Embedded Non Volatile Memories

Danny Shum
Logic and Embedded Alliance Development
Infineon Technologies Inc., Hopewell Junction, New York, USA

2070 Rt.52, 33i, Hopewell Junction, NY 12533
email: danny.shum@infineon.com

Abstract

The embedded Non Volatile Memories (NVM) offer in-system reprogrammability along with low power, noise reduction, more reliable, and reduced system cost while maintain high speed, high density system performance over other NVM including stand-alone Flash memory. The extent of market opportunity is driven by product cost. For most applications, manufacturability is the primary concern for embedded NVM, not cell size. The design simplicity is a key factor when choosing cell design. Another area of concern for cell design is the integration compatibility of the logic core process to the additional NVM flow. In this paper we will review different NVM concepts from ROM to Flash, and to FeRAM, and their applications in support of the diverse and emerging consumer market.

Introduction

Embedded NVM typically consist of ROM (Read Only Memory), EPROM (Electrical Programmable ROM), EEPROM (Electrical Erasable PROM), and Flash EEPROM of Si-based memories, and FeRAM/MRAM of non-Si-based memories. Referred to as charge-trapping devices, principle of Si-based cell is to shift the transistor threshold voltage by adding electrons (-Q, high Vt state), remove charges (neutral state), or adding holes (+Q, low Vt state) from the storage node as illustrated in Fig. 2a. Ferroelectric is based on polarization of crystal that creates hysteresis, as illustrated in Fig. 2b. The advantages of embedded NVM are in broad areas of applications such as communications, consumer electronics, automotive, and computer peripheral. Key benefits include dense board designs with reduced number of parts, reduced system costs, higher system speed due to fast code access, reduced noise, in-system on-board re-programmability of code and data storage, lower power dissipation, improved reliability, real time data storage during application, and others. Figs. 1a-1b give few examples of the embedded NVM products.

NVM Concepts

Floating Gate (FG) Memories

Fig. 3a illustrates the simplest concept of ROM which consist of a one-transistor cell, program either by channel Vt implant or some sort of fuse solutions. In all cases, the program was done during manufacturing; hence, it does not offer re-programmability. But it offers the smallest cell size of $3F^2$ (2Fx4F, where F is the minimum feature size). Floating gate concept is to place floating node above the transistor channel area to offer flexibility (Fig. 3b). Typical cell size is greater than 12F$^2$ (3Fx4F) in which the lateral dimension needs extra 1F in wordline to add capacitive-coupling between control gate CG-to-FG. Any charge stored on the FG node shifts the transistor threshold. Obvious conflicting requirements exist on this type of concept: a reliable charge transport mechanism through the gate dielectrics, and reliable charge storage, in the same dielectrics. First mass-produced FG is EPROM, known as FAMOS (Floating gate Avalanche-Injection MOS) device, introduced by Frohman-Bentchkowsky in 1971 [1]. Its charge transport employs channel hot-electron (CHE) injection from an avalanche plasma in the drain region underneath the gate for programming (electron stored in FG high Vt state, Fig. 4b) and UV erase (Fig. 4a). Endurance is limited to few cycles due to hot electrons degradation in tunnel oxide. EPROM is a
step above ROM to offer flexibility to change charge, but not in-system. Typical erase time takes 20' of UV exposure and its dielectric thickness is over 30nm to prevent hot electron damage. The next step is to add electrical erasability, which is known as EEPROM. Fowler-Nordheim (FN) tunneling is the mechanism [2] that allows electron transport at high field while retaining a high barrier at low field for charge retention (Fig 4c).

FG Memory Products

Conventional FG cell technology falls into two categories: planar and non-planar. In planar case, the electric injection field is slightly above the tunnel oxide field. Examples are Flash EPROM, EEPROM in which electrons transport by FN tunneling through the thin tunnel dielectric ~100Å oxide. In nonplanar case, the electric injection field is much greater than the tunnel oxide field. The e-field is enhanced by the surface topography on the cathode, which bend the oxide conduction band for electron tunneling at low field. It is known as textured poly device. Examples are poly-poly tunneling of SST cell (Fig. 5a) [3]. Typical oxide thickness is in the range of 200-350Å. Flash EEPROM (byte programmability and block/sector erase) can be classified by programming: CHE injection and FN tunneling. In CHE injection the cell is known to as SIMOS (Stacked gate Injection MOS). The smallest cell is the IT structure (Fig. 4b) such as ETOX™4 (EPROM Tunnel Oxide) with asymmetric S/D junction. The abrupt drain junction is optimized for CHE programming while the graded source region is intended for source-erase type [5] with FN tunneling. This type of cell is usually small but often requires complicated algorithm or 2-step erase scheme [6] with added circuitry to control Vt distribution. The split-gate 1ST structure such as SCSG (Fig. 5b) is a relax cell in series with an enhancement transistor to offer design simplicity for over-erase Vt control without complicated algorithm or erase scheme [7]. The series transistor ensures a high immunity to drain turn-on, and put a limit on the minimum Lcell of the NVM transistor. This cell also eliminates or improves source-drain punch-through problems. Triple-poly structure with Source-Side Injection (SSI, Fig. 3c) is designed to reduce programming current from few mA to few µA for low power applications [8]. In FN tunneling the cell is referred to as FLOTOX (FLaunting gate THin Oxide, Fig. 5d), where the tunneling area is a small thin oxide region over the gate. A variation to incorporate thin oxide over the entire FG transistor channel area creates alternative memory cells with improved performance. This type of NVM cells is called FETMOS (FG Electron Tunneling MOS). The device structure is simple IT cell and is further divided into different memory organizations such as NAND (Fig. 6a) [9], NOR (Fig. 6b), AND (Fig. 6c) [10], DiNOR (Fig. 6d) [11], and other combinations for high density, low power applications. The split-gate 1ST (Fig. 6e) with an enhancement transistor in series to the FG cell [12] and 2T (Fig. 6f) with a select transistor per cell for byte-addressable function [13] are excellent choice for embedded applications without complicated algorithm or overhead peripheral circuitry for Vt control and still offer low power compatibility with FN tunneling W/E operations.

Non-Floating Gate Memories

MNOS (metal-nitride-oxide-silicon) is the first non-FG device that stored charges in discrete traps in the nitrile layer [14]. SONOS (Silicon-Oxide-Nitride-Oxide, Fig. 7) is the improved MNOS device to inhibit gate injection and to block charges injected from the top oxide-nitride interface, resulting in a higher trapping efficiency and more reliable memory cell. The device can be programmed by FN tunneling [15] or CHE injection [NROM, 16] or SSI [SPIN, 17], and erased by hot hole injection. The key advantage is the elimination of FG; thus, reduce the process complexity and compatible to most single polysilicon CMOS manufacture. The cell size is reduced to a potential 8F² by elimination of 1F from the coupling regions.

FeRAM (Ferro-electric RAM)

Ferro-electric is a new type of NVM whose operation is based on non-linear dielectric crystals. Lattice displacements of atoms induce a polarization of the crystal under an external E-field and to remain polarized after the field removal. The polarization can be reversed by applying an opposite polarity field. Thus, a non volatile capacitor is obtained in which stored information is based on polarization state rather than on stored charge [18]. The stored data can be read by sensing the interaction of a "read field" with the polarization state of the element. If a read voltage is applied to the capacitor of polarity opposite to the previous write voltage, the polarization state will switch to cause electrical hysteresis, giving rise to a large displacement current that sensed by Sense Amplifier (Fig. 8). Examples of some Ferroelectrics materials are listed below:

- Strontium Bismuth Tantulate (SBT)
- Lead Zirconate Titanate (PZT)
- Lead Titanate (PbTiO3)
- Strontium Bismuth Titanate
- Barium Strontium Titanate (BST)
- Barium Magnesium Fluoride (BaMgF4)

Some of the advantages of FeRAM over conventional Si-based NVM are extended read/write cycles in 10¹²-10¹⁵ ranges, low program voltage (3-5V), and fast read/write time in ns.
Process Issues

A double-polysilicon stacked gate CMOS process is needed to fabricate NVMs. The first poly is used as the FG and the second poly as the CG. A high quality thin tunnel oxide, below 100Å, is formed between FG and Si with junctions overlap the gate edge to reduce substrate current. Post-gate poly process is another critical step. The post-annealing temperature can significantly affect the quality of the thin tunnel oxide. Local oxide thinning and oxide surface roughness at the poly/oxide interface are too influenced by post-gate temperature treatment. They are the direct consequence of the grain growth of the poly gate and viscous flow of the oxide, which are enhanced with rising annealing temperature and time. In addition, poly gate doping level and species play a major role in local variation of tunneling current. ONO interpoly dielectrics are formed between FG and CG to isolate FG from electron leakage or to block hole injection from CG. Typical thickness is in the 200Å range with scaling down to 120-150Å [19]. A high selective RIE polysilicon gate etch is another key process development to achieve high manufacturability. The double-polysilicon self-aligned (SA) gate structure requires a high precision stacked-gate etch, which is an anisotropic dry etching process, new to most wafer manufacture [20]. The RIE process etches through the stacked structure CG-ONO-FG and stop at thin tunnel oxide to avoid Si damage from etching. An alternative is non-SA gate structure which simplify etch process and avoid alignment sensitivity of the FG length. The penalty is a few percent cell area increase and higher bitline capacitance, but the process is more compatible to the baseline. Well formation is another critical area that requires MeV implant of N-band region to isolate array P-well from the P-type substrate. For 0.25um and below generations, standard shallow trench isolation (STI) process or modified scheme with more rounded corner is common practice among the industry. HV module is formed with conventional dual gate oxide scheme, resist patterning and wet etch. HV gate oxide thickness is in the range of 150-250Å, pending on the cell type. Figure 8 shows the cross-section along the WL and the BL directions [21]. The pictures illustrate many process issues such as gate oxide thinning around STI or FG corner, STI oxide thinning through the stacked gate cell, general oxide qualities and poly doping, and many others [22].

FeRAM processing presents contamination issues of ferroelectric materials to conventional CMOS manufacturing, new tools, different electrodes of Pt, Ir, Ru and their barriers that are critical to integration, polycrystalline thin film and its properties that are a function of morphology, stoichiometry, and thickness. The process must be compatible with ILD, metallization, forming gas and other BEOL processing steps. Other issues are modeling of aging effects, device, Spice simulations, general CAD and analog circuit support.

Challenges of Embedded NVMs

Several process approaches have been discussed. The main stream is either stand-alone approach for low cost and simplify process modules development, or logic based approach for high performance and simplify design reuse. The challenge for logic based is the addition of high voltage transistors (dual gate oxide) and Flash cells (tunnel oxide and ONO dielectrics) and the additional thermal budget that pleases the NVM reliability but prove to be nightmare for logic transistor design of maintaining the same transistors characteristics. Logic based process must use modular approach for compatibility of core logic and embedded memory process [21]. The challenge for stand-alone is how to convert the single gate oxide, buried P-channel transistors, and polyoxide process to high performance dual gate oxide, dual work function, salicide process for the high speed peripheral circuitry that is required by the system design.

Reliability Issues

The issues are wide open that include topics such as tunnel oxide defect density and burn-in screening, ONO trapping density, cell design, endurance, retention, stress induced (dielectrics) leakage current (SLIC) through trap center after endurance, charge induced dielectric breakdown and others. Tunnel oxide inline monitor is to exclude the extrinsic defect density after deposition in addition to wafer level reliability (WLR) at the end of process. WLR includes traditional TDB on special test structures, in addition to endurance and retention tests on actual product macros through random sample selection by number of wafers and number of lots per week. Recent improvement of tunnel oxide with nitridation of the preoxide to form N2O oxynitride can improve endurance characteristics well beyond 108 cycles range [23]. Data retention is lost of stored information due to FG charge leakage. Typically below 1V charge loss requires less than 1 electron per day of leakage. Endurance is hard bit fails due to repeat operations of write-read-erase. Program disturbs as in Fig. 6b cause adjacent cell high leakage along the same BL, resulting in read error if initial Vt-state is low. Short channel effect, punch-through, or create hot hole injection.

BEOL processing issues of multi-level metallization (MLM), Cu wire, HDP oxide, low K dielectric, and how in dealing with the cell endurance have just begun. Electrostatic charging of the WL, boron density of BPSG and hydrogen content of IMD all will have reliability effects on the cell retention.
Most semiconductor components are tested by voltage and temperature acceleration assuming some level of activation. NVM has the same reliability problems and follows the same acceleration as the semiconductor components. The "classical" burn-in screening is by high temperature acceleration such as 150°C to exclude the hard bit fails. Soft fails are the abnormal "moving bits" that screen with lower temperature such as below 80°C [21]. Other issues include erratic bits, endurance fails, and program disturbs. The work-around solutions to aid process reliability to solve the soft fails are the implementation of design-in tests such as ECC, recovery algorithms or on-chip failure corrections. The exact implementation will depend on company traditions.

Embedded NVM Module Design

Several key areas of design must enter into considerations. Array architecture is the number one priority that include program or erase disturbs, high performance and low stand-by power, erase block size variations, unit block design flexible for different array size implementations, and chip area optimization. Several analog block of designs are the basis for memory: sense amplifier and reference circuits, WL and BL drivers and decoders, global control and glue logic for bus interface, charge pumps and DC power system, high voltage circuits and level shifters [24].

Embedded NVM Test Challenge

Issues are cost concern that embedded NVM must be tested along with high speed logics using a high speed tester that is more expensive and less efficient for NVM than a specialized memory tester. Test cost for embedded NVM can be a significant portion of the device cost. The test time could be reduced through the write time reduction such as design for multiple-word or page mode programming, or simultaneous program of multiple NVM modules on-chip. A highly reliable process and a robust design can lead to significant tests required. Use of on-board CPU and RAM to perform self-test and other tasks can reduce test code development [25].

Conclusion

We have shown that embedded NVM parts offer performance, cost advantages and more robust system solutions over stand-alone NVM. Micro-controller with embedded NVMs will gain market share while stand-alone will have difficulty to match performance. There are different NVM concepts (ROM...Flash) established. The applications define the most practicable one. The extent of market opportunity is driven by product cost. The parts are expected to operate at -40°C and extend to 85°C, suitable for a wide range of wireless communication products or smart-card products. For most applications, manufacturability is the primary concern, not cell size, followed by ease of design solutions and test methodologies for embedded NVM. FeRAM is a challenging technology concept but also extreme interesting because of the potentials in support of the diverse and emerging consumer market.

Acknowledgments

The author wishes to thank Josef Winneker, Horst Gant, and Franz Neppel for their management support. Special thanks to Georg Tempel for technical discussions.

References

[25] Nonvolatile Semiconductor Memory Technology
Fig. 1a Example of Embedded NVM Product with 2Mb Flash and 4Kbyte Dataflash.

Fig. 1b Embedded ProASIC Technology (courtesy of Gatefield Corp.) Flash-based FPGA Solution.

Fig. 2 Principle function of cell operations: a) FG based NVM:

Fig. 2b FeRAM.

Fig. 3 a) ROM; b) EPROM.

Fig. 4 EPROM Si/SiO₂ Energy Band Diagram and its transport mechanism. a) neutral state; b) program operation. c) erase operation.
5c) Triple poly SSI cell  
5d) FLOTOX EEPROM cell

NAND

Fig 6a) NAND architecture, 16 bit per unit cell in series

5V D0 0V D1
W0 -9V W1 0V
common
Source

6b) NOR architecture, common source
6c) AND architecture

6d) DiNOR cell  
6e) Siemens 1.5T split-gate cell edge programming

6f) 2T EEPROM

Fig 7 MNOS  
a) SONOS  
b) SPIN

Fig.7 Principle of SBT capacitor

Fig.8 Cross-section of the cell along a) WL  
b) BL.