ICMTD

2007

2nd International Conference on Memory Technology and Design

May 7-10 2007
Giens, France

Proceedings

Technical Co-Sponsor:

The Institute of Electrical and Electronics Engineers, Inc

IEEE Electron Device Society

2nd International Conference on Memory Technology and Design

May 7 – 10 2007

Giens, France

Proceedings

Organized by:

IMEC (Belgium), Catholic University of Leuven (Belgium) and L2MP (France)

Technical Co-Sponsor:

The Institute of Electrical and Electronics Engineers (IEEE), IEEE Electron Device Society

Sponsored by:

ARCSIS, ATMEL, LEXAR, NXP Semiconductors, STMicroelectronics, Web-Feet Research

The papers in this book reflect the authors' opinions and are published as submitted and without change.

Additional copies may be ordered from:

L2MP-Polytech
IMT Techopôle de Château Gombert
Cedex 20
13451 MARSEILLE

Email: ICMTD-contact@ICMTD.com

Website : www.ICMTD.com

Printed in France by the University Paul CEZANNE (Aix-Marseille III)

ICMTD-2007

WELCOME TO ICMTD-2007

It is our pleasure to welcome you at the 2nd International Conference on Memory Technology and Design **ICMTD-2007**, which is held from May 7th-10th 2007, again on the Peninsula of Giens at the Mediterranean coast of France.

This conference -this year being organized by the Interuniversity Microelectronics Center (IMEC), Leuven, Belgium, the Provence Materials and Microelectronics Laboratory (L2MP), Marseille, France, and the Catholic University of Leuven, Belgium- was originally created to provide an international forum for presentations and discussions on recent developments in Memory Technology and Design. All aspects of memory devices, circuits, process technologies, materials and other related research are within the scope of the conference. These three days of presentations, in oral presentations and panel discussions, provide extensive opportunities for technical information exchange. Furthermore we expect that the conference settings and social events (incl. an excursion to the Porquerolles Island) will further stimulate informal communication among participants.

49 papers have been selected for oral presentations and have been organized in 8 sessions, 6 of which being organized in parallel. Additionally, these sessions include 11 invited papers from experts in the field:

- **Memory market update: shifting dynamics**
  C. Hirst *(Gartner Dataquest)*
- **Living with the DRAMification of NAND - How to survive the Flash Price Wars**
  A. Niebel *(Webfeet Research)*
- **Secure memories: dream or reality?**
  I. Verbauwhede *(KUL)*
- **From memory component to memory systems**
  D. Keitel-Schulz *(Qimonda)*
- **A designer's perspective on future memory architectures for software defined radios**
  P. Marchal *(IMEC)*
- **Smart cards: technologies and products**
  R. Zambrano *(ST Incard)*
- **Phase-Change Memory – Present and Future**
  H.-L. Lung *(Macronix)*
- **Current limitations of floating gate NVM and new alternatives**
  A. Bergemont *(Maxim Integrated Products)*
- **Magnetic RAM for embedded memory in SoC**
  S. Ueno *(Renesas)*
- **Copper oxide resistive switching for non-volatile memory applications**
  T.-N. Fang *(Spansion)*
- **Real-time soft-error rate testing of semiconductor memories on the european test platform ASTEP**
  J.-L. Autran *(L2MP)*

One session deals with the exciting topic of FinFet-based Flash technology and is organized as an open workshop of the IST-FinFlash project nr. FP6-016917.

Besides these paper sessions, we also invite you to participate in 3 panel discussions

- **Charge-based versus resistance-based non-volatile memory**
  Moderator: D. Wouters *(IMEC)*
  Panelists: J. Park *(Samsung)*, H.-L. Lung *(Macronix*), E. Prinz *(Freescale)*, R. Waser *(RWTH Aachen)*, S. Ueno *(Renesas)*
- **SOI and new memory opportunities**
  Moderator: C. Hirst *(Gartner Dataquest)*
  Panelists: D. Somasekhar *(Intel)*, S. Natarajan *(Emerging Memory Technologies)*, K. Itoh *(Hitachi)*, P. Fazan *(Innovative Silicon),* M. Shaheen *(SOITEC)*
- **Memory design in 45nm and beyond: how to survive the technology scaling**
  Moderator: W. Dehaene *(KUL)*
  Panelists: P. Marchal *(IMEC)*, D. Keitel-Schulz *(Qimonda)*, I. Verbauwhede *(KUL)*, D. Heslinga *(NXP Semiconductors)*

The conference also provides a limited number of posters from various memory-related European projects which can be discussed with the authors during coffee and lunch breaks.

We are confident that this conference will be an exciting event for each attendee and we would like to thank all participants for their valuable contributions to the conference.

Furthermore, we would like to express our gratitude to the Scientific/Technical committee for their review of contributed papers and to our sponsors for their financial and technical support.

Finally we would like to thank all who contributed to the organisation and implementation of the **ICMTD-2007** conference.

We wish you a pleasant and fruitful conference,

Jan Van Houdt, 2007 General Chairman

On behalf of the organization and steering committees

ICMTD 2007 Organization:

Wim DEHAENE ..... Catholic University of Leuven, Belgium (Leuven)

Pascal MASSON..... L2MP, France (Marseille)

Jan VAN HOUDT ..... IMEC, Belgium (Leuven)

Dirk WOUTERS ..... IMEC, Belgium (Leuven)

ICMTD 2007 Communication:

Anne DE SMET..... Momentum, Belgium (Leuven)

Liesbet MASSANT..... IMEC, Belgium (Leuven)

SCIENTIFIC - TECHNICAL COMMITTEE

Atila ALVANDPOUR .....Linköping University, Sweden (Linköping)
Karen ATTENBOROUGH.....NXP Semiconductors, Belgium (Leuven)
Lofti BENAMMAR .....Atmel, Rousset (France)
Jean-Michel DAGA .....Atmel, France (Rousset)
Barbara DE SALVO .....CEA/LETI, France (Grenoble)
Wim DEHAENE.....Catholic University of Leuven, Belgium (Leuven)
Pierre FAZAN.....Innovative Silicon Inc., Switzerland (Lauzane)
Albert FAZIO .....Intel, USA (Santa Clara)
William J. GALLAGHER .....IBM, USA (Yorktown Heights)
Eric GERRITSEN .....NXP Semiconductors, France (Crolles)
Yasuo INOUE .....Renesas, JAPAN (Tokyo)
Rajiv JOSHI.....IBM, USA
Doris KEITEL-SCHULZ.....Qimonda, Germany(Dresden)
Zoran KRIVOKAPIC.....AMD, USA
Pol MARCHAL.....IMEC, Belgium (Leuven)
Pascal MASSON .....L2MP, France (Marseille)
Pascale MAZOYER.....STMicroelectronics, France (Crolles)
Christophe MULLER.....L2MP, France (Toulon)
Jaekwan PARK.....Samsung, Korea
Agostino PIROVANO .....STMicroelectronics, Italy(Agrate)
Erwin J. PRINZ.....Freescale Technology, USA (Austin)
Yakov ROIZIN .....Tower, Israel (Migdal Haemek)
Kaushik ROY.....University of Purdue, USA (Purdue)
George SAMACHISA .....SanDisk, USA (Santa Clara)
Michiel VAN DUUREN.....NXP Semiconductors, Belgium
Jan VAN HOUDT.....IMEC, Belgium (Leuven)
Rainer WASER.....RWTH Aachen, Germany (Aachen)
Josef WILLER.....Qimonda, Germany (Aachen)
Dirk WOUTERS.....IMEC, Belgium (Leuven)

ICMTD STEERING - COMMITTEE

Atila ALVANDPOUR .....Linköping University, Sweden (Linköping)
Jean-Michel DAGA .....Atmel, France (Rousset)
Eric GERRITSEN .....NXP Semiconductors, France (Crolles)
Pascal MASSON .....L2MP, France (Marseille)
Pascale MAZOYER.....STMicroelectronics, France (Crolles)
Jaekwan PARK.....Samsung, Korea
Erwin J. PRINZ .....Freescale Technology, USA (Austin)
Jan VAN HOUDT.....IMEC, Belgium (Leuven)

TABLE OF CONTENT

SESSION A : *Invited Talks*

SESSION B : *Phase Change Memory*

SESSION C : *FinFlash*

SESSION D : *Floating Gate*

SESSION E : *RRAM & DRAM*

SESSION F : *SRAM & Process Variability*

SESSION G : *Charge Trapping*

SESSION H : *High-κFlash & Nano-crystals*

Memory Market Update: Shifting Dynamics

Clare Hirst, Gartner Dataquest

It is clear that the immediate future of the memory market is centered on the fortunes of DRAM and flash memory technology.

It is not so clear how changing supply and demand dynamics will impact the market and how well the competing vendors are positioned for success.

And with new, emerging memory technologies constantly under development, there is always the possibility of even greater challenges ahead.

Charge-based versus Resistance-based Non-Volatile Memory

Flash, which is the standard Non-volatile Memory technology today, is based on the storage of charge in the floating gate of an MOS gate -stack, and by that controlling the transistor threshold voltage. However, further reduction of the charge (number of electrons) as needed for the further technology scaling may be limited by leakage current, while cell geometry reduction may induce important electrostatic effects. To cope with arising scaling issues, new alternative Flash concepts are proposed based on different charge storage media, as nanocrystals or silicon nitride.

On the other hand, different new emerging memory concepts are proposed that are based on the switching of the resistance state (high or low resistive) in a material. Examples are magnetic RAM, Phase-Change RAM, Conductive Bridging RAM, and Oxide Resistive switching RAM. Resistance-based memories are thought of as being much more scalable, since, at least in principle, resistance based concepts do not show these limits for the further scaling of the memory element (from the viewpoint of information storage and layout). However, other fundamental scaling bottlenecks may arise, as e.g. the finite size of current filaments.

The topic of this panel session is to discuss the prospects of scaling for both types of memories. Also, besides geometrical cell size scaling, the possibility of multi-level (or multi-bit) operation will be addressed for the different concepts.

To cover those questions and this topic, this panel will be moderated by **Dirk WOUTERS (IMEC)**. The following panelists will share their view and answer the questions of the audience:

H.L.	LUNG	Macronix
J.	PARK	Samsung
E.	PRINZ	Freescale
S.	UENO	Renesas
R.	WASER	RWTH Aachen

SOI for new memory opportunities

Food for thought!

Now that embedded memory in many Systems-on-Chip takes up far more than 50% of the die area it would be appropriate to refer to these chips as memory devices with embedded logic.

This consideration serves to indicate how new opportunities to bridge the logic/memory gap could bring breakthrough changes in future system performance. This need for changes is further enhanced by the move towards multi-core processors, requiring even more memory integration.

Silicon-on-Insulator (SOI) is one of the enablers to bring about these changes.

An example of such a SOI based memory opportunity is the capacitorless DRAM, highlighted during ICMTD2005. Using the floating body effect, a bit is stored with just one transistor and without additional capacitor, as in a standard DRAM.

Another recent example, presented at ISSCC2006, combines a deep trench DRAM capacitor on SOI, suppressing this floating body effect, thereby increasing density and speed, at the expense of retention time thus power consumption.

Next to increased area density and reduced junction leakage, other specific benefits often attributed to Memory-on-Insulator are the improved immunity to soft error rate and the reduced process variability due to reduced substrate coupling.

These and other benefits should be discussed in the context of factors like cost, wafer supply and design investment, also including the choice between embedded (in System-on-Chip) versus stand-alone (on System-in-Package) memory options, and considering the evolution of these factors in the near future.

On a different note it can be mentioned that future MEMS solutions for data storage may extensively use SOI substrates for proper patterning of the mechanical read/write elements.

The panel is intended to give further “food for thought” and to evaluate the “appetite for SOI” of future memory technologies.

Moderator: C. HIRST.....Gartner Dataquest, UK

Panelists:
P. FAZAN .....Innovative Silicon, Switzerland
K. ITOH.....HITACHI, Japan
S. NATARAJAN.....Emerging Memory Technologies, Canada
M. SHAHEEN.....SOITEC, France
D. SOMASEKHAR.....Intel, USA

Memory design in 45nm and beyond: how to survive the technology scaling?

This panel addresses the topic of continued scaling and its impact on memory design. Will it be possible to make a stable RAM cell in 45 nm and beyond? Does a commonly used metric like noise margin still make sense in that context? Do we have to move to different memory architectures and cells or is the classic architecture based on the 6T cell still fine? The opinions of the different experts in this domain will first be summarized after which a discussion with the audience will follow.

To cover those questions and this topic, this panel will be moderated by **Wim DEHAENE (KUL)**. The following panelists will share their view and answer the questions of the audience:

D.	Heslinga.....	NXP Semiconductors
D.	Keitel-Schulz.....	Qimonda
P.	Marchal.....	IMEC
I.	Verbauwhede.....	KUL

),

SESSION A

Invited talks

Living with the DRAMification of NAND – 'How to survive the Flash Price Wars'

Alan Niebel, Web-Feet Research

Ever since 1991 when Flash came into the memory market, there has always been the question of whether Flash will ever catch up to DRAM in volume and suffer the same volatile fate of the DRAM boom and bust cycles. Flash memory has a different market structure than DRAM, where NOR and NAND Flash are found in over 160 different applications, while DRAM is consumed in only a handful. Mainly computers and recently cell phones are the predominant applications for DRAM that use commodity DRAM produced by the big four vendors.

NAND Flash has 'come of age' in 2004 growing to $7.3 billion and $12 billion in 2005, thereby establishing itself as an essential component used in the growing mobile consumer market. MP3 Players, Flash cards in Digital Cameras, USB Flash Drives, and emerging storage in cell phones has solidified the market consumption for NAND. In 2006, NAND bit consumption exceeded DRAM production.

In 2006, Samsung began high volume production of MLC NAND joining Toshiba and SanDisk who have produced over five generations of MLC NAND. Hynix and IMFT (Intel Micron Flash Technologies) are now also producing MLC NAND. In late 2006, Samsung and Hynix realized that there was an oversupply of NAND due to hit the market in early 2007. Nor, did they see any new applications that would consume the additional capacity coming into the market in the first half of 2007. Consequently, Hynix and Samsung shifted any additional capacity slated for NAND, back to DRAM production, which was supposed to keep NAND at current 'oversupply' levels for most of 2007.

Pricing of NAND also took a major decline at the end of 2006 and into the first two months of 2007. Overall 2006 NAND price declines came in around 62%, which is far higher than the market usually sustains at 30-50% per year. With the excess supply in the NAND market and Samsung shipping very slow NAND components at discounted prices at the end of 2006, prices continued to decline aggressively (nearly 50% in some densities) in the first two months of 2007. By the end of Q1, NAND prices when extrapolated for the year were showing an overall decline of over 60+%, which makes the NAND market look similar to the commoditized DRAM market. The NAND prices declining at a high rate do open up new applications, but if this rate is above 50+% then the market will not survive.

Although the NAND market is suffering from pricing woes due to oversupply, this market is not a repeat of the old DRAM business model. Granted that the large memory manufacturers like Samsung and Hynix and a lesser extent Micron are controlling how much supply they bring out in either DRAM or NAND, the NAND storage market will quickly change its dynamic to prevent a DRAMification of NAND. As the market grows too big for a few manufacturers to control, the inherent aspects of the NAND technology and market dynamics will self-correct.

First, the various types of NAND architectures (1-bit/cell, 2-bit, 3- or 4-bit/cell) have different performance and cost parameters that apply to the different market requirements: media storage, cell phone storage, and computing storage. Second, new applications are entering the market that should consume large quantities of NAND like the iPhone, video iPod, gaming players, GPS, video/movie storage and SSD computing. These new applications always face the challenge of when will they be brought into the market and how fast they will be adopted by the various consumer usage models in different regions and age groups. Over time, NAND will quickly recover its balance and new demand applications will help it to continue its growth until the technology runs out of advancement, but then another technology will take its place in 2015 or earlier.

Headquartered in Monterey, California, Web-Feet Research provides business consulting and market research services in the memory and storage markets, with emphasis on Flash memory components, Flash cards, Embedded Flash Drives, SSDs and small form factor HDD for mobile applications.

Secure Memories: dream or reality?

Ingrid Verbauwhede
ESAT/COSIC, Katholieke Universiteit Leuven

Abstract

The security of an embedded system should not depend on the obscurity of the algorithm or the design. It should depend on the secrecy of the key. This makes the storage of keys and other sensitive data the Achilles' heel of the system. In this presentation we will give an overview of possible attacks on embedded systems. More specifically we will focus on side channel leakage attacks on the memory part of these systems. Countermeasures and novel research directions will be indicated.

1 Introduction

Well designed security systems are based on Kerckhoffs's assumption [8]. It states that the security of a system should depend on the secrecy of the key, not the obscurity of the algorithm or the design.

Designers of cryptographic algorithms start from the assumption that an attacker knows all details about the algorithm except for the value of the secret key. This means that most widely used algorithms for commercial applications are publicly known and have been scrutenized by cryptanalysts. An example is the recent AES algorithm. The selection of Rijndael as the next Advanced Encryption Standard came after a call for proposals from NIST and a multi-round public evaluation.

From an implementation viewpoint, it also means that the hardware or software design, the architecture and/or implementation are assumed to be known to the attacker. The advantage of this approach is that the sensitive part is well defined and confined. E.g. when the secret is accidently disclosed, one need only to replace the key. However, this makes the storage of keys and other sensitive parts the Achilles' heel of the system.

The remainder of this paper discusses research issues related to secure storage. It does not claim to solve all problems. Because total security simply does not exist and is not practicle. For a well designed security system it will cost more to break into the system than the possible benefits that can be obtained.

The rest of this overview is organized as follows. We will first discuss the security model of embedded systems. Then we will give a short overview of active and passive attacks on embedded systems. Our focus will be on the passive, so-called side channel attacks. To store sensitive material such as unreplacable biometric information, novel techniques such as fuzzy vaults and PUFs are presented.

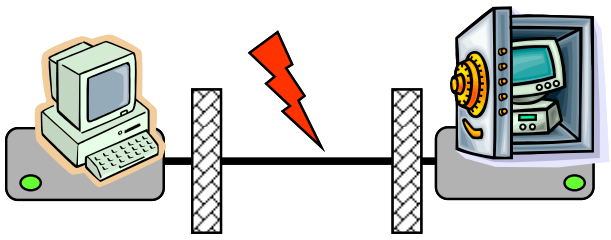


Figure 1: Traditional attack model

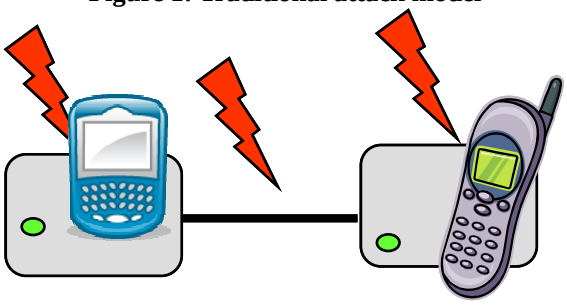


Figure 2: Attack model on embedded devices

2 Secure embedded systems

Processing and storage of data, both sensitive and non-sensitive has moved from centralized computers and servers to distributed wireless mobile devices. This changes the security model as indicated in Figures 1 and 2. In the old model, sensitive data is stored in desktops and servers, which are usually in locked offices or server rooms. This means that the communication channel is the main point of attack. Strong cryptographic algorithms and protocols can provide protection. In the new model, the attacker has access to both the channel and the end-nodes, or terminals. From a security viewpoint the model moves from a black box to a gray box.

In a traditional black box attack, the attacker will use techniques such as known- or chosen-plaintext attacks, linear or differential attacks on the algorithms or brute force key guesses. When the terminal is accessible to the attacker, she has many more option to derive information from the device.

Attacks to embedded devices are classified as either

active or passive attacks. A typical memory related active attack is the so-called buffer overflow or memory overwritting attacks. With this attack, an argument or entry into a program takes more memory than allocated for it. The malicious code sits at the end of the overlong input and might be executed by an unsuspicious other program. Countermeasures for these attacks are available [2].

The passive attacks that make use of the information leaked from these gray boxes are called side channel information leaks and side channel information attacks. As they are non intrusive and often don't need expensive equipment. They are a real threat as they can go undetected for a long time.

3 Side channel information leakage

Integrated circuits are at the heart of embedded systems. The implementation of the cryptographic algorithms and the processing of data and keys discloses possibly a lot of information. There are many sources of passive information leaks from integrated circuits. These side-channel leaks are subdivided in main categories.

3.1. Timing attacks

A first important category are timing attacks. From the execution time or response time of an integrated circuit, extra information can be derived. Consider the following simple example of a modular exponentiation algorithm by Kocher [9]. This algorithm evaluates $R = y^k \bmod n$, with k being the equivalent of a private key.

```
modexp(in k, in y, out R_{w-1}) {
    s_0 = 1
    for j = 0 to w – 1
        if (bit j of k) is 1 then
            R_j = (s_j . y) mod n
        else
            R_j = s_j
        S_{j+1} = (R_j)^2 mod n
    end for
```

When this algorithm executes in software, the value of k decides, bit by bit, which branches of the if-then-else statement execute. These branches are easy to distinguish since the multiplication $(s_j.y) \bmod n$ requires more computations on the processor than the simple assignment of s_j. An attacker could use the execution timing or the power profile of the processor to obtain the secret value k.

3.2. Cache attacks

Cache attacks are one class of timing attacks that are of particular concern for memory architecture designers. E.g.. a software implementation of the AES (Advanced Encryption Standard) cipher typically uses look-up tables. Bernstein, Osvik [3][10] and other authors have pointed out that the lookup tables used in AES (S-boxes) are a timing side-channel into the roundkey. Depending on the presence of a S-box entry into the processors cache, the execution time of the AES algorithm shows small variations.

A solution is to write constant executing time software. This is really a challenge as most memory architectures have a very time varying behaviour. Caches and multiple level of caches have been introduced to accelerate the average execution time by keeping recently used or repeatedly used data local to the processing units. More levels of cache will leak even more information.

3.3. Power attacks

Power attacks are a second major category of side channel attacks. CMOS technology is the technology of choice for low power integrated circuits. Its main advantage is that it only draws current when the circuit is active (apart from static leakage currents that become more and more a problem in deep-submicron technologies). Thus the activity or the processing of data can be monitored by monitoring the power supply. In this case the easiest attack points are the storage elements, in most cases registers or flip-flops to store intermediate data or keys. Multiple encryptions are monitored and the power profile is captured. The power consumption of storage devices is correlated to the toggle count of the flip-flops. I.e. the power consumption is correlated to the Hamming distance between the currently stored and the next values is registers. Experimental results show that less than 5000 encryptions on an 0.18 µm CMOS standard cell implementation of the AES algoroithm are sufficient to disclose the key of 128 bits [5]. For standard cell implementations, countermeasures exist by introducing e.g. dynamic differential logic styles or masked logic styles. The total power and area budget will however go up by a factor 2 to 4 for e.g. the WDDL logic style [14] and a factor 10 for the MDPL masked logic style [11].

3.4. Electromagnetic radiation

Electromagnetic radiation is yet another source of information [1]. High speed clocks with steep rise and fall times, have many harmonics. These clock harmonics get unintentionally modulated with information that depends on the operations on the integrated circuit. These high frequency signals can be captured and demodulated without being even close to the integrated circuit under attack.

3.5. Optical attacks

Optical attacks are not a passive, but semi-invasive attack. For memory devices, optical attacks are of particular concern. In [12] Skorobogatov shows how a flashlight from a camera or a laser pointer can change the state in a SRAM memory. This can then be combined with e.g. a fault attack to derive the key from the device.

3.6. Data reminance in semiconductor memories

Gutmann gives an extensive overview in [4] on possible sources of data reminance in volatile and non-volatile memories. Even when the memories have been erased, effects such as hot-carrier and electromigration leave information behind. Similarly, writing and erasing non-volatile memories leaves carriers behind. Combined

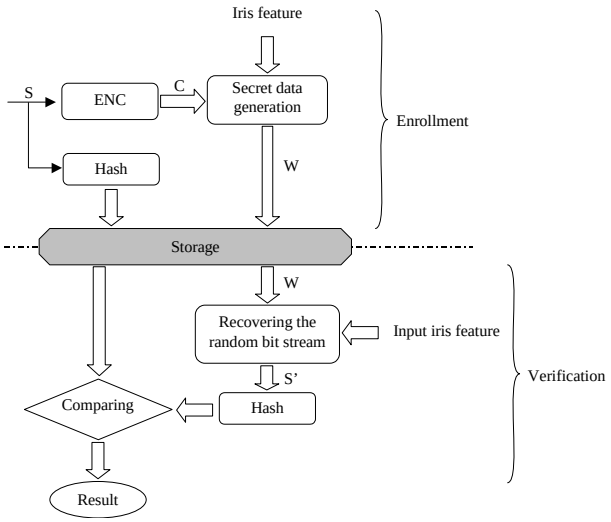


Figure 3: Fuzzy vault based iris verification

with semi-invasive techniques, some information can be obtained [13]. Gutmann gives a few practical guidelines for a secure use of semiconductor memories.

4 Unconventional storage

The storage of some classes of data is very critical and once disclosed cannot be corrected. This is the case for biometric information. Fingerprints, retina scans or the DNA of a person, have the advantage that it is unique from one person to another. So, it can be readily used to identify someone. However, in many cases it requires that reference biometric data is stored in a database. E.g. an employer can keep a database of fingerprints to give his employees access to the building and office areas. But even biometric applications are moving towards embedded portable devices such as laptops, key chains, door openers, and so on. With enough effort, using active and passive attacks, this information can be pulled out of the embedded device and might be compromised.

To address this issue, new fuzzy vault or helper data based approaches are proposed [6][7]. The idea is illustrated in Figure 3 for an iris verification system [16]. The biometric data itself is not stored but it is used to lock and unlock a vault. The secret in the vault is an encoded random generated number. During actual use of the device, the iris features are used to unlock, i.e. recover the random bit stream, which is then compared to a hashed version.

Physically Uncloneable Devices (or PUFs) are another unusual 'storage' of keys. The idea is to use special coatings as in [15] or other unique features of the integrated circuit and to derive from it a key or identification which is unique for the device. Other research tries to derive this from process variations between devices. E.g. for RFID tags it is considered too expensive to write a unique identification number into every device. For many counterfeiting applications, the security requires that the secret can be read but not be modified [15].

5 Conclusion

The ultimate secure memory does not exist. Many attacks are possible and countermeasures are proposed but many new attacks will appear. Therefore, secure storage will only be provided when looking at the global picture: including system, architecture, logic, circuit and physical protection mechanisms.

Acknowledgement

This work was partially supported by FWO projects G.0475.05 and G.0300.07 and funds from the K.U.Leuven.

References

[1] D. Agrawal, B. Archambeault, J. Rao, P. Rohatgi, "The EM Side-Channels(s)," CHES 2002, LNCS 2523, pp. 29-45, 2003

[2] R. Anderson, "Security Engineering: A guide to building dependable and distributed systems," Wiley Computer Publishing, 2001.

[3] D.J. Bernstein, "Cache-timing attacks on AES," preprint, 2005, online at http://cr.yp.to/papers.html

[4] P. Gutmann, "Data remanence in semiconductor devices," 10th USENIX symposium, August 2001.

[5] D. Hwang, K. Tiri, A. Hodjat, B. Lai, S. Yang, P. Schaumont, and I. Verbauwhede, "AES-Based Security Coprocessor IC in 0.18-um CMOS with Resistance to Differential Power Analysis Side-Channel Attacks," IEEE Journal of Solid-State Circuits 41(4), pp. 781-792, 2006.

[6] Juels, A. and Sudan, M., "A Fuzzy Vault Scheme," IEEE Int. Symp. on Information Theory, pp. 408-13, 2002, Piscataway, NJ.

[7] Linnartz, J-P. and Tuyls, P., "New Shielding Functions to Enhance Privacy and Prevent Misuse of Biometric Templates," 4th Int. Conf. on Audio- and Video-Based Personal Authentication, pp. 393-402, 2003, Guildford, U.K., Springer Verlag LNCS 2688.

[8] A. Kerckhoffs, "La cryptographie militaire, " Journal des sciences militaires, vol. IX, pp. 5–83, Jan. 1883, pp. 161–191, Feb. 1883

[9] P. Kocher, "Timing attacks on implementations of Diffie-Hellman, RSA, DSS and other systems," Proc. CRYPTO '96, Lecture Notes on Computer Science, 1109:104:113, Springer-Verlag, 1996.

[10] D. Osvik, A. Shamir, E. Tromer, "Cache Attacks and Countermeasures: the Case of AES," Proc CT-RSA, LNCS 3860, 1—20, Springer, 2006.

[11] T. Popp, S, Mangard, "Masked Dual-Rail Pre-charge logic: DPA-resistance without routing constraints," CHES-2005, LNCS 3659, pp. 172-186, 2005.

[12] S. Skorobogatov, R. Anderson, "Optical fault induction attacks," CHES 2002, LNCS 2523, pp. 2-12, 2002.

[13] S. Skorobogatov, "Data Remanence in Flash Memory Devices.,"CHES-2005, LNCS 3659, Springer-Verlag, pp.339-353

[14] K. Tiri, and I. Verbauwhede, "A Digital Design Flow for Secure Integrated Circuits," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 25(7), pp. 1197-1208, 2006.

[15] P. Tuyls, and L. Batina, "RFID-Tags for Anti-Counterfeiting," In Topics in Cryptology - CT-RSA 2006, LNCS 3860, Springer-Verlag, pp. 115-131, 2006.

[16] S. Yang, and I. Verbauwhede, "Secure Iris Verification," In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007), May 2007.

From Memory Component to Memory System

Doris Keitel-Schulz
Qimonda, QFL, Munich (doris.keitel-schulz@qimonda.com)

Abstract

Scaling of technologies and increasing challenges concerning clock speed, interface speed, active and standby power consumption, memory density and memory reliability require special solutions on chip and on system level.
In this paper we want to show the different areas for innovation and optimization to built reliable memories and memory systems.

1. Introduction

As the technological difference between SOCs and memory designs is ever increasing, System partitioning creates new kinds of memory solutions. These new memories are on the one side new application specific components, on the other side new memory sub-systems and systems. In contrast to the memory dies the memory systems and sub systems contain also logic devices, passives, boards and software in addition to the memory die.

2. Application specific memories

Typically each technology started with one memory architecture. I.e. DRAM started with asynchronous DRAM followed by EDO, later synchronous DRAMs have been developed which now are optimized for different application areas as shown in Fig. 1. The same already happened for NOR Flash and is just starting for NAND Flash.

DRAM	SDRAM DDR1,2,3,.. High Speed Low Power Graphics
NOR	Consumer, Mobile Computing Communication
NAND	Large page SLC Small page MLC DDR

Fig. 1: Memory types in Flash and DRAM technologies

Main product segments for memories are PC and Server, Graphics, Consumer and Mobile applications as shown in Fig.2.
Innovations for PC, server and graphics require high speed design of the interface and data path of memories [1,2,3]. Consumer and Mobile need power saving techniques in operation and stand by like temperature sensors, perfect matching of I/O drive capabilities, self timing of refresh operations and lowering of supply voltages.

Fig.2: Application areas for memories

To enable higher densities and wider buses, solutions like Multi Chip Packages and modules are developed.

3. Multi Chip Packages

Multi chip packages, and package on package devices are on the one side a means to increase density putting 2 or more dies of the same memory into one package.

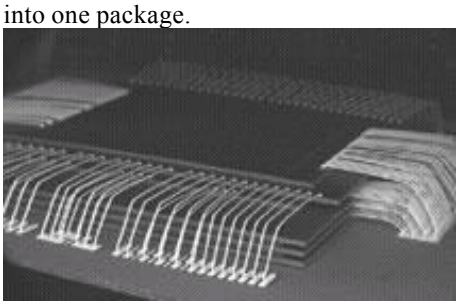


Fig. 3: Multi Chip Package example

On the other hand multi chip packages are used in i.e. mobile devices to host different kind of

memories. To accommodate several dies in a package, thermal and mechanical properties have to be simulated thoroughly and feed back to the chip designer. In addition new wafer thinning and package technologies have to be developed and improved to manage up to 8die stacking.
A typical example for a multi die MCP is the memory architecture in mobile phones as shown in Fig.4.
Historically the Baseband part of a mobile phone consisted of at least a DSP and µController with a memory bus and a memory controller. Connected to the memory bus is a Code Flash (NOR), Data Flash (NAND) and working memory (mainly DRAM). The code for the µC is stored in the NOR Flash [5], the data Flash [6] is used to store i.e. pictures or downloadable content. This methodology is called 'Execute in Place', as the code can be operated directly out of the NOR [4].
Today however, NAND Flash is significantly less expensive than NOR flash, thus the architecture changed to 'Shadowing'. In this case the NOR Flash is replaced by less expensive but slower NAND Flash. To compensate for the lower speed of the NAND, the code is shadowed at power up into the faster working memory, which has to be increased accordingly and is executed from there.

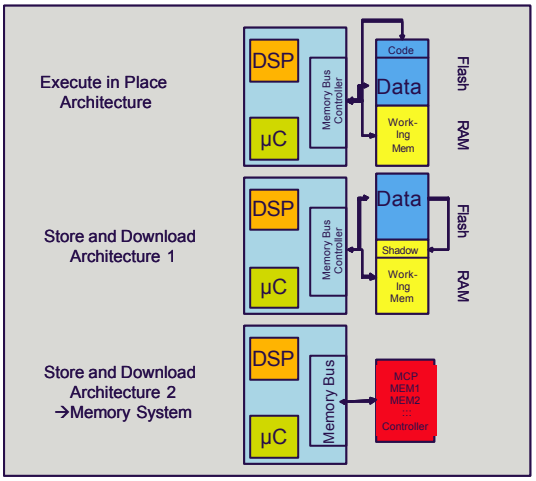


Fig. 4: Different possible memory architectures for baseband applications

4. Memory Systems and Sub-Systems

The next step of integration is the memory system in package. In this case the memory controller will be stacked together with all the memories. Via the controller interface to the bus the access to the application is defined. The controller also takes care, that an 'error free' memory operation is guaranteed. For this purpose Error Detection is introduced on memory dies and Error Correction is handled by the controller.

Another true memory system is a solid state hard disk. As NAND Flash devices will approach 16Gb in the near future, SSD with 32 to 64GB will be feasible. The principle architecture of the SSD shown in Fig. 5 is very close to a magnetic hard disk. The innovations necessary are in the field of Error correction and memory interface to the Flash Controller to reach the same reliability and even better performance values as the HDD. To obtain the optimum features, the error correction of the controller and the error correction capabilities of the flash die need to be optimized to each other. As an example the number of bits reserved for error correction on die needs to be exactly balanced with the capabilities of the controller. The selection of suitable Error correction codes and their adaptation is one of the main innovation areas, as semiconductor storage behaves significantly different as magnetic storage concerning failure modes and BER.

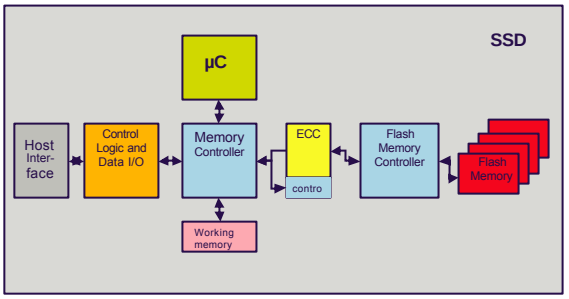


Fig. 5: Schematic of a Solid State Disk

After system optimization the advantages of the SSD versus the HDD are demonstrated in the low power characteristics, the form factor, the better random read characteristics, the higher temperature range and the higher shock resistivity.

5. Summary

Solid state memories are becoming more and more application specific. As the technologies for memories and logic devices are significantly different in performance, supply voltage and transistor density, system partitioning focuses more and more on logic and memory systems or subsystems. The development of memory systems and subsystems requires an in depth understanding of the target application especially of memory buses and interfaces, low power techniques, packaging development and system reliability methodologies like error correction.

[1] C. Yoo et al, A 1,8V 700Mb/s/pin 512Mb DDRII SDRAM with on die termination an off chip driver calibration, JSSC vol. 39 June 2004

[2] T. Matano et al, A 1-Gb/s/pin 512Mb DDR II SDR using a digital DLL and a slew rate controlled output buffer, JSSC vol. 38, May 2003
[3] H. Pilo et al, A 5,6nm random cycle 144Mb DRAM with 1,4Gb/s/pin and DDR3 SDRAM interface, JSSC vol.38, Nov 2003
[4] Lars Wehmeyer, Peter Marwedel, Fast, efficient and predictable memory accesses, Springer 2006
[5] T. Ogura et al, A 1,8-V 256Mb Multilevel cell NOR flash memory with BGO function
[6] A 56nm CMOS 99-mm² 8Gb Multi-Level NAND flash memory with 10-MB/s program throughput, JSSC vol 42, Jan 2007

A Designer's Perspective On Future Memory Architectures For Software Defined Radios

P. Marchal [a], B. Bougard [a], A. Papanikolaou [a], M. Miranda [a], F. Catthoor [a b] and W. Dehaene [b]

[a] IMEC, Belgium, marchal@imec.be
[b] ESAT, Kuleuven, Belgium

Abstract

Data-rich wireless communication terminals, such as smartphones, are integrating an increasing number of wireless access standards. To limit the manufacturing cost, both academia and industry are investigating the concept of software-defined radios (SDR). In this paper, we focus on the design challenges for building the memory architecture of such SDR platforms. Particularly, we advocate the use of advanced processing technologies for reducing cost, and limit the power consumption with widely distributed, still reliable and testable, memory architectures and by applying better-than-worst-case design methods for coping with process uncertainties, inherent to nano-scaled technologies.

1. Introduction

A strong growth exists for wireless data-rich services (such as video telephony, photo editing, multi-platform games, etc.), with expected income to constitute about 80% of the total wireless revenues around 2010. Serving this market segment is key for wireless system providers, but requires mobile platforms on which computing and communication are fully converged. These converged platforms should be capable of accessing multiple networks (WLAN, WPAN, WMAN, cellular) and, at the same time, provide support for several multimedia services (music, video, 3D games etc.). Their design entails the following two critical challenges: (1) limiting the product cost and (2) achieving the desired performance within the battery power constraints.

Flexible radios, s.a. Software Defined Radios (SDR) and CMOS scaling limit cost, but challenge energy efficiency. In a classical multi-mode design, the silicon real estate increases proportional to the growing number of supported standards. The *Software Defined Radio concept*, where a variety of radio access standards are implemented as software modules on a generic hardware platform, allows for a more efficient silicon usage and thus a lower cost. Moreover, as the SDR platform is programmable, it provides a scalable platform for future product generations, and thus reduces the NRE and time-to-market. Unfortunately, such flexibility inherently comes with an important energy penalty [1] that architects typically limit by introducing *extremely distributed, parallel, processing architectures.* In section 2, we will discuss the consequences of these design innovations on the memory architecture, based on the analysis of the SDR platform we have been developing.

Better than worst-case design methods for facing the nano-scale devil. Besides these architectural innovations, designers should benefit from Moore's law to further reduce the manufacturing cost. However, manufacturing uncertainties in nano-scale technologies cause functional and parametric yield loss (section 3). Current design practice cannot cope efficiently with the increasing number of design corners and their widening distributions, thereby increasing system's area/power penalty and thus jeopardizing scaling benefits. In memories the problem is aggravated by the many critical paths and minimum sized transistors[1]. Designers are therefore in great need of better-than-worst-case design techniques (BTWC) to maximize cost savings while minimizing power consumption, particularly for memories. We examine two possible BTWC design routes: (1) statistical design optimisation to avoid the accumulation of design margins (section 5) and (2) extensions for self-adaptive architectures, enabling the system to tune its available hardware to the actual (performance/power) requirements (section 6). Finally, we conclude the paper.

2. Low-Power High-Performance SDR

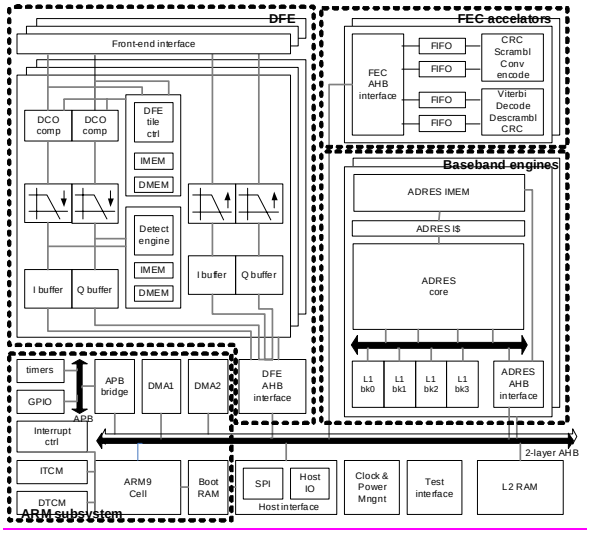
Many architecture styles have already been proposed for size weight and power constraint SDR platforms [15-19]. On most platforms, physical layer inner and medium access control functionalities are jointly implemented. It is commonly admitted that *heterogeneous multi-processor* SOC is the most efficient approach. Such architectures are described in [15, 18]. The SOC is generally articulated around a CPU (e.g., a ARM core) to which slave processing entities (PE) are appended. PEs are typically designed keeping in mind the most important characteristics of wireless physical layer processing: *high data level parallelism* (DLP) and *data flow dominance*. In the considered platform, as depicted in Figure 1, besides a ARM9 CPU and its peripherals, three types of PE are implemented with different programmability versus energy-efficiency trade-off:

Baseband processing engine –Very long instruction word (VLIW) instruction set processors with SIMD (Single Instruction – Multiple Data) functional units are mostly considered to exploit the data level parallelism with limited instruction fetching overhead [18,19]. Besides, data flow dominance is often exploited in coarse grain re-configurable arrays (CGA) [20,21]. The SDR platform embeds two hybrid baseband processors combining CGA and SIMD features [22]. The latter are each associated with a 4-bank data scratchpad with full cross-bar and memory access conflict arbitration. A limited number of units can be operated in VLIW mode, accepting arbitrary C code (glue code), fetching instruction through a 32K 128-bit wide instruction cache. When in array mode, DSP kernels are executed while keeping configuration into local buffers that are configured through direct memory access (DMA). Each can sustain a computing load of 50GOPS, comprises 34 memory instances and consumes 80 mW in VLIW mode and 260mW in kernel mode, including all memories and interfaces.

FEC accelerators – Forward error correction typically requires 10x more computing power than inner modem processing. Moreover, it mostly relies on a limited set of well-known algorithms (Viterbi, RS, turbo, belief-propagation). Therefore, configurable application specific VLSI architectures are usually considered. Our platform embeds two FEC accelerators with configurable convolution encoder, Viterbi decoder, (de)scrambler and CRC calculation/check support.

[1] Pelgrom's law states that smaller transistors are more sensitive to process variations [10]

Each comprises 6 memory macros, mostly in FIFO configuration.



Digital Front-End – the computing power of the CGA baseband engine coupled to its high level of programmability makes it a good choice for inner modem processing. However, these features are not strictly required for packet detection functions. Since those have a high duty cycle (they are active in idle phase), higher energy efficiency cores are needed. Three digital front-end tiles are implemented. Each is associated with a signal path from a multi-antenna analog front-end. A single tile comprises a transmit section that buffer, over-sample and forward the I/Q samples to the ADC, and a receive section where packet detection is implemented. The receiver path is supervised by a tile controller where automatic gain control, DC offset compensation and power detection are implemented. When signal power is detected, the samples are down-sampled (specifically designed filters) and buffered. Besides, a programmable detection engine is activated where time-domain synchronization is implemented. A DFE tile can generate a SOC level interrupt upon effective signal detection, waking up the ARM subsystem, which then ordinates the baseband processing using the baseband engines and the FEC accelerators. The DFE counts 18 memory macros.

Next to these three type of processing unit and the ARM subsystem with a 256KB L2 memory, instruction and data caches, instruction and data tightly coupled and a boot RAM complete the design.

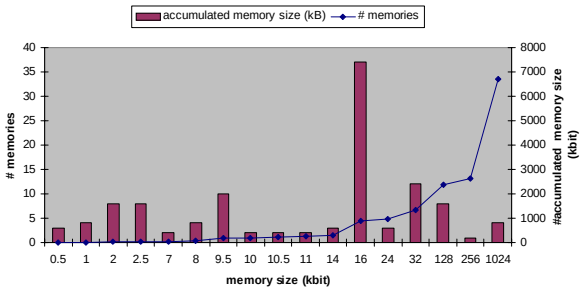


Figure 1 Memory Distribution

In Fig. 2 we summarize the on-chip memory usage. The SDR platform contains in total 113 memories. Most memories are rather small, having a size less than 256kb, and are distributed across the platform. They mostly act as buffer between, scratchpad and/or cache memory near the processing elements. Being very intensively used, they are also responsible for most of the memory hierarchies' dynamic power consumption. Moreover, they occupy up to 40% of the systems' memory footprint. Optimizing their usage with software-cleaning methods [14] and innovative circuit design is therefore mandatory for limiting power and area. Besides, the scaling momentum should be maintained for further reducing cost, even if it entails challenging the nano-scale devil.

3. The Nano-Scale Devil

Whereas technology scaling made systems more performing, less power consuming and cheaper, these times of happy scaling have come to an end. *Manufacturing variations are becoming worse with every new technology generations.* Scaling not only exacerbates existing uncertainties such as litho-related variations, line edge roughness, random dopant fluctuations, but also introduces new ones such as device

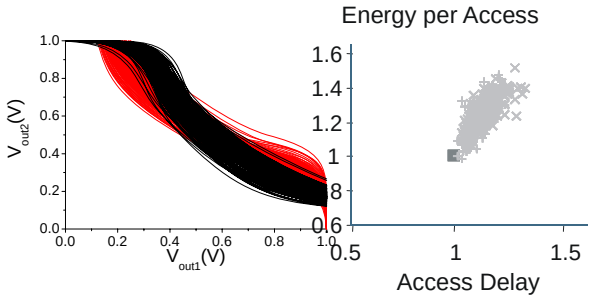


Figure 2 Impact of manufacturing uncertainties on SRAM stability and energy/delay

degradation (due to NBTI and electro-migration) [6][7]. Manufacturing uncertainties cause the electrical performance of devices and interconnect to deviate from the nominal case., thereby causing *functional yield loss.* Consider the inverter transfer characteristics of the 6-tor cells of a memory in Fig. 1-left. Due to process variations in many butterfly curves no SNM is remaining, i.e. these cells are incapable of retaining data. Besides, these uncertainties strongly influence the timing and energy consumption of circuits, thereby *complicating timing and power closure.* In Fig 1-right, we show the energy-delay pairs for 200 samples of the same 8kbit memory simulated in *65nm PTM* technology assuming σVt=0.1Vt. Variations up to 40% in performance and 45% in energy compared to the nominal case were measured. As the operating voltage is reduced (and thus while scaling), circuits' sensitivity to these variations becomes even worse.

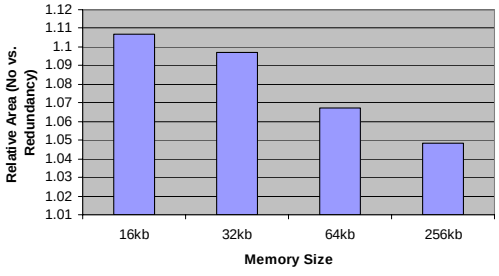


Figure 3 Area (and the corresponding power) penalty for redundancy increases for smaller memories

Existing design practice guard-bands each source of variations such that the functionality of the manufactured silicon is guaranteed. Each device in every gate of the design is assumed to operate in the worst possible way. Design margins are accumulated to make this design yielding. In reality, it is extremely unlikely that this worst-case situation ever occurs. Whereas this worst-case design approach was acceptable above

130nm, in nano-scale technologies the accumulation of safety margins at all levels of the design flow result in over-design, and thus come at the expense of extra power and area. Existing solutions for coping with uncertainties, such as redundancy are very effective for large memories, but result in an important cost penalty when applied on these small memories. In Fig. 4 the relative area overhead of redundancy in function of the memory size is depicted for a memory generator targeting a 90nm technology. Clearly, the overhead becomes larger for smaller memories as those encountered in the SDR platform, or in wireless platforms in general.

4. Driving Out the Nano-scale Devil

Systems designed following this worst-case design approach only marginally benefit from scaling, particularly in the sub-45nm regime. Designers are thus facing the nano-scale devil In response, both academia and industry have been developing novel circuit and architecture solutions for coping with these variations more efficiently. We discern three different approaches:

- Reducing variations with *design for manufacturing techniques* (DFM). The source of variations is modelled and consequently targeted by minor layout and circuit modifications. Design for manufacturing techniques typically target litho and etching induced variations (e.g., [13] and start ups such as PDF solutions).
- *Statistical analysis and optimisation techniques* Rather than accumulating design margins to ensure a yielding design even for the highly unlikely, worst-case corner (see section 3), statistical methods estimate the design's actual yield and steer the optimisation such that it achieves a desired yield target. Today, most statistical methods target the standard cell designs (e.g., industrial solutions are provided by Magma, Synopsys, Cadence and Extreme) In section 5, we discuss a similar technique for memory design.
- *Self-adaptive systems* are systems that can counter process variations at run-time. The idea is that they can monitor the chip's performance, identify slow circuits and locally increase/repair these defective circuit. It can be designed for the actual case rather than the worst-case. We will introduce the challenges for building self-adaptive systems and demonstrate a self-adaptive memory architecture in section 5.

As many more sources of uncertainties exist apart from the litho/etching-related ones, the classical DFM techniques must be complemented with statistical and preferably self-adaptive design techniques to build yielding memory architectures for SDR platforms. In the next two sections, we overview both techniques in more detail.

5. Statistical Design Optimization

Exploiting statistical models of the manufacturing uncertainties during design may limit design margins or guard bands. In classic design procedures, manufacturing variations are dealt with by assuming a worst case situation and adapting the design parameters accordingly. From that point on normal deterministic analysis and optimisation methods are used. As explained above, accumulation of design margins affects the energy-delay trade-off too much.

Alternatively statistical variations can be taken along in the design. This is done based on two main principles:

- The design or optimization criteria formulated as yield targets rather than performance, energy or robustness targets. This means e.g. that the design criterion is no longer a static noise margin of 100 mV. But a static noise margin of 100 mV in 99.7 % of the cases.
- The optimization algorithm is directed by statistical sensitivities rather than the sensitivity of the parameter itself. This means that the gradient of the yield as function of the design parameters (e.g. sizing) is used in the optimization.

These principles can only be applied if two prerequisites are fulfilled:

- When using statistical design optimisation, there will be a certain yield loss. It is mandatory that defect samples can be distinguished from the working silicon during product test.
- The statistical distribution of the crucial parameters must be known. This means that statistical variations of threshold voltage, mobility, subthreshold slope and so on must be characterized prior to the design.

A statistical design methodology was applied to an SRAM cell in [2]. As an illustration, one of the results is repeated in figure X, where an SRAM cell is optimised. The optimisation objective is to size the cell such that cell area is minimized while meeting a signal noise margin target. The statistical design approach clearly outperforms the worst-case one, using on average 15% less area while achieving a similar SNM. Similar statistical design optimisations have been proposed by [8][9]. In future work, we plan to extend these concepts to entire memories.

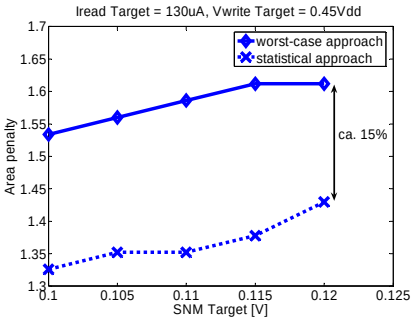


Figure : Statistical aware design reduces area for a desired SNM design target [2]

6. Towards Self-adaptive Architectures

An alternative approach is to go for systems that can operate with uncertainties and, hence, that are optimized for the typical case. These systems will measure uncertainties and compensate for them at run-time [11][12]. The logic components of such *self-adaptive* systems are illustrated in Fig. 5. It requires:

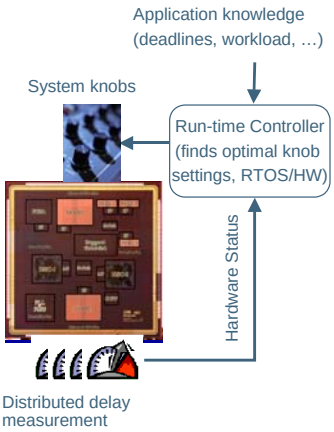


Figure 4 Logical components of a self-adaptive system

Uncertainty tolerant circuits. Circuits that can tolerate significant process and environment induced uncertainties and remain functionally correct over a wide performance range.

Delay monitors. The performance of the above circuits depends on the level of uncertainties. To guarantee throughput, the system should be able to monitor itself. Simple circuits for performance monitoring must be introduced.

Knobs. Besides monitoring its performance, the system should be able to adapt/tune itself at run-time. These knobs should therefore be integrated into the memories blocks. In [3][5], we have introduced as a promising option.

System controller. It should steer the knobs based on the information provided by the monitors. It should match the system's performance with the application's needs while minimizing the power consumption.

These concepts have been integrated into a self-adaptive memory hierarchy of a DAB receiver [4]. Preliminary results indicate a possible energy reduction of 30% compared to a worst-case design assuming a 65nm PTM technology. Many challenges remain to introduce self-adaptive systems in the design flow. At the circuit level, more efficient (in terms of area/cost) monitor circuits and novel knobs with a large performance range are required. At the system-level, a method should be developed to convert existing designs into self-adaptive ones. This technique should introduce monitors and knobs, while trading off their benefits with the extra area overhead.

7. Conclusions

Data-rich services require low-cost mobile platforms capable of accessing multiple standards. Rather than adding an extra radio for every other standard, academia and industry are exploring the benefits of flexible radios. To limit the power cost of flexibility, designers strongly rely on (1) parallelism and (2) the use of distributed memory architectures. Besides, technology scaling is used to further limit the manufacturing cost. Better-than-worst-case design methods should be applied for dealing with the inherent process uncertainties. Two possible BTWC-methods have been presented: statistical memory sizing and self-adaptive systems. Both techniques allow to trade-off energy/yield/performance of systems. Many challenge and opportunities remain to extend both design techniques.

References

[1] T. Claasen et al., "High Speed Not the Only Way to Exploit the Intrinsic Computational Power of Silicon", *Proc. ISSCC*, pp.22-25, Feb., 1999

[2] E. Grossar et al., "Statistically aware SRAM memory array design", *Proc. ISQED*, pp.-, March, 2006

[3] H. Wang et al., "Variable tapered pareto buffer design and implementation allowing run-time configuration for low-power embedded SRAMs", *IEEE Trans. VLSI Systems 13(10)*, pp. 1127-1135, 2005

[4] A. Papanikolaou et al., "A system-level methodology for fully compensating process variability impact of memory organizations in periodic applications", *Proc. CODES+ISSS* 2005, p117-122

[5] E. Karl et al., "Timing Error Correction Techniques for Voltage Scalable On-Chip Memories", *Proc. Symp. Circuits and Systems*, pp 3563-, 2005

[6] Groeseneken G., "Recent Trends in Reliability Assessment of Advanced CMOS Technologies", *Proc. of IEEE Intl. Conf. on Microelectronic Test Structures*, pp81-88, vol. 18, April, 2005

[7] McPherson J.W., "Scaling-Induced Reductions in CMOS Reliability Margins and the Escalating Need for Increased Design-In Reliability Efforts", *Proc. ISQED* 2001

[8] S. Mukhopadhyay et al, "Statistical design and optimization of SRAM Cell for Yield Enhancement", *Proc. ICCAD*, 10-13, 2004

[9] Kanj et al.,"Mixture Importance Sampling and Its Application to the Analysis of SRAM Designs in Presence of Rare Failure Events", *Proc. DAC*, pp. 69-, 2006

[10] M.J. Pelgrom, A.C.J., "Duinmaijer and A.P.G. Webers. Matching Properties of MOS Transistors," *IEEE J. Solid State Circuits*, vol. 24, no 5, Oct. 1989, pp. 1433-1440.

[11] D. Ernst et al., "Razor: A Low-Power Pipeline Based on Circuit-Level Timing Speculation", *Proc. Symp. Micro*, 2003[12] D. Sylvester, D. Blaauw and E. Karl, "ElastIC: An Adaptive Self-Healing Architecture for Unpredictable Silicon", IEEE D&T, Vol. 23, No. 6, pp 484-490

[13] W. Stephen, W. Maly and A. Strojwas, "VLSI Design for Manufacturing: Yield Enhancement", Springer, ISBN: 978-0-7923-9054-1, 1990

[14] F. Catthoor et al., "Custom memory management methodology", Kluwer, ISBN: 978-0-7923-8288-1, 1998

[15] B. Bougard et al., "A scalable programmable baseband platform for energy-efficient reactive software-defined radio," *Cognitive Radio Conference and related topics*, IEEE, Ed. Mykonos, Greece, 2006.

[16] G. Desoli and E. Filippi, "An outlook on the evolution of mobile terminals: from monolithic to modular multiradio, multiapplication platforms," *Circuits and Systems Magazine, IEEE*, Vol. 6, pp. 17-29, 2006

[17] J. Glossner et al., "A software-defined communications baseband design," *Communications Magazine, IEEE*, Vol. 41, pp. 120-128, 2003

[18] L. Yuan et al., "SODA: A Low-power Architecture For Software Radio," pp. 89-101, 2006

[19] K. Van Berkel et al., "Vector Processing as an Enabler for Software Defined Radio in Handsets for 3G+WLAN Onwards," *SDR Forum Technical Conference*, 2004, pp. 125-130.

[20] A. Lodi et al., "XiSystem: a XiRisc-based SoC with reconfigurable IO module," *Solid-State Circuits, IEEE Journal of*, Vol. 41, pp. 85-96, 2006.

[21]H. Singh et al., "MorphoSys: an integrated reconfigurable system for data-parallel and computation-intensive applications," *Computers, IEEE Transactions on*, Vol. 49, pp. 465-481, 2000

[22] D. Novo et al., "Energy-Performance Exploration of a CGA-based SDR Processor," *SDR Forum Technical Conference*. Orlando, FL., 2006

Smart Cards: Technologies and Products

R. Zambrano [a], E. Toscano [a] and A. Conte [b]

[a] STIncard, z. i. Marcianise sud – 81025 Marcianise CE - Italy
[b] STMicroelectronics, stradale Primosole, 50 - 95121 Catania CT - Italy

Abstract

Smart Cards are among the most pervasive products in today's world. SIM cards, disposable phone cards, credit and debit cards, RF ID, e-passports, e-government, e-purse, pay-TV, all these products and applications are based on a plastic card with an embedded Si chip featuring high level of security (almost impossible to duplicate, very difficult to simulate) and complete portability (on chip read/write capability allows easy storage and retrieval of data), enabling, thanks to embedded security algorithms, exchange of highly sensitive data (such as money transactions and personal records) with large infrastructures (ATMs, GSM networks, ID databases).

In this paper we will review some aspects related to the specific features of these products, with a specific section dedicated to the technology aspects.

1. Introduction: Generalities

The origins of IC Cards can be traced back to disposable phone cards first used in France, then in Western Europe and now worldwide. The contact card is the most commonly seen: although 8 contacts are defined, only 5 are normally used. The contactless card receives operating power through an inductive loop using low frequency electromagnetic radiation. Most contact cards contain a simple integrated circuit, although some products are developed that use more than one chip.

Smart Cards are used for portable storage and retrieval of data, hence the need for Non Volatile Memory (NVM) on board. Request for byte alterability of the memory content has resulted in use of EEPROM being predominant over other solutions. The request for higher storage densities and lower cost will dictate (at least for high end products) a shift towards Flash memories.

Then comes the need for other functionalities: a microcontroller (MCU) is key in getting the system working (incidentally, this justifies the "Smart" label), ROM (to store the operating system) and RAM (to correctly operate the MCU) come immediately after. The control logic must also ensure against fraudulent use. It needs to prevent unauthorized access to the EEPROM where data are stored, must act quickly to ensure transactions are completed within the allowed time slots, and must prevent any type of sniffing from intruders. Accordingly, security firewalls (covering the ROM and the EEPROM) are integrated, with other blocks dedicated to cryptography and unpredictable number generation. The schematic block diagram is schematized in figure 1.

A feature common to basically all Smart Cards, that at the same time makes them different from many other volume products is the personalization, i.e. each unit carries some unique traits that make it different from all the others.

In case of SIM cards these "traits" are phone number, PINs and PUKs. When banking applications are involved we start dealing with actual names of people, in case of health cards, e-passports and other ID applications it's personal (and very sensitive) data that are stored on the cards and are different from all other cards issued.

Almost any card stores different secret keys for authentication, performing cryptography algorithms.

Handling these data requires large databases, and specific security measures, such as complete traceability of materials (printed plastics, holograms, signature panels, pre-personalized devices and cards) and careful management of physical and SW keys for secure exchange of data between the institutions that issue the cards (telephone operators, pay per view operators, banks, governments, …) and the card manufacturer.

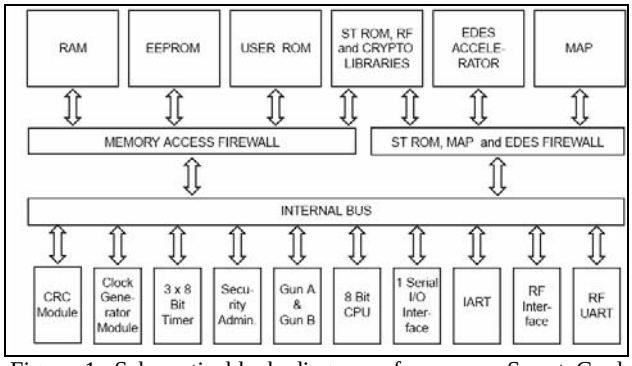


Figure 1: Schematic block diagram of a secure Smart Card (ST19WR66 architecture)

2. Applications and device requirements

The largest market today is for SIM cards. According to data from specialised institutions in 2006 more than 1.7B units were sold, a spectacular 45% increase over 2005. The projections for 2007 and 2008 are, respectively, 2.0 and 2.2 B units, still a healthy growth rate of more than 10%, with some initial signs of saturation.

Most of the SIM cards are using a 8 or, more frequently 16 or 32 bit MCU, need a rather large ROM (up to 500 kBytes) to store the card operating system, a 2 to 8 kBytes of RAM to allow fast execution of code, and anywhere from 16 to 256 kBytes of EEPROM, used to store specific product features (customizations), personalized data, and variable data. A closer look at the

memory density reveals that the lion share is for 32 and 64 kBytes EEPROM arrays (together 67% of total), low end products (16 kBytes and below) command 17%. The growth of high end products (128 kBytes and above) has been slower than expected, in 2006 the share has been close to 6%. Although the volumes are growing, the revenues are not going in the same direction (the market is actually commoditizing devices that are very customized).

Cards for banking applications normally use low density (2÷8 kBytes) EEPROMs. These products have a distinctive look (the "artwork"), printed with a glossy finishing (they represent the institution that issues the card, and are a real marketing tool carried for long time by the end user). The MCU must ensure high cryptography standards. The magnetic band on the back ensures backward compatibility with existing infrastructure (ATMs, POS).

An emerging application is that related to storage of personal data, such as ID (contact and contactless modes), biometrics, health records. The request for security is similar to that of banking products, and so are the demands to the MCU and memory.

Phone and loyalty cards products feature captivating artworks (sometimes actively searched by card collectors): these products normally require 1÷2 kBytes EEPROM arrays in combination with 8 bit MCU. Cards addressing the transportation sector and e-purse applications use the same HW resources, but for these products contactless is the main operation mode.

3. SW features

Smart Card operating system and software applications are getting larger and more complex. As a metric, we can look at the executable code of an operating system for the mobile market embedded in the ROM of a SIM. In the not so distant past (10 years ago) this was in the 20-50 kBytes range, and normally totally developed in assembler language with a proprietary architecture having poor logical segregation / separation between the software platform and the software application layer. Today the code takes anywhere from 100 to 500 kBytes. It is quite common to provide Smart Cards ready to download even remotely (post issuance, when cards are in the hand of the final users) applications written by third parties in standard languages (e.g. Java Card).

Some years ago the interoperability concept applied mainly to hardware and to the low level interface protocols to ensure the requirement of inter working between any Smart Card with any terminal. Today the interoperability paradigm is mainly applied to platforms, to API and applications to ensure correct operation of the same software application on Smart Cards provided by different vendors on different hardware/software platforms.

A software platform is today an open and complex system, very difficult to secure, which, on the other hand is the enabling tool to communicate with systems that permits exchange/protection of values or secrets. Service providers such as telecom operators and banks require the strongest protections against fraudulent transactions and rely on the Smart Card system (HW & SW) as the key element to secure any valuable transaction.

The architecture of a Smart Card system has to be viewed as a complex system that can be broken into functional blocks with different security requirements and functions. Ensure the SW highest security level while at the same time providing complex functionalities, flexible upgradeability and a behaviour nearest as possible to an open system is one of the strongest challenge for software designers.

A poorly designed software layer can compromise security even on the strongest protected hardware device. For example, if software passes secret information (e.g. supposed to be used only into internal calculation) out of the protected hardware environment because of a bug, not only a logic functional bug is shown, but a potential security hole is open and no hardware countermeasures can prevent the problem.

Secure software makes sense only on secure hardware. It is simple to demonstrate the concept with an example: if a secret information is correctly processed by the software but the current absorbed by the device or the electromagnetic radiation coming out from the device or the time needed by the hardware to process instructions depends on the value of the secret information and not only on the length of this secret information, a malicious "side-channel" attack can reconstruct the secret monitoring the power consumption, the radiated energy or the time needed to process the unknown secret information that is supposed to be well hidden.

The security level of the Smart Card is demonstrated and certified according to rules fixed by international standard committees. Best examples are the security certifications made accordingly to the ISO/IEC 15408 (Common Criteria) standard or accordingly to de facto standards stated by Visa or MasterCard to certify banking produces.

The Smart Card industry applies sophisticated software life cycle that have to guarantee secure complex product while the time available for development is often really short. Short time to design new software platforms as requested by the fast evolving SIM market and severe security requirements are two basic constraints difficult to find together in fields different from Smart Cards.

4. Technology

Devices for Smart Cards account for about 1% of the total semiconductor market, or approximately 10% of the MOS MCU market. Viable technologies for Smart Cards have to address the integration of high performance CMOS logic with NVM cells, not an easy feat as the requirements of the first fairly often are in conflict with the requirements of the latter and vice-versa.

CMOS logic uses thin (or ultrathin) gate oxides, low supply voltages, and relies on Place & Route tools to ensure efficient use of real estate Si area and proper

MCU functionality. On the other hand, NVM require on chip generation and management of high voltages and negative voltages, hence the need for thicker gate oxides. When we consider also tunnel oxide the overall count can easily go up to 3 or 4 different active oxides, with at least two levels of polySi. Field oxides too must be separated, because the logic integration density can be limited by latch-up (and then requires high thickness), while the memory array is very demanding on density. The match is better when the metallization layers are considered: efficient MCU designs can be made with 4 levels of metal, in the memory arrays the first 3 layers are used for interconnection, the last one can be used as screen (anti-tampering) against optical intrusions and μ-probing).

Portability dictates low power consumption, both during normal operation and in stand-by mode (then the need for low leakage currents). Exchange of data normally requires memories with high granularity (at the Byte level). Last but not least, we have the request of low cost, more and more pronounced as the market is not putting extra premiums on high density products. Once the technology is in place, added value from design comes from optimized device architecture and layout, rationalizing use of periphery and contributing to reduction of the overall device area.

Most of these requirements are addressed by using EEPROMs, thanks to their robustness and flexibility, demonstrated also when the very challenging requirements imposed by RF application are considered. The EEPROM cell has been scaled aggressively even beyond the limits that were supposed to be in place only a few years ago, several products are available today with 130 nm technologies.

Devices belonging to the 64 kBytes family (actual densities are 68 to 72 kBytes) were using approximately 10 mm^2 in 180 nm technology, with the ROM, EEPROM (or Flash) memory blocks and the CPU each taking approximately 25% of total, and the remaining 25% for RAM, other circuitry (including charge pumps) and bonding pads. Die size has been reduced by as much as 50% thanks to use of 130 nm technology, with the EEPROM using a larger portion of area compared to the previous technology generation.

As a matter of fact, EEPROM scaling is not easily carried out in the sub-100 nm regime. Reliability issues prevent significant scaling of tunnel oxide thickness, maintaining need for high voltages. The request for higher density devices could prove not economically viable relying only on this memory cell. No surprise then, the EEPROM-dominated scenario is being questioned by new entries.

Until a few years ago it was believed the ferroelectric memories (FeRAMs) could play a major role. These devices were particularly attractive when one was considering their reduced power consumption and their speed during the write cycle: contactless cards were the ones supposed to use FeRAMs first. As an additional benefit, FeRAMs have the possibility of operating almost like EEPROM and RAM cells, with a consequent simplification of process architecture. However, industrialization and scaling have been proved to be very difficult, leaving FeRAMs confined to low density tags manufactured in 0.35 μm (or less advanced) technology.

More interesting (and potentially able to grab a large share of the market) is the introduction of Flash NOR. Although devices using “variations” of Flash cells are already commercialized, the “standard” version is not yet widely used because of complications at design level (complexity of “Algorithms and State Machines” to control the modify operations, difficulty in ensuring byte alterability of the memory content). On the other hand Flash is much easier to scale, requires lower voltages to operate, and is really the only viable option for cards requiring more than 256 kBytes. Drawbacks with Flash NOR are related to the power consumption during the write operations and the need for bulk (or sector) erase. As a matter of fact the best granularity achievable is at the word level.

A desirable feature (that could become the key factor) is the possibility of using the Flash to replace part of the ROM. This will allow the card manufacturer an additional degree of freedom when planning improvements and/or new releases of the operating system, confining the “stable” code within the ROM and using the Flash to store the “variable” content as well as the customizations (now stored in the EEPROM). The dominance of EEPROM will probably be over when the migration from 130 to 90 nm technology generation will be complete. Device manufacturers could make the shift towards Flash faster if they will be willing to offer a single memory cut at the same cost as the low density EEPROM-based products.

Flash NAND are of interest when two different chips are put together in the same SIM card. The resulting device can store Gbit of data, although not in secure mode.

Looking further ahead we may consider Phase Change Memories (PCMs) as the only other valid option today (MRAMs cost/performance look unfavourable). Although there is no volume production yet for such devices, the investigations carried out by several large Integrated Device Manufacturers and other institutions have demonstrated that PCMs have nice scaling properties, can be integrated rather easily with advanced CMOS logic, and can be arranged in arrays with Byte granularity. These highly desirable features could make PCMs battle for market share vs. EEPROMs and Flash already at the 90 nm technology node.

5. Conclusions

Smart Cards are a class of products that we encounter in our daily life in a variety of applications: security and portability are the key features, achieved with proper integration of software and hardware components.

The commercial success of these products has enticed many players to enter the market, where the main entry barriers are the capability of developing high quality SW, the ability to make it compatible with HW platforms in order to meet the high security level

demanded by the market. The management of personalization data that represent the uniqueness of each Smart Card is another key ingredient for success.

At the device level the barriers of entry are higher. The availability of a technology that successfully integrates CMOS logic with different types of memories comes first, immediately followed by the system know how and then the ability to generate designs that strike the best bargains between so many different parameters to optimize.

The ability to shrink EEPROM cells down to 130 nm technology has ensured dominance for devices that use this memory type to store customization code and user parameters. The future, starting from devices in 90 nm technology, will see a shift towards other solutions, with Flash NOR already gaining ground and PCMs as the long term leading candidate.

SESSION B

Phase Change Memory

Phase-Change Memory – Present and Future

Hsiang-Lan Lung[a], Matt Breitwisch[c], Thomas Happ[b], and Chung Lam[c]

IBM Qimonda Macronix PCRAM Joint Project

IBM T.J. Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, N.Y. 10598, USA

[a]Macronix, [b]Qimonda, [c]IBM

Tel:+1-914-945-3828, email: lunghl@us.ibm.com

Abstract

This paper reviews the current development status of Phase-Change Memory (PCM), discusses an advanced scaling demonstration of this technology, presents a prospective view of future possible applications, and discusses the development road map for Phase-change Memory.

Introduction

Since the first PCM paper was published in 1968 [1], advancements towards realizable applications where slow to come. Thirty years passed before new possibilities for this technology were rediscovered after the invention of the CDROM and advancements in phase-change material. The lead development in phase-change memory was published 2001 [2]. This paper marks the beginnings of the phase-change memory development competition. Within the next five years, hundreds of papers and patents were published. The motivations driving this competition and the reasons why vast resources are being devoted to developing PCM have become clear. In comparison to the alternative well established non-volatile memories, PCM is scalable, has lower voltage operation, has lower power consumption, has lower fabrication costs, and has a fast programming speed. These demonstrated benefits are convincing the industry that this technology has a chance to replace NOR Flash, NAND Flash, and DRAM.

In this paper, we first discuss the current development status of the PCM technology. Next, the most advanced scaling demonstration results are reported. Finally, we discuss possible future applications and the development roadmap for PCM.

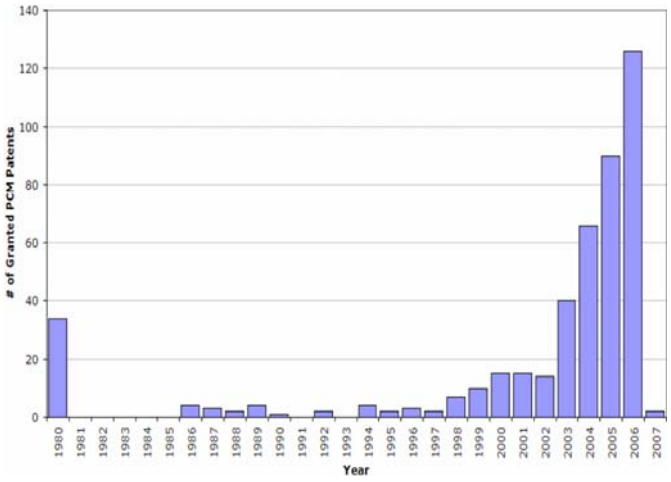


Fig. 1 US granted PCM patents distribution from 1990 to Jan 2007.

Current Status of Phase-change Memory

Patent distribution: Fig. 1 shows the distribution of PCM related patents from 1980 through the beginning of 2007. The amount of PCM related patents begins to increase exponentially between 2002 and 2006, indicating that PCM is heavily pursued within the semiconductor industry.

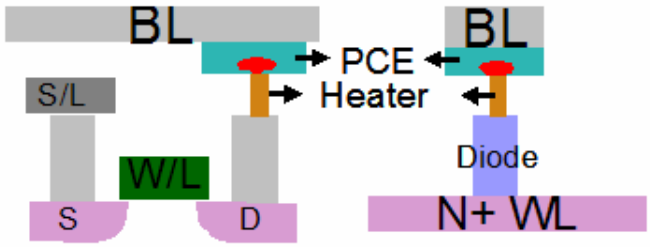


Fig. 2 Mushroom type PCM memory cell structures. Left: FET select structure. Right: diode select structure.

Cell Structure: Most of the major players are currently using a "mushroom" structure as their memory cell which is comprised of a bottom heater in contact with the phase-change material. Fig. 2 shows the mushroom cell structures with Field Effect Transistor (FET) or diode as the select device. The FET option provides a higher on/off ratio, a simpler integration scheme, but a larger memory cell size in order to supply the necessary programming current, whereas the diode select device can provides higher programming current which leads to smaller memory cell size.

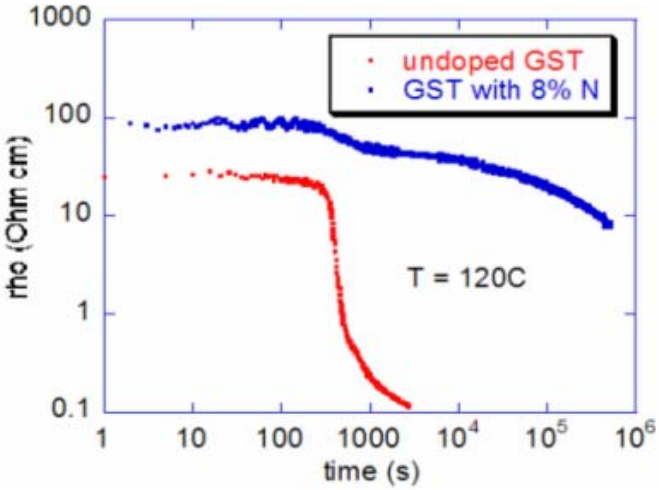


Fig. 3 Resistivity vs. annealing time for undoped and doped GST.

Material: $Ge_2Sb_2Te_5$ (GST) is the industry standard phase-change material. The material properties are improved by doping the $Ge_2Sb_2Te_5$ with nitrogen. The incorporation of nitrogen into the material inhibits grain growth thereby

increasing the retention of the amorphous state (Fig. 3) and also increases the resistivity of the material resulting in a reduction of the crystalline-to-amorphous programming current.

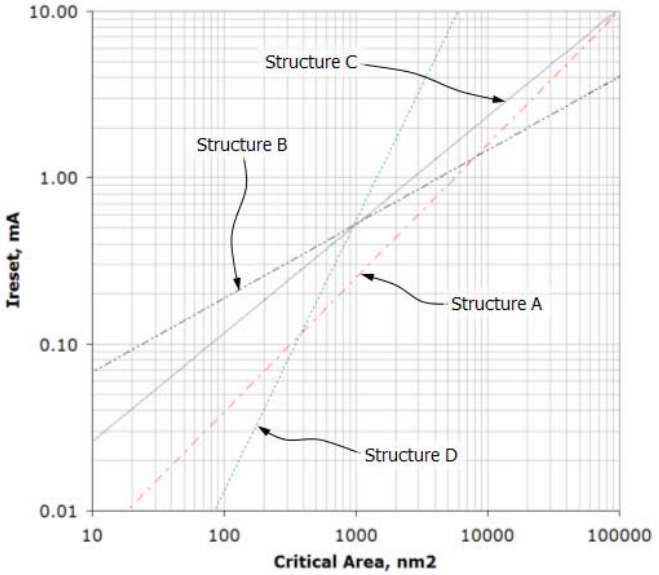


Fig. 4 RESET current vs. critical area for different memory cell structures.

Programming current: The major challenge for PCM technology is to reduce the RESET current of the memory element. The RESET current limits the size of the driving device (either diode or FET), as the driving device must be large enough to support the programming of the memory element. Many different cell structures have been proposed [3, 4] to reduce the RESET current. The RESET current can be reduced by decreasing the bottom heater dimension and by decreasing the volume of the region of the phase-change material which undergoes switching. Other methods of reducing the RESET current include increasing the resistivity of the bottom heater and incorporating additional dopants into the phase-change material [5]. Fig. 4 shows the RESET current vs. critical area of four different cell structures. Each cell structure exhibits a different dependence on the critical dimension, indicating that the choice of cell structure is critical and strongly influences how the RESET current will scale for the memory element in future technologies.

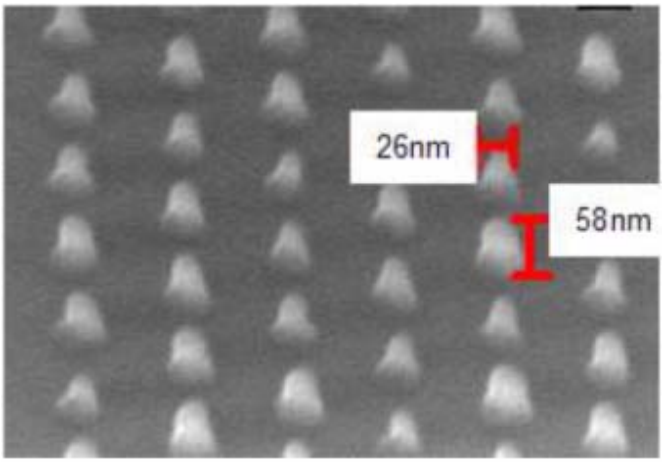


Fig. 5 Small phase-change material islands fabricated by e-beam lithography.

Phase-change Memory Scaling

One of the most attractive properties of PCM is the scalability of this technology. Fig. 5 shows nano-scale GST islands fabricated by e-beam lithography. X-ray diffraction results show that the phase-change material can crystallize with the islands having a diameter of approximately 25nm.

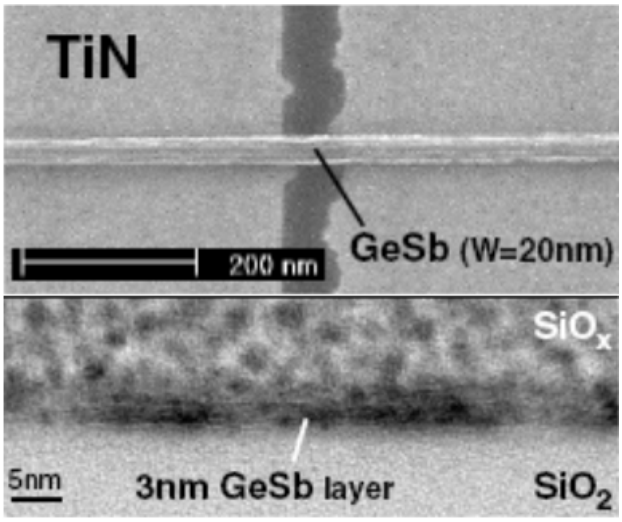


Fig. 6 TEM of a 20nm width, 3nm thick GeSb nano-line on top of a pair of TiN bottom electrodes forms the "bridge" cell.

Fig. 6 shows the smallest phase-change device reported to date. A very thin (3nm) and narrow (20nm) doped GeSb line was fabricated above of a pair of TiN bottom electrodes. The memory element can be switched with RESET currents as low as 80uA and shows SET-RESET cycling up to 10^5 without fail. Further experimental results show that the scaling of this structure is limited by the e-beam lithography and not by the memory structure itself.

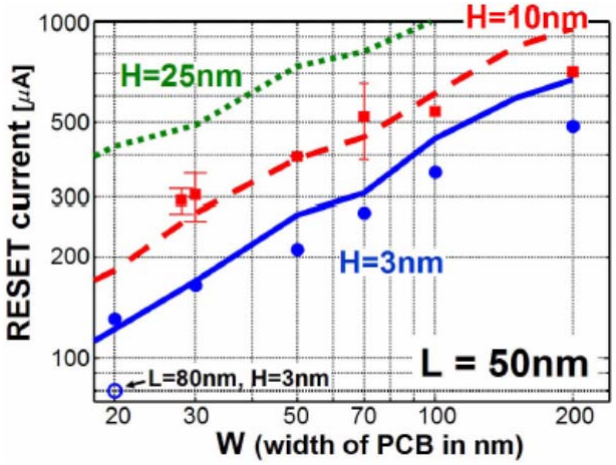


Fig. 7 RESET current vs. line width of a bridge-like memory cell.

Fig. 7 shows the RESET current as a function of the phase-change material width for different values of the thickness, with a bottom electrode separation of 50nm. Ultra-thin phase-change material was investigated in order to further understand the scaling limitation of phase-change material. As seen in Fig. 8, GeSb thin film thickness thinner than 1.1nm still shows crystallization after annealing. The fcc phase will disappear but the hcp phase remains. This indicates the GeSb can be scaled down to 3nm and continue to change phase.

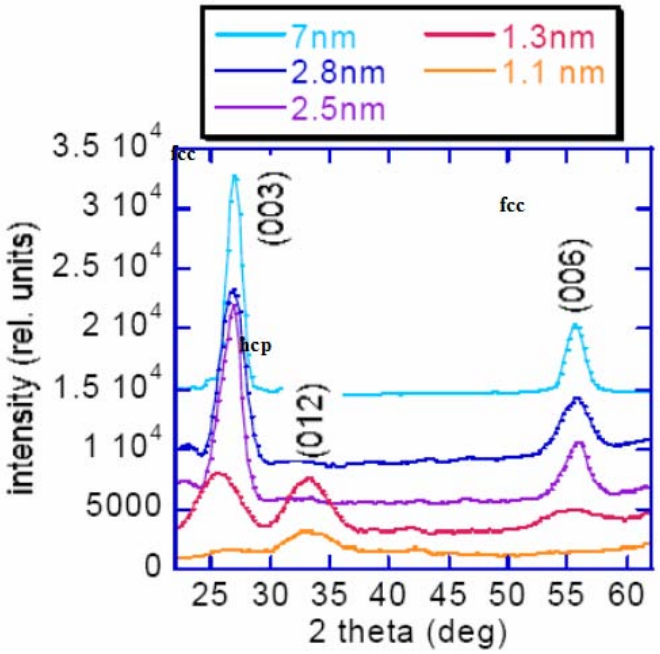


Fig. 8 XRD theta-2theta scans for thin GeSb films after 430ºC annealing.

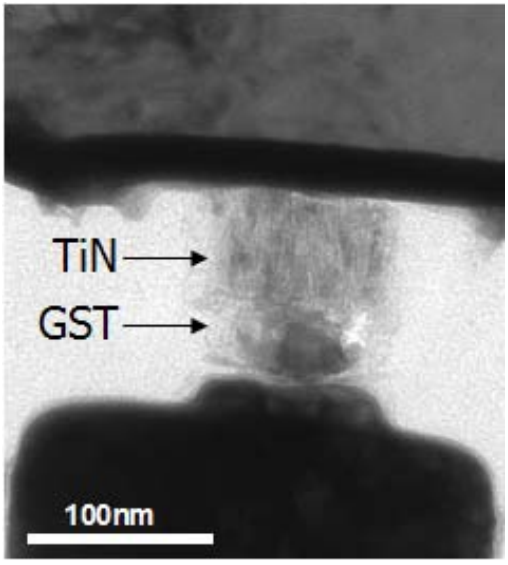


Fig. 9 One mask only needed phase-change memory cell.

Future of Phase-change Memory

Application: Stand-alone and embedded NOR Flash probably will be the first technology replaced by PCM. For embedded Flash memory, typical cell size is around 15~20F^2. PCM cell with FET select device has similar cell size at 90nm node [6]. PCM does not need high voltage devices, requires less complex process, and has much faster programming time. Fig. 9 shows a one-mask only PCM cell which can provide a much lower cost solution for embedded Flash application. The top electrode and phase-change material are deposited after the front-end of the line processes are completed and before the back-end of the line processes begin. The only extra mask is used for patterning the phase-change memory element. Stand-alone NOR Flash faces serious scaling issues beyond 45nm node. Diode select PCM cell [7] has demonstrated cell size smaller than 6F^2 at 90nm capable of supplying greater than 1.5mA for RESET operation. The diode select PCM cell is also shown to scale beyond 22nm.

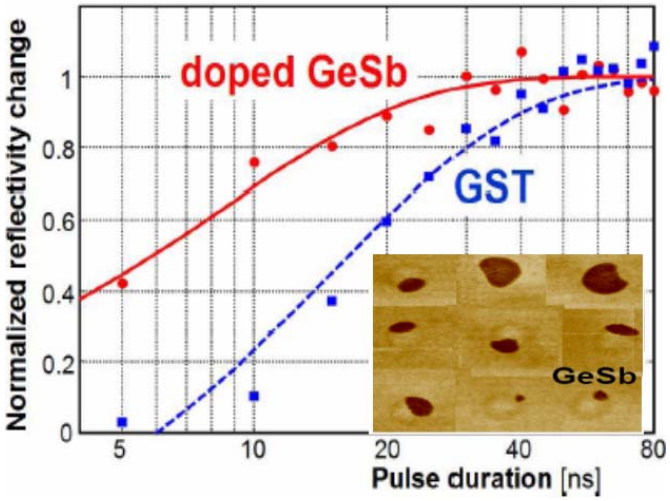


Fig. 10 Normalized reflectivity change as a function of optical pulse duration for doped GeSb. The inset shows the AFM imaging of partially crystallized doped GeSb optical spots.

It will be challenging for PCM to compete with DRAM due to DRAM's fast programming speed and high write endurance requirement. The SET (transition from amorphous to crystalline) performance of the programming operation is the issue here. By changing the phase-change material from GST to GeSb, the phase-change SET speed can be improved from 16ns to below 5ns with respect to 0.4 reflectivity change as shown in Fig. 10 in an optical demonstration. Hence, by utilizing a faster phase-change material PCM has a chance to meet the performance requirements of DRAM, however, the 10^{15} cycling requirement is still a large hurdle for PCM to overcome.

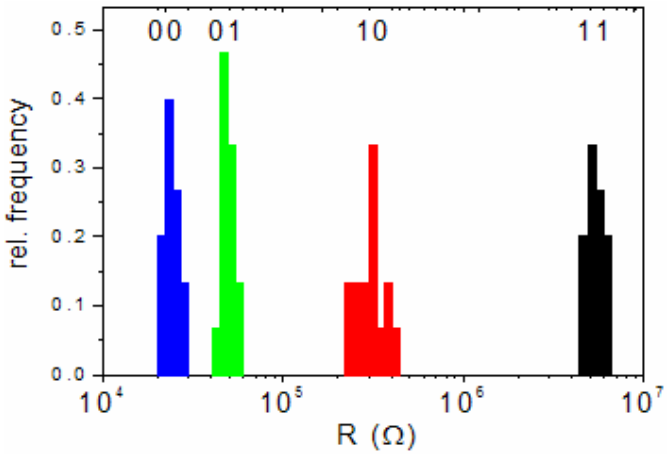


Fig. 11 4 level 2-bits-per-cell demonstration at set state.

NAND Flash has achieved the highest density and smallest memory cell size compared to all other semiconductor memories. Furthermore, the predicted scaling limit for NAND Flash is expected to extend beyond the 25nm node. For PCM to compete with NAND FLASH, the cell size needs to be smaller then 4F^2, which requires the use of a diode as the driving device. A diode of this cell size in the 25nm node requires the RESET current of the memory element to be less than 150uA. Another "must have" feature in order for PCM to compete with NAND is the capability of

multiple bits per cell operation. Fig. 11 shows a 2-bits-per-cell demonstration of a PCM cell. Four different resistance states are programmed with only one-shot program condition for each state. The key for a successful Multiple-Level-Cell (MLC) PCM is the ability to control the resistance distribution of each state.

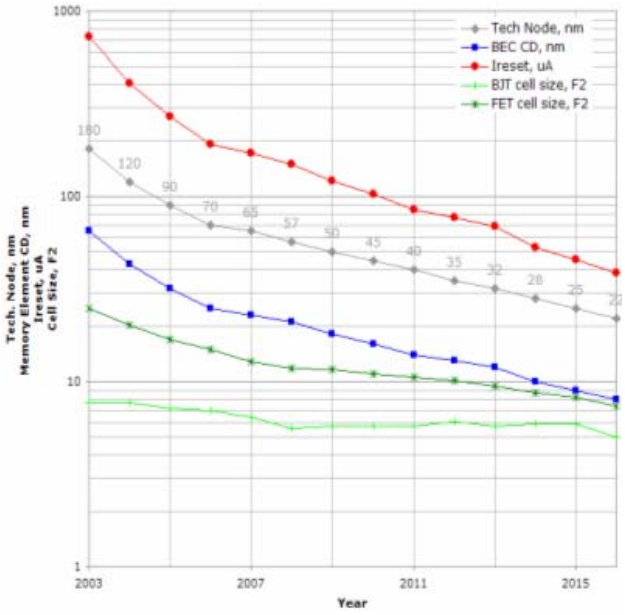


Fig. 12 PCM technology road map.

Technology development roadmap: Fig. 12 shows the estimated PCM technology roadmap. The diode (BJT) select array will always have a smaller cell size than the FET select array. 40nm node will be achieved around 2011 and will have a cell size of approximately $5F^2$ with a bottom heater diameter of approximately 15nm.

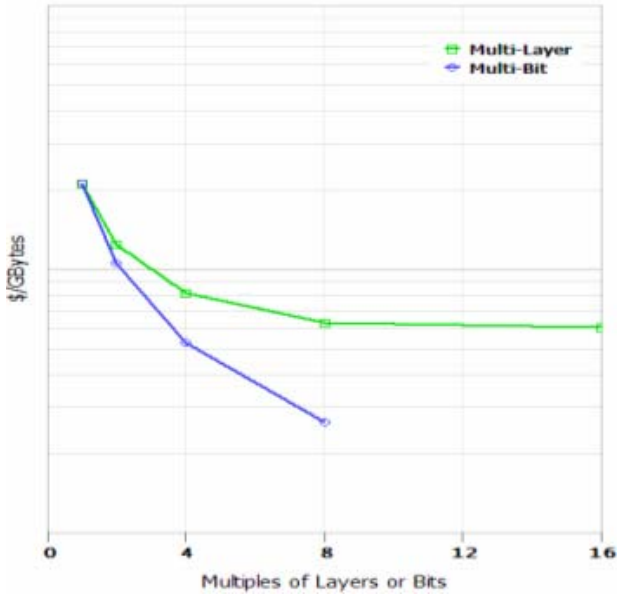


Fig. 13 Giga-bits cost comparison of the MLC approach and the multiple-layer approach for future PCM technologies.

There are two way to further reduce the memory cell size: one is MLC, and another is to have multiple layers of memory elements stacked on top of one another. Fig. 13 compares the benefits of these two approaches. Its shows the multi-layer's benefits will saturate after 8 layers, whereas the cost of MLC approach continues to decrease with increasing number of bits. These results further emphasized the importance of MLC operation for PCM in the future.

Memory Element Structure			
Contact CD		Phase Change CD	
Isolation/Select Device			
Bipolar		FET	
Bipolar	Diode	Planar	3D
Material Engineering			
Material Search			
Doping		Resistivity modulation	
		Crystallization Temp	
		Melting point	

Table 1 Main PCM future research and development items.

Table 1 shows the key research and development items. The structure of the memory element is crucial. The size of the heater and phase-change memory CD needs to be reduced while maintaining good control to minimize variations; this is necessary to achieve tight RESET and SET resistance distributions. Development of a diode or 3D transistor is needed for increasing the driving current capability, which ultimately limits the cell size for a given memory element. Finally, materials research needs to continue to study the effect of doping on resistively and re-crystallization which in turn affects the retention characteristic.

Conclusion

PCM has been receiving a great deal of attention in recent years as it provides a possible way to continue scaling for and to reduce the cost of future semiconductor memory technologies. Scaling of the PCM element down to at least 20nm appears to be feasible without showing serious problems. Decreasing the heater and phase-change element CD, developing high current support diode, developing MLC operation, and developing advanced phase-change materials are the most important ways to make PCM successful.

Acknowledgements

Authors would like to thank the PCRAM joint project team members and the expert processing support from the Microelectronic Research Line at IBM Watson Research Center. Valuable discussions with W. Gallagher, R. Liu, and G. Mueller are gratefully acknowledged

References

[1] S. R. Ovshinsky, "Reversible Electrical Switching Phenomena in Disordered Structures," Phys. Rev. Lett. Vol. 21, 1968, p. 1450.
[2] S. Lai, T. Lowrey, "OUM - A 180 nm nonvolatile memory cell element technology for stand-alone and embedded applications," 36.5, IEDM Tech. Dig., 2001.
[3] Y. C. Chen, et. al., "Ultra-Thin Phase-Change Bridge Memory Device using GeSb," 30.3, IEDM Tech. Dig., 2006.
[4] T. D. Happ, et. al., "Novel One-Mask Self-Heating Pillar Phase-change Memory," Symp. VLSI Tech.. 2006.
[5] W. Czubatyj et al, "Current Reduction in Ovonic Memory Devices," E*PCOS 2006.
[6] S. J. Ahn et al, "Highly Reliable 50nm Contact Cell Technology for 256Mb PRAM," Symp. VLSI Tech. 2005.
[7] J.H. Oh et al, "Full Integration of Highly Manufacturable 512Mb PRAM based on 90nm Technology," IEDM Dig. 2006.

Heater Electrode Engineering and Analysis of Series Resistance in Phase Change Memory

C.W. Jeong, D.H. Kang, D.W. Ha, Y.J. Song, J.H. Oh, J.H. Kong, J.H. Yoo, J.H. Park, K.C. Ryoo, D.W. Lim, S.S. Park, J.I. Kim, Y.T. Oh, J.S. Kim, J.M. Shin, Jaehyun Park, Y. Fai, Y.T. Kim*, G.H. Koh, G.T. Jeong , H.S. Jeong, and Kinam Kim

Advanced Technology Development Team 2, Samsung Electronics Co., Ltd, Yongin-City, Korea
*CAE Team, Samsung Electronics Co., Ltd, Yongin-City, Korea
chris.jeong@samsung.com

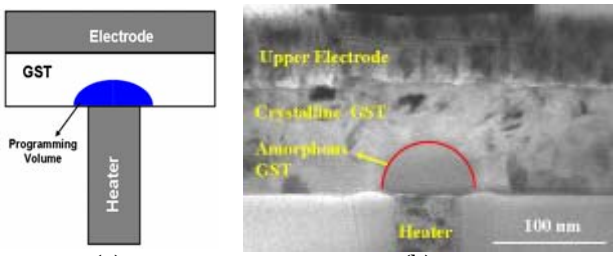
Abstract

We evaluated the limit of scaling bottom electrode contact(BEC) heater size and high resistivity heater to reduce programming current. It was found that the resistivity of heater should be increased for reducing programming current below the heater size of about 50nm without any undesirable increase of resistance of the crystalline state(SET state, Rset). It was shown in the numerical simulations that the dissipated heat loss through BEC during melting GST was decreased in the increase of resistivity of heater. In addition, we analyzed the resistance components contributing to the total set resistance. It was observed that the undesired sharp increase of Rset as the BEC size decreases below 50nm was attributed to the resistance component of GST-BEC interface. In the case of high resistivity heater, the contributions of both incomplete crystallization and heater itself were enhanced.

1. Introduction

The phase change random access memory (PRAM) has been investigated because of its non-volatility, relatively high endurance, and good scalability. It was reported that 512Mb PRAM was successfully developed by adopting 90nm diode technology.[1] Since it is essential to reduce the programming current (or reset current) for the high density PRAM, there are several structural and compositional efforts in the storage module.[2-4] In addition, it is very important to maintain stable cell uniformity for reliable operation for high density PRAM products. In order to control the cell uniformity within very small process variation, novel process technology was proposed. [5]

In this work, we investigate the scalability of BEC heater size reduction and evaluate the highly resistive heater electrode for improving writing current and set resistance, respectively. Figure 1 shows schematic structure of the test device and TEM picture of cell storage module. We prepared test samples with planar type BEC, ranging from 32 nm to 85 nm and with different heater electrode of a designed resistivity from 0.2 to 4 mΩcm. Typical R-I curve was obtained by measuring the resistance sweeping the height of the voltage pulse through a load resistance with the pulse width of 500 ns. With the aid of numerical simulations, we evaluate the heating efficiency of heater electrode with higher resistivity. Finally, we analyze the resistance components contributing to the total set resistance(minimum resistance) for individual cases of the heater size scaling and the higher resistivity electrode for scaling programming current.



(a) (b)

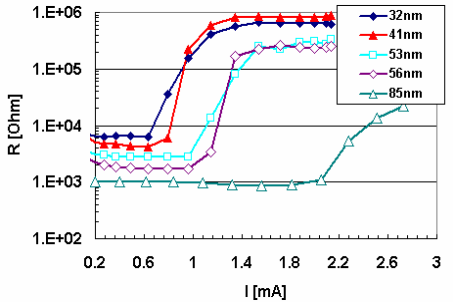
Fig. 1: (a) Schematic structure of the test device (b) TEM picture.

2. Heater Electrode Size Scaling

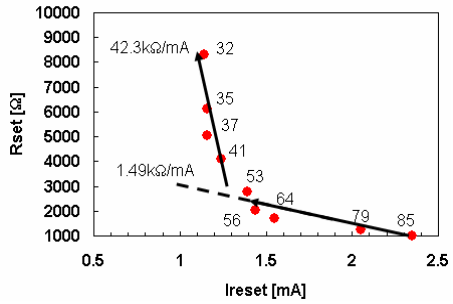
The scalability of the programming current as a function of heater size has been experimentally presented in Figure 2(a). Ireset and Rset are defined as the currents where the resistance reaches 100kΩ and the minimum resistance, respectively. Reduced contact size increases the local current density and joule heating in GST, leading to reduce not only the switching current, but also the dissipated heat flux through heater itself. A similar result was numerically obtained in the same structure.[6]

However, the reduced contact size brings up undesirable Rset increase. Figure 2(b) shows the correlation between Ireset and Rset. The general trend of Ireset and Rset can be described by Ireset $\propto$ L and Rset $\propto$ 1/L, where L is the contact size. This correlation curve shows the effectiveness of a item for the programming current reduction. For further discussion, we defined parameter S as the amount of increase of Rset per 1mA Ireset reduction.

In the range of 32~41nm contact size, S significantly increases to 42.3kΩ/mA, while S was 1.49kΩ/mA in the range of 53~85nm contact size. Therefore, it is impossible to reduce the reset current without any appreciable increase of Rset only by controlling the contact size below around 50nm contact size.



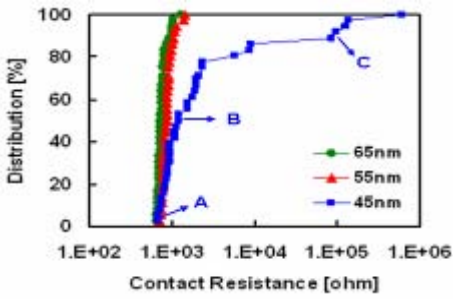
(a)



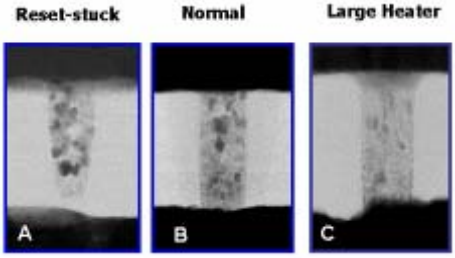
(b)

Fig. 2: (a) Typical R-I curves of test devices with different contact size (b) I_{RESET} vs. R_{SET} curve.

The saturation behavior of Ireset might be attributed to the following reasons First, the smaller contact size, the smaller programming volume, the larger surface area-to-programming volume($\propto$ 1/L), therefore the more dissipated heat loss through the surface of programming volume and the less efficiency of joule heating. Second, the smaller contact size, the larger surface area-to-heater volume ratio between heater and dielectric, therefore the more dissipated heat loss through the surface and the less efficiency of joule heating.



(a)



(b)

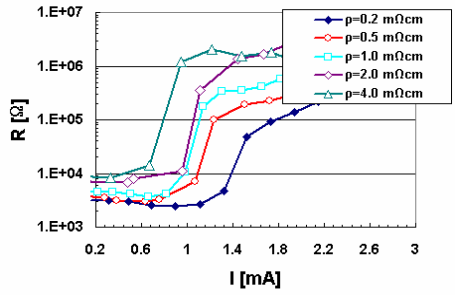
Fig. 3: (a) BEC-Top Electrode contact resistance distribution with different contact size (b) TEM analysis for the failing point.

In Figure 3(a), three curves show contact resistance distributions for three different contact sizes. In order to avoid the GST-BEC interfacial resistance effect, we made samples without GST deposition process to monitor BEC contact process issue alone. Below 45nm, the distribution of contact resistance becomes broad and can be hardly controlled. The fail bits were analyzed by TEM as shown in Figure 3(b).

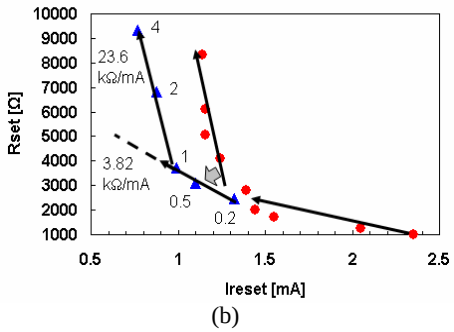
3. Heater Electrode Resistivity Engineering

The heater electrode with higher electrical resistivity was used to maintain the contact size over about 50nm for avoiding the BEC process issues, and, at the same time, to reduce the programming current. The test samples with different electrical resistivities of 0.2, 0.5, 1, 2, and 4 mΩcm were used and the contact size of all samples were controlled at around 55nm in order to avoid undesirable Rset increase as previously discussed in Figure 3. Typical transition curves(R-I curves) as a function of the different heater resistivity was presented in Figure 4(a). In the 1 mΩcm-heater electrode, we have obtained functional cells operating at 1.0mA without major increase of Rset.

In order to determine the effectiveness of Ireset scaling by the higher resistivity heater, Figure 4(b) shows Ireset-Rset curve in comparison with that of heater size scaling as already presented in Figure 2(b). It was observed that the increase in heater reisitiviy was effective up to 1 mΩcm without much penalty of Rset, which is indicated by the S parameter value of 3.82kΩ/mA. Further higher resistivity over 2mΩcm caused an unwanted Rset increase, which is also indicated by high S value of 23.6kΩ/mA.



(a)



(b)

Fig. 4: (a) Typical R-I curves of test devices samples with different heater resistivity(ρ) of 0.2, 1, 2, and 4 mΩcm (b) corresponding I_{RESET} vs R_{SET} in comparison with the case of heater dimension scaling.

Figure 5 shows the results of numerical simulation for the heater electrode with various electrical resistivity.

Material parameters used in the simulation were summarized in the Table 1, where thermal conductivity and resistivity were related with Wiedemann-Franz law. It was found that the dissipated heat loss during melting the GST, i.e. reset operation, was significantly reduced in the case using higher resistivity heater electrode. It was believed that the high heater resistivity resulted in more joule heating in the region of heater electrode, thus giving rise to the reduction of dissipated heat loss through the heater. It can be easily observed by temperature profile of cells with different resistivity of 0.2 and 2mΩcm during programming in the inset figures.

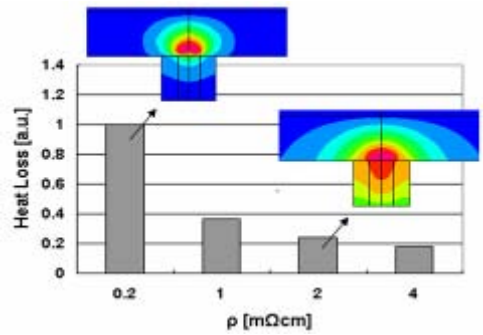


Fig. 5: Dissipated heat flux comparison of test devices with different heater resistivity (ρ) of 0.2, 1, 2, and 4 mΩcm. Inset figures depict temperature profiles of test devices with heater resistivity (ρ) of 0.2 and 2mΩcm.

Table 1: Material Parameters used in the numerical simulation.

Material	Electrical resistivity ρ (Ω-cm)	Density (g/cm³)	Thermal Conductivity κ (J/cm-K-s)	Specific Heat (J/cm³-K)
GST (xtal)	8.00E-04	6.2	0.018	1.2
Heater	0.2E-3~4E-3	5.4	0.13~0.0065	3.235
SiOx	1.00E+14	2.33	0.014	3.1
SiNx	1.00E+14	3.44	0.16~0.33	0.58

4. Analysis of Series Set Resistance

It is essential to analyze the resistance components contributing to the total set resistance for further scaling down programming current without the penalty of Rset increase. As illustrated in Figure 6, the total set resistance was assumed as

$$R_{TOTAL} = R_{HEATER} + R_{INTERFACE} + R_O$$

where R_{HEATER} is the resistance of heater itself, $R_{INTERFACE}$ consists of R_{SP} and R_C, where R_{SP} is the spreading resistance above the heater and R_C is the contact resistance, and R_O accounts for the resistance of top electrode contact, which is assumed to be negligible due to the large contact area. R_{SP} includes R_{PGM} and R_R, where R_{PGM} is GST resistance due to the incomplete crystallization and R_R is residual GST resistance due to remaining crystalline GST

Thermal crystallization was carried out to exclude incomplete crystallization which might be occurred during the electrical crystallization process accompanying threshold switching. Annealing process was executed for 1 hour at 250℃ to guarantee the complete crystallization. In order to experimentally separate the R_{HEATER} from $R_{INTERFACE}$, the test samples without GST deposition process were made.

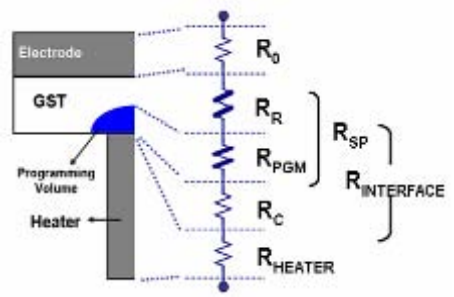


Fig. 6: Schematic representation of cell structure and series resistance components contributing to the total set resistance

Figure 7 shows the relative contribution of resistance components to the total resistance as a function of heater size. It was clearly shown that $R_{INTERFACE}$ significantly increases with decreasing the contact size and therefore relative contribution of $R_{INTERFACE}$ to the total resistance increases. In this geometry, $R_{INTERFACE}$ can be approximated by [7]:

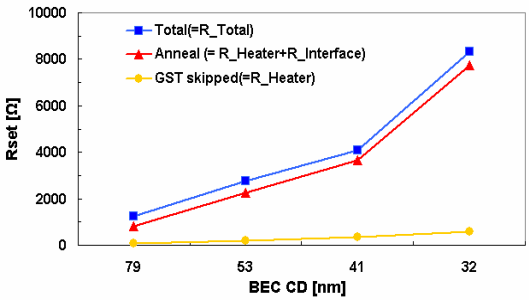
$$R_{INTERFACE} = R_C + R_{SP} = \frac{\rho_C}{\pi (L/2)^2} + \frac{\rho_{GST}}{\pi L} \arctan\left(\frac{4t}{L}\right) \quad (1)$$

where ρ_C is contact resistivity, L is contact size, ρ_{GST} and t are resistivity and thickness of GST, respectively. From Eq. (1), it was found that R_C contributes nearly 90% to the $R_{INTERFACE}$. From a plot of R_C vs. $1/\pi(L/2)^2$, ρ_C was 5.0×10^{-8} Ω cm^2. The R_{HEATER} component behaves in the expected ways by $R_{HEATER} \propto 1/L^2$, while R_{PGM} caused by incomplete crystallization was not significantly changed as the contact size decreased.

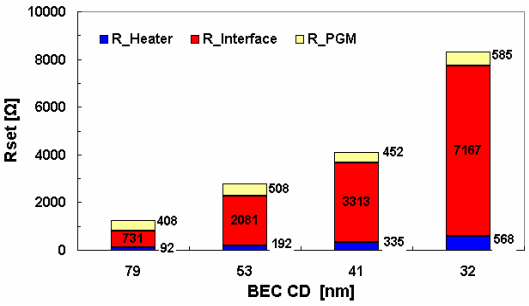
Figure 8 shows relative contribution of resistance components to the total resistance as a function of heater resistivity. In the contrary to the case of heater size scaling, the portion of $R_{INTERFACE}$ contribution to the total resistance became weaker. On the other hand, R_{GST} and R_{HEATER} portions significantly increase with the higher heater resistivity. (and therefore relative contribution of both R_{GST} and R_{HEATER} to the total resistance increases.) From Eq. (1), ρ_C increased from 4.1×10^{-8} to 5.3×10^{-8}Ω cm^2 with the heater resistivity. It was believed that the barrier height (Φ_B) between heater and GST(p-type) increased with heat resistivity, and thereby ρ_C increased because ρ_C was given by[8]:

$$\rho_C = \partial V / \partial J\big|_{V=0} \propto e^{q\Phi_B / E_0}$$

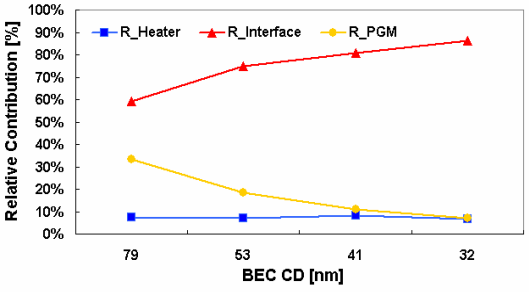
An optimized electrical crystallization method and localized high resistive layer at BEC-GST interface could reduce unwanted increase of both R_{GST} and R_{HEATER}, while $R_{INTERFACE}$ was considered hard to control.



(a)

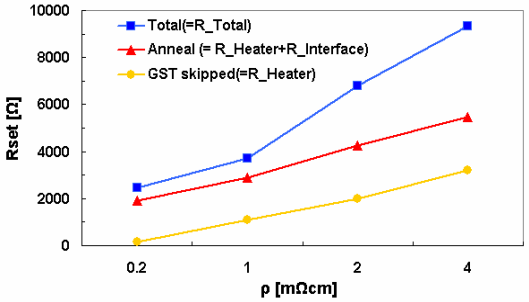


(b)

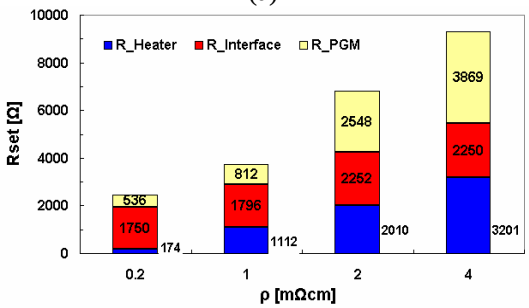


(c)

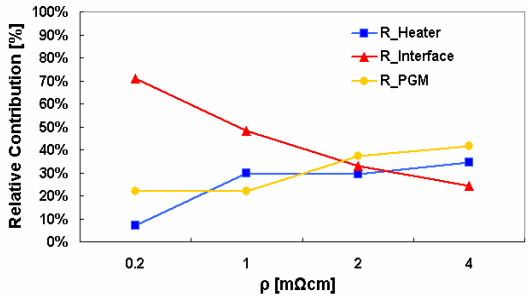
Fig. 7: Relative contribution of resistance components to the total resistance as a function of heater size. (a) Resistance vs. heater size for the case of electrical crystallization, thermal crystallization, and top electrode-heater contact resistance in GST deposition skipped process, (b) resistance of each components obtained from(a), and (c) relative contribution of each components.



(a)



(b)



(c)

Fig. 8: Relative contribution of resistance components to the total resistance as a function of heater resistivity(ρ). (a)Resistance vs. heater resistivity for the case of electrical crystallization, thermal crystallization, and top electrode-heater contact resistance in GST deposition skipped process, (b)resistance of each components obtained from(a), and (c)relative contribution of each components

5. Conclusions

In order to reduce programming current and develop high-density PRAM, we evaluated the limit of scaling heater size and the feasibility of high resistivity heater. Also, we analyzed the resistance components contributing to the total set resistance. It was found that high resistivity heater should be employed to reduce programming current and avoid undesirable increase of Rset, since the higher resistive heater electrode was more efficient for joule heating. It was observed that the sharp increase of Rset below the heater size 50nm was attributed to the resistance component of GST-BEC interface. In the case of high resistive heater, the contribution of resistance component of both incomplete crystallization and heater itself increased. Finally, we concluded that heater engineering was more promising way than heater size reduction for programming current reduction because an optimized electrical crystallization could reduce unwanted increase of R_{PGM} and localized heater engineering could maintain highly efficient joule heating without much increase of R_{HEATER}.

References:

[1] J.H. Oh et al., IEDM Tech. Dig.,49 (2006).
[2] Y.N. Hwang et al., IEDM Tech. Dig.,893 (2003)..
[3] N. Matsuzaki et al., IEDM Tech. Dig.,758 (2005).
[4] C.W. Jeong, et al, *Proceedings of NVSMW*, 28 (*2004).*
[5] S.J. Ahn et al., Symposium onVLSI Tech. Dig., 18 (2005).
[6]Y. T. Kim et al, Proceedings of SSDM, 244 (2004).
[7] R.H. Cox and H. Strack, Solid-State Electron. 10, 1213 (1967).
[8] A. Y. C. Yu, Solid-State Electron. 13, 239 (1970).

Effects of the crystallization statistics on programming distributions in phase-change memory arrays

D. Mantegazza[a], D.Ielmini[a], A.Pirovano[b], A.L.Lacaita[a], E.Varesi [b], F.Pellizzer[b], and R.Bez[b].

[a] DEI, Politecnico di Milano, piazza L. da Vinci 32, 20133 Milano, Italy, mantegaz@elet.polimi.it
[b] STMicroelectronics, FTM, Advanced R&D, NVMTD, Agrate Brianza, Italy

Abstract

For a reliable operation of phase change memory (PCM) arrays, the cell programming characteristic has to be carefully analyzed on the statistical level. The quenching operation was already found to critically control the reset distributions [1], but the impact of chalcogenide crystallization was never addressed. This paper shows a statistical characterization of quenching and crystallization characteristics for PCM arrays. A correlation between crystallization (at both high and low temperatures) and quenching behaviour of cells is found, allowing to describe the programming distribution uniquely in terms of the statistics of crystallization times.

1. Introduction

The phase change memory (PCM) is a non-volatile memory device, which relies on the phase change properties of a chalcogenide material, usually $Ge_2Sb_2Te_5$ (GST). The cell can be programmed in two logic states, corresponding to a highly-resistive amorphous phase (reset state) and to a highly-conductive crystalline phase (set state). The amorphous phase is achieved by melting (620°C) and rapidly cooling (quenching) a portion of the GST layer, while the crystalline state is obtained by rapid annealing below the melting point. Both programming operations are achieved via fast electrical pulses generating heat in the chalcogenide material, while read out is performed by low-voltage sensing of the cell resistance. A distinctive advantage of PCM is the large resistance window between the two states: the set-state resistance is usually few kΩ, while the reset state is above 1MΩ. At the array level, the resistance window is generally affected by a non-zero width of reset and set state distributions [1-3]. For a careful understanding and prediction of programmed distribution in large arrays, the statistics of the set/reset behaviour of cells have to be characterized.

This work provides a statistical investigation of the reset and set characteristics for PCM arrays. The role of the quenching time in reducing the reset-distribution tail is first quantitatively analyzed. Similarly, the impact of the set time in the transition from the reset to the set state is studied. Critical quenching and set times are defined and statistically characterized. A clear correlation is found between critical quenching and set times, indicating that the programming behaviour of the cell are uniquely controlled by the crystallization properties of the active chalcogenide material. The impact of the programming current on the critical set time is analyzed for typical and tail cells in the array. The correlation between fast and long-term crystallization kinetics is finally addressed.

2. Reset-state statistics

Fig. 1 shows measured resistance distributions for a set and a reset resistance distribution collected over a 2kb μTrench sub-array [3,4]. The reset distribution was measured after applying a non-optimized reset pulse with 60ns-long quenching time t_q, defined as the duration of the falling edge of the pulse. The set distribution displays no tail, while the reset distribution is affected by a low-resistance tail, thus indicating the presence of a sub-population of cells with non-optimal reset performance. As already pointed out in [1], the tightness of the reset distribution can be enhanced by reducing t_q: Fig. 2 shows the resistance distribution for the reset state

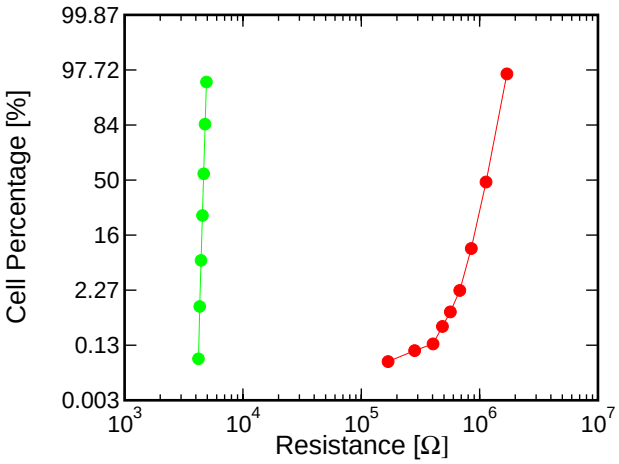


Fig. 1: Set (green) and non-optimized reset (red) resistance

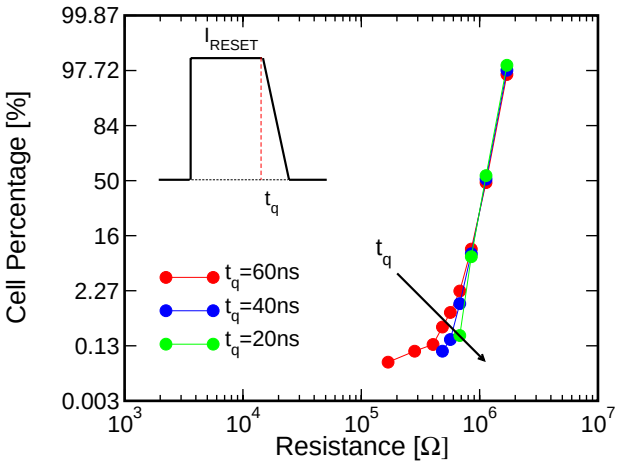


Fig. 2: Reset tail modulation by the quenching time t_q. Note that at t_q=20ns the reset tail disappears.

It should be noted that the change of t_q in Fig. 2 impacts only a small minority of cells in the reset tail, while the main distribution of reset resistance remains unaffected. This indicates that, although the value t_q=60ns is still sufficient for the majority of cells for the formation of a high-resistive amorphous phase, tail cells require a faster quenching process. For a more complete description of the quenching behaviour of cells in our sub-array, we characterized the resistance of any cell after the application of a reset pulse with fixed current (I_{reset}=1mA) and increasing t_q, in the range 2ns-2µs. Fig. 3 shows the measured resistance as a function of t_q for two cells, which were selected in the main (intrinsic) or in the tail region of the distribution in Fig. 1, respectively. For a sufficiently short quenching time (t_q<30ns), equal resistances are obtained for the intrinsic and tail cells. As t_q is increased, the tail-cell resistance drops to a relatively small value, which corresponds to a crystalline phase. This is because, at the atomic level, particles are allowed to experience several atomic configurations during the transition from the liquid to the solid phase, thus the thermodynamically-stable crystalline phase can be sampled and established [5].

From the quenching characteristic in Fig. 3, a critical quenching time $t_{q,crit}$ can be defined for any cell, as the minimum t_q for which a resistance lower than a resistance threshold R=$10^5\Omega$ is obtained. This parameter can be used to characterize the quenching properties of the PCM cell. A shorter $t_{q,crit}$ indicates that the cell is relatively weak with respect to crystallization during the quenching operation. Cells with lower $t_{q,crit}$ clearly belong to the reset tail in Figs. 1 and 2.

3. Crystallization dynamics

From the previous discussion, it is clear that cells in the reset tail have a faster crystallization dynamics during the transition from the liquid to the super-cooled liquid phase. For a deeper understanding of the crystallization behavior of the cell, we focused on the set

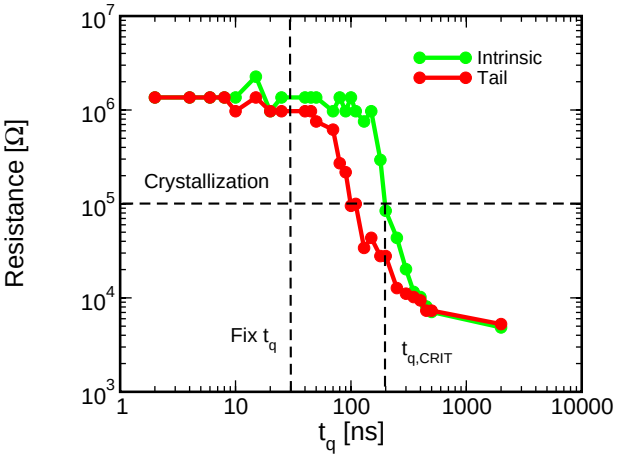


Fig. 3: Measured resistance as a function of t_q, after the application of a reset pulse with variable t_q. Choosing a crystallization threshold a critical quenching time $t_{q,crit}$ can be collected.

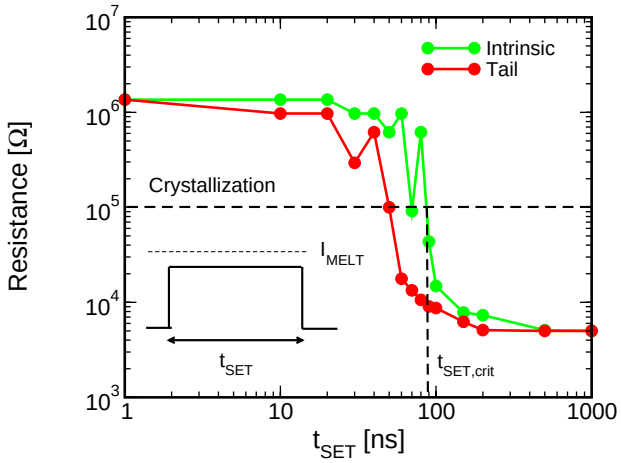


Fig. 4: Programmed resistance versus the box set pulse duration t_{SET}. A reset pulse with 2n falling edge is applied before each set pulse. Choosing a crystallization threshold a critical set time $t_{SET,crit}$ can be collected. The noise is due to crystalline grains nucleation statistics.

Fig. 4 shows the resistance measured after a set pulse of a duration t_{SET}, for the same intrinsic and tail cells of Fig 3. The cells were programmed with an optimized reset pulse (t_q=2ns), in order to have the same initial resistance level. A square set pulse below the melting current (see inset in the figure) was applied and the resulting resistance value was collected. For increasing t_{SET}, the crystalline fraction within the amorphous region increases, and the cell resistance consequently decreases. Similarly to Fig. 3, we define a critical set time $t_{SET,crit}$, which is the minimum set time for the resistance in the characteristic of Fig. 4 to drop below $10^5\Omega$. It is important to note that the tail cell displays a fast crystallization behavior in both Figs. 3 and 4, where both $t_{q,crit}$ and $t_{SET,crit}$ are lower than for the intrinsic cells. This demonstrates that both crystallization from the solid amorphous phase and crystallization during the quenching transition are controlled by the same physical parameter of the chalcogenide material.

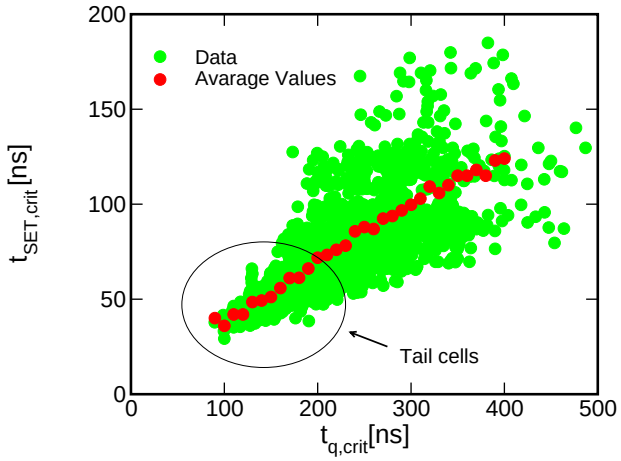


Fig. 5: Correlation between the critical quenching $t_{q,crit}$ and critical set time $t_{SET,crit}$ extracted over the 2kb statistics. Considering the average values a linear correlation can be noticed.

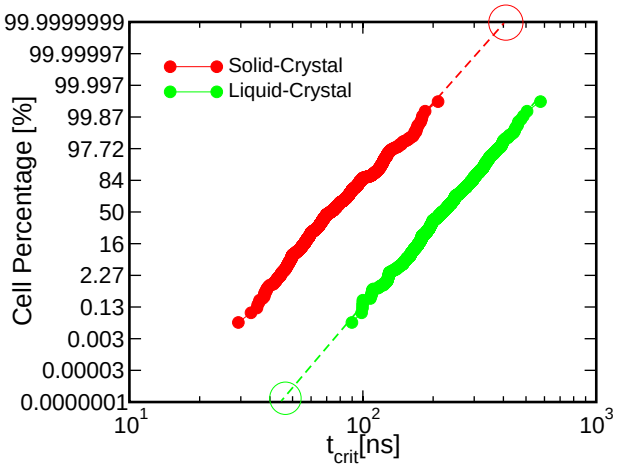


Fig. 6: Critical quenching $t_{q,crit}$ and critical set time $t_{SET,crit}$ extracted over the 2kb statistics.

4. Crystallization statistics

Quenching and set characteristics as in Figs. 3 and 4 were collected for any cell in the sub-array, allowing for a statistical analysis of the crystallization properties in our sample. Fig. 5 shows a scatter plot of and $t_{SET,crit}$ as a function of $t_{q,crit}$, clearly indicating an almost linear correlation between the two parameters. This result indicates that the quenching behavior of one cell can be, to a first order, predicted based on the setting characteristic, and *vice versa*. The location of reset-tail cells, occupying the short $t_{q,crit}$, short $t_{SET,crit}$ region, is shown in the figure.

Fig. 6 shows the measured cumulative distribution for $t_{q,crit}$ and $t_{SET,crit}$. An apparent log-normal shape of the distributions is obtained, demonstrating that *no anomalous crystallization behavior affects our cells*, at least within the statistical range investigated here (down to about 0.05%). Referring to the results in Fig. 1, it has to be pointed out that the measured reset tail does not correspond to any anomalous crystallization behavior, but simply results from the log-normal distribution of $t_{q,crit}$ (Fig. 6), combined with the highly non-linear quenching characteristics (Fig. 3). Namely, the log-normal distribution of $t_{q,crit}$ is distorted by the non linear R-t_q characteristics. The corresponding reset tail thus should not be viewed as a true anomalous tail in the array.

From the distributions in Fig. 6, a maximum t_q and a minimum t_{SET} for reliable operation in a 1Gb large array can be extrapolated. In particular, a minimum set time of 400ns can be estimated from the figure, which applies to a square-pulse, non-optimized set operation. It should be noted, in fact, that previous analysis has shown that a reliable set operation can be obtained by a more efficient set pulse of 130ns, by careful optimization of the set waveform [4]. On the other hand, a worst-case t_q of about 45ns is obtained from the extrapolation in Fig. 6.

4. Current-dependence of set operation

Given the strong correlation between $t_{q,crit}$ and $t_{SET,crit}$, a comprehensive characterization of the cell programming behavior can be obtained only by the set characteristics, with no need for an extensive quenching analysis. To provide a complete analysis of the set properties, however, also the dependence on the set current, which was kept constant in Fig. 4, has to be considered. The impact of the set current is shown in Fig. 7, where R-I curves were collected for increasing set time for the same cell. Before any current pulse, an optimum reset pulse (t_q=2ns) is applied in order to achieve a good amorphous volume. From the figure, a critical set current can be characterized, which allows for the cell resistance to drop below the threshold resistance of $10^5\Omega$ for a specific t_{SET}. In particular, from Fig. 7, the critical set current decreases for increasing t_{SET}: in fact, to provide the same crystalline fraction with a longer set pulse, a lower temperature will be required [6], corresponding to a lower current inducing Joule heating in the amorphous region of the cell.

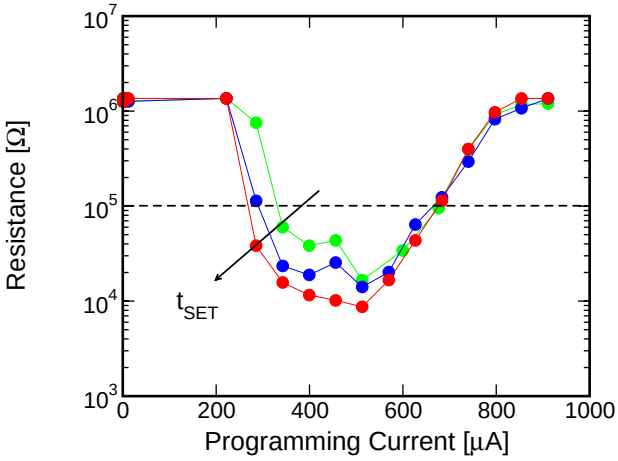


Fig. 7: Programming characteristics for three different programming times. In particular an optimized reset pulse with t_q=2ns is applied before each box programming pulse. The duration of the box programming pulse is t_{SET}.

Fig. 8 shows the scatter plot of critical set currents as a function of t_{SET}, resulting from measured characteristics as in Fig. 7 for an intrinsic and a reset-tail cell. As already pointed out, the current provides a figure of merit for the maximum temperature within the active volume of the cell. Thus, Fig. 8 is similar to the TTT (time-temperature-transformation) diagram which have been previously used to describe the crystallization kinetics in $Ge_2Sb_2Te_5$ for optical recording applications [6]. From Fig. 8, the reset-tail cell displays a faster crystallization for any set current applied. Equivalently, for the same t_{SET} the current required to reduce the cell resistance to $10^5\Omega$ will be smaller for the reset-tail cell, than for the intrinsic cell.

Also shown in the figure is the critical current to obtain $10^5\Omega$ for the set-reset transition (right edge in Fig. 7). In the latter case, the transition current is negligibly affected by the set time, resulting in the flat scatter plot above I_{melt} in Fig. 8. For the reset-tail cell, this critical current is larger than for the intrinsic cell: this is because the tail cell is affected by a fast crystallization kinetics also during the quenching operation. As a result, to achieve the same threshold resistance of $10^5\Omega$ at the same quenching time t_q=10ns, a larger programming current is required by the tail cell.

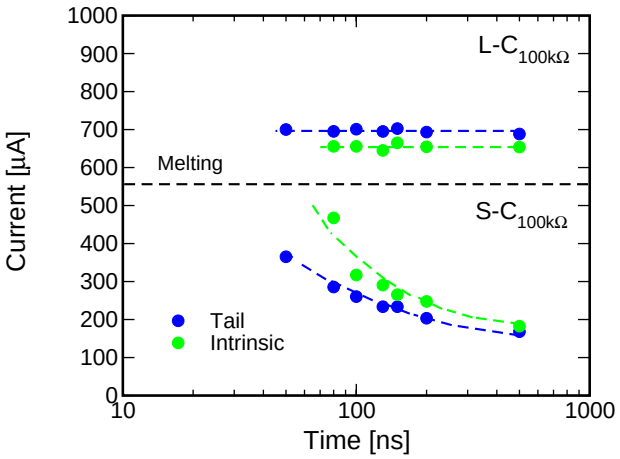


Fig. 8: Iso-resistance curves (100kΩ) obtained with a current pulse which amplitude in on y-axis and which duration on the x-axis. The quenching time is fixed and of 10ns. The tail cell is clearly faster in the crystallization process.

5. Long-term crystallization

Crystallization kinetics affects both the programming performance of the cell and the long-term stability of the amorphous phase. In fact, the amorphous volume within a reset-state cell can be transformed into the crystalline phase through low-temperature crystallization in the long term [7]. To study the link between the high-temperature set properties and the long-term, low-temperature crystallization in the cell, we performed an annealing experiment where cells were initially programmed in the reset state with t_q=2ns at room temperature, then baked at 250°C for 60s.

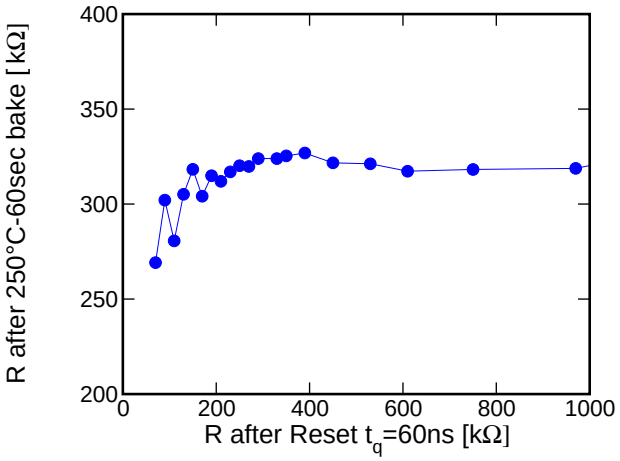


Fig. 9 Resistance after a bake experiment of 250°C-60sec as a function of resistance after a reset pulse with t_q= 60ns.

Fig. 9 shows the correlation between the measured resistance at the end of the annealing and the resistance of the cells after a non optimized reset operation, where a quenching time of t_q=60ns was applied in order to isolate a clear reset tail. A correlation is found, indicating that low-temperature and high-temperature crystallization are intimately related. Our analysis indicates that a first-order prediction of long-term crystallization for the cell can be provided by fast, high-temperature results such as those shown in Figs. 6 and 8. Note also that the long-term crystallization behaviour from Fig. 9 display a relatively small spread, as compared to the distribution slope of $t_{q,crit}$ and $t_{set,crit}$ in Fig. 6. This may allow for a better sensitivity in comparing different crystallization kinetics in PCM cells.

6. Conclusions

The quenching and set behaviour of a 2kb PCM sub-array were studied. It was found that the critical quenching and set times for our cells are strongly correlated, allowing to describe both programming performances (set-reset and reset-set transitions) uniquely based on the crystallization property of the active material. No anomalous crystallization behaviour was observed, indicating that reset tail phenomenon simply results from non linear quenching characteristics. We finally show that low-temperature and high-temperature crystallization kinetics are also intimately correlated, which may allow for a first order estimation of long-term crystallization in the cell only based on a fast, high temperature characterization of the array cells.

References

[1] D. Mantegazza, D. Ielmini, A. Pirovano, B. Gleixner, A. L. Lacaita, E. Varesi, F. Pellizzer and R. Bez, "Electrical characterization of anomalous cells in phase change memory arrays," *IEDM Tech. Dig.*, 53-56, 2006.

[2] K. Kim and S. J. Ahn, "Reliability investigations for manufacturable high density PRAM," *Int. Reliability Physics Symp.*, 157-162, 2005.

[3] F. Pellizzer, A. Pirovano, F. Ottogalli, M. Magistretti, M. Scaravaggi, P. Zuliani, M. Tosi, A. Benvenuti, P. Besana, S. Cadeo, T. Marangon, R. Morandi, R. Piva, A. Spandre, R. Zonca, A. Modelli, E. Varesi, T. Lowrey, A. Lacaita, G. Casagrande, P. Cappelletti, and R. Bez, "Novel μtrench phase-change memory cell for embedded and stand-alone non-volatile memory applications," *Symp. VLSI Tech. Dig.*, 18-19, 2004.

[4] F. Bedeschi, R. Bez, C. Boffino, E. Bonizzoni, E. C. Buda, G. Casagrande, L. Costa, M. Ferraro, R. Gastaldi, O. Khouri, F. Ottogalli, F. Peellizzer, A. Pirovano, C. Resta, G. Torelli and M. Tosi, "4Mb MOSFET-selected mtrench phase-change memory experimental chip," *IEEE J. Solid-State Circuit*, **40**, 1557-1565, 2005.

[5] P. G. Debenedetti and F. H. Stillinger, "Supercooled liquids and the glass transition," Nature **410**, 259-267, 2001.

[6] C. A. Volkert and M. Wuttig, "Modeling of laser pulsed heating and quenching in optical data storage media," J. Appl. Phys. **86**, 1808-1816, 1999.

[7] U. Russo, D. Ielmini, A. Redaelli and A. L. Lacaita, "Intrinsic data retention in nanoscaled phase-change memories – Part I: Monte Carlo model for crystallization and percolation," IEEE Trans. Electron Devices **53**, 3032-3039, 2006.

A Low Power PRAM using a Power-Dependant Data Inversion Scheme

Byung-Do Yang [a], Jae-Eun Lee [a], Jang-Su Kim [a], Jin-Kuk Yun [a], Seung-Yun Lee [b], Young-Sam Park [b], Sung-Min Yoon [b], and Byoung-Gon Yu [b]

[a] Chungbuk National University, Gaeshin-dong, Cheongju, Chungbuk, Korea, bdyang@cbnu.ac.kr

[b] Electronics and Telecommunication Research Institute (ETRI), Gajeong-dong, Yuseong-gu, Daejeon, Korea

Abstract

A low power PRAM using a power-dependant data inversion (PDI) scheme is proposed. The PRAM consumes large write power because large write currents are required for a long time. The PDI circuit compares two power consumptions to store the original data and its inverted data, and then it stores less power consuming data. Although the PDI scheme needs an additional inversion bit per data, the maximum and average powers of the PDI can be under 50% and 37.5% of the conventional write scheme, respectively. The average power for storing 8bit data is under 41%, due to the inversion bit. The 1K-bit PRAM chip with 128×8bits was implemented with a 0.8μm CMOS technology with a 0.5μm GST cell.

1. Introduction

Phase-change random access memory (PRAM) is an attractive non-volatile memory. PRAM has many advantages such as random access, non-volatility, good scalability, fast read time, moderately fast write time, good endurance for repetitive writing, and compatibility with CMOS process [1]. PRAM is much faster than Flash memory because PRAM can write any byte data, but Flash memory writes data in block unit with a complicated time consuming process [2]. PRAM is smaller than SRAM, and it does not consume standby power like as DRAM and SRAM. Therefore, PRAM is very attractive for low power mobile applications.

Fig. 1 shows the basic structure of the implemented PRAM unit-cell. The PRAM cell consists of an access transistor and a storage element of chalcogenide alloy (GST: $Ge_2Sb_2Te_5$). The GST is connected to a bit line. When the PRAM cell is selected, the word line turns on the access transistor. The GST has two resistances according to the stored values (‘0’ or ‘1’). At SET state, the GST has low resistance storing ‘0’. At RESET state, the GST has high resistance storing ‘1’. The PRAM utilizes the reversible phase-change phenomena between crystalline state (SET) and amorphous state (RESET) by electrical resistive joule heating. To make the SET state, the GST is heated by the SET current during the SET time, as shown in Fig. 2. To make the RESET state, the GST is heated by the RESET current during the RESET time.

The PRAM write power is significantly large because the write currents are large and the write times are considerably long. It can limit the PRAM applications for mobile devices.

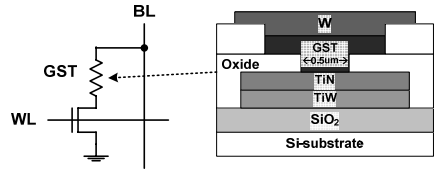


Fig. 1: Basic structure of PRAM unit-cell

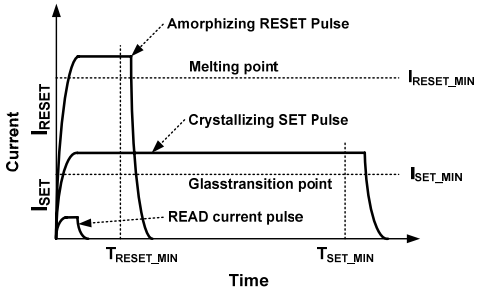


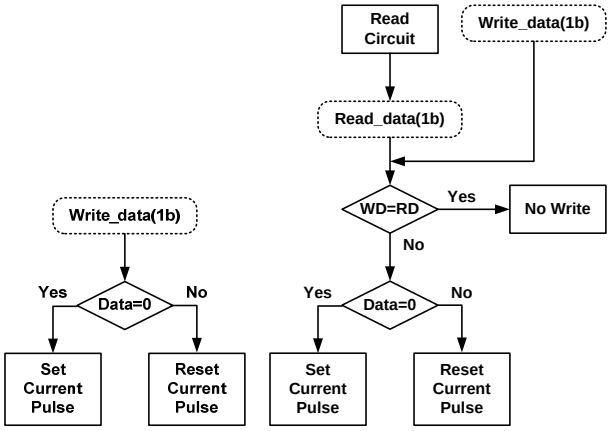
Fig. 2: Current pulses during the read, set, reset operations

Recently, the low power PRAM using a data-comparison write (DCW) scheme was proposed to reduce the write power [3]. The DCW circuit reads stored values from PRAM cells during write operation, and then it writes into the PRAM cells where the input and stored values are different. If the PRAM cell value does not change, it does not consume the write power. The average transition probability for each PRAM cell is 1/2. Therefore, the DCW scheme can reduce the write power consumption to a half.

In this paper, a low power PRAM using a power-dependant data inversion (PDI) scheme is proposed to reduce the write power. Basically, the PDI scheme uses the DCW scheme. In the DCW scheme, the average data transition probability is 1/2. The transition probability can be reduced by applying the data inversion technique used in the bus-invert coding [4]. The PDI uses this data inversion to reduce transition probability. However, the write power of the PRAM is not directly proportional to the transition probability. The power consumptions for storing ‘1’ and ‘0’ are significantly different. Therefore, the PDI circuit compares two power consumptions for storing the original data and its inverted data, and then it stores less power consuming data. This can minimise the write power consumption.

This paper is organized as follows. Section 2 introduces the concept and circuit implementation of the PDI scheme. Section 3 shows the chip implementation. Finally, Section 4 concludes this paper.

2. Power-dependant data inversion scheme



(a) (b)

Fig. 3: Flowchart of (a) Direct write (b) Data-comparison write

Cell Data Transition	Direct Write		DCW [3]	
	Power	***Probability***	***Power***	***Probability***
0 → 0	P_{SET}	1/4	0	1/4
0 → 1	P_{RESET}	1/4	P_{RESET}	1/4
1 → 0	P_{SET}	1/4	P_{SET}	1/4
1 → 1	P_{RESET}	1/4	0	1/4
Average Power	**$(P_{SET} + P_{RESET})/2$**		**$(P_{SET} + P_{RESET})/4$**	

Table 1: Power comparison of direct write and DCW

Fig. 3(a) shows a flowchart of the conventional direct write scheme. It always writes 1bit data on the selected PRAM cell. If input value is '0', the SET operation consumes the SET power (P_{SET}). If input value is '1', the RESET operation consumes the RESET power (P_{RESET}). There are four cell data transition cases (0→0, 0→1, 1→0, 1→1), as shown in Table 1. When the probabilities of four cases are 1/4, the average power of the conventional direct write scheme is $(P_{SET} + P_{RESET})/2$.

Fig. 3(b) shows a flowchart of the data-comparison write (DCW) scheme [3]. The DCW scheme performs the read operation before the write operation to know the previously stored value in the selected PRAM cell. If the input and stored values are the same, no write operation is performed. If not, the write operation is performed. The DCW scheme does not consume the write power for two cases (0→0, 1→1). Therefore, its average power becomes $(P_{SET} + P_{RESET})/4$.

The proposed power-dependant data inversion (PDI) scheme further reduces the write power in the DCW scheme. The PDI scheme also uses the DCW scheme, in which the maximum and average data transition probabilities are 1 and 1/2, respectively. The transition probabilities can be reduced to 0.5 and 0.375 by using bus-invert coding (BIC) [4]. In the BIC, each data code needs an extra bit which is called *invert*. If the number of transition bits in an *n*-bit data is over than *n*/2, *invert*=1 and the *n*-bit values are inverted. If not, *invert*=0 and the *n*-bit values are not inverted. Table 2 shows the write power comparison of various write schemes. We assume that the SET power (P_{SET}) is larger than the RESET power (P_{RESET}) and the number of data bits (*n*) is infinite. If *n*=8, the average data transition probability of the BIC becomes 0.41 from 0.375, respectively [4].

The PDI uses both DCW and BIC. However, the PDI inverts the data according to the real power consumption, whereas the BIC considers only the number of transition bits. The PDI circuit calculates the power consumptions for storing the original data and its inverted data, and then it stores less power consuming data.

The PRAM cell consumes the SET power (P_{SET}) and the RESET power (P_{RESET}). Typically, P_{SET} is larger than P_{RESET}. Where $\alpha=P_{SET}/P_{RESET}$, the maximum and average powers of the PDI depend on α. When α>1, those of the PDI are less than the BIC.

	Direct Write	**DCW [3]**	**BIC [4]**	**PDI**
Max. Power	**P_{SET}**	**P_{SET}**	**$P_{SET}/2$**	**-**
	1	**1**	**0.5**	**≤ 0.5**
Average Power @ $n=\infty$ (n=8)	**$(P_{SET} + P_{RESET})/2$**	**$(P_{SET} + P_{RESET})/4$**	**$(P_{SET} + P_{RESET})/4 \times 0.75$**	**-**
	1	**0.5**	**0.375 (0.41)**	**≤ 0.375 (≤0.41)**

Table 2: Write power comparison of various write schemes

	8bit data ($P_{SET}=\alpha$ & $P_{RESET}=1$)	**Invert bit**	**# of Transition**	**Total Power (α=5)**
Read data	0001,0111	0	-	-
Write data	0001,1000	0	-	-
Inverted data	1110,0111	1	-	-
Direct Write	**ααα1,1ααα**	**-**	**-**	**6α+2=32**
DCW	**0000,1ααα**	**-**	**-**	**3α+1=16**
Non-inverted	0000,1ααα	0(0)	4	3α+1=16
Inverted	11αα,0000	1(1)	5	2α+3=13
BIC	**Non-inverted**	**0**	**4**	**3α+1=16**
PDI	**Inverted**	**1**	**5**	**2α+3=13**

Table 3: Write power calculation example with 8bit data ($P_{SET}=\alpha=5$ & $P_{RESET}=1$)

α	**Direct Write**	**DCW**	**PDI**
1	1	0.5	0.41
2	1	0.5	0.40
3	1	0.5	0.39
4	1	0.5	0.39
5	1	0.5	0.38
6	1	0.5	0.38
7	1	0.5	0.37
8	1	0.5	0.37
9	1	0.5	0.37
10	1	0.5	0.37

Table 4: Normalized average write power comparison with 8bit data $\alpha=P_{SET}/P_{RESET}$

Table 3 shows a write power calculation example with 8bit data ($P_{SET}=\alpha=5$ & $P_{RESET}=1$). The DCW consumes power only for the transition bits. The BIC counts the number of transition bits for the non-inverted and inverted data. However, the PDI calculates the write powers for the non-inverted and inverted data. In this example, the non-inverted and inverted data have 4 and 5 transition bits but their power consumptions are 16 and 13, respectively. Therefore, the BIC writes the non-

inverted data, but the PDI writes the inverted data. In this case, the PDI consumes less power than the BIC.

Table 4 shows the normalized average write power comparison with 8bit data and $\alpha=P_{SET}/P_{RESET}$. As α increases, the PDI saves more write power.

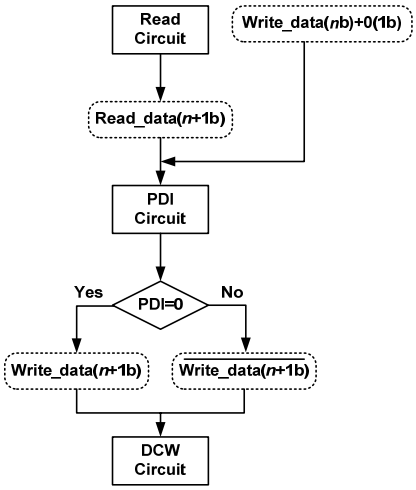


Fig. 4: Flowchart of the PDI scheme

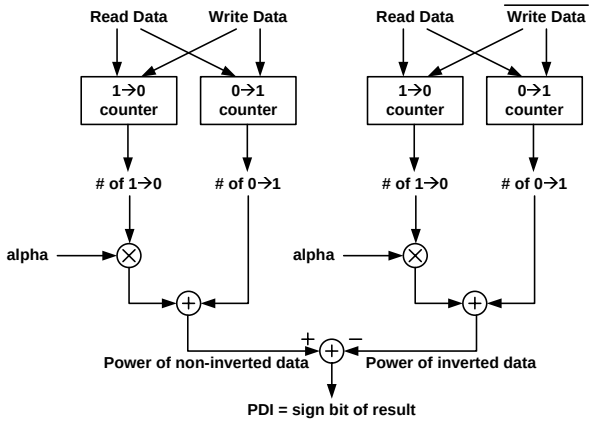


Fig. 5: PDI circuit

Fig. 4 shows a flowchart of PDI scheme. The stored *n*+1 bit *read_data* (*n* bit data + one invert bit) comes from the read circuit. The *n*+1 bit *write_data* (*n* bit data + one invert bit=0) comes from I/O circuits. As shown in Fig. 5, the PDI circuit calculates the write powers of the non-inverted and inverted data, and then compares which data consumes less power. If the non-inverted data consumes less power, *PDI*=0 (the result of PDI circuit) and the non-inverted data is stored. If not, *PDI*=1 and the inverted data is stored.

The calculated powers in the PDI circuit are normalized by P_{RESET}. Therefore, 1→0 transition power ($P_{1\to0}$)=α and 0→1 transition power ($P_{0\to1}$)=1. The number of 1→0 transition ($N_{1\to0}$) is multiplied by α, and then it is added to the number of 0→1 transition ($N_{0\to1}$). The result becomes total power ($\alpha\times N_{1\to0}+N_{0\to1}$). Two powers of the non-inverted and inverted data are calculated and compared. The PDI circuit consists of four bit-transition counters, two multipliers, two adders, and one subtractor (or comparator). The multipliers can be implemented by a few adders and shifters because α is constant.

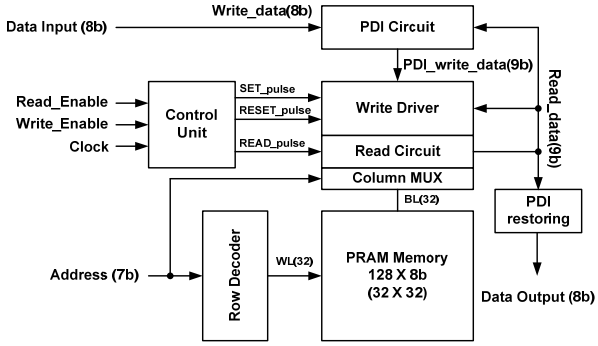


Fig. 6: Simplified block diagram of the PDI-PRAM

Fig. 6 shows the simplified block diagram of the power-dependant data inversion PRAM (PDI-PRAM). The PDI circuit generates the 9bit *PDI_write_data* from 8bit input data. The input address selects 9 cells with a selected word line and 9 selected bit lines (8bit data + one invert bit). The pulse generator makes 3 timing pulse signals (*READ_pulse*, *SET_pulse*, *RESET_pulse*) with 3 external signals (*Read_enable*, *Write_enable*, *Clock*). The read circuit and write driver perform the read and write operations with 3 timing pulse signals and the write data. The read circuit reads 9bit *read_data*. The PDI restoring circuit restores the 9bit *read_data* to 8bit original data by inverting the inverted *read_data*.

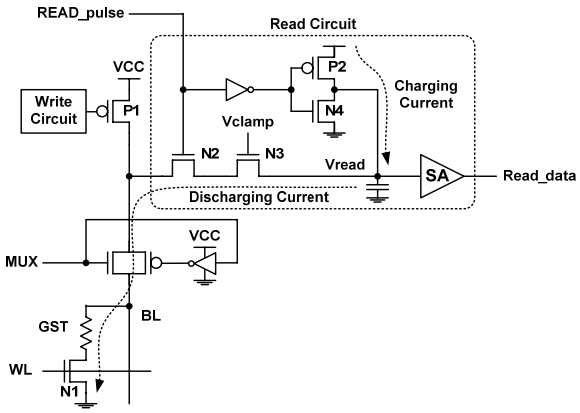


Fig. 7: Read circuit of the PDI-PRAM

Fig. 7 shows the read circuit of the PDI-PRAM. A PRAM cell is selected by an enabled word line (WL). A bit line (BL) is connected to the read and write circuits by a MUX. During the read operation, the *P2* transistor supplies the small read current into the selected bit line. To prevent unintentional write, the bit line is clamped by the *N3* transistor with V_{clamp}. The voltage of the bit line remains sufficiently lower than the threshold voltage of GST cell. When the GST has high resistance, the voltage of the sense amplifier input (*Vread*) rises to V_{DD}. When the GST has low resistance, *Vread* falls to ground.

Fig. 8 shows the write driver. The write circuit supplies the SET and RESET current pulses to changes the GST resistance. The PRAM uses two power supply voltages V_{CC} and V_{DD}. High voltage V_{CC} is required to supply large SET and RESET currents into the GST cell. The MUX signals also use V_{CC} to pass the write currents with high voltage. Low voltage V_{DD} is used for most of circuits except for the write current related circuits.

Fig. 9 shows waveforms of the write operation. It uses the DCW scheme. The selected cell data is sensed in the read circuit by the read pulse. The DCW circuit in Fig. 8 generates the SET or RESET signal for the write circuit only when the write data and read data are different.

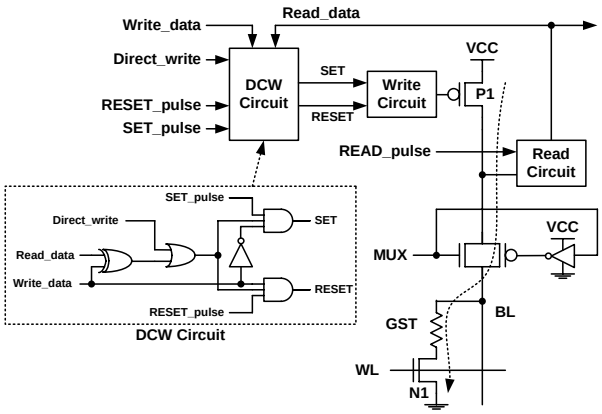


Fig. 8: Write driver of the PDI-PRAM

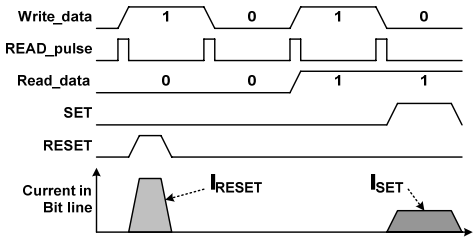


Fig. 9: Waveforms of write operation

3. Chip Implementation

The 1K-bit PRAM test chip with 128×8bits was implemented with a 0.8μm CMOS technology with a 0.5μm GST cell. Fig. 10 shows the chip photograph. The core area of the test chip is 2.4mm^2. The features of the test chip are tabulated at Table 5. The PRAM chip uses two power supply voltages V_{DD}=5V for logic circuits and V_{CC}=14V for SET and RESET currents.

Fig. 11 shows the measured waveforms. After the 10ns read pulse, if the write data is different from the read data, 4.5mA SET current or 16mA RESET current is supplied into the GST cell during 1μs or 50ns. The read time is 10ns and the read energy is only 74pJ/bit, whereas the write time is about 1μs and the SET and RESET energies are 64nJ/bit and 12nJ/bit, respectively.

4. Conclusion

In this paper, a low power PRAM using a power-dependant data inversion (PDI) scheme is proposed. To reduce the write power, the PDI scheme uses both DCW and BIC techniques. Also, it utilizes the fact that the power consumptions for storing '1' and '0' are different. The PDI circuit compares two power consumptions to store the original data and its inverted data, and then it stores less power consuming data. Although the PDI scheme needs an additional inversion bit per data, the maximum and average powers of the PDI can be under 50% and 37.5% of the conventional write scheme, respectively. The average power for storing 8bit data is under 41%, due to the inversion bit. The 1K-bit PRAM chip with 128×8bits was implemented with a 0.8μm CMOS technology with a 0.5μm GST cell.

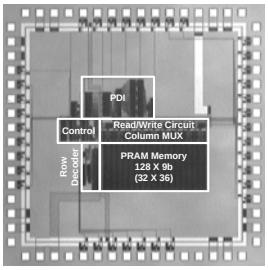


Fig. 10: Photograph of the PDI-PRAM chip

Technology	0.8μm CMOS process with 0.5μm GST cell and 2 metals
Organization	128 × 8 bits (Internally 9bit)
Supply Voltage	V_{DD} = 5V
	V_{CC} = 14V
Read Time	10 ns
Write Time	SET = 1000 ns @ 4.5mA
	RESET = 50 ns @ 16mA
Chip Core Area	2.4 mm^2
Energy / bit	READ = 74pJ
	SET = 64nJ
	RESET = 12nJ

Table 5: Features of the PDI-PRAM chip

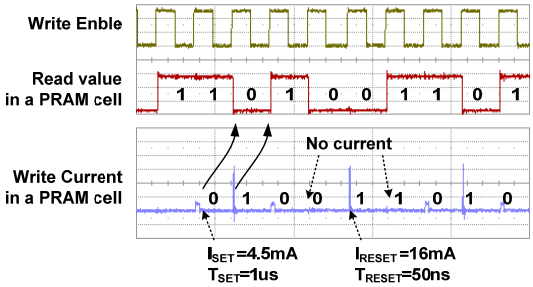


Fig. 11: Measured waveforms of the PDI -PRAM

5. Acknowledgement

This work was supported by the Regional Research Center Program of the Ministry of Education and Human Resources Development in Korea. Authors thank IDEC for CAD tool support.

References

[1] Hyung-rok Oh, et al., "Enhanced Write Performance of a 64-Mb Phase-Change Random Access Memory," *IEEE J. Solid-State Circuits*, Vol. 42, No. 1, pp. 122-126, Jan. 2006.

[2] Woo Yeong Cho, et al., "A 0.18-μm 3.0-V 64-Mb Nonvolatile Phase-Transition Rrandom Access Memory (PRAM)," *IEEE J. Solid-State Circuits*, Vol. 40, No. 1, pp. 293-300, Jan. 2005.

[3] Byung-Do Yang, et al., "A Low Power Phase-Change Random Access Memory using a Data-Comparison Write Scheme," IEEE *International Symposium on Circuits and Systems*, May 2007, accepted.

[4] Mircea R. Stan and Wayne P. Burleson, "Bus-Invert Coding for Low-Power I/O," *IEEE Transactions on Very Large Scale Integration Systems*, Vol. 3, No. 1, pp. 49-58, Mar. 1995.

Threshold switching in doped SbTe phase change line cells

Friso J. Jedema [a], Job van der Wagt [a], Micha A.A. in 't Zandt [a], Rob A.M. Wolters [a], Bas W.S.M.M. Ketelaars [b], Romain Delhougne [c], David Tio Castro [c], Dirk J. Gravesteijn [c] and K. Attenborough [c]

[a] NXP Semiconductors, High Tech Campus 4, 5656 AE Eindhoven, The Netherlands, friso.jedema@nxp.com
[b] Philips Research, High Tech Campus 4, 5656 AE Eindhoven, The Netherlands

[c] NXP Semiconductors, Kapeldreef 75, 3001 Leuven, Belgium

Abstract

Threshold switching is an essential property of a phase change memory cell. In this work, the threshold switching of doped SbTe phase change line cells is studied. It is observed that the time scale of the switching event is very short, typically less than 1 ns. The magnitude of the threshold voltage is observed to be dependent on the amorphous line resistance, line length and delay time between a reset and set pulse. Interestingly, a finite threshold voltage is determined at infinitely small line lengths. A change in the amorphous state resistance and threshold voltage is observed when the lines are exposed to large reset currents.

1. Introduction

Phase change random access memory (PCRAM) is a candidate to replace current Flash memory technology. Its main advantages are scalability, programming speed, less lithographic masks and smaller cell size. Two concepts of PCRAM are being pursued today: the Ovonyx Unified Memory (OUM) concept [1,2] and the line concept [3,4]. In the line concept the melting of the phase change material occurs where the cross-sectional area of the line is the smallest. The lateral cell design has several advantages compared to the OUM concept. First, all heating occurs within the phase change line rather than at a metal/phase change interface. The hottest part of the line can therefore be fully surrounded by a low heat conductance dielectric material. Secondly, the current cross-sectional area can be reduced by varying the film thickness of the phase change material, thereby offering more aggressive possibilities for reducing programming current. Lastly, as the memory is scaled down to smaller dimensions, the resistance of the cell remains constant. Since the programming current scales with decreasing line width, the constant cell resistance causes a simultaneous reduction in the programming voltage.

For a PCRAM to function, three regimes in the available voltage window have to be defined to enable independent reading, set programming to obtain the low resistive crystalline set state and reset programming to obtain the high resistive amorphous reset state. For PCRAM, the minimal set voltage is determined by the so-called threshold voltage (V_T) of the phase change (PC) material. For voltages smaller than V_T, a PCRAM cell in the reset state does not conduct the required current to be programmed into the set state. The threshold voltage is therefore an important parameter for the PCRAM memory cell. In this paper the magnitude of threshold voltages of doped SbTe phase change line cells are determined and its dependence on the amorphous state resistance, line length and time are studied in detail.

2. Device fabrication

The devices were fabricated on (100) Si wafers with a 500 nm grown thermal oxide layer. First, TiW bottom electrodes are deposited and patterned by standard optical lithography. Subsequently an oxide layer is deposited and the surface is planarized by a chemical mechanical polishing (CMP) step.

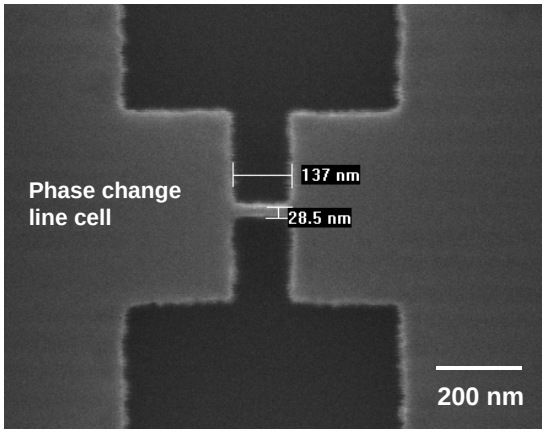


Fig. 1: Top view scanning electron microscope picture of a 25 -125 nm (designed) line cell.

After CMP, a 20 nm PC film is sputter deposited. Single line cells are then patterned by e-beam lithography and Ar plasma etching, using a Hydrogen SilsesQuioxane (HSQ) hard mask. In Fig. 1 a typical single line cell is shown with a designed width (W) of 25 nm and a designed length (L) of 125 nm. The line is connected to phase change contact flaps of several microns long and wide, which are lying on top of the TiW bottom electrode. Prior to the sputter deposition of the PC film, the TiW electrodes were cleaned by an in-situ sputter etch. After e-beam patterning the phase change lines are passivated by oxide.

3. Reset current sweep

In Fig. 2 reset current sweeps are shown for a 25-100 nm line cell. A load resistance of 3.3 kΩ was placed in series with the sample resistance. A reset sweep starts with a measurement of the initial crystalline resistance, typically around 1.2 kΩ. Subsequently, reset pulses of 50 ns duration with a leading edge and trailing edge of 4 ns ,see inset of Fig. 2, are applied with increasing amplitude. After each reset pulse, the resistance is measured with a delay time of approximately 1 second. Once the sample is programmed into the amorphous

reset state, the sample is brought back to its crystalline set state by applying a set pulse of 300 ns duration, having leading and trailing edges of 100 ns.

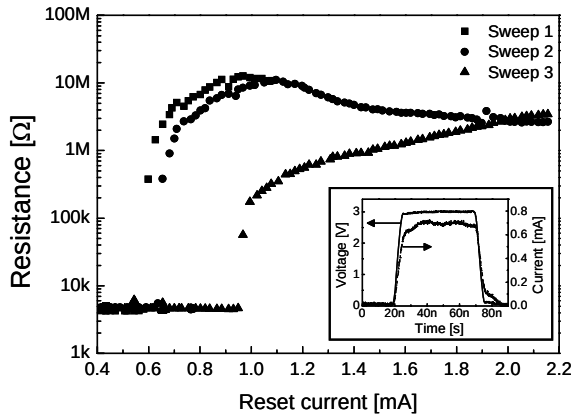


Fig. 2: Reset current sweeps of a 25-100 nm line. Sweeps 1, 2 and 3 are taken directly after each other and show that the line has changed under application of large reset currents. The inset shows an example of a 50ns reset pulse with an amplitude of 3 Volt and 0.7 mA. The measurement time in between the points of sweep 1,2 and 3 is approximately 1 second.

A first sweep (sweep 1) was started at a reset current of 0.4 mA and stopped at 1.12 mA. The data points of sweep 1 show that a minimal reset current of 0.59 mA is needed to program the 25-100 line into the amorphous reset state. From that point on, a continuous increase of the amorphous state resistance is observed with increasing reset currents. This is attributed to a continuously increasing size of the amorphous spot in the phase change line. A maximum in the reset resistance is observed for ~ 0.95 mA. At this point transmission electron microscope (TEM) images confirmed that the line is fully amorphous. Subsequently, a decrease in the amorphous resistance of the line is observed when applying reset currents larger than 0.95 mA, up to 1.12 mA. After sweep 1, a second sweep (sweep 2) was started at 0.4 mA and stopped at 2.15 mA. A slight change is observed in the minimal reset current which is now shifted to a larger value of 0.66 mA. The maximum of the amorphous state resistance has changed also, as well as its position on the horizontal axis in Fig 2., being equal to the end point amorphous state resistance of sweep 1. The observation that the cell has changed is further confirmed by a third reset current sweep (sweep 3). Here the minimal reset current is 0.98 mA and a monotonous reset resistance increase is observed, up to a reset current of 2.15 mA, the point were sweep 2 ended. From Fig. 2 it can be concluded that the line cell cannot sustain large reset currents without under going significant changes. As will be shown in the next 2 paragraphs, the change in the amorphous reset resistance is also accompanied by a change in the magnitude of the threshold voltage.

4. Determination of the threshold voltage

The magnitude of the threshold voltage can be determined from the I-V characteristics of the set pulse. Four typical I-V curves for a 25-100 nm sample are plotted in Fig. 3. The I-V curves of Fig 3. were obtained using similar set pulses as applied in sweep 1 of Fig. 2, having a 300 ns duration with leading and trailing edges of 100 ns. The voltage plotted in Fig. 3 is the voltage of the pulse as measured across the sample and a 3.3 kΩ load resistor. The current is obtained by measuring the voltage across a 50 Ω input resistance of the oscilloscope, connected in series with the sample and the load resistor. The time between two consecutive measuring points of the I-V plots in Fig. 3 is 0.1 ns.

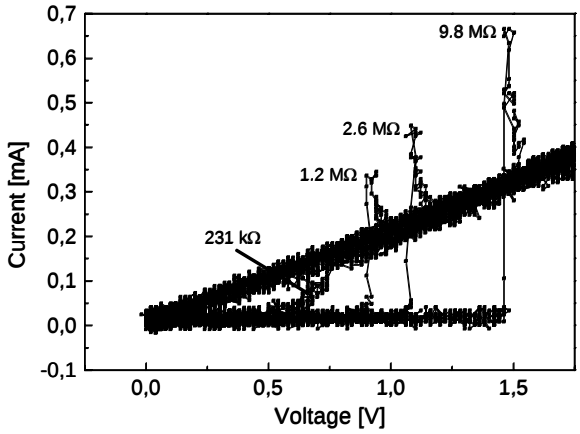


Fig. 3: I-V characteristics of a 25-100 nm line, obtained from a set pulse of 300ns duration, having 100 ns leading and trailing edges. No current is flowing until the voltage reaches the threshold voltage, after which a rapid increase occurs within 1 ns and the line switches into its crystalline resistance state. The labels indicate the amorphous state resistance before the set pulse was given. The time between 2 consecutive measuring points in all curves is 0.1 ns.

Figure 3 shows a gradual development of the threshold switching effect as the magnitude of the amorphous state resistance is increased from 231 kΩ to 9.8 MΩ. It can be seen that the threshold switching takes place on a time scale of less than 1 ns. As this time scale is much shorter than the typical RC delay times in memory circuits, the line cell might be modelled as an instant switchable resistor. The peaks in I-V's at the moment of switching are attributed to parasitic capacitances in the measurement circuit. Without the 3.3 kΩ load resistor, the peaks in I-V's are absent.

5. V_T dependence on the amorphous R

In Fig. 4 the magnitude of the threshold voltages are plotted as a function of the amorphous state resistance for the same sweeps of the 25-100 line cell, as plotted in Fig. 2. An almost linear relation is observed. The threshold voltage has a maximum of 1.8 Volt, at an amorphous state resistance of 12.4 MΩ. This resistance corresponds to the maximum resistance at a reset current of 0.98 mA in sweep 1 of Fig. 2, where the line was determined to be fully amorphised by a TEM image.

Figure 4 further shows that the change in the amorphous state resistance, as earlier observed in Fig. 2, is also accompanied by a change in the magnitude of the threshold voltage. For sweep 2 the threshold voltage is slightly increased for the same amorphous resistances as compared to sweep 1. However, for sweep 3, the threshold voltages are doubled for similar measured amorphous state resistances in sweep 1.

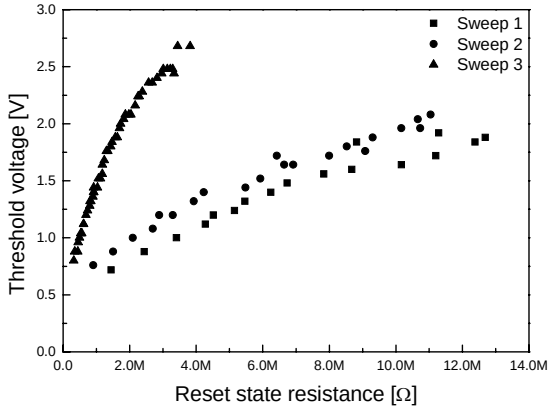


Fig. 4: Threshold voltage dependence as a function of the amorphous reset resistance of a 25-100 nm line. The amorphous reset resistances on the horizontal axis, correspond to the amorphous reset resistances on the vertical axis of Fig.2. Note that in this plot the threshold voltage was determined approximately 1 second after a reset pulse was applied.

As the threshold voltage should be minimized for obtaining a proper set regime in the available voltage window, one should therefore avoid exposing the line cell to too high reset currents.

6. V_T dependence on line length

From Fig. 2 it can be established that a line is completely amorphised when the amorphous reset resistance is maximal at a particular reset current. Knowing the amorphous reset resistance at this reset current, one can subsequently determine the magnitude of the threshold voltage for a completely amorphised line from Fig. 4.

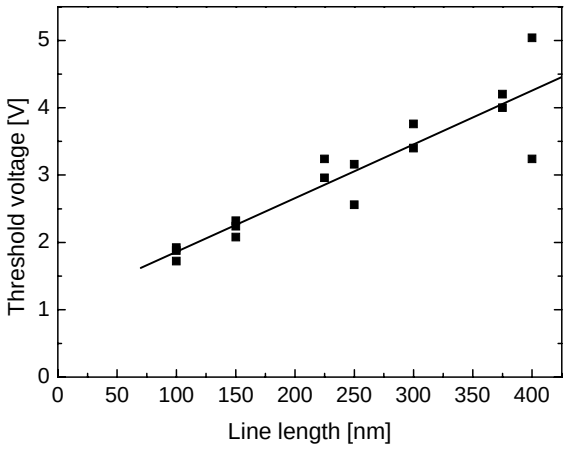


Fig. 5: Threshold voltage dependence as a function of line length. The square data points indicate the threshold voltage for completely amorphised lines, as indicated by a maximum in the amorphous state resistance in reset current sweeps, similar to sweep 1 in Fig. 2. The solid line represents a linear fit to the data points. Note that the threshold voltage was determined at approximately 1 second after a reset pulse was given.

For the 25-100 nm line of Fig. 2 and Fig.4, the threshold voltage for a completely amorphised line is determined to be 1.8 Volt. By performing similar experiments for longer line lengths, one can obtain the threshold voltage as a function of line length. This was performed for line cells with lengths ranging from 100 nm to 400 nm and the result is plotted in Fig. 5. Applying a linear fit to the data points in Fig. 5 yields the following relation:

$$V_T = 8 \times L + 1 \quad (1)$$

Here V_T is the threshold voltage in Volts and L is the line length in μm. Note that Eq. 1 is valid for threshold voltages obtained at a time of approximately 1 second after a reset pulse has been given. Equation 1 shows that there exists a threshold electric field of about 8 V/μm. But it also shows that the threshold voltage has a finite positive offset of about 1 Volt. At present the origin of this offset is unclear. However, one explanation for the offset voltage could be that it is related to the band gap of the amorphous phase change material. The band gap represents the minimal energy a charge carrier has to acquire to be excited from the valence into the conduction band. This energy scale is relevant for the process of impact ionization, which is reported to be responsible for the threshold switching in GST 225 phase change materials [5,6].

7. V_T time dependence

In the previous paragraphs, the magnitude of the threshold voltage was measured at a time of approximately 1 second after a reset pulse has been given. However, the threshold voltage is also found to be dependent on the time between the reset pulse and the consecutive applied set pulse. Therefore an experiment with a series of reset-set pulses was carried out, having a variable delay time between the reset and set pulse. The delay time is defined here as the time between the end of the reset pulse and the start of the set pulse. In Fig. 6 a typical oscilloscope recording of a reset-set waveform of a 75-300 nm line cell is plotted with a delay time of 10 ns. Figure 6 is showing the recorded voltage waveform as well as the recorded current waveform.

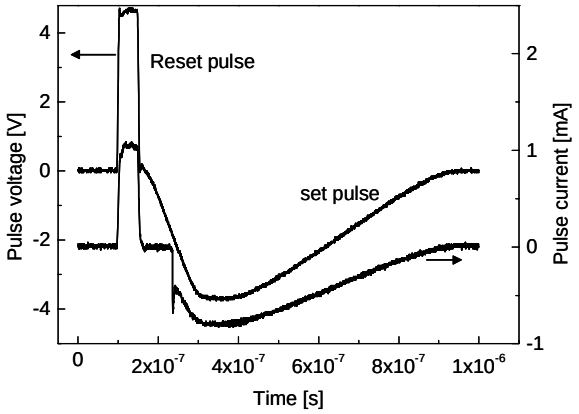


Fig. 6: Oscilloscope waveforms of a typical reset-set pulse of a 75-300 nm line cell. The reset pulse has an amplitude of 4.7 Volt and 1.04 mA. The set pulse has an amplitude of 3.8 Volt and 0.8 mA. The delay time between the end of the reset pulse and the start of the set pulse is 10 ns. The pulse voltage is the voltage across the sample and a 3.3 kΩ load resistor, whereas the pulse current is obtained by measuring the voltage across a 50 Ω input resistor of the oscilloscope, in series connected with the sample and the 3.3 kΩ load resistor.

A clear threshold switch is observed in Fig. 6 when the set pulse has reached –1.9 Volt. The negative voltages in set pulse were used to be able selectively record only the set part of the reset-set pulse with the oscilloscope, for longer delay times.

In Fig. 7 the threshold voltage is plotted as a function of the delay time, for two different reset-set pulses. One reset-set pulse had a 50 ns reset pulse with an amplitude of 3.7 V and 0.8 mA, resulting in a 1 MΩ amorphous reset state. A second reset-set pulse had a 50 ns reset pulse with an amplitude of 4.7 V and 1.04 mA, resulting in a 5 MΩ amorphous reset state. The threshold voltages are obtained from similar waveform plots as shown in Fig. 6. From Fig. 7 it can be observed that longer reset-set delay times lead to a monotonous increase in the magnitude of the threshold voltage. A similar behavior has also been observed for the GST 225 phase change material, and is reported to be caused by intrinsic trap dynamics in the amorphous state [5].

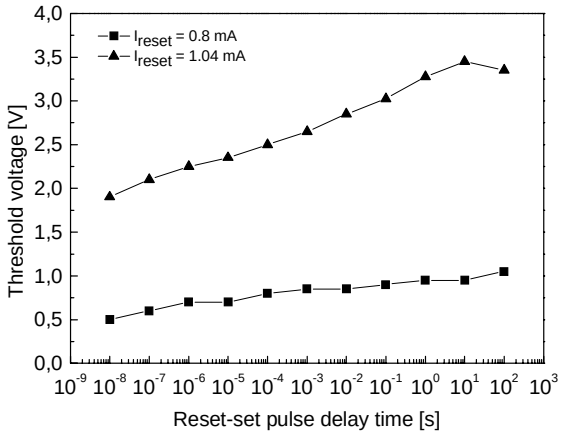


Fig. 7: Threshold voltage of a 75-300 nm line cell, as a function of delay time between the end of a reset pulse and the start of a set pulse. The square points are the measured threshold voltages after applying a 50 ns reset pulse with an amplitude of 3.7 V and 0.8 mA. The triangular points represent the measured threshold voltages after applying a 50 ns reset pulse with an amplitude of 4.7 V and 1.04 mA. The set pulses were in both case the same and had an amplitude of 3.7 V and 0,8 mA with a duration of 700 ns, having a 100 ns leading edge and a 500 ns trailing edge. The solid lines serve as a guide to the eye.

Figure 7 furthermore indicates that a line cell can be read out as quick as 110 ns after a reset pulse has been applied, provided that the read out voltage is less than 0.5 Volt. Note that only at shorter delay times than 1 second the magnitude of the threshold voltage is lower than the offset voltage as given by Eq. 1. At a delay time of 1 second the threshold voltage is 0.9 Volt, in close agreement with the offset voltage as predicted by Eq. 1.

8. V_T read out stability

For proper operation of a PCRAM cell, it is necessary that the amorphous reset state can withstand many read outs without being destructed. Therefore a read endurance test was performed on a 25-125 line cell, as shown in Fig. 8. First the line cell was programmed to 1.6 MΩ. Subsequently, 50 ns read out pulses of 0.7 Volt were applied, as shown in the inset of Fig. 8. After 1E+09 read out pulses no significant change occurred. Only a slight increase of the amorphous state resistance is observed from the initial 1.6 MΩ to a value of 1.8 MΩ. After the 1E+09 read out cycles, a 2 Volt set pulse was applied and the threshold voltage was determined to be 1.1 Volt. This shows that the amorphous reset state is robust against many consecutive read out pulses with a voltage of only 0.4 Volt below the actual threshold voltage of the amorphous reset state.

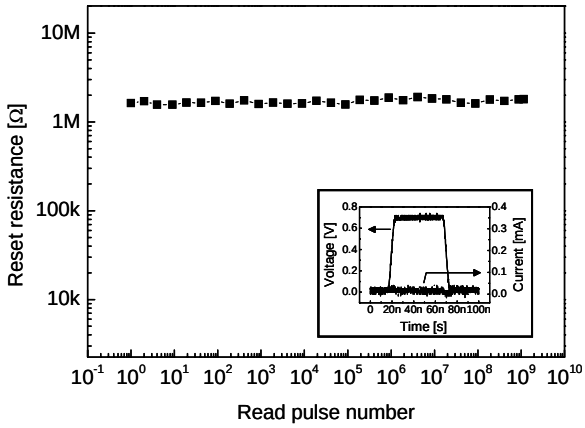


Fig. 8: Read endurance of a 25-125 nm line, as a function of the number of applied read out pulses. Prior to the endurance test, the line was programmed into an amorphous reset state of 1.6 MΩ. The inset shows the voltage and current waveform recordings of the 50 ns read out pulse with a amplitude of 0.7 V. After 1E+09 read out pulses the threshold voltage was determined to be 1.1 Volt.

9. Conclusions

Doped SbTe line cells show a threshold switching at a time scale of less than 1 ns. This time scale is shorter than typical RC delay times of memory circuits. The threshold voltage is dependent on the amorphous state resistance, line length and time. An offset in the threshold voltage is observed at infinitely small line lengths. The amorphous reset state is robust against many read out pulses. When applying large reset currents to a line cell, lower amorphous resistances and larger threshold voltages are observed.

References

[1] F. Pellizzer et al., "A 90 nm Phase Change Memory Technology for Stand-alone Non-Volatile Memory Applications", *Symp. On VLSI Tech.*, 2006.

[2] S.L. Cho, et. al., "Highly scalable on-axis confined cell structure for high density PRAM beyond 256Mb", *Symp. VLSI Tech.*, 2005.

[3] M.H.R. Lankhorst, W.S.M.M. Ketelaars, R.A.M. Wolters, "Low-Cost and Nanoscale Non-Volatile Memory Concept for Future Silicon Chips", *Nature Materials* 4, 347-352 (2005).

[4] Y.C. Chen et al., "Ultra-Thin Phase-Change Bridge Memory Device Using GeSb", *IEDM Tech. Dig.*, 2006.

[5] D. Ielmini, A.L. Lacaita, and D. Mantegazza, "Recovery and Drift Dynamics of Resistance and Threshold Voltages in Phase-Change Memories", IEEE Transactions on electron devices, Vol. 54, No. 2, February 2007

[6] A. Pirovano, A.L. Lacaita, A. Benvenuti, F. Pellizzer and R. Bez, "Electronic Switching in Phase-Change Memories", IEEE Transactions on electron devices, Vol. 51, No. 3, March 20

Geometry and material optimization for programming current scaling in phase-change memory

U. Russo, A. Redaelli, D. Ielmini, A. L. Lacaita

Dipartimento di elettronica e informazione, Politecnico di Milano, and IUNET, p. L. da Vinci 32, 20133 Milano,
email: russo@elet.polimi.it

Abstract

The reduction of cell programming current is a major challenge for phase change memory in order to allow cell-area scaling and large parallelism in array programming. This paper addresses the problem for vertical lance and ring cells. The trade-off with sensing issue is analyzed and it is demonstrated to be not a main concern down to the 16 nm technology node. Geometrical design and material engineering are proposed, providing optimum programming performance.

1. Introduction

The phase change memory (PCM) is attracting large interest as a possible new technology for non-volatile memories [1]. Although the physical concept of PCM is known from several decades, some issues have to be solved for the technology to become competitive with today's NAND and NOR memories. One major concern is related to the relatively-high programming current, which currently limits the size of the MOS selection transistor and the overall area scaling of the cell [2]. Although it has been recognized that the programming current can be decreased by geometrical cell-size scaling [2, 3], methods to optimize the programming current keeping the same size of the bottom electrode and the same cell resistance have not been proposed. Also, contact-area scaling increases cell resistance, thus increasing the cell readout time and raising a potential readout issue.

This work addresses the programming current optimization and scaling and the programming-read trade-off issues in details. We first compare simulation results to measured electrical characteristics for a reference 90 nm vertical lance cell [4], in order to validate our a numerical electro-thermal model for PCM device simulations [5]. Then, optimization and scaling of write-read performances are analyzed for scaled cells down to the 16 nm technology node. Both isotropic and non-isotropic scaling approaches are considered and compared. Further cell-performance optimization is then investigated, comparing both different vertical cell structures (lance-type [4] and ring-type [6]) and improved materials for both the chalcogenide layer and the bottom electrode.

2. Experimental data and simulations

Fig. 1 shows the SEM cross-section of the reference cell in this work, namely a 90 nm PCM cell employing $Ge_2Sb_2Te_5$ (GST) as the phase-change material and a vertical lance-type structure [4]. The doped-TiN bottom electrode, or heater, is cylindrical shaped and displays sub-litho contact area with the GST layer above. The cell operation is based on the phase change of the GST material in the so-called programmable volume, close to the heater. Phase change is achieved via Joule heating through electrical pulses: if Joule heating is enough to melt the GST material, a transition to the amorphous phase is obtained. The crystalline phase can be recovered by a softer Joule heating through solid-state nucleation and growth. Readout of the cell resistance enables to distinguish between the amorphous and the crystalline phase, with high and low resistance respectively.

Fig. 2 shows measured *I-V* characteristics for the cell in the *set* state, corresponding to the crystalline phase of the GST. Readout is usually performed at low voltage, yielding the resistance R_{set} in Fig. 2. On the other hand, the on-resistance R_{on} largely impacts the programming operation, since it determines Joule-heat generation within the cell. Note that while R_{set} is given by the series contributions of the GST layer and of the heater, R_{on} is dominated by the latter contribution. In fact, GST becomes largely conductive at relatively high voltage, mainly due to a large thermal generation of carriers. Also shown in Fig. 2 are calculations by our electro-thermal model for electrical and programming characteristics of PCM cells, indicating a good agreement with data.

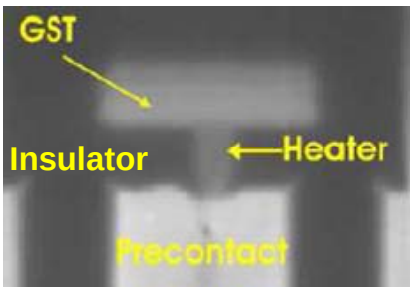


Fig. 1: SEM cross section of the vertical cell in the 90nm technology, which was used as reference cell to calibrate our electro-thermal model for PCM cells.

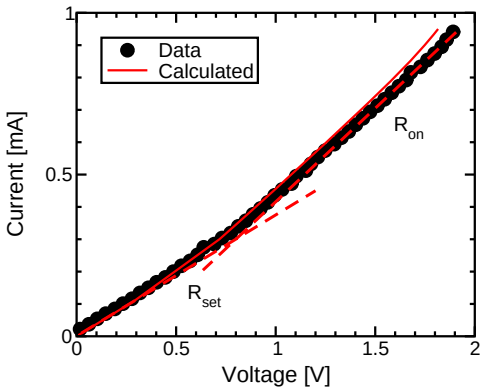


Fig. 2: Measured and calculated electrical *I-V* characteristic of a PCM lance cell in the crystalline, or *set*, state. The low field, R_{set}, and high field resistances, R_{on}, are marked.

Fig. 3 shows the measured and calculated *R-I* characteristics, where the programmed resistance *R* after the programming pulse is plotted as a function of the current pulse *I*, always applied to a cell in the set state. From the figure, *R* remains equal to R_{set} for small currents, then sharply increases for currents above about 520μA. This current marks the condition where Joule heating is sufficiently large to bring the GST temperature above the melting point, and is called melting current I_m. For increasing current above I_m, the programmed amorphous volume and the cell *R* increase [5]. Note that, as in Fig. 2, a good agreement between calculations and data is achieved, which validates our electro-thermal model for the purpose of numerical investigation of the optimization and scaling of the PCM cell in the following.

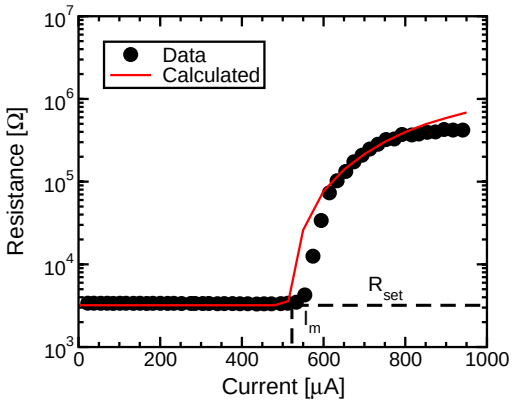


Fig. 3: Measured and calculated programming *R-I* characteristic. The programmed resistance increases over the initial set-state value as *I* exceeds the melting current I_m.

3. Programming current optimization

The parameters I_m and R_{set} in Figs. 2 and 3 are the main parameters affecting the programming and read efficiency of the cell, respectively. In fact, the cell is usually reset with a current I_{reset} larger than I_m, in order to achieve a sufficient resistance window between set and reset states. Thus I_m has to be minimized for minimum reset current consumption. On the other hand, a low R_{set} is required if appreciable current is needed when a sensing voltage of a few hundreds of mV is applied during the readout operation. Therefore, we investigated a means for optimizing the programming current by changing the cell geometry at a fixed lithographical node (i.e. maintaining a constant bottom contact diameter ϕ) and at a fixed resistance R_{set} value, thus not affecting negatively the readout performances.

Fig. 4 shows the simulation methodology in this work: the heater length L_h and the chalcogenide thickness L_c were changed with fixed ϕ and R_{set}, while the melting current I_m and the resistance R_{set} were calculated. It is clear from the figure that, as L_h is increased, L_c is correspondingly decreased to maintain the same R_{set}. Also shown in the figure are the temperature maps at melting condition: to maximize the efficiency of the programming operation, the hot spot has to be located close to the interface between heater and GST (Fig. 4b), since the resistance in the *R-I* curve is effectively increased only if the GST phase in the programmable volume is changed. As L_h is increased, the hot spot moves into the bottom contact (Fig. 4c); increasing the chalcogenide thickness will instead move the melting point deep into the GST (Fig. 4a).

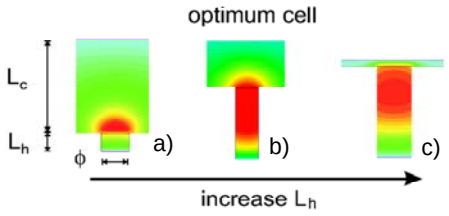


Fig. 4: Simulation methodology for studying the optimization of programming current. For a given cell diameter ϕ, heater and GST height L_h and L_c are changed at constant R_{set}, while the programming current is evaluated to search for minimum-I_m cell geometry.

Fig. 5 shows the calculated optimization curves for ϕ=30 nm, which is suitable for a technology node of *F*=45 nm. Calculations are reported for increasing L_h (hence decreasing L_c) and for R_{set}=2, 3, 4 and 5kΩ. Clearly, I_m decreases for increasing R_{set}, raising a trade-off between programming and sensing requirements. On the other hand, for each R_{set} a minimum I_m can be found, which corresponds to the optimized cell geometry. Note that the width of the U-shaped curve increases for larger R_{set}, as L_h and L_c increase and the thermal and electrical boundaries move away from the active region. This is further demonstrated in Fig. 6, where the optimum cell geometries are mapped as a function of L_h and L_c. Here, a cell geometry is considered optimum if its melting current is within 10 μA from the minimum I_m. From the figure, the optimum region spreads over for increasing L_h and L_c (hence increasing R_{set}), thus leaving a profitably larger degree of freedom in the cell design. Note that upper limits to L_h and L_c are likely to be dictated by manufacturability constraints. These may possibly arise from *a)* the technological capability of realizing the heater pillar by filling with TiN a pore with a large L_h/ϕ aspect ratio, and *b)* limitations in the maximum reliable thickness of the patterned GST layer. Fig. 6 reports suitable L_h and L_c constrains for the 45 nm node.

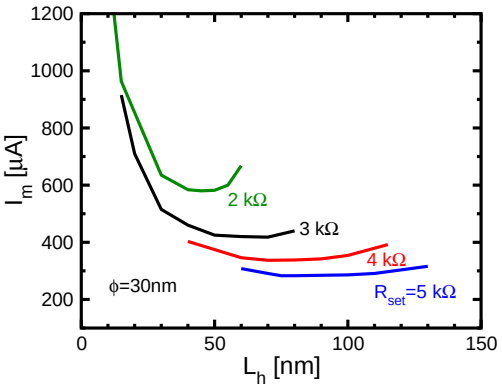


Fig. 5: Optimization curves for a lance cell with ϕ=30 nm. Calculated I_m is reported for several constant R_{set} values as a function of L_h. Note that when L_h is increased, L_c is decreased (see Fig. 4).

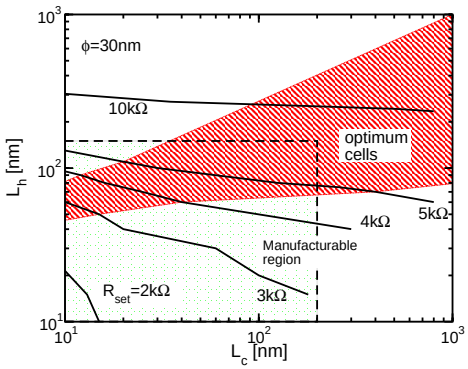


Fig. 6: Values of L_h and L_c corresponding to optimum cell geometry (in red), defined as a cell with I_m within 10 μA from the minimum melting current at the same R_{set}. L_c and L_h values yielding the same R_{set} are shown (solid iso-R_{set} curves). A suitable manufacturability region is also shown.

4. Scaling performances

The evolution of the programming and read parameters for the technology nodes from F=90 to 16 nm was studied by numerical simulations. The 90 nm optimized cell was taken as a reference, then two different scaling solutions were considered: an *isotropic* scaling, where all geometric dimensions are scaled by the same factor, and a *non-isotropic* scaling, where ϕ is reduced while keeping constant L_h and L_c.

Fig. 7 shows programming voltage V_m and current I_m as a function of F, for both scaling approaches. Non-isotropic scaling reduces I_m more aggressively than isotropic scaling. This is because electrical and thermal resistances, hence the Joule heating efficiency, increase more rapidly for increasingly downscaled cell. At the same time, the increase in electrical resistance also leads to a significant V_m increase with scaling, which is negative for low-voltage applications. Instead V_m remains constant with the isotropic scaling approach.

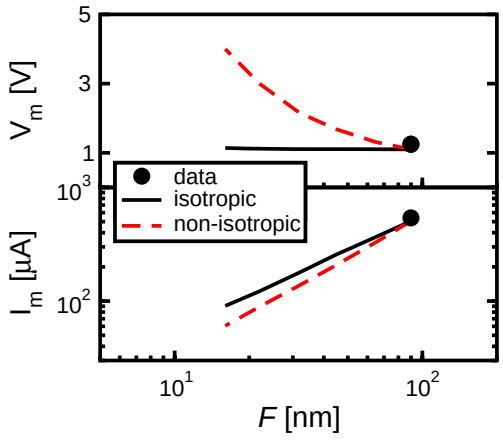


Fig. 7: Calculated programming voltage and current as a function of the technology node in a lance cell, for isotropic and non-isotropic scaling.

Note in Fig. 7 that, from the technological standpoint, the non-isotropic scaling approach may not be straightforwardly manufacturable, since for particularly small F the aspect ratio L_h/ϕ steeply increases. Hence, non isotropic scaling should be considered as an upper, theoretical limit for aggressive I_m scaling.

As already pointed out, the reduction of I_m by scaling is accompanied by an increase of R_{set}, thus degrading the read performance of the cell. This is shown in Fig. 8, which is a scatter plot of calculated I_m and R_{set} for all technological nodes from 90 to 16nm. From the I_m-R_{set} plot, it is clear that the best trade-off between write and read performance upon scaling is obtained by isotropic scaling, which was found to preserve the optimum feature of the 90nm cell. For isotropic scaling, a reduction of I_m results in an equal increase of R_{set}. Thus the slope in the I_m-R_{set} plot is -1, meaning that a constant R_{set}·I_m product is obtained. Non isotropic scaling, although providing a faster I_m downscaling (Fig. 7), is affected by a lower slope (-0.7) in the I_m-R_{set}, which indicates that the R_{set} increase is higher than the I_m reduction. We note however that R_{set} remains below 100 kΩ for both scaling approaches, thus ensuring a read current of at least few μA. This suggests that, even for non-isotropic scaling, the degradation of readout performance for the lance cell should be within the acceptable limits even at the 16 nm technology node.

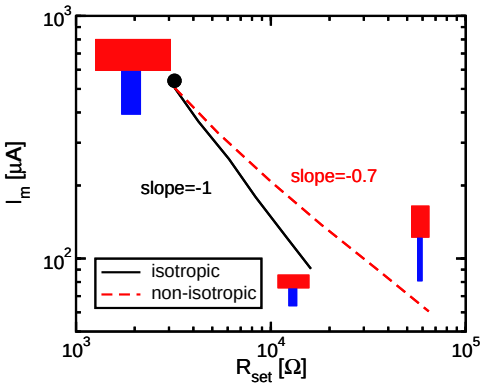


Fig. 8: Melting current I_m as a function of set resistance R_{set}, for lance cell and both isotropic and anisotropic scaling.

5. Geometry and material optimization

In order to further scale the programming current in a PCM cell, different cell geometries or accurate material engineering can be employed. A ring-type structure has been recently reported to reduce the programming current with respect to the lance cell [6]. In a ring-type cell, the pore in the insulator defining the bottom contact is not completely filled with TiN, but only a thin film is deposited to obtain a ring-shaped heater, which is then filled with insulator. This results in reducing the contact area with GST from a full circle to a thin circular section. Fig. 9 compares cross sections and temperature map of the simulated ring (a) and lance cell (b). Both simulations refer to an applied current equal to I_m, which was about 380 μA for the ring and 520 μA for the lance cell. The reduced I_m of the former structure arises from the smaller contact area, inducing larger current density and Joule heating. However, the increased heater resistance also shifts downward the hot spot (see Fig. 4c for comparison), leading to a slightly non-optimized cell from the standpoint of the program-read trade-off.

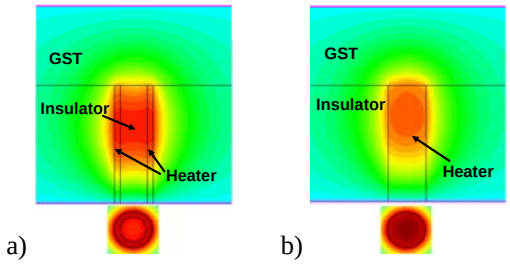


Fig. 9: Schematic vertical and horizontal cross-sections of a ring (a) and lance cell (b). Reported temperature maps are calculated at melting condition.

A comparison between lance and ring programming performances is provided in Fig. 10, reporting calculated I_m as a function of F for both cells and for isotropic and non-isotropic scaling. For both scaling approaches, the ring thickness (i.e., the thickness of the thin film defining the heater) was supposed to scale from 8 nm at F=90 nm to 3 nm at F=16 nm. The figure demonstrates the lower programming current needed by the ring cell, especially for larger nodes F. However, for small F the heater-GST contact area scales slower than in the lance cell, resulting in a slower decrease of I_m. In particular, the TiN film thickness becomes comparable with the heater diameter at the 16 nm node, making the ring structure very similar to a lance.

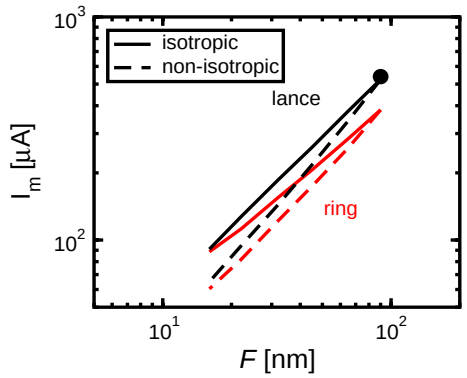


Fig. 10: Calculated programming current as a F compared in lance and ring-type cell, for isotropic and non-isotropic scaling.

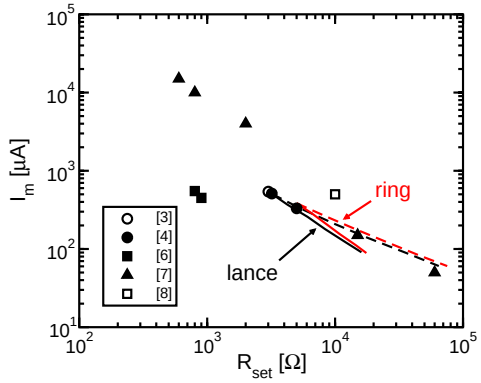


Fig. 11: I_m-R_{set} scatter plot for simulated lance and ring cells (lines), and published measured cells (symbols).

Fig. 11 shows calculated I_m as a function of R_{set}, comparing the write-read performances of lance and ring cell for the above mentioned technology nodes. For ring cell a slightly larger $I_m \cdot R_{set}$ product is achieved, due to the above mentioned non-optimized thermal profile and increased thermal losses from heater to insulator. However, the comparison with published I_m-R_{set} data from different manufacturers (symbols in the figure), probably employing slightly different materials, suggests that the $I_m \cdot R_{set}$ product can be best tuned with accurate material engineering.

The impact of material engineering in shown in Fig. 12, reporting I_m-R_{set} scatter plot of optimized 90 nm lance cells employing different materials for heater and chalcogenide. With respect to fitting parameters, the electrical resistivity ρ_c was first increased (as previously observed for N-doped GST [6]), then the thermal conductivity κ_c was reduced by the same factor, accounting for possible correlation between electrical and thermal conduction. The same approach was finally applied to the heater. From the figure, a ρ increase reduces I_m, due to increased heat generation. The most evident variation is obtained through ρ_h, since the heater preserves its high resistivity also at high temperature. On the other side, the increase of ρ_c has a larger effect on R_{set} than on I_m, worsening the $I_m \cdot R_{set}$ product. Finally, a reduction of κ has a similar impact in both materials in reducing I_m, thanks to enhanced thermal isolation of the programmable volume.

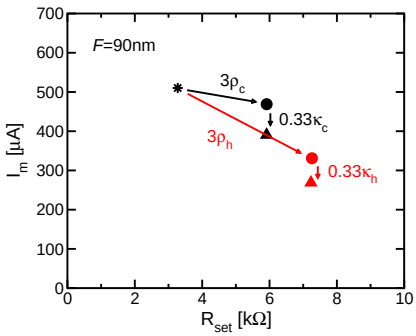


Fig. 12 I_m-R_{set} scatter plot of optimized 90 nm lance cells employing different materials for heater and chalcogenide

5. Conclusions

The PCM cell optimization issue has been analyzed in this work, outlining the role of optimum cell design for program and read cell performances. The scaling of programming current has been quantitatively addressed for lance and ring cells down to the 16nm technology node, at which the set resistance has been demonstrated to still guarantee good readout properties. Finally, geometrical and material optimization have been demonstrated for further current reduction.

Acknowledgments: This work has been partially supported by the EU within the FP6 project CAMELS (IST-3-017406).

References

[1] S. Lai, IEDM Tech. Dig., 255-258, 2003.
[2] Y. N. Hwang et al., IEDM Tech. Dig., 893-896, 2003
[3] A. Pirovano et al., IEDM Tech. Dig., 699-702, 2003.
[4] F. Pellizzer et al., VLSI Tech. Symp., 122-123, 2006.
[5] A. L. Lacaita, et al., IEDM Tech. Dig., 911-914, 2004.
[6] S. J. Ahn et al., VLSI Tech. Symp., 98-99, 2005.
[7] N. Takaura et al IEDM Tech. Dig., 897-900, 2003.
[8] T. Happ et al., VLSI Tech. Symp., 120-121, 2006.

Composition variations of nitrogen doped Ge-Sb-Te thin films and their read/write properties for phase change memories

Hyunseok Lim [a], Dohyung Kim [a,c], Gyuhwan Oh [a], Shin-Jae Kang [a], Nak-Hyun Lim [a], Yongho Ha [a], Junsoo Bae [a], Jaehee Oh [b], Insun Park [a], Hyeon-Deok Lee [a] and Joo-Tae Moon [a]

[a] Process Development Team, Semiconductor Institute, Samsung Electronics, Yongin 446-711, Korea
[b] Advanced Technology Development Team 2, Semiconductor Institute, Samsung Electronics, Yongin 446-711, Korea
[c] anselmus.kim@samsung.com

Abstract

The composition variations of nitrogen doped Ge-Sb-Te thin films for the phase change memory have been investigated as a function of the nitrogen concentration. The Ge-Sb-Te films are deposited in sputter process. The nitrogen concentration varies from 0 to 6.8%. The compositional point of nitrogen doped Ge-Sb-Te film varies perpendicular to the pseudo binary Sb_2Te_3-GeTe line from $Ge_2Sb_2Te_5$ to near GeSbTe. This is due to the volatilization of Te atoms at initial growth stage. The grain size steadily decreases with ranging up to 2.5% of nitrogen concentration and then it saturates. The XPS shows Ge_3N_4 formations. The endurance and retention properties of nitrogen doped Ge-Sb-Te films have also been investigated. The overwrite cycles of 10^6 has been achieved and memory cells sustain their bits for 36 hours at 140°C corresponding to 10 years at around 90°C.

1. Introduction

Non-volatile memories, especially flash memories, are fascinating technology for the future integrated circuits [1, 2]. However their long programming time and degradations are obstacles to their commercial success. Phase change memories based on chalcogenide materials such as Ge-Sb-Te alloy [3-5] are a prominent candidate for overcoming the limitations. The operating principle of phase change memories is based on the resistance difference between amorphous and crystalline states. The phase change between two states is reversible and fast (within several tens of nanoseconds). Recently it has been reported that the nitrogen doping enhances overwrite cyclability and decreases the operating current. Doped nitrogen into Ge-Sb-Te refines the grain size and enhances the overwrite cyclability [6-8]. And also it decreases the operating current owing to increasing resistance of crystalline state [9-11]. Most efforts with Ge-Sb-Te films are based on the composition ratio of 2:2:5 (Ge:Sb:Te) because it shows fast phase transition and stable structure.

In this article, the compositional and structural variations of $Ge_2Sb_2Te_5$ thin films as a function of nitrogen doping concentration have been investigated. Normally the composition variations occur along the pseudo binary line of Sb_2Te_3-GeTe (e.g. $Ge_2Sb_2Te_5$, $Ge_1Sb_2Te_4$ and $Ge_1Sb_4Te_7$) [12-18]. However in this study it varies nearly perpendicular to the pseudo binary line of Sb_2Te_3-GeTe from $Ge_2Sb_2Te_5$ to near GeSbTe with increase of the nitrogen concentration. In order to explain the reason of compositional variations, the

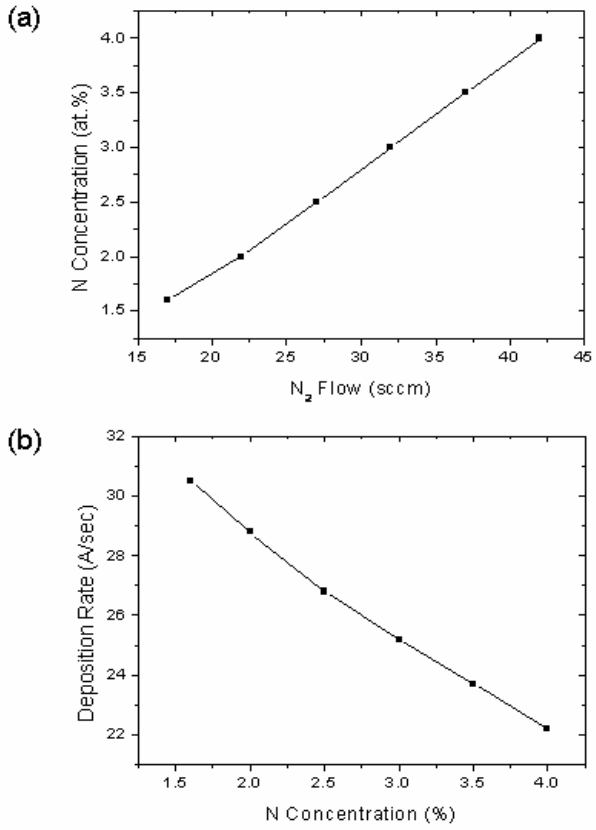


Fig. 1: (a) Nitrogen concentration and (b) deposition rate as a function of N_2 flow rate. The nitrogen concentration of 6.8% (70 sccm of N_2 flow) is not displayed in these figures.

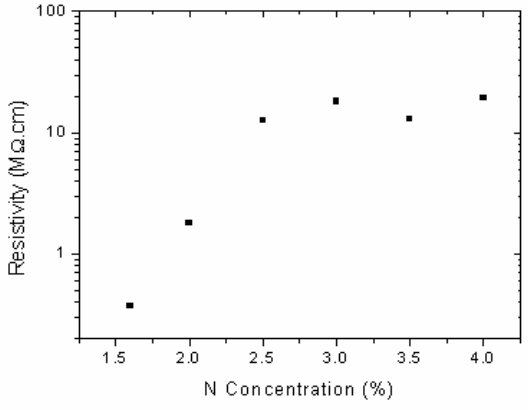


Fig. 2: The resistivity as a function of nitrogen concentration. It increases with nitrogen concentration up to 2.5% and then saturates.

structural and bonding properties have been investigated by using x-ray diffraction (XRD) and x-ray photoelectron spectroscope (XPS). Finally the capacitor module with the nitrogen doped Ge-Sb-Te films has been fabricated and its endurance and retention properties have been investigated.

2. Experiments

Nitrogen doped Ge-Sb-Te (N-GST) films with a thickness of 100nm are deposited by sputtering of a $Ge_2Sb_2Te_5$ target with N_2/Ar flow on Si substrates. The nitrogen atomic concentration of N-GST films ranges from 0.0% (undoped) to 6.8%. It is linearly proportional to the N_2 flow rate (Fig. 1a). At the same time the deposition rate decreases with increasing of N_2 flow rate (Fig. 1b). This is because the doping nitrogen retards the film formation. The resistivity of the film increases with increasing of nitrogen concentration up to 2.5% and then it saturates (Fig. 2). It means that the films are deposited as amorphous state in high nitrogen concentration (>2.5%) at given deposition temperature (i.e. 300°C). This is because the crystallization temperature increases with increasing of the nitrogen concentration [10] and the as-deposited structures of the Ge-Sb-Te films are governed with the deposition temperature.

3. Results and discussion

The chemical composition of the deposited films is determined by using x-ray fluorescence (XRF). The composition of undoped Ge-Sb-Te film is 22:22:56 at.% (Ge:Sb:Te). When the nitrogen concentration increases, the composition ratios of Ge and Sb steadily increase and that of Te decreases (Fig. 3a). Fig. 3b is the phase diagram. The composition point of nitrogen doped Ge-Sb-Te film varies nearly perpendicular to the pseudo binary line of Sb2Te3-GeTe from $Ge_2Sb_2Te_5$ to near GeSbTe with increase of the nitrogen concentration.

In order to investigate their structural variations, x-ray diffraction (XRD) analysis is used. Fig. 4 is the x-ray diffraction patterns of as-deposited Ge-Sb-Te films for the different nitrogen concentrations. For undoped Ge-Sb-Te, hexagonal peak has been detected at 39° indicating plane index (00<u>18</u>) [12]. Whereas for nitrogen doped Ge-Sb-Te, there is no hexagonal peak. Most diffraction peaks indicate that the as-deposited Ge-Sb-Te films have face centred cubic (FCC) structures. It seems that the doping nitrogen atoms prevent Ge-Sb-Te films forming hexagonal structures. The FCC structure is more profitable for the read/write properties than hexagonal structure.

In order to analyze detailed crystalline properties, (111) and (220) plane indices have been decomposed. The (111) and (220) peaks appear at 26° and 42° respectively. Fig. 5 is the intensities and widths of decomposed (111) and (220) peaks as a function of nitrogen concentration. When the nitrogen concentration increases, the peak intensities decrease. It means that the portion of amorphous state becomes higher. In other words the crystallization temperature increases when the nitrogen doping concentration increases as we confirmed

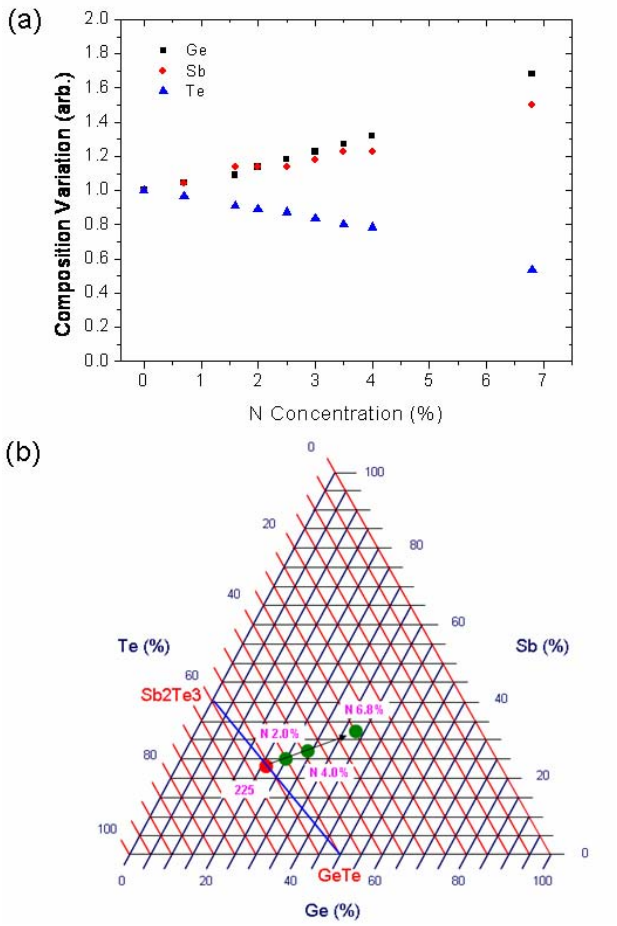


Fig. 3: (a) The chemical composition variations as a function of nitrogen concentration and (b) their phase diagram.

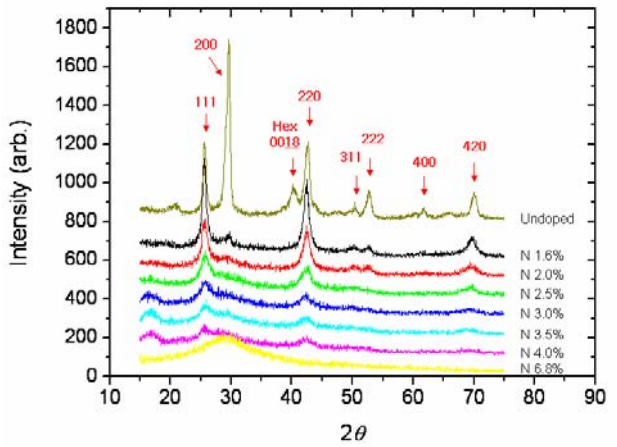


Fig. 4 : X-ray diffraction patterns of as-deposited Ge-Sb-Te films for the different nitrogen concentrations.

by their resistivity. The peak width of (220) steadily broadens.

Whereas that of (111) broadens up 2.5% of the nitrogen concentration and then it saturates. It indicates that the crystalline structure of (111) direction distorts first in the amorphous phase, but that of (220) remains its form and then steadily distorted.

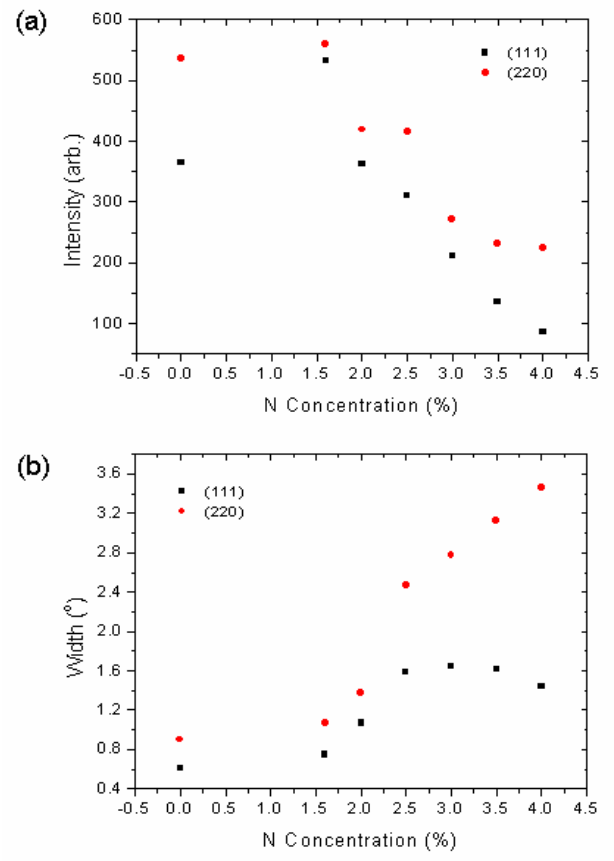


Fig. 5 : (a) The intensities of (111) and (220) peaks and (b) the peak widths as a function of nitrogen concentration.

The grain sizes are estimated from the Scherrer formula with the full width half maximum (FWHM) [19].

$$t = \frac{0.9\lambda}{\sqrt{w^2 - w_0^2}\cos\theta} \quad , \qquad (1)$$

where λ is the x-ray wavelength (i.e. 0.01 – 10nm) and θ is the incident angle with the reflecting planes; w and w_0 are peak widths of the given concentration and an instrumental broadness respectively. The principal x-ray wavelength of this study and the instrumental broadness can be estimated from undoped Ge-Sb-Te films. The scanning electron microscope (SEM) images show that the maximum grain size of undoped Ge-Sb-Te films is about 1μm and the major size is 500nm or less. Fig. 6 is the calculated maximum grain size. It decreases with increasing of nitrogen concentration and then saturates from 2.5% of nitrogen concentration. The major grain size of highly doped (> 3.0% of N) Ge-Sb-Te films is estimated 100 – 200nm. The saturation of grain size is not due to the phase transition, because the width of (220) steadily increases with increasing of the nitrogen concentration regardless of the phase transition. The decrease of grain size can be explained that nitrogen atoms are mainly located at the interstitial site [7, 20] and they prevent grains growing.

Fig. 7 is the XPS spectra of N 1s indicating normally

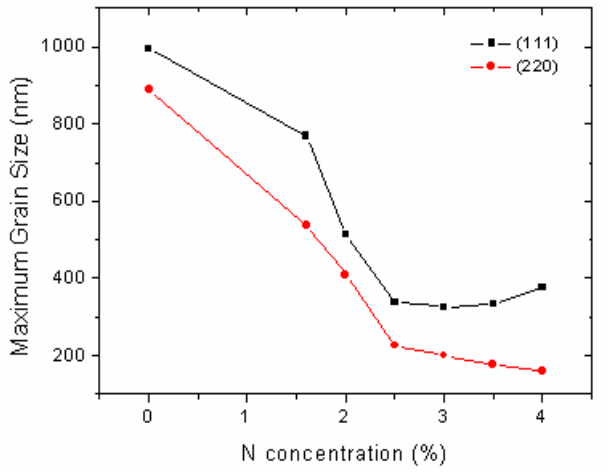


Fig. 6: Maximum grain size as a function of nitrogen concentration for the different crystal planes. The major grain size is less than half of the maximum.

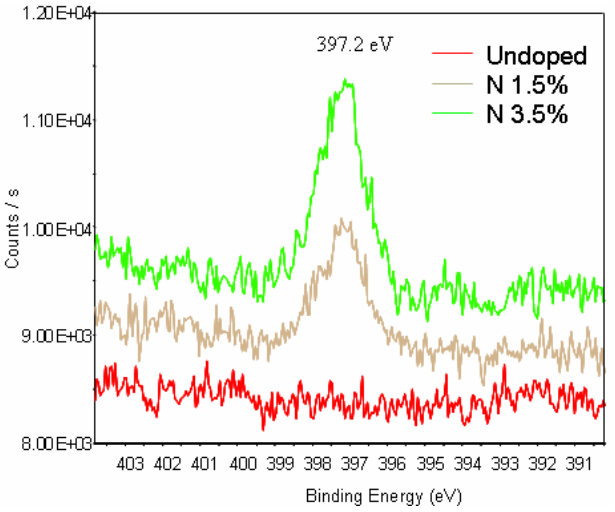


Fig. 7 : XPS spectra of N 1s indication Ge_3N_4

Ge_3N_4 [21]. The intensity of the spectrum increases with increasing of nitrogen doping concentration. At 572.5eV indicating Ge-Te formation, it slightly shifts toward to higher energy when the nitrogen concentration increases. It may be due to replacing of Te atoms with very small amount of nitrogen atoms [11] or bonding of them to vacant site of FCC structures [20]. Any other differences of spectra for doped Ge-Sb-Te films with undoped one could not be found.

Then where is diminished Te atoms when the nitrogen concentration increases ? Table 1 is the melting temperature T_{melt} of Ge-Sb-Te. That of Te is the lowest, and it can drop below 300°C (i.e. deposition temperature) when the particle size decreases below 100nm [22]. When the amount of doped nitrogen atoms increases, they retard forming stable Ge-Sb-Te structures (see Fig. 1b) and prevent grains growing (Fig. 6). Therefore some Te atoms of which melting temperature is lower than the deposition temperature volatilize at initial growth stage.

	$Ge_2Sb_2Te_5$	Ge	Sb	Te
T_{melt}	560°C	938.3°C	630°C	449.5°C

Table 1: Melting temperature T_{melt} of Ge-Sb-Te [23-24].

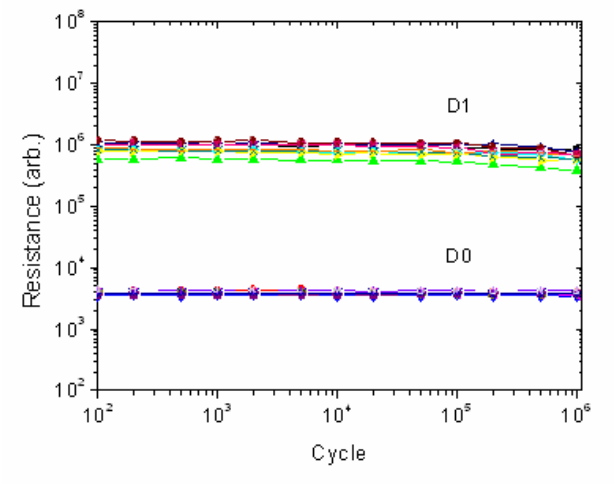


Fig. 8 : Endurance of D0 and D1 bits.

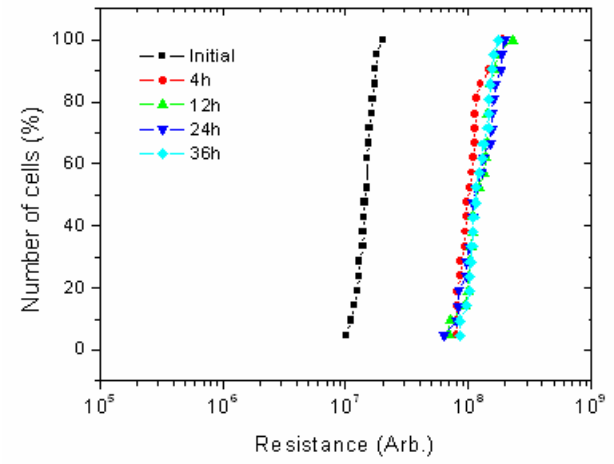


Fig. 9 : Retention characteristics. The capacitor module has been burned for 36 hours at 140°C.

The capacitor module with 3.5% of nitrogen doped Ge-Sb-Te films has been fabricated. The doping concentration has been selected in consideration of the grain size. Detailed module structure has been described elsewhere [8, 25, 26]. Fig. 8 shows the endurance of D0 and D1 bits. The overwrite cycles of 10^6 has been achieved. Whereas undoped Ge-Sb-Te fails below 10^6 overwrite cycles. It seems that the grain size is strongly related to the endurance. When the grain size decreases, the resistance of the film increases and the operating current decreases. So the Ge-Sb-Te films are less damaged from current during cycles.

Fig. 9 shows the retention characteristics. The capacitor module has been burned at 140°C. For undoped $Ge_2Sb_2Te_5$, they change to crystalline phase in 20 mins at 140°C [23]. When nitrogen doping concentration increases, the crystallization temperature increases (see Fig. 2). Therefore they can maintain their bits at higher temperature. In our case, the memory cells sustain their bits for 36 hours. It guarantees the cells to sustain their memorized states for 10 years at around 90°C.

4. Conclusion

The compositional variations of nitrogen doped Ge-Sb-Te thin films for the phase change memory have been investigated as a function of the nitrogen concentration. The compositional point of nitrogen doped Ge-Sb-Te film varies perpendicular to the pseudo binary Sb_2Te_3-GeTe line from $Ge_2Sb_2Te_5$ to near GeSbTe with increase of the nitrogen concentration. This is because doped nitrogen atoms bonded with Ge and at the interstitial state retard the film formation, and then Te atoms volatilize due to its low melting point. The small grain size at high nitrogen doping concentration confirmed by XRD supports this explanation. The XPS shows the formation of Ge_3N_4. It is believed that the doping nitrogen atoms are mainly located at the interstitial sites and also form Ge_3N_4. The endurance and retention properties of nitrogen doped Ge-Sb-Te films have also been investigated. The overwrite cycles of 10^6 has been achieved and memory cells sustain their bits for 36 hours at 140°C corresponding to 10 years at around 90°C

References

[1] M. Wuttig, Nature Materials **4**, 265 (2005).
[2] M. H. R. Lankhorst et al., Nature Materials **4**, 347 (2005).
[3] S. R. Ovshinsky, Phys. Rev. Lett. **21**, 1450 (1968).
[4] A. E. Owen and J. M. Roberson, IEEE Trans. Electron Devices **ED-20**, 105 (1973).
[5] A. Pirovano et al., IEEE Trans. Electron Devices **51**, 452 (2004).
[6] R. Kojima et al., Jpn. J. Appl. Phys. **37**, 2098 (1998).
[7] T. H. Jeong et al., Jpn. J. Appl. Phys **39**, 2775 (2000).
[8] H. Horii et al., Symposium on VLSI technology 2003, p. 177.
[9] S. M. Kim et al., Jpn. J. Appl. Phys. **44**, L208 (2005).
[10] H. Seo et al., Jpn. J. Appl. Phys. **39**, 745 (2000).
[11] B. Liu et al., Thin Solid Films **478**, 49 (2005).
[12] N. Yamada et al., J. Appl. Phys. **69**, 2849 (1991).
[13] N. Kato et al., Jpn. J. Appl. Phys. **38**, 1707 (1999).
[14] K. Nakayama et al., Jpn. J. Appl. Phys. **39**, 6157 (2000).
[15] S. Kyrsta et al., Appl. Surf. Sci. **179**, 55 (2001).
[16] M. Laurenzis et al., IEE Proc.-Sci. Meas. Technol. **151**, 394 (2004).
[17] S. Privitera et al., J. Appl. Phys. **94**, 4409 (2003).
[18] J. H. Coombs et al., J. Appl. Phys. **78**, 4906 (1995).
[19] B. D. Cullity. Element of x-ray diffraction. 2nd ed. Massachusetts: Addison-Wesley; 1978, p. 102; Special lectures of Korea Research Institute of Standards and Science.
[20] K. Kim et al., Appl. Phys. Lett. **89**, 243520 (2006).
[21] Y. Kim et al., J. Appl. Phys **100**, 083502 (2006).
[22] A. Moisala et al., J. Phys. Cond. Mat. **15**, S3011 (2003).
[23] D. Kim et al., Jpn. J. Appl. Phys. **42**, 5107 (2003).
[24] D. R. Lide. CRC handbook of chemistry and physics. 79th ed.: CRC press (1998)
[25] J. H. Oh et al., International Electron Devices Meeting 2007, S2P6
[26] Y. J. Song et al., Symposium on VLSI technology 2006, p. 118.

SESSION C

FinFlash

FinFET SONOS non-volatile memory arrays

D. S. Golubović, N. Akil, M. van Duuren, A. H. Miranda and R. van Schaijk

NXP Semiconductors, Kapeldreef 75, B - 3001 Leuven, Belgium

Abstract

FinFET SONOS non-volatile memories with a minimum fin width of 15 nm and gate length of 20 nm have been fabricated by using DUV 193 nm optical lithography combined with dry etch. By measuring the programme/erase curves of NOR arrays with 256 bits, the cumulative V_T – distributions, as well as their endurance and retention properties, we demonstrate that FinFETs in combination with the SONOS concept offer a highly scalable solution for high-density non-volatile memories.

1. Introduction

Multi-gate FETs and in particular FinFETs have a better immunity to short channel effects, a steeper sub-threshold slope and potentially higher channel mobility compared to their planar counterparts, which makes them a promising candidate for further CMOS scaling [1,2].

Currently available floating gate (FG) non-volatile Flash memories (NVMs) face scaling challenges related to high programme/erase voltages, reduced capacitive coupling between the control and floating gate, cross-coupling of adjacent memory transistors due to a smaller pitch, punch-through associated with channel hot electron programming and so forth [3, 4, 5]. Given that the demand for high density Flash memories has been steadily increasing over the past years, a great deal of effort has been focused into finding a scalable replacement for FG NVMs.

Silicon-Oxide-Nitride-Oxide (SONOS) memories have extensively been investigated over the past couple of years as a scalable alternative for FG Flash memories [4, 5]. The main advantages of the SONOS concept are reduced programme/erase voltages, a better endurance, absence of erratic tail bits, a higher radiation tolerance and, in view of embedded applications, better CMOS compatibility, owing to its single poly integration. Recently, a 32 Gb SONOS-based NAND Flash memory has been demonstrated [6].

A combination of the inherent scaling advantages of the FinFET concept with the SONOS scalability in the context of NVMs may well provide an efficient route for deep sub-micrometre Flash scaling [3, 7].

In this paper we report on tri-gate FinFET SONOS memory arrays in NOR configuration, which have been fabricated by using advanced, but standard manufacturing processing steps for the 90 nm CMOS generation and beyond. This makes it possible to assess the manufacturability of FinFET SONOS devices and establish whether the combination of the FinFET and SONOS scaling concepts improves the scaling perspective of high density NVMs. Equally importantly, the fabrication of NVMs with standard manufacturing steps is particularly favourable for embedded Flash applications, where baseline compatibility facilitates easier integration and lower fabrication costs.

FinFET SONOS NOR arrays with 256 bits, having 40 nm wide fins and 50 nm gate length, have been investigated. In view of low power applications, the arrays were programmed /erased using direct tunnelling. In addition to the programme/erase (P/E) characteristics, the endurance, retention, as well as cumulative V_T – distributions are also shown. We demonstrate that FinFETs in combination with SONOS are a prominent candidate for deep sub - micrometre scaling of NVMs.

2. Fabrication

A 60 nm high fin on an 8” SOI wafer with 400 nm buried oxide was defined using 193 nm DUV lithography and dry etching. The width of the fin was shrunk around 80 nm compared to the lithographically defined width by means of resist trim. After the surface of the fin was healed by sacrificial oxidation and Hydrogen anneal to remove the damage caused by the dry etch step, as well as to round its corners, an Oxide-Nitride-Oxide (ONO) tri-layer and 100 nm poly-Si gate were deposited. The thickness of the bottom (tunnel) oxide is t_{ox} = 2 nm, the thicknesses of the Nitride trapping layer equals 6 nm and the top Oxide thickness is 8 nm. The fin was left undoped, whereas the poly-Si was implanted with Phosphorus prior to the gate definition. Similarly to the fins, the gates were patterned using 193 nm DUV lithography and dry etching, which in this case included a trimming step of approximately 50 nm. The Nitride spacer formation was followed by the source/drain implantation and Ni - silicidation. The FinFET memory arrays were contacted up to the metal-1 level with standard W-contact and Al-metallisation modules. The poly-Si gates are used as the wordlines in the NOR arrays, common source lines are realized with silicided active lines, whereas the Al-lines serve as the bitlines.

Fig. 1 shows a top view scanning electron microscope image of the active lines in an array after the fin formation. The white bar denotes 200 nm. The wide horizontal Si - lines in Fig. 1 are the common source lines, whereas the round features along the fins are used for the drain contacts. In the array shown in Fig. 1 the width of the fin has been reduced from lithographically defined 120 nm to 40 nm.

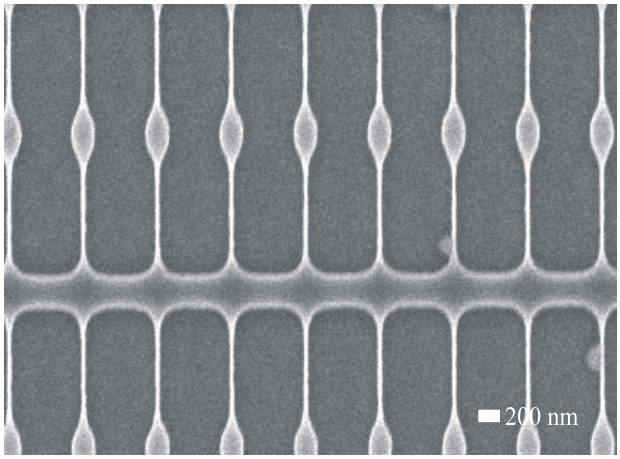


Fig. 1 : Top view scanning electron microscope image of the active lines in an array after the fin formation. Wide horizontal lines are used as common source lines, whereas the round features along the fins are intended for drain contacts. The white bar denotes 200 nm.

3. Results and discussion

Throughout the paper the electric characteristics of the FinFET SONOS devices have been investigated on NOR mini-arrays with 256 bits. The width of the fins in the selected arrays is 40 nm, whereas the gate length is 50 nm.

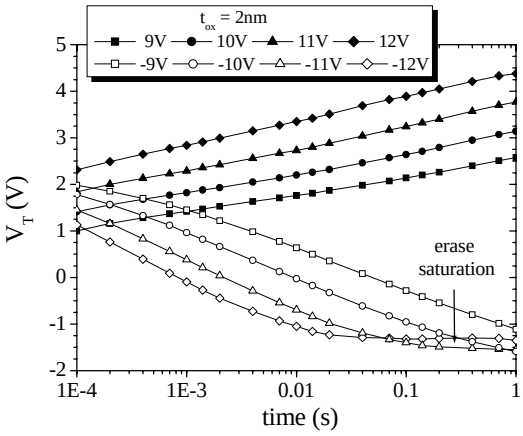


Fig. 2: Programme/erase curves of a FinFET SONOS mini-array with 256 bits. The width of the fins is 40 nm, whereas the gate length equals 50 nm. The programme/erase voltages from ±9 V to ±12 V were used. Plotted is the average V_T over the entire arrays.

Fig. 2 shows programme/erase (P/E) curves of a FinFET SONOS mini-array. Programmed threshold voltages (V_T) are given with filled symbols, whereas the erased threshold voltages are shown with open symbols. The devices were programmed/erased with voltages ranging from ±9 V to ±12 V by direct tunneling of carriers through the bottom oxide, in order to facilitate low power operation. The threshold voltages were measured using a current criterion of 5 μA at a drain voltage of 0.5 V. Unselected wordlines were biased at a negative voltage in order to prevent over-erase. A V_T – window of around 2 V is obtained by using 10ms long pulses of ±10 V. An increase in the erase voltage does not proportionally decrease the erased V_T, due to the well - know phenomenon of erase saturation, which is present in SONOS devices with n-type gates, and sets in when, under a negative voltage applied to the gate, the tunneling of holes from the channel to the Nitride becomes compensated by the tunneling of electrons from the poly-Si gate to the Nitride layer. The erase saturation of FinFET SONOS arrays is similar to that of planar SONOS devices with the same ONO stack [4,5].

Fig. 3 shows the cumulative V_T – distributions of the fresh, programmed and erased FinFET SONOS arrays. The horizontal arrow indicates the available V_T - window. The arrays were programmed by 10 ms long 10 V and erased by 100 ms, -10 V pulses. The P/E V_T – distributions were measured shortly after the programme/erase. A very good V_T – uniformity indicates a good process control. Furthermore, in line with known SONOS features, no erroneous tail bits are observed.

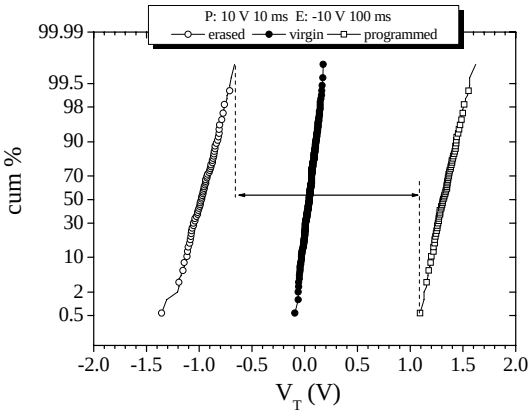


Fig. 3: Cumulative V_T – distribution of 256 bit mini-arrays programmed by 10 ms long 10 V pulses and erased by -10 V 100 ms pulses. The arrow indicates the V_T – window.

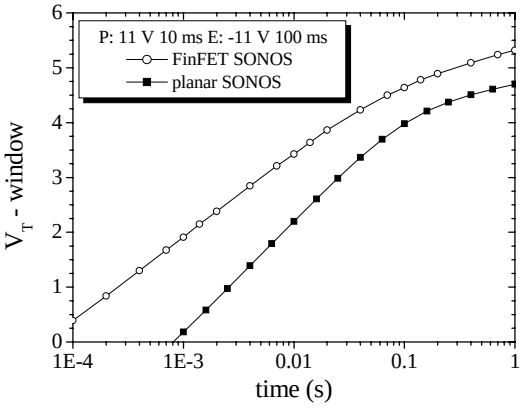


Fig. 4: A comparison of the V_T – window of the FinFET SONOS (open symbols) and planar SONOS (filled symbols) programmed/erased by ±11 V.

Fig. 4 comparatively shows the difference between programmed and erased threshold voltages (V_T – window) of the FinFET SONOS devices (open symbols) and planar SONOS devices (filled symbols) with the same ONO – stack obtained by using ±11 V P/E pulses. The FinFET SONOS arrays program/erase much

faster than planar SONOS devices with the same ONO – stack.

Fig. 5 shows the endurance of a mini-array, measured up to $3 \cdot 10^6$ cycles. The mean (filled symbols), maximum (open symbols) and minimum (crossed symbols) values of the programmed and erased threshold voltages are presented. 10 ms long 10 V pulses were used to write the devices, whereas 100 ms pulses of -10 V were used to erase the array. Note that the erase voltage has been chosen so as to avoid erase saturation (see P/E curves in Fig. 2), since this is essential, both for planar and FinFET SONOS devices, to obtain a good endurance. The good endurance characteristics of planar and FinFET SONOS devices are a result of the fact that the programming occurs through a thin bottom oxide, whose degradation is predominantly caused by break down, as opposed to the FG devices with a thicker (~ 8 nm) tunneling oxide, which gradually wear out due the charge transport through the tunnel oxide.

Fig. 6a) and 6b) show the room temperature retention data. Curves given by squares were obtained by using 10 ms 10 V pulses for writing, curves with triangles show the retention after programming with 10 ms long 12 V pulses, whereas the retention in the erased states (circles) is measured upon applying 100 ms, -10 V pulses.

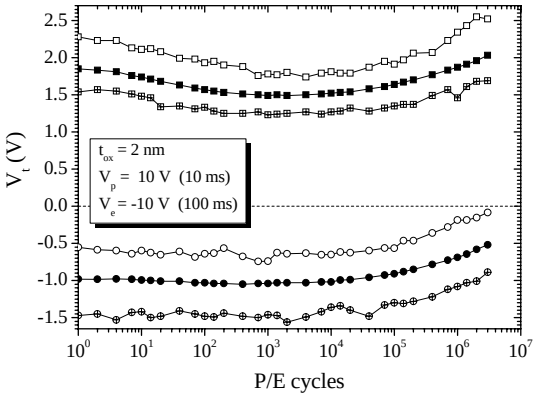


Fig. 5: Endurance curves up to $3 \cdot 10^6$ cycles of 256 bit mini-array programmed with 10 ms pulses of 10 V and erased by -10 V pulses of 100 ms.

In Fig. 6a) the retention data after 1 programme/erase cycle is shown, whereas Fig. 6b) shows the retention after 10^5 cycles. The average V_T – values are given by symbols, whereas the vertical dotted line marks the 10 year limit. The retention has been measured during approximately two days. Both curves have been extrapolated to 10 years, as shown by dashed lines in Fig. 6a) and 6b). Retention curves in Fig. 6a) show that, similarly to the planar SONOS devices, the remaining average V_T – window after 10 years, depends upon the initial difference in the programmed and erased threshold voltages. The initial V_T – window of 1. 9 V obtained by ±10 V is extrapolated to reduce down to 0.2 V after 10 years, whereas the initial V_T – window of 2.8 V obtained by using 12 V programming pulses reduces after 10 years to 0.7 V, which is more than three times greater. This behaviour is caused by the character of the charge trapping in ONO - layers, and the fact that the charge in the Nitride layer, located close to the thin bottom oxide, escapes by direct tunneling shortly after programming. Thus, the retention is predominantly determined by the charge captured deeper in the Nitride layer and consequently, the greater the initial V_T – window, the better the retention. Fig. 6b) shows that the cycling has virtually no detrimental influence on the retention. The same retention features have been observed in the planar SONOS devices, as well.

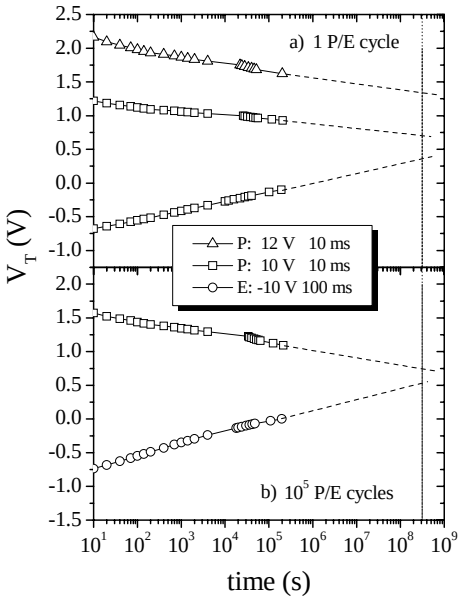


Fig. 6: a) The room temperature retention of 256 bit mini-array after 1 P/E cycles. The curves with squares are obtained by 10 ms long 10 V pulses, whereas the curves with triangles are obtained after programming with 10 ms 12 V pulses. Curves with circles show the retention in the erased state upon applying 100 ms -10 V pulse. b) The room temperature retention after 10^5 cycles. The programming was carried out by 10 ms 10 V pulses and erase by 100 ms -10 V pulses.

4. Conclusions

256 bit NOR FinFET SONOS arrays have been fabricated using standard, commercially available processing steps. Programme/erase characteristics, along with the cumulative V_T – distributions, as well as endurance and retention of the devices with 40 nm wide fins and 50 nm gate length have been presented. It has been demonstrated that the FinFET concept can be efficiently combined with the scaling potential of SONOS memories for deep sub - micrometre Flash memory scaling.

References

[1] H. S. P. Wong, IBM J. Res. Dev. **46**, 133 (2002).
[2] R. Ritzenthaler *et al.*, *Proceedings of ESSDERC '05*, 81 (2005).
[3] S. K. Sung *et al.*, IEEE Tran. Nanotech. **5**, 174 (2006).
[4] N. Akil *et al.*, IEEE Tran. Elec. Dev. **52**, 492 (2005).
[5] R. van Schaijk *et al.*, Solid State Elec. **49**, 1849 (2005).
[6] Y. Park *et al.*, *IEDM Tech. Digest*, 29 (2006).
[7] F. Hofmann *et al.*, Solid. State. Electron. **49**, 1799 (2005).

Corner enhancement of FNT program/erase operations in nitride storage FinFLASH devices.

L. Breuil, M. Rosmeulen, J. Loo, A. Furnémont, L. Haspeslagh and J. Van Houdt

IMEC Kapeldreef 75, 3001 Leuven, Belgium
Breuil@imec.be

Abstract

This paper investigates the Program / Erase behavior by Fowler Nordheim Tunneling of nitride storage FinFLASH devices. Transient characteristics and IdVg characteristics of the devices in program and erase states show that carrier injection takes place preferably at the corner parts of the fin where the field is enhanced.

1. Introduction

The FinFLASH cell architecture is a promising solution for sub-50nm NAND Flash memory scaling due to its strong electrostatic gate control over the channel region, which improves the short channel effect behavior of the transistor and increase the read current [1-4]. Furthermore, the use of an ONO stack for charge storage removes the floating gate coupling and SILC issues. However, as the charge is stored in a non-conductive medium in a non planar structure, inhomogeneous charge distributions can be expected because of enhancement of the injecting field on corner regions of the fin, as suggested in [5]. In this paper, we investigate the program / erase behavior by Fowler-Nordheim Tunneling (FNT) of a nitride storage FinFLASH device, and show evidence for preferential operations at the corners of the fin.

2. Device structure and processing

The FinFLASH device consists of a FinFET cell where the gate oxide is replaced by an ONO stack for charge storage in the nitride layer. Wafers with a 65 nm SOI layer on 150 nm Buried Oxide are used. The Fin is patterned using a TEOS Hard Mask (HM) of 60 nm. The fins are 60 nm high, and a fin width down to 30 nm can be obtained by HM trimming. Several dimensions for the fin width are available on mask. A curing of the fin is performed in order to reduce its roughness and for corners rounding. Well implantation is done by a triple boron implant. The ONO stack is then deposited. A 4nm Tunnel oxide (Tox) is deposited by ISSG in order to have good top-sidewall uniformity. The nitride layer of 5 nm is deposited by LPCVD, and the top (blocking) oxide is deposited with a thickness of 5nm by LPCVD, followed by a wet re-oxidation.
The gate is formed by deposition of 200 nm a-Si, which is etched back to 100 nm. This procedure allows a smoother topography. Then, trimming of the HM (60nm TEOS) is performed. A broad range of Lg dimensions is available on mask, and Lg down to 50 nm can be obtained. After etching of the gate, the ONO stack is etched using a Wet-Dry-Wet etch sequence. Then, halos and extensions are formed. The implant conditions were tuned by TCAD toward a sufficiently high fin Vt in narrow fins, and a conformal junction in the width cross section of the fin. Then, nitride spacers are formed with a thickness of 100 nm, and the HDD are implanted. Silicidation is performed with 5 nm Ni, followed by PMD, Contact and Metal 1 modules. Fig. 1 shows a SEM picture of a fin of 30nm width and 60nm height (left), and a cross TEM picture of the ONO layer. Fig. 2 shows a SEM picture of a FinFLASH device after spacer module.

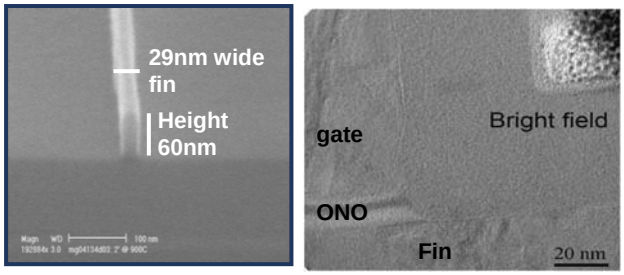


Fig. 1: SEM picture after fin definition (left). Cross TEM picture after ONO etching (right).

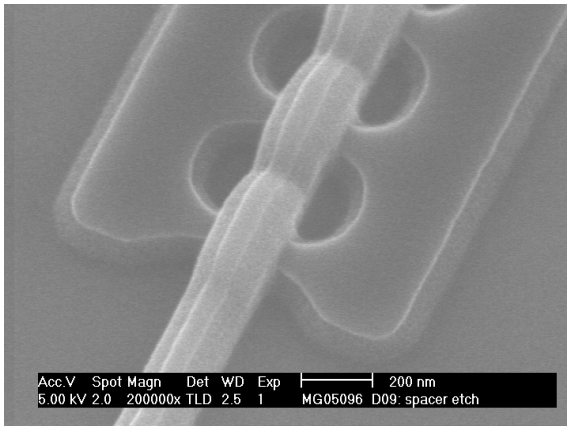


Fig. 2: SEM picture of a multi-fingers FinFLASH cell after spacer etch.

3. FinFLASH transistor characteristics

The FinFLASH structure allows greatly improving the Short Channel Effects (SCE) characteristics. Fig. 3 shows the IdVg curves of devices of 50nm gate lengths for different fin widths. Both sub-threshold slope and DIBL are strongly improved when the fin width is reduced. Despite the good behaviour of a Lg=50nm and W=30nm cell, this study was performed on Lg=100nm devices, in order to be able to investigate devices with greater fin widths, which would show bad characteristics at Lg=50nm. The characteristics of the Lg=100nm devices are shown in fig. 4. A decrease of the Vt is

observed as the fin width decreases. This is due probably to channel dopants diffusing out of the fin during processing.

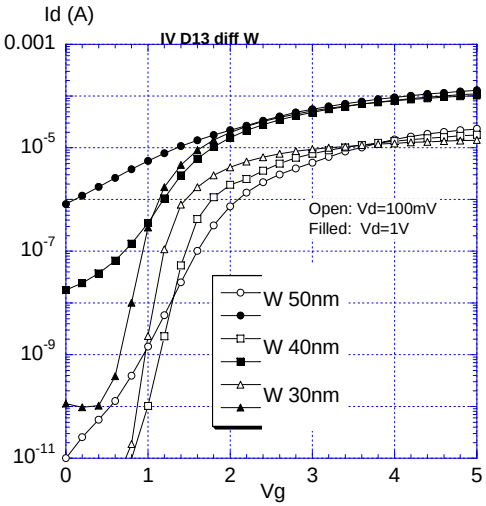


Fig. 3: IdVg characteristics of Lg=50nm FinFLASH cell as a function of the fin width W. Open symbols: Vd=100mV, Filled symbols: Vd=1V.

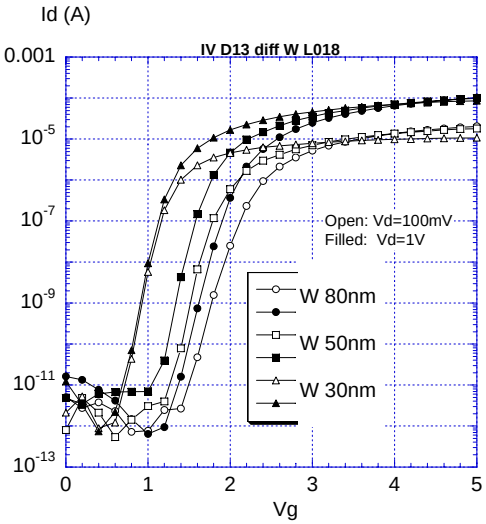


Fig. 4: IdVg characteristics of Lg=100nm FinFLASH cell as a function of the fin width W. Open symbols: Vd=100mV, Filled symbols: Vd=1V.

4. FNT program / erase behavior

The devices were programmed and erased by Fowler-Nordheim Tunneling (FNT), using ± 14V gate voltage pulses with durations of 100ms. Fig. 5 shows the IdVg characteristics of the device in virgin, programmed and erased states, for drain voltages of 100mV and 1V. The characteristic of the programmed state is identical to the one of the virgin state, with only a parallel shift. However, the characteristic of the erased state is strongly distorted, and the ON current of the cell is reduced, although the cell has been erased back to the same Vt as the virgin state. This suggests an inhomogeneous charge distribution in the nitride after erase.

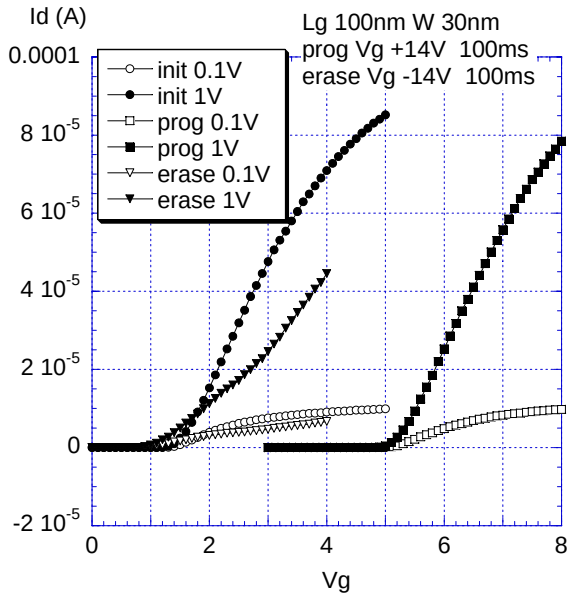


Fig. 5: IdVg characteristics of the FinFLASH cell in virgin, programmed and erased states

Transient program / erase characteristics have been performed alternatively on a fresh cell. The program and erase curves are shown on fig. 6 and 7, respectively. It is clearly seen that the first programming operation is slower than the subsequent ones, although the same Vt is obtained after a sufficiently long time. The erase curves however, have all the same shapes. A saturation of the erase Vt is observed.

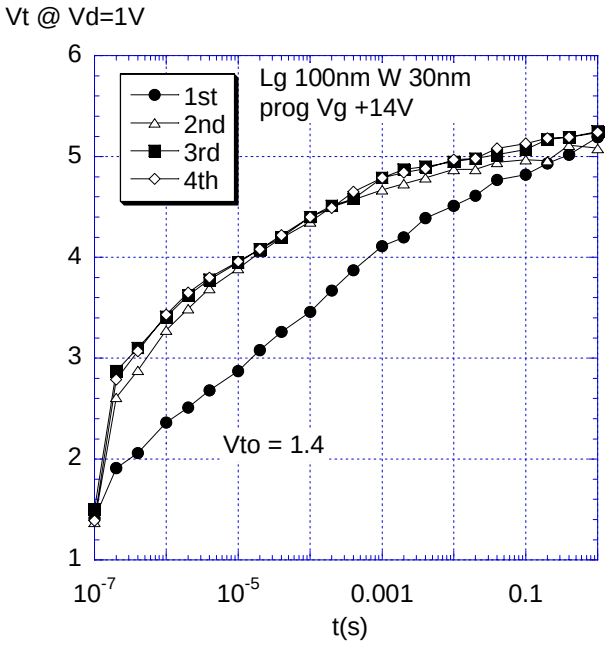


Fig. 6: Successive programming characteristics.

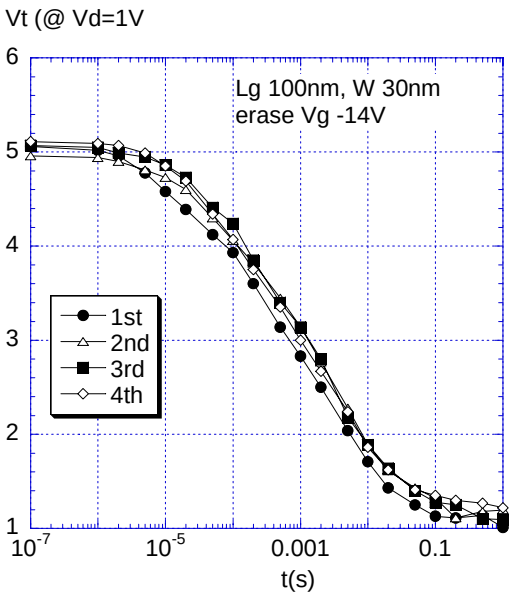


Fig. 7: Successive erase characteristics

The IdVg characteristic of this device has been measured after each transient, and is presented in fig. 8. All characteristics are reproducible, except that the virgin characteristic can not be recovered upon erase operations.

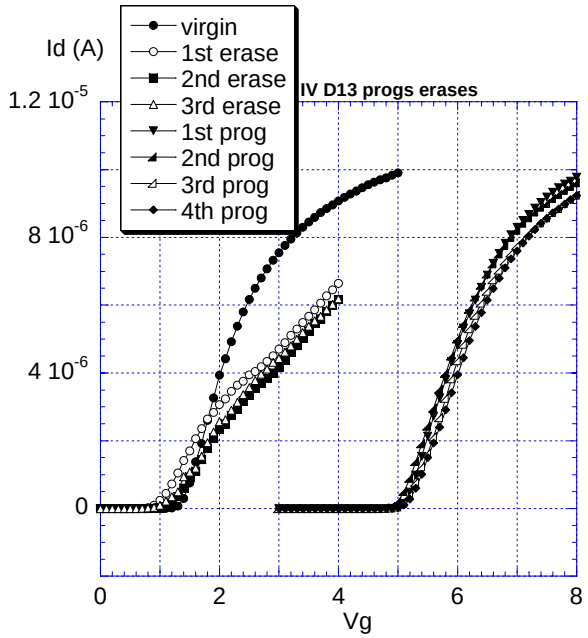


Fig. 8: IdVg characteristics of the Fin FLASH devices after alternative program / erase operations.

The observations of fig. 6, 7 and 8 can be explained by preferential FNT program / erase at the corners of the fin. Starting from a fresh device, the entire nitride layer needs to be filled by electrons in order to raise the Vt of the cell. The phenomenon is limited by filling of the planar parts of the fin where injection is slower. Then, for erase, it is sufficient to erase the parts at the corner of the fin to lower the Vt. However, this leads to an inhomogeneous charge distribution in the width section of the device with electrons remaining on the planar parts. Therefore, we see a distortion of the IdVg characteristic for all erase states, and a reduction of the ON current. Then, for the subsequent program operation, electrons need to be injected only in the corner parts of the fin. This is why the subsequent programming operations are always faster than the first one.

5. Effect of Fin width.

The IdVg characteristics of virgin and erased states have been further compared on devices with different fin width of 50, and 80 nm (fig. 9). The maximum current is higher in the wider device in virgin states because of larger channel contribution. In erased states however, the maximum currents are similar. This shows that mostly the corners contribute to the channel current in the erased state, so that the difference in fin width does not play a role anymore.

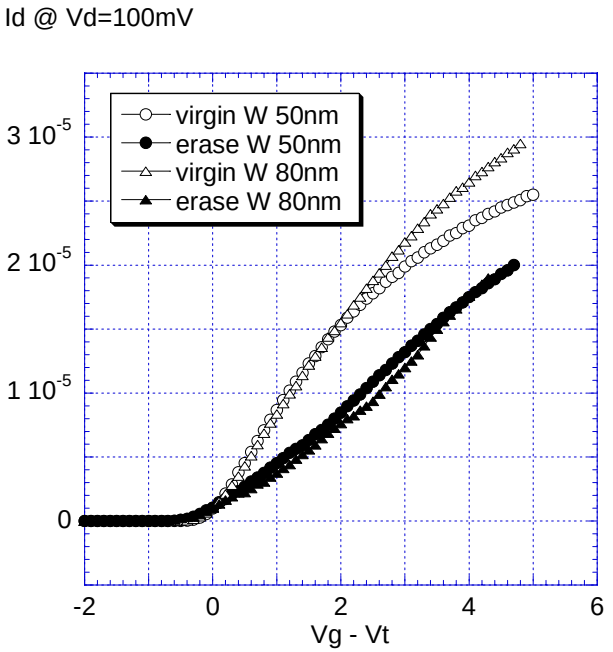


Fig. 9: IdVg characteristics of FinFLASH devices of 50nm and 80nm fin width in virgin and erase states.

6. Conclusion

FNT operation of FinFLASH devices using nitride storage has been investigated. These devices show a strong improvement of short channel effects and efficient program / erase. The analysis of these operations shows that carrier injection occurs preferably at the corners where the injecting field should be more important. This leads to an inhomogeneous carrier distribution is erased state, but homogeneous electrons distribution in the program state. This situation is expected to bring retention issues, which will be further investigated.

Acknowledgements

This work has been performed within the IST FinFLASH project.

References

[1] S-K. Sung et al. IEEE proceedings of VLSI **2006**.
[2] P. Xuan et al. IEEE proceedings of IEDM, **2003**

[3] S-K Sung et al. Nanotechnology, IEEE Transactions on Volume 5, Issue 3, May **2006** Page(s):174 - 179

[4] Y. Ahn et al. IEEE proceedings of VLSI **2006**

[5] G. Fiori et al. Nanotechnology, IEEE Transactions on. Volume 4, Issue 3, May **2005** Page(s):326 - 330.

Program / erase characteristics of ultra-scaled Si Nanocrystal FINFLASH memories

S. Lombardo[a], C. Gerardi[b], D. Corso[a], G. Cina'[b], E. Tripiciano[b], V. Ancarani [b], C. Bongiorno[a], E. Rimini[a] and M. Melanotte[b]

[a] CNR-IMM, Stradale Primosole 50, 95121 Catania, Italy. E-mail: salvatore.lombardo@imm.cnr.it
[b] STMicroelectronics, Stradale Primosole 50, 95121 Catania, Italy.

Abstract

Si nanocrystal FINFLASH memory cells were realized on SOI substrates. Ultra-scaled devices down to 20 nm fin width exhibit excellent program / erase characteristics at low voltage. The main results are shown and discussed.

1. Introduction

Memory cells for Flash-type applications with a FinFET architecture (FINFLASH) are a promising approach for Flash scaling. In fact, the FinFET double and trigate architectures allow very large drive currents and superior performances in terms of short channel effects. In addition, the use of discrete storage nodes in planar Flash-like devices permits a reduction of floating gate interferences, reduced thickness of tunnel and control dielectrics, and therefore lower operation voltages [1]. Hence, the implementation of the discrete trap memory concept with the FinFET architecture represents a viable approach for Flash scaling [2-4].

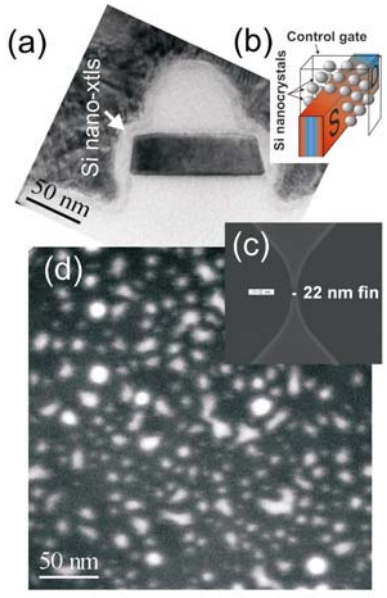


Fig. 1: (a) is a cross-sectional TEM micrograph orthogonal to the Fin direction (see (b)) of a Si nano-xtl FINFLASH. (c) is an SEM plan-view image showing a 22 nm fin active area, and (d) is an energy-filtered plan-view TEM image of a Si nanocrystal layer.

We have realized ultra-scaled discrete trap FINFLASH cells in which the storage medium was realized by chemical vapor deposition (CVD) of Si nanocrystals on the tunnel oxide grown of the fin sides. In this work we report on the program / erase performances of these devices.

2. Device fabrication

FINFLASH memory cells were realized on SOI wafers. The FinFET double gate structure was obtained by using a thick top dielectric, so that inversion channel is effectively formed at the fin sides. The charge storage medium consists in an array of Si nanocrystals formed by CVD. Tunnel and control oxide have a 4 and 10 nm EOT, respectively, and either a high temperature oxide (HTO) or an oxide-nitride-oxide (ONO) were used as control dielectric. In some cases boron halo implants were used to locally increase the channel electric field. Fig. 1(a) reports a cross-sectional TEM image of the structure along the direction orthogonal to the fin (see the inset (b)). The thicker top oxide and the Si nanocrystal layer are clearly evident. Fin widths and channel lengths down to 20 nm (Fig. 1(c)) and 40 nm, respectively, were defined. Fig. 1(d) shows an example of a Si nanocrystal layer used in the FINFLASH devices. Various CVD conditions were examined with Si nanocrystal average radii in the range between 3 and 4 nm, and Si dot surface coverages between 8 and 19 %.

3. Device characteristics

Devices are functional, very robust with respect to short channel effects, and with very large drive currents. Details on these aspects can be found in [5].

Program / erase characteristics depend strongly on the structure of the nanocrystals and on the type of dielectrics used in the gate stack. Figs. 2 and 3 show examples of Fowler-Nordheim tunneling program / erase characteristics of FINFLASH cells of comparable size (≈40 nm fin width, W, and ≈200 nm channel length, Lch). Left and right plots refer, respectively, to devices with low and high Si coverage, and with HTO and ONO control dielectrics. When HTO and low Si dot surface coverage are used, the maximum obtainable threshold voltage (VT) window is reduced, since by increasing program voltage above ≈15 V, the program characteristics worsen.

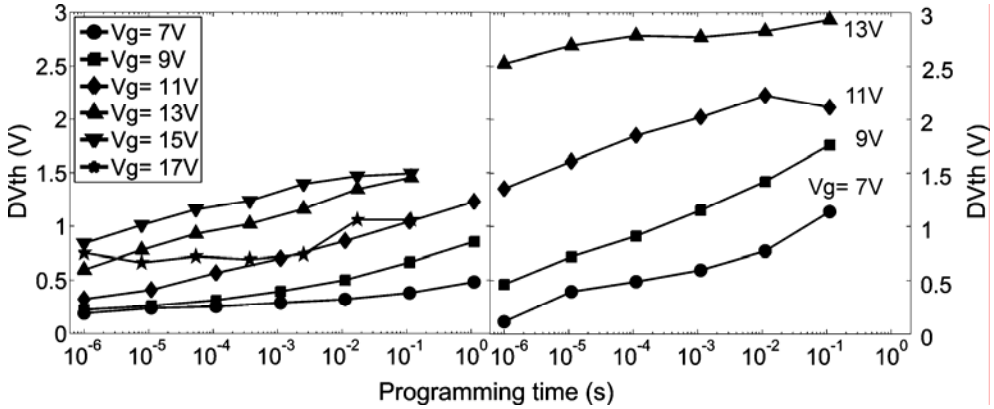


Fig. 2: Fowler-Nordheim tunneling program characteristics of FINFLASH cells of comparable size (about 40 nm fin width, W, and 200 nm channel length, Lch). Left and right plots refer, respectively, to devices with low and high Si coverage, and with HTO and ONO control dielectrics.

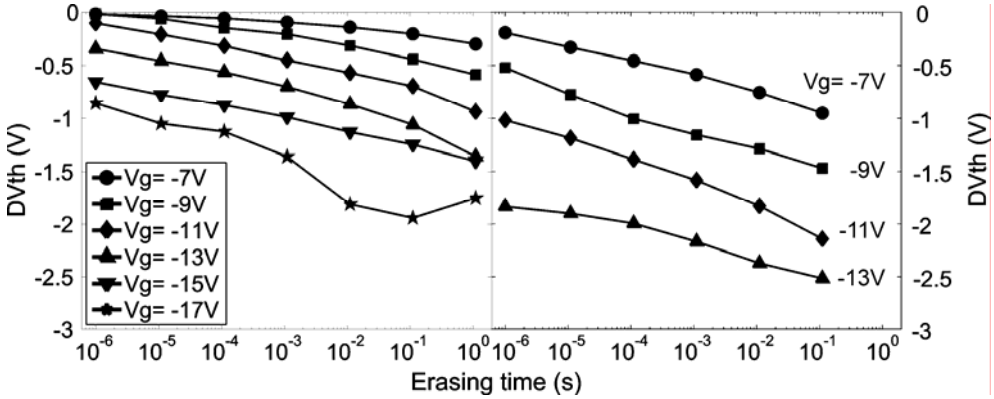


Fig. 3: Fowler-Nordheim tunneling erase characteristics of the FINFLASH cells of Fig. 2.

The worsening at high voltage is attributed to top side injection of holes as the n+ poly-Si gate goes into inversion, or to valence band electron injection from the Si dots into the gate. To improve the VT window, the increase of surface coverage and also the use of ONO play a very positive role, as demonstrated by the data of Figs. 2 and 3, right plots. An improved VT window is evident, with high speed and low voltage operation possible, down to values as low as 13 V.

In addition to excellent performances for Fowler-Nordheim program / erase, also channel hot electron injection (CHEI) programming is possible, even in the case of ultra-short channels, as shown below.

Fig. 4 reports an example of the dependence on time and bias of channel hot electron programming in the case of a FINFLASH with low nanocrystal density, no halo, and HTO control oxide. (a), (b), and (c) refer to the case of a device with Lch ≈ 260 nm and W ≈ 60 nm with VDS equal to 3, 4, and 5 V, respectively during the programming pulse. It is evident that even at low voltage, below the theoretical value of 3.2 V corresponding to the SiO_2-Si barrier height due to conduction band offset, CHEI programming is possible. This suggests the occurrence of programming by warm electrons, i.e., by tunneling of carriers warmed by the relatively high drain voltage.

On the other hand, by increasing the voltage at VDS = 5 V, programming efficiency noticeably improves. We propose that such effect is produced by the combination of two phenomena: an avalanche breakdown at the drain and a floating body effect. Note that simply the effect of a floating body bias would go in the opposite direction of reducing the programming efficiency, since as VDS is increased the body bias would increase thus reducing the CHE efficiency improvement due to the VDS increase. Fig. 4(c) indicates the opposite, i.e. an efficiency improvement at VDS = 5 V above ≈ 1 ms pulses. This is consistent with another observation: Fig. 5 shows the time dependence of the drain current under various bias conditions. At low VDS values (below 4 V) ID decreases with time. This indicates the occurrence of channel hot / warm electron programming, since as the devices gets programmed, VT shifts in the positive direction and as a consequence ID decreases. However, at VDS = 5 V, though at low times the ID decrease trend is still observed, for times of the order of 1 ms and above the situation changes. ID has an abrupt increase due to the build-up of an avalanche at the drain. So, the sudden improvement of the CHEI programming at VDS = 5 V (Fig. 4(c)) is concomitant to a drain avalanche. The relatively long characteristic time necessary for the avalanche build-up, of the order

of 1 ms, is consistent with an effect of hole recombination [6], which decreases the body bias, increases the effective drain-body voltage drop, and triggers the drain avalanche.

Another aspect concerns the dependence of CHEI on geometry: among the various parameters, the most important is the channel length. Fig. 4(d) reports the case of a cell with Lch ≈ 80 nm and W ≈ 60 nm programmed with VDS of 4 V. These characteristics can be readily compared to the case of Fig. 4(b), where Lch ≈ 260 nm. The short channel device is much more efficient in CHEI programming.

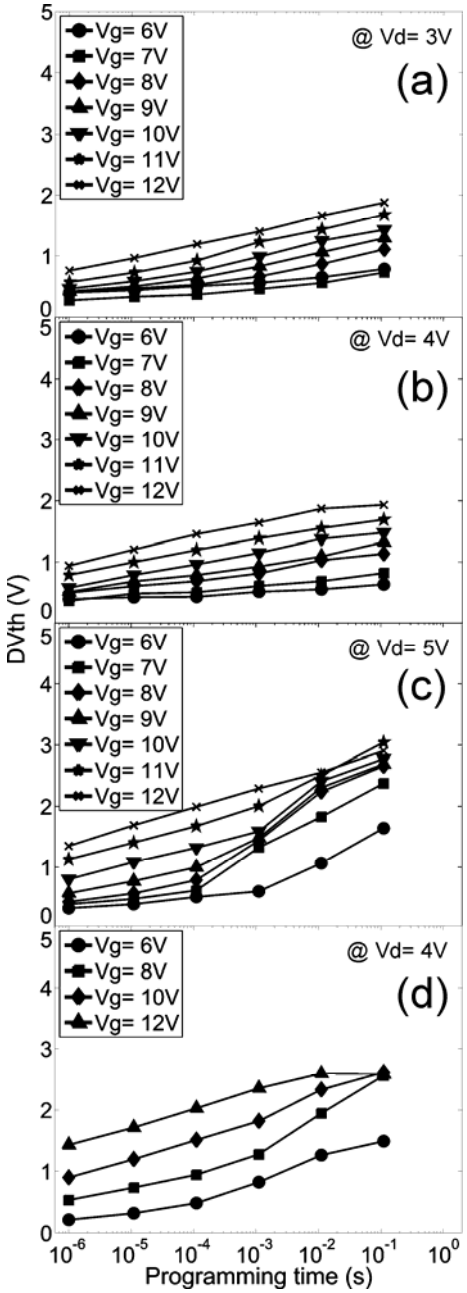


Fig. 4: Time and bias dependence of CHEI programming in a FINFLASH cell with low nanocrystal density, no halo, and HTO control oxide. (a), (b), and (c) refer to a device with Lch ≈ 260 nm and W ≈ 60 nm with VDS equal to 3, 4, and 5 V, respectively. (d) reports the case of a cell with Lch ≈ 80 nm and W ≈ 60 nm programmed with VDS of 4 V.

By some optimization of the parameters (nanocrystal deposition, use of appropriate halo doping, ONO control dielectric, etc.) we can obtain further noticeable improvements of the CHEI programming characteristics. As an example, Fig. 6 shows the CHEI programming characteristics of a cell with 40 nm channel length, 20 nm channel width, 20 nm fin height, halo doping, medium density nanocrystals. In this case operation at low voltage is possible. VT shifts in excess of 3 V with 1 μs pulses can be achieved with VDS of 4 V and VG of 8 V.

The noticeable CHEI programming efficiency does not impact the drain disturb robustness of the ultra-scaled cell. Fig. 7 shows drain read disturbs measured in the cell of Fig. 6 (Lch ≈ 40 nm, W ≈ 20 nm, 20 nm fin height), with very good retention of the bit state (programmed or erased) under extensive stress time.

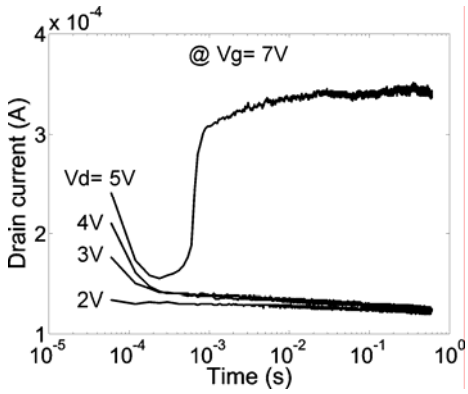


Fig. 5: time dependence of the drain current under various bias conditions. At low VDS ID decreases with time, indicating channel hot electron / warm electron programming. At VDS = 5 V, for times of the order of 1 ms and above ID has an abrupt increase due to avalanche at the drain.

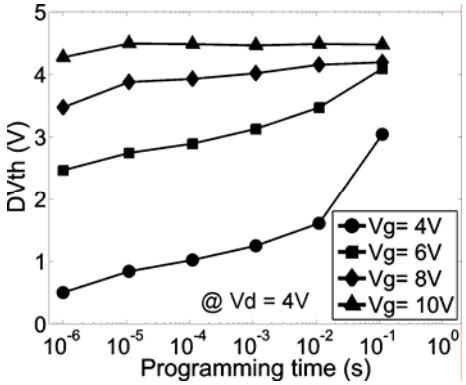


Fig. 6: CHEI programming of a cell with 40 nm channel length, 20 nm channel width, 20 nm fin height, halo doping, and medium density nanocrystals.

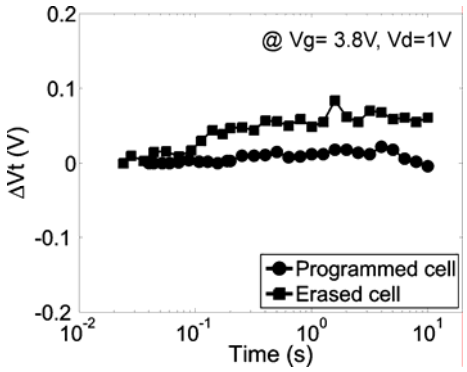


Fig. 7: drain read disturbs measured in the cell of Fig. 6.

4. Conclusions

We have investigated Si nanocrystal FINFLASH cells with minimum sizes down to 40 nm channel length and 20 nm fin width. Data of program / erase characteristics are discussed and it is shown a noticeable potential for Fowler-Nordheim tunneling program / erase at very low voltage and ability to CHEI programming for ultra-short channels.

We gratefully acknowledge L. Baldi, F. Giarrizzo, D. Mello, P. Barbera, C. Garozzo, and R. Puglisi for the support. Work funded by the FP6-FINFLASH EU Project.

References

[1] B. De Salvo et al., IEEE TDMR, 377 (2004).
[2] Peiqi Xuan et al. IEDM 2003.
[3] Chang Woo Oh et al. IEDM 2004
[4] M. Specht et al. VLSI Tech. Symp 2004
[5] C. Gerardi et al. submitt. VLSI Tech. Symp 2007
[6] S.M. Sze ed., *High Speed Semiconductor devices* (Wiley, 1990), 193

Physical Insights on Design of SONOS FinFETs Programmed with Channel Tunneling

Francesca Nardi, Giuseppe Iannaccone

Dipartimento di Ingegneria dell'Informazione: Elettronica, Informatica, Telecomunicazioni,
Università di Pisa, Via Caruso 16, 56122, Pisa, Italy, g.iannaccone@iet.unipi.it

Abstract

We investigate charge injection and storage in non-volatile memories based on the FinFET structure and on a silicon nitride storage layer (SONOS), programmed with channel injection (Fowler-Nordheim or direct tunneling). We demonstrate a procedure for performing an accurate time-dependent three-dimensional simulation of the program operation. However, the presence of a non planar interface between the silicon fin and the gate oxide in trigate FinFETs poses a few issues. During program with zero V_{ds} (channel tunnelling), the electric field in the tunnel oxide, the tunnel current density, and the stored charge density are not uniform. This can deteriorate the electrostatic behavior of the device and the program window, and lead to a non-effective utilization of the storage layer.

The non-uniform injection is a critical issue for FinFlash reliability and operation, and must therefore have a decisive impact on device design choices. In this work we have used 3D time-dependent simulations of the program operation of SONOS FinFETs based on Fowler-Nordheim (FN) or Direct Tunnelling (DT) to gain physical insights on their design. We focus in particular on understanding the effects of tunnel oxide thickness, fin height, edge curvature radius, and of the presence of a third gate, determining the region of the design space that allow us to maximize the program window.

1. Introduction

The use of multiple gate FETs is considered one of the most promising solutions to push further the scaling of flash memories beyond the 32 nm technology node [1] due to the improved electrostatic control of the channel. Moreover, multiple gate structures allow larger drive currents, which can be very useful to improve memory access time and programming speed. The storage medium wrapped on three sides of the channel considerably increases program efficiency. It has also been shown that the use of discrete-trap memory cells could be a promising approach to overcome limitations to further scaling down the conventional fresh cell architecture [2]. They ensure inherently lower voltage, good scalability properties and improved reliability, since a leakage path due to a single defect in the oxide leads to the loss of only a small fraction of the stored charge. In particular, the use of silicon nitride as discrete storage medium (the so-called SONOS structure) is attractive because the fabrication process is simple and the total equivalent gate oxide thickness is much thinner than in the case of a conventional flash memory, allowing smaller program and erase voltage and better electrostatic properties. For these reasons, nitride-based flash memories based on the FinFET architecture are promising both for embedded high-density flash memory applications [3] and for high-density NAND Flash technologies [4]. Indeed, it has been demonstrated that SONOS FinFET memory devices show excellent scaling perspectives down to a channel length of 20 nm [5].

In this paper, we investigate the non uniformity of charge injection and storage in non-volatile multiple gate memories with discrete storage nodes programmed with channel tunneling mechanisms (i.e., Fowler-Nordheim, or Direct Tunneling).

Non uniform charge storage has already been noticed in silicon-on-insulator (SOI) nanocrystal memories [6]. A preferential injection of electrons during the program operation from regions close to the edges of the SOI channel due to the reduced barrier height and increased electric field in the tunnel oxide is confirmed by experiments. As a consequence, charge is mainly stored in the discrete storage nodes in the oxide region surrounding the edges.

We investigate such issue in SONOS FinFET structures considering the reliability issues it poses and its impact on the design of the device architecture. To obtain an accurate quantitative evaluation of the charging process we have setup a dedicated time-dependent 3D simulation of the entire program operation by customizing and adding post-processing steps to a commercial TCAD tool [7].

2. Device Structure

We have considered the FinFET structure shown in Fig.1. The fin cross section has a width W of 20 nm, and uniform acceptor doping of $5\text{x}10^{17}\,\text{cm}^{-3}$. The gate length L is 70 nm. Source and drain are n+-doped ($N_D = 10^{20}\,\text{cm}^{-3}$) and the gate is made with n+ polysilicon ($N_D = 10^{20}\,\text{cm}^{-3}$). The thickness of the nitride layer is 5 nm, and the thickness of the control oxide is 5 nm. Different values are considered for edge curvature radius r, fin height H, tunnel oxide thickness t_{ox}.

The electric field in the tunnel oxide is largely enhanced near the edges of the fin, representing a significant reliability problem. Figure 2 shows the electric field component in the direction perpendicular to the silicon fin and the tunnelling current density as a function of position along the interface in the central transversal cross section.

The local tunnel current density along the direction perpendicular to the interface has been computed for each point at the silicon/tunnel oxide interface considering the local actual shape of the barrier [8]. Since oxide reliability is essentially limited by the device regions most stressed by the electric field, field enhancement can pose serious reliability problems. For this reason, we have evaluated all the different memory structures for the same maximum value of the electric field in the oxide, 10 MV/cm, i.e., for the same level of electric field stress. This means that different gate voltages are applied to the different structures during the program operation. If, as is often the case, program voltages are determined by other specifications, the control oxide thickness can be adjusted in order to comply with both the given program voltage and the maximum tolerable electric field.

3. Time-dependent simulation

In order to evaluate quantitatively the distribution of injected charge during programming, we have developed a procedure for the time-dependent simulation of the program operation, based on a commercial TCAD tool and ad-hoc codes. The simulation flow is illustrated in Fig.3. The local injected tunnel current varies as a function of time during programming, as charge stored in the nitride layer modifies the potential profile of the FinFET structure. Such variation is more pronounced near the edges, where charge is accumulated at a faster rate. In Figures 4 and 5 results of the simulation of the program operation are shown for the considered structure in the case of DT injection (when t_{ox} = 2.4 nm) and FN injection (t_{ox} = 4 nm), respectively. In the case of FN injection the stored charge density in the curved regions is three orders of magnitude larger than in the flat regions. Such difference is reduced to one order of magnitude in the case of DT.

The transfer characteristics of the programmed cell exhibits a two-step transition from subthreshold to inversion, due to the fact that the channel is first formed under the flat lateral faces (low threshold voltage region), then under the edges (high threshold voltage region). (Fig 6).

The simulation procedure allows us to draw a series of considerations for the design of SONOS FinFETs and in general of FinFET memories with discrete storage nodes.

- *DT programming should be preferred*, since it ensures a much more uniform charge storage than FN tunnelling and better use of the storage layer. However DT is obtained with a tunnel oxide thinner than 3 nm, that can cause retention problems. Such issue can be addressed using a two-layer tunnel dielectric, adding a second dielectric layer with smaller energy gap, that has no effect on the program operation for the same maximum electric field but largely increases retention.
- In the case of a trigate FinFET memory, *short fins should be preferred*, in order to minimize the flat regions, which are not effectively programmed and behave as a parasitic low threshold voltage transistor. The time-dependent current density is much more uniform in shorter fins, as can be seen in Figs. 5 and 9, where fin height is varied from 10 to 63 nm. As can be seen in Fig. 10, the shorter the fin, the larger the threshold voltage shift obtained for the same program condition. Let us stress the fact that short fins are easier to fabricate.
- Obviously, for trigate structures it is important to *maximize the edge curvature* radius, in order to reduce the electric field enhancement at the edges. (Figs. 11-12) An alternative possibility is to use a double gate structure, which enables one to use in an effective way the storage layer in the flat regions, by removing electric field crowding at the edges. Fig. 13 shows that *for the same program condition and fin height, the double gate FET memory (structure E in Fig. 11) provides a larger program voltage*.

In table 1 threshold voltage shifts are compared for the structures considered in the same program conditions. Structures D (Fig. 8) and E offer the largest threshold voltage window. Of course, the choice between a short fin with a trigate structure (D) and a double gate structure (E) must be based on additional considerations that are also valid for FinFETs to be used for logic applications such as control of short channel effects and manufacturability [9].

4. Conclusion

We have implemented a procedure for the 3D time-dependent simulation of the program operation of SONOS FinFETs to be programmed with FN or direct tunneling. We have shown that *time dependent* simulations are necessary to obtain accurate results. Based on such simulations, we have derived a set of design criteria that allow us to improve the utilization of the storage layer and to maximize the threshold voltage window for given program conditions. Such criteria are not always consistent with those obtained for FinFETs to be used for logic.

Acknowledgment

This work has been supported by the EU VI FP Project FinFlash (contract FP6-016917).

References

[1] International Technology Roadmap for Semiconductors, 2005 edition, PIDS, p. 41.

[2] B. De Salvo et al., IEEE Trans. Device and Materials Reliab., **4**, 377 (2004).

[3] P. Xuan, M. She, B. Harteneck, A. Liddle, J. Bokor and T. J. King, IEDM Tech. Dig., 609 (2003).

[4] M. Specht et al., VLSI Tech. Dig., 244 (2004).

[5] H. Hofmann et al. **49**, 1799 (2005).

[6] G. Fiori, G. Iannaccone, G. Molas and B. De Salvo, Appl. Phys. Lett., **86**, 113502-1 (2005).

[7] SILVACO – ATLAS User's Manual

[8] M. Depas, B. Vermeire, P. W. Mertens, R. L. Van Meirhaeghe, and M. M. Heyns, Solid-State Electronics **38**, 1465 (1995).

[9] J.G. Fossum et al., IEDM Tech. Dig., 613 (2004)

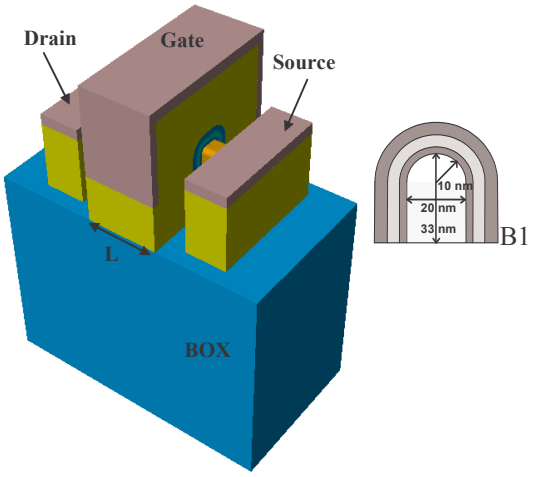


Figure 1: Structure of the SONOS FinFET under investigation. Inset: Fin cross section.

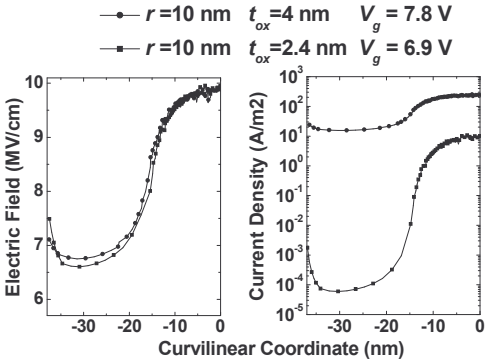


Figure 2: Electric field component in the direction perpendicular to the silicon fin – tunnel oxide interface as a function of position along the interface in the central transversal cross section (left). The curvilinear coordinate is zero at the point on the axis of symmetry of the fin. The structure is symmetric, therefore only negative values of the curvilinear coordinate are shown.

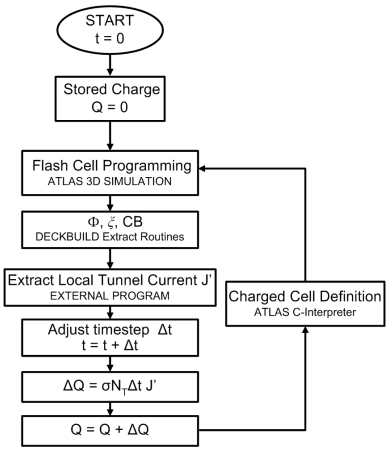


Figure 3: Flow diagram of the procedure for the time-dependent simulation of nitride based Flash memories

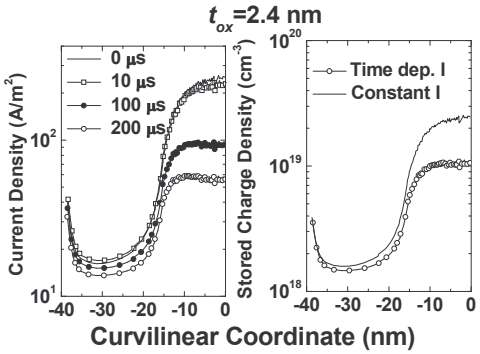


Figure 4: Local injected current density (left) and charge density in the nitride layer (right) as a function of the curvilinear coordinate for different program times. Structure B1 with r = 10 nm, t_{ox} = 2.4 nm, V_g = 7.8 V.

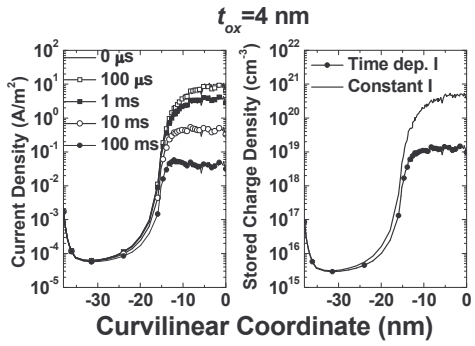


Figure 5: Local injected current density (left) and charge density in the nitride layer (right) as a function of the curvilinear coordinate for different program times. Structure B1 with r = 10 nm, t_{ox} = 4 nm, V_g = 7.8 V.

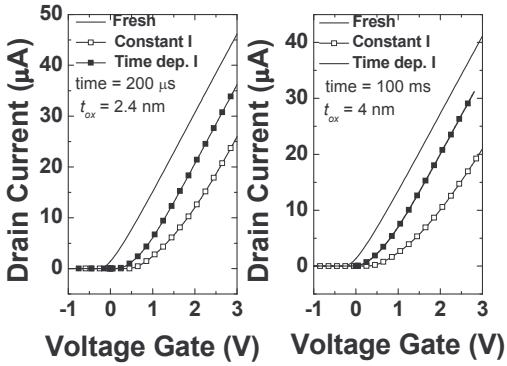


Figure 6: Transfer characteristics (V_{ds} = 1 V) of the fresh and programmed SONOS FinFET obtained with constant injected current and with the time-dependent simulation: (left) tunnel oxide thickness of 2.4 nm and program time of 200 µs, (right) tunnel oxide thickness of 4 nm and program time of 100 ms

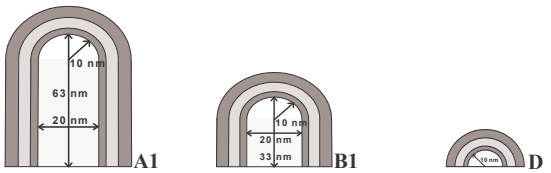


Figure 7: Considered cross section to evaluate the effect or the fin height: the structures A1, B1, and D, have the same fin width (20 nm), edge curvature radius (10 nm), and different fin heights, of 63, 33, and 10 nm, respectively.

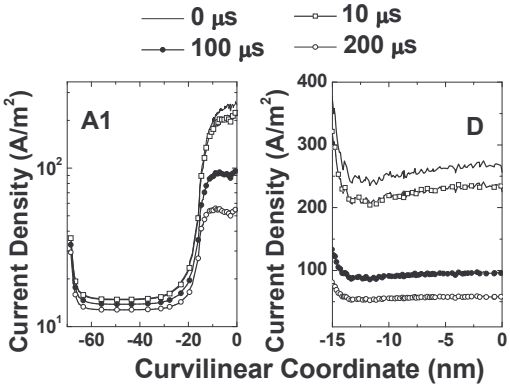


Figure 8: Local injected current density as a function of the curvilinear coordinate for various program times in a DT regime. Considered structures: A1 with t_{ox} = 2.4 nm, H = 63 nm, r = 10 nm, V_g = 6.9 V (left); D with t_{ox} = 2.4 nm, H = 10 nm, r = 10 nm, V_g = 6.9 V (right).

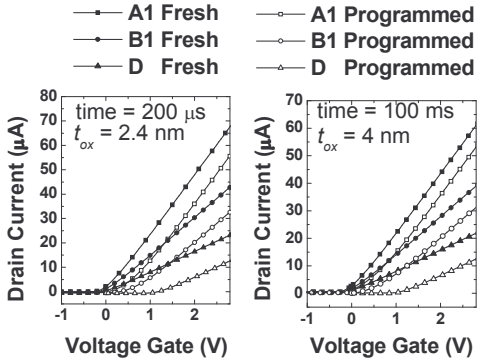


Figure 9: Transfer characteristics (V_{ds} = 1 V) of the fresh and programmed SONOS FinFET memories as a function of fin height for two different tunnel oxide thickness implying DT (left) and FN (right) regime, respectively. Comparison for different heights and the same curvature radius.

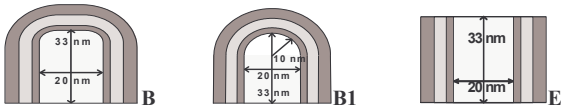


Figure 10: Considered cross section to evaluate the effect of the fin curvature and top gate. The structures B, B1, have the same fin width (20 nm) and different edge curvature radius (5 and 10 nm). Devices have the same height (33 nm).

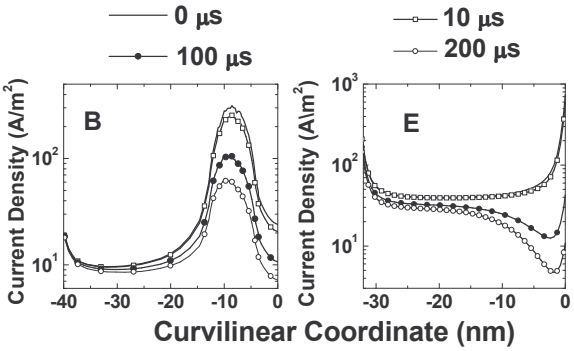


Figure 11: Local injected current density as a function of the curvilinear coordinate for various program times in a DT regime. Considered structures: B with t_{ox} = 2.4 nm, H = 33 nm, r = 5 nm, V_g = 6.2 V (left); E with t_{ox} = 2.4 nm, H =33 nm, V_g = 7.8 V (right).

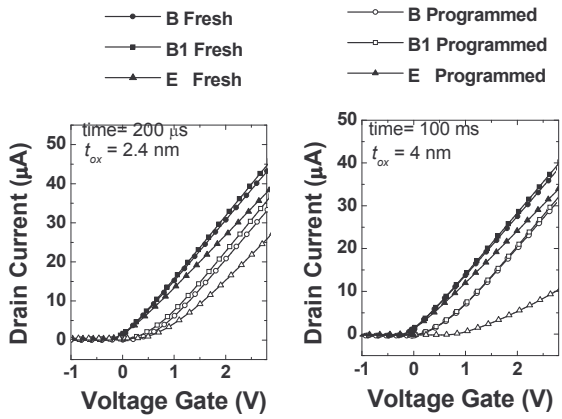


Figure 12: Transfer characteristics (V_{ds} = 1 V) of the fresh and programmed SONOS FinFET memories as a function of fin height for two different tunnel oxide thicknesses implying DT (left) and FN (right) regime, respectively. Comparison for the same fin height and different curvature radius.

Structure	t_{ox} = 2.4 nm		t_{ox} = 4 nm	
	ΔV_t (mV)		ΔV_t (mV)	
	I_d=10 μA	I_d=10 nA	I_d=10 μA	I_d=10 nA
A1	446	247	300	65
B1	633	443	509	280
D	973	1197	1200	1161
B	595	309	522	268
E	819	553	1905	950

Table 1: Computed threshold voltage shift for the considered devices in both subthreshold and strong inversion regime. In subthreshold the threshold voltage shift has been defined at I_d = 10 nA; in strong inversion it has been defined at I_d = 10 μA.

Study of Programming Characteristics of 4-bit SONOS Flash Memory Using 3-Dimensional Transient Simulation

Jang-Gn Yun, Yoon Kim, Il Han Park, Seongjae Cho, Jung Hoon Lee, Gil Sung Lee, Doo-Hyun Kim, Dong Hua Lee, Se-Hwan Park, Jong-Duk Lee, and Byung-Gook Park

Inter-University Semiconductor Research Center (ISRC) and School of Electrical Engineering and Computer Science, Seoul National University, San 56-1, Sillim-dong, Gwanak-ku, Seoul 151-742, Republic of Korea,
E-mail address: jgyun7@snu.ac.kr

Abstract

Systematic investigation of programming characteristics of 4-bit SONOS flash memory has been carried out through 3-dimensional transient simulation. Programming speed dependence on the bias condition is simulated by changing applied voltages on gate1 and drain. Enhanced programming speed is achieved under higher bias conditions. Programming disturbance is characterized in the devices with different fin widths. As the fin width decreases, more programming disturbance on the opposite word-line (WL) across the channel is observed.

1. Introduction

A 4-bit SONOS flash memory is 3-dimensional structure and it has 4 storage nodes as depicted in Fig. 1 [1]. Making use of the charge separation by the Si-channel [2] and the non-conductive nitride layer [3], 4-bit/cell operation is possible. It has two channels and they are governed by two side-gates. Random programming/reading is achieved with appropriate bias modulation.

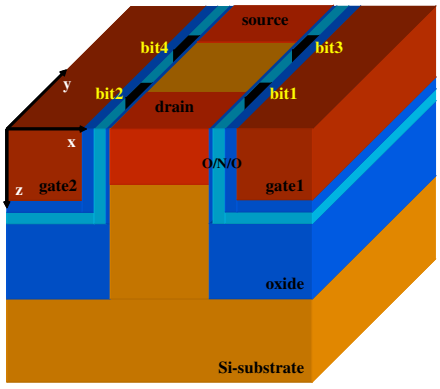


Fig. 1: Structure of 3-dimentional 4-bit SONOS flash memory. 4 storage nodes can be distinguished by the Si-fin and the non-conductive nitride layer.

In this device, each bit is programmed by channel hot electron injection (CHEI) mechanism as shown in Fig. 2. Hot electrons are generated by impact ionization on the drain-side depletion region and fast programming is feasible through their multiplication process [4].

By controlling bias conditions, injected electrons can be localized in the nitride layer on the vicinity of drain junction. In this study, programming properties of 4-bit SONOS memory are investigated with different bias conditions and fin widths (W) using 3-dimensional transient simulation [5].

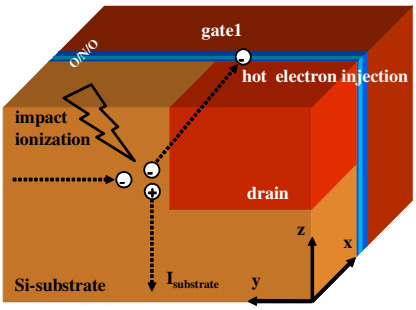


Fig. 2: Impact ionization induced CHE and hole generation near the drain-side depletion region. Generated holes contribute to $I_{substrate}$ while hot electrons are injected in the charge trapping layer.

2. Design of 4-bit SONOS Cell for Programming Transient Simulation

Si-box as a charge trapping node is placed in nitride layer to simulate the effect of localized charges on the channel as illustrated in Fig. 3.

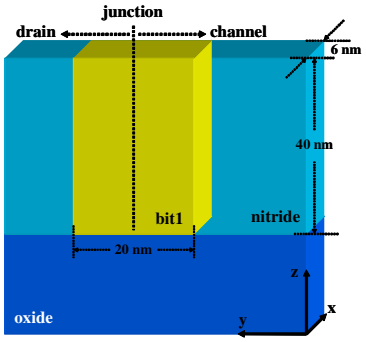


Fig. 3: Si-box in the nitride layer used for localized charge trapping node.

To program bit1, drain voltage (V_D) is elevated to its final value until the programming time reaches to 10 ps with fixed gate1 voltage (V_{gate1}) as shown in Fig. 4.

To analyze the dependence of programming properties on the bias condition, V_{gate1} and drain voltage

(V_D) are varied as V_{gate1}= 4, 5, 6 V and V_D= 1.5, 2.0, 2.5 V, respectively. Different fin widths of W= 20, 30, 40, and 50 nm are used for the characterization of programming disturbance.

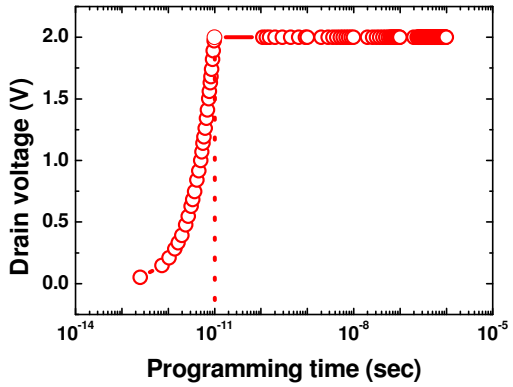
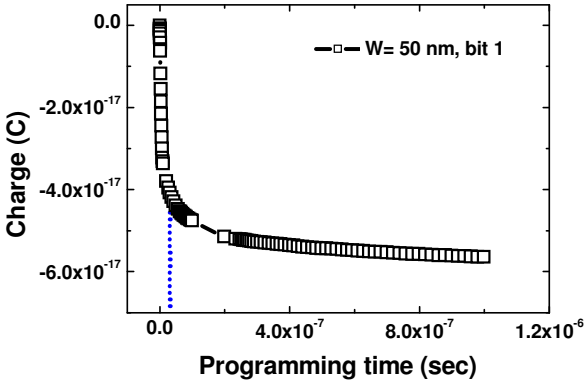


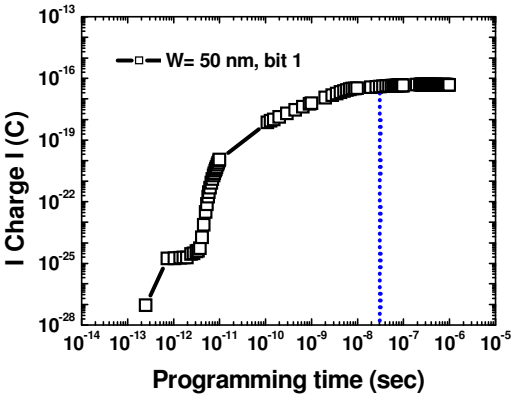
Fig. 4: V_D transient as a function of programming time with fixed V_{gate1}.

3. Results and Discussion

By the aid of transient simulation, the integrated charges in bit1 are changed with programming time as shown in Fig. 5. At first, charges are integrated rapidly but it tends to saturate soon.



(a)



(b)

Fig. 5: Variation of integrated charges in bit1 during the programming operation (V_{gate1}= 5 V, V_D= 2 V). (a) Linear and (b) log scale.

Fig. 6 shows the threshold voltage (V_{th}) as a function of programming time. As time goes by, V_{th} increases because locally trapped charges induce the variation of potential barrier near the drain junction [6].

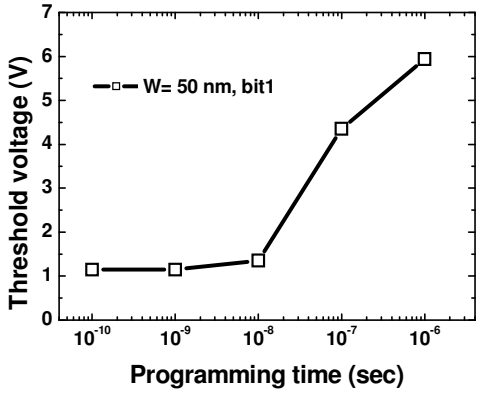


Fig. 6: Programmed V_{th} characteristic in bit1 (V_{gate1}= 5 V, V_D= 2 V).

Substrate current ($I_{substrate}$) can be used as a parameter of the impact ionization rate. As the impact ionization and the accompanying multiplication increase, more holes are generated and thus more $I_{substrate}$ flows [4]. $I_{substrate}$ calculated with programming time shows dramatic change at about programming time= 30 ns as in the inset of Fig. 7. After the time, hole generation rate shows gradual diminution.

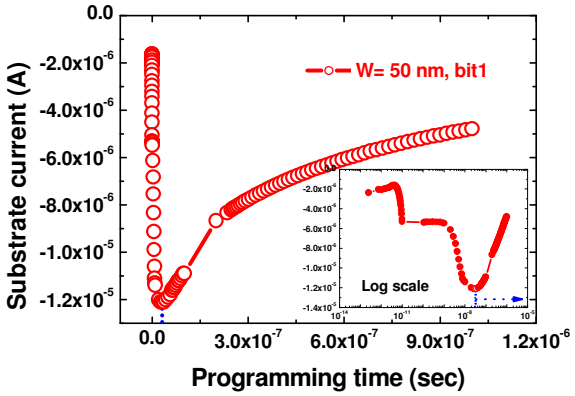


Fig. 7: $I_{substrate}$ as a function of programming time.

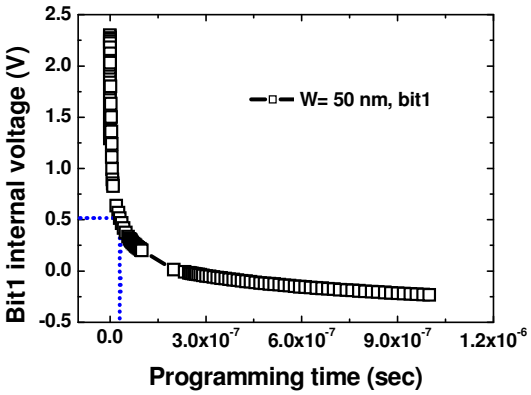
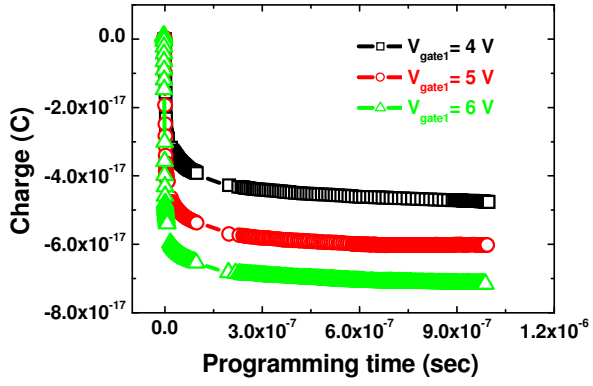


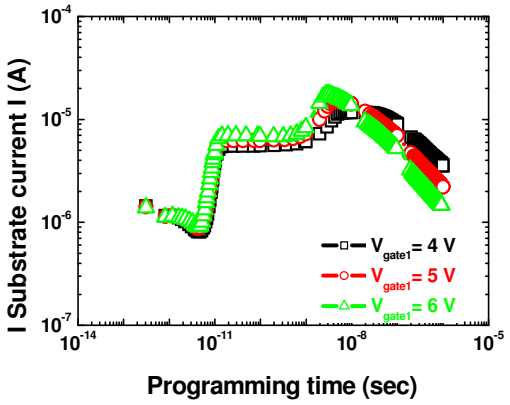
Fig. 8: Internal voltage change in bit1.

From Fig. 8, it is because of the bit1 internal voltage reduction as the stored electrons increase. As the charges involved in the impact ionization process are reduced due to the internal voltage decrease in bit1, lesser holes are generated near the drain junction. The rapid increment of electron injection starts to slow down after the time as indicated in Fig. 5.

Programming speed dependence on the bias condition is characterized by changing V_{gate1} and V_D. In Figs. 9 and 10, the effect of V_{gate1} on the programming speed is characterized by simulating the charge, $I_{substrate}$ and V_{th} vs. programming time.



(a)



(b)

Fig. 9: Transient (a) charge and (b) $I_{substrate}$ variation with different V_{gate1} (W= 20 nm, V_D= 2 V).

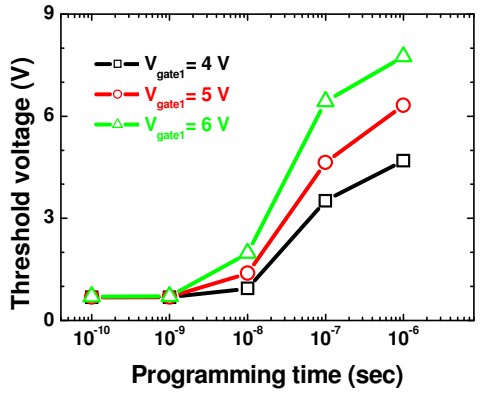
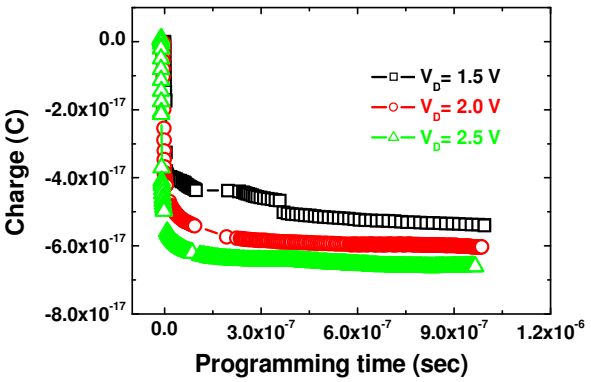


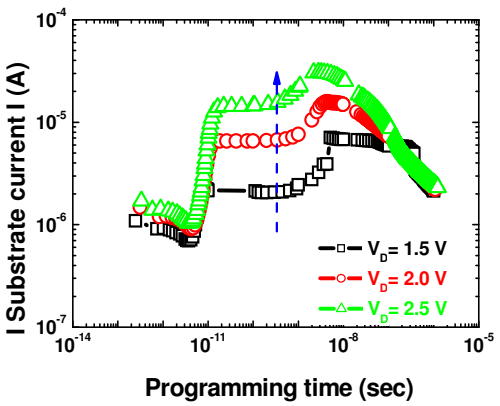
Fig. 10: V_{th} shift as a function of programming time with different V_{gate1} (W= 20 nm, V_D= 2 V).

Although it still shows slowing down of the charge injection in a short period of time, increase of the transverse electric field with higher V_{gate1} leads to more electron injection at the same programming time as shown in Fig. 9(a). In Fig. 9(b), on the other hand, just small increment of $I_{substrate}$ is observed with higher V_{gate1}. These mean that V_{gate1} is not a dominant factor for the impact ionization but it influences the charge injection efficiency. Faster V_{th} shift with higher V_{gate1} is resulted because of the stronger transverse electric field as shown in Fig. 10.

In Figs. 11 and 12, V_D is changed with fixed V_{gate1} to characterize the programming speed dependence on V_D.



(a)



(b)

Fig. 11: Transient (a) charge and (b) $I_{substrate}$ variation with different V_D (W= 20 nm, V_{gate1}= 5 V).

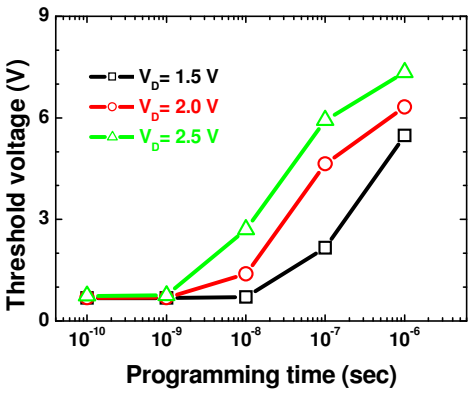


Fig. 12: V_{th} shift as a function of programming time with different V_D (W= 20 nm, V_{gate1}= 5 V).

As expected, injected charges increase as V_D rises as shown in Fig. 11(a). Although the transient charge variation characteristic is similar with that of Fig. 9(a), $I_{substrate}$ shows different peculiarity in comparison with Fig. 9(b). With higher V_D, $I_{substrate}$ increases dramatically as shown in Fig. 11(b).

As V_D increases, the lateral electric field increases and more impact ionization takes place in the drain-side depletion region. This generates more holes and thus $I_{substrate}$ increases with higher V_D. That is, V_D affects on the CHE generation rather than enhancing the charge injection efficiency. Rapid programming characteristic is also observed with higher V_D on account of the increased impact ionization as shown in Fig. 12.

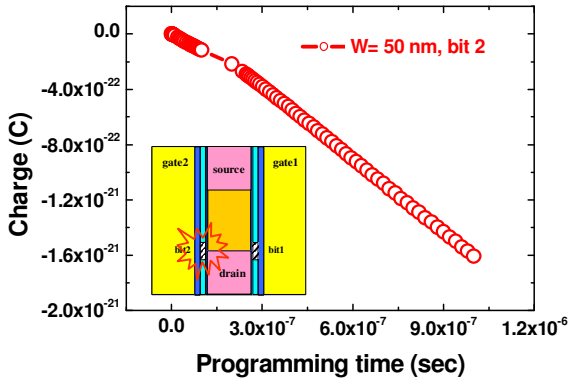


Fig. 13: Charge injection in bit2 during the bit1 programming (V_{gate1}= 5 V, V_D= 2 V, V_{gate2}= V_S= 0 V).

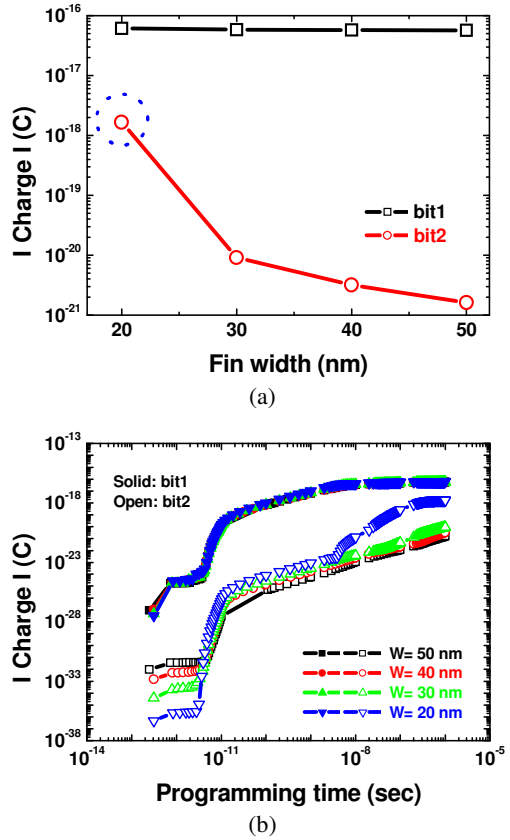


Fig. 14: Programming disturbance characteristics (a) according to the fin width at the programming time= 1 μs and (b) as a function of programming time (V_{gate1}= 5 V, V_D= 2 V, V_{gate2}= V_S= 0 V).

In conventional NOR-type flash memory, gate programming disturbance is brought about in the unselected cell sharing same WL mainly due to the high WL voltage [7]. Another programming disturbance can be taken place in the device such as double SONOS memory [8] which has a common conduction medium. In 4-bit SONOS memory, two gates share one Si-fin as a chnnel. Therefore, the bit2 can also be programmed while programming bit1 even though gate2 voltage (V_{gate2}) is grounded as shown in Fig. 13. Linear increase of injected charges in bit2 during the bit1 programming is observed as the programming time increases.

Fig. 14(a) shows programming disturbance characteristics according to the different fin widths at the programming time = 1 μs. The programming disturbance on the opposite WL across the fin becomes stronger as the two gates get closer. Especially, drastic increase of injected charges in bit2 is observed in the device with W = 20 nm. Special care is needed to reduce the programming disturbance as scaling down the device dimension, especially in the devices which have separated multiple gates for the multibit operation.

4. Conclusions

Charge injection characteristics in 4-bit SONOS memory are studied using 3-dimensional transient simulation. CHE generation and corresponding charge injection into the storage node is delayed because the accumulation of injected charges leads to internal voltage increment. Fast programming in bit1 is achieved by increasing V_{gate1} and V_D. V_{gate1} affects on the charge injection efficiency while CHE generation is mainly dominated by V_D. The programming disturbance on the opposite WL becomes stronger as the fin width decreases.

References

[1] J. G. Yun, Y. Kim, I. H. Park, S. J. Cho, J. H. Lee, D. H. Kim, G. S. Lee, J. Y. Song, J. D. Lee, and B. G. Park, IEEE Nanotechnology Materials and Devices Conference 2006, 214 (2006).

[2] S. J. Cho, I. H. Park, T. H. Kim, J. S. Kim, K. W. Song, J. D. Lee, H. C. Shin, and B. G. Park, IEEE Trans. Nanotechnol., **5**, 180 (2006).

[3] F. Hofmann, M. Specht, U. Dorda, R. Kommling, L. Dreeskornfeld, J. Kretz, M. Stadele, W. Rosner, L. Risch, Solid-State Electron., **49**, 1799 (2005).

[4] P. Pavan, R. Bez, P. Olivo, and E. Zanoni, Proceedings of the IEEE, **85**, 1248 (1997).

[5] ATLAS User's Manual, software version 5. 10. R, SILVACO Internatinoal, Santa Clara, CA, (2005).

[6] E. Lusky, Y. S. Diamand, I. Bloom, and B. Eitan, IEEE Electron Device Lett., **22**, 556 (2001).

[7] D. Ielmini, A. Ghetti, A. S. Spinelli, and A. Visconti, IEEE Trans. Electron Devices, **53**, 668 (2006).

[8] C. W. Oh, S. H. Kim, N. Y. Kim, Y. L. Choi, K. H. Lee, B. S. Kim, N. M. Cho, S. B. Kim, D. W. Kim, D. Park, and B. I. Ryu, VLSI Tech. Dig., 50 (2006).

Investigation of the impacts of channel length, fin width on Si-NC SOI-FinFlash memory characteristics

C. Jahan, J. Razafindramora, L. Perniola, M. Gély, C. Vizioz, A. Toffoli, F. Allain,
S. Lombardo*, C. Bongiorno*, G.Reimbold, F.Boulanger, B. De Salvo and S. Deleonibus

CEA/LETI-Minatec, 38054 Grenoble, France, Tel: +33 438785192, Fax: +33 438785459, carine.jahan@cea.fr
CNR-IMM, Catania – Italy

Abstract

This paper presents the technological process and electrical behaviour of Silicon Nano-Crystal (Si-NC) FinFlash memories fabricated on Silicon On Insulator (SOI) substrates. We study ultra-scaled memory devices (with channel length, L_G, and fin width, W_{FIN}, down to few deca-nanometers), with Si-NC storage nodes fabricated either by LPCVD or by annealing of Silicon-Rich-Oxide, under different electrical configurations (NAND and NOR schemes).

1. Introduction

Tri-gate FinFlash memory devices [1, 2] are one of the most promising solutions to solve scaling problems of floating-gate Flash memories, both for stand-alone and embedded applications. In fact, the discrete storage node approaches, i.e. Silicon Nano-Crystal [3] or SONOS technologies, conjugated to the novel 3D FinFET architecture [4] offer the possibility of scaled gate dielectrics, implying scaled operating voltages, along with short channel effect immunity and higher sensing current drivability. In this work, we report on the program/erase performances of Si Nano-Crystal FinFlash devices, with a particular attention to the dependence of the electrical characteristics on device geometries (L_G, W_{FIN}).

2. Device Fabrication

FinFlash devices were fabricated on SOI wafers, the process being based upon a reference FinFET flow [4]. E-beam lithography and resist trimming are used to pattern both the fin and the control gate. Sidewall oxidation is carried out to round fin corners and obtain smaller widths. The fins are 30nm high and fin widths down to 10nm are obtained. After fin patterning and boron channel implantation, gate stack deposition is performed. A 5nm-thick thermal SiO_2 is grown followed by the storage layer deposition. Two different techniques have been used to develop the Si-NC layers: either direct Si LPCVD deposition or annealing of Silicon Rich Oxide (SiO_x). Fig.1 shows Scanning Electron Microscopy (SEM) pictures of the fin structures covered by LPCVD Si-NC. Then, the blocking dielectric (8nm-thick HTO) is deposited, followed by the 100nm N^+ Poly-Si control gate. Fig.2 shows a cross Transmission Electron Microscopy (TEM) picture of the 20nm wide fin after gate stack deposition. After the gate etching and extension implants, a 50nm-thick nitride spacer is done. Then raised Souce/Drain are epitaxially grown in order to decrease the series resistance, prior to HDD implants. Finally, a 1050°C spike anneal is used, followed by Nickel silicidation and classical BEOL process. A broad range of fin widths W_{FIN} and gate lengths L_G dimensions have been obtained.

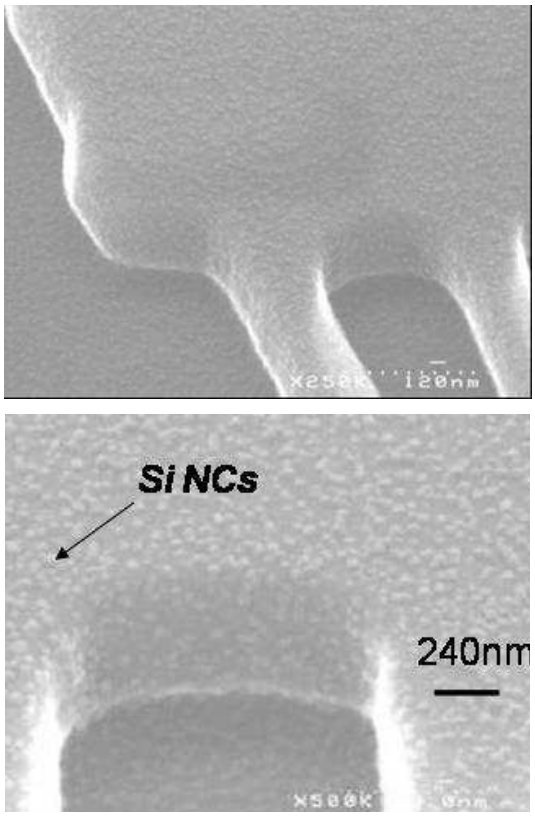


Fig.1: SEM pictures of LPCVD Si-NCs deposited on fin structures.

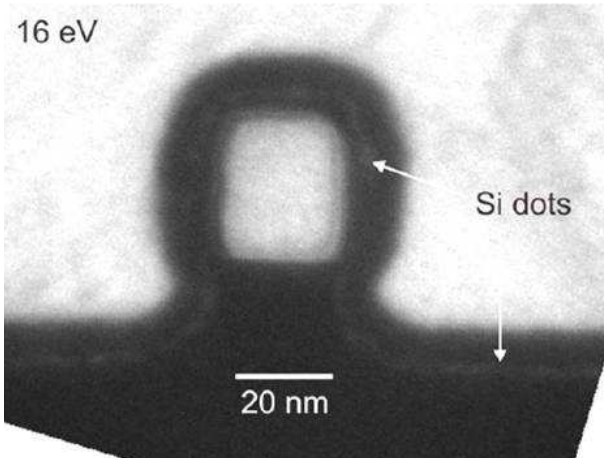


Fig. 2: Cross TEM picture of fin after gate stack deposition including 5nm tunnel oxide, Si-NCs layer, 8nm top oxide and 100 nm Poly-Si gate.

3. Electrical Characteristics

Device electrostatics – Finflash devices exhibits good short channel control. Fig.3 shows the Id(Vg) curves in SiO_x Si-NC FinFlash devices with 20nm and 30nm fin widths, respectively, and different gate lengths. We observe that the short channel performance of our devices is further improved by reducing the fin width: a better electrostatic gate control over the channel is obtained for narrower fins. Devices with small aspect ratio (e.g. W_{FIN}/L_G=20/70nm) provide sub 100mV/dec Subthreshold Slope and sub 0.1V/V DIBL (not shown here).

Note that, in our process flow, the channel doping was targeted to achieve positive threshold voltage for fresh cells. The negativity of Vth here observed may be due to the extremely scaled geometry of devices that we are considering: the doping level of the fin is not able to sufficiently raise the Vth.

NAND Characteristics *(Fowler-Nordheim Write/Erase)* – The FinFlash cells were programmed and erased by Fowler Nordheim tunnelling. We started our analysis by investigating the dependence of the FN program/erase characteristics on the nanocrystals technologies, by comparing the performances of Finflash cells with Si-NCs fabricated by LPCVD or by phase separation of SiO_x.

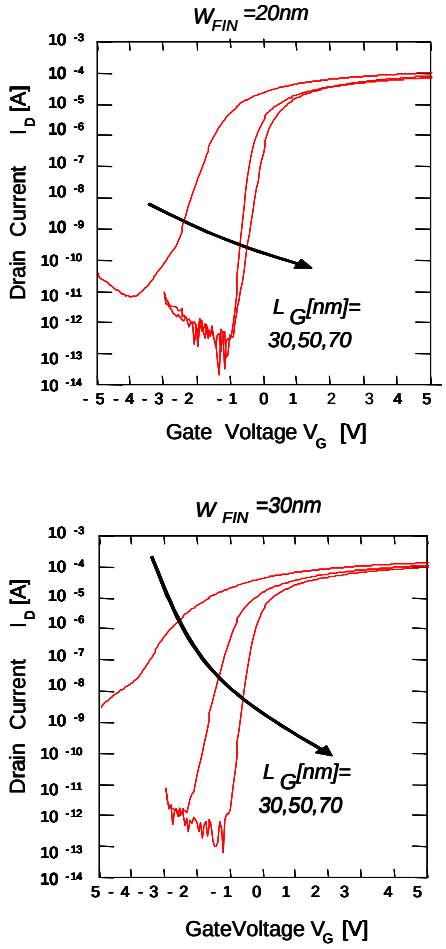


Fig.3: Id(Vg) characteristics of FinFlash devices with W_{FIN}=20nm (up) and W_{FIN}=30nm (down), with different gate lengths L_G (30nm, 50nm, 70nm). Vd=1V.

Fig.4 shows a comparison of the Id(Vg) characteristics. Indeed, devices exhibit the same behaviour in term of drive current and subthreshold slope, but it clearly appears that devices containing LPCVD Si-NCs are more effective in FN/FN write/erase mode. Note that in both cases, ΔV_{th} larger than 2V are achieved with few ms programming times.

We also investigate the impact of the FinFlash geometry on FN characteristics. Id(Vg) characteristics of FinFlash devices with equal L_G (=70nm) and different fin widths W_{FIN} (i.e. 30nm and 50nm), written and erased in FN, are shown in Fig.5. We can observe that the ΔV_{th} window is enhanced by reducing W_{FIN}. This behaviour has been previously theoretically explained [5] by the preferential injection of electrons at the corners of the fin, rather than a uniform injection, which indeed creates an effective "bottleneck" to channel conduction in narrowest fins.

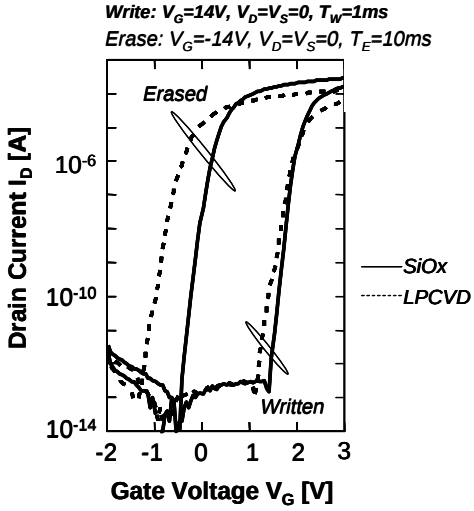


Fig.4: Comparison between Id(Vg) characteristics of FinFlash cells with LPCVD or SiOx Si-NCs, in FN/FN written and erased states.

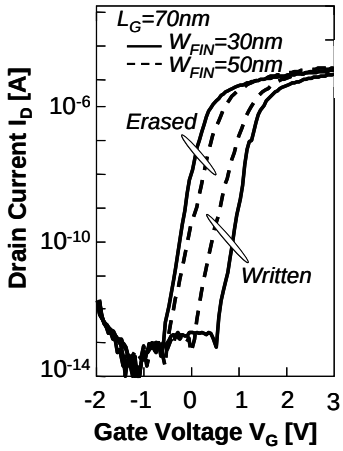


Fig.5: Id(Vg) characteristics of FinFlash cell with L_G=70nm and different W_{FIN}: 30nm (solid lines); 50 nm (dotted lines), in written and erased states. FN/FN program/erase conditions are: V_G=14V, Vd=Vs=0V, T_w=1ms; and Vg=-14V, Vd=Vs=0V, T_E=10ms, respectively.

NOR Characteristics *(Channel Hot Electron Write / Fowler-Nordheim Erase)* – We perform Channel Hot Electron (CHE) Injection and Fowler-Nordheim (FN) erasing on FinFlash devices with different geometries. Fig.6 shows the V_{th} transient characteristics of SiOx Si-NC cells with W_{FIN}=20nm and different L_G (50nm and 70nm). In both cases, quite large ΔV_{th} (~2V) can be obtained with low operating voltages and short programming times. We can also observe the dependence of the programming window ΔV_{th} on the gate length L_G: writing by CHE is more effective with small gate lengths. Indeed, the understanding of this phenomenon will require further experimental and theoretical investigations.

Retention – Data retention measurements at high temperature (T=150°C) have been carried on a scaled SiO_x Si-NC FinFlash device (W_{FIN}=40nm, L_G=70nm) after HE/FN write/erase. Fig.7 shows that a ~1V programming window remains after 10 years.

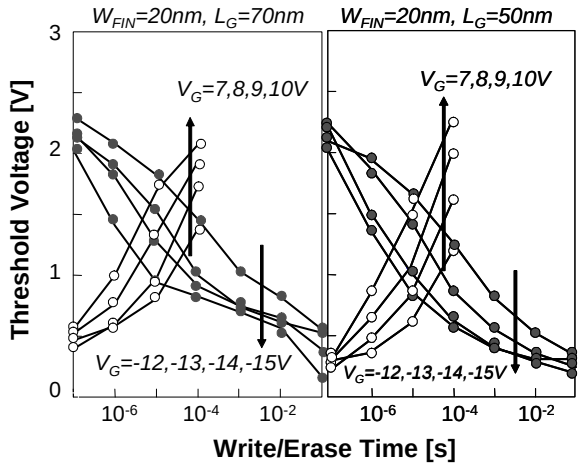


Fig.6: Transient Vth characteristics, in CHE/FN write/erase mode, of SiOx Si-NC Finflash cells with W_{FIN}=20 nm and two different gate lengths L_G: 70nm (left) and 50nm (right). During writing: Vd=2.5V and Vg= 7, 8, 9, 10V.

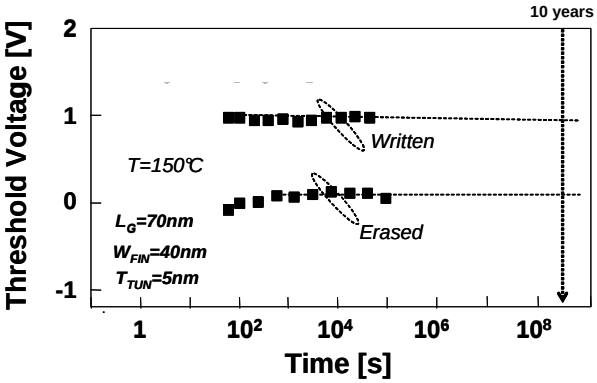


Fig.7: Data retention @ 150°C of a SiO_x Si-NC FinFlash device (W_{FIN}=40nm and L_G=70nm).

4. Conclusions

SOI Finflash cells with Si-NC storage nodes, fabricated either by LPCVD or by annealing of Silicon-Rich-Oxide, have been fabricated. Electrical characterization has shown that NAND and NOR functionalities are achievable in ultra small devices (with W_{FIN} and L_G in the few deca-nanometer range). Significant programming windows are obtained with low operating voltages and short programming times. Acceptable charge loss at high temperature has also been demonstrated. These results further prove the high interest of the FinFlash architecture for future stand-alone and embedded memory applications.

Acknowledgements

This work has been partly founded by the European IST-NMP FinFlash project.

References

[1] S. H.Lee et al., *Tech. Dig. of IEDM 2006*, p.33.

[2] C. Friederich et al., *Tech. Dig. of IEDM 2006*, p.963.

[3] B. de Salvo et al., *Tech. Dig. IEDM 2003*, p. 597.

[4] C. Jahan et al., *Symp. on VLSI Tech.* 2005, p.112.

[5] G. Fiori et al. *IEEE Trans. On Nanotech. 2005*, 326.

SESSION D

Floating Gate

Current limitations of floating gate NVM and new alternatives

Albert Bergemont, TRD Director
Maxim Integrated Products
3725 North First Street
San Jose, CA, 95134, USA

Introduction

Floating gate Flash memory scaling has been tremendous in the last 15 years. As a result, the overall Flash market has grown at an unprecedented pace, mainly driven by exploding customer demand for mobile mass storage applications such as digital cameras, MP3 players and cell phones. In such a large commodity market, the course for further Flash cell scaling has led to tremendous TRD activity in the last few years with NAND density doubling almost every year. Although it is expected that Flash will continue to scale, there are several physical limitations to confront and downscaling beyond 45nm is a concern, especially for floating gate architectures. The Aim of this paper is to review those limits and present the most promising emerging Flash technologies for mass storage applications.

High density Stand Alone NVM

Among floating gate Flash architectures, NAND flash memory has the smallest cell footprint due to its simple one-transistor structure and Source/Bit line contacts common to multiple cells within a string. With its high programming throughput and highest density, NAND is the dominant technology for data storage. On the other hand, NOR random access Flash has inherently better access time. Even with a larger (~2X) cell footprint, NOR stays the mainstream technology for applications requiring storage of codes and parameters and more generally execution in place. [Table 1]

Flash scaling is not only limited by photolithography (critical dimensions and overlay), but by reliability: NVM reliability has been and continues to be the biggest issue to be resolved. It is limited by specifications such as:
- Numbers of P/E cycles
- P/E disturbs between adjacent cells/sectors or pages
- 10 years read disturbs
- 10 years retention time after cycling

Note that those specifications are different whether the part is going into a consumer, industrial or automotive application (mainly numbers of cycles, temperature operation and FIT rate).
Those reliability concerns have driven the following limitations:

- Vertical dimensions scaling, tunnel oxide and inter-poly dielectric (ONO) have reached their lower thickness limit (table 2) leading to severe difficulties to maintain a decent gate coupling ratio (fig1 [1]). A high-k dielectric as an inter-poly layer could provide a solution against coupling ratio degradation.
- With the reduction of dimensions, the tolerance to charge loss reduces, due to the overall reduction of cell charge capacitance (less than 100 electrons at 45nm node) (fig3 [1]). Again high-k inter-poly dielectric would help, especially for MLC.
- Cell-to-cell interference, i.e. crosstalk between adjacent cells, requires the introduction of low-K dielectric between adjacent stacked gates and a reduction of the floating gate height, which in return affects the coupling ratio… (fig2,4[2])

Overall, it appears that the floating gate technology has reached its scaling limits; unsurprisingly, a lot of new NVM concepts have emerged in the last few years. Most promising are charge-trap (CT) NVM, phase change memory (PCM), three dimensional memory, MRAM, FinFet, etc… Among these, three Flash concepts appear most promising at this time:

- CT Flash (NROM [3], TANOS [1,4]), where charges are stored within a thin ONO layer (SONOS): not only the nitride layer is able to store many electrons but those cells are free from cell-to-cell interferences (Non floating gate). NROM is used in a NOR configuration (Virtual ground contact less array) whether TANOS (fig 5,6,7 [1,4]) is used in a NAND configuration; it is questionable if NROM will be able to handle further scaling due to the lateral redistribution of charges during bake. TANOS is more promising, as charges are

trapped uniformly through the channel, but will suffer further bottom oxide scaling.

- PCM, using a current-induced Joule effect in a chalcogenide alloy [5]: main issues for further scaling are the reduction of programming current with dimensions as well as avoiding thermal cross-talk.
- 3-D vertically stackable memory seems to be the ultimate paradigm as it retains higher die efficiency at smaller die sizes, by placing support circuitry under the memory array; it will be the most plausible way beyond the 20 nm node. Few concepts have been reported so far (Sandisk and Samsung), both using SONOS type in a NAND organization. [6,7]

Embedded NVM

Although the bulk of the market is in the mass storage market, embedded NVM is finding increasing use in a wide array of ICs with applications ranging from a few bits (analog trimming) to a few Megabits for data/code storage. All presently available EEPROM or FLASH require process modifications to the CMOS baseline, such as tunnel oxide, inter-poly oxide, dual poly gate, thicker oxides and additional implant steps. The result is costly (around $300-400 added), especially for applications where a few bits are required. The last few years saw the emergence of low cost, friendly embeddable NVM solutions such as OTPs (one time programmable) [8,9] and MTPs (multi time programmable) [10-12]. Those architectures have the benefits of being cheap at the expense of NVM performances such as number of cycles and data retention. For industrial applications (such as automotive), requiring tough cycling and retention temperature, the floating gate technologies are still prevailing, although more costly. One candidate for their replacement is nano dots NVM, with inherent easy integration with conventional CMOS processes and robustness to charge retention bake, as the distributed nature of charge storage makes it more robust (in a conventional floating gate NVM, a weak spot is fatal).

Conclusion

For mass storage applications, floating gate NVM will see new alternatives in mass production by the end of the decade: CT-cells, PCM in the short term, followed by 3-D vertically stackable memories after 2010.
For embedded applications, FG solutions will survive but might be displaced later by nano-crystals memories.

References

[1] S-M. Yung et al, "Three Dimensionally stacked NAND Flash memory technology using stacking single crystal Si layers on ILD and TANOS structure for beyond 30nm node", NVSM 2006, pp37-40.
[2] K. Kim et al, "Future outlook of NAND Flash technology for 40nm and beyond", NVSM 2006, pp9-11.
[3] B. Eitan et al, "NROM: a novel localized trapping, 2-bit nonvolatile memory cell", IEE El Dev lett 2000, pp543-545.
[4] C.H Lee et al, "A novel SONOS structure of SiO_2/SiN/ Al_2O_3 with TaN metal gate for multi giga Bit Flash memories", IEDM 2003
[5] R. Bez et al, "Chalcogenide Phase Change Memory: scalable NVM for the next decade?", NVSM 2006, pp12-14.
[6] A. Walker et al, "3D TFT-SONOS Memory Cell for ultra-high-Density file storage applications", Symp on VLSI technology, 2003
[7] K. Park et al, "Highly manufacturable 32Gb Multi-Level NAND Flash Memory with 0.0098um^2 cell size using TANOS (Si-Oxide-Al_2O_3-TaN) Cell technology", IEDMM 2006, pp29-32.
[8] A. Bergemont et al, "A Non volatile memory device with true CMOS compatibility", NVSM 2000.
[9] A. Bergemont and A. Kalnitsky, US Patents 6130848, 6081451, 6055185, 6137722
[10] A. Bergemont and A. Kalnitsky, US Patents 6137721, 6137723
[11] Y. Ma et al, "Reliability of pFET EEPROM with 70A tunnel oxide manufactured in generic logic CMOS process", NVSM 2000.
[12] H.M. Lee et al, "NeoFlash- True logic Single Poly Flash Memory technology"", NVSM 2006, pp15-16.

	NAND	NOR
Read Access	Serial	Random
P/E Mechanisms	FN/FN	CHE/FN
Cell Size (F^2)	5-6	10--12
Throughput	10MB/sec	0.5MB/sec

Table 1: NAND/NOR attributes

	TOSHIBA IEDM 2000	SAMSUNG IEDM 2001	SAMSUNG IEDM 2002	SAMSUNG IEDM 2004
F Minimum dimension (um)	0.14	0.12	0.09	0.06
NAND Channel length (um)	0.14	0.12	0.09	0.063
NAND Tunnel oxide (A)	90	70	70	60
NAND Width	0.140	0.100	0.080	0.063
NAND Wing (um)	0.060	0.060	0.040	0.0085
NAND Interpoly (A)	150	145	145	145
Stacked gate spacing (WL)	0.18	0.12	0.09	0.063
FG poly Spacing	0.08	0.06	0.05	0.05
NAND Prog Vpp (V)	16	19	17	17
Gamma	0.67	0.66	0.64	0.57
Cell size Y	0.32	0.24	0.18	0.126
Cell size X	0.34	0.28	0.21	0.13
Cell size	0.1088	0.0672	0.0378	0.0164
NAND Cell size (F2)	5.55	4.67	4.67	4.55

Table 2: NAND Scaling

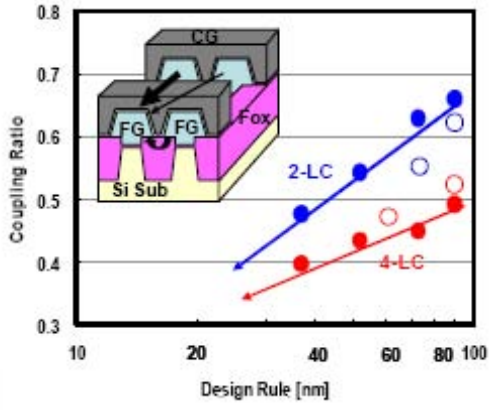


Fig 1: Coupling ratio degradation with DR Scaling [1]

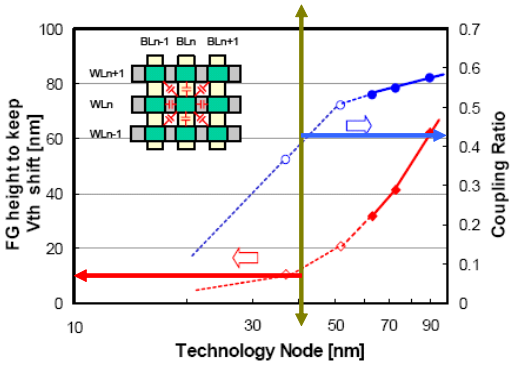


Figure 3. Coupling ratio trend as a function of technology node.

Fig 2: Coupling ratio vs. FG Height [2]

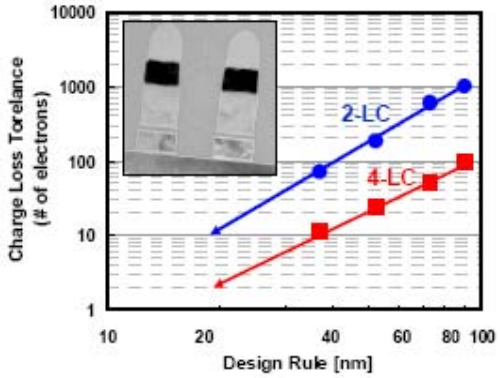


Fig 3: Charge loss tolerance with DR Scaling [1]

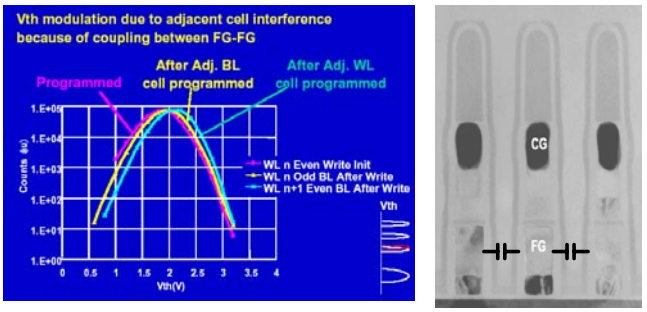


Fig 4: Cell to Cell interference issue

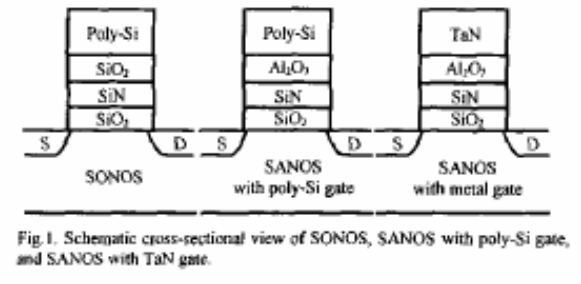


Fig. 1. Schematic cross-sectional view of SONOS, SANOS with poly-Si gate, and SANOS with TaN gate.

Fig 5: From SONOS to TANOS [4]

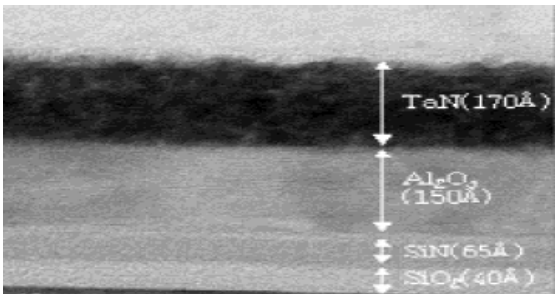


Fig 6: TANOS TEM [1]

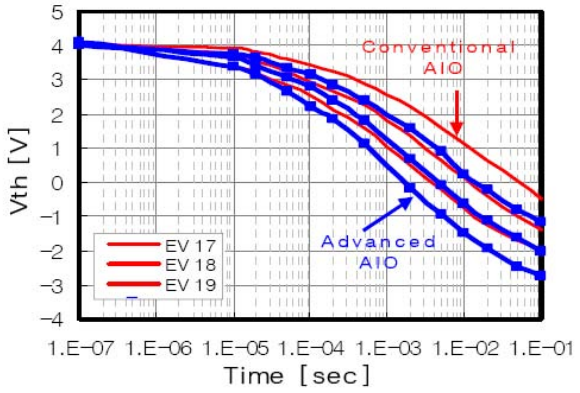


Fig 7: TANOS Erase characteristic [1]

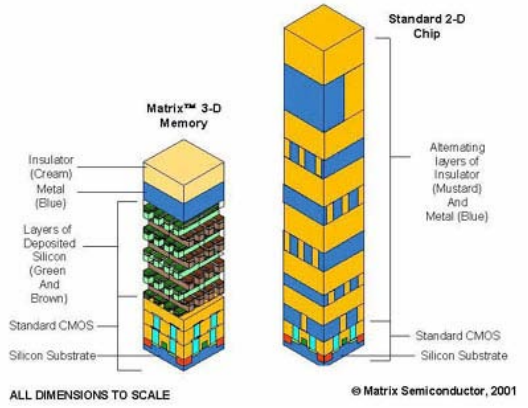


Fig 8: 3D stacked memory

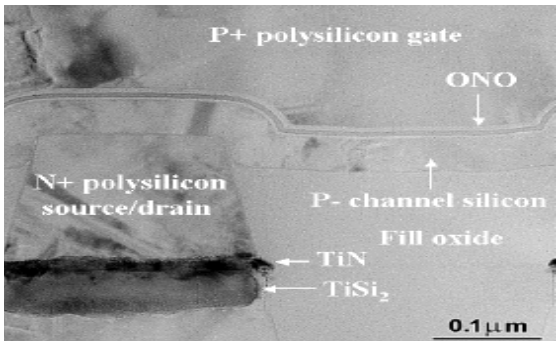


Fig 9: 3D stacked SONOS TFTs [6]

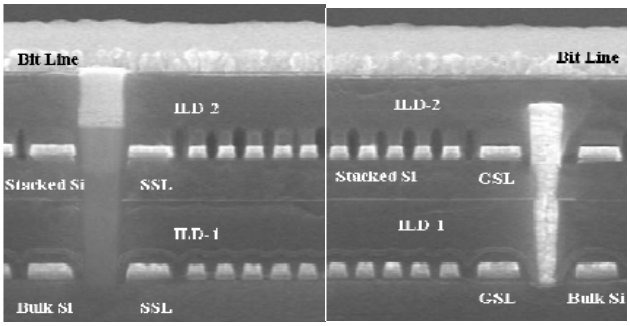


Fig 10: 2 level stacked TANOS [7]

	1E5 cycle Count	Byte P/E	Page P/E	High Density	Design Complexity	Low Cost	Zero Cost
EEPROM	Yes	Yes	Yes				
FLASH	Yes		Yes	Yes	Yes		
MTP	1K					Yes	
OTP	1 shot						Yes

Table 3: Embedded memory attributes

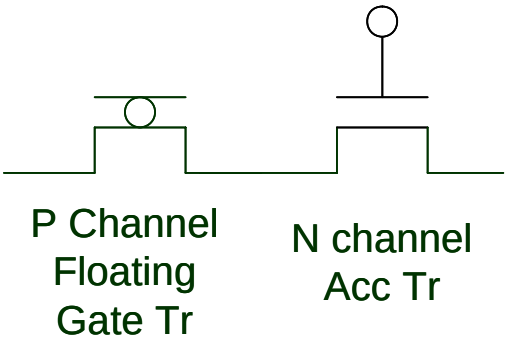


Fig 10: Free OTP Concept [8]

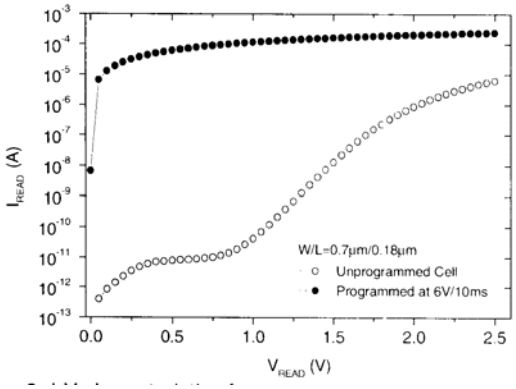


Fig 11: Free OTP Programming [8]

The Moving Bits: Generation and Annealing

Samir Mouhoubi [a], Thierry Yao [b], Antony Lowe [b], Pierre Gassot [b], Frederic Lalande [a]

[a] L2MP – Polytech'Marseille UMR CNRS 6137 IMT-Technopôle de Château-Gombert 13451 Marseille Cedex 20
[b] AMI-Semiconductor Belgium BVBA Westerring 15, B-9700 - Oudenaarde, Belgium

Abstract

The aim of this paper is to bring new information about the temperature behavior of the so called Moving Bits. A set of dedicated experiments allows us to build a global understanding of the mechanisms responsible for generating these bits as well as the parameters able to cure these defects.

We show that High Temperature impacts the Data retention in opposite ways: enhancement of the leakage in the short term, and annealing of the moving bits in the long term. We pointed out the difficulties that it brings to the industry to define worst case conditions.

The work performed makes it possible to offer practical solutions for the industry. Indeed, with the information gathered we know that to reduce the bit failure rates of the memory devices, it is better to: 1) avoid erasing at high temperatures, 2) anneal Moving Bits by performing bakes on erased cells, and 3) 1 week bake is optimum to obtain the best curing results.

1. Introduction

The reliability of Non Volatile Memory (NVM) devices has been studied by the microelectronics industry for several years. Data retention of floating gate (FG) based non-volatile memories (NVM) attracted a lot of attention since it is shown that stress-induced leakage current (SILC) through the dielectric layers isolating the FG may cause dramatic failure. Indeed, SILC can generate moving bits (MB), whose threshold voltage drifts excessively, leading to failure in a relatively short period.

MBs are a small population of extrinsic bits in a large memory array. Their characteristics have been extensively studied. Thus, it is known that MB can start and stop drifting. Moreover, programming and erasing generate more MBs. Furthermore, a high temperature bake can anneal out MBs if the tunnel oxide is thicker than 8nm. In a previous article, we have shown that the leakage current is thermally activated with an activation energy of 0.3eV. The goal of the present paper is to complete the picture with additional data about the impact of the temperature on the generation and the annealing of the MBs.

2. Methodology and material

To study MBs, several methods have been used by different authors: The slope of the Cumulative Distribution Functions (CDFs) of the cell's V_T [1] or the equivalent cell principal [2][3][4]. Other authors use the average of the 10 or 100 fastest bits (those of the CDF tail extremity) [5]. For this study we have developed the following method: MBs will be described by the number of bits failing at a given failure criterion (fig. 1). We define a failing bit as the MB whose threshold voltage V_T crosses a fixed value. The main parameter that will be used to evaluate the MB character is the Bit Failure Rate (BFR); which represents the ratio between the number of failing bits and the total number of bits in the array [6].

In this study we use memory devices of 1.4Mb. All the bits are programmed and erased once before starting the data retention test.

3. Moving bits generation

Programming and erasing is known to degrade the oxide and therefore to increase the number of MBs [7][6]. **Figure 2** shows that the BFR is almost 2 orders of magnitude higher when comparing devices cycled 100x and 1x. The aim of the current section is to show the impact of temperature on the degradation due to programming and erasing.

To investigate the effect of the cycling temperature on the BFR, several memory devices were cycled 100 times at different temperatures: -45°C, 0°C, 25°C, 85°C, 100°C, 125°C and 155°C. After that, all the devices were kept at room temperature and their V_T was read out periodically (Data Retention or DR test). The results show that increasing the cycling temperature increases slightly the BFR (fig. 3). Indeed, data have been fitted with an Arrhenius law using an activation energy of 0.05eV. The impact of cycling temperature is relatively small compared to the effect of cycling itself; however, it is large enough to double the BFR when increasing the temperature from 25°C to 155°C.

In the previous paragraph, we showed that the higher the number of cycles, the worse the results in terms of BFR. We found also that cycling at high temperature is slightly worse than cycling at RT. In this paragraph, we aim to determine in which proportion Erasing and Programming at HT degrades the oxide (since cycling is nothing but Erase and Program operations).

Thus, four groups of devices are prepared by cycling the devices so that the programming temperature is independent of the erasing one. The devices of each group are characterized by the temperature parameters T_{prog} and T_{era}, where each temperature parameter is either 25ºC (referred to as L in fig. 4) or 155ºC (referred to as H). With this convention, the label "HL" represents devices that were programmed at High temperature and erased at Low temperature. From fig. 4 one can conclude that the degradation at HT of the tunnel oxide is mainly related to the erasing (BFR at HH about 3 times higher than BFR at LL).

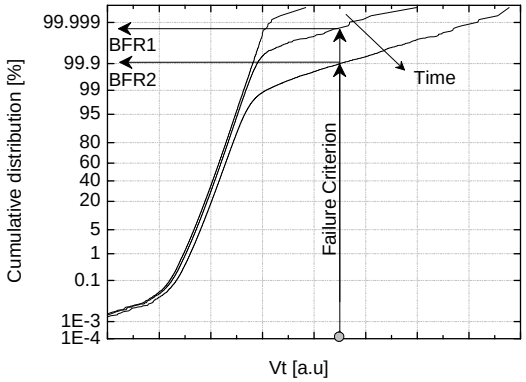


Figure 1: Example of a cumulative distribution showing the evolution of the Vts of a memory array with time. The BFRs are extracted at the intersection of the tails with the failure criterion.

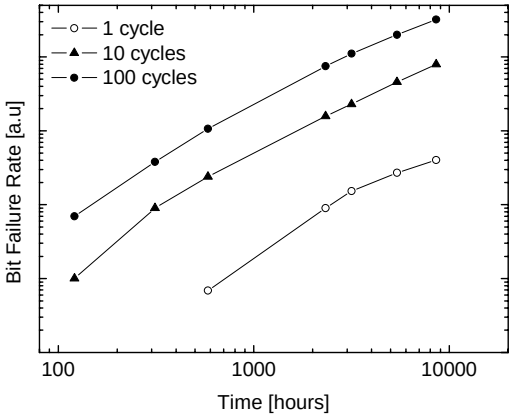


Figure 2: Evolution of BFR when increasing the number of P/E cycles (data retention at room temperature). Cycling increases the number of failing bits.

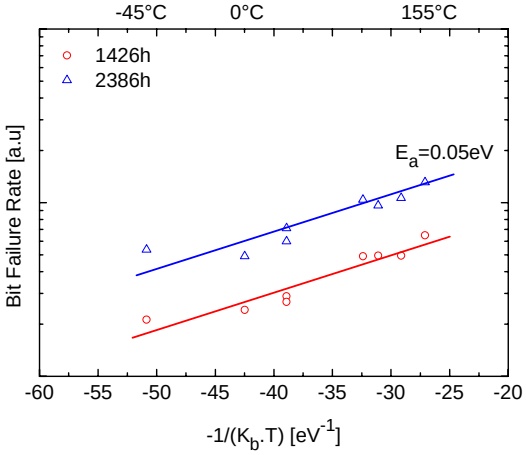


Figure 3: Bit Failure Rate evolution versus Temperature of cycling. Devices cycled 100x.

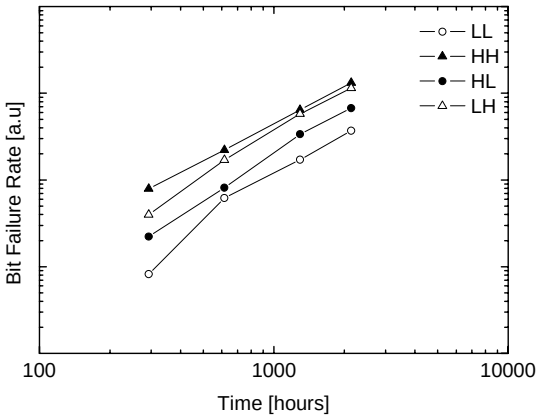


Figure 4: Bit Failure Rate depending on P/E temperature. Program/Erase = PE=LL, HH, HL, LH where "L" stands for "Low temperature" and "H" for "High temperature".

4. Temperature annealing

The impact of temperature on the dynamic of MB appearing was investigated by several authors [4][5][8]. They pointed out the existence of a worst functioning temperature namely 60°C-80°C and an annealing of the defects above 150°C. We showed, in our previous article [6] that this worse case temperature is time dependent and results from two opposite mechanisms: the temperature activation of the SILC and the temperature activated annealing of the defects that cause the SILC. By means of a dedicated method the two effects were dissociated, making possible the calculation of the activation energy (Ea) of the leakage mechanism. The value of Ea found (0.3eV) is independent of the FG voltage and the number of program/erase cycles (fig. 5). We showed also that an annealing occurs at a temperature as low as 60°C but needs more time to be observed.

To conclude, a 150°C functioning temperature is known to anneal the oxide defects and hence to reduce the number of MBs. A lower temperature will need more time to anneal the defects; a higher one will not be acceptable since most of the plastic packages used by industry are meant to sustain a temperature lower than 150°C. So, in the next sub-sections we will focus on the impact of 150°C prebake (combined with other parameters) on the reduction of the MB population. Thus, we will point out the MB annealing according to the duration of the prebake at HT and the charge contained in the FG during the prebake.

4.1 High temperature pre-bake duration

The aim of this Pre-Bake at 150°C is to determine the optimum duration that anneals most of the defects. Several devices were pre-baked for: 1 day, 3 days, 1 week and 2 weeks. They were then set to the same electrical state and put in a DR test. Notice that devices without pre-bake are used as a reference. The results are quite interesting: increasing the pre-bake duration reduces the BFR. For example, a pre-bake of 1 week reduces the BFR by a factor of 3 compared to a pre-bake of 1 day (fig. 6). The figure shows also that the BFR saturates after 1 week of pre-bake. That means the optimum pre-bake duration is about 1 week.

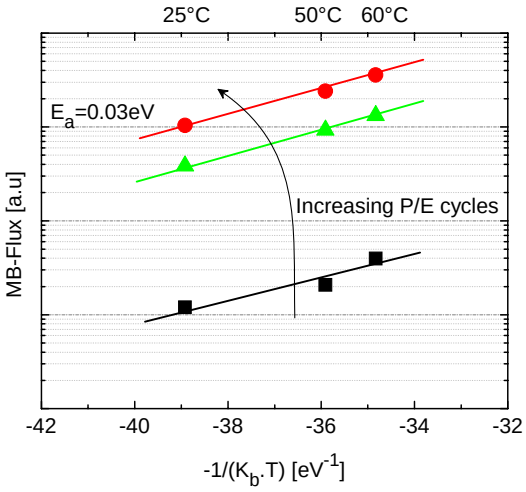


Figure 5: Moving bit flux versus the inverse of the thermal energy. Calculation of the activation energy for three levels of P/E cycles: 1, 10^2 and 10^4. The activation energy Ea is independent from the cycling level (parallel lines).

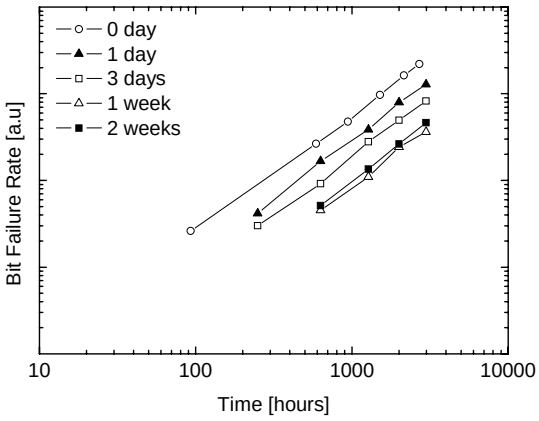


Figure 6: Bit Failure Rate of devices for different pre-bake durations.

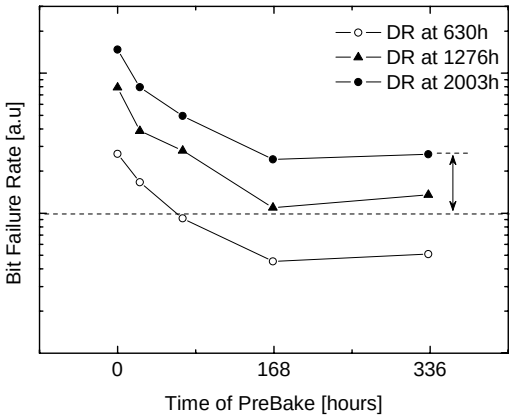


Figure 7: Evolution of BFR with time of prebake (for devices up to 2003h of DR).

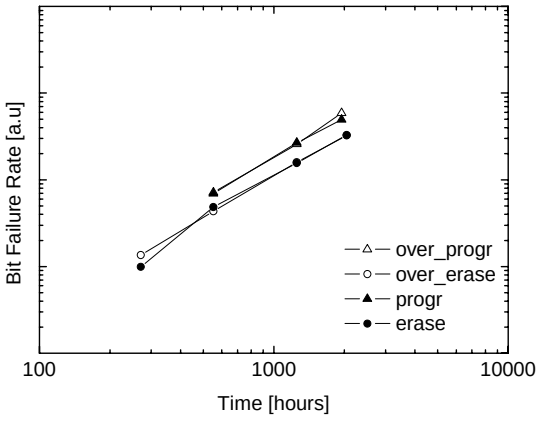


Figure 8: Bit Failure Rate of devices containing different amount and/or nature of charge in the FGs.

Besides, fig. 7 displays the evolution of the BFR according to pre-bake duration at 630h and 1276h of DR. It is interesting to mention that 1 week of pre-bake reduces the BFR of non-baked devices by almost 1 order of magnitude (about 8 times). Notice also that 1 week of pre-bake brings the BFR of devices cycled 100 times almost to the BFR of a fresh device (93% of the failing bits were cured, which is a very good result).

4.2 HT pre-bake and FG charge

In this sub-section, we aim to investigate the effect of pre-bake on the BFR when the FGs of the cells do not contain the same amount and/or nature of charges. Thus, 4 groups were formed with devices that were put to the state: over_programmed (channel in accumulation), programmed (flat band voltage), under_erased and over_erased (inversion). The devices were then baked at 150°C for 1 week (the best condition to reduce the BFR, as mentioned above). Following the bake, the devices were all put into the same state, after which a regular DR test was performed. During 12 weeks, the BFR of these devices is monitored. As depicted in Figure 8, there is no impact of the amount of FG charges on the pre-bake cure effect (under-erased versus over-erased devices and programmed versus over-programmed).

It is also interesting to mention that the pre-bake of devices in the programmed state is less efficient in annealing than pre-bake of devices in the erased state.

5. Discussion

In **section** 3, we showed that at HT erasing is more degrading than programming. For the HiMOS™ cell we used in this study, erasing is done by ejecting electrons from the FG. In that mode, the substrate is in accumulation and hot holes are generated via band-to-band tunneling at the erase junction [9][10][11]. It is also known that the release of the hydrogen atoms that passivates the dangling bonds at the Si/SiO2 interface is

thermally accelerated [12][13]. Thus, more oxide-bulk defects are generated as well as interface states. Therefore, there could be a link between Moving Bits and the release of interfacial H^0.

As we demonstrated in **section 4**, the V_T drifts observed are the result of two opposing mechanisms: 1) the temperature accelerated leakage through the oxides of a given population of bits, and 2) the temperature cure of the defects causing the leakage. In practice, this raises a real difficulty for the industry. For example, which temperature to use for DR tests? After what duration the annealing lowers the BFR?

The difficulty of separating the annealing of MBs and the leakage enhancement is due to the use of the V_T CDFs that treat the MBs in a statistical way. This is the only manner to study the MBs. Indeed, tracking the MBs individually has no meaning since a single program/erase cycle will completely change the "Moving character" of the cells (MBs disappear and others appear). The only constant is the population itself: it remains unchanged, meaning that the number of MBs will always be the same (even the individuals have changed). One of the solutions that could be used by industry is to determine a worst-case temperature based on a short-term test. Long-term predictions will be pessimistic, but at least within the specifications.

In **section 4.1** we showed that the BFR curves reach a flat level after 1 week of pre-bake (this level is slightly higher than that of fresh devices). That could be explained by one of these theories: 1) the small population that is represented by this level is a kind of MB that HT cannot anneal (meaning that we have at least two different populations of MB [8]), or 2) the defect density after long pre-bake is so low that it is reset to approximately the same level by the single program/erase cycle performed before the start of the DR test.

In **section 4.2**, we investigated the impact of the FG charge on the cure effect of baking. It turns out that the bake is less efficient when the FG is negative (channel in accumulation) than when it is positive (channel in inversion). That brings some interesting information about the physics of the SILC since it suggests that charged defects play a role in the MB issue.

6. Conclusion

This paper completes the former studies on MB issues and deals more specifically with the aspects related to temperature behavior. The high temperature has two effects: 1) it enhances the leakage of the FGs,(in the short term) and 2) it cures the oxide defects (in the long term).

We show in this study that the optimum pre-bake duration at 150°C is 1 week. Furthermore, the pre-bake is more efficient on the erased state (negative FG state). Finally, the most degrading condition during the lifetime of the memory devices is showed to be the erase operation at HT.

Our study shows that the thermal history of devices must be under control to draw the right conclusions out of experiments and to make the correct predictions. Indeed, the cycling temperature as well as the temperature before the data retention test can significantly impact the results of a test. Finally, pessimistic predictions can be made by choosing the right temperatures during cycling and before and during DR test.

Acknowledgement

The authors acknowledge the IWT for the support to that work.

References

[1] A. Scarpa, in Int. Rel. Workshop, (2000).
[2] D. Ielmini, IEEE Electron Dev Lett; **23**, 40–42 (2002).
[3] A. Hoefler, in Int. Rel. Phys. Symp, pp 21-25, (2002).
[4] L.-C. Hu, in Int. Rel. Phys. Symp, pp. 643-644 (2004).
[5] H. Kameyama et al., 38th Int. Rel. Phys. Symp., pp. 194-199 (2000).
[6] S. Mouhoubi, T.Yao, in JNCS, Accepted for publication, ref 6-SiO2 C-2R1 (2006).
[7] T. Vermeulen, T. Yao and al, Eur. Sol. Sta. Dev. Res. Conf., pp. 269-272 (2004).
[8] A. Modelli, F. Gilardoni et al, in Int. Rel. Phys. Symp, pp. 61-66 (2001).
[9] D. Ielmini and al, IEEE, Trans on Dev, vol 49, NO.10, October, pp 1723-1728 (2006).
[10] M. Suhail and al, IEEE, IRPS, Dallas, pp 439-440 (2002).
[11] P. Hanmant and al, IEEE, IRPS, Dallas, pp 7-20 (2002).
[12] N. Bhat, M. Cao, in IEEE, TED, vol. 44, NO. 7, pp 1102-1108 (1997).
[13] S. Mahapatra, B. Kumar, in IEEE, TED, vol. 51, NO.9, pp 1371-1379 (2004).

Improvement of Retention and V_{th} Window in Flash Memory Device through Optimization of Floating Gate Doping

Chen Shen, Jing Pu, Ming Fu Li, and Byung Jin Cho
Silicon Nano Device Lab, Dept. of Electrical & Computer Engineering, National University of Singapore, Singapore 119260
Tel: 65-6516-6470 Fax: 65-6516-1103 email: *elebjcho@nus.edu.sg*

Abstract

We propose to use p-type doped floating gate with optimized doping concentration, which demonstrated improvement in retention and larger program/erase V_{th} window, especially for smaller capacitance coupling ratio cell which is important for future scaled Flash memory cells.

Introduction

As flash memory scales down, two pressing issues emerged, namely the need for multilevel operation, and a dramatic drop in the area of IPD layer. At 45nm technology node and above, coupling ratio may drop to 0.3-0.4 if current ONO stack is used as IPD, which results in significant loss in program/erase voltage window [1,2]. In this report, we assess one possible route of engineering floating gate (FG) that may enable further scaling of FG type Flash memories, which is the lightly doped p-type FG. The proposed cell structure demonstrated improved operation P/E window, especially for very low coupling ratio (~0.3), and much improved retention which is important for multi-level operation.

Lightly doped p-type floating gate (p^- FG)

I. Device operation principle

To understand the device operation principle of the proposed p^- FG, device simulation is performed. A custom PDE solver is used to solve 1D poisson equation, with charge-balance boundary condition applied to FG. Electron and hole direct tunneling is calculated using WKB approximation. Fig. 1 shows the band diagram of programmed state (V_{th}=3V) from simulation, with V_g=0. In the case of p-type FG, Fermi-level of FG aligns to the bandgap of Si substrate which greatly reduces the electron leakage from FG to substrate. The tunneling barrier is also high for valence band electrons to escape from FG. Calculated direct tunneling current of p-type FG is 7 orders of magnitude lower in retention state than that of n-type FG. However, it is reported that heavily doped p^+ FG shows very slow erase due to increased barrier height against tunnel oxide [3]. To overcome such problem while maintaining the advantage of good retention property of p^+ FG, here we suggest the use of lightly doped p-type FG instead. If p-type FG doping concentration is carefully selected so that the inversion occurs at the tunnel oxide interface during erase as shown in Fig. 2 (c), there are sufficient electrons in conduction band at the interface, by which similar erase characteristics as n-type FG can be achieved. The very short carrier lifetime in polysilicon will guarantee the amply supply of electrons through thermal generation process. When negative gate voltage reduces, the electron concentration at the interface quickly reduces as shown in Fig. 3, thereby we can ensure the improved retention as the case of p^+ FG. Fig. 4 shows the erasing speed as a function of p-type FG doping concentration. Erase time exponentially increases when doping is above 4×10^{19} cm^{-3}, however, below 3×10^{19} cm^{-3}, we can achieve almost the same erase speed as that of n-type FG if additional 1V higher erase voltage is used. This additional 1 V requirement is due to the voltage drop across the depletion region in p^- FG during erase. Another important advantage of p^- FG is the improvement in program/erase V_{th} window when the area of IPD becomes small as shown in Fig. 5. This V_{th} window improvement is mainly due to the stronger ability of p-type FG to store electrons, which manifests in much higher saturation V_{th} in programming. Therefore, the use of p-type FG would be preferred for cells with small IPD area, which is likely the case for scaled flash memory beyond 45 nm technology node.

II. Experimental Results

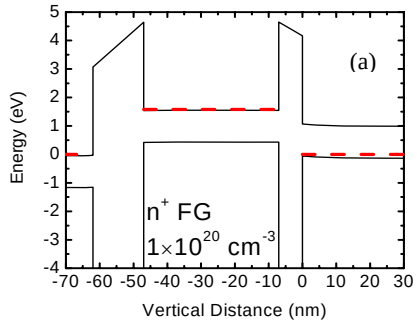
Memory transistors with different types of FG dopings were fabricated to verify the above proposal. The results in Figs. 6 and 7 confirms that p-type FG samples show much improved retention performance both before and after P/E cycling., compared to conventional n-type FG. The p^- and p^+ FGs show no clear difference in retention property. The programming speed is weakly dependent on p-type doping, and p^- FG has comparable programming speed to n-type FG (Fig. 8). Erase speed depends very significantly on doping concentration in p-type FG, as predicted in the simulation. Lightly doped p-type FG has erase speed as fast as n-type FG (Fig. 9). The trend is well agreed with simulation results. The coupling ratio (CR) of the cells are varied by varying the over-lapping area between control gate and floating gate (Fig. 10). When the area of the IPD layer is reduced, coupling ratio drops, and P/E voltage window reduces. However, p-type FG exhibits much larger V_{th} window even at very small CR (Fig. 11). At CR=0.3, n-type FG sample totally lost V_{th} window, while p- FG still maintain V_{th} window as large as 10 V. This is an important advantage for future flash devices with small IPD area, which can defer the introduction of high-K for IPD.

Conclusion

We have demonstrated that retention in Flash memory device can be significantly improved by FG engineering. These techniques can increase the lifespan of FG type Flash memory devices, without modification of tunneling oxide or IPD.

References

[1] ITRS, online at http://public.itrs.net, ed. 2005.
[2] K. Kim, *IEDM 2005*, pp.333-336.
[3] K. Kuo, *IEEE TED*, 51(2):282 –285, 2004.



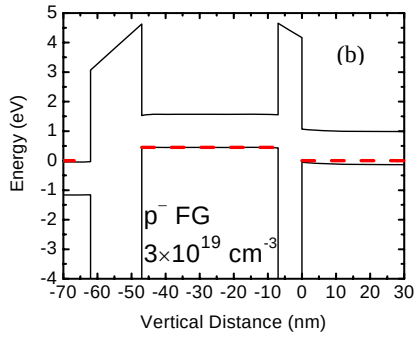
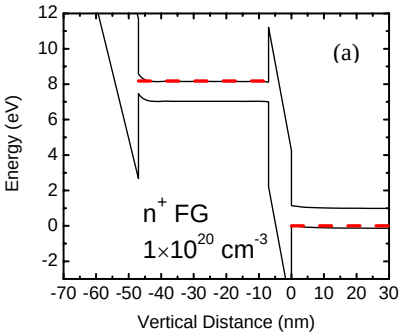
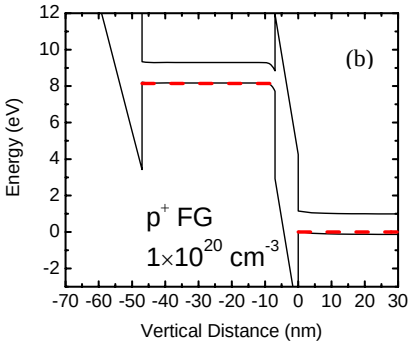


Fig. 1. Band diagram of Flash memory transistor under retention for a) n-type FG and b) p-type FG. Charge loss in p-type gate is much less due to higher barrier height for valance electrons and alignment of E_F of FG to the bandgap of Si substrate.





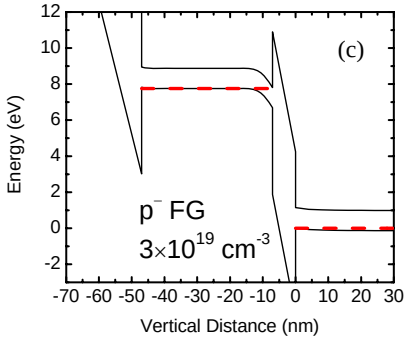


Fig. 2. Band diagram of Flash memory transistor during erasing when Vg = -20V, for a) n^+-type FG (1×10^{20}cm^{-3}), b) p^+-type FG (1×10^{20} cm^{-3}), c) p^--type FG with lower doping (3×10^{19} cm^{-3}). With lower p-type doping, the p-type FG can be inverted at tunnel oxide interface, which allows efficient erase.

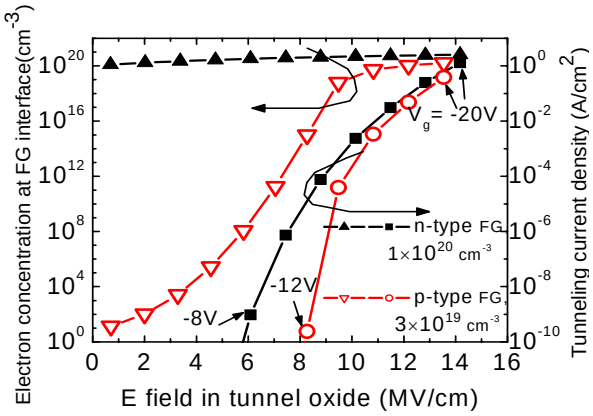


Fig. 3. Simulated electron concentration in FG at tunnel oxide interface and tunneling current as a function of the E-filed in tunnel oxide. Tunneling current is better controlled by E-field (steeper slope) if p-type FG is used, due to the modulation of electron density by E-field.

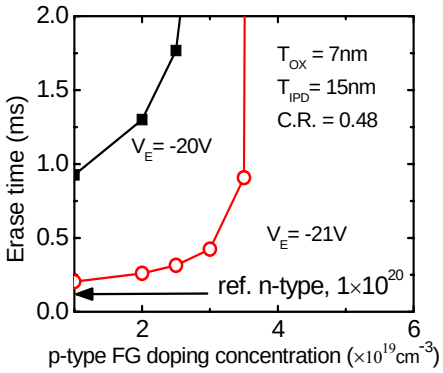
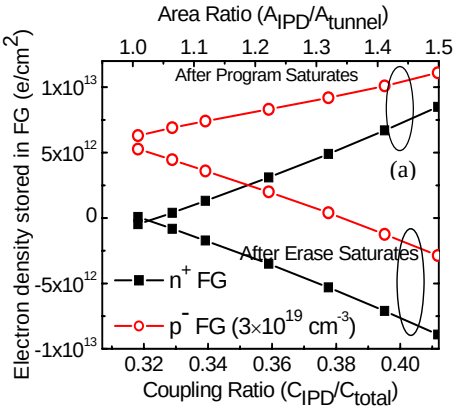


Fig. 4. Simulated erase speed as a function of p-type FG doping. Low p-type doping in FG is desired for fast erase. Erase time is defined as the time required to reach V_{th}=-3V.



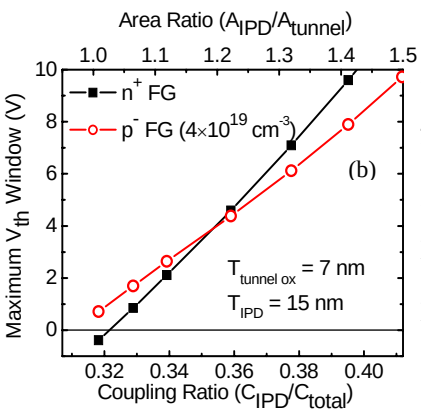


Fig. 5. Simulated maximum program /erase window for n-type and p-type FGs. For memory cell with very small coupling ratio, cell with n-type FG totally loses V_{th} window, while p-type FG cell maintains a positive operation window. This is because p-type FG is able to held more electrons during programming.

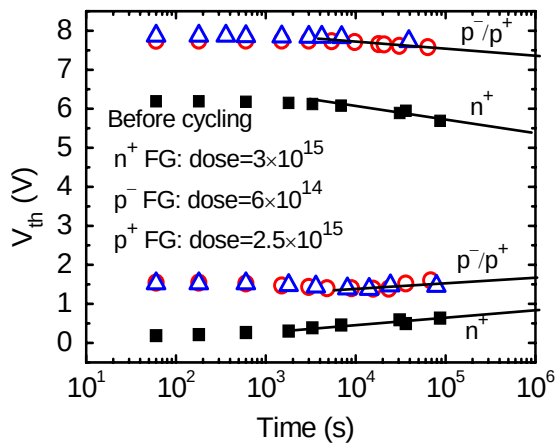


Fig. 6. Experimental retention performance of the n-type and p-type FG samples, before program/erase cycling. P-type FGs show better retention.

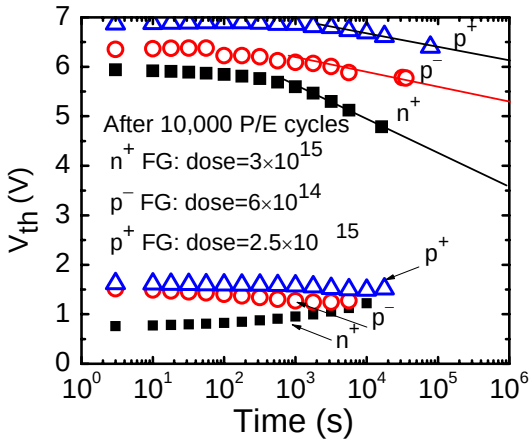


Fig. 7. After 10^4 program/erase cycles, p-type FGs again show much better retention than the n-type FG.

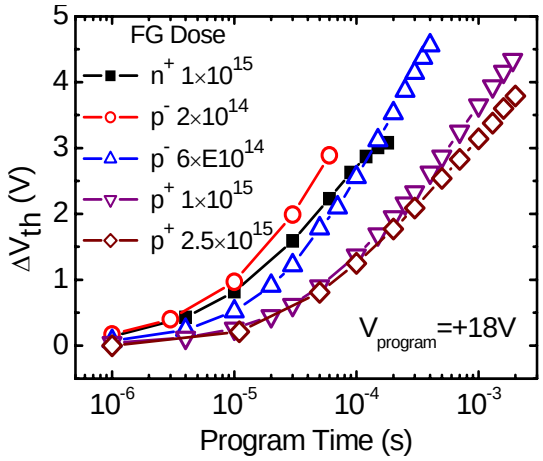


Fig. 8. Program speed for different FG doping. It is seen that p^- FG has comparable program speed to n^+ FG.

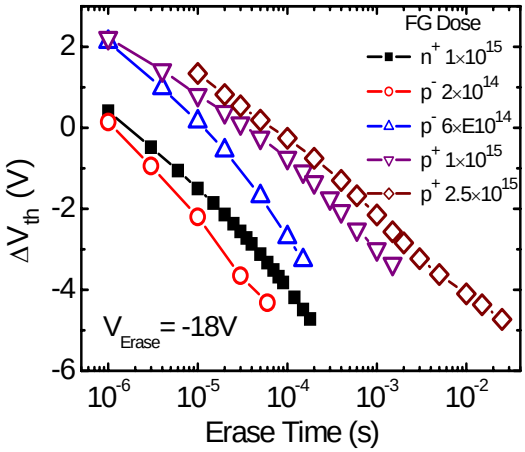


Fig. 9. Erase speed for different FG doping. Lightly doped p^- FG shows comparable erase speed to n-type FG, while increasing p-type doping concentration significantly slows down the erase.

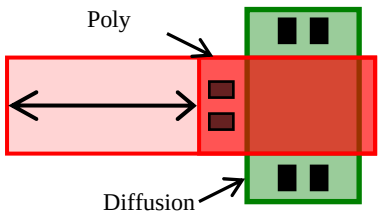


Fig 10. The variation of coupling ratio in the test pattern was done by stretching the area of poly mask on top of field oxide.

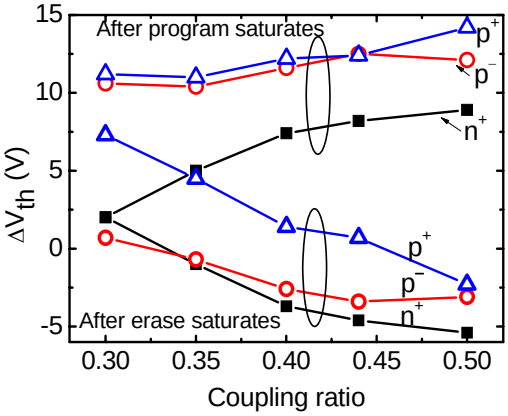


Fig. 11. The maximum program/erase window for different FG and coupling ratio. When CR = 0.3, n^+ FG totally loses V_{th} window, while p^- FG maintains 10V V_{th} window.

A single-poly NVM based on a CMOS inverter with a common floating gate

Y. Roizin, A.Fenigstein, V.Kairys, Z. Kuritsky, A.Lahav
Tower Semiconductor Ltd., P. O. Box 619, Migdal HaEmek, 23105, Israel

Abstract

A new single polysilicon logic nonvolatile Flash memory based on a CMOS inverter with a common floating gate was developed and integrated into the standard 0.18um process without additional masks. The memory cell is programmed and erased by band-to-band tunneling (BBT) electrons and holes and has a select transistor for data read-out. The main advantage of the proposed approach is low currents and voltages in all operational regimes. Experimental results confirming high endurance and retention of the proposed memory are provided. Applications requiring low-cost small/middle size embedded memory with high reliability are targeted.

Introduction

The embedded Flash or EEPROM memories are used for system configuration, security, backing up, etc., and general data storage The memory size in embedded applications is usually lower than in stand-alone Flash memories. Low cost and high reliability of the end products are mandatory [1,2]. Memories of this type are typically single-poly solutions that employ Fowler-Nordheim (F-N) injection mechanism in programming and erase [3,4,5]. The control gates of single poly EEPROM memories are usually formed in the substrate and designed as separate capacitors that occupy relatively large areas. The F-N injection starts at voltages much exceeding Vdd, so that memory chip designs include special high voltage transistors and charge pumps to generate high voltages. Hot channel electrons and holes were also used for programming the single–poly devices [6-8]. Though the corresponding memory cells can be significantly scaled down in some cases, large programming currents are their evident limitation in low power applications. BBT tunneling of electrons from the p-channel transistor was used for programming of single-poly memory in [5]. This allowed strong decrease of voltages and currents in programming. Nevertheless, high voltages were still needed in erase. In most single-poly designs sense amplifiers are used in the read-out circuits to compare the current of the memory transistor with the current of a reference cell. Sense amplifiers are energy consuming and imply limitations in applications that require ultra-low power operation, such as RFID systems and memories with enhanced security.

In the proposed Complementary Flash ("C-Flash") memory, both programming and erase are performed by BBT with low voltages and currents [9]. The read-out scheme is "logical" (without a sense amplifier). In this paper we focus on the C-Flash memory cells employed in the developed 16kBit Array-TEGs with no special high voltage devices.

C-Flash Device Configuration and Operation

The C-Flash memory cell is schematically shown in Fig. 1. It consists of a CMOS inverter with a common floating gate and an NMOS select transistor. The PMOS and NMOS transistors of the inverter share the same FG and CG. In the read-out mode, the cell is selected by choosing the word line (WL) select transistor and the output bit line (BL). Programming and erase are performed by applying voltages V_{CG}, V_{PS} and V_{NS}. The typical operation conditions are listed in Table 1. In programming, the electrons are created in the drain region of the p-channel transistor by the BBT mechanism, accelerated in the lateral field and injected into the common FG. Erasing is done in a similar way by BBT holes from the n-channel part of the inverter [9]. BBT programming allows to decrease the absolute values of the employed voltages below the 5V level.

When electrons are injected into the FG, V_T's of PMOS and NMOS transistors are increased, and the transfer curve of the floating gate CMOS inverter shifts as shown in Fig. 2. The transfer curve shows the voltage at the BL as a function of the V_{CG} potential. Under a fixed V_{CG_Read} chosen between the erase and program transfer curves, the output of the inverter is clamped by V_{PS} through the on-state PMOS transistor. In this case, the NMOS part of the C-Flash cell is in the off-state. In the erased state of the inverter, the PMOS part is in the off-state and the NMOS is in the on-state. Thus, we have V_{NS} at the output of the inverter. The output of the cell is controlled by V_{PS} and V_{NS} and is not affected by the positions of the transfer curves at the V_{CG} axis as long as they do not cross the $V_{CG\ Read}$ level. The maximum programming current (at the beginning of the programming pulse) is below $5*10^{-7}$ A/cell and the maximum erase current is of the order of 10^{-7} A/cell for the design of the C-Flash cell shown in Fig.3. This is significantly lower compared with corresponding values in most Flash memory designs. The corresponding programming and erase characteristics for selected and unselected cells are presented in Fig.4. The shift of transfer curves between the program and erase states is of the order of ~ 3.5V. The discussed memory allows single shot programming and erase without verify operation.

Device Manufacturing and Optimization

The devices described in this paper were manufactured using the standard Tower Semiconductor Ltd. 0.18um CMOS technology (already including a deep n-well mask). After the

formation of the bottom oxide (~7 nm CMOS HV GOX) and polysilicon deposition, FG, CG and CMOS gates were patterned with the same mask (see layout in Fig. 3). The drain fields in the N-channel and P-channel transistors of the C-Flash were enhanced by changing the spatial position of the LDD implants in the C-Flash area compared with the core CMOS technology. Only the existing LDD implant masks were used. An interdigitated pattern of FG and CG formed the CG capacitance with vertical wall plates. During the spacer formation (oxide plus nitride) and pre-metal dielectric deposition, the gap between FG and CG poly lines was mostly filled with silicon nitride, thus forming the CG dielectric with the effective dielectric constant k~5.5 (estimated from device simulations). The dimensions of NMOS and PMOS transistors (ratios of channel width to channel length) in the design shown in Fig.3 are 0.42/0.32um and 0.42/0.28um, respectively. The intrinsic threshold voltages (measured with interconnected FG and CG) are 0.7V and –0.7V for the n-channel and p-channel parts of the C-Flash.

The unit cell area in Fig 3, b is ~26um^2. However, it can be reduced to less than ~15 um^2 by layout optimization. This optimization takes into account peculiarities of C-Flash operation. There is a trade-off between CG capacitance to the FG, capacitance of the FG to the n-well and p-well and operation voltages. For example, placing the FG over the n-well allows to shift the transfer curves to lower voltages and thus decrease the CG voltage in the read-out.

The HV (70A GOX) transistors of the core 0.18um process work at 5.5V for ~1h in a wide temperature range without significant degradation. This ensures more than 10^5 program/erase cycles. Thus, there was no need for special HV devices in the process flow.

The transfer curves for increasing cycling numbers and voltages at the output of the cell are presented in Fig. 5a. Though there are shifts of transfer curves connected with electron and hole trapping in the gate oxide of the n-channel and p-channel transistors (typical for all EEPROM memories that use hot carriers for programming and erase), there are no changes at the output (at the BL) of the cycled cell (Fig.5b) After the 250^{o}C/1hour bake of 100K cycled cell, there was still a reliable readout that corresponded to the window in the positions of the C-Flash inverter transfer curves of ~ 1V.

The characteristics of the 16 Kbit C-Flash Array TEG are presented in Fig.6. The spread of transfer curve positions does not exceed 0.5V both in programmed and erased states. The endurance/retention results of the memory arrays are consistent with those of the single cells.

Summary and Conclusions.

A new CMOS logic memory (C-Flash) featuring low voltage and low current in all operational regimes has been proposed, fabricated in the single-poly embodiment and verified at the level of 16kBit Array-TEG. A standard CMOS technology with no additional masks was employed. The memory demonstrated good endurance/retention performance (passed standard 10k cycle endurance/retention tests). The proposed memory can be useful in applications that require small/intermediate memory density, ultra-low power consumption in all operational regimes and high endurance/retention .

References

[1] C. Dary and P. Gendrier, IEEE MTDT *D*igest. 2002, p. 143.

[2] L. Chang, C. Kuo, C. Hu, A. Kalnitsky, ISDRS Digest, 1999, pp. 50-57.

[3] Jaroslav Raszka, Manik Advani, Vipin Tiwari, Laura Varisco, Narbeh Der Hacobian, Anurag Mittal, Michael Han, Al Shirdel, Alexander Shubat, 2004 IEEE International Solid-State Circuits Conference, presentation 2., 2004, pp.46-47.

[4] Yanjun Ma, Troy Gilliland, Bin Wang, Ron Paulsen, Alberto Pesavento, C. -H. Wang, Hoc Nguyen, Todd Humes, and Chris Diorio, IEEE Transact. Device and Materials Reliability, **4** 353 –358 (2004).

[5] Ted Chang, Kevin Huang, Binh Li , Mike Chen, Al Kwok, Alex Wang and Nader Radjy IEEE NVSMW, Monterey, CA. 1998.

[6] B. Song, "Flash EEPROM Cell and Manufacturing Methods Thereof", U.S. Patent 5,616,942, 1997.

[7] J-C. Lee, J-S. Kim and S. Kim, Journal of Korean Physical Society**, 41,** pp. 846-850 (2002).

[8] Kung-Hong Lee and Ya-Chin King, Symposium on VLSI technology, Kyoto, 2003, pp.93-94

[9] Y. Roizin, "Complementary Non-volatile memory cell", U.S. Patent, 6,788,576, 2004.

Table1. Typical operation conditions of CFLASH memory cell

	Program selected cell	Program unselected cell	Erase	Read
$\mathbf{V_{Nsel}}$	0V	0V	0V	V_{DD}
$\mathbf{V_{CG}}$	5V	-3V	-5V	1.5 V
$\mathbf{V_{NS}}$	Floating	Floating	5V	0V
$\mathbf{V_{P\text{-}Well}}$	V_{SS}	Vss	V_{SS}	V_{SS}
$\mathbf{V_{PS}}$	-5V	-5V	Float	1V
$\mathbf{V_{N\text{-}Well}}$	0	0	0	1V
Typical time	50 μs	-	3 ms	Depends on application

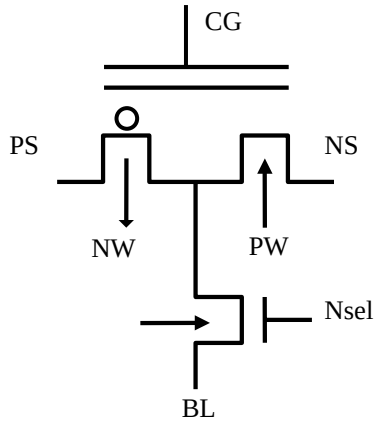


Fig.1. Schematics of CFLASH cell

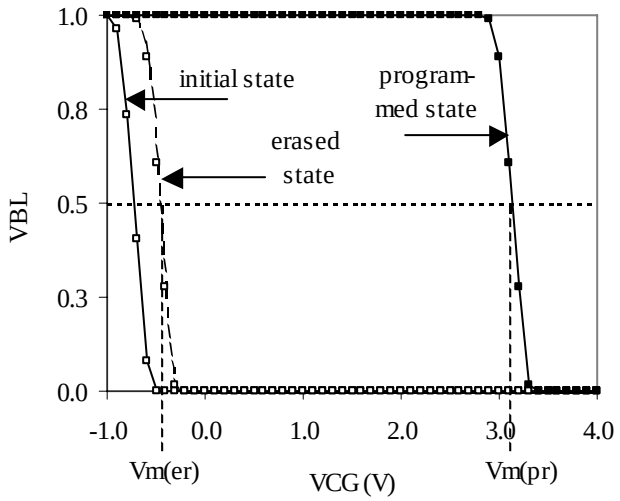


Fig.2. Transfer curves of program/erase.

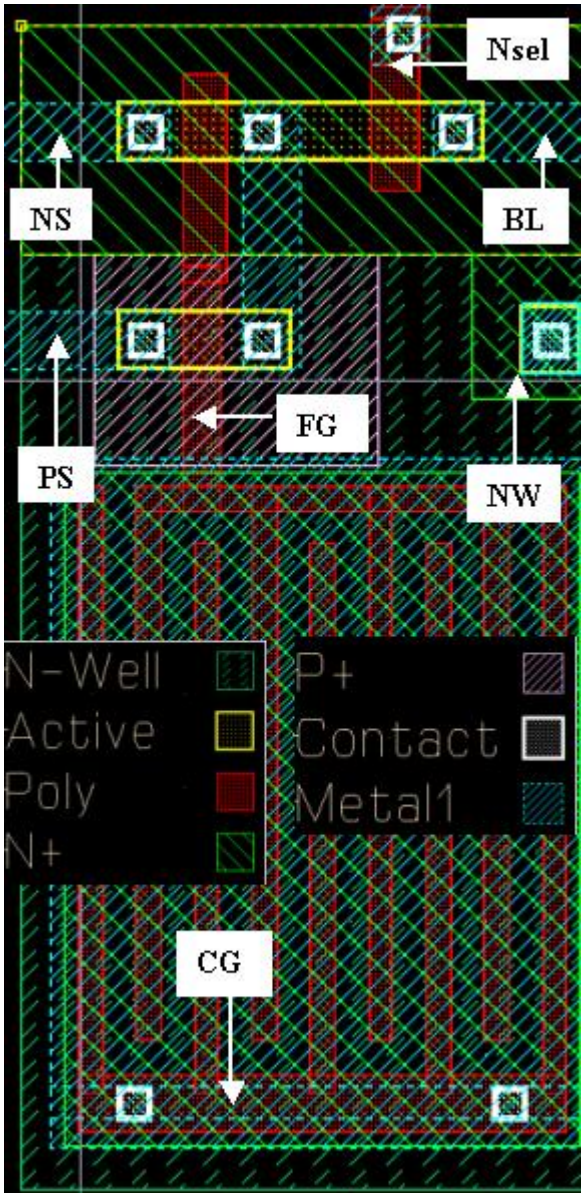


Fig.3. CFLASH cell layout

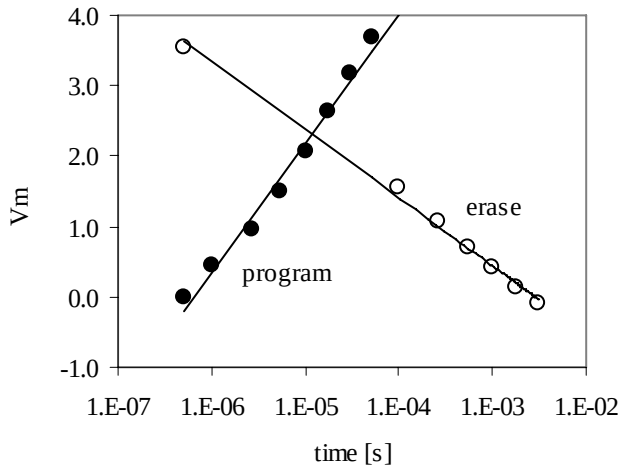


Fig.4. Program\Erase characteristics of CFLASH cell

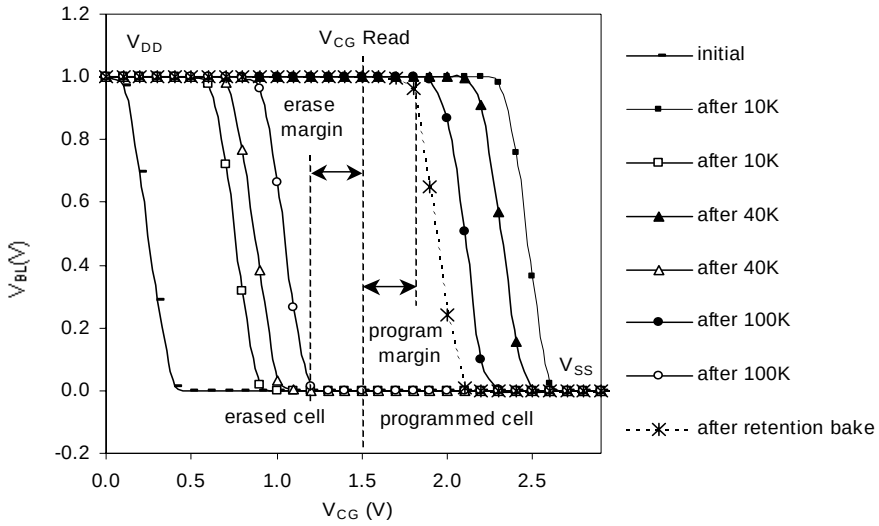


Fig.5a The transfer curves for increasing cycling numbers**.**

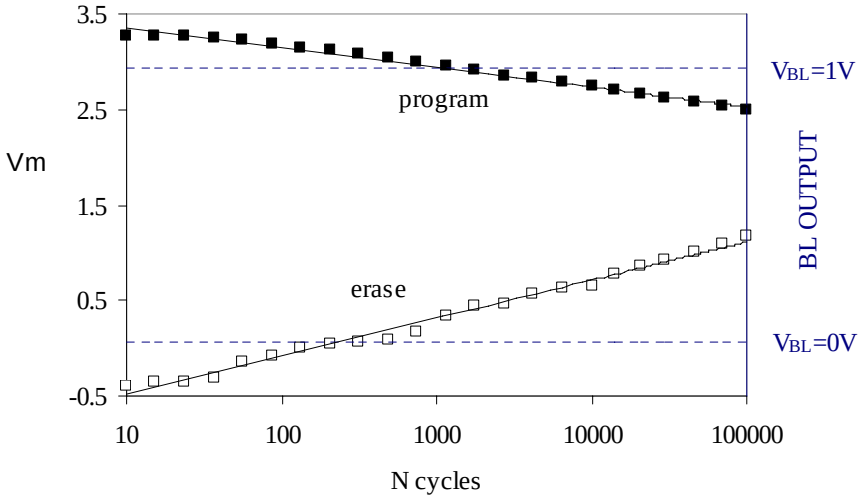


Fig.5b Vm and BL output at cycling.

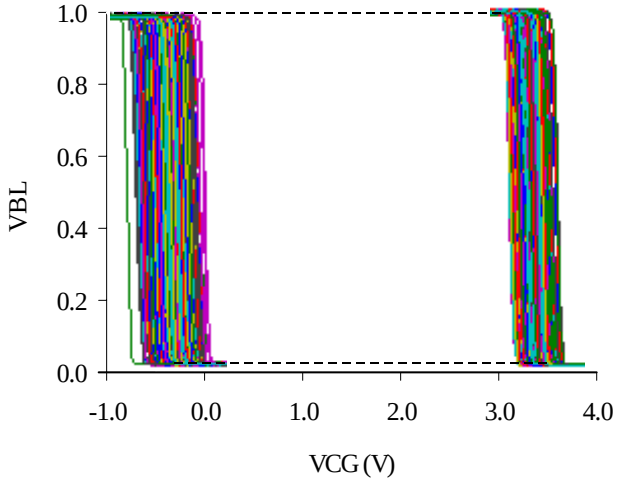


Fig. 6. The program\erase characteristics of the 16 Kbit Array TEG.

Introduction of HC (Hemi Cylindrical)-FET for Development of NAND CTF (Charge Trap Flash) Cell with 76nm pitch Technology

Sangyong Park, Byungjoon Hwang, Hyungkyu Park, Yunkyoung Lee, Sunghyun Kwon, Kwangseok Lee, Minjung Kim, Jooyoung Kim, Donghwa Kwak, Yongsik Yim, Jaekwan Park, Keonsoo Kim, Kinam Kim

Advance Technology Advanced Technology Development Team, Semiconductor R&D Center, Memory Business, Samsung Electronics Co., Ltd. Hwasung-City, Kyungki-Do, Korea, 449-900
Phone:+82-31-208-2004, Fax :+82-31-209-3274, e-mail: fuzzycom.hwang@samsung.com

Abstract

As the size of NAND CTF (Charge Trap Flash) cell is scaled, the degradation of cell performances such as the operational speed and cell window becomes severe. These degradations are caused by the scaled channel width and the reduced charge capacity of charge trapping layer, etc. In this paper, we propose a novel CTF cell transistor, called HC (Hemi Cylindrical)-FET. Using 76nm pitch technology, the effective active width of HC-FET is increased by 60% and the coupling ratio is improved by 34% compared to those of planar type cell. The improvement of the electrical characteristic with the HC-FET structure is observed in this study.

1. Introduction

In the NAND CTF memory, the coupling ratio of the cell is determined by the ratio of the capacitance between the charge trapping layer and the control gate, and the capacitance between the charge trapping layer and the Si-body [1][2]. In order to develop the NAND CTF (Charge Trap Flash) memory with the high density, we should consider a novel structure of cell transistor with a high coupling ratio, and HC (Hemi Cylindrical) – FET is suggested in this paper. Fig. 1 and 2 show three different types of CTF cell transistors and their coupling ratios as a function of active width. The coupling ratio of HC-FET is 1.34 times larger than those of planar transistor and Fin-FET due to its rounded channel shape. The coupling ratio of Fin-FET [3][4] shows the same value with the planar transistor from 60nm design rule, because the side gate of Fin-FET does not contribute to the coupling ratio anymore when its active space becomes shorter than 60nm as shown in Fig. 2. In this paper, the process integrated technology for the HC-FET with TANOS (Si/SiO_2 /SiN/Al_2O_3/TaN) [2] will be introduced, and the electrical data will be compared to that of planar transistor.

2. Fabrication and Experimental

The HC-FET type CTF cell is fabricated by similar process flow with the conventional TANOS cell except the active rounding process. The brief summary of process flow is shown in Fig. 3. First, STI (shallow Trench Isolation) process is performed with the depth of 2000Å and filled up a SiO2 in the field between actives. The hemi-cylindrical active is made by 750℃ annealing treatment with additional gases after 250Å of field oxide is recessed by wet etching. Then, TANOS (Si/SiO_2 35Å/SiN70Å/Al_2O_3 200Å/TaN50Å) stack is deposited and tungsten is used as the word line material. To define word line, followed by etch process of the TaN layer above active and in the valley between the prominent rounded actives. Fig. 4 is the TEM image of the planar (1) and HC-FET (2) NAND CTF memory with TANOS structure respectively.

3. Results and Discussion

The estimated channel width of HC-FET cell is approximately 59nm, 60% increased value compared to the 38nm active width. So, the on-cell current is 1.1uA/cell (1 order higher than that of planar cell) and the off-cell current is 5.9nA/cell (1 order lower) as illustrated in Fig. 5 and Fig. 6. The increased coupling ratio improves the sub-threshold voltage (~140mV/dec) drastically and lowers the program voltage about 0.05V compared to planar transistor as shown in Fig. 7 and Fig. 8 respectively. And the variation of the on-cell current is similar to that of planar cell (Fig. 9), but off-cell current shows the wide distribution compared to that of planar cell (Fig. 10). This can be explained by the wide distribution of the HC-FET channel width. These results mean HC-FET has advantages in improving controllability and cell speed, due to increased coupling ratio and effective active width.

4. Conclusion

To overcome many critical issues in developing NAND CTF (Charge Trap Flash) cell using 76nm pitch technology (38nm node), we propose a novel transistor, HC (Hemi Cylindrical)-FET in this paper. The effective active width of HC-FET is increased by 60% and also the coupling ratio between the control gate and Si-body is improved by 34% compared to that of the planar transistor. So we have developed the manufacturable HC-FET with improved electrical characteristics (10~15%).

References

[1] Kinam Kim et al., “Future Outlook of NAND Flash Technology for 40nm Node and Beyond”, IEEE NVSMW, pp. 9~11, 2006.

[2] Chang-Hyun Lee et al., “Multi-Level NAND Flash Memory with 63nm-node TANOS(Si-Oxide-SiN-Al2O3-TaN) Cell Structure”, VLSI Technology Digest of Technical papers, pp. 21~22, 2006.

[3] Suk-Kang Sung et al., “Fully integrated SONOS flash memory cell array with BT(body tied)-FinFET structure”, nanotechnology IEEE transaction, pp. 174~ 179, 2006.

[4] Donggun Park et al., "3-dimentional nano-CMOS transistors to overcome scaling limits", SSICT, pp. 35~40, 2004.

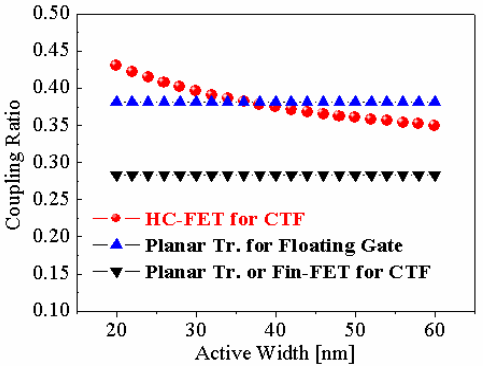
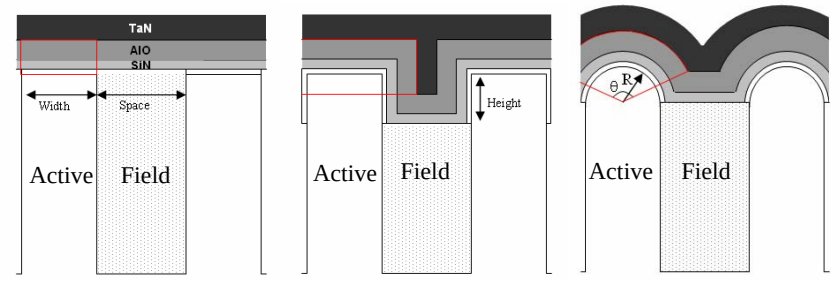


Fig. 1 Comparison of coupling ratio between control gate and Si-body of HC-FET, planar Tr. and Fin-FET for CTF(Charge Trap Flash) respectively.

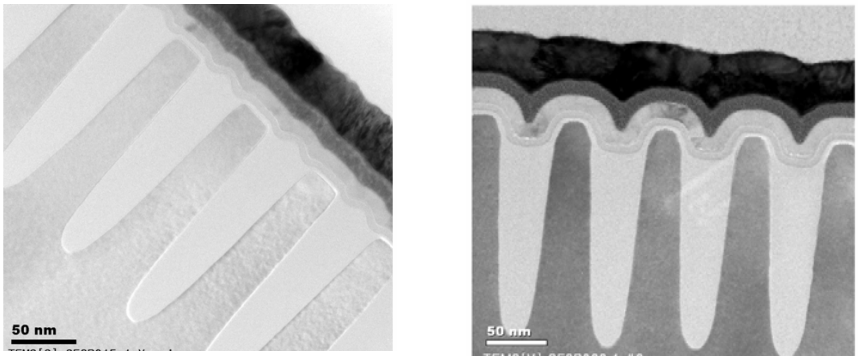


(1) Planar transistor (2) Fin-FET (3) HC-FET

Fig. 2 Structures of planar transistor, Fin-FET and HC-FET by using TANOS ($Si/SiO_2/SiN/Al_2O_3/TaN$) respectively.

- Active patterning: Bar : Space=38nm : 38nm
- Shallow Trench Isolation: 2000Å
- SiO_2 Gapfill
- **Field SiO_2 recess process: 250Å-wet etching**
- **Anneal process for rounding: 750℃ with addtional gases**
- Deposition of TANOS (Si/SiO_2 35Å/SiN70Å/ Al_2O_3 200Å/TaN50Å)

Fig. 3 The process sequence of HC-FET for 32ea NAND Flash cells-string.



(1) Planar transistor (2) HC-FET

Fig. 4 TEM Images .of planar transistor and HC-FET with 76nm pitch (38nm node) using TANOS (Si/SiO_2 35Å/SiN70Å/Al_2O_3 200Å/TaN50Å) respectively

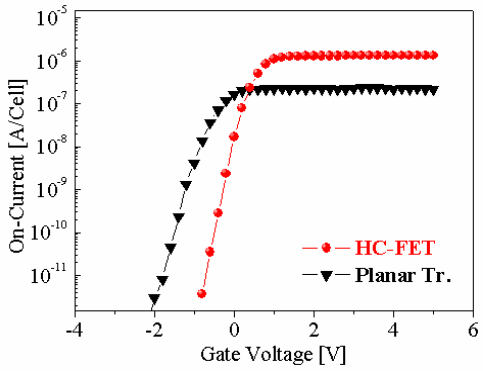


Fig. 5 Comparison of V_G-I_D curve between planar transistor and HC-FET

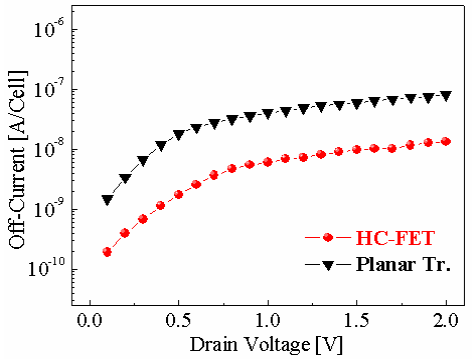


Fig. 6 Comparison of off state leakage current between planar transistor and HC-FET

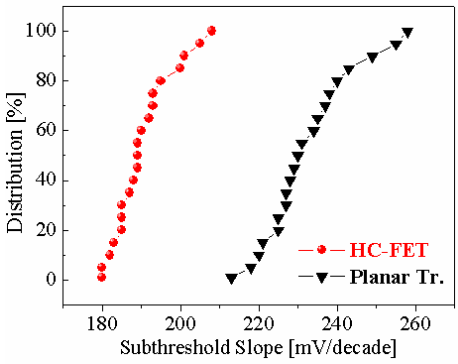


Fig. 7 Distribution of sub-threshold slope of planar transistor and HC-FET

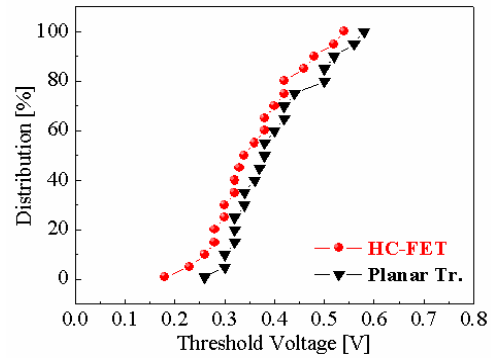


Fig. 8 Distribution of threshold voltage of planar transistor and HC-FET

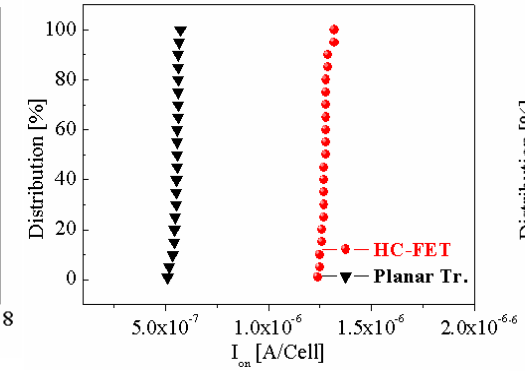


Fig. 9 Distribution of on state current of planar transistor and HC-FET @V_D=1V

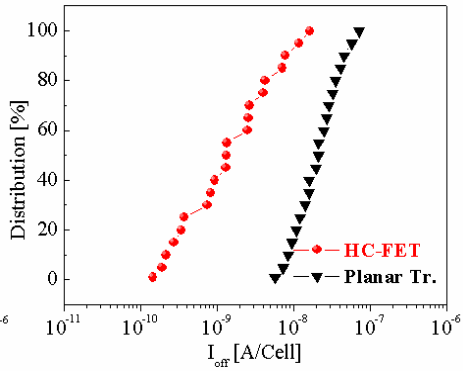


Fig. 10 Distribution of off state leakage current of planar transistor and HC-FET @V_D=1V

A self-synchronized, 1V operation read circuitry for high speed advanced embedded flash memories

Jimmy Fort, Jean-Michel Daga

Library and Design Tools Department, ATMEL, 13790 Rousset, France

Abstract

We propose a single ended self-synchronized read architecture to be used for very low voltage (1.2V+/-10%) embedded flash memories. The bit line variation due to the cell current is amplified to bias a latch circuitry before activation. All the basic operations are synchronized using a dedicated circuitry. 16ns random access time has been simulated on a 8Mb, 0.13µm embedded flash memory, using the proposed approach.

1. Introduction

Flash based solutions are becoming more and more popular [1] as they fit a new trend in producing a wide range of products with high performance, in smaller quantities, requiring faster time to market. Flash re-programmability provides a maximum of flexibility for code development. Shortened development time and faster introduction of new applications becomes possible. If the additional process cost of the flash solution can be balanced by the added flexibility, performance degradations compared to a ROM solution must be minimized as much as possible. For code management, execution in place from the flash memory is a must, removing the need for additional expensive cache memory. Starting from 0.13µm technologies, fetch operation beyond 100MHz is desirable on high end MCUs. In order to enhance flash read access performances, 128-bit architecture solutions based on the pre-fetch buffer concept have been proposed [2]. However, these solutions are not fully deterministic, and require code linearity in order to guarantee optimal performances. The overall system performance in case of many interruptions, such as in industrial control for example, is not predictable. Deterministic, high-speed system operation can be achieved by decreasing as much as possible the penalty due to the random access latency in case of jumps. Among the various parameters to be considered to optimize the random access time, such as the memory sectors size to minimize bit line and word line capacitances, fast current sensing operation is mandatory. With advanced technologies, very low voltage (down to 1V) sensing circuitry has to be developed. Various sensing schemes have been proposed for flash. They are fully asynchronous, and their function is twofold: firstly precharge the bit line voltage to a clamped value, and secondly sense the current flowing through the bit line and proceed to a fast current to voltage conversion. The main differences between the proposed circuits are in the way the current to voltage conversion is done. Some circuits use a comparator and require a dummy bit line connected to a reference cell [3,4], other circuits are single ended, and do not require any reference cell [5]. A fast solution based on a cell current amplification followed by a current comparison has been proposed in [6]. However, even if low voltage circuits have been investigated [7], no solution for 1V flash operation has been proposed. On the other hand, high-speed synchronous sense amplifiers based on a latch circuitry have been used for years on SRAM memories [8,9]. These circuits are well suited for low voltage operation, but they are fully complementary, as they amplify the voltage difference between the bit line (BL) and the inverted bit line (BLB). In this paper, we propose a single ended synchronized architecture to be used for low voltage embedded flash. Here, only the bit line variation is amplified to generate the latch DC voltage imbalance before activation. In addition, this structure allows direct precharge to a voltage close to VDD, which is acceptable for disturb purpose, as VDD is in the 1.2V+/-10% range. A description of the proposed sense amplifier is given in part 2 of the paper. Then a description of the sequencing architecture is provided in part 3. Part 4 is dedicated to simulation results, while the last section of the paper is dedicated to conclusions.

2. Description of the sense amplifier

2.1. Basic operation

The core architecture of the sense amplifier is depicted in Fig. 3. It is composed of a basic latch circuit (N1 and N2 transistors), and the circuitry used to bias the latch before the latch operation is activated (P1, P2, P4, P5 devices, R1 and R2 resistors, switch and dummy capacitance).

The switch is ON during the precharge and latch biasing steps. A current can flow through the R1 and R2 resistances of equivalent value (R1=R2=R) and the switch. At the end of the precharge operation, the nodes BL and CL (connected to the dummy capacitance) are set to their precharge voltage.

This precharge voltage on node BL is equal to V_{DD}-R_P.Ibias2, where Ibias2 is the current flowing through the P2 transistor, and Rp the equivalent resistance of the P5 (P4) transistor.

On node CL, the precharge voltage is equal to V_{DD}-R_P.Ibias1, where Ibias1 is the current flowing through the P1 transistor. Both voltages can be made very close to VDD by correct sizing of the devices.

Due to the current imbalance (Ibias1≠Ibias2) in the structure, a current I_{Rinit} is flowing through the R1 and R2 resistances and the switch. This current sets the DC biasing conditions of the latch at the end of the precharge operation.

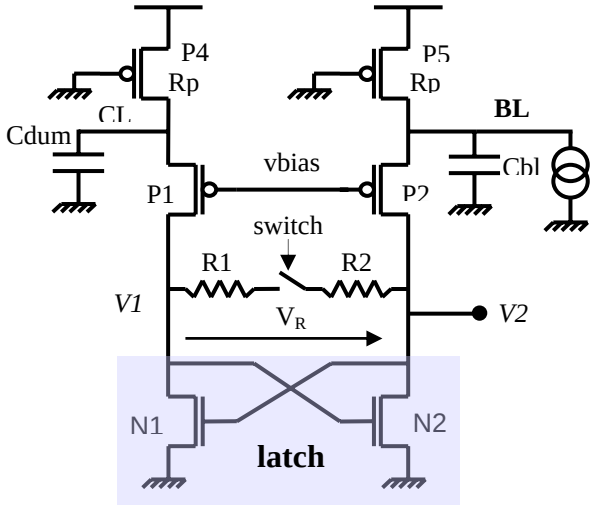


Fig. 1: Illustration of the proposed sensing circuitry (latch + latch biasing circuit).

The DC biasing condition is given by: V_{Rinit}=V2-V1=(2R+R_{switch}).I_{Rinit}.
The imbalance current in the structure is obtained by correct sizing of the devices. For example, the drive of the transistor P2 can be tuned to be larger than the drive of the transistor P1 ($W_{P2}>W_{P1}$, $L_{P2}=L_{P1}$). In this case, a positive differential DC voltage is obtained at the inputs of the latch, V1 and V2.
Now, let's consider the memory cell current flowing in the BL. This current (Icell) generates a voltage variation at the BL node that can be explained as:

$$\Delta V_{BL} = -\left(\frac{R_P}{R_P \cdot gm_{P2} + 1}\right) \cdot Icell \qquad [1]$$

R_p is the equivalent resistance of the transistor P5 biased in linear mode, and gm_{p2} the transconductance of the transistor P2 biased in saturation mode.
The CL node is stable to its precharge value.
The voltage variation on the BL generates an amplified variation at the inputs of the latch thanks to the biasing circuitry. The variation of the differential voltage V_R due to the cell current can be expressed as:

$$\Delta V_R \approx -\frac{gm_{P2}}{\left(gm_{P2} + G_P\right) \cdot \left(\frac{G \cdot gm_{N2}}{gm_{N1}} - gm_{N2} + G\right)} \cdot Icell$$

$$= f \cdot Icell$$

[2]

G_P=1/R_P, gm_{N1} and gm_{N2} are the transconductances of N1 and N2 devices respectively. G that can be expressed as:

$$G = \frac{1}{2R + R_{switch}} \approx \frac{1}{2R} \qquad [3]$$

R=R1=R2, and R_{switch} is the equivalent resistance of the switch. It can be made about negligible compared to R.
From these equations, we can extract the following condition to achieve a correct functionality:

$$G \geq \frac{gm_{N1} \cdot gm_{N2}}{gm_{N1} + gm_{N2}} \qquad [4]$$

This expression is used for the correct sizing of the resistance R. At the end of the step 2, the DC biasing conditions at the input of the latch are:

$$V_R = V_{Rinit} - \Delta V_R = V_{Rinit} - f \cdot Icell \qquad [5]$$

Once the inputs of the latch are correctly biased to the DC V_R differential value, the latch circuitry can be activated. To activate the latch, the switch must be set OFF. Then the latch switches according to its initial DC input conditions given by V_R. Latch switching operation is very fast due to its positive feedback.
Assuming that there is no mismatch in the latch circuitry (N1 and N2 devices perfectly identical) the theoretical condition to get a correct latch switching operation is:

$$|V_R| \geq 0 \qquad [6]$$

$|V_R|$ is the absolute value of the differential voltage V_R.
However, when considering the mismatching on N1 and N2 devices the practical condition is:

$$|V_R| \geq 3 \cdot \sigma_{VTN} \qquad [7]$$

VTN is the threshold voltage of the N1 and N2 devices.
This condition ensures that the latch will switch correctly in the direction imposed by the DC conditions. If V_R is negative ie V2<V1 (cell ON), then V2 will switch to 'L' value while V1 goes to 'H'. If V_R is positive ie V1>V2 (cell OFF), then V2 will switch to 'H' value while V1 goes to 'L'.
Using equation 7, it is possible to derive the conditions to be fulfilled by the memory cell current to be correctly sensed and converted to by the latch. We have:

$$3 \cdot \sigma_{VTN} < V_{Rinit} - f \cdot Icell < -3 \cdot \sigma_{VTN} \qquad [8]$$

Resulting in the following conditions for the current:

Condition 1: $$Icell > \frac{3\sigma_{VTN} + V_{Rinit}}{f} = I_{L1} \qquad [9]$$

Condition 2: $$Icell < \frac{3\sigma_{VTN} - V_{Rinit}}{f} = I_{L2} \qquad [10]$$

If the condition 1 is respected, V2 will switch to 'L' when the latch is activated. If condition 2 is fulfilled, the latch will switch in the opposite direction and V2 will be set 'H'. Now, if the memory cell current is within I_{L2} and I_{L1} limits, the sense circuitry behavior is not guaranteed due to the mismatching of the devices, and the latch output is unknown. So conditions 1 and 2 must be fulfilled to ensure the correct functionally, as illustrated in Figure 2.

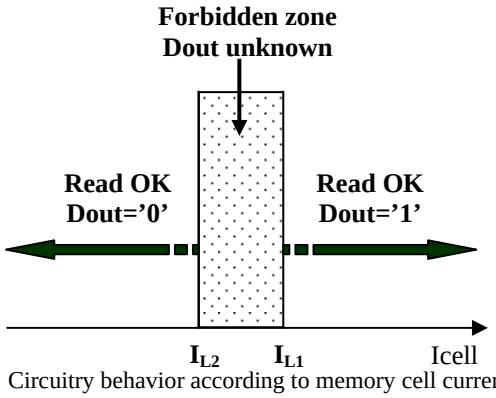


Fig. 2: Circuitry behavior according to memory cell current.

2.2 Practical implementation

A practical implementation of the sensing circuit including the precharge devices is given in Figure 3. In addition to Figure 1, the transistors P6 and P7 have been added. They are controlled by the *prech* signal, and are switched ON during the bit line precharge, OFF once the END of the precharge is detected.

P10, P11 and N10 devices are used to generate the bias voltage on the gate of the transistors P1 and P2. N10 transistor gate is controlled by the output of the inverter 5, which input (*rdn* signal) is low during read. The switch of Figure 1 is implemented using P20 and N20 devices in parallel, and is controlled by the *latch* and *latchn* (inverse of latch) signals. Two identical structures are connected to the output nodes of the sensing latch *V1* and *V2*, in order to guarantee a good matching of the capacitances on these nodes. The voltage on node *V2* is transferred to the *dout* node once the sensing operation has been performed, when the latch switching operation is completed. The output switches on both *V1* and *V2* nodes must be set OFF during the sensing operation. Data transfer to the output is obtained by turning the switch (N22, P22) ON (signal latchd set to '0', latchdn to '1'). This operation is delayed until the latch switching operation has been fully completed.

3. Sequencing architecture

In order to properly use the sense amplifier structure described in part 2, correct sequencing of the different operations must be done using a dedicated circuitry. Because the memory interface is asynchronous, address transition detection is used to internally synchronize all the sequences starting from the address transition.
The detection of the address transition starts the read operation. The read sequence can be divided in 5 major steps that are internally timed:

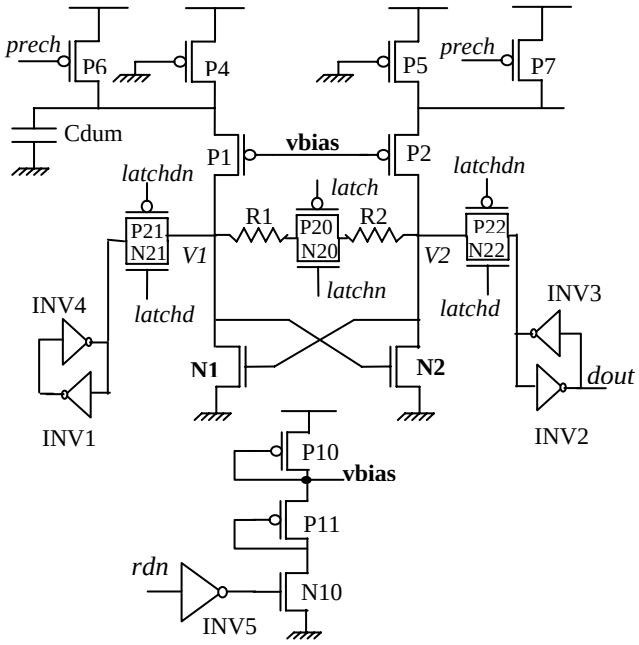


Fig. 3: Practical implementation of the sense circuitry including precharge and dout data path.

1) Initialization of the dummy word line and bit line: dummy bit line and word line are discharged.
2) Precharge, memory cell activation: the bit line and word lines are decoded and precharged, in parallel to the dummy word line and bit line. The end of the precharge operation is detected thanks to the detection circuitry connected at the output of the dummy bit line and word line. This detection ends the precharge operation. At this time, the bit line precharge circuitry is turned OFF.
3) Current to voltage conversion on the bit line, biasing of the latch circuitry: once the precharge circuitry is OFF, the memory cell is correctly biased for read. The cell will drive a current according to its programmed state. The bit line voltage variation is proportional to the cell current. This is a current to voltage conversion on the bit line. This voltage variation on the bit line is amplified by the sense biasing circuitry in order to properly bias the latch circuitry before activation. Then, the sense circuitry is ready for the latch operation. A dedicated circuitry is used to generate the delay necessary to perform this operation.
4) Latch activation and switching: at the end of the previous step, the latch is properly biased, and can be activated. Depending on its initial biasing conditions, the latch will switch in one direction or the other. This operation is irreversible. If the cell current is high enough, the voltage variation on the bit line is amplified in such a way that the resulting DC biasing on the latch makes it to switch in one direction. If the current is lower than a given limit, the structure will impose biasing conditions resulting in the latch switching in the opposite direction.
5) Sense to data out data path: this is the delay generated by the output data path.

The access time can be derived from:

$$T_{acc} = T_{Init} + T_{prech} + T_{latchBias} + T_{latch} + T_{dout} \qquad [11]$$

Figure 4 illustrates the typical sequencing of the different signals involved in the read process.

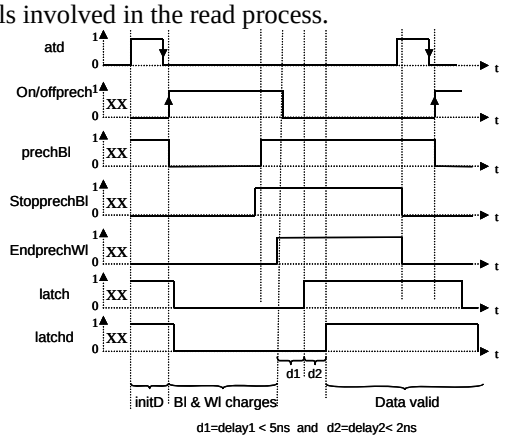


Fig. 4: Sequencing signals chronogram.

When an address transition is detected, atd is set 'H', and goes to 'L' value after an internally controlled delay. Note that atd stays to 'H' value as long as the input address bits are toggling. The pulse on the atd signal is used to discharge the dummy bit line and word line, during the initialization phase. Once the initialization phase has been performed, the precharge operation starts. During the precharge operation, latch signal is set 'L' in order to have the switch of Figure 1 ON. PrechBL and latchd signal are 'L' as well. Once the dummy bit line has reached the desired level, the level detection circuitry sets the signal stopprechBl to 'H' value, and the prechBL signal goes 'H' to switch OFF the precharge device (P7 in Figure 3). At this time, the bit line is set to the correct value for read (closed to VDD). Now, the structure has to guarantee that the word line has reached the correct voltage level

necessary to bias the memory cell in the right conditions. The word line level detection circuit drives the EndprechWl signal to 'H' once the dummy word line has reached the correct level, which means that the previous condition is realized. Once the bit line and word line precharge operations are completed (On/offprech='L'), the sensing operation can start. A first internally generated delay d1 is used to bias the latch circuit, according to the BL voltage variation due to the cell current. Assuming constant DC initial conditions on the latch, delay d1 is directly proportional to the BL capacitance. Once the latch biasing is completed, latch signal goes 'H', and the latch operation is initiated. This operation is very fast, and must be completed during the d2 delay. At the end of the sensing operation, latchd signal goes 'H'. A detailed sequencing circuitry is provided in Figure 5. This circuitry is used to generate the sequencing signals of Figure 4.

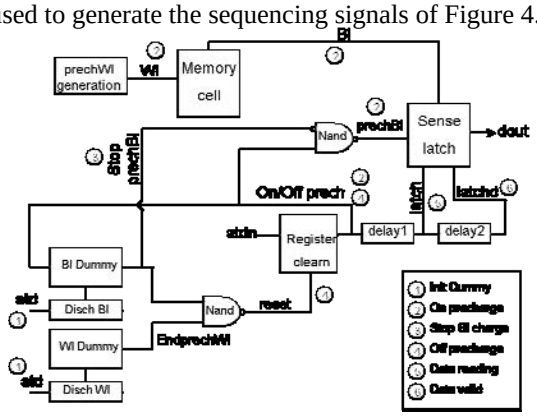


Fig. 5: Sequencing circuitry.

4. Simulation results

Simulations were performed on a 8Mb embedded flash memory, implemented on a 0.13μm embedded flash process. The sectoring of the memory was done as following: 256 bytes per page, 1024 pages per global bit line, and 128 pages per local bit line. The global/local bit line scheme allows bit line to bit line coupling capacitance minimization. The simulated sense current trip point, and random access time are reported on Table 1. The intrinsic current trip point is derived from a DC simulation, and is equal to the cell current resulting in a voltage V_R=0V. As shown, the trip point is stable with the temperature, and increases with VDD. Transient simulations using the more critical pattern were performed to simulate the access time. Assuming that σ_{vtnmos}=2.5mV, and $(W.L)_{N1,N2}=4\mu m^2$ (devices of Figure 3), a $|V_R|$ (see equation 7) of 20mV was chosen to avoid any mismatching issues with the latch. Figure 6 illustrates the analog behavior of the bit line voltage, and latch input nodes during the read operation at 1.2V. The bit line voltage rises up to 1.1V during precharge. During the latch DC biasing step, the bit line voltage decreases, V2 decreases and V1 increases. Once the condition V1-V2≥20mV is achieved, the latch is activated, and switches very fast thanks to the positive feedback.

A typical access time of 16ns was obtained using the proposed memory sectoring, and $|V_R|$=20mV, which is above the $3\sigma_{vtn}$ (7.5mV) condition of equation 7. Further access time improvement could be possible by speeding up BL and WL precharge (impact on area), or by relaxing the condition on VR (impact on reliability).

	T=-40°C		**T=27°C**		**T=85°C**	
V_{DD} (V)	T_{acc} (ns)	I_{tp} (μA)	T_{acc} (ns)	I_{tp} (μA)	T_{acc} (ns)	I_{tp} (μA)
1	18	1.8	20	1.8	22	1.8
1.2	14	2.6	16	2.6	18	2.6
1.4	13	3.6	15	3.6	17	3.6

Table 1: Simulated access time and sense current trip point for different TEMP&V_{DD} conditions.

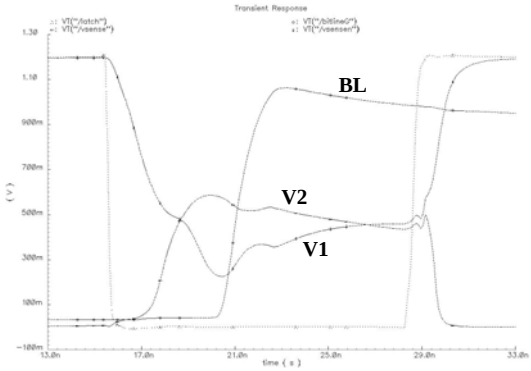


Fig. 6: Simulated waveforms (bit line, V1 and V2 latch input nodes).

5. Conclusion

A new read architecture allowing 1V, high-speed read operation on embedded flash memories has been proposed [10]. The sensing operation uses a high-speed latch structure, which is coupled to a biasing circuitry allowing single ended operation based on the amplification of the bit line voltage variation. In comparison with most of the existing solutions that are asynchronous, the proposed solution needs a sequencing of the read operations starting from the address transition. This is done using a synchronization circuitry, making the memory locally synchronous for read. Simulation results are very promising and show that 16ns read operation can be achieved in typical read conditions on large blocks.

References

[1] "World Microcontrollers Market", Frost & Sullivan, 2005

[2] M. Combe and JM Daga, "Design of high-speed 128-bit embedded flash memories allowing in place executionof the code", Solid-State Electronics 49, 1867-1874, 2005

[3] Takeshi Nakayama et al, "A 60ns 16Mb Flash EEPROM with program and erase sequence controller", IEEE J. Solid-State Circuits, vol. 26, no. 11, 1600-1605, 1991

[4] P.Cappelletti, "Flash Memories", Kluwer Academic Publishers, ISBN 0-7923-8487-3, 1999

[5] Donald Yuen Yu et all, "High-speed sense amplifier having variable current level trip point", US patent no. 5666310, 1997

[6] JM Daga et all, "Single-ended current sense amplifier", US patent no. 6608787, 2003

[7] N. Otsuka et all, "Circuits techniques for 1.5V power supply flash memory", IEEE J. Solid-State Circuits, vol. 32, no. 8, 1217-1230, 1997

[8] T. Seki, et al., "A 6-ns 1-Mb CMOS SRAM with Latched Sense Amplifier" IEEE J. of Solid-State Circuits, vol. 28, no. 4, 1993

[9] T. P. Haraszti, "High Performance CMOS Sense Amplifier," US patent no. 4169233, Sep. 1979

[10] J. Fort, filed patent number US11/463,391 on 8/9/2006

A low voltage, low power, highly reliable, multi-purpose, cost-competitive embedded non-volatile memory in 90nm node

Guoqiao Tao[a], Erik van der Vegt[b], Jean-Pierre Carrère[c], Florence Larman[b], Dick Boter[a], and Do Dormans[a]

[a]NXP semiconductors, Gerstweg 2, 6534 AE Nijmegen, The Netherlands

[b]NXP semiconductors, 850 rue Jean Monnet, 38921 Crolles cedex, France

[c]STMicroelectronics, 850 rue Jean Monnet, 38921 Crolles cedex, France

Guoqiao.Tao@NXP.COM

Abstract

In answer to the needs in various application areas, a state-of-the-art embedded NVM technology in the 90nm node has been developed in a 300mm fab and is presented in this paper. By utilizing the proven 2T-FNFN-NOR device concept, a cost-competitive, low voltage, low power technology is realized with very good P/E endurance and data retention.

1. Introduction

In many application areas, such as: chip-cards, automotive electronics, and a variety of hand-held applications, embedded Non-Volatile Memories (NVM) are required, together with special application requirements like low voltage low power (for contact-less or hand-held applications), high endurance (for EEPROM data memory), wide temperature range (for automotive applications), etc.[1-3]. In answering to the market needs and to stay cost competitive, NXP has developed a low voltage, low power, highly reliable, multi-purpose and cost competitive embedded non-volatile memory technology option on the 90nm CMOS platform in the Crolles2 alliance 300mm fab. This paper reports the process and device architecture, the challenges encountered, and the reliability results.

2. Device architecture

The 2T-FNFN-NOR device architecture [1,2] is applied. This architecture has been proven to give competitive compact module size for embedded applications [4]. The two transistor cell with an access gate at the source side in an NOR array, enables the low voltage low power operation and the high reliable uniform FN tunnelling both for programming and erasure. A schematic and TEM cross section of such a cell is shown in figure 1 and 2 respectively.

To satisfy the specific requirements in various applications, different array structures are developed by varying the floating gate size while keeping a fixed size in the channel length direction. Small floating gate are implemented for high-density flash applications, while wide floating gates are applied for fast programming/erasure full-featured EEPROM applications in addition to by-segmentation [1,2].

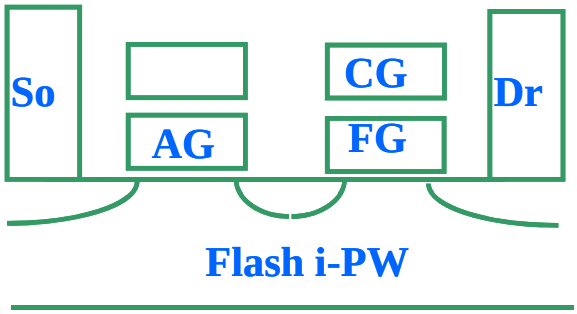


Fig. 1: A schematic cross section of the two-transistor cell. Control gate (CG) and floating gate (FG) are on the right side near the drain (Dr) while the access gate (AG) is on the left side near the source (So).

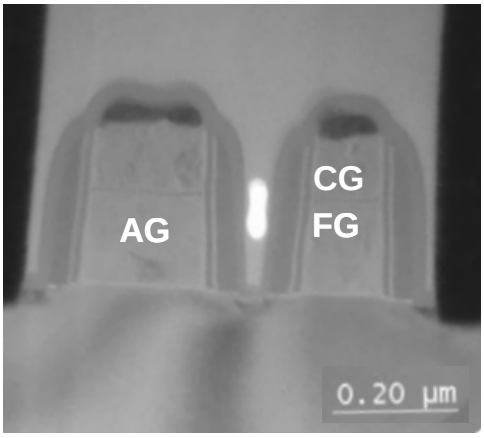


Fig. 2: TEM cross-section of the two-transistor cell in 90nm node.

3. Process architecture and features

The process architecture is similar to that in earlier process generations [1]. The embedded NVM process steps are modularised with low temperature budget to minimize the dopant diffusion. The key features of the NVM part are: a 8.5nm ISSG (In-Situ Steam Generation) RTO tunnel oxide [5] with post oxidation nitridation,

and a 6/5/5 nm ONO stack (13.5nm oxide equivalent) as inter-poly-dielectric. DUV resist with hard-mask is employed in flash cell patterning.

• Field isolation (STI)
• Well formation (logic, triple, high-voltage)
• Dual logic oxides and gate deposition
• Tunnel oxidation and floating gate formation
• ONO formation
• High-voltage oxidation and control-gate formation
• Gate patterning and S/D implantations
• 6- or 7-layer copper / low-k back-end

Table 1: Process architecture.

4. Results

Figure 3 shows the programming and erasure speed of a typical flash cell by FN tunnelling at 14.5V. With 1ms programming and 100ms erasure, a nominal Vt window of nearly 4V is achieved.

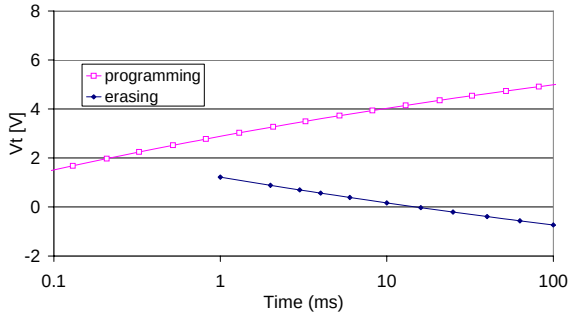


Fig. 3 Programming and erasure speed of a typical flash cell by FN tunneling at 14.5V.

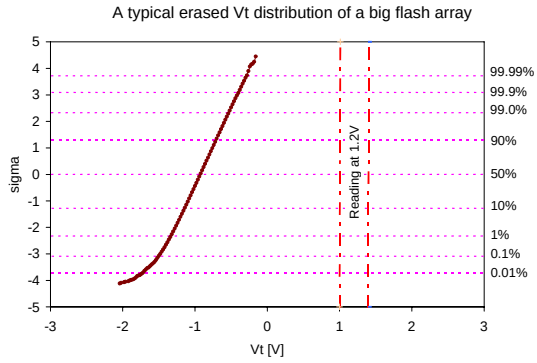


Fig. 4: Cumulative distribution of cell Vt's at erased state, with a single shot of 100ms –14.5V pulse, showing enough margin to the read condition even for the worst case cell.

The capability of low-voltage read (at Vdd=1.2V) is ensured by a good margin between the erased Vt and the nominal read condition (1.2V on control gate). Figure 4 shows the erased Vt distribution of the flash array, indicating a margin of more than 1V for the worst-case cell.

The flash array has a very good program/erasure characteristic, thanks to the uniform FN tunnelling mechanisms. Figure 5 shows the endurance curve of the nominal flash cell with single pulse programming/erasure. With an initial Vt window of 4V, more than 1 million cycles is achieved without significant window closure. Further analysis of the endurance characteristics reveals that the evolution of the Vt values follows a -root law, as shown in figure 6. This indicates that the window closure is very slow, and well controlled [3].

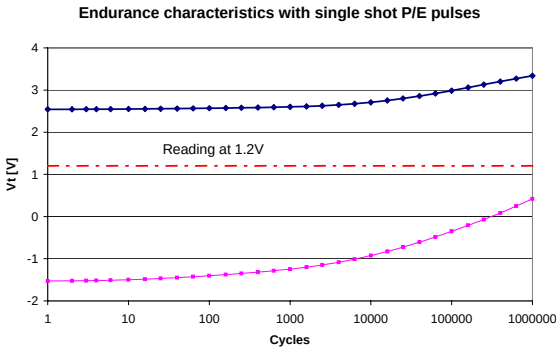


Fig. 5: Endurance characteristics of the flash cell.

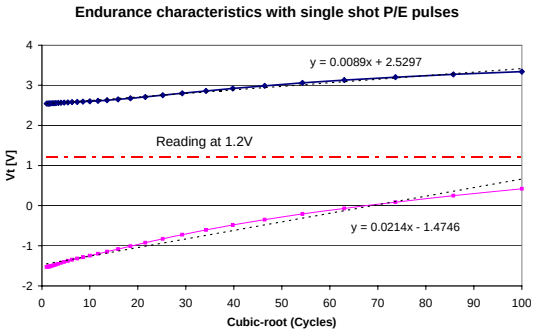


Fig. 6: Data of figure 5, but plotted against cubic-root of cycles on X-as.

The low-temperature data retention (LTDR), especially after extensive P/E cycles, caused by stress-induced leakage current (SILC), is addressed by accelerated gate stress measurements on large Flash arrays. The results are plotted in figure 7 for a flash die after 100,000 cycles. These results are comparable to the performance of mature production processes in previous generations [6,7].

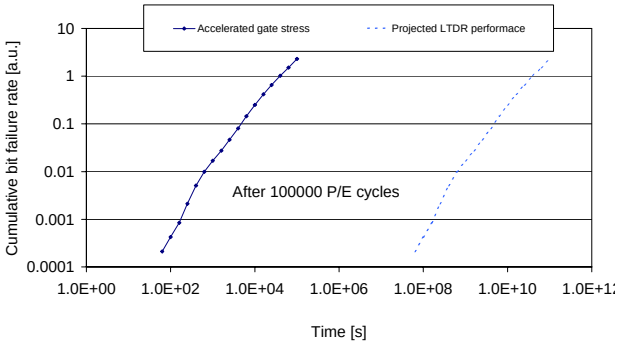


Fig. 7: Accelerated gate stress measurements and projected low-temperature data retention performance for flash arrays after 100000 P/E cycles.

5. Summary

A state-of-the-art embedded NVM technology in the 90nm node has been presented. By utilizing the proven 2T-FNFN-NOR device concept, a cost-competitive, low voltage, low power technology is realized with very good P/E endurance and data retention. This technology is suitable for multi-application areas.

References

[1] D. Dormans et al., "Low-Voltage Embedded Flash-EEPROM in 0.18μm CMOS", in SSDM2001. pp 540-541.

[2] D. Dormans et al., "Cost-competitive embedded non-volatile technology for Flash and EEPROM applications", in ICMTD 2005. pp 161-162.

[3] G. Tao, et al., "Experimental Study of Temperature Dependence of Program/Erase Endurance of Embedded Flash Memories with 2T-FNFN Device Architecture". IEEE International Reliability Workshop 2006, pp 76-79.

[4] T. Ditewig et al., ISSCC 2001, p. 2.4.

[5]. C. Dijkstra et al, "ISSG RTO (In-Situ Steam Generation Rapid Thermal Oxidation) grown Tunnel Oxide for Improved Reliability of Flash Memories", in NVSMW 2003, pp. 85-86.

[6] G. Tao, et. al., "Device architecture and reliability aspects of a novel 1.22μm2 EEPROM cell in 0.18μm node", In INFOS 2003 Insulating Films on Semiconductors, paper WS3-4..

[7] G. Tao, et. al., "Reliability aspects of advanced embedded floating-gate non-volatile memories with uniform channel FN tunneling for both program and erase"; IEEE Non-volatile semiconductor memory workshop 2001.

DATA RETENTION RELIABILITY OF P+ POLY FLOATING GATE MEMORIES IN LOGIC CMOS PROCESSES

Yanjun Ma, Rui Deng, Bin Wang, Andy Horch, and Ron Paulsen
Impinj Inc.
701 N. 34th St; Seattle, WA 98103
206-834-1054; fax: 206-517-5262; e-mail: yanjunma@impinj.com

ABSTRACT

Long term (up to two years) low temperature bake data from p+ floating gate memories with ~70Å tunnel oxide manufactured in logic CMOS processes are presented for the first time. The tail bit behaviors, including bake time, cycling, and temperature dependences, are shown to be similar to those of n+ floating gate reported in the literature. We then apply the models developed for n+ floating gates to show that p+ floating gates have an inherent advantage in terms of data retention. The data retention advantage enables the scaling of tunnel oxides to below 80 Å for p+ poly based floating gate memories for small bit count embedded applications.

1. INTRODUCTION

While the first floating gate NVM was pFET based[1], nFET based floating gate MOSFETs have been the overwhelming choice in commercial EEPROM and flash products. It is worth noting that the tunnel oxide thickness for the nFET based devices has been limited to > 80 Å due to stress induced leakage current (SILC). Recently, there has been a resurgence in interest in pFET based floating gate memories, especially for embedded applications[2-8]. For example, we previously reported the intrinsic data retention reliability of a pFET based EEPROM in logic CMOS process targeted at small bit count applications [6-8]. Chung, et al. discussed the performance of a p+ gate nFET EEPROM [3] and reviewed the advantage of p-channel n+ floating gate memory in the area of programming speed and write disturb[4]. Yet there have been no studies on the impact of tail bits on data retention reliability of p+ floating gate pFET based NVM. In this work we report long term bake results of such memories manufactured in 0.18μm and 0.25μm standard logic processes. The results are compared to those observed for n+ floating gate memories and then analyzed using a tunneling model developed for nFET flash arrays. The data retention advantage of p+ floating gate memories is then discussed.

2. TEST VEHICLE

Our pFET based NVM performance and intrinsic reliability have been described previously.[6-8] Figure 1 shows a pMOS-based memory cell. The cell differs from typical NVM cells in that it is differential. The difference in floating-gate charge on nodes Fg_0 and Fg_1 is reflected as a difference in channel currents in the read transistors M_1 and M_2, which is resolved during readout using a current-sense amplifier. We use Fowler-Nordheim (FN) tunneling to erase both sides of the cell, and impact-ionized hot-electron injection (IHEI)[6-7] or FN tunneling[8] to write the appropriate data state. A differential cell is advantageous since with proper design it has built-in redundancy – even if the floating gate on one side leaks, the cell should maintain its logical state if there other side does not leak. Essentially the failure probability of a differential cell is the product of the failure probability of the floating gates Fg_0 and Fg_1. This greatly reduces the susceptibility of a differential cell against the negative impact of localized defects, such as traps, that causes stress induced leakage current (SILC).

The memory test chips were fabricated at a leading foundry using its standard single poly 0.18 and 0.25μm logic CMOS processes. The nominal gate oxide thickness of the 3.3V I/O transistor used in the memory cell range from 62 to 65Å. A number of test chips were used for this paper with memories of 256b to 16Kb in size.

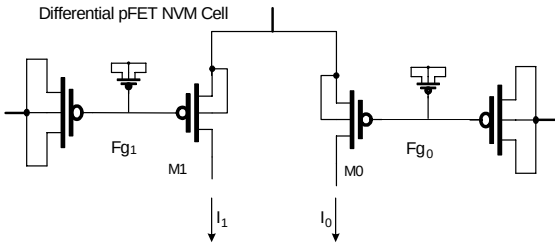


Fig 1. pMOS based memory cell stores the logic state as the difference in charge between Fg_0 and Fg_1.

3. RESULTS

High temperature bake data has been reported previously [6-8] to show that the intrinsic data retention is excellent. Here we only present results on low temperature data retention bakes to focus on the behavior of cycling induced tail bits. We use the storage window defined by the cell current difference between the two sides of the differential cell as the measure of data retention reliability. Figure 2 shows the distribution of the storage window of nearly forty 16Kb chips with 1K cycles as a function of bake time at 85°C. The distribution is virtually identical to those that have been reported in the literature for nFET floating gate NVM [9-13].

More specifically we observed the following:

a. Log time dependence - Figure 3 shows the movement of the fastest leaking bits from a number of 16Kb chips. While a log time dependence is clearly seen over most of the range, it appears that the cell current loss is slowing

down (but did not stop) over time. This deviation from ln(t) dependence for the cell current does not mean that we are observing a different behavior as that observed in Ref. [10-11] where it has been demonstrated that Vt ~ ln(t). The reason is as follows: the cell current Icell ~ $(Vfg-Vt)^2$, which yields ΔIcell ~ (Vfg-Vt) ΔVfg. The cell current change is then expected to slow down as Vfg is approaching Vt. The above relationship can be re-written as $\frac{1}{\sqrt{I_c}}\frac{dI_c}{dt} \propto \frac{dV_{fg}}{dt}$. In Fig. 4, $\frac{1}{\sqrt{I_c}}\frac{dI_c}{dt}$ is plotted as a function of time, it is seen that the data can be fitted very well with the following $\frac{1}{\sqrt{I_c}}\frac{dI_c}{dt} \sim 1/t$. We then obtain, Vfg-Vfg0 ~ ln(t), the same relationship as in Ref. [10-11] that can be derived from a trap to trap direct tunneling model [11].

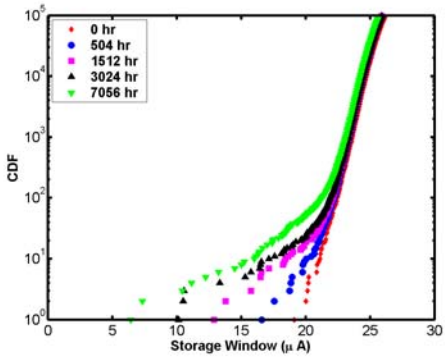


Fig 2. Cumulate distribution of nearly forty 16Kb memories as a function of storage window for different bake times at 85°C. These memories have been cycled 1K times prior to bake.

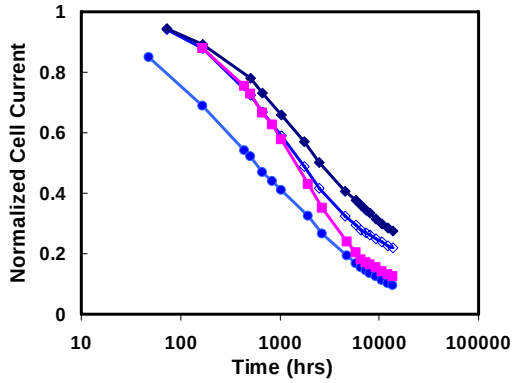


Fig. 3 cell current (normalized to programmed current) of fastest leaking bit from several memories as a function of time at 25°C. The parts have either 10K or 100K cycles.

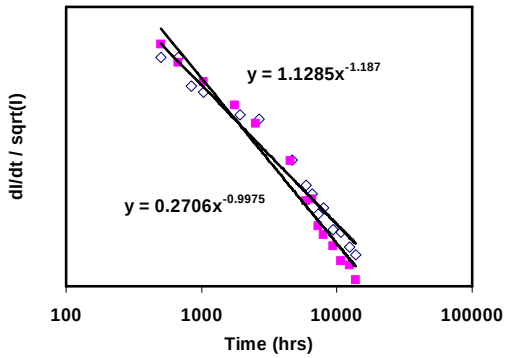


Fig 4. dIc/dt /sqrt(Ic) plot for the tail bits from Fig. 3

b. Temperature dependence – The number of tail bits and the leakage rate is observed to be only slightly dependent on temperature with an activation energy of less than 0.2eV. Figure 5 shows the window distribution of chips after 100K pre-cycling and 4500 hours of bake at 25°C, 85°C, and 150°C. It is clear that higher temperature does not cause more tailbits. In fact, it appears that 25 and 85°C are worse than baking at 150°C as far as tail-bit is concerned, agreeing with previous studies of n+ poly silicon memories.[12,13]

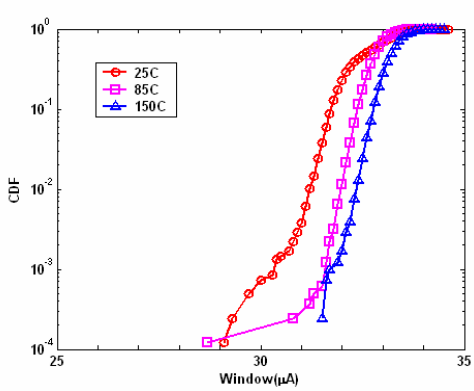


Fig. 5. Distribution of tail-bit for a memory with 62 Å tunnel oxide after baking at 25°C, 85°C, and 150°C. The arrays have been cycled 100K times prior to bake.

c. Cycle dependence – the number of tail bits increases with the number of pre-cycles. Preliminary data indicates a cycle dependence, if fitted to a power law, of index ~ 0.4-0.6.

These aspects (a-c) of the tail bit behavior are all similar to those observed for nFET flash memories [10-15].

d. Annealing of tail bits - Figure 6 shows the movement of some of the fastest tail bits from memories that have undergone 100K pre-cycles. We see that these bits suddenly stopped leaking, indicating annealing or deactivation of the defects causing these tail bits. Minimal annealing is observed at room temperature. Again this is similar to observed low temperature data retention tail bits observed on nFET floating gate NVMs [11,12].

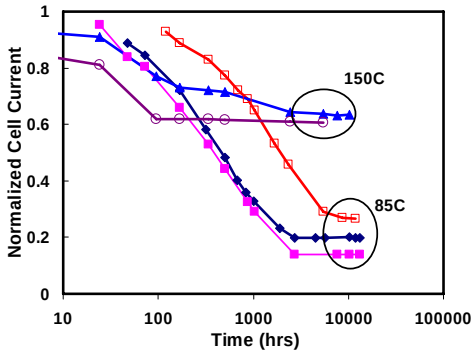


Fig. 6 cell current (normalized to programmed value) of fastest leaking bit from several memories as a function of time at 85°C and 150°C.

4. Retention of N+ vs P+ floating gate memory

Having established that the tail bit behavior of the p+ floating gate is similar to that of n+ floating gate, we now discuss retention characteristics of p+ floating gate using the large volume of literature and models/methods that already existed and developed on n+ poly silicon floating gate based on nFET.

First of all, one can easily estimate that amount of net electrons on a programmed floating gate (~10^5 e/μm^2) is far from enough to fill all the holes that existed on a p+ poly gate (density of 10^{19}-10^{20}/cm^3 or surface density ~10^7 /μm^2 of for 200nm thick poly silicon). As a result, for negatively charged p+ poly floating gate, the charges are excess electrons in the valence band, see Figure 7. Because of the low activation energy observed for the tail bit (point b above), it is unlikely the minority carrier in the conduction band of the p+ floating gate played an important role since that contribution should have an activation energy of around 1.1eV (the Si bandgap). Independent of the leakage mechanism, directly tunneling or trap assisted tunneling, one would then expect that the leakage would be much less for p+ poly-Si than the n+ poly-Si gate due to the increased barrier height. As an illustration, in Figure 8 the FN tunneling current density of PMOS and NMOS are compared. It is seen that PMOS (with p+ poly) have lower tunneling current, especially in the inversion condition. Although this picture does not directly correlate with the retention characteristics of floating gate memories, it does illustrate the role that the barrier height is playing in determining the tunneling current density under the same electrical field and oxide thickness conditions.

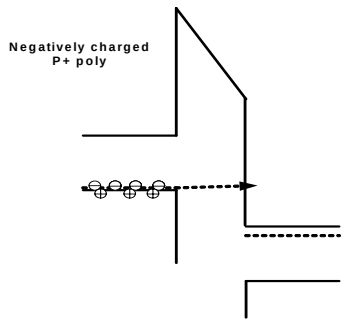


Figure 7. Band diagram illustrating the leakage from negatively charged p+ poly gate.

There are many models of TAT that can be adapted to describe the leakage from p+ poly gate.[15-18] Virtually all the models have a barrier height as an fundamental parameter. Using the model from Ref.[10] as an example, the leakage current is of $I_{leak} = I_0 e^{b_0 V_{ox}}$ form, where I_0 and b_0 have been given in terms of the electron effective mass, barrier height, and other fundamental constants. Using the relations given in Ref. [10], and substituting the barrier height to 4.2eV for the p+ floating gate, we estimated that the leakage current is 1000x less than that of n+ floating gate for a 70 Å tunnel oxide, assuming same effective mass for both cases.

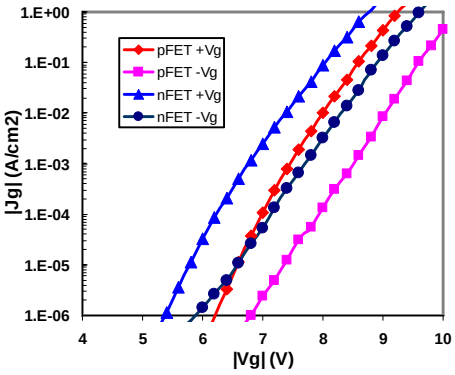


Figure 8 FN tunneling current density comparison of PMOS and NMOS in both accumulation and inversion conditions. The gate poly are doped p+ and n+, respectively.

From experimental point of view, p+ poly has also been shown to have much lower SILC than n+ poly floating gate.[19] This may be attributed to the observation that SILC is mainly due to traps located at about 2.6eV from the top of the valence band in the oxide, thus much closer to that of the silicon conduction band edge.[20] It has also been observed that the SILC for negative gate voltage is lower than for positive Vg for p+ floating gate. The former is the condition for the charge storage in one of the states in our NVM. This point illustrates the importance in choosing the right charge storage level in the floating gate to minimize the effect of SILC.

On the other hand, it is well known that the defect creation in oxide is strongly dependent on the applied voltage.[21] Since, as shown in Ref. [22], it takes about 1V higher voltage for the p+ poly PMOS device than n+ poly PMOS device to have the same tunnel current density, the tunneling from p+ poly floating gate may increase the degradation of the oxide. As such, the endurance performance of a p+ poly floating gate must be carefully monitored.[22] In general, P+ poly is well suited for moderate cycle count embedded NVM applications requiring scalability to thinner oxides.

5. Conclusions

In summary, we demonstrated with long term bake the similarities between the tail bit characteristics of p+ floating gate memories with that of reported behavior for the n+ floating gate. We are then able to use the models developed

for the n+ floating gate to show that p+ floating gate has an inherent advantage in terms of data retention over n+ floating gate. Thus with proper design, one should be able to take advantage of the increased barrier height of the p+ poly silicon floating gate to scale the tunnel oxide thickness to thinner than the ~80 Å limit of n+ based floating gate memories.

REFERENCES

[1] D. Frohman-Bentchkowsky, Appl. Phys. Lett. **18**, 332 (1971).
[2] R. Lin, Y. Wang, and C.C. Hsu, Proc. NVSMW, 27 (1998).
[3] S.S. Chung et al, Proc. VLSI Tech Symp. 19 (1999).
[4] T. Ohnakado and S. Satoh, IEEE Trans. Electron Dev. **47**, 1209 (2000).
[5] S.S. Chung, et al, Proc. IRPS, 67 (2001).
[6] A. Pesavento et al, Proc. NVSMW, 48 (2004).
[7] Y. Ma et al, IEEE Trans Dev. Mater. Rel, **4**, 353 (2004).
[8] Y. Ma et al, Proc. NVSMW 2006, 39 (2006).
[9] P.J. Kuhn, e t al, Proc. IRPS, 266 (2001).
[10] H. P. Belgal, et al, Proc. IRPS, 7 (2002).
[11] F. Schuler, et al, J.J. Appl. Phys. **41**, 1 (2002) and F. Schuler et al, Proc. IRPS, 26 (2002).
[12] G. Tempel et al, NVSMW, page 105 and L. Hwang, et al, page 108 (1999).
[13] G. Tao, et al, Micro. Engineering, **48**, 419 (1999).
[14] D. Ielmini, et al, IEEE Trans. Electron Device, **51**, 1288(2004).
[15] J. De Blauwe, et al, IEEE Trans. Electron Devices, **45**, 1745 (1998).
[16] S. Takagi, et al, IEEE Trans. Electron Devices, **46**, 348 (1999).
[17] D. Ielmini, et al, IEEE Trans. Electron Devices, **47**, 1258 (2000).
[18] R. Degraeve, et al, IEEE Trans. Electron Devices, **51**, 1392 (2002).
[19] V.E. Houtsma, et al, Tech. Digest IEDM, 457 (1999).
[20] J.Wu, L.F. Register, and E. Rosenbaum, Proc. IRPS 389 (1999).
[21] E. Y. Wu, et al, Tech. Digest IEDM, 541 (2000)
[22] B. Wang, et al, to be published in IRPS (2007).

SESSION E

RRAM & DRAM

Magnetic RAM for embedded memory in SoC

S. Ueno, K. Kuroiwa[(a)], T. Tsuji, H. Tanizaki[(b)], M. Shimizu and Y. Inoue

Renesas Technology Corp., 4-1, Mizuhara, Itami-shi, Hyogo, 664-0005, Japan
[(a)] Mitsubishi Electric Corp., 8-1-1, Tsukaguchi-honmachi, Amagasaki-shi, Hyogo, 661-8661, Japan
[(b)] Renesas Design Corp., 4-1, Mizuhara, Itami-shi, Hyogo, 664-0005, Japan

Tel: +81-72-787-2427, E-mail: ueno.shuichi@renesas.com

ABSTRACT

This paper presents MRAM technologies for a use of embedded memory in SoC. Random access operation is one of most important features for an embedded memory. MRAM shows a good performance for a random access operation with low supply voltage, except for reading speed. Therefore, both folded bitline architecture and optimizing methodology of MR/RA are presented for high-speed operation. Moreover, a 4MTJ-1Transistor type MRAM is proposed to reduce chip area without decreasing system performance, briefly.

1. INTRODUCTION

MRAM has been studied as one of universal memories due to its non-volatility, high-speed operation and unlimited writing/reading endurance [1,2,3,4,5,6]. This universality indicates that MRAM is suitable for an embedded memory in SoC, because most of recent SoC chips use several kinds of memories such as DRAM, SRAM, Flash and so on. MRAM is a promising device to get a big advantage by replacing these memories from viewpoints chip area and number of process step.

In this paper, we discuss MRAM as an embedded memory. A key of MRAM-challenge is fast random access time, which is an absolute requirement of embedded memory. We will show two techniques for achieving fast reading operation. We also introduce an approach for high density MRAM, briefly.

2. SAMPLES

Both 1MTJ-1Transitor (1MTJ-1T) and 4MTJ-1Transistor (4MTJ-1T) type 1Mbit-MRAMs are fabricated by using a 0.13μm CMOS technology with 4 level Cu damascene process as shown in Fig. 1. A MTJ is used as a local interconnect between 3rd. and 4th. metals. A strap connects MTJ and 3rd. metal locally. Moreover, a MTJ is patterned by reactive ion etching with hard mask. The size of MTJ is 0.26x0.44μm^2.

A MTJ stack consists of free and pinned layers, which are insulated by AlO_x. The pinned layer is composed of both synthetic-anti-ferromagnetic (SAF) structure and an anti-ferromagnetic layer of PtMn. We prepare a CoFeB as a magnetic material for free layer.

3. EMBEDDED MEMORY IN SOC

Most of SoCs use SRAMs and DRAMs as its embedded memories. These memories enable to be accessed randomly. Recently, Flash is often used due to its non-volatility. However, Flash is not a random access memory (RAM). Therefore, system architecture is very complicate. Then non-volatile RAM is demanded.

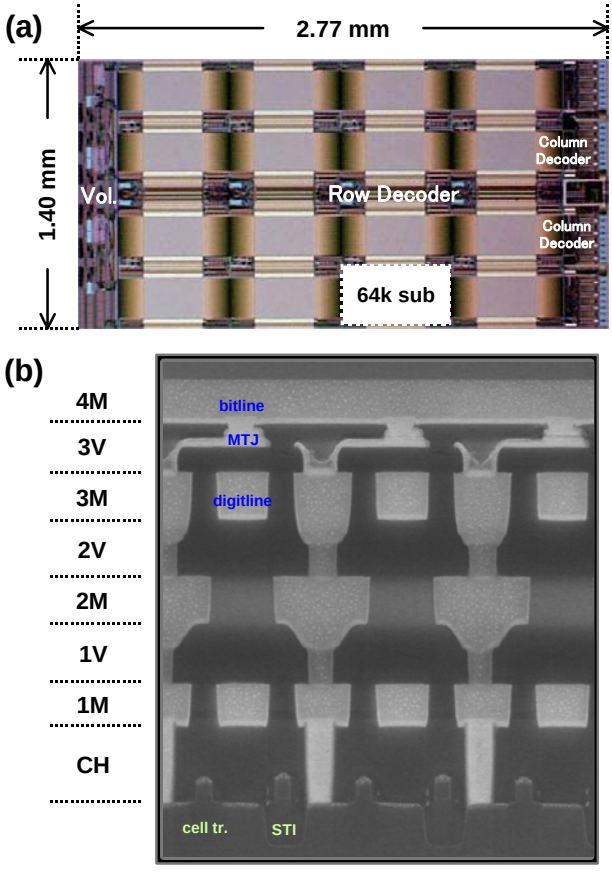


Fig. 1. Photograph of 1Mbit MRAM core (a) and SEM view of memory cell (b)

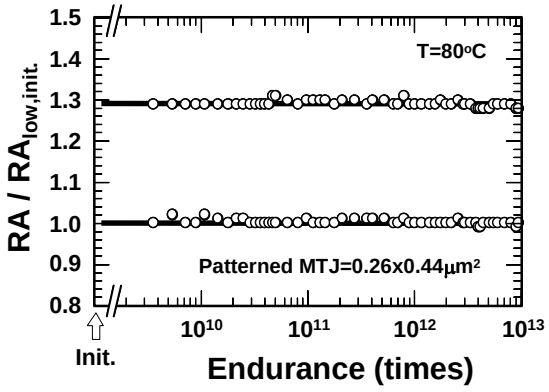


Fig. 2. High endurance characteristic of MRAM

We propose three features of RAM from a viewpoint of system architecture.

(1) No separate concepts for writing and erasing operation
(2) No limitation for reading/writing endurance
(3) Same number of clocks is used for reading and writing operation.

MRAM is easy to achieve request (1) because data0 and 1 can be written by changing direction of current flow at one write line. Request (2) is also easy to achieve as shown in Fig. 2, because write0/write1 cycling is completed by only changing spin polarization. Request (3) means that operation speed is decided by slower time between reading and writing operations.

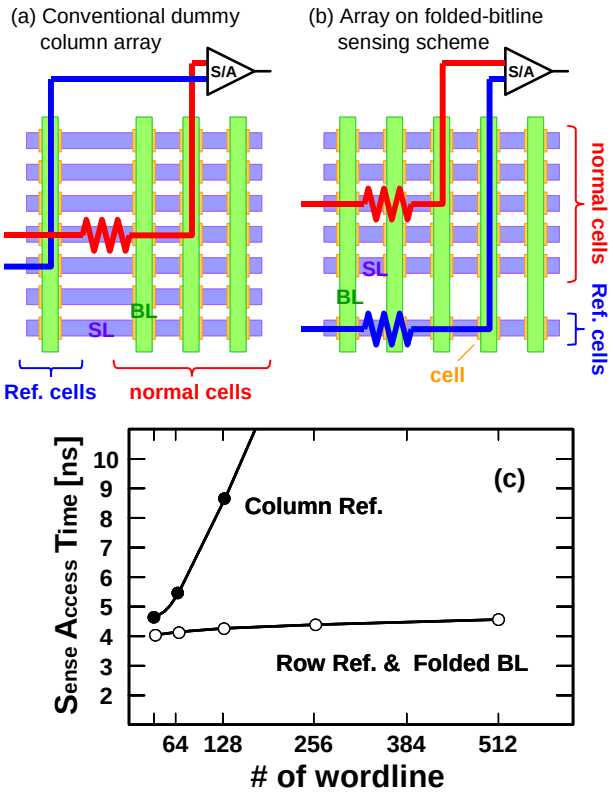


Fig. 3. Comparison between dummy column array (a), dummy row array with folded bitline sensing scheme and simulation result of comparing sensing time of two arrays (c)

Writing speed of MRAM is very fast, which can be observed below 3ns in our experiment. Therefore, reading time limits operation speed and needs to improve. We will describe two techniques to improve reading speed in the next session.

4. High Speed Reading Operation

A. *Folded bitline with dummy row architecture*

Generally, reference cells are arranged at one dummy bitline. This design has a large difference of resistance between two input current paths of sense amplifier as shown in blue and red lines of Fig.3 (a). This is because two paths have different length of CoSi n^+ sourceline, which has a larger resistance than metal wire. Folded bitline with dummy row architecture does not have such a problem, because the length of sourceline is almost same as shown in Fig.3 (b). Therefore, sensing speed does not depend on bit location on cell array as shown in Fig.3 (c). This means that a margin of reading speed can be set very small and fast reading speed can be achieved.

To realizing folded bitline with dummy row architecture, our cell array is designed as shown in Fig.4 [1]. MTJs are arrayed each cross point of bitline and digitline. Straps enlarge a different direction for each bitline. According to this alternate manner, a bitline does not connect to sourceline when a neighbor bitline is used for reading, because these bitlines don't share the same wordline. Therefore, a neighbor bitline can use for reference cell.

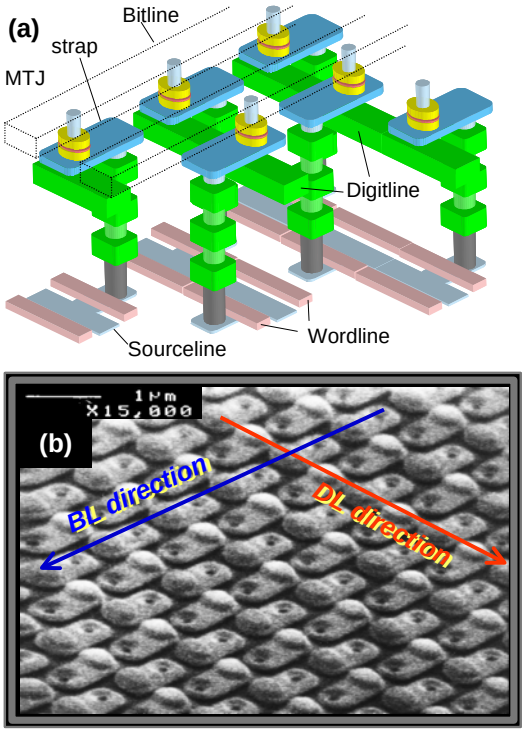


Fig. 4. Schematic image (a) and SEM image (b) of cell array on folded bitline with dummy row sensing scheme

B. *Optimization of MTJ*

Differential sensing scheme, generally used in memories, needs to large difference between two input-signals (currents) for fast reading operation speed. Therefore, increasing MR ratio and decreasing RA are two main methods for realizing fast reading time by MTJ optimization.

Recently, MR becomes very high by many developments of insulating materials and magnetic materials of free and pinned layers [7,8,9,10]. These are favor for high-speed reading and MR improvement will continue to be progressed. In this session, we will discuss an optimizing methodology after getting high MR by using a case of AlO_x insulator.

Low RA can be achieved by decreasing thickness of insulator. However, we should care for decreasing MR when reducing RA. Reported relationships between MR and RA in the study of MRAM [9,11,12] and magnetic Heads [13,14,15] are plotted in Fig.6. At region B in Fig.6, MR doesn't have dependence of RA. Therefore, reducing RA is directly increases reading speed. However, MR has strong dependence on RA and decreases with reducing RA in the region A. Reading speed may be decreased because small MR gives a seriously impact on the current, even if the RA is decreased. This means there exist optimum condition for reading speed in the region A.

To find the fastest reading condition, MR / RA relation in Fig.6 puts on the sensing time chart in Fig.5. In Fig.7, black lines are contour map to maintain same sensing time and dashed lines are MR / RA relation of MTJ. Red, blue and gray means (free layer / pinned layer) structure of (CoFeB /CoFeB), (CoFeB /CoFe) and (CoFe /CoFe) with AlO_x insulator respectively. Green line shows the region A. It should be noted that red and blue dashed lines touch black line at sensing time of 5.2ns as shown in star mark of Fig.7. Here is the optimum point. In our case, sensing time of 5.2ns is possible to achieve at MR of 55% and RA of 2kΩ μm^2. Moreover, optimum point indicates us MTJ stuck. Free layer should be used CoFeB (Red or Blue dashed lines in Fig. 7). Pinned layer can be chosen either CoFeB or CoFe to realize fastest operation. Furthermore, optimum condition exists not in region B where MR is the maximum for RA variation, but in region A where MR losses by decreasing RA.

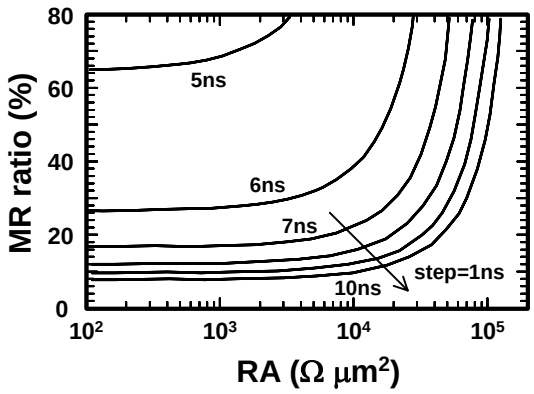


Fig. 5. MR and RA relationship to get same sensing time

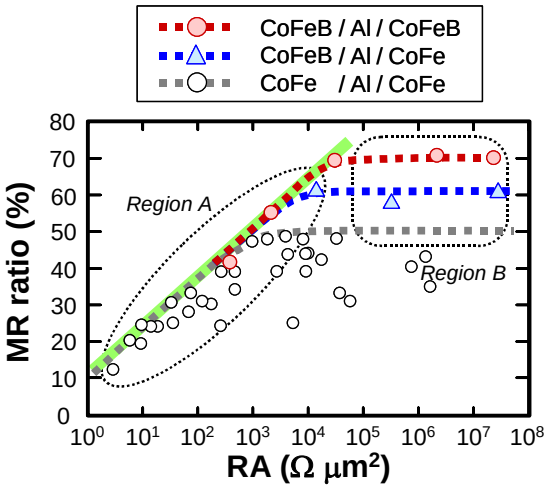


Fig. 6. Reported relationships between MR and RA

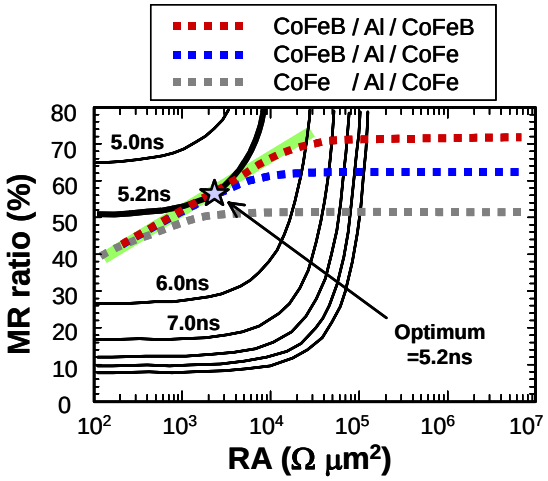


Fig. 7. Optimization of MR and RA

We can also observe an interesting result in Fig.7. MTJ with CoFe as pinned layer doesn't have higher MR than that with CoFeB because of material mismatch between free and pinned layers. However, sensing time is the same because MR at optimizing RA is same. This means that not only high MR but also small lowering with decrease of RA are important when developing materials to achieve high MR.

5. HIGH DENSITY MRAM

To realize a small cell size, a cross-point type MRAM has been proposed, which has cell size of 4F^2 (* F=metal layer) [16]. However, it is found that a cross talk problem is very serious influence on reading speed because a large sneak current hides the main signal.

We propose a 4MTJ-1T cell for another shrinking way of

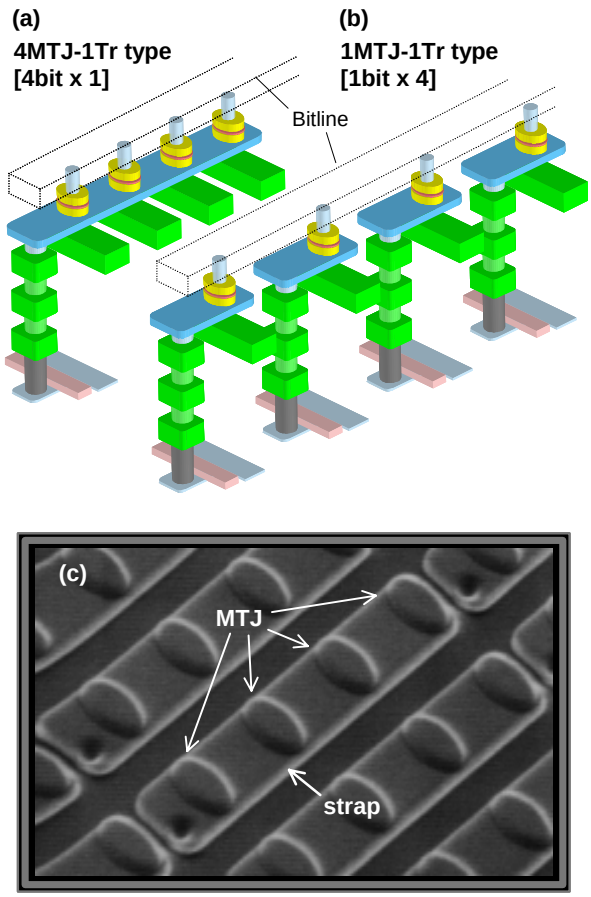


Fig. 8. Schematic image (a,b) and SEM image of 4MTJ-1Tr. cell structure

cell size [17]. 4 MTJ elements are arrayed on one strap and connected same bitline as shown in Fig.8. It can reduce 44% area comparing with that of a 1MTJ-1T cell. However, speed penalty is generated because self-reference sensing scheme is necessary to identify one bit among 4MTJs. Therefore, we propose an on-chip hierarchical memory scheme composed of fast 1MTJ-1T cell for cache memory and small 4MTJ-1T cell for large-capacity memory to reducing total chip size with keeping system operation speed high.

6. CONCLUSION

MRAM is suitable for embedded memory in SoC, because fast random access operation can be achieved with both folded bitline architecture and optimizing MR/RA for MTJ. We also proposed on-chip hierarchical memory scheme composed of 1MTJ-1T and 4MTJ-1T cells for decreasing area of chip without reducing system operation speed.

REFERENCES

[1] T.Tsuji, H.Tanizaki, M.Ishikawa, J.Otani, Y.Yamaguchi, S.Ueno, T.Oishi and H.Hidaka, "A 1.2V 1Mbit Embedded MRAM Core with Folded Bit-line Array Architecture", *in 2004 Sympo. on VLSI Circ.*, p.450, 2004

[2] S.Ueno, T.Eimori, T.Kuroiwa, H.Furuta, J.Tsuchimoto, S.Maejima, S.Iida, H.Ohshita, S.Hasegawa, S.Hirano, T.Yamaguchi, H.Kurisu, A.Yutani, N.Hashikawa, H.Maeda, Y.Ogawa, K.Kawabata, Y.Okumura, T.Tsuji, J.Ohtani, T.Tanizaki, Y.Yamaguchi, T.Ohishi, H.Hidaka, T.Takenaga, S.Beysen, H.Kobayashi, T.Oomori, T.Koga and Y.Ohji, "A 0.13mm MRAM with 0.26x0.44μm^2 MTJ optimized on universal MR-RA relation for 1.2V high-speed operation beyond 143MHz", *in Ext. abst. of IEDM 2004*, p.579, 2004

[3] R.W.Dave, G.Steiner, J.M.Slaughter, J.J.Sun, B.Craigo, S.Pietambaram, K.Smith, G.Grynkewich, M.DeHerrera, J.Akerman, and S.Tehrani, "MgO-Based Tunnel Junction Material for High-Speed Toggle Magnetic Random Access Memory", *IEEE Trans. on Magn.*, 42, p.1935, 2006

[4] H.Yoda, T.Kai, T.Inaba, Y.Iwata, N.Shimomura, S.Ikegawa, K.Tsuchida, Y.Asao, T.Kishi, T.Ueda, S.Takahashi, M.Nagamine, T.Kajiyama, M.Yoshikawa, M.Amano, T.Nagase, K.Hosotani, M.Nakayama, Y.Shimizu, H.Aikawa, K.Nishiyama, E.Kitagawa, R.Takizawa, Y.Ueda, M.Iwayama and K.Itagaki, "1.8 V Power Supply 16 Mb-MRAMs With 42.3% Array Efficiency", *IEEE trans. on Magn.*, 42, p.2724, 2006

[5] N.Sakimura, T.Sugibayashi, T.Honda, H.Honjo, S.Saito, T.Suzuki, N.Ishiwata, and S.Tahara, "MRAM Cell Technology for Over 500MHz SoC", *in 2006 Sympo. on VLSI Circ.*, p.108, 2006

[6] M. Motoyoshi, I. Yamamura, W. Ohtsuka, M. Shouji, H. Yamagishi, M. Nakamura, H. Yamada, K. Tai, T. Kikutani, T. Sagara, K. Moriyama, H. Mori, C. Fukamoto, M. Watanabe, R. Hachino, H. Kano, K. Bessho, H. Narisawa, M. Hosomi and N. Okazaki, "A study for 0.18 μm high-density MRAM", *in 2004 Sympo. on VLSI tech.*, p.450, 2004

[7] S.S.P.Parkin, C.Kaiser, A.Panchula, P.M.Rice, B.Hughes, M.Samant and S.H.Yang, "Giant tunneling magnetoresistance at room temperature with MgO(100) tunnel barriers", *Nature Mater.*, 3, p.862, 2004

[8] S.Yuasa, T.Nagahama, A.Fukushima, Y.Suzuki and K.Ando,"Giant room-temperature magnetoresistance in single-crystal Fe/MgO/Fe magnetic tunnel junction", *Nature Mater.*, 3, p.868, 2004

[9] M. Motoyoshi, K. Moriyama, H. Mori,C. Fukumoto, H. Itoh, H. Kano, K. Bessho, and H. Narisawe, "High-performance MRAM technology with an improved magnetic tunnel junction material", *in 2002 Sympo. on VLSI tech.*, p.212, 2002

[10] K.Tsunekawa, D.D.Djayaprawira, S.Yuasa, M.Nagai, H.Maehara, S.Yamagata, E.Okada, N.Watanabe, Y.Suzuki and K.ndo, "Huge Magnetoresistance and low junction resistance in magnetic tunnel junctions with crystalline MgO barrier", *IEEE Trans. on Magn.*, 42, p.103, 2006

[11] K. Tsunekawa, Y. Nagamine, H. Maehara, D. D. Djayaprawira, M. Nagai and N. Watanabe, "Over 60% TMR at room temperature in MTJ films prepared with surface modification process.", *in 9th Joint MMM/Intermag Conf. 2004*, BD-03, 2004

[12] D. Wang, C. Nordman, J. Daughton, Z. Qian and J. Fink, "70% TMR at room temperature for SDT sandwich junctions with CoFeB as free and pinned layers", *in 9th Joint MMM/Intermag Conf. 2004*, BD-02, 2004

[13] J. J. Sun, K. Shimazawa, N. Kasahara, K. Sato, S. Saruki, T. Kagami, O. Redon, S. Araki, H. Morita, and M. Matsuzaki, "Low resistance and high thermal stability of spin-dependent tunnel junctions with synthetic antiferromagnetic CoFe/Ru/CoFe pinned layers", *Appl. Phys. Lett.*, 76(17), p. 2424, 2000

[14] J. Fujikata, T. Ishi, S. Mori, K. Matsuda, K. Mori, H. Yokota, K. Hayashi, M. Nakada, A. Kamijo, and K. Ohashi, "Low resistance magnetic tunnel junctions and their interface structures", *J. Appl. Phys.*, 89(11), p.7558, 2001

[15] Bryan Oliver, Qing He, Xuefei Tang, and J. Nowak, "Dielectric breakdown in magnetic tunnel junctions having an ultrathin barrier"., *J. Appl. Phys.*, 91(7), p.4348, 2002

[16] N. Sakimura , T. Sugibayashi, T. Honda, S. Miura, H. Numata, H. Hada and S. Tahara, "A 512Kb cross-point cell MRAM", *in 2003 ISSCC Dig. of Tech. Papers*, p.278, 2003

[17] H.Tanizaki, T.Tsuji, J.Otani, Y.Yamaguchi, Y.Murai, H.Furuta, S.Ueno, T.Oishi, M.Hayashikoshi and H.Hidaka, "A high-density and high-speed 1T-4MTJ MRAM with voltage offset self-reference sensing scheme", *in 2006 ASSCC Dig. of Tech. Papers*, p.303, 2006

Performances of a ZrO_2 PEALD Dielectric for 45nm Embedded DRAM 3D MIM (Metal-Insulator-Metal) Stacked Capacitors

A. Berthelot [a] , C. Caillat [b] , H. Del-Puppo [c], B. Icard [d], E. Deloffre [b], N. Emonet [b], M. Gros-Jean [b], S. Barnola [d], C. Soonekindt [a], R. Pantel [b] and F. Lalanne [b]

[a] NXP Semiconductors, [b] STMicrolectronics, [c] Freescale Semiconductors, [d] CEA LETI-Minatec,

850, rue Jean Monnet, 38926 Crolles, France
Phone : +33 (0)4 38 92 20 27 ; Fax :+33 (0)4 38 92 29 51 ; email : audrey.berthelot@nxp.com

Abstract

For the first time, we report an evaluation of a 45 nm embedded DRAM (eDRAM) 3D Metal-Insulator-Metal (MIM) stacked capacitor integration with an aggressive cell size of 0.072µm². For this study, we have evaluated a ZrO_2 dielectric layer deposited by Plasma Enhanced Atomic Layer Deposition (PEALD) at a low temperature (250°C) and with a high throughput. We have obtained an extremely low Equivalent Oxide Thickness (EOT) of 6.7 Å. Moreover, this 6.7 Å EOT dielectric meets all the 45 nm eDRAM specifications: high capacitance value with leakage current density within specifications (<1fA/cell at +/-1Volt) and an extrapolated lifetime greater than 10 years (at Vdd=1V).

1. Introduction

Using embedded stacked DRAM memory in modern SOC (Silicon On Chip) systems is very attractive in term of chip size, speed and power consumption. Stack technology is divided into two families: COB (Capacitor Over Bit line) and CUB (Capacitor Under Bit line). The CUB architecture is very low cost: only three additional critical masks are needed to integrate the capacitor, while COB needs up to 6 critical masks.

In this paper we report for the first time an evaluation of a 45 nm eDRAM CUB stacked capacitor integration. With an aggressive cell size of 0.072µm², these new 45 nm eDRAM memory cuts are very competitive by comparison to embedded SRAM (eSRAM) since they are expected to be 3.5 times smaller than 45 nm eSRAM cuts.

For our 45 nm eDRAM MIM stacked capacitor architecture, the dielectric must meet the following requirements: a high dielectric constant, an EOT below 1 nm, a dielectric leakage current within specifications (<1fA/cell at +/-1 Volt) and an extrapolated lifetime > 10 years (at Vdd=1V). As previous works have demonstrated that ZrO_2 is the good candidate for 45 nm DRAM technology and below [1,2], this dielectric material has been chosen for this study on a real 45 nm eDRAM flow. ALD technology has already been chosen in production for deposition of the high-K dielectrics in memories due to excellent film quality and conformal step coverage [3]. However, a low throughput has been viewed as one major issue in term of mass production. For that reason, we have chosen to evaluate for the first time a PEALD ZrO_2 layer. Moreover, the deposition temperature of 250°C is low by comparison to a standard ZrO_2 ALD layer [1]: a lower thermal budget reduces the risk of device performance degradation. The excellent electrical characteristics of these 45 nm 3D MIM stacked capacitors are reported in this paper.

2. Integration

Experiments were performed on a real 45 nm 3D stacked CUB capacitor flow. Figure 1 presents a schematic cross section of our 45 nm TiN/ZrO_2/TiN MIM eDRAM cell. Figures 2a and 2b are TEM (Transmission Electron Microscopy) images of 45 nm stacked capacitors in both Y and X directions respectively.

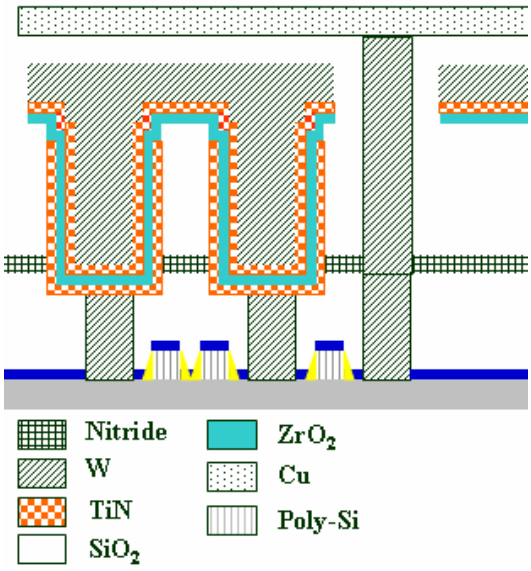


Fig. 1: Schematic cross section of CUB stacked embedded DRAM (Y direction).

In this early development phase, we have chosen to perform some critical lithography levels with ebeam process instead of standard optical one.

After the formation of the W plug (same as the CMOS process baseline), a SiN etch stop layer and a TEOS layer are then deposited. A cylindrical cavity (X=0.09µm; Y=0.3µm and H=0.3µm) is etched in the

oxide layer (figure 3a). The critical lateral distance between two adjacent capacitors is 70 nm. The TiN ALD bottom electrode is deposited and an Etch Back step is used to separate each electrode. After cleaning steps, the dielectric ZrO_2, the TiN top electrode and a W capping layers are deposited: our 3D capacitor is then formed (figure 4). A second specific lithography step is used to etch the W/TiN/ZrO_2 stack in order to allow the path of the specific eDRAM contact to connect the bitlines (1st Copper line) with the W plug (figure 3b). Despite a relatively high aspect ratio of this contact (about 8.5:1), TEM picture proves that the W filling is very good (figure 5): no void in W is observed in the centre of this contact in SEM top view images.

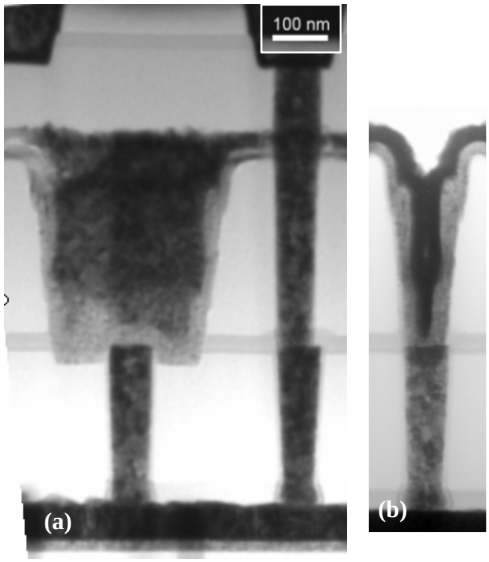


Fig. 2: TEM cross sections of 45 nm stacked eDRAM TiN/ZrO_2/TiN capacitors in (a) Y and (b) X directions.

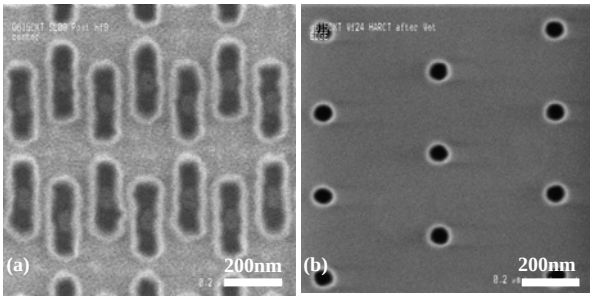


Fig. 3: Top view SEM (Scanning Electron Microscopy) images of (a) capacitors and (b) High Aspect Ratio Contacts after etch.

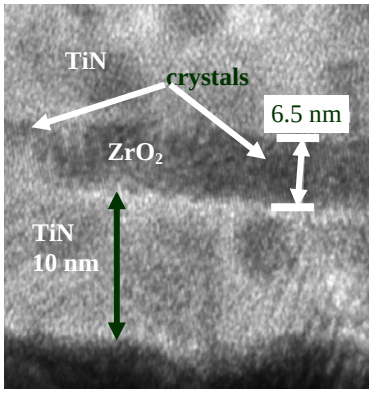


Fig. 4: High Resolution Transmission Electron Microscopy image of TiN/PEALD ZrO_2/TiN stack (EOT=6.7 Å).

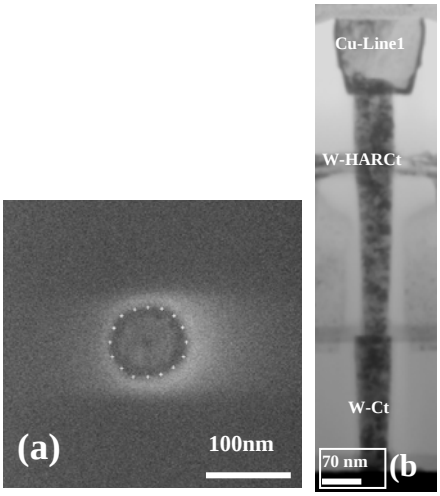


Fig. 5: (a) SEM Top view after CMP W and (b) Transmission Electron Microscopy image of High Aspect Ratio Contact.

3. Experiments

ZrO_2 film was deposited at 250°C by PEALD in a showerhead reactor, using a TEMAZr source. The elementary cycle is composed of source dispense/O_2 plasma/purge. PEALD technology uses shorter purge times than ALD allowing a throughput improvement. The thickness uniformity of the film is below 1% (1σ).

In this work, we have evaluated a 65 Å ZrO_2 PEALD film. The top and bottom electrodes are both 10 nm TiN layers, deposited by ALD at 400°C, using a $TiCl_4$ source and NH_3 as reactant. No additional anneal step is done after the TiN top electrode deposition.

An XRD analysis done on this 65 Å ZrO_2 layer does not give any crystalline information, because of a too thin film. Nevertheless, TEM images prove that the 65 Å ZrO_2 layer presents a polycrystalline structure despite a very low deposition temperature (figure 4). The step coverage of any plasma assisted technique can be a concern. In this study, thickness measurements based on TEM photographs show a dielectric thickness difference below 10 % between vertical and horizontal dielectric deposition. Thus, excellent step coverage is obtained.

4. Results and Discussion

Equivalent Oxide Thickness-Capacitance

Figure 6 shows the Equivalent Oxide Thickness of ZrO_2 ALD standard process and PEALD versus the physical thickness of the film. The EOT of this 65 Å PEALD ZrO_2 layer evaluated here is perfectly in line with the ZrO_2 ALD results. This new PEALD thin layer appears to be within the same regime as thicker standard ALD ZrO_2 film (between 80 Å and 120 Å) studied in a previous work [1]. TEM cross sections (figure 4) confirm that 65 Å PEALD layer presents the same polycrystalline structure as a thicker ALD ZrO_2 [1]. The threshold thickness for the ZrO_2 crystallization initialization on TiN ALD is then below 65 Å.

Moreover, the HRTEM images of 65 Å thick ZrO_2 PEALD film do not show any interfacial layer between ZrO_2 and TiN (figure 4). This is confirmed by electrical measurements, since the Y axis intercept is about 0.7 Å (figure 6): there is no measurable interfacial capacitance. The slope of the curve yields a ZrO_2 dielectric constant (K) of 45 for ZrO_2, which is a very high value by comparison to other results obtained for MIM applications [2,4]. As reported in the literature, dielectric constant of ZrO_2 is very dependent on crystalline structures [5] and a crystallized ZrO_2 film presents a higher dielectric constant than amorphous one [5].

Figure 7 presents the nearly flat C-V characteristics of ZrO_2 PEALD dielectric layer in the 45 nm MIM stacked capacitors. This TiN/ZrO_2/TiN stack offers a good Cmin/Cmax ratio over 90% within [-Vdd; +Vdd].

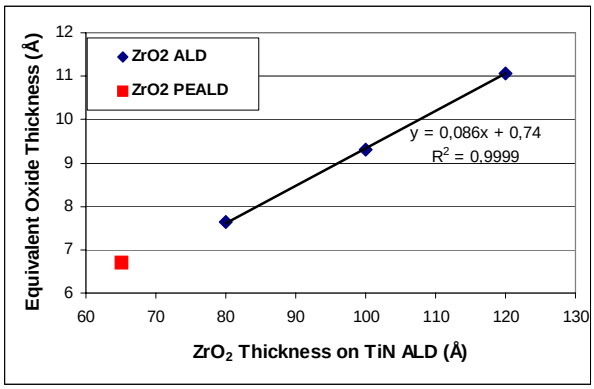


Fig. 6: Relationship between the TEM thickness and Equivalent Oxide Thickness of ZrO_2 ALD (from [1]) and PEALD films.

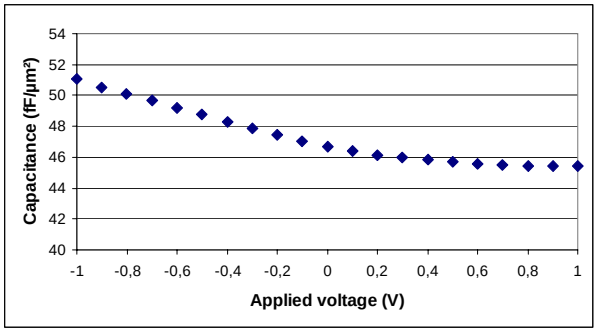


Fig. 7: C-V characteristic of 45nm stacked TiN / PEALD ZrO_2 / TiN capacitors at room temperature.

Dielectric Leakage Current

I-V characteristic of TiN/ZrO_2 PEALD/TiN capacitors at room temperature is presented on figure 8. This result shows that even though EOT went down to 6.7 Å, leakage current density is still within specifications (<1fA/cell @ +/-1 V). Nevertheless, a rapid rise in leakage is observed in low voltage range: this direct tunneling conduction is linked to the low dielectric thickness.

Figure 9 shows the dielectric leakage current density versus EOT at room temperature and at +/-Vdd (+/-1 V). The more rapid leakage rise for low EOT is only due to a change in the conduction mode (direct tunneling) due to a low thickness. Despite this point, PEALD ZrO_2 leakage current is in line with results obtained in previous work done on standard ZrO_2 ALD film in both negative and positive voltage.

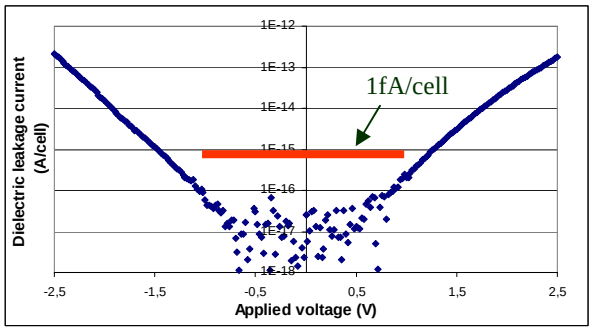


Fig. 8: I-V characteristic of TiN ALD / 65Å ZrO_2 PEALD / TiN ALD stacked capacitors at room temperature (400K cell structure).

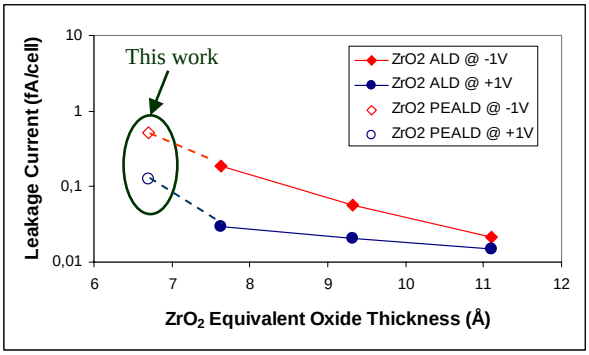


Fig. 9: Dielectric leakage at V=+1 / -1V and at room temperature versus EOT.

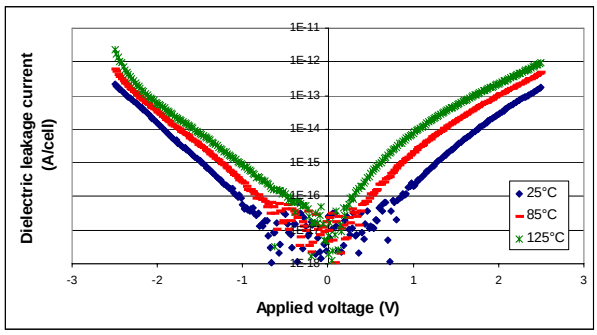


Fig. 10: I-V characteristics of TiN / 65Å PEALD ZrO_2 / TiN stacked capacitors at 25°, 85°C and 125°C.

The temperature dependency of I-V characteristics for the 65 Å ZrO_2 PEALD film is presented in figure 10. This film offers a good temperature stability since the leakage increase is less than 1.5 decade between 25°C and 125°C.

Voltage-to-breakdown

The leakage current density of the MIM 45 nm stack is evaluated over a large voltage range in order to determine the Voltage-to-Breakdown (Vbd). The positive Vbd is plotted in figure 11 and a comparison is done between ZrO_2 PEALD, ZrO_2 ALD and Al_2O_3 ALD (previous works [1]). A linear variation of Vbd is

observed for both dielectrics. Moreover, the breakdown behavior of ZrO_2 ALD and PEALD are perfectly in line, which confirms that the 65Å PEALD layer structure is very similar to the structure of thicker ALD layers. No particular unreliable behavior is expected for this thin PEALD film.

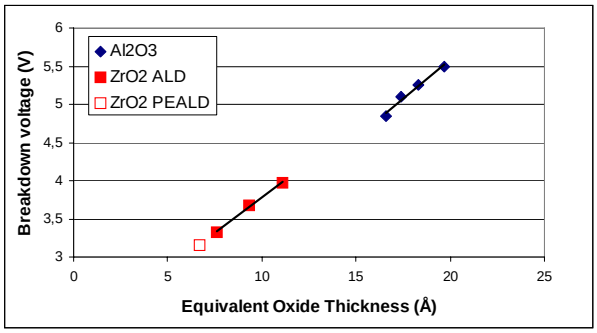


Fig. 11: Variation of positive breakdown voltage (Vbd) with EOT for ZrO_2 and Al_2O_3 at room temperature.

TDDB studies

Because of low Vbd values, the intrinsic oxide breakdown lifetime of this 65 Å ZrO_2 film is questionable. We have applied constant voltage stresses to the top electrode while the bottom electrode was grounded. As the positive voltage stress appears to be a slightly worse case (figure 8), this study was done for positive voltage stresses only.

Figure 12 shows a typical breakdown behavior of the thin 65 Å ZrO_2 film, which consists of time dependent soft breakdowns. Soft breakdown is considered to result from a weak localized path in the oxide between electrodes [6]. Leakage variations in the soft breakdown regions result from electrons trapping/detrapping mechanisms.

In this study, we have extracted from leakage versus stress time curves (I(t)) (example shown in figure 12) the Time Dependant Dielectric Soft Breakdown (TDDSBD), corresponding to the first Soft Breakdown event. For each applied voltage, the median value of TDDSBD is reported versus voltage in figure 13 which, thus, represents the intrinsic behavior of our 45 nm MIM capacitor. This result proves that an extrapolated lifetime greater than 10 years at use conditions (Vdd=1V) is obtained with this thin PEALD ZrO_2 film.

5. Conclusion

For the first time and despite a critical cell size of 0.072µm², the integration of 45 nm MIM stacked capacitors was successfully demonstrated on large DRAM monitoring structures. Moreover, in view of mass production, we have chosen to evaluate a high throughput PEALD ZrO_2 dielectric layer. This study demonstrates that this thin film can meet all the requirements needed for 45 nm eDRAM technology: an extremely low EOT of 6.7Å and a dielectric leakage current within specifications (<1fA/cell at +/- Vdd). Soft breakdowns are monitored versus stress time and an extrapolated lifetime greater than 10 years is found at +/- Vdd.

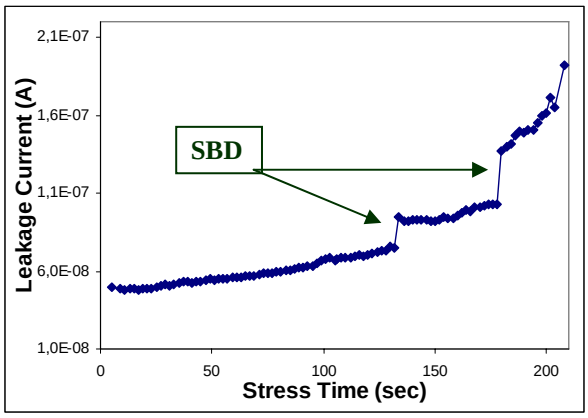


Fig. 12: Soft breakdowns (SBD) of 65 Å ZrO_2 film at V=2.8V at room temperature on a 40 K cell structure.

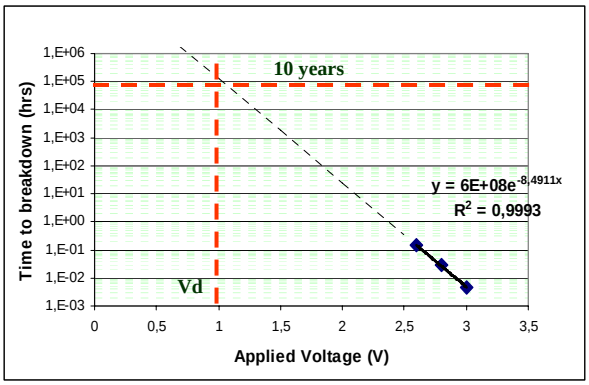


Fig. 13: Time Dependent Dielectric Soft Breakdown (TDDSBD) behavior of TiN (ALD) / 65 Å ZrO_2 (PEALD) / TiN(ALD) for 45 nm stacked capacitors (40 K cell structures) at room temperature.

Acknowledgement

PEALD ZrO_2 dielectric development was supported by ASM Phoenix. Part of this work was realized, within the Crolles2 Alliance project, in the framework of European Eureka program MEDEA+ T-126 (Bluberries) on embedded memories

References

[1] A. Berthelot et al., ESSDERC proceedings, 343 (2006)
[2] K-R Yoon, et al., SSDM, 188 (2005)
[3] Gerritsen, et al., Solid State Elec., **49**, 1767 (2005)
[4] G.D Wilk, et al., J. Appl. Phys., **89**, n°10, 5243, (2001)
[5] X. Zhao and D. Vanderbilt, Phys. Rev. B. **65**, 075105, (2002)
[6] M.K. Bera and C.K. Maiti, IPFA proceedings, 295 (2006)

Conductance switching behaviour of a phenol substituted bithiophene memory device

M. Caironi [a], D. Natali [a], M. Sampietro [a], C. Bertarelli [b], A. Bianco [b], E. Canesi [b], G. Zerbi [b]

[a] Dip. di Elettronica e Informazione, Politecnico di Milano, P.za L. da Vinci, 32 20133 Milano ITALY, marco.sampietro@polimi.it

[b] Dip. di Chimica, Materiali e Ing. Chimica, Politecnico di Milano, P.za L. da Vinci, 32 20133 Milano ITALY

Abstract

An extensive testing is presented on a new class of organic memory devices based on phenol substituted bithiophenes whose conductivity can be switched between two stable values upon electric pulses. The devices are shown to operate both in air and in vacuum with low programming voltages of about ±4V, to retain the information for largely more than 48 hours in each state and to sustain multiple write&erase cycles in excess to 100, without degradation in the active material volume as testified by infrared spectroscopy. The experienced drift of the off-current upon prolonged electrical stress is discussed and is shown to be a partially reversible process. In addition the device switching time was investigated and possible basic switching mechanisms are discussed.

1. Introduction

Organic materials are attractive for memory applications: they can be ideally scaled down to molecular size, they are processed at a temperature close to ambient temperature, they can be deposited in layered structures on a large number of substrates, silicon included. Therefore organic materials address correctly the issues of next generation memory devices, that is minimum size of the cell and 3-dimentional stacking for maximum density [1], without forgetting the lower cost of fabrication and the ecological advantage of a very low energetic budget during production.

These potentialities have since long solicited the investigation of memory effects in organic materials, leading to interesting results on both molecular [2] and bulk [3,4] bistable devices. Although questions still remains on the basic switching mechanisms [5-7] and on the role of the electrodes [8], organic-based devices have proved to merit attention.

In this paper we propose a voltage-driven conductance switching memory device based on an organic molecule. The paper is organized as follows: after reviewing the chemical structure of the adopted molecule and the device architecture in Sec. 2, the device operation is described in Sec. 3; the measurements of the retention time are reported in Sec. 4 and the effect of prolonged electrical stress is discussed in Sec. 5. In Sec. 6 the effect of film morphology on the device performance is reported and in Sec. 7 the device switching time is analyzed conjunction with a brief discussion of the possible involved mechanisms. In Sec. 8 some conclusions are drawn.

2. Active material and device realization

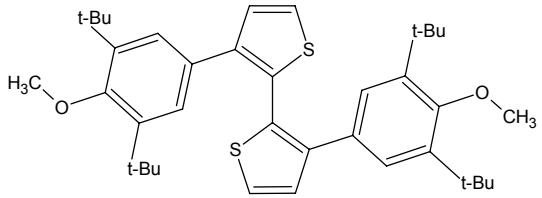


Fig. 1: Chemical structure of the DPBT molecule (*t*-Bu stands for *tert*-butyl groups).

The active material is a 3,3'-bis-(3,5-di-*tert*-butyl-4-methoxyphenyl)-2,2'-bithiophene, hereafter named DPBT, shown in the scheme of Fig.1. DPBT belongs to the general class of hindered phenol substituted bithiophenes. The molecule has a limited degree of planarity: DFT (Density Functional Theory) simulations show that the inter-ring torsional angle between the thiophene units is 54.3° and the angle between the thiophene-phenyl units is 40.3°. As a consequence the delocalization of π-electrons is limited and the resulting energy gap relatively large (absorption band at 270 nm). DPBT was synthesized by Suzuki coupling of 1-methoxy-2,6-di-*tert*-butyl-4-bromo benzene [9] with 3-thiophen boronic acid [10] to give 3-(3,5-di-*tert*-butyl-4-methoxyphenyl)thiophene. The resulting intermediate was first brominated in 2-position with *n*-bromosuccinimide in dimethylformamide and then reacted with buthyllithium and coupled with $CuCl_2$ to yield the targeted molecule.

The tested devices have a vertical sandwich structure. Starting from a 1 mm thick glass substrate with a 70 nm thick ITO layer ($R_{\square}$<20 Ω/□, Merck, DE), which acts as bottom electrode, DPBT was spin cast in ambient air from a chloroform solution (60 mg/ml) to yield a ≈400 nm thick film. Then samples were transferred in a vacuum chamber (≈10^{-6} mbar) to evaporate the top aluminium electrode (≈100 nm). The active area of devices is about 3 mm^2.

3. Conductance switching in air and in vacuum

DPBT-based devices can store one bit of information exploiting a voltage driven conductance switching effect. Devices are initially in a low conductive state, the OFF state, which is preserved when applying a moderate negative or positive voltage to the Al electrode. By applying a suitable negative voltage, the device switches to a more conductive state, the "ON state", which is retained until a suitable positive voltage restores the OFF

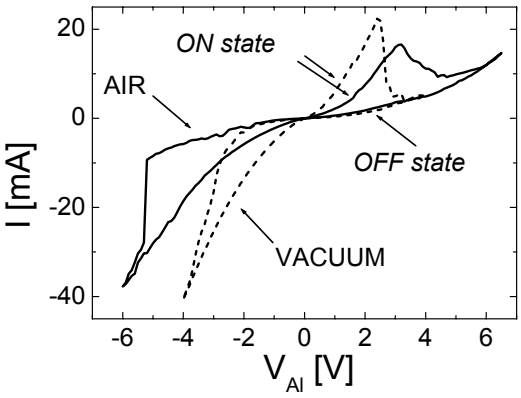


Fig. 2: I-V characteristic curves of the DPBT memory cell measured in ambient air and in vacuum (p≈10^{-6} mbar).

state. As shown in Fig. 2, this bistable behaviour was reproducibly measured not only in high vacuum (p=10^{-6} mbar) but also in ambient air. The ratio between the ON current and the OFF current can be as high as 100. The difference between vacuum and air operation stands in the magnitude of the operating voltages: in vacuum writing and erasing voltages are –3V and +2.5V respectively, whereas in air they are slightly higher in magnitude, in the range of about –5.5V and +4V respectively. This difference can be ascribed to a partial oxidation of the aluminium contact through defects during air operation [11].

4. Retention time of the memory cell

We tested the capability of the devices to retain the programmed conductive state. We forced the ON state and then probed the device state by applying every 10 s a 0.5V pulse 130ms long and measuring the resulting current. Afterward the device was brought to OFF state and checked with the same voltage pattern. The measurement scheme and results are reported in Fig. 3. Both the ON and OFF state are stable over the investigated time scale (64 hours for the ON state and 48 hours for the OFF state) with no noticeable sign of

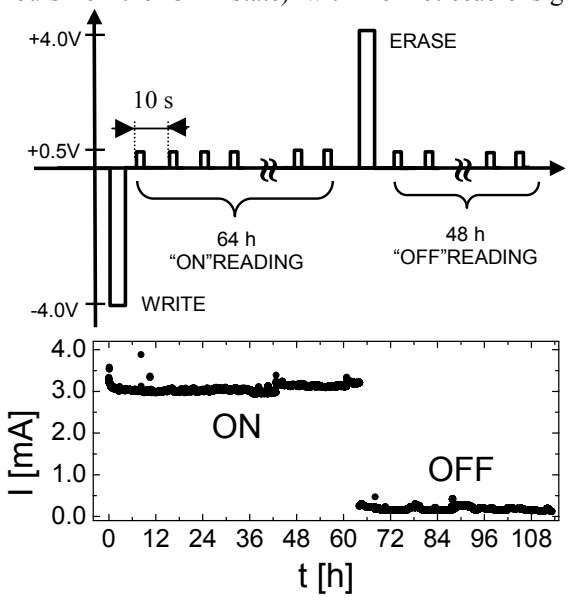


Fig. 3: Measurement of the retention time for the ON and the OFF states (p=10^{-6} mbar) of the DPBT memory devices.

degradation after more than 4 days and 40.000 consecutive electronic queries.

5. Endurance to Read&Write cycles

We tested the robustness of the devices by performing a series of consecutive write and erase cycles. This was done both in vacuum and in air. In each single cycle the voltage is swept from 0V to the threshold voltage for the OFF-to-ON transition, then to the threshold voltage for the ON-to-OFF transition, and finally back to 0V. The devices withstood the prolonged cycling both in vacuum and in air preserving their conductance switching behaviour. The main difference between air and vacuum consists in somewhat noisier current values for the latter. A common feature is that whereas the ON current is substantially constant and unaltered by the electrical stress, there is an increasing trend for the OFF current and a consequent degradation in the ON/OFF ratio. This is shown in Fig. 4 for measurements in vacuum. It is to be noted though that the degradation in the device performance subsequent to electrical stress is not an irreversible process. In fact, when a stressed device measured in vacuum is exposed to air for one hour and then it is brought back to vacuum, the OFF current was restored to lower values, similar to those recorded in the first cycles.

This is further supported by IR spectroscopy performed on the DPBT films. The analysis was carried out *in situ*, exploiting the high reflectivity of the ITO for IR frequencies that turns in a double transmission experiment. Fig. 5 displays the spectra for a freshly deposited DPBT film (solid line), just before aluminium deposition, and for a film which was stressed with hundreds of IV cycles (dotted line). In this latter case aluminium was mechanically delaminated to access the active material volume. The FT-IR spectrum of DPBT shows two regions of interest: one between 3000 and 2800 cm^{-1}, due to the symmetric and anti-symmetric stretching of C-H bonds, and one between 1500 and 800 cm^{-1}, mainly due

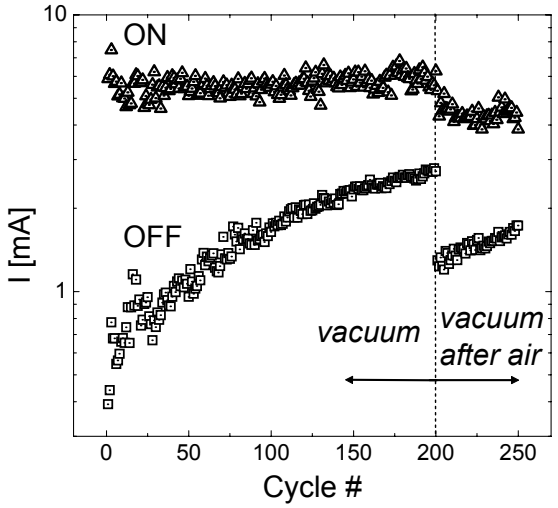


Fig. 4: Evolution of ON and OFF currents (read at 1V) upon prolonged cycling in vacuum. The last 50 cycles have been measured after breaking and restoring the vacuum.

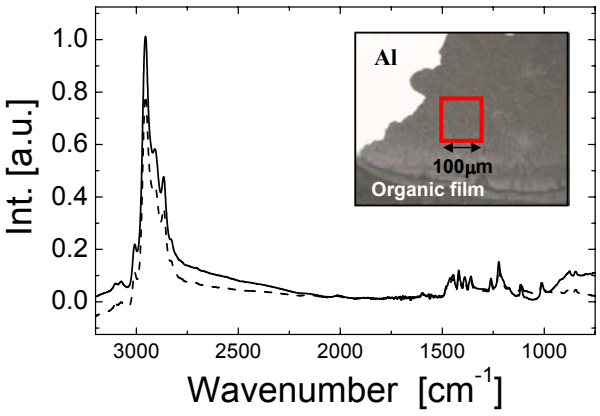


Fig. 5: FT-IR spectra measured on a freshly deposited DPBT film (solid line) and on an electrically stressed film (dotted line) of a delaminated device, which is shown in the inset with the indication of the analyzed area.

to normal modes localized on lateral functional groups or delocalized on the molecular backbone. The former contribute is very intense given the presence of *tert*-butyl groups. The latter is the most interesting one and we can look at it as the fingerprint of the molecule. As the spectra of the freshly deposited and of the delaminated one almost superimpose, it is clearly demonstrated that no chemical modifications have undergone with IV cycling and that the noticed increase of the OFF current can not be ascribed to an irreversible degradation occurring in the bulk of the organic material.

6. Effect of film morphology

We have investigated the role of the film morphology on the device performance. Freshly spin-cast films of DPBT are amorphous, but this state is metastable: in fact, if not covered by Al, the film evolves towards the formation of more ordered domains with a time scale of a few days. This was monitored by infrared spectroscopy by focusing on the normal mode related to the O-CH_3 stretching, which occurs at around 1000cm^{-1}: as the direction of this vibration is orthogonal to the plane where the phenyl ring lies, it is highly sensitive to the solid state arrangement of the molecules. In Fig. 6 spectra taken just after deposition (solid line) and one week after deposition (dotted line) are reported: the shift in the peaks of the latter spectrum indicates a higher degree of crystallinity for aged DPBT films.

The film crystallinity has an adverse impact on the performance of the memory cell: when Al is evaporated on freshly deposited films we observe high reproducibility of the electrical characteristics of devices, whereas if Al is deposited on crystalline films we hardly observe conductance switching phenomena and devices do not endure electrical stress in excess of a few cycles. Note that the deposition of the Al cathode prevents the crystallization process: in fact spectra of delaminated devices are similar to the ones of freshly deposited films even few weeks after cathode deposition.

7. Switching Time

The characteristic switching times of the devices were investigated by applying a voltage step to the cell

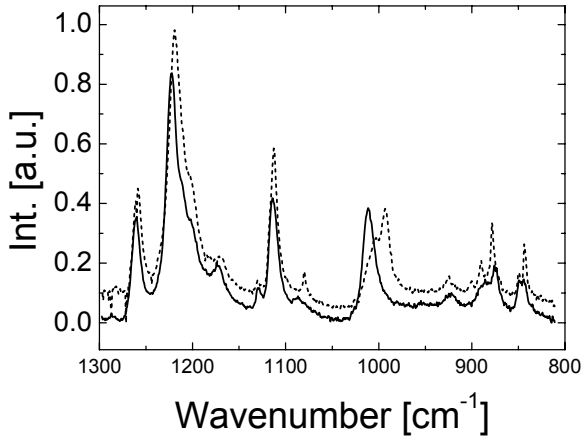


Fig. 6: FT-IR spectra for a freshly deposited film (solid line) and for a one week aged film (dotted line), which has a higher degree of crystallinity.

by means of a 50Ω terminated pulse generator and by measuring the voltage across the device.

For the ON-to-OFF transition the voltage across the device, following the application of a +5V pulse lasting 100ms, is shown in Fig. 7. The erasing transition (curve I) is characterized by two processes: an initial *delay* which lasts a few hundreds of microseconds, followed by the proper *switching* of the cell which is accomplished in about one hundred microseconds, as can be appreciated in the inset of Fig. 7. When consecutive writing and erasing pulses are applied to the cell (curves II and III of Fig.7), the switching time is substantially unaltered, whereas the delay time does not endure the electrical stress and shows an increase to few milliseconds for curve III.

No trends in consecutive cycles have been instead measured for the OFF-to-ON transition characterized by a delay of few hundreds microseconds followed by the proper switching lasting a few milliseconds.

The reported transition times are not compatible neither with an electronic process, that would occur on the scale of the nanoseconds [4], nor with a filamentary path formation as the one reported to occur on a microseconds time-scale for polymer matrixes [12].

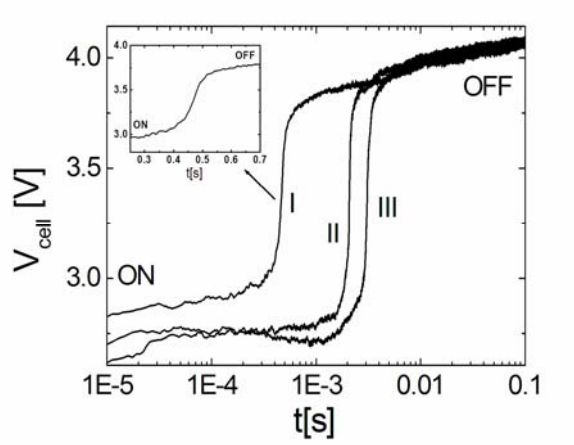


Fig. 7: Three consecutive ON-to-OFF transitions induced with +5V pulses lasting 100ms. Voltage across the cell is read by an oscilloscope. In the inset a zoom in linear scale of curve I is displayed.

Other phenomena that might explain the observed conductance switching are the presence of a thin layer of Al_2O_3 between the Al electrode and the active film which has been claimed to give rise to material-independent memory effects [5], and metal nanoparticles brought inside the active media during the thermal evaporation of the top contact [6, 13] In order to rule out these effects, we substituted aluminum with mercury: even though it is not an easy processable metal, it addresses the above mentioned issues since *i)* it is not subject to oxidation and *ii)* is cold deposited thus forming a sharper preformed contact with the organic semiconductor [14]. Measurements, performed both at a moderate vacuum of 40 mbar (to avoid Hg evaporation) and in air, show that also in this case the memory cell is characterized by a voltage driven conductance switching, and therefore indicate that the organic material does not act as a mere matrix, but is involved in the observed conductance switching [11].

Transition times in DPBT based devices are more similar to the one's reported for other phenomena such as conformational changes, occurring for example in ferroelectric polymers in tens of microseconds [15], atomic movements (millisecond time-scale in bistable rotaxanes [16]), molecular isomerization (photochromic dyarylethene [17], and dopants migration in a P3HT depletion layer [18]. Further investigations to clarify the observed behavior are under way.

8. Conclusions

In conclusion, we have reported on a simple two terminal device that can store information in a non-volatile way by means of two well distinct conductance states that can be programmed and read electronically. The device characteristics in term of thresholds (few volts), retention (few days), writing&erasing cycles (few hundreds), switching times (below milliseconds) together with its stability to open air conditions, prove that hindered phenol substituted bithiophenes merit further attention and investigation toward the realisation of scaled memory cells.

Acknowledgments

Authors wish to thank Davide Ulivi for his help during devices characterization. Funding of CEE under Nosce Memorias project and of Italian MIUR through FIRB RBNE033KMA is acknowledged.

References

[1] C. Pinnowe and T. Mikolajick, J. of the Electrochem. Soc. 151, K13 (2004).
[2] P. M. Mendes, A. H. Flood, and J. F. Stoddart, Appl. Phys. A **80**, 1197 (2005); C. Li, D. Zhang, X. Liu, S. Han, T. Tang, C. Zhou, W. Fan, J. Koehne, J. Han, M. Meyyappan, A. M. Rawlett, D. W. Price, and J. M. Tour, Appl. Phys. Lett. **82**, 645 (2003).
[3] S. Moller, C. Perlov, W. Jackson, C. Taussig, and S. R. Forrest, Nature (London) **426**, 166 (2003); A. Bandyopadhyay and A. J. Pal, Appl. Phys. Lett. **84**, 999 (2004); Q. Lai, Z. Zhu, Y. Chen, S. Patil and F. Wudl, Appl. Phys. Lett. **88**, 133515 (2006); D. Tondelier, K. Lmimouni, D. Vuillaume, C. Fery, and G. Haas, Appl. Phys. Lett. **85**, 5763 (2004); J. Chen and D. Ma, Appl. Phys. Lett. **87**, 023505 (2005); Y. Song, Q. D. Ling, C. Zhu, E. T. Kang, D. S. H. Chan, Y. H. Wang and D.-L. Kwong, IEEE Electron Device Lett. **27**, 154 (2006).
[4] Y. Yang, J. Ouyang, L. Ma, R. J.-H. Tseng, C.-W. Chu, Adv. Funct. Mater. **16**, 1001 (2006).
[5] M. Cölle, M. Büchel, and D. M. de Leeuw, Org. Electron. **7**, 305 (2006).
[6] L. D. Bozano, B. W. Kean, M. Beinhoff, K. R. Carter, P. M. Rice, J. C.Scott, Adv. Funct. Mater. **15**, 1933 (2005).
[7] M. J. Rozenberg, I. H. Inoue, and M. J. Sánchez, Phys. Rev. Lett. **92**, 178302 (2004).
[8] B. Mukherjee, and A. J. Pal, Org. Electron. **7**, 249 (2006).
[9] K. Takahashi, T. Suzuki, K. Akiyama, Y. Ikegami, and Y. Fukazawa, J. Am. Chem. Soc. **113**, 4576 (1991).
[10] S. O. Lawesson, Arkiv Kemi **11**, 373 (1957).
[11] M. Caironi, D. Natali, M. Sampietro, C. Bertarelli, A. Bianco, A. Dundulachi, E. Canesi, and G. Zerbi, Appl. Phys. Lett. **89**, 243519 (2006).
[12] Y. Segui, B. Ai, and H. Carchano. J. Appl. Phys. **47**, 140 (1976).
[13] J. G. Simmons and R. R. Verderber, Proc. R. Soc. London Ser. A, **301**, 77 (1967).
[14] G. G. Andersson, H. H. P. Gommans, A. W. Denier van der Gon, H. H. Brongersma, J. Appl. Phys. **93**, 3299 (2003); G. Nesher, A. Vilan, H. Cohen, D. Cahen, F. Amy, C. Chan, J. Hwang, A. Kahn, J. Phys. Chem. B **110**, 14363 (2006).
[15] W. Wang, T. Lee, and M. A. Reed, Phys. Rev. B **68**, 35416 (2003).
[16] Y. Chen, D. A. A. Ohlberg, X. Li, D. R. Stewart, R. S. Williams, J. O. Jeppesen, K. A. Nielsen, J. F. Stoddart, D. L. Olynick, and E. Anderson, Appl. Phys. Lett. **82**, 1610 (2003).
[17] T. Tsujioka and H. Kondo, Appl. Phys. Lett. **83**, 937 (2003).
[18] J. H. A. Smits, S. C. J. Meskers, R. A. J. Janssen, A. W. Marsman, and D. M. de Leeuw, Adv. Mater. **17**, 1169 (2005).

Improved CuTCNQ resistive non-volatile memories and a statistical study on their threshold voltage

J. Billen[*,°], R. Müller, J. Genoe, and P. Heremans[°]

IMEC vzw, MCP/PME, Kapeldreef 75, B-3001 Leuven, Belgium.

[°]Also with Katholieke Universiteit Leuven, ESAT/INSYS, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium.

[*]joris.billen@imec.be

Abstract

A significant improvement of electrical results for Cu/CuTCNQ/Al devices is achieved by an enhancement in CuTCNQ layer smoothness. A limited known amount of TCNQ is evaporated on a Cu substrate. The reaction to form CuTCNQ is forced in a second step. This allows a better growth control. Memories can be switched nearly 2000 times in pulsed measurements, with an extremely stable on-current and an on/off-ratio of over 100. We performed statistical studies on the threshold voltage distribution for switching and observe that for a 210 nm CuTCNQ layer this distribution can be approximated by a Gaussian, with peak at -/+ 4.2 V for switching on/off. Further we observe there is a trend that for switching thicker layers, higher voltages are required and that the switching voltage is temperature dependent.

1. Introduction

Scaling down of non-volatile FLASH memories is known to approach its physical limitations [1]. Although several organic materials showing bistable behavior have been suggested, the mechanism of switching is often unclear and reliability is problematic. Furthermore integration in the nanometer range is not straightforward and seldom performed.

CuTCNQ (Fig. 1) is an organometallic material known for exhibiting resistive switching in the nanosecond range [2]. The observed bistability is attributed to a partial transition from the ionic to the neutral phase and the material can be grown easily by a spontaneous reaction between Cu and TCNQ (7,7,8,8-tetracyano-p-quinodimethane) [2]. In the last few years electrical characteristics were improved by investigation of different growth mechanisms [3-6]. CuTCNQ seems to be of particular interest since it has been shown recently that it can be grown in 250 nm vias of a Cu back end-of-line process [5].

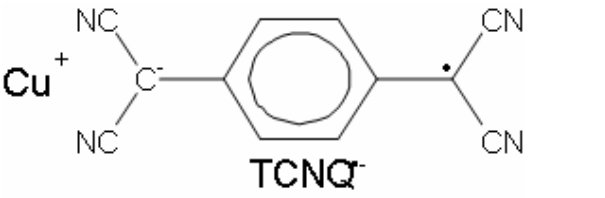


Fig. 1: Structure of CuTCNQ.

However, because of the spontaneous nature of formation of CuTCNQ, the reaction is often difficult to control. The resulting heterogeneous layers imply that electrical switching is limited to few 100 cycles. In this work we have focussed on a method that improves the smoothness of the CuTCNQ layer by depositing a known amount of TCNQ and forcing the reaction afterwards. This allows us to perform a detailed investigation of threshold voltage behaviour.

2. Sample preparation

Memories under consideration consisted of Cu lines (200-nm thick, 200-μm wide) attached by a TiW adhesion layer (15 nm) on a Si/SiO_2 (100 nm) substrate. Deposition of TCNQ was performed in an evaporation chamber and reaction to form CuTCNQ was forced in a second step. Finally a 100-nm thick, 200μm-wide Al top electrode was evaporated perpendicularly to the Cu lines, resulting in $(0.2\ mm)^2$ memory elements. Thicknesses were checked by SEM (5 kV acceleration voltage). Layer smoothness was quantified by acoustic mode AFM. The successive grown CuTCNQ in Fig. 2 shows a smooth, uniform, homogeneous layer of 210 nm. AFM shows an average roughness over an area of 10 x 10 μm of 21 nm. Before, smooth CuTCNQ layers were only obtained by coevaporation [3]. Our new result is a significant improvement compared to other growth methods as spontaneous electrolysis [2] and electrochemical growth [4] where polycrystals occur of micrometer size, or vapour phase growth [6] where growth in preferential direction results in nanowires up to 1 μm length. The observed roughness was then usually in the order of the film thickness [7].

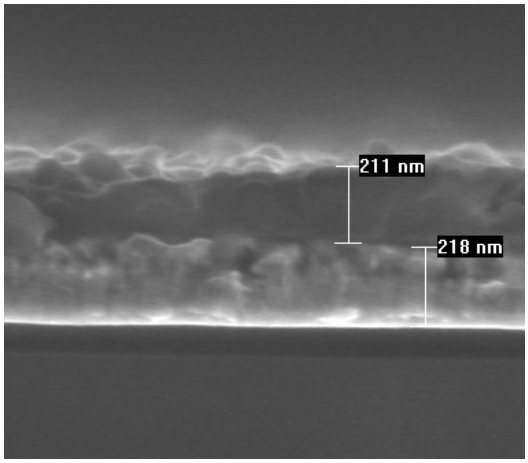


Fig. 2: Dense CuTCNQ layer (210 nm) grown from evaporated TCNQ on top of Si/SiO_2(100 nm)/TiW (15nm)/Cu (200 nm)-substrate.

3. Write/erase cycles

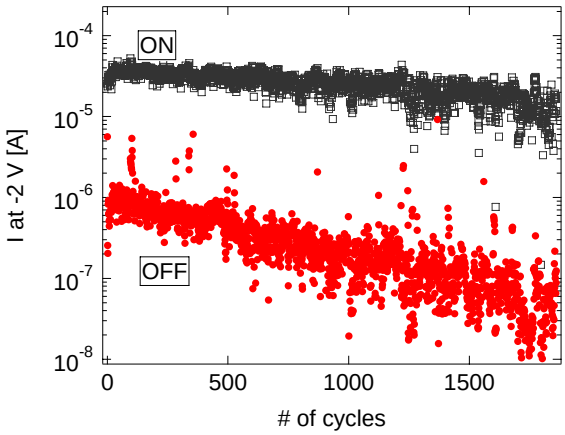


Fig. 3: On and off-currents at -2 V. Writing was performed by 500 ms pulses at -6 V, erasing by +6V (500 ms) applied to the memory in series with a 10 kΩ resistance.

Fig. 3 shows results of the current in two states at -2V applied for a 150 nm layer of CuTCNQ. Writing was performed by 500 ms pulses of -6 V, erasing by 500 ms pulses at +6 V. The signal was applied to the Al electrode while the Cu electrode was grounded. To limit the current during switching, the memory was protected by a series resistance of 10 kΩ and therefore the actual voltage across the memory will be somewhat lower than the applied voltage (it was measured to be around -/+ 4.5 V). The device shows an on/off-ratio between the two states of over 100 and an extremely stable on state around 0.02 mA (100 kΩ) for up to 2000 cycles. The initial off-state is around 1 μA (2 MΩ) and is decreasing over time. The dominant failure mechanism for these thin layers is a switch to a permanent on state. This is in contrast to previous results [6] where often an increase in off-current is observed. Here such a "closing window" is not observed.

The improvement is explained by the obtained smoothness of the layer. For the non-uniform layers reported previously, the deposited Al was in better contact at some parts of the layer. These regions were stressed more extensively in long-term performance tests and this leads to unreliable switching. For the obtained uniform layer this is less likely to occur.

4. Threshold voltage study

For the device shown in Fig. 2, the behaviour of the threshold voltage for on- and off-switching is examined. The *I(V)*-characteristics are measured applying 125 ms steps of 0.2 V increment, starting from 0 V → -6 V → +6 V → 0 V continuously for 250 times. Again the memory is protected by a resistor in series of 10 kΩ. Therefore we show the current against the measured voltage across the memory element itself (V_{mem}). The value of the series resistance determines the current that is reached in the high conductive state.

When extensive cycling is performed, the switching voltages will differ from experiment to experiment (inset Fig. 4). We are interested if higher threshold voltages will be required when the device is stressed for longer times and how threshold voltages are distributed statistically. The threshold voltage for on- (off-) switching is defined as the voltage where the current doubles (halves) for a 0.2 V increment. Fig. 4 shows how this corresponds to the observed switching for one typical cycle.

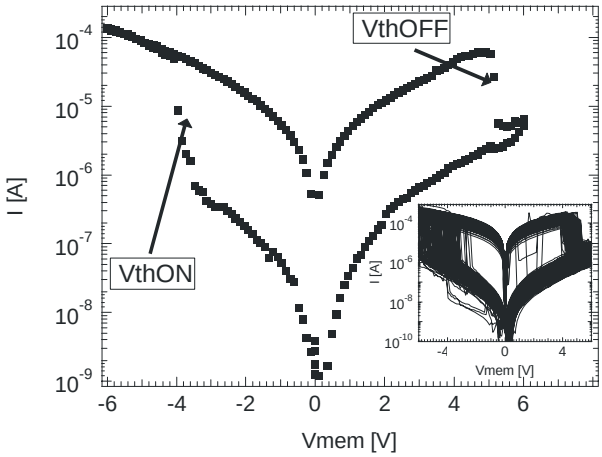


Fig. 4: *I(V)*-characteristic for Cu (200 nm)/CuTCNQ (210 nm) /Al (100 nm) element of area $(200\mu m)^2$ with threshold voltage for switching on (V_{thON}) and off (V_{thOFF}). The inset shows the data obtained for over 200 cycles.

The obtained switching voltages for switching on (V_{thON}) and off (V_{thOFF}) are shown in Fig. 5. Both increase in the first few cycles but remain fairly stable and symmetric compared to each other afterwards. Fig. 6 shows a histogram of the obtained threshold voltages approximated by a Gaussian. There is a peak at -4.2 V for on- and at +4.2 V for off-switching. Hence the average field for on- and off-switching is found to be at 20 MV/m for this 210 nm CuTCNQ device. We note that no on-switching occurred for voltages between 0 V and -2 V. As in all experiments, a small peak at low positive voltage (arrow in Fig. 5) was observed for switching off. Therefore a low negative reading voltage is mandatory.

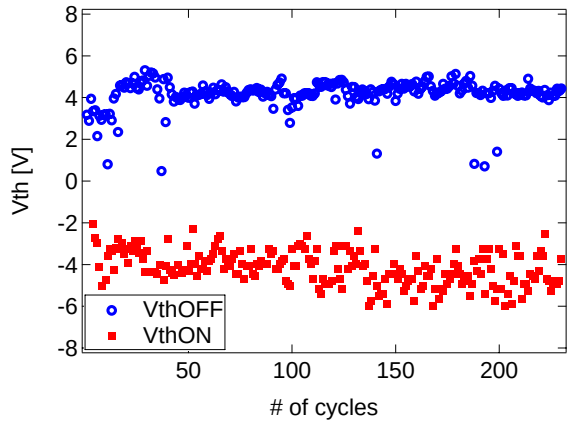


Fig. 5: Threshold voltages for switching on and off for 230 cycles for the device shown in Fig. 2.

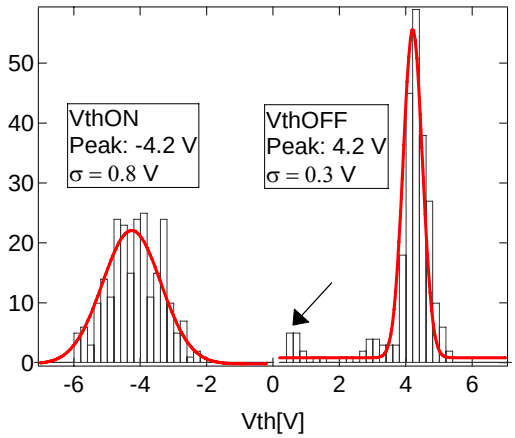


Fig. 6: Histogram of statistical distribution of threshold voltages for on and off switching approximated by Gaussian fits.

Besides the improved smoothness and enhanced electrical results, the described method allows us now to produce layers of different thicknesses. For those the threshold voltage distribution was examined in similar way. The average threshold voltage for on- (V_{thON}) and off- (V_{thOFF}) switching are displayed below in Table 1 together with the standard deviations of the resulting Gauss curves.

If we assume the distribution is indeed Gaussian, this suggests that a failure rate below 0.4 % will be reached when an excess voltage of 3σ (99.7% confidence interval) is applied on top of the mean value of the Gaussian. This means that the switching voltage can be brought down to -/+ 3.6 V for a 150 nm layer. -4.5 V (excess of 6σ) would be required for almost 100% confidence, comparable to the result in Fig. 3. For the 350 nm layer the required field seems higher than the value reported before. We note that for these thicker CuTCNQ layers also the average roughness increased to some extent. Even thicker layers (>400 nm) again resulted in the undesired formation of polycrystalline layers with higher peaks and lower valleys.

d_{CuTCNQ} [nm]	V_{thON} [V]	σ_{ON} [V]	V_{thOFF} [V]	σ_{OFF} [V]
150	-2.7	0.3	2.6	0.4
210	-4.2	0.8	4.2	0.3
350	-8.4	0.5	6.1	0.4

Table 1: Peak and width of the Gaussian from a statistical distribution study of threshold voltages for on- (V_{thON}) and off-switching (V_{thOFF}).

In the past it was already shown that Cu/CuTCNQ/Al samples still show switching at elevated temperatures [8]. We investigated the influence of *T* on switching voltage by measuring complete *IV*-curves at room temperature and at 150°C, shown in Fig. 7. For the first 100 cycles at room temperature this device showed a peak in its histogram at -4.5 V for V_{thON} and at 5.2 V for V_{thOFF}. At 150°C the peak for on-switching decreases to -1.8 V and to 0.6 V for off-switching. Further it was found that at low T (-188°C) the threshold voltage increased significantly (not shown here). We conclude the threshold voltage depends strongly on *T*.

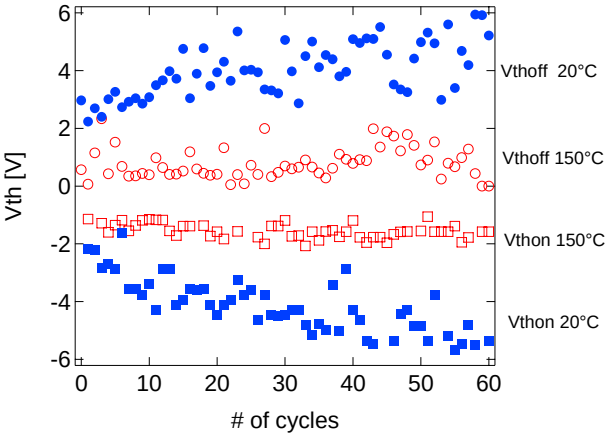


Fig. 7: Threshold voltages for switching on (squares) and off (circles) at room temperature (filled) and at 150°C (open).

4. Conclusions

In summary we have investigated a method for improving smoothness and uniformity of the CuTCNQ layer. The obtained electrical results are a significant improvement for Cu/CuTCNQ/Al memory elements. We were able to examine different layer thicknesses. A statistical study of the threshold voltage, points towards downscaling of the switching voltage with layer thickness and for a 150 nm layer, a -/+ 3.6 V pulsing sequence is suggested to obtain failure rate below 0.4%. Finally we observe that the threshold voltage depends strongly on temperature.

This research was performed within the framework of the NOSCE MEMORIAS project of the European commission (FP6-507934). The authors acknowledge D. Cheyns (IMEC) for SEM and C. Rolin (IMEC) for AFM measurements. JB thanks the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen) for financial support.

References

[1] R. Bez, A. Pirovano, Mater. Sci. Semicond. Process **7**, 349 (2004).
[2] R. Potember, T. Poehler, D. Cowan, Appl. Phys. Lett. **34,** 405 (1979).
[3] T. Oyamada, H, Tanaka, K. Matsushige, H. Sasabe, C. Adachi, Appl. Phys. Lett. **83**, 1252 (2003).
[4] R. Müller, J. Genoe, P. Heremans, Appl. Phys. Lett. **88,** 242105 (2006).
[5] R. Müller, S. De Jonge, K. Myny, D. Wouters, J. Genoe, P. Heremans, Appl. Phys. Lett. **89,** 223501 (2006).
[6] R. Müller, R. Naulaerts, J. Billen, J. Genoe, P. Heremans, Appl. Phys. Lett. (in press).
[7] J. Hoagland, X. Wang, K. Hipps, Chem. Mater. 1993, **5**, 54-60.
[8] R. Müller, J. Genoe, P. Heremans, 1st International Conference on Memory Technolgy and Design, Giens (F), May 2005.

The Influence of Different Electrode Materials on Resistive Switching in Cu:7,7,8,8-Tetracyanoquinodimethane Thin Films

Thorsten Kever, Ulrich Böttger and Rainer Waser

Institute of Materials in Electrical Engineering and Information Technology 2
RWTH Aachen University
D-52074 Aachen, Germany
kever@iwe.rwth-aachen.de

The occurrence of an electrical induced resistive switching effect in thin films of the metal-organic charge transfer complex system Cu:7,7,8,8-Tetracyanoquinodimethane (TCNQ) is long known. In this contribution, we will focus on the influence of the electrode materials on the switching effect. The investigated samples were prepared by physical vapor deposition (PVD). This process results in the formation of amorphous Cu:TCNQ thin films with a ratio of 1:1 of the metal and the organic compound.

Simple capacitor like test structures with different electrode materials were prepared with Cu:TCNQ thin films as an active layer. These simple memory cells were electrically and physically characterized. Depending on the used electrode materials different current- voltage characteristics could be observed.

1. Introduction

Materials with reversible conductance switching properties are desirable for future high-density non-volatile memory applications. Preferably, those materials should offer high resistance ratios between off and on states, a non destructive readout and low process complexity, and CMOS compability.

Various materials have been proposed for use in such systems. In this study, we will focus on the switching phenomena in metal-organic charge transfer complex thin films consisting of copper as metal donor and Tetracyanoquinodimethane (TCNQ) as organic acceptor. Resistive switching effects in Cu:TCNQ were first reported by Potember et al. [1]. They observed a current-controlled, bistable electrical switching.

In this study we focus on the influence of the electrode materials on the resistive switching properties of simple, capacitor like test structures with Cu:TCNQ thin films as active layers.

2. Sample Preparation

All samples analyzed in this study were prepared on oxidized silicon wafer pieces (25.4 mm x 25.4 mm). The memory cells are constructed as simple capacitor like structures as shown in Fig. 1.

The bottom electrode is deposited on a thin NiCr adhesion layer. Different electrode materials were used. The active layer in this set up is a Cu:TCNQ thin films with a thickness of around 150 nm. This layer is deposited via a physical vapor deposition (PVD) process. This method results in the formation of amorphous Cu:TCNQ thin films. In Fig. 1 a scanning electron microscope image of the sample structure without top electrode is shown.

The top electrode material is deposited through a shadow mask with circular openings between 75 µm and 1 mm in diameter. The deposition process and the physical characterization of the Cu:TCNQ layer are described in more details elsewhere [2].

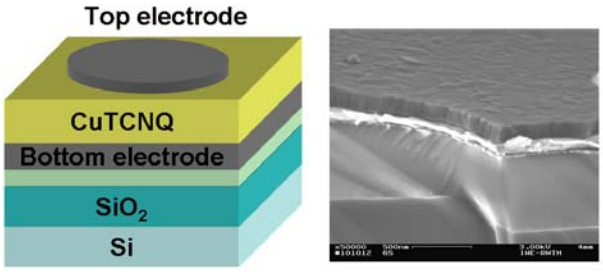


Fig. 1: Schematic test structure setup (left), and SEM image of a sample without top electrode (right).

3. Experimental Results

The standard memory cell setup for our measurements uses thermally evaporated copper as bottom electrode material and aluminium as top electrode material. This configuration yields the most reliable resistive switching characteristics. Our previous research on the nature of the switching effect was carried out on samples with this standard set up [3].

Typical resistive switching parameters for this standard configuration are high resistance states of $R_{OFF} \approx 5$ - 50 MΩ and low resistance states of $R_{ON} \approx 50$ - 500 kΩ. The switching thresholds for switching from R_{OFF} to R_{ON} are $V_{th,offon} \approx -3.5$ V and for switching back $V_{th,onoff} \approx +3$ V by application of the signal to the top electrode. A typical current voltage characteristic of such a memory cell is shown in Fig. 2. This measurement was performed with the current

compliance set to 5 µA in order to prevent the cell from breakdown. All samples were initially in the high resistance state. Those memory cells show fairly good endurance values, with life times of so far more than 10^4 cycles.

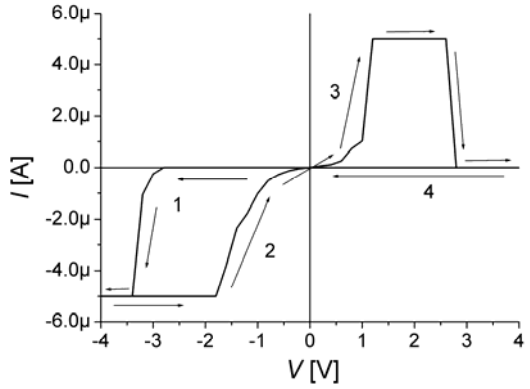


Fig. 2: Current- Voltage characteristic of a standard memory cell with thermally evaporated Cu bottom and Al top electrode. The Cu:TCNQ layer has a thickness of 130 nm.

The results of the electrical characterizations of different electrode setups are summarized in Table 1. The deposition of the top electrode material by sputtering led to shorted memory cells, even for very low sputtering energies. We could obtain non shorted samples only with thermally evaporated top electrodes.

We could only observe resistive switching in this simple capacitor like structure with Cu:TCNQ as active layer with Al top electrodes. All attempts using either thermally evaporated Au or Ag led to a very low resistance state, which could not be switched. An important point to take into consideration is the fact that all electrical measurements were made under ambient conditions. This means, the top electrode is exposed to air prior to the measurements. As Hoagland et al. have published previously, both the inner and outer surface of the thin Al top electrode will oxidize quickly under these conditions [4]. This suggests, that the formation of a thin Al_2O_3 layer at the interface between the top electrode and the Cu:TCNQ layer plays an important role in the resistive switching mechanism.

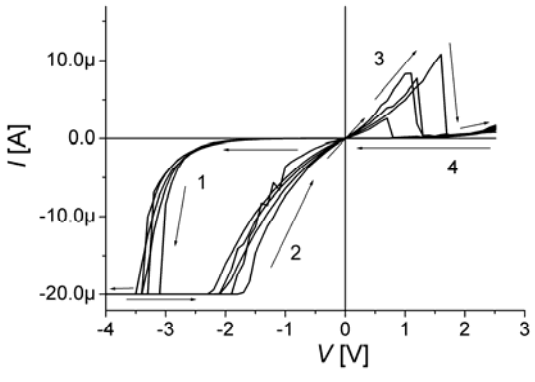


Fig. 3: Current- Voltage characteristic of a memory cell with thermally evaporated Au bottom and Al top electrode. The Cu:TCNQ layer has a thickness of 130 nm.

Bottom electrode	Top electrode	Electrical characteristics
Cu, thermally evaporated	Al, thermally evaporated	R_{OFF} ≈ 5 - 50 MΩ, R_{ON} ≈ 50 - 500 kΩ, $V_{th,offon}$ ≈ -3.5 V, $V_{th,onoff}$ ≈ +3 V, good stability and reproducibility
Cu, thermally evaporated	Au, thermally evaporated	no switching, low resistance values of the memory cells
Cu, thermally evaporated	Ag, thermally evaporated	no switching, low resistance values of the memory cells
Cu, thermally evaporated	Pt, sputtered	all tested samples short due to the sputter process
Pt, sputtered	Al, thermally evaporated	no switching, very high resistance values of the memory cells up to breakdown at high voltages
Au, thermally evaporated	Al, thermally evaporated	R_{OFF} ≈ 1 MΩ-10 MΩ, R_{ON} ≈ 10 - 100 kΩ, $V_{th,offon}$ ≈ -3.5 V, $V_{th,onoff}$ ≈ +1.5 V, poor stability and poor reproducibility
Al, thermally evaporated	Al, thermally evaporated	R_{OFF} ≈ 100 MΩ-1 GΩ, R_{ON} ≈ 1 - 10 MΩ, $V_{th,offon}$ ≈ -4.5 V, $V_{th,onoff}$ ≈ +4 V, poor stability and poor reproducibility

Table 1: The investigated electrode set ups.

The choice of the bottom electrode material is not as crucial as it is with the top electrode. Although memory cells with copper showed best stability and reproducibility, memory cells with Au and Al showed also resistive switching. A possible reason for the better results with Cu bottom electrodes is a favorable interconnection between Cu and Cu:TCNQ compared to other materials.

A current voltage characteristic for a sample with Au bottom electrode is shown in Fig. 3. The conductivity in this case is a bit higher compared to the standard samples with Cu bottom electrodes. Also the switching voltage threshold from on-state to off-state is roughly halved while $V_{th,offon}$ remains more or less unchanged.

Memory cells with Al bottom and top electrode showed also resistive switching behavior with higher resistance values for both, the off- and on-state and increased switching voltage thresholds compared to the standard samples. The observed resistive switching is asymmetrical even so the setup of the memory cell looks symmetrical on paper (Al-Cu:TCNQ-Al). However, this structure is not strictly symmetrical due to different interface between the electrodes and the Cu:TCNQ layer. Also, as described earlier the oxidation of the top electrode is another reason for the asymmetry.

4. Conclusion

The electrical induced resistive switching effect in Cu:TCNQ thin films is strongly dependent on the electrode materials and the resulting interfaces. Samples with Cu bottom and Al top electrodes showed the best resistive switching behavior. The usage of aluminum as top electrode material was in our case mandatory to fabricate working memory cells. In contrast, other materials (Au, Al) used as bottom electrodes resulted also in resistively switching samples.

References

[1] R. S. Potember, T. O. Poehler, D. O. Cowan, Appl. Phys. Lett. **34**, 405 (1979).

[2] T. Kever, C. Nauenheim, U. Böttger, R. Waser, Thin Solid Films **515**, 1893 (2006).

[3] T. Kever, B. Klopstra, U. Böttger, R. Waser, J. Appl. Phys., unpublished.

[4] J. J. Hoagland, X. D. Wang, K. W. Hipps, Chem. Mater. **5**, 54 (1993).

Copper Oxide Resistive Switching for Non-Volatile Memory Applications

Tzu-Ning Fang, Swaroop Kaza, Sameer Haddad, An Chen, Yi-Ching (Jean) Wu, Zhida Lan, Steven Avanzino, Dongxiang Liao, Chakku Gopalan, Sara Mahdavi, Matthew Buynoski, Christie Marrian, Michael VanBuskirk and Masao Taguchi

Advanced Memory Development Group, Spansion Inc.
915 DeGuigne Drive, P.O.Box 3453, MS 177, Sunnyvale, CA 94088-3453, USA
Tel: (408) 616-8620, Fax: (408)616-8401, E-mail: tzu-ning.fang@spansion.com

Abstract

A Metal-Insulator-Metal (MIM) device based on a Cu_2O insulator has demonstrated excellent memory characteristics with a process fully compatible to conventional CMOS technology. In this paper, we report the electrical characteristics of the memory cell. Space-Charge-limited-Conduction (SCLC) was proposed to describe the resistive switching of Cu_2O (MIM) structures. We will also discuss the electrode effect which significantly impacts on the electrical properties of Cu_2O memory cells. A thermal erase model of $Cu/Cu_2O/Ni$ has been proposed and verified with temperature dependency and a power calculation.

1. Introduction

Portable products associated with multimedia applications, Web browsing, video conferencing, 3D and interactive gaming create higher demands on memory requirements, providing new opportunities for emerging memory technologies. Two-terminal resistive devices based upon chalcogenide phase-change materials, perovskite oxides, organic polymers, and metal oxides have drawn a lot of attention recently because of the potential to build a memory combining the nonvolatile feature with fast operating speed. Among these devices, binary metal oxides are particularly attractive due to simpler structures, low current, fast operating speed as well as compatibility with conventional CMOS processing [1-4]. In this paper, we present the electrical characteristics of the Cu_2O resistive switching memory. We also report on the dependence of device characteristics on the top electrode and propose an erase mechanism based on Joule heating.

2. Memory Cell Structure

Fig. 1 (a) shows a sectional view of a 64Kb memory test array with Cu_2O memory elements integrated in a standard CMOS process. The bottom electrode is defined by the Cu via which is electrochemically-deposited (ECD) and planarized by standard CMP. Copper oxide is thermally grown on top of the 0.18 μm Cu via as shown in Fig. 1 (b). The top electrode is then deposited and patterned with a subtractive etch process. NMOS transistors are connected to the Cu vias to select and control the memory cells.

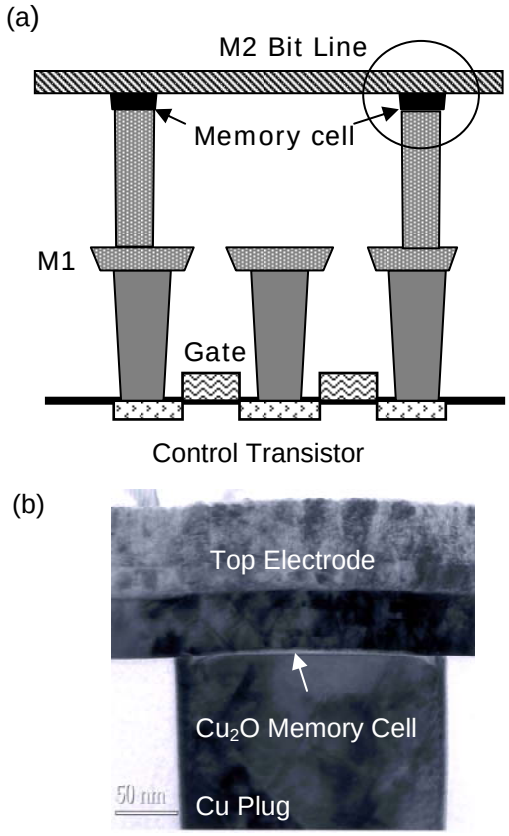
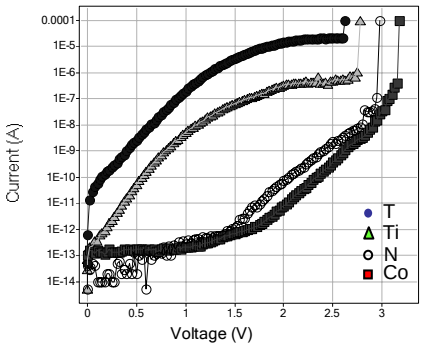


Fig. 1: (a) Schematic of Cu_2O MIM cells in a memory array. (b) Cross sectional TEM image of MIM memory cell. The memory cell is built on top of a 0.18 μm Cu via. Cu_2O thickness is 60-80Å.

3. Electrode Effect

The Cu_2O MIM structures have been shown to have low current leakage due to compensation of multiple trap levels, where the traps could be Cu and O vacancies [3-5]. The intrinsic trap levels would, however, be modified when additional oxygen vacancy defects are generated due to reaction with electrodes. Fig. 2 shows OFF-state current leakage of Cu_2O MIM cells with various top electrodes. Current leakage is seen to increase as a consequence of chemical reaction between the top electrode metal and the Cu_2O. Considering the free energy of formation or stability of the electrode metal oxides, the reactivity increases in the order Ni/Co<Ti< Ta. The low leakage for a Ni or Co top



MO_x	ΔG_f (kcal/mole)
Cu2O	-35
CoO	-51
NiO	-51.7
TiO2	-212
Ta2O5	-457

Fig. 2: IV characteristics of Cu/Cu_2O/TE (top electrode) memory cell with various top electrodes. Positive bias is applied to top electrodes. The table shows heat of formation of the metal oxides involved.

electrode Cu_2O cell indicates minimal reaction at the interface and Cu_2O stoichiometry is maintained.

4. Memory Cell Characteristics

OFF state leakage of the Cu2O memory cell has shown characteristics matching SCLC and Frenkel-Poole (FP) emission [3,5]. As shown in Fig. 3, OFF state leakage has strong temperature dependence and can be fitted to the FP relationship. The switching characteristic of Cu_2O memory cells, with Ni and Ti electrodes, is illustrated in Fig. 4. Programming operation for memory cells with Ni or Ti top electrodes is similar. Switching from a high-resistance state ("OFF") to a low-resistance state ("ON") state occurs at the trap-filled-limit voltage (V_{TFL}) [3,5]. The final ON-state resistance is controlled by the current limit determined by the transistor gate voltage (Vg). Fig. 5 shows ON-current dependence on the Vg in programming for the Cu_2O cells with Ti electrode. The maximum switching current can be as low as 45 μA. Switching at even lower current (<10μA) was also achieved on arrays processed with different memory stacks, indicating potential for a low-power operation.

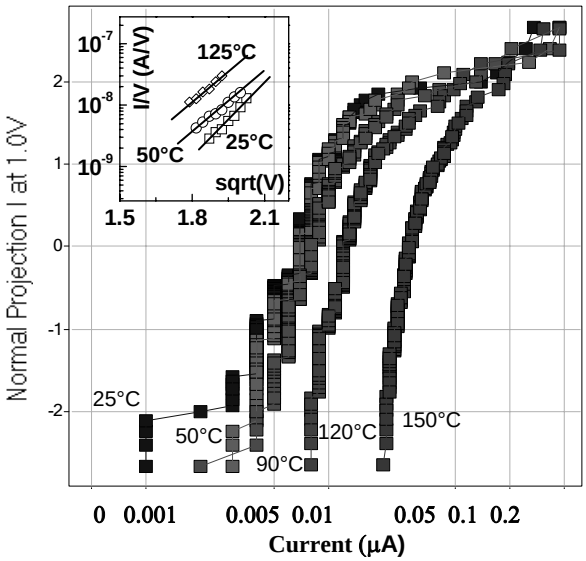


Fig. 3: Distribution of OFF state current at 1V from 25°C to 150°C; the inset shows the Frenkel-Poole component by I-V fitting at 25°C, 50°C and 125°C.

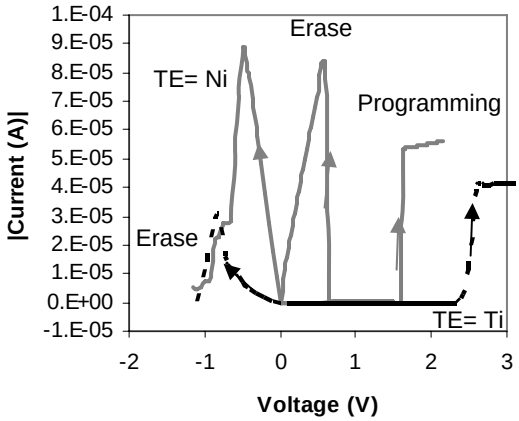


Fig. 4: Switching characteristics for cells with Ni and Ti electrodes. Voltage is applied to the top electrode for programming. Cells with Ti electrodes can only be erased with reversed field, while Ni electrode cells can be erased with either polarity.

Erase characteristics of the MIM cells, in Fig. 4, depend significantly on the top electrode. In the case of a Ni electrode, memory cells require higher current with larger gate voltage to erase. Further, the cells can be erased with either polarity, which implies trap levels are symmetric to both electrodes. By contrast, the Cu/Cu_2O/Ti memory cells require lower current and can only be erased by reverse polarity i.e., opposite to that of programming. This implies asymmetric trap levels due to an interface reaction. The detailed erase model will be proposed and discussed in the next section.

Fig. 6 shows the AC switching characteristics for both program and erase for Cu/Cu_2O/Ni memory cells. The applied waveforms consist of a program/erase pulse followed by a read, to verify the cell state. Both programming and erase operations can be completed within 100ns duration.

ON-state retention time is proportional to $\exp(\Delta E_t/kT)$, where ΔE_t is trap depth [5]. Thus, the thermal dependence of the ON-state resistance has a characteristic temperature of data loss, related to the trap depth. Fig. 7 shows an accelerated temperature test, ON-

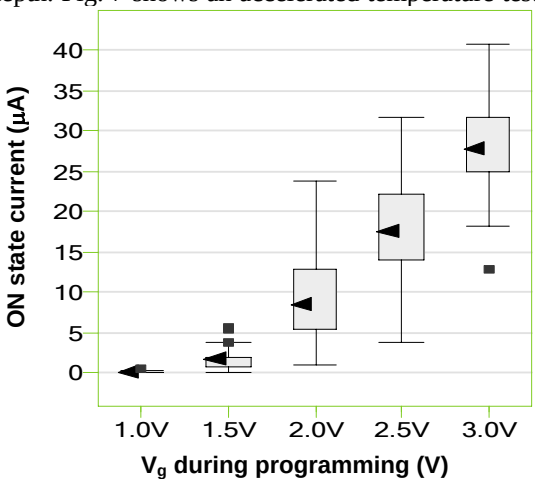


Fig.5 : ON-current as a function of the transistor V_g during programming; the box shows data range and the triangle symbol indicates the mean value of the distribution.

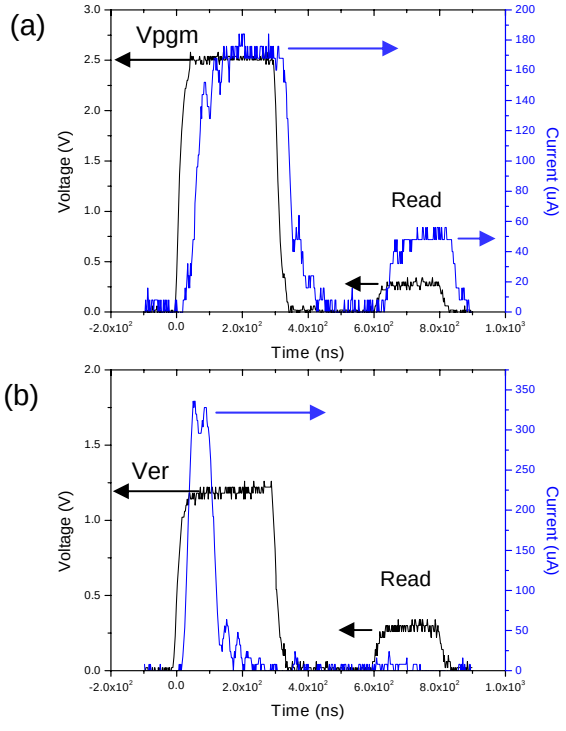


Fig. 6: AC response of (a) Programming (Vg=3V) and (b) erase with 300 ns pulses. Cell ON/OFF state resistances are obtained from the read pulse.

state resistance following annealing at various temperatures for cells with Ni and Ti electrodes. Cells were programmed with a similar current limit of 30-50μA. A significant portion of the Ti top electrode cells retain their ON-state to at least 250°C, while the cells with Ni electrode lose their state at <150°C. This indicates the reaction of the Ti electrode with Cu_2O modifies the intrinsic Cu_2O trap levels towards the deeper levels.

5. Joule Heating Effect and Thermal Erase Model

The ON-state temperature dependence of $Cu/Cu_2O/Ni$ memory cells is strongly correlated to the programming conditions as shown in Fig. 8. The critical temperature for an ON-OFF transition increases as the programming current limit increases. A similar correlation exists in the erase operation. The IV curve is shown in Fig. 9(a), while Fig. 9(b) shows the calculated power during erase. The maximum power occurs at the erase point and the value of this erase power increases with the programming current. Hence the joule heating temperature at the erase point increases with programming current. This is the same trend seen with the erase temperature in Fig. 8. This strong correlation suggests Joule heating is the dominant effect for the erase event in $Cu/Cu_2O/Ni$ cells.

Fig. 9(b) shows that the power through the cell reaches a maximum at the erase point and then drops dramatically. This "quench" effect corresponds to the ON-OFF transition in Fig. 9(a), which is followed by a significant decrease in current through the cell. The cells erase to an intermediate OFF-state (~100nA) compared with fresh cell state (~1pA). This intermediate OFF-state

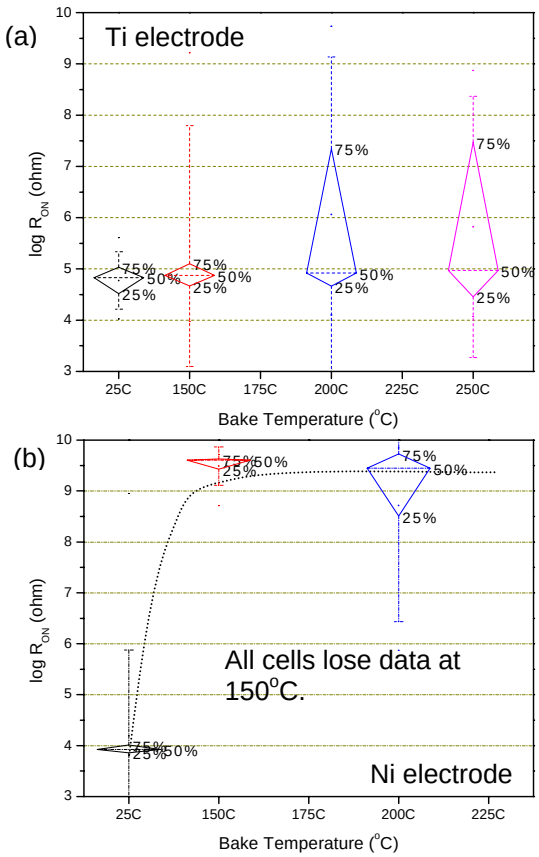


Fig. 7: Temperature dependence of ON-state resistance (a) $Cu/Cu_2O/Ti$ and (b) $Cu/Cu_2O/Ni$ memory cells. Ni top electrode cells lose data at ~150°C, while Ti electrode cells retain up to 250°C. Cells are programmed with a current limit of (a) 30μA and (b) 50 μA.

is also evident during ambient temperature erase, shown in Fig 8. Further increase in ambient temperature causes the cells to be erased to the fresh cell state.

The model was further investigated by comparing the "erase" power with the "end of programming" power for several ON-states. Erase power is calculated at the erase point, while "end of programming" power is calculated using R_{on} from read pulse. Fig. 10 consists of calculated "erase" and "end-of-program" power with

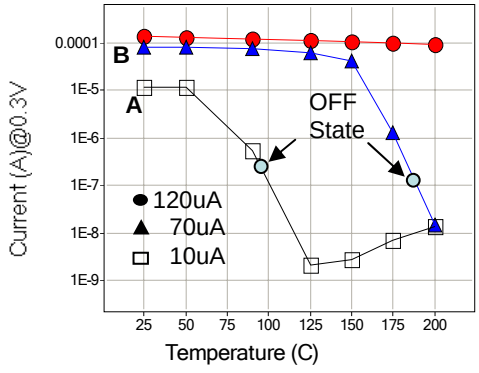


Fig. 8: Temperature dependence of $Cu/Cu_2O/Ni$ cells programmed with different current limits. Current drop indicates ON-OFF (erase) transition. Higher program current limit requires higher erase transition

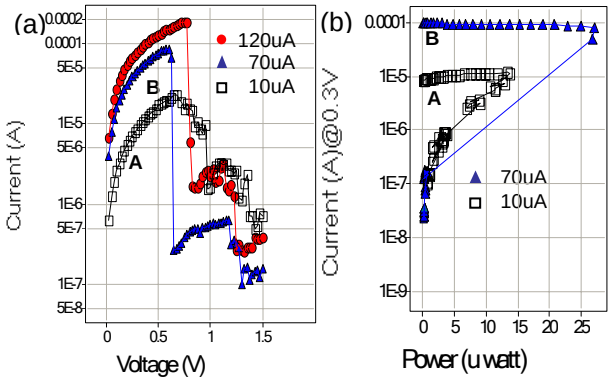


Fig. 9: (a) Erase IV curves of $Cu/Cu_2O/Ni$ cells programmed with different current limits. (b) Joule heating power corresponding to the erase curves A and B in (a).

respect to the memory cell ON-state and shows erase power consistently increases with programming Vg. Voltage across the cell during erase is plotted in Fig. 11 and shows no such dependence on Vg. This suggests thermal assisted de-trapping to be the dominant erase process. Also, power at "end of programming" approaches, but never crosses-over the erase power-R_{on} curve. This agrees with a thermal erase model; no ON-state exists when the "end-of-programming" power is high enough to initiate thermal de-trapping. Thus, R_{on} has an upper bound in the erase power trend line and gives a predictable cell R_{on} range for a given transistor. Further evidence for the thermal erase model in $Cu/Cu_2O/Ni$ cells can be seen in Fig. 12, which shows a decrease in erase power with an increase in ambient temperature of erase [7].

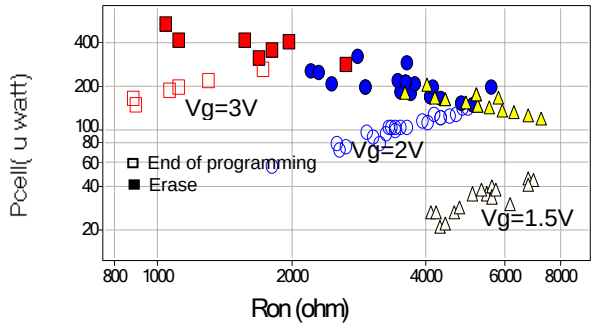


Fig. 10: Erase power and "end-of-programming" power versus the cell ON- state resistance for $Cu/Cu_2O/Ni$ cells. Cell ON-state resistance is determined by Vg.

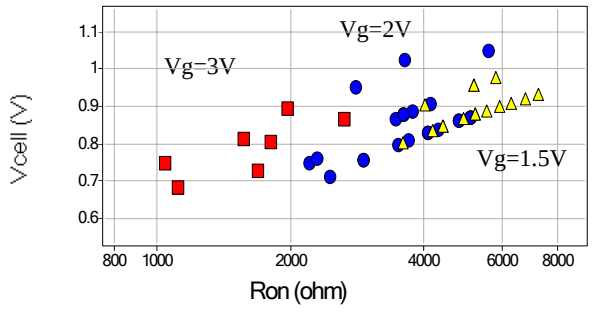


Fig. 11: Voltage across the cell during erase versus R_{on} (ON-state resistance).

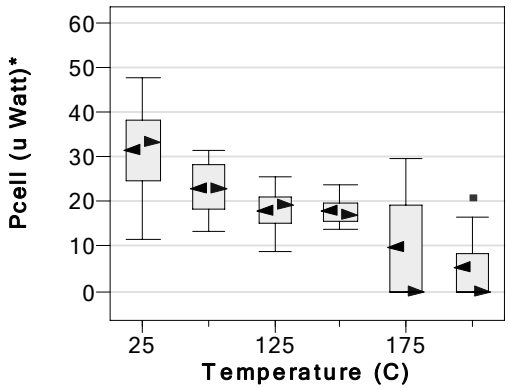


Fig. 12: Erase power dependence on ambient temperature. All cells initially programmed at room temperature. "0" power indicates cells erased by temperature alone. The box shows the data range and the triangle indicates the mean and median.

The thermal erase model does not, however, apply to $Cu/Cu_2O/Ti$ MIM erase characteristics, shown in Fig. 4. Power calculations show "erase" power to be less than the power at "end of programming", suggesting that Joule heating is not the dominant effect in the erase operation [4].

6. Summary

Device characteristics of Cu_2O MIM resistive memory structures show a strong dependence on the top electrode. Cells with Ni electrodes exhibit erase in the same field direction as programming which enables ease of operation and integration. Power calculations during erase and temperature dependence of erase imply a joule heating model for the erase mechanism. Low currents, a practical ON/OFF resistance range, high operation speed and low operation temperature demonstrate suitability for future memory applications.

References

(1). A. Beck, J. G. Bednorz, Ch. Gerber, C. Rossel, and D. Widmer, "Reproducible switching effect in thin oxide films for memory applications", *Appl. Phys. Lett.* Vol. 77(1), p.139, 2000.

(2). I.G. Baek, M.S. Lee, S. Seo, M.J. Lee, D.H. Seo, S. Suh, J.C. Park, S.O. Park, H.S. Kim, I.K. Yoo, U-In Chung, and J.T. Moon, "Highly scalable non-volatile resistive memory using simple binary oxide driven by asymmetric unipolar voltage pulses", *IEDM Tech. Dig.* p.587, 2004.

(3). An, Chen, S. Haddad, Y.C. Wu, T.-N. Fang, Z. Lan, S. Avanzino, S. Pangrle, M. Buynoski, M. Rathor, W. Cai, N. Tripsas, C. Bill, M. Vanbuskirk, and M. Taguchi, " Non-volatile resistive switching for advanced memory", *IEDM Tech. Dig.* p.746, 2005.

(4). T.-N Fang, S. Kaza, S. Haddad, A. Chen, Y.-C. Wu, Z. Lan, S. Avanzino, D. Liao, C. Gopalan, S. Choi, S. Mahdavi, M. Buynoski, C. Bill, M. Vanbuskirk, and M. Taguchi, " Erase Mechanism for Copper Oxide Resistive Switching Memory Cells with Nickel Electrode", *IEDM Tech. Dig.* p.789, 2006.

(5). A.E. Rakhshani, "Thermostimulated impurity conduction in characterization of electrodeposited Cu_2O films", *J. Appl. Phys.* Vol. 69(4), p.2365, 1991. A.E. Rakhshani, "The role of space-charge-limited-current conduction in evaluation of the electrical properties of thin Cu_2O films", *J. Appl. Phys.* Vol. 69, p. 2365, 1991.

(6). M.A. Lampert and P. Mark, "Current injection in solids", Academic press, 1970.

(7). R. Scheuerer, K.F. Renk, E. Schomburg, and W. Wegscheider, " Nonlinear superlattice transport limited by Joule heating", *J. Appl. Phys.* Vol. 92, p. 6043, 2002.

Resistive switching and microstructure of NiO binary oxide films developed for OxRRAM non-volatile memories

L. Courtade[a], Ch. Turquat[a], Ch. Muller[a], D. Goguenheim[b], J.G. Lisoni[c], L. Goux[c], and D.J. Wouters[c]

[a] L2MP, Laboratoire Matériaux et Microélectronique de Provence, UMR CNRS 6137, Université du Sud Toulon Var, BP 20132, F-83957 La Garde, France; [b] L2MP, ISEN-Toulon, Place Georges Pompidou, F-83000 Toulon, France
[c] IMEC, Interuniversity MicroElectronics Center, Kapeldreef 75, B-3001 Leuven, Belgium

Abstract

Oxide Resistive Random Access Memories (OxRRAM) are discussed for future high density non-volatile memory chips. NiO and other simple binary transition metal oxides have recently attracted lots of attention for their resistive switching behavior. In most cases, polycrystalline oxide films are deposited by reactive sputtering on conductive substrates to form bi-stable Metal/Resistive oxide/Metal (MRM) structures. In this paper, an alternative way is explored to obtain NiO films from the controlled oxidation of a Ni metallic film. Different oxidizing conditions were evaluated to prevent the complete consumption of the Ni film used as bottom electrode. Electrical and microstructural analyzes were performed to apprehend the influence of the process parameters on the switching behavior.

1. Introduction

Bi-stable resistive switching phenomena controlled by external currents or voltages attract a lot of attention for future high-density non-volatile memory devices. The proposed resistive memory materials range from organic materials, e.g. Cu-TCNQ,[1] to inorganic, *e.g.* chalcogenide alloys, perovskite-type oxides, or transition metal oxides.[2–5] Typical current-voltage I(V) characteristics of Metal/Resistive oxide/Metal (MRM) structures exhibit a drastic change in resistance between a high resistance state (*i.e.* OFF state) and a low resistance state (*i.e.* ON state).

Among simple transition metal oxides, nickel oxide, NiO, is a promising candidate for non-volatile memory devices due to its compatibility with standard CMOS (Complementary Metal-Oxide-Semiconductor) process.[2, 6] Resistance switching in a NiO crystalline film was observed in 1964 by Gibbons *et al.*[2] and the switching mechanism was explained by the reversible formation/rupture of filamentary conductive paths at the interfaces between metallic electrodes and NiO film.[2,7,8]

In most cases, NiO films are deposited by dc reactive sputtering on conductive substrates to fabricate MRM structures.[8,9] In this paper, an alternative way is explored: NiO-based MRM structures were produced from the oxidation in an Rapid Thermal Annealing (RTA) furnace of a blanket Ni metallic film used as bottom electrode. This approach is similar to the one developed by Chen *et al.* to form Cu/CuO_x bi-stable stacks.[10] Several process parameters, such as oxidation time, temperature and oxygen partial pressure, were monitored to control the nickel oxidation and to produce bi-stable NiO films. To complement electrical testing, the MRM structures were investigated by x-ray diffraction mainly to apprehend the Ni oxidation kinetics and transmission electron microscopy (TEM) to observe the stack microstructure and film interfaces.

2. Switching behavior

2.1. Switching characteristic

Current-Voltage characteristics were measured with AixACCT TF Analyzer 2000 system using either triangular waveform or staircase voltage ramp. A typical I(V) characteristic of a Pt/NiO/Ni structure is shown in Fig. 1. The measurements were performed on 100 nm thick Ni layer oxidized in pure O_2 at 400°C for 30 seconds. The high current in bias ranging from –5 to 4.2 V indicates the ON state. When the bias voltage reached 4.2 V, the current suddenly decreases and the structure irreversibly switches to the high resistance state (OFF). Additionally, low (R_{ON}) and high (R_{OFF}) resistances were deduced from the slope of I(V) characteristics.

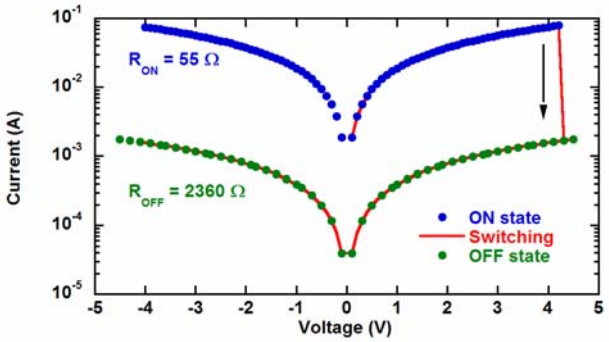


Fig. 1: Typical I(V) switching characteristic.

2.2. Influence of temperature and annealing time

Systematic I(V) measurements were performed on samples coated with a 24 or 100 nm Ni layer, oxidized in pure O_2 at different temperatures (200, 300 and 400°C), with RTA times ranging from 10 seconds to 5 minutes. Fig. 2a shows the evolution of the threshold voltage V_{th} as a function of annealing times for various temperatures. For samples with an initial 100 nm thick Ni layer, the largest V_{th} voltages were obtained with an annealing at 400°C. In contrast, no switching was observed for samples annealed at 200°C. For samples with initial 24 nm thick Ni layer, V_{th} follows the same tendency. It has to be noted that no switching was observed for Ni films annealed at 400°C for times above 30 seconds. Besides, V_{th} increases up to 20% along with annealing time.

The R_{OFF}/R_{ON} ratio remains quite constant whatever the annealing time: ≈ 50 for samples with 100 nm Ni layer and ≈ 15 for samples with 24 nm Ni layer (Fig. 2b).

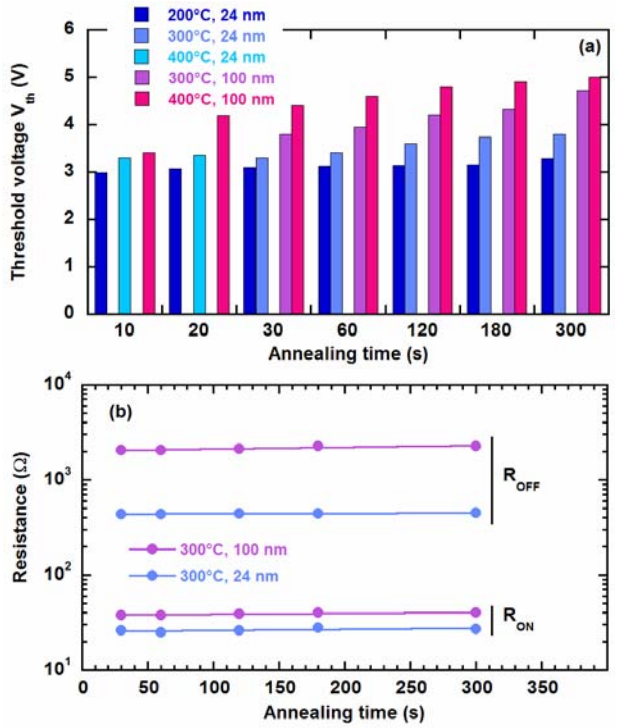


Fig. 2: (a) Evolution *versus* annealing time of the threshold voltage (V_{th}) for Pt/NiO/Ni MRM structures. (b) Annealing time-dependent variation of low (R_{ON}) and high (R_{OFF}) resistances of 24 and 100 nm thick Ni layers oxidized at 300°C in pure O_2.

2.3. Influence of oxidizing atmosphere

I(V) characteristics were measured on samples coated with an initial Ni layer of 24 or 100 nm thickness and annealed under different oxidizing atmospheres (20 and 500 ppm of O_2; pure O_2) using RTA at 200 and 400°C for 30 seconds and 3 minutes. The main trend is the augmentation of V_{th} with increasing oxygen partial pressure (Fig. 3). V_{th} varies from 2.2 V for a 24 nm thick Ni layer annealed at 200°C for 30 seconds under 20 ppm of O_2 to 5 V for a 100 nm thick Ni layer annealed at 400°C for 180 seconds in pure oxygen. It is worth to note that in few cases no switching was observed (*e.g.* 24 nm thick Ni layer with annealing time larger than 180 seconds whatever the oxygen partial pressure). Once again, the R_{OFF}/R_{ON} ratio remains unchanged whatever the oxidizing atmosphere. Hence, Fig. 3 helps to apprehend the experimental conditions required to form bi-stable NiO films with switching voltages compatible with memory applications.

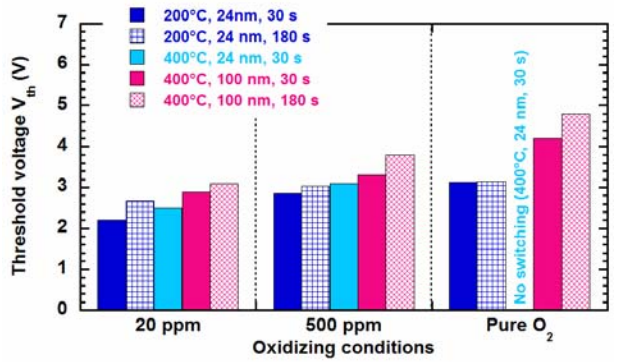


Fig. 3: Distribution of threshold voltages (V_{th}) for Pt/NiO/Ni MRM structures produced under different oxidizing atmospheres (RTA at 200 and 400°C for 30 and 180 seconds).

3. Microstructure of bi-stable NiO films

3.1. Microstructure of stacks Pt/NiO/Ni/SiO_2

TEM experiments were performed on sample coated with a 100 nm thick Ni layer, annealed in pure O_2 at 400°C for 30 seconds and covered with a Pt top electrode. The observation of a cross-section in imaging mode shows a stack of several layers with different contrasts (Fig. 4). As shown in Fig. 4, there is a large spread of NiO thickness typically ranging from 40 to 75 nm. Micro-diffraction experiments (not shown here) have suggested that the Ni layer is preferentially oriented with the [111] crystallographic direction parallel to the substrate's normal.

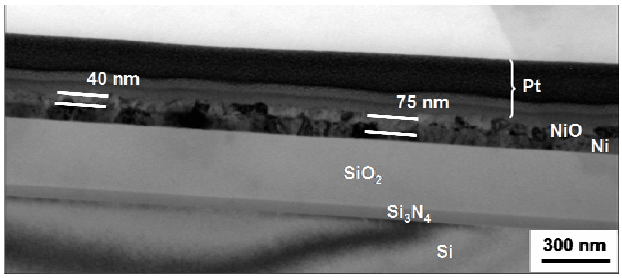


Fig. 4: TEM cross section of stack Pt/NiO/Ni/SiO_2/Si_3N_4/Si.

3.2. Texture of Ni and NiO films

X-ray texture analyzes have confirmed the electron micro-diffraction results. Indeed, the strong [111] texture of the Ni substrate (Fig. 5a) since the maximum intensity of the $(111)_{Ni}$ Bragg reflection is at the center of the pole figure. On the other hand, the $\{200\}_{Ni}$ pole figure (Fig. 5b) shows a random orientation of [200] directions around the substrate's normal. Thus, these analyzes confirm the strong [111] fibre texture of the Ni layer.

A similar analysis was performed on the NiO film after oxidation of a 100 nm thick Ni layer at 400°C for 120 seconds in pure O_2. The $\{111\}_{NiO}$ pole figure clearly indicates a texture of the NiO layer along [111] direction (Fig. 5c). Consequently, the crystallographic orientation of NiO grains appears to be conditioned by the texture of the underlying Ni film since the NiO film preferentially grows with [111] direction parallel to the substrate's normal.

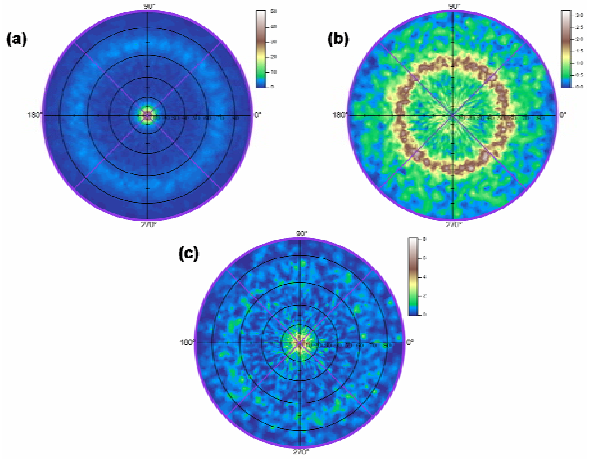


Fig. 5: (a) $\{111\}_{Ni}$ and (b) $\{200\}_{Ni}$ pole figures of a non-oxidized 100 nm thick Ni layer. (c) $\{111\}_{NiO}$ pole figure of NiO film obtained from oxidation of a 100 nm thick Ni layer (400°C, 120 seconds, pure O_2).

3.3. Oxidation kinetics of Ni layer

Ex situ x-ray diffraction was used to check the presence of both Ni and NiO phases and to apprehend the progressive oxidation of the nickel layer. These analyzes have shown that the $(111)_{NiO}$ reflection is clearly observed for an initial 100 nm thick Ni layer oxidized at 300 and 400°C in pure oxygen. The absence of NiO phase at 200°C indicates that Ni film is not oxidized. In contrast, for an initial 24 nm thick Ni film, the $(111)_{Ni}$ diffraction peak is totally absent at 400°C with annealing times larger than 20 seconds. Hence, in this latter case, the Ni film is fully consumed, the subsequent absence of bottom electrode impeding electrical characterization of NiO films. On the contrary, both Ni and NiO phases were detected for lower temperatures (*i.e.* 200 and 300°C).

A detailed analysis was performed on x-ray diffraction patterns collected on a 100 nm thick Ni film oxidized at 400°C in pure O_2, with RTA times ranging from 10 seconds to 30 minutes. In Fig. 6, concomitant diffracted intensity variations of both Ni (decrease) and NiO (increase) phases are observed with increasing oxidation times. These features indicate that the Ni metallic film is progressively consumed as the NiO film grows.

Second, the whole oxidation kinetics of an initial 100 nm thick Ni layer was followed thanks to *in situ* time-dependent x-ray diffraction experiments performed at isotherms 400 and 500°C in pure O_2. Before the oxidation step, the Ni metallic films were first heated under vacuum and the temperature-dependent evolution of the $(111)_{Ni}$ Bragg reflection measured during heating indicated a further crystallization of Ni layer.[11] Afterward, the time-dependent x-ray diffraction patterns were collected in pure O_2 at constant temperature for several hours. In agreement with the evolution obtained from *ex situ* x-ray diffraction experiments (Fig. 6), concomitant intensity variations of both Ni and NiO phases were observed with increasing oxidation times (Fig. 7 shows the time-dependent evolution at 400°C). This evolution indicates that the Ni film is progressively consumed as the NiO film grows up to the total consumption of the Ni layer. The same kind of evolution was observed at 500°C with faster oxidation kinetics.

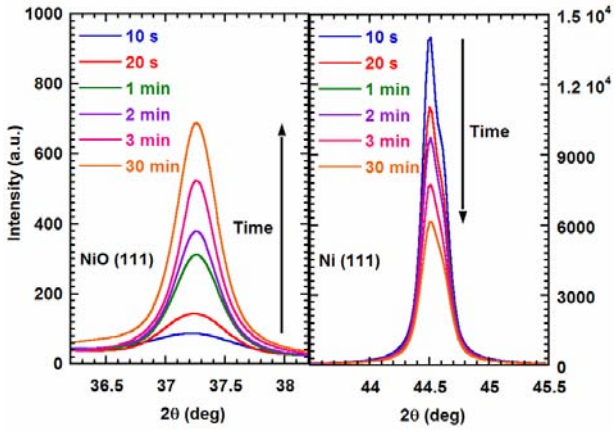


Fig. 6: Evolution *versus* oxidation time of $(111)_{NiO}$ (on left) and $(111)_{Ni}$ (on right) Bragg reflections (initial 100 nm Ni layer annealed at 400°C in pure oxygen).

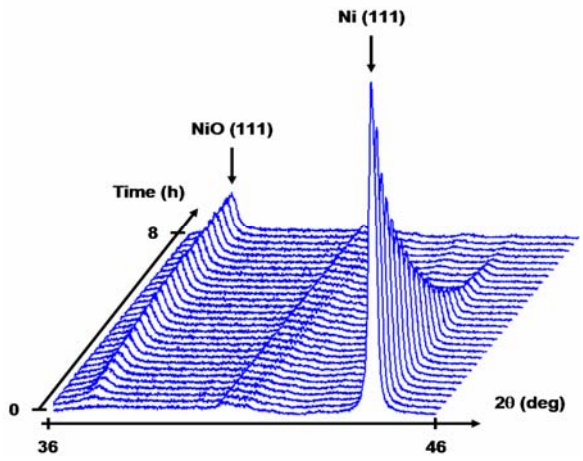


Fig. 7: Evolution *versus* oxidation time of $(111)_{Ni}$ and $(111)_{NiO}$ Bragg reflections during Ni oxidation at 400°C in pure O_2.

These *in situ* experiments enable apprehending the isothermal oxidation kinetics of Ni layer and the conditions necessary to avoid complete consumption of Ni layer used as bottom electrode. Besides, the growth kinetics of NiO phase was described using a nucleation/growth Avrami-type model.[11]

4. Discussion

In this section, both microstructural and electrical characteristics of Pt/NiO/Ni MRM structures are discussed in relation with the experimental conditions.

First, for samples with a 100 nm Ni layer, an annealing at 300°C or above in pure oxygen is required to at least partially oxidize the metallic film and to fabricate NiO-based switching structures. This result was confirmed by *ex situ* x-ray diffraction showing that Ni films are not oxidized at 200°C (with subsequent absence of switching behavior). On the other hand, for samples with a 24 nm Ni layer, the same experimental conditions result in rather different electrical behavior. At 400°C, whatever the oxygen partial pressure, the annealing time has to be limited to 30 seconds to prevent the total oxidation of Ni layer (checked by x-ray diffraction) with subsequent consumption of bottom electrode. For lower temperatures (*i.e.* 200 and 300°C), the Ni layer is only partially oxidized whatever the annealing time and oxygen partial pressure. Furthermore, the threshold voltage may be tailored in monitoring time and O_2 partial pressure.

Electrical testing presented in section 2 has revealed major differences in the switching behavior which may be linked to the Ni microstructure before oxidation step. Indeed, it might be surprising that a resistive switching occurs on a 24 nm Ni layer oxidized at 200°C whereas it is not observed on a 100 nm Ni film annealed in the same conditions. This peculiar behavior may find its origin in the thickness-dependent microstructure of the Ni layer which influences the oxidation kinetics.[11] X-ray diffraction profiles of $(111)_{Ni}$ reflection showed that the initial 100 nm Ni layer presents a stronger texture along [111] direction as compared to the 24 nm Ni layer. Moreover, the 24 nm Ni layer presents much smaller crystallites (≈ 10 nm) as compared to those of the 100 nm Ni layer (≈ 40 nm). The NiO film growth

depends on both texture and crystallite size of the initial Ni metallic layer. Consequently, the non-oxidation of the 100 nm thick Ni film at 200°C (with subsequent absence of switching) may be explained (i) by a stronger $[111]_{Ni}$ texture that increases the resistance to oxidation and (ii) by larger crystallites that decrease the amount of grain boundaries and limit the oxygen diffusion.

Besides, the comparison of time-dependent growth of NiO crystallites has confirmed once more the influence of Ni microstructure on the NiO growth.[11] When the Ni metallic film is annealed under vacuum prior to oxidation, a further crystallization occurs with an enhancement of $[111]_{Ni}$ texture and a growth of Ni crystallites. In contrast, the RTA conditions lead to a rather different situation with two simultaneous competitive mechanisms: the Ni crystallization with subsequent changes in the microstructure and the growth of NiO film with progressive consumption of the Ni layer. Thus, the different "history" of the Ni layer before oxidation may explain the faster kinetics observed in RTA conditions when the Ni layer is directly subjected to the oxidizing atmosphere without special pre-treatment. Thus, in agreement with previous works,[12-15] Ni surface morphology and its thermal treatment prior to oxidation radically change the NiO film growth kinetics.

As compared to sputtered films reported in literature, NiO films obtained from Ni oxidation are initially in ON state without special electro-forming generally required to reach conductive state. To explain such a behavior, it may be proposed that the present conditions lead to the formation of Ni-excess oxide films despite quite long annealing times in pure O_2. Indeed, varying oxygen content in sputtering gas mixture, Park *et al.* have shown drastic modifications of electrical properties of Ni_xO films from a metallic behavior (Ni-excess films) at low oxygen content (< 5%) to a monostable threshold switching (Ni-deficient films) at high oxygen content (> 20%), the memory switching region being limited to intermediate oxygen contents.[16] However, the switching from ON to OFF states indicates an intermediate situation between metallic and oxidic behaviors.[16]

Another unexpected feature is the irreversible switching, the MRM structures remaining in OFF state after the first switching. This irreversibility may be certainly linked to the roughness of NiO/Ni and Pt/NiO interfaces. Indeed, TEM observations have revealed that the Ni metal surface is covered by a continuous but irregular oxide layer (*cf.* Fig. 4). This latter characteristic may have a crucial role on the reversibility of switching since several authors have attributed switching to the rupture of filamentary conductive paths near the interface.[2,7] During oxidation, the growth of the NiO oxide film is rather non-uniform as a consequence of different growth rates for different Ni grain orientations.[17,18] Besides, in previous works, Haugsrud has mentioned the existence of microfissures within the NiO oxide obtained from Ni oxidation at high temperatures.[15,19] Based on these results, one may envisage the existence of such microfissures within the NiO film explaining the difficulty to re-form the filamentary conductive paths. Such microfissures may also lead to a partial delamination at the interface between NiO and Ni films.

5. Conclusion

Resistive switching phenomena have been demonstrated in Pt/NiO/Ni structures with NiO films obtained from Ni metallic layer oxidation. Various process parameters of Rapid Thermal Annealing route were tested to achieve oxidation. Thermal treatments were selected to oxidize the metallic film with conditions (i) preventing the complete consumption of Ni film used as bottom electrode and (ii) producing bi-stable oxide films. In these experimental conditions, the as-grown NiO films were initially in the low resistance ON state without special electro-forming usually required. The fabrication of Ni-excess oxide films may explain the initial metallic behavior. Besides, above the threshold voltage, MRM structures irreversibly switched into a high resistance OFF state. This feature may be linked to the roughness of NiO/Ni and Pt/NiO interfaces due to non-stabilized Ni film microstructure prior to oxidation.

In the perspective of memory devices, further microstructural and electrical analyzes are required to decrease the threshold voltage and to improve the reversibility and the cyclability of the memory element. Currently, new conditions are evaluated in order to control Ni crystallization prior to oxidation and an alternative route for oxidation is explored (plasma treatment under N_2O or O_2). Finally, fully integrated bi-stable NiO/Ni stacks in via structures will be fabricated and the stability of these structures will be checked in conditions close to those used in the back-end process.

References

[1] R. Müller *et al.*, Solid-State Electronics 50(4), 601 (2006).
[2] J. F. Gibbons *et al.*, Solid State Electronics 7, 785 (1964).
[3] B. J. Choi *et al.*, J. Appl. Phys. 98(3), 033715 (2005).
[4] C. Rohde *et al.*, Appl. Phys. Lett. 86(26), 262907 (2005).
[5] L. Courtade *et al.*, IEEE Proceedings of Non Volatile Memory Technology Symposium, p. 94 (2006).
[6] S. Seo *et al.*, Appl. Phys. Lett. 85(23), 5655 (2004).
[7] I. H. Inoue *et al.*, IEEE Proceedings of the Non Volatile Memory Technology Symposium, p. 131 (2005).
[8] D. C. Kim *et al.*, Appl. Phys. Lett. 88(20), 202102 (2006).
[9] G. Baek *et al.*, Int. Electron Devices Meeting Tech. Dig., p. 587 (2004).
[10] An Chen *et al*, International Electron Devices Meeting Tech. Dig., p. 746 (2005).
[11] L. Courtade *et al.*, J. Appl. Phys., submitted.
[12] M. J. Graham *et al.*, J. Electrochem. Soc. 119, 879 (1972).
[13] M. J. Graham *et al.*, J. Electrochem. Soc. 119, 1523 (1973).
[14] L. Berry *et al.*, Mem. Sci. Rev. Metall. LXV, 9, 651 (1968).
[15] R. Haugsrud, Corrosion Science 45, 211 (2003).
[16] Jae-Wan Park *et al.*, J. Vac. Sci. Technol. B 24(5), 2205 (2006).
[17] R. Peraldi *et al.*, Oxidation of Metals 58, 275 (2002).
[18] R. Peraldi *et al.*, Materials at High Temperatures 20, 649 (2003).
[19] R. Haugsrud, Corrosion Science 45, 1289 (2003).

Switching between two high-resistive states in Cu/chalcogenide/W structures for application in non-volatile memories

Ludovic Goux[1], Judit G. Lisoni[1], Thomas Gille[1,2], Kristin De Meyer[2], Karen Attenborough[3], and Dirk J. Wouters[1]

[1]IMEC, Silicon Process and Device Technology Division, Kapeldreef 75, B-3001 Leuven, Belgium,
[2]ESAT/INSYS, KU Leuven, Kasteelpark Arenberg 10, Leuven, 3001, Belgium
[3]Research, NXP Semiconductors, Kapeldreef 75, 3001 Leuven, Belgium
ludovic.goux@imec.be

Abstract

This work shows promising electrical switching properties of a Cu/chalcogenide/W stack for non-volatile memory applications requiring low-power operation. Simple test-cells are fabricated by sputtering a doped-SbTe chalcogenide glass onto W bottom electrode followed by top electrode formation by sputtering of Cu dots. Electrical results indicate that the cells switch reproducibly between two high-resistive states and suggest that switching takes place at the Cu-chalcogenide interface, where a high resistive layer can be formed using a low writing voltage of 0.4 V. While switching current on large devices are high, extrapolation of resistances in both states to smaller device dimensions may lead to the decrease of the writing current down to the μA range, or below, while maintaining a resistance ratio around one decade. This is to our knowledge the first report showing electrical switching of this type in a Cu/chalcogenide-based structure.

1. Introduction

Today Flash technology dominates the market of non-volatile memories, but it is expected that this technology will face severe scaling problems beyond the 32 nm technology node due to fundamental physical limitations [1]. This situation has favored open competition involving several emerging technologies which hold the promise to be more scalable. In this respect resistive-switching memories, based on the voltage- or current-induced change of resistance of the active material, are among the most promising technologies. For compatibility with high density levels, the resistive-switching material should not only exhibit a sufficient and controllable memory window with scaling, but it should show low-current switching as well as low-power operation in particular. In this context, the programmable metallization cell (PMC) memory technology is promising because recently it has been demonstrated that the cell is switchable using only a voltage of ~0.2 V and a switching current down to 10 μA [2]. The switching mechanism of the cell is the electrolytic formation and rupture of a conductive filament within a glass, usually a chalcogenide glass, which promises scalability provided that the density of filaments can be controlled.

In the present work, we explore the switching properties of a structure composed of a doped-SbTe chalcogenide glass sandwiched between the metals Cu and W. We characterized the electrical switching properties of large test-cells between two high-resistive states and we discuss the possible switching mechanism. Based on that, we show that the extrapolation of the reproducible switching of the cells might lead to low-power operation for small devices.

2. Experimental

The test-cells were realized as follows. The W bottom electrode was chemical-vapor deposited and planarized on SiO_2 coated Si substrates as in standard CMOS integration process flows. Then, amorphous doped-SbTe chalcogenide layer was deposited by sputtering at room temperature. The thickness of the layer was varied between 20 and 500 nm. Finally, Cu electrodes of various sizes down to a diameter of 150 μm were sputtered through a shadow mask, also at room temperature. For additional reference, samples were made with (sputtered) Pt top electrode dots instead of Cu. Note that after cell fabrication the chalcogenide layer is still in the amorphous state.

Switching was tested by measuring current-voltage I(V) characteristics using a conventional setup HP4156. Voltage was swept to the Cu electrode in the range -1 to 1 V using voltage steps of 0.1 V and a delay time of 50 ms between each step.

3. Basic electrical properties

Figure 1 shows a typical I(V) characteristics of a Cu/chalcogenide/W cell with a chalcogenide layer thickness of 500 nm. Initially, the cell is in a very high resistive state, called OFF_{init} hereafter (see Figure 1). However the cell switches to a less resistive state (OFF) after applying a voltage of around +0.8 V.

Once the cell is in the OFF state, it can be switched reversibly at low voltage between two different states called ON and OFF hereafter. The switching from ON to OFF occurs for a voltage of +0.2 V applied to the Cu electrode, and the reverse switching from OFF to ON occurs for a voltage of -0.4 V (see Figure 1). This reverse switching was also observed for V ~ +1 V, which indicates also possible unipolar memory functionality, however this unipolar switching proved to be less reproducible and less uniform than the bipolar switching shown in Figure 1.

After switching to a given state, the reading of the state by sweeping up to 0.1 V in amplitude confirmed the non-volatile nature of the switching. In this low reading-voltage range, a resistance ratio of ~5 was extracted from the traces in Figure 1. For all samples, the resistances of both states together with the switching voltages were observed to be very stable over several tens of cycles (see the few cycle traces shown in Figure 1).

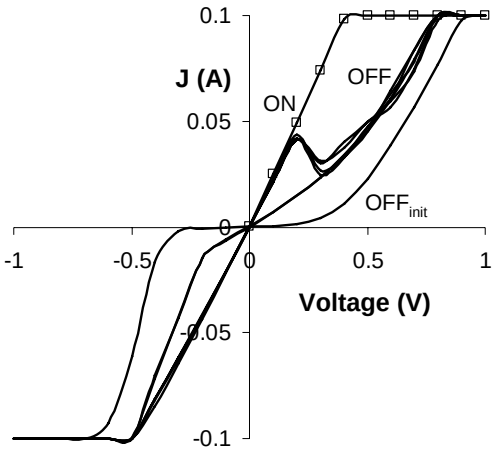


Fig. 1: Typical I(V) characteristic of Cu/doped-SbTe/W test-cell (voltage is applied to the Cu top electrode), together with the I(V) trace obtained for a Pt/doped-SbTe/W cell (open squares); capacitor area is 0.85 mm^2 for both cells.

Figure 1 also shows the trace that is obtained if Pt top electrode dots are sputtered instead of Cu. The I(V) trace follows the same trace as for the Cu/chalcogenide/W cell in the ON state, but does not switch to higher resistive state. This indicates that Cu is needed for the cell to have the property to switch, probably in relation with a possible interaction between Cu and the chalcogenide layer. As for the Pt/chalcogenide/W cell, the linear voltage dependence of the current indicates that the Pt/chalcogenide/W cell behaves as a resistor. However, if we assume that only the amorphous chalcogenide layer contributes to this resistance-like trace the extracted resistivity is of the order of 1000 Ω.cm, which is at least two orders of magnitude higher than the resistivity of the chalcogenide material in the amorphous state. This suggests the existence of a high-resistive interfacial layer in series with the chalcogenide layer both for Pt/chalcogenide/W and Cu/chalcogenide/W cells. In the latter case, the switching to the OFF state indicates that a voltage-induced change of resistance (toward higher resistance values) takes place, presumably at the Cu/chalcogenide interface. Similarly to the Pt/chalcogenide/W cell, the linear voltage dependence of the current in the ON state points to a resistance behavior of the Cu/chalcogenide/W cell. However, the dependence is clearly non-linear both in the OFF and OFF$_{init}$ states, which might suggest that the series interfacial resistance is voltage dependent, as it is for a space-charge layer.

Among possible switching mechanisms, Cu may be partly diffused in the chalcogenide layer, and may be alternatively either oxidized when applying +0.2 V or reduced when applying -0.4 V to the Cu electrode. This mechanism is similar to the voltage-controlled formation and rupture of Ag-based electrodeposit in PMC cells as described by Kozicki [2]. The difference is that in our cell there is no formation of Cu electrodeposit in the chalcogenide glass, because this would lead to a more conductive state than the resistive state of the chalcogenide glass. In our cell, the hypothetical Cu reduction would take place on the electrode surface and results in the ON state, while the Cu oxidation would further increase the base cell-resistance and result in the OFF state.

An alternative scenario to explain the switching might be that the Cu-chalcogenide interaction results in a trap-rich interfacial layer, whereby the application of a voltage would control the filling and emptying of the traps and control then the resistance of the layer. For instance, a positive voltage of +0.2 V may induce trap emptying over a certain thickness at the interface, accounting for the switching to the OFF state. The non-linear voltage dependence of the current may suggest that the thickness over which traps are emptied is voltage-modulated as for a space-charge layer. On the other hand, a voltage of -0.4 V or +0.8 V would induce trap filling and recovery of a resistor-like ohmic behavior, that is to say without any space-charge layer.

First retention tests suggest that the OFF state is more stable (retention over several weeks) than the ON state (some events of resistance increase were seen after less than 1 hour). In a trap-based scenario, these results suggest that traps tend to empty over time, which is also consistent with the stability of the initial state OFF$_{init}$ in as-prepared test-cells.

Further detailed electrical characterizations together with local physical analyses of the Cu/chalcogenide interface are currently under further investigation to help elucidate the actual switching mechanism.

4. Effect of the decrease of the chalcogenide-layer thickness

We studied the influence of the chalcogenide-layer thickness on the switching parameters of the cell.

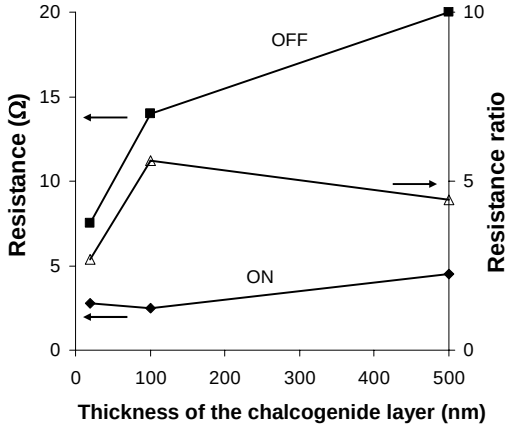


Fig. 2: Resistance plots of the Cu/doped-SbTe/W, extracted from both the OFF and the ON states, depending on the chalcogenide thickness; capacitor area is 0.85 mm^2.

Figure 2 shows the extracted resistance values in the two high-resistive states for different thickness values of the doped-SbTe layer from 500 nm down to 20 nm.

The plots show slight resistance decrease with the decrease of the thickness in the ON state, but significantly more important resistance reduction in the OFF state. However, these variations remain limited and support the hypothesis that the extracted resistance is mostly controlled by a high-resistive interfacial layer.

5. Effect of the decrease of the cell area

We also investigated the cell-size dependence of switching parameters, comparing the switching traces obtained both for 0.85 mm^2 and 0.085 mm^2 large Cu/chalcogenide/W cells, where the chalcogenide layer thickness was 500 nm. Both the resistance in the ON state and the switching current were usually similar for different cell sizes. However the resistance in the OFF state was systematically increased with the decrease of the cell size. This decrease is, however, lower than expected from the cell-size reduction assuming a simple resistor model. This might point to non-uniformities of the resistance, whereby the actual situation might be better modeled with parallel resistances of lower values distributed over the cell area. Regarding the cell-size independence of the resistance in the ON state, results suggest that switching occurs locally over the cell area, over part of the parallel resistances only, as for a filamentary switching. These results further confirm that the control of the high-resistive interfacial layer should be further optimized. Furthermore, the resistance ratio extracted was 5 for the large cell and more than 10 for the small cell.

Figure 3 shows the resistance points measured for our cells in the ON and the OFF states (squares). As a comparison, additional plots give the resistance that would be obtained as a function of the cell area considering only the resistivity of the chalcogenide layer (triangles). In addition, typical PMC resistance window is shown for 40 nm wide pores (black circles) [2]. As our measurements suggest the presence of an interfacial series resistance, the ON-OFF resistance values of the scaled devices should always give higher values than values calculated only from the chalcogenide resistivity. Hence, the extrapolation of the measured plots to small pore dimensions suggests that the resistance in the resistance window of our cell might be similar to the window of a PMC cell, and that the resistance in the ON state may even be higher than for a PMC cell (see Figure 3), which can also be deduced from the switching mechanism. This would mean that the operating current could be of the order of 1 μA or lower. If the circuitry is then able to sense and amplify such a low current, distinguishing it from the current of the OFF state (possibly down to the nA range), the scaled cell could thus be used as a low-power memory cell. Hence, to summarize, the promise of scaling this resistive-switching cell is that it may have low read and write currents down to the μA range or below, a write voltage that can be as low as for PMC cells, while resistance window is at least of the order of a decade.

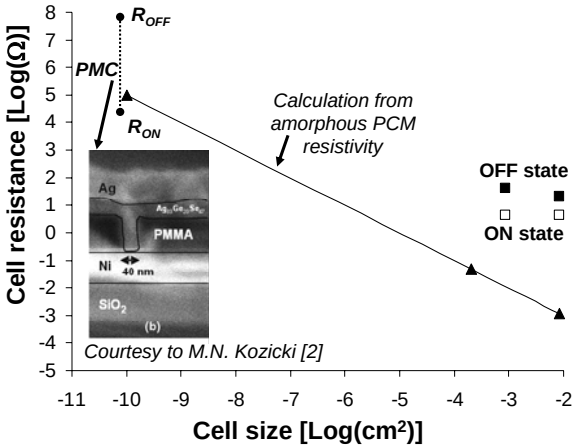


Fig. 3: Cell-size dependence of the measured resistances in both the ON (open squares) and OFF (full squares) states, and of the resistance as calculated from the resistivity of the chalcogenide layer (full triangles), for 500 nm thick chalcogenide layers; a typical resistance window of a PMC cell (from Ref. 2) for small pore opening is also shown.

6. Conclusion

We have explored the switching properties of a new structure composed of a doped-SbTe chalcogenide glass sandwiched between the metals Cu and W. We have demonstrated reproducible electrical switching between two high-resistive states for large test-cells. Extrapolation of the results suggests low-power switching operation of smaller devices, keeping still sufficient memory window. The switching mechanism between two high-resistive states may in itself prove a better route than a switching mechanism based on the formation/rupture of conductive filaments, because it may lead to lower current and better control of the current uniformity for small devices.

References

[1] International Technology Roadmap for Semiconductors, 2005

[2] M.N. Kozicki, IEEE Transactions on Nanotechnology, 4(3) (2005)

1TBulk eDRAM a reliable concept for nanometre scale high density and low power applications

S. Puget [1], G. Bossu[2], C. Guerin[2], R.Ranica[2], A.Villaret[2], P.Masson[3], J-M. Portal[3], R.Bouchakour[3], P.Mazoyer[2], V. Huard[1], T.Skotnicki[2]

[1]NXP, 850 rue Jean Monnet 38926 Crolles, France
[2]STMicroelectronics, 850 rue Jean Monnet 38926 Crolles, France
Contact: Phone +33 4 76 92 26 27, email: pascale.mazoyer@st.com
[3]L2MP UMR-CNRS 6137, IMT Technopôle de Château Gombert, 13451 Marseille Cedex 1, France

Abstract

Capacitor-less eDRAM cell appears to be an interesting candidate for future embedded memory generations. A particular attention is paid to the 1TBulk architecture in terms of bias operations, power consumption scalability and reliability. Thin and thick gate oxide device are analysed and compared. 1TBulk is found to be very promising for eDRAM low power applications. The thin gate oxide device 3.3 nm shows interesting reliability performance which allows the perspective of a very dense architecture integration.

1. Introduction

Many studies have tended to show that 1T-eDRAM cells, based on the floating body effect are interesting solutions compared to the standard 1T/1C cell for eDRAM applications. The 1T-DRAM concept was demonstrated on different CMOS platforms: bulk silicon [1-2-3-4], PDSOI [5-6], FDSOI [7] and Independent double gate [8-9]. In this paper a deep analysis of the 1TBulk cell is given. The interest of this architecture is the low cell area that can be achieved (10 F^2) with a very simple process fully compatible with standard CMOS logic technology. These characteristics are mandatory for low cost and high density embedded applications.

For this analysis different 1TBulk devices have been realized based on 90 nm technological platform: GO1 3.3 nm gate oxide device and a thicker gate oxide device GO2 5.3 nm. The cross-section, **Fig.1**, presents the GO1 87 nm device studied. Memory mechanisms are impact ionization for write, and forward biasing of source-body junction for erase operation [1]. **Fig.2** shows the kink effect typically observed on floating body devices.

The target of this paper is to demonstrate the interest of a thin gate oxide for the 1TBulk application. A peculiar attention is given on reliability when gate oxide is exposed to hot carriers during the write operation.

2. Gate length scaling

Tab.1 reports the bias memory operations for long (Lg = 169 nm) and short (Lg = 87 nm) 3.3 nm gate oxide devices. **Fig.3** shows higher memory effect amplitude for short device [2]. **Fig.4** gives the number of stored charges (Qmem) normalized by the gate length. This charge evaluation is based on the model [4] for the two different gate lengths. The model has been calibrated with impact ionisation current and junction leakages. Charges are stored in the depletion width (Wp) of each junction (source, drain, gate and N-buried) and are given by (1):

$$\left|\delta Q(Vb)\right| = qNaS\left|Wp_{eq} - Wp_{"1"/"0"}(Vb)\right| \quad (1)$$

for each junction.

Tab.2 gives the stored charges for each device and the corresponding floating body bias (Vb) respectively for states 1 and 0. It is observed that the number of stored charges is lower for the short gate length. But less charge have a larger effect on memory amplitude. Indeed the floating body bias shift (ΔVbret) is higher for short device than for long device.

3. Bias operation optimisation

In order to maximize the memory effect amplitude, we have analysed the different mechanisms involved in write and read operations. To reduce voltage sources in the application, gate bias (Vg) will be the same for read and write operations. This is mandatory to save circuit area.

Write operation: Impact ionization measurements, **Fig.5**, have been carried out on similar devices without N-buried implant. Vg will be chosen high enough to guaranty a sufficient impact ionization current level. To achieve a short write time, Vg is situated above the threshold voltage. **Fig.6** shows that the maximum of the memory amplitude is not observed for the maximum of impact ionization current (with Vg read = Vg write). In addition, for a given gate length, memory amplitude increases when Vg write decreases. Besides, **Fig.7** reports the impact ionization current efficiency: Isub/Id vs. Vg for a given Vd. When Vg decreases, efficiency increases and is quite independent of gate length.

Vg (read = write) will be chosen above the threshold voltage and below the maximum of the impact ionization current.

Vd write operation will be analyzed later in the next paragraph (Cf. &4 Speed).

Read operation: Impact of Vd read is illustrated in **Fig.8**. Memory effect amplitude increases when Vd read increases. But when impact ionization occurs, a parasitic write appears and state 0 disappears. **Fig.9** illustrates the increase of the memory amplitude with Vd read before parasitic write. For a same stored charge, ΔIs read increases with Vd.

Vd read will be chosen below the significant level of impact ionization current.

4. Speed

The write time τ is defined by (2):

$$Qmem = \int_0^{\tau} (Isub - Ileakage)dt \quad (2)$$

Qmem is the charge brought by the impact ionisation mechanism less charge flowing through the junction during the state 1 programming operation.

Tab.3 recalls memory operation bias for GO2 device and accumulated charge corresponding to theses conditions. The value of write time calculated for GO1, **Fig.10**, and GO2, **Fig.11**, will be optimistic. Indeed the evolution of diode leakage and impact ionization current has not been considered during all the programming time but only at the beginning. **Fig.10** and **Fig.11** allow to compare both devices.

Fig.10 shows GO1 device write time. 5 ns are achieved for Vd located in the grey part of the graph: Vd must be higher than 1.9 V. **Fig.11** shows GO2 device write time. 5 ns are achieved for Vd higher than 2.7 V.

One way to minimize the write time is optimisation of the junction profile. More abrupt the junctions will be, more efficient the impact ionisation will be. The limit is given by the corresponding junction leakage.

Fig.12 compares Isub for two gate lengths of GO1 devices. When gate length decreases, write time decreases according to increase of impact ionisation current shown in **Fig.13.**

As the maximum of impact ionization efficiency is obtained for short device and low Vg write, **Fig.7**, write time optimisation will be easier when devices will be shrunk across technological node evolution.

5. Consumption

Consumption is evaluated in terms of current for the different memory operations.

For reading conditions, **Fig.14,** read current is more important for GO1 (39µA) than for GO2 (17µA) devices, whenever a similar memory amplitude is measured. Read consumption will increase with the technological shrink.

For write operation, consumption is equivalent for all gate oxide thicknesses if we do not consider write time. But the faster write time for GO1 compared to GO2 device allows to decrease consumption for thin gate oxide devices (**Tab.4**). Write operation consumption is lower in GO1 than in GO2.

The consumption during the erase operation is very low, since in this case the transistor is not open.

1TBulk device seems to be indicated for high speed and low power applications.

6. Retention

Fig.15 shows that retention is not affected by the reduction of gate length. Even if stored charges in GO1 (Lg=87 nm) are less important than for Lg=169nm, the retention is not degraded. **Fig.16** and **Fig.17** show that retention is not affected by reduction of oxide thickness. Mechanisms involved in data loss are intrinsic to the junction. It depends on doping profile and quality of junctions. 10ms are measured on both devices (thick and thin gate oxide) at 85°C. This median value is one order magnitude higher compared to standard 1T/1C eDRAM characteristics. This good median retention time will allow to reduce refresh cycle and consequently the standby power consumption.

7. Reliability

1TBulk hot carrier reliability is investigated for both GO1 and GO2 oxide devices. **Fig.18** summarizes the main results obtained. It shows the relative degradation of reading current versus number of cycles under optimal write stress conditions. GO1 devices present a better reliability with at minimum $10^{14\text{-}15}$ cycles allowed whenever the GO2 devices are limited to $10^{13\text{-}14}$ cycles.

8. Conclusion and perspectives

This study has shown that, to reach a good memory amplitude and a fast write time, a compromise in terms of bias is required.

Similar memory amplitude is shown on thick and thin gate oxide devices. But, a smallest gate oxide provides a faster write time, with lower voltage supply. Retention time does not depend on gate oxide thickness, and is compatible with eDRAM requirements. Reliability is improved for GO1 device.

Thin gate oxide device for 1TBulk application seems to be the most indicated device for 45nm and below.

References

[1] R. Ranica et al., VLSI Tech. Dig., 2004, p.128-9
[2] R. Ranica et al., VLSI Tech. Dig., 2005, p.38-3B
[3] P. Malinge et al, VLSI Circuit, 2005, p.358-23
[4] A. Villaret et al., Micr Eng., Vol.72, 2004, p. 413-9
[5] S. Okhonin et al., IEEE Int. SOI Conf., 2001, p.153-4
[6] T. Shino et al., VLSI Tech. Dig., 2004, p.132-3
[7] C. Kuo et al., IEDM Tech. Dig., 2002, p.843-6
[8] C. Kuo et al., IEDM Tech. Dig., 2002, p.843-6
[9] I. Ban et al, IEDM Tech Dig., 2006, p -21

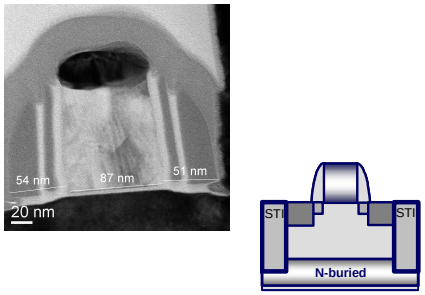


Fig.1: Cross section view of device fabricated with 90 nm technology (Lg=87nm, Tox=3.3nm (GO2), W=0.16µm).

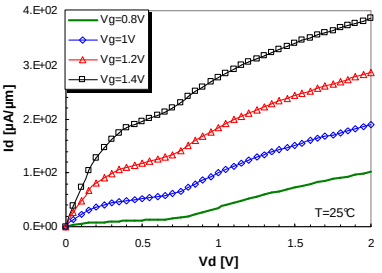


Fig.2: Id(Vd) characteristics at variable Vg show p-well auto-biasing for Lg=87nm device.

	Write	Erase	Read	Hold
Vd	+1.8V	-1.2V	+0.4V	0V
Vg	+0.8V	-1.2V	= Vg write	0V
Vs	0V	0V	0V	0V
$V_{N\text{-buried}}$	+0.6V	+0.6V	+0.6V	+0.6V

Tab.1: Memory operation bias 87 and 169 nm GO1 devices. Threshold voltage is 0.5V. All the conditions have been chosen to reduce the number of voltage sources.

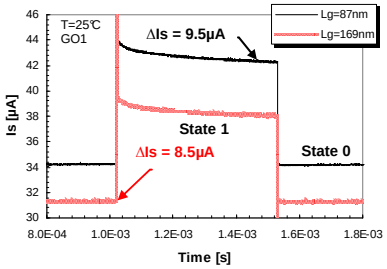


Fig.3: Source current shift on devices with two different gate length (Lg 169nm and 87nm). Write operation is more efficient on short devices.

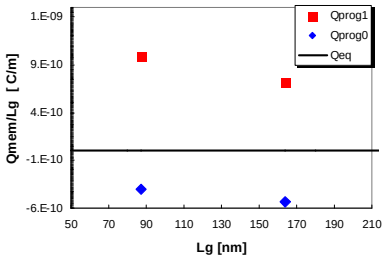


Fig.4: Memory charge evaluation for GO1 devices calculated by the model [4] after state "1" and state "0" programming for Lg=87nm et Lg=169nm.

Gate length	169nm	87nm
Qprog1 : « 1 » Cell hole nbr	+766	+534
Qprog0 : « 0 » Cell electron nbr	-560	-217
Qmem=Qprog1-Qprog0	1327	752
Qmem [C]	$2{,}12.10^{-16}$	$1{,}20.10^{-16}$
Vbret « 1 » [V]	0.217	0.275
Vbret « 0 » [V]	-0.186	-0.167
ΔVbret [V]	O.403	O.413

Tab.2: Memory charge evaluation for GO1 devices and floating body bias calculated by the model [4].

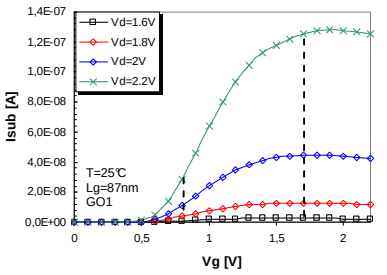


Fig.5: Ionization current Isub(Vg) for different Vd write (Lg=87nm, W=0.16µm).

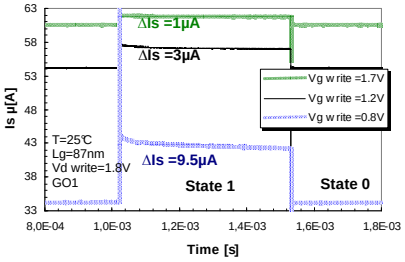


Fig.6: Source current shift on 87nm device for three different write gate bias (Vg write =1.7V corresponds to the maximum of the impact ionization current).

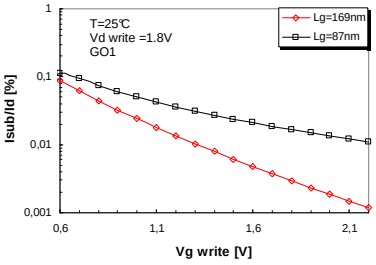


Fig.7: Efficiency (Isub/Id) for two different gate lengths (Lg 169nm and 87nm) Efficiency increases with short gate length and with Vg write decrease.

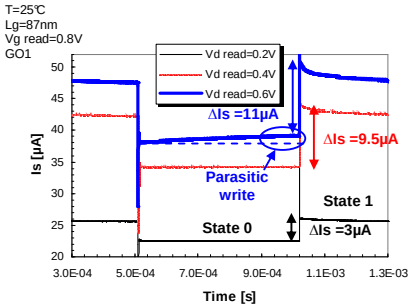


Fig.8: Read current margin between state 1 and state 0 (Lg=87nm) with different Vd read. Source current shift shows a more efficient memory effect for Vd read= 0.4V without deterioration of state 0.

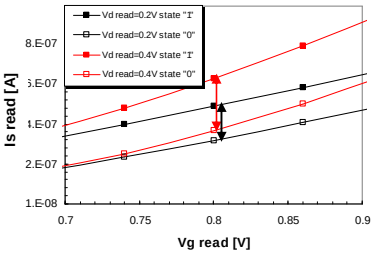


Fig.9: Is read vs. Vd read. Non calibrated model. For the same stored memory charge Is read is more important at higher Vd read.

	Write	Erase	Read	Hold
Vd	+2.5V	-1.2V	+0.4V	0V
Vg	+0.8V	-1.2V	= Vg write	0V
Vs	0V	0V	0V	0V
VN-buried	+0.6V	+0.6V	+0.6V	+0.6V

(a)

Qprog1: « 1 » Cell hole number	+625
Qprog0: « 0 » Cell electron number	-569
Qmem=Qprog1-Qprog0	1195
Qmem [C]	$1{,}9.10^{-16}$

(b)

Tab.3: (a) Memory operations bias GO2 device. Threshold voltage is 0.5V. All the conditions have been chosen to reduce the number of voltage sources. Device characteristics are Lg=164nm, Tox=5.3nm W=0.16µm. (b) Memory charge evaluation for GO2 device calculated by the model [4].

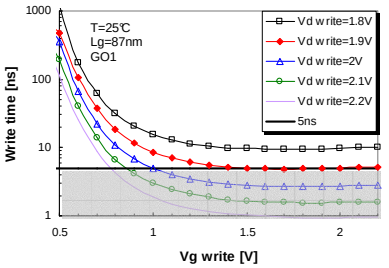


Fig.10: Write time for GO1 shows 5ns are required to write state 1 with Vd write superior to 1.9V and Vg write=1.4V (at beginning of state 1 programming).

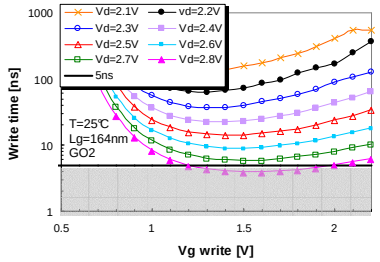


Fig.11: Write time for GO2 shows few nanoseconds are required to write sate 1 with Vd write superior to 2.7V (at beginning of state 1 programming).

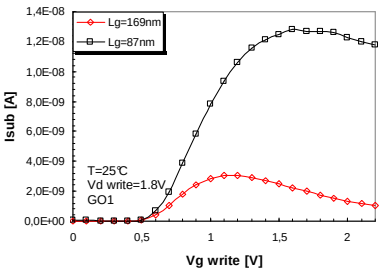


Fig.12: Ionization current Isub(Vg) on devices with two different gate lengths (Lg 169nm and 87nm) shows at the same Vg write a higher ionization current for the shorter device.

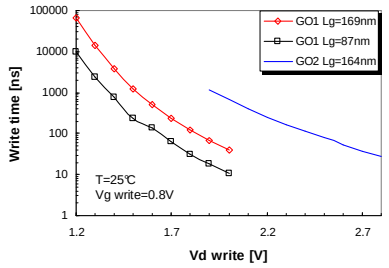


Fig.13: Write time for GO1 and GO2 devices. A shorter write time is noticed on thin gate oxide devices an specifically with shorter gate length (for bias presented respectively in Tab.1 and Tab.3 (a)).

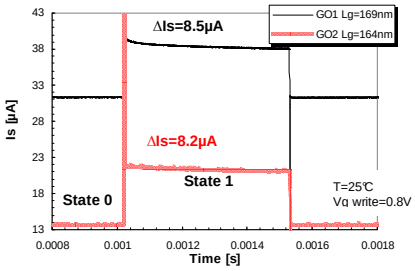


Fig.14: Read current margin between state 1 and state 0 for two devices with GO1 and G02 gate oxide (GO1 Lg=169nm, GO2 Lg=164nm, for bias presented respectively in Tab.1 and Tab.3 (a)).

	Read			
	Vds [V]	Is [μA]	P [mW]	Write time [ns]
GO1	0.4	39	0.016	
GO2	0.4	17	0.007	
	Write			
GO1	1.8	69	0.012	31
GO2	2.5	49	0.012	81

Tab.4: Read and write consumption evaluation for GO1 (Lg=67nm) and GO2 (Lg=164nm).

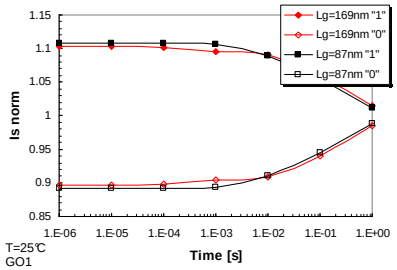


Fig.16: Normalized retention characteristics for GO1 devices. Retention is not linked to gate length.

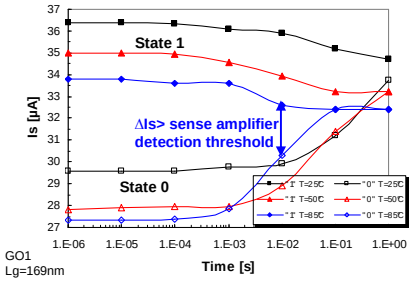


Fig.16: GO1 Retention characteristics show a retention time over 100ms at 25°C and over 10ms at 85°C.

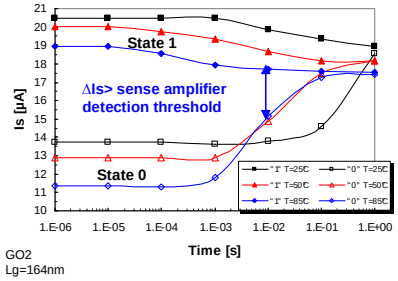


Fig.17: GO2 Retention characteristics show a retention time over 100ms at 25°C and over 10ms at 85°C.

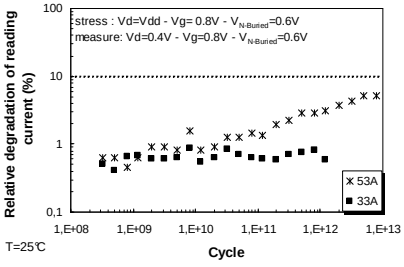


Fig.18: Reliability: relative degradation of reading current versus the number of write cycles for GO1 (Lg=169nm) and GO2 (Lg=164nm) devices.

SESSION F

SRAM & Process Variability

Real-time Soft-Error Rate Testing of Semiconductor Memories on the European Test Platform ASTEP* *(invited)*

J.L. Autran[a,b], P. Roche[c], G. Gasiot[c], D. Munteanu[a], T. Parrassin[c], J. Borel[d], J.P. Schoellkopf[c]

[a] L2MP, UMR CNRS 6137, 49 rue Joliot Curie, BP 146 – F-13384 Marseille Cedex 13, France
Phone: +33 4 96 13 97 17 – Fax : + 33 4 96 13 97 09 – E-mail: jean-luc.autran@l2mp.fr
[b] Also with Institut Universitaire de France (IUF)
[c] STMicroelectronics, 850 rue Jean Monnet – F-38926 Crolles Cedex, France
[d] JB R&D, Ferrière, F-05250 Saint-Etienne en Dévoluy, France

Abstract

The "Altitude SEE Test European Platform" (ASTEP*) is dedicated to real-time soft-error rate (SER) testing of semiconductor memories. The platform, located in the French Alps on the "Plateau de Bure" at 2552m, has been operational since March 2006. This test facility includes a proprietary automatic test equipment specially designed for static memory (SRAM) testing and secured remote control operation via internet. Real-time SER measurements on 3.6 Gbit of SRAMs manufactured in CMOS 130 nm technology are reported, as well as the comparison between real-time and accelerated SER. Finally, project perspectives for CMOS 65nm SRAMs and real-time in situ neutron monitoring are presented.

1. Context and project milestones

Since terrestrial cosmic-rays have been identified to be at the origin of soft errors in modern integrated circuits, the estimation of soft error rates (SER) is rapidly becoming a major consideration for reliability aspects at device, circuit and system levels [1]. For deep submicron technologies, SER of chips is becoming a vital customer issue and its determination is still an open challenge, not only to investigate and understand technology sensitivity but also to extrapolate the trends for future generations of circuits.

Different experimental and simulation approaches are known to estimate SER: accelerated testing using alpha, neutron or proton source/beam, life testing under natural environments, modeling and software simulation at device or circuit level, combination of experimental/simulation approaches [1-2]. In contrast with accelerated testing which is relatively easy to conduct, cheaper and fast (a few hours/days is generally sufficient to obtain confident results), life testing is clearly time consuming and expensive. But it appears as the unique experimental solution to accurately estimate SER, ensuring that the test does not introduce artificial results due, for example, to beam uniformity or fluctuations, dosimetry errors, chip disorientation or difference in spectrum (largely introduced by the cut-off energy of the accelerator which is always well below cosmic ray energies). Life testing can also address SER at system level for complex electronic solutions and, installed in an underground site, provide an efficient method of monitoring for radioactive contamination. On the contrary, when based at altitude to increase the flux of particles (primarily neutrons), SER by life testing can be accelerated by a factor ~2-15 depending on the earth location of the test site.

ASTEP, Plateau de Bure, France		
Latitude (°N)		44.6
Longitude (°E)		5.9
Elevation (m)		2550
Atm. depth (g/cm^2)		757
Cutoff rigidity y (Gy)		5.0
Relative	Active Sun low	5.76
	Quiet Sun peak	6.66
	Average	**6.21**

Table 1. Location and main environment characteristics of the ASTEP Platform (After Ref. [2]).

Figure 1. Aerial view of the Plateau de Bure Observatory. The ASTEP platform is hosted in Building POM2 indicated in the figure (Photo courtesy of IRAM).

The project of an "Altitude SEE Test European Platform", located in the French Alps and simultaneously opened to industrials and laboratories to conduct, in the same place, qualification tests or research works, was initiated by STMicroelectronics and JB R&D in 2001. The ASTEP consortium was created in 2003 and operated by the CNRS and L2MP laboratory in 2004, in the framework of a multi-partner European-national funding program. At the end of 2004, Bertin Technology (Aix-en-Provence, France) was selected as

* The ASTEP project (www.astep.eu) is operated by L2MP-CNRS, in collaboration with STMicroelectronics, JB R&D and with the technical support of IRAM. This work is supported by the European Commission, the Provence Alpes Côte d'Azur Regional Council, the Hautes Alpes Department Council, the City of Saint-Etienne en Dévoluy, STMicroelectronics, the Centre National de la Recherche Scientifique, the Université de Provence and the Institut Universitaire de France.

the industrial integrator of the automatic tester equipment. The construction of the machine was conducted in 2005, including one trimester of tests and qualifications and the preparation of the test area on the Plateau de Bure. Finally, the equipment was installed in altitude, on the Plateau de Bure, in February/March 2006 and the first campaign of SER life testing officially began on March 2006. The aim of this paper is precisely to present an overview of this project and to report the first experimental results obtained during this first year operating period.

2. The ASTEP platform

The ASTEP platform is located in the French Alps at 2,552m of elevation. The installation is hosted by the Institute for Radio-astronomy at Millimeter Wavelengths (IRAM) on the Plateau de Bure in the Dévoluy Mountains. Fig. 1 shows a general aerial view of the observatory on the Plateau: the ASTEP platform is installed in an ancient radio-telescope building reconverted into an altitude laboratory platform (one floor standard concrete slab building). The main environment characteristics of the ASTEP platform are summarized in Table 1. Since 2006, this test location has been referenced in the latest release of JEDEC Standard JESD89A [3]. Data of Table 1 corresponds to values of Table A3.B in Ref. [3].

The ASTEP masterpiece is a specially designed and universal SRAM automatic test equipment (ATE), capable of monitoring several thousands of synchronous/asynchronous SRAM memories and performing all requested operations such as writing/reading data to the chips, comparing the output data to the written data and recording details on the different detected errors. The ATE hardware and software have been designed and developed by BERTIN Technologies (Aix-en-Provence, France), in collaboration with L2MP-CNRS and STMicroelectronics. The design of both the hardware and software components of the system follows all the specifications of the JEDEC SER test standard [3]. Fig. 2 shows a schematic representation of the whole test equipment. This system is divided in two identical

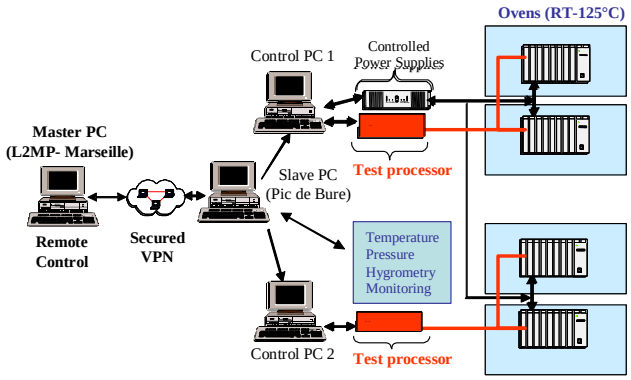


Figure 2. Schematic description of the automatic test equipment (ATE) designed and built to perform real-time SER tests on static memories. The ATE is divided in two networked subsystems (Astep1 and Astep2), capable of monitoring two memory racks (640 test chip per rack) placed in temperature-controlled ovens.

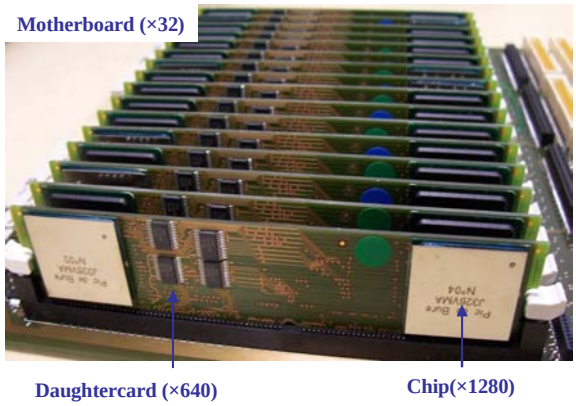


Figure 3. Detail of one motherboard containing 40 daughtercards and 80 test chips (2 per daughtercard). The complete ATE system is designed to test a maximum of 1280 chips dispatched on 32 motherboards and 640 daughtercards.

subsystems, Astep1 and Astep2, capable of independently performing an automatic survey of two racks of memories placed into temperature-controlled ovens (RT-125°C). Each rack contains 8 motherboards; each motherboard connects 40 daughtercards and each daughtercard drives 2 test chips. The daughterboards (and consequently the chips) are oriented face horizontally. The current consumption of each daughtercard, i.e. of each pair of devices, is real-time monitored on 4 supply lines on a total of 5 for Single Event Latchup (SEL) detection. This explains the relative complexity of the electronic circuitry developed at the level of each daughtercard (current monitoring) and for each motherboard (the system is able to disconnect a given daughtercard in case of abnormal power consumption; the current intensity threshold is directly controlled by the software interface as an input parameter of the test). Fig. 3 shows a zoom of a motherboard with its daughtercards and test chips. The maximum capacity of the ATE in the current configuration is 1280 test chips. Of course, the system is modular and it will be able to received additional racks for future experiments. The ATE can be completely controlled in remote control mode via a virtual private network (VPN) on internet. This allows the users to perform all control operations during a real-time experiment and to access all the data in real-time.

The test procedure implemented in the ATE control software allows the discrimination of the following error types as a function write/read and rewrite/reread operation results: "*Transient Soft Error*" (TSE), sometimes called dynamic read error; "*Static Soft Error*" (SSE), i.e. classical bit flip; "*Single-Event Hard Error*" (SHE), sometimes referred as stuck bit. When an error is detected, the software increments the data-log file with all information concerning the fail: date and time, type of error (TSE, SSE or SHE), written pattern and read pattern, round number (i.e. scan number of all memories from the beginning of the experiment), testchip identification (rack, board, slot and chip numbers) and logical address. An additional software is used to perform a physical mapping of the bit flips. This mapping is used to discriminate logical multi-bit upsets (MBU) from physical multi-cell upsets (MCU), as recently recommended by [3].

3. Results on 130nm SRAMs

The first real-time testing campaign has been currently performed on bulk static memories (SRAM) fabricated by STMicroelectronics using a CMOS 130 nm commercial technology (PBSG-free) process (6-transistor cell with a bit cell area of 2.50 μm^2.). This technology was extensively characterized in previous works using alpha irradiations (with both ST and L2MP setups) and neutron accelerated SER tests performed with two continuous neutron sources available in North-America (TRIUMF and LANSCE facilities). The use of a mature and well-characterized technology is important for the validation of ATE functionalities and data comparison between accelerated and life time testing. The test chip is composed of different memory cuts; only a single cut of 4 MBit has been used (i.e. connected) for the present measurement campaign. The ATE was initially loaded at 72% of its maximal capacity with 492 test chips in the Astep1 subsystem and 424 chips in Astep2, respectively. This represents a total of 912 chips, i.e. 3,664 MBit under test.

Fig. 4 shows the cumulative fail number versus test hours for the two subsystems Astep1 and Astep2 during the period March 31 – November 26, 2006. These events correspond to Static Soft Errors (SSE) detected under nominal test conditions (V_{DD} = 1.2 V, room temperature and checkerboard pattern for all devices). Because of the installation and maintenance operations in the test room, the two subsystems were subjected to interruptions during this period. However, the number of MBit×h cumulated during these eight months reaches 15 617 920 MBit×h, which gives an excellent confidence interval on the extrapolated SER, as shown in the following. A total of 72 fails was detected, including 67 single event upsets and 5 MBU. These later involved 2 physical adjacent bit cells in all cases.

From data of Fig. 4, we estimated real-time SER, shown in Fig. 5, using the following expression:

$$SER = \frac{N_r}{AF \times \Sigma_r} \times 10^9 \text{ (FIT/MBit)} \qquad (1)$$

where N_r is the number of errors observed at time T_r , AF is the acceleration factor of the test location and Σ_r is the number of MBit×h cumulated at time T_r. The acceleration factor value is considered equal to AF=6.21 for the ASTEP platform. It corresponds to the estimated relative neutron flux (average value) of the Plateau de Bure with respect to the reference flux of New-York City, as reported in Table A.3-B of Ref. [3].

95% confidence interval upper and lower limits [3] are also indicated in Fig. 5. An average value of 750 FIT/MBit is obtained, with lower and upper confidence limits equal to 610 and 900 FIT/MBit, respectively. In addition, Fig. 5 shows that the convergence of SER vs. test hours is reached in approximately 2000-2500 h, i.e. for 4-6×10^6 MBit×h. Beyond this duration, the SER remains constant around 750 FIT/MBit.

In complement to real-time characterization, both neutrons and alpha irradiations (under nominal test conditions) were performed on several test chips issued from the same technological lot. Neutron experiments

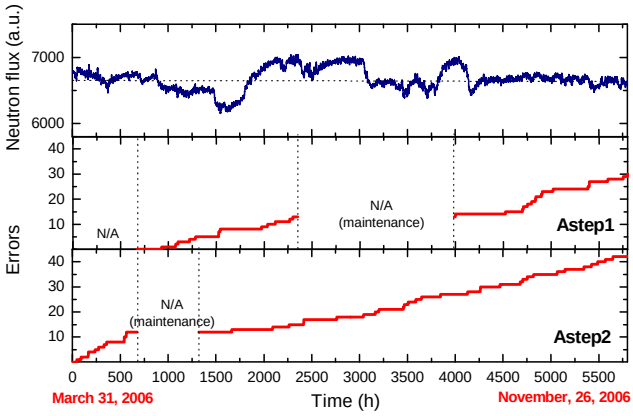


Figure 4. Cumulative fail numbers versus test hours for the two boards Astep1 and Astep2 of the ATE. The experiment started on March 31, 2006 and stopped on November 26, 2006 under nominal test conditions (V_{DD} = 1.2V, room temperature, checkerboard). The real time data of the NM64 neutron monitors Jungfraujoch is also indicated (Jungfraujoch neutron monitor data were kindly provided by the Cosmic Ray Group, Physikalisches Institut, University of Bern, Switzerland).

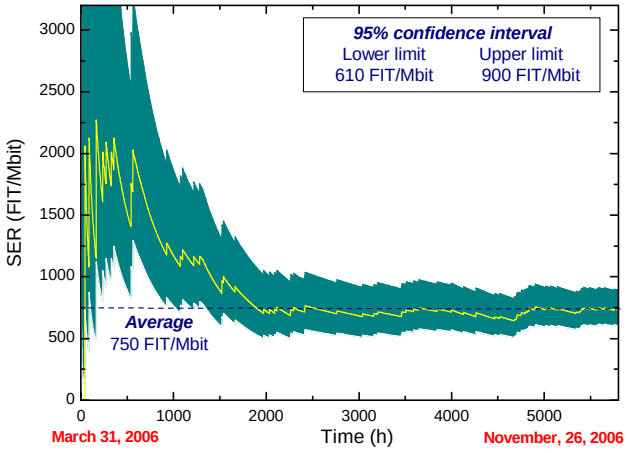


Figure 5. Estimated real-time SER (FIT/MBit) versus test hours calculated from data of Fig. 7. 95% confidence intervals are also indicated. The average acceleration factor value of 6.21, given in Ref. [1], was used to estimate this real-time SER. The erratic character and high value of the SER in the first ~1000 hours are due to the very low number of cumulated fails during this period, introducing a large error in the evaluation of the ratio N_r/Σ_r in Eq. (1).

were performed using the continuous spectrum sources available at both the Los Alamos Neutron Science Center (LANSCE) and at the Tri-University Meson Facility at Vancouver, (TRIUMF). Additional alpha-irradiation measurements were also performed using two different Am^{241} sources at ST and at L2MP laboratory. SER values obtained with these two setups agreed within ±10%. Fig. 6 shows the normalized accelerated results for this commercial CMOS 130nm SRAM. Average values of 380 FIT/Mbit and 665 FIT/Mbit have been obtained for alpha-SER and neutron-SER, respectively.

If we try now to compare, as in [4], these accelerated and real-time test results, one have to reminder that the SER given by the altitude experiment (i.e. 750 FIT/MBit) must be corrected from the impact of alpha contamination affecting all tested devices. In other words, this signifies that a certain number of fails detected by the ATE during the test period have been induced by alpha particles and not by neutron interactions. Supposing a real-time alpha-SER equal to

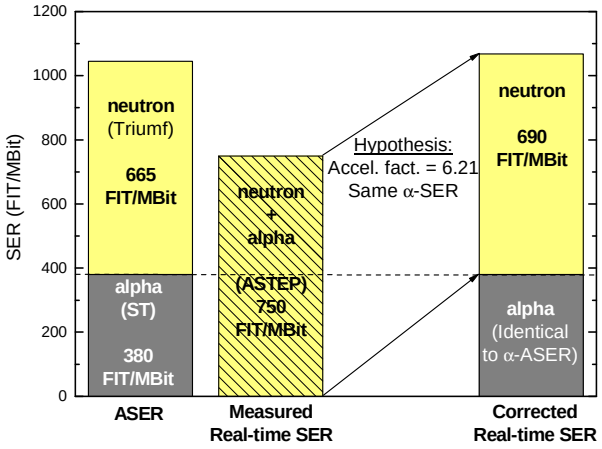


Figure 6. Comparison between the accelerated (ASER) and the real-time SER estimated in Fig. 5. Accelerated tests were performed at Triumf facility for neutron-ASER and at both ST-Crolles and L2MP-Marseille for alpha-ASER characterization, respectively. Real-time SER data of Fig. 5 is corrected from the contribution of alpha disintegrations, considering a same alpha-SER for both accelerated and real-time experiments.

the value given by accelerated tests (i.e. 380 FIT/Mbit) and taking into account the acceleration factor (AF=6.21) of the test location only for neutron-induced fails, we obtain a corrected neutron failure-in-time of 690 FIT/Mbit. This corrected value is very close to the accelerated neutron-SER equal to 665 FIT/Mbit, demonstrating, in this case, a very good agreement between the two estimation methods within the experimental error margins.

Three Dimensional (3D) Monte-Carlo simulations used to predict neutron-SER were also performed for this SRAM circuit, using a proprietary radiation simulator developed by STMicroelectronics [5-6]. This simulation code considers the exact layout of the memory cell and calibrated TCAD results as inputs. Simulation results for nominal conditions (V_{DD} = 1.2 V, room temperature) give a typical neutron failure-in-time of 700 FIT/Mbit. This value is very close to those corresponding to both accelerated and real-time experiments.

4. Future experiments and project perspectives

Since February 2007, a new measurement campaign has been started with the same 130 nm SRAM circuits. The objective is now to quantify the impact of several key-parameters, i.e. the power supply voltage, the test temperature (up to 125°C) and the configuration of the written pattern, on the real-time SER. Another objective is also to further compare real-time and accelerated tests for which data is already available on this test vehicle.

Beyond this work currently in progress, STMicroelectronics, with its industrial partners NXP (formerly Philips Semiconductor) and Freescale (formerly Motorola), will test a new circuit on the ASTEP platform within the following months. This circuit is a CMOS 65nm test vehicle for library qualification and process monitoring. It contains 8.5 Mb of single port SRAM (bit cell area of 0.525 µm^2) already characterized this year from an accelerated-test point-of-view with neutrons at LANSCE and TRIUMF, as well as with alphas at STMicroelectronics. About 1000 test chips, representing a memory capacity up to 8 Gbit, will be mounted in the automatic test equipment by the beginning of Q2 2007.

The last near-term perspective for the ASTEP platform is the development, in 2007, of an *in situ* neutron monitor, installed close to the test equipment (directly in the test room) to try to correlate the observed circuit fails with the total neutron flux incident on the experimental area. This equipment, based on high pressure He^3 neutron proportional counters (LND type 253109), will be completed, also in 2007, by a large array CCD camera for imaging (and quantitatively analyzing) neutron-Silicon interactions. Finally, a high sensitivity neutron spectrometer will be constructed and installed in 2008-2009 for an ultimate characterization of the ASTEP radiation environment [7].

Acknowlegments

The authors would like to thank their colleagues Jean-Marc Drevon, Grégory Wauters and Dominique Delhom from BERTIN Technologies (Aix-en-Provence) for their major technical contribution to the development and construction of the automatic test equipment. The logistical support of the Institute for Radioastronomy at Millimeter Wavelengths (IRAM) is also gratefully acknowledged. Special thanks are due to Bertrand Gautier (IRAM, Station Manager of the Plateau de Bure Observatory) for his continuous support.

References

[1] J.F. Ziegler, H. Puchner, SER – History, Trends and Challenges, Cypress Semiconductor, 2004. See also references therein.

[2] P. Roche, "Year-in-Review on radiation-induced Soft Error Rate", tutorial at IEEE International Reliability Physics Symposium, San Jose, USA, March 2006.

[3] JEDEC Standard Measurement and Reporting of Alpha Particles and Terrestrial Cosmic Ray-Induced Soft Errors in Semiconductor Devices, JESD89 Arlington, VA: JEDEC Solid State Technology Association.

[4] H. Kobayashi, H. Usuki, K. Shiraishi, H. Tsuchiya, N. Kawamoto, G. Merchant, J. Kase, "Comparison Between Neutron-Induced System-SER And Accelerated-SER in SRAMs", Proceedings of the IEEE International Reliability Physics Symposium, Phoenix, USA, pp. 288-293, 2004.

[5] P. Roche, G. Gasiot, "Impacts of Front-End and Middle-End Process Modifications on Terrestrial Soft Error Rate", IEEE Transactions on Device and Materials Reliability, Volume 5, N°3, pp. 382-396, 2005.

[6] P. Roche, G. Gasiot et al., "Comparisons of Soft Error Rate for SRAMs in Commercial SOI and Bulk below the 130 nm Technology Node", IEEE Transactions on Nuclear Science, Volume 50, N°6, pp. 2046-2054, 2003.

[7] All information and updates are available on the website of the ASTEP platform: www.astep.eu

Low Voltage SRAM with Noble Cell Bias Technique to Increase Static Noise Margin

Yeonbae Chung, Seung-Ho Song, Yoon-Joo Eom, and Sang-Won Shim

School of Electrical Engineering and Computer Science, Kyungpook National University
1370 Sankyug-Dong, Book-Gu, Daegu, Republic of Korea
e-mail: ybchung@ee.knu.ac.kr

Abstract

This work presents a low voltage SRAM technique based on dual-boosted cell array. For each read/write cycle, the wordline and cell power node of selected SRAM cells are internally boosted into $1.5V_{DD}$ and $2V_{DD}$, respectively. This technique enhances the read static noise margin (SNM) to a sufficient amount, even at sub-1V supply voltage. It also improves the SRAM circuit speed owing to an increase of the cell read-out current. A 0.18-μm CMOS 256-Kbit SRAM test chip has been implemented with the proposed scheme, which demonstrated: 1) 0.8 V operation with 50 MHz while consuming a power of 65 μW/MHz; and 2) a reduction by 87 % in bit-error rate while operating with 43 % higher clock frequency compared with that of conventional SRAM.

1. Introduction

As mobile electronic systems become popular, power consumption is a major concern in VLSI chip design. Although various techniques to reduce the power dissipation have been developed [1], lowering of the supply voltage is the most effective way. The SRAM is an important intellectual property block and occupies a large area in SoC. However, the performance of SRAM is greatly affected by the operating voltage. The static noise margin (SNM), a measure of the SRAM cell stability, deteriorates with reduction of the supply voltage [2]. It causes to increase the fail-bit rate of SRAM cell array. In addition, the SRAM cell current for data detection is also reduced, which degrades the operation speed of SRAM. There have been several attempts to overcome the degraded operating margin and degraded cell current due to lowering the supply voltage [3]-[7]. In this work, we propose and demonstrate a dual-boosted cell based SRAM which can enhance both cell stability and cell current to a sufficient amount.

2. Boosted-Cell-Array Based SRAM

Fig. 1 shows a schematic diagram of 6-T SRAM cell and the proposed bias conditions on the memory cell. Although the SRAM cell stability is certainly important during standby mode, the SNM during read operation represents a more significant limitation to SRAM operation [2]. In addition, both read SNM and cell read-out current value are in an inverse correlation [5]. The main idea to improve both cell operating margin and circuit speed is the following.

During the read operation, boosting the wordline

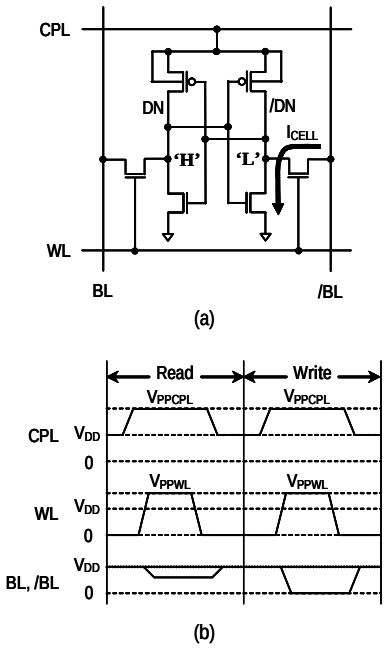


Fig. 1. 6-T SRAM cell: (a) configuration, (b) proposed bias conditions for read/write cycle. (DN: data-node, /DN: /data-node)

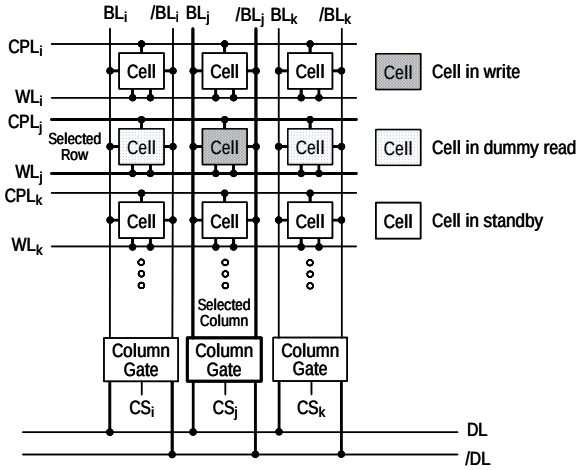


Fig. 2. Status of memory cells in array during write operation. (CS: column-select, DL: dataline, /DL: /dataline)

(WL) voltage of selected cells above supply voltage increases the driving capability of NMOS access transistor. Thus it increases the cell read-out current (I_{CELL}), resulting in reduced bitline (BL) delay time [3]. But the read SNM decreases. To compensate the reduced SNM, a higher voltage than the WL level is applied to the cell power line (CPL), which improves the cell read stability with enlarged SNM. In this work, the WL

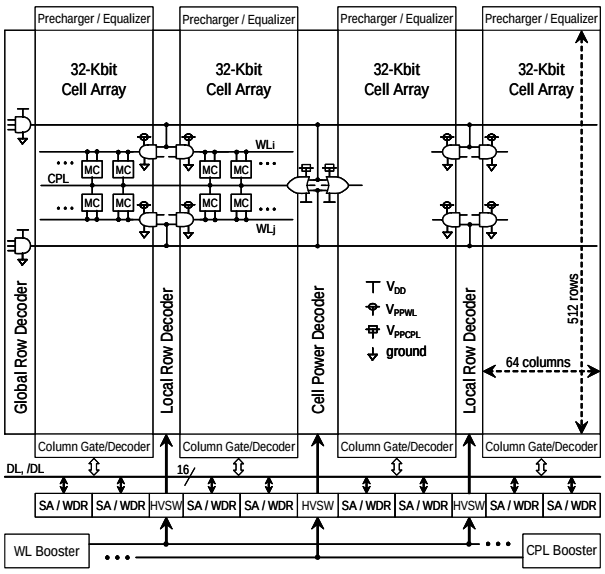


Fig. 3. Configuration of a memory block. (MC: memory cell, HVSW: high-voltage switch, SA: sense amplifier, WDR: write driver)

boosting level (V_{PPWL}) and CPL boosting level (V_{PPCPL}) are chosen to be $1.5V_{DD}$ and $2V_{DD}$, respectively. Meanwhile, to achieve a good write operation in low voltage SRAM, an access transistor with strong driving capability and relatively weaker pull-up transistor are desired [7]. However, during the write operation as shown in Fig. 2, only one of interleaved columns is selected for write while cells from the remaining columns on a selected row will experience a dummy read operation. To achieve the best read and write margin on the cells from a selected row, the same bias conditions are attempted to the memory cells as that of read operation. Boosting the CPL voltage to $2V_{DD}$ increases the read SNM of cells on the dummy read operation, but disturbs the write operation of cell on the selected column because the conductance of PMOS pull-up transistor becomes larger [4]. To resolve it, a boosting voltage of $1.5V_{DD}$ is applied to the WL, which improves driving capability of the access transistor. With increased driving capability of access NMOS, the cell internal node with data 'high' state can be driven closer to the ground through the bitline during write, which makes cell flip state easily.

Fig. 3 shows a memory array configuration based on the proposed dual-boosted cell technique. The WL booster providing $1.5V_{DD}$ is composed of a simple circuitry with one boosting capacitor [8]. The cell array design has been optimized for both boosting performance and area efficiency. One array block has four sub-blocks, each containing 512-row × 64-column. Each WL couples 64 cells. Each CPL is shared among two up and down cells and runs parallel to the WL, coupling 256 cells within two adjacent sub-blocks. High-voltage switch (HVSW) decoded by block address supplies the boosting voltage to the selected sub-block, which limits the overall power consumption due to voltage boosting only to a selected sub-block level.

Fig. 4 shows the CPL boosting circuit providing $2V_{DD}$. It consists of two stages of boosting circuitry

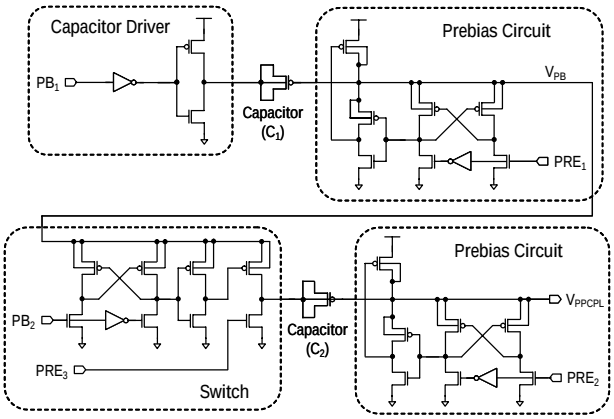


Fig. 4. Configuration of CPL booster.

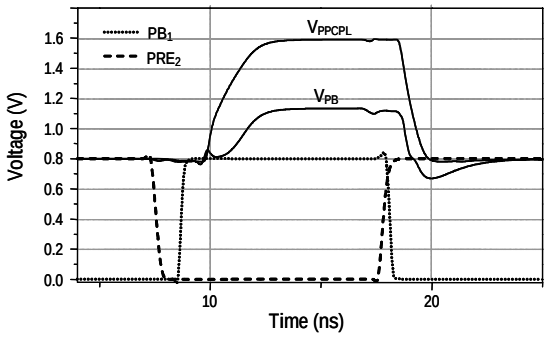


Fig. 5. Simulation waveforms of CPL booster at V_{DD} = 0.8 V.

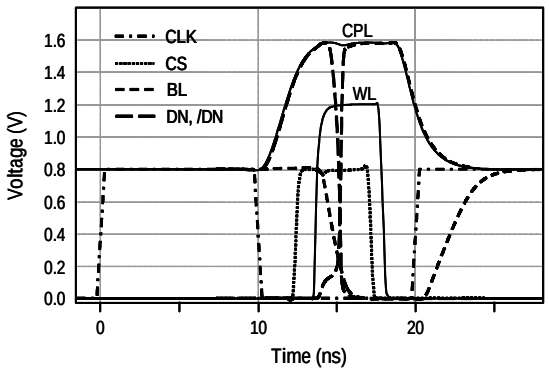


Fig. 6. Simulation waveforms for write at V_{DD} = 0.8 V and 20 ns cycle. (CLK: clock signal)

which are serially connected by switch. As shown in Fig. 5, the voltage level of CPL is settled down to the target value in 5 ns after kicking boosting capacitors by signal PB_1.

Fig. 6 shows circuit simulations for write cycle at 0.8 V. The cycle time is 20 ns. The internally boosted levels of CPL and WL are 1.6 V and 1.2 V, respectively. Even though the voltage level of CPL is higher than that of WL, the cell internal nodes (DN, /DN) flip the state in 2 ns after wordline access.

3. Experimental Results

The proposed techniques has been designed in a 256-Kbit SRAM, and fabricated with 0.18 µm CMOS process technology. Fig. 7 shows the CAD plot and Fig. 8 shows the chip photograph. The organization is 32K-word × 8-bit. The macro size is 1520 µm × 1490 µm =

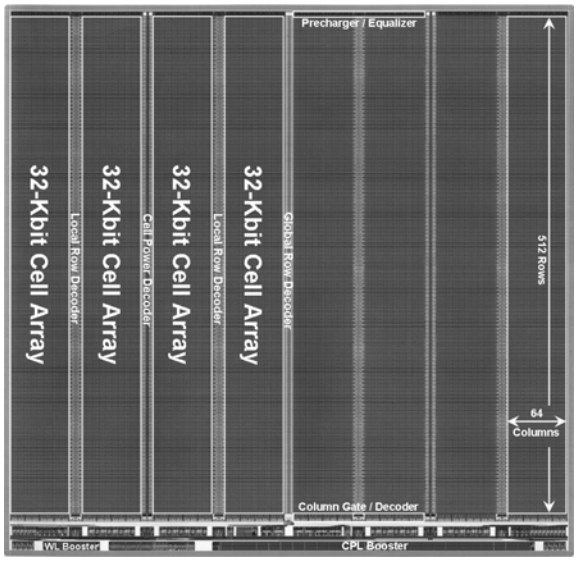


Fig. 7. CAD plot of 256-Kbit SRAM macro.

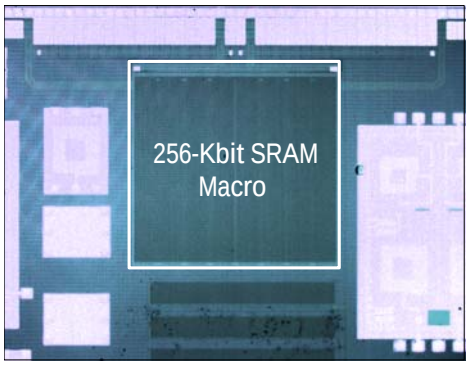


Fig. 8. Chip microphotograph.

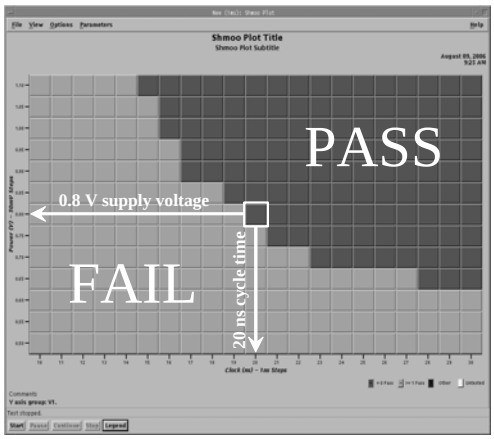


Fig. 9. Shmoo plot.

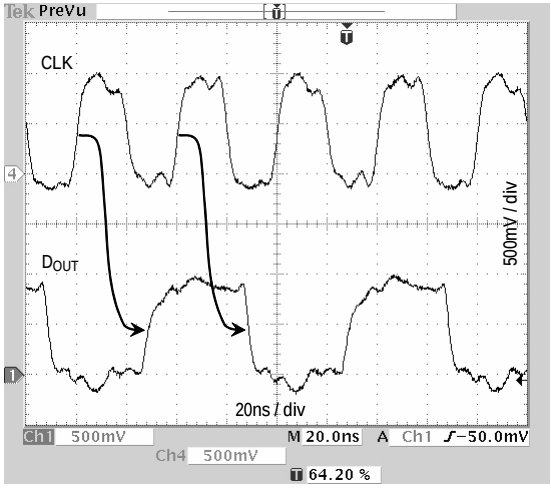


Fig. 10. Measured waveforms for read at V_{DD} = 0.8 V and 40 ns clock cycle.

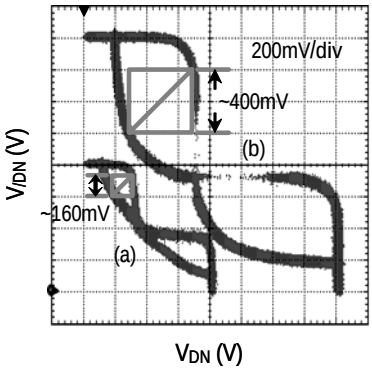


Fig. 11. Measured butterfly curves at V_{DD} = 0.8 V: (a) without cell boosting, (b) V_{PPWL} = 1.2 V and V_{PPCPL} = 1.6 V.

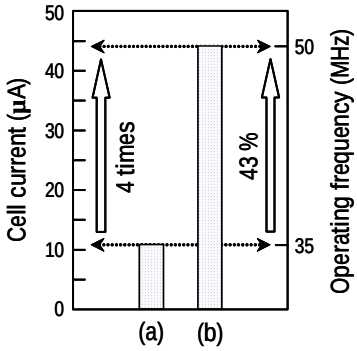


Fig. 12. Measured cell read-out current at V_{DD} = 0.8 V: (a) without cell boosting, (b) V_{PPWL} = 1.2 V and V_{PPCPL} = 1.6 V.

2.26 mm^2. Fig. 9 shows the relation between V_{DD} and cycle time. The SRAM achieves a 50-MHz operating frequency at 0.8-V power supply. The power dissipation is 65 μW/MHz. Fig. 10 shows the measured waveform of SRAM data out for 40 ns clock cycle.

Fig. 11 shows the measured butterfly curves for both conventional and proposed bias technique. At the supply voltage of 0.8 V, the read SNM is found to be ~160 mV in the conventional no boosting cell. By boosting the voltage level of WL and CPL to 1.2 V and 1.6 V respectively, the read SNM is drastically increased to ~400 mV. It also increases the cell read-out current about 4 times as shown in Fig. 12. The improvement on SRAM operating frequency is 43 % owing to the increase of read cell current. It was confirmed by the circuit simulation.

The primary benefit of the proposed technique is to reduce bit failure induced by read and write margin. It has been also measured from fabricated chips as shown in Fig. 13. At 0.8-V operation, about 87 % reduction in number of fail bits has been achieved by boosting the cell dually.

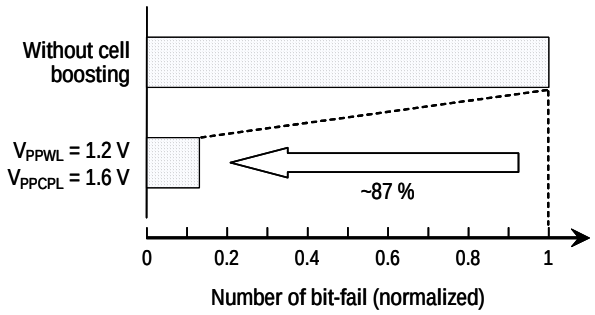


Fig. 13. Measurements of bit-fail count at V_{DD} = 0.8 V.

4. Conclusion

In order to improve the cell stability and the SRAM circuit speed encountered with low voltage SRAM, we have proposed a dual-boosted cell-array technique which can enhance both cell SNM and cell read-out current. For each read/write cycle, the wordline and cell power node of selected SRAM cells are internally boosted into $1.5V_{DD}$ and $2V_{DD}$, respectively. A 256-Kbit SRAM test chip with the proposed scheme has been fabricated in a 0.18-µm CMOS process, and demonstrated: 1) 0.8 V operation with 50 MHz while consuming a power of 65 µW/MHz; 2) 400 mV read SNM and 44 µA cell current at 0.8 V power supply; and 3) a reduction by 87 % in bit-error rate and 43 % higher chip operating frequency compared with that of conventional SRAM. Since the memory chip yield is often determined by the failure rate of memory cells, the proposed technique will be able to provide a significant improvement in the manufacturing die yield.

Acknowledgment

This work was supported by the Korea Research Foundation Grant (KRF-2006-331-D00321) and the IC Design Education Center in Korea.

Reference

[1] K. Itoh, K. Sasaki, and Y. Nakagome, "Trends on low-power RAM circuit technologies", *Proceedings of The IEEE*, vol. 83, no. 4, Apr. 1995, pp. 524-543.

[2] E. Seevinck, F. J. List, and J. Lohstroh, "Static-noise margin analysis of MOS SRAM cells", *IEEE J. Solid-State Circuits*, vol. SC-22, no. 5, Oct. 1987, pp. 748-754.

[3] H. Morimura and N. Shibata, "A step-down boosted-wordline scheme for 1-V battery-operated fast SRAM's", *IEEE J. Solid-State Circuits*, vol. 33, no. 8, Aug. 1998, pp. 1220-1227.

[4] M. Yamaoka, K. Osada, and K. Ishibashi, "0.4-V logic-library-friendly SRAM array using rectangular-diffusion cell and delta-boosted-array voltage scheme", *IEEE J. Solid-State Circuits*, vol. 39, no. 6, Jun. 2004, pp. 934-940.

[5] K. Takeda et al., "A read-static-noise-margin-free SRAM cell for low-VDD and high-speed applications", *IEEE J. Solid-State Circuits*, vol. 41, no. 1, Jan. 2006, pp. 113-121.

[6] K. Zhang et al., "A 3-GHz 70-Mb SRAM in 65-nm CMOS technology with integrated column-based dynamic power supply", *IEEE J. Solid-State Circuits*, vol. 41, no. 1, Jan. 2006, pp. 146-151.

[7] M. Yamaoka et al., "90-nm process-variation adaptive embedded SRAM modules with power-line-floating write technique", *IEEE J. Solid-State Circuits*, vol. 41, no. 3, Mar. 2006, pp. 705-711.

[8] T. Tanzawa and S. Atsumi, "Optimization of word-line booster circuits for low-voltage flash memories", *IEEE J. Solid-State Circuits*, vol. 34, no. 8, Aug. 1999, pp. 1091-1098.

A Noise-margin monitor for SRAMs

Peter Geens and Wim Dehaene

KU Leuven, ESAT-MICAS, Kasteelpark Arenberg 10, B-3001 Heverlee, Belgium
{peter.geens,wim.dehaene}@esat.kuleuven.be

Abstract

In current and future technologies leakage components play a key role in the total power-consumption of circuits. For SRAMs, this effect is even more noticeable due to the large portion of "non-active" cells in the matrix. To resolve this, a secondary sleep supply has been introduced into the system with savings of up to a factor of 10 in power-consumption [1][2]. In sub-100nm technologies inter and intra-die variability has a high impact on SRAM performance. In these highly variable environments the minimal voltage at which data can be reliably stored differs from die to die and even cell to cell. This work presents a monitor and regulation system that guarantees the lowest possible value for the "sleep"-voltage while maintaining data with an externally controlled minimal noise margin. This system enables die to die minimisation of the leakage currents, while maintaining the stored data, without costly calibration after manufacturing.

1. Introduction

The evolution towards making systems more mobile, by increasing battery life and the simultaneous demand for more functionality, has made power consumption the key specification for digital electronics design. Not only is the active power consumption under close investigation, the stand-by consumption has become the subject of research too. In the newly emerging bulk CMOS deep-submicron technologies needed to meet the functionality requirements, stand-by power consumption is dominated by leakage currents.

In the current state-of-the-art systems-on-chip, SRAMs are taking up the bulk of the area. This evolution makes that the power consumption is dominated by the SRAM and its interfacing. In stand-by mode the leakage currents originating in the SRAMs are the defining factors in the stand-by power consumption, due to the high number of leaking devices. It is imperative to control and reduce those leakage currents to be able to fulfil the stringent power consumption specifications for mobile communications.

In recent CMOS technology, there is a trend in decrease of the bulk-effect and increase of the Drain Induced Barrier Lowering (DIBL) effect on subthreshold leakage currents. The thinning of the gate oxides with every generation has also increased the gate-leakage currents [6]. This leakage component depends exponentially on the gate voltage. The relation between supply voltage and leakage currents has been reported earlier many times, e.g. in [1]-[5]. Both subthreshold and gate leakage currents depend exponentially on the supply voltage as both the voltage on the gate and across the transistor are linked closely to the supply voltage. This relationship can be exploited in circuits, including SRAM, to reduce the leakage currents by introducing a lower supply.

With almost every new technology generation the number of metal layers increases. This in turn facilitates routing more signals and supplies across the chips.

These evolutions have made introducing a secondary lower supply into the SRAM matrix an attractive and feasible option to reduce leakage currents during stand-by phases. Whether this second supply is referred to as a sleepy or drowsy supply [1][3][10] is of no consequence to the goal that is pursued: minimising the leakage currents through the exponential relation between supply and leakage [1] while maintaining the data in the matrix. How to find this point of minimum leakage and a novel way to do so on chip are the subject of this paper.

The paper is organised as follows. First the currently known methods to find the "sleepy" supply voltage in open literature are described. The mathematical foundations of the proposed solution are explained in the part thereafter, including the discussion on the bit-integrity measures that will be used. The algorithmic implementation of the theory and its optimisation will be the subject of the fourth part. In the final part the conclusion will be drawn and future work will be announced.

2. Known solutions

Design time solutions.

Based on elaborate Monte-Carlo simulations or mathematical derivations of the minimum hold voltage for a cell, a sleepy supply voltage can be found that guarantees holding the data while leakage reduction is maximised. [15]

However, in today's deep submicron technologies the influence of process variations has a major impact on device characteristics and hence also on the full system. In realistic environments such variability can not be ignored. This results in large safety margins to be taken into account to have reliable operation and yield, when worst case scenarios are considered. These margins translate to a higher applied sleepy voltage, leading to a more stable cell but less reduction in leakage currents. Because the margins have to compensate for the worst case and can not be altered later; there will be a large number of cases where more reduction would have been possible.

Time-dependent variations such as temperature or over time degrading of the materials can not be compensated without further increasing the margin. This eventually leads to almost always a suboptimal solution with regards to leakage reduction. Using a statistical approach as published in [12] can reduce the designed overhead partly.

Solutions based on calibration

A first step in having a higher reduction in leakage currents while maintaining data, is to have a calibration step during testing to find an optimal sleep supply voltage. This has to be done on die to die basis which would increase the, costly, test time. The benefit to this approach is the higher savings on leakage that can be achieved and allows every die to be near its optimal performance point under controlled circumstances. Again, time-dependent variations of the systems performance are not compensated.

Real time on-chip solutions

The above solutions still have one major weakness, the lack of being able to compensate for time-dependent variations. It is clear only an on-chip technique can give a solution in this area.

One such way would be to implement a series of banks containing a significant replicated part of the system, that could foretell failure depending on supply levels and, as they are part of the same system, time-dependencies or process variations. In [9] this is done for the flip-flops of a digital system under the name of "canary flip-flops". A similar system could easily be derived for SRAMs where the banks could consist of matrix cells.

While such a monitor has its merits in achieving some form of real-time control, there is the issue whether it is representative for the behaviour of the system, more specifically the matrix. These banks would namely be designed to be different from the matrix-cells, both in form and function. In processes where not only the inter-die variation or process corners, but also the intra-die variation is of great importance both circuits will react differently on the same variations. This is a source for under-performance in leakage reduction as it can only be compensated by increasing the margin on the applied voltage.

3. Proposed solution

General concept

If the system has to be able to compensate for process variations but also time-dependent variations reliably, a real-time monitoring solution is needed that gives more than a go-no go decision.

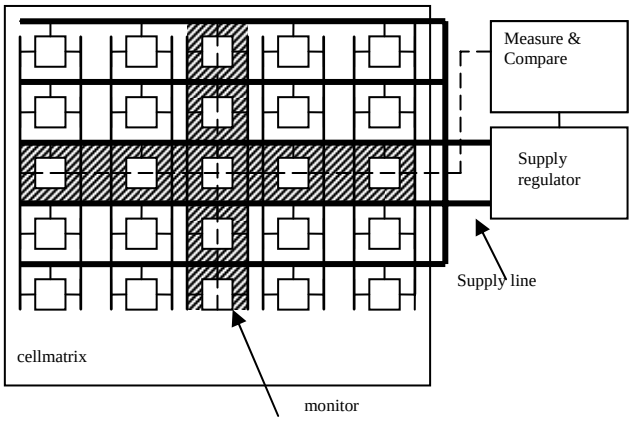


Fig. 1 system overview

The system depicted in Fig. 1 consists of 2 main parts, an observable entity for the bit integrity parameter and a measurement part to extract the value of the parameter that in turn returns a reference signal for the generation of the secondary supply. This reference signal can be used for a DC-DC converter or any other supply regulation circuit.

The system has an external reference input that sets the minimal value the bit integrity parameter should have.

Bit integrity parameter.

The Static Noise Margin (SNM) as defined by Seevinck [8] has been the standard bit integrity parameter for read conditions for many years. The SNM under hold (SNMh) would be the logical extension of this definition (Fig. 2) to measure data retention capability under non-access conditions. It could be measured with the same setup as the traditional SNM with the only difference being the off-state of the pass-transistors (Fig. 3).

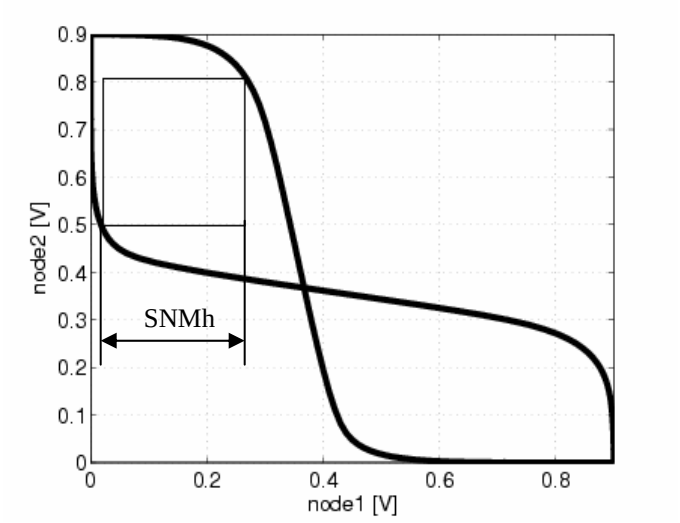


Fig. 2 butterfly curve under hold

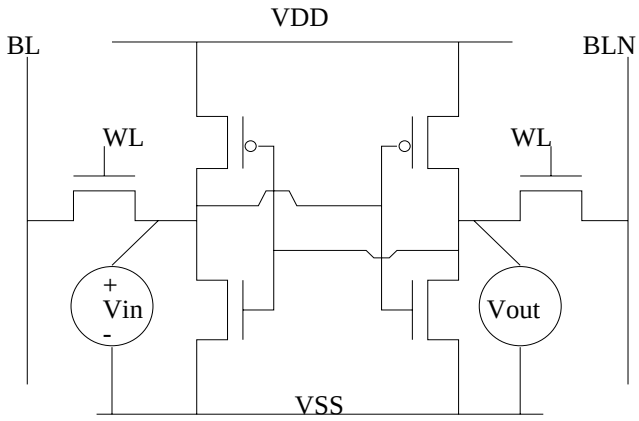


Fig. 3 measurement setup

Theoretical background

SNMh is a mini-max criterion for the eye-opening of the butterfly curve as depicted in figure (butterfly). Let f and g be the functions that describe the DC-transfer curves of the cell-invertors. In this case f_{45} and g_{45} are f and g rotated over in the coordinate axes rotated over 45 degrees. The eye-opening function, h, would then be equal to formula (1)

$$h = f_{45} - g_{45} \tag{1}$$

SNMh is an extreme of this function, which can be found by deriving the function h to x. This in turn leads to the conclusion that SNMh can be found at the point where the derivatives of f_{45} and g_{45} are equal (form. (2)).

$$\frac{dh}{dz} = \frac{df_{45}}{dz} - \frac{dg_{45}}{dz} = 0$$
$$\Downarrow \tag{2}$$
$$\frac{df_{45}}{dz} = \frac{dg_{45}}{dz}$$

This requirement is invariant under rotation, so the SNMh can be measured on the butterfly curves in the points where the derivative of both curves is equal, x1 and x2 on Fig. 4.

4. Implementation

Implementation of the algorithm

An implementation of the mathematical derivation, as described in the previous section, is feasible but can be optimised further. It has to be noted that an absolute equality can never be reached when depending on analog measurements.

The actual implemented algorithm as depicted in the following flow-graph (Fig. 5), consists of 2 main stages. The first one will search the point on the second curve over a 45 degree translation corresponding with the current measurement point. The second stage implements a binary search algorithm to find the actual maximum difference. This difference is the SNM. One optimisation step is further added. When a measured value returns a value that is bigger than the reference value supplied to the system, it is clear the voltage on the cells can be lowered, even if the actual SNMh hasn't been found yet.

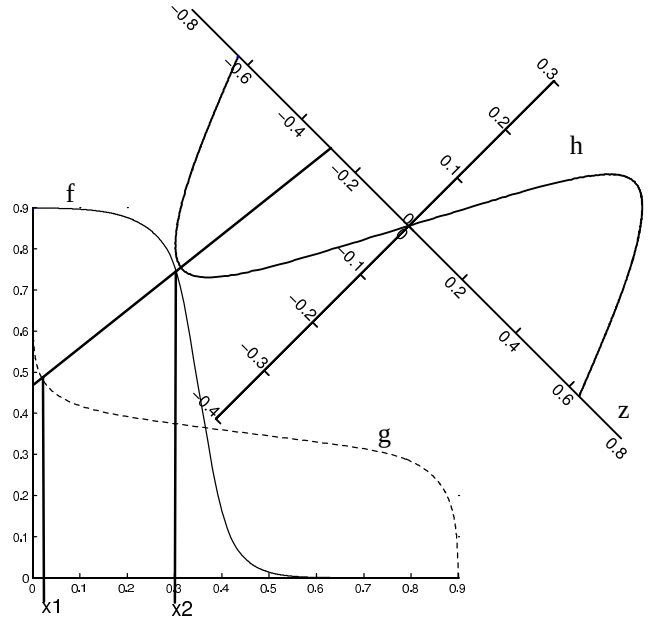


Fig. 4 butterfly transformations

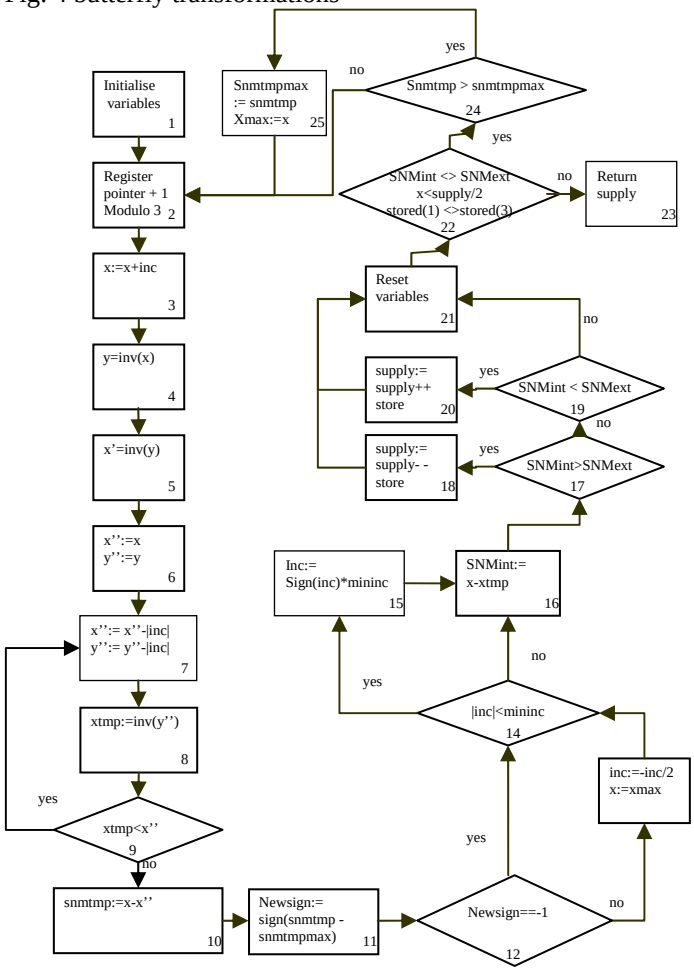


Fig. 5 algorithm flow graph

The algorithm can run on dedicated hardware or use spare cycles of an already present microprocessor. Its accuracy will greatly depend on the measured values. The biggest influence on this precision will be the reliability and representativeness of the monitor cell.

Implementation of the monitor cell

In the high variability environments this system has to operate, knowing the nominal point of operation of the die is of utmost importance. In a first step this requires the cells that are going to be monitored to be as close as possible to the actual matrix cells, both in circuit-behaviour as geometrically. In a second step intra-die process variations should be compensated. Pelgrom's law [11] states, that mismatch of a device parameter is inversely proportional to the square root of the area of the device. Increasing the cell area of the monitor cells is however not an option as it would contradict the first step. The solution to having a large area of the devices and preserving their geometric similarity is to put many monitor cells in parallel, see hashed region on fig. 1. In that way their layout and DC-characteristics stay the same but benefit from the averaging over a large area. This is a technique that has been successfully employed in high precision current-steered digital-to-analog converters.[14]

This approach reduces the variability on the monitor cells by the square root of the number of cells in parallel and hence allows having reliable measurements.

A margin to accommodate for the difference between the actual cells and the monitor cells has to be included into the reference SNMh. This margin can be calculated based on the variability data obtained from the process vendors. Due to the use of the above monitor configuration this margin is a square root of 2 smaller than in the case of a single monitor cell.

5. Conclusion

This paper presents a novel noise-margin monitoring system for SRAMs that allows maximising the reduction of leakage power consumption. By implementing a monitor cell that is geometrically and electrically close to the matrix cells, combined with a real-time measurement and control system, it is possible to minimise the sleep supply voltage while maintaining the data integrity. The mismatch of the monitor circuit is reduced so it is representative for the nominal operation point of the system by using several cells in parallel in accordance with Pelgrom's law.

Future work will consist of an implementation of this system in a 90nm CMOS technology, including an on-chip solution to generate the secondary supply.

6. Acknowledgements

The systems presented in this paper are subject to the pending parent [13] This research was sponsored by the TAD project within IMEC, Belgium.

References

[1] P. Geens, W. Dehaene, "A small granular control system for SRAMs", Journal of Solid-State Electronics, vol. 49, November 2005, pp 1776-1782.

[2] F. R. Saliba, H. Kawaguchi, T. Sakurai, "Experimental verification of Row-by-Row Variable VDD Scheme Reducing 95% Active Leakage Power of SRAMs", Symposium of VLSI Circuits Digest of Technical Papers, 2005, pp. 162-165

[3] K. Roy, S. Mukhopadhyay, H. Mahmoodi-Meimand, "Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits", IEEE Proceedings of the IEEE, vol 91, no. 2, February 2003, pp. 305-327

[4] M. Drazdziulis, P. Larson-Edefors, D. Eckerbert and H. Eriksson, "A power Cut-Off Technique for Gate Leakage Suppression"', IEEE proc. of ESSCIRC 2004, pp. 171-174

[5] K. Nii, Y. Tsukamoto, T. Yoshizawa, S. Imaoka, Y. Tamagami, T. Suzuki, A. Shibayama, H. Makino and S. Iwade, ``A 90-nm Low-Power 32-kB Embedded SRAM With Gate Leakage Suppression Circuit for Mobile Applications", IEEE Journal of Solid-State Circuits, vol. 39, no. 4,april 2004, pp. 684-693

[6] R.W. Mann et al., "Ultralow-power SRAM technology", IBM Journal on Research \& Development, VOL.47, NO. 5/6, September/November 2003

[7] N. Kim et al., "Circuit and Microarchitectural Techniques for Reducing Cache Leakage Power", IEEE Trans. on VLSI, vol. 12, no. 2, feb. 2004

[8] E. Seevinck, F. List, J. Lohstroh, "Static Noise Margin Analysis of MOS SRAM cells", IEEE Journal of Solid-State Circuits, vol. sc-22, no. 5, october 1987

[9] B. Calhoun, A. Chandrakasan, "Standby Power Reduction Using Dynamic Voltage Scaling and Canary Flip-Flop Structures", IEEE Journal of Solid State Circuits, vol. 39, no.9, september 2004

[10] K. Zhang et al., "SRAM design on 65-nm CMOS Technology With Dynamic Sleep Transistor for Leakage Reduction", IEEE Journal of Solid State Circuits, vol. 40, no 4, April 2005

[11] Pelgrom M.; Duinmaijer A; Welbers A, IEEE Journal of Solid-State Circuits, "Matching properties of MOS transistors",Vol 24, no. 5, October 1989,pp 1433 - 1439

[12] E. Grossar, M. Stucchi, K. Maex, W. Dehaene, "Read Stability and Write-Ability Analysis of SRAM cells for Nanometer Technologies",IEEE journal of Solid State Circuits, vol. 41, no 11, November 2006,pp 2577-2588

[13] EU patent submission, 22th January 2007

[14] Miki T., et al., "An 80 MHz, 8-bit CMOS D/A converter", IEEE Journal of Solid State Circuits, vol. 21, no. 6, June 1986, pp 983-988

[15] Hulfang Q.; Yu C.; Markovic, D.; Vladimirescu, A.; Rabaey, J., "SRAM leakage suppression by minimizing standby supply voltage", IEEE Proc. Of 5th International Symposium on Quality Electronic Design, 2004, pp 55 - 60

A Variability Tolerant Embedded SRAM Offering Runtime Selectable Energy/Delay Figures

Hua Wang [a,b] , Miguel Miranda [a] , Peter Geens [b] , Wim Dehaene [b] , Francky Catthoor [a,b],

[a] IMEC, Kapeldreef 75, Leuven, 3001, Belgium; wanghua, miguel,catthoor@imec.be
[b] K.U.Leuven, Kasteelpark Arenberg 10, Leuven, 3001 Belgium; pgeens, wim.dehaene@esat.kuleuven.be

Abstract

This paper presents the circuit and layout design of an 8KB low power self-timed embedded SRAM capable of providing at runtime Energy/Delay (E/D) trade-offs that are robust to variability effects. Experimental results from post-layout extraction at 130nm technology have shown that the implemented SRAM can effectively provide a trade-off range of about 40% in both energy and delay with low area overhead. Such range can be utilized at runtime by the system to help reduce the overall system energy consumption.

1. Introduction

Trends in miniaturization and autonomy in future technologies (e.g., bio- and nano-technology) will increase the need for Ultra-Low Power (ULP) SoC while ensuring reliability of their operation. The most energy/delay critical components in modern SoCs are embedded memories, both for data and instruction/configuration storage [1]. Given the dynamic workload conditions in most applications running on the same system, it is therefore essential to enable the scalability of power/performance figures within the individual blocks, e.g., Layer-1 (L1) memories, to just meet the varying application timing constraints with lowest energy consumption. In this way, the limited battery life can be maximally utilized to improve user experience. To achieve this, it is important to gracefully trade energy and execution time while executing the application. In data-path design, this is well handled by applying the techniques like dynamic Vdd scaling. But for SRAM design, this is not that obvious. In this paper, we will discuss the low level design of an SRAM where the trade-off is achieved by introducing ``knobs" in the memory. This allows the application to control the performance of the memory, hence saving energy when faster than needed nominal execution times are not needed for real-time operation. For instance, when used together with system level dynamic task scheduling techniques, different tasks of the application can be mapped to either high speed functional units (FU) and SRAMs or low power energy efficient FU and SRAMs depending on the timing constraint so that the overall system energy consumption can be effectively reduced.

"Knobs" mostly seen in current state-of-the-art SRAMs are the supply voltage tuning [2] and back-gate biasing controlling techniques [3]. Indeed, they are effective in trading the performance for the dynamic and/or static power consumption of the circuits. However, as technology scales down, the margin available for these parameters is clearly decreasing [2, 3], thus leaving little room for the E/D trade-off. For instance, the diminishing body effect through scaling is bringing more difficulties in the threshold voltage control techniques. On the other hand, Vdd tuning still remains powerful on leakage power control, up to a factor of 10 saving in SRAM has been observed in previous work [4]. This is range is much smaller for dynamic energy savings due to the fact that Vdd has to be high enough to avoid the impact from process variability effects. Therefore, we believe that other novel techniques have to be added on top of these conventional "knobs" to maintain a good SRAM E/D trade-off range at nanometer technology nodes.

For typical small size SRAMs, which are largely located in the L1 layer, we have observed that the memory cell matrix is not the only dominant component of their dynamic energy and delay. The peripheral circuitry contributes considerably to these two figures. It has been reported that for small size SRAMs (<128Kbit) the decoder and the wordline buffers are responsible for about half of the energy and delay of the memory [5]. Interestingly, the buffers present in these blocks are the actual circuits that account for this amount of delay and energy. This is simply due to the fact that nearly all the fanout (interconnect capacitance and logic gate capacitance) of the decoder and wordline driver have to be driven by them. Obviously, bringing in E/D trade-off in these buffers can surely enable a good trade-off range for the SRAM. Because of this, we believe that applying the runtime Pareto buffer design techniques [6] onto such buffers would help to produce satisfactory trade-offs at the SRAM level. By Pareto buffer we understand a variable tapered buffer configuration such that for any given energy budget no other configuration exists offering smaller delay. At the same time for any given delay constraint no other buffer exists offering lower energy than the Pareto optimal configuration [6]. In this sense, the SRAM equipped with such type of buffer becomes a runtime configurable memory capable of operating at several optimized energy/delay points. Note that the area of SoCs are mainly dominated by those L2 and L3 memories, hence the extra area overhead introduced by the L1 runtime configurable SRAMs can be largely tolerated at the SoC level. Up till now, this is the first work that reports the low level design of a runtime configurable SRAM in the low power context. Previous researches on this subject are mostly done at the reconfigurable architecture level targeting at computation efficiencies [7].

Apart from the challenge of the increasing low power need, the advance of process technology also brings new uncertainty issues to the circuit design practice. One of such problems is the random process variability effects

caused by the difficulties of accurately controlling process parameters, e.g., doping profile, line edge roughness, etc, when manufacture the chips in Deep Submicron (DSM) era. Such imperfections directly increase the un-predictability of transistor threshold voltage and drain currents. Hence the random variability effect greatly influences the circuit's functionality as well as parametric figures [8]. Embedded SRAM is one of the victims suffering such effect. Previous work already shows, apart from functional problems, SRAM delay and energy can greatly drift away from their nominally designed points [8]. A conventional way of tackling such drift is to confine it by adding design margins such that the circuit parameters only vary within a tolerable small range [9]. However, such design method has the drawback of inducing significant amount of overhead in performance, power and area when variability increases [9, 10]. Although margin-based design will survive in a limited range for future DSM circuit design, we feel the necessity to improve circuit's tolerance to variability effect via other design techniques. On top of this, we believe that enabling the SRAM circuit to self adapt to its internal parametric variation can effectively improve their robustness to the uncertainty effects and eventually ensure a functionality correctness. Therefore, we have decided to introduce self-timed control units and interface into our SRAM design.

The following sections describe the design of SRAM architecture plus some important sub-circuits and end up with experimental results and a conclusion.

2. SRAM architecture and timing

The SRAM we used is typically found in the L1 hierarchy of the low power embedded systems. In our case, it assumes a typical bit-width of 16-bit wide with one matrix composed of 256 rows and 256 columns.

As shown in Figure 1, the SRAM is essentially composed of the row and column decoders, cell matrix, wordline drivers and timing control circuit plus self-timed interface. Apart from this, external latches are attached at the address and data I/O to store information as well as to shield SRAM internal circuits from the external dynamic address lines. Besides, additional delay measurement circuit and calibration circuit are also added to help benchmark the design on silicon. The post silicon energy measurement of the SRAM will be carried out via the power supply pins for the internal blocks.

The row decoder and column decoder are both designed using static CMOS logic to avoid floating nodes that are vulnerable to leakage problem. Moreover, given the number of wordlines (256 in total) to be decoded, a two-stage NOR-NAND decoding style is used for the row decoder to reduce fanin. In this way, the addresses are pre-decoded via the NOR gates and dispatched via the buffers after the pre-decoder to activate one of the NAND gates responsible for asserting its associated wordline. Obviously, a large fanout composed of interconnect capacitance and logic gate capacitance is present for the predecoder buffers. As SRAM does not always have to operate at full speed due to application timing variations, conventional high speed buffers used to drive the load will lead to energy overheads when the high speed is not necessary. Therefore, instead of regular CMOS buffers, two-option runtime configurable Pareto buffers offering E/D trade-offs are used to drive such load. This is also the case for the wordline drivers where large load from the wordline and pass transistors of the cells is also expected.

To even more reduce dynamic energy consumption and improve access time in the matrix, a hierarchical wordline is used which consists of global wordline and sub-wordlines. In addition, leakage power in the matrix is suppressed following state-of-the-art voltage scaling techniques on those standby cells [4]. In this way, only the accessed cells are supplied with a normal voltage. For the rest, a sleep voltage is applied that significantly reduces the standby power.

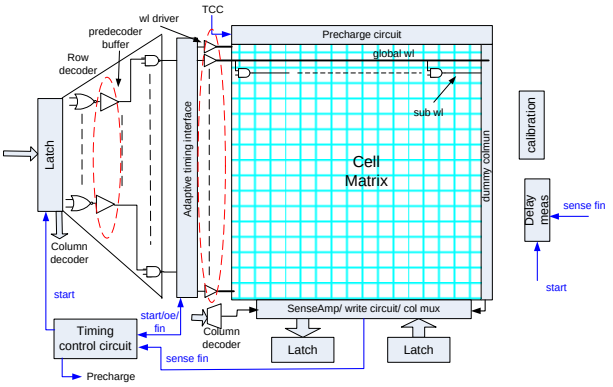


Fig.1 SRAM architecture

The overall access of the SRAM is fully controlled via the timing control circuit (TCC) together with the adaptive self-timed interface in between row decoder and wordline buffers. The self-timed flavour of the memory is depicted in Figure 2.

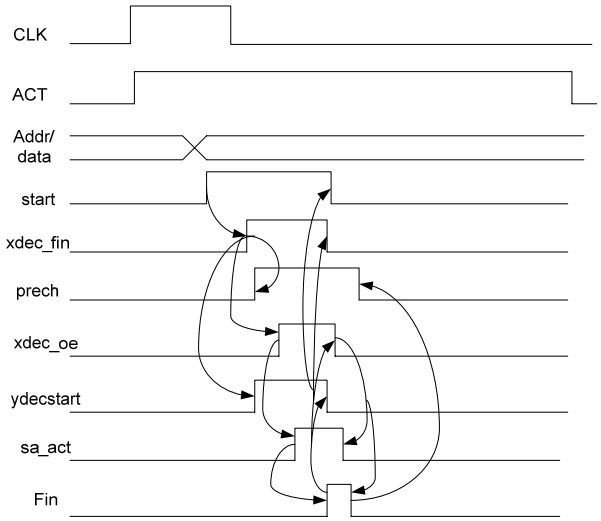


Fig.2: Timing diagram of the SRAM in a read cycle

In a read cycle, the external clock and activation signal are sent to the timing control circuit indicating a new cycle of access. The TCC acknowledges this event by asserting the access start event which then latches the input address and resets the timing interface. Row and column decoding processes start automatically until one

of the postdecoder has the decoded output (active low). Given its simpler and smaller structure, the column decoder usually decodes way faster than row decoding hence its delay is shadowed in the row decoder delay. The timing interface detects this event and informs the TCC that the row decoding process has finished. TCC then disables the precharge signal on the corresponding column selected already by the column decoder and approves the decoder output enable event back to the interface, which is sent via a runtime configurable Pareto buffer. On receiving this approval, the interface enables corresponding wordline driver hence access to the matrix starts. A dummy column at the other of end of the matrix monitors the swing on the bitline and enables sense amplifier (SA) at appropriate time. This enable signal, after some calibrated delay, is transferred as a sensing finished signal back to the TCC. The control circuit then disables decoder output enable signal hence grounds the active wordline, cuts off the column mux bridging SA and bitline and in the end restores the precharge signal. The control sequence of a write cycle is largely similar to the read one.

According to the control manner, it is obvious that performance variation in the row decoder as well as that in the matrix will largely not impact the functionality of the entire SRAM. Hence its robustness is assured for process variability and runtime E/D operation point reconfiguration.

3. Design of key circuits

The runtime configurable Pareto buffers present in the row decoder, wordline driver and decoder output enable buffer are implemented to be able to provide two different driving capabilities (high speed and low power) following the style as shown in Figure 3. In fact, the implementation is composed of a regular low power inverter chain (lower shaded branch) and a special inverter chain terminated by a tri-state driver (upper branch). The low power part is always operating while the upper branch is controlled via the runtime switchable enable signal to operate only in high speed mode. In that case, the two branches together realize the driving capability of a regular high speed inverter chain. Such implementation has the advantage that the normal high speed buffer sizes are shared between the lower and upper branches. In this way, the extra load from the tri-state driver present to the low power part is small and thus only leads to small energy and delay overhead for that part. Moreover, such an implementation has small area overhead (about 5% at layout level) compared to a conventional high speed buffer normally found in such places. This is indeed negligible at the SoC level. The actual configuration decisions (number of stage and associated sizes) on such runtime configurable Pareto buffer is obtained following the methodology depicted in [6], where a set of Pareto optimum energy/delay configurations is first explored according to the extracted load condition and two of them are then selected for the final implementation.

A sub-block of the delay variation tolerant adaptive timing interface after the row decoder is implemented as shown in Figure 4. In total 256 such blocks are attached at each output of the row decoder. However, only one of them will be activated during access hence consuming only a small amount of energy. As previously mentioned, such interface is used to detect the decoded signal from the row decoder and communicate with TCC to help disable/enable a set of important signals so as to make sure SRAM access is always carried out in a reliable way. In the end, it fulfils this requirement by operating in an event-based manner.

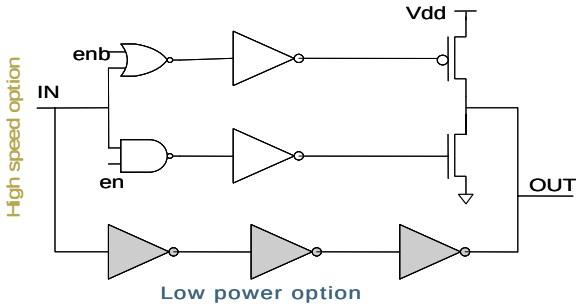


Fig. 3: Runtime configurable buffer implementation

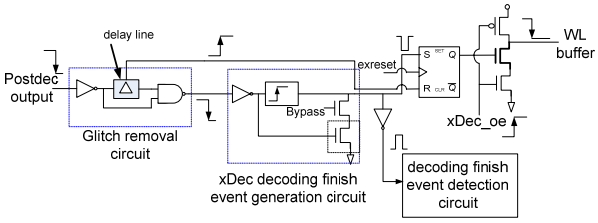


Fig.4 Adaptive timing interface

As shown in the schematic, the decoded signal from the post row decoder will first be filtered through a glitch removal circuit to protect following event detection circuits. The amount of glitch to be removed is obtained via Monte-Carlo simulations at several process corners. Passing through the glitch removal circuit, the decoded signal arrives at the row decoding finish event generation circuit which generates a short pulse upon the falling edge of the decoded signal. This short pulse is used to resemble the decoding finished event to the detection circuit. Meanwhile, it sets the output of the following SR latch (see Fig.4) to prepare enabling wordline driver when the decoder output enable is approved by the TCC. The decoding finished event generation circuit also has a bypass input to help generate the event pulse when the same row decoder output is accessed in consecutive cycles. In this situation, the input addresses will not change hence no decoded falling edge will be generated at the output of the post row decoder. For this, an address transition detection (ATD) circuit at the input of the decoder will generate the bypass signal upon the decoding start event hence the finished event pulse will still be generated only at the same output.

The detection of the decoding finished event from the 256 outputs is carried out in a special finish event detection circuit implemented using the fast wired-OR structure as shown in Figure 5. There, the final decoding finished pulse is generated after two levels of detection. Such levelization helps reduce wiring complexity and improve detection speed. After the final decoding finished signal arrived at TCC, the precharge on the corresponding column is disabled in the matrix and

decoder output is enabled thus activates the corresponding wordline driver.

As previously mentioned, the deactivation of the wordline driver is carried out when TCC received the sensing finished event from the matrix.

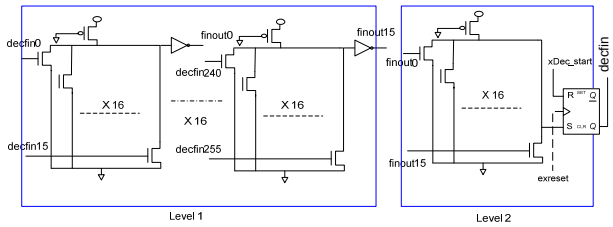


Fig.5: Decoding event detection circuit

4. Experimental results

The design of the SRAM is finally implemented with our in-house IMEC 130nm platform technology for tape-out as shown in Figure 6. Due to process deadline issues, the layout was not fully optimized. However, it does not significantly affect the effectiveness of the overall design method as well as the final results. This is already confirmed in a second design that is on-going at this moment.

Transistor level simulation with extracted parasitic from the layout has been performed at the full SRAM level. Results have shown that the SRAM can robustly operate at 450MHz with 3.6mW power consumption as well as 310MHz with 1.72mW power consumption. From the energy/delay trade-off point of view, the implementation offers a good range of about 40%. These two different operation modes can be selected on the fly in between memory accesses via the runtime configurable buffers. The adaptive timing interface plus the timing control circuit can effectively ensure a correct SRAM functionally through the changes. The area overhead induced by timing control interface is less than 5% according to the final layout estimation.

Given this type of SRAM, high level scheduling techniques can take full advantage of it, and together with the dynamic task mapping techniques, to even more reduce overall power consumption of the embedded system[10].

5. Conclusion

This paper presents the design of a variability tolerant embedded SRAM offering very good E/D trade-off operating points that can be selected at runtime. The SRAM utilizes the runtime configurable buffer together with a self-timed timing interface to achieve robust operation under varying performance figures. A range of about 40% in both energy and delay is obtained via the implementation at 130nm technology. The achieved range in SRAM, coupled with system level techniques, can increase the dynamic E/D operating range of the embedded system hence helps to aggressively suppress system energy consumption.

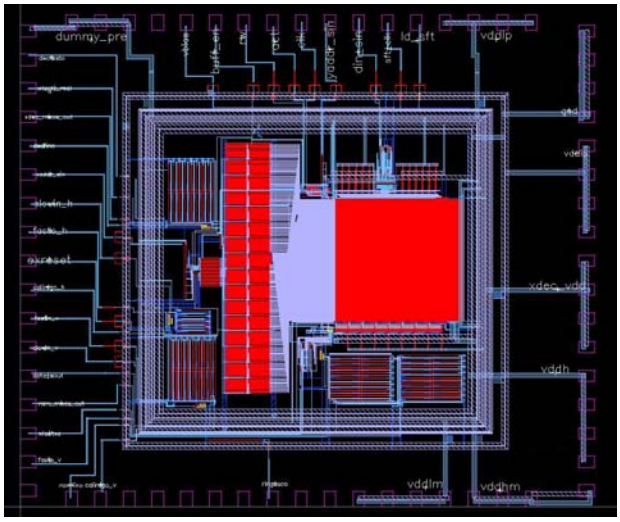


Fig.6 Layout of the SRAM design

References

[1] L.Benini, et.al, “System-level power optimization techniques and tools”, ACM Trans. on Design Automation for Embedded Systems,2000
[2] A. Srivastava,et.al, “Concurrent Sizing, Vdd and Vth Assignment for low power design”, Proc. DATE,2004
[3] V. Arnim, et.al, “Efficiency of body biasing in 90nm CMOS for low power digital circuits”, Proc. ESSCIRC,2004
[4] P.Geens, et.al, “A small granular controlled leakage reduction system for SRAMs”,Proc. ICMTD,2005
[5] B. Amrutur,et.al, “Speed and power scaling of SRAM”, JSSC,vol.35, issue 2, 2002
[6] H.Wang,et.al, “Variable tapered pareto buffer design and implementation allowing run-time configuration for low-power embedded SRAMs”, TVLSI,2005
[7] K.Mai,et.al, “Smart memories:A modular reconfigurable architecture", Proc. ISCA,2000
[8] H.Wang, et.al, “Systematic Analysis of Energy and Delay Impact of Very Deep Submicron Process Variability Effects in Embedded SRAM Modules”, Proc. DATE,2005
[9] R. Heald, et.al, “Variability in sub-100nm SRAM designs”, Proc.ICCAD,2004
[10] A.Papaniko,et.al, “A system level methodology for fully compensating process variability impact of memory organizations in periodic applications”, Proc. CODES,2005

Protection of embedded memory systems - a comprehensive solution

Riccardo Mariani[a], Federico Colucci[a], Peter Fuhrmann[b]

[a] Yogitech SpA, via Lenin 132/p, 56017 San Martino a Ulmiano, Pisa, Italy
[b] Philips Research Laboratories, Aachen, Germany

Contacting author: riccardo.mariani@yogitech.com

Abstract

Protection of embedded memories against faults brings costs into the loop such area, performance and/or power overheads. Complexity of modern systems requires having flexible protection schemes and architectural options capable to offer a wide set of tradeoffs to the chip architect and to the system engineer. This paper describes a fault supervisor mastering the flexibility in protection schemes and architectural variants for embedded memories. Twos solutions are presented especially suited for multi-banks and multi-masters memory systems, such a "two-memories"/"shared" architecture allowing a flexible partitioning of memory in different regions with selectable protection levels, and a distributed memory protection unit protecting the memory against SW faults. The fault supervisor has been certified by TÜV-SÜD in accordance with IEC 61508 norm for safety related electronic systems. Results are based on a reference 32-bit MCU platform for automotive applications by NXP.

1. Introduction

Major types of faults affecting embedded memories are permanent faults such stuck-at, bridging, transient faults such soft-errors [1], addressing faults, delay faults and other faults such data retention (for non volatile memories), modelling uncertainties and so forth. Concerning radiation effects, trends are for 1350 FIT/Mb for SRAM, 60 FIT/Mb for DRAMs [2].

Despite the huge literature concerning fault detection and correction, faults occurring in embedded memories are still a nightmare for many applications: automotive [3], biomedics [4], routers / workstations / data processing [5] and solid state disks [6]. In fact, with microcontrollers (MCU) becoming part of many safety-relevant systems, the integrity of the memory sub-system is no longer an availability aspect only, but related to safety requirements determined by the application and its criticality. In addition, the handling of embedded memory for cost-sensitive, demanding markets such automotive requires careful consideration of several aspects related to performance and overhead. Finally, the introduction of fault detection/correction and other integrity means for the memory sub-system should not compromise the scalability and flexibility requirements coming along with a microcontroller platform approach. MCU products and different members of a product family often vary in the amount of embedded memory and its organization. For a semiconductor manufacturer this means that the approach for protected embedded memory must support scalable memory arrays but also different memory configurations. For instance, memory architectures with multiple instances of SRAM arrays sharing a single memory controller might be driven from chip layout restrictions or other technology reasons. The multitude of possible configurations poses the strong requirement for a suitable memory supervisor to support this flexibility by architecture.

In the framework of the approach already presented in [7], this paper presents a fault supervisor for memory sub-systems and its improvement to fulfil the requirements of multi-banks and multi-masters memory architecture. The "two-memories architecture" allows the reduction of memory code overhead due to data protection by enabling the partitioning of memory into different regions with selectable protection levels. The "distributed MPU" access protection function located at the memory allows an efficient memory protection when multiple masters are accessing, i.e. sharing the same memory.

2. The fault supervisor

A schematic diagram of the fault supervisor and how it is connected to the memory sub-system (i.e. embedded memory plus memory controller) is presented in Figure 1. It uses Error Detection and Correction Coding (ECC) reinforced by additional techniques. Some of them, such scrubbing, have been already described in [1]: this paper concentrates on the improvements to fulfil the performance requirements and to support multi-banks and multi-masters memory architecture. The memory fault supervisor is composed by three functional units:

- F-MEM: it interfaces the memory array and it hosts the coder/decoder and the scrubbing features
- MCE: it interfaces the F-MEM with the memory controller and with the bus, providing a Direct Memory Access (DMA) for F-MEM scrubbing feature as also the MPU functionality
- F-IF: it provides the interface with the external sub-systems, e.g. the CPU, it includes the controller to generate the corresponding alarms, the logger and a configuration block.

It includes as well a feature named "fast-track" enabling highest operating frequency maintaining same level of ECC protection, i.e. no additional wait cycles are required. It is also very useful for memory systems with a low probability of failures or with a limited amount of data or instructions that should be executed with a higher level protection: in fact with fast-track, most of the application can be run at the highest speed.

With the fast-track, during a write access the data coming from the memory controller is written in a write buffer together its address. The coder in the next cycle

will take it and will write the data and proper code in the memory. During a read access, the data is read by the memory controller from the data memory without passing through the decoder: the decoder operates in parallel. If the decoder detects an error, it generates a "SW Recoverable Error": an error has been detected, but due to the presence of fast-track the system took the wrong data, so an error confinement or error decontamination procedure is necessary.

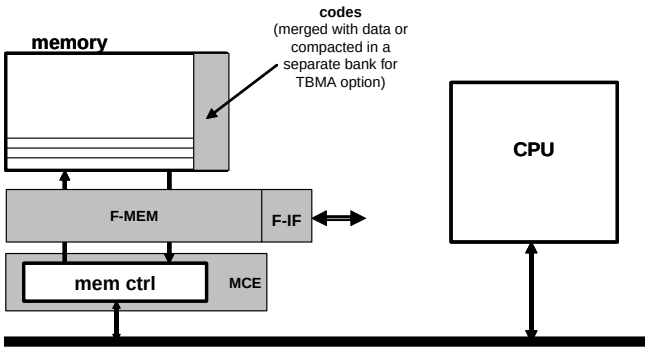


Figure 1: The memory sub-system and the memory fault supervisor

Error "confinement" is a mixed HW-SW procedure. This procedure avoids that the wrong data "contaminates" the bus master that requested such data (e.g. the CPU) by stopping the fetch of this data before it enters inside the processing unit. In other words, the error is "confined" at the periphery of the processing unit. This procedure allows a complete recovery of the hazardous situation. This procedure must be supported by a proper connection of the SW Recoverable Error. In many cases the implementation of an error confinement procedure is not needed, i.e. no recovery action is required. Error decontamination is sufficient, i.e. a procedure is implemented that accepts wrong data entering the master and that then acts in a way to "decontaminate" the processing unit from this error as much as possible. This procedure can be of different types: an interrupt is generated, and a dedicated except handler is executed to restart a full block (i.e. to repeat the algorithm) or to reset the system. Alternatively, the SW Recoverable Error alarm is read by an external diagnostic unit and this unit brings the system in a fail-safe or fail-silent configuration.

As a proof of concept, the fault supervisor has been used in a reference 32-bit MCU platform for automotive applications by NXP [8]. The main aspects related to the platform integration were related to the strict bus timing requirements including the signalling of detected faults as well as the data and signal consistency for the fast-track option. This integration resulted in three different implementations: two versions of the fast-track technique and one without but including configurable latency architecture of the fault supervisor with a "latency hiding" technique embedded to reduce the cycle penalty as much as possible. In the first "extreme fast-track" implementation, a proper connection was done between the fault supervisor and the CPU. With this connection, during a read access with errors, a prefetch (in case the read data was an instruction) or data (in case the read data was not an instruction) abort is automatically executed by the system processor. The prefetch and data exception handlers have been modified to implement a certain list of actions, which are anyway typical for exception routines. This modification is rather simple and consists in very few instructions.

The "extreme fast-track" implementation is possible only if the master has enough "intelligence" to handle such fault confinement procedure. If the master is simpler, like a DMA controller, two different architectures are used:

- a "light fast-track" implementation is used, where the detection and correction is not executed in parallel. In case of a single fault, the CPU will fetch the corrected data without any problem. In case of a multiple error, an unrecoverable flag will be issued and this can be sent to a central supervisor, i.e. a centralized fault supervision unit that can stop the execution and proceed with the proper failure control action. In such architecture, no error flag is needed to generate the error response and this allows a faster bus interface implementation.
- the memory fault supervisor is pipelined adding one or two pipeline stages. To avoid a systematic latency in the operation, a "latency hiding" architecture is implemented: a one/two cycles of penalty are introduced only in case of a non-sequential access to the memory sub-system, while in case of sequential accesses (i.e. burst accesses), no additional latency is introduced.
-
- Address faults (both stuck-at or address delay faults or similar) are one of the required fault models to be covered by the IEC 61508 norm for safety related electronic systems [9]. To fulfil that, the memory fault supervisor coder shall receive as input both data and addresses. During a read access, the decoder will compare the code bits extracted from the memory with the code bits computed "ex-novo" by using the data read from the memory and the address coming from the memory controller. However, such implementation of addresses protection comes with some extra costs in terms of fault supervisor gate count, access time and extra memory bits for each word needed to include the addresses in the code word. In some applications this cost can be too high As it will be shown in the following section, the "two-memories architecture" embeds an addresses protection mechanism compliant with IEC 61508 and allowing lower costs at the same time.
-

2. The Two-Memories Architecture (TMA)

The TMA architecture allows the reduction of memory overhead due to ECC protection, by enabling the partitioning of the memory sub-system in different regions with three selectable protection levels: no protection, parity or ECC.

The TMA architecture is presented in figure 2: a mapper decodes the incoming address and decides which protection level has to be applied. The codes are packed separately from the data. Therefore, an additional memory (called "code memory") has been introduced.

The codes are packed by the mapper to reduce the code memory area.

In case the memory controller supports banked memory architecture, a banked two-memories architecture (BTMA) is possible. In such architecture one or more banks are used to share data and codes instead of having two different data and code memories. An input called "Code Aperture Control" (CAC) is used to allocate part of data banks to store protection codes. The code memory is located at the end of the available memory space and SW access to that area is prevented. The behaviour of the resulting memory sub-system is the following: if a data read or write access is done above the CAC, the memory fault supervisor will generate alarms; if a data read or write access is done below the CAC, the access is granted.

Concerning the code overhead, the use of the BTMA architecture allowed a very efficient memory partition, as shown in the table 1. To be noted that one of the interesting benefit of BTMA is that the same memory bank can be used both for data and code, allowing a very flexible architecture.

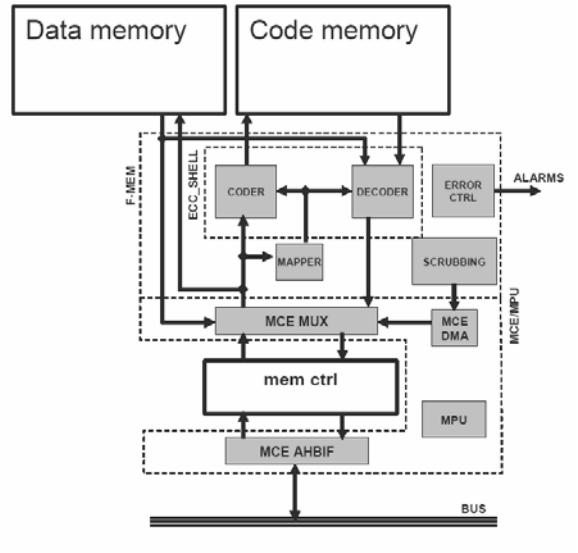


Figure 2: TMA architecture

Space for codes (CAC)	Kb protected with ECC	Kb protected with parity
2	1,5	6
4	3	12
8	6	24
16	12	32

Table 1: code overhead with BTMA architecture

Both TMA and BTMA options are able to detect – even without the address coding as previously specified – the memory array address faults as required by IEC61508 for SIL3 as also other addresses-related faults such the ones in the TMA/BTMA mapper. This is due to the following reasons:

- the code memory (or the banks hosting the codes, as in BTMA) is physically different than the data memory: a physical fault affecting the address lines or the muxing/demuxing address nets of the data memory will very unlikely determine exactly the same error at the same time also in the code memory. The consequence of that is the data will be decoded with a code corresponding to a different data and so it will be detected. The same if the address fault occurs in the code memory: the data will be decoded with a code corresponding to a different data or with an inconsistent code.
- the code memory (or the banks hosting the codes, as in BTMA) is driven by a different address respect the one used for the data memory. This different address is generated by the TMA/BTMA mapper: faults occurring in the mapper will determine the code to be read (or stored and then read) from an inconsistent address and so the error will be detected.

The TMA architecture can be expanded to a "shared memory" architecture, where the code memory is shared between different memories. With such architecture, if two or more data memories are write/read accessed at the same time, the access to the code memory is managed with a fixed (but SW configurable) priority, i.e. the data memory with the highest priority will complete the cycle without any delay while the others will wait its turn. The combined use of shared memory and fast-track assures the best cost-speed tradeoff.

3. The distributed MPU

The "distributed" MPU functionality protects the memory sub-system against SW faults in particular for multi-master systems (figure 3). The memory is divided in a number of pages associated with attributes and permissions. The MCE block of the fault supervisor uses signals from the bus to discriminate these attributes and permissions. The behavior of the resulting memory sub-system is the following. During a write or read cycle, the MPU functions in the MCE checks if the access is permitted or not. If it is permitted, the data is written in the data memory or read from it. If acess is prohibited, alarms are generated, the memory content is not changed, and the read data is fixed to an HW configurable value determined by the accessing master.

With this function, an additional level of protection is provided. Potential faults covered are wrong master accessing the memory, wrong addressing while accessing a page, wrong permissions while accessing a page, wrong attributes while accessing a page.

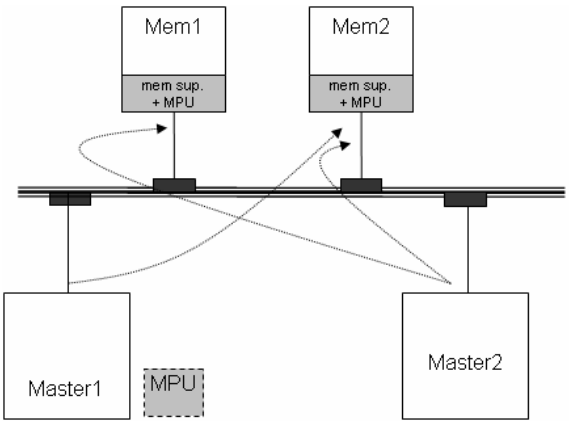


Figure 3: Distributed MPU

4. The validation process

A strict and measurable validation flow is important in order to cross check the protection features of a fault supervisor. The proposed methodology uses a validation flow based on a mix of tools, of which the main one is a simulation-based fault injector together with a fault simulator.

The fault injector tool (see figure 4) is built on top of a state-of-art functional verification tool [10]. By integrating fault injection with functional verification, it is possible to set up a fault injection flow that solves many of the issues that affect most of the environments presented in literature. Thanks to the interaction with the functional verification tool, verification components available on the market can be easily reused as a workload to inject faults, obtaining at same time design validation and reliability evaluation. The use of a standard language enables an easy and configurable way to model the faults. The engine of the coverage-driven functional verification tool allows to uniquely correlate Workload, Operational Profiles, Fault List, and final measures. The Operational Profile (OP) is a collection of information about all relevant fault-free system activities. The purpose of the OP is to better understand the situation in which the system or the application will be used, and then analyze this information to ensure that only faults which can produce an error are selected during the fault list generation process. In this way the generated fault list is compacted and non trivial.

Different fault models can be injected, from the lowest to the highest level: transient faults, bridging faults, stuck-at faults, physical-level faults such noise, bus errors, register mismatches. Moreover, the user can easily model its proper faults.

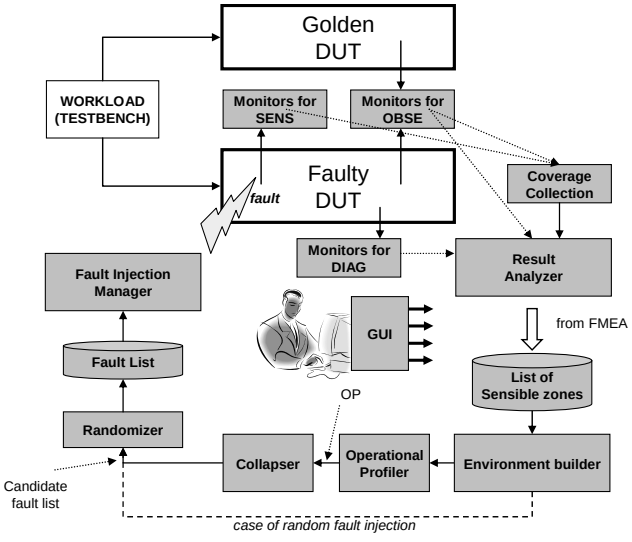


Figure 4: The Fault Injector

5. Conclusions

A first implementation of the presented memory supervisor has been accomplished. The work has been done under the supervision of TÜV-SÜD: the final result of the assessment is that the proposed fault supervisor can be used in SIL3 systems. It has finally got the official TÜV-SÜD certificate according to IEC 61508 [10] for SIL3 safety integrity level.

The proposed approach is valid also for other types of embedded memories, such as DRAM, EEPROM and FLASH. Moreover, it can easily integrated and combined with Built-in-Self-Test and Built-in-Self-Repair [11] or with transistor-level solutions addressing specific faults such for instance soft-errors [12].

In conclusion, it has been shown that the extensive efforts spent in arriving at a certifiable memory fault supervisor are not in contrast to the efficient integration to a microcontroller platform. The support of flexible data protection schemes and scalable memory subsystem architectures addresses at the same time the cost-efficiency requirements of a system provider and a silicon manufacturer respectively.

References

[1] R. Baumann, “Silicon Amnesia: Radiation Induced Soft Errors”, RADECS 2001 Short Course, September 2001

[2] International Technology Roadmap for Semiconductors, 2006 edition

[3] “Cosmic rays damage automotive electronics”, Actel, Automotive DesignLine, May 2006

[4] “Cancer Radiotherapy Equipment As A Cause Of Soft-errors In Electronic Equipment”, Medtronic, IEEE paper, Sept. 2005

[5] “Soft Memory Errors and Their Effect on Sun Fire™ Systems” by SUN Microsystems, April 2002

[6] M-Systems , Electronics Products, Feb 2006

[7] R. Mariani, G. Boschi, “A System Level Approach for Embedded Memory Robustness” JSSE Special Issue: Papers selected from the 1st International Conference on Memory Technology and Design - ICMTD’05

[8] R. Mariani, P. Fuhrmann, B. Vittorelli, “Cost-effective Approach to Error Detection for an Embedded Automotive Platform”, 2006-01-0837, SAE 2006 World Congress & Exhibition, April 2006, Detroit, MI, USA

[9] CEI International Standard IEC 61508, 1998-2000

[10] http://www.cadence.com/products/functional_ver

[11] Yervant Zorian, Embedded Memory Test and Repair: Infrastructure IP for SOC Yield, Proceedings of the 2002 IEEE International Test Conference, p.340, October 07-10, 2002

[12] P. Roche et al, “High-Density SRAM robust to radiation-induced soft-errors in 90nm CMOS technologies”, ICMDT 2005 conference

Bit Cell Leakage-Aware SRAM Sense Amplifier Activation Schemes

T. Song [a], K. Lim [a], G. Kim [b], I. Son [b], and J. Laskar [a]

[a] Georgia Institute of Technology, 85 5th St. NW, Atlanta, GA 30308, song@gatech.edu

[b] Samsung Electronics Co., San #24 Nongseo-Ri, Giheung-Eup, Yongin-City, Gyeonggi-Do, Korea 449-711

Abstract

This paper presents new static random access memory (SRAM) sense amplifier (SA) activation schemes that can be adopted under bit cell leakage-worst conditions. The proposed bit cell leakage-aware (BLA) SA activation schemes do not degrade the speed of SRAM under normal conditions, but slow down internal speed for bit lines to be evaluated only under hot temperature and fast process conditions. A 1.7-ns access 1-Mb SRAM macro was realized adopting these new schemes using a SEC 0.13-um, 1.2V CMOS process.

1. Introduction

With the scaling of technology, static random access memory (SRAM) macros occupy a large portion in system-on-chip (SoC) design. As SRAM size increases, the power and speed of SRAM macros have become more important in chip design. However, the inevitable large resistance and capacitance of cell arrays in large-capacity memory degrade SRAM performance. Moreover, the leakage current of stacked cell arrays can worsen performance under specific conditions.

Previously, leakage reduction schemes for SRAM macros have been reported [1-4]. These schemes modify the cell array architecture having additional peripheral circuits that control internal voltages, which reduce cell array leakage current in standby or dynamic states. However, the trade-off for such approaches is increased circuit complexity and area overhead for leakage current saving. In addition, these schemes have an effect on SRAM under normal as well as worst conditions to degrade the performance.

Meanwhile, another technique was reported as reducing leakage current while having better storage capability in the cell array [5]. This scheme uses replica-cell characteristics to control internal voltages for cell latches. However, mismatches between cell-modeling peripheral blocks and cell array blocks may not reduce leakage current. Moreover, they can worsen the stability of the cells.

In this paper, novel approaches to control the leakage current in bit cell arrays are proposed. These works take advantage of the characteristics that bit cell leakage current does not always affect SRAM speed. In this case, leakage current can be considered only under the specific conditions of a hot temperature and fast process (HF). Therefore, sense amplifier (SA) activation timing delay under HF conditions does not affect SRAM speed under other conditions. These approaches result in eliminating necessary trade-offs between SRAM speed and leakage current control schemes. In Section 2, the need for a bit cell leakage-aware (BLA) SA activation circuit is examined. Section 3 presents three novel types of SA activation circuits that trace bit cell array leakage current under HF conditions. Finally, comparative SPICE-level simulation results and conclusions are examined.

2. Leakage current effect on SRAM

Fig. 1 shows the conventional 6T SRAM cell array schematic having the data that cause maximum leakage current to affect the sensing margin (SM). As SRAM has to store and read the data at every case, the worst data condition should be examined carefully. If one bit cell is activated and the other bit cells store the opposite data to the activated cell, the activated cell suffers from SM reduction as shown in Table 1. We assume that the precharge voltage is VDD, no. of cells in array is N, static voltage drop made by bit cell subthreshold current I_{off} is ΔV_{off}, dynamic voltage drop made by activated bit cell current I_{on} is ΔV_{on}, and SM ΔV is $V_{bitb} - V_{bit}$.

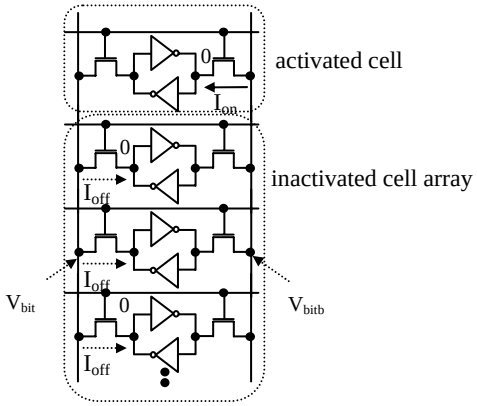


Fig. 1. 6T SRAM cell array having data to make maximum leakage current

	Minimum Leakage	Maximum Leakage
No. of I_{off} in bit lines	(N - 1) * 1/2	N - 1
No. of I_{off} in bitb lines	(N - 1) * 1/2	0
V_{bit}	VDD – 1/2 * (N - 1)*ΔV_{off}	VDD – (N - 1) * ΔV_{off}
V_{bitb}	VDD – 1/2 * (N - 1)*ΔV_{off} - ΔV_{on}	VDD - ΔV_{on}
SM ΔV	ΔV_{on}	ΔV_{on}– (N - 1) * ΔV_{off}

Table 1. Analysis of voltage of bit lines and SM under maximum and minimum leakage data structure.

ΔV_{on} is decided by I_{on}, bit cell array loading as capacitance and resistance, and I_{off}. However, ΔV_{off} made by I_{off} in the maximum leakage data structure affects the SM more than in the minimum leakage structure by the ΔV_{bit} drop of $(N-1)*\Delta V_{off}$. Especially, udner hot temperature and fast process conditions, the SM reduction amount is not negligible.

Fig. 2 shows the corner simulation results for both I_{on} and the sum of I_{off}. The delta value means I_{on} minus the sum of I_{off}, which affects SM. In cold temperature or SS (both PMOS and NMOS are slow) conditions, I_{off}_sum is not large enough to affect I_{on}. However, in both hot temperature and FF (both PMOS and NMOS are fast) conditions, I_{on} decreases because of I_{off}.

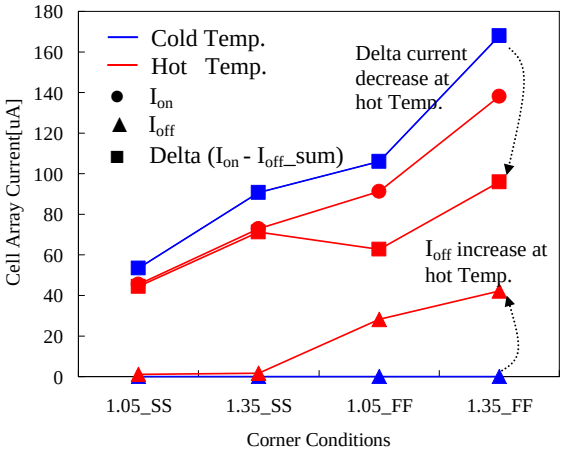


Fig. 2. 1-Mb SRAM cell array current

Fig. 3 shows the SA margin under the same conditions as above. As we predict from Fig. 2, the SA margin is reduced at hot temperature and FF conditions. If we assume that the minimum SA margin required to meet the SA specification is 50mV, the SRAM which was made in the FF process cannot be read accurately in a hot temperature environment. In these cases where the SA margin is reduced, the SRAM designer has to delay SA activation time for bit lines to be evaluated and read through the SA, which also causes speed degradation under other conditions also.

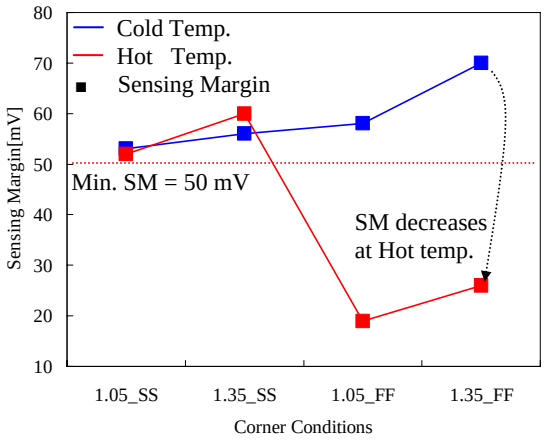


Fig. 3. 1-Mb SRAM SM

Therefore, this paper suggests novel SA activation schemes which trace I_{off} and are adjusted under certain conditions without hurting SRAM performances under normal conditions.

Fig. 4 shows a conceptual diagram of the BLA SA activation scheme (pulse generator). It is different from the previous one in that a leakage monitoring part is inserted in the pulse generator. I_{off} controls internal resistance and capacitance of the pulse generator so that SA activation can be delayed for bit lines to be evaluated. In this paper, an active resistance control scheme, mos capacitance control scheme, and coupling capacitance control scheme are suggested.

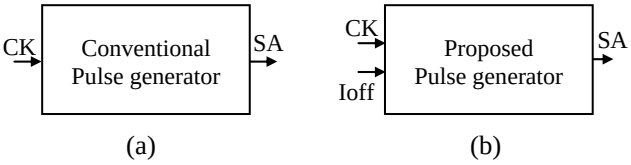


Fig. 4. (a) Conventional pulse generator (b) proposed pulse generator adopting I_{off}

3. Bit cell leakage-aware (BLA) pulse generator

Fig. 5 shows a pulse generator to be adjusted by BLA active resistance. The unit delay for the pulse generator is composed of active resistance and capacitance. The transistor of active resistance is gated by I_{off}_in, which is made from the I_{off} generator. In normal cases, I_{off} is not comparable, so the pulse generator makes delay adequate for sensing. However, in the case of hot temperature and FF conditions, I_{off}_in decreases lower than the precharge voltage, VDD, because of excessive I_{off}. If this voltage is inserted into basic delay cells, the delay cell is delayed according to the basic equation of I_{ds} vs. V_g.

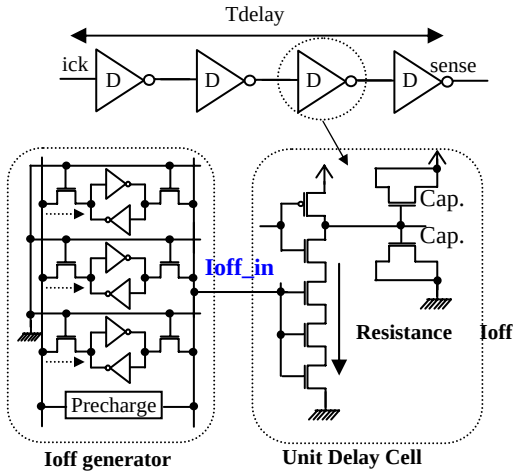
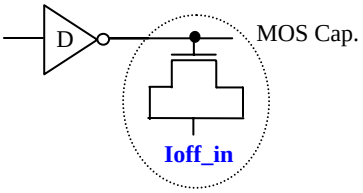
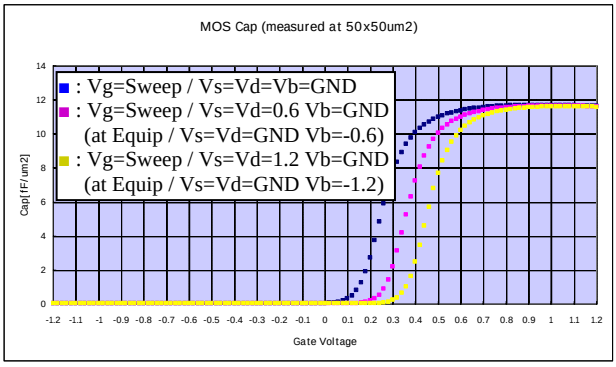


Fig. 5. SRAM pulse generator adopting I_{off} to control active resistance

Fig. 6 (a) (b) presents an adaptable pulse generator using mos gate capacitance. Because mos gate capacitance is variable according to bias of source, if we use I_{off}_in as a bias of source, we can adjust mos cap. as much as I_{off}. Fig. 6(b) shows the extracted mos capacitance under various conditions. If I_{off}_in becomes lower than 1.2V, we can use a mos cap. larger than that under normal conditions. In that case, the unit delay cell of the pulse generator slows down, and a larger SM can be obtained under HF conditions.



(a)



Vd=Vs	Vg	MOSCAP	Ratio
1.2V	0.4V	2fF	100%
0.6V		6.5fF	325%
1.2V	0.5V	8fF	100%
0.6V		10fF	125%

(b)

Fig. 6. (a) Pulse generator using BLA mos cap. (b) Extracted values of mos capacitance

Finally, Fig. 7 shows the BLA pulse generator using the coupling capacitance of the adjacent bit lines. This scheme uses leakage current to control the delay of the pulse generator dynamically. This is composed of an I_{off} generator, coupling capacitance generator, and pulse generator. An I_{off} generated from a stacked dummy cell array is induced into the coupling capacitance generator. Under HF conditions, if I_{off} is generated excessively, the voltage of I_{off}_in becomes low enough to trigger the coupling generator PMOS (CGP). If CGP operates, the dl_couple signal is discharged to the lower level, and it causes the falling of the adjacent bit line (dlread). According to the voltage of dlread coupled by the dl_couple signal, the pulse generator can be delayed. Therefore, timing delay under HF conditions does not hurt the speed performance under normal conditions.

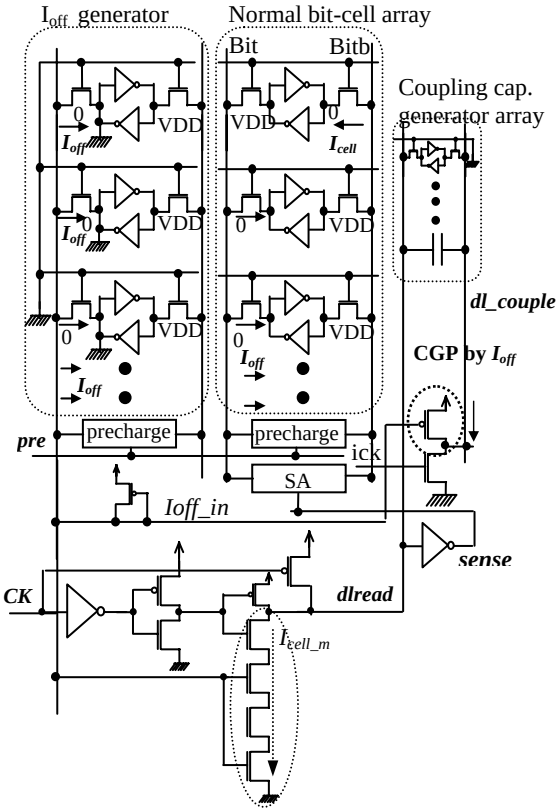


Fig. 7. BLA pulse generator using coupling capacitance of adjacent bit lines.

Fig. 8 shows the simulation results of the BLA pulse generator using bit line coupling capacitance. These results show that it makes the delay of the pulse generator at HF (Hot temp. and Fast process), but does not at HS (Hot temp. and Slow process). Therefore, the worst time delay is still decided at HS.

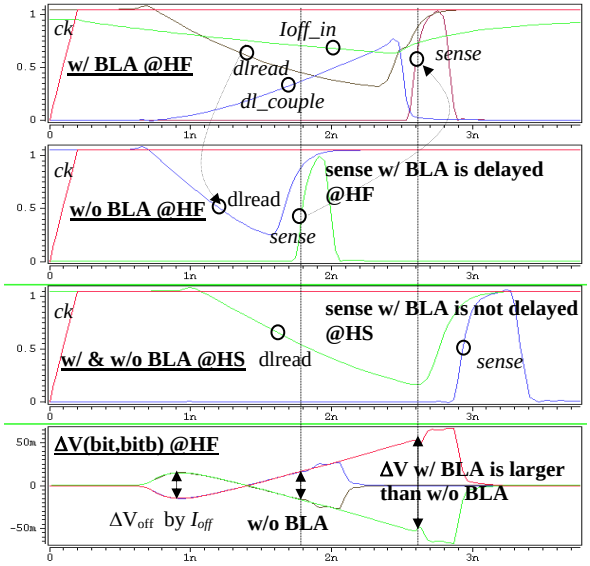


Fig. 8. Simulation results of BLA pulse generator using coupling capacitance of adjacent bit lines (HF=Hot Temp.&FF, HS=Hot Temp.&SS)

4. Conclusions

In this paper, novel pulse generators were examined that can be applicable to SRAM cache memory under HF conditions. Each method uses an I_{off} generator adopting bit cell subthreshold leakage current under hot temperature, fast process conditions. Pulse generator using I_{off} can control the active resistance of delay cell and mos gate capacitance. Finally, coupling capacitance of adjacent bit lines were used to make pulse generator delay dynamically in proportion to I_{off} under HF conditions. Fig. 9 shows the comparison of SMs w/ and w/o BLA schemes. It shows that the SM w/ BLA is larger than w/o BLA by 368%, and does not affect SRAM speed under normal conditions even with this benefit. Fig. 10 shows the layout of a 1-Mb SRAM w/ BLA. The area is 3.24 mm2, and the BLA has an area overhead under 0.5%.

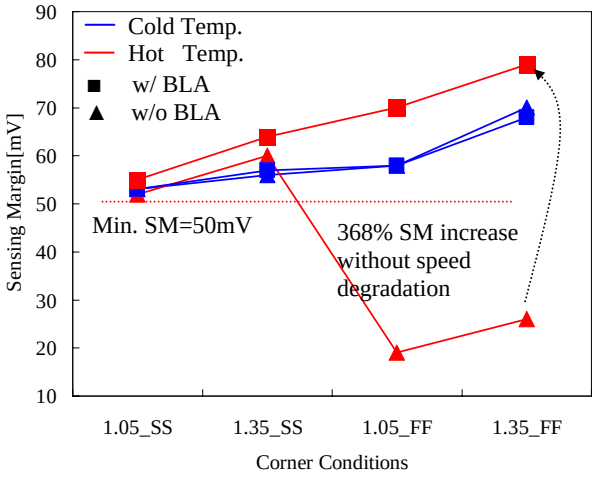


Fig. 9. Comparison of SM under the worst conditions

Acknowledgment

The authors wish to acknowledge Dr. F. Bien for his technical discussions.

References

[1] F. Frustaci et al., "Techniques for Leakage Energy Reduction in Deep Submicrometer Cache Memories," IEEE Trans. On VLSI Systems, vol. 14, no. 11, pp. 1238-1249, Nov. 2006.

[2] D.Ho et al., "Ultra-Low Power 90nm 6T SRAM Cell for Wireless Sensor Network Applications," IEEE ISCAS 2006, May 2006.

[3] C. Hyung-il Kim et al., "A Forward Body-Biased Low-Leakage SRAM Cache: Device, Circuit and Architecture Considerations," IEEE Trans. On VLSI Systems, vol. 13, no. 3, pp. 349-357, March 2005.

[4] M.Yamaoka et al., "A 300-MHz 25-uA/Mb-Leakage On-Chip SRAM Module Featuring Process-Variation Immunity and Low-Leakage-Active Mode for Mobiel-Phone Application Processor," IEEE J. Solid-State Circuits, vol. 40, no. 1, pp. 187-193, Jan. 2005.

[5] Y. Takeyama et al., "A Low Leakage SRAM Macro With Replica Cell Biasing Scheme," IEEE J. Solid-State Circuits, vol. 41, no. 4, pp. 815-822, April 2005.

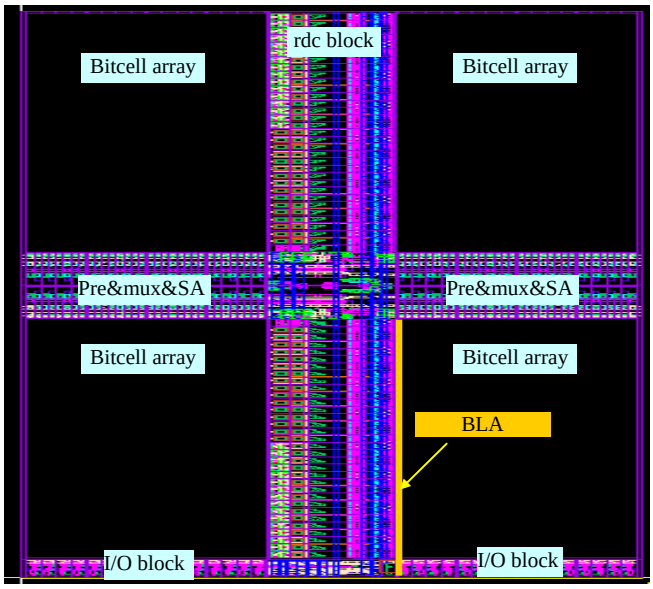


Fig. 10. Layout of 1-Mb SRAM w/ BLA

A 128Kb 5T SRAM in 0.18µm CMOS

[1]Stefan Andersson, [1]Ingvar Carlson, [1,2]Sreedhar Natarajan, [1]Atila Alvandpour

[1]Division of Electronic Devices, Linköping University, Sweden
[2]Emerging Memory Technologies Inc., Canada

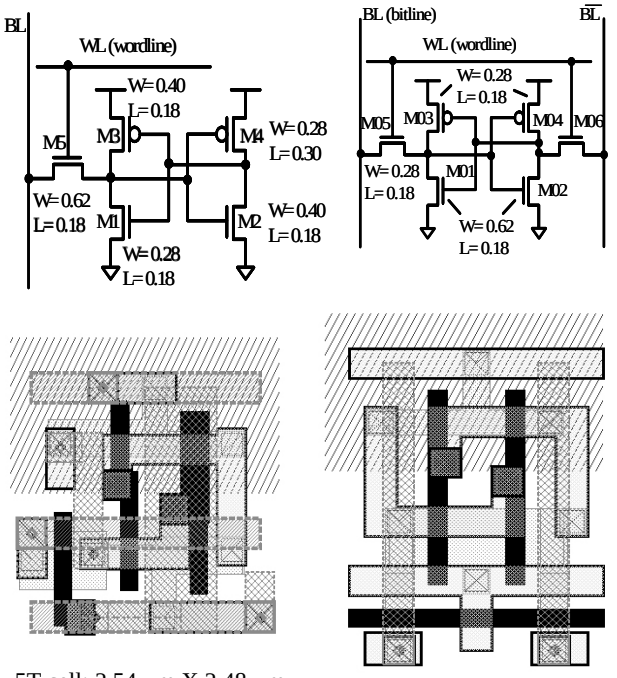
Abstract - This paper presents chip measurement results of a high density, fully static 128Kb on-chip cache utilizing 5-transistor single-bitline memory cells in a standard 0.18µm CMOS technology. Compared to a 128Kb 6T SRAM in the same process, the 5T SRAM has 23% smaller area and 4X lower bitline leakage. Despite the single bitline, a differential sensing scheme results in a comparable read-time of 360ps, at 1.8V, and a 6T-compatible read/write scheme. Furthermore, operation of every single memory cell has been successfully tested over a supply voltage range of 0-9V-to-1.8V.

I. INTRODUCTION

Mobile applications and advanced microprocessors demand for increasingly large on-chip memories with low power consumption and low standby leakage in standard CMOS technologies. This has not been satisfied with (i) embedded DRAM's or 4T SRAM's [1] due to increased manufacturing cost, or with (ii) planar DRAM's, which have not been proven to be viable for high-yield, high-volume microprocessors. Therefore, conventional 6T-cell SRAM's are the main choice for today's on-chip cache applications. This paper presents a high density, fully static 128Kb on-chip cache utilizing 5-transistor (5T) single-bitline memory cells in a standard 0.18µm CMOS technology. Compared to a 128Kb 6T SRAM in the same process, the 5T SRAM has 23% smaller area and 4X lower bitline leakage. Correct and robust write operation across the process corners has been ensured by an intermediate bitline precharge voltage and the corresponding memory cell sizing. Despite the single bitline, a differential sensing scheme results in a comparable read-time (360ps, at 1.8V) and a 6T-compatible read/write scheme without requiring any extra signal or access to any additional cell nodes such as those in [2]. The basic operation of the proposed 5T SRAM was discussed in [3] using circuit simulations only. This paper describes the 5T 128Kb cache based on chip measurement results, where the operation of every single memory cell has been successfully tested over a supply voltage range of 0.9V-to-1.8V.

II. 5T SRAM CELL AND MEMORY ORGANIZATION

Figure 1 shows the schematic and layout of the proposed 5T-cell and the conventional 6T-cell in a standard 0.18µm, 1.8V CMOS technology. The 5T-cell has only one access transistor 'M5' and a single bitline 'BL'. Writing of '1' or '0' into the 5T-cell is performed single-ended by driving the bitline to Vcc or Vss respectively, while the wordline is asserted at Vcc. The writability of the cell is ensured by a different cell sizing strategy (Figure 1). The trip-point of the inverter M2-M4 has been decreased, while the trip-point of the inverter M1-M3 has been increased. Furthermore, the pass-transistor M5 is sized to support both write and read operation.



5T cell: 2.54 µm X 2.48 µm

6T cell: 2.56µm X 3.12 µm

Figure 1: Circuit topology and layout of the proposed 5T SRAM cell and the conventional 6T cell in a standard 0.18µm, 1.8V CMOS technology using "logic design rules".

Since the 5T SRAM cell is writable at $V_{BL}=V_{WL}=V_{CC}$, a non-destructive read operation requires a bitline precharge voltage, Vpc, where Vss < Vpc < Vcc. Figure 2 shows a post layout simulation example of a write operation, and Figure 3 shows simulation results for possible bitline precharge levels (340mV-860mV) for which the 5T-cell (Figure1) supports correct read/write operation across process corners at 110°C.

The sizing of the 5T-cell results in a symmetrical and balanced memory cell at onset of a read operation, when the bitline is precharged to Vpc=600mV.

We utilized an external supply voltage for Vpc to explore the impact of different Vpc levels. However, a low power DC-DC conversion based on selectively precharging bitlines to Vcc or Vss followed by equalization was proposed in [3].

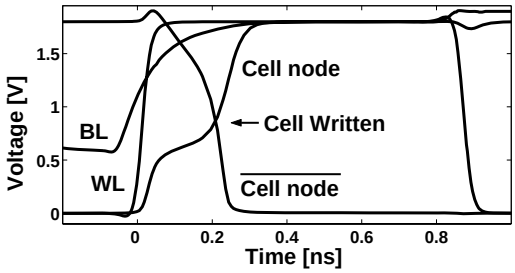


Figure 2: Post layout simulation of a local bitline write ('1') operation of the´128Kb 5T-cell memory

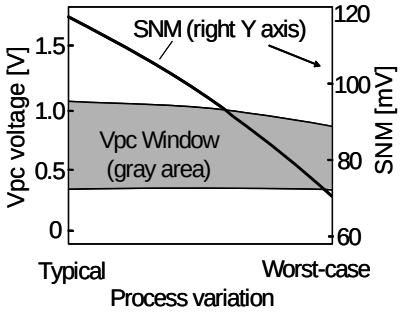


Figure 3: 5T-cell static noise margin, SNM (on right Y axis), and possible bitline precharge voltage levels across worst case process corners (left Y axis).

The stability of the 5T-cell has been verified by evaluating the Static Noise Margin [4]. The SNM analysis was performed by ramping up noise voltage sources while the wordline is asserted and the bitlines are maintained at the precharged voltage, which is Vcc=1.8V for 6T-cell, and Vpc=600mV for 5T-cell [3] (Figure 4).

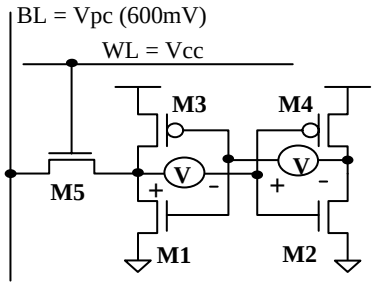


Figure 4: Static noise margin (SNM) evaluation setup for 5T SRAM cell.

Figure 3 shows also the 5T-cell SNM from 117mV to 70mV across process corners. Compared to the 6T-cell, the SNM of the 5T-cell is about 50% lower. The reduced SNM is an unfavorable trade-off. However, the SNM is sufficient to support correct functionality of the 5T-cell across the process corners.

Figure 5 shows the organization of the 5T-cell 128Kb SRAM. Despite the single bitline, a conventional differential sense amplifier has been utilized for the read operation. Figure 6 shows the sense amplifier. For clarity, only one pair (of eight) of column selectors is shown. The lower bitline precharge voltage (*Vpc*) allows us to use NMOS devices for column selectors. The second differential input of the sense amplifier is connected to a separate and equalized bitline. The cells on the second bitline are not accessed simultaneously.

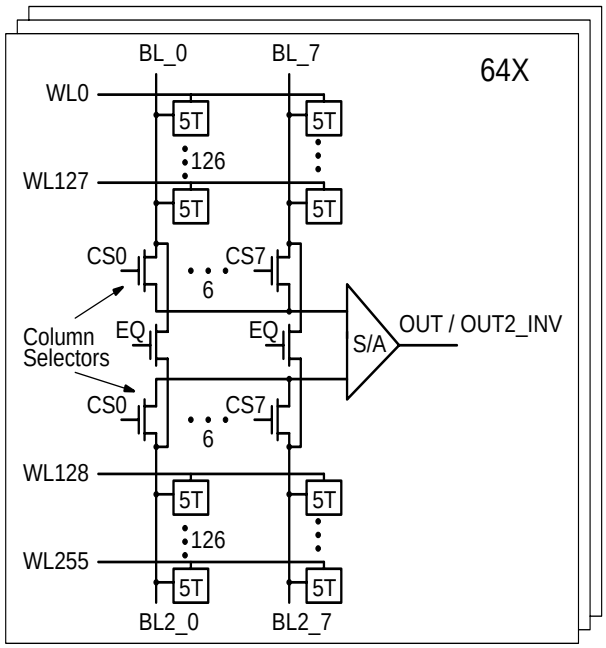


Figure 5: Organization of the 5T-cell 128Kb SRAM.

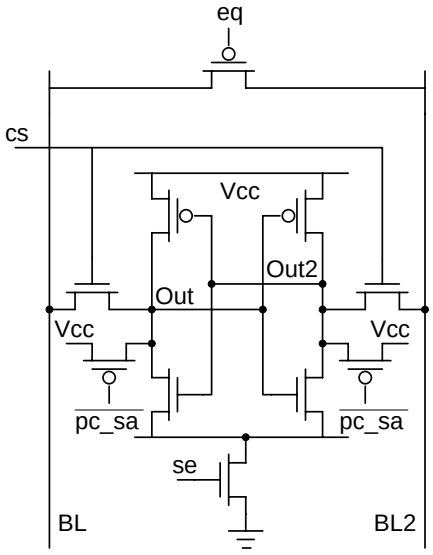


Figure 6: Differential sense amplifier

Figure 7 shows the read operation scheme for the 128Kb memory, which is similar to that of a conventional 6T SRAM. The main difference is the bitline precharge level, which is *Vpc*=600mV. Figure 8 shows the post layout simulation example of a local read operation, where the bitline is shared between 128 memory cells.

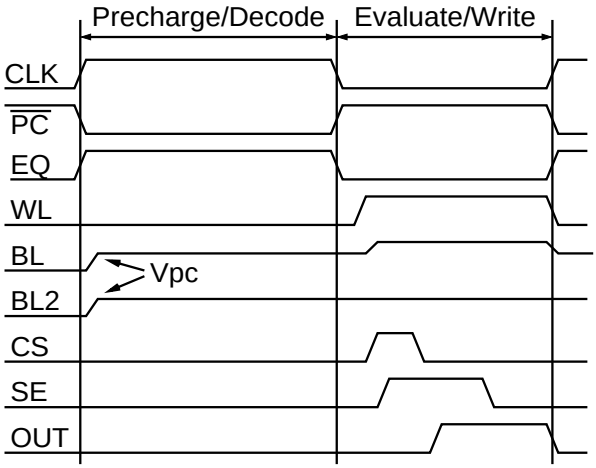


Figure 7: Read/write operation scheme for the 5T 128Kb memory, and the transistor schematic of the Sense Amplifier.

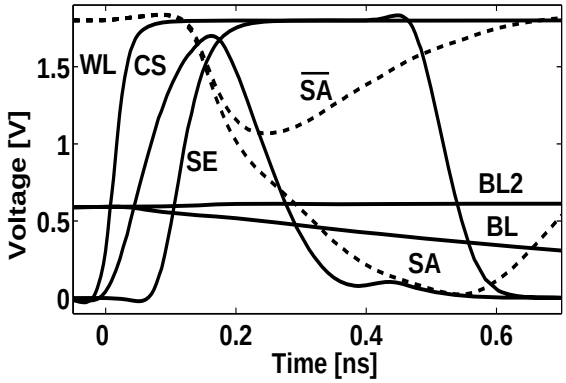


Figure 8: Post layout simulation of a local bitline read operation of the 128Kb 5T-cell memory

III. Comparisons and Chip Measurement Results

The standard 0.18µm CMOS technology supported "logic" design rules. For accurate comparisons, a conventional 6T-cell 128Kb SRAM was also designed with complete chip layout. After the post-layout comparisons, only the 5T SRAM was taped-out for fabrication. Both 5T and 6T 128Kb SRAM's have been partitioned in 16 banks of 128 rows X 64 columns. The 5T-cell utilizes a metal3 wordline and the 6T-cell has a poly wordline stitched with metal 3.

The 5T-cell has 21% smaller area (2.54µm X 2.48µm vs. 2.56µm X 3.12µm for the 6T-cell), while the 5T 128Kb memory array has a 23% smaller area (0.88mm^2 vs. 1.15mm^2 for the 6T 128Kb array). The post-layout simulations at 1.8V, 110^{o}C showed comparable read/write performance (421ps/191ps for the 5T 128Kb, and 499ps/135ps for 6T 128Kb), while measured read-delay for the 5T 128Kb SRAM is 360ps at 1.8V, 40^{o}C, and a Vpc of 600mV.

Furthermore, the single bitline 5T-cell leakage was 32% lower (4.78nA vs. 7.08nA for 6T-cell). In addition, precharging the single bitline to Vpc=600mV resulted in 4X lower bitline-leakage/cell (0.8nA vs. 3.16nA for 6T-cell) at 110^{o}C. This is due to the lower V_{DS} over the wordline pass transistors.

Based on the encouraging post-layout comparisons, the 5T-cell 128Kb cache was fabricated in the standard 0.18µm, 1.8V, CMOS technology. The chip photograph is shown in Figure 9.

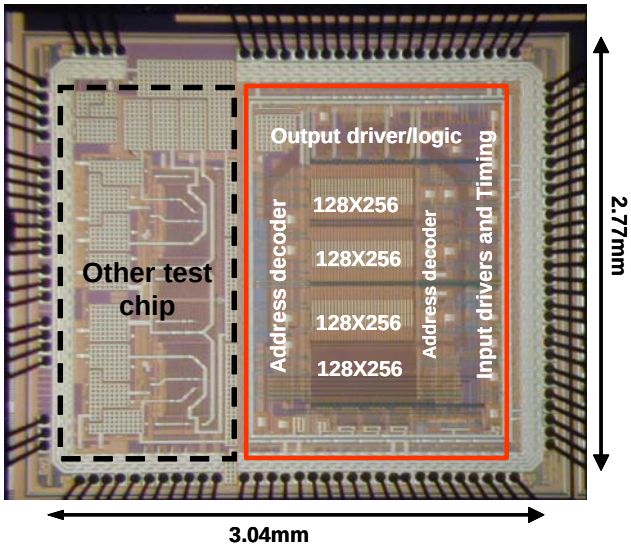


Figure 9: Chip photograph; Process: 0.18µm CMOS, cell area (with logic design rule): 6.30µm^2, memory area: 0.88mm^2.

Major efforts were made to achieve a fully testable and controllable read/write timing circuitry, where every single memory cell has been addressed and accessed for read and write. A single internal clock pulse traverses through controllable delay elements and logic generating all control signals for read/write operations. For accurate performance measurements, the internal clock pulse is triggered and controlled by an external signal generator, and the chip was directly bonded on a PCB with 50Ω terminated I/0's. The measurements included read/write test of every single memory cell at 40^{o}C, where all the memory cells have been fully functional. Figure 10 shows a measurement example of a successive write and read operation of '0' to '1', and '1' to '0'.

Figure 11 shows measured read delay of 360ps at 1.8V, 40^{o}C, and a Vpc of 600mV. Figure 12 shows voltage scalability of the 5T 128Kb memory over a Vcc range of 0.9V-to-1.8V corresponding to a Vpc range of 330mV-to-600mV.

Figure 13 shows the relative insensitivity of the 5T SRAM to potential variations in Vpc over a range of 1.2V.

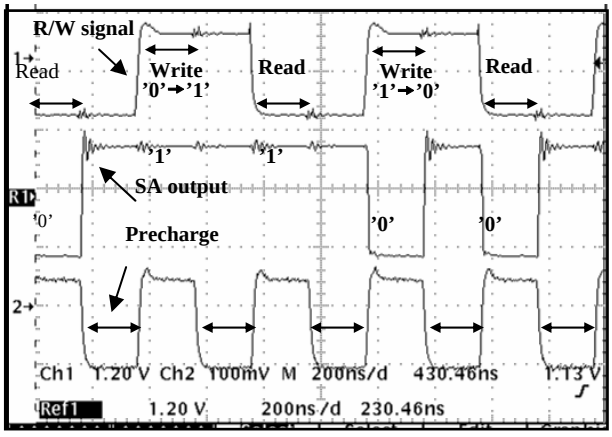


Figure 10: Measurement; Successive write/read operation of '0'→'1'→'0'

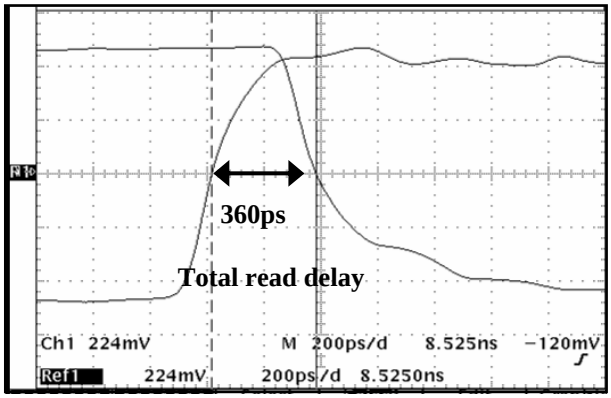


Figure 11: Measurment; 360ps total read delay at 1.8V, 40^{o}C.

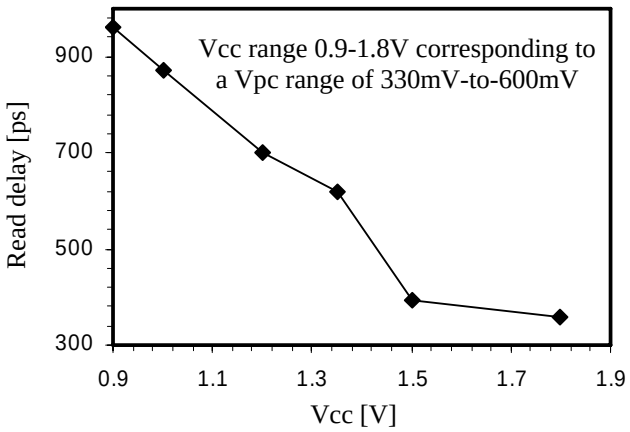


Figure 12: Measurement; Voltage scalability of the 5T 128Kb SRAM.

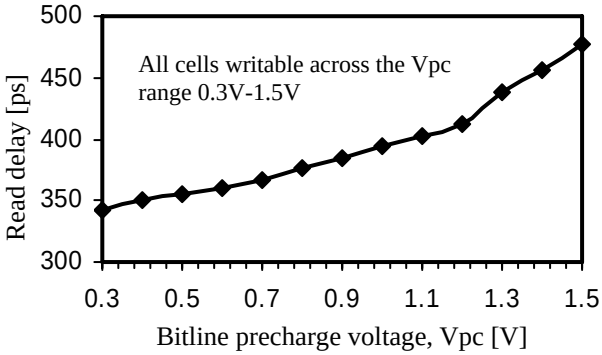


Figure 13: Measurement; Relative insensitivity of the 5T memory's read and writability to potential bitline precharge voltage variations.

IV. Conclusions

We presented chip measurement results of a high density, fully static 128Kb on-chip cache utilizing 5-transistor single-bitline memory cells in a standard 0.18µm CMOS technology. Operation of every single memory cell has been successfully tested over a supply voltage range of 1V-to-1.8V. For a non-destructive read operation, the bitline is precharged to an intermediate voltage Vpc= *600mV*. Measurement results show that the memory is relatively insensitive to variations in the bitline precharge voltage. Despite the single bitline, a differential sensing scheme results in a comparable read-time of 360ps, at 1.8V, and a 6T-compatible read/write scheme. Compared to a 128Kb 6T SRAM in the same process, the 5T SRAM has 23% smaller area and 4X lower bitline leakage.

Acknowledgments

The authors would like to thank: Arta Alvandpour for PCB design. Dr. Dinesh Somasekhar, Dr. Ram Krishnamurthy, Dr. Vivek De, and Shekhar Borkar from Intel Corporation and Professor Christer Svensson from Linköping University for useful discussions and support.

References

[1]. A. Kotabe, et. al, "A low power four-transistor SRAM cell with a stacked vertical poly-silicon PMOS and a dual-word-voltage scheme", IEEE J. Solid-State Circuits, vol. 40, pp. 870-875, April, 2005.

[2]. H. Tran, et. al, "Demonstration of 5T SRAM and 6T Dual-Port RAM Cell Arrays," Symposium on VLSI Circuits, pp. 68-69, Jun, 1996.

[3]. I. Carlson, S. Andersson, S. Natarajan, A. Alvandpour, " A high density, low power 5T SRAM for embedded Caches", European Solid-State Circuits Conference 2004, pp. 215-218.

[4]. E. Seevinck, F. J. List and J. Lohstroh, "Static-Noise Margin Analysis of MOS SRAM Cells," IEEE JSSC, VOL. SC-22, NO.5, pp.748-754, Oct. 1987.

SESSION G

Charge Trapping

Sub-lithographical Shrink of Twin Flash™ Memory Cells to the 32 nm Technology Node

M. F. Beug[a], R. Knoefler[a], C. Ludwig[a], R. Hagenbeck[a], T. Müller[a], S. Riedel[a], M. Isler[a], M. Strassburg[a], T. Höhr[a], T. Mikolajick[b] and K.-H. Küsters[a]

[a] Qimonda Technologies GmbH & Co. KG, D-01099 Dresden, Germany

[b]Chair of Electronic- and Sensor-Materials, TU Bergakademie Freiberg, 09596 Freiberg, Germany

Abstract

The extended scalability of Twin Flash memory cells down to 32.5 nm half pitch is demonstrated in a conventional planar cell layout. Starting with 65 nm line space array and doubling the number of word lines, a cell size of 0.0112 µm² can be achieved. This corresponds to bit sizes of 0.0056 µm² and 0.0028 µm² for SLC and MLC mode, respectively. It was found that the proposed aggressive shrinking of the cell spacing in word line direction results in a cross talk of 300 mV when both neighboring cells are programmed to the highest MLC level. It is reported for the first time that cross talk in charge trapping memory cells becomes an issue when the cell spacing approaches the 20 nm mark.

1. Introduction

The Twin Flash or NROM [1] cell is an alternative to other non-volatile memory cells due to its competitive cost position and bit sizes. The localised charge storage in the ONO (SiO_2/SiN/SiO_2) trapping layer allows to store up to four bits per cell, when four charge levels (MLC) are realized above source and drain, respectively [2]. But the storage of two separated charge distributions also limits the length scaling of the Twin Flash cell. To overcome this scaling limit we demonstrate here for the first time a sub-lithographical shrink in the word line (WL) dimension or width direction for Twin Flash cells. The resulting cell size in relation to the smallest directly printed lithographical structure of 65 nm is $2.65F^2$. In relation to the final effective WL half-pitch of 32.5 nm the corresponding cell size would be $10.6F^2$ which results in $2.65F^2$/bit in the MLC mode.

2. Sample preparation

The samples were fabricated based on 65 nm test structures, which were used as a reference. The word line patterning for future 32.5 nm Twin Flash cells uses a 65 nm line space array as shown schematically in Fig. 1(a). The 65 nm lines are trimmed by an isotropic etch process to the target thickness of 45 nm (Fig. 1(b)). After 20 nm spacer deposition and etch (Fig. 1(c)) the final number or word lines (Fig. 1(d)) can be defined before this structure is transferred into the hard mask which is used for WL patterning. A fabricated Twin Flash cell with 45 nm channel width can be seen in the SEM micrograph of Fig. 1(e). The cell length of these cells was 100 nm. The 20 nm wide WL stack etch was developed and successfully carried out on specific test structures. The asymmetric division of available space (45 nm cell width and only 20 nm WL space) was chosen to keep the cell read current as high as possible.

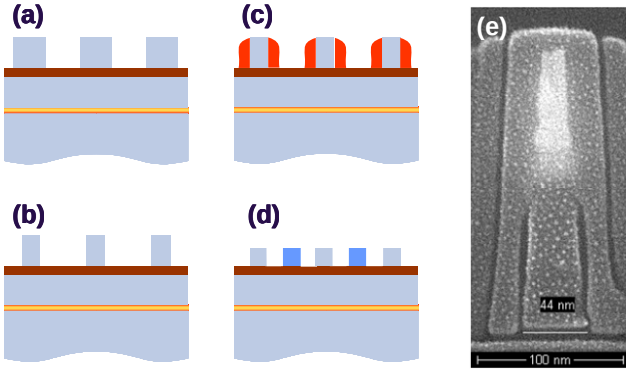


Fig. 1: Schematic process flow for word line patterning using a method to double the number of word lines (Pitch Fragmentation) starting from a 65nm line space array (a)-(d). (e) SEM micrograph of a cross section through Pitch Fragmentation Twin Flash cell with 45 nm width.

3. Electrical characteristics

The functionality of the 45 nm wide Twin Flash cells was investigated using the standard characterisation methods known from previous technologies [4],[5]. Program and erase operation is schematically illustrated in Fig. 2. The drain voltage dependent program (Fig. 3(a)) and erase (Fig. 3(b)) curves show normal behaviour.

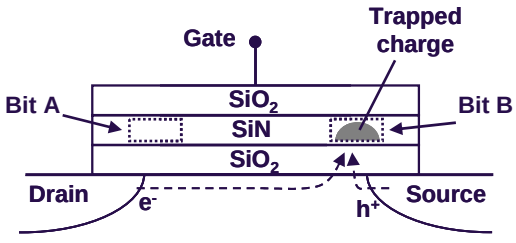


Fig. 2: Schematic picture of program and erase operation (Bit B) of a Twin Flash device by channel hot electrons and hot holes, respectively.

In this example Bit B was programmed and erased, as shown in Fig. 2. Additionally Bit A was readout to judge the second bit effect [5].

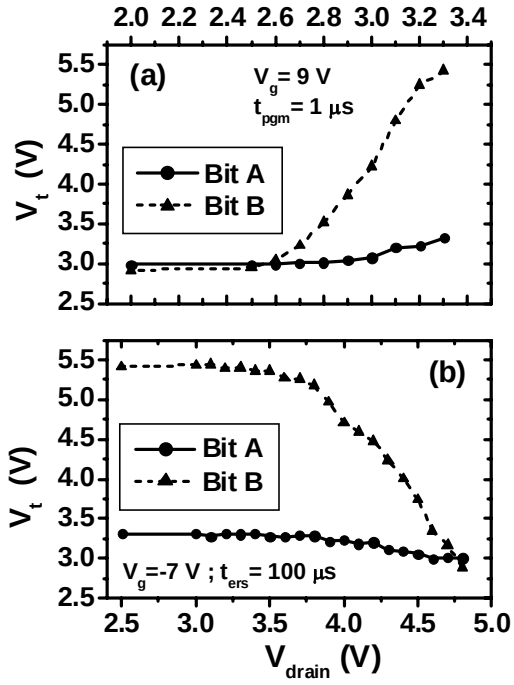


Fig. 3: Program and erase characteristic of 45 nm wide Twin Flash cell. Bit B is programmed and erased. The influence of programming Bit B on the readout of Bit A can be seen (second bit effect [5]).

As a consequence of the narrow cell width and relatively high channel doping the initial Vt level is about 3 V. The program erase cycling results of the 45 nm wide devices is depicted in Fig. 4, which shows the endurance up to 100k cycling. The maximum drain voltages of the last drain stepping pulses to reach the corresponding program or erase levels can be seen in Fig. 4(b). Fig. 4(a) and (b) indicate very good endurance behaviour.

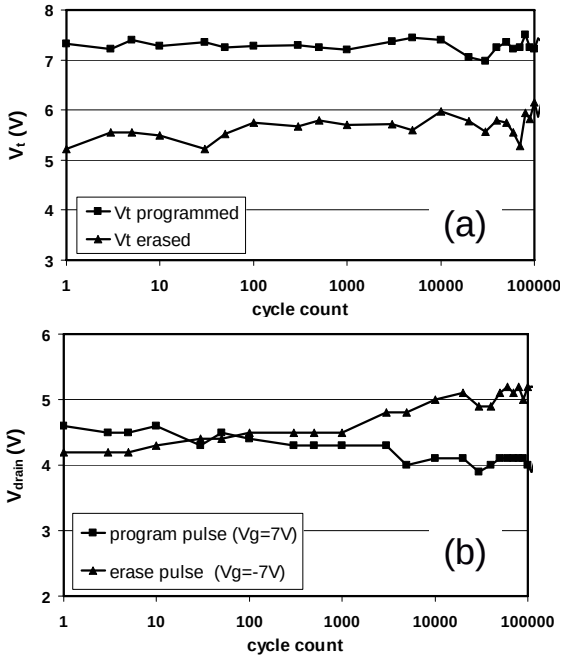


Fig. 4: Endurance results of 45 nm wide Twin Flash cell.

To evaluate the functionality of our target structure with 45 nm Twin Flash cells with 20 nm WL spacing, we investigated the influence of cross-talk between neighbouring WLs. Measurements were done at 65 nm wide cells with 65 nm WL space. As indicated in the inset of Fig. 5 the neighbour bits 1 and 5 (WL1 and WL3) were programmed to a threshold voltage shift of ΔV_{tN}=4 V and the influence on the Vt of the cell in between the two programmed cells (WL2, bit3) was investigated. The threshold voltage of bit3 was shifted by 100 mV for given conditions which is consistent with simulation results as can be seen in Fig. 5.

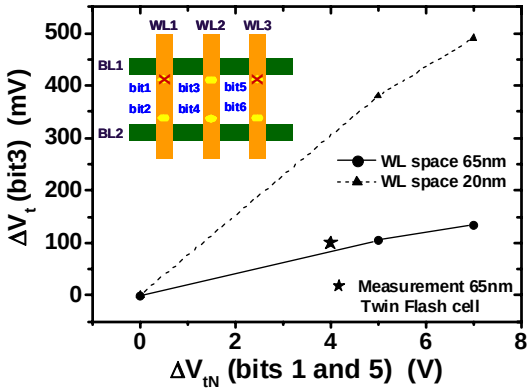


Fig. 5: Threshold voltage shift of the active cell (bit3) as function of threshold voltage shift ΔV_{tN} of the two neighboring bits (bit1 and bit5) for different word line spacing values shown for simulations and measurement.

As mentioned we investigated cross talk due to neighbour bit programming in 2D simulations. Fig. 6 shows the 2D structure with WL space of 20 nm for neighbour WL voltages of 0 V and -5 V which corresponds to neighbour bit programming threshold voltage shifts (ΔV_{tN}) of 0 V and +5 V, respectively. The influence on the inversion strength at center position of the channel (at gate bias Vg=3 V) can be clearly observed. As an equivalent threshold voltage criterion in this 2D structure we used the gate voltage which is needed to generate a fixed surface inversion charge density in the middle of the channel.

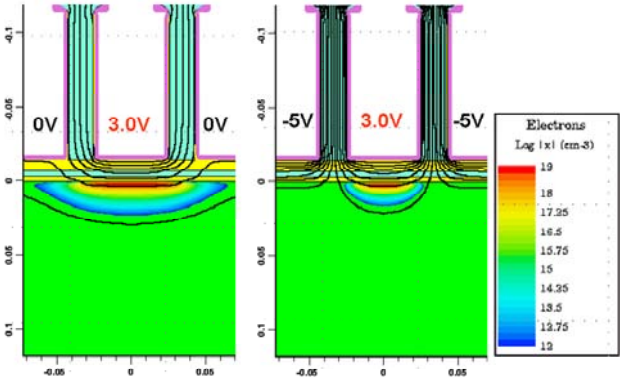


Fig. 6: Influence of the neighbor bit programming state on the inversion layer electron density of the active memory cell for 20 nm word line spacing extracted from 2D simulations.

From Fig. 5 can be seen that for 20 nm WL space and a MLC capable threshold voltage window of 4 V the resulting V_t shift is about 300 mV. This value can be

reduced by successive programming of the WLs and additionally by programming algorithm, but it has to be pointed out that the same effect takes place in all charge trapping memory cells of comparable dimension.

4. Conclusions

We investigated the shrinkability of Twin Flash cells down to the 32 nm half pitch technology node. The presented 45 nm wide Twin Flash cells are the smallest shown up to now. We conclude that a 65 nm pitch can be realized using the shown 45 nm wide cells together with an aggressive word line spacing of only 20 nm. The narrow Twin Flash cells show good endurance behavior. We further show for the first time that for such small word line spacing of 20 nm, cross talk between charge trapping memory cells becomes an issue. This effect is comparable to the FG-FG coupling in floating gate NAND non-volatile memory cells. For a programmed MLC threshold voltage window of 4 V for both neighboring bits we determined a maximum Vt shift of 300 mV. This value can be minimized by programming algorithm which assures the MLC functionality of the presented 32 nm half pitch Twin Flash approach. But it demonstrates that cross talk can not be neglected in charge trapping memory devices below 30nm half pitch.

5. Acknowledgements

This work was financially supported by the Federal Ministry of Education and Research of the Federal Republic of Germany (Project No 01M3160). The authors are responsible for the content of the paper.

References

[1] B. Eitan et al., IEEE Electron Dev. Lett., **21**(11), 543(2000).
[2] B. Eitan et al., IEDM, 547 (2005).
[3] J. Willer et al., VLSI, 76 (2004).
[4] N. Nagel et al., VLSI, 120 (2005).
[5] E. G. Stein von Kamienski et al., NVMTS, 5 (2005).

An embedded spacer-trapping memory in the CMOS technology

E.Pikhay [a], ***Y. Roizin*** [a], ***A.Fenigstein*** [a], ***A. Heiman*** [a] , ***E. Aloni*** [a], ***G. Rosenman*** [b]

[a] *Tower Semiconductor Ltd., P. O. Box 619, Migdal HaEmek, 23105, Israel*
[b] *Department of Electrical Engineering-Physical Electronics, Tel Aviv University, Ramat-Aviv 69978, Israel*

Abstract

A novel two-bit per cell non-volatile memory device was introduced into the regular CMOS core technology by minor modifications of the standard process flow. The memory cell utilizes charge trapping in the Silicon Nitride spacers of a standard MOS transistor for storing information (spacer Flash or S-Flash). The device is programmed by channel induced secondary electrons and erased with hot holes created in the drain junction. The proposed memory is intended for high density and low cost applications.

Introduction

The standard MOS transistors often include spacers of silicon nitride, a material with high density of deep traps. For a certain transistor design, trapping of electrons and holes in these spacers results in significant changes of the threshold voltage Vt [1,2]. In a standard CMOS manufacturing process, the nitride of spacers is placed on a thick (~200A) CVD oxide (Fig.1a). Decreasing of the bottom oxide (BOX) (Fig.1b) thickness and skipping of LDD extensions enhances trapping of electrons in the nitride, thus converting the standard n-channel MOS transistor into a non-volatile memory cell. The trapped in the spacer electron charge is monitored in a reverse read-out scheme (read-out in the direction reverse to programming[3], Fig.1b). Two bits of information (in the spacers from both sides of the gate) may be stored in a single memory cell, similar to NROM memories[3]. The cell can be erased by hot holes, generated by band-to-band tunneling (BBT) mechanism in the drain junction. In this paper we present an embedded memory of the discussed type introduced into the Tower Semiconductor 0.18um core CMOS process flow with only two additional masks.

S-Flash memory cell

The work on the S-Flash memory cell included transistor optimization and development of adequate programming/erase algorithms. The parameters influencing the cell performance are presented in Fig.1c. Source/drain junction engineering (the junction position and lateral field enhancing implants) was performed and the influence of BOX and sidewall (SOX) oxide thicknesses were studied. As already mentioned, decreasing the BOX thickness enhances the electron injection. Nevertheless, T_{BOX} must remain above the tunneling thickness (>50 Å [4]) to ensure the retention of the trapped charge. Another important parameter is the channel length overlayed by the trapping media (the overlap region). It strongly influences the operation window (the difference between Vt in programmed and erased states). To reach the sufficient operation window, the n+ source/drain implant dose was decreased compared to the corresponding doses in the core CMOS and the thickness of the nitride spacer was increased. Similar to the BOX, the sidewall oxide must guarantee the retention of the trapped charge. Nevertheless, increasing of the sidewall oxide thickness results in a strong deterioration of the erase performance. SOX thickness was optimized and fixed at the level of ~50A. The thickness of the gate oxide (GOX) remained as in the high voltage core CMOS transistor (~70 Å).

Channel Induced Secondary Electrons (CHISEL) mechanism is known as one of the most efficient for programming of the EEPROM cells [5]. Unfortunately, this mechanism can not be used in standard two-bit per cell NROM devices due to hot electron trapping in the channel far from the drain, which influences the Vt of the second bit. Since there is no trapping media in the middle of the S-Flash transistor channel, we could use a version of the CHISEL mechanism for programming. For this purpose, the memory array was placed into the isolated P-well (existing in the Tower Semiconductor core 0.18um CMOS) which could be biased with respect to the substrate.

Experimental results

The Id-Vg curves of the fresh and programmed memory cells are presented in (Fig.2a). Programming was performed by applying positive voltage pulses to the gate and drain and negative voltage pulses V_b to the P-well. The CHISEL mechanism was efficient even for low voltages and relatively short pulses. For the programming time τ_p= 1msec and voltages V_g~3V ; V_d ~ 5V ; V_b ~ -5V , the window above 1V is achieved (Fig.2a) and 2-bit per cell operation is demonstrated (Fig.2b). Programming with different V_b shows that efficiency of charge injection is dramatically improved by the P-well negative bias (Fig.2c). A dramatic decrease of the programming current I_d without reducing the programming rate (to I_d levels below 10uA) is achieved by lowering the gate voltage to ~2V, (Fig.2d,e).

In the retention tests (one time programming), the cells were baked at 250C for 1 hour. Only ~20% decrease of the memory window margin was observed.

Erasing with different gate (Fig.3a) and drain (Fig.3b) voltages showed that both have strong influence on the erase efficiency. In the endurance tests, the memory cell V_t was measured in both programmed and erased states. Fig.4a shows that the memory window remained constant. There was no need in V_d (V_g) adaptive algorithms usually employed in NROM memories[5]. The cells cycled to different numbers were left in the programmed state and retention test (1hour/250C) was performed. The V_t loss after the bake is shown in Fig.4b. After 100 program/erasing cycles there is still more than 300mV operation margin left.

The fabricated Array-TEG used the periphery circuits developed for Tower Semiconductor NROM memories and had an H-array geometry discussed in[7]. The initial Vt distribution is shown in Fig.5. The results of Array measurements (memory window, program/erase times, retention etc.) are consistent with those measured for single cells. No disturb issues different from those known for NROM memories were revealed in the studied range of cycles and employed voltages.

Discussion

The developed memory differs from similar solutions that employ trapping at the sides of the polysilicon gate electrode[1,2] by (i) special drain engineering; (ii) integration scheme and topology, (iii) programming mechanism. The overlap region (Fig.1c) was chosen to provide maximum window in identical programming conditions. For too large overlap distances (of the order of the spacer width) programming and erase are not efficient due to the loss of gate control over the channel. Too small overlaps do not allow sufficient charge trapping since the main part of the nitride spacer is situated over the strongly doped N+ region.

The endurance and retention limitations related to the BOX are similar to that of the standard NROM devices [8]. SOX thickness is also critical for the erase performance (Fig. 3c). Vertical field stimulates the BBT hole generation and lateral field is necessary for hole heating. It is important to note that Vg has a small influence on the programming rate. The gate voltage must reach the transistor threshold to supply channel electrons generating CHISEL. In contrast to NROM, where high fields in the dielectric stack are a must to facilitate the hot electron injection, in the employed programming scheme the electrons overcome the regular (or even increased after charge trapping) Si-SiO_2 barrier by acquiring energy in the biased to |Vd -Vb| drain junction. This junction is not overlapped by the gate electrode and has lower N+ doping, compared with the core CMOS devices, which results in the significantly higher breakdown voltages.

Summary and conclusions

A novel non-volatile memory was embedded into the standard CMOS process with minimum additional operations (only two non-critical masks added). The principle of charge trapping in the Nitride spacers of slightly modified n-channel transistors was verified at the level of 2Mb Array TEG. The memory cells were optimized for the maximum operation window and efficiency of programming and erase with minimum programming voltages and currents. Programming and erase with voltages below |5|V, excluded the need of special high voltage periphery circuitry.

References

1. M. Fukuda, T. Nakanashi and Y. Nara IEEE ED Letters , **24**, 490-492 (2003).
2. H. Iwata and A. Shibata, US Patent 7.164. 167, 2007.
3. B. Eitan, P. Pavan, I. Bloom, E. Aloni, A. Frommer, D. Finzi, IEEE ED Lett., **21**, 543-545 (2000).
4. Y. Roizin, E. Pikhay and M. Gutman, IEEE ED Letters, **26**, 35-37, (2005).
5. J. D. Bude, M.R Pinto, R.K Smith, IEEE Transact. ED **47**, 1873-1881 (2000).
6. Y.Roizin, E.Pikhay, M. Lisiansky, A. Heiman, E. Alon, E. Aloni and A. Fenigstein, IEEE NVSMW, Monterey, CA, 2006 , pp.74-75.
7. J. Kim, US Patent 6,765,259, 2004
8. Y. Roizin, R.Daniel, S. Greenberg, M. Gutman, M.Lisiansky, V. Kairys, E. Pikhay, P.Zisman., IEEE NVSMW Monterey, CA, 2004, pp.85-87.

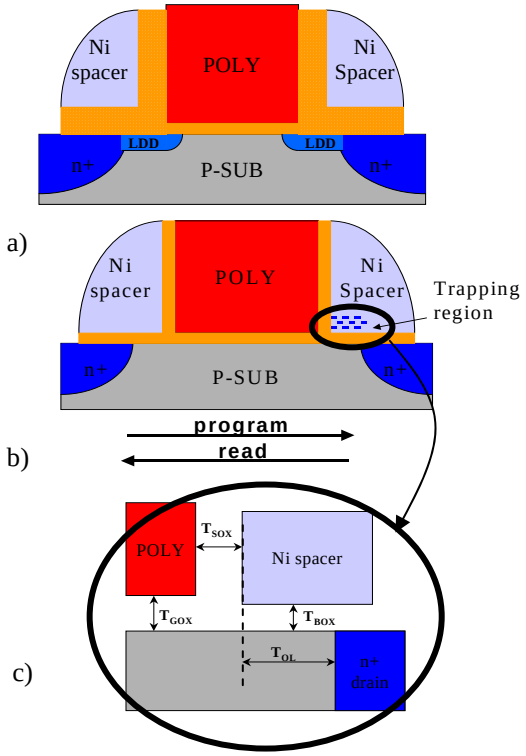
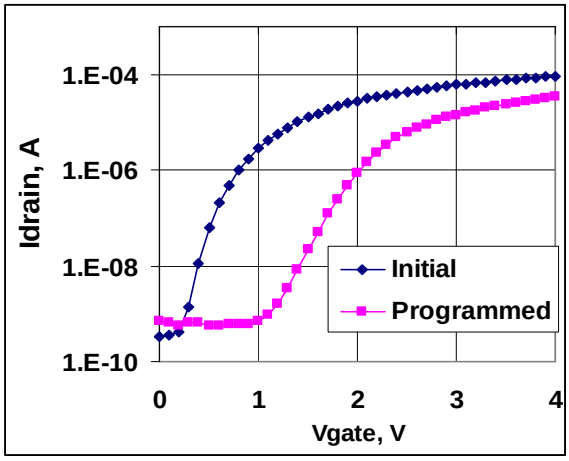
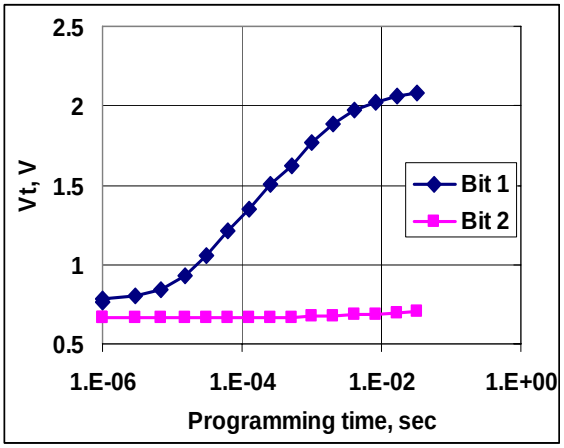


Fig.1. Schematic cross-section of the n-channel MOS transistor: a) standard; b) modified (no thick oxide under the nitride spacer and no LDD); c) trapping region thickness, T_{BOX} – bottom oxide thickness, T_{GOX} – gate – major parameters. T_{SOX} – sidewall oxide oxide thickness, T_{OL} – overlap region length.

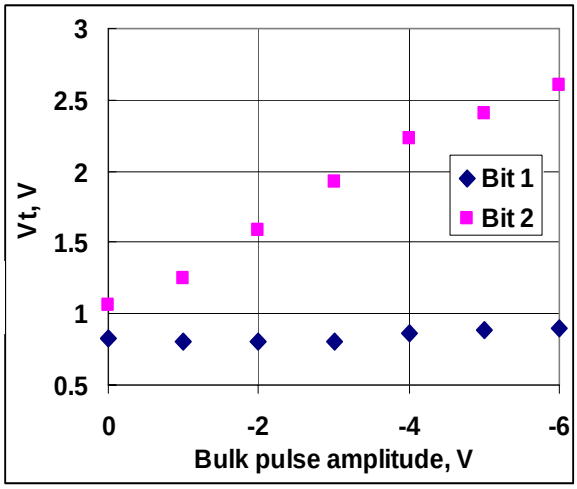
a



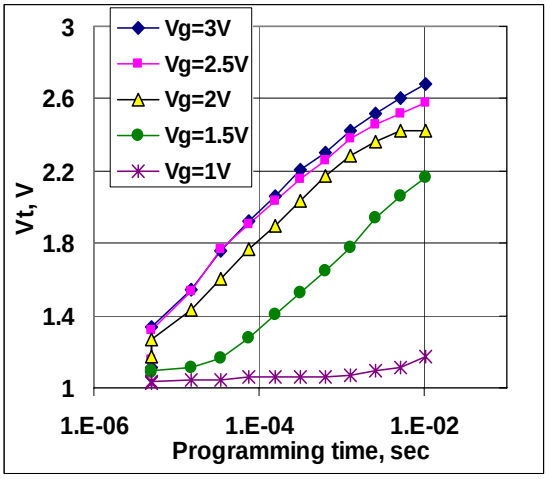
b



c



d



e

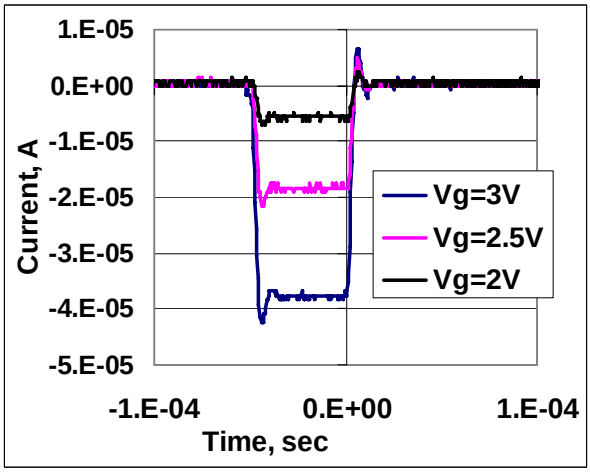
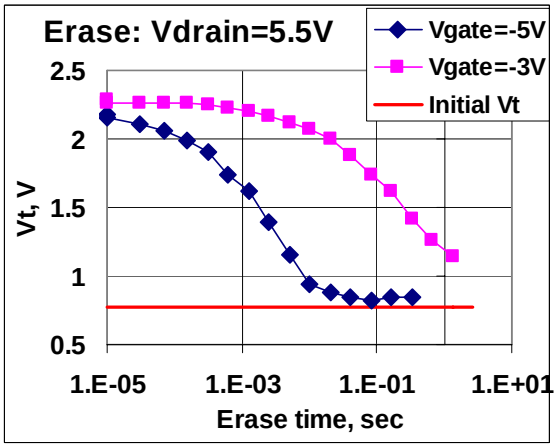


Fig.2. Programming by CHISEL mechanism: a) initial and programmed states; b) programming operation consists of pulses with increasing duration ; Vd=5V, Vg=3V, Vb=-5V, Vs=0, each point corresponds to Vt after a single pulse; c) dependence of the programming efficiency on bulk voltage, each point corresponds to Vt after the complete programming operation (Vdr=5V, Vg=3V, τ_p=1msec); d) programming by different gate voltages (Vd=5V, Vb=-5V, Vs=0); e) measuring the source-drain current at programming shown at (d).

a



b

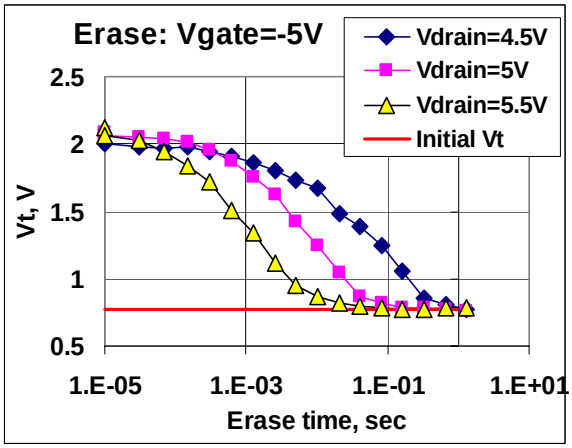
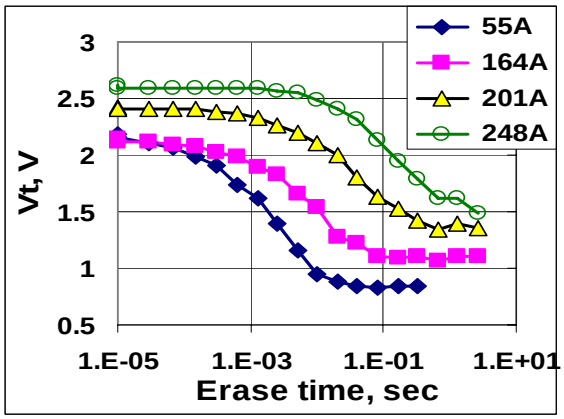
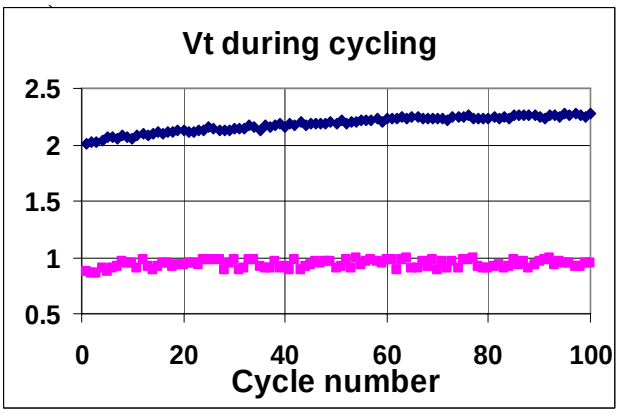


Fig.3. Erase performance as a function of gate (a) and drain (b) voltages and the thickness of sidewall oxide (c),Vd = 5.5V, Vg = -5V.

c



a



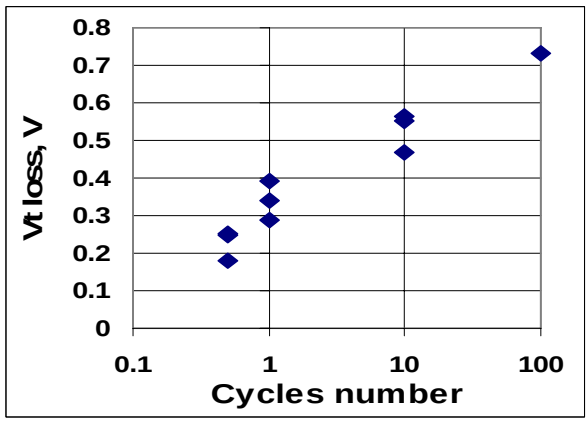


Fig.4. Endurance experiment: after performing the P/E cycling (OTP, 1, 10 and 100 cycles) the cells were programmed and baked at 250C for 1hour. The Vt loss after the bake is summarized in this graph.

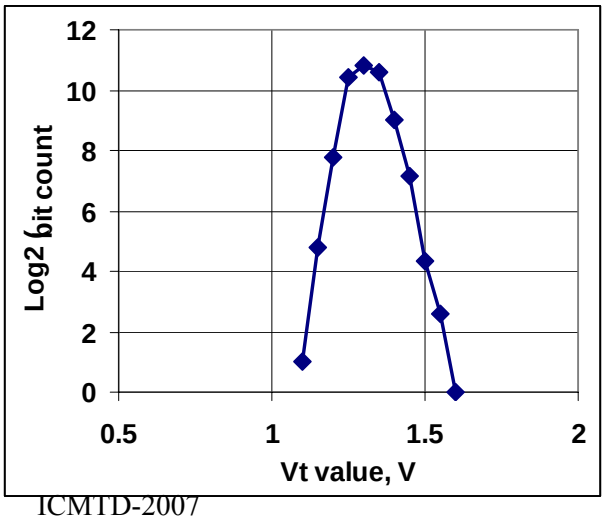


Fig.5. Vt distribution of the erased S-array sector (5680 bits). Measurements with a MOSAID tester. Sensing of Vt is performed at 12uA.

Comparison of DPT (Double Patterning Technology) vs. R (Reversal)-DPT using Off-set spacer for Bit-line contacts of 76nm pitch on NAND Flash cell

Jang-Ho Park, Byungjoon Hwang, Jaehwang Shim, Kwangseok Lee, Sunghyun Kwon, Sang-Yong Park, Donghwa Kwak, Jaekwan Park, Keonsoo Kim, Kinam Kim

Advance Technology Advanced Technology Development Team, Semiconductor R&D Center, Memory Business, Samsung Electronics Co., Ltd. Hwasung-City, Kyungki-Do, Korea, 449-900
Phone:+82-31-208-2004, Fax :+82-31-209-3274, e-mail: fuzzycom.hwang@samsung.com

Abstract

As the design rule of NAND Flash with high density is scaled down, it is required to form a small bit-line contact with the low resistance, as well as the low junction leakage current due to the borderless contact. In this paper, we introduce two novel processes, DPT (Double Patterning Technology) and R (Reversal)-DPT process, to form 38nm small size contact by using 193nm ArF lithography equipment. The R-DPT process using off-set spacer is suggested for NAND Flash device with 76nm pitch technology to minimize the contact resistance and the variation of contact resistance.

1. Introduction

It is the crucial issue to obtain a small bit-line contact with low resistance, low variation of resistance and low junction leakage for high density NAND Flash memory as shown in Fig.1 [1][2][3]. Furthermore, as the design rule is reduced, the contact patterning is one of the major limitation factors [4][5]. In order to overcome limitation, DTP (Double Patterning Technology) and R (Reversal)-DPT process are suggested to form small contact by using the 193nm ArF lithography equipment. We will introduce the process sequence of DPT and R-DPT process respectively and show the contact resistance variation of both processes caused by the process variation. Thus, a novel process, the R-DPT is proposed by providing excellent contact resistance and small variation.

2. Fabrication and Experimental

The process sequences of DPT and R-DPT to form 38nm bit-line contact are compared in Fig. 2. The common process sequence of both is as follows; (1) SiO_2 (~550nm) for ILD (Inter Dielectric Layer) and 1st poly-Si layer (~150nm) as a hard mask for odd patterns are deposited. And 38nm dot-type and 114nm space are defined in DPT and 38nm line and 114nm space are defined in R-DPT respectively by using 193nm ArF lithography and RIE etching equipment. (2) ALD (Atomic Layer Deposition)-SiO_2 (~38nm) for an off set spacer and 2nd poly-Si (~100nm) layer for even patterns are deposited. And then (3) the etch-back of the 2nd poly-Si is proceeded to form even patterns between the odd patterns. In DTP process, ALD-SiO_2 and ILD-SiO_2 layer between 1st poly-Si layer and 2nd poly-Si layer are etched back (~200nm) to form a pillar as a hard mask. Photo-resist is filled in the etched region. The pillar is etched using photo-resist. Finally, (4-1) the remained ILD-SiO_2 layer is etched to form the contact using selective etchant, and then (5-1) the doped poly-Si is filled in the contact hole after deposition of Ti/TiN as a glue layer, and the poly-Si is etched back and tungsten is filled as a final material of bit-line contact. (6-1) tungsten is flattened by CMP (Chemical Mechanical Polishing) process. In new scheme, R-DPT process, off-set spacer (ALD-SiO_2) and ILD-SiO_2 (~5500nm) are etched using the additional photo step and 1st and 2nd poly-Si (hard mask) as shown in the plane view SEM image of (4-2) in fig.2. And then the following steps are same as the DTP process as shown in Fig. 2.

3. Results and Discussion

Table 1 shows the comparison of process complexity and process induced CD variation of DPT and R-DTP. The R-DPT uses one more photo mask, less than 2 etch steps and half time in CD variation compared with those of DTP. Furthermore, there is an inherent problem which is the size difference between the odd and even contacts in the DPT process. If the odd contacts become bigger than the target, the even contacts which are made by the odd contacts are too small. However the size of even and odd contacts is uniform in the R-DPT process. Because ALD-SiO_2 layer with a good uniformity is etched to form the contacts. Fig. 4 and fig. 5 show the co-relation between the odd and even contact size and the distribution of contact size respectively. In addition, we could obtain the better contact resistance, junction leakage current and the distribution of them using the R-DPT process in fig. 6, 7 and 8.

4. Conclusion

The reduced active area forces to form small bit-line contacts with low resistance, low variation of contact resistance and junction leakage current due to a borderless contact for scaling down of a design rule to develop a high density NAND Flash device. In this paper, a novel process, the R (Reversal)-DPT (Double Patterning Technology) process has been developed to make 38nm small size contact with 76nm pitch by using 193nm ArF lithography equipment.

References

[1] ITRS (International Technology Roadmap for Semiconductors), www.itrs.net, 2005
[2] Kinam Kim et al., "Future Outlook of NAND Flash Technology for 40nm Node and Beyond", IEEE NVSMW, pp. 9~11, 2006.
[3] Jung-Dal Choi et al., "A 0.15 um NAND Flash Technology with 0.11um^2 Cell for 1 Gbit Flash Memory",IEDM, pp. 767~770, 2000.
[4] D.H.Kim et al., "Borderless Contact Leakage Induced Standby Current Failure on Sub-0.15um CMOS Device", IPFA, pp. 165~168, 2001.
[5] B.J.Hwang et al., "Development of viscous PR flow technology for 0.26um contact pitch on 0.84um2 SRAM cell", ESSDERC, pp. 391~394, 2003.

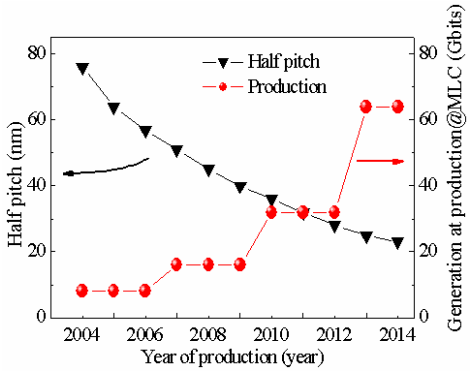


Fig. 1 Road map of NAND Flash [4]

Item		DPT	R-DTP
Process Complexity	Photo Step	1ea	2ea
	Etch Step	24ea	19ea
	CVD Step	4ea	4ea
	CMP Step	1ea	1ea
CD Uniformity		38nm±8 (3*ℓ+1*m)	38nm±4 (1*ℓ+1*m)

ℓ=2nm: CD variation@Photo/Etch, m=2nm: thickness variation@Off-set spacer

Table 1 Comparison of the process complexity and the process induced CD variation of DPT and R-DPT.

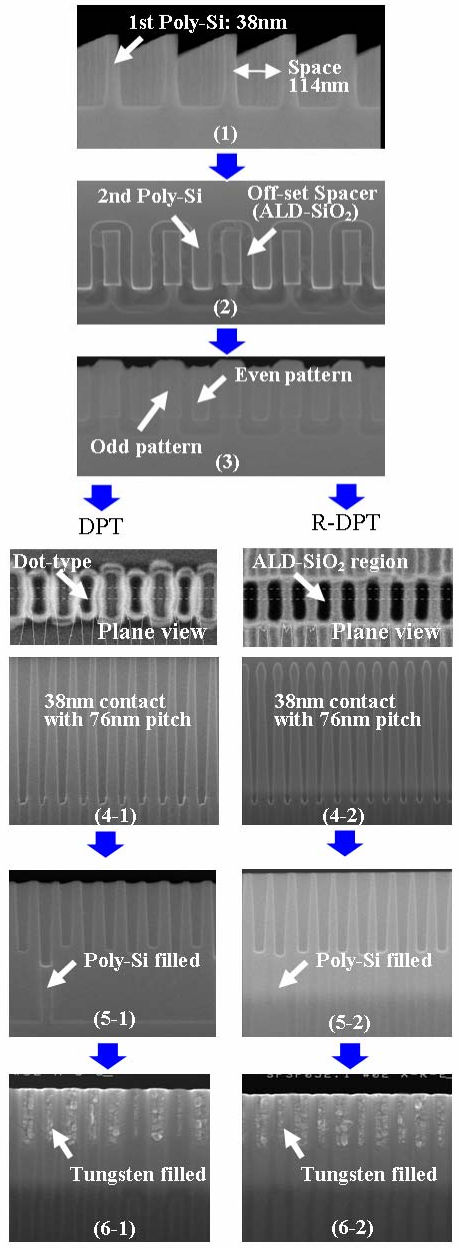


Fig. 2 The process sequence of DPT and R-DPT with plane view SEM images and cross section SEM images of 38nm contact with 76nm-pitch

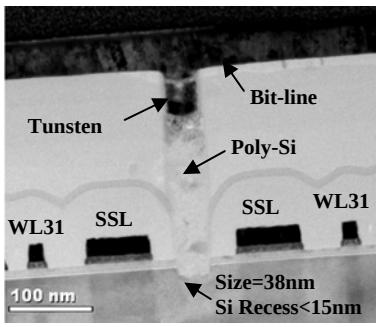


Fig. 3 The Y-axis cross section TEM image 38nm contact between SSL and SSL

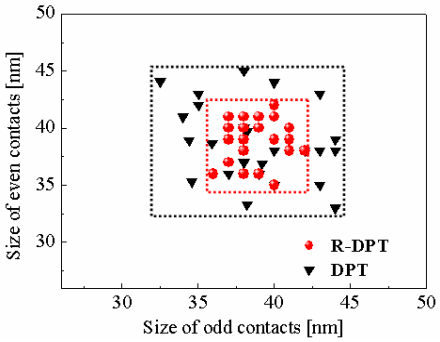


Fig. 4 Size co-relation of odd and even contacts of DPT and R- DPT

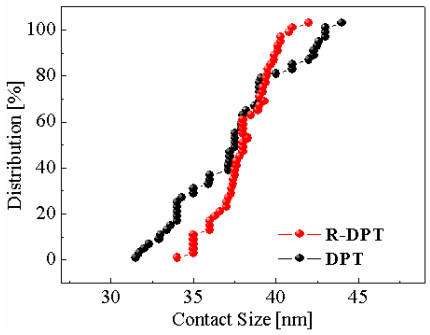


Fig. 5 Distribution of contacts size of DPT and R- DPT

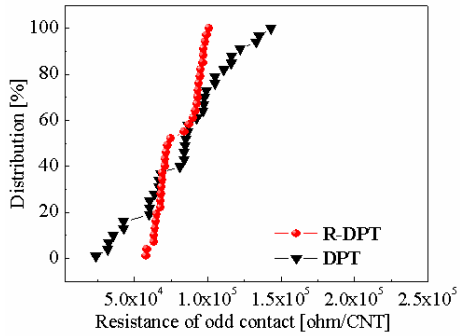


Fig. 6 Distribution of odd contact resistance of DPT and R- DPT

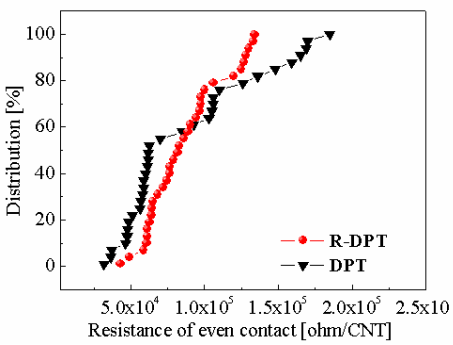


Fig. 7 Distribution of even contact resistance of DPT and R- DPT

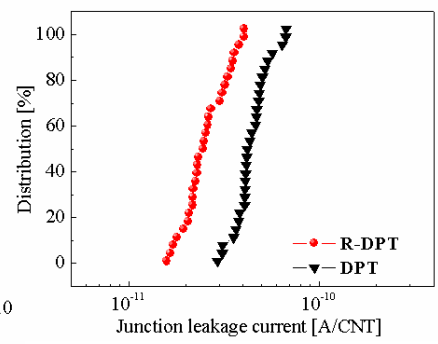


Fig. 8 Distribution of junction leakage current of DPT and R- DPT

Depletion 2-Transistor-SONOS Flash memories with zero gate voltage read out

Nader Akil, Michiel van Duuren, Do Dormans*, Dick Boter*, Almudena Huerta Miranda, Dušan Golubović, Rob van Schaijk*, and Michiel Slotboom*

NXP Semiconductors, Research, Kapeldreef 75, B-3001 Leuven, Belgium Email:Nader.Akil@nxp.com

**NXP Semiconductors, Gerstweg 2, 6534AE Nijmegen, The Netherlands*

Abstract

In this paper we present a solution to eliminate the gate read disturb encountered in N-type SONOS (silicon-oxide-nitride-oxide-silicon) flash memories. The time to failure under read disturb is extended from a few weeks to 10 years with this method. Moreover, read out with a gate voltage of zero volt becomes possible avoiding voltage boosting during read, which minimizes power consumption and reduces silicon area. This is achieved by using a 2-Tramsistor-SONOS (2T-SONOS) bit-cell with a depletion mode NMOS SONOS transistor. We demonstrate the concept on embedded 2T-NOR flash memory arrays in 0.18µm CMOS technology. By a careful design of the channel profiles, the depletion concept is scalable to sub 100nm gate length for future generations of flash memories.

1. Introduction

Non-volatile memories (NVM) based on charge trapping layers are considered potential candidates to replace floating gate NVM, thanks to their ease of integration in CMOS baseline and moderate program and erase voltages [1], [2]. SONOS NVM's are operated with the direct tunneling (DT) mechanism for programming and erasing and can offer low power [3] and high endurance [4] options for embedded flash memories. Among the major issues encountered in SONOS are erase saturation [3] and gate read disturb [4]. The first one is caused by electron back tunneling from the gate to the nitride which causes a leveling off of the threshold voltage (V_T) during erase and limits the size of the programming window [5]. The gate read disturb is due to the relatively thin tunnel oxide (~2 to 3nm) used in SONOS. A positive gate voltage (V_g) during read operation causes, on the long term, soft programming of the erased cells, thus destructing the stored data. To suppress the read disturb, low or even zero gate voltage during read out operation is beneficial [6].

In this paper we show that when depletion NMOS SONOS is used, a symmetric programming window centered around zero volt can be obtained. The device can be read with zero gate voltage to eliminate the gate disturb, and to reduce the periphery area and power consumption. In order to prevent over-erase issues we use a 2T-cell in which an enhancement MOS selection device in series with each depletion SONOS transistor. Subsequently, we show that the depletion concept can be scaled for future flash memory generations when the channel doping profiles are carefully designed.

2. Device description

2T-SONOS devices, as shown in fig. 1, in NOR architecture are considered in this paper. Next to every SONOS memory transistor an access gate (AG) is added in series to enable low voltage read and avoid over-erase. The bottom oxide, nitride and top oxide thickness of the ONO stack considered in this paper are 2, 6 and 8nm respectively (equivalent oxide thickness is approximately 13nm). Two types of SONOS devices are investigated here: enhancement NMOS SONOS (reference) and depletion NMOS SONOS. Both devices have an n-type gate. Fig. 2 illustrates SIMS profiles of the channel doping of a depletion n-type SONOS transistor. Shallow n-type doping (e.g. Arsenic) is implanted in a p-type substrate to lower the natural threshold voltage (V_{Tn}) to negative values, while a high Boron dose is implanted deeper and acts as anti-punch through (APT), to suppress short channel effects. The shallow Arsenic implant is done blanket in the flash memory area. An extra masked p-type implant is used to transform the AG transistor into an enhancement transistor. This is needed for proper selection during read and program inhibit operations, where the AG has to isolate the bitline voltage (e.g. 0.5V for read and 5V for program-inhibit) for non-selected cells [7]. The AG overdope implants are done with the same mask which is applied to remove the ONO from the AG channel, as shown in fig. 3. The SONOS 2T cell is embedded in a 0.18µm process [8]. The gate length of the CG and AG transistors is 0.24µm as is their width.

3. Results and discussion

The transfer characteristics of enhancement and depletion SONOS devices are compared in fig. 4 at a read bitline (BL) voltage of 0.5V. The V_{Tn} is shifted down by 1.6V in case of depletion devices at the expense of some degradation of the sub-threshold swing which is about 160mV/decade for depletion SONOS compared to 95mV/decade for enhancement SONOS. The program and erase curves of both enhancement and depletion devices are compared in fig. 5. A shift of the P/E curves is observed without affecting the shape of the curves. As shown in fig. 5, a symmetric programming window around 0V of about 3V can be obtained in case of depletion SONOS only. Therefore, the depletion SONOS device can be read at a gate voltage V_{CG}=0V instead of ~1.6V for enhancement SONOS. This will give significant improvements in device performance. The direct

impact is reduction of the gate disturb during reading. Fig. 6 illustrates the on-shelf retention and read disturb on both average programmed and erased states of 64k enhancement SONOS cells. The on-shelf retention shows a small remaining window after 10 years. In fact, this is true as far as the device is not read for a long time during the lifetime of the memory. However, in some situations the gate read voltage can be present continuously on the CG during a substantial part of the lifetime of the memory. In this case the gate read disturb is a major issue as shown in figure 6. After only a few weeks of gate stress of 1.6V (middle of the programming window at t=0s) the average of erased cells starts to be read as programmed state and give a reading error. In case of depletion SONOS, the on-shelf retention and gate read disturb are identical since in both cases 0V is present on the CG. Fig. 7 shows the read disturb of average programmed and erased 64k depletion SONOS. The decay of the P/E curves is no longer disturbed by the read voltage, and the device can be continuously read for 10 years (extrapolated) without suffering from read disturb.

Besides, the depletion concept also helps reducing the area of the periphery. In fact, in advanced CMOS nodes like the 90nm and beyond, the supply voltage is below 1.2V. To read with V_G=1.6V (in case of enhancement) there is a need for a boosting circuitry which consumes a large silicon area to keep fast read access (in the nano-second range) and consumes power [9]. In case of depletion SONOS, the boosting circuitry for reading is not needed.

As stated above, the AG channel has been over-doped by Boron to transform it to an enhancement device. Fig. 8 and 9 show, respectively, the transfer characteristics and the Drain-Induced-Barrier-Lowering (DIBL) behavior of the over-doped AG compared to a reference enhancement AG. No significant change in the AG characteristics between the two devices is found.

One main drawback of the depletion concept is the limited scaling of the CG length due to short channel effects. To overcome this issue, the channel profiles have to be redesigned using halo implants as shown in fig. 10. The simulated V_{Tn} roll-off of depletion and improved depletion CG with halo's versus CG channel length are compared in fig. 11. The depletion devices with and without halo implants show similar V_{Tn} for long channels, but with halo implants V_{Tn} is nearly constant for devices down to 100nm gate length. As shown in fig. 12 depletion+halo's device still has negative V_{Tn} with acceptable[1] sub-threshold swing (360mV/decade). Accordingly, the depletion device with halo implants shows good CG length scaling and can be used for sub 100nm SONOS flash memories.

4. Conclusion

We have shown that using depletion SONOS instead of enhancement SONOS in 2T-NOR embedded flash can increase the time to failure under read disturb from a few weeks to 10 years. We have demonstrated the depletion SONOS concept in embedded flash memories in a 0.18 μm CMOS process. Depletion SONOS can be read with V_{CG}=0V, therefore no need for a boosting circuitry when the supply voltage is scaled down in advanced CMOS generations, which reduces the periphery area and power consumption. The depletion concept can be combined with halo's to allow scaling of the gate length beyond 100nm for advanced embedded flash memory generations.

[1]Remember that the CG is not used for selection but only to discriminate between 0 and 1 states.

References

[1] M-K Seo, et al., *IEEE Journal of Solid-State Circuits*, vol. 40, No. 4, pp. 877-883, April 2005.

[2] M. H. White, et al., *IEEE Trans. Components, Packaging, and Manufacturing Technology*, Part A, vol. 20, No. 2, pp. 190-195, June 1997.

[3] R. van Schaijk, et al., *Solid State Electronics*, vol. 49, pp.1849-1856, November 2005.

[4] B. de Salvo, et al., *IEEE Trans. Device Materials Reliability*, vol. 4, No. 3, pp. 377-389, September 2004.

[5] H. Bachhofer, et al., *Journal of Applied Physics*, vol. 89, No.5, pp. 2791-2800, March 2001.

[6] C.T. Swift, et al., *in Proc. IEDM*, 2002, pp.927-930.

[7] N. Akil, et al., *IEEE Trans. Elec. Devices*, vol.52, No.4, pp. 492-499 April 2005.

[8] D. Dormans, et al. *in Proc. SSDM*, 2001, pp. 540-541.

[9] G. Palumbo et al., IEEE Trans. Circ. Syst.-part I, vol. 49, No. 11, pp.1535-1542, November 2002.

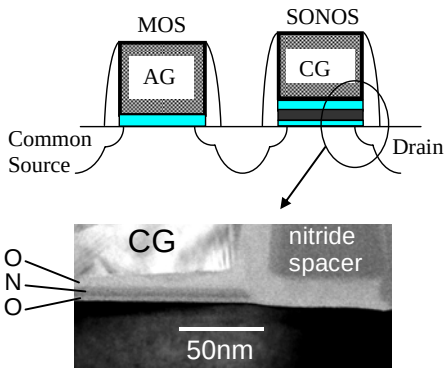


Fig.1: 2D cross section of a 2T SONOS cell. The drain is connected to a metal bitline. The TEM picture gives a magnification of the ONO stack at the gate edge.

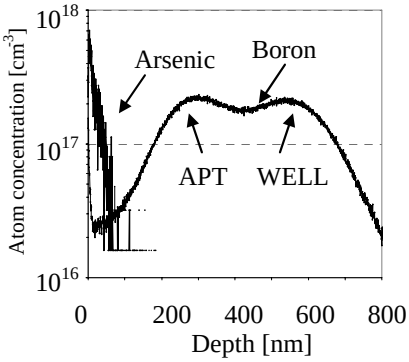


Fig.2: SIMS of the doping profiles below the CG of an NMOS depletion SONOS transistor. Depth=0nm corresponds to the substrate surface.

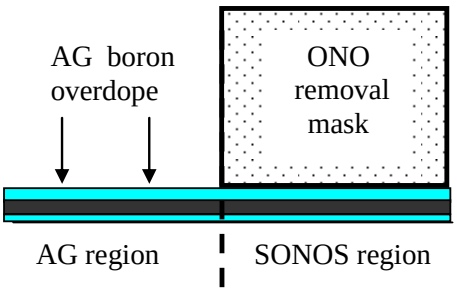


Fig.3: The AG overdope is done using the ONO removal mask just before ONO removal from the AG area.

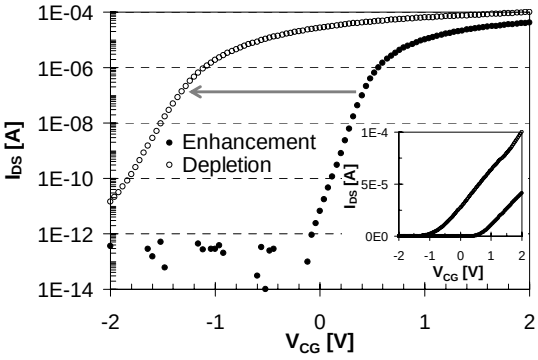


Fig.4: Transfer characteristics of enhancement and depletion SONOS devices with V_{DS}=0.5V. The data are plotted on a linear scale in the inset graph.

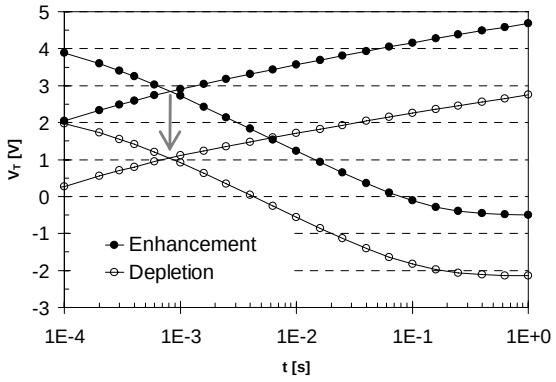


Fig. 5: Program/erase curves of enhancement and depletion SONOS. The data represent the average V_T of 4k cells. P: V_{CG}=+11V, E: V_{CG}=-11V.

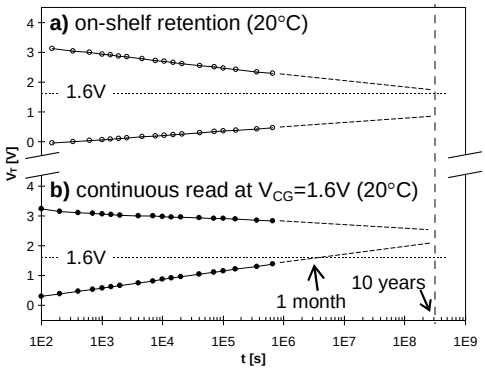


Fig.6: a) On shelf retention and b) gate disturb on programmed and erased enhancement SONOS cells in 2T-NOR architecture. The data are average V_T of 64k cells. P: V_{cg}=+11V for 10ms, E: Vcg=-11V for 100ms. The gate disturb is done with V_{cg}=1.6V.

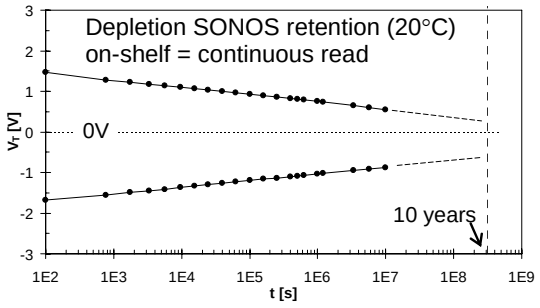


Fig. 7: Continuous read disturb with V_{CG}=0V of programmed and erased depletion SONOS cells. The data are average V_T of 64k cells. P: Vcg=+11V for 10ms, E: Vcg=-11V for 100ms.

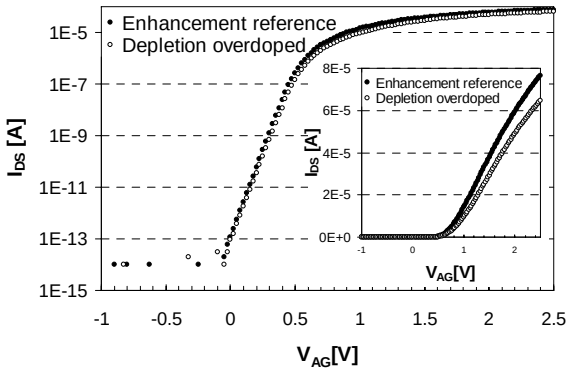


Fig. 8: Transfer characteristics at V_{DS}=0.5V of a standard enhancement AG and over-doped (depletion transformed to enhancement) AG.

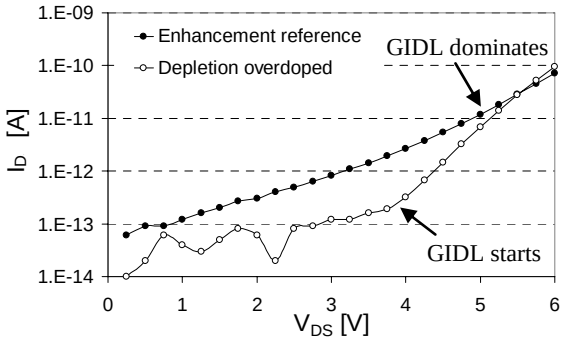


Fig. 9: DIBL behavior of the enhancement and the over-doped (depletion transformed to enhancement) AG transistors at V_{AG}=0V. The arrows indicate where the Gate-Induced-Drain-Leakage (GIDL) start and where it becomes dominant.

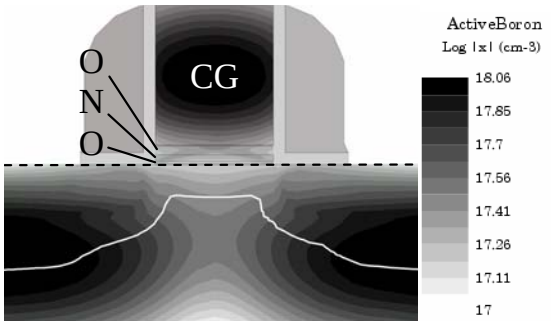


Fig. 10: 2D process simulation of a depletion NMOS SONOS device with boron halo implants to improve short channel effects. The white lines represent the metallurgical junction positions.

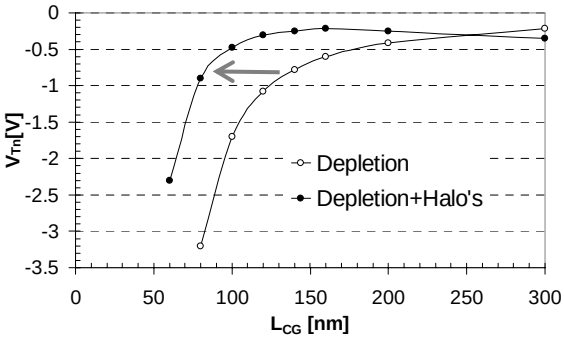


Fig. 11: Simulated V_{Tn} roll-off versus L_{CG} for the depletion and the depletion+halo's devices at V_{DS}=0.5V

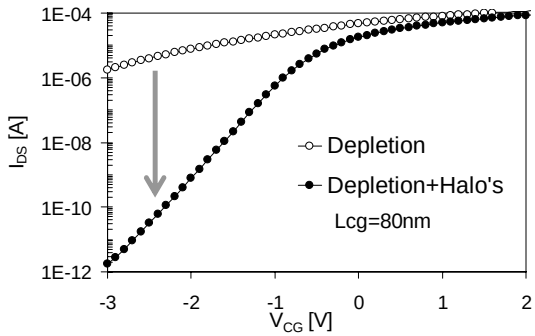


Fig. 12: Simulation of the transfer characteristics of 80nm CG SONOS device.

Physical understanding and modeling of SANOS retention in programmed state

Arnaud Furnémont[a,b], Maarten Rosmeulen[a], Antonio Cacciato[a], Laurent Breuil[a], Jan Van Houdt[a], Kristin De Meyer[a,b] and Herman Maes[a,b]

[a] Interuniversity MicroElectronics Center, Kapeldreef 75, B-3001 Heverlee, Belgium.

[b] K. U. Leuven, ESAT, Kasteelpark Arenberg 10, B-3001 Heverlee, Belgium.

E-mail: furnem@imec.be Phone: +32-16-288-779

Abstract

In this paper, the retention in programmed state of the SANOS technology is for the first time accurately analyzed and modeled. Firstly, the retention is studied on capacitors to determine the main retention mechanisms. The electron detrapping in the silicon nitride, followed by tunneling through the aluminum oxide is found to be the dominant mechanism causing the retention loss. The modeling of this effect reproduces the observed temperature, gate workfunction and window dependency. Secondly, these results are applied to scaled devices where the retention is dominated by the same mechanisms. The slight difference in the retention loss between capacitors and devices is explained by process-induced fixed charges in the stack leading to a different field distribution in the gate dielectric.

1. Introduction

Nowadays, the Flash memory market is experiencing a tremendous growth, mainly due to the success of portable electronics such as digital camera's and mp3 players. This success is very challenging for the research field because it coincides with reaching the scaling limits of the traditional floating gate technology. Other concepts must be developed, e. g., SANOS ($Si/Al_2O_3/Si_3N_4/SiO_2/Si$) memories, which may soon replace the floating gate for NAND applications, due to their excellent performance, scalability and their process simplicity [1-2]. Before mass production applications, a complete understanding of the device needs to be built in order to define the specifications and the limits of the device, but also to go past these limits and optimize the concept.

In particular, the retention is the most important feature of a memory device. At the same time, it is the most difficult to study because the specification of 10 years is not reproducible in laboratory. The purpose of this paper is thus to build a physical understanding of the retention in SANOS devices, by determining the dominant mechanisms and the parameters which influence the device behavior. Firstly, only the vertical component of retention will be studied using 50 by 50 μm^2 capacitor devices (ANO = 10/5/4 nm). Capacitors with ONO gate dielectric (6/5/4 nm) will also be useful to isolate retention mechanisms due to the aluminum oxide. Secondly, scaled devices will be measured to check the validity of the results on real cells.

2. Retention in capacitors

Different mechanisms can explain the retention loss in programmed SANOS capacitors, as shown in figure 1: electron tunneling from the nitride to the channel (I), electron tunneling from the nitride to the gate (II), hole tunneling from the channel or the gate (III and IV), and thermal electron detrapping from the nitride, followed by tunneling to the channel (V) or to the gate (VI). If the aluminum oxide is assumed to be without a large number of defects, the tunneling of holes or electrons through the top oxide (mechanisms II and IV) is unlikely compared to the tunneling through the bottom oxide (I and III), because of its physical thickness.

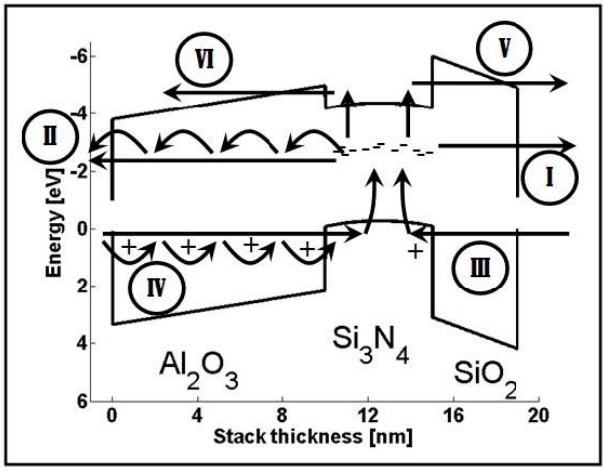


Fig. 1: Band diagram of SANOS in programmed state. The retention can be driven by six different mechanisms: electron tunneling from the nitride to the channel (I), electron tunneling from the nitride to the gate (II), holes tunneling from the channel or the gate (III and IV), thermal electron detrapping from the nitride, followed by tunneling to the channel (V) or to the gate (VI).

Figure 2 shows the strong temperature dependence of the retention in SANOS. Hence, the dominant mechanisms cannot be the tunneling through the bottom oxide (I and III) because direct tunneling has limited temperature dependence. However, the room temperature loss is not negligible, which would be expected if only thermal detrapping determines the retention loss. The comparison with the SONOS retention gives further information (fig. 3). Contrary to SANOS, the SONOS retention does not depend on the temperature. Besides, this SONOS retention coincides with the room temperature retention of SANOS. The SANOS retention is thus the addition of two components. Firstly, a non-temperature dependent mechanism, occurring also in SONOS, can be explained by mechanisms I or III of figure 1. This also proves the validity of the first assumption: mechanisms II and IV can not be the cause of this retention loss because it occurs also in SONOS devices. Secondly, a temperature dependent component

is driven by the detrapping of electrons, followed by tunneling through the top oxide barrier (VI). Indeed, the detrapped electrons can not escape through the SiO_2 (V) because this component is not visible in SONOS.

It is very difficult to separate mechanisms I and III from measurements. However, it is well-known that nitride memories are mainly erased by injection of holes from the P-substrate, and not by electron tunneling through the bottom oxide [3]. The non-temperature dependent mechanism causing the retention loss at low temperature is thus likely the mechanism III of figure 1.

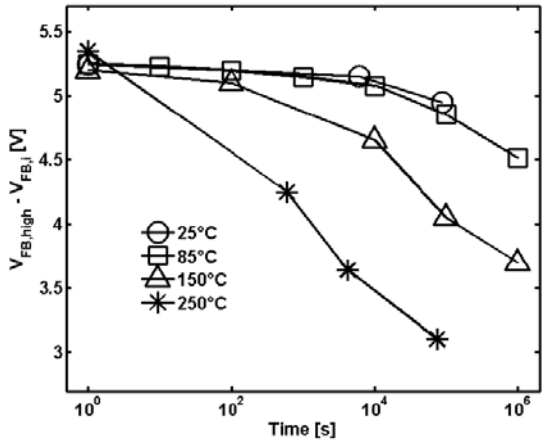


Fig. 2: Retention at different temperatures for a window of 5 V above the initial flatband voltage in SANOS capacitors.

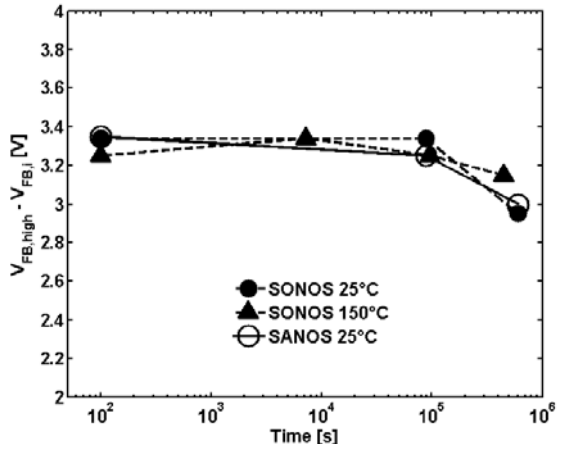


Fig. 3: Retention at room temperature (RT) for SANOS and SONOS capacitors, and at 150°C for SONOS. The SONOS retention does not vary with the temperature and is similar to the SANOS RT retention.

As shown in figure 3, this element of retention is not an important issue for the SANOS device. Temperature dependent mechanism on the contrary can be unacceptable for some applications, especially for multi-level memory cells which need large windows and very tight threshold voltage distributions [1]. The modeling of the mechanism can be performed if the phenomena explained in figure 4 are taken into account. Due to the electrons trapped in the nitride layer, the field pushes some detrapped electrons towards the Al_2O_3 barrier, and some other towards the SiO2, depending on their position in the nitride layer. The electrons close to the bottom oxide can not escape due to the high conduction band of the SiO_2. On the contrary, the other electrons, pushed to the Al_2O_3 barrier, can easily escape. Firstly because the difference between Al_2O_3 and Si_3N_4 conduction bands is limited. Secondly, because the band bending in the nitride creates a triangular potential close to the oxide, leading to a quantization of the available energy state, which still lowers the effective barrier seen by the electrons. Besides, the mean free path of an electron in silicon nitride is long, as explained in [4]. Another phenomenon is that for a symmetric device, exactly half of the detrapped electrons will go towards the Al_2O_3 barrier. However, this will not be the case with different Fermi levels between the gate and the channel (fig. 4B), because it induces an extra field component through the stack.

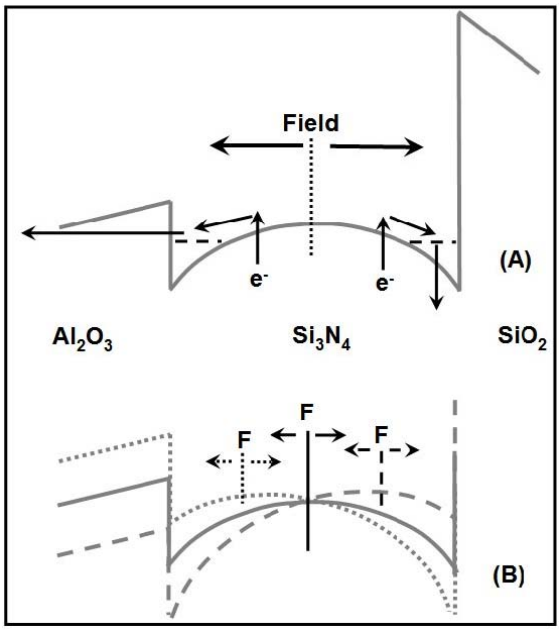


Fig. 4: Schematics of the electron detrapping during retention. Three phenomena have to be taken into account to understand this retention mechanism. Firstly, the electric field will force detrapped electrons towards the top oxide or towards the bottom oxide barrier depending on the proximity of this barrier. As the Al_2O_3 conduction band is much lower than in the SiO_2 case, the electrons close to the Al_2O_3 can escape easily, which is not the case for the electrons close to the SiO_2 barrier (A). Secondly, the tunneling through the Al_2O_3 is made easier by the triangular potential, which induces a quantization of the energy levels (A). Thirdly, different Fermi levels or workfunctions between the gate and the channel induce a different distribution of the field (B). This leads to different fraction of electron pushed to the top or the bottom barrier.

This last effect is illustrated in figure 5. Both P- and N-type gates are implemented in the SANOS capacitors, and their retention is compared. Due to boron diffusion, the workfunction of the P-type gate is estimated to be close to mid-gap, which gives a Fermi level offset of 1 eV between the channel and the N-type gate, and 0.5 eV between the channel and the P-type gate. Clearly, the P-type gate device has a better retention than the N-type gate device, if the window is wide enough, i. e., if the band bending is large enough to have the phenomenon, described in figure 4B.

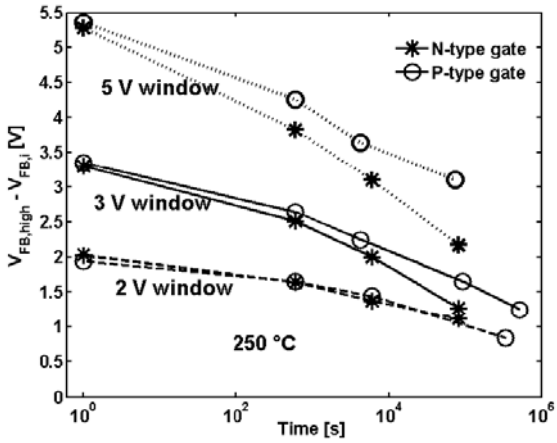


Fig. 5: SANOS capacitor retention with N- and P-type gates for different windows. The P-type gate cells have a significantly better retention for high windows.

Figures 6 and 7 show the agreement between the retention measurements and the model of this retention with both varying temperature and window. In this model, a gaussian distribution is assumed for the nitride trap energy, with 1.8 eV for the mean and 0.27 eV for the variance [5]. The field direction is calculated from the nitride charge (assumed to be uniformly distributed) and the difference between gate and channel workfunctions. The tunneling through the aluminum oxide is not the limiting factor, as proven by the temperature dependence, and the tunneling probability is considered to be 1. The mechanism III is taken into account by adding the room temperature retention to the model results. In figure 5, there is no difference between the N- and P-type gate device retention for small window because the flatband voltage shift is enough in both cases to have 100% of detrapped electrons directed to the Al_2O_3 (95% and 80% for N- and P-type gate devices respectively for a 3 V window, and 65% and 80% for a 5 V window).

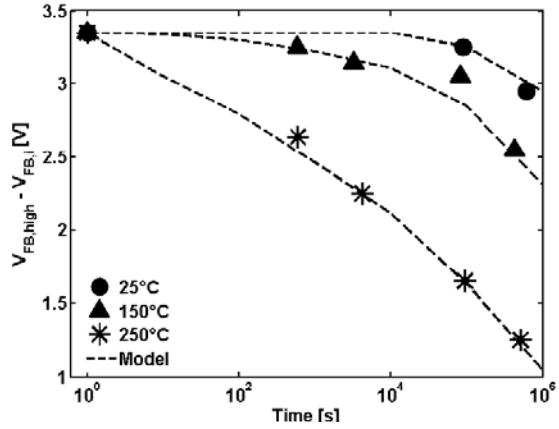


Fig. 6: The retention at different temperatures in SANOS can be reproduced by the model. The dominant mechanism is the detrapping from the channel and the tunneling through the Al_2O_3. The three phenomena explained in figure 4 are taken into account in the model. The temperature dependence is limited by the thermal detrapping.

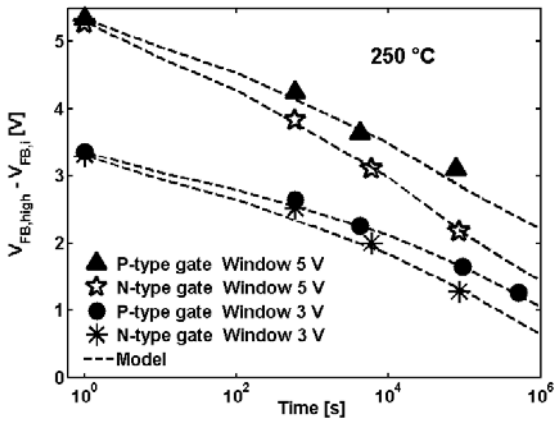


Fig. 7: The retention for N- and P-type gate SANOS can be reproduced by the model. The dominant mechanism is the detrapping from the channel and the tunneling through the Al_2O_3. The three phenomena explained in figure 4 are taken into account in the model. The gate type dependence is explained by the offset between the gate and the channel Fermi levels, which leads to a vertical field leading more or less the detrapped electrons driven towards the Al_2O_3 barrier.

3. Retention in devices

The retention in scaled devices (with the same gate stack) is measured and compared with the capacitor results in figure 8. At room temperature, the retention is similar, which means that the mechanism III of figure 1 is also occurring in scaled devices. But when the window increases, a discrepancy between both results appears. The thermal detrapping remains a major effect in the devices whatever the gate type or the window. However, not only a fraction but all the detrapped electrons seem to disappear. Three possibilities have to be considered. Firstly, the electrons directed towards the bottom oxide barrier can laterally spread to reach nitride regions out of the active area. Indeed, the nitride is not cut along the width and along the length, which gives place for the electrons. Secondly, the programming operation can be non-uniform due to possible non-uniformities in the channel doping level and the stack thickness, which gives the possibility of a lateral redistribution. Third, the process may induce interface or fixed traps in the aluminum oxide, which would lead to a field distribution such as every de trapped electron is directed towards the top oxide barrier, and none towards the high bottom oxide barrier.

To discriminate these three theories, a retention test is performed on devices with varying width and length (fig. 9). The thermal detrapping model is also applied, with the simple assumption that every detrapped electron escapes from the gate stack. The model fits with the measurements, and still more important, no significant W and L dependence can be observed, which shows that there are no border effects in the device. The uniformity of the programming operation is checked by measuring I-V curves during programming (fig. 10). As all subthreshold slopes are perfectly parallel, the uniformity is confirmed. The conclusion is that the vertical field is high enough to direct any detrapped electron towards the

top oxide barrier, which allows this electron to escape by tunneling through the Al_2O_3.

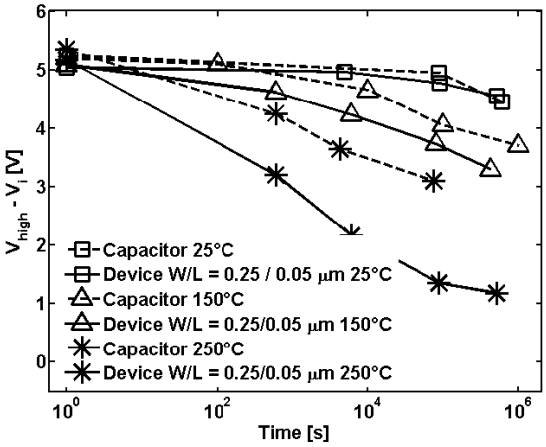


Fig. 8: Measurements of retention for SANOS capacitors (ΔV_{FB}) and scaled devices (ΔV_{TH}). Both show similar characteristics at room temperature, but the cells show worse retention for higher temperatures.

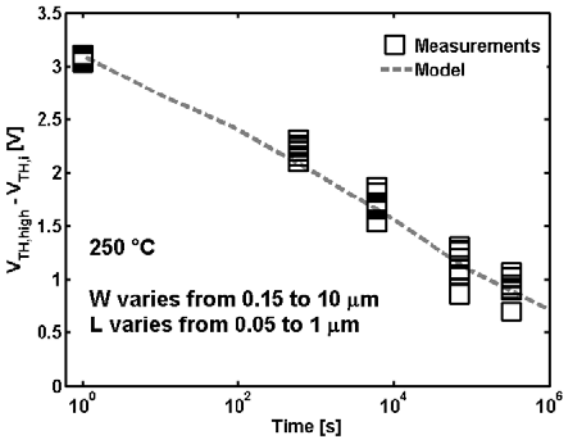


Fig. 9: The SANOS cell retention does not depend on the width and the length of the device. The dominant mechanism is the thermal detrapping from the nitride, followed by the escape of every detrapped electron.

4. Discussion

The retention is SANOS-type memories is mainly driven by the thermal detrapping, followed by the tunneling of the charges through the aluminum oxide. To improve this retention, two alternatives can exist. Either the electrons have to be prevented to detrap, or the detrapped electrons must be confined in the nitride layer on top of the channel. The first solution necessitates to find a better trapping material than the silicon nitride. Alternatively, it is maybe possible to increase the top oxide barrier in order to prevent the tunneling, or to increase sufficiently the gate work function to push the detrapped electrons towards the silicon oxide barrier. This result can also be achieved by adding negative charge in the aluminum oxide layer. But even if the electrons cannot escape vertically anymore, they can still laterally spread, which can only be prevented by cutting the nitride layer exactly on top of the channel.

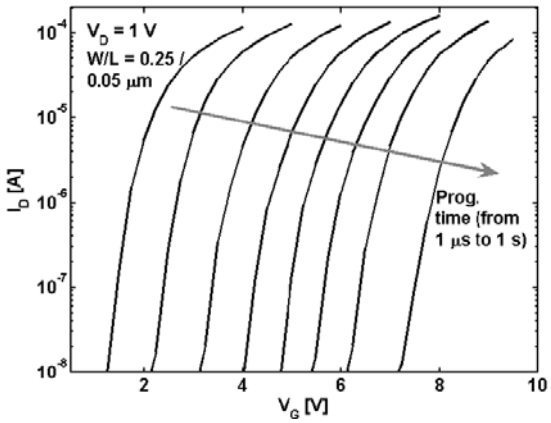


Fig. 10: Drain current versus gate voltage measured after different programming times in a scaled device, from 1 μs to 1 s with 1 decade step. All the subthreshold slopes are parallel, which indicates a fully uniform programming operation.

5. Conclusion

The retention loss in SANOS-type devices can be split into two different components. Holes tunneling through the bottom oxide towards the nitride induces a slight retention loss independent of the temperature. A second mechanism dominant at high temperature is the thermal detrapping of electrons in silicon nitride, followed by the tunneling of these electrons through the aluminum oxide barrier. For the first time, a model with simple assumptions allows to reproduce the retention for different gate types, windows and temperature. This model allows to predict the long term retention performance but also to determine the key parameters in order to improve the retention.

Acknowledgements

This work was carried out within the framework of the IMEC Industrial Affiliation Program on Advanced Flash Memory together with Infineon Technologies, Intel Corp., Micron Technology and Samsung Electronics. The authors would like to thank Luc Haspeslagh and Joeri De Vos for their help in the device fabrication.

References

[1] Y. Park, J. Choi, C. Kang, C. Lee, Y. Shin, B. Choi, J. Kim, S. Jeon, J. Sel, J. Park, K. Choi, T. Yoo, J. Sim, and K. Kim, A highly manufacturable 32-Gb multi-level NAND flash memory with 0.0098 μm2 cell size using TANOS cell technology", IEDM Tech. Dig., pp. 29-32 (2006).

[2] K. Kim and J. Choi, "Future outlook of NAND Flash technology for 40nm node and beyond", NVSMW Proc., pp. 9-11 (2006).

[3] M. White,. D. Adams, and J. Bu, "On the go with SONOS", IEEE Circuits and Devices, Vol. 16, pp. 22-31 (2000).

[4] A. Furnémont, M. Rosmeulen, J. Van Houdt, K. De Meyer, and H. Maes, "Model for electron redistribution in silicon nitride", proc. ESSDERC, pp. 447-450 (2006).

[5] A. Furnémont, M. Rosmeulen, K. van der Zanden, J. Van Houdt, K. De Meyer, and H. Maes, "Physical modeling of retention in localized trapping nitride memory devices", IEDM Tech. Dig., pp. 397-400 (2006).

40nm TANOS (TaN - Al_2O_3 - SiN - Oxide - Si) Cell Technologies for High Density NAND Flash Memory Applications

Bonghyun Choi, Youngwoo Park, Changseok Kang, Changhyun Lee, Yoochoel Shin, Juhyung Kim, Sanghun Jeon, Jongsun Sel, Jintaek Park, Jaesung Sim, Chungil Hyun, Wonseok Jung, Beomjin Kim, Sun-kyu Hwang, Youngjae Kim, Janghyun You, Joo-heon Kang, and Jungdal Choi

Semiconductor R&D Center, Memory Business, Samsung Electronics Co., Ltd.
San #16, Banwol - Dong, Hwasung - City, Kyunggi-Do, Korea, 445-701

Phone: +82-31-208-2762 Fax: +82-31-208-4799 E-mail: bonghyun.choi@samsung.com

Abstract

40nm TANOS cell technologies for 32Gb multi–level NAND flash memory have been successfully demonstrated. The advanced blocking oxide in the TANOS memory cell and high numerical aperture (N.A.) photolithography for patterning are key technologies.

1. Introduction

Floating-gate (FG) cell has been used as a memory cell, which stored data in the flash memory device since its invention. However, below 50nm design rule, the FG faces several difficult problems such as cell-to-cell interference, reduced coupling ratio, and process complexity. As an alternative way to overcome the difficulties in the FG technology with scale-down of design rule, a TANOS NAND cell has been developed using 40nm design rule [1]. In this previous work [1], fully working 32Gb NAND flash cells were demonstrated. In addition, it was noted that cell characteristics such as program and erase speed are highly dependent on the property of the Al_2O_3 blocking oxide. Further improvement of erase speed is necessary to suppress charge loss.

In this paper, we report that cell characteristic was further improved by adopting a sophisticated post-deposition treatment after Al_2O_3 deposition.

2. Experimental

A TANOS cell was fabricated using the integration scheme as shown in table.1. A dielectric composite of SiO_2/SiN/Al_2O_3/TaN was adopted for the TANOS NAND cells. Word line and bit line were patterned with a pitch of 90nm as shown in Fig. 1. Immersion lithography technology with high N.A. was used to pattern critical layers such as active, word line, bit line contact and bit line. The cross-sectional views of fabricated 32Gb NAND flash with the TANOS cell structure along bit line direction are presented in Fig. 2. Between the bit line contact and the common source line contact, 32 cells and two select transistors are connected in series. Note that two select transistors have the same structures as the cell transistors.

Fig. 3 shows the cross-sectional SEM and TEM images of Shallow Trench Isolation (STI) profiles along the word line direction. Adjacent cell transistors were isolated by using STI. The depth of STI is about 200 nm and the active width is about 40 nm. The TANOS cell composite was composed of 40Å-thick SiO_2, 70Å-thick SiN, 150Å-thick Al_2O_3 and 170Å-thick TaN as a tunnel oxide, a trap layer, a blocking oxide, and a control gate, respectively. The SiN and Al_2O_3 of the TANOS cell composite was formed by LPCVD and ALD method respectively. W/WN layer with a low sheet resistance below 5Ω/□ were deposited on the TaN to reduce gate resistance. After the patterning of word lines, subsequent BEOL process was performed by conventional NAND flash technologies as shown in table. 1.

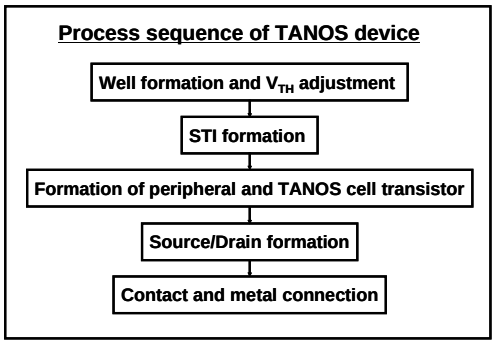


Table. 1. Fabrication process of TANOS-NAND flash memory.

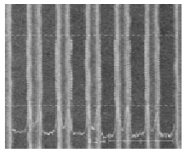


Fig. 1a. Active.

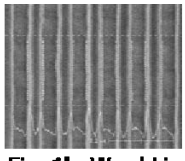


Fig. 1b. Word Line.

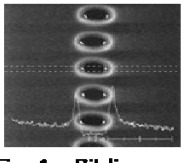


Fig. 1c. Bit line contact.

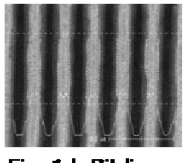


Fig. 1d. Bit line.

Fig. 1. Top-view images of critical lithography layers.

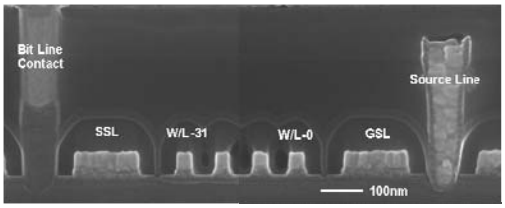


Fig. 2. Cross-sectional SEM images of a TANOS cell string including select string line (SSL) and ground select line (GSL).

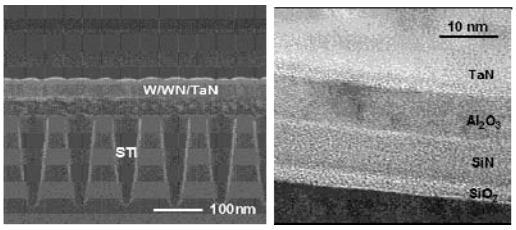


Fig. 3. Cross-sectional SEM image of TANOS STI profile and TEM photograph of a cell composite.

3. Result and discussion

Fig. 4 shows high temperature storage (HTS) characteristics of TANOS cells as a function of blocking oxide thickness. With increasing the blocking Al_2O_3 thickness, HTS characteristics both for no cycle and 1k cycling were dramatically improved. The improvement can be mainly attributed to the reduction of tunnelling current through the blocking Al_2O_3 layer during retention mode at high temperature of 200°C.

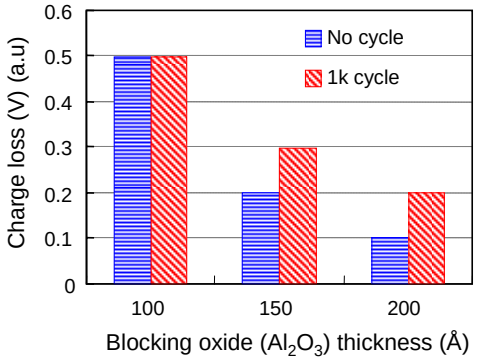


Fig. 4. Charge loss characteristics of TANOS cells as a function of the conventional blocking oxide (Al_2O_3) thickness.

As shown in Fig. 5, it is found that charge loss of the TANOS cell is highly dependent on the erase voltage (Verase) ; the lower Verase results in the smaller charge loss. In this respect, lowering Verase is desirable to reduce charge loss.

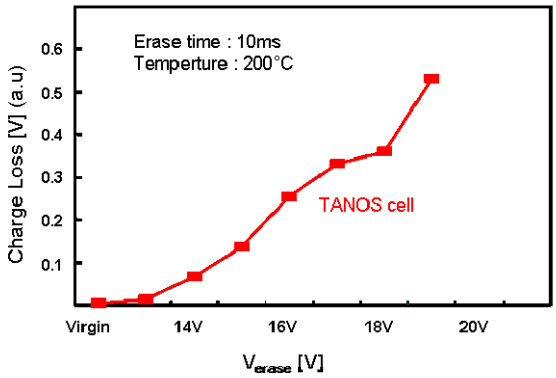


Fig. 5. Charge loss characteristics of TANOS-NAND cells as a function of the erase voltage.

Sufficient operation window of threshold voltage (> 6V) should be achieved for realizing the MLC operation. To obtain fast erase speed for MLC operation in the case of conventional TANOS process, high Verase (>20V) is essential for the thickness of TANOS structure. However, this high Verase is unacceptable in the respect of HTS requirement.

As a solution to improve erase speed, we are proposing an advanced Al_2O_3 process. In Fig. 6, erase speed of TANOS-NAND cells was compared to two Al_2O_3 processes: one is a conventional process reported in [2], the other is the advanced Al_2O_3 process adopted the sophisticated post-deposition treatment after Al_2O_3 deposition. Erase threshold of -1V was achieved at erase time of 10ms using 19V in the conventional process, while erase threshold of -2.8V was obtained for the same erase condition in the advanced Al_2O_3 process.

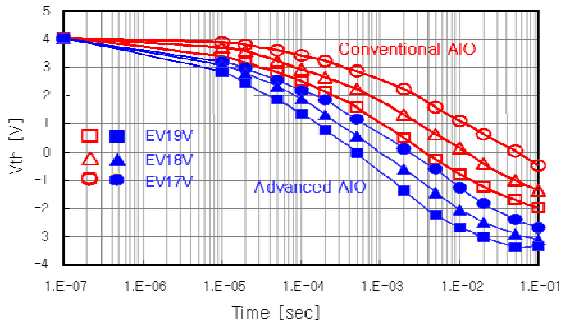


Fig. 6. Erase characteristics of TANOS-NAND cells for two different Al_2O_3 conditions.

The improvement of the erase speed in the advanced Al_2O_3 process can be explained by the reduced leakage current compared to the conventional one as plotted in Fig. 7. The advanced Al_2O_3 process shows approximately 1 order smaller leakage current density both in lower electric field and in higher electric field regions than the conventional one. Therefore, the back tunnelling through advanced Al_2O_3 is significantly reduced at the erase operation, which results in faster erase speed in the advanced Al_2O_3.

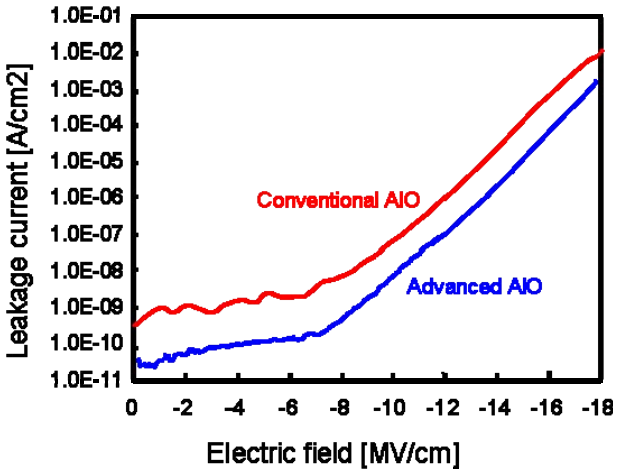


Fig. 7. Leakage characteristics of Al_2O_3 as a blocking oxide.

Due to the relaxed erase condition by the reduction of erase speed, endurance characteristics were improved in the advanced Al_2O_3 process [data not shown].

As shown in Fig. 8, Vth distribution with a wide enough gap between states for multi-level operation was achieved even at a 40nm TANOS NAND cell. The Vth distribution is similar to one of a 63nm TANOS NAND cell reported before [3].

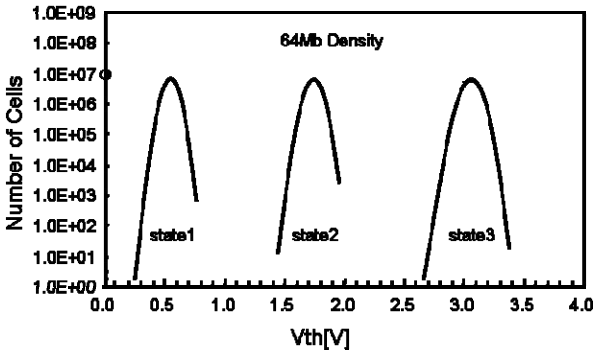


Fig. 8. Cell Vth distribution of 64Mb cells by multi-level cell programming.

4. Conclusion

A 40nm cell technology for multi-level flash memory applications was developed based on the TANOS cell structure. As a key technology of TANOS NAND cell, an advanced Al_2O_3 was obtained by the process optimization of post-deposition treatment. By using the advanced Al_2O_3 with lower leakage current during erase operation, better erase characteristics and excellent HTS characteristics were achieved. These advanced technologies give highly manufacturable process margin for high density NAND flash memories with 40nm design rule and beyond.

References

[1] Youngwoo Park, "Highly Manufacturable 32Gb Multi-Level NAND Flash Memory with TANOS (Si-Oxide-Al_2O_3-TaN) Cell", IEDM Tech.Dig., 29 (2006).
[2] Yoochoel Shin, et al., "A Novel NAND-type MONOS Memory using 63nm Process Technology for Multi-Gigabit Flash EEPROMs", IEDM Tech. Dig., 337 (2005).
[3] Chang-Hyun Lee, "Multi-Level NAND Flash Memory with 63nm-node TANOS (Oxide-SiN-Al_2O_3-TaN) Cell Structure", VLSI Tech. Dig., 26 (2006).

Effect of Al_2O_3 morphology on the erase saturation performance in SANOS-type memory cells

A. Cacciato , A. Furnémont, L. Breuil, J. De Vos, L. Haspeslagh, J. Van Houdt

IMEC, Kapeldreef 75, B-3001 Leuven, Belgium, E-mail: cacciato@imec.be

Abstract

In this paper we use both short-loop capacitors and fully integrated cells to evaluate the effect of the post-Al_2O_3 deposition thermal treatment on the erase saturation performance of $Al_2O_3/Si_3N_4/SiO_2$ stacks. It is found that the temperature of the post deposition anneal is a very critical parameter to fully exploit the beneficial effect of Al_2O_3 as blocking layers. In particular, in order the layer to be effective in reducing the erase saturation effect, the temperature should be high enough to cause a complete crystallization of the Al_2O_3 film. In this case, an improvement of 4 V in the saturated threshold voltage of the erased cells is observed.

1. Introduction

The SONOS-type cell, for its excellent scalability and process simplicity, is the candidate to push the scaling roadmap for FLASH memories beyond the intrinsic limit imposed on floating-gate flash memories by the electrostatic interference between adjacent cells [1].

In particular, combined with a high-k top dielectric and a multilevel programming scheme, the SONOS concept could pave the way towards the 32 and 22 nm nodes [2]. The use of a high-k dielectric reduces the electric field across the top dielectric, thus suppressing the unwanted FN gate injection current during the erase operation. This, in turns, allows a thicker tunnel oxide, thus contributing to improve data retention.

Among the possible high-k materials, Al_2O_3 has already been found to be able to significantly improve the erase operation [3, 4], guaranteeing at the same time excellent endurance and sufficient bake retention [5].

However, the post-deposition annealing treatment is known to have a strong influence on the transformation kinetics of the Al_2O_3 film from the amorphous to the crystalline phase [6]. This phase change could have a relevant impact on the erase performance of the SANOS devices.

In this paper we use both short-loop capacitors and fully integrated cells to evaluate the effect of the post-Al_2O_3 deposition thermal treatment on the erase saturation performance of $Al_2O_3/Si_3N_4/SiO_2$ stacks. In particular the effect of annealing temperature and ambient are analyzed.

2. Device fabrication

Poly-Si/Al_2O_3/Si_3N_4/SiO_2/c-Si (SANOS) and Poly-Si/SiO_2/Si_3N_4/SiO_2/c-Si (SONOS) stacks have been manufactured using ISSG oxidation to grow the bottom SiO_2 layer, a standard LPCVD process to deposit the Si_3N_4 charge trapping layer and a HTO oxide as top dielectric in the SONOS stack. For the SANOS stack, atomic layer deposition carried out at 300 ^oC (with H_2O as precursor) was used to deposit the Al_2O_3 films. A densification anneal was carried out immediately after Al_2O_3 deposition. The Al_2O_3 thickness was varied to take into account the densification of the layer upon annealing [6, 7] so to achieve the target thickness of 10 nm, independently from the annealing temperature. For example, to obtain a 10 nm layer upon annealing at 1000 ^oC a 12 nm film was deposited. In the following, except when explicitly reported, all the SANOS data refers to devices that received a post-annealing treatment of 1000 ^oC for 60 s in N_2 ambient. The thicknesses of the ANO and ONO stacks used for this work are 10/5/4 nm and 6/5/4 nm, respectively, ensuring an EOT of about 10 nm for both stacks.

After the gate stack formation, the in-situ doped, 100 nm poly-Si gate was deposited. Both n+-type (Phosphorus doped) and p+-type (Boron doped) poly-Si gates are evaluated.

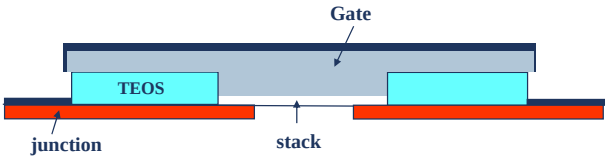


Fig. 1: Schematic cross-section of the SONOS/SANOS capacitors

The electrical characterization of the nitride stacks is carried out either on large (50x50 μm^2) capacitors or full integrated memory cells. Capacitors were manufactured by using a short-loop flow (three masks and no STI processing). The capacitor area was patterned by opening a window on a 300 nm TEOS, which is then wet etched to avoid damaging the Si surface. After window opening, the stack and the poly-Si gate were deposited and patterned. Unlike in the fully integrated cells, in the short-loop capacitors edge effect are expected to play no role on the performance of the charge-trapping stack. In fact, in the case of capacitors, the gate stack etch lands on the 300 nm-thick TEOS (over-etch is therefore not critical) and there is a large overlap of the poly over the capacitor area (> 1.5 μm), Finally, after poly-Si

patterning, TEOS is removed from the junction and salicidation is carried out (the schematic of the capacitor cross-section is reported in Fig. 1).

The integrated memory cells were manufactured using a 130 nm CMOS platform where the standard gate oxide has been replaced by the ANO (or ONO) stack. Gate lengths down to 50 nm were obtained by applying both hard mask and resist tripping during gate patterning.

3. Electrical characterization

Capacitors and cells are electrically characterized by measuring the program/erase (P/E) transients of the flat band (V_{fb}) and threshold (V_{th}) voltage, respectively. The program and erase voltages are 16 V and -18 V.

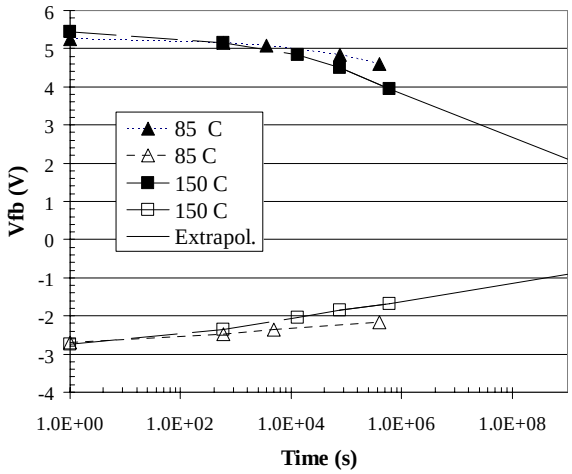
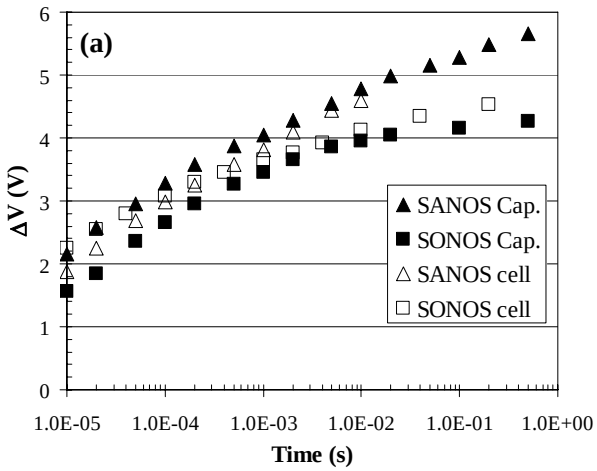


Fig. 2: Data retention at 85 °C and 150 °C of SANOS capacitors with p-type gate.

As shown in Fig. 2, large (about 8 V) P/E windows are obtained in the case of SANOS stacks. The P/E window decreases upon bake time. However, good data retention is achieved, extrapolated data suggesting that it would still be as large as 3 V after 10 years bake at 150 °C. This result proves the quality of the deposited stacks.



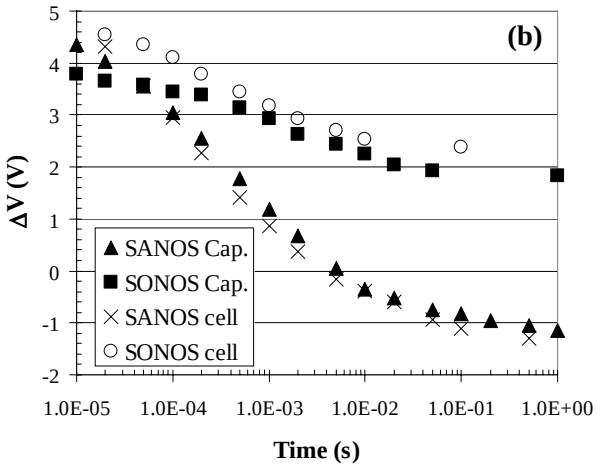


Fig. 3: V_{fb} and V_{th} shift during the program (a) and erase (b) operation for SANOS and SONOS stacks with p-type gate. The program and erase voltages are 16V and -18V respectively.

In Fig. 3 the shift of V_{fb} (for capacitors) and V_{th} (for cells) as a function of the program (a) and erase (b) time is shown for SANOS and SONOS stacks with p-type gate. As expected [3], the most relevant difference between SONOS and SANOS stack is observed during the erase operation. In particular, data in Fig. 3b show that the erase operation for SONOS devices saturates at 2 V. Therefore it is not even possible to reach the intrinsic level for this kind of stack. For a SANOS stack, on the other hand, the erase operation saturates at about -1 V, showing that the presence of Al_2O_3 causes an increase of the P/E window of ≈ 3 V. It should also be noted that no significant difference is observed in the figure between cells and capacitors. This demonstrates that the more complex processing in the case of the memory cells does not modify the intrinsic P/E properties of the stack, thus ensuring that the cell results reported in this paper are not affected by process marginalities (like, for example severe over-etch during the gate etch).

4. Effect of Al_2O_3 morphology

Figure 4 compares the erase transients of p-type gate SANOS capacitors for two different PDA temperatures (1000 °C versus 700 °C) and two different annealing ambients (N_2 versus O_2). For all variants the annealing time was 60 s. Results clearly show a decrease of the saturated V_{fb} of about 4 V when the annealing temperature increases from 700 °C to 1000 °C. It is to be noted that, the erase performance of the 700 °C variant is ≈ 1 V worse that on standard SONOS capacitors (compare data in Fig. 4 and Fig. 3b). The advantage of using Al_2O_3 instead of SiO_2 as blocking layer is therefore completely lost for low-temperature PDA treatments.

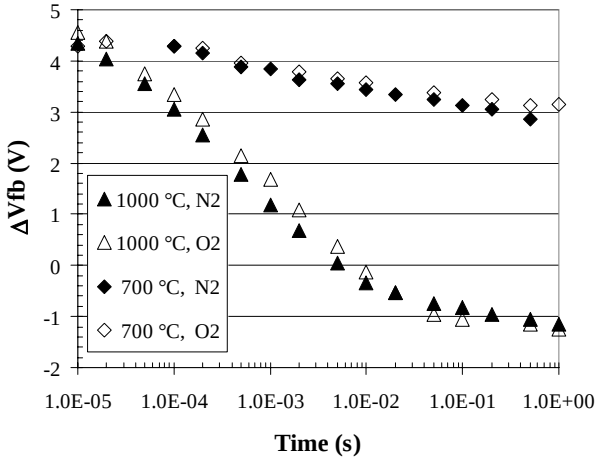


Fig. 4: Effect of the post-Al_2O_3 deposition annealing treatment (temperature and ambient) on the erase transient of SANOS capacitors with p-type gate. The program and erase voltages are 16V and -18V respectively.

The huge impact of the annealing temperature on the erase saturation performance of the SANOS capacitors cannot be explained by a change in the dielectric constant of the layer upon crystallization. In fact, this change has been reported to be negligible [7]. Most probably, the shift of the Al_2O_3 conduction band upwards by ≈ 0.5 eV upon crystallization reported in [6] is one of the causes instead. Indeed, such a shift would increase the potential barrier between Al_2O_3 and Si_3N_4 thus reducing electron tunnelling from the gate into the nitride layer, which is the root cause of the erase saturation effect.

Oxygen is known to slow down the crystallization of Al_2O_3 films [6]. However, unlike for the annealing temperature, no significant influence of the annealing ambient on erase saturation is observed. This indicates that at 1000 °C the amorphous/crystalline transformation occurs too fast to be significantly slowed down by the presence of oxygen.

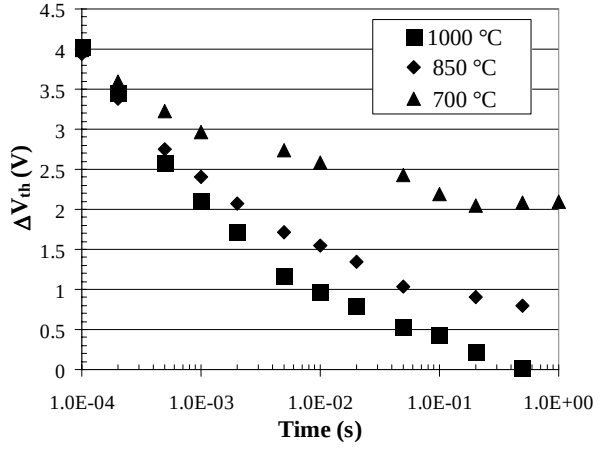


Fig. 5: Effect of the post-Al_2O_3 deposition annealing temperature on the erase transient of SANOS transistors (W/L = 0.25/0.12) with n-type gate. The program and erase voltages are 16V and -18V respectively.

Data on fully integrated cells (Fig. 5) add further evidence to the importance of the Al_2O_3 phase on the erase saturation. The figure shows the erase transients measured on n-type gate, SANOS cells for three different PDA temperatures: 1000 °C, 850 °C and 700 °C. The PDA anneal was carried out in N_2 ambient for all the variants.

Similarly to what observed in the SANOS capacitors, also in the case of fully integrated cells, the erase saturation improves by increasing the PDA temperature. It can also be noted that an annealing at 850 °C for 60 s is not sufficient to fully transform the amorphous Al_2O_3 layer into a crystalline film.

Unlike for SANOS capacitors, SANOS cells received a spike anneal at 1030 °C after S/D implantation. The poor erase saturation performance of the 700 °C variant in Fig. 5 clearly indicates that this anneal is not sufficient to crystallize the Al_2O_3 layer, demonstrating that the amorphous-crystalline transformation of Al_2O_3 is retarded when the film is capped by poly-Si.

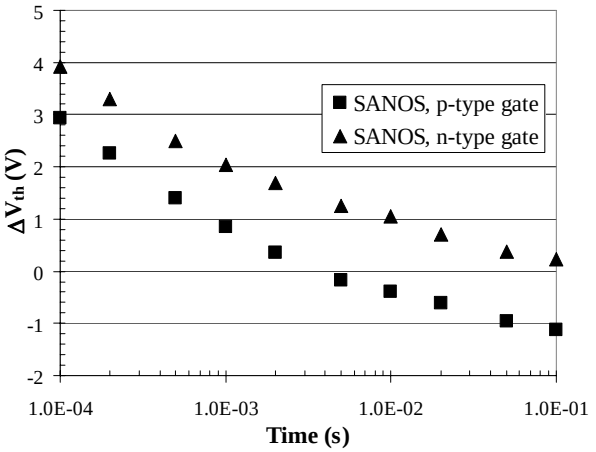


Fig. 6: Threshold voltage versus erase time for SANOS transistors (W/L = 0.25/0.12) with p-type and n-type gate. The program and erase voltages are 16V and -18V respectively.

The saturated erased V_{th}'s in Fig. 5 are about 1 V higher than those reported in Fig. 4 (see Fig. 6 for a direct comparison). Because data in Fig. 5 refer to n-type gate SANOS cells whereas those in Fig. 4 to p-type gate cells, this difference could be explained by the shift of the Fermi level towards the Si valence band caused by the change from n-type to p-type doping. However, the smaller impact of this shift ($\approx$ 1V) compared to that caused by the Al_2O_3 crystallization ($\approx$ 4 V) suggests that unlike expected from deposition conditions (the doping concentration in the in-situ doped poly-Si is about 10^{21} cm^{-3}), poly-Si in not degenerated. In fact, for degenerated Si a change from n-type to p-type should induce a change of the barrier for electron injection from the gate higher than the change of 0.5 eV caused by the amorphous/crystalline transformation in Al_2O_3.

SIMS analysis on fully-processed wafers showed that the doping concentration in the poly-Si is $\approx 6x10^{20}$ cm^{-3}. This concentration is sufficient to obtain a degenerated poly-Si. Therefore, the small improvement of erase

saturation in p-type gate cells compared to n-type gate cells is not the consequence of doping out-diffusion during processing but of poor electrical activation or others poly-Si/Al_2O_3 interface effects.

4. Conclusions

We have shown that both on SANOS capacitors and fully integrated cells, the temperature of the post deposition anneal is a very critical parameter to fully exploit the beneficial effect of Al_2O_3 as blocking layers in SONOS-like stacks. In particular, it is argued that, in order the layer to be effective in reducing the erase saturation effect, the temperature should be high enough to cause a complete crystallization of the Al_2O_3 film. In this case, an improvement of 4 V in the saturated threshold voltage of the erased cells is observed.

Acknowledgments

This work was performed under IMEC's Industrial Affiliation Program on Advanced Memory Technology together with Intel Corp. and Infineon Technologies

References

[1] M. White, *IEEE Circuits and Devices*, **16**, (2000) 22;
[2] J. Van Houdt *ICICDT Proceedings*, p. 43 (2006);
[3] C.H. Lee, S.H. Hur, Y. S. Shin, J. H. Choi, D. G. Park, and K. Kim , *SSDM Proceedings*, p. 162 (2002) ;
[4] C. H. Lee, K. I. Choi, M. K. Cho, Y. H. Song, K. C. Park, and K Kim. – *IEDM Tech. Dig*., (2003), p. 26.5.1
[5] Y. Shin, C. Lee, S. Hur, J. Choi, K. Kim, *IEEE Non-Volatile Memory workshop Proceedings*, (2003), 58 ;
[6] V.V. Afanas'ev, A. Stesmans, B.J. Mrstik and C. Zhao, *Appl. Phys.Lett*. **81**, 1678 (2002);
[7] B. Govoreanu, D.P. Brunco, L. Haspeslagh, J. De Vos, D. Ruiz Aguado, P. Blomme, G. Puzzilli, K. van der Zande, F. Irrera, J. van Houdt, presentation AM G1.4 to the *Material Research Society Spring Meeting* (2006).

SESSION H

High-κ Flash & Nano-crystals

A Systematic Study of High-K Interpoly Dielectric Structures for Floating Gate Flash Memory Devices

Lu Zhang, Wei He, Daniel S.H. Chan, and Byung Jin Cho
Silicon Nano Device Lab, Dept. of Electrical & Computer Engineering, National University of Singapore, Singapore 119260
Tel: 65-6516-6470 Fax: 65-6516-1103 email: elebjcho@nus.edu.sg

Abstract

A systematic simulation and experimental study is presented on application of high-K interpoly dielectric (IPD) for floating gate type Flash memory devices involving a variety of materials and structural combinations. A general guideline for the optimization of high-k IPD is proposed.

1. Introduction

The trend of aggressive scaling and low voltage operation for floating gate type Flash memory devices will soon require a reduction of the IPD layer thickness down to 6 ~ 8 nm EOT, as indicated by ITRS in Fig. 1. In addition, the loss of sidewall capacitance coupling between control and floating gates in further scaled devices will significantly drop the coupling ratio (CR) as shown in Fig. 2. This requires an even thinner IPD to ensure enough CR. However, current ONO stack starts failing in 10 ~ 12 nm range due to excessive leakage current [1]-[3]. Therefore, it is imperative to introduce high K dielectrics for the IPD layer, so as to scale down its EOT for good coupling ratio while maintaining low leakage current. Despite the urgency of this requirement, there are only a few research reports of high-K IPD for floating gate type Flash memory [3]-[6]. In this work, a systematic and comprehensive study on the suitability of using a variety of materials and structural combinations as the IPD is described.

2. Results and Discussion

Hf and Al-based high-K oxides have been extensively studied recently for use as thin gate dielectric. However, the application of high-K for IPD in Flash memory has a number of different technical issues because the thickness range and the operating voltage are much thicker and higher (6 ~ 10 times). It is known that increasing the high-K thickness degrades its thermal stability. Fig. 3(a) shows the thermal stability of a HfO_2 layer with an EOT of 8.8nm suitable for Flash memory IPD. The single layer HfO_2 shows a rapid degradation after high temperature annealing, which is attributed to poly-crystallization and film stress build-up. The use of Tb-doped HfO_2 was previously reported to result in a lower leakage current in RF MIM capacitors [7]. Since the previous work did not include a high temperature process, we have evaluated thermal stability of 4% Tb-doped HfO_2 with an EOT of 6.6 nm. The result shows that Tb-doped HfO_2 is a possible candidate for IPD application, exhibiting reduced leakage current and improved thermal stability as shown in Fig. 3(b). However, HfO_2 – Al_2O_3 – HfO_2 triple layer dielectrics with a similar total EOT exhibits more improved thermal stability after up to 950°C as shown in Fig. 4 and lower leakage current at high fields, compared to single layer high-K dielectrics after the high temperature annealing as shown in Fig. 5. Since multiple layer high-K structures are likely to be required, it is important to find out the best combination of materials and structures. For this purpose, tunneling current in a triple layer dielectric structure was simulated, starting with a fixed total physical thickness of 19 nm, while varying the ratio between the physical thicknesses of the blocking oxide layer and middle layer. Following the current ONO structure, we first evaluated the high-low-high barrier height combination. The simulation was done using MEDICI. The result in Fig. 6 shows that even though the Al_2O_3–HfO_2–Al_2O_3 structure shows a lower leakage current than SiO_2–HfO_2–SiO_2, the leakage current in both cases increases monotonically with decreasing EOT. The unchanging slope of the I *vs*. EOT plot indicates that the leakage current is governed by the tunneling through the blocking oxide layer only. Contrary to the belief that the blocking layer should have a higher barrier height, however, the HfO_2–Al_2O_3–HfO_2 stack, which has a low-high-low energy barrier structure, shows lower leakage current than a high-low-high barrier structure and, more importantly, has an optimum thickness ratio which can provide minimum leakage current for a fixed total physical thickness as shown in Fig. 7. Simulation of a structure with a fixed total EOT (6.5nm) also shows that low-high-low barrier structure has a lower leakage current and there is an optimum thickness ratio for minimum tunneling current as shown in Fig. 8. Simulated band diagrams under high voltage in Fig. 9 helps in understanding the reason for a higher leakage current in the high-low-high barrier structure. Direct experimental comparison in Fig. 10 also agrees with the trend of simulation result.

Based on such considerations, we examined the feasibility of material variation for each layer in low-high-low barrier high-K stack. Replacement of the middle Al_2O_3 layer with HfLaO (Hf:La = 50%:50%) gives an improvement in leakage current at high field as shown in Fig. 11. In HfLaO formation, higher % of La showed better performance. This is attributed to improved thermal stability for higher % of La in HfLaO film which is confirmed by TEM results in Fig. 12. Feasibility of AlLaO was examined, as well, but AlLaO always showed properties inferior to HfLaO as shown in Fig. 13. In addition, thick AlLaO often showed delamination problem after high temperature annealing due to poor adhesion and large film stress. Variation of blocking layer material was also investigated. While HfAlO does not improve the leakage current, 4%-Tb doped HfO_2 blocking layer reduces the leakage current quite effectively as shown in Fig. 14.

However, regardless of the different materials and structural combinations used, all the dielectric stacks still show leakage current much higher than simulation results. Further simulation reveals that very thin interfacial layer formed between high-K and polysilicon FG plays a key role in determining the leakage current. Even a 0.2 ~ 0.5 nm SiO_2 interfacial layer can increase leakage current dramatically at high voltage as shown in Fig. 15. Such low-K interfacial layer is possibly formed on polysilicon FG during ALD process or subsequent annealing. Under high voltage bias, as depicted in Fig. 16, a significant voltage drop happens on the low-K interfacial layer, leading to a great reduction in tunneling distance. In the gate dielectric case, the presence of low-K interfacial layer would not change total tunneling distance so significantly because of the low operating voltage and relatively thin total physical thickness, as shown in Fig. 17. This indicates that the control of the interfacial layer is the key factor to achieve the low leakage current for high-K IPD. Several ways to reduce the formation of interfacial layer have been evaluated. Figure 18 shows that heavy nitridation of polysilicon through 1 ATM NH_3 anneal or N_2 plasma is an effective way. Poly-SiGe shows even lower leakage current as shown in Fig. 19, because Ge inhibits formation of the oxide layer. Unlike polysilicon FG, leakage current of poly-SiGe FG drops with annealing temperature. Leakage current with poly-SiGe electrode is comparable with TaN electrode which indicates minimal formation of interfacial layer even after high temperature annealing as shown in Fig. 20.

3. Conclusion

It is shown that multi-layer high-K dielectric structures can be successfully used for IPD layer for next generation Flash memory device. It is also shown that the leakage current can be greatly suppressed by proper bandgap engineering, doping of Lanthanide element into high-K dielectrics and interfacial layer control.

References

[1] ITRS, ed. 2005

[2]S. Mori et al., IEEE TED, Vol 43, No. 1, p. 47, 1996

[3] Y. Yamaguchi, et al., Symp. on VLSI Tech, p. 85, 1993. [4] YY. Chen et al., IEEE conf. on Emerging Tech.-Nanoelectronics, p. 463, 2006.

[5] B.Govoreanu et al., Solid-State Electronics, vol. 49, p. 1841, 2005.

[6]M. Alessandri et al., 208th ECS Meeting, p 975, 2005.

[7] SJ Kim, et al, IEEE EDL, Vol. 24, p. 442, 2003.

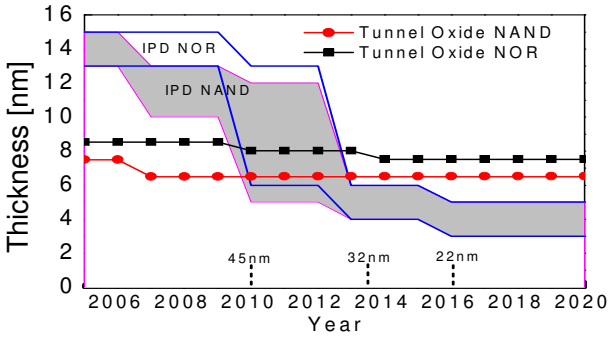


Fig. 1. ITRS 2005 prediction on electrical thickness of IPD and tunneling oxide for both NAND- and NOR-Flash memory devices.

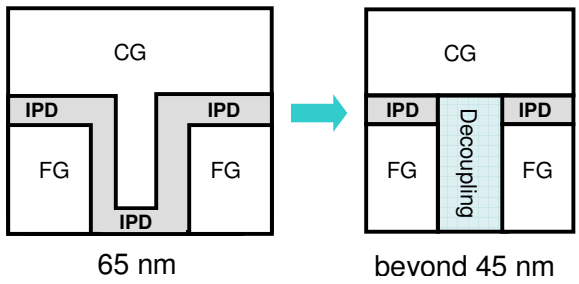


Fig. 2. Spacing between two adjacent gates stacks in Flash Memory is too close to have control gate overlapping the vertical side wall of FG beyond 45nm technology node, which leads to a significant reduction in coupling ratio. To compensate the capacitance loss, a dramatic drop in EOT is required.

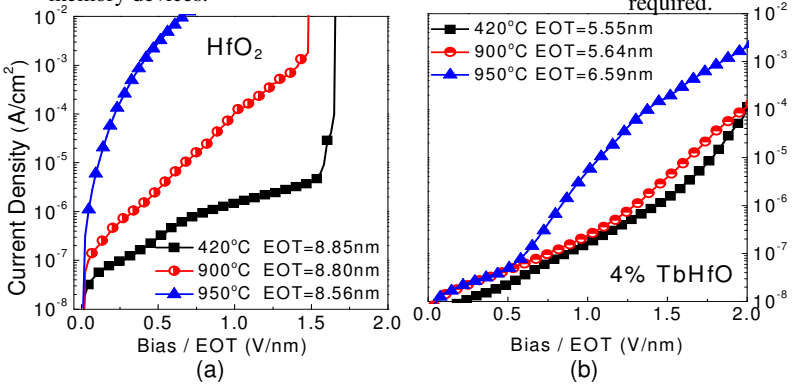


Fig. 3. Thermal stability of single layer (a) HfO_2 and (b) 4% Tb-doped HfO_2 IPD. 4% Tb-doped HfO_2 exhibits improved thermal stability up to 900°C and lower leakage current. Both FG and CG were TaN.

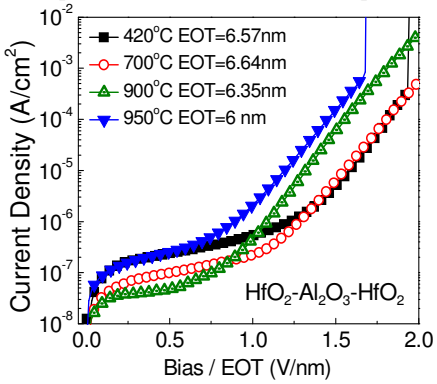


Fig. 4. Thermal stability of HfO_2-Al_2O_3-HfO_2 multilayer IPD. Multi-layer high-K stack with a similar total EOT shows better thermal stability. Both FG and CG were TaN.

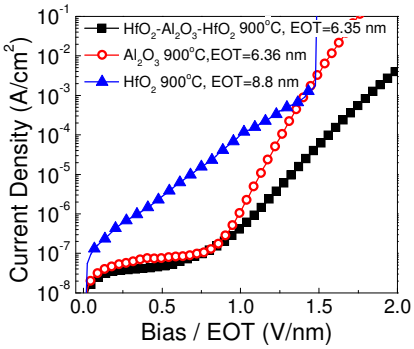


Fig. 5. Leakage current comparison between single layer Al_2O_3 , HfO_2 and multi-layer HfO_2 - Al_2O_3 - HfO_2 structure after 900°C annealing. Both FG and CG were TaN.

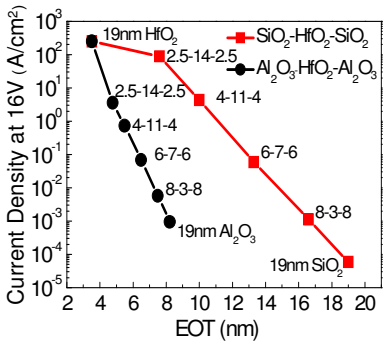


Fig. 6. Tunneling current simulation result for SiO_2-HfO_2-SiO_2 and Al_2O_3-HfO_2-Al_2O_3. Total physical thickness is fixed at 19 nm, and the middle layer thickness ratio is varied. Tunneling electrons are injected from polysilicon FG.

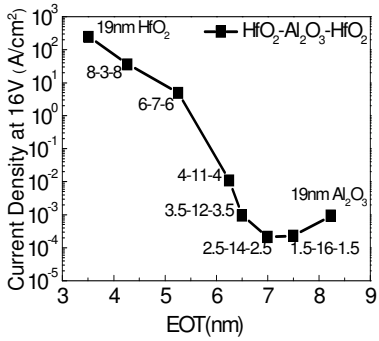


Fig. 7. Tunneling current simulation result for HfO_2-Al_2O_3-HfO_2. Total physical thickness is fixed at 19 nm, and the middle layer thickness ratio is varied. Tunneling electrons are injected from polysilicon FG.

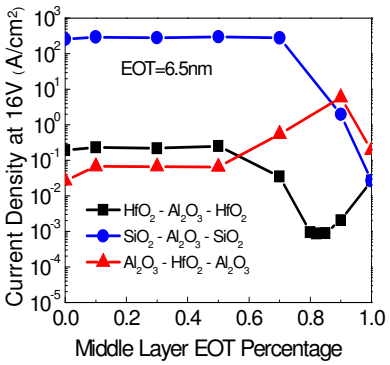


Fig. 8. Tunneling current simulation results for fixed EOT of 6.5nm. The middle layer EOT ratio is varied. Tunneling electrons are injected from polysilicon FG.

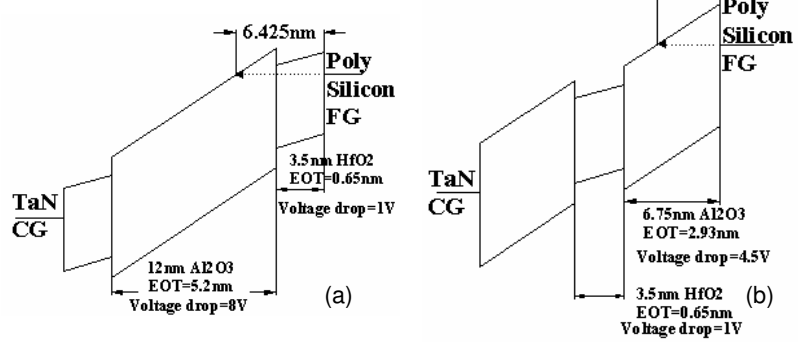


Fig. 9. Simulated Band diagram under +10V for (a) HfO_2 – Al_2O_3 – HfO_2 structure (b) Al_2O_3 – HfO_2 – Al_2O_3 structure. High-Low-High barrier structure has shorter tunneling distance than Low-High-Low barrier structure.

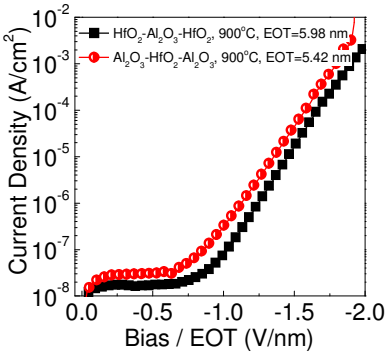


Fig. 10. Leakage current comparison between HfO_2 – Al_2O_3 – HfO_2 and Al_2O_3 – HfO_2– Al_2O_3 structure.

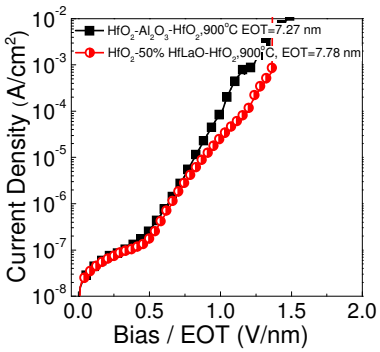


Fig. 11. Leakage current comparison by changing the middle Al_2O_3 layer with HfLaO. Electron injection from polysilicon FG.

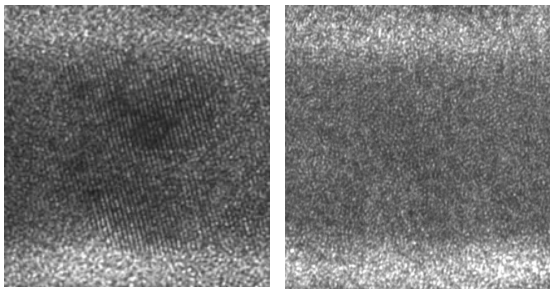


Fig. 12. TEM picture of HfLaO after 900°C 30s anneal for (a) 15% La in HfLaO (Hf:La = 85%:15%) and (b) 50% La in HfLaO. HfLaO (Hf:La = 50%:50%) remains amorphous even after after 900°C.

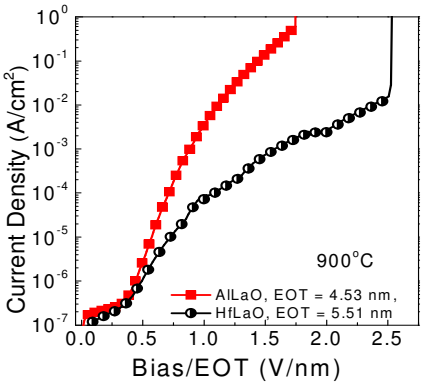


Fig. 13. Leakage current comparison between HfLaO and AlLaO single layer dielectrics

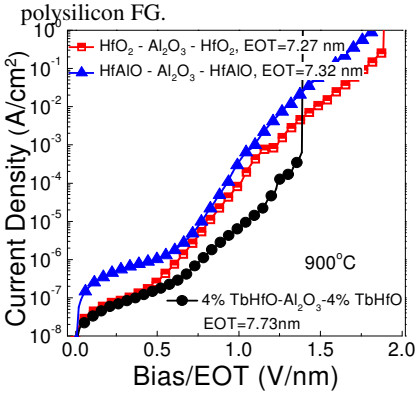


Fig. 14. Leakage current comparison by replacing HfO_2 blocking layer with HfAlO or 4% Tb-doped HfO_2.

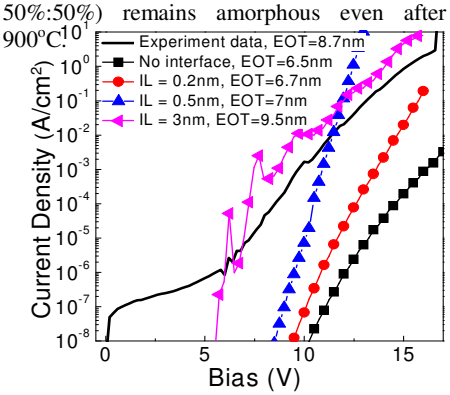


Fig. 15. Comparison between the simulation results and experimental data. Adding a thin SiO_2 interfacial layer increases the tunneling current significantly at high voltage even though the total EOT is increased.

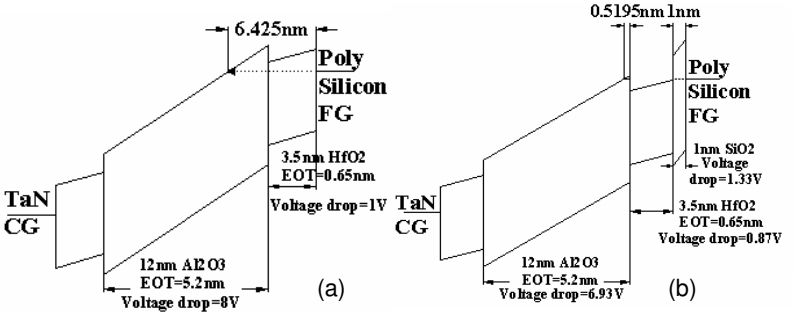


Fig. 16. Simulated band diagram (a) without interfacial layer and (b) with 1 nm SiO_2 interfacial layer between high-K IPD stack (HfO_2 – Al_2O_3 – HfO_2) and polysilicon FG. The band diagram explains the role of thin interfacial layer for the dramatic increase of tunneling current.

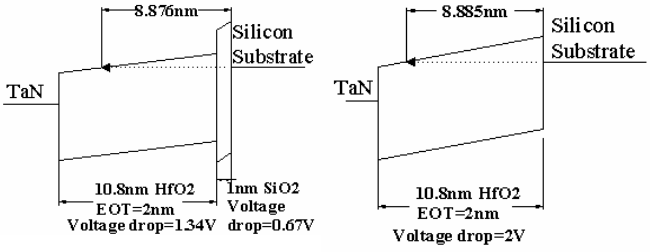


Fig. 17. Simulated band diagram of TaN/HfO_2 gate stack MOS capacitor under +2V bias (a) with 1 nm SiO_2 interfacial layer; (b) without interfacial layer. Due to thinner total thickness and low bias, the presence of thin interfacial layer does not change the tunneling distance much.

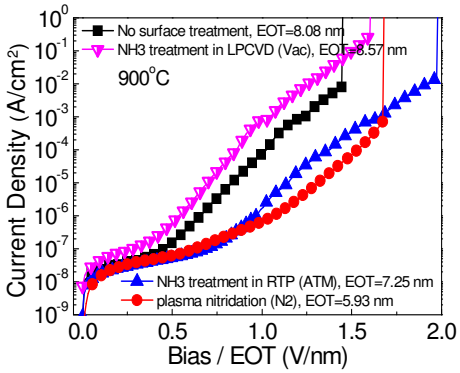


Fig. 18. Comparison of polysilicon FG surface treatment techniques. Heavy nitridation of polysilicon helps to reduce the leakage current by reduction of IL growth. HfO_2 – Al_2O_3 – HfO_2 IPD stacks are used.

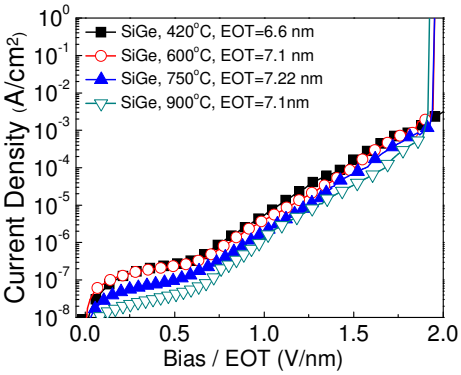


Fig. 19. Use of SiGe floating gate is effective in suppressing IL growth, leading to lower leakage current. HfO_2 – Al_2O_3 – HfO_2 IPD stacks are used.

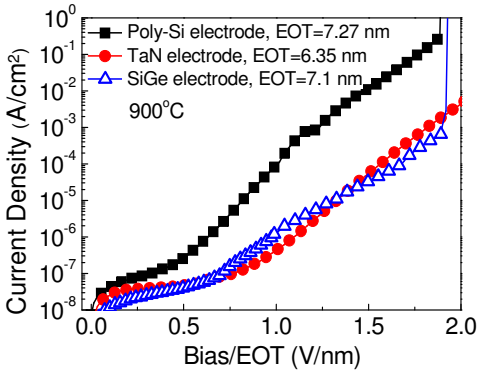


Fig. 20. Leakage current comparison on three different floating gates for HfO_2 – Al_2O_3 – HfO_2 IPD stack

Use of Al_2O_3 as Inter-Poly Dielectric in a Production proven 130nm embedded Flash Technology

R. Kakoschke[a], L. Pescini[a], J.R. Power[b], K. van der Zanden[c] , E.-O. Andersen[b], Y. Gong[b], R. Allinger[a]

[a] Infineon Technologies AG, Munich
Phone: +49-89-234-44309, Fax: +49-89-234-9555634, e-mail: ronald.kakoschke@infineon.com,
[b] Infineon Technologies Dresden GmbH & Co. OHG
[c] Infineon Technologies, affiliated at IMEC

Abstract

We have successfully integrated 2Mb arrays with SiO_2/Al_2O_3 stacks as inter-poly dielectric (IPD) fabricated in a proven 130nm eFlash technology. Gate stack write/erase high voltages (HV) can be reduced by 3V. Write/erase distributions show evidence of bit pinning which can be explained by barrier lowering along Al_2O_3 grain boundaries. Reliability assessment of the 2Mb array reveals promising data retention and cycle endurance measurements indicating no charge trapping in the high-k IPD. Despite several integration issues, these results demonstrate the high potential of Al_2O_3 IPDs in embedded Flash technologies.

1. Introduction

NAND Flash memory scaling has pulled ahead of DRAM and continues shrinking with a 3-year technology cycle [1]. Keeping this pace is challenging for embedded applications, since CMOS functionality cannot be compromised. Both cell array and HV circuitry have to be compatible with the platform CMOS technology. Additionally, manufacturing costs are of key importance. The overall goal is to minimize chip costs through reducing chip size and/or process complexity while still meeting product specifications.

In systems with embedded NVM, the memory array occupies only a certain fraction of the total die area. For this reason, the cell size shrink allows only a marginal increase in process complexity; otherwise the area benefit is lost. However, if the high write/erase voltages could be reduced, the area consuming HV peripheral Flash circuitry may also be shrunk along with the cell array. For this purpose, replacement of the traditional ONO IPD with a high-k dielectric material represents a promising approach for achieving a considerable shrink without impacting process complexity. Here, Al_2O_3 is a good candidate, being a well-known high-k material that has already reached the required level of maturity [2]. Recently, integration and data retention were demonstrated on single cells [3] and small arrays [4]. In this paper, we present the successful integration of Al_2O_3 as an IPD material into a qualified embedded Flash process using a 2Mb memory array as demonstrator.

2. Device Description

Based on our conventional 130nm eFlash technology, we have integrated 2Mb arrays of 1-transistor UCP flash cells with high-k stacks replacing the conventional ONO IPD. After active area and HV well formation the tunnel oxide (8.5nm) is grown and floating gate poly deposited. After poly pre-structuring the IPD is deposited and removed in selected areas. Then the gate oxide is formed and control gate poly (equivalent to transistor gate poly) deposited, followed by stack gate etch and side wall oxidation. After extension and S/D implants the process is completed by 4 layers Cu-metallization.

Al_2O_3 was deposited by ALD using trimethylaluminum (TMA) and ozone precursors. Post deposition annealing (PDA) was done at 1000°C for 20s. We realized various high-k IPD stacks with SiO_2 bottom oxides below Al_2O_3 to provide sufficient data retention [3]. The chosen bottom oxide thickness ranged from 1nm to 5.5nm, while the Al_2O_3 thickness was adjusted to achieve the target equivalent oxide thickness (EOT) from 6nm to 8.5nm. The EOT of the ONO reference was 16nm.

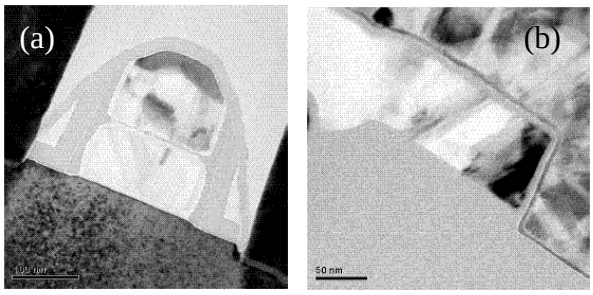


Fig. 1: TEM images of 1T UCP cell with SiO_2/Al_2O_3 as IPD along bit line (a) and word line (b).

The unit processes for a high-k dedicated flow have been developed. Fig. 1a,b show TEM images after final processing. A bird's beak is evident at the control gate/ floating gate edge (Fig. 1a). The deposited Al_2O_3 layer is conform around the floating gate but not smooth (Fig. 1b). Although the potential of the high-k IPD can be investigated with such a gate stack, further optimisation is required.

3. Electrical Results and Discussion

Figs. 2 and 3 show V_{th} distributions obtained with two different IPD stacks: SiO_2/Al_2O_3 3nm/6.5nm and 4nm/10.4nm. These distributions have been selected as representative for all measured distributions (not shown here) and reveal the potentials and issues of Al_2O_3 IPDs. Fig. 2 shows the V_{th} distributions with SiO_2/Al_2O_3 3nm/6.5nm for write (a) and for erase (b). A large fraction of pinned bits in both the written and erased state is observed. Fig. 3 shows the evolution of the V_{th}

distributions with increasing write (a) and erase (b) voltages. Compared to Fig. 2, a clear improvement is observed by increasing the thicknesses to 4nm/10.4nm. The distributions become more symmetric. Although tail bits are still present, their number is reduced to the order of ppm for write (Fig. 3a). For erase, the tail bits are uncovered at higher erase voltages (Fig. 3b). They are more numerous (some 0.01%) than for write. Bitmaps show that these bits are randomly distributed throughout the array.

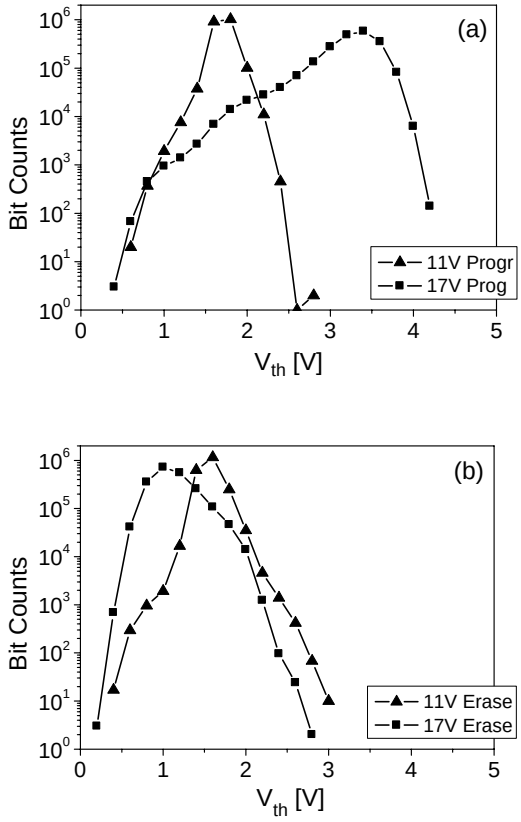


Fig. 2: V_{th} distributions of 2Mb array after write (a) and erase (b) with 11V and 17V total stack voltage. IPD: SiO_2 /Al_2O_3 3nm/6.5nm. Pinning is observed in both program and erase distributions.

By plotting the distribution maximum versus applied total stack voltage, the funnel curves shown in Fig. 4 are realized. Write/erase times were 1ms. Due to improved coupling ratio, the peak programming voltages are reduced by 2.5V (Fig. 4b) and 3.0V (Fig. 4c) with respect to the reference group (Fig. 4a). This is achieved although part of the coupling ratio gain is lost due to the bird's beak at the control gate-IPD-floating gate edge (Fig. 1a). At the onset of write/erase, the slope of both write and erase branches is 1 (1V increase in V_{th} for 1V increase in stack voltage). However, Fig. 4c shows early write and erase saturation which is also visible in the write branch of Fig. 4b. Due to the thin IPD, electrons tunneling to (from) the floating gate are compensated by current through the IPD. The asymmetry for write and erase is attributed to the Variot effect [5] within the IPD. For the same voltage drop in either direction, the tunneling barrier is larger for erase (reverse-Variot case) than for write (Variot case).

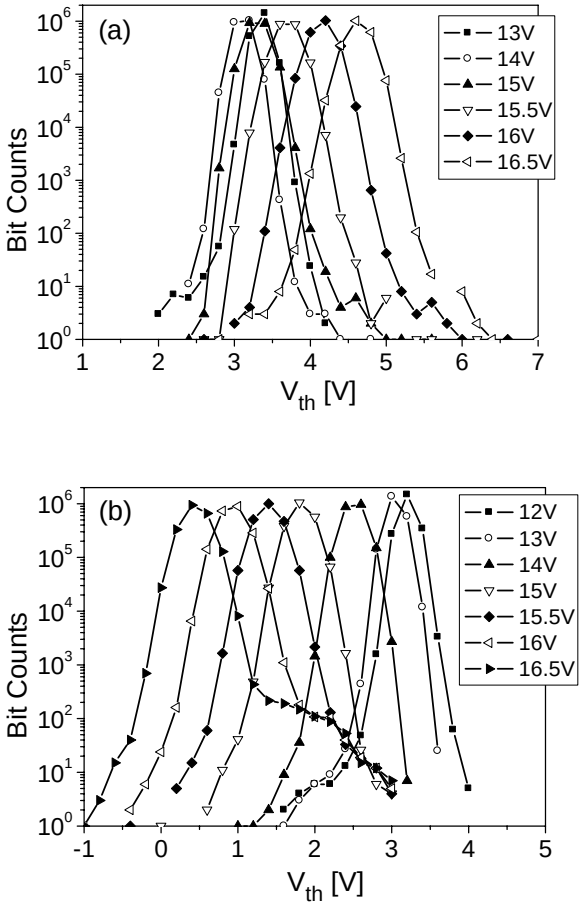


Fig. 3a,b: V_{th} distributions of 2Mb array, 12-16.5V; (a) write, (b) erase. IPD= SiO_2 /Al_2O_3 4 nm/10.4 nm. Some pinning is observed in the erase distributions.

In order to explain the tail bits in the erase distribution (see Fig. 3b), we assume that defects induce low barrier current paths from the control gate through the Al_2O_3 to the bottom oxide interface. When ramping the erase voltage, Fowler-Nordheim tunnelling from the floating gate to the channel sets in. With further ramping, the electric field across the IPD increases and some current paths through the Al_2O_3 layer open. Consequently, the electric field across the IPD bottom oxide increases and finally a leakage current through the complete IPD occurs. A current equilibrium is achieved, i.e., electrons tunnelling through the tunnel oxide out of the floating gate are replaced by electrons from the control gate. A further V_{th} shift will therefore not occur, resulting in effectively pinned bits. The distribution of barrier heights for defect induced current paths determines the distribution of pinned bits. Concluding from Fig. 3b, low barriers are rare, higher barriers are more frequent. However, the probability for a leakage path depends not only on the Al_2O_3 thickness, but also on the potential drop across the total IPD, that is, the bottom oxide and Al_2O_3 layer together. We note that a single leakage path somewhere in the IPD is sufficient to cause pinning.

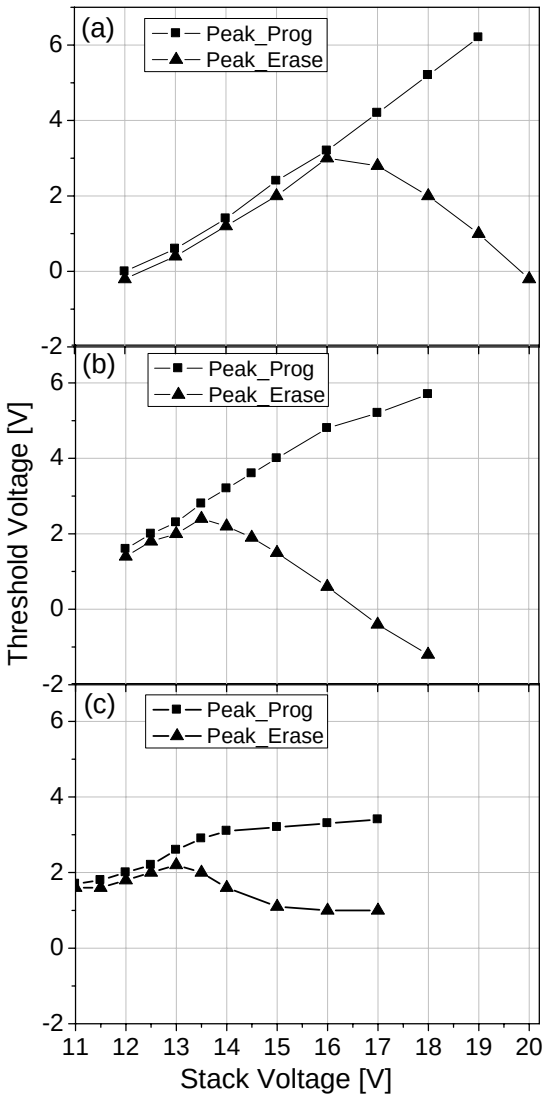


Fig. 4: Write/erase characteristics (a) ONO reference, (b) IPD: SiO_2/Al_2O_3 1nm/11.5nm and (c) IPD: SiO_2/Al_2O_3 3nm/6.5nm. Write/erase threshold voltages are measured with $V_{d,read}$ = 1.2V, I_d = 1µA. Write/erase time is 1ms. The x axis represents the total stack voltage $|V_G - V_{SD}|$ required for write and erase.

Alternatively, the pinning effect could be attributed to charge trapping. In this case, many trapped charges would be required to explain the large V_{th} differences. However, as can be seen in Fig. 5, there is no indication that trapping in the IPD contributes to a V_{th} shift, even after 1E5 write/erase cycles. Only a slight shift of V_{th} is observed, which is also present in the ONO reference group. It is attributed to charge trapping in the SiO_2 tunnel oxide.

In Fig 6a the data retention results of SiO_2/Al_2O_3 1nm/11.5nm IPD cells are shown. Charge loss occurs already at room temperature. For SiO_2/Al_2O_3 4nm/10.4nm, however, no charge loss can be observed up to 12 day at room temperature (Fig. 6b). It is important to note that the pinned bits appearing in the erase distribution also show good retention although these are assumed to be defect related. For erase or write, the potential difference between control gate and floating gate is higher than during storage, i.e., leakage channels can be activated during erase but remain inactive during storage, fitting with the barrier leakage model discussed above. Also we note in general that only fairly narrow distributions show good data retention.

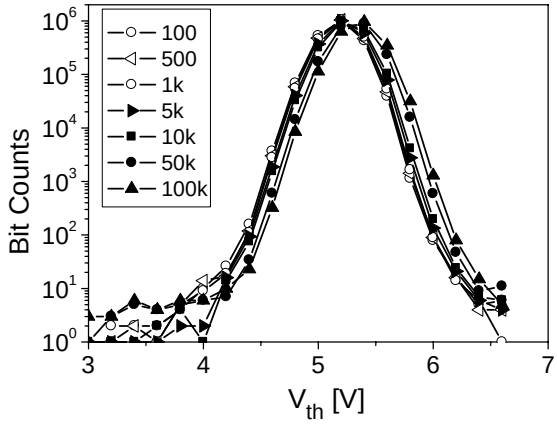


Fig. 5: 2Mb array endurance with up to 100k cycles for cells with SiO_2/Al_2O_3 4nm/10.4nm IPD stack. The endurance is good, with only 0.2V shift in V_{th} after 100k cycles. Such a shift appears also in the ONO reference and is attributed to trapping in the tunnel oxide. No indication of charge trapping within the IPD is observed.

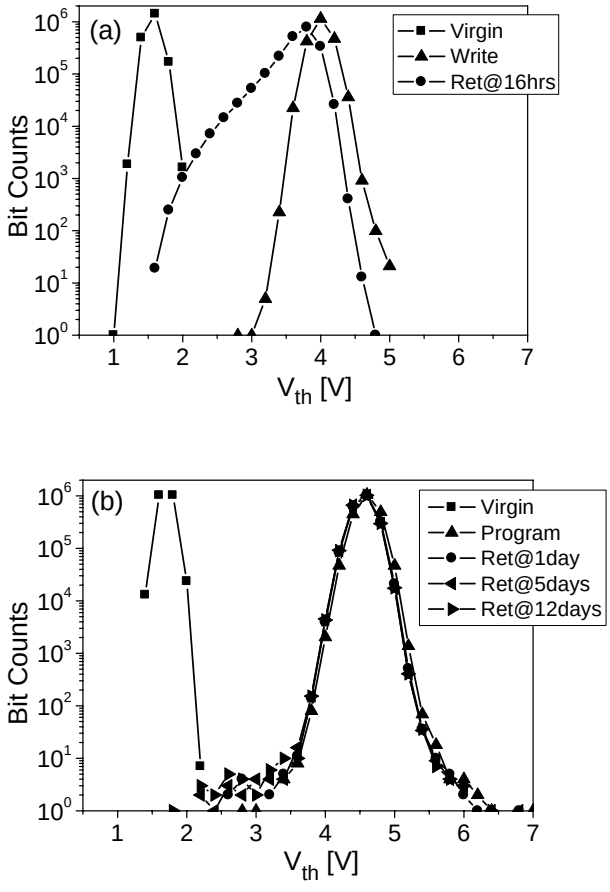


Fig. 6: 2Mb array data retention at room temperature. (a) IPD: SiO_2/Al_2O_3 1nm/11.5nm. Charge loss is observed already after 16 hrs. (b) IPD: Al_2O_3/SiO_2 4nm/10.4nm. Negligible charge loss is measured after 12 days.

We observe strong variations in the V_{th} distributions and data retention behaviour from chip to chip across the wafer. We attribute such variations to the fact that processing steps around the gate stack complex have not yet been optimized. Wet cleans, side wall oxide formation and PDA need to be trimmed to accommodate the properties of the high-k material. For example, the HV reduction was compromised by a loss of coupling ratio due to the formation of a bird's beak at the control gate/floating gate edge arising from too aggressive etch processing of the Al_2O_3 IPD after stack formation. In some cases, this undesired effect can even completely compensate the whole coupling ratio gain obtained by the Al_2O_3 IPD. The smaller coupling ratio, in turn, requires use of a higher write/erase potential, which favours pinning. By optimising such processing we expect to achieve higher coupling ratios together with reduced pinning.

4. Conclusion

We have integrated 2Mb arrays of 1-transistor UCP flash cells with high-k stacks replacing the conventional ONO IPD in our 130nm eFlash technology with a high-k IPD. In comparison with the ONO IPD reference, the write/erase voltage can be reduced by approximately 3V. A careful design of IPD layer thicknesses is necessary to avoid early write/erase saturation, minimize tail bits and achieve good data retention. For the SiO_2/Al_2O_3 4nm/10.4nm devices no indication of early saturation was found and very good data retention could be demonstrated after 12 days at room temperature. The significant bird's beak observed at the stack edge and presence of tail bits in the measured distributions indicate that several integration issues must be solved before the true potential of the high-k IPD can be realized. Although the potential of the high-k IPD could be demonstrated, the full benefit may only be realised through further optimizing and fine tuning unit processes and process flow.

Acknowledgements

The authors would like to thank especially H. Bernhardt and Th. Hecht (Qimonda AG) for aluminum oxide deposition and valuable inputs. Many thanks are also extended to the embedded flash technology development group, R. Strenz, A. Gratz, W. Langheinrich, M. Röhrich and G. Tempel for fruitful discussions and for providing decisive assistance and E. Sommer for process-flow logistics.

References

[1] ITRS 2006 update (http://www.itrs.net/Links/2006Update/FinalToPost/00_ExecSum2006Update.pdf, p4-5)

[2] W. Mueller, et al, "Challenges for the DRAM Cell Scaling to 40nm", in IEDM Tech. Dig. 2005, pp. 347 - 350.

[3] Dirk Wellekens, et al, "Al_2O_3 based Flash Interpoly Dielectrics: a comparative retention study", Proc. ESSDERC 2006

[4] Almudena Huerta Miranda, et al, "Reliability Comparison of Al_2O_3 and HfSiON for use as Interpoly Dielectric in Flash Arrays", Proc. ESSDERC 2006, pp. 234 - 237

[5] B. Govoreanu, D. Brunco, J. Van Houdt, "Scaling Down the Interpoly Dielectric for Next Generation Flash Memory: Challenges and Opportunities", ICMTD 2005, pp. 211 - 214

Investigation of aggressively scaled $HfAlO_x$-based interpoly dielectric stacks for sub-45 nm nonvolatile memory technologies

B. Govoreanu, D. Wellekens, L. Haspeslagh, D.P. Brunco[a], J. De Vos, D. Ruiz Aguado[*], P. Blomme, K. van der Zanden[b], J. Van Houdt

IMEC Leuven, RDO/PT Division, Kapeldreef 75, B-3001 Leuven, Belgium
[a]Intel Corp. Assignee, Kapeldreef 75, B-3001 Leuven, Belgium
[b]Infineon Technologies Assignee, Kapeldreef 75, B-3001 Leuven, Belgium
[*]*also with* K.U. Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

Contact: bogdan.govoreanu@imec.be, Tel. +32-16-281337

Abstract

This paper discusses the performance and reliability of aggressively scaled $HfAlO_x$-based interpoly dielectric stacks in combination with high-workfunction metal gates for sub-45 nm non-volatile memory technologies. It is shown that a less than 5 nm EOT IPD stack can provide a large window, while operating at moderate program/erase voltages and has excellent retention, with an extrapolated 10-year retention window of about 3 V at 150 C. The impact of the process sequence and metal gate material is discussed as well, suggesting directions for further improvement.

1. Introduction

Introduction of the high-k dielectrics in non-volatile memory (NVM) technology is a must. Recent works report on the successful integration of high-k materials as interpoly dielectrics (IPD's), using HfSiON [1] or dual-layer SiO_2/Al_2O_3 stacks [2], with electrical thicknesses (EOT's) in the range of 5-10 nm, targeting embedded memory applications.

The need for dense non-volatile memory arrays requires cell planarization for sub-45 nm technology nodes. Such an architectural change calls for a dramatic reduction of the electrical thickness of the IPD to below 5 nm. This will eventually compensate for the loss of the sidewall coupling capacitance and restore the coupling factor to an acceptable value of around 0.6. It is suggested that this target is achievable by an integral approach, combining high-k IPD's with high-workfunction (high-W_m) metal gates [3].

In this work, we discuss the performance of $HfAlO_x$-based IPD's with various control gates. It is shown that the material gives a very large program/erase (P/E) window, while keeping the operating voltages at acceptable levels. Furthermore, it shows very good room temperature retention, with minor window closure, even after ~10^7 s storage time. The impact of the metal gate material and deposition process is discussed, suggesting points of attention for further improvement.

2. Test structures and splits

Stacked gate capacitors with tunnel oxide and high-k IPD's have been fabricated in a simplified process. This approach allows for a quick screening of the "intrinsic" material performance, while minimizing the possible impact of the processing steps following the high-k deposition, which could affect the material performance.

2.1. Process flow and test structures

The process flow is summarized in Fig. 1. A thick TEOS layer is deposited on n-type Si wafers. After a gate window etch, a wet tunnel oxide of 8.5 nm is grown. The floating gate (FG) is formed by an in-situ Phosphorous-doped polysilicon deposition, followed by a chemical-mechanical polishing (CMP) step, resulting in a planar poly top-surface lining up with the TEOS layer. The IPD stack consists of either a thin, almost 1 nm SiO_2-like layer, formed after a specific IMEC-clean [4], or an HTO layer and the $HfAlO_x$ layer. The high-k layer has been formed by atomic layer deposition (ALCVD) in a Polygon 8200 cluster, by depositing Al_2O_3 and HfO_2 in a 1:1 cycle ratio. The control gate (CG) is finally formed by depositing either an n^+-type poly-Si layer or a metal gate, followed by an etch stopping in the thick TEOS layer. Processing is completed with a sintering step.

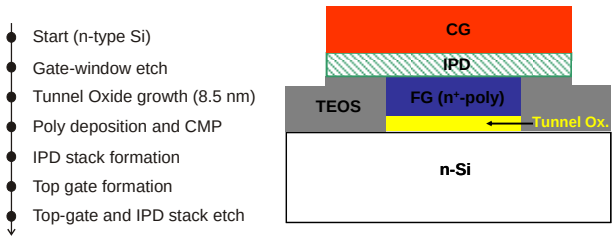


Fig. 1: Simplified sequence of the process flow (left) and schematic drawing of the stacked gate test structures (right).

2.2. Process splits

The $HfAlO_x$ is subjected to a post-deposition anneal (PDA) at 800 °C, for 60 s, in an N_2 ambient. At this temperature, the material remains amorphous, consequence of the increased thermal stability [5] induced by the mixing the corresponding binary oxides.

The top gate is deposited either as TiN, TaN or n^+ poly-Si. The TiN is deposited either by ALD or by

sputtering from ionized metal plasma (IMP) [6], a process that is shown to produce a more uniform TiN, with a better interface to the underlying dielectric. The TaN is deposited by physical vapour deposition (PVD), while the poly gate formation is completed with a two-step silicidation process.

Since the metal gate deposition is much colder compared to its poly counterpart, a degas process may be performed just before the metal gate deposition, in order to desorb water residues in the high-k. Presence of water molecules or OH^- radicals in the high-k is a potential source of defects in the material and may affect the high-k/gate interface, as well. Splits with different degas temperatures (in the range of 350 °C to 500 °C) and times were considered and compared against a reference split without degas.

3. Results and Discussion

3.1. Performance

Large area capacitor structures were programmed (erased) by applying a positive (negative) CG pulse, which allows electrons tunnelling from the accumulated n-Si substrate (n^+-poly FG) through the tunnel oxide. The flatband voltage (V_{FB}) shift in high-frequency capacitance-voltage (HFCV) curves was used as a monitor for the V_{FB} window closure. Raw CV data (Fig. 2) taken for a split with an IPD consisting of 1 nm SiO_2 / 12 nm $HfAlO_x$ and a TiN gate demonstrate a large P/E window of up to 7 V. The programming times are in the range of 1-10 ms, while a deep erase can be carried out in 10-100 ms. The excellent immunity to erase saturation is given by the midgap to p-type character of the metal gate, which determines a large barrier height at the CG/high-k interface. It is also observed that the FB voltage shift does not lead to any distortion in the CV curves and only shows little hysteresis, even for the highest P/E voltages.

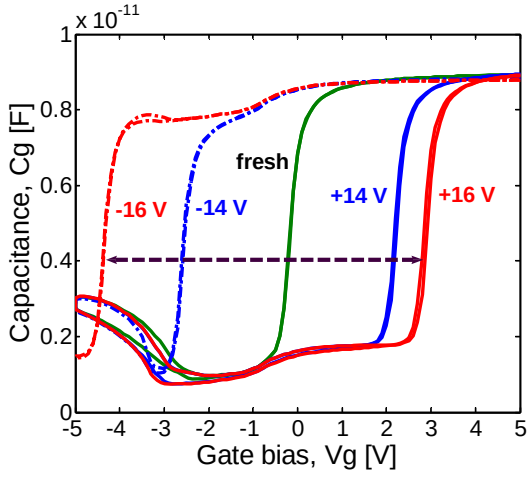


Fig. 2: CV curves for a fresh device and after P/E show large V_{FB} window, of up to 7 V, for P/E pulses of +/-16 V, 10 ms long. The small humps observed on all traces (either at the top of bottom of the CV curves), are due to the parasitic bondpad capacitance.

3.2. Process impact: the short time scale behaviour

Double-sweep HFCV curves were taken before and right after applying a positive programming pulse to several splits. The splits reported here are summarized in Table 1. The hysteresis is measured at the FB voltage and its cumulative distribution (CD) shows rather small values, of less than ~5 mV, for fresh devices (Fig. 3). It indicates, however, some charge trapping in the IPD, due to the disturbance produced during the CV measurement.

Table 1: Summary of the discussed splits. $HfAlO_x$ deposition is always followed by a PDA at 800 °C, for 1 min in N_2. 1 nm SiO_2 resulted after an IMEC clean, while 3 nm SiO_2 is HTO.

Split	IPD Stack (SiO_2/ $HfAlO_x$)	Degas	Top Gate
(S1) - ○	3 nm / 7 nm	n/a	n^+-poly
(S2) - □	1 nm / 12 nm	180 s, 500 C	IMP TiN
(S3) - ◊	1 nm / 12 nm	180 s, 350 C	PVD TaN
(S4) - Δ	1 nm / 12 nm	None	ALD TiN

After a programming pulse is applied, the hysteresis increases to up to 20 mV, suggesting that somewhat more charge is trapped in the IPD during the programming pulse. It is however noticed that the hysteresis of the splits subjected to a degas step prior to the CG formation remains lower (up to ~10 mV) compared to a non-degassed split. Within the time resolution of the HFCV measurement, this shows the beneficial effect of the degas step, which reduced the density of the shallow defects present in the high-k dielectric. For comparison, an n^+-poly CG split is also shown.

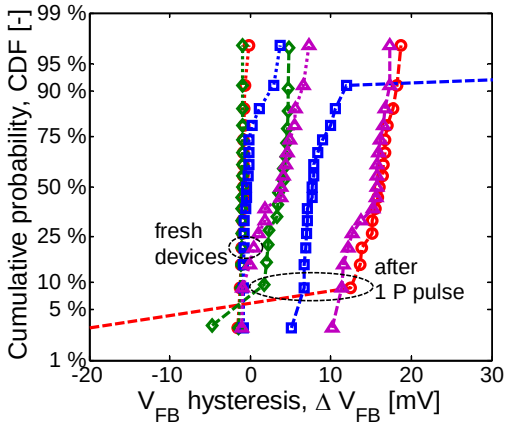


Fig. 3: Cumulative distribution of the FB voltage hysteresis of fresh and once programmed devices, for different IPD splits. All splits with metal gate have an IPD consisting of 1 nm SiO_2 and 12 nm $HfAlO_x$. The FB voltage is taken as the gate voltage corresponding to an arbitrary set reference capacitance level. The considered splits are summarized in Table 1. Corresponding symbols are as summarized in Table 1: (S1) – ○; (S2) – □; (S3) – ◊; (S4) – Δ. The dotted lines are for fresh samples, while dashed lines are for once-programmed samples.

3.3. Room-temperature retention

Room-temperature (RT) retention monitoring (Fig. 4) showed a very robust material, with well-behaved CD,

where a -4.5 V shift of the median FB voltage of erased cells persisted for more than 5.10^6 s (~2 months).

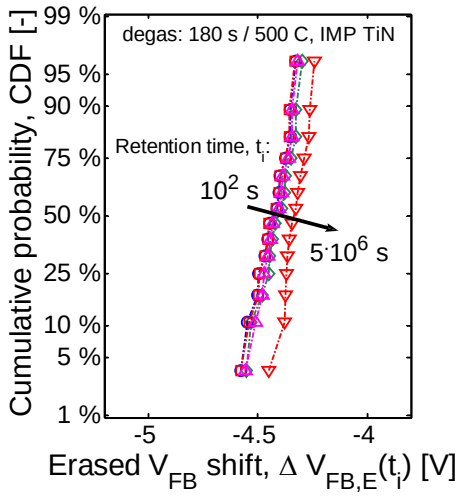


Fig. 4: CD of FB voltage shift for erased devices with an IPD of 1 nm SiO_2 / 12 nm $HfAlO_x$. The samples were kept at room temperature storage conditions. The Erased V_{FB} shift after a retention time t_i is defined as: $\Delta V_{FB,E}(t_i) = V_{FB,E}(t_i) - V_{FB}^{(fresh)}$.

The average retention curves for both P/E states (Fig. 5) show the following: in the high-V_{FB} state, there is a small (up to 0.2 V) initial window closure, evidenced more in the split with poly gate. This is believed to be due to detrapping from the high-k dielectric and correlates with larger hysteresis observed for the poly split (Fig. 3). The initial window closure is not observed on the low-V_{FB} state, which suggests again presence of relatively shallow traps in the high-k dielectric.

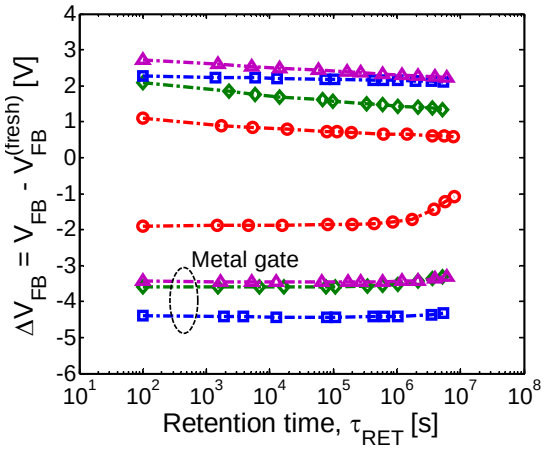


Fig. 5: Comparative retention for 4 different $HfAlO_x$ IPD splits, as summarized in Table 1. Corresponding symbols are as in Table 1: (S1) – ○; (S2) – □; (S3) – ◊; (S4) – Δ.

The retention behaviour of the considered splits is explained as follows: during the programming operation, the parasitic leakage through the IPD stack favours the filling of the shallow traps, due to injection from the poly FG coupled with a parasitic VARIOT effect [7], which also limits the achievable high-V_{FB} level. On the contrary, during erasing, the shallow traps are less accessible for electrons arriving from the top gate, due to the "misalignment" between their energy levels and the Fermi level in the CG. This energy level difference increases from n^+-poly gates towards p-type metal gates, as illustrated in Fig. 6. The split with an IMP TiN gate shows the best RT retention. The TaN-gate split, although subject to degas, has somewhat more window closure, which may be attributed to the difference in the effective workfunction. This is lower for TaN (closer to midgap), compared to TiN, hence reducing the misalignment to trap levels in the high-k. This explanation is also consistent with the larger P/E window obtained for the TiN gate (Fig. 5). The ALD TiN split also shows a larger decay of the high-V_{FB} state, attributed to a worse dielectric quality, when no degas step is performed. More analysis is ongoing in order to clarify and validate this qualitative understanding.

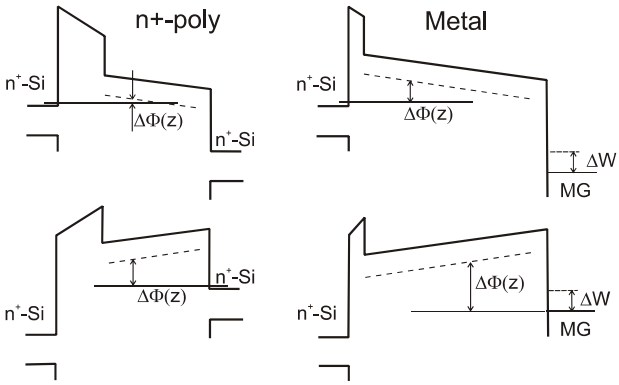


Fig. 6: Schematic representation of the band diagrams for cases corresponding to n^+-poly splits (S1) – left compared to high-W_m metal gates (S2)-(S4) – right. The "mismatch" between the depth of a shallow trap and the average injection level is lower for (S1) compared to (S2)-(S4), causing the leakage to be higher. ΔW denotes the difference between the Fermi levels in the n^+-poly and the p-type/midgap metal gate.

The window closure results (Fig. 7) summarize the comparison of the RT retention of the considered splits. The 1 nm SiO_2 / 12 nm $HfAlO_x$ / IMP TiN split had less than 4 % closure of the V_{FB}-window after ~2 months of RT storage, with a remaining average V_{FB}-window of 6.45 V.

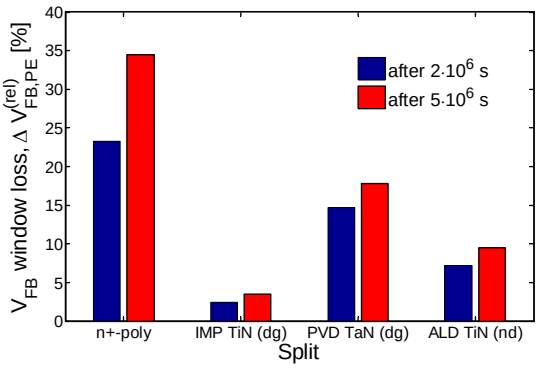


Fig. 7: Relative V_{FB}-window loss for the splits shown in Fig. 6, after 2.10^6 s (~3 weeks) and 5.10^6 s (~2 months) storage at room temperature.

3.4. The high-temperature retention

High-temperature retention tests up to 250 C were carried out, in order to accelerate the FG charge loss/gain. It was found that temperature acceleration for $HfAlO_x$-based IPD's is very strong for all considered splits, as opposite to the case of Al_2O_3-based IPD's, which showed rather weak temperature acceleration. However, due to the large initial P/E window, after a bake test of ~5.10^6 s at 250 °C, a V_{FB}-window of almost 2 V is still observed, while a 150 °C retention test suggests a 10-year extrapolated window of slightly more than 3 V (Fig. 8). This result propels the combination $HfAlO_x$/high-W_m MG as one of the most promising candidates for a high-k based IPD in sub-45 nm FG Flash technology.

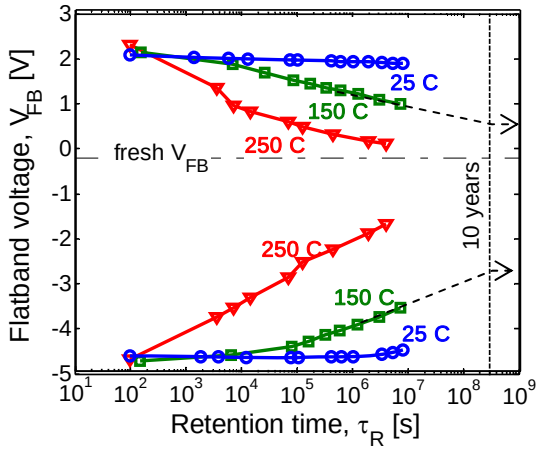


Fig. 8: Room and high-temperature retention corresponding to the $HfAlO_x$ IPD with IMP TiN gate (split (S2) in Table 1).

The strong temperature acceleration suggests that the dominant leakage mechanism is due to rather shallow traps, which are shown to determine a low-field conduction significantly increasing with temperature [8]. In order to characterize the temperature acceleration of the charge loss/gain, the time corresponding to 1 V window shift with respect to the initial V_{FB} value (right after Program or Erase) has been extracted. The extraction assumes either measured (for high-temperature) or extrapolated (RT) data. Assuming an Arrhenius law holds, an activation energy of 1.27 eV (Erased state) or 1.26 eV (Programmed state) has been extracted.

An inverse-temperature law has been also used in order to get the maximum temperature for a remaining window of 3 V after 10 years of retention. The degassed split with IMP TiN gate shows a maximum temperature of ~162 °C. Although an extrapolation, the confidence in this result is also supported by the data in Fig. 8, where after a retention test at 150 °C for a time of about $7{\cdot}10^6$ s (~ 3 months), a V_{FB} window of more than 4 V is still measured.

4. Conclusion

In summary, we have shown that IPD's based on $HfAlO_x$ combined with high-workfunction top-metal gate appear as a promising combination for targeting 5 nm EOT and below. Large V_{FB} window of more than 6 V is achievable, using P/E voltages that do not exceed ±16 V. RT retention shows only little window closure, while accelerated retention tests project a larger than 3 V window at 150 °C. These results might as well be exploited towards multilevel cell NAND Flash. More improvement can be achieved by further process tuning, in order to improve the high-k material quality and to adjust the relevant properties of the high-k/CG interface.

Acknowledgments – This work has been carried out within the framework of the IMEC Industrial Affiliation Program (IIAP) on Flash Memory, partnered by Infineon Technologies Intel Corporation, Micron and Samsung.

References

[1] M. van Duuren, R. van Schaijk, M. Slotboom, P. Tello, P. Goarin, N. Akil, F. Neuilly, Z. Rittersma and A. Huerta: *Performance and Reliability of 2-Transistor FN/FN Flash Arrays with Hafnium Based High-K Inter-Poly Dielectrics for Embedded NVM*, in Proc. NVSM Workshop, pp. 48-49, Monterey (CA), USA, 2006.

[2] D. Wellekens, P. Blomme, B. Govoreanu, J. De Vos, L. Haspeslagh, J. Van Houdt, D.P. Brunco, K. van der Zanden: *Al_2O_3-based Flash interpoly dielectrics: a comparative retention study*, in Proc. ESSDERC 2006, Montreux (Switzerland), 2006.

[3] B. Govoreanu, D.P. Brunco, J. Van Houdt: *Scaling down the interpoly dielectrics for nex generation Flash memory: challenges and opportunities*, Solid-St. Electronics, **49**(11): 1841-1847, 2005.

[4] M.M. Heyns, T. Bearda, I. Cornelissen, S. De Gendt, R. Degraeve, G. Groeseneken, C. Kenens, D.M. Knotter, L.M. Loewenstein, P.W. Mertens, S. Mertens, M. Meuris, T. Nigam, M. Schaekers, I. Teerlinck, W. Vandervorst, R. Vos, and K. Wolke: *Cost-effective cleaning and high-quality thin-gate oxides*, IBM J. Research and Development **43**(3): 339-350, 1999.

[5] H.Y. Yu, N. Wu, M.F. Li, C. Zhu and B.J. Cho: *Thermal stability of $(HfO_2)_x(Al_2O_3)_{1-x}$ on Si*, Appl. Phys. Lett. **81**(19): 3618-3620, 2002.

[6] Y. Tanaka, E. Kim, J. Forster, Z. Xu: *Properties of titanium nitride film deposited by ionized metal plasma source*, J. Vac. Sci. Tech. **B 17**(2): 416-422, 1999.

[7] B. Govoreanu, P. Blomme, M. Rosmeulen, J. Van Houdt, K. De Meyer: *VARIOT: A novel multilayer tunnel barrier concept for low-voltage nonvolatile memory devices*, IEEE El. Dev. Lett., **24**(2): 99-101, 2003.

[8] B. Govoreanu, D. Wellekens, L. Haspeslagh, J. De Vos, J. Van Houdt: Investigation of *Investigation of the low-field leakage through high-k interpoly dielectric stacks and its impact on nonvolatile memory data retention*, in IEDM Tech. Dig, pp. 479-482, San Francisco, USA, 2006.

High-Quality Aluminum-Oxide Tunnel Barriers for Scalable, Floating-Gate Random-Access Memories (FGRAM)

Xueqing Liu, Vijay Patel, Zhongkui Tan, Konstantin K. Likharev, and James E. Lukens

Stony Brook University, Stony Brook, NY 11794-3800, U.S.A.
E-mail: *klikharev@notes.cc.sunysb.edu*

Abstract

We have demonstrated all-metallic tunnel junctions based on rf-plasma-grown aluminum oxide layers, which enable scalable, floating-gate memory cells with 20-ns-scale write time, 1-s-scale retention time, low operating voltage (3.0-3.5 V), and high endurance in high electric fields (up to 10^{11} write cycles). We believe that such memories may be suitable for some (and after some improvement, most) RAM applications.

1. Introduction

In the course of the continuing search for the "perfect" (scalable, non-volatile, random-access) memory, our group had suggested [1, 2] the concept of NOVORAM – a floating-gate memory based on quantum-mechanical tunneling of electrons through specially crafted layered barriers. According to calculations, at the optimum choice of the potential barrier heights (conduction band offsets) of the layers and their dielectric constants κ, the transparency of such "crested" barriers can be changed by more than 16 orders of magnitude by merely doubling the voltage applied to the barrier, i.e. much faster than barriers made of any known uniform insulator.[1] Such high sensitivity would enable a fast and scalable floating-gate RAM with the cell structure shown in Fig. 1 [1]. Its main difference from the usual non-volatile memories is that in order to suppress the barrier deterioration by hot carriers from the MOSFET channel, the Fowler-Nordheim tunneling responsible for write/erase operations is moved to the back of the floating gate, while the gate oxide is kept thick enough to suppress tunneling at all times.

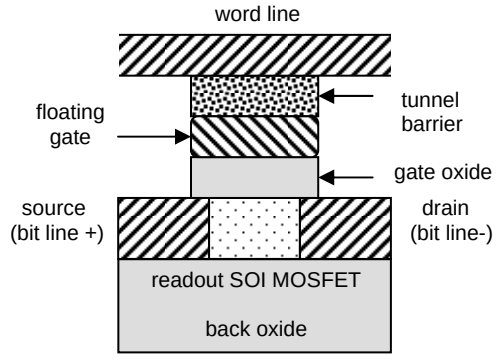


Fig. 1: Memory cell structure of NOVORAM and FGRAM [2].

[1] The difference of κ alone may also provide a transparency steepness improvement [3], though for realistic values of parameters this effect it weaker than that of the tunnel barrier height difference.

The later experimental work has shown that layered barriers made of several material combinations (including $Si_3N_4/SiO_2/Si_3N_4$ [4, 5], SiO_2/ZrO_2 [6], $HfON/Si_3N_4$ [7], and SiO_2/AlO_x [8]) can indeed improve the barrier transport sensitivity to voltage in comparison with the traditional SiO_2 barriers. Unfortunately, to the best of our knowledge, the conductivity change ranges demonstrated so far have not been sufficient for the full implementation of the NOVORAM concept. In particular, the attempts by our group to combine different species of aluminum oxide (for example, thermally-grown and plasma grown AlO_x [9]) to form crested barriers so far have not been successful. However, in the course of this work we have found a way to fabricate quasi-uniform aluminum oxide layers with very high transport properties, including high endurance to electric fields in excess of 10 MV/cm, and extremely high values of charge-to-breakdown (close to 10^6 C/cm^2). These properties may be used in what we call Floating-Gate Random-Access Memories (FGRAM) with the cell structure similar to NOVORAM (Fig. 1).[2] Essentially the only difference of the memory operation is the necessity to refresh the FGRAM contents exactly as this is currently done in DRAM. Our AlO_x barriers may provide the retention time (of the order of 1 s) necessary for this operation.

2. Fabrication

The barriers have been fabricated in nearly the same way as those described in Ref. 9. Briefly, thin (10-nm-scale) aluminum films have been dc-sputtered either directly on oxidized silicon wafers or on a sub-layer of a different metal. Immediately after their deposition, the films have been oxidized in an rf plasma discharge, with power from 10 to 250 W at oxygen pressure in the range from 15 to 75 mtorr. (The results shown below correspond to the lower ends of these ranges.) Immediately after the oxidation (without a vacuum break) the junctions have been sealed with a metallic counter-electrode. Such in-situ fabrication results in highly reproducible junctions, with conductivity scaling well with the junction area A (which ranged from 3×3 to 300×300 μm^2). In order to improve the junction quality, in particular their endurance in high electric field, after the lithographic area definition, they have been subjected to rapid thermal post-annealing (RTA) for 10 to 180 seconds at temperatures from 300 to 550°C.

[2] This opportunity was briefly mentioned in Ref. 10.

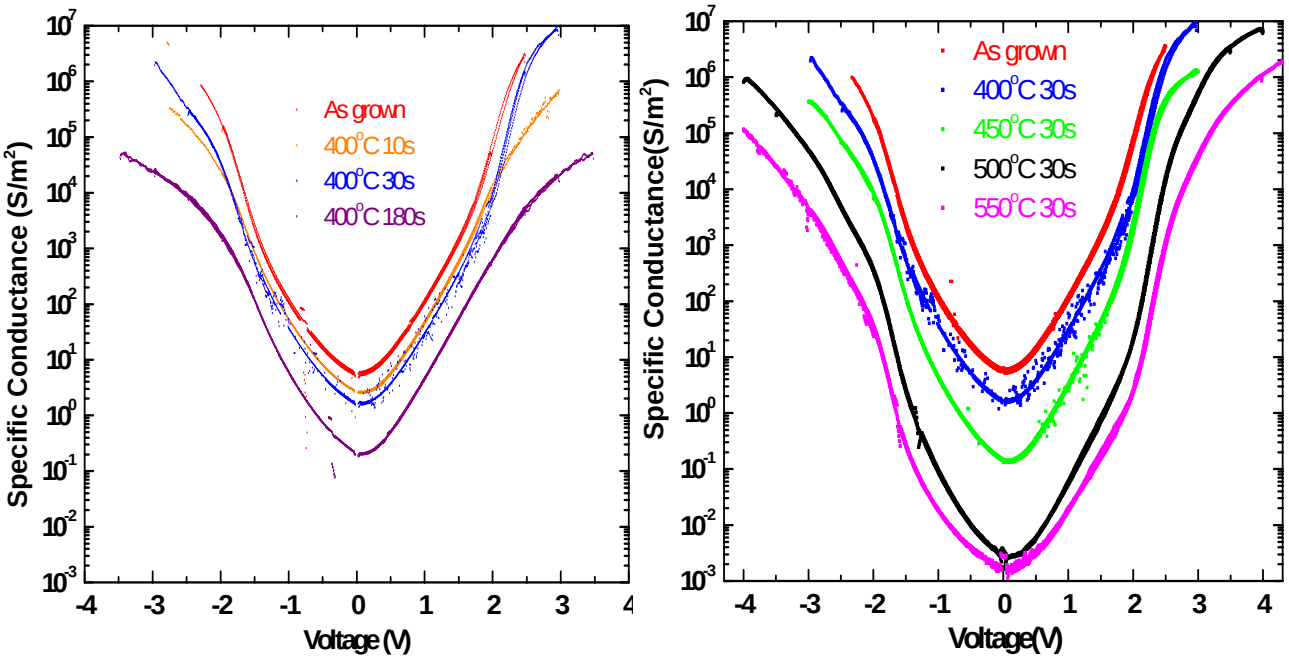


Fig. 2 : Specific differential conductance $G \equiv A^{-1}(dI/dV)$ of junctions from wafer CB17 (rf power 10 W, oxidation time 10 minutes) as a function of applied voltage, for various durations and temperatures of the rapid thermal post-annealing. The data are for T = 4.2 K, but the room temperature results are close, with the only exception of somewhat lower breakdown voltage.

3. Experimental results

Throughout this range of fabrication conditions, the junctions show steep *I-V* curves (Fig. 2) with very weak temperature dependence (similar to that shown in Fig. 2 of Ref. 9), which can only be explained by direct tunneling[3] of electrons through the whole AlO_x layer. Moreover, the fitting of the curves with the "microscopic" (non-WKB) theory [9] has shown that the results may be reasonably well described by tunneling through a uniform potential barrier with a height (depending on the exact fabrication parameters) from 2.0 to 2.4 eV and an effective thickness $d_{ef} = (m_{ef}/m_0)^{1/2}d$ from 1.75 to 2.5 nm. The estimates of the effective mass m_{ef} of the carriers using the junction capacitance measurements [9], as well as high-resolution TEM (courtesy by Dr. Y. Zhu, Brookhaven National Laboratory) show that the physical thickness d of the barriers is in reasonable correspondence with d_{ef}, with the ratio m_{ef}/m_0 somewhere between 0.3 and 0.5.

As Fig. 2 shows, the rapid thermal annealing results in a dramatic improvement of the junction endurance to high electric field. In particular it increases the breakdown dc fields above 10 MV/cm at room temperature (and above 15 MV/cm at 4.2K), i.e. substantially beyond those for the best SiO_2 layers we are aware of.

Another striking feature of these junctions is their high charge-to-breakdown Q_{BD} which (for some fabrication parameters) exceeds 10^5 C/cm^2, the number to be compared with ~10^1 C/cm^2 for typical SiO_2 layers used in flash memories. Figure 3 shows a more adequate figure-of-merit, the maximum number of write cycles $N \equiv Q_{BD}/CV$, plotted versus the calculated write time scale $\tau \equiv CV/I(V)$, where C is the junction capacitance (for our samples, between 1.5 and 2.0 μF/cm^2), V is the (high) applied voltage, and $I(V)$ the current corresponding to this voltage.

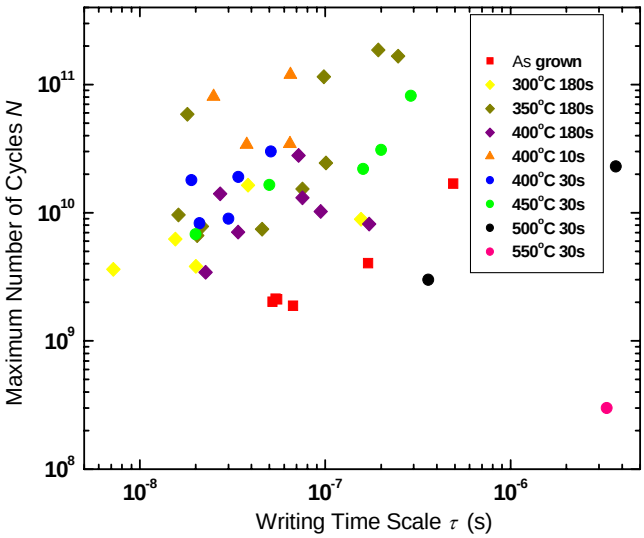


Fig. 3 : Field endurance of junctions from wafer CB17 for several RTA parameter sets, at room temperature.

One can see that at semi-optimized post-processing, the junctions can combine a 20-ns-scale write time (acceptable for most applications currently using stand-alone DRAM chips) with ~10^{11} write cycles and ~1-second-scale retention time $\tau_R \equiv C/G(0)$. We believe that these parameters enable the application of FGRAM, based on such tunnel barriers, for at least some RAM

[3] We use this term to describe all ranges of applied voltage, including the Fowler-Nordheim regime.

applications, though the further increase of N may be still desirable. Our plans are to continue the optimization of fabrication parameters to achieve this goal.

4. Conclusion

To summarize, we have shown that such simple, CMOS-compatible fabrication steps as plasma oxidation of aluminum with rapid thermal post-annealing of the resulting junctions enable the implementation of fast, scalable floating-gate memories which may be suitable for at least some RAM applications. We believe that such memories, after a modest improvement, may become the RAM of choice for integrated circuits beyond the 32-nm ITRS technology node.[4]

5. Acknowledgments

This work was supported in part by AFOSR. Useful discussions with E. Cimpoiasu, J. Cosgrove, T. P. Ma, S. Tolpygo, and X. W. Wang, as well as the generous help by Y. Zhu and his BNL group with HRTEM, are gratefully acknowledged.

References

[1] K. Likharev, "Layered tunnel barriers for nonvolatile memory devices". *Appl. Phys. Lett.*, vol. 73, pp. 2137-2139, Nov. 1998.

[2] K. K. Likharev, "NOVORAM: A new concept for fast, bit-addressable nonvolatile memory based on crested barriers", *IEEE Circuits and Devices*, vol. 16, pp. 16-21, June 2000.

[3] B. Govoreanu, P. Blomme, M. Rosmeulen, J. Van Houdt, and K. De Mayer, "VARIOT: A novel multilayer tunnel barrier concept for low-voltage nonvolatile memory devices", *IEEE Electron Dev. Lett.*, vol. 24, pp. 99-101, Feb. 2003.

[4] S. J. Baik , S. Choi, U-I. Chung, and J. T. Moon, "Engineering on tunnel barrier and dot surface in Si nanocrystal memories", *Solid-State Electronics*, vol. 48, pp. 1475–1481, Sep. 2004.

[5] S. H. Hong, J. H. Jang, T. J. Park, D. S. Jeong, M. Kim, C. S. Hwang, and J. Y. Won, "Improvement of the current-voltage characteristics of a tunneling dielectric by adopting a Si_3N_4 /SiO_2/Si_3N_4 multilayer for flash memory application", *Appl. Phys. Lett.*, vol. 87, art. 152106, Oct. 2005.

[6] B. Govoreanu, P. Blomme, J. Van Houdt, and K. De Meyer, "Enhanced tunneling current effect for nonvolatile memory applications", *Jpn. J. Appl. Phys., Part 1*, vol. 42, pp. 2020–2024, Apr. 2003.

[7] Y. Liu, S. I. Shim, F. C. Yeh, X. W. Wang, and T. P. Ma, "Barrier engineering for non-volatile memory application", Report 3.2 at *SISC 2006* (San Diego, CA, December 2006), and to be published.

[8] J. Cosgrove, private communication (2006).

[9] E. Cimpoiasu, S. K. Tolpygo, X. Liu, N. Simonian, J. E. Lukens, K. K. Likharev, R. F. Klie and Y. Zhu, "Aluminum oxide layers as possible components for layered tunnel barriers", *J. Appl. Phys*., vol. 96, pp. 1085-1093, July 2004.

[10] K. K. Likharev, "Electronics below 10 nm", in: J. Greer *et al.* (eds.), *Nano and Giga Challenges in Microelectronics*, Elsevier, Amsterdam, 2003, pp. 27-68.

[11] J. C. Scott, "Is there an immortal memory?", *Science*, vol. 304, pp. 62-63, Apr. 2004.

[12] D. B. Strukov and K. K. Likharev, "Defect-tolerant architectures for nanoelectronic crossbar memories", *J. of Nanoscience and Nanotechnology*, vol. 7, pp. 151-167, Jan. 200

[4] Their main competition may be resistive memories [11] with their more compact memory cells ($4F^2$ vs. $6F^2$), especially in their hybrid ("CMOL") version [12].

Intrinsic fixed charge and trapping properties of HfAlO interpoly dielectric layers

M. Bocquet [a b], G. Molas [a], H. Grampeix [a], J. Buckley [a], F. Martin [a], J. P. Colonna [a], M. Gély [a], G. Pananakakis [b], G. Ghibaudo [b], B. De Salvo [a], S. Deleonibus [a]

[a] CEA-LETI, 17 rue des Martyrs 38054 Grenoble Cedex 9, France, marc.bocquet@cea.fr
[b] IMEP-CNRS/INPG, MINATEC – INPG – 3, Parvis Louis Néel 38016 Grenoble, France

Abstract

The objective of this work is to investigate the fixed charge and the trapping properties of SiO_2 – HfAlO – SiO_2 (OHO) tri-layer stacks as interpoly dielectrics of NVMs. This study focuses on the key role played by the HfAlO composition. We show that the intrinsic fixed charge content increases with the Al concentration, while the trapping capabilities, during a gate stress, increases with the Hf ratio of the compound. We argue also that the charge trapping happening during a gate stress is mainly located at the high-k interface rather than in the volume. Retention characteristics are also shown. Finally, the experimental data are explained through a model based on a SRH approach.

1. Introduction

In order to meet the performance requirements of future generations of Flash memory [1], one of the nearest major improvements will concern the scaling of the InterPoly Dielectric (IPD) stack. For the 45nm and 35nm nodes, in order to compensate the loss of the vertical sidewalls of the poly-Si floating gate and keep high the coupling ratio [2, 3], the IPD thickness should be reduced. In a previous work [4], we proposed HfAlO high-k materials to replace the nitride layer in ONO interpoly dielectric stacks for future Flash memories, arguing the advantages both in terms of coupling and insulating properties. We showed that the leakage current was strongly governed by the trapping in the high-k layer, with a strong temperature activation. Indeed, a Poole-Frenkel conduction, probably assisted by the traps in the HfAlO layer, was identified.

In this paper, we further developed the analysis of HfAlO high-k materials, embedded in OHO stacks, by focusing on the trapping properties and fixed charges. In particular, we investigate: *(1)* the intrinsic negative fixed charge in the oxides, *(2)* the trapping phenomena which take place in the high-k dielectrics with various compositions and thicknesses during a gate stress, *(3)* the retention properties of these layers, and *(4)* finally, we present simulations based on a Shockley Read Hall (SRH) approach which allow us to model the electron trapping in the OHO stacks.

2. Experimental results

2.1 Sample description

The schematic of the triple layer capacitors studied in this work is shown in Figure 1. The high-k films were sandwiched between two HTO (High Thermal Oxide) deposited at 730°C, with a thickness of 4nm.

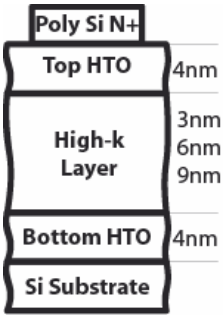


Fig. 1: Schematic showing the capacitor device stack studied in this work. Various high-k were investigated: HfO_2, Al_2O_3 and HfAlO with different Hf:Al ratios (9:1, 1:4 and 1:9).

Three different high-k materials, deposited by ALCVD, were studied: Hafnium Oxide (HfO_2), Aluminum Oxide (Al_2O_3) and Hafnium Aluminate (HfAlO). In HfAlO films, the Hf concentrations were controlled by the $HfCl_4$:$Al(CH_3)_3$ deposition cycle ratio, which are respectively: 9:1 (94% of Hf), 1:4 (31% of Hf), and 1:9 (27% of Hf). Different high-k physical thicknesses ranging between 3nm and 9nm were fabricated by controlling the number of ALD deposition cycles.

2.2 Intrinsic negative fixed charge

In this section we investigate the intrinsic fixed charge of the HfAlO layers.

Figure 2 plots the capacitance-voltage characteristics of the studied OHO triple layers in the virgin state.

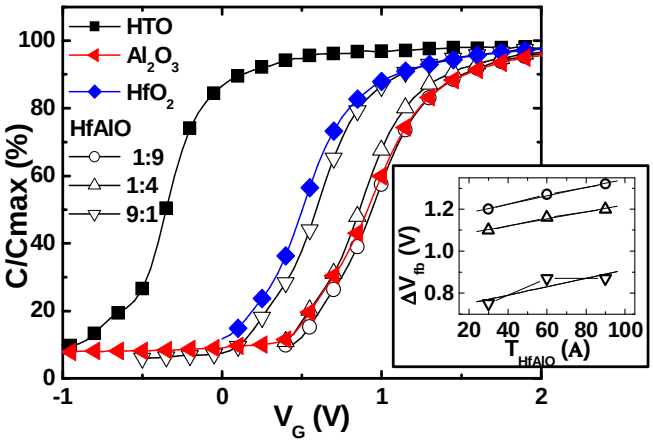


Fig. 2: C–V_G characteristics (in the virgin state) of OHO stacks with various high-k materials (Al_2O_3, HfO_2 and HfAlO with different compositions) compared to a 10nm HTO reference device. Inset: Flatband voltage shifts as a function of the HfAlO physical thickness.

We can observe that the flatband voltages of OHO samples are shifted compared to the 10nm-thick HTO reference sample, due to the presence of intrinsic fixed charge in the high-k layers. The shift increases monotonically when the Al concentration of the HfAlO alloy increases.

The origin of the intrinsic negative fixed charge in HfAlO materials is nowadays still not clear [5]. In amorphous Al_2O_3 layer, it was suggested that the Al_2O_3 could be dissociated into $(AlO_{4/2})^-$ and Al^{3+} [6] and that, at the SiO_2/Al_2O_3 interface, the charge compensation does not take place.

The inset of Figure 2 shows that ΔV_{FB} is a linear function of the HfAlO thickness, which suggests a surface rather than a volume distribution of the fixed charge according to Equation 1, being in agreement with previous works reported in the literature [7-9].

$$\Delta V_{fb} = \frac{t_T}{\varepsilon_{SiO_2}} \cdot (\sigma + t_H \cdot \rho) + \sigma \cdot \frac{t_H}{\varepsilon_H} + \rho \cdot \frac{t_H^2}{2 \cdot \varepsilon_H} \quad (1)$$

with :

- σ : charge density at high-k/Bottom HTO interface.
- ρ : volume charge density in high-k.
- t_H : thickness of high-k layer.
- t_T : thickness of HTO top layer.
- ε_H : high-k dielectric constant.
- ε_{SiO2} : SiO_2 dielectric constant.

Table 1 summarizes the number of equivalent fixed charge localised at the bottom HTO/High-k interface, calculated from the V_{FB} shifts.

HfAlO composition	Number of fixed charge nb/cm^2
9:1	$3 \cdot 10^{12}$
1:4	$3.5 \cdot 10^{12}$
1:9	$4 \cdot 10^{12}$

Table 1: Number of equivalent fixed charge (extracted from the characteristics reported in Figure 2) localised at the bottom HTO /High-k interface.

2.3 Trapping properties

In this section, we investigate in more details the charge trapping phenomena of OHO samples during a gate stress.

The gate current density as a function of the electric field is reported in Figure 3. On the same graph is also reported the time evolution of the gate current at constant gate bias (J_G–Time) on virgin devices. The hysteretic behaviour, and the continuous decreasing of the leakage current with the elapsing time demonstrate that trapping phenomena take place in the high-k materials. Note that the charge trapping could be an issue for IPD applications, as it may degrade the reliability of the memory, generating threshold voltage instabilities.

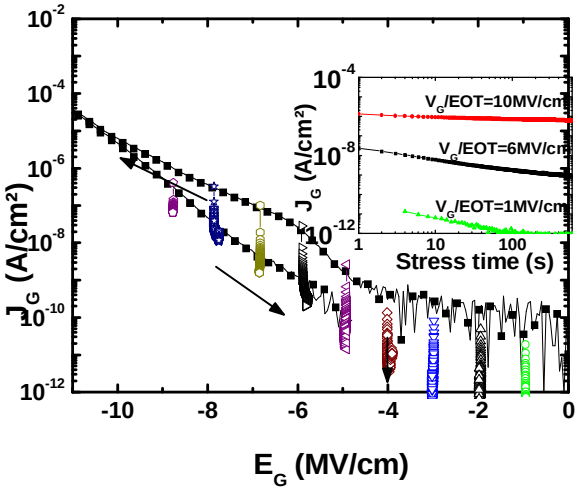


Fig. 3: Current density J_G versus equivalent electric field E_G of HTO /HfAlO 9:1 – 9nm/HTO stack. J_G–time measurements on virgin devices are also represented for different electric fields. Inset: J_G–time measurements performed at different applied electric field (1MV/cm, 6MV/cm and 10MV/cm) as a function of time.

To evaluate more precisely the trapping capabilities of the interpoly stacks, we monitored the evolution of the flatband voltages as a function of time when the devices were submitted to different gate stresses (Figure 4). A continuous V_{FB} shift is observed, showing the progressive electron trapping in the stack as the stress time increases. It clearly appears that for a given stress condition, the trapping capability increases with the Hf concentration. As already reported in the literature [4, 10], this result could be correlated with the crystalline structure of the high-k materials: the larger the Hf concentration, the more crystalline the layer, and hence, the higher the trapping capability.

Based on these measurements, we extracted the charges trapped in the gate stack after programming. We assume that the charges are localized at the interface between the Bottom HTO and the HfAlO layer. Indeed, the further the traps are from the cathode, the slower the charging kinetics. Making this assumption, it appears that the extracted trapped charge value does not depend on the HfAlO thickness (Figure 5). This confirms (see Equation 1) that charges are mainly trapped at the high-k interface and that the bulk contribution is negligible at the first order.

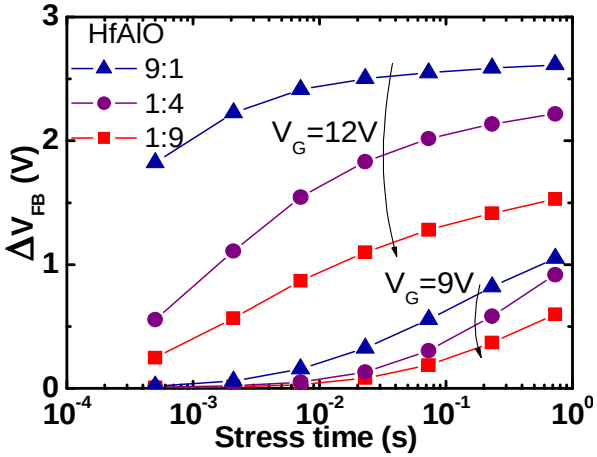


Fig. 4: Programming characteristics of OHO samples with various HfAlO compositions. The HfAlO thickness is 6nm.

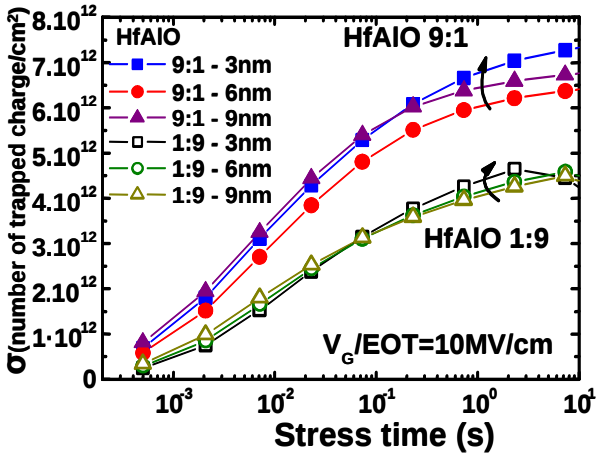


Fig. 5: Trapped charges in OHO samples, with various compositions and thicknesses of the HfAlO layer, as extracted from the programming characteristics. The stressing conditions are performed at constant V_G/EOT. We assume that the charges are localized at the interface between the Bottom HTO and the HfAlO layer.

2.4 Retention characteristics

In this section, we investigate the dynamic discharging of the charges previously trapped in the OHO layers by a writing stress. Figure 6 plots the room temperature retention characteristics of OHO samples with different compositions of the HfAlO layer. We observe that for the Hf-rich sample, the electron discharging rate is quite fast, whereas the HfAlO 1:4 and 1:9 keep most of the charge after 10^6s.

Figure 6 also shows the strong temperature activation. Indeed, the charge loss is strongly accelerated at 85°C (within a factor 2 for the 9:1 sample). Nevertheless, the trend observed at room temperature is still conserved, i.e.: the 1:4 and 1:9 HfAlO layers exhibit the same charge decay, while for the 9:1 HfAlO sample, only 50% of trapped charge remains after 10^6s. These observations are consistent with experiments performed on SONOS-like structures with high-k trapping layers [11].

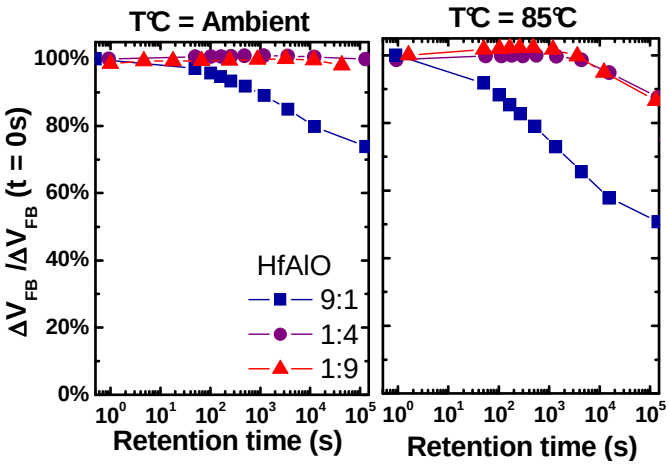


Fig. 6: Room temperature and 85°C retention characteristics of OHO samples with various compositions of HfAlO. The HfAlO thickness is 6nm. The programming conditions are fixed to have an initial flatband voltage shift of 1.5V for each sample.

3. Modelling

In this part, we introduce an analytical model to qualitatively explain the trapping characteristics of our IPD stacks. To this aim, we use the SRH model presented in [10], and we focus on the Hf-rich (9:1) OHO samples.

The simulations are performed assuming that:

- The trapped charge is localized at the bottom HTO / HfAlO interface, which is consistent with our experimental data. Note that in a more realistic approach, we should take into account the charge trapped in the HfAlO bulk, characterised by slower trapping time constants.
- The thermalization of electrons in the SiO_2 layer close to the cathode is neglected. In fact, in our range of programming voltages, the electron paths in the conduction band of the HTO, after tunneling, is inferior to 1nm (Figure 7), which is indeed shorter than the mean free path of electrons in SiO_2 [12]. In other words, we assume that the electron energy remains constant till the trapping in HTO/HfAlO interface states happens.
- The gap (E_G=5.65eV) and the permittivity (εr=17) of 9:1 HfAlO were extracted by ellipsometry and by C–V_G measurements, respectively. We also consider that the bottom and top HTO are 4.5nm and 5nm thick respectively, which is in agreement with TEM observations.
- The shift between the conduction band of Si and HfAlO, ΔE_C, and the trapping cross section, σ, were fixed based on literature data: ΔE_C=2eV [13,14], $\sigma=10^{-18}cm^2$ [15,10].
- N_{st}, the trap density, is adjusted to fit the experimental saturation level.

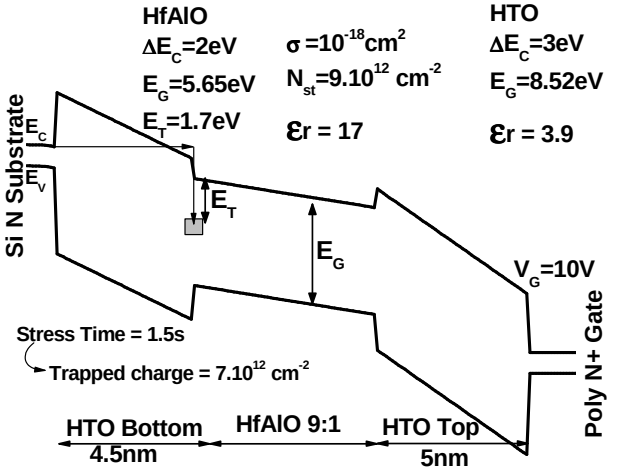


Fig. 7: Energy band diagram of OHO stack at V_G=10V simulated in this work. Fitting parameters are indicated. The HfAlO thickness is 6nm, the concentration is 9:1. The charges trapped in the HTO-HfAlO interface are responsible for the different electric field values in the bottom and top oxide layers.

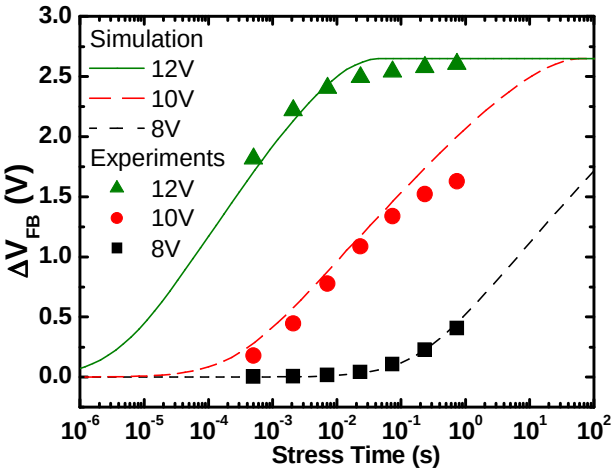


Fig. 8: Modeling of the trapping characteristics of OHO stack (HfAlO thickness is 6nm, concentration is 9:1), based on structure and parameters reported in Figure 7.

Based on these assumptions, we use the following equations to simulate the programming characteristics:

$$\Delta Vth = \frac{N_{st} \times ft}{Ct}$$

$$\frac{dft}{dt} = (1 - ft) \times (cn + ep) - ft \times (en + cp)$$

- cn/en and cp/ep are the electron and hole capture/emission rates which govern carrier exchanges between the traps and substrate.
- ft: trap occupation probability according to Shockley-Read-Hall statistics model [16].
- Ct: trap to gate coupling capacitance.

Figure 8 shows the experimental and the modelling trapping characteristics. We observe a very good correlation between the simulation and the experimental data for the three programming voltages, which validates our theoretical approach.

4. Conclusions

In this paper we investigated the intrinsic fixed charge and trapping phenomena happening under stress of HfAlO based interpoly dielectric stacks. We demonstrated that the fixed charge content increases with the Al concentration of the HfAlO layer. We showed that the trapping capability when the device is submitted to a constant voltage stress increases as the Hf ratio of the compound increases. Based on programming measurements, we proved that in our devices the electron trapping mainly occurs at the first interface, between the bottom HTO and the HfAlO layer, rather than in the volume of the high-k dielectrics. We also observed that the discharging rate of the previously trapped charges is more important for Hf-rich alloys. Finally, an analytical model based on a SRH approach allowed us to fit our experimental data and to extract the main trapping parameters of HfAlO high-k materials.

Acknowledgments

Part of this work was supported by the MEDEA+ NEMESYS project.

References

[1] http://www.itrs.net/Common/2005ITRS/Home2005.htm
[2] M. Alessandri *et al*, Proc. Of the 208th ECS Meeting, (2005).
[3] J. V. Houdt, IRPS Tech. Dig., pp. 234-239 (2005).
[4] G. Molas *et al*., Proc. of ESSDERC, pp. 242-245 (2006).
[5] J. H. Lee *et al* , Tech. Dig. of IEDM (2000).
[6] G. Lucovsky and J.C. Phillips, Appl. Surf. Sci. **166**, , pp. 497-503, (2000).
[7] S. J. Lee *et al*, IEEE Elec. Dev Lett **24** (2), pp. 105-107 (2003).
[8] S. H. Bae, C. H. Lee, R. Clark, and D. L. Kwong, IEEE Elec. Dev Lett **24** (9), pp. 556-558 (2003).
[9] J. Buckley *et al*, Microelectronic Engineering **80**, pp. 210–213 (2005).
[10] J. Buckley *et al*., Tech. Dig. of IEDM, (2006).
[11] Y. N. Tan *et al*., Tech. Dig. of IEDM, pp. 889-892 (2004).
[12] O. Brière, K. Barla, A. Halimaoui and G. Ghibaudo, Solid-State Electronics 41(7), pp. 987-990 (1997).
[13] A. Gehring and S. Selberherr, IEEE Trans. on Device and Materials Reliability **4**, (2004).
[14] G. D. Wilk *et al*, Journ. of Appl. Phys. **89**, pp. 5243-5275 (2001).
[15] A. Fernandes et al., Proc. of ESSDERC, pp. 139-142 (2001).
[16] D. Ielmini et al., IEEE Trans. on Elec. Dev. **47** (6), pp.1258-1265 (2000).

A fully planar Stacked Gate Flash Technology with T-shaped Floating Gate for increased cell coupling ratio.

Joeri De Vos, Luc Haspeslagh, Pieter Blomme, Marc Demand,
Katia Devriendt, Frank Vleugels, Dirk Wellekens, Jan Van Houdt

IMEC, Kapeldreef 75, B-3001 Leuven, Belgium, E-mail: Joeri.DeVos@imec.be

Abstract

In this paper we demonstrate the process integration of a planar stacked gate process with T-shaped floating gate. This concept further builds on the scalable stacked gate technology presented in [1]. However, the T-shaped cell allows higher coupling ratios without sacrificing the beneficial planar character of the structure and is therefore perfectly suited for interpoly dielectrics with moderate k-value (such as the well-established Al_2O_3). The higher coupling ratio lowers program and erase voltages and makes the interpoly layer less sensitive to cycling induced trapping.

1. Introduction

For scaling down floating gate (FG) based memory to 45nm and beyond, we proposed in [1] a planar structure which uses high-k material as interpoly dielectric (IPD). A planar structure is beneficial for controlling the different layer depositions without step-coverage issues. A structure without topography also results in a larger process window for gate patterning. All these are major add-ons to non-planar structures, where the control gate (CG) is wrapped around the FG [2-4].

For replacement of the traditional ONO layer as interpoly dielectric different possible high-k materials are reported in literature [5-7]. Among those, aluminum oxide is a good candidate because it is not only known as a stable and mature material, but also has a high tunnel barrier enabling 10 year data retention times at high temperature (125ºC) [6, 8]. For high-k interpoly dielectrics with rather low k-value like Al_2O_3, the coupling ratio of the structure presented in [1] is however limited to 50-60%, resulting in a higher sensitivity to P/E cycling induced charge trapping. Therefore, a sufficiently high coupling ratio is required, which is always obtained by wrapping the CG around the FG, leading to a non-planar structure.

In this paper we propose a novel structure that allows a large coupling ratio without giving up the beneficial planarity of the structure in [1]. An important parameter in the coupling ratio of a floating gate cell is the ratio of the CG-to-FG capacitor area and the FG-to-substrate capacitor area. In the structure presented in [1], the area ratio equals 1. In the stacked gate process with a modified T-shaped floating gate that we propose in this paper, the area of the CG-to-FG capacitor is enlarged, while keeping the area of FG-to-substrate capacitor unchanged. This leads to a higher coupling ratio, which decreases program and erase voltages and makes the cell more reliable. Although the cell area is increasing with the introduction of the T-shaped FG, this adjusted process scheme is well suited for those applications where planarization is more relevant than aggressive area scaling.

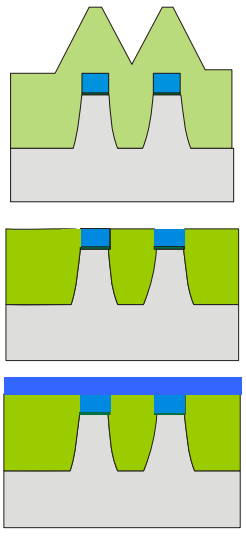
2. Process scheme

In figure 1 the process flow of the stacked gate cell with T-shaped FG is presented.

The oxide CMP process of the STI formation is optimized to prevent oxide from remaining on top of the nitride layer. The oxide etch-back in the STI module is skipped.

A 115nm PECVD nitride layer is deposited and patterned by means of 193nm lithography. This extra mask is an oversized version of the mask used for patterning active area. It will determine the field overlap of the floating gate underneath the control gate. The coupling ratio will increase with increasing overlap.

After resist strip, 200nm of HDP oxide is deposited, followed by a second oxide polish. A cross section SEM picture at this point is shown in figure 2.



STI module
HDP oxide deposition

Oxide CMP
No oxide etch back
No nitride removal

Nitride deposition

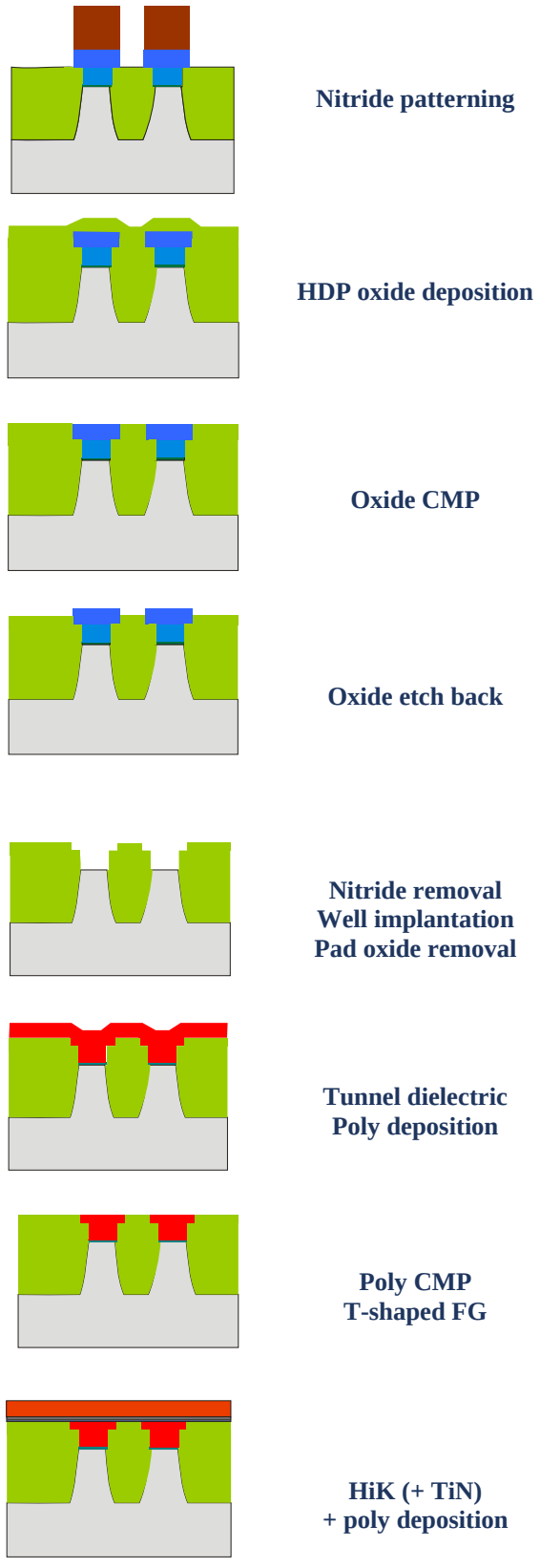


Fig. 1 Overview of the stacked gate etch process flow with T-shaped floating gate.

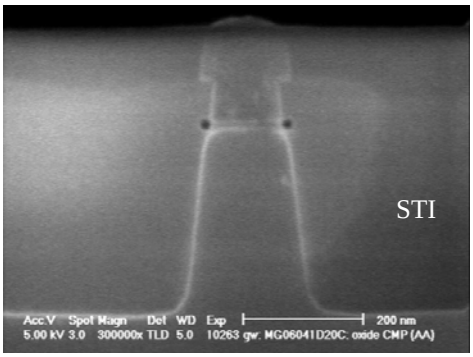


Fig. 2 T-shaped structure formed in sacrificial nitride layer (picture taken after short HF etch to enhance contrast).

Further in the process flow, the nitride layer will be replaced by poly to form the floating gate of the cell. In order to remain in the process window of our standard stacked gate etch process [1], the total thickness of the FG should be limited to 100nm. Therefore, the HDP oxide is etched back at this point by 90nm by means of wet HF etching.

After wet nitride removal, the wells are implanted. Then either the tunnel oxide is grown or either the VARIOT [9] stack is deposited, followed by an in-situ phosphorous doped poly deposition. Finally, a well-controlled poly CMP yields the T-shaped form of the FG. The bottom part of the FG is still self-aligned to the active area. In the cross-section SEM picture of figure 3 the T-shaped FG is shown. The total thickness is about 100nm (80nm of the self-aligned part). In this figure, it can be seen that the planarity of the structure is maintained when introducing the T-shaped FG.

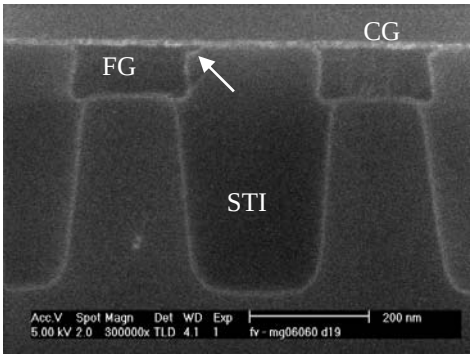


Fig. 3 Cross-section SEM after FG CMP. The arrow indicates the 30nm field overlap of the floating gate.

After defining the FG, the IPD layer, poly and a PECVD oxide are deposited on a flat structure, the last layer acting as a hard mask. The T-shaped process flow is dedicated to investigate ALCVD Al_2O_3 in combination with a bottom oxide. Chemical oxides, high temperature oxides (HTO) and ISSG oxides as bottom oxide are under investigation.

At this point also a thin TiN metal layer can be included before poly deposition, which allows a larger threshold voltage window, because of reduced erase saturation [10].

Further process steps are copied from the standard stacked gate process. Gate patterning is performed with 193nm lithography. On the active area, the entire double poly stack has to be etched, whereas on the field area only the top electrode and the IPD layer are etched. Therefore high selectivity towards field oxide is needed when etching the bottom poly layer.

A hard mask based approach with trimming is used to pattern the stack. The top and bottom poly gates (CG and FG) are etched selectively towards the underlying layer using a conventional poly gate etch recipe. BCl_3 is used to etch the Al_2O_3 layer.

After the stacked etch, a dry removal of the hard mask is performed, followed by an ISSG poly sidewall reoxidation.

Then the LDD extensions are implanted and spacers are formed. HDD implantations, salicidation, Cu metallization and a 20 minutes forming gas anneal at 420°C finalize the processing.

3. Electrical results

As already mentioned in the introduction, the extra mask needed to form the T-shaped floating is an oversized version of the active mask. The minimum oversize to assure good alignment to the active mask is 30nm at each side. On the test mask various oversize lengths between 30nm and 400nm were included. The nominal width of the device is 150nm.

In figure 4 the measured voltages to program/erase the device to 2V above/below the intrinsic Vt in a timeframe of 1ms are plotted as a function of the coupling ratio. The interpoly dielectric consists of 7nm Al_2O_3 on top of a 5nm HTO bottom layer. Below the FG, conventional 8.5nm tunnel oxide was replaced by a bidirectional VARIOT stack [10] consisting of 2nm SiO_2/8nm Al_2O_3/2nm SiO_2. It is clear that the introduction of the T-shaped FG has a positive influence on the applied voltages. The enlarged coupling ratio also reduces the electric field across the IPD, resulting in more reliable devices.

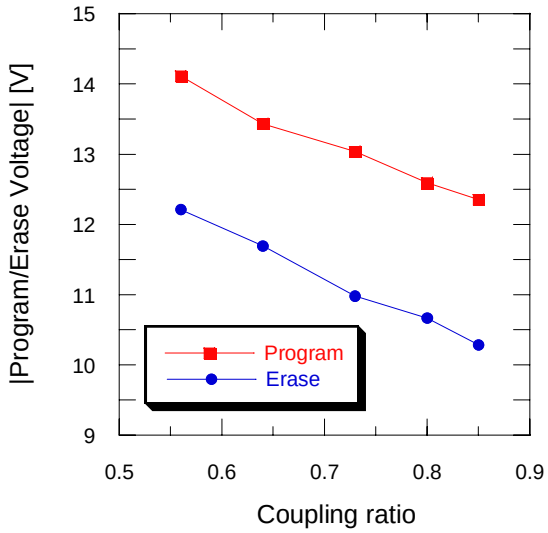


Fig.4 Measured program and erase voltage dependency on the coupling ratio. The coupling ratio was calculated starting from the FG overlap on field.

4. Conclusion

We proposed a stacked gate technology with a T-shaped floating gate. This technology is well suited for interpoly dielectric materials with moderate k-values like Al_2O_3, as it yields an increased coupling ratio, while maintaining all the advantages of a fully planar structure at the expense of only 1 additional mask. An increased coupling ratio results in a reduction of the program/erase voltages and a more reliable interpoly dielectric.

Acknowledgements

This work was performed under IMEC's Industrial Affiliation Program on Advanced Memory Technology together with Intel Corp., Infineon Technologies, Micron Technology and Samsung Electronics.

References

[1] J. De Vos et al., ICICDT 2006, pp.21-24
[2] W. Kwon et al. , Extended Abstracts of the SSDM 2005 pp.448-449
[3] C. Park et al. , VLSI Technology 2004, pp.238-239
[4] J. Park et al., IEDM 2004, pp.873-876
[5] M. van Duuren et al., ICICDT 2006, pp.36-39
[6] A. Miranda et al., ESSDERC 2006, pp.234-237
[7] G. Molas et al., ESSDERC 2006, pp.234-237
[8] P. Blomme et al., NVSMW 2006, pp.52-53
[9] B. Govoreanu et al., EDL 2003 Vol. 24, No 2, pp. 99-101
[10] D. Wellekens et al., ESSDERC 2006, pp.238-241

On the localization of the trapped charges in Silicon nanocrystal NOR Flash devices

S.Jacob[a,b], L.Perniola[b], B.De Salvo[b], E.Jalaguier[b], G.Festes[a], R.Coppard[a], F.Boulanger[b], S.Deleonibus[b]

[a] ATMEL Rousset, Zone industrielle, 13790 Rousset, France
[b] CEA-LETI, 17 rue des Martyrs, 38054 Grenoble Cedex 9, France, *stephanie.jacob@cea.fr*

Abstract

In this work, we present a thorough investigation of the localization of the trapped charges in Silicon nanocrystal (Si-NC) devices written by Hot Electron (HE) injection (i.e. NOR configuration). In particular, we study the influence of the applied gate- and bulk-biases, as a function of the writing time. Exhaustive experiments are explained through a comprehensive electrostatic analytical model, suitable for discrete-trap memories, and compared to 2D dynamic TCAD simulations. We argue that the trapped charge location remains nearly constant when the gate stress bias is raised, whereas the charge centroid shifts from the drain junction toward the channel as the stressing time increases or as a negative voltage is applied to the bulk.

1. Introduction

According to the ITRS [1], conventional Flash devices will have to face drastic scaling down in the next years [2]. In this context, one of the most critical challenges to be solved will be the reduction of the tunnel oxide thickness while keeping data retention. A promising solution consists in replacing the continuous poly-Si floating gate of standard Flash devices with discrete storage nodes made by nanometer-sized Si-NCs [2, 3]. In a NOR configuration, the device is written by injecting channel hot electrons close to the drain junction. The injected charges are trapped in the Si-NCs located above the drain/channel metallurgical junction, and they do not diffuse all over the channel as in conventional poly-Si floating gate memories, due to the discrete nature of the Si-NCs. Consequently, the localization of the trapped electrons in Si-NCs is a critical point which should be considered in view of the technological optimization of the memory cell [4]. Simulations of the HE injection current under different bias conditions in discrete trap memories have been addressed in the past, especially for NROM devices [5, 6], and physics of the Si-NC cell has been addressed through semi-analytical models [7, 8, 9]. Nevertheless, the comprehension of the localization of the injected charges in Si-NC NOR memories requires deeper investigations.

In this work, we present an exhaustive set of experimental data of Si-NC NOR devices. A comprehensive analytical model suitable for discrete-trap memories [10, 11] is then applied to the obtained data, to extract the trapped charge density and the charged region length. Finally, the results are compared to 2D numerical TCAD dynamic simulations.

2. Experimental results

Electrical tests have been performed on NMOS memory cells with a layer of LPCVD (Low Pressure Chemical Vapor Deposition) Silicon nanocrystals acting as floating gates. The mean diameter of nanocrystals is about 5 nm and the density is 1E12 dots/cm^2. The tunnel and top oxide dielectrics are thermal SiO_2 and HTO (High Temperature Oxide), with a thickness of 4 nm and 10 nm, respectively. The gate length and width of the cell are 0.23µm and 0.16µm, respectively (corresponding to the ATMEL 0.13µm NOR Flash technology node) (see Fig. 1).

We have performed hot electron writing tests to study the influence of the gate and bulk voltages as a function of the stressing time, on the programming characteristics. The written threshold voltage has been read in the forward (i.e. while applying V_d= 1V and V_s=0 V) and reverse mode (i.e. while applying V_d=0 V and V_s=1 V). Indeed, for discrete-trap memories, as the charges are confined near the drain junction, the forward threshold voltage (V_{thF}) is lower than the reverse threshold voltage (V_{thR}). In fact, during the forward read, the trapped charges in Si-NCs are screened by the applied drain voltage and the effect of the charge pocket on the channel potential is reduced, giving rise to a lower V_{thF}. On the contrary, the reverse read is more sensitive to the effect of the negative trapped charges which strongly decrease the channel conduction [12, 13]. As the length of the trapped charge region increases toward the channel, the difference between the reverse and forward V_{th} decreases.

Firstly, writing tests have been performed for different stressing gate voltages (6, 8 and 10V) and different times (10µs and 100µs). Fig. 2 shows the evolution of the written V_{th} of the cell read in forward and reverse modes, as a function of the gate stress. As the stressing gate voltage increases, the V_{th} increases quite linearly. Moreover, the difference between the reverse and forward V_{th} remains constant.

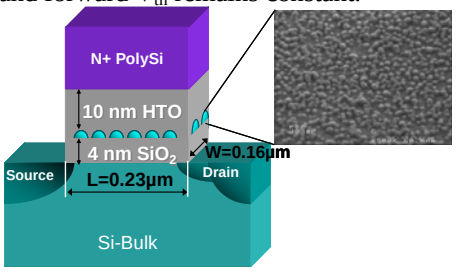


Fig. 1: Schema of the Si-NC devices studied in this work (a SEM picture of Si-NC layer is shown).

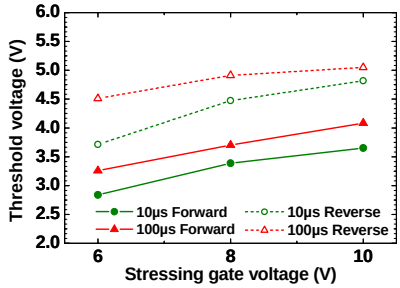


Fig. 2: Dependence of the V_{th} on the writing gate voltage for two stressing times (t_{stress}=10µs or 100µs). Vd_{stress}=5V, V_{bulk}=0V, V_s=0V. The device is read at $|V_{ds}|$=1V and V_{th}@I_{ds}=100nA.

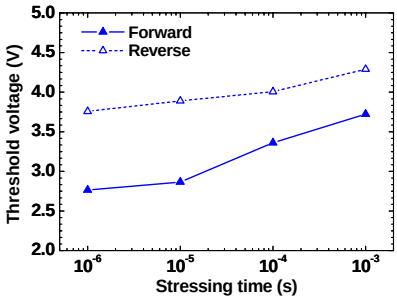


Fig. 3: Dependence of the V_{th} on the stressing time Vg_{stress}=10V, Vd_{stress}=3.5V, V_{bulk}=0V, V_s=0V. The device is read at $|V_{ds}|$=1V and V_{th}@I_{ds}=100nA.

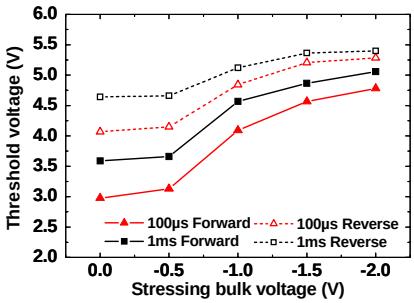


Fig. 4: Dependence of the V_{th} on the applied bulk voltage for two different times (t_{stress}=10µs or 100µs). Vg_{stress}=8V, Vd_{stress}=3.5V, V_s=0V. The device is read at $|V_{ds}|$=1V and V_{th}@I_{ds}=100nA.

Secondly, the impact of the stressing time has been studied more in details by exploring a larger range (from 1µs to 1ms). The forward read curve on Fig.3 shows the saturation of the threshold voltage versus time, while we see that the difference between reverse and forward states becomes smaller for longer stress times.

Finally, we have observed the effect of applying a negative bulk bias during programming. This is known as the CHISEL (CHannel Initiated Secondary Electrons) phenomenon [14]. It has been shown in the literature [5] that the secondary electrons coming from the bulk are injected over the middle of channel, whereas the electrons coming from the CHE (Channel Hot Electrons) mechanism are injected close to the drain junction. Experimentally, we see in Fig. 4 that larger programming windows are obtained with increasing |Vb|, while the difference between reverse and forward states becomes smaller. Furthermore, it can be noticed that for our devices the programming window attains a saturation level at V_{bulk}=-1.5V.

3. Analytical modeling

An analytical model well describing the electrostatic of discrete trap memories (presented in details in [11]), has been used in this work to understand how the Id-Vg characteristics of the written devices are influenced by the trapped charge region length. The model is based on the computation of the surface potential of non-uniformly charged cells and is able to provide an analytical formula for the subthreshold slope and the threshold voltage of memory devices.

The first step consists in giving an analytical expression of the surface potential along the channel of a virgin cell (see Fig. 5):

$$\Psi_S(y) = (\Psi_r - \Psi_L)\frac{\sinh(y/\lambda)}{\sinh(L/\lambda)} + (\Psi_l - \Psi_L)\frac{\sinh[(L-y)/\lambda]}{\sinh(L/\lambda)} + \Psi_L \quad (1)$$

Where y=0 (y=L) corresponds to the source (drain) contact, $\Psi_l = V_{bi} + V_s - V_b$, and $\Psi_r = V_{bi} + V_d - V_b$, where V_{bi} is the built-in voltage at the drain-bulk and source bulk junctions, V_s is the source voltage, V_d is the drain voltage, V_b is the bulk voltage used for Id-Vg reading and Ψ_L is the long channel surface potential [15].

The parameter λ is defined as:

$$\lambda \equiv \sqrt{\frac{\varepsilon_{Si} t_{eox} X_{dep}}{\varepsilon_{ox}\eta}} \quad (2)$$

where t_{eox} is the equivalent oxide thickness of the total memory gate stack, X_{dep} is the space charge region thickness and $\varepsilon_{Si(ox)}$ is the silicon (oxide) permittivity. η is a fitting parameter (normally calibrated by fitting the surface potential of the virgin cell with numerical simulations).

As a second step, the electrostatic effect of the charge is considered via the superposition principle. The charge region localized near the drain junction is assumed to be uniform, while the region over the rest of the channel is free of charge. The pocket of charge is described in terms of two effective parameters, an effective charged length L_2 and an effective charge density Q. In the 2-D structure (see Fig. 5), the perturbative potential Ψ_{Sp}, is obtained by integrating the elementary contributions along the vertical direction x and longitudinal direction y [16], yielding:

$$\Psi_{Sp}(y) = \frac{-\rho}{\pi(\varepsilon_{Si}+\varepsilon_{ox})} \times \int_{L_1}^{L_1+L_2} dy' \int_{t_1}^{t_1+t_m} \ln\left(\sqrt{(y-y')^2+x^2}\right) dx \quad (3)$$

where ρ is the density of injected charge per unit volume and t_1 is the tunnel oxide thickness, t_m is the height of the charge region and L_1=L–L_2. Finally, the surface potential $\Psi_{Stot} = \Psi_S + \Psi_{Sp}$.

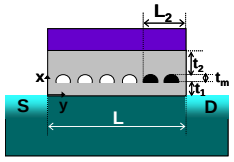


Fig. 5: Sketch of device to be considered for calculation of surface potential.

4. Data interpretation

The analytical model presented in Section III is here used to extract information on the distribution of the trapped charges in the written devices. In Fig. 6, the experimental and analytical transfer characteristics in the virgin and written states for the forward and reverse readings have been plotted. A good agreement is achieved between measurements and the model.

Experiments presented in Section II, give the transfer characteristics and the values of the threshold voltage of the virgin cell V_{thinit}, as well as V_{thF} and V_{thR} for the written devices. Then, the two quantities defined as:

$$\Delta V_{thtot}=V_{thR}-V_{thinit}$$

and

$$\Delta V_{RF}=V_{thR}-V_{thF}$$

can be computed.

Based on the analytical model, it is possible to create a map with the contour plots of ΔV_{thtot} and ΔV_{RF} as a function of the density of electrons Q and of the charged length L_2. By introducing on this map the experimental values of ΔV_{thtot} and ΔV_{RF} in different writing conditions, it is possible to have a description of the trapped charge region through Q and L_2.

Fig. 7 shows the fitting of data provided in Fig. 2. As we can see, while raising the gate stress voltage, the density of trapped electrons Q increases, while the charged length L_2 remains the same (about 50nm) for the three stressing gate voltages.

Secondly, in Fig. 8, the influence of the writing time shown in Fig. 3 has been fitted with the contour plots. It clearly appears that the charged length L_2 increases with the stressing time, showing that the injected charges shift toward the channel during the stress. L_2 increases of about 25nm as the time increases from 1µs to 1ms.

Finally, in Fig. 9, we have introduced in the contour plots the data shown in Fig. 4, concerning the dependence of the programming window with respect to the writing bulk voltage. We observe a large increasing of the injected charge Q as V_{bulk} decreases from 0 to -1V and then the density of charge begins to saturate. A similar behavior is observed for the charged length. L_2 increases from 45nm to 60nm as V_{bulk} decreases from 0V to -1V and then remains constant.

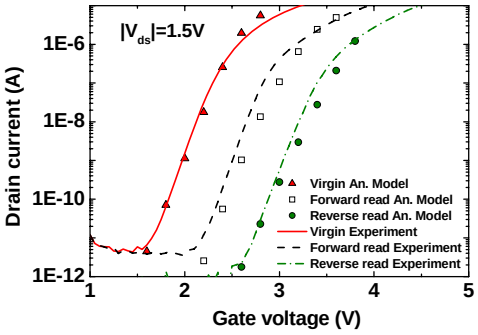


Fig. 6: Analytical model vs experiments for virgin and written characteristics ($|V_{dsread}|$=1.5V). Writing conditions are Vg_{stress}=8V, Vd_{stress}=5V, V_{bulk}=0V and t_{stress}=10µs. Parameters used for the analytical model are L_2= 66 nm, Q=2.44E12 electrons.cm^{-2}.

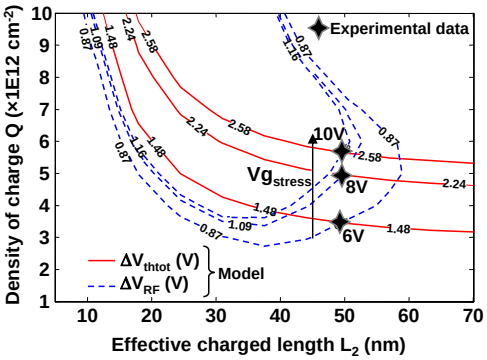


Fig. 7 : Contour plots of ΔV_{thtot} (solid lines) and ΔV_{RF} (dashed lines) versus the effective charged region length L_2 and the density of trapped charge Q. Data corresponding to different stressing gate voltages (6V, 8V, 10V) while Vd_{stress}=5V, V_{bulk}=0V, V_s=0V and t_{stress}=10µs (see Fig. 2), are reported.

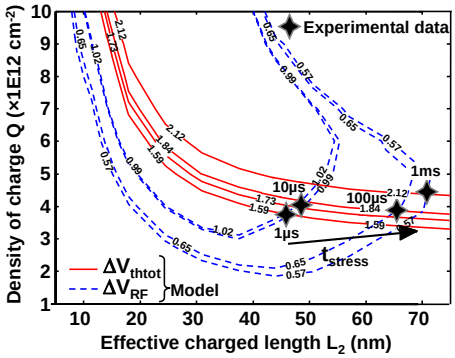


Fig. 8: Contour plots of ΔV_{thtot} (solid lines) and ΔV_{RF} (dashed lines) versus the effective charged region length L_2 and the density of trapped charge Q. Data corresponding to different stressing times (1µs, 10µs, 100µs, 1ms) while Vg_{stress}=10V, Vd_{stress}=3.5V,V_{bulk}=0V and V_s=0V (see Fig. 3), are reported.

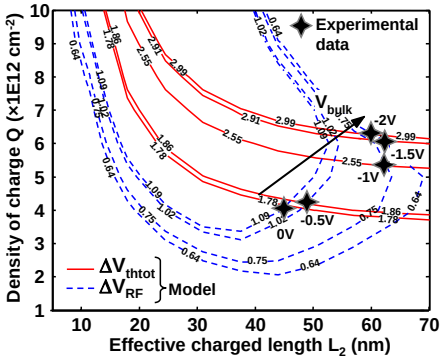


Fig. 9: Contour plots of ΔV_{thtot} (solid lines) and ΔV_{RF} (dashed lines) versus the effective charged region length L_2 and the density of trapped charge Q. Data corresponding to different writing bulk voltages (0V, -0.5V, -1V, -1.5V, -2V), while Vg_{stress}=8, Vd_{stress}=3.5V, V_s=0V and t_{stress}=100µs (see Fig. 4), are reported.

4. TCAD simulation results

2D TCAD simulations of a nanocrystal memory have been performed with commercial tools [17] in order to validate the previous presented results. The structure used in the device simulator, including doping profiles, was obtained by 2D process simulations. Nanocrystals were approximated as uniformly distributed metallic cubes with 5nm edge and a density of 1E12 dots/cm² (see Fig. 10).

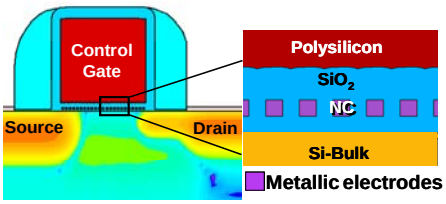


Fig. 10: View of the 2D simulated structure.

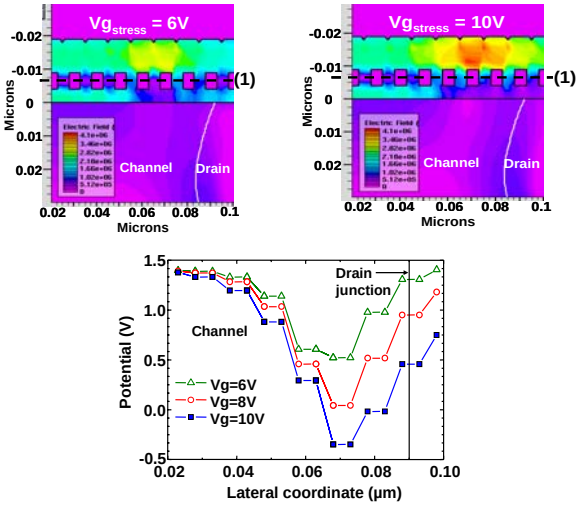


Fig.11: Up: Electric field in 2D device for different writing gate voltages, while Vd_{stress}=5V, t_{stress}=10µs, V_{bulk}=0V, V_s=0V. Down: Potential cuts in the dots along the cutline (1).

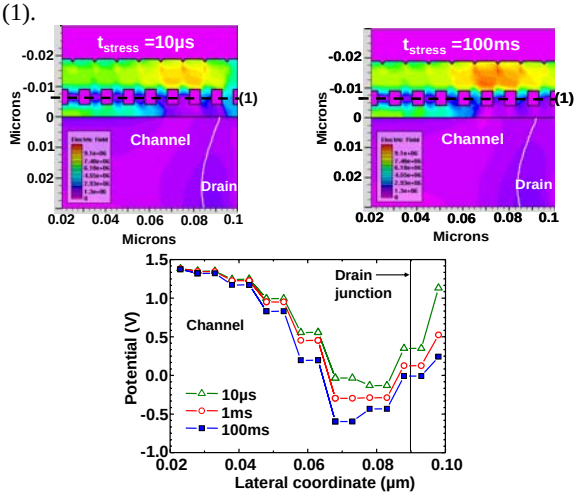


Fig.12: Up: Electric field in 2D device for different writing times, while Vg_{stress}=10V, Vd_{stress}=3.5V, V_{bulk}=0V, V_s=0V. Down: Potential cuts in the dots along the cutline (1).

Dynamic hot electron injection simulations have been performed with the energy balance model [18, 19] for the carrier transport and the lucky electron model [20] for the injection current calculation. In a previous work [4], we have shown that the 2D simulations provided a good agreement with the experimental results as a function of the stressing gate voltages and writing times.
Fig. 11 represents the potential in nanocrystals for the three writing conditions corresponding to Fig. 2. It shows that the injected electrons keep at the same channel lateral location for different Vg stresses, in agreement with the analytical results shown in Fig. 7.
Then we have plotted on Fig. 12 the potential in nanocrystals at different writing times, corresponding to data shown in Fig.3. We observe that the charge moves toward the channel as the stressing time increases, which once again is consistent with the increasing of the effective charged length L_2 found with the analytical model (see Fig. 8).

4. Conclusions

In this paper, we have presented experimental results of Si-NC cells written in hot electron injection configurations. These results have been correlated to an analytical model, which has allowed us to extract both the trapped charge density and the trapped charge length in Si-NCs after programming. Finally, 2D dynamic TCAD simulations of the hot electron injection have been performed to validate the previously presented results.
We have shown that, during a hot electron writing stress, the gate bias has a nearly linear influence on the injected charge density, whereas the writing time and the stressing bulk bias lead to a saturation of the injected charge density. Moreover, the localization of the trapped charges remains the same (close to the drain junction) while increasing the gate bias, whereas the charges are shifted toward the channel as the stressing time increases or as a negative bulk voltage is applied.

Acknowledgements - This work has been done within the "EREVNA" joined development program between ATMEL and CEA-LETI.

References

[1] International Technology Roadmap for Semi-conductors, 2005 Edition: http://www.itrs.net/Links/2005ITRS/Home2005.htm
[2] B. De Salvo et al., IEEE Trans. Device Mater. Reliability, 4, 377 (2004).
[3] R. Muralidhar et al., IEEE IEDM Tech. Dig., 601 (2003).
[4] S. Jacob et al., Proc. of Non-Volatile Memory Technology Symposium, 31 (2006).
[5] G. Ingrosso, L. Selmi and E. Sangiorgi, Proc. of ESSDERC, 187, (2002).
[6] R. Hagenbeck et al., Proc. of SISPAD, 322 (2006).
[7] B. De Salvo et al., IEEE Trans. El. Dev., 48, 8, 1789, (2001).
[8] A. Campera et al., Solid State Electronics, 49, 1745 (2005).
[9] C. Compagnoni et al., IEEE Trans. El. Dev., 52, 4, 569, (2005).
[10] L. Perniola et al., IEEE IEDM Tech. Dig. (2005).
[11] L. Perniola, G. Iannacone and G. Ghibaudo, IEEE Tech. Nano., 5, 4, 373 (2006).
[12] B. De Salvo et al., IEEE IEDM Tech. Dig., 597 (2003).
[13] L. Larcher, G. Verzellesi, P. Pavan, E. Lusky, I. Bloom, and B. Eitan, IEEE Trans. El. Dev., 49, 11, 1939 (2002).
[14] J.D. Bude et al., IEEE IEDM Tech. Dig., 279 (1997).
[15] Y. Divides, Operation and modeling of the MOS Transistor, 2nd ed. New York: Columbia Univ., 74 (1999).
[16] L.D. Landau, E.M. Lifshitz and L.P. Pitaevskii, Electrodynamics of Continuous Media, 2nd ed. London, U.K.: Butterworth-Heinemann (1984).
[17] http://www.silvaco.com/.
[18] R.Stratton, Phys. Rev., 126, 6 (1962).
[19] R.Stratton, IEEE Trans. El. Dev., 12, 1288 (1972).
[20] Tam, S et al, IEEE Trans. El. Dev., 31, 9 (1984).

Electrostatics and its effect on spatial distribution of tunnel current in metal Nanocrystal flash memories

Aneesh Nainani[a], Arunanshu Roy[a], P.K. Singh[a], G. Mukhopadhyay[b], Juzer Vasi[a]

[a] Department of Electrical Engineering, IIT-Bombay, Mumbai, India -40076
[b] Department of Physics, IIT-Bombay, Mumbai, India -40076

Abstract

In this paper we present an analytical formulation of the electrostatics of metal based Nanocrystal (NC) flash memory, and study its effect on tunnelling probabilities. We establish that asymmetry in field distribution resulting from the electrostatics enhances the field near the NC. A spatial distribution of tunnelling probabilities is presented for the first time. This analytical formulation can be easily coupled with the Schrödinger's equation to describe the Program/Erase dynamics of the NC, greatly reducing the computational time.

1. Introduction

Discrete charge storage based Nonvolatile memories such as metal-oxide-nitride-oxide-silicon (MONOS), silicon-oxide-nitride-oxide-silicon (SONOS), and Nanocrystal (NC) have been widely studied as a possible direct adaptation for the conventional flash memories to meet scaling requirements as per the ITRS Roadmap. NC memory devices made of Si, Ge, and metal dots have recently received attention as promising candidates to replace conventional FG flash [1-3]. While Si and Ge dot devices show barrier-lowering problem due to quantum confinement [4], metal dot devices do not show this problem due to large density of states around the Fermi level in metals. Metal dot memories are shown to have better charge capacity and better retention characteristics than Si NC memory in comparable gate stacks [5] due to deeper quantum well. A lot of work has been done on the materials and fabrication aspects of metal NC devices to improve performance [6-7].

While quite a few attempts have been made towards modeling of NC based Flash memories [4, 8], little literature exists [9] for modeling the electrostatics due to presence of metal NCs in the gate stack and no attempt has been made to model its effect on the tunneling currents between the NC and substrate and between two neighboring NCs. Most of the studies [4, 8] use one dimensional (1D) Fowler-Nordheim and direct tunneling expressions which is inaccurate due to two reasons. Firstly, the 1D formulation assumes that the variation in electric field is only along one co-ordinate axis, which is violated due to the discrete nature of the NCs. Secondly, the application of a 1D expression assumes dots to be cubical in shape, which is unrealistic. Also if the distribution of these cubes is made random as expected in a realistic process, the faces of the cubes between which tunneling occurs are not aligned with respect to each other, in which case its not wise to apply the 1D formulation for the inter dot tunneling. On the other hand using spherical dots overcomes both the problems stated above, but few attempts have been made to model the electrostatics of such a system and its effect on tunneling completely.

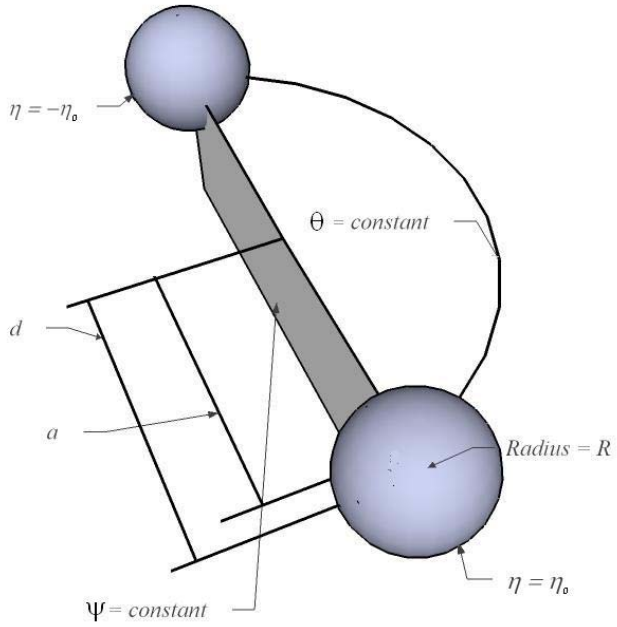


Fig. 1 : Bispherical Coordinate System. $a = \sqrt{d^2 - R^2}$, where d is the distance of the center of the sphere from the xy-plane and R is the radius of the sphere. See Equations (1)-(4).

Some of the recent studies [10] give a full quantum mechanical treatment to the problem, where the electron energy levels and wave functions in a NC, depending on the bias voltage are calculated through a finite element method. This solution is computationally very intensive and unsuitable for a system having more than a couple of dots.

We present a new approach where exact analytical solutions of the Laplace's Equation are obtained for (a) two spheres at any given voltages, and (b) a sphere and a plane, by employing Bispherical Coordinates. Principle of least action and the Wentzel–Kramers–Brillouin (WKB) approximation is used over this potential distribution to obtain tunneling probabilities. Polarization of fields which results in the enhancement of electric field in vicinity of NC is established and its effect on tunneling is examined, which appears to be another factor in support of metal based NC memories. A spatial distribution of tunneling probabilities is presented for the first time.

2. Proposed Approach

A typical gate stack of NC-based memories consists of a distribution of spherical NCs formed by self assembly on top of a tunneling dielectric deposited on a silicon substrate. A positional distribution in NCs is evident from the STEM images [9]. The NCs are covered with control dielectric with a metal gate on top. We approximate the substrate as a ground plane. This is quite accurate in the inversion and accumulation regimes, i.e., during *P*/*E* operations, and sets the upper bounds for electric fields in the depletion region. Since the main focus of this work is to study the electrostatics and its effect on tunneling between the two NCs or the NC and substrate, the analysis is restricted to a pair of conductors, ignoring any image effect. The 3D electrostatics from the interaction of a metal NC and substrate or from the interaction of two metal NC's has an analytical formulation in Bispherical Coordinates. Using the notation of [11], the bispherical coordinates are defined as

$$x = \frac{a \sin\theta \cos\phi}{\cosh\eta - \cos\theta} \quad (1)$$

$$y = \frac{a \sin\theta \sin\phi}{\cosh\eta - \cos\theta} \quad (2)$$

$$z = \frac{a \sin\eta}{\cosh\eta - \cos\theta} \quad (3)$$

Surfaces of constant η are given by

$$x^2 + y^2 + (z - a\coth\eta)^2 = \frac{a^2}{\sinh^2\eta} \quad (4)$$

and surfaces of constant θ are given by apples for $\theta > \pi/2$ and lemons for $\theta < \pi/2$ as shown in Fig. 2.

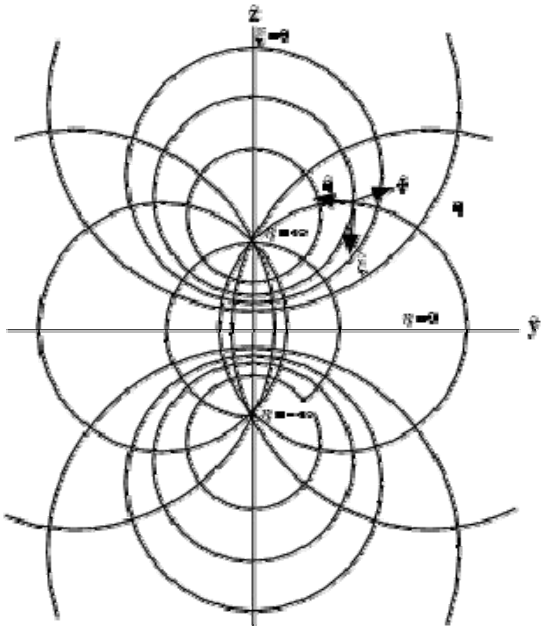


Fig 2: Surfaces of constant η and θ. Surfaces of constant θ are given by apples for $\theta > \pi/2$ and lemons for $\theta < \pi/2$.

While solving Laplace's Equation $\nabla^2\psi = 0$ for spheres at constant potentials, it is convenient to use Bispherical Coordinates, which allows easy implementation of the boundary condition. By setting $\psi = \sqrt{\cosh\mu - \cosh\eta}\, F$. The equation yields:

$$\frac{\partial^2 F}{\partial\eta^2} + \frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial F}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2 F}{\partial\phi^2} - \frac{1}{4}F = 0 \quad (5)$$

This separates into three simple differential equations (detailed analysis is given in [11]). The typical solution is given by

$$\psi = \sqrt{\cosh\eta - \cos\theta}\times \sum_{n=0}^{\infty}\left[A_n e^{(n+\frac{1}{2})\eta} + B_n e^{-(n+\frac{1}{2})\eta}\right]P_n(\cos\theta) \quad (6)$$

where P_n are the Legendre polynomials. A_n and B_n are determined by applying appropriate boundary conditions. For spheres at potentials $\Psi(\eta_0)$ and $\Psi(-\eta_0)$:

$$A_n = \sqrt{2}\left[\frac{\psi(-\eta_0) - \psi(\eta_0)e^{2(n+\frac{1}{2})\eta_0}}{1 - e^{4(n+\frac{1}{2})\eta_0}}\right] \quad (7)$$

$$B_n = \sqrt{2}\left[\frac{\psi(\eta_0) - \psi(-\eta_0)e^{2(n+\frac{1}{2})\eta_0}}{1 - e^{4(n+\frac{1}{2})\eta_0}}\right] \quad (8)$$

where $\eta = \eta_0$ corresponds to the surfaces of the NCs and is related to the geometry by [Fig.1] :

$$\eta_0 = \sinh^{-1}\left(\frac{a}{R}\right) \quad (9)$$

For $\Psi(\eta_0) = -\Psi(-\eta_0) = V_0$, the positive half of the potential distribution is similar to that between a sphere held at potential V_0 and the grounded plane. Fig. 3, shows such a distribution for V_0 = 1V, R =2.5nm, d= 4nm.

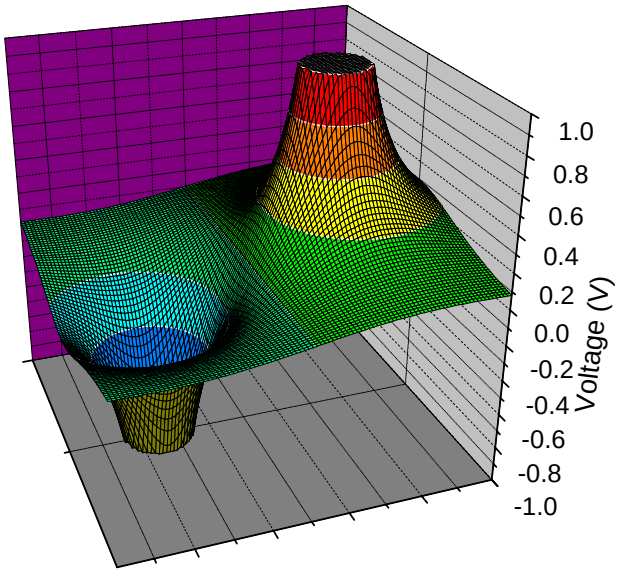


Fig. 3: Potential distribution for the system as shown in Fig.1, with two spheres at potentials 1V and -1V respectively.

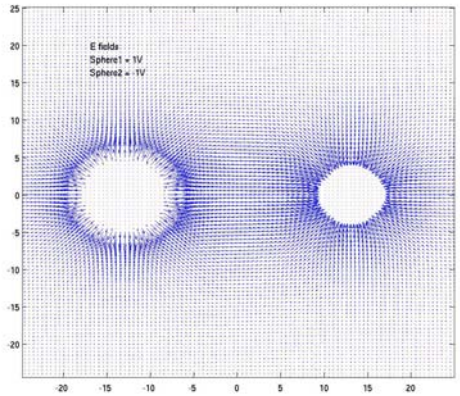
3. Results and Discussion

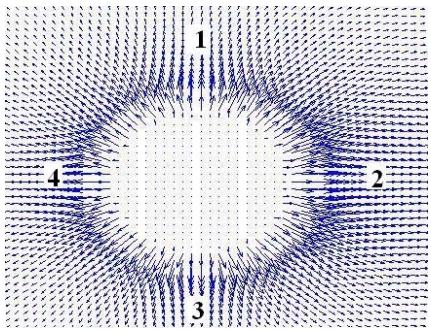
Fig. 4(a) shows the electric field resulting from the potential distribution given in Fig. 3 assuming SiO_2 as the dielectric medium. Polarization of electric field that can be attributed to the discrete nature of NCs is depicted in

Fig.4 (b). It is observed that contribution from the higher order Legendre terms in (6) cannot be ignored, which results in higher orders of polarization than just a simple dipole as suggested by [12]. Four regions of polarization are visibly identifiable in Fig.4 (b).

As visible in Fig.3 the potential distribution is skewed near the NC, which results in significantly higher fields near the vicinity of the NC. This is further illustrated in Fig. 5 and Fig. 6 where potentials and fields are plotted directly along the z-axis from substrate to the bottom end of the NC.

(a)





(b)

Fig. 4(a) Electric Fields resulting from potential distribution in Fig.3; (b) Polarization of Electric Field – Four centers of polarization are marked as 1,2,3,4

The polarization and enhancement of field near the NC is expected to have a significant effect on the Fowler-Nordheim and direct tunneling. The tunneling probabilities are calculated using WKB approximation over the path of least action. The tunneling coefficient is given by

$$T = \left(\frac{\left| dV(\mathrm{r})/ds \right|_{s=s2}}{\left| dV(\mathrm{r})/ds \right|_{s=s1}} \right) \exp(-T_{12}) \qquad (10)$$

where T_{12} is given by

$$T_{12} = \frac{2}{\hbar} \int_{s1}^{s2} ds \sqrt{2m[V(\mathrm{r}) - E]} \qquad (11)$$

The integration is done along the path of least action (s1 to s2) calculated from the field distribution using Newton's equation of motion; which results from a semi classical approximation as explained in [13] . (s1-s2) is the corresponding arc length.

This tunneling coefficient determines the contribution of a trajectory to the tunneling current. Fig. 7 shows the tunneling coefficient along the substrate axis for the potential distribution described in Fig.3, for a particle with energy level *E=0* as per (9)-(10). There are two important observations to be made at this point. First, the polarization, which leads to significant enhancement of fields in the vicinity of the NCs results in higher tunneling coefficient near the bottom end of the NC dot, this is much higher than the tunneling coefficient due to 1D tunneling formulation, which is also shown in Fig. 7. We believe that this high tunneling probability in the vicinity of the bottom end of the NC is responsible for large memory window at much lower control gate bias. A memory window of 1V was observed for program voltage of 5V for Tungsten NCs embedded in HfAlO as seen in Fig.8 (fabrication details in [14]) and similar results have been reported by other groups [9], while the memory window is observed to be absent for Si based NC in similar gate stacks [9]. Secondly, though most of the tunneling is concentrated in the region just under the NC, the total tunneling probability reflected by the area under the bell shaped curve is more than that from the 1D formulation reflected by area under the rectangle. Thus using 1D formulation will result in underestimation of tunneling currents.

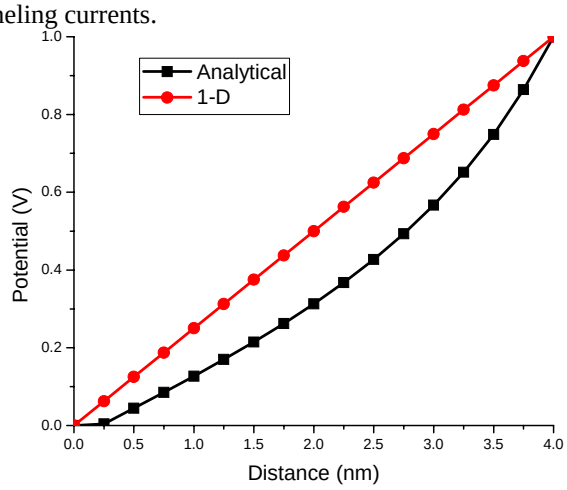


Fig.5: Potential along the z-axis from the substrate towards the bottom of the NC. The potential due to 1D electrostatics is shown by the straight line.

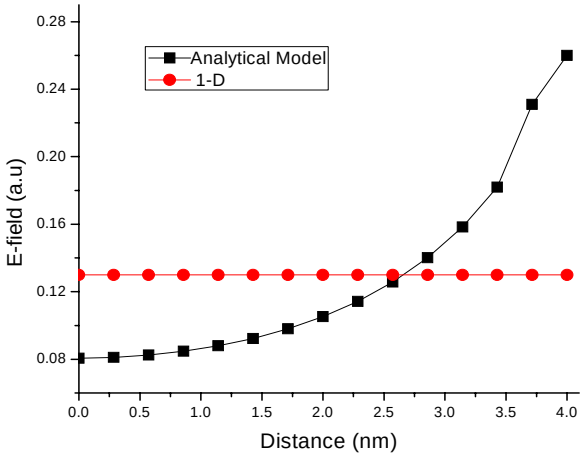


Fig. 6: Electric field along the z-axis from the substrate towards the bottom of the NC. The field due to 1D electrostatics is constant. Note the enhancement of field near the NC.

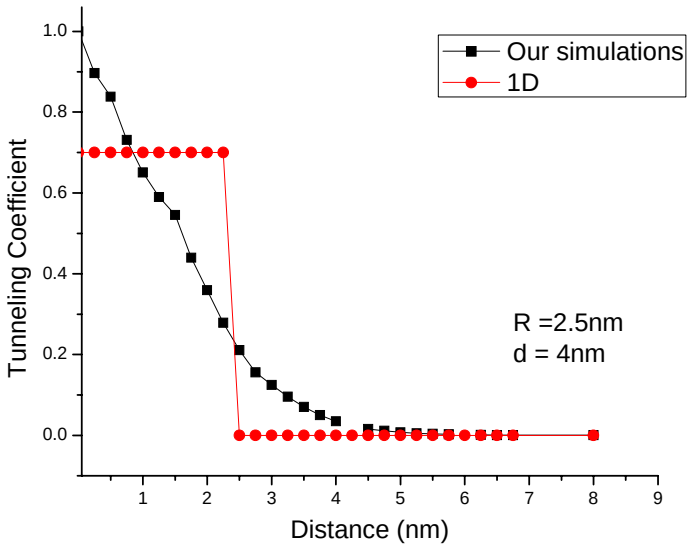


Fig 7: Tunneling coefficient at the substrate for *E=0*, total tunneling is represented by area under the curve. The rectangle represents the tunneling used in a 1D formulation.

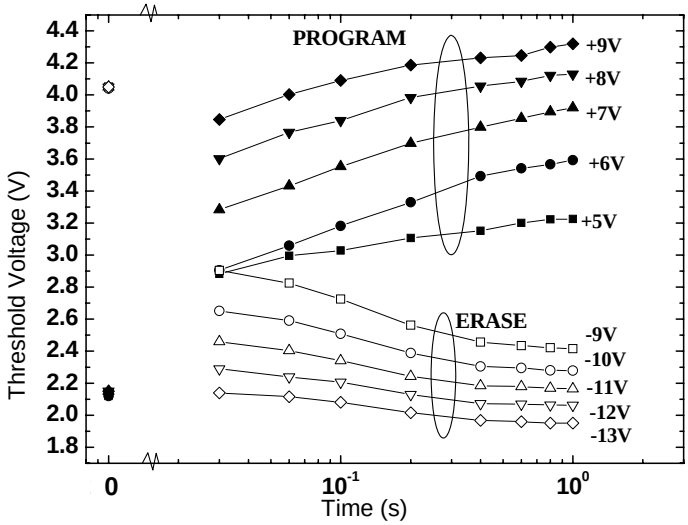


Fig 8: P/E characteristics under NAND operation. Memory window of 1V is observed at 5V programming (Fabrication details in [14]).

4. Conclusion

An analytical solution for the electrostatics in metal NC based memories is presented. It is shown that the electric field significantly enhances in vicinity of the NC, and this enhances the tunneling probability in paths located near the bottom of the NC. This establishes the electrostatic advantage of metal NC based memories as compared to Si NC based memories. The tunneling calculations based on 1D are shown to be insufficient to describe the system. To describe charging/discharging of the NC, coupled solution of the electrostatic with Schrödinger's equation is required. Now that we have a description of electrostatics in a functional form, which reduces the computation time significantly, this becomes feasible; calculations along this line are in progress and will be reported elsewhere.

References

[1] Jing-Hao Chen *et al*, "Nonvolatile Flash Memory Device Using Ge Nanocrystals Embedded in HfAlO High-k Tunneling and Control Oxides: Device Fabrication and Electrical Performance", *IEEE Trans on Electron Devices*, **51**, pp. 1840 – 1848, Nov 2004.

[2] J D Blauwe, "A novel aerosol nanocrystal floating gate device for nonvolatile memory application", *IEDM Tech Dig 2000*, pp. 683-686.

[3] Zengtao Liu *et al*, "Metal Nanocrystal Memories – Part I: Device Design and Fabrication", *IEEE Trans on Electron Devices*, **49**, pp. 1606-1613, Sept 2002.

[4] Min She, Tsu-Jae King, "Impact of crystal size and tunnel dielectric on semiconductor nanocrystal memory performance", *IEEE Trans on Electron Devices*, **50**, pp. 1934 - 1940, Sept. 2003.

[5] C. Lee, A. Gorur-Seetharam, and E. C. Kan, "Operational and reliability comparison of discrete-storage nonvolatile memories: Advantages of single- and double-layer metal nanocrystals", *IEDM Tech. Dig* 2003, pp. 557–560.

[6] Zengtao Liu *et al*, "Metal Nanocrystal Memories – Part II: Electrical Characteristics", *IEEE Trans on Electron Devices*, **49**, pp. 1614-1622, Sept 2002

[7] J. J. Lee *et al*, "Metal nanocrystal memory with high-k tunneling barrier for improved data retention", *IEEE Trans on Electron Devices*, **52**, pp. 507-511, Apr. 2005.

[8] C. M. Compagnoni, D. Ielmini, A. S. Spinelli, A. L. Lacaita, "Modeling of tunneling P/E for nanocrystal memories", *IEEE Tran on. Electron Devices*, **52**, pp. 569-576, April 2005.

[9] Chungho Lee, Udayan Ganguly *et al*, "Asymmetric Electric Field Enhancement in Nanocrystal Memories", *IEEE Electron Device Letters*, **26**, pp. 879-881, Dec 2005.

[10] A.S. Cordan, Y. Leroy, B. Leriche, "Electrostatic coupling between nanocrystals in a quantum flash memory", *Solid State Electronics*, **50**, Feb 2006.

[11] P.M. Morse and H.Feshbach, "Methods of Theoretical Physics: Part-2" , (New York: McGraw-Hill, 1953), pp. 1298

[12] Udayan Ganguly *et al*, "Three dimensional analytical modeling of nanocrystal memory electrostatics", *Journal of Applied Physics*, **99**, Jun 2006.

[13] Z.H. Huang *et al* , "Wentzel-Kramers-Brillouin method in multidimensional tunneling", *Physical Review*, **41**, pp. 31-41, Jan 1990.

[14] S.K. Samanta, P.K. Singh *et al*, "Enhancement of memory window in short channel non-volatile memory devices using double layer tungsten nanocrystals**",** *IEDM Tech Dig 2005*, pp. 170-173.

LIST OF AUTHORS

NOTES

NOTES

NOTES

NOTES

NOTES

NOTES

NOTES

NOTES

NOTES

NOTES

The ICMTD-2007 is organised by IMEC (Belgium), Catholic University of Leuven (Belgium) and L2MP (France):

The ICMTD-2007 is organised with financial and logistics support from the following organisations:

ORGANISATION

Wim DEHAENE
Catholic Univ. of Leuven
Heverlee, Belgium

Pascal MASSON
L2MP
Marseille, France

Jan VAN HOUDT
IMEC
Heverlee, Belgium

Dirk WOUTERS
IMEC
Heverlee, Belgium

COMMUNICATION

Anne DE SMET
Momentum
Leuven, Belgium

Liesbet MASSANT
IMEC
Heverlee, Belgium

SPONSORS