The Zen of Nonvolatile Memories

Erwin J. Prinz
Freescale Semiconductor, Inc.
6501 William Cannon Drive West
Austin, Texas 78735, U.S.A.
(512) 895-8443
Erwin.Prinz@Freescale.com

ABSTRACT

Silicon technology based nonvolatile memories (NVM) have achieved widespread adoption for code and data storage applications. In the last 30 years, the traditional floating gate bitcell has been scaled following Moore's law, but recently scaling limits have been encountered which will require alternative solutions after the 65 nm technology node. Both evolutionary and novel solutions are being pursued in the industry. While the traditional floating gate technology will scale to the 65 nm node, novel device structures and array architectures will be needed past that node.

Categories and Subject Descriptors

B.7.1 [**Integrated Circuits**]: Types and Design Styles, Advanced technologies, Memory technologies, Microprocessors and microcomputers, VLSI (very large scale integration)

General Terms

Measurement, Design, Reliability, Experimentation, Security

Keywords

Nonvolatile memories; floating gate; SONOS; nanocrystal; MRAM; phase change memory; FeRAM

1. INTRODUCTION

Silicon technology based nonvolatile memories (NVM) retain information without consuming power. In a system, they can assume the function of a computer hard drive, storing a few bytes up to a few Gigabytes of code and/or data. Prominent examples of applications enabled by NVM are cellular phones (NOR Flash), MP3 players and digital still cameras (NAND Flash), and microcontrollers (embedded NOR Flash). Due to the explosive growth of these applications, there will be more NVM bits shipped in 2006 than DRAM bits [12].

While end users demand fast system write and erase times, fast read access times, and high reliability with respect to data retention, NVM technologists study storage methods and media, bitcell designs, and array architectures and supporting circuits, with the goal of increasing the NVM storage capacity (density) by reducing the size of the NVM bitcell following Moore's law.

Compared to SRAM memories, traditional NVM have had long write/erase times and limited endurance, therefore research efforts are also geared at closing in on the "universal memory" which is nonvolatile and can be read, written, and erased as fast as a SRAM, while also having the smaller bitcell size of, e.g., a DRAM [2]. Since all NVM concepts devised so far fall short of at least some of the "universal memory" features, a diverse group of memory concepts is in production for the mainstream and niche markets.

The maximum die size which can be profitably manufactured at a given technology node has been around 1–2 cm^2, limited by random defects. Silicon technology scaling follows Moore's law resulting in a doubling of the number of transistors every 2-3 years. The available transistor budget has been used to build dedicated memory chips ("standalone" memories) with the largest die size compatible with the transistor budget, or to build systems-on-a-chip ("SoC"), in which only a fraction of the transistors is used for the NVM.

Table 1 shows selected parameters for NVM built in 90 nm technology. While the size of the bitcell is aggressively reduced in standalone memories at the expense of a more complex manufacturing process, larger cell sizes achieved with a less complex process are acceptable for microcontroller SoC's, in which the NVM processing is added to a high performance CMOS process.

As an example for Moore's law, Figure 1 shows how the increasing transistor budget has been utilized to increase the performance, and complexity, of microcontrollers used in automotive applications.

In this paper, various aspects of NVM technology are discussed in light of Moore's law. While the traditional, floating gate based, NVM technology will scale to the 65 nm node, new bitcell approaches are needed to satisfy Moore's law past 65 nm. For each chosen approach, circuit designs fully adapted to the bitcell features will be needed.

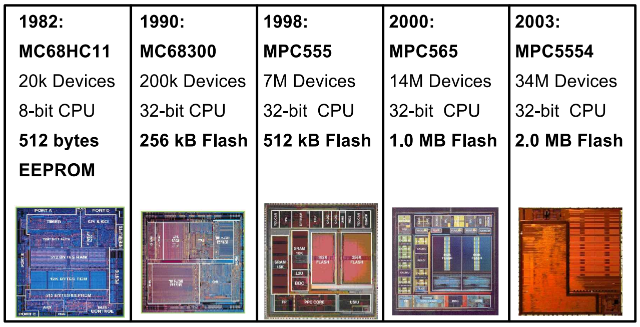
2. FLOATING GATE NVM

Silicon-based nonvolatile memory technology through the last 30 years has been dominated by an approach in which charge is stored in the gate dielectric of a MOSFET to vary

DAC 2006, July 24–28, 2006, San Francisco, California, USA.
Copyright 2006 ACM 1-59593-381-6/06/0007 ...$5.00.

	Embedded NVM (NOR)	Standalone NOR	Standalone NAND
Typical Application	Micro-controller	Cellular Phone	MP3 Player, Digital Camera
Typical Density (90nm technology)	1-32 Mb	256-512 Mb	2 Gb
Typical % of Chip Area	5% - 50%	100%	100%
Bitcell Size (90nm)	$0.18\ \mu\text{m}^2$	$0.09\ \mu\text{m}^2$	$0.05\ \mu\text{m}^2$
Random Access Read Time	20 ns	50-100 ns	15 µs
Process Complexity	low	highest	high

Table 1: Some features of 90 nm NVM technologies.



1982:	1990:	1998:	2000:	2003:
MC68HC11	**MC68300**	**MPC555**	**MPC565**	**MPC5554**
20k Devices	200k Devices	7M Devices	14M Devices	34M Devices
8-bit CPU	32-bit CPU	32-bit CPU	32-bit CPU	32-bit CPU
512 bytes EEPROM	**256 kB Flash**	**512 kB Flash**	**1.0 MB Flash**	**2.0 MB Flash**

Figure 1: Evolution of microcontrollers with embedded NVM.

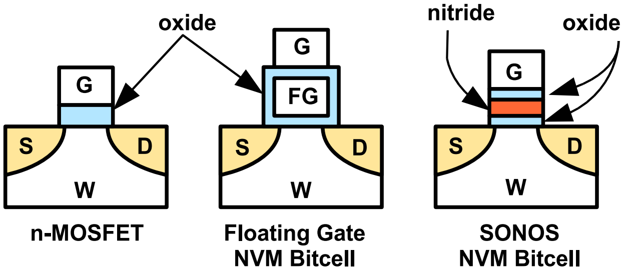


Figure 2: Operating principle of traditional NVM based on charge storage on floating gates or in nitride traps (SONOS), embedded into the gate oxide of a MOSFET.

its threshold voltage, as shown in Figure 2 [6]. The charge storage location can be a volume of doped (electrically conducting) silicon fully surrounded by silicon dioxide insulator, or a nonconducting film of nitride. To change the logic level of the bitcell, electrons or holes have to be moved through the surrounding insulator onto the charge storage location. This is achieved by applying a combination of high voltages, compared to the bitcell read voltages or the logic circuit V_{DD}, to the bitcell terminals. A 1-transistor NOR bitcell, e.g., is biased to obtain hot electron injection of electrons from the channel onto the floating gate for the write operation, and quantum-mechanical Fowler-Nordheim tunneling of electrons from the floating gate to the well for the erase operation, as shown in Figure 3.

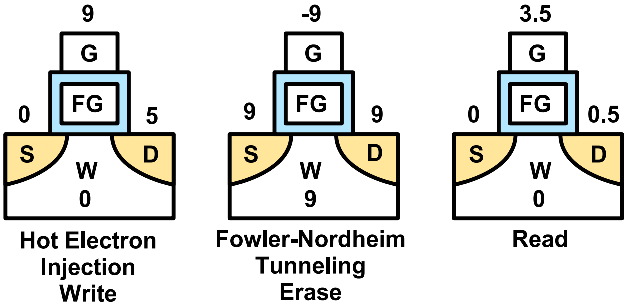


Figure 3: Voltages applied to NOR floating gate bitcell for write, erase, and read operations. S, G, D, W denote source, gate, drain, and well, respectively.

The topology of the metal lines connecting the row and column decode circuits in the periphery to the block of bitcells is governed by constraints of not affecting the charge state of the neighbors of the bitcell to be written or erased. In Flash EEPROM, writing a random bit is possible, but an entire block of bitcells is erased. This compromise in functionality results in a more efficient bitcell consisting of only one memory transistor, compared to a memory in which each byte can be erased independently (EEPROM).

The two dominant array architectures are NOR and NAND, optimized for fast read access and highest density respectively.

Figure 4 shows the NOR architecture in which each bitcell is connected on the source side to a common source, and on the drain side to a metal bitline via a contact, resulting in a low impedance path of the row and column decode circuits to the bitcell for fast read with a typical read current of 10 μA.

In the NAND architecture shown in Figure 5, 16 to 64 bitcells are connected in series eliminating the area required for a dedicated contact per bitcell, but inserting 15 to 63 bitcells into the read path resulting in low read current of 100 nA and therefore slow random read access.

The NVM bitcell endurance with respect to write/erase cycling is limited by degradation of the gate insulators from the charge transport at high electric fields, typically to 100k to 1M write/erase cycles. Data retention of floating gate Flash EEPROM arrays can exceed 20 years with sub-1 part per million failure rates in the field, even at extreme ambient temperatures of -40 through 125°C.

Within the industry, the reliability of floating gate NVM technology, including data retention after write/erase cy-

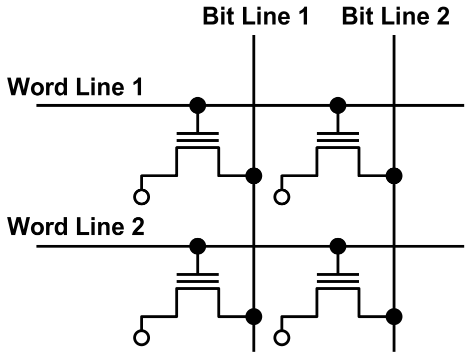


Figure 4: NOR array architecture.

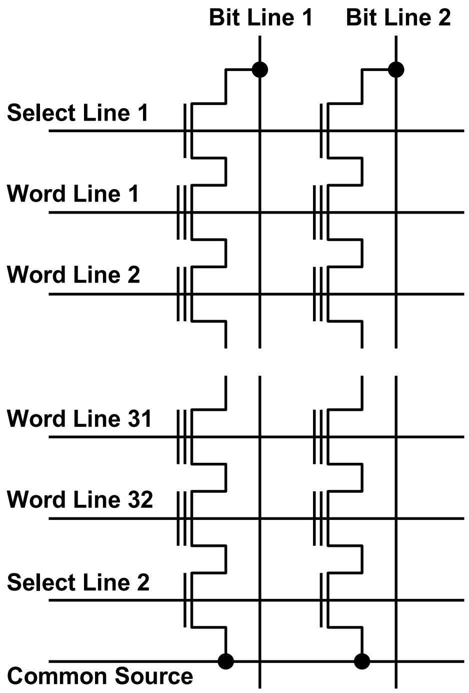


Figure 5: NAND array architecture.

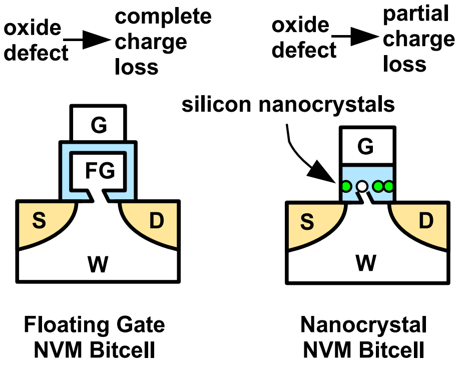


Figure 6: Schematic showing that one oxide defect can discharge the entire floating gate, whereas if localized charge is stored, only partial discharge occurs.

cling, has been characterized extensively, and it has been found that the thickness of the insulator surrounding the floating gate cannot be reduced to below ≈ 10 nm for applications demanding highest reliability, because write/erase cycling can cause leakage paths in a small fraction of bitcells in thinner bottom oxides, and in a (conducting) floating gate, a single oxide defect can result in the entire charge leaking out at that location as shown in Figure 6 [9].

This scaling limitation implies that the write/erase voltages cannot be reduced any further, and therefore the area occupied by the high voltage transistors used in the decode circuits and charge pumps does not scale, limiting the Flash module area gain from reducing the bitcell physical feature size. In fact, since the required read access times decrease due to increasing CPU operating frequencies, embedded high performance NVM module area may be limited by peripheral rather than bitcell area.

In the 1-transistor NOR architecture, the threshold voltages of both the charged (written) and discharged (erased) bitcells have to be greater than 0V to switch off erased, low threshold voltage bitcells, schematically shown in Figure 7. This implies that the read voltage is around 3.5V and has to be supplied by a wordline boost or a charge pump circuit. When bitcell feature size is reduced without reducing the tunnel oxide thickness, the threshold voltage distribution widths can increase, resulting in a smaller voltage window for the read operation. Therefore, the accuracy of the read reference circuits must increase with bitcell size reduction, requiring good matching to the NVM bitcells over supply voltage and temperature operating ranges.

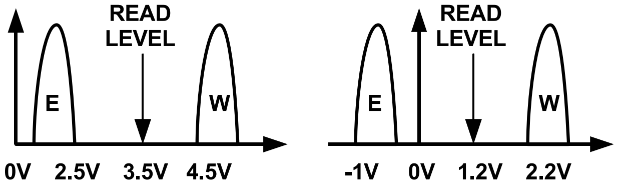


Figure 7: Schematic comparison of 1-transistor floating gate, and nanocrystal split gate bitcell threshold voltage distributions.

Alternative array architectures based on “split gate” bitcells in which the channel area between source and drain is controlled by two gates in series, a memory gate and a select gate, have been devised to lower the wordline read voltage. The logic transistor gate is employed to switch the bitcell off rather than the memory gate, therefore the memory transistor threshold voltage window (erased and written state) can be moved towards more negative voltages, as shown in Figure 7. An additional advantage of this structure is that for hot electron programming, the write current can be limited by the select gate while a vertical field is applied at the control gate, which greatly increases electron injection efficiency and results in a 100 × lower write current, as shown in Figure 8 [7]. Split gate bitcells based on the floating gate approach are widely used for embedded Flash modules in microcontrollers.

3. NITRIDE CHARGE STORAGE

Charge storage in a thin nitride layer embedded into the gate insulator of a MOSFET was demonstrated in the same

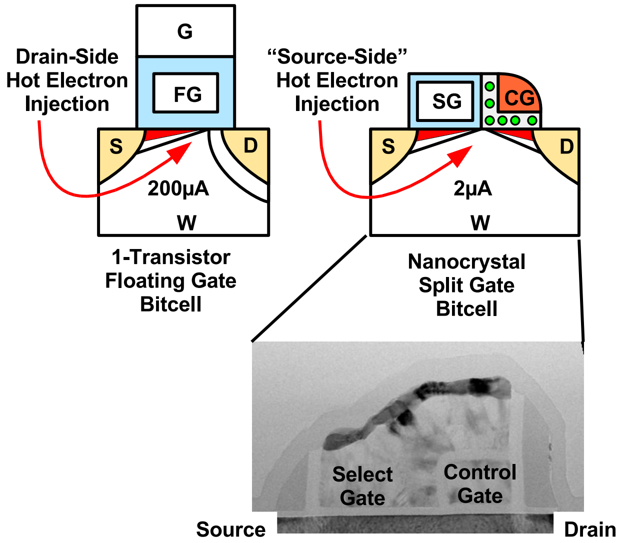


Figure 8: Schematic comparison of 1-transistor floating gate, and nanocrystal split gate bitcells; and transmission electron micrograph of nanocrystal split gate bitcell.

year as the floating gate principle [21]. Within the nitride layer, electrons and holes can be stored in localized traps, with negligible lateral conduction. Since a single defect in the bottom insulator cannot discharge the entire bitcell, a thinner bottom oxide can be used, resulting in lower operating voltages of ± 6V. The resulting structure, named "SONOS" (silicon-oxide-nitride-oxide-silicon) was initially built with a 2 nm-thin bottom oxide, resulting in the need for a multi-transistor bitcell or a more complex wiring in the array compared to the floating gate arrays, which relegated this NVM approach to niche markets. Recently, this structure has been enhanced with a thicker, 7 nm bottom oxide, and operated in a virtual ground array with hot electron injection programming and hot hole injection erase (see Figure 9). Instead of uniformly charging the nitride, only packets of charge are stored close to the source/drain terminals, resulting in the capability to store 2 to 4 bits in one bitcell [4]. To control the threshold voltage distributions during write/erase cycling, a microcontroller has been embedded on this stand-alone memory. Stand-alone memories based on nitride storage have entered volume production, replacing floating gate NOR modules in code and data storage applications [20].

4. NANOCRYSTAL CHARGE STORAGE

A layer of silicon nanocrystals of 5–20 nm diameter, sandwiched between two silicon dioxide layers, results in a charge storage medium with no lateral conductivity, as in nitride, but with deeper charge storage energy levels similar to a floating gate, see Figure 10. A single oxide defect will only result in insignificant charge loss out of one nanocrystal, therefore the dielectrics surrounding the nanocrystals can be scaled compared to floating gate technology. Initially, thin bottom oxides in combination with vertical direct tunneling operation were envisioned [19], but improved data retention is obtained with a 5 to 7 nm-thick bottom oxide, hot electron

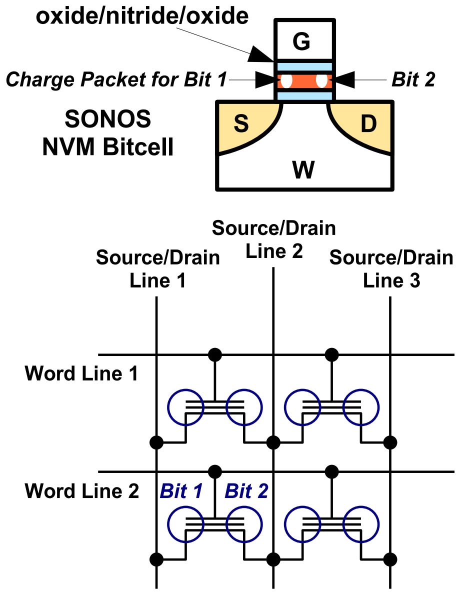


Figure 9: Bitcell, and virtual ground array schematics of 1-transistor bitcells for 2- or 4-bit per cell storage, using localized charge in a nitride film.

injection write, and Fowler-Nordheim tunneling erase [11], with a structure as shown in Figure 11. Advances in rapid thermal chemical vapor deposition of silicon have enabled the growth of silicon nanocrystals with high uniformity and reproduceability, see Figure 12.

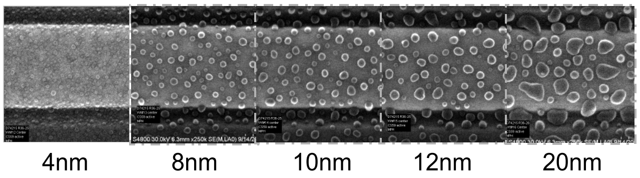


Figure 10: Growth of silicon nanocrystals on silicon dioxide bottom dielectric, vs. mean silicon thickness.

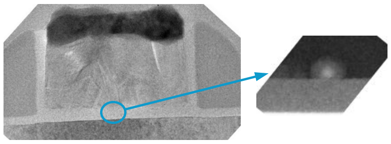


Figure 11: NVM bitcell with silicon nanocrystals sandwiched between two silicon dioxide layers.

The unique properties of silicon nanocrystals can be exploited either in a 1-transistor NOR architecture, resulting in a write/erase voltage reduction compared to floating gate technology from ± 9V to ± 6V, or in a split gate NOR architecture, fully optimized for small decode and charge pump areas (Figure 8) [22]. Multi-bit storage in a nanocrystal split gate bitcell also has been demonstrated [13, 15]. Productization of this technology is expected at the 65 nm technology node.

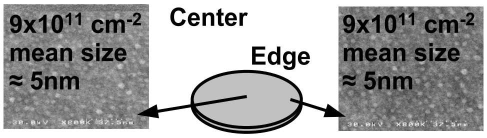


Figure 12: Top view SEM pictures of nanocrystal layer of center and edge of 8-inch wafer, showing uniform deposition.

In addition to changing the charge storage layer in the MOSFET gate oxide, other elements of the bitcell require innovation. To reduce short channel effects in scaled bitcells, dopants with low diffusion coefficients can be used to form very abrupt retrograde profiles for source/drain junctions and wells, resulting in enhanced scalability, and in tighter control of the bitcell threshold voltage. Figure 13 shows a schematic of a bitcell where the conventional boron well doping has been replaced by a background doping of phosphorus, and a peak of indium. The low diffusion coefficient of indium during subsequent thermal wafer processing results in the desired abrupt profile. Figure 14 shows the achieved improvement in the short channel effect, the dependence of threshold voltage on gate length [17].

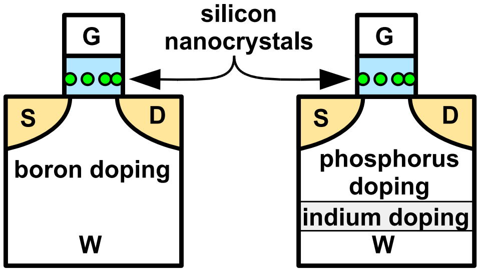


Figure 13: Schematic showing a conventional boron-doped well profile, and a steeply retrograde profile formed with indium/phosphorus doping.

5. EMERGING NVM CONCEPTS

While the paradigm of achieving nonvolatility by storing charge in the gate dielectric of a MOSFET is enabling a significant fraction of the semiconductor industry, the drawbacks of limited endurance, slow write/erase times, and high operating voltages have motivated the search for alternative methods of nonvolatile information storage compatible with mainstream silicon technology.

5.1 Magnetic Random Access Memory (MRAM)

In MRAM, information is stored in a magnetic tunnel junction consisting of a "pinned" magnetic layer, a aluminum oxide tunnel dielectric, and a "free" magnetic layer. The electrical resistance to vertical current flow perpendicular to the layers through the tunnel dielectric depends on the magnetization of the "free" layer with respect to the "pinned" layer, with the parallel orientation having a lower resistance than the anti-parallel orientation [18]. The magnetization in the "free" layer is switched by the combined magnetic field caused by currents flowing in two perpendicular conductors located close to the tunnel junction, as shown Figure 15.

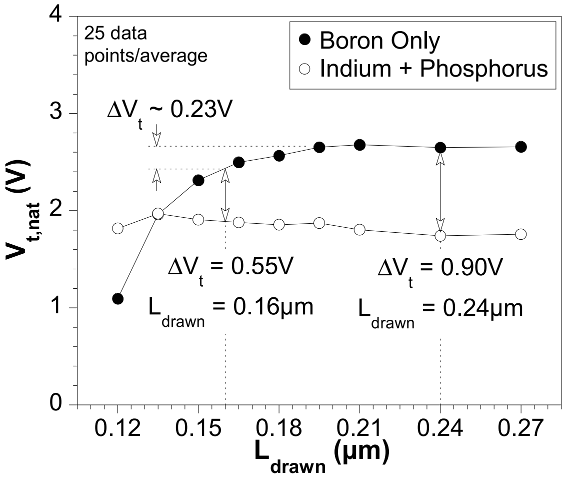
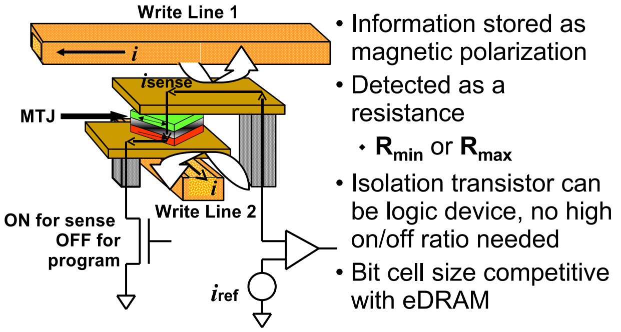


Figure 14: Dependence of uncharged threshold voltage of nanocrystal memory transistor on drawn channel length with conventional boron doping, and improvement achieved with indium/phosphorus retrograde wells.

The magnetic tunnel junction is typically embedded between two upper metalization layers of a memory, above the isolation transistors. Contrary to conventional floating gate NVM, the isolation transistor in the bitcell can be a low voltage logic device, and the row and column decoders are also constructed from low voltage transistors. Writing is fast, and there is no practical endurance limit. Read access time is competitive with SRAM (except the fastest) at a fraction of the cell size. Figure 16 is a die photo of a 0.18 μm technology 4 Mb memory, as well as selected memory properties [3, 1]. Alternative materials to increase the magnetoresistance ratio have been explored on 90 nm technology bitcells, indicating further scalability of the MRAM bitcell [16].

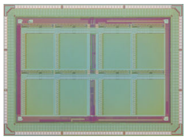


- Information stored as magnetic polarization
- Detected as a resistance
  - R$_{min}$ or R$_{max}$
- Isolation transistor can be logic device, no high on/off ratio needed
- Bit cell size competitive with eDRAM

Figure 15: Operating principle of MRAM.

5.2 Ferroelectric Memory (FeRAM)

In ferroelectric memories, information is stored as the polarization of certain dielectric materials (lead-zirconate-titanate "PZT" compounds) which can be switched by applying an external electric field to a capacitor containing the ferroelectric material. The memory is read by apply-



256K x16 Organization
3.0-3.6V Supply Voltage
Read cycle 25ns and 35ns
Write cycle 25ns and 35ns
Operating temperature 0-70°C
Data Retention >10 Years

Figure 16: 4 Mb MRAM memory die photo and properties.

ing a voltage to the capacitor and sensing the displacement current which depends on the polarization of the capacitor. High endurance has been obtained, and a 64 Mb memory has been demonstrated [5].

5.3 Phase Change Memory (PCM)

In phase change memories, information is stored as the phase (amorphous or crystalline) of a small volume of chalcogenide material [14]. The conversion from one to the other phase is achieved by heating a resistor in contact with the material in a well-controlled thermal cycle. If the material is melted and then cools down fast, the amorphous state is "frozen" despite the fact that the crystalline state is energetically favorable. Conversely, if the material is melted and then allowed to "anneal" at a temperature below the melting point, it reaches a crystalline state. The resistance through the material is 100 × lower for the crystalline state. High endurance of 10^{12} write/erase cycles has been achieved [10], and a 256 Mb PRAM has been demonstrated [8].

6. CONCLUSIONS

Scaling of the conventional floating gate nonvolatile memories is limited by reliability constraints, leading to the exploration and productization of alternative NVM concepts. The traditional paradigm of storing charge in the gate dielectric of a MOSFET is being extended by utilizing nanocrystal storage layers. Alternative NVM concepts are being pursued to more closely approximate a "universal" memory.

7. ACKNOWLEDGMENTS

Valuable inputs by Thomas Jew, Jon Slaughter, Ronnie Syzdek, and Saied Tehrani are gratefully acknowledged.

8. REFERENCES

[1] T. Andre et al. A 4 Mb 0.18 μm 1T1MTJ Toggle MRAM memory. In *2004 International Solid State Circuits Conference.*, 2004.

[2] W. Brown and J. Brewer. *Nonvolatile Semiconductor Memory Technology.* IEEE Press, New York, 1998.

[3] M. Durlam et al. A 0.18 μm 4 Mb Toggling MRAM. In *2003 International Electron Devices Meeting, Washington, D.C.*, 2003.

[4] B. Eitan et al. 4-bit per Cell NROM Reliability. In *2005 International Electron Devices Meeting, Washington, D.C.*, 2005.

[5] K. Hoya. A 64 Mb Chain FeRAM with Quad-BL architecture and 200MB/s Burst Mode. In *2006 International Solid-State Circuits Conference, San Francisco, CA.*, 2006.

[6] D. Kahng and S. Sze. A Floating Gate and its Application to Memory Devices. *Bell Syst. Tech. J.*, 46:1288, 1967.

[7] M. Kamiya et al. EPROM Cell with High Injection Efficiency. In *1982 International Electron Devices Meeting, San Francisco, CA.*, 1982.

[8] S. Kang et al. A 0.1 μm 1.8 V 256 Mb 66 MHz Synchronous Burst PRAM. In *2006 International Solid-State Circuits Conference, San Francisco, CA.*, 2006.

[9] P. Kuhn et al. A Reliability Methodology for Low Temperature Data Retention in Floating Gate Nonvolatile Memories. In *2001 International Reliability Physics Symposium, Orlando, FL.*, 2001.

[10] S. Lai. OUM - A 180 nm Nonvolatile Memory Cell Element Technology for Standalone and Embedded Applications. In *2001 International Electron Devices Meeting, Washington, D.C.*, 2001.

[11] R. Muralidhar et al. A 6V Embedded 90 nm Silicon Nanocrystal Nonvolatile Memory. In *2003 International Electron Devices Meeting, Washington, D.C.*, 2003.

[12] A. Niebel. Business Outlook for the Nonvolatile Memory Market. In *2006 Nonvolatile Semiconductor Memory Workshop, Monterey, CA.*, 2006.

[13] T. Osabe et al. Charge-Injection Length in Silicon Nanocrystal Memory Cells. In *2004 Symposium on VLSI Technology.*, 2004.

[14] S. Ovshinsky. Reversible Electrical Switching Phenomena in Disordered Structures. *Phys. Rev. Lett.*, 21:1450, 1968.

[15] E. Prinz et al. A 90 nm Embedded 2-Bit Per Cell Nanocrystal Flash EEPROM. In *2006 Nonvolatile Semiconductor Memory Workshop, Monterey, CA.*, 2006.

[16] J. Slaughter et al. High Speed Toggle MRAM with MgO-Based Tunnel Junctions. In *2005 International Electron Devices Meeting, Washington, D.C.*, 2005.

[17] C. Swift et al. Gate Disturb Reduction in a Silicon Nanocrystal Flash EEPROM by Means of Natural Threshold Voltage Reduction. In *2006 Nonvolatile Semiconductor Memory Workshop, Monterey, CA.*, 2006.

[18] S. Tehrani et al. Magnetoresistive Random Access Memory Using Magnetic Tunnel Junctions. *Proc. IEEE*, 91(5), 2003.

[19] S. Tiwari. Volatile and Non-Volatile Memories in Silicon with Nanocrystal Storage. In *1995 International Electron Devices Meeting, Washington, D.C.*, 1995.

[20] M. van Buskirk. MirrorBit^{TM} Technology, Past, Present, and Future: The On-going Scaling of Nitride-based Flash Memory. In *2006 Nonvolatile Semiconductor Memory Workshop, Monterey, CA.*, 2006.

[21] H. Wegener et al. The Variable Threshold Transistor, a New Electrically Alterable Non-Destructive Read Only Storage Device. In *1967 International Electron Devices Meeting, Washington, D.C.*, 1967.

[22] J. Yater et al. 90 nm Split-Gate Nanocrystal Non-Volatile Memory with Reduced Threshold Voltage. In *2006 Nonvolatile Semiconductor Memory Workshop, Monterey, CA.*, 2006.