JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

CHALLENGES IN THE DOWNSCALING OF SEMICONDUCTOR MEMORY

Saji Joseph
Department of Physics
Pavanatma College, Murickassery, Kerala, India

Abstract : Requirements, challenges and the scaling trends of the memory system over the years as well as for future applications are discussed. Some promising design directions to overcome challenges posed by scaling of volatile and nonvolatile memory are reviewed. Introduction of high-density 6T SRAM cells and high mobility channel (HMC) is expected to improve the scalability of SRAM. Current 1T1C DRAM structure may extend for another few years, but the scaling may stop below 10nm. So, to continue DRAM scaling, cells with vertical gate and capacitorless DRAM cells are to be adopted. Presently, feasibility of multi-bit storage technology and multi-layer integration (MLC) are being explored for the scaling of flash memory. However, several emerging non-charge based storage memories such as FeRAM, MRAM, PCRAM, and ReRAM are also being investigated due to the difficulty in stacking too many layers in MLC.

Index Terms - **SRAM, DRAM, flash, FeRAM, ferroelectric memory.**

I. INTRODUCTION

Since the introduction of the integrated circuit chips in the early 1960s, the number of processing elements in a single chip has multiplied many folds. This has resulted in surprisingly high processing power for present day computational systems. During the last sixty years, the semiconducting industry was able to reduce significantly the size of the Metal-Oxide-Semiconductor-Field-Effect-Transistor (MOSFET), which is the basic building block of all IC chips. In fact, from the days of first integrated circuit in the 60s, the transistor dimensions have been steadily decreasing obeying what is popularly known as Moore's law [1] which states that the number of transistors in a dense integrated circuit doubles about every one and a half years resulting in low power consumption and higher processing speed. The semiconducting industry successfully developed many new lithographic techniques during this period, which caused this surprising growth in processing power. The progress in lithographic techniques after 2010 includes the development of 193 nm ArF immersion with multiple patterning [2], self-aligned double and quadruple patterning (SADP and SAQP) [3], [4] and Extreme Ultraviolet (EUV) lithography technique [5]**.** As a result of these developments and the consequent miniaturization, the feature size of the transistor has reduced from about 10000nm in 1972 to as small as 7nm in 2018. However, from this extremely scaled dimensions, further reduction in device feature size, known as system scaling, enabled by Moore’s scaling is being increasingly challenged due to the quantum effects that fundamentally affect the device characteristics as well as the inadequacy of resources such as power and interconnect bandwidth under the requirements of perpetual interaction between big data and instant data [6]. Instant data generation requires devices with extremely low power consumption coupled with an “always-on” feature. Big data requires very high computing power, large data transfer bandwidth and high memory resources to generate the service and information that clients need. The goal of the semiconductor industry is to meet these emerging market requirements by being able to continue the progress in overall performance at reduced power and cost.

Till 2015, the International Technology Roadmap for Semiconductors (ITRS) [7] was produced annually from 1998 by a team of semiconductor industry experts from major semiconductor manufacturing nations. Its primary objective was to provide guidance to researchers in initiating innovations in various areas of technology. In December 2012 the International Roadmap Committee decided to reorganize itself in order to address the renewed requirements and challenges of the evolving microelectronics industry and identified several key areas to be addressed in future devices, including system integration, system connectivity, heterogeneous integration, diversification, miniaturization and search for alternate devices that are not based on variation of MOS technology such as spintronic or magnetic devices. As the goals of ITRS were reaching beyond the limits of CMOS based technologies, in 2016 the International Roadmap for Devices and Systems (IRDS) [8] was initiated as the successor of ITRS. It consists of a set of predictions about likely developments in electronic devices and systems in future considering the key elements of present day digital logic technology platform such as speed, power, density, and cost. The More Moore roadmap of IRDS 2017 [6] provides physical, electrical, and reliability requirements for logic and memory technologies to sustain power, performance, area and cost scaling for big data, mobility, and cloud based applications. The More Moore roadmap 2017 expects 15% increase in operating frequency at scaled supply voltage, 30% less energy consumption per switching at a given performance, 30% reduction in chip area footprint and 15% reduction in die cost while expecting an increase of 30% in wafer cost every two years for the next fifteen years.

Future applications anticipate the simultaneous achievement of all the forecasted targets in frequency improvement, supply voltage scaling, power consumption reduction, chip area reduction and cost cutting. However, many of these requirements are mutually competing. For instance, the expected size reduction requires more sophisticated process technologies demanding high cost fabrication units. . Thus, the More Moore roadmap of IRDS 2017 can be possible only by adopting new channel materials, introduction of novel device architecture, advancements in contact engineering and device isolation. For instance, present day Si and III-V compound semiconductor channels of transistors which suffer from threshold voltage roll off, band-to-band tunnelling, leakage currents and large investments in establishing fabrication infrastructure are expected to be replaced with high mobility channels like SiGe and Ge. Difficulty to control interconnect resistance, electromigration (EM), and time-dependent dielectric-breakdown (TDDB) limits are to be expected at least after 2022. The interconnect material used in present day device is Copper. As interconnect resistance Cu in the ultrascaled device has already entered an exponential increase regime, new barrier materials are required. TDDB is putting a limit on the dielectric thickness, forcing a slow-down in the permittivity (κ-value) scaling. However, even if these advancements are realized, ground rule scaling of the past is expected to slow down and saturate around 2028. Beyond 2028, transition to 3D integration and use of beyond CMOS devices for complementary System-on-Chip (SoC) functions are required.

Scaling the memory unit of a computational system also faces these challenges, as the memory chips are usually built from MOSFETs and capacitors. This article reviews the requirements, challenges and the scaling trends of this key element of the digital logic technology, memory scaling.

II. Various Memory Units

Memory units of a computational system are either used for permanent storage of data or for temporary storage of instructions before and after execution by the CPU. The desirable characteristics of an ideal memory unit of a computational system are fast READ/WRITE (R/W) speeds, low energy consumption during R/W operation, very long data retention ability, high R/W cycling endurance, low operating voltage and excellent dimension scalability [9]. Unfortunately, many of these characteristics are mutually exclusive. As a result, any single practical memory device does not possess all these characteristics. Hence the memory subsystem of the computational system consists of many components. One component is the cache memory for which data transfer from the CPU to other units are extremely fast due to their small R/W latency. Static random-access memory (SRAM) is commonly used as the cache memory of computers. Another component is called the main memory which is used to store programs and data during the run time operations of the CPU. The main memory should have low energy consumption and very high read speed. Dynamic random-access memory (DRAM) is commonly used as the main memory. The third memory component is for the permanent storage of data and hence must have very high memory capacity. Flash memory is the popular storage of the present computational system. SRAM and DRAM are volatile (VM) while flash is non-volatile (NVM). The difference in the memory technologies arises due to the difference in charge storage mechanism; SRAM stores the charges at the storage nodes of the transistor latch, DRAM stores the charges at the cell capacitor, and flash stores the charges at the floating gate of the transistor.

III. Volatile Memory Scaling

3.1. SRAM Scaling

SRAM is used as cache memory due to high R/W speeds. SRAM is not a non-volatile memory though it can retain the information as long as the power is on. The chip density of SRAM less as it uses latches (usually 4T or 6T cells) to store data [10] [11]. Density of SRAM cells increased in different technology nodes in the past owing mainly to the reduction in transistor size and device optimization [12]. Degradation of cell metrics such as threshold voltage variations as a consequence of channel length reduction and the new process variability issues such as fin thickness variation involved with the emerging device geometries like FINFETs are expected to pose challenges in the scaling of SRAM designs in future [13] [14] [15]. The requirement to maintain adequate noise margins and to control soft-error rates in the presence of random threshold voltage fluctuations has slowed down SRAM scaling after 2017. From 2012 to 2017 the bit cell area was reduced from about 0.09 μm2 to about 0.02 μm2, whereas from 2017 the cell size has not been scaled down appreciably. This inability of the actual SRAM scaling (32x improvement in chip density from 90nm to 10nm node) to keep up with the ideal scaling (81x improvement in chip density from 90 to 10nm node) has resulted in significant cumulative loss in effective technology scaling benefits.

3.2. DRAM Scaling

A DRAM chip density is higher than SRAM, because it consists of 1T1C cells. Its READ speed is high and has lower operating voltages in comparison with SRAM, but its WRITE speed is less. As the data retention time is only about 100ms in DRAM, its data refresh time is very small, leading to significant power consumption.

Historically, unlike SRAM scaling, DRAM memory technology has followed more or less the ideal scaling curve achieving anticipated cost per bit reduction and high-speed accompanied with low-power requirements [16]. DRAM process sizes shrank significantly in the past until 2016. For instance from 2008 DRAM cell node levels were shrinking continuously from 40nm-class (4x node level) to 10nm class (1x node level) till 2016 when DRAM was being introduced with many new technologies such as 193 nm argon fluoride (ArF) immersion, high-NA lithography with double patterning technology, improved cell FET technology including fin type transistor [17] [18] [19] and buried word line/cell FET technology [20]. However, future DRAM scaling below 1znm technology poses many challenges [21] such as cost-per-bit reduction, curbing of cell disturb margin etc. As a result, after 2016 the scaling rate has reduced and DRAM cells are still shipped at the 1xnm node level. From 2016, DRAM products below 20 nm regime are labeled as 1xnm node (17nm to 19nm), 1ynm (14nm to 16nm), and 1znm (11nm to 13nm). In 2016 Samsung announced the mass production of the industry's first 10nm class, 8Gb, DDR4, 18nm DRAM chips after overcoming technical challenges in DRAM scaling using ArF (argon fluoride) immersion lithography. It was 30% faster and consumed less power than the 2xnm device. SK Hynix has announced the 16Gb, DDR5 DRAM using 1ynm process technology that offers ultra-high speed and high density with reduced power consumption (30% less) and increased data transfer rate (60% higher) as compared to DDR4 to be released after 2018.

Thus, it is clear that DRAM scaling after 2016 in the 1xnm node regime progresses in incremental steps, sometimes nanometre by nanometre. One of the challenges in DRAM scaling is the worsening Row hammer mechanism [22]. Row hammer is a wordline disturbance issue due to the accumulation of electrons at the interface of wordline that occurs when one wordline of DRAM chip is continuously addressed leading to increased cell charge gain and error rates. DRAM cell capacitance decreases with scaling [6], leading to the degradation of the refresh time performance as the refresh time is proportionate to the cell capacitance. Therefore,

efforts to improve the cell capacitance in the 1T1C cell to maintain refresh time at the desired level is a key concern in future DRAM scaling below 1xnm technology [16].

1T1C DRAM structure may last for another few years, but the scaling may stop below 10nm. So the industry is looking at geometries such as the vertical gate, capacitorless 1T DRAM cell and several other next-generation memory types such as phase-change memory (PCM), ReRAM and STT-MRAM that could replace the conventional 1T1C DRAM structure.

IV. Non-volatile Memory Scaling

Non-volatile memory (NVM) may be divided into two categories—Charge based flash memories that store the charges at the floating gate of the transistor, and non-charge-based-storage memories (ferroelectric or FeRAM, magnetic or MRAM, phase-change or PCRAM, and resistive or ReRAM).

4.1. Flash Memory Scaling

Flash memory, based on EEPROM technology has been the most popular choice for non-volatile applications because of its highest chip density among all memory types due to its 1T cell structure [23] and its compatibility with the current CMOS process. A flash memory cell is simply a MOSFET cell, except that it has stacked gate structure where a poly-silicon floating gate [24]or Silicon Nitride charge trap layer [25] is sandwiched between a tunnel oxide and an inter-poly oxide to form a charge storage layer. Flash memory scaling in the past has lagged behind the CMOS logic device scaling, since the effective oxide thickness (EOT) of the gate stack has to be very large to control the leakage current during read and write.

Depending on the way the memory cells are organized flash memory is classified in to NOR flash and NAND flash. In NOR Flash, one end of each memory cell is connected to the source line and the other end directly to a bit line resembling a NOR Gate. In NAND Flash, a number of memory cells are connected in series similar to a NAND gate. As a result of this cell organization, NAND flash may be erased, written, and read in blocks, while in NOR flash a single machine word could be written, erased or read independently.

Though the scaling of flash cell in terms of its feature size is difficult due to the difficulty in scaling EOT of the gate stack, there are other possibilities to increase the flash cell density. For instance, a single flash memory cell can be designed to store multiple bits of data rather than the conventional HIGH and LOW levels, a property known as the multi-bit per cell storage [26]. Multi-bit storage technology increases memory density and thus has become a significant step in the scaling of flash. Multi-bit storage is a possibility in flash because many distinct threshold voltage (VT) states can be achieved in a transistor by controlling the amount of charge stored in its floating gate. Each flash memory cell with transistor having four different threshold voltage VT states cells can store two bits of data. Such flash memory chips having two-bits/cell have already been commercialized. The feasibility of four-bits/cell flash memory device which requires transistors with eight different VT states is being explored by the industry vigorously [27].

Another possibility to achieve even higher density and lower cost per chip is to adopt multi-layer integration [28] [29] [30]. Using this multi-level cell (MLC) technology SSDs up to 100 TB capacity with up to 96 layers are being shipped presently by flash chip manufacturers [6]. However, stacking too many layers may produce high stress that bends the wafer, limiting the applicability of MLC technology in future, and it is expected that flash chips with a maximum of 192 to 256 layers are possible [6].

4.2. Ferroelectric Memory Scaling

Several emerging non-flash, non-volatile memories that are not based on charge storage such as FeRAM, MRAM, PCRAM, and ReRAM are being explored vigorously to overcome the scaling limitations of the flash memory [6]. Here, we discuss the scaling trends of ferroelectric memory only since a detailed discussion on other emerging memory types can be found in the IRDS report 2017 [6], [8]. There are two types of ferroelectric memory. One is called Ferroelectric RAM (FRAM or FeRAM) whose memory cell generally consists of a ferroelectric capacitor and a selection transistor, similar to a DRAM cell. The other is called a ferroelectric FET (FeFET) with a MOSFET like structure having a gate stack layer consisting of a ferroelectric layer and an electrode. Both store information as a polarization state of the ferroelectric material.

FeRAM has 1T1C cells [31] similar in construction to DRAM. Non-volatility in the 1T1C cell is achieved by using a ferroelectric layer with large residual polarization, such as lead zirconate titanate PbZrxTi1-xO3 (PZT) or strontium bismuth tantalate Sr1-yBi2+xTa2O9 (SBT), instead of the dielectric layer in the capacitor of the DRAM cell. The atoms in the ferroelectric layer change polarity in an electric field. After the electric field is removed, the atoms remain in the polarized state due to dielectric hysteresis, which makes the materials non-volatile and the state of the memory is preserved. FeRAM is considered as a potential alternative to the flash memory that suffers from many scaling issues. FeRAM has lower power consumption, faster write cycle, larger R/W endurance (as high as 1014 cycles) and longer data retention times of more than 10 years at +85 °C in comparison with flash memory chips. However, FeRAM's read process is destructive necessitating a write-after-read architecture.

In 1998 Evans et.al. and Eaton et.al. independently reported practical Si-based FeRAMs [32] [33]. Significant improvement in read out method using non-destructive readout through pulses of UV radiation was achieved in 1991 at NASA's Jet Propulsion Laboratory (JPL). Fujitsu started FeRAM mass production in 1999 with La-doped PZT (PLZT) as the dielectric layer. In 1996, Samsung introduced a 4 Mb FeRAM fabricated using NMOS logic followed by SK Hynix in 1998. Texas Instruments was able to embed FeRAM cells using two additional masking steps during conventional CMOS semiconductor manufacture enabling the integration of FeRAM onto microcontrollers at reduced cost. The first commercial product using FeRAM was Sony's Play Station2 Memory Card released in 2000, the microcontroller of which contained embedded FeRAM fabricated using a 500nm CMOS process. In 2001, Texas Instruments (TI) and Ramtron developed FeRAM test chips with an improved 130nm process. In 2005, Fujitsu and Seiko-Epson jointly developed the 180nm FeRAM process. In 2015, Fujitsu announced new ferroelectric capacitor fabrication technology to realize the mass production of 1T1C 4Mb FeRAM with LCSPZT (La, Sr and Ca doped PZT) as the ferroelectric layer [34]. Several other companies, including Ramtron and Texas Instruments, are producing FeRAM at a large scale and also investing for the improvement of FeRAM scalability.

FeFET, a MOSFET with HfO2-based ferroelectric materials as the gate insulator layer, is an emerging technology having high scaling potentials because of their compatibility to the CMOS process, high switching speeds, low power consumption, and nondestructive readout characteristics [35] [36] [37] [38]. Based on the gate stack composition, FeFET structures can be classified in to (1) MFS (metal ferroelectric semiconductor) in which ferroelectric layer F is directly deposited on top of the semiconductor S [39] [40] [41] (2) MFIS (metal ferroelectric insulator semiconductor) in which a thin buffer layer (I) is introduced between the F and S and (3) MFMIS (metal ferroelectric metal insulator semiconductor) in which floating conductive layer M is introduced between F and L.

In FeFETs with MFS stack, oxide based channel materials are usually adopted instead of Si to alleviate the possibility of the inter-diffusion of elements between the F and the substrate and their chemical reactions. Many Ferroelectric/Channel combinations such as PZT /ZnO [42], PZT / indium tin oxide (ITO) [43], PZT /Indium gallium zinc oxide (IGZO) [44], HZO / IGZO, HZO/ tungsten oxide (WO) have been reported in the past few years with minimum inter-diffusion between the ferroelectric layer and the channel material. The MFIS gate stack structure solves the issue of inter-diffusion by introducing a thin insulator layer between the ferroelectric and the channel. High performance MFIS-FeFETs [35] have been reported recently. In MFMIS gate stack structure, the ferroelectric layer is sandwiched between two conductive layers for improving data retention potential of MFIS. HfO2-based MFMIS FeFETs such as Pt/ Al:HfO2 /TiN/SiO2/Si, with promising scalability have been reported recently.

V. Conclusions

Size scaling and device optimization have resulted in substantial increase in SRAM cell density for different technology nodes in the past. However, continued channel length reduction and the associated process variability issues are expected to pose challenges in the scaling of SRAM designs in future. Introduction of high-density 6T SRAM cells and High Mobility Channel is expected to improve logic density, area savings and drive current gain. DRAM cell node levels were shrinking continuously from 40nm-class to 10nm class till 2016 due to the introduction of many new technologies like 193 nm argon fluoride (ArF) immersion, high-NA lithography with double patterning technology and improved cell FET technology. DRAM scaling below 1znm technology poses many challenges such as cost-per-bit reduction, curbing of cell disturb margin etc. As a result, after 2016 the scaling rate has decreased. DRAM cells with vertical gate, capacitorless DRAM cells and emerging memory types such as phase-change memory (PCM), ReRAM and STT-MRAM are being explored to continue the scaling of DRAM. Multi-bit storage technology and multi-layer integration are crucial in the scaling of flash memory due to the difficulty in scaling EOT of the gate stack. However, applicability of MLC technology may be limited due to the difficulty in stacking too many layers. Hence, several emerging non-charge based storage memories such as FeRAM, MRAM, PCRAM, and ReRAM are being explored vigorously to overcome the scaling limitations of the flash memory. Non-volatility in FeRAM is achieved by using a ferroelectric layer with large residual polarization in the capacitor of the 1T1C cell. FeFET, with HfO2-based ferroelectric materials as the gate insulator layer, is another emerging technology with high scaling potentials to replace flash. Many FeFETs with MFS, MFIS and MFMIS gate stack structures with promising scalability have been reported recently, which are expected to solve the scalability issue of flash memory.

References

[1] Gordon E Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, no. 8, pp. 33-35, April 1965.

[2] H Yaegashi, "Pattern fidelity control in multi-patterning towards 7nm node," in *2016 IEEE 16th International Conference on Nanotechnology (IEEE-NANO)*, 2016, pp. 452-455.

[3] Ma Yuangsheng et al., "Self-aligned double patterning (SADP) compliant design flow," in *Proc. SPIE 8327, Design for Manufacturability through Design-Process Integration VI*, vol. 8327, 2012.

[4] Jing Jiang, Souvik Chakrabarty, Mufei Yu, and Obe Christopher K, "Metal oxide nanoparticle photoresists for EUV patterning," *Journal of Photopolymer Science and Technology*, vol. 27, no. 5, pp. 663--666, 2014.

[5] BJ Rice, "Extreme ultraviolet (EUV) lithography," in *Nanolithography*.: Elsevier, 2014, pp. 42--79.

[6] https://irds.ieee.org/editions/2017/more-moore.

[7] http://www.itrs2.net/itrs-reports.html.

[8] https://irds.ieee.org/editions/2017.

[9] Chen An, "A review of emerging non-volatile memory (NVM) technologies and applications," *Solid-State Electronics*, vol. 125, pp. 25--38, 2016.

[10] Liao Chu-Feng et al., "Zero static-power 4T SRAM with self-inhibit resistive switching load by pure CMOS logic process," in *2016 IEEE International Electron Devices Meeting (IEDM)*, 2016, pp. 428-431.

[11] C Lage, J.D.Hayden, and C Subramanian, "Advanced SRAM technology-the race between 4T and 6T cells," in *International Electron Devices Meeting. Technical Digest*, 1996, pp. 271-274.

[12] Yamauch Hiroyuki, "Embedded SRAM trend in nano-scale CMOS," in *2007 IEEE International Workshop on Memory Technology, Design and Testing*, 2007, pp. 19-22.

[13] Kawasaki H et al., "Challenges and solutions of FinFET integration in an SRAM cell and a logic circuit for 22 nm node and beyond," in *2009 IEEE International Electron Devices Meeting (IEDM)*, 2009, pp. 1-4.

[14] Chuang Ching-Te, Saibal Mukhopadhyay, Kim Jae-Joon, Kim Keunwoo, and Rahul Rao, "High-performance SRAM in nanoscale CMOS: Design challenges and techniques," in *2007 IEEE International Workshop on Memory Technology, Design and Testing*, 2007, pp. 4-12.

[15] Qazi Masood, Sinangil Mahmut, and Chandrakasan Anantha, "Challenges and Directions for Low-Voltage SRAM," *IEEE Design Test of Computers*, vol. 28, pp. 32-43, 2011.

[16] Mutlu Onur, "Memory scaling: A systems architecture perspective," in *2013 5th IEEE International Memory Workshop*, 2013, pp. 21--25.

[17] Kim Ji-Young et al., "The breakthrough in data retention time of DRAM using Recess-Channel-Array Transistor (RCAT) for 88 nm feature size and beyond," in *2003 Symposium on VLSI Technology. Digest of Technical Papers (IEEE Cat. No. 03CH37407)*, 2003, pp. 11-12.

[18] Kim JV et al., "S-RCAT (sphere-shaped-recess-channel-array transistor) technology for 70nm DRAM feature size and beyond," in *Digest of Technical Papers. 2005 Symposium on VLSI Technology, 2005*, 2005, pp. 34-35.

[19] Park Sung-Wook et al., "Highly scalable saddle-Fin (S-Fin) transistor for sub-50nm DRAM technology," in *2006 Symposium on VLSI Technology, 2006. Digest of Technical Papers*, 2006, pp. 32-33.

[20] Schloesser T et al., "6F 2 buried wordline DRAM cell for 40nm and beyond," in *2008 IEEE International Electron Devices Meeting*, 2008, pp. 1-4.

[21] Mandelman Jack A et al., "Challenges and future directions for the scaling of dynamic random-access memory (DRAM)," *IBM Journal of Research and Development*, vol. 46, no. 2.3, pp. 187--212, 2002.

[22] Kim Yoongu et al., "Flipping bits in memory without accessing them: An experimental study of DRAM disturbance errors," *ACM SIGARCH Computer Architecture News*, vol. 42, no. 3, pp. 361--372, 2014.

[23] Kim Dong-Won, Prins Freek E, d Kim Taehoon, Kwong Dim-Lee, and Banerjee Sanjay, "Charge retention characteristics of SiGe quantum dot flash memories," in *60th DRC. Conference Digest Device Research Conference*, 2002, pp. 151--152.

[24] Blomme Pieter, Govoreanu Bogdan, Van Houdt Jan, and De Meyer Christina, "A novel low voltage memory device with an Engineered SiO2/High-k tunneling barrier," in *Non-Volatile Semiconductor Memory Workshop-NVSMW, Location: Leuven Belgium*, 2003, pp. 93--94.

[25] Konstantin K Likharev, "Riding the crest of a new wave in memory [NOVORAM]," *IEEE Circuits and Devices Magazine*, vol. 16, no. 4, pp. 16--21, 2000.

[26] Sugizaki T. et al., "Novel multi-bit SONOS type flash memory using a high-k charge trapping layer," in *2003 Symposium on VLSI Technology. Digest of Technical Papers (IEEE Cat. No.03CH37407)*, 2003, pp. 27-28.

[27] Y. Polansky et al., "A 4b/cell NROM 1Gb Data-Storage Memory," in *2006 IEEE International Solid State Circuits Conference - Digest of Technical Papers*, 2006, pp. 448-458.

[28] Park Ki-Tae et al., "A 45nm 4Gb 3-dimensional double-stacked multi-level NAND flash memory with shared bitline structure," in *2008 IEEE International Solid-State Circuits Conference-Digest of Technical Papers*, 2008, pp. 9--10.

[29] Maeda Takashi et al., "Multi-stacked 1G cell/layer pipe-shaped BiCS flash memory," in *2009 Symposium on VLSI Circuits*, 2009, pp. 22--23.

[30] Park Ki-Tae et al., "Three-Dimensional 128 Gb MLC Vertical nand Flash Memory With 24-WL Stacked Layers and 50 MB/s High-Speed Programming," *Solid-State Circuits, IEEE Journal of*, vol. 50, pp. 204-213, 2015.

[31] Ishiwara Hiroshi, "Ferroelectric random access memories," *Journal of nanoscience and nanotechnology*, vol. 12, no. 10, pp. 7619--7627.

[32] Evans Joseph T and Womack Richard, "An experimental 512-bit nonvolatile memory with ferroelectric storage cell," *IEEE journal of solid-state circuits*, vol. 23, no. 5, pp. 1171--1175, 1988.

[33] Eaton S Sheffield, "A ferroelectric nonvolatile memory," in *Dig. Tech. Pap. IEEE Int. Solid-State Circuit Conf., 1988*, vol. 31, 1988, pp. 130--131.

[34] Eshita Takashi et al., "Development of ferroelectric RAM (FRAM) for mass production," in *2014 Joint IEEE International Symposium on the Applications of Ferroelectric, International Workshop on Acoustic Transduction Materials and Devices \& Workshop on Piezoresponse Force Microscopy*, 2014, pp. 1-3.

[35] Trentzsch M et al., "A 28nm HKMG super low power embedded NVM technology based on ferroelectric FETs," in *IEEE International Electron Devices Meeting (IEDM)*, 2017, pp. 11-16.

[36] George Sumitha et al., "Nonvolatile memory design based on ferroelectric FETs," in *Proceedings of the 53rd Annual Design Automation Conference*, 2016, pp. 1-6.

[37] Muller J et al., "Ferroelectric Zr0.5Hf0.5O2 thin films for nonvolatile memory applications," *Applied Physics Letters*, vol. 9, no. 11, p. 112901, 2011.

[38] J. Muller, D. Brauhaus, U. Schroder, U. Bottger T. S. Boscke, "Ferroelectricity in hafnium oxide thin films," *Applied Physics Letters*, vol. 99, no. 10, p. 102903, 2011.

[39] Shu-Yau Wu, "A new ferroelectric memory device, metal-ferroelectric-semiconductor transistor," *IEEE Transactions on Electron Devices*, vol. 21, no. 8, pp. 499--504, 1974.

[40] Sugibuchi K, Kurogi Y, and Endo N, "Ferroelectric field-effect memory device using Bi4Ti3O12 film," *Journal of Applied Physics*, vol. 46, no. 7, pp. 2877--2881, 1975.

[41] Rost Timothy A, Lin He, and Rabson Thomas A, "Ferroelectric switching of a field-effect transistor with a lithium niobate gate insulator," *Applied physics letters*, vol. 59, no. 27, pp. 3654--3656, 1991.

[42] Zhang X et al., "Zno field-effect transistors with lead-zirconate-titanate ferroelectric gate," *Materials Research Innovations*, vol. 19, no. 2, pp. S2-181-S2-184, 2015.

[43] Minh DH, Loi NV, Duc NH, and Trinh BNQ, "Low-temperature PZT thin-film ferroelectric memories fabricated on SiO2/Si and glass substrates," *Journal of Science: Advanced Materials and Devices*, vol. 1, no. 1, pp. 75--79, 2016.

[44] Nomura Kenji et al., "Room-temperature fabrication of transparent flexible thin-film transistors using amorphous oxide semiconductors," *Nature*, vol. 432, no. 7016, pp. 488--492, 2004.