[illegible]

Takashi Kono
Renesas Electronics Corporation, Japan

ABSTRACT

As an enabler of sustainable and innovative progress in wide range of applications from automotive to IoT, embedded non-volatile memory (eNVM) including embedded Flash (eFlash) has been and will be playing key roles. To meet more stringent and/or diversified requirements while overcoming many challenges along with memory cell scaling, hierarchical optimization in eNVM design from memory cell to hard macro and system level has greater significance. By taking MONOS based eFlash design as a preferred example, this paper will survey the latest eNVM design in hierarchical manner as well as future directions of eNVMs including emerging memories.

INTRODUCTION

In the upcoming IoT era, smarter societies will be created based on the concept of "Cyber-Physical System (CPS)", which consists of cloud systems to process vast amount of data with more advanced learning technologies and a huge number of edge devices connected via networks mutually and to cloud systems for sensing and actuating "things" in physical world. In the cars with ADAS (Advanced Driver Assistance System) or future autonomous cars, detailed information around cars such as the positions of human, other cars and obstacles are collected by several types of sensors and analyzed based on complicated algorithm and such reference information in cloud systems as maps and weather condition. Finally, proper vehicle controls (moving, stopping, turning) are determined within very stringent time constraint and fed back to actual electrical/mechanical components in a car. In smarter factories supported by machine-to-machine (M2M) interconnection, the condition data in production lines are collected by and shared among edge devices with sensors. Some of the data are sent to cloud systems and analyzed by high level algorithm using such techniques as machine learning. According to the analysis results on local and cloud level, the necessary actions are automatically determined and fed back to proper edge devices in production lines so that they can prevent troubles from occurring and maximize output efficiency.

As interfaces to "physical" world via sensing, communication and feedback to control, each edge device such as a microcontroller unit (MCU) has its own embedded system so as to meet the requirements of real-time operation, low power consumption, reliability, security and safety. Embedded non-volatile memories (eNVMs) as represented by embedded Flash (eFlash) are one of the most essential technologies in embedded systems because of programmability, non-volatility and embeddedness [1]. Especially, eFlash has been and will be widely used in many applications from automotive to consumer and newly emerging applications such as medical and wearable. Each application has specific requirements to eFlash. For example, very stringent reliability under harsh junction temperature up to 170 °C is strongly required in automotive applications. Meanwhile, in low power applications, eFlash is expected to be a key enabler for efficient intermittent operations.

To sufficiently satisfy application-oriented requirements, hierarchical and systematic approach is essential in eFlash design. Fig. 1 describes a hierarchical eFlash system design composed by three levels. Level-1 is related to memory cells. Based on the requirements from applications, the type of memory cell should be carefully selected because it dominantly defines the baseline of total electrical characteristics and reliability. Level-2 focuses on eFlash hard macros, which include the design of memory array architecture and such critical circuits as sense amplifiers, high voltage generators, and so on. Electrical characteristics and data reliability of an eFlash system can be strongly dependent on the design quality at Level-2. Level-3 includes the design techniques at system level. Each functional block in an eFlash system should share proper roles to enhance and achieve target specifications as a system.

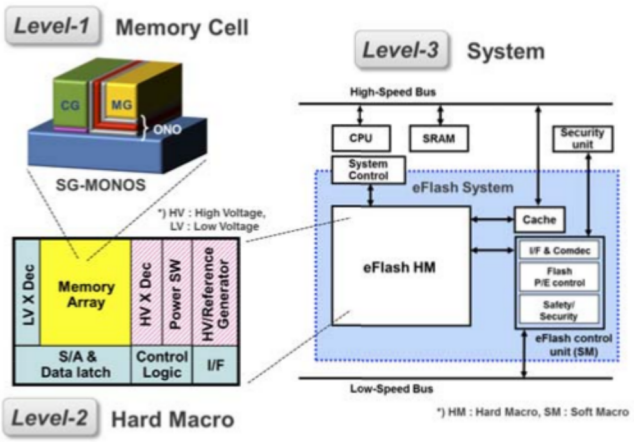

FIGURE 1. HIERARCHY IN EMBEDDED FLASH SYSTEM DESIGN

HIERARCHICAL eFLASH SYSTEM DESIGN

Level-1: Choice of eFlash Memory Cell

In the history of flash memory R&D, many cell types have been proposed for stand-alone and embedded uses. For stand-alone uses such as NAND, smaller cell size is important to achieve lower bit cost. On the other hand, reliability and performance often take the first priority for embedded uses. This is why split-gate structure is actively accepted in eFlash memory cells. Split-gate structure offers the following attractive advantages; a) source-side hot carrier injection capability for fast programming with very small program cell current (~1uA/cell), b) fast readout path design with low-voltage logic transistors thanks to low voltage bias at bit-lines (BLs) in all operations, and c) flexible Vth level setting (lower than 0V) for erased cells without any concerns about depletion issues during read operations. Fig. 2 shows the major eFlash memory cells for MCU applications. The main factors for an eFlash memory cell to survive scaling races beyond 90nm are 1) high scalability without degrading reliability and performance, and 2) high affinity with baseline logic process. As for 2), gate height of logic transistors has been getting lower along with scaling, which makes it difficult to integrate eFlash memory cells with high cell height into advanced baseline logic process. Furthermore, in FinFET process, gate structure of logic

978-1-5090-3969-2/17/$31.00 ©2017 IEEE

transistors becomes three-dimensional. Therefore, affinity with baseline logic process is much more important in advanced nodes.

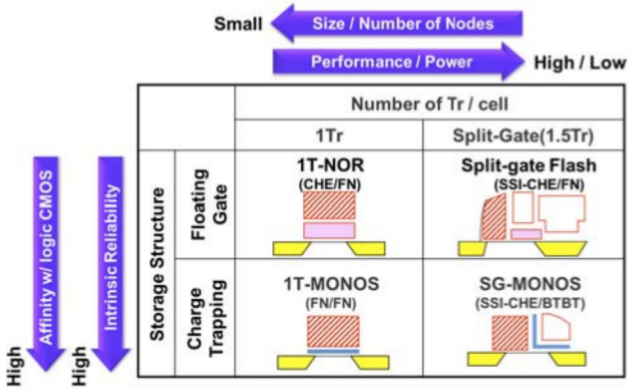

FIGURE 2. CONVERGENCE OF EFLASH MEMORY CELL

Given this drastic evolution of logic process, charge trapping (CT) type eFlash memory cells have great advantages over floating gate (FG) type eFlash memory cells. In CT type cells, charges are stored by the traps which are separately located in thin nitride film and at the interface of nitride and oxide. This means that CT type cells are intrinsically more reliable in terms of data retention thanks to better tolerance against defects. In addition, owing to thin film storage, CT type cells have lower cell height, resulting in higher affinity (better embeddedness) with advanced logic process. Split-gate (SG) MONOS memory cell, a kind of CT type cells combined with split-gate structure, has successfully proven the advantages of CT type cells in terms of scalability and affinity with advanced logic process. It can meet very stringent requirements of performance and reliability from automotive applications [2,3] and demonstrates excellent scalability down to 16nm/14nm and beyond with FinFET structure [4] as shown in Fig. 3.

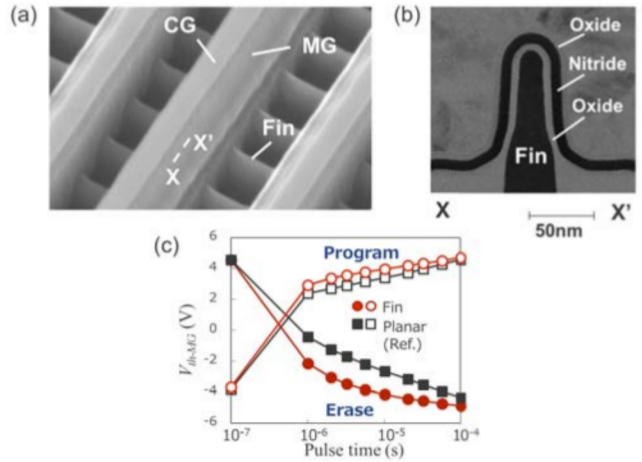

FIGURE 3. 16NM FINFET SG-MONOS [4] (A) BIRD'S EYE VIEW, (B) CROSS-SECTIONAL VIEW, AND (C) P/E CHARACTERISTICS

Level-2: Memory Array & Peripheral Circuit Design

The total performance, reliability and cost of an eFlash system is largely determined by not only memory cells but also design concept and circuit techniques for eFlash hard macros (HMs). Especially when targeting high-end applications such as automotive and industry, memory array architecture and related peripheral circuits should be carefully studied, selected, and properly implemented in eFlash HMs so that the intrinsic potential of eFlash memory cells can be fully derived and further enhanced. At the same time, reliability of peripheral circuits should be taken into account in advanced process node. Cost competitiveness is also of great concern and can be boosted by not only HM area reduction but also improved testability. Proper DFT (Design for Testing) effectively reduces testing time and screens potential failures after shipment, resulting in better cost competitiveness and customers' trust.

eFlash memory array architecture should be determined in deep consideration of target applications and cost. The interval of WL stich regions and/or the number of cells per BL have potent influence on random read access performance. Regarding reliability, program disturb is one of the big concerns. In general, the less cells are connected to the same nodes to which high voltages are applied during program operations, the less vulnerable these cells are to program disturb, resulting in better reliability. Read disturb is another concern especially in one transistor type cells. One transistor MONOS (1T-MONOS) [5] has realized read disturb-free memory array by introducing proper voltage settings in read operation and meets reliability requirements from automotive applications.

As for the importance of peripheral circuit design, a readout path including sense amplifiers (S/As) and reference voltage/current generators are greatly responsible for total performance and reliability of eFlash HMs. A sensitive readout path can correctly read out the data from memory cells with narrow Vth window or even after long retention time. Moreover, it can realize less electrical stress on memory cells and better reliability. The necessary Vth window or cell current difference is largely determined by sensitivity (or "offset") of S/As and temperature dependency mismatch between cell current and reference current. Many ideas about S/As equipped with analog-type offset cancelling scheme have been proposed [6,7]. However, they are not always suitable for fast read operation over 100 MHz because offset cancelling operation is needed in each read cycle and causes relatively large timing overhead in case of read operation over 100MHz. SA with digital offset cancellation (SA-DOC) in Fig. 4 is one of the promising solutions to overcome this problem [2]. 40nm SG-MONOS achieved 160MHz random read operation for code flash under junction temperature (Tj) of 170°C by effective combination of SA-DOC and split-gate structure cell. Furthermore, 28nm SG-MONOS achieved not only 200MHz random read operations for code flash but also over 1M cycle endurance for data flash thanks to intrinsically high reliability of memory cells and stress-mitigation design techniques [3].

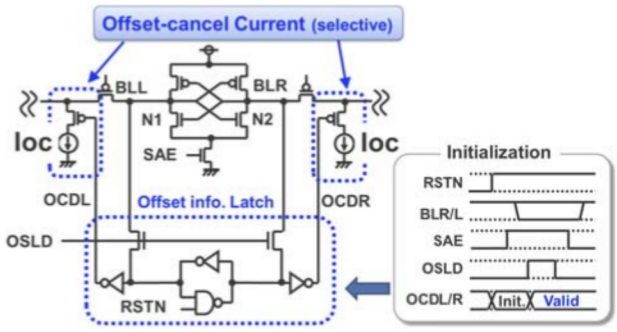

FIGURE 4. S/A WITH DIGITAL OFFSET CANCELLATION

"Process friendly" peripheral circuit design is also of great importance in eFlash HM design with advanced eFlash technology integrated in standard logic CMOS process. In general, not only the thickness of logic CMOS transistor's gate oxide but also that of dielectric films between interconnect metal layers get thinner along with scaling. This may cause the degradation of interconnect metal

time dependent dielectric breakdown (IM-TDDB) lifetime in advanced eFlash process. To tackle this issue, 28nm SG-MONOS eFlash HM adopted temperature-adaptive step pulse erase control (TASPEC) [3]. Given opposing temperature dependence of IM-TDDB lifetime (worst at high temperature) and erase speed (worst at low temperature), TASPEC adaptively controls the maximum erase voltages and avoid degrading IM-TDDB lifetime at high temperature.

Level-3: System Level Design

An eFlash system mainly consists of one or multiple eFlash HMs and several dedicated soft macros (SMs). These SMs contribute to enhancing and diversifying system level specifications. Especially, they play key roles in satisfying some of the strong requirements to embedded systems in IoT era; 1) low energy/power consumption, 2) in-field software (SW) upgrade and 3) security.

Intermittent operation is a key technology to achieve low energy consumption in embedded systems for such applications as real-time environment or biometric monitoring. Efficient system power management based on power gating and evacuation of processed data and system status to non-volatile memories (or SRAM in retention mode) are indispensable. 1T-MONOS based embedded system, as shown in Fig. 5, has high affinity with low energy intermittent operation [5]. Energy for programming is very low thanks to FN programming and optimized high voltage generation techniques. Idling P/E management unit (IPMEU) takes over P/E management and on-chip power control after other elements including CPU and SRAM transit to standby state and achieves drastic reduction of energy consumption as a system.

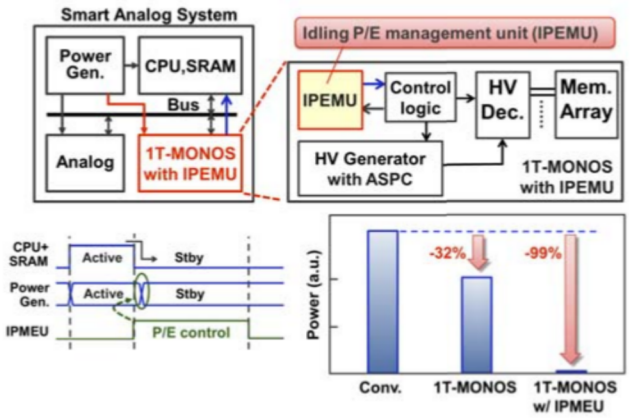

FIGURE 5. SMART ANALOG SYSTEM WITH 1T-MONOS TECHNOLOGY

Some applications need eNVM systems which simultaneously achieve low power consumption and high performance. One of the examples is human-body attached motion sensing application in the field of sports, which pursues higher sampling rate and smaller form factor with small battery. Memory hierarchy composed by dedicated cache and eNVM with related power management can meet these requirements. Fig. 6 shows the sensing system based on 40nm low power MCU with low power parallel cache system and SG-MONOS flash [8]. Low power parallel cache system features the addition of small cache, which has the same latency as large one and much lower current consumption owing to flip-flop based design. Given that most of instruction codes consist of a loop which can be stored in small cache, hit efficiency to small cache is quite high, resulting in ultra-low power operation. Since both cache operate in parallel, there is no timing penalty when small cache misses and large cache hits. In addition, during the processing status which needs larger amount of instruction codes, instruction codes are copied to large cache and the power supply to SG-MONOS is turned off. The proposed system achieved energy efficiency of 20uA/MHz at 200MHz.

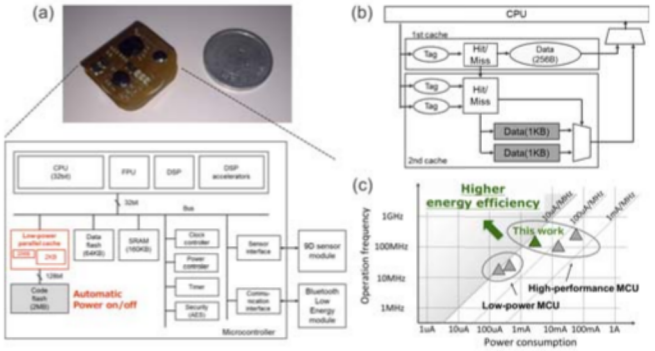

FIGURE 6. 40NM LOW POWER MCU FOR SENSING APPLICATIONS [8] (A) SYSTEM BLOCK DIAGRAM, (B) LOW POWER PARALLEL CACHE, AND (C) POWER EFFICIENCY

Capability of in-field SW upgrade is one of the original benefits brought to MCUs by introduction of eFlash in place of mask ROM. With the advent of IoT era in which everything is mutually connected via internet, in-field SW upgradability is coming into a new phase of wireless-communicated program updates called OTA (Over-The-Air). SW upgrade by OTA will prevail in many applications as a key technique to reduce maintenance cost and prolong product lifetime. For example, in near-future automotive societies, big benefits will be brought by OTA-based SW upgrade to both car owners (no need to bring cars to dealers) and dealers (get cost reduction and better customer satisfaction). OTA-based SW upgrade makes new demands to eFlash system. First, larger flash capacity (~2x) is needed for backup storage during update, which can motivate further scaling of eFlash memory cells. In addition, proper address translation or switch scheme should be implemented so that the blocks storing updated codes are safely accessible while the access to those storing old codes is prohibited after the update is completed. EMI reduction remedies may be required in eFlash system to prevent noise on power line by P/E operations from causing malfunctions or reducing operational margin in the eFlash system itself and/or upper level systems [3]. Security issues in OTA are also of great concern.

The importance of security has been and will be drastically increasing for secure real-time data collection and transmission among connected devices and systems. In eFlash systems, dedicated function blocks in eFlash control unit serve as "gate keepers" by monitoring access to eFlash system and preventing illegal data fetch and tampering by authentication. In addition, eFlash HM can offer one-time programmable (OTP) area or access-inhibit area to avoid overwriting important data. Conventionally, security functions were implemented based on each vendors' security policy. However, in response to the recent gain of security risk in connected world, common security policies are being discussed and developed (e.g. EVITA [9] and AUTOSAR [10] for automotive).

PROSPECT OF EMERGING MEMORIES AS eNVM

Onset of full-scale CPS and its continuous evolution will drastically change our societies and make them much smarter. This technology trend will definitely and significantly impact on the requirements to eNVMs in edge devices. Given total energy consumption by huge amount of edge devices in systems, more advanced architectures such as "normally-off" computation (Fig.7) are strongly expected for much lower energy consumption [11]. In addition, higher endurance and low cost are other key factors so that

edge devices spread widely throughout the society and collect various kinds of information about "things".

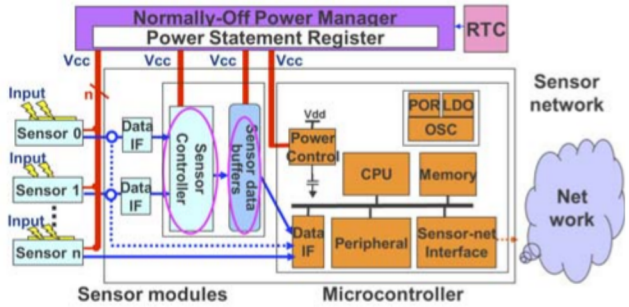

FIGURE 7. EXAMPLE OF NORMALLY-OFF ARCHITECTURE

Emerging memories (e.g. STT-MRAM, PCM and ReRAM) have been intensively studied for more than 10 years not only for stand-alone uses such as storage class memory (SCM) but also for embedded uses. Especially, as eNVM candidates, they are expected to be a key enabler for normally-off operations with no DC current during standby period and minimum timing overhead in transition from standby to active and vice versus. In addition, they have an attractive potential for higher endurance, which can widen eNVM application area. Moreover, only a few additional masks in BEOL are needed to baseline CMOS process. As an example, perpendicular STT-MRAM has recently attracted much attention as a leading candidate to replace embedded SRAMs for cache usages based on its potential of non-volatility, short write time (<50ns) and high endurance (>10^8 cycles) [12].

In reality, emerging memories are facing some technical challenges which prevent them from being adopted in commercially available products. Major issues are commonly seen in data retention at high temperature, program current scaling, quality of critical films and variability. Given their MCU applications, the data stored in embedded emerging memories (eEMs) must be retained not only under normal operation temperature profile but also at solder reflow conditions (e.g. 260°C, 3min.), because boot code programs and system trimming parameters are generally stored in eEMs prior to shipment and board assembly on user sites. Therefore, data retention at high temperature is one of the key requirements in MCU applications. As for program current scaling, program current of emerging memories is typically around 100uA/bit or more, which is much larger than that of eFlash (~1uA/bit in SG-MONOS). The tradeoff between program current and retention makes it difficult to reduce program current without degrading retention capability [13, 14]. This also prevents the scaling of access transistor size in 1T1R type cell, leading to another difficulty in cell size scalability. Cross point type cell can solve this issue in some applications, but R&D to find proper selector/diode devices still has a long way to go [15]. Quality of critical films such as magnetic tunnel junction (MTJ) film in STT-MRAM or metal oxide in ReRAM hasn't yet reached sufficient level and limit endurance characteristics. From design viewpoint, small magneto-resistance (MR) ratio and read disturb in STT-MRAM, large resistance drift in PCM and random telegraph noise (RTN) / noise-induced resistance broadening in ReRAM pose critical challenges for read circuit design. Further investigations to solve these technical issues are strongly needed from both sides of material engineering and circuit/system design.

CONCLUSION

As a key enabler to realize smarter society, eNVMs represented by eFlash will continue to play key roles in full-scale CPS. To satisfy more diversified and stringent requirements from edge devices in wider range of applications, the importance of hierarchical design approach for eNVM system is increasing. In this respect, SG-MONOS and 1T-MONOS technologies are the preferable examples. SG-MONOS is leading eNVM races and satisfies the requirements from automotive applications with proven memory cell scalability and design techniques to enhance performance and reliability. The embedded systems with 1T-MONOS and SG-MONOS technologies successfully achieved low energy consumption and/or high energy efficiency based on well-chosen system level design. Looking ahead to the future, there are strong expectations toward emerging memories as game changers in embedded systems for ideal normally-off operations and practically infinite endurance. Currently, several technical issues including reliability stand in their way. Further hard works in emerging memory R&D to solve them will open up a new horizon to their practical use and the whole eNVM applications.

REFERENCES

[1] H. Hidaka, et al., "Embedded Flash Memory for Embedded Systems", to be published from Springer.

[2] T. Kono, et al., "40nm Embedded SG-MONOS Flash Macros for Automotive with 160MHz Random Access for Code and Endurance Over 10M Cycles for Data", ISSCC Dig. Tech. Papers, pp. 212-213, 2013.

[3] Y. Taito, et al., "A 28nm Embedded SG-MONOS Flash Macro for Automotive Achieving 200MHz Read Operation and 2.0MB/s Write Throughput at Tj of 170°C", ISSCC Dig. Tech. Papers, pp. 132-133, 2015.

[4] S. Tsuda, et al., "First Demonstration of FinFET Split-Gate MONOS for High-Speed and Highly-Reliable Embedded Flash in 16/14nm-node and beyond", IEDM Tech. Dig., pp. 280-283, 2016.

[5] H. Mitani, et al., "A 90nm Embedded 1T-MONOS Flash Macro for Automotive Applications with 0.07mJ/8kB Rewrite Energy and Endurance Over 100M Cycles Under Tj of 175°C", ISSCC Dig. Tech. Papers, pp. 140-141, 2016.

[6] M.-F. Chang, et al., "An offset-tolerant current-sampling-based sense amplifier for sub-100nA-cell-current nonvolatile memory", ISSCC Dig. Tech. Papers, pp. 206-208, 2011.

[7] M. Jefremow, et al., "Time-Differential Sense Amplifier for Sub-80mV Bitline Voltage Embedded STT-MRAM in 40nm CMOS", ISSCC Dig. Tech. Papers, pp. 216-218, 2013.

[8] M. Nakajima, et al., "A 20uA/MHz at 200MHz Microcontroller with Low Power Memory Access Scheme for Small Sensing Nodes", in IEEE Symposium on Low-Power and High-Speed Chips, COOL CHIPS XIX, 2016.

[9] http://evita-project.org/

[10] http://www.autosar.org/

[11] M. Hayashikoshi, et al., "Low-Power Multi-Sensor System with Normally-off Sensing Technology for IoT Applications", International SoC Design Conference, pp.195-196, 2016.

[12] Y.-J. Song, et al., "Highly Functional and Reliable 8Mb STT-MRAM Embedded in 28nm Logic", IEDM Tech. Dig., pp. 663-666, 2016.

[13] A. V. Khvalkovskiy, et al., "Basic principles of STT-MRAM cell operation in memory arrays", J. Phys. D, Appl. Phys., vol. 46, no. 7, Feb. 2013, Art. no. 074001.

[14] R. Yasuhara, et al., "Consideration of Conductive Filament for Realization of Low-current and Highly-reliable TaOx ReRAM", IMW, pp.34-37, 2013.

[15] R. Aluguri, et al., "Overview of Selector Devices for 3-D Stackable Cross Point RRAM Arrays", IEEE J. Electron Devices Society, vol.4, no.5, pp. 294-306, 2016.