash cell with a vertical storage gate. This structure moves the channel-length constraint into the vertical (z) direction, allowing further x–y scaling to occur.

Capacitor Network Scaling Challenges and Alternatives

Maintaining adequate coupling of the control gate to the floating gate is a key aspect in scaling the cell channel. This is required to keep the erase voltage as low as possible, avoid erase saturation, and provide control of the channel for reading and programming. Erase saturation occurs when the field across the interpoly dielectric is equal to the field in the tunnel oxide, so that for each electron that is removed (erased) from the floating gate, another electron is added through the IPD. As the cell scales and self-aligned techniques are used for the floating gate, maintaining control of the channel requires a thinner IPD between the control gate and the floating gate, approaching the thickness of the tunneling dielectric. The requirement of an IPD is to provide good capacitive coupling to the floating gate from the control gate, while minimizing any leakage through the dielectric. Unlike the tunnel oxide, this dielectric stack is not expected to support charge transfer during programming and erase operations. For the last decade, the IPD has been composed of an ONO sandwich film. ONO scaling has been achieved by thinning the dielectrics while maintaining control of the thickness. This scaling has also been enabled by improving the quality, in terms of the number of imperfections, of the oxide on the floating gate, which can be aided by reducing the floating-gate polysilicon surface roughness.

An alternative approach would be to replace the IPD with a higher-dielectric-constant material as compared to SiO_2. To maintain the high capacitive coupling improvement, a metal gate would be required to reduce the series capacitance from gate depletion. Since this film is not expected to transport charge (as is the case with the tunnel oxide), it has fewer constraints and it may be easier to find a candidate to meet the cell requirements, as compared with replacing the tunnel oxide. Possible options include Al_2O_3 with a TaN gate that has been proposed for a SONOS-type memory device,[9] that could in principle be deployed in a floating-gate device as well. However, the gate electrode should be chosen consistent with the high-κ film to make sure the energy barrier is engineered to reduce carrier injection.

Another scaling limitation of a capacitor network is the communication of two adjacent cells through the capacitance between the cells. As the spacing between

Figure 8. Gate-length scaling projection. L_{gate} is the gate length; L_{eff} is the effective electrical gate length.

floating gates is reduced, coupling between floating gates increases; thus, the data stored in one cell can influence the operation of an adjacent cell (Figure 10). Communication between adjacent cells may be reduced through a reduction in the thickness of the floating gate and control gate, thus reducing the area of the capacitor. Alternative nonconducting floating gate structures are being actively investigated in the industry as ways to replace the normally conducting floating gate with nonconducting floating nodes. The two main thrusts in this area are silicon nitride storage[10] and nanocrystal storage, including metal-based and semiconductor nanocrystals.[11] While these approaches offer the possibility of replacing the floating gate, both require significant changes in processing, cell operation (program, erase, read) and structure in order to achieve a viable memory. As such, these approaches face significant barriers to becoming mainstream alternatives to floating gate storage.

Tunnel Oxide Scaling Challenges and Alternatives

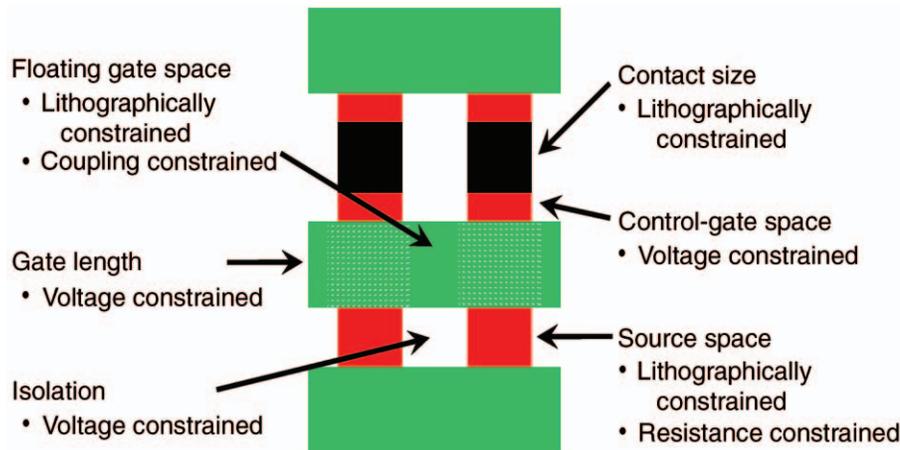Scaling of the tunneling oxide is a primary limitation to cell scaling. The tunnel

Figure 6. Flash cell layout and major scaling constraints. The lighter-green areas are flash cell gates.
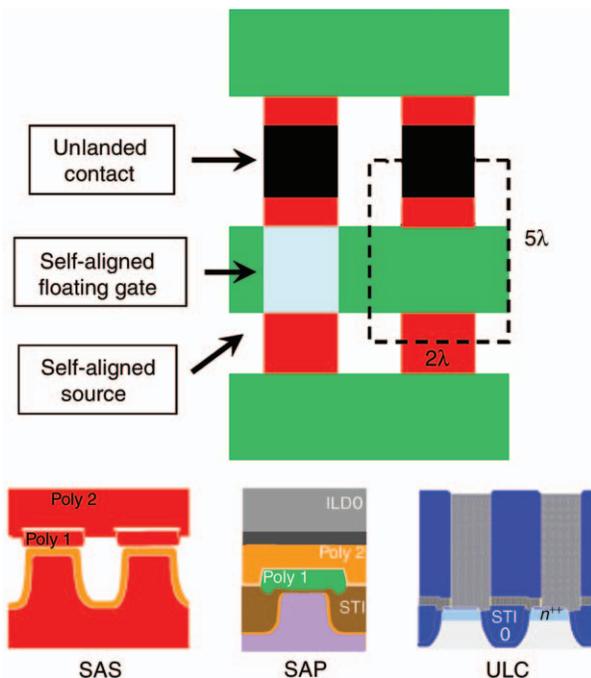

Figure 7. Self-alignment techniques. Use of a self-aligned source (SAS), self-aligned poly (SAP) floating gate, and unlanded contact (ULC) eliminates most lithography registration components of the flash cell area. The dotted box represents one unit memory cell within an array. λ is the minimum lithography feature size; ILD = interlayer dielectric; STI = shallow trench isolation.
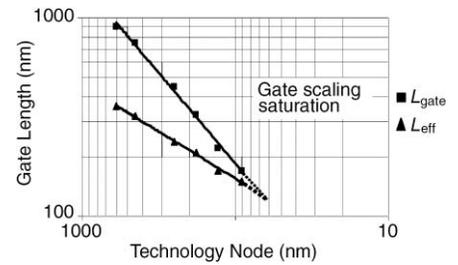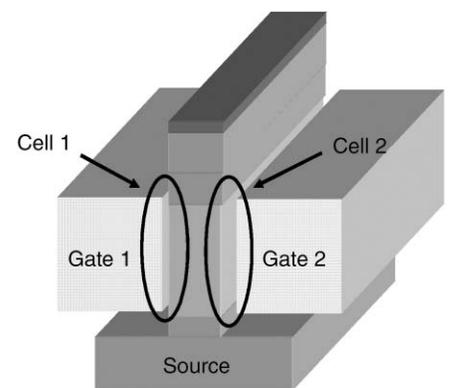

Figure 9. Three-dimensional "finfet" (fin-shaped field-effect transistor) flash structure with a vertical storage transistor.
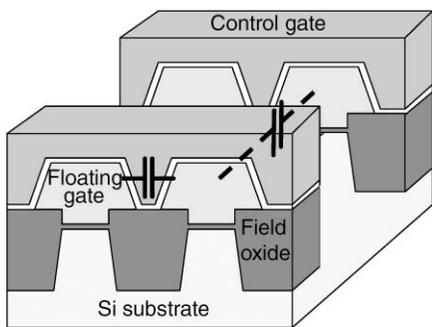
Figure 10. Capacitive coupling between neighboring cells.

oxide needs to scale to enable channel-length scaling, but this is not possible, due to data retention requirements. The charge retention requirements set a lower limit for tunnel oxide thickness of about 80 Å, due to stress-induced low-field tunneling,[12] as shown in Figure 11. In fact, the International Technology Roadmap for Semiconductors (ITRS) revealed a tunnel oxide scaling roadmap that was made more conservative in the 2001 edition due to the challenges in going below 80 Å. It is becoming clear that to enable a significant reduction below 80 Å, a significant change in the SiO_2 dielectric is required. Alternative materials are being explored to allow further scaling of the tunneling dielectric. In particular, "crested barrier"[13,14] composite films may have promise in enabling the design of a more optimal barrier structure. With SiO_2, or any material with a uniform barrier, the high transparency at electric fields necessary for fast program and erase operations (~10 MV/cm) cannot be achieved along with low transparency at fields (1–3 MV/cm) where good retention is required. The reason for this is that the barrier height remains fixed regardless of the applied electric field, and field emission occurs due to thinning of the barrier width (Figure 12a). If a triangular or crested barrier approach is used (Figure 12b), then the
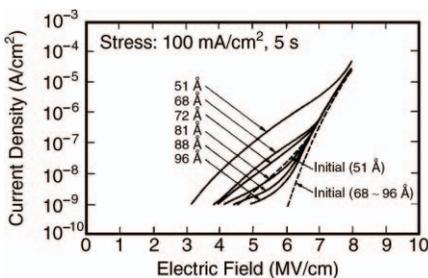

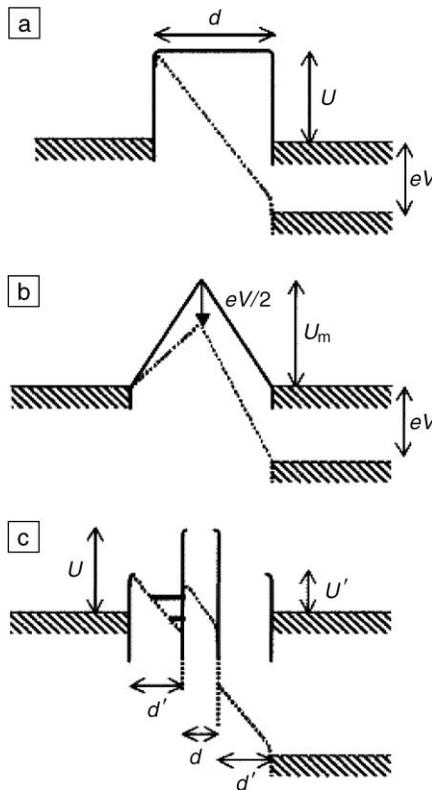Figure 11. Stress-induced leakage current in thin oxides.[12]

Figure 12. (a) Classical single barrier shows that tunneling is modulated by the barrier width. With barrier engineering, it is possible to have (b) a lower barrier under high field and (c) a higher barrier under low field.[14] U = energy barrier height; d = distance.

barrier height reduces under the high-field condition when charge transport is desired, but increases again when a low field is present, reducing the low-field leakage current and improving retention.

Summary

Flash memory devices are now pushing into the sub-100-nm lithography regime. Device scaling is becoming challenging due to the high electric fields required for the programming and erase operations and the stringent leakage requirements for long-term charge storage. Overcoming these limitations will require innovations in cell structure and device materials. Three-dimensional structures and self-alignment techniques can address the physical scaling issues. High-dielectric-constant and crested barrier (or barrier-engineered) materials can address the dielectric scaling issues. It is projected that flash scaling can progress at the current rate through at least the end of the decade (2010) using techniques that are available today or projected to be available in the near future.

References

1. W. Brown and J. Brewer, *Nonvolatile Semiconductor Memory Technology: A Comprehensive Guide to Understanding and Using NVSM Devices* (IEEE Press, New York, 1998).
2. V.N. Kynett, A. Baker, M.L. Fandrick, G.P. Hoekstra, O. Jungroth, J.A. Kreifels, S. Wells, and M.D. Winston, "An In-System Reprogrammable 256 K CMOS Flash Memory," *Tech. Dig. IEEE Int. Solid-State Circuits Conf.* (1988) p. 132.
3. M. Bauer, R. Alexis, G. Atwood, B. Baltar, A. Fazio, K. Frary, M. Hensel, M. Ishac, J. Javanifard, M. Landgraf, D. Leak, K. Loe, D. Mills, P. Ruby, R. Rozman, S. Sweha, S. Talreja, and K. Wojciechowski, "A Multilevel-Cell 32 Mb Flash Memory," *Tech. Dig. IEEE Int. Solid-State Circuits Conf.* (1995) p. 132.
4. S. Tam, P.K. Ko, and C. Hu, "Lucky Electron Model of Channel Hot Electron Injection in MOSFETs," *IEEE Trans. Electron. Dev.* (September 1984).
5. M. Lenzlinger and E.H. Snow, *J. Appl. Phys.* **40** (1) (January 1967) p. 278.
6. A. Fazio, S. Keeney, and S. Lai, *Intel Technol. J.* (May 2002), accessible at developer.intel.com/technology/itj/2002/volume06issue02/ (accessed October 2004).
7. M. She, T.-J. King, C. Hu, W. Zhu, Z. Luo, J.-P. Han, and T.-P. Ma, *Proc. 2001 Int. Semicond. Dev. Res. Symp.* (2001) p. 641.
8. H.B. Pein and J.D. Plummer, *Tech. Dig. 1993 IEEE Int. Electron. Dev. Meet.* (1993) p. 11.
9. C.-H. Lee, K.-I. Choi, M.-K. Cho, Y.-H. Song, K.-C. Park, and K. Kim, "A Novel SONOS Structure of $SiO_2SiNAl_2O_3$ with TaN Metal Gate for Multi-Gigabit Flash Memories," *Tech. Dig. 2003 IEEE Int. Electron. Dev. Meet.* (2003).
10. B. Eitan, P. Pavan, I. Bloom, E. Aloni, A. Frommer, and D. Finzi, *Electron. Dev. Lett.* **21** (11) (2000) p. 543.
11. J. De Blauwe, "Nanocrystal Nonvolatile Memory Devices," *IEEE Trans. Nanotechnol.* **1** (1) (March 2002) p. 72.
12. K. Naruke, S. Taguchi, and M. Wada, *IEDM Tech. Dig.* (1988) p. 424.
13. A. Korotkov and K. Likharev, *IEDM Tech. Dig.* (1999) p. 223.
14. K.K. Likharev, IEEE Circuits Dev. Mag. **16** (4) (July 2000) p. 16. □