Flash memories: Successes and challenges

S. K. Lai

Flash memory grew from a simple concept in the early 1980s to a technology that generated close to $23 billion in worldwide revenue in 2007, and this represents one of the many success stories in the semiconductor industry. This success was made possible by the continuous innovation of the industry along many different fronts. In this paper, the history, the basic science, and the successes of flash memories are briefly presented. Flash memories have followed the Moore's Law scaling trend for which finer line widths, achieved by improved lithographic resolution, enable more memory bits to be produced for the same silicon area, reducing cost per bit. When looking toward the future, significant challenges exist to the continued scaling of flash memories. In this paper, I discuss possible areas that need development in order to overcome some of the size-scaling challenges. Innovations are expected to continue in the industry, and flash memories will continue to follow the historical trend in cost reduction of semiconductor memories through the rest of this decade.

Introduction

Flash memory was first conceived [1] as a functional improvement to UV-erased (ultraviolet-erased) EPROM (Erasable Programmable Read Only Memory). With electrical flash erase, it is possible to reprogram the read only memory *in situ* without removing the memory from a system. The first significant high-volume application for flash memory was BIOS (Basic Input/Output System) memory in personal computers—that is, the memory that activates the computer when the system is first turned on. The flash memory business flourished when the memory was adopted as the standard memory in cell phones, in which the memory enabled just-in-time loading of the latest program code as the last step in manufacturing, and program bugs could be fixed without taking the phone apart. The flash memory optimized for program code execution is called *NOR flash memory* [2–4] (**Figure 1**). In NOR flash memory, each cell resembles a standard MOSFET, except that the cell has two gates, stacked vertically, instead of just one. Each NOR memory cell is connected to the common drain connection called a *bitline* and can be read from directly giving the fast read performance that is necessary for fast program execution. In order to decrease the cost of flash memory, *NAND flash memory* (Figure 1) was invented [5]. In NAND flash memory, the memory cells are connected in series with 16 or 32 memory cells connected to the bitline and source line through two select transistors. [In Figure 1, the source line is connected to the ground through SG(S).] Because cell contact area represents about 30% of unit cell area, this serial cell approach gives smaller cell size and lower die cost compared to NOR memory. The tradeoff is slower read performance because the read current is lower when using serial transistors. The NAND memory business flourished with the growth in popularity of digital cameras, for which NAND memory cards provided a convenient low-cost media for picture storage. The slow read speed is not an issue for such applications. This application was followed by the ubiquitous USB (Universal Serial Bus) drives and MP3 (MPEG-1 Audio Layer 3) players. An emerging new application that will drive more growth for NAND memory is solid-state disks to replace disk drives in notebook computers. The growth of flash memory over the years was driven by the relentless memory cost reduction through Moore's Law; the price for flash memory dropped from approximately $80,000 per gigabyte in 1987 for NOR flash to approximately $10 per gigabyte in 2007 for NAND flash. In this paper, I discuss the innovations that enable the continuous size scaling and cost reduction. Maintaining

0018-8646/08/$5.00 © 2008 IBM

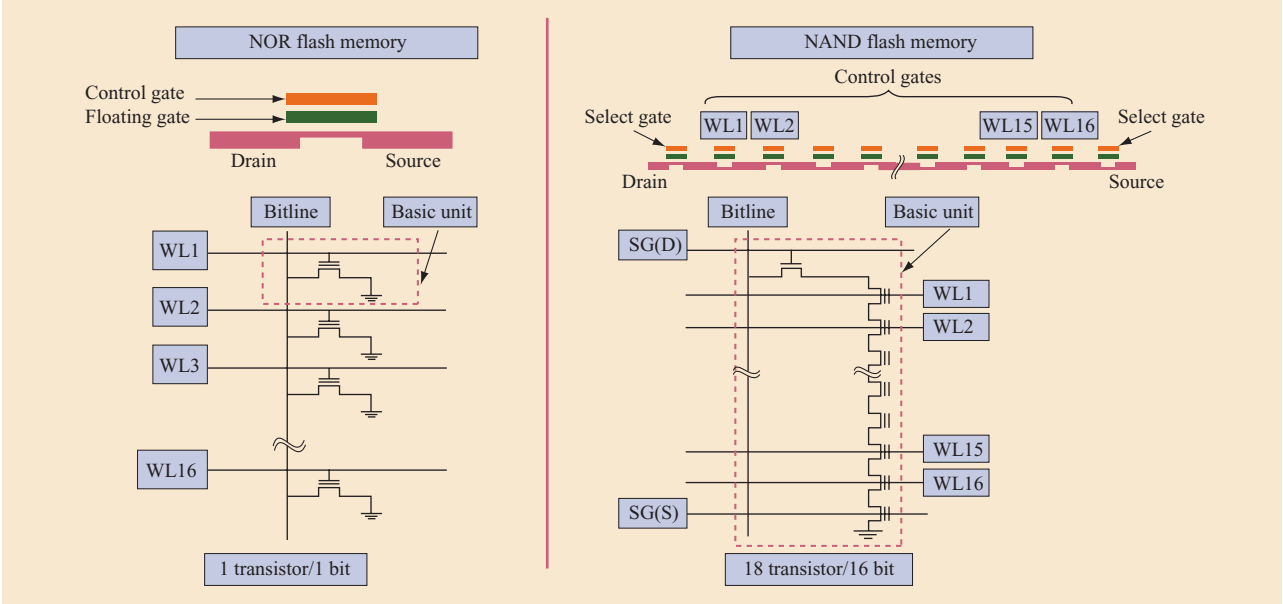

Figure 1

Schematic cross-sections and circuit diagrams for NOR and NAND flash memory. In NOR memory, the basic unit is one memory transistor. For NAND memory, the basic unit is 16 memory and 2 select transistors. [WL: wordline; SG(D): select gate drain; SG(S): select gate source.]

the pace of Moore's Law scaling in the future will be increasingly difficult, but through innovative device design, introduction of new materials, and memory-error management, I anticipate that flash memories will be economically viable for the next few generations, going beyond the 45-nm lithography node generation to generations that are smaller than the 40-nm lithography node.

Basics of NOR and NAND flash memories

As mentioned, the cell cross-sections and schematic diagrams of NOR and NAND flash memories are shown in Figure 1. NOR flash memory was an evolutionary development from EPROM, which is a one-transistor, highly scalable, electrically programmable memory. Programming of an EPROM cell is selective, and both gate and drain of the floating-gate transistor must be biased high in order to generate channel hot electrons that are injected into the floating gate near the drain region. Erasure of EPROM is accomplished by UV light, which excites the floating-gate electrons to sufficient energy to overcome the silicon–silicon dioxide barrier. The buildup in electric field drives electrons from the floating gate. A quartz window is required in the package so that UV light can illuminate the memory cells, and an EPROM package is, thus, more bulky and expensive than NOR and NAND memories. In the development of NOR flash memories [2–4], the first technology decision was to keep the basic EPROM structure and program mechanism. The challenge was to find a way to electrically erase the EPROM. The solutions were twofold. First, the gate-oxide thickness was reduced to approximately 10 nm (from ~25 nm), allowing Fowler–Nordheim electron tunneling [6] to occur when the oxide is biased in order to generate an electric field higher than 10 MV/cm. Second, the source–junction breakdown voltage was increased by grading of the junction with a phosphorous implant, so that more than 10 V can be applied to the junction for an electrical erase. The drain junction was optimized for efficient hot-electron programming; for example, the drain junction is a shallower junction with lower breakdown voltage. By separating functions between source and drain (i.e., the drain is for the program function and the source is for the erase function), the cell can be better optimized for high reliability. With only simple modifications to the manufacturing of EPROM, NOR flash benefited from the EPROM manufacturing learning curve and in a short time achieved comparable cost in manufacturing.

NAND flash memory [5] was invented to further lower the cost of flash memories by reducing the number of contacts when 16 or 32 floating-gate transistors are placed in series (Figure 1). With the memory transistors in series, it was not practical to program individual transistors by channel hot electrons unless very high voltages were applied. Instead, the floating-gate transistors are

programmed and erased by direct Fowler–Nordheim tunneling of electrons between the floating gate and channel under high applied voltages [7]. In this case, the voltages required for direct tunneling are greater than 20 V applied to the top control gate or to the P-well underneath the channel. With the NAND architecture, the read current is much lower, yielding slower reads compared to NOR architectures. On the other hand, with program and erase by Fowler–Nordheim tunneling, the program and erase current required is much lower than that for channel hot electrons in NOR memories, allowing for programming of many bits in parallel and giving a high write bandwidth. Also, with lower voltage drop across each transistor channel for read-only functions, NAND memories allow for much more aggressive scaling of the channel length.

Successes of flash memory scaling

The size scaling of both NOR and NAND flash memories for the last 20-plus years has been a testament to the ingenuity and hard work of some of the best scientists and engineers in the semiconductor industry. Flash memory cells are the most highly electrically stressed transistors in high-volume manufacturing. The program and erase processes involve either a high electric field across the channel of the memory transistor or a high electric field across the tunnel dielectric. For a typical product, the memory cells could be programmed and erased up to one million times. After program and erase of the memory for a number of cycles, the memory cell is expected to store the charge for more than 10 years. For each technology node represented by a new generation of lithographic capability, a team of device, process, design, product, and reliability engineers work closely together to release a new generation of products in about 2 years. At Intel Corporation, the NOR flash technology is called ETOX** (EPROM Tunnel Oxide). The memory cell size for ETOX decreased from 36 μm^2 in 1986 for the 1.5-μm lithography node to 0.0457 μm^2 in 2006 for the 65-nm node. This represents a 788-fold reduction in cell area in 20 years. Furthermore, all of the 65-nm products are based on multilevel cell technology in which two bits of information are stored per memory cell [8–10], giving an effective cell size of 0.0228 μm^2. This represents an effective cell size reduction of 1,579 times in ten generations.

Even though the basic driving force of Moore's Law arises from lithography capabilities, as denoted by the technology nodes, a number of significant innovations are required to meet the scaling goal of 50% reduction in cell size every 2 years. The key innovations for each technology node are listed in **Table 1** [11].

One key aspect of the cell size reduction involves the reduction of the nonactive space in the memory cell.

Table 1 Innovations in NOR scaling with respect to the technology node. [The StrataFlash** multilevel cell (MLC) flash memory technology (from Intel Corporation) stores two or three bits per cell, rather than one bit.]

Technology node	*Key innovation*
1.5 μm	Established flash memory as a viable technology
1.0 μm	Improvement of isolation region overlap to reduce erase variation
	Program and erase cycling reliability established
0.8 μm	Recessed LOCOS (local isolation of silicon)
0.6 μm	Self-aligned source
	Scaled array field oxide
0.4 μm	Negative gate erase
	Intel StrataFlash memory
0.25 μm	Trench isolation
	First time that self-aligned silicide is used
0.18 μm	Self-aligned floating gate
	Unlanded contacts
	Multiple periphery gate oxides
0.13 μm	Channel erase
	Dual trench
	Dual-gate spacer
90 nm	Copper interconnect
65 nm	StrataFlash-only products

These spaces are the layout areas that are not part of the active transistor, and examples include spacing between transistors and the spacing from the transistor gate to the contact. One method to achieve size reduction includes the use of self-aligned technologies in multiple dimensions [12]. One example of self-aligned technologies is a self-aligned source: The connection between memory cells in the source region is created by using the poly (i.e., polysilicon) wordline as an etch mask (self-aligned to wordline). Another example is a self-aligned floating gate. Here, the floating gate is self-aligned to the isolation region between active transistors using a chemical mechanical polish process. One final example is an unlanded contact, in which the contact is allowed to overlap the isolation, reducing bitline space. These self-aligned techniques allow layout space reduction without adding lithographic layers and without requiring additional alignment tolerances, both of which contribute to smaller cell area and lower cost.

High-volume NAND flash memory production started about 10 years after the production of NOR flash memories in the mid-1990s. The basic cell structure scaling of NAND memory is similar to that of NOR memory, for which technology innovations and self-aligned techniques, combined with lithographic advances, permit a significant reduction in cell size. Many of the techniques listed in Table 1 are also deployed for NAND memory. The important new innovation for NAND memory that makes the memory workable is a new usage model that allows bit failures during operation. The bit errors are corrected by error-detection and error-correction techniques. The delay time of error correction is not a problem for most NAND applications (such as memory cards, USB drives, and MP3 players) for which the relatively slow read gives more than sufficient time for errors to be detected and corrected. Typically, a controller is used to manage the errors in multiple chips. The controller also manages a process called *wear-leveling* in which data is written uniformly across multiple blocks and multiple chips before the process returns to the same location. This guarantees that any single block is not excessively written, which could cause premature failure after approximately one million erase and write cycles.

The bit error rate requirement is different for NOR flash memory in which the memory is used for program execution. There is little time for errors to be detected and corrected during the read cycle. The acceptability of shipping products with bad bits in NAND memory greatly reduces the product and reliability engineering efforts before product shipments, allowing for faster development cycles. With the shorter development cycle for NAND, the memory densities of products have been increasing by twofold every year for the last ten years. The unit memory cell size for 8 Gb at a 63-nm node is 0.0164 μm^2, and only 0.0082 μm^2 if multilevel cell capability [13] is included. Compared to the first NOR memory in 1986, this represents a reduction of 4,390 times in cell size. A further improvement in manufacturing efficiency comes from the wafer size increasing from 150 mm in 1986 to 300 mm in 2006. The cell-size and wafer-size changes collectively account for the approximately 8,000 times reduction in cost per gigabyte noted in the introduction.

Future scaling challenges

NAND and NOR flash memories have had great success with respect to memory cell-size reduction and the corresponding product cost reduction. Looking toward the future, we expect significant scaling challenges. The next few sections describe some of the technical challenges and possible solutions. In most cases, even though innovations will exist to facilitate scaling, increasingly, they will involve a significant increase in complexity or the use of expensive new manufacturing tools. Thus, the scaling limit in the future may depend more on economics than on purely technical issues. This cost issue is complex and is not addressed in this discussion.

Broadly speaking, there are three key areas of challenges for flash memory scaling: 1) *physical scaling*, which is primarily defined by lithography and the cell layout design; 2) *electrical scaling*, which is primarily defined by the program/erase/read voltage requirements; and 3) *reliability scaling*, which is primarily defined by the fundamental physics of the program, erase, and storage mechanisms.

The three scaling challenges and possible solutions are now discussed.

Physical cell-scaling challenges

In all cases, lithography is still the main factor affecting Moore's Law scaling. NAND memory, with its very regular layout consisting of straight lines and spaces, allows for the use of many optical enhancement technologies, which greatly facilitate the continued use of conventional optical lithography. Consequently, NAND memory leads the industry with the tightest lithographic pitch of any silicon memory products. However, conventional lithography with i-line and immersion technology is affected by a manufacturing limit at approximately 40 nm. To go beyond conventional lithography, new techniques such as self-aligned double patterning have been reported [14]. By using spacers on two sides of the lithographic line, the space between lines can be further subdivided, effectively improving the resolution. Given this feature, the ability to define minimum line and space is extended to nearly 20 nm and will mostly likely not be the limiting factor for scaling. Note that this technique adds exposure and other process steps that increase the cost of pattern definition. In the case of NOR memory, as far as the 65-nm generation, the layout of the memory cell has 45-degree angle structures around source contacts that are not conducive to optical enhancement techniques [15]. To improve NOR scaling, a self-aligned contact technology is being developed for the 45-nm lithographic node [15], which gives two significant improvements. First, the self-aligned contact reduces the contact area, a scaling limiter for NOR flash memory. Second, the new layout consists of straight lines only similar to NAND, making it easier to implement optical enhancement techniques.

To summarize, in both NOR and NAND memories, it is possible to continue to reduce the physical size of the cell to dimensions close to 20 nm. The true limiters of cell-size reduction involve electrical and reliability requirements, which is discussed in the following sections.

Electrical cell-scaling challenges

For NOR flash memories, a primary scaling limitation for the cell is the high voltage required during the programming operations, which in turn limits the minimum channel length. To achieve hot carrier channel programming, a voltage of more than 4 V is required from the drain to the source to produce electrons of sufficient energy to overcome the 3.2-eV Si-to-SiO_2 barrier height [12]. Therefore, the minimum gate length will be limited to the channel length that can withstand the required programming voltage. For NAND flash, the transistor channel is used for read only, requiring a much lower drain-to-source voltage and, therefore, a shorter channel length limit.

Three-dimensional cell structures are one way to address the gate length scaling constraint. Both above-silicon fin structures [12, 16] and below-silicon U-shaped structures [14] have been reported. These structures move the channel length constraint into the Z direction, allowing further X/Y scaling to occur. This permits further area scaling while maintaining the total channel length required. Recent experimental results reported for NAND, involving the hemi-cylindrical FET (HCFET) [14], show superior transistor characteristics down to the 38-nm node.

Another significant scaling limitation involves maintaining adequate coupling of the control gate to the floating gate. A high coupling ratio is required to provide adequate control of the channel used for reads. As the cell scales in size and self-aligned techniques are used for the floating gate, maintaining control of the channel requires a thinner inter-poly dielectric between the control gate and the floating gate. One possible solution is the use of a high-k dielectric (i.e., a dielectric material with a high dielectric constant). It must be emphasized that the requirement of a high-k dielectric for the inter-polysilicon layer is different from a high-k dielectric for the transistor gate. The inter-poly dielectric has to be optimized for no leakage current under low-field charge storage, whereas a gate dielectric can have a small leakage current.

Another scaling limitation involves the coupling of two adjacent cells through the capacitance between the cells [12, 17, 18]. As the spacing between floating gates is reduced, the floating-gate to floating-gate coupling increases. The data stored in one cell can influence the operation of an adjacent cell. Different solutions exist to address this problem, including reducing the size of the floating gate, electrical screening of the floating gate, or special read biases to compensate for the coupling. In the case of NAND memory, the most promising approach is to replace the floating gate with either floating traps [14, 17–19] or floating conducting islands [18, 20] that function as charge storage layers. In such cases, the capacitive coupling between adjacent cell charge storage layers is greatly reduced. However, this is not a possible solution for floating-gate NOR memory, because in order to move across the transistor channel, NOR memory relies on a conducting charge storage layer to redistribute the channel hot electron charge injected in the drain area.

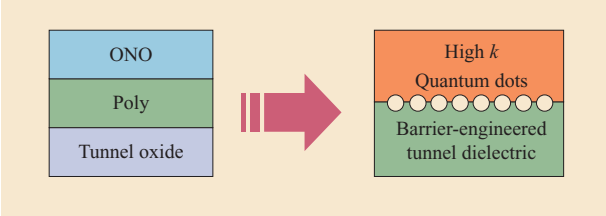

Figure 2

High-level depiction of floating-gate transistor improvement. The ONO dielectric in the traditional transistor (left) is replaced by a high-k insulating dielectric (right). A floating-gate (*poly* in the diagram) is replaced by either floating traps or quantum dots. A tunnel oxide is replaced by a barrier-engineered multilayer dielectric. In each case, new materials are involved.

To summarize, the general concept of extending scaling limit is shown in **Figure 2**. First, ONO (oxide-nitride-oxide) scaling can be extended by the use of a high-k dielectric. Second, the polycrystalline floating gate can be replaced by either floating traps or floating quantum dots. Last, alternative materials can be explored to allow further improvement of the tunneling dielectric. In an ideal case, the tunnel dielectric can be engineered so that the retention is not compromised at low electric fields while the tunneling probability is enhanced at high fields [21]. One proposal reported involves the VARIOT (variable oxide thickness) composite films [22] in which dielectrics of different barrier heights are layered to give the desired tunneling properties. This may be made possible with the improvement in atomic-layer deposition processes and the demonstration of trap-free dielectric films [23]. Note that in all of the above examples, new materials are involved. The introduction of new materials in semiconductor manufacturing is a key part of the innovations that enabled Moore's Law scaling to continue for so many generations.

Reliability scaling challenges

One of the most significant innovations in both NOR and NAND flash memories is multilevel cell (MLC) technology: the storage of more than one bit in a single flash cell [8–10, 13]. This is possible for flash memory because of the analog nature of charge storage in the floating gate, which allows for subdividing the amount of stored charge into small increments. When this is coupled with the superior retention characteristics of the floating

gate, it is possible to accurately determine the charge state after a long period of time. The weakness of MLC arises because the separation between charge states is less for MLC technologies compared to SLC (single-level cell) technologies, resulting in a higher sensitivity to cell degradation mechanisms. To achieve stable storage, it is important to properly control the write and erase operations, using special MLC charge-placement algorithms, to reduce the damage of the tunnel dielectric by reducing the applied fields and controlling how the fields are increased or decreased with the controller rate during write and erase. A further enhancement is the use of error-management techniques, such as error correction, which can recover data or prevent errors.

As the memory cell is scaled, the cell capacitance is decreased, resulting in the decrease of charge stored [12, 18]. For NOR flash [12] with a larger memory cell layout, the number of stored electrons is approximately 1,000 for the 45-nm node, while for NAND flash [18], it is less than 500. In this case, for two-bit-per-cell with four-level MLC technology, the number of electrons per level is just more than 100. While the numbers of stored electrons decrease with each new lithography node, the defect charge leakage mechanisms causing charge loss remain the same. Thus, the impact of each defect on the cell-threshold voltages is proportionally larger for each new node, manifesting as faster threshold voltage drops and an increase in error rates. One method of mitigation involves the improvement of the tunnel dielectric to make it more resistant to defect generation by the introduction of nitrogen into silicon dioxide [24]. Another method is to replace floating gates with either floating traps such as silicon nitride or floating dots such as silicon islands or metal nanodots [14, 17–20]. With discrete charge storage, the impact of a defect is limited only to charge stored in its proximity and not to leaking of the conducting floating gate. However, the discrete traps or islands may store a smaller number of electrons compared to a floating gate, further exacerbating the decrease in stored electrons for each storage level. Even though this problem exists in both NOR and NAND memories, for actual products, it is a larger challenge for NOR flash memories because for NOR flash used in program execution, data errors will result in system failure, and the fast-read requirement does not give much time for error detection and corrections. For NAND flash used in secondary data storage, it is possible to implement extensive error corrections through sophisticated data controllers. With the error rate increasing with scaling, it is possible to implement in the data controllers increasing sophisticated error correction techniques that have been developed for the disk drive industry, and which have made disk drives one of the more reliable storage devices.

Summary

Flash memory devices are now entering the sub-50-nm lithography regime. Flash cell scaling is always challenging because of the high electric fields required for the program and erase operations and the stringent leakage requirements for long-term charge storage. The electric field requirements impose restrictions on the physical scaling of the memory; the diminishing number of stored electrons imposes restrictions on reliability. Overcoming these limitations will require innovations in cell structures and device materials, as well as innovations relating to how the memory is used and managed as part of the overall system. Three-dimensional structures and self-alignment techniques can address the physical scaling issues. High-dielectric-constant insulators and barrier engineering of the tunnel barrier can address the dielectric scaling issues. Floating traps and floating islands can address coupling and dielectric defect issues. Finally, system-level error correction and detection can address some of the reliability issues. I project 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. F. Masuoka, M. Asano, H. Iwahashi, T. Komuro, and S. Tanaka, "A New Flash E^2PROM Cell Using Triple Polysilicon Technology," *Proceedings of the International Electron Devices Meeting*, Vol. 30, 1984, pp. 464–467.
2. S. Mukherjee, T. Chang, R. Pang, M. Knecht, and D. Hu, "A Single Transistor EEPROM Cell and Its Implementation in a 512K CMOS EEPROM," *Proceedings of the International Electron Devices Meeting*, Vol. 31, 1985, pp. 616–619.
3. V. N. Kynett, A. Baker, M. Fandrich, G. Hoekstra, O. Jungroth, J. Kreifels, and S. Wells, "An In-System Reprogrammable 256K CMOS Flash Memory," *Solid-State Circuits Conference, Digest of Technical Papers*, February 17–19, 1988, pp. 132–133.
4. S. Tam, S. Sachdev, M. Chi, G. Verma, L. Ziller, G. Tsau, S. Lai, and B. Dham, "A High Density CMOS 1-T Electrically Erasable Non-volatile (Flash) Memory Technology," *Symposium on VLSI Technology, Digest of Technical Papers*, 1988, pp. 31–32.
5. F. Masuoka, M. Momodomi, Y. Iwata, and R. Shirota, "New Ultra High Density EPROM and Flash EEPROM with NAND Structure Cell," *IEDM Technical Digest*, 1987, pp. 552–555.
6. M. Lenzlinger and E. H. Snow, "Fowler-Nordheim Tunneling into Thermally Grown SiO_2," *J. Applied Phys.* **40**, No. 1, 278–283 (1967).
7. S. Aritome, R. Shirota, R. Kirisawa, T. Endoh, R. Nakayama, K. Sakui, and F. Masuoka, "A Reliable Bi-polarity Write/Erase Technology in Flash EEPROMs," *Proceedings of the International Electron Devices Meeting, IEDM Technical Digest*, December 9–12, 1990, pp. 111–114.
8. M. Bauer, R. Alexis, G. Atwood, B. Baltar, A. Fazio, K. Frary, M. Hensel, et al., "A Multilevel-Cell 32Mb Flash Memory," *Solid-State Circuits Conference, Digest of Technical Papers*, 1995, pp. 132–133.

9. G. Atwood, A. Fazio, D. Mills, and B. Reaves, "Intel StrataFlash™ Memory Technology Overview," *Intel Technol. J.* **1**, No. 2, Q4 (1997).
10. A. Fazio and M. Bauer, "Intel StrataFlash™ Memory Technology Development and Implementation," *Intel Technol. J.* **1**, No. 2, Q4 (1997).
11. A. Fazio, S. Keeney, and S. Lai, "ETOX™ Flash Memory Technology: Scaling and Integration Challenges," *Intel Technol. J.* **6**, No. 2, 23–30 (2002).
12. G. Atwood, "Future Directions and Challenges for ETox Flash Memory Scaling," *IEEE Trans. Device Material Rel.* **4**, No. 3, 301–305 (2004).
13. T.-S. Jung, Y.-J. Choi, K.-D. Suh, B.-H. Suh, J.-K. Kim, Y.-H. Lim, Y.-N. Koh, et al., "A 3.3 V 128 Mb Multi-Level NAND Flash Memory for Mass Storage Applications," *Solid-State Circuits Conference, Digest of Technical Papers*, San Francisco, CA, February 1996, pp. 32–33.
14. D. Kwak, J. Park, K. Kim, Y. Yim, S. Ahn, Y. Park, J. Kim, et al., "Integration Technology of 30nm Generation Multi-Level NAND Flash for 64Gb NAND Flash Memory," *Symposium on VLSI Technology, Digest of Technical Papers*, Kyoto, Japan, 2007, pp. 12–13.
15. M. Wei, R. Banerjee, L. Zhang, A. Masad, S. Reidy, J. Ahn, H. Chao, et al., "A Scalable Self-Aligned Contact NOR Flash Technology," *Symposium on VLSI Technology, Digest of Technical Papers*, Kyoto, Japan, 2007, pp. 226–227.
16. H. B. Pein and J. D. Plummer, "Performance of the 3-D Sidewall Flash EPROM Cell," *Proceedings of the International Electron Devices Meeting, IEDM Technical Digest*, December 5–8, 1993, pp. 11–14.
17. K. Kim, "Technology for Sub-50nm DRAM and NAND Flash Manufacturing," *Proceedings of the International Electron Devices Meeting, IEDM Technical Digest*, December 5–7, 2005, pp. 323–326.
18. K. Prall, "Scaling Non-Volatile Memory below 30nm," *Non-Volatile Semiconductor Memory Workshop*, August 26–30, 2007, pp. 5–10.
19. Y. Shin, J. Choi, C. Kang, C. Lee, K.-T. Park, J.-S. Lee, J. Sel, et al., "A Novel NAND-Type MONOS Memory Using 63 nm Process Technology for Multi-Gigabit Flash EEPROMs," *IEEE Electron Devices Meeting, IEDM Technical Digest*, December 5–7, 2005, pp. 327–330.
20. J. De Blauwe, "Nanocrystal Nonvolatile Memory Devices," *IEEE Trans. Nanotechnol.* **1**, No. 1, 72–77 (2002).
21. K. K. Likharev, "Layered Tunnel Barriers for Nonvolatile Memory Devices," *Appl. Phys. Lett.* **73**, No. 15, 2137–2139 (1998).
22. B. Govoreanu, P. Blomme, M. Rosmeulen, J. Van Houdt, and K. De Meyer, "VARIOT: A Novel Multilayer Tunnel Barrier Concept for Low-Voltage Nonvolatile Memory Devices," *IEEE Electron Device Lett.* **24**, No. 2, 99–101 (2003).
23. D. Wang, T.-P. Ma, J. W. Golz, B. L. Halpern, and J. J. Schmitt, "High-Quality MNS Capacitors Prepared by Jet Vapor Deposition at Room Temperature," *IEEE Electron Device Lett.* **13**, No. 9, 482–484 (1992).
24. T. Chang, H. Jones, C. Jenq, W. Johnson, J. Lee, S. Lai, and V. Dham, "Oxidized-Nitridized Oxide (ONO) for High Performance EEPROMs," *International Electron Devices Meeting*, 1982, p. 810.

Received September 13, 2007; accepted for publication October 8, 2007; Internet publication July 10, 2008

Stefan K. Lai *Ovonyx, Inc., 2956 Waterview Drive, Rochester Hills, Michigan 48331 (slai@ovonyx.com)*. Dr. Lai received a B.S. degree in applied physics from the California Institute of Technology in 1973 and a Ph.D. degree in applied quantum physics from Yale University in 1979. After graduation, Dr. Lai joined the IBM Thomas J. Watson Research Center as a member of the technical staff. He joined Intel Corporation in 1982 as Program Manager for scalable E2PROM components. He co-invented the EPROM tunnel oxide (ETOX) flash memory cell, which has become an industry standard. Dr. Lai retired from Intel at the end of 2006. His last job at Intel was Vice President of the Flash Memory Group and CTO of California Technology and Manufacturing. Currently, he is Vice President, Business Development, Ovonyx, Inc. In his role, he is responsible for engaging the semiconductor industry to develop phase-change memory based on chalcogenide alloys for use in mainstream nonvolatile memory products. Dr. Lai was recognized as an IEEE Fellow in 1998 for his research on the properties of silicon MOS interfaces and the development of flash EPROM memory. He is the recipient of the 2008 IEEE Andrew S. Grove Award for his contribution to multiple generations of flash memories.