

I
C
M
T
D
2005

1st International Conference on Memory Technology and Design

May 21-24 2005
Giens, France

Proceedings

Technical Co-Sponsor:



The Institute of Electrical and Electronics Engineers, Inc



IEEE Electron Device Society

1st International Conference on Memory Technology and Design

May 21 – 24 2005

Giens, France

Proceedings

Organized by:

Provence Materials and Microelectronics Laboratory (L2MP) and associated institutes: CNRS, Paul Cézanne University, Provence University, Sud Toulon-Var University, Polytech'Marseille, ISEN

Technical Co-Sponsor:

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

Sponsored by:

ATMEL, CREMSI, GDR FIFA (Films Ferroélectriques et Applications), Philips Semiconductors, Région Provence-Alpes-Côtes d'Azur (PACA), STMicroelectronics

Copyright © 2005 by the Provence Materials and Microelectronics Laboratory (L2MP)

All rights reserved

Copyright: abstracting is permitted with credit to the source

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

Additional copies may be ordered from:

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

Tel: (33) 4 91 05 47 79

Fax: (33) 4 91 05 47 82

Email: ICMTD-contact@L2MP.fr

WELCOME TO ICMTD-2005

We take great pleasure in welcoming you to the very first International Conference on Memory Technology and Design **ICMTD-2005**, which is held from 21 to 24 of May 2005 on the Peninsula of Giens on the Mediterranean coast of France.

This conference, organized by the Provence Materials and Microelectronics Laboratory L2MP, provides an international forum for presentation and discussion on recent development on Memory Technology and Design. All aspects of memory devices, circuits, processing technologies, materials and other related research are within the scope of the conference. These three days of presentations, in oral and poster sessions, provide extensive opportunities for technical information exchange. Furthermore we expect that the conference setting and social events will stimulate informal communication among participants.

Sixty-six papers were selected for this first edition, 35 for oral presentation and 31 for poster presentation. In addition 7 invited speakers will present a large variety of subjects from DRAM cells to Polymer Memories:

- **Pierre FAZAN** - Innovative Silicon: DRAM cell trends
- **William GALLAGHER** - IBM: Recent advances in MRAM technology field-effect transistors
- **Johan HENDRICKX** - Philips Semiconductors: Embedded memory design: challenge or nightmare
- **Franz HOFMANN** - Infineon Technology: NVM based on FinFet device structures
- **Veli LIPPONEN** - Nokia: Memories: status and trends
- **Jean-Pierre NOZIERES** - SPINTEC: Thermo-magnetic MRAM
- **Agostino PIROVANO** - STMicroelectronics: Alternatives to Conventional Non-Volatile Memory

To complete these oral and poster presentations, we invite you to participate in:

- an evening lecture about "System-in-Package advances for future mainstream memory technologies" animated by **Lluís PARIS** from ATMOS Corporation.
- a panel discussion on "Does a universal memory make sense? " moderated by **Richard GORDON** from Gartner Dataquest.
- a round table to conclude the two sessions dedicated to MRAM memory.
- an evening lecture about "High-mobility solution-processed polymer ferroelectrics" given by **Ronald NABER** from Groningen University.
- a lecture, opening the gala dinner, given by **Prof. Bernard SOUMIREU** from the University of Provence on mechanisms of Human Memory

We are confident that this conference will be an exciting event for each attendee and we would like to thank all participants to the conference. Your presence demonstrates the large interest in the area of memory technology.

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 those who have contributed to the organisation and implementation of the **ICMTD-2005** conference.

We hope you will enjoy **ICMTD-2005** as much as we have enjoyed organizing it ,

the organization committee

Conference creation and organisation:

Rachid BOUCHAKOUR..... L2MP, France (Marseille)
Eric GERRITSEN *Philips Semiconductors*, France (Crolles)
Pascal MASSON L2MP, France (Marseille)
Pascale MAZOYER..... *STMicroelectronics*, France (Crolles)
Christophe MULLER..... L2MP, France (Toulon)

Communication / Diffusion:

Corinne JOACHIM *CREMSI*, France (Rousset)
Cathy PAITEL..... L2MP, France (Marseille)

SCIENTIFIC - TECHNICAL COMMITTEE

Atila ALVANDPOUR	Linköping University, Sweden (Linköping)
Roberto BEZ	STMicroelectronics, Italy (Agrate)
Rachid BOUCHAKOUR	L2MP, France (Marseille)
Eleonore DAEMEN	Atmel, France (Rousset)
Jean-Michel DAGA	Atmel, France (Rousset)
Barbara DESALVO	CEA/LETI, France (Grenoble)
Pierre FAZAN	EPFL, Switzerland (Lauzane)
Richard FERRANT	ALTIS Semiconductor, France (Corbeil Essonnes)
William J. GALLAGHER	IBM, USA (Yorktown Heights)
Cosimo GERARDI	STMicroelectronics, Italy (Catania)
Eric GERRITSEN	Philips Semiconductors, France (Crolles)
Gérard GHIBAUDO	INPG-IMEP, France (Grenoble)
Johan HENDRICKX	Philips Semiconductors, Netherlands (Eindhoven)
Franz HOFMANN	Infineon Technologies, Germany (Munich)
Giuseppe IANNACCONE	University of Pisa, Italy (Pisa)
Markku LESKELÄ	University of Helsinki, Finland (Helsinki)
Veli LIPPONEN	Nokia, Finland (Tampere)
Frans LIST	Philips Semiconductors, Netherlands (Eindhoven)
Salvatore LOMBARDO	IMN, Italy (Catania)
Michel MADORE	Atmel, France (Rousset)
Pascal MASSON	L2MP, France (Marseille)
Pascale MAZOYER	STMicroelectronics, France (Crolles)
Jean-Michel MIRABEL	STMicroelectronics, France (Rousset)
Christophe MULLER	L2MP, France (Toulon)
Jean-Pierre NOZIERES	SPINTEC, France (Grenoble)
Anthony PEAKER	University of Manchester, United Kingdom (Manchester)
Agostino PIROVANO	STMicroelectronics, Italy (Agrate)
Carole PLOSSU	INSA of Lyon-LPM, France (Lyon)
Erwin J. PRINZ	Freescale Technology, USA (Austin)
Didier SAVE	GEMPLUS, France (Gémenos)
Alain SCHUHL	LPM, France (Nancy)
Danny SHUM	Infineon Technologies, Germany (Dresden)
Michael SPECHT	Infineon Technologies, Germany (Munich)
Sandip TIWARI	Cornell University, USA (Ithaca)
Yeur-Luen TU	TSMC, Taiwan (Hsin-Chu)
Michiel VAN DUUREN	Philips research, Belgium (Leuven)
Jan VAN HOUDT	IMEC, Belgium (Leuven)
Dirk WOUTERS	IMEC, Belgium (Leuven)
Thierry YAO	AMI Semiconductor, Belgium (Oudenaarde)

TABLE OF CONTENT

PANEL DISCUSSION: Does a universal memory make sense ?	5
Moderator: R. Gordon	
Panellist: S. Okhonin, A. Pirovano, E. Prinz, J. Van Houdt	
EVENING LECTURE: "SIP Advances for Future Mainstream Memory Technologies"	7
L. Paris	
EVENING LECTURE: "High-mobility solution processed polymer ferroelectric field-effect transistors"	9
R.C.G. Naber, C. Tanase, P.W.M. Blom, G.H. Gelinck, A.W. Marsman, F.J. Touwslager, S. Setayesh, D.M. De Leeuw	
ROUND TABLE: "MRAM"	11
Moderators: N. Jumelet (<i>ConsulHtech</i>), T.J. Kroppenwnicki (<i>AMAT</i>), M. Madore (<i>Atmel</i>), K. Ounadjela (<i>Cypress</i>), J. Slaughter (<i>Freescale</i>)	
INVITED: "On Human Memory"	13
B. Soumireu	
INVITED: "Memories – Status & Trends"	N.A.
V. Lipponen	

SESSION A : Silicon nano-crystal based high density memory devices

A-1	INVITED : "Alternatives to conventional Non-Volatile memory: status and perspective"	17
	A. Pirovano & R. Bez	
A-2	Influence of Silicon Nanocrystal Size and Density on the performance of non-volatile memory arrays	21
	R.A. Rao, H.P. Gasquet, R.F. Steimle, G. Rinkenberger, S. Straub, R. Muralidhar, S.G.H. Anderson, J.A.Yater, J.C. Ledezma, J. Hamilton, B. Acred, C.T. Swift, B. Hradsky, J. Peschke, M. Sadd, E.J. Prinz, K.M. Chang, B.E. White Jr.	
A-3	Single-electron phenomena in ultra-scaled floating-gate devices and their impact on electrical characteristics	25
	D. Deleruyelle, G. Molas, B. De Salvo, M. Gely, D. Lafond	
A-4	Si nanocrystals by ultra-low-energy ion beam-synthesis for nonvolatile memory applications	29
	C. Bonafos, H. Coffin, S. Schamm, M. Carrada, N. Cherkashin, G.B. Assayag, P. Dimitrakakis, P. Normand, M. Perego, M. Fanciulli, Claverie	
A-5	Time-dependent simulation of the program and erase operations of nanocrystal Flash memories	33
	A. Campera, G. Iannaccone	
A-6	A model for the channel potential of charge-trapping memories and its implications for device scaling	37
	M. Sadd, S.A. Anderson, B. Hradsky, R. Muralidhar, E.J. Prinz, R. Rao, S. Straub, R.F. Steimle, C.T. Swift, B.E. White, J.A. Yater	
AP-1	In-line characterization of silicon nano-crystals grown on high-k tunnel dielectrics	N.A.
	E. Nolot, P. Mur, S. Favier, M.L. Villani	
AP-2	Discrete multilevel charging description of silicon nano-crystals memories	41
	C. Busseret, S. Ferraton, L. Montès, J. Zimmermann	
AP-3	On the saturation mechanism in the Ge nanocrystals based non volatile memory	45
	M. Kanoun, C. Busseret, T. Baron, A. Souifi	
AP-4	Long charge retention time in novel memory devices based on semiconductor quantum dots	49
	C. Balocco, S.W. Lin, M. Missous, A.M. Song	

SESSION B : DRAM, SRAM and Soft-Error-Rate

B-1	INVITED : "DRAM cell trends"	55
	P. Fazan	
B-2	1T-Bulk DRAM cell with improved performances: the way to scaling	59
	R. Ranica, A. Villaret, P. Malinge, P. Candelier, P. Masson, R. Bouchakour, P. Mazoyer, T. Skotnicki	
B-3	FinFET based zero-capacitor DRAM (Z-RAM) cell for sub 45 memory generations	63
	S. Okhonin, M. Nagoga, P. Fazan, L. Mathew, B.-Y. Nguen, H.-H. Chen, T. Stephens	
B-4	Evolution of stacked-capacitor technology for embedded DRAM applications	67
	N. Emonet, N. Jourdan, A. Berthelot, M. Piazza, D. Fraboulet, L. Gabette, C. Caillat, M. Charleux, B. Boeck, J. Regolini, D. Dutartre, V. Huard, F. Monsieur, P. Mazoyer, E. Gerritsen	
B-5	INVITED: "Embedded memory design : challenge or nightmare"	71
	J.Hendrickx	
B-6	A small granular controlled leakage reduction system for SRAMs	73
	P. Geens, W. Dehaene	
B-7	Impact of process variations on the alpha-particle-induced SER of embedded SRAMs	77
	T. Heijmen, B. Kruseman	
B-8	High density SRAM robust to radiation-induced soft errors in 90 nm CMOS technology	81
	P. Roche, F. Jacquet, G. Gasiot, C. Caillat, B. Borot, J.P. Schoellkopf	
BP-1	A model to explain the C-V curve of DRAM capacitors with silicon electrodes and trapping dielectrics	85
	L. Lopez, P. Masson, D. Née, R. Bouchakour	
BP-2	Timed verification of the SPSMALL memory	89
	M. Baclet, R. Chevallier	
BP-3	Device simulation study of SEU in SRAMs based on double-gate MOSFETs	93
	K. Castellani-Coulié, D. Munteanu, J.-L. Autran, V. Ferlet-Cavrois, P. Paillet, P. Masson	
BP-4	A system-level approach for memory robustness	97
	R. Mariani, G. Boschi, A. Ricca	

SESSION C : MRAM 1/2

C-1	INVITED: "Recent advances in MRAM technology"	103
	B. Gallagher	
C-2	Toggle MRAM performance and reliability	105
	S. Tehrani	
C-3	Low power MRAM design based on new switching mechanisms	107
	H. Yoda, T. Kai, M. Nakayama, S. Ikegawa, Y. Asao, H. Hada, N. Ishiwata, S. Tahara	
C-4	Development strategy for writing margin of high density MRAM	109
	J. E. Lee, S. C. Oh, H.-J. Kim, Y. K. Ha, J. S. Bae, K. T. Nam, S. O. Park, H. S. Kim, U-In Chung and J. T. Moon	
C-5	State of the art in Magnetic Random Access Memory	111
	K. Ounadjela	
C-6	INVITED: "Thermo assisted MRAM for low power applications"	113
	O. Redon, M. Kerekes, R. Sousa, L. Prejbeanu, H. Sibuet, F. Ponthennier, A. Persico, J.P. Nozières	

SESSION D : MRAM 2/2

D-1	Switching speed in high density MRAM based on precessional or spin-transfer switching	117
	T. Devolder, C. Mauroury, H.W. Schumacher, C. Chappert, C. Maufroid, J.-V. Kim, R. Fournel, Y. Suzuki, B., C., K. Yagami . Tulapurkar	
D-2	Switching issues for high density MRAM	N.A.
	T. Kim	
D-3	Giant Tunneling Magnetoresistance and Tunneling Spin Polarization in Magnetic Tunnel Junction Memory Cells using MgO Tunnel Barriers	121
	S. Parkin	
D-4	Preparation of TMR multiplayer for MRAM production	123
	W. Maass, B. Ocker, J. Langer, Y. Huai	

D-5	Current status and issues for the patterning of MRAM devices	127
	T.J. Kropenwnicki, R. Patz, J. Stokes, A. Paterson, T. Panagopoulos, J. Holland	
DP-1	Non-contact measurement of the magnetic properties of MRAM wafers	129
	W. VanDrent, F. Vajda	
DP-2	Spin-wave dynamics of a square magnetic bit in remanence	131
	S.O. Demokritov, V.E. Demidov, B. Hillebrands, Ch. Back	
DP-3	Magnetic tunnel junction device model for MRAM and hybrid magnetic-CMOS components design	135
	V. Javerliac, F. Bernard-Granger, Li. Prejbeanu	
DP-4	Brillouin light scattering observation of the damon-eshbach mode in a nano-granular (SiO₂)Co/GaAs film	137
	A.A. Stashkevich, Y. Roussigné, G. Viau, N.A. Sergeeva, M.P. Kostylev, L.V. Lutsev, A.I. Stognij, N.N. Novitskii	
DP-5	Dynamic heating studies in thermo-magnetic MRAM junctions	139
	M. Kerekes, R. Sousa, O. Redon, L. Prejbeanu, U. Ebels, P. Xavier, B. Dieny, J.P. Nozière	
DP-6	Thermal and non-thermal ultrafast magnetization control new avenues in magnetic memory?	141
	F. Hansteen, A.V. Kimel, A. Kirilyuk, D. Stanciu, R.V. Pisarev, Th. Rasing	
DP-7	Magnetic Random Access Memories integrated in silicon (the MEMIS project)	145
	C. Duluard, A. Bsiesy, V. Safarov, A. Filipe, A. Francinelli, H. Achard	

SESSION E : NV Innovative architectures and materials

E-1	INVITED : "NVM based on FinFet device structures"	149
	<u>E. Hofmann</u> , M. Specht, U. Dorda, R. Kömmling, L. Dreeskornfeld, J. Kretz, M. Staedele, W. Rösner, L. Risch	
E-2	Modeling of a double-gate FinFlash memory	153
	S. Jacob, B. De Salvo, G. Le Carval, S. Deleonibus	
E-3	3D simulation study of gate and noise coupling in advanced floating gate non volatile memories	157
	A. Ghetti, L. Bortesi, A. Mauri, L. Vendrame	
E-4	Cost-competitive embedded non-volatile technology for Flash and EEPROM applications	161
	D. Dormans, D. Boter, J. Dubois, J. Garbe	
E-5	Composition control and ferroelectric properties of sidewall Sr_{0.8}Bi_{2.2}Ta₂O₉ in integrated 3-dimensional ferroelectric capacitors	163
	L. Goux, J.G. Lisoni, M. Schwitters, V. Paraschiv, D. Maes, L. Haspeslagh, D.J. Wouters, N. Menou, Ch. Turquat, V. Madigou, C. Muller, R. Zambrano	
E-6	Technological advances for memory applications : crystalline high-K gate dielectrics and alternatively doped gates	167
	Y. Stefanov, R. Komaragiri, U. Schwalke	
E-7	A two-terminal semiconductor memory made in a single nanolithography step	171
	A.M. Song, M. Missous, W. Seifert	
E-8	Development of atomic layer deposition processes for multicomponent oxide dielectrics	175
	M. Vehkamäki, M. Ritala, M. Leskelä	
EP-1	A macro model of programmable metallization cell devices	177
	N.E. Gilbert, C. Gopalan, M. N. Kozicki	
EP-2	CuTCNQ based organic non-volatile memories: downscaling, stress tests, and temperature effect on I-V curves	181
	R. Müller, J. Genoe, P. Heremans	
EP-3	Organic electrically bistable materials for non-volatile memory applications	185
	A. Pirovano, S. Conoci, R. Sotgiu, S. Petralia, F. Buonocore	
EP-4	Growth at low temperature of very high-K material on silicon	189
	M. Detalle, E. Fribourg-Blanc, E. Cattani, D. Remiens	
EP-5	Analysis of US patents (1966-Novembre 2004) in phase charge memory	191
	S. D. Savransky, T.C. Wei	
EP-6	Effects of threshold switching and parasitic capacitance in the programming transient of chalcogenide phase-change memories	195
	D. Ielmini, D. Mantegazza, A.L. Lacaita, A. Pirovano, F. Pellizzer	
EP-7	Economics of Alternative Nonvolatile-Memories	199
	R. Andrei	

SESSION F : *Memories and high- κ dielectrics*

F-1	Use of W damascene as true metal1 in a 130 nm flash technology S. Louwers	203
F-2	Investigation of SiO₂/HfO₂ gate stacks for application to non-volatile memory devices J. Buckley, B. De Salvo, G. Ghibaudo, M. Gely, J.F. Damlencourt, F. Martin, G. Nicotra, S. Deleonibus	207
F-3	Scaling down the interpoly dielectric for next generation flash memory : challenges and opportunities B. Govoreanu, D. Brunco, J. Van Houdt	211
F-4	ISSG RTO (In-situ Steam Generation Rapid Thermal Oxidation) grown tunnel oxide on advanced FLOTOX E²PROM memories G. Festes, S. Angle, J.-M. Bédérine, M. Berrebi, M. Cerisier, E. Daemen, A. Dumas, M. Grégoire, E. Hugonnard, P. Rohr, E. Serres, A. Solere	215
F-5	SONOS flash memories with HfO₂ or HfSiON R. Van Schaijk, M. Van Duuren, F. Neuilly, W. Baks, A. H. Miranda, M. Slotboom, N. Akil, P.G. Tello	219
F-6	Fully compatible novel SNONOS structure for improved electrical performance in NAND flash memories J.H. Han, J.K. Kim, S.H. Lee, S. Jeon, C. Kim	223
F-7	Scaling effects in dual-bit split-gate nitride memory devices L. Breuil, L. Haspeslagh, M. Lorenzini, J. De Vos, J. Van houdt	227
F-8	Low voltage and low power embedded 2T-SONOS flash memories M. Slotboom, D. Dormans, M. Van Duuren, R. Van Schaijk, N. Akil, R. Beurze	231
FP-1	Retention reliability improvement using nitride with varying trap density in SONOS type non-volatile memory J.H. Kim, J.H. Han, Y.S. Min, M.K. Kim, Y.S. Jeong, S. Jeon, J.W. Hyun, S.H. Lee, H.S. Chae, S.D. Chae, C. Kim	235
FP-2	Flash memory with high-k tunnel dielectrics : comparative retention study P. Blomme, A. Akheyar, J. Van Houdt, K. De Meyer	239
FP-3	A 4 MegaBit, 128-bit data embeded flash memory allowing high-speed in place code execution M. Combe, J.-M. Daga	243
FP-4	Lifetime model and procedure for failure rate prediction of moving bits in case of room temperature retention T. Yao, D. Medjahed, P. Gassot, A. Lowe, P. Cacharelis	247
FP-5	EEPROM retention time extrapolation from floating gate SILC measurements S. Burignat, N. Baboux, C. Plossu, P. Boivin	251
FP-6	A new reliable P/E methodology for fully fowler-nordheim operating non volatile memories F. Irrera, G. Puzzilli	255
FP-7	STI engineering for high program & erase performance of STI-bounded nitride storage flash memory cells M. Isler, J.-M. Schley, S. Riedel, T. Mikolajick, C. Ludwig, K.-H. Kuesters, G. Tempel, J.-U. Sachse, P. Deconinck, R. Mikalo, R. Reichelt, N. Schulse, E. Stein v. Kamienski, M. Strassburg, J. Willer, F. Lau, R. Hagenbeck, P. Haibach	259
FP-8	Low temperature postmetallisation annealing of high-k dielectrics Y.F. Loo, S. Taylor, P. Taechakumput, A.C. Jones, P.R. Chalker, L.M. Smith	263
FP-9	EEPROM failure analysis methodology : can programmed charges be measured directly ? C. De Nardi, R. Desplats, P. Perdu, F. Beaudoin, J.-L. Gauffier <i>Laboratoire d'Expertise commun CNES et TES (France), LNMO (France)</i>	267
FP-10	Flash memory cell : an automated diagnosis tool for geometric failures B. SAILLET, J.-M. Portal, D. Née	271
FP-11	A novel embedded OTP NVM using standar foundry CMOS logic technology B.M. Fliesler, G. Rosendale, J. Wang, C. Ng, Z.S. Liu, J. Peng	275

Panel discussion: Does a universal memory make sense?

In this panel, we will cover the viability of a universal memory technology. Among others, the panelists will answer the following questions: Does it make sense to have such a universal memory at the technology level and at the system level? If it does make sense, is it realistic to have technologies ready for such memories, and if yes, by when? Among others, are new materials needed to realize such a universal storage medium and what are those materials? What are the manufacturing challenges expected to integrate those materials? Finally what are the leading memory technologies that could be a candidate?

To cover those questions and this provoking topic, this panel will be moderated by **Richard Gordon**, research VP at Gartner Dataquest. The following panelists will share their view and answer the questions of the audience:

A. Pirovano ST Microelectronics (Italy)
E. Prinz Freescale (USA)
J. Van Houdt IMEC (Belgium)
S. Okhonin Innovative Silicon (Switzerland)
S. Sivaram Matrix Semiconductor

SIP Advances for Future Mainstream Memory Technologies

L. Paris

Emerging Memory Technologies Inc., 349 Terry Fox Drive, Kanata, ON K2K 2V6, CANADA
email: lparis@emergingmemory.com

The semiconductor industry is at crossroads in dealing with technological scaling, SOC compatibility and deciding between multiple new memory technologies. Most new memory technologies being researched in the market requires that they be SOC compatible., but will SOC really last that long given its disadvantages and competition against SIP. SIP (system

in a package) is gaining momentum over the last couple of years with its reduced cost and design-reuse making it a viable option in future. This talk will outline the SIP trends compared to SOC and will address the emerging memory technological issues addressing the package(embedded) compatibility.

High-mobility solution-processed polymer ferroelectric field-effect transistors

R.C.G. Naber^a, B. de Boer^a, P.W.M. Blom^a and D.M. de Leeuw^b

^a Materials Science Centre, University of Groningen, Nijenborgh 4, 9747 AG Groningen, The Netherlands, r.c.g.naber@rug.nl

^b Philips Research Centre, Professor Holstlaan 4, 5656 AA Eindhoven, The Netherlands

Abstract

We demonstrate the successful integration of a high-mobility solution-processed organic semiconductor into a polymer ferroelectric field-effect transistor (FeFET). The increased mobility facilitates a low-voltage read-out operation because the same drain current response can be obtained for lower drain voltages. Read-out with low drain voltages is required for ferroelectric field-effect transistors with very thin ferroelectric insulator thicknesses because higher voltages can affect the polarization state of the ferroelectric.

1. Introduction

Organic bistable devices are currently in the focus point of much research effort because of their deemed application as a low-cost memory solution [1-6]. This cost reduction is largely obtained from the fact that the active materials can be processed from solution, which makes it suitable for simple, large-scale production methods. We recently demonstrated a solution-processed polymer FeFET that has a memory functionality due to the attenuation of the semiconductor channel charge density by the ferroelectric polarization of the gate dielectric [7]. A schematic cross-section of the prepared devices is shown in Fig. 1. In order to further improve the performance of this device architecture, one important parameter is the choice of the semiconductor material. We here report on the use of solution-processed regioregular poly(3-hexylthiophene) (rr-P3HT) as opposed to the use of a poly(*p*-phenylene vinylene) (PPV) derivative in our previous report. While the latter is an amorphous material, P3HT has been shown to crystallize and form a high mobility phase due to the enhanced molecular orbital overlap in the crystalline phase [8]. The increased mobility can have a number of advantages such as a reduced programming time [9], but the present motivation is to enable a drain voltage reduction in transfer curve measurements on FeFETs with very thin ferroelectric insulator layers.

2. Results and discussion

Fig. 2 shows the transfer curve of an FeFET with unannealed P3HT. The curve is qualitatively similar to our previous results with PPV, as observed by the high on-state drain current and a similar gate field where switching occurs. The on-state channel conductance

$$G = \frac{L \cdot I_D}{W \cdot V_D} \quad (1)$$

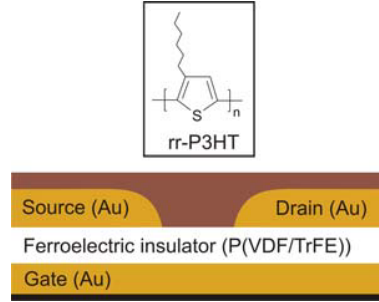


Fig. 1: Schematic cross-section of the polymer ferroelectric field-effect transistor and the chemical formula of poly(3-hexylthiophene) that is used for the organic semiconductor top layer.

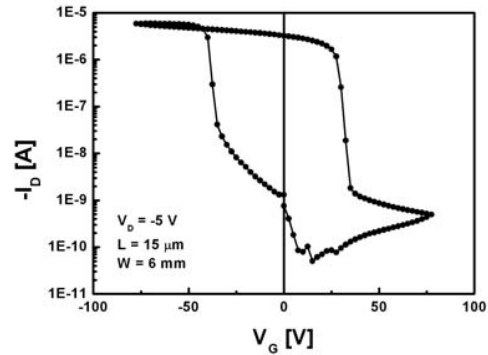


Fig. 2: FeFET with regioregular P3HT, without an annealing step for the semiconductor. The ferroelectric insulator layer thickness was 0.85 μ m.

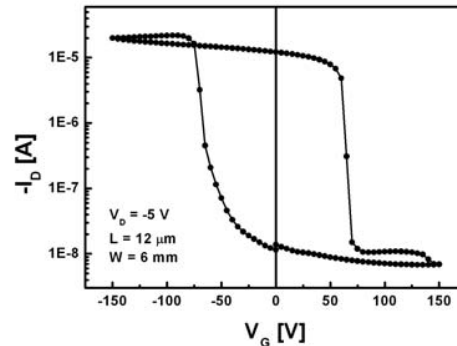


Fig. 3: FeFET with regioregular P3HT, with a semiconductor annealing step at 138°C. The ferroelectric insulator layer thickness was 1.7 μ m.

at zero gate bias is $1.6 \cdot 10^{-9}$ S, which is 2 times larger than the conductance of $8 \cdot 10^{-10}$ S, as previously obtained with PPV (see Fig. 4 in [6]). The on/off ratio of 2500 is large, but it is lower than the 10^4 ratio value obtained with PPV, due to the relatively high off-state conductance. We attribute this off-state conductance to the presence of a switch-on voltage V_{so} that is observed in many organic semiconductors [10].

Fig. 3 shows the transfer curve for an annealed P3HT layer. The annealing temperature was chosen to be the highest possible. For this device, it is limited by the ferroelectric polymer melting temperature of $\sim 155^\circ\text{C}$ and therefore the annealing temperature was chosen to be slightly below this value, at 138°C .

Comparison between Figs. 2 and 3 shows no qualitative differences, as observed by the similar on/off ratio and the gate insulator field where switching occurs. The on-state channel conductance at zero gate bias is higher however, by a factor of 3. This is consistent with the expected mobility increase due to annealing-induced crystallization. Compared to the results with PPV, the overall on-state conductance is increased by a factor of 6.

It may be possible to further increase the on-state conductance by optimizing the processing of the P3HT [8], but the current results already allow for a significant drain voltage decrease, whilst retaining the same drain current signal.

3. Conclusion

We demonstrated the successful integration of regioregular poly(3-hexylthiophene) into a polymer ferroelectric field-effect transistor. The obtained mobility increase facilitates a low-voltage read-out operation.

Experimental

Bottom-gate transistors with gold electrodes were prepared upon clean glass substrates. The bottom electrodes were applied by shadow mask evaporation of a thin chromium adhesion layer and gold. Subsequently, poly(vinylidene fluoride/ trifluoroethylene) 65/35 mol% random copolymer (Solvay Duphar, Belgium) was applied by spin casting from filtered 2-butanone solutions. These ferroelectric polymer films were annealed in a vacuum oven at 138°C to enhance the crystallinity. Gold source-drain electrodes were

deposited with shadow mask evaporation, using a wire to create the channel. Regioregular poly(3-hexylthiophene) (rr-P3HT, electronic grade; 98.5 % regioregular; Rieke Metals, Inc.) was purified after purchase by the following procedure. It was dissolved in distilled toluene, dedoped with hydrazine at 60°C and precipitated in methanol. The fraction collected was Soxhlet extracted for at least 64 hrs with methanol, *n*-hexane, CH_2Cl_2 and finally with CHCl_3 . The chloroform fraction was precipitated in methanol, dried under vacuum and stored in the glovebox under N_2 atmosphere. The purified polymer semiconductor was dissolved in chloroform and spin cast onto the FeFET layer stack in a N_2 filled glovebox. Some of the devices were annealed again in a vacuum oven at 138°C to raise the crystallinity of the P3HT. Field-effect measurements on transistors were performed in dark and vacuum using a Keithley 4200 semiconductor analyzer. Polymer layer thicknesses were determined using a Dektak profilometer.

Acknowledgement

We acknowledge financial support by the Dutch Science Foundation NWO/FOM.

References

- [1] S. Möller, C. Perlov, W. Jackson, C. Taussig and S.R. Forrest, *Nature* **426**, 166 (2003).
- [2] A.W. Marsman, C.M. Hart, G.H. Gelinck, T.C.T. Geuns and D.M. de Leeuw, *J. Mater. Res.* **19**, 2057 (2004).
- [3] R. Schroeder, L.A. Majewski, M. Voigt and M. Grell, *IEEE Electr. Dev. Lett.* **26**, 69 (2005).
- [4] Y. Matsuo, T. Ijichi, H. Yamada, J. Hatori and S. Ikehata, *Centr. Eur. J. Phys.* **2**, 357 (2004).
- [5] T.B. Singh, N. Marjanović, G.J. Matt, N.S. Sariciftci, R. Schwödiauer and S. Bauer, *Appl. Phys. Lett.* **85**, 5409 (2004).
- [6] J. Ouyang, C. Chu, C.R. Szmanda, L. Ma and Y. Yang, *Nature Mater.* **3**, 918 (2004).
- [7] R.C.G. Naber, C. Tanase, P.W.M. Blom, G.H. Gelinck, A.W. Marsman, F.J. Touwslager, S. Setayesh and D.M. de Leeuw, *Nature Mater.* **4**, 243 (2005).
- [8] Z. Bao, A. Dodabalapur and A.J. Lovinger, *Appl. Phys. Lett.* **69**, 4108 (1996).
- [9] M.W.J. Prins, S.E. Zinnemers, J.F.M. Cillessen and J.B. Giesbers, *Appl. Phys. Lett.* **70**, 458 (1997).
- [10] E.J. Meijer, C. Tanase, P.W.M. Blom, E. van Veenendaal, B.H. Huisman, D.M. de Leeuw and T.M. Klapwijk, *Appl. Phys. Lett.* **80**, 3838 (2002).

Round table: MRAM

Recent developments in the physics of spin electronics have enabled the emergence of a new class of non-volatile memories, Magnetic Random Access Memories (MRAM). MRAM combine on paper all the virtues of current silicon-based memories: speed, density, non-volatility, low power consumption, radiation hardness and endurance. MRAM is therefore viewed as one, if not the only, potential candidate for the long awaited “universal memory”.

There is however still a long road to go and the first targeted applications are those of embedded Flash, which, despite being non-volatile, have a high power consumption, are sluggish at write and are prone to aging effects. Major Integrated Circuit manufacturers have started extensive research and development efforts and expect to launch the first product by the end of the year.

Despite all its promises, MRAM is still in infancy and large R&D efforts are still required both at the academic and industrial levels. It is the scope of this conference sessions (C & D) to gather all the players in the field on order to exchange on the state of the art and identify the remaining key challenges for industrialization. This round table will conclude the two sessions dedicated to this memory.

Moderators:

N. Jumelet	ConsulHtech
T.J. Kropenwnicki	AMAT
M. Madore	Atmel
K. Ounadjela	Cypress
J. Slaughter	Freescall

On human memory

Bernard SOUMIREU-MOURAT

University of Provence / CNRS UMR 6149

Memory in mammals – especially in humans – can be defined as processing and storing information across time. Two points are essential, a content in terms of information, and elapsing time. As already told by ARISTOTE, Memory can be modified through the time. But also, human Memory is quite ambiguous, as most of the highest cognitive functions, and RIBOT wrote one century ago that « La Mémoire est par essence un fait biologique, et par accident un fait psychologique ».

The general view for the neurobiological basis of memory is the one developed by Donald HEBB 60 years ago. Memory traces would be supported by specified neuronal networks ;

it means that a specific set of neurons- and specified synapses on these neurons – would correspond to each memory trace. It must be told that evidently, a neuron and even a specified synapse of this neuron might be involved in a lot of different networks.

The specification of the network would result from prolonged activation of the network during the acquisition of information, and retrieval would result from activation of this same network. HEBB made the hypothesis that synapses were the locus of the modifications allowing to identify the network. Years later, a mechanism was identified by BLISS and LOMO – the LTP or long-term potentiation – and it is presently the best candidate for specifying the synapses on the network during memory processes.

LTP is a long-term facilitation of synaptic efficiency and is characterized by its 3 properties : specificity, associativity, and cooperativity.

In fact, human memory is complex and heterogeneous. At the end of the nineteenth century, William JAMES made the first distinction defining short-term memory (STM) and long-term memory (LTM). Presently several types of (long-term) memory

are distinguished; a largely admitted classification is the one synthetized by Larry SQUIRE, with mainly declarative memory and non-declarative memory (including procedural).

According to the type of memory concerned, different cerebral structures seem to be mainly involved, but, considering human pathology (amnesia) and experimental data (in rodents and monkeys), one structure has been especially studied - the hippocampus. Hippocampus is part of the medial temporal lobe, and part of the Papez's circuit (or HMT circuit). This large loop puts in relation the hippocampus and the neocortex (where is the half of the whole neurons), so that it could work as a comparator of information, with specialized cells reacting when the stimuli are either new or old (already known).

At the same time, through relations between hippocampus and several subcortical structures as Amygdala, Hypothalamus and Reticular Formation, the Hippocampus would evaluate the on-going information as significant or not in terms of emotions, motivation, ...

When the arriving information is new and significant, the hippocampus would initiate the specification of the neuronal network by inducing LTP on the concerned synapses of the activated network, so that this network would be durably marked. Durably means it would last as long as the memory trace, which may be modified across time, depending on new events or changes in the organism or the surroundings of the subject, but which also may be forgotten.

This leads to a lot of interactions between Memory mechanisms *per se* and other functional aspects of the brain, which may modulate Memory.

SESSION A

Silicon nano-crystal based high density memory devices

Chairmen: **Erwin J. PRINZ**, Freescale Technology, USA (Austin)
 Jan VAN HOUDT, IMEC, Belgium (Leuven)

Alternatives to Conventional Non-Volatile Memory: Status and Perspective

Agostino Pirovano and Roberto Bez

STMicroelectronics, NVMTD, via C. Olivetti 2, 20041 Agrate Brianza (MI), Italy

Abstract

For more than ten years Flash memory technology has been able to follow the evolution of the semiconductor roadmap, becoming the non-volatile memory (NVM) market mainstream. Although it is expected that Flash will continue to scale with the same trend for next technology nodes, there are physical limitations to be faced and the downscaling beyond 45 nm generation is still considered critical. Alternative concepts are thus being developed. Aim of this paper is to present the potentialities of the most promising emerging NVMs and the issues that they must still face to become competitive with Flash and succeed in transforming their superior performance into a sizeable market share.

Keywords: Non-volatile memories, Flash, FeRAM, MRAM, Phase-change memories, probe storage.

1. Introduction

Looking back to the evolution of the silicon era, the demand for memory devices capable to support the CMOS logic has been one of the most striking issues for the semiconductor industries. However, the ideal storage system has never been discovered. Several concepts and technologies have been thus exploited to patch this shortcoming, none of them able to satisfy all the requirements at the same time. The development of memory technologies has been thus driven by the growth of applications demanding a specific memory concept that could better satisfy their requirements. As a proof of this, the most relevant phenomenon of this past decade in the field of semiconductor memories has been the explosive growth of the non-volatile memory (NVM) market, driven by cellular phones and other types of

electronic portable equipment and, in recent years, the explosive growth of the Flash NAND market due to the increasing demand for data storage capabilities.

The demand for a NVM product is thus strictly related to the appearance of a driving application that can exploit its potentialities. However, the acquisition of a leading competitive position based on performance advantages cannot be sustained without the capability of memory devices to scale down and reduce their costs. In this view, despite it is expected that in the coming years portable systems will continue to demand even more NVMs, the exploitation of these opportunities is mainly related to the Flash capability to scale more and more. But, although logic devices will continue their evolution simply scaling the original planar MOS concept at least down to the 22 nm technology node [1], existing NVM devices will soon face fundamental limits to shrink their dimensions. In this scenario, emerging NVM concept can try to exploit their performance advantages and better scaling capabilities to enter in the actual Flash-centric memory market and eventually become the leading NVM technology. Aim of this paper is to discuss the status and the perspective for Flash technology to be the NVM mainstream in the next decade continuing its scaling-based cost-reduction path, and how emerging concept and novel memory technologies could off-set the current market creating new opportunities.

2. Flash technology perspectives

Recent technological developments have definitely confirmed the division of the NVM market into two mainstreams based on different applications of the Flash technology concept.

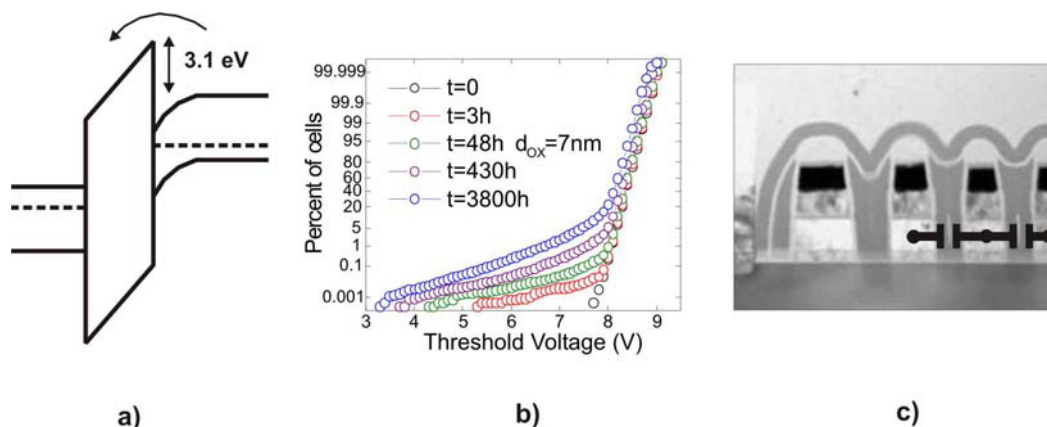


Fig. 1: Schematic description of scaling limits for Flash technology. The high barrier at the Si/SiO₂ limits the voltage scaling (a), while the oxide thickness reduction must face the reliability constraints (b). Finally, capacitive coupling between adjacent cell cannot be neglected in scaled devices (c).

By exploiting its multi-level capabilities (MLC) and the fast random access for execution in place applications, NOR memories are a cost effective solution to stack code and data storage into a single chip. With a cell size close to the minimum litho defined feature of $4F^2$ and MLC capabilities, NAND technology is expected to be the lowest cost NVM and the best solution for high-density data storage for the rest of this decade.

Throughout the history of the development of Flash, the cell size has been reduced through a combination of lithography reduction and self-alignment techniques [2]. This trend will continue in next-generation devices [3], with an increasing importance of self-alignment techniques to compensate the difficulties to scale down the lithographic registration dimensions. But, as Flash scales into the sub 100nm regime, challenges arise from physical limitations related to the high voltages needed by the programming and erase mechanisms, the stringent charge storage requirements of the dielectrics, and the capacitive coupling between adjacent cells (Fig.1). The difficulties to scale down voltages for Flash programming and erasing is also a problem for the development of System-on-Chip architecture and embedded high performance portable systems, where the voltage and current requirements to achieve fast program/erase operations does not allow an easier integration in standard CMOS logic and to meet low-power specifications. These issues will become critical to scale down the cell size below the 65nm technology node [3] and they must be faced with more aggressive solutions than in the past, with innovations in the cell structure, dielectric materials, and overall system architecture management.

FinFET and three-dimensional (3D) cell structures are possible ways to address the gate length scaling constraint [4]. This structure moves the channel length constraint into the vertical direction, allowing further area scaling to occur. Moreover, lower channel doping concentrations and higher driving current capabilities are expected for 3D devices, resulting into an improvement of drain disturbance immunity and better performance.

High- k dielectric as well as barrier engineered composite films promise to allow the design of a more optimal barrier structure, to reduce the oxide thickness and the energy barrier height. Another valuable solution to overcome Flash memory scaling limitations is based on the “discrete-trap” mechanism, where the floating gate is replaced by many discrete trapping centers, which can be made by natural traps in an appropriate insulator (a nitride layer in SONOS memories) or by semiconductor (usually silicon) or metallic nano-crystals [5,6]. However, while oxide damage and resultant charge loss can be an intrinsic effect, the primary source of reliability consideration for Flash memories is extrinsic (a few bits out of millions) charge loss or gain [7], and the introduction of novel dielectric material will require a deep investigation of their extrinsic behaviors. Moreover, although Flash technology has provided a valuable solution for several application for portable and

mobile equipments, there are intrinsic limits in the performance achievable and that would be likely removed. In this scenario, floating-gate memories based on different dielectric could only slightly improved these limitations, thus providing a solution for scaling without a significant improvement of performance.

With the scaling of the NVM devices, particularly beyond the 65 nm technology node, all the reliability issues could become more critical. From this point of view, design of fault-tolerant architectures and system management techniques will help to continue the scaling path. Proper control of the write and erase operations with charge placement algorithms can reduce the damage of the tunnel dielectric by reducing the applied fields and controlling how the fields are ramped up and down during write and erase; error management techniques, such as error correction, can recover data or prevent errors; capacitive coupling effects could be compensated through how adjacent bits are written and read so that data stored in one cell is accounted for in the reading or writing of an adjacent cell. Despite these solutions can impact the overall performance, looking back to the historical development of floating-gate memories (e.g., multi-level capabilities is an easy way to increase the storage density without adding technological steps), system management appears to be the easier way to continue the cost reduction trend for Flash memories. It is thus expected that Flash scaling will progress at the current rate through at least the end of the decade with the 45nm generation using techniques that are available today or projected to be available in the near future without twisting the current Flash concept.

3. Emerging NVM technologies

Flash technology is characterized by a compact structure in which the selecting-element and the storage-element are merged in an MOS-like architecture. The resulting full compatibility with the CMOS technology and the compact device size have made Flash technology the cheaper solution for stand-alone and embedded memory applications. However, discordant requirements to shrink the MOS structure preserving good selecting and storage capabilities are making Flash scaling more and more difficult. A possibility for enhancing the scalability of a memory architecture is to decouple the selector from the storage element, thus allowing to follow the already difficult CMOS scaling without additional constrains, and move the memory element in the back-end part of the process integration, thus maintaining an acceptable cell size and simplifying its embeddability. Moreover, even if in the long term the cost advantage is important, better performance can speed-up a novel technology introduction. In fact, a NVM with low power and low voltage capabilities, bit granularity, fast operations and higher endurance would be a potentially game-changer for system designers.

	Flash		FeRAM	MRAM	PCM	Probe Storage
Cell type	NOR 1T	NAND 1T	1T/1C	1T/1R	1T/1R	AFM based
Cell size (F ²)	10	5	30-100	30-50	8-16	0.4 (no litho)
Endurance write/read	10 ⁶ /∞		10 ¹² /10 ¹²	>10 ¹⁴ /∞	10 ¹² /∞	10 ⁵ -10 ¹² / 10 ⁷ -∞
Read time (random)	60ns	60ns/serial	40+80ns (read + write destructive read)	30ns	60ns	2-20ms (access time)
Write time (byte)	1μs	200μs/page		30ns	10ns	0.1-1 ms for each tip <1μs/bit
Erase time (byte)	1s/sector	2ms/block		30ns	150ns	
Scalability	Fair		Poor	Poor	Good	Very good
Scalability limits	Tunnel oxide, HV		Capacitor	Current Density	Litho	None
Multi-bit capability	Yes		No	No	Yes	No
Relative Cost x bit	Medium	Low	High	High	Medium	Very low
Maturity	Very high		Medium	Low	Low	Very low

Tab.1: Comparison between the expected properties of the emerging NVM alternatives.

Currently, a total of more than 30 NVM technologies and technology variations compete for a piece of the fast growing NVM market, many of them aiming to replace DRAMs. However, long term survival will be ensured only by those who will succeed in transforming technological superiority into a sizeable market share. Among this alternatives, the most promising NVM concepts being pursued are Ferroelectric memories (FeRAM), which exploits the permanent polarization of a ferroelectric material, like Pb(Zr,Ti)O₃ (PZT) or SrBi₂Ta₂O₉ (SBT), with a cell composed by one transistor and one capacitor, Magnetoresistive memories (MRAM), which uses the electrical resistance change in a magnetic tunnel junction to indicate the memory state, and Phase Change Memories (PCM), based on an electro-thermally induced phase-change transition in a chalcogenide alloy (e.g., Ge₂Sb₂Te₅). A comparison between the expected performance for each technology is reported in Tab.1. Though all these emerging NVM competitors have superior performance with respect to Flash technology, all of them have still to demonstrate high volume manufacturability on the leading edge technology node. Once demonstrated it, only the technology able to scale beyond Flash will be successful in becoming the NVM mainstream.

FeRAM is the only alternative NVM that has been commercialized so far, even if in a relaxed technology at very low density. However, several challenging technology problems, mainly related to new materials and new manufacturing technologies, are still limiting their diffusion. An impeding factor for FeRAM scaling is the decreasing of the sensing signal with the capacitor area shrinking [8]. Their scaling will require therefore technologies to at least maintain or increase the capacitor area, probably following the approach experienced in DRAM evolution, i.e., three dimensional capacitors. This requirement would make harder the already critical compatibility of the ferroelectric film characteristics with the CMOS process. Moreover, recent trends [9,10] suggest that FeRAMs could be a valuable solution only for high-performance small-density memories in embedded applications, where their fast operation capabilities can compensate their large area occupation.

MRAM has the potential to be competitive with all existing semiconductor memories and, through its

unique properties, provide new functionalities. The key attributes of MRAM technology are high-speed operations combined with an effectively unlimited read-write endurance [11]. However, MRAM development has faced many challenges that impact the manufacturability of high-density devices. Moreover, although its far superior performance with respect to any existing or forecasted memory technology, the major MRAM disadvantage is the high write current. While this technology has enough read current to guarantee a fast access time, it requires a very large write current (mA range), which increase power consumption. The write current is expected to increase with MRAM device shrinking, thus posing a fundamental scalability limitation for this technology, arising in unsustainable power consumption and reliability issues.

The Phase-Change Memory technology is one of the best candidate to cover several NVM application fields, matching both the high density as well as the high performance specifications. Despite this technology dates back to the '70s, its application for NVM has known renewed efforts in these years [12], showing the capabilities to reach the maturity for manufacturability [13,14]. Actually, there are no known physical limitations to scale down PCM to CMOS limit. The major concern for their perspectives comes from the large current consumption for programming operation.

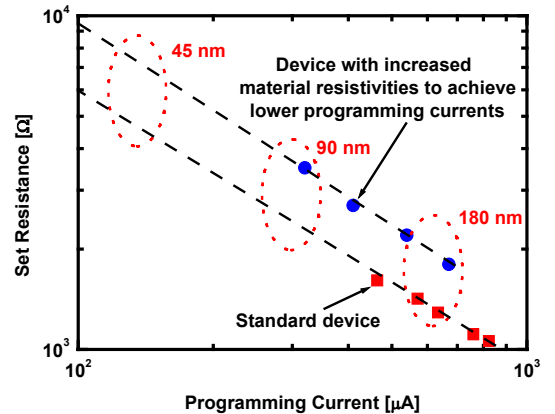


Fig. 2: PCM scaling trends for the programming current and the set resistance. Note that also at the 45 nm node a set resistance lower than 10kΩ is expected.

This current is expected to scale at least linearly with the technology node, but more aggressive scenarios have been forecasted [15]. Another concern is related to the increasing of the minimum programmable resistance by decreasing the device dimensions, thus reducing the read current and correspondingly the fast sensing capability. Fig.2 reports some experimental data and the expected scaling trends for the programming current and the corresponding set resistance. Moreover, despite the minimum cell dimension so far reported is about $12F^2$, thus being comparable to NOR Flash size, the multi-level capabilities intrinsic in the resistance variation of phase-change materials could be exploited to become cost-competitive for data storage applications.

As reported in Tab.1, these emerging NVMs are characterized by larger cell size with respect to the NAND Flash technology, thus being more expensive. A novel technology that could provide a cost-advantage for mass-storage and directly compete with NAND Flash for high-density NVM is constituted by the probe storage approach, also called atomic force microscope (AFM)-based data storage [16]. The probe storage concept relies on a "seek-and-scan" mechanism, typically actuated through MEMS devices and obtained with an array of probes that address individual bits on a particular media surface. The data density is a function of both bit size, that depends on the media and on the read/write mechanism, and of the bit pitch, strictly related to the bit size and to the scanning resolution. Despite not suitable for high-performance applications and embedded systems, this technology has a potentially ultrahigh density, terabit/inch² capacity, small form factor, and high data rate. Several approaches employing different storage materials and write/read techniques have been so far proposed[16-18]. However, the media integration presents the same challenging issues as for the lithography-dependent emerging NVMs: material contamination in the fabrication process and degradation due to back-end thermal cycles and to the mechanical wear-out. The main advantage of the probe storage concept relies on its independency from lithography, thus not following Moore's law and its constraints for MOS scaling. It results that there are no known scaling limitations down to the nanometers range, even if the research activity is still in the discovery phase and the complex full integration of the system has not yet been demonstrated. In any case, the development of probe storage technology could have interesting applications in several fields, ranging from lithography to molecular manipulation and nano-storage material development,

thus tracing the route for the post-Moore's law NVMs development.

4. Conclusions

Actually, Flash technology is the mainstream in the NVM market. To continue this trend, innovations in cell structure, dielectric materials, and overall system architecture management will be mandatory. It is expected that Flash scaling will progress at the current rate through at least the end of the decade with the 45nm generation using techniques that are available today or projected to be available in the near future without twisting the current Flash concept. In the mean time, several alternatives are competing to replace Flash providing better performance and wider scaling perspective. Among these emerging technologies and due to its suitability to be exploited in both high-performance embedded systems and data storage applications, PCMs are the most promising concept to eventually become the mainstream in the NVM market.

References

- [1] P. Gargini, Intel nanotechnology virtual open house (2004).
- [2] A. Fazio *et al.*, Intel Tech. Journal, **6**, n. 2 (2002).
- [3] G. Atwood, IEEE Trans. on Dev. Mat. and Rel., **4**, 301-305 (2004).
- [4] H. B. Pein and J. D. Plummer, IEDM Tech. Dig. 2001, 11-14 (1993).
- [5] B. DeSalvo *et al.*, IEEE Trans. on Dev. Mat. and Rel., **4**, 377-389 (2004).
- [6] M. Takata *et al.*, IEDM Tech. Dig. 2003, 553-556 (2003).
- [7] P. Cappelletti *et al.*, IEDM Tech. Dig. 2004, 489-492 (2004).
- [8] K. Kim and G.-H Koh, Proc. MIEL 2004, 377-384, (2004).
- [9] M. S. Choi *et al.*, Proc. ISSCC 2002, 162-163 (2002).
- [10] H. J. Joo *et al.*, Symp. on VLSI Tech., 148-149 (2004).
- [11] S. Tehrani *et al.*, Proc. of the IEEE, **91**, 703-714 (2003).
- [12] S. Lai and T. Lowrey., IEDM Tech. Dig. 2001, 803-806 (2001).
- [13] F. Pellizzer *et al.*, Symp. on VLSI Tech., 18-19 (2004).
- [14] S. J. Ahn *et al.*, IEDM Tech. Dig. 2004, 907-910 (2004).
- [15] A. Pirovano *et al.*, IEDM Tech. Dig. 2003, 699-702 (2003).
- [16] P. Vettiger *et al.*, IBM J. Res. Develop., **44**, 323-340 (2000).
- [17] C.S. Lee *et al.*, Applied Phys. Lett., **83**, n. 23, 4839-4842 (2003).
- [18] F. Gidon *et al.*, Applied Phys. Lett., **85**, n. 26, 6392-6395 (2004).

Influence of Silicon Nanocrystal Size and Density on the Performance of Non-Volatile Memory Arrays

R. A. Rao^a, H.P. Gasquet, R.F. Steimle, G. Rinkenberger, S. Straub, R. Muralidhar, S.G.H. Anderson, J.A. Yater, J.C. Ledezma, John Hamilton, B. Acred, C.T. Swift, B. Hradsky, J. Peschke, M. Sadd, E.J. Prinz, K.M. Chang and B.E. White Jr.

^a Technology Solutions Organization, Freescale Semiconductor, 3501 Ed Bluestein Drive, Austin, TX 78737, USA
Rajesh.Rao@Freescale.Com

Abstract

Silicon nanocrystal memories offer opportunities for voltage scaling and process simplification for embedded non-volatile memories. While electrically isolated nanocrystals mitigate charge loss through oxide defects, the impact of nanocrystal size and density characteristics as well as statistical fluctuations on memory arrays is not well understood. This paper shows that memory window, read disturb and data retention at temperature are roughly insensitive over a broad range of nanocrystal characteristics. Further, data from mega-bit arrays shows that nanocrystal fluctuations effects are small.

1. Introduction

Silicon nanocrystal memories (Figure 1) break up the floating gate of a non-volatile memory into discrete electrically isolated charge storage elements and thereby mitigate charge loss through tunnel oxide defects [1]. This permits scaling of tunnel oxide and operating voltages. This feature in addition to process simplifications over floating gate technology, make it an attractive for embedded NVM applications [2]. There are two main device related concerns in this memory. First, the impact of nanocrystal characteristics such as size and density on memory bit performance has not been elucidated. Secondly, since these nanocrystals are typically formed by random nucleation and growth processes, impact of fluctuations on memory arrays is not well understood. This paper presents for the first time a detailed study of the effects of nanocrystals size, number density and area fraction on memory bit cell performance. Additionally, by optimized CVD processes that results in self-organized nucleation and growth of nanocrystals [3], it is shown that impact of fluctuations in memory arrays is small.

The threshold voltage shift due to electron storage in nanocrystal memory devices is proportional to the number density of nanocrystals and for single electron storage can be of the order of 0.4-0.6V [1], which forces multiple electron storage to obtain a memory window of the order of 2-3V. However, with multiple electron storage in nanocrystals, charge confinement or Coulomb blockade effects become important and manifest as raised electron energy levels in nanocrystals, which, in turn can influence memory cell performance. For example, for a 5nm nanocrystal with 5 electrons, the 5th electron single particle energy level is about 0.5eV

above the silicon conduction band edge reducing the oxide barrier from 3.1eV to about 2.6eV. This raising of energy levels can influence program, erase and retention characteristics and as such, it is necessary to understand the impact of nanocrystal size and density on the memory window. The fractional coverage of nanocrystals is also important as it determines the capture cross section for injected electrons and impacts reliability.

2. Memory Bit Cells

Silicon nanocrystal memory bitcells were produced using a conventional 0.13 μ m CMOS process flow. After tunnel oxidation, silicon nanocrystals were deposited using low pressure CVD. Si island growth during CVD on SiO₂ is believed to proceed by atomistic nucleation, with a critical size of between 1 to 4 atoms [4]. A typical nucleation curve as is shown in Figure 2 consists of the initial incubation phase during which Si adatom concentration builds up to a point where nucleation can occur, a second nucleation phase during which adatom encounters leads formation of fresh nuclei, a third saturation phase where nanocrystals grow in size at approximately constant density and a final coalescence phase during which contiguous nanocrystals impinge by growth resulting in decreased nanocrystal density. The saturation phase is a consequence of the merging of surface adatom depletion zones around nanocrystals (a consequence of diffusion concentration field around nanocrystals) that arrests nucleation. By varying the temperature, precursor partial pressure and the deposition time, the density and size of nanocrystals can be controlled as shown in Figure 3 in which we have fabricated nanocrystals of similar density and different sizes (cases a, and b) and also varied size and density at approximately constant area fraction (cases b and c). Figure 4 shows an EFTEM cross section of the gate stack of the memory device and isolated nanocrystals are clearly seen.

Figure 5 depicts the bitcell performance of three devices [0.24 μ m (W) x 0.24 μ m (L)] with identical bottom oxide (45A SiO₂) and top oxide (120A SiO₂) but different nanocrystal depositions shown in figure 4. The hot carrier injection (V_g=6V, V_d=3.5V and V_b=-2V) program speed and saturation threshold voltage depend on the nanocrystal number density and mean size [Fig 5(a)]. In all cases, the programmed state V_t eventually

saturates when the rate of charge injection into the nanocrystals due to HCI is balanced by the rate of charge removal from nanocrystals due to FN tunneling through the control dielectric. As the mean nanocrystal size is reduced at fixed density, reduction in capture cross section and Coulomb blockade result in lower program speed and saturation V_t . At very low nanocrystal densities ($2 \times 10^{11}/\text{cm}^2$), approximately 15 electrons have to be squeezed into every nanocrystal to achieve the same memory effect. This results in a high field across the top oxide during program [Fig 5(b)], which in turn causes a lower program speed and saturation V_t during program to the case with nanocrystal density of $8 \times 10^{11}/\text{cm}^2$ and mean size of 62Å. The greater raising of energy levels in smaller nanocrystals results in faster erase [Fig 5(b)] and is again a manifestation of Coulomb blockade effect. Interestingly enough, although the device with the smallest nanocrystals erases fast initially, Coulomb blockade effects decrease as the electrons in the nanocrystals are removed and in all cases, the erase saturates at the same value.

The effect of nanocrystal size and density on charge loss through the top oxide was probed during READ disturb measurements of the program state using a 1V gate overdrive [Fig 6]. A charge loss of $< 0.5V$ was observed even after 10^5 seconds and it does not appear to be sensitive to nanocrystal size, possibly, because charge loss from nanocrystals may be masked by detrapping of any electrons trapped in the dielectrics during program/erase cycles. The nanocrystal areal coverage and size also determine the potential for 2-bit per cell applications as well as reliability in single bit per cell operation [5]. The program charge is more localized near the drain edge for the device with smaller areal coverage of nanocrystals (Fig 7). As the coverage and size increases to greater than about 35%, the program charge spreads from drain to source edge, reducing potential for 2-bit/cell application.

3. Memory Array Results

Silicon nanocrystal memory cells were integrated into a 4 Mbit non-volatile memory array. [6] Silicon nanocrystals were deposited using the CVD process described earlier. Deposition times were varied from wafer-to-wafer to achieve a range of nanocrystal densities and sizes. Before and after deposition, the film in the NVM areas was measured using a KLA-Tencor spectroscopic ellipsometer. By subtracting the measurement before the nanocrystals were deposited from the measurement after they were deposited an equivalent nanocrystal “thickness” was determined at various sites on the wafers. Although, there is not a one to one correspondence this thickness and nanocrystal size and density, this technique, under fixed deposition conditions, allows one to observe the progress of nanocrystal deposition with time as well as a measure of deposition non-uniformity across the wafer. The technique also allows for process control of the nanocrystal deposition without resorting to high-

resolution and high-magnification SEM techniques. Shorter deposition times give a thinner equivalent thickness than longer deposition times. After nanocrystal deposition and passivation by growth of a nitrided oxide shell around nanocrystals [6], a high temperature oxide of a thickness of 10 nm was then deposited as the control dielectric. The rest of the processing closely followed a conventional 130 nm CMOS flow.

The wafers were then screened using a J750 Teradyne test system to determine the functionality of the 4 Mbit arrays. Each die was subjected to the following test/measurement sequence: 1) Out-of-fab threshold voltage distributions, 2) screening of failing die using an erase pulse, 3) programming using a 6 V gate, 3.5 V drain and $-2V$ well condition for 65 microseconds and 4) erasing at $-7V$ gate, $+7V$ well, for 200 msec.

Passing die with depositions of 20, 30, 40 and 60 angstroms equivalent thickness were selected for accelerated lifetime tests at 200 deg C temperature. To mitigate the effect of any nanocrystal non-uniformities across the wafer, die were selected from specific areas of different wafers to match inline ellipsometric data. The individual arrays were programmed in a physical checkerboard pattern using a 6 V gate, 4 V drain, $-2.4 V$ well pulse for 10 microseconds to put all adjacent bits in opposite data states. After the initial programming, distributions of both the programmed and erased distributions were measured. No evidence was seen that the programming of one-half of the array had any influence on the erased half. This test reveals that even in large arrays, there is no leakage path from a programmed cell to an unprogrammed cell. The physical underpinning for this is that the nanocrystals are at least partially self-organized during formation, a process that keeps them a minimum separation from each other [3]. Thus nanocrystals do not coalesce to the point of causing bit-to-bit interactions across a relatively wide process window.

To further support this conclusion, we subjected these units to a 200C unbiased bake stress to accelerate data retention loss and lateral charge spreading. This data was then analysed by process split with results described below. Shown in Figure 8 is a typical histogram of V_t averaged across 8 blocks (512k bits) comprising one 4MB array showing the evolution of V_t at several time intervals during the stress. V_t distribution data was collected with a 100mV resolution in gate bias.

Nanocrystal coalescence is expected to enhance lateral charge spreading, thus degrading data retention performance due to hopping conduction. This should be manifested as a charge loss from the program state as charge is redistributed over the trench isolation, reducing impact to bit cell V_t . In the most severe case lateral spreading could cause charge to cross into a neighboring bit in the erase state and cause charge gain and corresponding increase in bit cell V_t . Since nanocrystal coalescence occurs randomly across the surface, not all bits in the array would be affected. One would expect

affected bits to exhibit worse data retention behavior than the intrinsic population, thus forming extrinsic tails that encroach on the useable V_t window for information storage. The absence of extrinsic tails in Figure 8 implies that nanocrystal significant coalescence is not observed over this size range.

Shown in Figure 9 are V_t trajectories for the least erased and least programmed bits (worst case) averaged over a sample of blocks from each process split. Figure 9 shows that data retention behavior is unaffected by nanocrystal size during the first 1000 hours of bake stress showing that the data retention is weakly dependent on nanocrystal characteristics. Further, it supports the absence of coalesced networks that may manifest as tails. In an attempt to look for other effects with nanocrystal size, the width of the V_t distribution for each data state versus nanocrystal size, were extracted as shown in Figures 10 and 11. V_t distribution widths for both data state are essentially flat up to 40Å. The program state shows some distribution broadening at 60Å nanocrystal size that may be caused by increased statistical variation of bit cells with larger nanocrystals. As nanocrystals become larger, sporadic coalescence effects result in larger program distribution widths as such coalescence effects reduce Coulomb blockade and delay program saturation.

Figure 12 depicts an increasing memory window with increasing nanocrystal size. Since a 10 microsecond pulse is insufficient to reach saturation, we can conclude from Figure 5a that differences in the V_t window versus nanocrystal size is a manifestation of programming speed differences associated with the nanocrystal sizes. Small nanocrystals which are equivalent to our 20 angstrom “thickness” samples program slower than larger nanocrystals due to decreased capture cross section and increased gate disturb.

4. Summary

This paper demonstrates shows that Coulomb blockade effects although present, do not degrade memory characteristics as long nanocrystal size and density are adequate. Further, partially self-assembled growth of nanocrystals during CVD also ensures well separated nanocrystals in memory bits over a range of nanocrystal sizes and leads to excellent reliability without tails in array distributions as well as tight V_t distributions.

5. References

- [1] S. Tiwari, et al, Appl. Phys. Lett. **68**, pp 137, (1996).
- [2] R. Muralidhar et al, IEDM Tech Digest, **26.2.1.**, (2003)
- [3] R. A. Puglisi et al. Surface Science **550**,119 (2004)
- [4] T. Kamins, “Polycrystalline Silicon for Integrated Circuits and Displays,” Kluwer Academic Publishers, Massachusetts, (1998).
- [5] B. Eitan et al, Intl. Conf. Solid State Devices and Materials, Tokyo,(1999)
- [6] R. Steimle, et al. IEEE NVSM Workshop, 73, (2004)

6. Figures

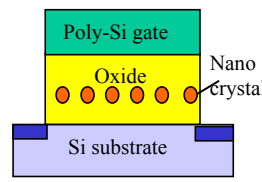


Fig. 1: Nanocrystal Memory Cell

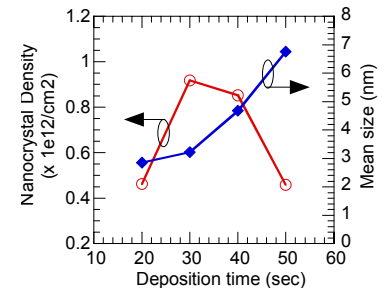


Fig. 2. Nucleation curve of Si nanocrystals deposited by CVD on SiO₂ showing evolution of nanocrystal density and size with deposition time

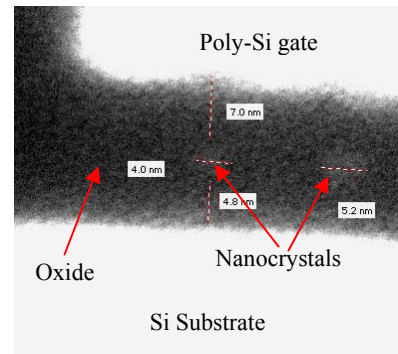


Fig. 3. EFTEM cross-section of the gate stack of a memory bitcell with 45Å bottom oxide, isolated nanocrystals and 80Å top oxide.

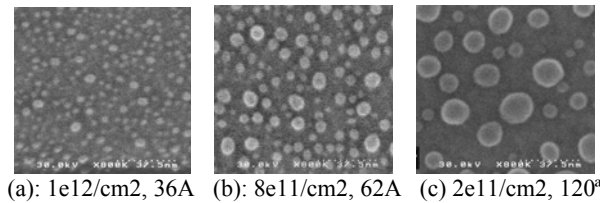


Fig. 4. SEM images of nanocrystals deposited on SiO₂ surface at three different conditions to produce different number density and size distribution

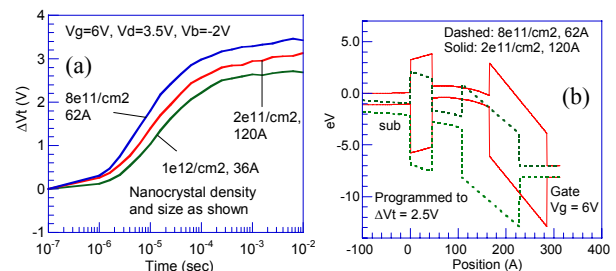


Fig. 5. (a) HCI program and (b) band diagrams during programming for devices with 45Å bottom oxide and 120Å top oxide, with different nanocrystal depositions shown in fig 4

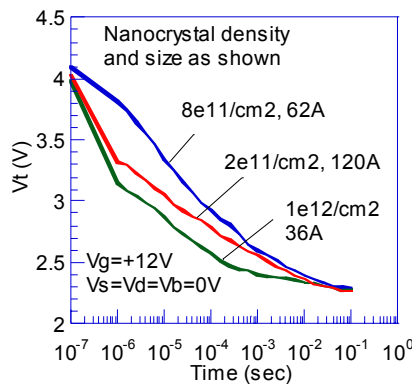


Fig 5(c). FN erase speed under positive gate bias for devices with 45A bottom oxide and 120A top oxide, with different nanocrystal depositions shown in fig 3

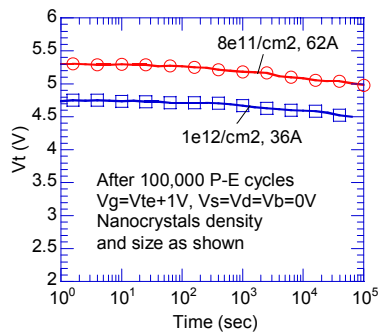


Fig 6. Read disturb of unselected bit in high V_t state at a gate voltage of 1V over the erase V_t for memory bitcells with 45A bottom oxide, 120A top oxide and nanocrystal density and size as shown

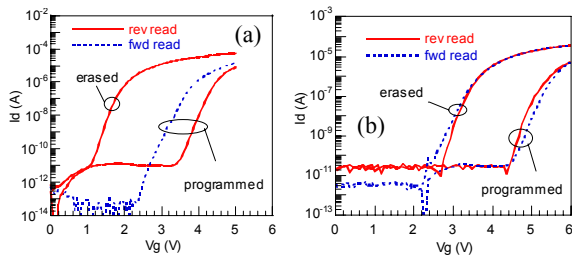


Fig. 7. Program charge localization as measured by forward and reverse read I_d - V_g for programmed memory bitcells with (a) 18% areal coverage of nanocrystals and (b) > 45% areal coverage of nanocrystals

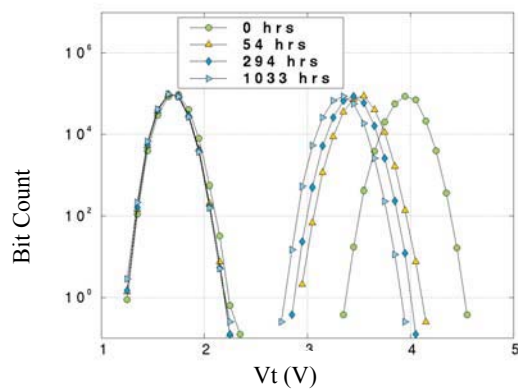


Fig 8: 200C data retention bake of a 4Mb memory array showing the evolution of V_t distribution of programmed and erased bits with time.

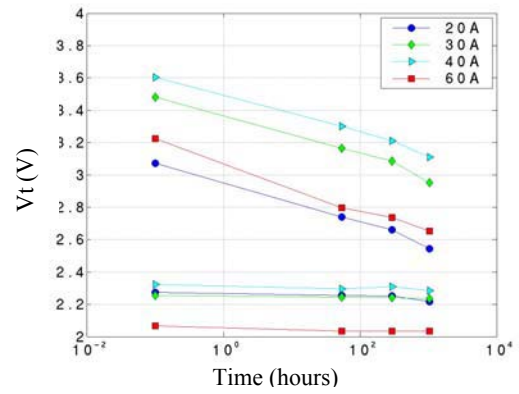


Fig. 9: Effect of nanocrystal size on data retention of least programmed and least erased bits from memory arrays

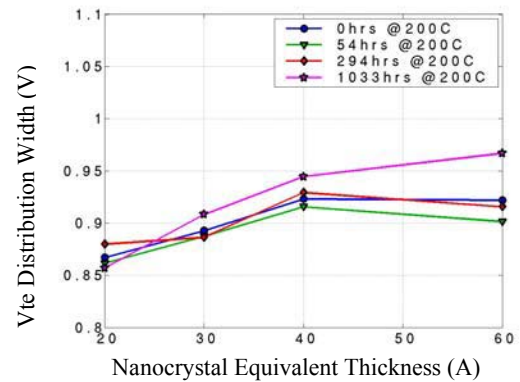


Fig 10. Dependence of the erase state V_t distribution width on nanocrystal size

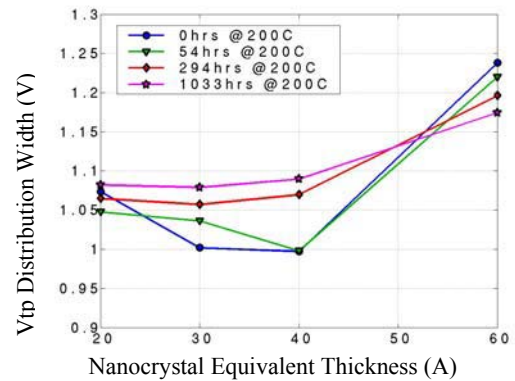


Fig 11. Dependence of the program state V_t distribution width on nanocrystal size

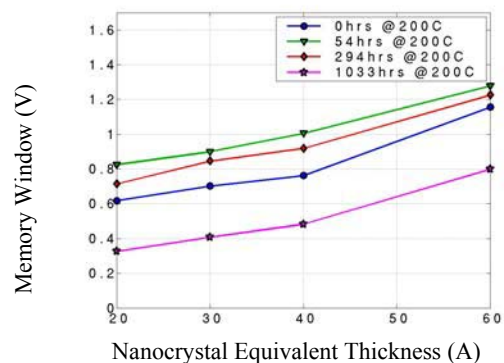


Fig 12. Dependence of memory window on nanocrystal size

Single-electron phenomena in ultra-scaled floating-gate devices and their impact on electrical characteristics

D. Deleruyelle^(a), G. Molas^(b), B. De Salvo^(b), M. Gely^(b), D. Lafond^(b)

Corresponding author: Deleruyelle@L2MP.fr

^a L2MP – Université de Provence / IMT Technopôle de Château Gombert / 13451 Marseille Cedex 20 Affiliation,

^b CEA-Leti-D2NT-LNDE, 17 Avenue des martyrs, 38054 Grenoble Cedex 9

Abstract

Pursuing the continuous downscaling of Flash memory devices now faces many challenging issues such as scaling the active dielectrics [1] (i.e. the tunnel oxide and the interpoly dielectric) and maintaining parasitic coupling between adjacent cells below critical values [2]. Recent works have highlighted other fundamental physical limits related to single-electron phenomena [3]. In this work, single-electron events are firstly evidenced in ultra-scaled floating-gate memory devices with active areas as small as 40×30nm². Then, the impact of these phenomena on retention characteristics has been quantified and discussed in the second part of this work.

1. Introduction

As the dimensions of Flash memory devices are reduced between each technological node, the number of electrons stored in the floating-gate representing a bit is also reduced. Indeed, the number of electrons (N_{elec}), inducing a threshold voltage shift ΔV_{TH} is given by: $N_{elec} \equiv C_G \times \Delta V_{TH}$ where C_G is the interpoly capacitance. Considering the schematic representation of an elementary Flash memory cell (Fig.1-b), N_{elec} can be obtained as function of the gate length and width (L, W), the floating gate height (H) and the interpoly dielectric EOT (t_{IPD}):

$$N_{elec} \equiv \frac{\epsilon_{SiO_2}(WL + 2HL)}{t_{IPD}} \times \Delta V_{TH} \quad (1)$$

Fig.1-a shows the evolution of N_{elec} for different works presented in the literature and according to the ITRS projections for Flash memories considering a threshold voltage shift of 3V [4]. As we can see, the most advanced cells, corresponding to the 65nm node [5] only requires few hundred of electrons to store a bit. This number will continue to decrease at a rate of $\sim 0.77\times$ between each Flash technology node, thus entering in the few electron area for deca-nanometer sized devices. Therefore, it will become urgent to study few-electron phenomena for future generations of Flash memory cells.

2. Device Fabrication

Figs.2-a and 2-c&d show a schematic representation and TEM observations of the devices studied in this work. This technology was developed on 8" SOI wafers where active areas were firstly thinned to a thickness of 15nm. The silicon layer was then

oxidized to form a 5nm thick gate oxide and covered by a first 25nm thick n⁺-polysilicon. A first hybrid Deep-UV (DUV)/e-beam lithography followed by an etching is then performed to define the SOI channel and the floating gate width. The control oxide was obtained by the deposition of a HTO layer. Finally the deposition of a 100nm-thick n⁺-polysilicon layer followed by a second hybrid lithography and etching defines the control gate of the devices. As shown in Fig.2-a, devices with gate length/width as small as 30nm were obtained with this fabrication process. Table in Fig.2-b summarizes the technological characteristics of the studied memory devices.

3. Electrical Results and discussion

4.1. Transfer characteristics

Single-electron effects were firstly evidenced on transfer characteristics. Fig.3 shows a hysteresis measurement performed on a 50×50nm² memory cell. As we can see, abrupt steps appears in I_D - V_G curves corresponding to discrete threshold voltage shifts due to single electron charging/discharging from the control gate to the floating gate of the device during the gate voltage sweep.

4.2. I_D -time evolution

The discrete nature of the charging/discharging events has also been observed on I_D -time measurements. In Fig.4, the transfer characteristics of a 40×30nm² cell are shown both in the initial state and after application of a writing stress. As shown in the inset of Fig.4, the time evolution of the drain current between the two states, at constant gate voltage, reveals that I_D decreases by abrupt steps corresponding each to a single electron transfer to the floating gate of the device. At this stage, the value of the discrete threshold voltage shift, δV_{TH} , can be calculated from the sub-threshold slope, S_s , of the I_D - V_G curve and the values of I_D before and after the electron transfer (resp. I_{D1} & I_{D2}) by:

$$\delta V_{TH} = 1/S_s \times \ln\left(\frac{I_{D1}}{I_{D2}}\right) \quad (2)$$

4.3. Evolution of δV_{TH} with the cell dimension

In order to quantify the evolution of δV_{TH} with the dimension of the cell, I_D -time measurements, similar to those of Fig.4, were performed on devices with surfaces varying from 30×30nm² to 100×100nm². As we can see on Fig.5, the amplitude of current steps (and thus the value of δV_{TH} itself), related to single-electron charging, increases as $W \times L$ is reduced.

Following the same procedure for a large number of devices and using equ.2, the values of δV_{TH} have been extracted for various $W \times L$ and reported in Fig.6. Results indicate that δV_{TH} values as high as 15mV per electron for the smallest area devices ($30 \times 30 \text{ nm}^2$) can be obtained, in good agreement with theoretical estimations.

4.4. Impact on retention characteristics

Finally the influence of single-electron phenomena on some electrical characteristics has been studied. In particular, Fig.7 shows retention characteristics evaluated by I_D -time measurements after a writing sequence. Due to the stochastic nature of single-electron tunnelling [6], the retention time, T_{RET} , here defined by the time to reach a fixed current value, becomes a stochastic variable and thus retention measurements do not give reproducible results. This result emphasizes that single-electron effects affect the intrinsic reliability of ultra-scaled memory devices. As shown in Fig.8, repeating several times the same measurement on the same device shows that T_{RET} is distributed around a mean value.

4. Conclusions

In this work, single-electron phenomena were evidenced on ultra-scaled floating-gate memory devices. It has been shown that the amplitude of single electron phenomena increases with reducing the surface of memory devices. Due to the stochastic nature of single-charge tunnelling, it has been finally shown that these phenomena introduce dispersions in the electrical characteristics of memory devices that inherently call the intrinsic reliability of floating-gate devices into question in the deca-nanometer range.

5. References

- [1] G. Atwood, "Future Directions and Challenges for ETox Flash Memory Scaling", IEEE Trans. On Dev. And Mat. Rel., vol. 4, n°3, pp. 301-205, 2004
- [2] J.D. Lee et al., "Effects of Floating-Gate Interference on NAND Flash Memory Cell Operation", IEEE Electron Dev. Lett., vol. 23, n°5, pp. 264-266, 2004
- [3] G. Molas et al., "Impact of few electron phenomena on floating-gate memory reliability", IEDM 2004
- [4] International Technology Roadmap for Semiconductors 2003, <http://public.itrs.net/Files/2003ITRS/Home2003.htm>
- [5] J.H. Park et al., "8Gb MLC (Multi-Level Cell) NAND Flash Memory using 63nm Process Technology", IEDM 2004.
- [6] K. Yano et al., "Single-Electron Memory for Giga-to-Tera Bit Storage", Proc. of IEEE, vol. 87, n°4, pp.633-651, 1999

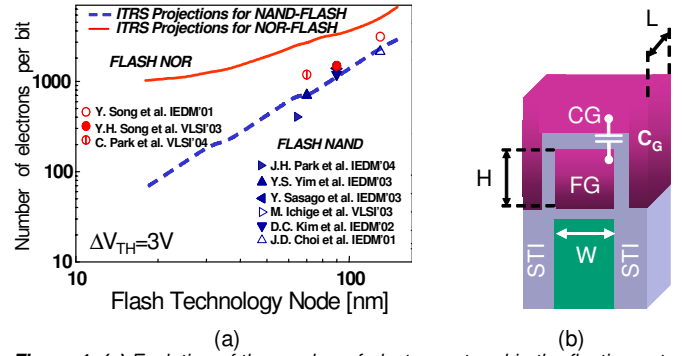


Figure 1. (a) Evolution of the number of electrons stored in the floating gate of a Flash-E²PROM cell to induce a V_{TH} shift of 3V. This number has been calculated for different cells presented in the literature and extrapolated according to the ITRS projections. **(b)** Schematic representation of an elementary Flash memory cell

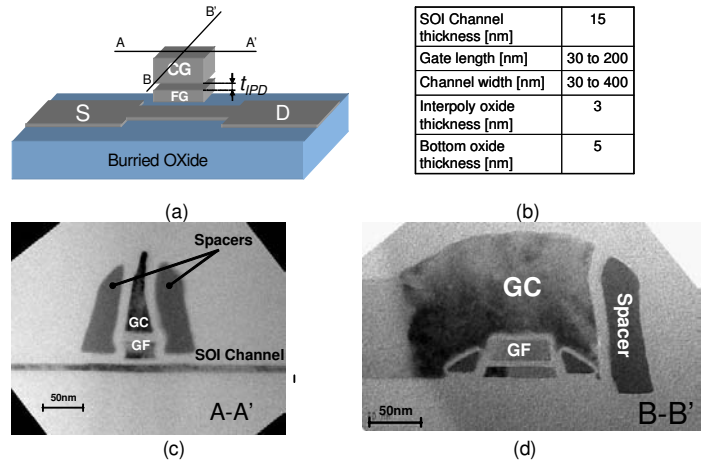


Figure 2. (a) Schematic representation of the devices studied in this work **(b)** Table containing the main characteristics of the studied devices. **(c)** TEM cross sections of a memory device in a direction parallel and **(d)** perpendicular to the channel direction (L is $\sim 30 \text{ nm}$)

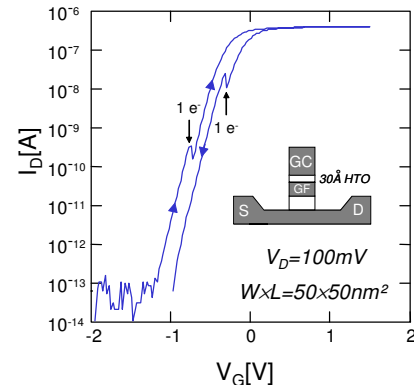


Figure 3. Transfer characteristics of a $50 \times 50 \text{ nm}^2$ memory device. Abrupt steps appear during the hysteresis measurement evidencing single-electron charging/discharging during the gate voltage sweep.

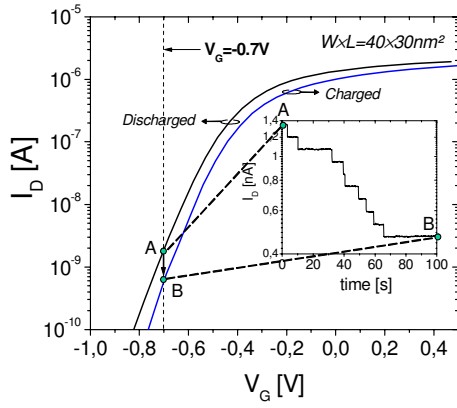


Figure 4. I_D - V_G characteristics of a $40 \times 30 \text{ nm}^2$ memory device initially discharged and after application of writing stress ($1 \text{ s} @ V_{GW} = -0.7 \text{ V}$). The figure in inset shows that the time evolution of the drain current at constant gate voltage between the erased state ("A") and written state ("B") reveals that charging is performed by single electron transfer.

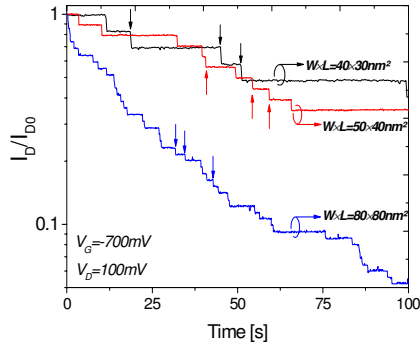


Figure 5. Time evolution of the drain current during a writing phase for memory devices with various surfaces. Arrows on the curves indicates some single electron effects.

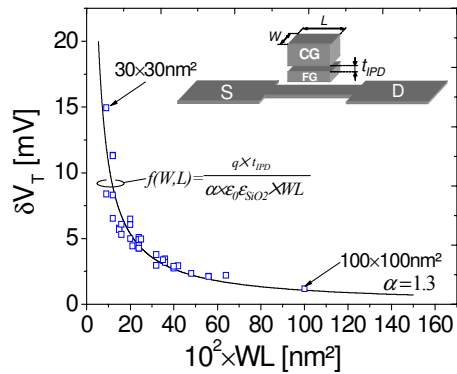


Figure 6. Evolution of the single-electron threshold voltage shift (δV_{TH}) extracted from I_D -time measurements for devices with various surfaces. The continuous line corresponds to a theoretical estimation of δV_{TH} considering the schematic representation of the cell presented in the inset.

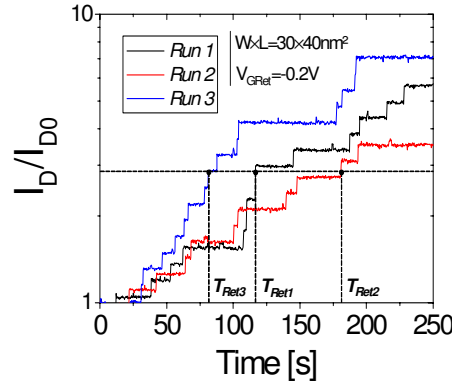


Figure 7. Retention characteristics of a $30 \times 40 \text{ nm}^2$ cell evaluated by I_D -time measurements after a writing sequence ($1 \text{ s} @ -5 \text{ V}$) for 3 successive runs

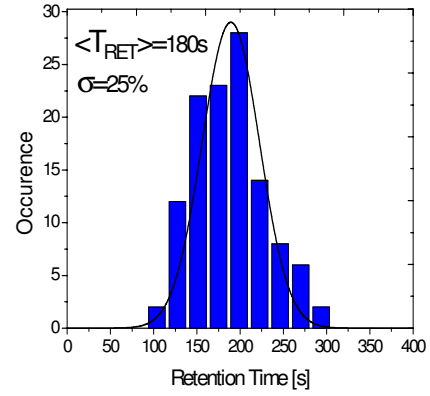


Figure 8. Distribution of the retention time extracted from I_D -time measurements.

Si nanocrystals by ultra-low-energy ion beam-synthesis for nonvolatile memory applications

C. Bonafos^a, H. Coffin^a, S. Schamm^a, M. Carrada^a, N. Cherkashin^a, G. Ben Assayag^a, P. Dimitrakis^b, P. Normand^b, M. Perego^c, M. Fanciulli^c and Claverie^a

^a CEMES-CNRS, 29 rue J. Marvig, 31055, Toulouse, France, bonafos@cemes.fr

^b Institute of Microelectronics, NCSR 'Demokritos', 15310 Aghia Praskevi, Greece

^c Laboratorio MDM - INFN Agrate-Italy.

Abstract

In this work, we show how to manipulate the depth-position, size and surface density of two dimensional arrays of Si ncs embedded in thin (≤ 10 nm) SiO₂ layers, fabricated by ultra-low energy (≤ 1 keV) ion implantation and subsequent annealing. Particular emphasis is placed upon the influence of implantation and annealing conditions under oxidizing ambient on the ncs characteristics (position, size, density) and on the charge storage properties of associated MOS structures.

1. Introduction

By exploiting the benefits (robustness and fault-tolerance) of a charge storage distribution, nanocrystal MOS memory devices offer an attractive alternative for overcoming the scaling limitations of conventional floating-gate semiconductor memories [1,2]. For achieving reliable devices and avoiding fluctuations in device performance not only the nanocrystals (ncs) characteristics (size, density, interspacing) but also their location inside the gate dielectric have to be well controlled. In particular, fine control of the thickness of the dielectric layer (injection oxide) separating the ncs and the substrate is absolutely required since a change of less than 1 nm affects dramatically the programming properties (write/erase times and voltages) and data retention of the devices. Among the different technological routes [1-4] explored the last few years for generating nanocrystals in the gate oxide of MOS devices, the ion-beam-synthesis (IBS) technique has received substantial attention due to its flexibility and manufacturing advantages. The potential of this technique for nc-based-memories operating at low voltages has been recently enhanced through the synthesis in the ultra-low-energy (ULE) regime (typically 1keV) of single Si-nc layers in thin SiO₂ films (≤ 10 nm) by ULE-IBS. Specific experimental methods have been recently developed to characterize the ncs populations. They include using Transmission Electron Microscopy (TEM) Fresnel imaging for evaluating the distances and widths of interest [6], High Resolution TEM and spatially resolved Electron Energy Loss Spectroscopy (EELS) using the spectrum-imaging mode of a Scanning Transmission Electron Microscope (STEM) to measure the size distribution and density of

the ncs population [7]. We have used these techniques to study the influence of implantation parameters (energy, dose), annealing ambient (N₂ versus N₂+1.5% O₂) and oxide thickness on the ncs characteristics and understand how the latter ones affect the charge storage properties of associated MOS devices.

2. Experimental details

Four sets of samples have been examined. All sample details are summarized in Table 1. Each set is characterized by 3 parameters: nominal oxide thickness, implantation energy and dose. In Set 1, the dose was kept constant and the energy was varying from 0.65 to 5 keV. In Set 2, the Si implantations were carried out using different dose-energy combination extracted from TRIDYN [8] simulations in such a manner that the peak of concentration of implanted atoms always reached 35 at.% (in SiO₂). The oxide thickness and the implantation dose were varied for respectively Set 3 and 4. All the samples were annealed at 950°C for 30 minutes under N₂. A fifth set of samples (Set 5) was dedicated to the study of the effect of the annealing ambient. It consists in 7 nm SiO₂ layers, implanted at 1 keV with a dose of 2×10^{16} cm⁻² and annealed at 950°C for 30 min under N₂ and then oxidised at 950°C for 30 min and 120 min, under N₂+1.5%O₂. Specimens were prepared for TEM observations using the standard procedure involving mechanical polishing and ion milling. Cross sectional images were taken using a CM30 Philips TEM operating at 300 keV using strongly underfocused bright-field imaging conditions (Fresnel contrast). Spatially resolved EELS using the spectrum imaging mode of a scanning TEM was performed on plane-view specimen to measure the size distribution, surface density and coverage of the ncs. Charge storage of selected samples was evaluated through high-frequency capacitance-voltage measurements of Al gate capacitors.

Sample	Set 1	Set 2	Set 3	Set 4
t _{ox} (nm)	10	10	10, 7, 5	10
E (keV)	0.65, 1, 2, 3, 5	1, 3, 5	1	1
Dose (cm ⁻²)	10 ¹⁶	10 ¹⁶ 2.1x10 ¹⁶ 3.2x10 ¹⁶	10 ¹⁶	5x10 ¹⁵ 8x10 ¹⁵ 10 ¹⁶ 1.5x10 ¹⁶ 2x10 ¹⁶

Table 1: Experimental conditions for the examined sets of samples.

3. Structural characterization

a) Effect of implantation energy and oxide thickness on ncs location

For tuning the location of the ncs layer within the SiO₂ layer, our first attempt was to vary the implantation energy according to the conditions of Set 1 samples. As the ion energy increases from 1 to 3 keV, the ncs layer shifts deeper in the SiO₂ layer, i.e., closer to the Si/SiO₂ interface (see Fig. 1a). Surprisingly, no ncs are detected for 5 keV implantation and HREM imaging (not shown) reveals an unusual roughness [9] of the SiO₂/Si interface. In order to compensate the decrease of the Si peak concentration with increasing ion energy, the experiment was repeated with ion doses adjusted to obtain a constant peak concentration of 35 at. % Si excess (Set 2 samples). Again a nanocrystal layer is formed and shifts closer to the SiO₂/Si interface when the implantation energy increases from 1 to 3 keV. Moreover, a rough interface is found at 3 keV and multiple twins and grains appear underneath the interface at 5 keV. Another way to decrease the injection distance is to keep fixed implantation conditions and decrease the nominal oxide thickness. TEM observations indicate that three parameters remain constant when decreasing oxide thickness (Fig. 1b, Set 3 samples): the depth position of the ncs band with respect to the surface, its width and the overall swelling of the SiO₂ layer. For the 5 nm-thick oxide, an injection distance as small as 2 nm can be obtained. Two regions of excess Si corresponding to the Gaussian like implantation profile and error-function-like profile at the interface due to ion mixing have been detected by ToF-SIMS measurements of these samples (see [9, 10]). These regions are always well separated in the case of implantation at low energy (1 keV) even for SiO₂ layers as thin as 5 nm. On the contrary, they overlap when increasing the implantation energy up to 5 keV. TRIDYN simulations combined with experimental data gives a safe explanation regarding the unique technological route of obtaining Si nanocrystals at distances smaller than 5 nm from the channel [8, 9]. The formation of nanocrystals requires that the interface mixing due to collisional damage does not overlap with the range profile to the extent that there is no more a local maximum of excess Si buried in the SiO₂ layer.

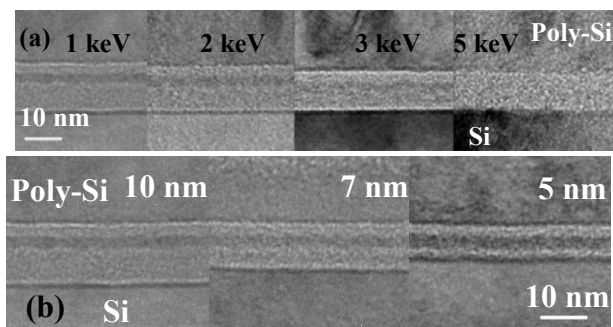


Fig. 1: Series of XTEM images of (a) Set 1 and (b) Set 3, taken under Fresnel conditions

b) Effect of the implanted dose

The ncs morphology strongly depends on the implanted dose. The formation of Si ncs in a 10 nm-thick oxide at fixed implantation energy (1 keV) has been monitored by varying the implanted dose (Set 4 samples) and examined using spatially resolved EELS using STEM spectrum imaging (Fig. 2). Three main characteristics, -size, density, surface coverage- have been extracted from these images after special software treatment: 1) no ncs are detected for the lowest dose; 2) a dense array (3.3x10¹² cm⁻², 15% surface coverage) of 2.7 nm-diameter sphere-like ncs is formed at medium dose; 3) a connected array of elongated ncs with larger size (9.3 nm) and smaller density (10¹¹ cm⁻², 44% surface coverage) is created at highest dose. An expansion of the whole SiO₂ thickness that increases with the implanted dose is measured by XTEM [5]. The observed structural features as a function of the implantation dose agree qualitatively with the ones predicted by kinetic 3D lattice Monte Carlo simulations [11]. Indeed, at low dose we observe the growth of sphere-like and separated ncs while at higher doses spinodal decomposition with creation of elongated and connected Si structures takes place. It should be emphasized here that the experimental doses needed for ncs formation are twice higher than that used in simulations. A probable reason for that is the partial loss of Si implanted dose as a result of O₂ penetration through the damage induced by implantation and further oxidation [11].

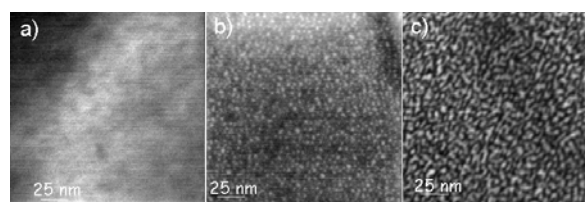


Fig. 2: EELS/STEM images of samples Set 4 showing the effect of dose on Si ncs morphology: a) 5x10¹⁵ cm⁻²; b) 10¹⁶ cm⁻²; c) 2x10¹⁶ cm⁻².

c) Effect of the annealing ambient

The integrity of the oxide is affected by the presence of implantation and excess Si-related defects. To restore the quality of the SiO₂ layer and achieve ncs-based-MOS devices operating at low voltages, a post-implantation annealing treatment in diluted oxygen environment was proposed [12]. For this reason, we have studied a series of samples (Set 5) implanted at high dose and low energy, and subsequently annealed under N₂ atmosphere for 30 min or under diluted oxygen (N₂+1.5%O₂) during 30 or 120 min. XTEM imaging (Fig. 3) reveals additional swelling of the control oxide after annealing under oxidizing atmosphere. For long annealing times (120min, Fig. 3c), a substantial increase of the injection oxide is also observed. As it can be seen on the EELS/STEM images (Fig. 3d), coalesced elongated-Si ncs are formed after N₂ annealing. By annealing the samples in oxidizing ambient during 30 minutes, separation of coalesced ncs occurs (Fig. 3e). A 120 min annealing time in diluted oxygen leads to well-separated

sphere-like particles with smaller densities and sizes (Fig. 3 f).

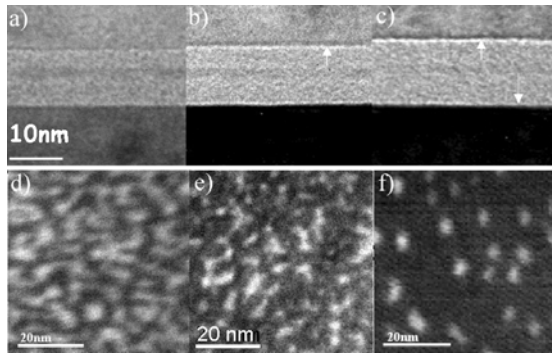


Fig. 3: XTEM (top) and EELS/STEM (down) images of samples Set 5 showing the effect of annealing ambient on Si ncs: (a) and (d) 950°C N₂ 30 min; (b) and (e) 950°C 30 min, N₂+1.5%O₂; (c) and (f) 950°C 120 min, N₂+1.5%O₂.

A systematic experimental study and the modeling of the Si ncs oxidation has been performed and is detailed in ref [12]. We have chosen to study the oxidation process in the case of low implanted dose (10^{16} cm⁻²) for which a high density of well-separated spherical ncs are observed after annealing under N₂. In this case, the average size of the ncs (square symbols in Fig. 4) slowly decreases from 2,9 nm $\pm$ 0.3 nm for the non-oxidised sample (N₂, 1050°C, 30 min) to 1,8 nm $\pm$ 0.2 nm for the sample annealed under N₂+O₂ during 240 min. Meanwhile, the ncs density (not shown) decreases by a factor of 10 from 5 min to 240 min indicating that a large part of the population of ncs is totally consumed. Thus, the number of Si atoms contained within the ncs decreases from 1.4×10^{15} at/cm² after 5 min to 10^{14} at/cm² after 240 min of annealing. For modeling the oxidation process of the Si ncs, we have used the classical extension of the Deal and Grove model to a 3-D geometry with appropriate boundary conditions in spherical coordinates as proposed by Omachi et al [13]. As SiO₂ has a very large viscosity, the non-planar viscous deformation of the oxide produces large stresses. Two types of stress are generated that slow down the oxidation process. One is the normal compressive stress σ generated at the Si/SiO₂ interface as the newly formed oxide pushes out the old oxide and faces a resistance from the normal stress perpendicular to the surface. This stress makes the reaction at the silicon surface more difficult and affects the surface reaction rate constant which decreases as the normal stress increases [13, 14]. On the other hand, a new stress, in tension, is also created in the bulk of the old oxide which is stretched and being pushed out. This stress affects the O₂ diffusivity and solubility in SiO₂ but also the SiO₂ viscosity that increases under pressure [13, 14]. The oxidation rate has been calculated by taking into account these different stress effects and the time evolution of the oxidation rate has been solved by means of a Runge- Kutta method [12]. We obtain a good agreement between our model and the experimental data (see Fig. 4) for a viscosity of 7×10^{14} poises i. e., close to the one measured for dry oxides at this temperature, rendering our model realistic.

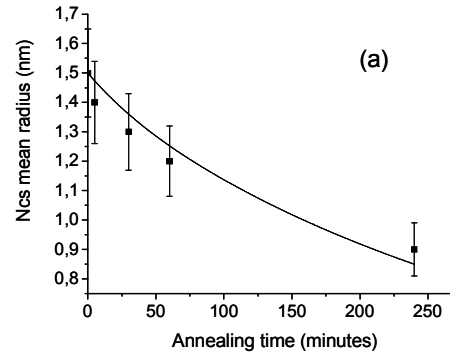


Fig. 4: Evolution of the Si ncs mean radius versus the annealing time under N₂+O₂ measured on EFTEM images. In plain line are plotted the model predictions.

4. Electrical characterization

The structural characteristics of the ncs layers are strongly related to the electrical properties of the associated MOS structures [15-17]. As the implantation energy increases, the electric field needed for effective charging decreases and the magnitude of the flat-band voltage shift, ΔV_{FB} , increases, as shown in Fig. 5a. While such a result can be directly correlated to the location of the ncs layer (see Fig. 1a), it is believed that the implantation-related-damage including the excess of silicon atoms in the injection oxide contributes substantially to the significant charging observed at small electric fields despite the large injection distances. The same behavior (not shown) is observed when decreasing the initial oxide thickness (see Fig. 1b). The increase of the implanted dose up to 10^{16} cm⁻² also leads to an enlargement of the memory window (Fig. 5b). Such a result is obviously linked to the increase of the ncs density (see Fig. 2) but also to the increase of the concentration of excess silicon in the injection oxide. Further increase of the implantation dose up to 1.5×10^{16} cm⁻² leads to the formation of a coalesced array of ncs (Fig. 2c) and to a decrease of the memory window [16]. For studying the effect of oxidizing ambient, samples with implantation conditions of Set 5 have been annealed under the following conditions: (A1) 950°C/30 min/N₂; (A2) 950°C/15min/N₂ plus 950°C/15min/N₂ +1.5%O₂; (A3)950°C/30min/N₂+1.5%O₂. In comparison with A1 treatment, the magnitude of ΔV_{FB} as a function of the gate voltage sweep strongly increases after A2 and A3 thermal treatments (see Fig. 5c). Such an increase is not only due to the improved quality of the overall oxide and enhanced thickness of the control oxide as reported in [15] but also to the separation (see Fig. 3) and passivation of the initially connected ncs [18].

By exploiting the flexibility of the ULE-IBS-related fabrication routes briefly described in this study, Si-ncs single-transistor memory cells aiming at EEPROM-like [15-17] and NVRAM-like applications have been successfully fabricated.

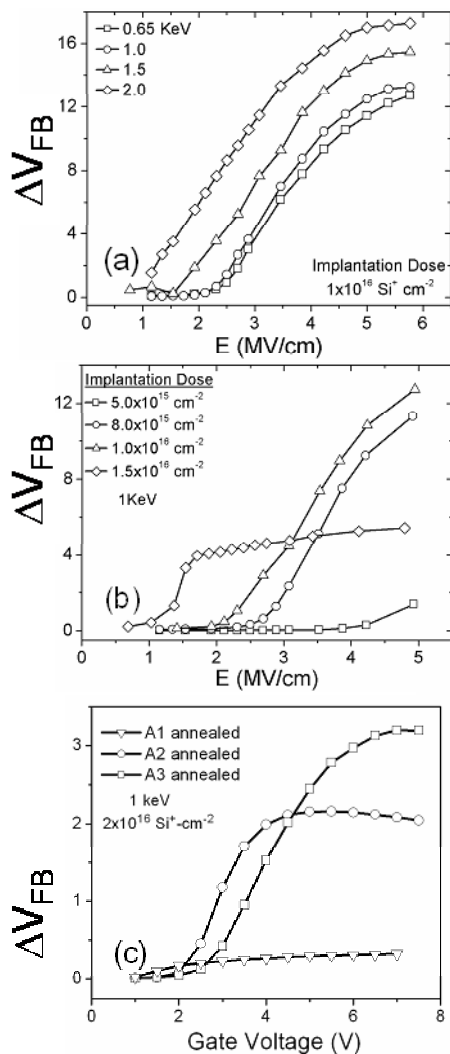


Fig. 5: Flat-band voltage shift (ΔV_{FB}) as a function of the applied electric field (a, b) and gate voltage round sweep (c) for 10-nm-thick oxide layer implanted with Si (a) at different energies and (b) increasing implantation dose and (c) 7-nm-thick oxide layer implanted with 2×10^{16} Si/cm² at 1 keV and subsequently annealed following the A1, A2, or A3 treatment. For (a) and (b) samples a 15 nm-thick oxide was deposited before annealing.

5. Conclusion

In this work, we have shown how to control the characteristics of 2D arrays of Si ncs embedded in thin SiO₂ layers. The injection distance can be tuned from 8 to 2 nm by a judicious combination of ion beam energy and initial SiO₂ thickness. The ncs surface density and coverage can be controlled by the implantation dose even if the efficient range is small. An increase of the memory window is observed when decreasing the injection oxide thickness, increasing ncs the density up to the coalescence threshold and annealing under oxidizing ambient. During such oxidation conditions the oxide integrity is restored and a separation of the connected ncs takes place, making it possible to control the ncs density, mean size and surface coverage. An extended spherical Deal-Grove model for self-limiting

oxidation of embedded Si ncs has been developed and compared to the experimental results. The evolution of the Si ncs size under oxidation that we experimentally observe can be fitted by taking into account a stress-induced retardation process.

6. Acknowledgements

This work was supported by the EC through the Growth project G5RD/2000/00320-NEON (Nanoparticles for Electronics). The authors want to thank V. Soncini from ST Microelectronics Agrate for the oxidized wafers, M. Tencé and C. Colliex from LPS/Orsay for the PEELS/STEM images and M. Respaud from LNMO-INSa, Toulouse for its “numerical” support.

References

- [1] S. Tiwari, F. Rana, H. I. Hanafi, A. Hartstein, E. F. Crabbé and K. Chan, Appl. Phys. Lett. **68**, 1377 (1996).
- [2] P. Normand, E. Kapetanakis, P. Dimitrakakis, D. Skarlatos, K. Beltsios, D. Tsoukalas, C. Bonafos, G. Ben Assayag, N. Cherkashin, A. Claverie et al., Nucl. Inst. and Meth. in Phys. Res. B **216**, 228 (2004).
- [3] G. Ammendola, M. Vulpio, M. Bileci, N. Nastasi, C. Gerardi, G. Renna, I. Crupi, G. Nicotra, and S. Lombardo J. Vac. Sci. Technol. B **20**, 2075 (2002).
- [4] Y. C. King, T. J. King and C. Hu, IEEE Trans. Electron Devices Meet. ED-48, 696 (2001).
- [5] M. Carrada, N. Cherkashin, C. Bonafos, G. Ben Assayag, D. Chassaing, P. Normand, D. Tsoukalas, V. Soncini and A. Claverie, Mat. Sc. Eng. B **101**, 204 (2003).
- [6] G. Ben Assayag, C. Bonafos, M. Carrada, P. Normand, D. Tsoukalas and A. Claverie, Appl. Phys. Lett. **82**, 200 (2003).
- [7] C. Jeanguillaume, C. Colliex, Ultramicroscopy **28**, 252 (1989).
- [8] W. Möller and W. Eckstein, Nucl. Inst. and Meth. in Phys. Res. B **2**, 814 (1984).
- [9] C. Bonafos, M. Carrada, N. Cherkashin, H. Coffin, D. Chassaing, G. Ben Assayag and A. Claverie, J. Appl. Phys. **95**, 5696 (2004).
- [10] M. Perego, S. Ferrari, M. Fanciulli, G. Ben Assayag, C. Bonafos, M. Carrada and A. Claverie, J. Appl. Phys. **95**, 257 (2004).
- [11] T. Müller, C. Bonafos, K.-H. Heinig, M. Tencé, H. Coffin, N. Cherkashin, G. Ben Assayag, S. Schamm, G. Zanchi, C. Colliex, W. Möller, A. Claverie, Appl. Phys. Lett. **85**, 2373 (2004).
- [12] H. Coffin, C. Bonafos et al., proceeding of MRS 2004 Fall meeting, in press.
- [13] J. Omachi, R. Nakamura, K. Nishiguchi and S Oda., MRS Symp. Proc. 638 (2001).
- [14] D. B. Kao, J. P. McVittie, W. D. Nix, and K. C. Saraswat, IEEE Trans. Electron Devices, ED-35, 25 (1988).
- [15] P. Normand, E. Kapetanakis, P. Dimitrakakis, D. Tsoukalas, K. Beltsios, N. Cherkashin, C. Bonafos, H. Coffin, G. Benassayag, A. Claverie et al, Appl. Phys. Lett. **83**, 168 (2003).
- [16] P. Dimitrakakis, E. Kapetanakis, P. Normand, D. Skarlatos, D. Tsoukalas, K. Beltsios, A. Claverie, G. Benassayag et al., Mat. Sc. and Eng. B. **101**, 14 (2003).
- [17] P. Dimitrakakis, E. Kapetanakis, D. Tsoukalas, D. Skarlatos, C. Bonafos, G. Ben Assayag, A. Claverie, M. Perego, M. Fanciulli, V. Soncini, R. Sotgiu, A. Agarwal, M. Ameen, P. Normand, Solid State Electronics **48**, 1511 (2004).
- [18] A. Wellner, V. Paillard, H. Coffin, N. Cherkashin, C. Bonafos, J. Appl. Phys. **96**, 2403 (2004).

Time-dependent simulation of the program and erase operations of nanocrystal Flash memories

A. Campera, G. Iannaccone

Dipartimento di Ingegneria dell'Informazione: Elettronica, Informatica, Telecomunicazioni
Università degli Studi di Pisa, Via Caruso, 56122, Pisa, Italia
{andrea.campera, g.iannaccone}@iet.unipi.it

Abstract

We have investigated, by means of numerical simulations, the charging and discharging dynamics of flash memories based on silicon nanocrystals during program and erase operations. Comparison between simulations and experiments available in the literature provides useful insights of the storing mechanisms. In particular, they allow us to rule out that electrons are stored in the conduction band of silicon nanocrystals, and confirm the hypothesis that electrons are actually trapped in localized states in the silicon gap, probably located at the interface between the nanocrystals and the surrounding silicon oxide. Our simulation code is based on a one-dimensional Poisson-Schroedinger solver which allows the computation of the tunnelling current and of generation and recombination rates between the terminals and localized states in the dielectric layer, which has been successfully used in the investigation of trap-assisted tunnelling responsible for SILCs. Time dependent simulations of the program/erase operations are based on a description of the nanocrystal memory dynamics in terms of a master equation. Comparison with experiments suggests that the dominant traps responsible for charge storage are located in the silicon gap at an energy close to the valence band edge.

1. Introduction

Silicon nanocrystal memories [1] are investigated as a promising alternative to conventional Flash memories [2,3]. Indeed, the discrete nature of the storage elements reduces the impact of Stress-Induced-Leakage-Currents (SILCs) on data retention [4] and therefore enables the use of a thinner tunnel oxide, which in turn implies higher programming current, lower program/erase times, and at the same time lower program and erase voltages.

One of the main open issues for a correct understanding of the operation of nanocrystal memories is where electrons are trapped. Some authors accept the physical picture of electrons trapped in confined states in the nanocrystal conduction band [4-8]. Other papers conclude that electrons are instead mostly stored in localized traps at the interface between silicon nanocrystals and silicon oxide [9-11]. The issue is obviously relevant also from a technological point of view, since it determines which aspects must be more closely addressed for improving data retention and for reducing program/erase times.

In this paper, we address such issue by means of numerical simulations, and show that only the mechanism according to which electrons are stored in

traps localized in the silicon gap at the nanocrystal-silicon oxide interface is compatible with the well known experimental results, such as for example those presented in Ref. [3].

It is always very hard to draw drastic conclusions on an experiment by means of a numerical simulation, which is based on a model that must obviously be rather simplified and idealized with respect to the actual system. Nevertheless, we shall show that results are rather robust with respect to varying parameters, and enable us to claim with some confidence that electrons are not stored in the nanocrystal conduction band.

2. Physical model

For the study of the dynamics of nanocrystal charging and discharging we have adapted a one-dimensional model [12], that was originally proposed for the study of trap-assisted tunnelling responsible for SILCs in MOS capacitors, which treats tunnelling into and from the nanocrystals in terms of generation/recombination processes.

Here, we want to address with a unique model two very different cases from the physical point of view:

- In the first case, we have a quantum dot defined by the electron confinement due to the large gap of surrounding silicon oxide layer. The electrochemical potential of the dot is a function of the bias conditions and of the number of electrons in the dot, due to single electron charging.
- The second case correspond to several electrically insulated traps placed at the silicon-silicon oxide interface of each nanocrystal. Let us assume that Coulomb repulsion allows at most one electron per trap. In this case, we can still see the nanocrystal as a unique system, its electrochemical potential being the minimum energy required to place an electron in one of the traps. For simplicity, we assume that all energy traps are identical, and that when an electron is added to the system, the energy of all trap states increases by a given charging energy.

Here we briefly report the basic concepts of our model (further details are available in [12]). Let us stress our simplifying assumptions that *i*) the nanocrystals are uniformly distributed, *ii*) charged dots do not interact with each other, that *iii*) when an electron is added to the nanocrystal system, its electrochemical potential increases by a factor $\Delta\mu = q^2/C$, where q is the electron charge and C is the capacitance between the nanocrystal and the electrodes, that can be approximated by

$$C = C_1 + C_2 = \frac{\pi \epsilon_{Si} d^2}{2t_1 + \frac{\epsilon_{ox}}{\epsilon_{Si}} d} + \frac{\pi \epsilon_{Si} d^2}{2t_2 + \frac{\epsilon_{ox}}{\epsilon_{Si}} d},$$

where d is the nanocrystal diameter, t_1 is the tunnel oxide thickness, and t_2 is the control oxide thickness, and $\epsilon_{Si}(\epsilon_{ox})$ is the silicon (oxide) dielectric constant. Here we discard the effect of quantum confinement on the addition energies.

We can refer to the 1D profile of conduction and valence bands in the vertical direction shown in Fig. 1. For each applied gate voltage, the band profile is computed with a Poisson-Schrödinger solver, which takes into account quantum confinement at the emitter, mass anisotropy in silicon conduction band and light and heavy holes. The computation is performed assuming that the tunnelling current is so low that the oxide separates two regions in local equilibrium with two different Fermi energies. Let us also consider the electrochemical potential of the nanocrystal μ_{NC} .

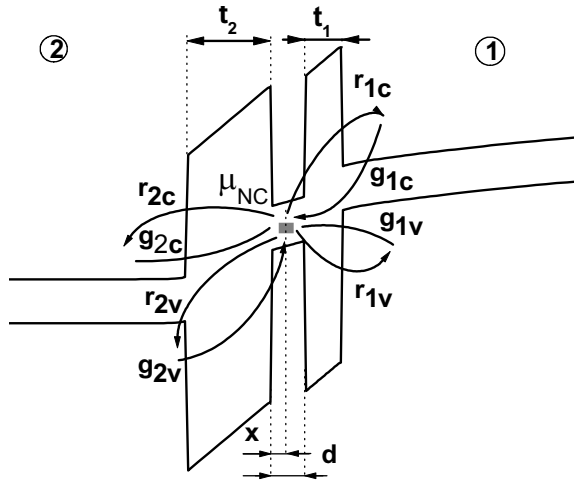


Fig. 1: Band profile of the structure used in this work and generation and recombination rates. μ_{NC} indicates the electrochemical potential of the nanocrystal.

We call generation rate the transition rate from an electrode to the nanocrystal, and recombination rate the transition rate from the nanocrystal to one electrode [12]. As we can see in Fig. 1 there are eight different generation and recombination rates, with an obvious meaning of the subscripts. We do not enter into the detail of the expressions here, referring the interested reader to [12].

Let us only stress that fact that we allow for inelastic transitions, from a state $|\beta\rangle$ in a contact to a (Kohn-Sham) state $|\alpha\rangle$ in the dot, by writing the transition rate, according to Fermi “golden rule”, as:

$$\nu_{\beta \rightarrow \alpha} = \frac{2\pi}{\hbar} |M(\beta, \alpha)|^2 h_{\Gamma}(E_{\alpha} - E_{\beta}) \quad (1)$$

where

$$h_{\Gamma}(E_{\alpha} - E_{\beta}) = \frac{\Gamma/\pi}{(E_{\alpha} - E_{\beta})^2 + \Gamma^2}$$

The larger Γ , the larger degree of inelasticity of the transitions. This assumption corresponds to define a trap cross section per unit energy as:

$$\sigma_{\alpha, \beta} = k \cdot h_{\Gamma}(E_{\alpha} - E_{\beta}) \quad (2)$$

where k is an unknown constant surface, to be determined via a fitting with experiments.

The computation of all generation and recombination rates can therefore be done with a unique unknown multiplying coefficient, i.e. k . By setting for the moment k to 1 m^2 , all time quantities will be obtained in units of seconds divided by $(k/1 \text{ m}^2)$.

Let us indicate with $g(n)$ the total generation rate of the n -th electron, i.e. the probability per unit time that the n -th electron is added to the nanocrystal summed over all possible transitions. Similarly, we define $r(n)$ as the total recombination rate of the n -th electron. If $P(n, t)$ is the probability that n electrons are in the nanocrystal at time t , we can write the following master equation:

$$\frac{dP(n, t)}{dt} = r(n+1)P(n+1, t) + g(n)P(n-1, t) - P(n, t)[r(n) + g(n+1)] \quad (3)$$

During the charging process (program operation), Equation (3) is numerically solved with the following boundary conditions:

$$\begin{cases} P(n, 0) = 0 & \forall n \geq 1 \\ P(0, 0) = 1 \end{cases}, \quad (4)$$

corresponding to zero electrons in the dot at time zero. We can therefore obtain the average number of electrons in the dot as a function of time as

$$\langle n(t) \rangle = \sum_{n=0}^{\infty} nP(n, t), \quad (5)$$

and the average threshold voltage shift as a function of time as

$$\langle \Delta V_T(t) \rangle = \gamma \langle n(t) \rangle \quad (6)$$

where

$$\gamma = \frac{q\rho_{NC}}{\epsilon_{ox}} \left(\frac{d\epsilon_{ox}}{4\epsilon_{Si}} + t_2 \right).$$

For the simulation of the discharging process (erase operation) the boundary conditions are:

$$\begin{cases} P(m, 0) = 1 \\ P(n, 0) = 0 & \forall n < m \end{cases} \quad (7)$$

where m is the initial number of electrons inside the dot, that we choose so that $m\gamma$ is equal to the initially programmed threshold voltage.

3. Results

We apply our model to a typical structure for which experiments of program/erase operations are available. In particular, consider the structure indicated with B in

Ref. [3]: it has control oxide thickness of 8 nm, tunnel oxide thickness 3.5 nm, average nanocrystal diameter 3.2 nm, nanocrystal density $2 \times 10^{11} \text{ cm}^{-2}$. The dependence of the threshold voltage on the voltage applied to the control gate for program and erase operations is shown in Fig. 2 [Fig.7 of Ref. [3]]. In the simulations, we consider a tunnelling effective mass of electrons in silicon oxide of $0.5 m_0$, and of holes of $0.3 m_0$.

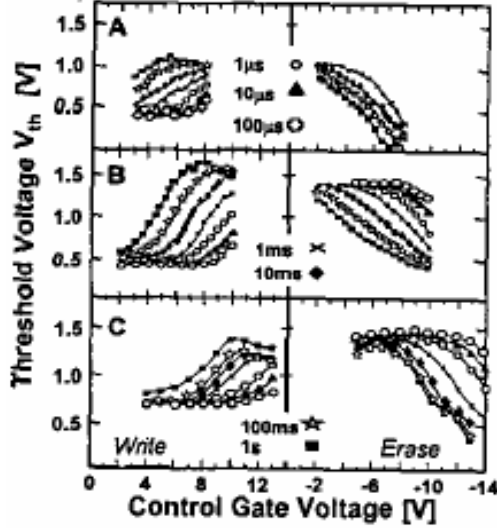


Fig. 2: Here we reproduce Fig. 7 of Ref. [12]. The experimental results corresponding to the structure we are considering in our paper are those indicated with B.

Let us first consider the results from a simulation performed considering the electrons in the conduction band. The single electron levels correspond to the ground state of the dot. As we can see in Fig. 3, simulations are very different from experiments: first, the peak of ΔV_T is obtained for gate voltages smaller than 4 V, while we can see in Fig. 2 that in experiments the peak of ΔV_T is obtained for gate voltages larger than 7 V. This behaviour is very robust with respect to variations of the values of meaningful parameters of the model. In addition, in Fig. 3 it is clear that the steady state is reached more quickly for gate voltages smaller than the peak, while in Fig. 2 is reached more quickly for gate voltages larger than the peak.

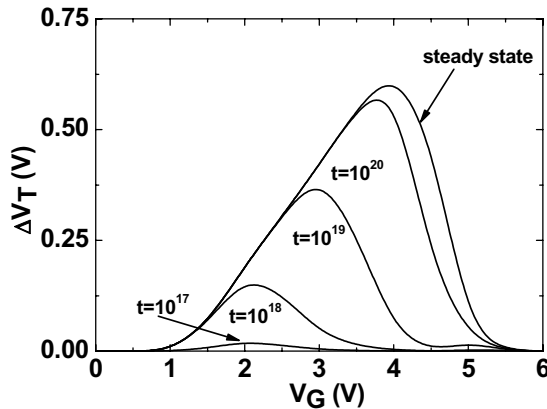


Fig. 3: Average threshold voltage obtained considering electrons in the conduction band of the dot. Time is in arbitrary unit.

The main reason for such behaviour is that the recombination rate r_{ic} dominates over all other rates already for a gate voltage of 1 V.

We have therefore investigated the program/erase behaviour for a series of trap positions and single electron energy levels, in particular shown in the inset of Fig. 4. Here we do not have the possibility to show all results: however, results very close to the experiments are obtained for trap energies 1 eV below the silicon conduction band and positioned in the center or towards the control oxide. The associated write characteristics for traps located in the center are shown in Fig. 5. Let us stress again the fact that the characteristics exhibit very evident qualitative changes as we modify trap position and energy, which cannot be recovered by simply adjusting the model parameters.

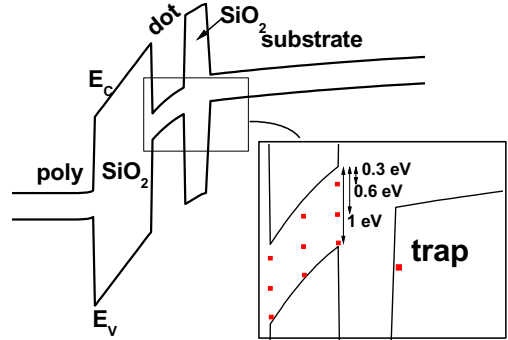


Fig. 4: Band profiles and (inset) nine trap positions and energies considered in the simulation.

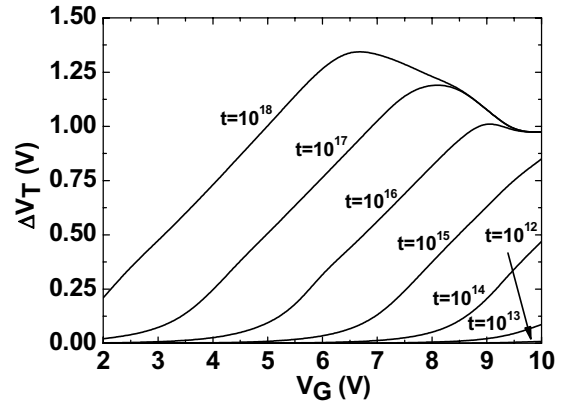


Fig. 5: Average threshold voltage versus gate programming voltage for different write times. Time is in arbitrary units.

For example, some models [8] take into account the shape of the dot by multiplying the transitions rates by appropriate geometry factors. However, we show in Fig. 6 that if we multiply the recombination rates by a factor β of 5 or $1/5$, leaving generation rates unaltered, the characteristics do not change from a qualitative point of view, because the dominant terms are represented by the tunnelling probabilities, that are exponentially dependent on the energy and the applied gate voltage.

As far as the erase operation is concerned, we may consider that two concurrent processes are possible: *i*) trapped electrons can tunnel out to the substrate and *ii*) trapped electrons can first be detrapped into the silicon conduction band, and then can tunnel out to the substrate. The relative importance of the two processes depends on the efficiency of the detrapping mechanism.

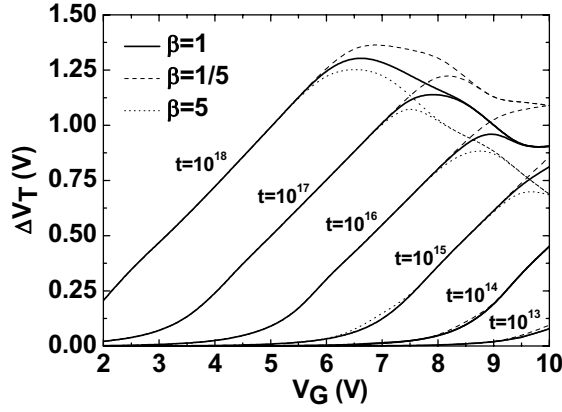


Fig. 7: Comparison between the write characteristics obtained for $\beta = 5, 1, \frac{1}{5}$. Time is in arbitrary units.

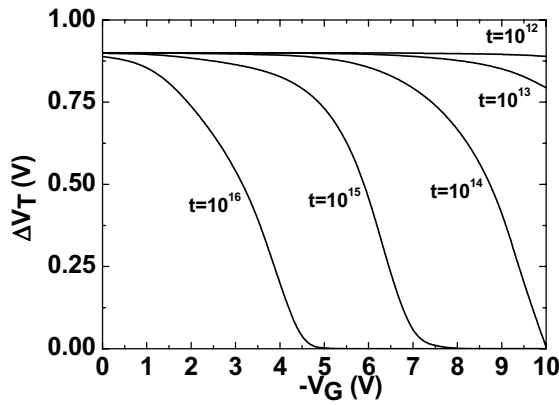


Fig. 8: Threshold voltage shift as a function of $-V_G$ for different erase times (in arbitrary units). Traps are located at the center of the nanocrystal layer, 1 eV below silicon conduction band.

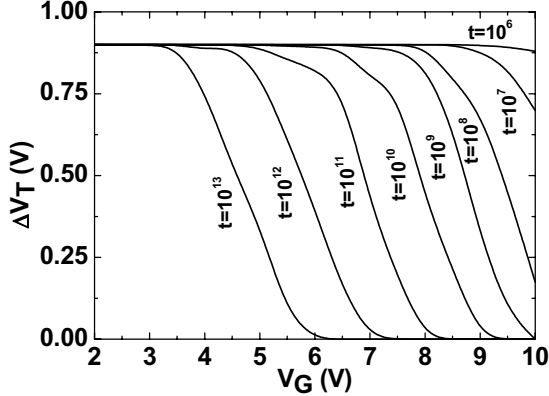


Fig. 9: Threshold voltage shift as a function of $-V_G$ for different erase times (in arbitrary units). Electrons are located in the nanocrystal conduction band.

However, we can consider the two limiting cases of inefficient detrapping, in which only process i is possible, and of fully efficient detrapping, in which all trapped electrons are moved to the conduction band of the silicon nanocrystal, and from there tunnel to the device substrate.

The erase characteristics associated to the former case are shown in Fig. 8: the initial ΔV_T is set to 0.9 V in agreement with Fig. 2. The erase characteristics associated to the latter case are shown in Fig. 9. The

actual situation is likely to be intermediate between the two, as it is the one of experiments in Fig. 2, but any conclusion requires further investigation, in order to evaluate detrapping efficiency.

4. Conclusion

Comparison of numerical simulations of program/erase characteristics of nanocrystal memories with experiments published in the literature allows us to exclude that electrons are stored in quantum confined states in the nanocrystal conduction band. Even if the model used is rather simplified and idealized, the results are robust enough to rule out such possibility.

We have investigated other possible positions and energies for traps located at the nanocrystal-silicon oxide interface, and have obtained very interesting results from the qualitative and quantitative point of view when traps are located 1 eV below silicon conduction band. Comparison with experiments allows us to obtain the value of the constant k and therefore to express the computed time in seconds, instead of arbitrary units. We obtain $k \approx 10^{17} \text{ m}^2 = 10^{13} \text{ cm}^2$, which means that the arbitrary units used for time are 10^{-17} s .

As a final remark, let us stress the fact that, notwithstanding its simplicity, the proposed model is, among those presented in the literature, the one that provides the more complete qualitative and quantitative agreement with experiments. Further investigations are needed, in order to achieve quantitative agreement with experiments also for the erase operation and for a broader set of experimental results.

Acknowledgements

The present work has been supported by IEIIT-CNR through the FIRB-MIUR project “Sistemi miniaturizzati per elettronica e fotonica”.

References

- [1] S. Tiwary *et al.*, IEDM Tech. Dig. Papers, 521 (1995).
- [2] J. De Blauwe, IEEE Trans. Nanotechnol. **1**, 72 (2002).
- [3] B. De Salvo *et al.*, IEDM, 597 (2003).
- [4] De Salvo *et al.*, IEEE Trans. Electron Devices **48**, 1719 (2001).
- [5] F. Rana, S. Tiwari, J.J. Welser, Superlatt. Microstruct **23**, 757 (1998).
- [6] J.S. de Sousa, A.V. Thean, J.P. Leburton and V.N. Freire, J. Appl. Phys. **92**, 6182 (2002).
- [7] G. Iannaccone, P. Coli, Appl. Phys. Lett.
- [8] C. Monzio Compagnoni *et al.*, IEDM, 549 (2003).
- [9] Y. H. Kwon *et al.* Appl. Phys. Lett. **80**, 2502 (2002).
- [10] E. Kapetanakis *et al.* Appl. Phys. Lett. **77**, 3450 (2000).
- [11] M. She and T.J. King, IEEE Trans. Electron Devices **50**, 1934 (2003).
- [12] G. Iannaccone, F. Crupi, B. Neri and S. Lombardo, IEEE Trans. Electron Devices **50**, 1363 (2003).
- [13] Y.T. Hou, M.F. Li, Y. Jin Proc. 6th Int. Conference of Solid-State and Integrated-Circuit Technology, **2**, pp. 22-25 (2001).
- [14] K.N. Yang *et al.*, IEEE Trans. Electron Devices **47**, 2161 (2000).

A Model for the Channel Potential of Charge-Trapping Memories and Its Implications for Device Scaling

M. Sodd, S.A. Anderson, B. Hradsky, R. Muralidhar, E.J. Prinz, R. Rao, S. Straub, R. F. Steimle, C.T. Swift, B. E. White, J.A. Yater

Freescale Semiconductor, 6501 W William Cannon Dr, Austin, TX 78735 USA, M.Sodd@freescale.com

Abstract

We extend a quasi two dimensional model for the channel potential at threshold as a function of an arbitrary distribution of trapped charge. The model agrees well with detailed numerical simulations, as long as the effective channel length is adjusted to model the uneven turn-on in the device. This formulation of the channel potential reveals that the trapped charge affects the threshold voltage through an averaging over lengths long enough that it should reduce the effect of variations in the density of trapped charge that is expected in memory devices at small dimensions.

1. Introduction

To achieve fast programming times and to enable multiple bit storage, a number of non-volatile device technologies utilize channel hot-electron programming into a dielectric storage medium, often either nitride [1] or nano-crystals [2,3]. The programmed charge distributions in such devices are non-uniform, with the charge typically confined within 20 nm of the drain [4]. In addition, in a short-channel device the threshold voltage (V_t) will respond more to drain bias than will the uncharged device, and the V_t will be asymmetric with respect to current direction--a property that forms the basis for multiple bit storage [1]. These properties may be modelled analytically with a quasi two dimensional calculation for the channel potential [5], as shown in [6] for the case of two-level charge distribution that is locally constant over the source and drain regions.

The narrow charged region present in these technologies leads to questions of how much the device characteristics will vary given natural deviations in programming characteristics. This issue is of particular concern in the nano-crystal device because the number and placement of the nano-crystals themselves is at least partially random [7,8]. In modern short-channel devices, however, charge in the dielectric stack affects the channel indirectly through an averaging process that mitigates the impact of charge variation on the channel potential. We develop an extension to this analytical model that comprehends this average effect of a non-uniform charge distribution of any shape, and use it to better understand the scaling of these charge trapping devices.

2. Analytical Model

Using the coordinate system shown in Fig. 1, the channel surface potential $\psi(y)$ in the quasi two-dimensional approximation is [5],

$$\ell^2 \frac{\partial^2 \psi(y)}{\partial^2 y} + V_{gs} - V_{t0} + 2\phi_F - \psi(y) = 0 \quad (1)$$

Here, $V_{t0} = t_{eff} q N_{sub} X_{dep} / \epsilon_{ox} + V_{fb} + 2\phi_F$ is the long channel threshold voltage, and $\ell^2 = \frac{X_{dep} \epsilon_{si}}{\epsilon_{ox} \eta} t_{eff}$ is a characteristic length scale. Here $\eta \sim 1$, but may be adjusted to account for variations in the depletion region [5]. This equation derives the lateral variation in $\psi(y)$ from the pill-box structure of depth X_{dep} , as shown in Fig. 1.

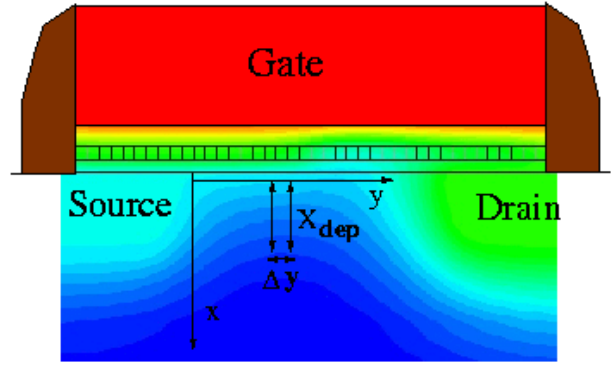


Fig. 1: Coordinate system for Eqs. 1-5.

By including a local threshold voltage shift, $\Delta V_t(y)$ a modified version of this equation includes the effect of non-uniform charge trapping,

$$\ell^2 \frac{\partial^2 \psi(y)}{\partial^2 y} + V_{gs} - V_{t0} - \Delta V_t(y) + 2\phi_F - \psi(y) = 0 \quad (2)$$

Here $\Delta V_t(y)$ is the local threshold voltage shift due to trapped charge at y .

This differential equation has the solution,

$$\begin{aligned} \psi(y) = & V^0(y) + [V_{bi} + V_{ds} - V^0(L_{eff})] \frac{\sinh(y/\ell)}{\sinh(L_{eff}/\ell)} \\ & + [V_{bi} - V^0(0)] \frac{\sinh((L_{eff} - y)/\ell)}{\sinh(L_{eff}/\ell)} \end{aligned} \quad (3)$$

where,

$$V^0(y) = V_{gs} - V_{T0} + 2\phi_F - \Delta V_{ave}(y), \quad (4)$$

$$\Delta V_{ave}(y) = \frac{1}{2\ell} \left[\int_0^y e^{-(y-y')/\ell} \Delta V_t(y') dy' + \int_y^{L_{eff}} e^{(y-y')/\ell} \Delta V_t(y') dy' \right] \quad (5)$$

Given this expression for the channel potential in terms of the trapped charge, the threshold voltage is determined by solving for the gate bias at which the potential minimum is at $2\phi_F$ [5].

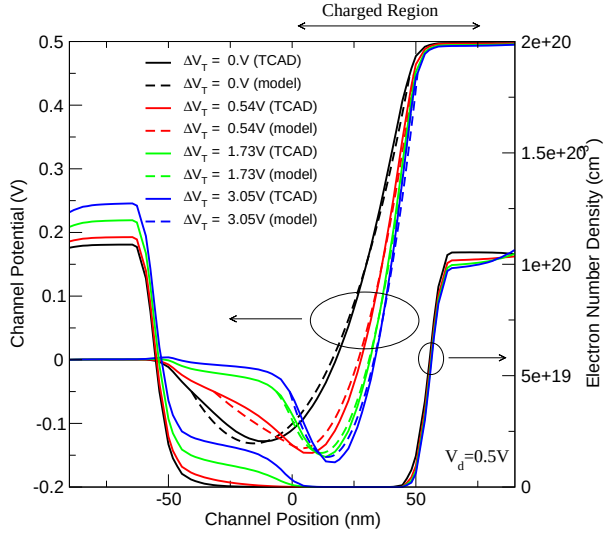


Fig. 2: Comparison between analytical model in Eqs. 3-5 and a numerical TCAD simulation with charge stored on the drain side and $V_{ds}=0.5V$. The device shown has an $L_{eff} = 100$ nm, $N_{sub} = 5.5 \times 10^{17} \text{ cm}^{-3}$, $T_{bot} = 50 \text{ \AA}$, $T_{top} = 100 \text{ \AA}$. The trapped charge is set to $0.2, 5, 8, \times 10^{12} \text{ cm}^{-2}$, which for uniform doping across the channel would result in V_t shifts of 0V, 1V, 2.5V, and 4V, respectively.

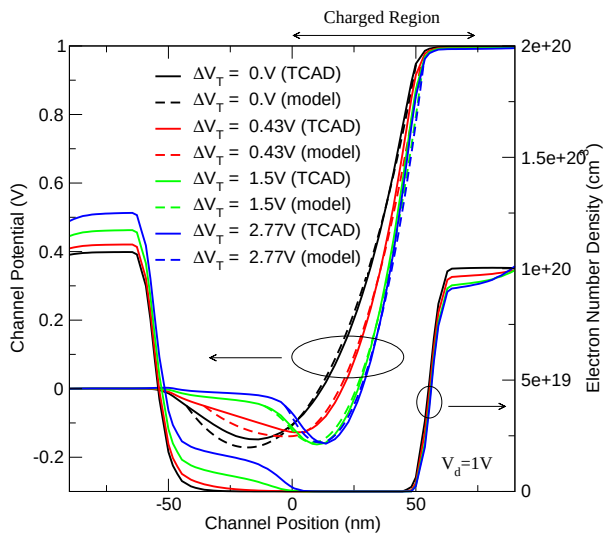


Fig. 3: Device of Fig. 2 with charge stored on the drain side and $V_{ds}=1V$.

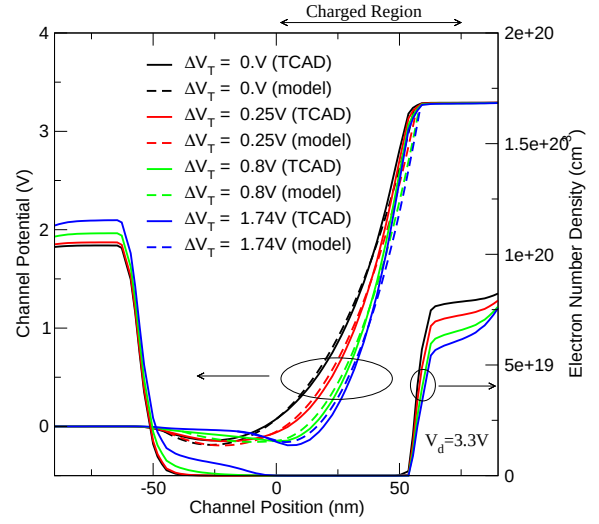


Fig. 4: Device of Fig. 2 with charge stored on the drain side and $V_{ds}=3.3V$.

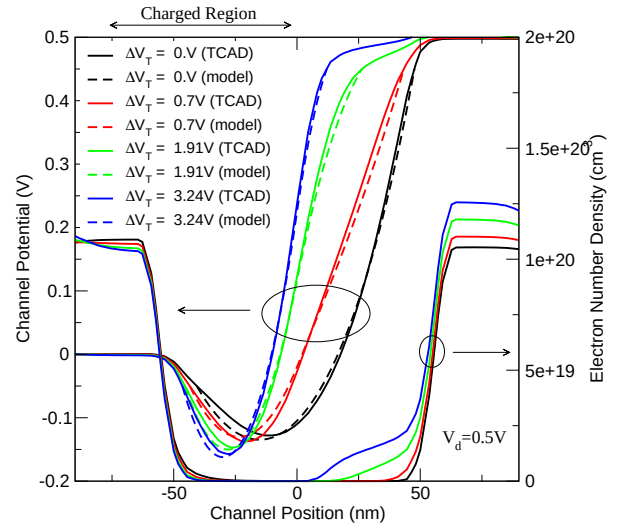


Fig. 5: Device of Fig. 2 with charge stored on the source side and $V_{ds}=0.5V$.

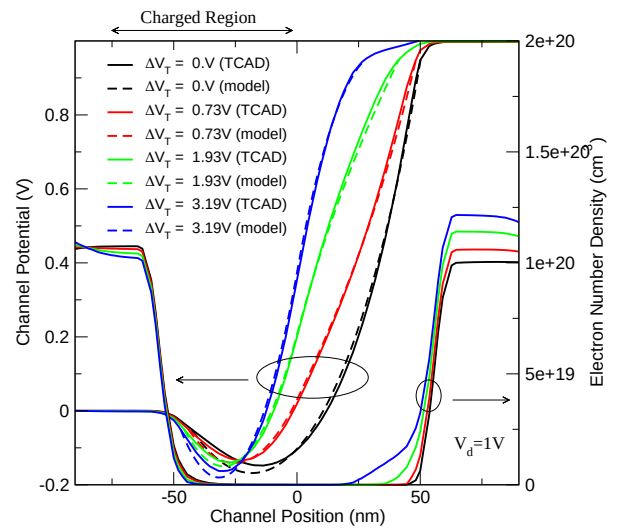


Fig. 6: Device of Fig. 2 with charge stored on the source side and $V_{ds}=1V$.

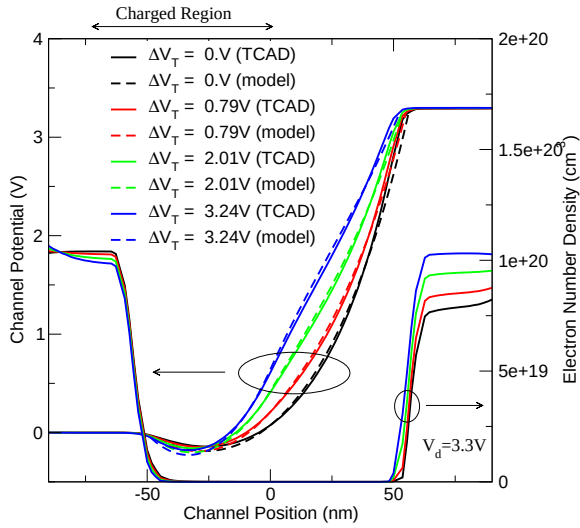


Fig. 7: Device of Fig. 2 with charge stored on the source side and $V_{ds}=3.3V$.

3. Numerical Comparison

Figs. 2-7 show a comparison between the channel potential from Eq. 3 and from detailed numerical calculations with a two dimensional device simulator [9]. Here the gate bias is placed at $V_{gs}=V_t$, and the drain bias and charging condition are as indicated. The analytical model and numerical calculations agree very well across the range of drain bias and charging conditions.

At a low drain bias $V_{ds} = 0.5V$, Figs. 2 and 5 show that the uncharged side of the device increasingly begins to invert as the other side becomes increasingly charged. Since the gate is placed here at $V_{gs}=V_t$, the neutral side sees an increasing gate over-drive relative to the neutral V_t of $V_{gs} - V_{t0} = \Delta V_t$, causing the uncharged region increasingly to invert. Within the model of Eq. 3, at threshold, this behaviour is effectively modelled as if the source or drain had extended further into the channel, so that the L_{eff} decreases. Taking into account this variation of the "effective" L_{eff} , and all other parameters set to their theoretical values, the value of the fit parameter η differs from 1 by less than 20% throughout the range shown.

Fig. 8 shows the variation of L_{eff} through the conditions in Fig. 2-7. It is seen that for source-side storage at high drain bias, the L_{eff} varies little, and Fig. 7 shows little electron density entering the drain side of the channel. In this case, the high drain to gate bias prevents inversion on the uncharged drain side--the device is effectively "pinched off" even at its threshold voltage.

4. Device scaling

A new feature of the model in Eq. 3 is that the effect of trapped charge is felt by the channel through $\Delta V_{ave}(y)$ that averages over the trapped charge distribution with a Green's function with a decay length of ℓ . For the typical case in Figs. 2-7, the length $\ell \sim 50$ nm. This effect results from the integration of the second-order differential Eq. 2, which also includes the competing effects of the boundary conditions at the source and drain.

For devices that trap charge in a gate dielectric, this result suggests that a small variation in trap charge distribution may not have a disproportionate effect on the device characteristics. In particular for nano-crystal memories, since the mean spacing ~ 10 nm [2] between nano-crystals is less than ℓ , the channel average in Eq. 3 is insensitive to variations in the nano-crystal placement. Such a property may become critical to controlling threshold voltage distributions as the device dimension scales, and the number of charged nano-crystals in each transistor can fall below 100 [8].

5. Conclusion

In addition to providing these insights, the formulation in Eq. 3 may be used to derive the threshold voltage as a function of drain bias and polarity, providing for a means to characterize the charge distribution [6]. The strength of this formulation is an understanding of the variation in "effective" L_{eff} and the sensitivity of charge trapping memories to variations in the trapped charge density. Intuitively, such averaging process should dominate over percolation leakage [7] as long as the device operates at a short enough channel and high enough drain bias. In that regime, the device conduction is modulated by source barrier, as opposed to the case in the linear regime in which the conduction is a simple function of the average channel charge immediately underneath the trapped charge. Since memory devices operate in the former regime, they should benefit from these reduced device fluctuation as the technology scales to smaller dimensions.

References

- [1] B. Eitan, *et. al. IEEE Elec. Dev. Lett.* **21** 543 (2000).
- [2] R. Muralidhar, *et. al. Trans. IEDM* **26**.2.1 (2003).
- [3] B. De Salvo, *et. al. Trans. IEDM* **26**.1.1 (2003).
- [4] E. Lusky, *et. al. IEEE Elec. Dev. Lett.* **22** 556 (2001).
- [5] Z. Liu, *et. al. IEEE Trans. Elec. Dev.* **40** 86 (1993).
- [6] L. Perniola, *et. al. Proc. 34th ESSDRC* 249 (2004).
- [7] C.M. Compagnoni, *et. al. Proc. IEEE Int. Rel. Phys. Symp.* 509 (2004).
- [8] L. Perniola, *et. al. IEEE Trans. Nanotech.* **2** 277 (2003).
- [9] ISE TCAD Manual, release 8.5.

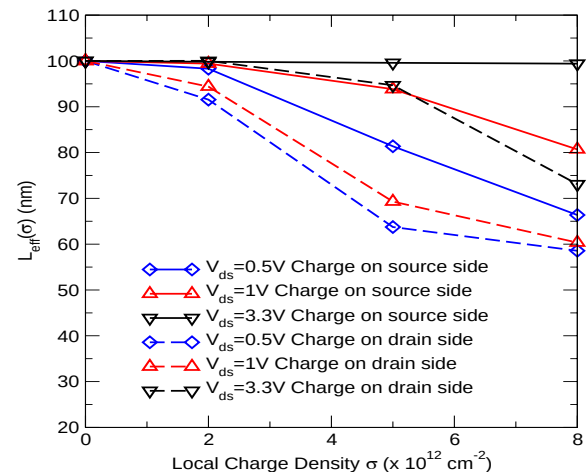


Fig. 8: Variation of L_{eff} with in Eq. 3 with peak charge density σ .

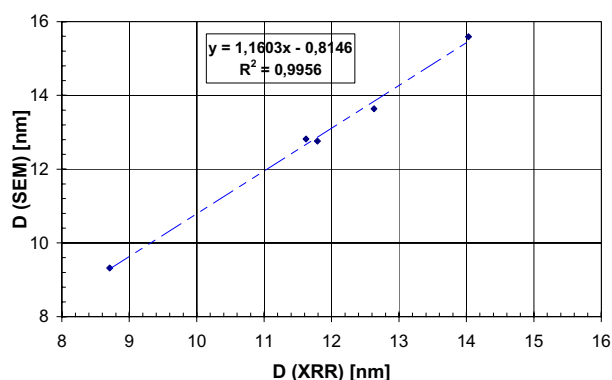


Fig.7 8: SEM vs. XRR dot size D

6. Conclusion

An in-line dot size distribution characterization protocol was tested, coupling density evaluation by light scattering analysis and dot size determination by X-Ray Reflectivity measurement. This fast (<10 minutes per wafer) in-line protocol gives very encouraging results, showing a very good correlation with SEM off-line measurement.

Interesting results, in terms of dots density and size, were obtained. The dot density varies between 10^{11} and $6 \times 10^{11} \text{ cm}^{-2}$, however with a wide size distribution, according to the SEM images. These results show the necessity of an in depth study for reliable in-line characterization, and optimise the test protocol for the control of the density and size distribution of nc-Si.

This work has been carried out, in the frame of CEA-LETI / CPMA collaboration, with PLATO Organization teams and tools

References

- [1] Impact of few electron phenomena on floating-gate memory reliability
Gabriel Molas, Damien Deleruyelle, Barbara De Salvo, Gérard Ghibaudo, Marc Gely, Stéphanie Jacob, Dominique Lafond, Simon Deleonibus, IEEE-International-Electron-Devices-Meeting-2004. 2004: Conference-Paper
- [2] Lee, J.J.; Wang, X.; Bai, W.; Lu, N.; Liu, J.; and Kwong, D.L.; Theoretical and Experimental Investigation of Si Nanocrystal Memory Device with HfO₂ High-k Tunneling Dielectric; Symposium of VLSI Technology Digest of Technical Papers, 2003
- [3] J. De Blauwe et al., A novel, aerosol-nanocrystal floating-gate for non-volatile memory applications, IEDM, 2000
- [4] A.Fernandes, B.DeSalvo, T.Baron, J.F.Damlencourta, A.M.Papona, D.Lafonda, D.Mariollea, B.Guillaumote, P.Bessonc, P.Massonb, G.Ghibaudo, G.Pananakakis, F.Martina and S.Haukkae, Memory Characteristics of Si Quantum Dot Devices with SiO₂/ALD Al₂O₃ Tunneling Dielectrics, IEDM 2001

[5] Baron,-T.; Fernandes,-A.; Damlencourt,-J.-F.; De-Salvo,-B.; Martin,-F.; Mazen,-F.; Haukka,-S.; Growth of Si nanocrystals on alumina and integration in memory devices; Applied-Physics-Letters. 9 June 2003; **82** (23): 4151-3.

[6] Bertin,-F.; Baron,-T.; Mariolle,-D.; Martin,-F.; Chabli,-A.; Dupuy,-M., Characterization of deposited nanocrystalline silicon by spectroscopic ellipsometry, Physica-Status-Solidi-A. 16 Sept. 1999; **175**(1): 405-12.

[7] Nicotra,-G.; Lombardo,-S.; Spinella,-C.; Ammendola,-G.; Gerardi,-C.; Demuro,-C.; Observation of the nucleation kinetics of Si quantum dots on SiO₂/sub 2/ by energy filtered transmission electron microscopy; Applied-Surface-Science. 31 Jan. 2003; **205**(1-4): 304-8.

[8] P. Donnadiou ; M. Verdier ; G. Berthomé ; P. Mur ; Phase retrieval applied to TEM images of nanodots; The thirteenth European Microscopy Congress – Anvers – Belgium – 22-27/08/2004 – Conference paper

DISCRETE MULTILEVEL CHARGING DESCRIPTION OF SILICON NANO-CRYSTALS MEMORIES

C. Busseret ^a, S. Ferraton ^b, L. Montès ^b, J. Zimmermann ^b

^a LPM, 20 av. A. EINSTEIN, 69621 Villeurbanne, France

^b IMEP, 23 rue des Martyrs, 38016 Grenoble, France

Phone: +33-476-856-035 E-mail: christophe.busseret@insa-lyon.fr

Abstract

We propose a new model to better describe nano-floating gate memories. The granular approach that we have developed overcomes the disadvantage of previous continuous floating gate descriptors : a fraction of an electron or hole inside the dots. Without any macroscopic approximations, our model improves greatly the charging kinetics simulations.

1. Introduction

With the objective to turn the limits of device downscaling away, novel structures have to be imagined, fabricated and optimized. In the field of the Non Volatile Memory devices, a solution consists in replacing the continuous Si-poly floating gate by a planar distribution of Silicon nanocrystals (Si-nc), each Si-nc acting as a storage node [1] (Figure 1). In this kind of device an improvement of reliability, plus a read/write process at moderate voltage are expected.

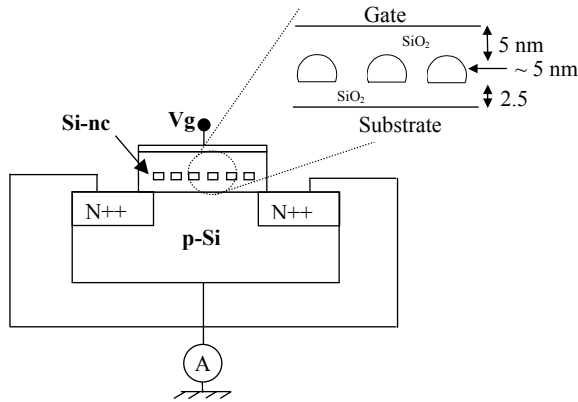


Figure 1 : Cross sectional schematic of a Non Volatile Memory device with a 2D layer of Si-nc acting as the floating gate.

2. The Floating gate model

We have measured the charging current at different gate voltages after a 10s negative stress. This was possible thanks to an original technique based on the feedback charge method [2]. The charging kinetics of the Si-nc have been rather accurately described by De Salvo [3] using a model borrowed from the conventional floating gate memory. However this approach is not suitable to incorporate the granular effects expected with Si-nc (electrons charging one by one) and this may involve important mistakes as can be seen in Figure 2. Indeed, the positive current observed at $V_g = -1.5V$ is only explained by more than one hole per dot after the -2V stress while C-V measurements have shown that this positive charge value is widely overestimated.

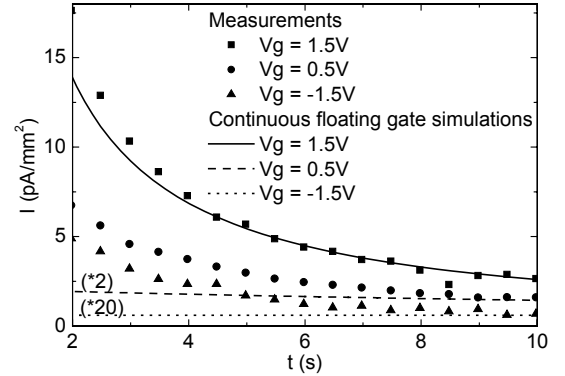


Figure 2 : Charging current measurements and simulations with a continuous floating gate model. Every curve can not be fitted with the same fitting parameters. We have focused here on $V_g = 1.5V$. The charging current at $-1.5V$ can be approached only with an erroneous amount of positive charges at the beginning. The previous 10 seconds stress at $-2V$ induces less than 0.5 hole per Si-nc.

3. The granular model

The aims of this work is therefore to propose an accurate granular model. The main disadvantage of the continuous floating gate model is that a fraction of an electron -or an hole- can be stored inside a dot. This of course is unacceptable. Therefore we suggest here a discrete charging levels description involving only integral numbers of carriers. The mathematical approach that we have developed allows a generalization to a N-level description. As shown in Figure 3a we will focus here on a three states Si-nc model: one hole (-1), one electron (1) or neutral (0). The arrows present the various paths with the corresponding charge quantities. n_i , J_i are the number of dots and the current density probabilities in the state i . S is the effective area of the dots related to the Si-nc diameters. The loading kinetics are described by three differential equations (Figure 3b) to be analytically resolved.

The Gauss law is used to calculate the different electric fields F_L , F_0 and F_I observed in the three states (Figure 4). These fields involve the tunnel current density probabilities J_i through the thin tunnel oxide. The initial conditions introduced in the solver are attributed to the negative stress applied before the measurement : a proportion x , in %, of the dots are originally occupied by one hole, $(100-x)$ are in the neutral state. Of course, there is no negative dot at the beginning. Moreover, we have introduced a Si-nc size distribution (Figure 5) in agreement with Scanning Electron Microscope and electrical ICMTD-2005

measurements [4]. The probability to be charged by a hole just after the negative stress is more important for the smallest dots because the electric field in the tunnel oxide is larger. The distribution of the dots in state -1 after the -2V stress period is given in Figure 5. Taking into account the different charging paths and the Si-nc size dispersion, the measured current J is calculated.

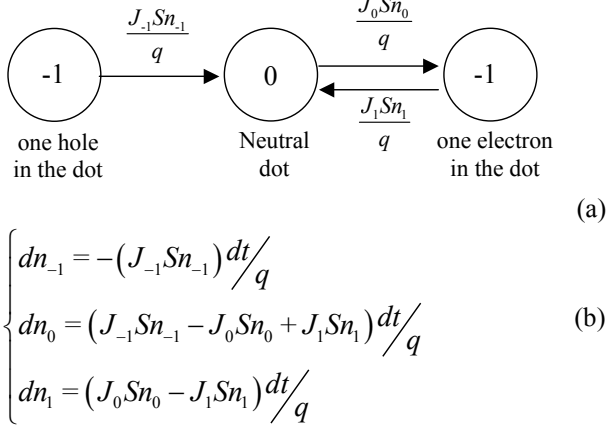


Figure 3 : (a) The three load states and the available transitions with their carriers flux at positive gate bias. (b) The fundamentals differential equations of the kinetics.

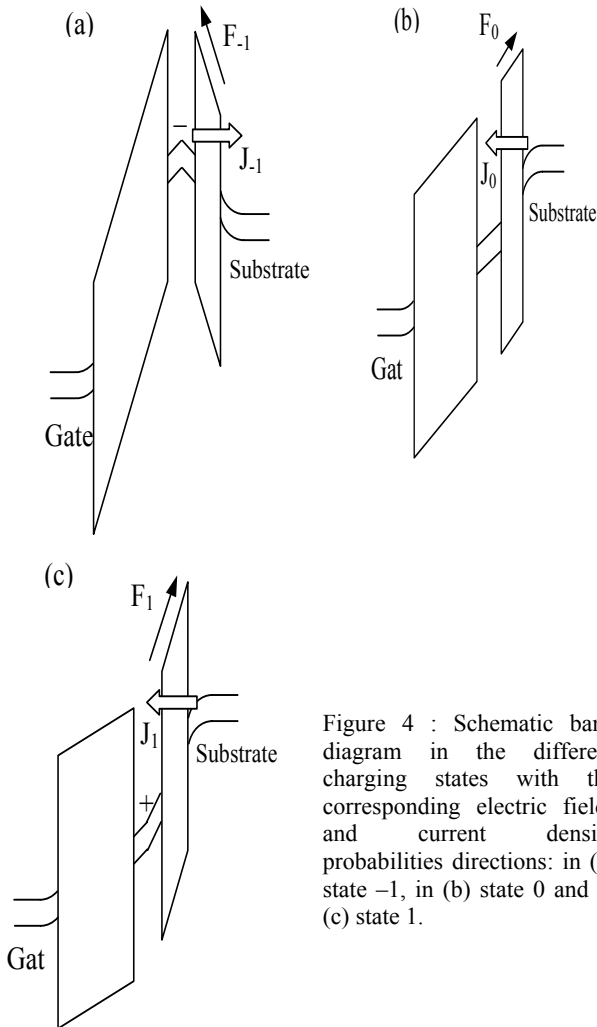


Figure 4 : Schematic band diagram in the different charging states with the corresponding electric fields and current density probabilities directions: in (a) state -1, in (b) state 0 and in (c) state 1.

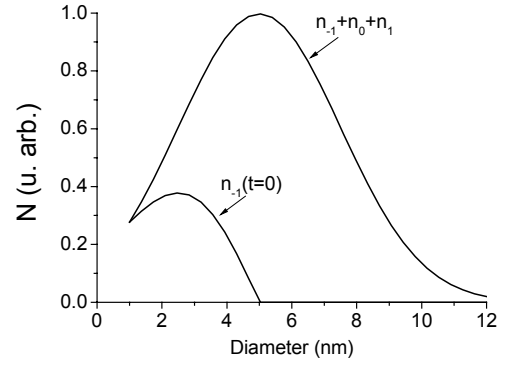


Figure 5 : Dot size distribution of the total population ($n_{-1}+n_0+n_1$) and the dot with one hole just before the current measurement.

Figure 6 shows the simulation results. The discrete multilevel charging approach describes almost exactly the charging kinetics with $x=20\%$ of the dots occupied by a hole at the beginning. The granular description of the stored charge can accurately model the local increase of the electric field under occupied dots and then the charging/discharging current.

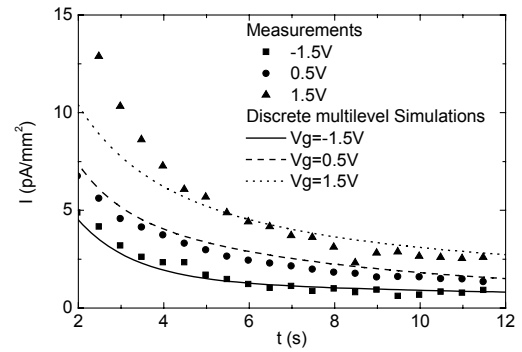


Figure 6 : Charging current measurements, and simulations using our discrete multilevel model.

To conclude, we demonstrate with our new nano-floating gate model that Si-nc kinetics are related to the number of dots occupied by one charge and not to the average number of charge per dot. This simple remark is very fruitful and enables an accurate description of nanocrystals memory with granular effects unpredictable with previous approaches. The interest of our model is enhanced by analytical leading equations that can be easily added in a simulator and directly applied to memory devices. It is fast and easy to handle.

[1] S. Tiwari, F. Rana, H. Hanafi, A. Hartstein, E. F. Crabbe, K. Chan, A silicon nanocrystals based memory, *Applied Physics Letters*, 68, 1996
[2] S. Ferraton, L. Montès, A. Souifi, J. Zimmermann, Charge dynamics of silicon nanocrystals in MOS capacitors, *Microelectronic Engineering*, 73-74, 2004

- [3] B. De Salvo, G. Ghibaudo, G. Pananakakis, *et al.*, Experimental and Theoretical Investigation of Nano-Crystal and Nitride-Trap Memory Devices, *IEEE Transactions on Electron Devices*, 48, 2001
- [4] T. Baron , F. Mazen , C. Busseret , A. Souifi , P. Mur, *et al.*, Nucleation control of CVD growth silicon nanocrystals for quantum devices, *Microelectronic Engineering*, 61-62, 2002

On the saturation mechanism in the Ge nanocrystals based non volatile memory

M. KANOUN^a, C. BUSSERET^a, T. BARON^b and A. SOUIFI^a

^a Laboratoire de Physique de La Matière, UMR-CNRS 5511, INSA De Lyon, Bât. Blaise Pascal, 7 avenue Jean Capelle 69621 Villeurbanne cedex, France

^b LTM/CNRS, LETI/DTS CEA/GRE, 17 Avenue des martyrs 38 054 Grenoble cedex 9, France

Abstract

The hole charging mechanism in germanium nanocrystals (nc-Ge) embedded in SiO₂ matrix fabricated by low pressure chemical vapour deposition is investigated by the means of capacitance – voltage (C-V) analysis. The charging kinetic shows a logarithmic behaviour and a saturation phenomenon for different gate voltage stress. For lower gate bias stress, the saturation of the carrier in the nc-Ge seems to be depending on electric field in the oxide. On the other hand, we have demonstrated that for higher gate voltage stress, the hole's number per dot at saturation becomes constant and depends only on the diameter of the nc-Ge. Finally we propose a model to simulate the saturation mechanism which permits us to extract the nanocrystals size and density.

1. Introduction

Recently many works have reported the interest of Flash memory embedded nanocrystals [1-3]. Indeed conventional memories with poly-Si floating gate are slow to program and have poor endurance because of the thick injection oxide [4]. In order to improve their characteristics, Tiwari *et al.* have proposed memories containing nanocrystals. This allows thinner oxide, and hence, lower consumption, better endurance, and faster writing/erasing time [1,5]. After the first proposal of memory containing Si nanocrystals (nc-Si), many works have confirmed that Non Volatile Memory (NVM) can be achieved using PMOS or NMOS transistor [1,2,3]. In order to improve data retention one can replace nc-Si by nc-Ge to take advantage of the smaller energy gap. Indeed King *et al* [6] have recently reported superior properties for nc-Ge based memory over those based on Si, in terms of writing/erasing time. Isolated nc-Ge in SiO₂ layers can be obtained by various techniques such as Ge implantation into SiO₂ films [7], oxidation and reduction of Ge/Si islands [8], UV-assisted oxidation of Si_{1-x}Ge_x alloys [9], and rapid thermal oxidation (RTO) of nc-Ge [10, 11, 12] or (Ge + SiO₂) [13]. In spite of the fact that nc-Ge based memory gives more promising characteristics, the works on

charging/discharging phenomena are focused on nc-Si based memory [14].

In this paper, we propose an extensive study of the saturation mechanism in the germanium nanocrystals fabricated by Low Pressure Chemical Vapor deposition (LPCVD). Constant gate voltage stress combined with capacitance versus voltage (C-V) have been done in order to evaluate the charging kinetic and to study the saturation mechanism. Finally we have proposed a model to simulate the saturation effect observed in the nc-Ge.

2. Experiments details

The substrates used in this study are p-type 200 mm Si (100) wafers covered by a thermal SiO₂ layer (tunnel oxide). Si nuclei and nc-Ge were grown in a single reduced pressure chemical vapour deposition (CVD) reactor designed by Applied Materials, in which the SiH₄ and GeH₄ gaseous precursors are diluted in H₂. The nc-Ge are grown using a two-step process. First, the SiO₂ surface is functionalized by the deposition of Si nuclei using SiH₄ as a gaseous precursor. Second, once the Si nuclei are deposited, we stop the SiH₄ gas flow, and GeH₄ is introduced to selectively grown on the Si nuclei [15]. Figure 1 shows a typical SEM image of nc-Ge embedded in SiO₂ using this process. The samples investigated in this study are labeled A, B, C, D and they are described in table 1. For the electrical measurements we have deposited a 250 nm-thick aluminum gate in order to obtain MOS capacitors. The devices have a square shape with surfaces ranging from 100 x 100 μm² to 400 x 400 μm². C-V and I-V curves have been recorded using respectively a HP 4284 and a HP 4156 respectively.

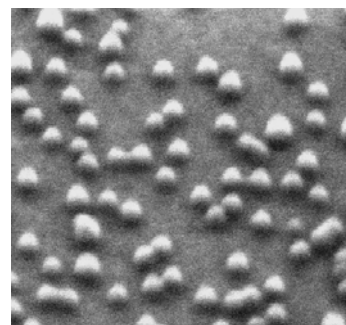


Fig.1: SEM picture of typical Ge quantum dots deposited on SiO₂ matrix with a density of $2.7 \times 10^{11} \text{ cm}^{-2}$. The mean sizes is equal to 8.5 nm

Sample	A	B	C	D
$d_{TO}(nm)$	2	2.5	2.5	2.5
$d_{CO}(nm)$	10	10	10	10
$d_{dot}(nm)$	8.5	8.5	8.5	6
$D (cm^{-2})$	$6 \times 10^{11} cm^{-2}$	$6 \times 10^{11} cm^{-2}$	$2.7 \times 10^{11} cm^{-2}$	$2.7 \times 10^{11} cm^{-2}$

Table 1: description of samples investigated in this study. D the nanocrystals density, d_{dot} the dot diameter d_{TO} and d_{CO} are the tunnel and control oxide thickness respectively.

2. Results

Figure 1 shows that a negative stress bias ($V_g > V_{FB}$) shifts the C-V characteristics on the right which indicates a hole charging in the nc-Ge. A positive stress shifts the C-V curves on the left in agreement with electrons trapping in the dots. The exact ΔV_{FB} for n charges per Ge dot is given by the following equation [1]:

$$\Delta V_{FB} = \frac{-q \cdot n \cdot D}{\epsilon_{ox}} \cdot \left(d_{CO} + \frac{1}{2} \cdot \frac{\epsilon_{ox}}{\epsilon_{Ge}} \cdot d_{dot} \right) \quad (1)$$

Where d_{CO} is the control oxide thickness, D is the density of the Ge dots, d_{dot} is the diameter of a Ge dot, ϵ_{Ge} and ϵ_{ox} are dielectric permittivities of Ge and SiO_2 respectively, and q is the magnitude of the electronic charge. We note that the reference sample without dot do not exhibit any flat band shift.

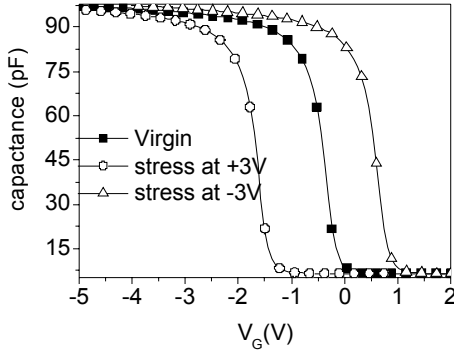


Fig. 2: C-V measurements after different stress. A negative stress shifts the C-V to the left indicating a hole charging of nc-Ge. A positive stress allows an electron charging.

In figure 3 we present the charging kinetic of positive carriers at different stress bias. We note that the kinetic is logarithmic, and it is faster when the stress gate voltage is larger. Figure 3 also shows that for a low bias stress ($V_g = -2V, -3V$ and $-4V$) the saturation is obtained after a longer time and the number of stored holes per dot is lower. The saturation is probably due to the trapped charges which produce a repulsive electric field. As we can observe on the same figure, the charge saturation is the same for $V_g = -6V$ and $-7V$. More generally, we have observed that above a certain gate bias stress, there is a limitation of the total carriers in the dots. We can distinguish two saturation domains: low voltage and high voltage stress.

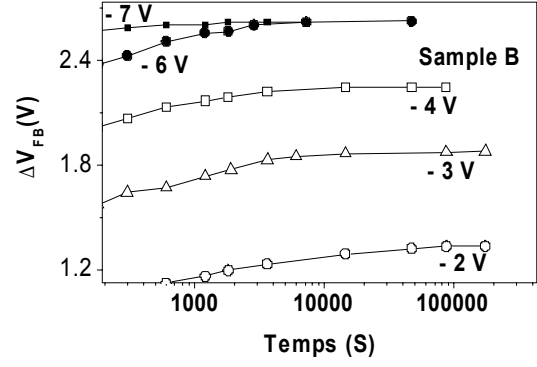


Fig. 3: Charging kinetic for sample A ($d_{dot} = 8.5$ nm, $d_{tun} = 2.5$ nm and $D = 6 \times 10^{11} cm^{-2}$) for different stress gate voltage.

3. Discussion

3.1 High voltage stress.

From our experiments, the saturation can be explained by two phenomena.

- Leakage current from the gate. Indeed when we apply a stress at a high voltage, the barrier becomes triangular, and then electrons can tunnel by Fowler-Nordheim mechanisms from gate towards the control oxide.
- When new electric charges are added in the dots, the energy levels will increase [16,17] (figure 4). When we reach a certain number of charges, the shift of the energy levels becomes very important. In such conditions we can not add new charges at low voltages because of the high misalignment between substrate's and dot's energy levels. The saturation is in that case due to a Coulomb blockade phenomenon. In this condition the saturation mechanism depend only on the diameter of the dot.

The control oxide is 10 nm-thick, so that one must apply a gate voltage higher than 10 V in order to obtain Fowler-Nordheim conditions. In addition, if electrons escape from the gate, they will recombine with stored holes in the dots. In such conditions, the number of trapped holes should decrease [18]. Indeed, the electron injection process is faster than that of hole injection. The electron current density is therefore more important than the hole's one, and as a consequence most of the trapped holes will recombine with injected electrons. The result of this phenomenon is the decrease of the initial stored charge [18]. In our case the number of carrier at saturation remains constant for different gate biases, so that this first assumption is not in agreement with our results. In the second model, the charge saturation does not depend on the gate stress voltage. Therefore, this explanation is in better agreement with our experiment data. In addition, we have used gates biases well below the Fowler-Nordheim threshold value.

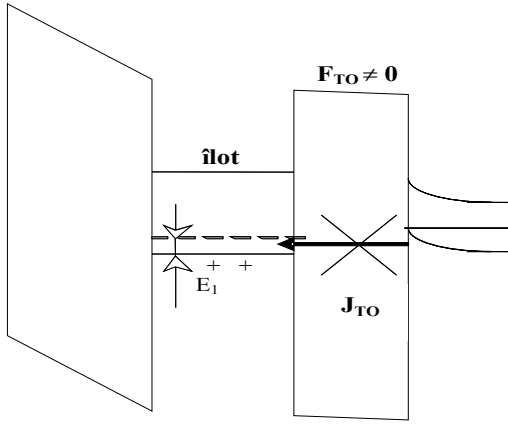


Fig.4 The band diagram explaining the saturation by Coulomb blockade.

In order to confirm this hypothesis we have investigated the saturation phenomenon for four different samples A, B, C, and D. As we can observe on the figure 5-b, the magnitude of the holes per dot at saturation remains the same for sample A ($d_{TO} = 2$ nm) and B ($d_{TO} = 2.5$ nm) which indicates that the tunnel oxide and then the electric field does not intervenes in this regime. On the other hand we clearly observe on figure 5-a that the number of holes at the saturation is lower for the smaller dot's diameter. Indeed, in a larger quantum dot the coulomb energy due to the adding of a charge is less important, so the misalignment is reached more difficulty. In addition those curves show clearly that it is not possible to stock more than 12 holes in a dot with $d_{dot} = 8.5$ nm and more than 9 holes with $d_{dot} = 6$ nm.

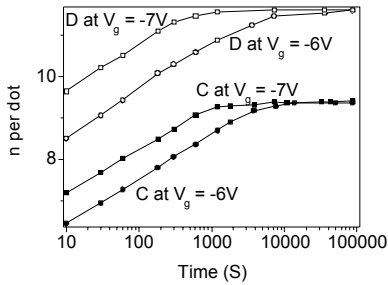


Fig.5-a: Charging kinetic for sample C ($d_{dot} = 6$ nm) and D ($d_{dot} = 8.5$ nm) showing that the saturation depends on the dot's diameter.

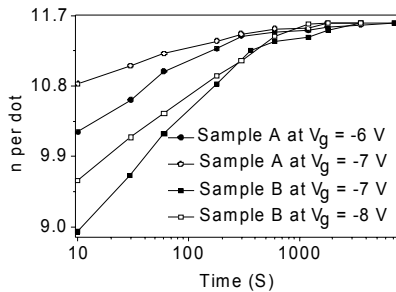


Fig.5-b: Charging kinetic for sample A ($d_{TO} = 2$ nm) and B ($d_{TO} = 2.5$ nm) showing that the saturation does not depend on the tunnel oxide thickness for the same dot's diameter.

3.2 Low voltage stress.

The charging effects in dots are attributed to the direct tunneling of the carriers through the tunnel oxide.

This current strongly depends on the electric field. When a hole is stocked in the dot it creates an electric field which screens the following of other charges from the substrate. In such conditions the total electric field decreases drastically. The saturation is obtained when this electric field becomes null and then the tunneling current is equal to zero. That's why the saturation charge increases with the electric stress applied to the gate. In fact, more charges are needed to compensate the higher electric field.

If we take into account of the charge stored in the nc-Ge, the electric field in the tunnel oxide is given by:

$$F_{TO}(V_g) = \frac{V_g - V_{fb} - \Psi_s - \frac{\Delta V_{FB}}{\alpha}}{d_{CO} + d_{TO} + \frac{\epsilon_{SiO_2}}{\epsilon_{Ge}} \cdot d_{dot}} = 0 \quad (2)$$

Where Ψ_s is the substrate band-bending and α is a geometric factor that corresponds to the area occupied by dots divided by the capacitor area. ϵ_{SiO_2} and ϵ_{Si} are the dielectric constant of the oxide and the silicon respectively.

As we said previously, the saturation takes place when there is no tunnel current through the tunnel oxide and the corresponding electric field is null. In such condition the equation (2) gives:

$$\Delta V_{fb/sat} = A \cdot V_g + B \quad (3)$$

With

$$\begin{cases} A = \alpha \\ B = -\alpha \cdot (V_{fb} + \Psi_s) \end{cases}$$

Now, the electric field in the tunnel oxide is null. So, the substrate band-bending equals zero and, the equation (3) is reduced to:

$$\Delta V_{fb/sat} = A \cdot V_g + B \quad \text{with} \quad \begin{cases} A = \alpha \\ B = -\alpha \cdot V_{fb} \end{cases} \quad (4)$$

3.3 Simulation results for low voltage stress.

If we refer to the equation (3), we conclude that the behaviour of the flat band voltage at saturation is linear as a function of the gate voltage stress value.

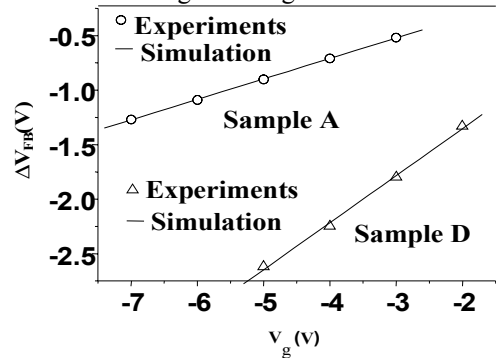


Fig.6: Flat-band shift ΔV_{fb} at saturation versus the gate bias V_g for samples A and D. The lines correspond to the simulation and the extracted parameter (slope) is given in table II.

The α parameter extracted from the slope of the figure 6 (α_{exp}) for the sample A and B is reported in the table 2 and compared to the one deduced from the SEM observations. We can note the good agreement of both values (extracted from the slope of the ΔV_{FB} (Vg) curve and from SEM analysis). This enables us to conclude that this model is accurate to describe the saturation process at low electric field. Moreover, this model allows to easily estimating the area occupied by the dots.

	A	D
α_{exp}	0.38	0.18
α_{SEM}	0.40	0.17

Table 2: Comparison of the parameters α obtained by charging curves (α_{exp}) and by microscopy measurement (α_{SEM}).

Conclusion:

In this paper, we have studied the saturation in the nc-Ge. We have demonstrated that for the saturation phenomenon we can distinguish two domains: at low gate bias stress attributed to the cancellation of the electric field through the tunnel oxide, and for high gate voltage stress attributed to a Coulomb blockade process. Finally we have proposed a model for the saturation at low V_g . The simulations results have shown good agreements between the experiment and the theoretical data. This model enables us to estimate the area occupied by the dots. Finally, as the measurement of the maximum number of stored holes allows extracting the dot's diameter, it is possible to extract the dot density by assuming a spherical shape.

References

- ¹S. Tiwari, F.Rana, H. Hanafi, A. Hrtstein, E. F. Crabbé, and K. Chan, Appl. Phys. Lett. **68**, 1377 (1996).
- ²K. Han, I. Kim, H. Shin, Proc. ISDRS'99 Charlottesville, VA, USA, 73 (1999).

- ³B. De Salvo, G. Ghibaudo, G. Pananakakis, B. Guillaumot and T. Baron, Superlattices and Microstructures, **28**, 339 (2000).
- ⁴Y. King, "thin dielectric technology and memory devices" PhD. Dissertation, Univ. California, Berkeley, CA, 1999.
- ⁵S. Tiwari, F.Rana, K. Chan, L. Shi, and H. Hanafi, Appl. Phys. Lett. **69**, 1232 (1996).
- ⁶Y. C. King, T. J. King, and C. Hu, IEEE trans. Electron Devices **48**, 696 (2001).
- ⁷H. I. Hanafi, S. Tiwari, I Khan, Electron Devices IEEE Transactions on, **43**, 9, 1553 (1996).
- ⁸T. Sass, V. Zela, A. Gustafsson, I. Pietzonka, and W. Seifert, Appl. Phys. Lett. **81**, 3455 (2002).
- ⁹V. Craciun, I. W. Boyd, A. H. Reader, W. J. Kersten, F. J. G. Hakkens, P. H. Oosting, and D. E. W. Vandenhoude, J. Appl. Phys. **75**, 1972 (1994).
- ¹⁰A. Kanjilal, J. Lundsgaard Hansen, P. Gaiduk, A. Nylandsted Larsen, N. Cherkashin, A. Claverie, P. Normand, E. Kapelanakis, D. Skarlatos, and D. Tsoukalas, Appl. Phys. Lett. **82**, 1212 (2003).
- ¹¹H. Fukuda, T. Kobayashi, T. Endoh, S. Nomura, A. Sakai and Y. Ueda, Applied Surface Science, **130-132**, 776 (1998).
- ¹²Y. Kim, H. J. Cheong, K. H. Park, T. H. Chung, H. J. Bark, J. Yi, S. H. Bang, J. H. Cho, Semicond. Sci. Technol. **17**, 1039 (2002).
- ¹³W. K. Choi, W. K. Chim, C. L. Heng, L. W. Teo, Vincent Ho, V. Ng, D. A. Antoniadis, and E. A. Fitzgerald, Appl. Phys. Lett. **80**, 2014 (2002).
- ¹⁴Yong Kim, Kyung Hwa Park, Tae Hun Chung, Hong Jun Bark, Jae-Yel Yi, Won Chel Choi, Eun Kyu Kim, Ju Wook Lee, and Jeong Yong Lee. Appl. Phys. Lett. **78**, 934 (2001).
- ¹⁵T. Baron, B. Pelissier, L. Perniola, F. Mazen, J. M. Hartmann, and G. Rolland. Appl. Phys. Lett. **83**, 1444 (2003).
- ¹⁶Johann Sée, Philippe Dollfus, Sylvie Galdin, and Patrice Hesto. J. Appl. Phys. **94**, 5053 (2003).
- ¹⁷G. Allan, C. Delerue, M. Lannoo, and E. Martin. Phys. Rev. B **52**, 11982–11988 (1995).
- ¹⁸C. Busseret, « Etudes optiques et électriques des propriétés électroniques de nano-cristaux de silicium pour composants mono-electroniques ». PhD dissertation INSA Lyon, France (1999).

Long Charge Retention Time in Novel Memory Devices Based on Semiconductor Quantum Dots

C. Balocco, S. W. Lin, M. Missous, and A. M. Song

School of Electrical and Electronic Engineering, University of Manchester, P. O. Box 88,

Manchester M60 1QD, United Kingdom

Email: A.Song@manchester.ac.uk

Abstract

Semiconductor quantum dot (QD) structures, grown by molecular beam epitaxy, have been used to fabricate a novel type of memory devices. Either one or five stacking layer(s) of quantum dots were embedded in otherwise standard high electron-mobility transistor structures. Memory operations were realised by electrons transferring from and to the two-dimensional electron layer. Different from previous attempts, however, our memory devices not only operated at room temperature but also were fully controlled by gate biases, without the need for light excitations to switch memory states. We have obtained charge retention times in the order of minutes, which are orders of magnitude longer than that of a standard DRAM. From the experimental results, we conclude that the long charge retention time can only be explained by charging and discharging deep levels that coexist in the quantum dot layer, rather than the intrinsic quantum dot levels.

The devices are, to the best of our knowledge, the first room-temperature, fully gate-bias-controlled memories based on InAs quantum dot structures. Because of the great flexibility of energy-band engineering in such compound semiconductor structures, much greater charge retention time or even non-volatile memories could be realised by improved designs.

1. Introduction

Self-assembled semiconductor quantum dots (QDs), created by epitaxial growth using Stranski–Krastanow mode, have been intensively studied because of their importance in device applications and considerable interests in fundamental physics. While so far most investigations focused on optoelectronic devices such as lasers, LED, and photodetectors, the possibilities to design QD-based memory devices have recently attracted more and more attention. Because of the single-electron charging/discharging effect in self-assembled QDs, the electric current and therefore power consumption required for memory operations could be substantially lower than in a conventional memory cell. Whereas a cluster of defects in a conventional dynamic memory or a flash memory could cause a rapid discharge of the whole capacitor plate or charge-storage floating gate, such a problem is absent in QD-based memory devices. An accidental discharge in one or more individual dots would not affect the charges in other dots since they are spatially separated from each other [1]. Moreover, these devices do not require hot electrons to

charge the QDs, allowing low operation voltages to be used [2].

A simple design of such memory devices is to embed quantum dots in the close vicinity to the two-dimensional electron gas (2DEG) in a high electron-mobility transistor (HEMT) structure, and charges may be controlled by gate biases to transfer between the quantum dot layer and 2DEG. In this way, the device fabrication and operations are entirely compatible to the established HEMT technology. Despite much effort, so far most such devices were demonstrated only at low temperatures [3–6]. Although Koike et al recently presented memory operations at room temperature, the devices required a light excitation to switch memory states [7]. From the application point of view, however, both room-temperature and fully electrically controlled memory operations are essential. Furthermore, compared with many measurements on hysteresis in electron density [3,4], source-drain current [5,7], and capacitance [8], hardly any time-domain measurement has been carried out to characterise the memory holding time, despite it being one of the most important parameters of memory devices.

In this work, we fabricated memory devices using a few different pseudomorphic InGaAs/GaAs/AlGaAs HEMT structures, each with one or five stacked layer(s) of InAs quantum dots. An atomic force microscope (AFM) was used to examine the dot density and room temperature photoluminescence (PL) measurements were carried out to study the carrier confinement energies in the dots, before the fabrications of the memory devices. Our experiments not only demonstrated both room-temperature and fully electrically controlled memory devices, which are, to the best of our knowledge, the first of this type, but also showed a remarkable memory holding time of tens of minutes in real-time characterisation measurements. Our gate-voltage dependence experiments also revealed that deep levels co-existed with the QDs and resulted in the long charge retention time.

2. Results and discussions

Figure 1 shows the schematics of the two wafers grown on rotating substrates in a Thermo VG V90 molecular beam epitaxy growth chamber. Note that wafer #1681 is an inverted, modulation-doped quantum well structure so that the stacked QD layers are much closer to the 2DEG than in wafer #1680. Figure 2 shows a well-defined QD PL peak at about 1015 nm at room

temperature, confirming the presence of the quantum dots. From AFM measurements on an uncapped control wafer (a typical image shown in the inset of Fig. 2), a dot density of about $6 \times 10^9 \text{ cm}^{-2}$ was obtained. In wafer #1681, largely because of the inverted structure, similar PL signals from the InGaAs quantum well were clearly observed. Two more wafers were also grown. In particular, wafer #1693 is similar to #1681 but has only a single QD layer and consists of a GaAs (rather than AlGaAs) barrier between the InGaAs quantum well and the QD layer.

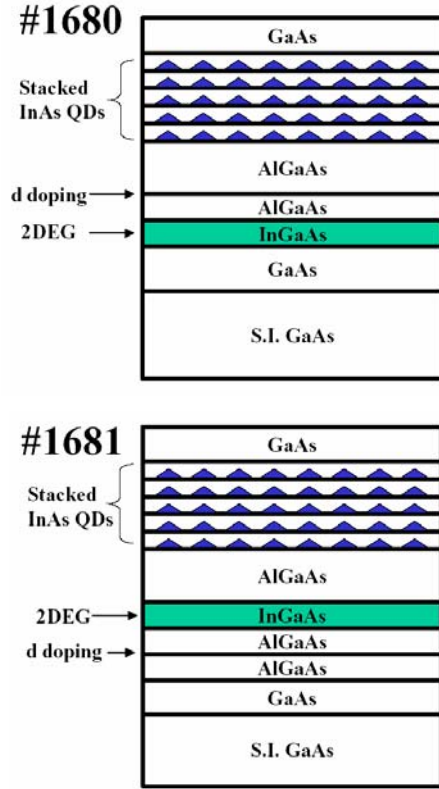


Fig. 1: Schematics of the structures of wafers #1680 and #1681.

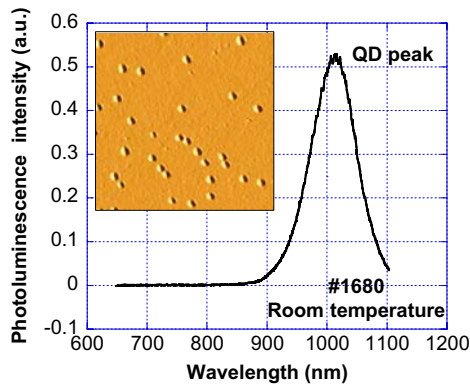


Fig. 2: Photoluminescence spectrum from as grown wafer #1680 at room temperature. The inset shows an atomic force micrograph of quantum dots on an uncapped sample.

The fabrication of the memory devices is essentially the same as that of a conventional HEMT, using a

standard photolithography combined with wet chemical etching. A large number of devices were fabricated from these wafers, with the gate length ranging from 2 μm to 5 μm and the gate width from 10 μm to 100 μm . The transistor characteristic of a typical device fabricated from wafer #1681 is shown in Fig. 3, which is similar to result of the devices from wafer #1680. However, in the memory operation measurements, devices from wafer #1681 showed a much stronger memory effect than these fabricated from wafer #1680. This is expected because of the difference in the distance between the QD layer and 2DEG [9]. In a control wafer where no QDs were embedded, no memory effect was observed.

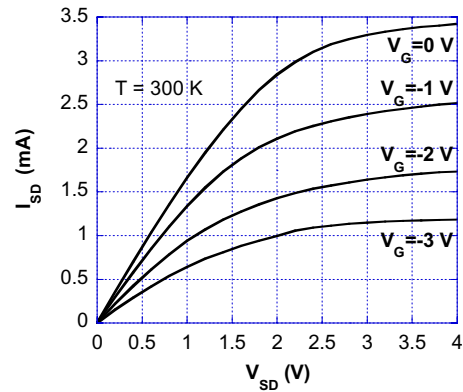


Fig. 3: Typical transistor characteristics of a device fabricated from wafer #1681 with a $5 \mu\text{m} \times 10 \mu\text{m}$ gate.

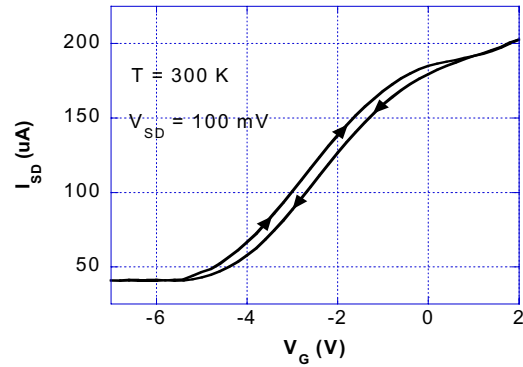


Fig. 4: Memory effect shown by the clear hysteresis in the I_{SD} vs V_G curve, which was obtained with a fixed $V_{SD} = 100 \text{ mV}$ at room temperature in a device fabricated from wafer #1681.

Figure 4 presents a pronounced hysteresis in the source-drain current, I_{SD} , versus the gate bias, V_G , curve, measured with a fixed source-drain voltage $V_{SD} = 100 \text{ mV}$ at room temperature. The presence of the hysteresis indicates a memory effect in the device by electron transferring between the 2DEG and QD layers. To better characterise the memory operations, particularly the charge retention time, we performed real-time experiments. The results of devices fabricated from wafers #1693 and #1681 are shown in Fig. 5. The two memory states, "1" and "0", were read by measuring the source-drain current I_{SD} at a source-drain voltage of 100 mV, after applying different gate voltage pulses. The source-drain current was higher, defined as memory state "1", after each

gate pulse of -4 V. If, however, a positive gate pulse of 2 V was applied, the memory state was switched to “0” with a lower source-drain current, as was observed in both devices in Fig. 5.

The higher source-drain current at memory state “1” was mainly due to a higher carrier density of the 2DEG, after the negative gate pulse of -4 V pushing electrons back from the quantum dot layer(s) in which a number of electrons were previously trapped. We note that after reaching memory state “1”, the source-drain current reduces with time, indicating that electrons tend to go back to the quantum dot layer(s) in the absence of a gate bias. The relaxation process in Fig. 5 allows for a good measure of the charge retention time, which is about tens of seconds and some minutes for devices fabricated from wafers #1693 and #1681, respectively. The AlGaAs barrier between the dot layer and 2DEG in #1681, which is higher than the GaAs barrier in #1693, clearly resulted in the longer charge retention time. Although this is not yet comparable to flash memories (which the devices were not yet designed for), the charge retention time is about three to four orders of magnitude longer than that of a typical DRAM. Because of the great flexibility of energy-band engineering in such compound semiconductor structures, much greater charge retention time or even non-volatile memories could be realised by improved designs.

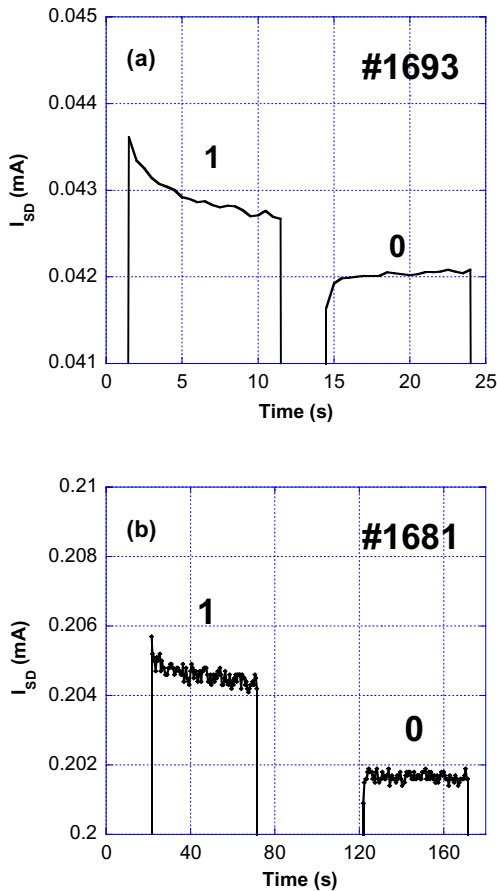


Fig. 5: The real-time measurements at room temperature, showing memory operations in devices made from wafers #1693 and #1681. The two memory states, “1” and “0”, were read by measuring the source-drain current I_{SD} at a source-drain voltage of 100 mV, after applying different gate voltage pulses to control memory states. After applying a gate pulse of

-4 V, the source-drain current was found higher, referred to as memory state “1”. After each gate pulse of +2 V, the memory state was switched to “0” or a lower drain current state.

The conductance difference between memory “1” state and “0” state, evaluated from Fig. 5, is between 4% and 2%. This is yet too small for any practical applications, but is predominantly due to the very high 2DEG density, around $1.5 \times 10^{12} \text{ cm}^{-2}$. It is much higher than the amount of electrons transferred to and from the QD layer. If the doping is reduced during the wafer growth to reach a 2DEG density of about $1 \times 10^{11} \text{ cm}^{-2}$, the conductance difference in the two memory states should reach the level of about or beyond 50%. It certainly can be further improved by increasing the density of the charge storage states in the QD layer(s). We expect that with optimised designs, it is quite feasible for the source-drain conductance to reach zero when the memory state is “0”, providing ideal operations.

One of the most important issues for future improvements as well as for device physics is to identify which electron states in the QD layer(s) are responsible for the memory operations. It has been suggested that intrinsic quantum dot states could lead to such memory effect [3-6]. However, it is noted that memory state “0” was stable in Fig. 5, indicating that electrons tend to leave the 2DEG and go to the QD layer. This means that the electronic states in the QD layer that are responsible for the memory effect are below the Fermi energy in the absence of a gate bias. A self-consistent calculation of the conduction-band profile was then carried out, and we found that the intrinsic quantum dot states were actually well above the Fermi energy, meaning that they did not contribute to the memory operations in these structures. Actually, the long charge retention time at room temperature can only be from rather deep states (at least 500 meV) rather than the intrinsic quantum dot states, which are only up to 200 meV below the conduction band bottom of the GaAs barrier. Systematic gate-bias dependence experiments recently showed that deep levels co-existing in the QD layer(s) were responsible for the memory operation and led to the long charge retention time [9,10]. The deep levels were most likely induced by the strain during the Stranski–Krastanow growth process due to the strong lattice mismatch [11].

3. Conclusions

We have, to the best of our knowledge, realised the first memory devices based on InAs QD structures that were not only fully electrically controlled at room temperature, but also fully compatible with the established HEMT technology. We have also discussed that both the charge retention time and the difference between the two memory states can be greatly improved, and the devices may be promising in real applications.

4. Acknowledgements

This work was supported by the Engineering and Physical Science Research Council (EPSRC) GR/S17741/01. We thank A. R. Peaker and O. Engström for valuable discussions.

References

- [1] S. Lombardo, B. D. Salvo, C. Gerardi, and T. Baron, *Microelectron. Eng.* **72**, 388 (2004).
- [2] I. Crupi, D. Corso, S. Lombardo, C. Gerardi, G. Ammendola, G. Nicotra, C. Spinella, E. Rimini, and M. Melanotte, *Mater. Sci. Eng.*, **C23**, 33 (2003).
- [3] G. Yusa and H. Sakaki, *Appl. Phys. Lett.* **70**, 345 (1997).
- [4] G. Yusa and H. Sakaki, *Superlatt. Microstr.* **25**, 247 (1999)
- [5] H. Kim, T. Noda, T. Kawazu, and H. Sakaki, *Jpn. J. Appl. Phys.* **39**, 7100 (2000).
- [6] A. Schliemann, L. Worschech, S. Reitzenstein, S. Kaiser, and A. Forchel, *Appl. Phys. Lett.* **81**, 2115 (2002).
- [7] K. Koike, K. Saitoh, S. Li, S. Sasa, M. Inoue, and M. Yano, *Appl. Phys. Lett.* **76**, 1464 (2000).
- [8] H. Son, J. Kim, M. Kim, and S. Hong, *Jpn. J. Appl. Phys.* **40**, 2801 (2001).
- [9] C. Balocco, A. M. Song, and M. Missous, *Appl. Phys. Lett.* **85**, 5911 (2004).
- [10] S. W. Lin, et al, unpublished.
- [11] C. Walther, J. Bollmann, H. Kissel, H. Kirmse, W. Neumann, and W. T. Masselink, *Physica B* 273–274, 971 (1999).

SESSION B

DRAM, SRAM and Soft-Error-Rate

Chairmen: **Atila ALVANDPOUR**, Linköping University, Sweden (Linköping)
 Pierre FAZAN, Innovative Silicon, Switzerland (Lauzane)

DRAM Cell Trends

Pierre C Fazan, S. Okhonin, M. Nagoga

Innovative Silicon Inc., PSE-B, CH-1015 Lausanne, Switzerland,

Tel. +41-21-693 8991, email: pfazan@z-ram.com

Abstract

In this paper we review the various trends in DRAM memory cell architecture evolution for both stand alone and embedded memory applications. The shift from planar to 3D capacitor structures and then to high k materials is illustrated. We then focus on the recently introduced capacitor-less DRAM cell as it appears as the next logical step in ultimate memory integration.

1. Introduction

For the last 30 years, the evolution of the basic 1T/1C Dynamic Random Access Memory (DRAM) bit cell was dominated by the capacitor requirements to keep a minimum cell capacitance above a certain threshold. The most common requirement is 25 to 30 fF per cell. To reach those values, very complex 3-dimensional (3D) structures were developed in the late 80's, including both trench and stacked capacitor options. At the same time, the use of silicon nitride rather than silicon dioxide as a dielectric was a step in the high k direction and very recently, the industry is moving to real high k materials to keep the capacitance value as features are scaled below 100 nm. These 3D capacitor structures and new materials, and the process complexity required to build or integrate them show that standard DRAM cells are running out of steam. A new era of opportunity is now appearing for simpler memory structures, such as the Zero-capacitor DRAM cell (Z-RAM cell). In this paper we will review the various trends in cell architecture for both stand alone and embedded memory applications.

2. Planar Capacitor Cells

The concept of a 1T/1C DRAM cell was proposed in 1968 by Dennard at IBM [1]. The first introduction of such cells was realized with 8 μm features ($F=8 \mu\text{m}$) for 4Kb NMOS DRAMs. At that time, the manufacturers used planar capacitors with silicon dioxide as a dielectric (see Fig.1), and the cell area was 1280 μm^2 or 20 F^2 . Shrinking this cell forced all memory makers to use silicon nitride as a capacitor dielectric for the 256 Kb generation, and to exploit the third dimension (3D capacitors) to build reliable 4 Mbit chips.

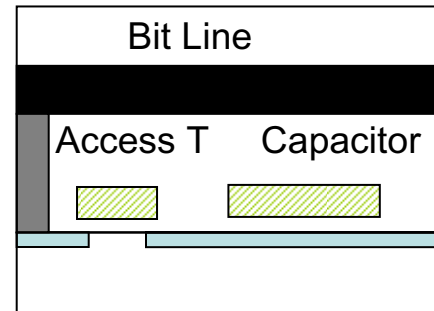


Fig. 1: Earlier planar DRAM cell cross section

3. 3D Stacked & Trench Cells

The 4 Mbit DRAM generation saw the introduction of two types of 3D capacitor structures: the stacked capacitor [2] and the trench capacitor [3] DRAM cell where by exploiting the vertical direction, the capacitor area could be kept to relatively high values, even though the projected cell area decreased from generation to generation. With such technologies, DRAM chips using features sizes F from 1 μm down to 100 nm were successfully manufactured.

The stacked capacitor memory cell structure exploited various capacitor electrode shapes to create area, but the best candidate proved to be the cylindrical shaped structure illustrated in Fig. 2. Adding rugged electrodes by exploiting Hemispherical Grain Silicon (HSG) [4] or exploiting Ta_2O_5 dielectrics [5] made this structure successful through many memory generations.

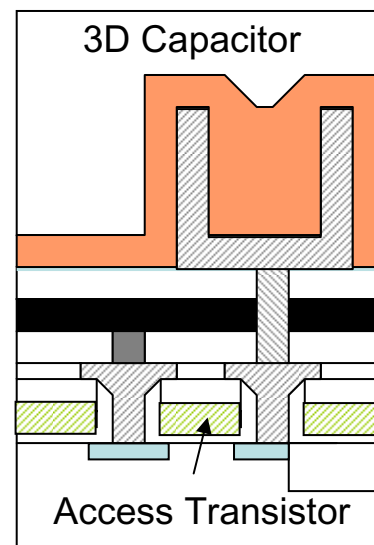


Fig. 2: Stacked DRAM cell cross section

The trench capacitor memory cell exploited various concepts with the common electrode first inside then outside the trench hole, and the most successful cell is illustrated in Fig. 3. It exploits a buried strap to connect the transistor to the storage node electrode [6] and has been used down to the 130 nm technology node. To scale further the trench cell, rugged and bottle shaped trenches [7] or the integration of a vertical access transistor in the trench have been considered [8].

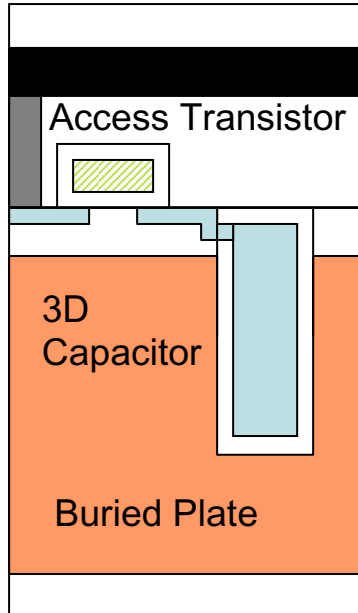


Fig. 3: Trench DRAM cell cross section

4. MIM & High K

At the 90 nm node and below, it is not clear how the standard 1T/1C DRAM capacitor can further scale. Most DRAM companies are developing MIM capacitor structures integrating high k dielectric materials [9] and it seems that very unrealistic capacitor aspect ratios (above 50:1 for trench structures) or contact aspect ratios (above 10:1) have to be manufactured to keep the targeted capacitance value. The cost and manufacturing challenges encountered with all those approaches are such that we can expect now a major change in DRAM cell architecture evolution.

5. Capacitor-Less Z-RAM Cell

In order to drastically ease memory integration and cost, and continue scaling for a few generations, Okhonin & al. recently introduced a capacitor-less DRAM structure [10-12]. Also called FBC (Floating Body Cell) [13,14] or Z-RAM cell (for Zero capacitor), it exploits the floating body effect of SOI transistors and is a pure single device memory structure. It can achieve the smallest area of any memory cell. Initially targeted for embedded applications, it doubles memory density when compared to DRAM and achieves 5 times the density of

SRAM without requiring 3D structures, new materials or anything else than a standard CMOS process. Integrated in stand alone memory arrays, it can achieve a cell area of only $4 F^2$. The Fig. 4 shows a 1 Mb embedded memory block designed with 90 nm rules.

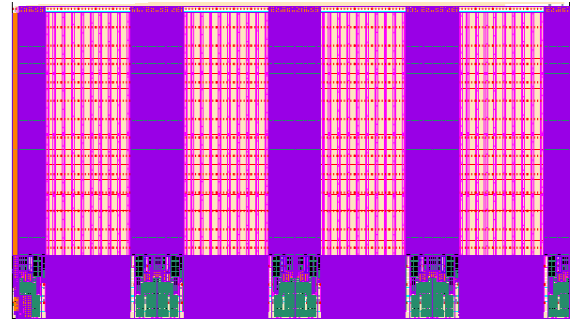


Fig. 4: 1 Mb macro of a Z-RAM memory

The scaling potentials of this technology have been demonstrated by its application to partially Depleted (PD), Fully Depleted (FD) and FinFET type of devices. This is illustrated in Fig. 5, 6 & 7 that show programming window characteristics of all 3 device types. We can therefore expect that the Z-RAM technology will allow dense DRAM memories to be extended for a long time in the sub 100 nm era.

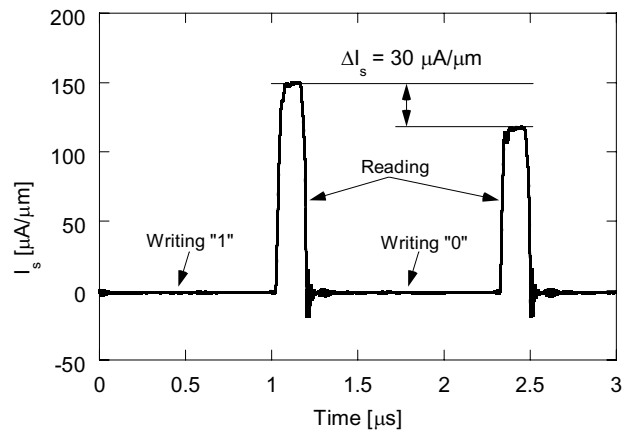


Fig. 5: Programming window of a PD Z-RAM cell

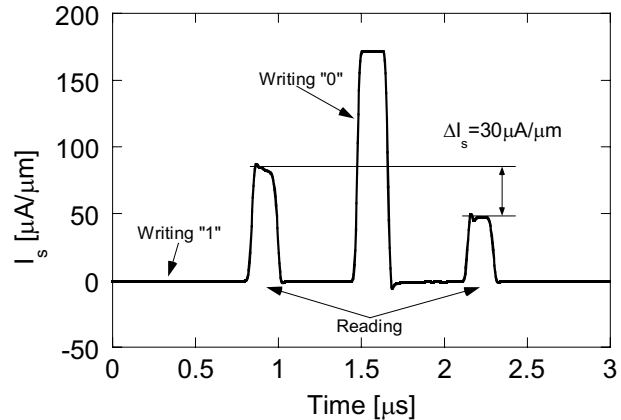


Fig. 6: Programming window of a FD Z-RAM cell

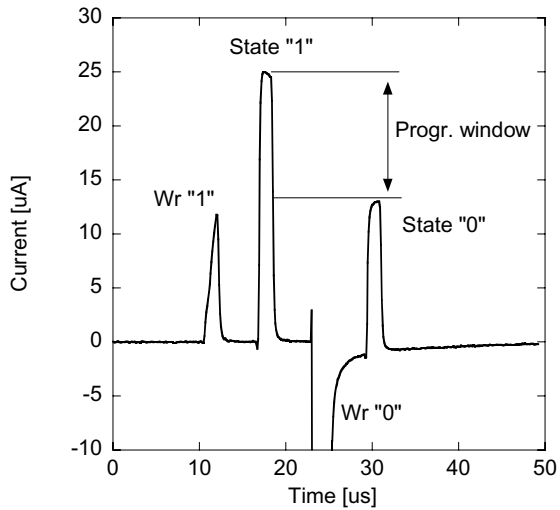


Fig. 7: Programming window of a FinFET Z-RAM cell

6. Conclusion

The DRAM cell structure experienced many changes over the last 30 years. An evolution from silicon dioxide to silicon nitride dielectric and from planar to 3D capacitor shapes occurred in the late 80's and early 90's. In the late 90's, a lot of creativity was devoted to make smart 3D capacitor shapes for both trench and stacked approaches. More recently, an evolution to high k dielectric has occurred. We believe that the next change will be focused on cost and manufacturability. With this respect, the simplicity and CMOS compatibility of the floating body Z-RAM cell makes it the best candidate for future scaled DRAMs for many generations to come.

References

- [1] B. Dennard, R.H. Dennard, US Patent 3,387,286, June 4, 1968.
- [2] M. Koyanagi et al., IEDM Tech. Digest, 348, 1978.
- [3] H. Sunami et al., IEDM Tech. Digest, 806, 1992.
- [4] P.C. Fazan et al., IEEE Electron Dev. Letters, 279, 1990.
- [5] P.C. Fazan et al., IEDM Tech. Digest, 263, 1992.
- [6] L. Nesbit et al., IEDM Tech. Digest, 627, 1993.
- [7] J. Amon et al., IEDM Tech. Digest, 73, 2004.
- [8] U. Gruening et al., IEDM Tech. Digest, 25, 1999.
- [9] S. Arai et al., IEDM Tech. Digest, 403, 2001.
- [10] S. Okhonin et al., Proc. IEEE Intl. SOI Conf., 153, 2001.
- [11] S. Okhonin et al., IEEE Electron Dev. Letters, 85, 2002.
- [12] P. Fazan et al., Proc. IEEE Intl. SOI Conf., 10, 2002.
- [13] T. Ohsawa et al., IEEE J. Solid-St. Circ., 1510, 2002.
- [14] T. Ohsawa et al., ISSCC Tech. Digest, 458, 2005.

1T-Bulk DRAM cell with improved performances: the way to scaling

R.Ranica, A.Villaret, P.Malinge, P.Candelier, P.Masson*, R.Bouchakour*, P.Mazoyer, T.Skotnicki

STMicroelectronics, 850 rue Jean Monnet 38926 Crolles, France

Contact: Ph. +33 4 76 92 64 12, e-mail: pascale.mazoyer@st.com

*L2MP UMR-CNRS 6137, IMT Technopôle de Château Gombert, 13451 Marseille Cedex 1, France

Abstract

As capacitor-less DRAM cell appears to be an interesting candidate for future embedded memory generations, we paid particular attention to performances improvement and scalability of the 1T-Bulk architecture. We analysed through our model and then measured on fabricated devices the influence of different cell parameters: floating body doping level, gate length and gate oxide thickness. 1T-Bulk cell is found to be very promising for eDRAM applications.

1. Introduction

DRAM industry investments are focused on memory cells with high density, low power consumption and low cost process for the Systems-On-Chip market. A simple architecture to be considered as a challenger of the conventional 1T/1C cell is the capacitor-less DRAM cell. In this innovative device, the charge is stored in the floating body of the transistor. Up to now, this concept has been demonstrated for several architectures: Bulk [1], PDSOI [2-3], FDSOI [4-5] and Double Gate [4-5]. In this paper we focused on the 1T-Bulk cell. The interest of this architecture is the low cell area that can be achieved ($10F^2$) with a very simple process fully compatible with standard CMOS logic technology. These characteristics are mandatory for low cost and high density embedded applications.

In this paper we investigated how the memory effect can be improved through model and experimental measurements on $0.13\mu\text{m}$ devices. Then we status on the ability of this architecture to be shrink. For this purpose, devices with scaled gate length were fabricated following standard CMOS 90nm technology.

2. Device fabrication

Two kinds of 1T-Bulk devices were fabricated: the first ones are realized with a $0.13\mu\text{m}$ technology ($L_g/W = 0.35/0.3\mu\text{m}$, gate oxide $65\text{\AA}$), the second ones with a 90nm technology ($L_g/W = 0.28/0.4\mu\text{m}$ and different gate oxide thicknesses). In both the case, the memory cell consists in an NMOS transistor presenting a floating body. The device's pwell is insulated laterally by STI (Shallow Trench isolation) and vertically by an n-buried layer (here called "niso"). Only few additional process steps are required to build this device, thus ensuring very low process cost and total compatibility with conventional CMOS logic. A cross-section picture of $0.13\mu\text{m}$ device is presented in Fig.1. The same mechanisms described in [1] are used for memory operations: impact ionization for write and forward biasing of source-body junction for erase operation.

Operating voltages for $0.13\mu\text{m}$ devices are summarized in Table I.

3. 1T-Bulk cell modelling

The evolution of the pwell potential VB during a sequence of memory operations is studied through our model [6]. Operating bias described in Table I are applied on drain, source, gate and n-buried layer (Fig.2). When the device is switched from a programming mode (write or erase) to the hold one, the limit conditions for Poisson equation change into the floating pwell. A new pwell potential is required to ensure the electrostatic equilibrium and thus all the Space Charge Areas (SCA) are modified (Fig.3). We named this step SCA reorganization, as showed in Fig.2. We will analyse the impact of intrinsic cell parameters such as the pwell doping at the Source and Drain (SD) junctions and at the niso junction.

4. Pwell doping impact on memory cell characteristics

If we consider the pwell doping in our device, two concentrations must be taken into account: the one at the SD-pwell junction (Na SD side) and the one at the niso-pwell junction (Na niso side). Pwell potential is given in Fig. 4 as a function on the first parameter (and at a constant Na niso side doping). We observe that the pwell doping at SD side influences the VB potential during all the memory operations. We will interest mainly on its impact during the reorganization step. We plotted the VB resulting after the program/hold switch for both state 1 and 0 as a function of the Na SD side doping (Fig.5). We observe that the VB loss during reorganization increases for both the memory states by increasing the Na SD side doping: this will result in a reduction of the memory effect amplitude. The impact of Na niso side pwell doping on memory operations is shown in Fig.6. Even in this case the doping level affects the VB during the reorganization step. We notice, in this case, that an increased pwell doping at the niso side will be suitable to enhance the memory window (Fig.7). The described VB modulation during reorganization depends essentially on the Na SD / Na niso rate. Indeed, if this rate remains constant, the VB loss during the program/hold switch does not change (Fig.8). As predicted by our model, the VB loss can be reduced by minimizing this rate, which is in accordance with presented results (Fig. 4-7).

5. Experimental results

In order to validate these theoretical results, different memory devices have been fabricated with $0.13\mu\text{m}$ technology by varying the pwell doping energy and dose. Moreover, the impact of NLDD doping on the memory

characteristics was also evaluated. The resulting read current shift is shown in Fig. 9 for all the devices. Operating voltages described in Table I are used. We observe that

- (i) by increasing the pwell energy the memory effect amplitude is enhanced
- (ii) by changing the pwell dose, the Na SD / Na niso rate increases, thus reducing the memory window
- (iii) by increasing the NLDD dose the memory effect amplitude increases

To clarify the pwell doping role in affecting the memory characteristics, 1D simulations of doping profiles are performed for all the presented devices. If we consider the pwell, we observe that a higher energy results in a lower Na SD side doping and a higher Na niso side doping (Fig.10). As previously demonstrated, this doping configuration is favourable to memory effect at both SD and niso sides. On the contrary, by increasing the pwell dose, both the SD and niso side pwell doping level increases (Fig.11). In this case, an optimum in the Na SD / Na niso rate must be found by adjusting the pwell dose in order to maximize the memory window. According to the presented electrical results, the best dose is our reference.

Devices with higher NLDD dose exhibit a read current difference higher than the reference one. This can be easily explained: a higher doping level at NLDD results in a higher longitudinal electric field in the saturation regime and thus in higher impact ionization rate, which is the mechanism used to program state 1.

Finally, we evaluated the role played by operating biases on the memory window, especially the voltage applied during erase operation. State 0 programming is particularly important because this state has the better retention characteristics over time [1]. A negative bias is applied on gate electrode during erase ($V_D = V_G$) in order to perform low power memory operations. Fig.12 shows that a more negative erase voltage could greatly enhance the memory, even if the VB loss during the reorganization is slightly increased (Fig.13). These considerations are confirmed again by experimental results. Fig. 14 shows that, by applying -2V during erase operation, the memory window in our devices is improved with respect to read current shifts obtained by using -1.2V erase voltage on gate and drain electrodes (Fig.9).

6. 1T-Bulk cell scaling

In order to status on memory cell behaviour towards scaling, devices with reduced gate length ($L_g = 0.28\mu\text{m}$) and different gate oxide ($T_{ox} = 65, 50$ and $30\text{\AA}$) were fabricated by using standard 90nm technology. For every sample, operating voltages were optimized to maximize the memory window according to the gate oxide thickness (Table II). The resulting source current difference is measured at different temperatures (25°C and 85°C). For example, the memory window of samples with $T_{ox} = 50\text{\AA}$ is represented in Fig.15. Fig. 16

compares the memory effect amplitude for all gate oxide thicknesses. After normalization of all the read currents, we notice that the memory window is improved by reducing the gate length. This is due to the fact that transistors with shorter gate length exhibit higher impact ionization rates.

If we consider the gate oxide influence, we notice that the memory effect amplitude is slightly reduced by decreasing the gate oxide thickness. This is easily explained through the substrate effect dependence on T_{ox} : $dV_{th}/dV_B \propto T_{ox} \cdot \sqrt{2\psi_B + V_B}$ [7]. But the main interest of these experimental results is that we demonstrated a significant memory effect even in samples with T_{ox} scaled down to $30\text{\AA}$. This is very promising for the device scaling because it will allow the fabrication of 1T-Bulk cells with smaller and smaller gate length.

Moreover, we wonder if doping concentrations impact 90nm devices as well as the $0.13\mu\text{m}$ ones. For this reason, devices with $0.28\mu\text{m}$ gate length and $50\text{\AA}$ gate oxide with different implant conditions were fabricated. Memory windows were compared to the reference (Fig.17): electrical measurements confirm the behaviour already presented for $0.13\mu\text{m}$ devices. It is worth to notice that these different doping concentrations don't affect the memory cells retention characteristics. Fig.18 shows the retention time at room temperature of the three 90nm samples: any difference is observed in the memory window behaviour as a function of time. These experimental results guarantee that doping modifications can be integrated to our architecture in order to improve the memory window without degrading retention time even for scaled 1T-Bulk cells.

7. Conclusions and perspectives

With the help of a dedicated model, we studied how the memory amplitude can be improved in our 1T-Bulk device. The doping concentration in the transistor's floating pwell is found to be one of the key parameters in this architecture. Devices with different implant conditions were fabricated with $0.13\mu\text{m}$ standard CMOS technology. Experimental results confirmed the model predictions and showed a memory window enhancement simply by adjusting the pwell doping. Then, the scalability of this concept is taken into account: devices with reduced gate length and different gate oxides were fabricated with 90nm technology. A greater memory window is exhibited, which is very promising for the future scaling of the 1T-Bulk DRAM cell down to 45nm technological node.

References

- [1] R. Ranica et al., VLSI Tech. Dig., 2004, p. 128-9
- [2] S. Okhonin et al., IEEE Intl. SOI Conf., 2001, p. 153-4
- [3] T. Shino et al., VLSI Tech. Dig., 2004, p. 132-3
- [4] R. Ranica et al., IEDM Tech. Dig., 2004, p. 277-80
- [5] C. Kuo et al., IEDM Tech. Dig., 2002, p. 843-6
- [6] A. Villaret et al., Micr. Eng., Vol.72, 2004, p. 434-9
- [7] Y. Taur, T.H. Ning, "Fundamentals of Modern VLSI devices"

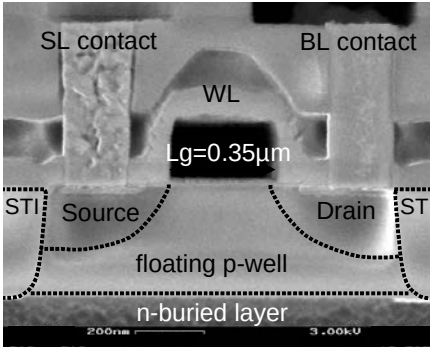


Fig.1: Cross section view of devices fabricated with 0.13μm technology. The niso layer is put into evidence through a chemical doping revelation.

0.13μm devices

	VD	VG	VS
Write	+2V	+1.5V	0V
Erase	0V	-1.2V	-1.2V
Read	+0.4V	+1.5V	0V
Hold	0V	0V	0V

Table I. Voltages applied during memory operations on drain, gate and source electrodes for 0.13μm 1T-Bulk cells. The niso layer is positively biased to 0.6V. Negative voltage is applied on gate during erase operation in order to reduce the power consumption.

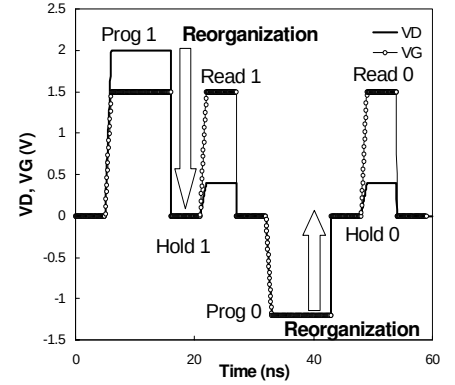


Fig.2: Simulated operating voltages. For simplicity, the erase bias is applied on drain instead of source.

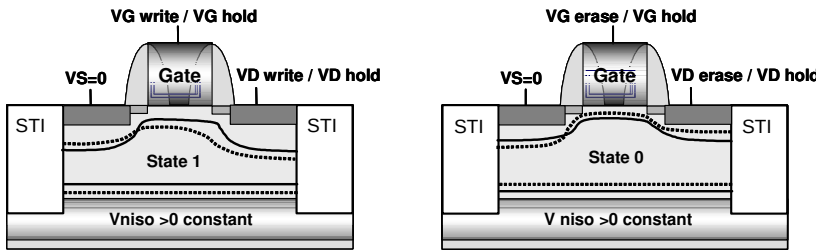


Fig.3: As reported in [6], the memory charge is stored in the SCA of 1T-Bulk device. We assumed that the total space charge is conserved during this phase. As a consequence, the SCA variation at drain and gate electrodes (whose voltage is modified during the switch) is balanced by a SCA modulation at source and niso junction. This SCA modulation is responsible for the VB variation from program to hold mode.

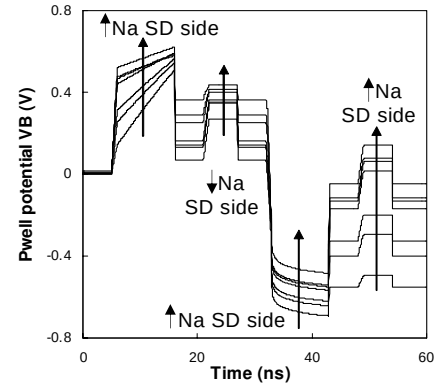


Fig.4: VB during memory operations vs. Na SD side doping with Na niso doping fixed.

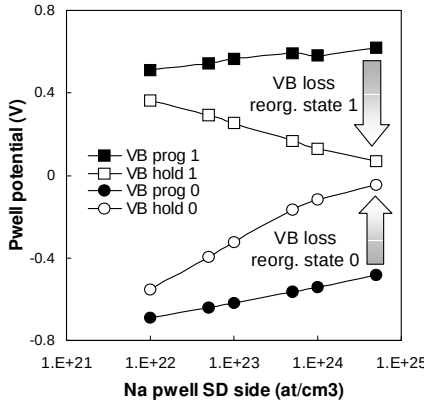


Fig.5: VB in program and hold mode for state 1 and 0 vs. Na SD side doping.

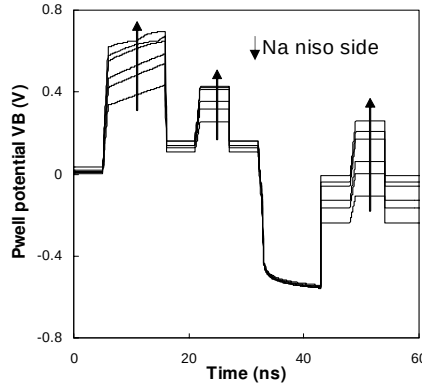


Fig.6: VB during memory operations vs. Na niso side doping with Na SD doping fixed.

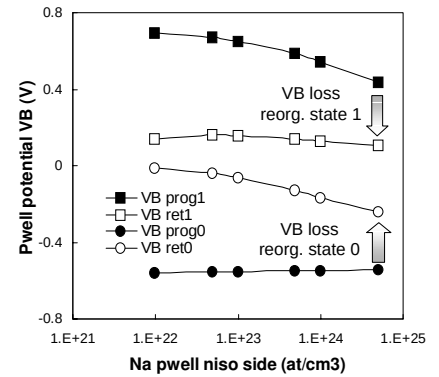


Fig.7: VB potential in program and hold mode for state 1 and 0 vs. Na niso side doping.

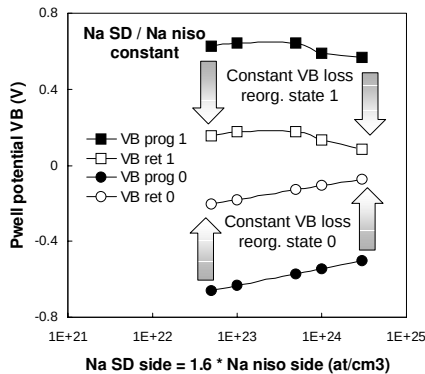


Fig.8: VB loss during reorganization remains constant when Na SD / Na niso is constant.

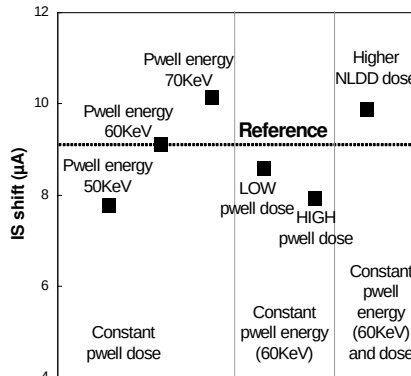


Fig.9: Memory effect amplitude for the different fabricated 1T-Bulk cells.

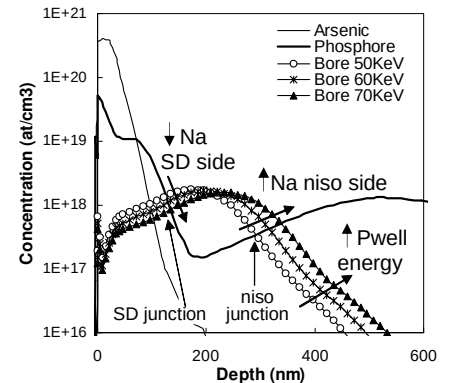


Fig.10: Simulated doping profiles in devices with different pwell energies.

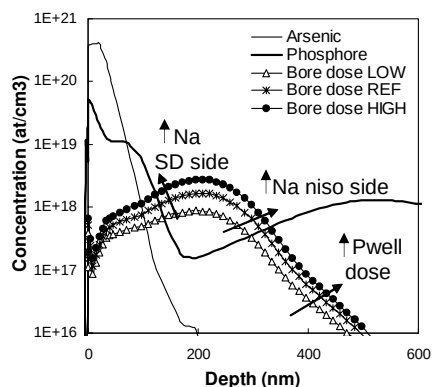


Fig.11: Simulated doping profiles in devices with different pwell doses.

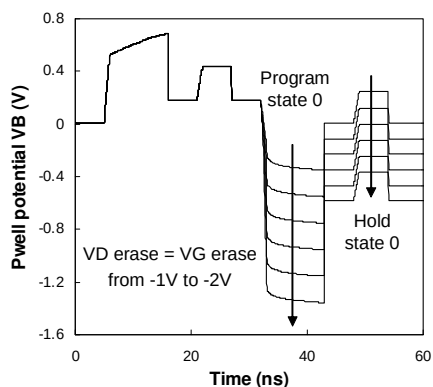


Fig.12: Impact of VD and VG erase biases on VB in program and hold mode for state 0.

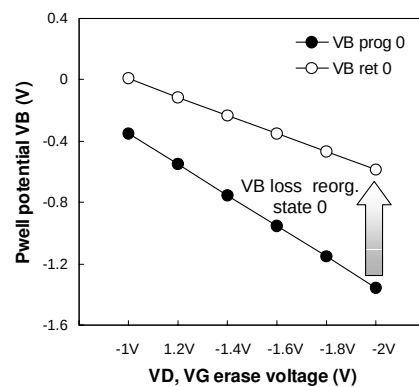


Fig.13: More negative erase voltages result in lower VB in program and hold mode.

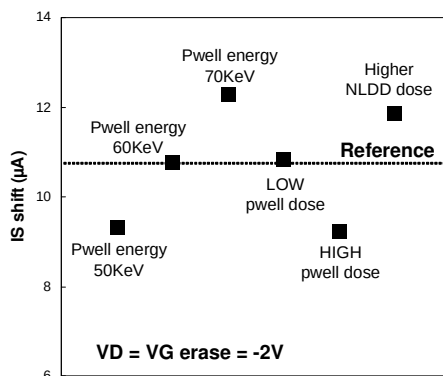


Fig.14: The read current difference is enhanced by applying a more negative erase voltage (-2V with respect to -1.2V) on gate and drain electrodes.

Tox 65-50Å	VD	VG	VS
Write	+2V	+1.2V	0V
Erase	0V	-1.4V	-1.4V
Read	+0.4V	+1.2V	0V
Hold	0V	0V	0V

Tox 28Å	VD	VG	VS
Write	+1.4V	+1V	0V
Erase	0V	-1.2V	-1.2V
Read	+0.4V	+1V	0V
Hold	0V	0V	0V

Table II. The same operating voltages were used for devices with Tox 65 and 50Å. Memory cells with Tox 28Å required lower biases ($L_g / W = 0.28 / 0.4\mu\text{m}$).

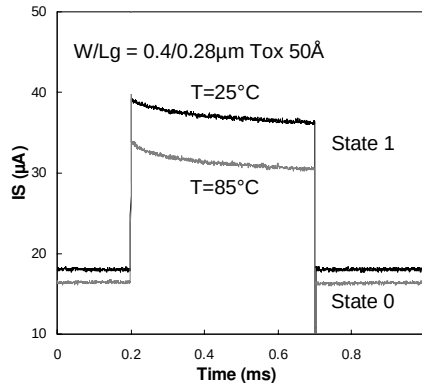


Fig.15: Read current margin between state 1 and state 0 for samples with gate oxide 50Å at two different temperatures.

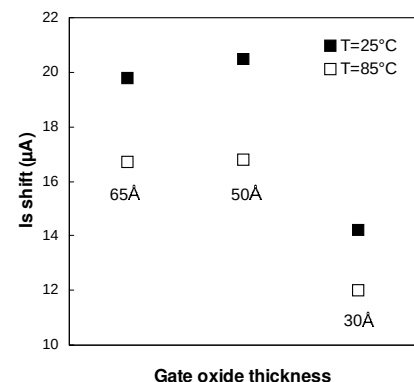


Fig.16: Memory window for all the 90nm fabricated samples. Greater amplitude is measured with respect to 0.13μm samples. This is partly due to the larger W (0.4 vs. 0.3μm), partly to the more negative erase voltage (-1.4 vs. -1.2V), but mainly to the shorter gate length (0.28 vs. 0.35μm).

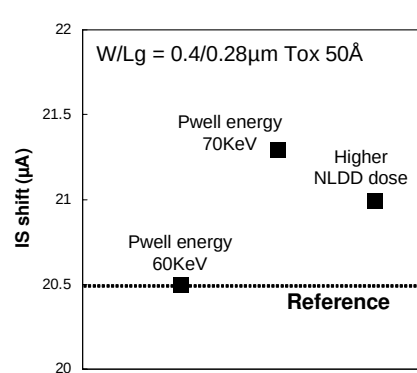


Fig.17: Source current difference for 90nm devices with different implant conditions. Only conditions supposed to give enhanced memory amplitude were repeated. Even in this case, higher pwell energy and higher NLDD doping give a better memory window.

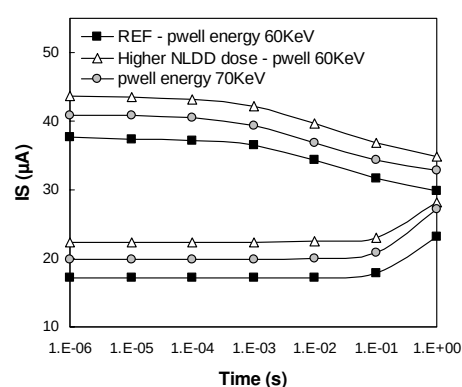


Fig.18: Retention characteristics at 25°C of 90nm devices with different doping conditions. The evolution of the read current margin with time shows no remarkable difference between the three samples.

FinFET based Zero-capacitor DRAM (Z-RAM) Cell for sub 45 nm Memory Generations

S. Okhonin¹, M. Nagoga¹, P. Fazan¹, L. Mathew², B. -Y. Nguen², H.-H. Chen², and T. Stephens²

¹ Innovative Silicon, PSE-B, CH-1015 Lausanne, Switzerland, ph: +41-21-6938992, e-mail: sokhonin@innovativesilicon.com

² Freescale Semiconductor Inc, 3501 Ed Bluestein Blvd. MD: K10, Austin TX-78721, U.S.A.

Abstract

We experimentally demonstrated for the first time a zero-capacitor RAM (Z-RAMTM) cell based on a CMOS double gate FinFET. Two operation modes and different programming techniques are described. A retention time of a few milliseconds is measured at room temperature that suits eDRAM specification. This FinFET based Z-RAM memory will allow the manufacture of low cost DRAMs and eDRAMs for 45 and sub 45nm generations.

Introduction

Dynamic Random Access Memory (DRAM) and embedded DRAM (eDRAM) manufacturers face a tremendous challenge to shrink the basic memory cell and the main issue in cell area reduction lies with the capacitor integration. In the recently introduced zero-capacitor floating body memory cell (Z-RAM), the conventional storage capacitor is replaced by the body capacitance of a SOI MOSFET [1-6]. The floating body charge affects the channel conductivity through the body effect and can be used to distinguish two states. The functionality of a Z-RAM memory array was already demonstrated on silicon using PD or FD SOI devices [7, 8]. To ensure the scalability of this new memory, it is mandatory to demonstrate its operation on advanced devices such as double-gate MOSFETs, FinFETs, etc. Recent simulations illustrate the memory effect in double-gate FinFET devices [9].

In this paper we experimentally demonstrate for the first time a Z-RAM cell based on a CMOS double gate FinFET also referred to as Vertical Multiple Independent Gate Field Effect Transistor (MIGFET) [10]. The devices comprise two gates on the two sides of a silicon fin (Figures 1, 2). These gates can independently control the channel and this provides a possibility to introduce two operation modes: in the first mode a constant bias is applied to the gate g2 to create a floating body storage node; in the second mode a constant bias V_{bg} is applied to the back gate (wafer substrate) to create a floating body storage node and the gates g1 and g2 are shorted. A V_{bg} value of 25V is used; it can be dramatically reduced by decreasing the buried oxide thickness.

Experimental

Silicon <100> wafers covered with 50nm nitride film were patterned into thin silicon channel region of 100nm height. The gate oxide of 24.7Å was thermally grown on these fins. Functional PMOS and NMOS transistors are obtained by implanting and annealing the source/drain regions. The devices used in this study (Fig. 1a,b) have undoped channel regions and doped polysilicon gates. P-channel transistors with the width of 110nm and channel length of 100nm were characterized.

Results and discussion

Normally the FinFET's body is fully depleted and this device is virtually free of kink effect. In the first operation mode (Fig. 2), by applying a positive voltage (V_{g2}) to the second gate, an accumulation layer can be created at the second interface. In this case the transistor shows a partially depleted behavior and a kink effect can be observed as presented in Figure 3. In a Z-RAM cell the state "1" corresponds to an excess of majority carriers in the transistor's body and the state "0" corresponds to their deficiency. The sensing is performed by measuring the channel current. To investigate the memory effect in the FinFET we write "1", read the source current, write "0", and read the source current again. The measured current versus time characteristic is shown in Figure 4. The difference between the currents, measured after write "1" and write "0" characterizes the memory effect. In Figure 4a, writing "1" is performed with the impact ionization mechanism. In this write operation, the transistor is in its ON state, consuming therefore power. In Figure 4b to write a "1" while keeping the transistor in its OFF state, we use the band to band tunneling mechanism. By applying positive gate and negative drain voltages, electrons are created in the floating body and the state "1" is generated. As no current flows through the cell during this writing, there is no power consumption. Writing "0" is performed by applying a negative voltage to the Gate 1. Figure 4b demonstrates the possibility of low power programming for the FinFET Z-RAM memory.

An important characteristic of a DRAM is its retention time. To perform its measurement we used the transient measurement setup described above. The evolution of the state "1" and state "0" with time is measured. The results are presented in Figure 5. Retention times of few milliseconds are observed at room temperature. These values suit typical embedded DRAM specification.

In the second operation mode (Fig. 6), a constant back gate bias is used to create an accumulation layer at the back interface. Measurements performed in this mode are shown on the Figures 7-9.

Conclusion

A Z-RAM memory cell based on a CMOS double gate FinFET is experimentally demonstrated for the first time. Two operation modes and different programming techniques are

introduced. A retention time of few milliseconds is measured at room temperature. This FinFET based Z-RAM memory will allow manufacturing of very low cost DRAMs and eDRAMs for 45 and sub 45-nm generations.

References

- [1] M.R. Tack et al, *IEEE Trans. Electron Dev.*, 1990, pp.1373-1382.
- [2] S. Okhonin et al, *IEEE Int. SOI Conf.*, 2001, pp. 153-154.
- [3] S. Okhonin et al, *IEEE Electron Device Letters*, 2002, pp. 85-87.
- [4] T. Ohsawa et al, *IEEE J. of Solid-State Circ.*, 2002, pp. 1510-1522.
- [5] P. Fazan et al, *IEEE Int. SOI Conf.*, 2002, pp. 10-13.
- [6] C. Kuo et al, *IEEE Trans. Electron Dev.*, 2003, pp.2408-2416.
- [7] K. Inoh et al, *Symp. VLSI Technology*, 2003, pp. 63-64.
- [8] T. Shino et al, *Symp. VLSI Technology*, 2004, pp. 132-133.
- [9] T. Tanaka et al, *IEEE IEDM*, 2004, pp. 919-922.
- [10] L. Mathew et al, *IEEE Int. SOI Conf.*, 2004, pp. 187-189.

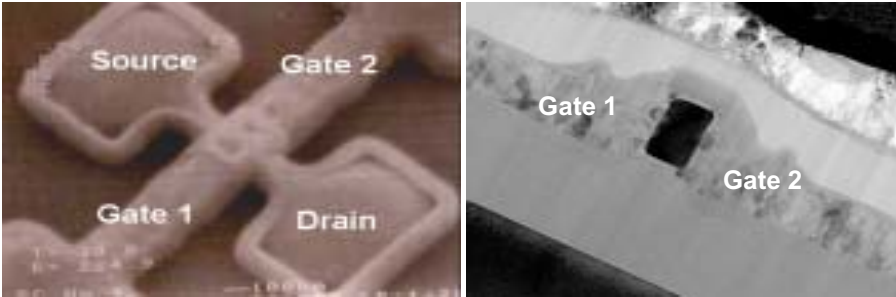


Figure 1. a) MIGFET top view, b) TEM cross-section.

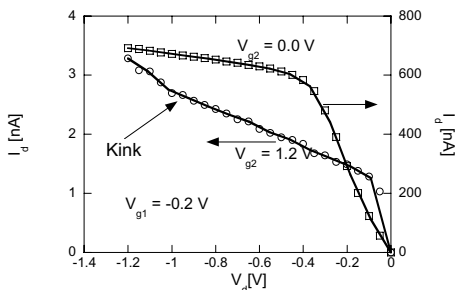


Figure 3. Kink effect observed at $V_{g2} = 1.2V$ in the first operation mode.

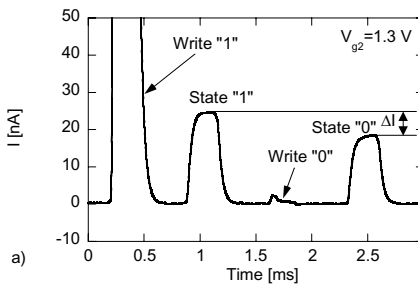


Figure 4. Source current time dependencies measured in the first operation mode. State "1" is written with the impact ionization mechanism (a) and with band to band tunneling (b).

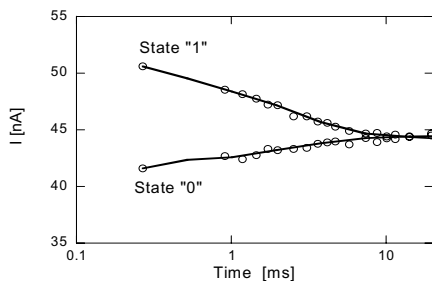
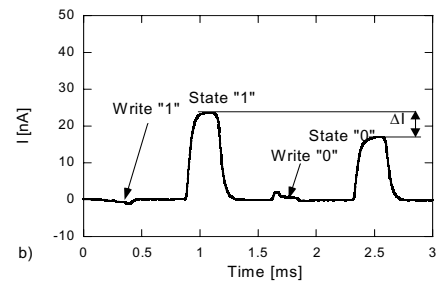


Figure 5. Retention characteristics of a Z-RAM memory cell on MIGFET at $T=300K$ in the first operation mode.

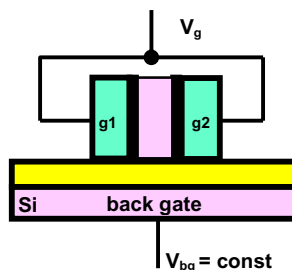


Figure 6. Second operation mode: the back gate is used to form the floating body storage node; gate 1 and gate 2 are connected.

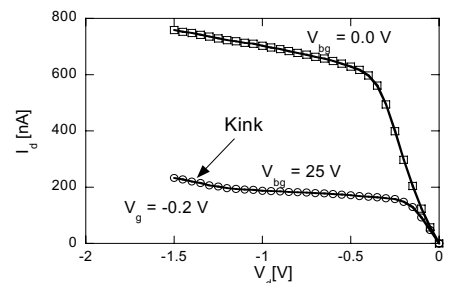


Figure 7. Kink effect observed at $V_{bg} = 25V$ in the second operation mode.

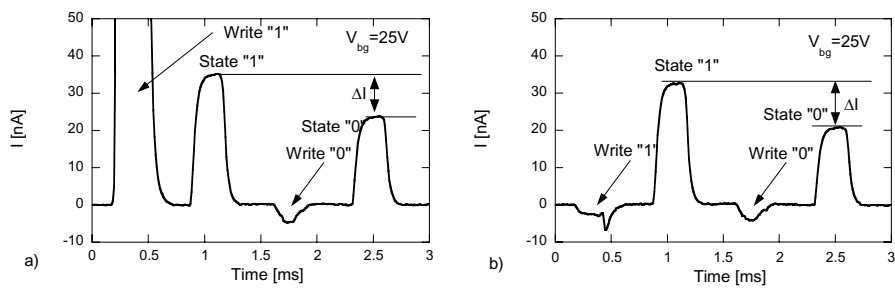


Figure 8. Source current time dependencies measured in the second operation mode. State "1" is written with the impact ionization mechanism (a) and with band to band tunneling (b).

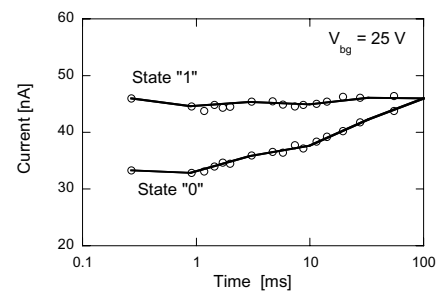


Figure 9. Retention characteristics of a Z-RAM memory cell on MIGFET at $T=300K$ in the second operation mode.

Evolution of stacked-capacitor technology for embedded-DRAM applications

Nicolas Emonet¹, Nicolas Jourdan¹, Audrey Berthelot², Marc Piazza¹, David Fraboulet[#], Laurence Gabette², Christian Caillat¹, Marion Charleux¹, Bruce Boeck³, Jorge Regolini¹, Didier Dutartre¹, Vincent Huard², Frederic Monsieur¹, Pascale Mazoyer¹ and Eric Gerritsen²

Crolles2 Alliance

¹ STMicroelectronics, ² Philips Semiconductors ³ Freescale Semiconductor
860 rue Jean Monnet, 38926 Crolles, France.

[#] CEA-LETI, rue des Martyrs, Grenoble, France

corresponding author: eric.gerritsen@philips.com

Abstract

For stacked-capacitor applications in DRAM, architecture, material choices and process technology are a crucial concern for each new technology node. An overview of capacitor technology evolution is presented from the early days of planar PIS (Poly/Insulator/Silicon) capacitors to the MIM (Metal/Insulator/Metal) capacitors used for today's 90nm node and beyond. Perspectives of MIM architectures, including ALD deposition technology for electrode materials and high-k dielectrics, are given.

1. Introduction

A typical DRAM cell consists of an access transistor and a storage capacitor (1T/1C). There are two types of storage capacitors, stacked-capacitors and trench capacitors, of which the latter offers highest density, at the expense of process complexity [1]. Trench capacitors are less suitable for integration of high-k dielectrics because they are formed before transistor formation with their associated high temperature anneals. For embedded-DRAM we have adopted stacked capacitors as being the most cost-competitive solution. Because stacked capacitors are formed after formation of the transistors the overall impact on performance of the CMOS core process is of primary concern. Capacitor processing at too high thermal budget will degrade transistor performance. In the first part of this paper, we will outline the general evolution of capacitor architectures. Next we will detail how capacitor materials and processing have evolved from furnace based polysilicon electrodes with (oxy)nitride dielectrics towards metal electrodes with high-k dielectrics, both processed by Atomic Layer Deposition (ALD) at moderate thermal budget favourable to embedded-DRAM applications.

2. Capacitor Architecture Evolution

Technical innovation in capacitor processing continued over the past 20 years in the industry, enabling DRAM scaling into the sub-100nm regime of today. Figure 1 illustrates the successive architectures adopted. In the following summary of DRAM capacitor technology since 1982, no distinction is made between stand-alone and embedded-DRAM as both are driven by similar concerns like high capacitance, high breakdown voltage and high capacitance density. Between 1982 and 1985, DRAM density increased from 64Kb to 1Mb [2]

using planar capacitors, from the early PIS solution with gate oxide dielectric to 2D-PIP (Poly/Insulator/Poly) capacitors allowing adjustment of dielectric thickness and capacitance. In 1985 the concept of stacked capacitors was introduced using 3-D capacitors and a triple poly [3].

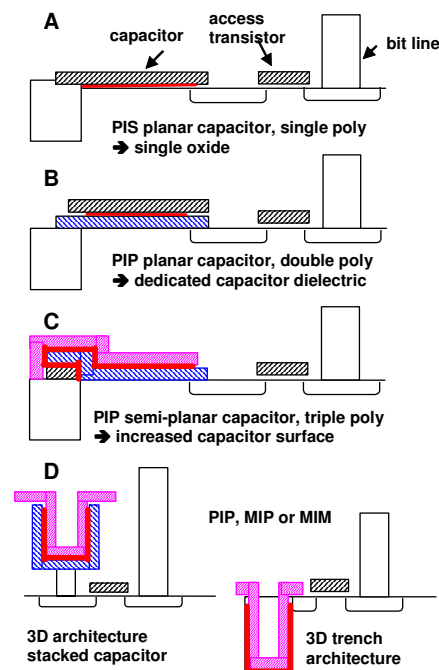


figure 1: DRAM architecture evolution from 1980 to 2005.

From 1986 to 1990, DRAM density is driven from 1Mb to 16Mb with design rules ranging from 0.8μm to 0.4μm. These cells are produced with 3D-architectures and SiO₂ dielectrics. The capacitor surface is extended in the third dimension to allow downscaling without adjustment of the SiO₂ dielectric.

From 1991 to 2000, DRAM evolution went from 64Mb towards the Gigabit era. Starting with a 0.4μm CMOS process using a stacked-capacitor cell, 1 Gb is demonstrated with 0.25 to 0.16μm CMOS. The third dimension of the cell is no longer sufficient to reduce lateral cell size. Si₃N₄ is added to SiO₂ to provide low capacitor leakage in a 3D-PIP architecture. SiO₂/Si₃N₄ dielectrics are optimised by increasing Si₃N₄ content and by reducing process temperature to ensure low leakage of the PIP capacitors.

Hemi-Spherical polysilicon Grains (HSG) are introduced as bottom-electrode in 0.18 and 0.15 μm technologies [4], increasing the effective surface, as illustrated in figure 2.

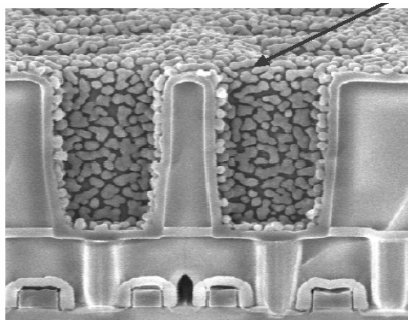


Fig. 2: HSG bottom electrode

Next efforts were focused on improving data-rate rather than bit density and especially on embedded applications with a pure Si_3N_4 dielectric.

From 2001 onwards, radical dielectric material changes appeared. Si_3N_4 is abandoned due to its high thermal budget. MIS stacked-capacitors using Ta_2O_5 or Al_2O_3 as high-k dielectric and metallic top electrodes are reported with 1Gb and 4Gb chips in 0.10 μm technology [5,6]. TiN is the mainstream electrode material whereas alternatives, like ruthenium, are being studied for higher workfunction and oxidation resistance [7].

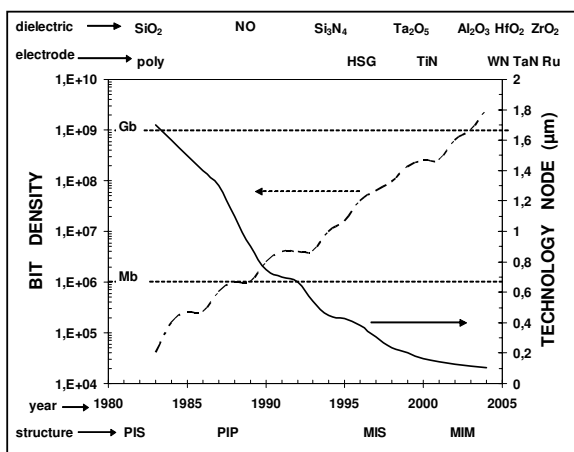


Fig. 3: Trend of DRAM production node, bit density, dielectrics and electrode materials.

DRAM enters into sub-100nm CMOS nodes focusing on design and application issues. New high-k dielectric materials, with low thermal budget, are introduced to replace Ta_2O_5 alongside with the introduction of MIM capacitors. Figure 3 summarizes the evolution of DRAM density, technology node, capacitor structure and dielectric materials.

3. Polysilicon Electrodes / Nitride Dielectrics

The use of polysilicon as electrode material was extended to 0.18 and 0.12 μm nodes by introducing hemispherical grains (HSG, see fig.2). This allowed an increase of the capacitor surface by a factor around 2 (fig.4). A further increase of capacitance was obtained by introducing a pure nitride dielectric, without oxide

(figure 5), where leakage is reduced by operation at lower voltage (V_{dd} from 1.8 to 1.2V)

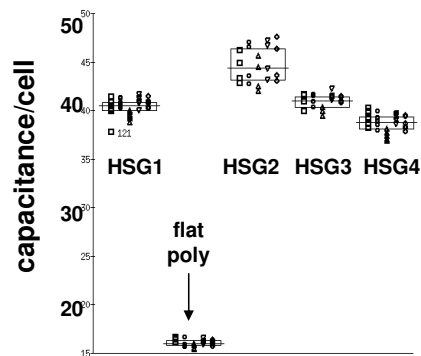


Fig. 4: Capacitance gain ($\sim 2.5\times$) achieved by introducing HSG. Note the robustness of the process, as the 4 variations in doping level and film thickness do not give significant changes.

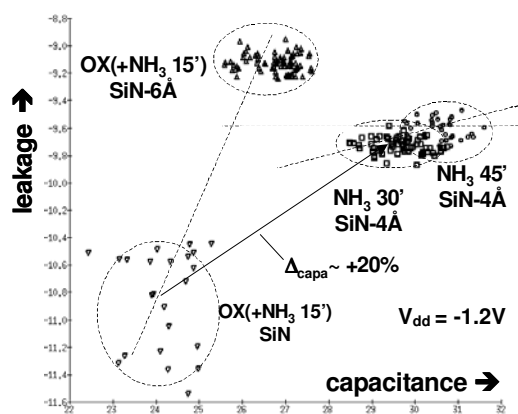


Fig. 5: Capacitance gain (+20%) when replacing oxide/nitride (ON) dielectric by a pure nitride (NN) dielectric.

However, since our embedded-DRAM uses stacked capacitors, formed after transistor formation, the use of furnace based HSG below 100nm is limited by its high thermal budget ($\sim 650^\circ\text{C}$) and the resulting degradation in performance of the logic core process.

Other limitations of HSG below the 100 nm node are :

- 1) Poly-depletion, with loss of capacitance giving an increase of several Ångströms in SiO_2 Equivalent-Oxide-Thickness (EOT).
 - 2) Formation of interfacial SiO_2 , further increasing EOT.
 - 3) Non-uniformity of the dopant level at the bottom and sidewall of the capacitor trench, giving additional depletion and increased resistance
- Poly-depletion, interfacial SiO_2 and high-resistivity can all be resolved by introducing metallic electrodes.
- 4) The relatively thick layer of HSG gives a significant reduction in effective surface area when shrinking the capacitor cell (figure 6), thus counteracting the increase due to the hemispherical grains. Whereas it is hard to reduce the thickness of HSG below 500Å, metal electrodes can easily be applied below 100Å thickness. Figure 6 shows that HSG electrodes in the 90nm node can give over 30% area loss, whereas this is less than 5% for thin TiN electrodes.

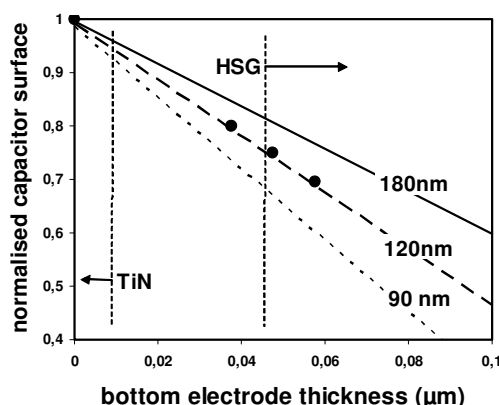


Fig. 6: Decrease of capacitor area due to thickness of bottom electrode for 180, 120 and 90nm nodes. This counteracts the area increase of the hemispherical grains

Similarly, the use of Si_3N_4 as capacitor dielectric beyond the 100nm node is limited by its high-thermal budget around 650°C. In addition, its dielectric constant is insufficient to meet the requirement on EOT which is set at 18, 13 and 9Å for the 90, 65 and 45nm nodes, respectively. Some high-k materials which have become available offer much higher permittivity, giving lower EOT, as well as higher bandgap, thus improving leakage performance and data retention (figure 7).

4. MIM Development and Perspectives

With the MIM capacitor cell established as concept of choice for our 90nm and 65nm e-DRAW technologies [8], our development effort has been focused on :

- choice of high-k dielectric and electrode materials: Al_2O_3 , HfO_2 and Ta_2O_5 as well as a $\text{Al}_2\text{O}_3/\text{Ta}_2\text{O}_5$ stack were evaluated as dielectric while different types of TiN were evaluated as electrode materials .
- choice of deposition technology: Metal Organic CVD (MO-CVD) and Atomic Layer Deposition (ALD) have been evaluated for dielectrics whereas PVD, CVD, MOCVD and ALD have been evaluated for electrodes.

Figure 7 gives an overview of the most commonly applied high-k dielectrics. It shows that nature is not generous on the trade-off between a high dielectric constant (needed for capacitance density) and a wide bandgap (needed for low capacitor leakage) . The figure also indicates the potential of atomic engineering by ALD to generate mixtures or nanolaminates of oxides with a more ideal behavior .

Ta_2O_5 has reached some maturity in DRAM processing and is mostly deposited by MO-CVD, which is suitable for mass production due to its reasonable deposition rate. The electrical (capacitance/leakage) performance of the deposited Ta_2O_5 film is critically dependent on a post-deposition anneal in oxygen or ozone, required to lower oxygen deficiency and carbon impurities in the film. Too aggressive anneals deteriorate interface quality and increase capacitor leakage

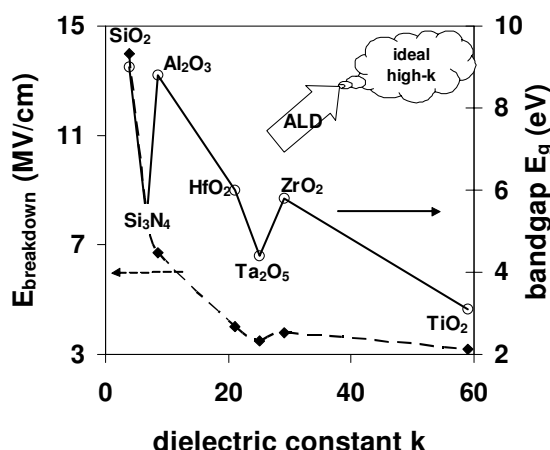


Fig.7: Variation of bandgap and breakdown field with dielectric constant for some common high-k dielectrics

I(V) characteristics of a Ta_2O_5 MIM capacitor are shown in figure 8 as a function of test temperature. A very strong increase of leakage is observed when raising the temperature up to 125°C, thus exceeding the specification of 10fA/cell (@+/- 0.6V).

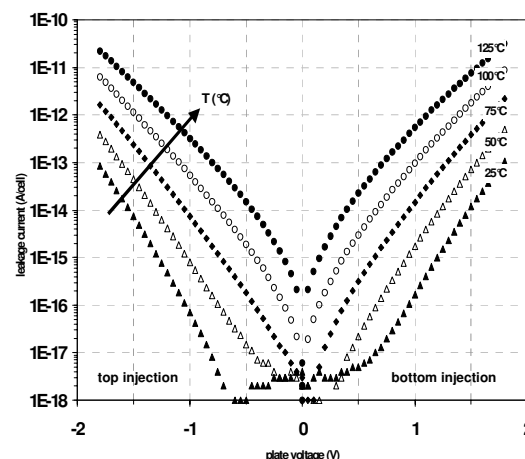


Fig. 8: Voltage/leakage characteristics of a 100 Å MOCVD- Ta_2O_5 MIM capacitor at temperatures from 25 °C to 125 °C.

In comparing Al_2O_3 to Ta_2O_5 , the first may offer only moderate dielectric constant, comparable to Si_3N_4 (figure 7), but it is much less reactive than Ta_2O_5 towards metal electrodes like TiN.

Most of all Al_2O_3 offers a wide bandgap, much larger than both Ta_2O_5 and Si_3N_4 , and comparable to that of SiO_2 (figure 7).

Al_2O_3 is deposited at around 400°C by ALD, with TMA as precursor and O_3 as reactant, with no anneal required.

Figure 9 illustrates how Al_2O_3 offers low capacitor leakage with excellent temperature stability. Figure 10 shows that the thickness dependence of the intrinsic Weibull slope for ALD- Al_2O_3 , as taken from TDDB curves, is in line with that of the highly reliable SiO_2 . According to percolation theory this implies that the type and size of the defects related to dielectric breakdown are similar for SiO_2 and Al_2O_3 .

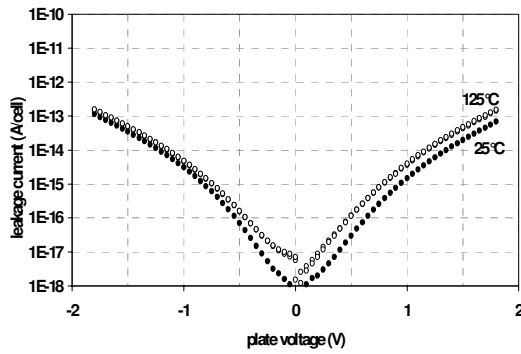


Fig. 9: I-V dependence at 25°C and 125°C for a MIM capacitor with 40 Å Al_2O_3 (17 Å EOT)

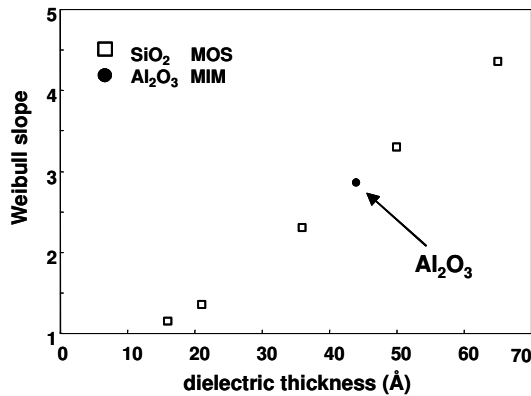


Fig. 10: Weibull slope of Al_2O_3 (circle) and SiO_2 (squares) as function of dielectric thickness

Hafnium oxide HfO_2 (with $k=21$ and $E_g \sim 6\text{eV}$) is evaluated in figure 11, improving further on capacitor performance as it combines the individual benefits of Ta_2O_5 (high $k \sim 25$) and Al_2O_3 (high $E_g \sim 8.8\text{eV}$).

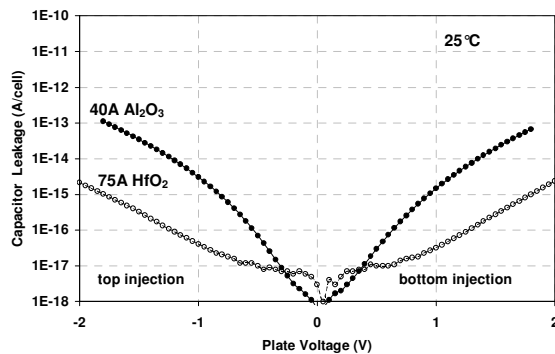


Fig. 11 I(V) curves for ALD- Al_2O_3 and ALD- HfO_2 MIM capacitors, both with an EOT of 17 Å.

For MIM processing TiN remains a preferred electrode material which we have deposited with various techniques. Sputter deposition (PVD) lacks conformality in the DRAM trenches. Standard CVD-deposition based on TiCl_4 is not ideal as it gives high-chlorine levels which can only be reduced at high deposition temperatures ($>600^\circ\text{C}$). MO-CVD type TiN contains large amounts ($\sim 10\text{at.}\%$) of residual carbon and oxygen. Figure 12 shows the capacitor I(V) behavior when TiN is deposited by ALD at low temperature (400°C), with TiCl_4 as precursor and NH_3 as reactant.

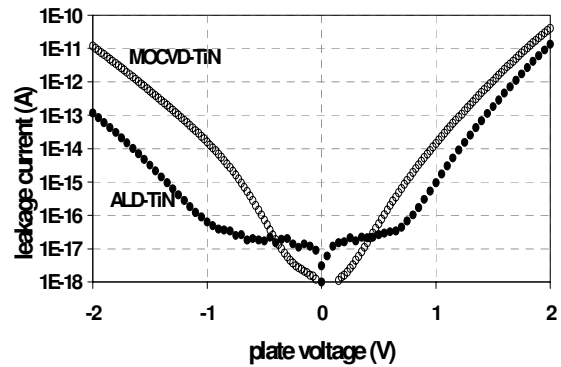


Fig. 12 I(V) dependence for an $\text{Al}_2\text{O}_3/\text{Ta}_2\text{O}_5$ bi-layer MIM capacitor comparing TiN top electrodes by MOCVD and ALD.

ALD-TiN is seen to give a massive decrease of leakage current compared to electrodes deposited by MO-CVD at the same temperature. Deposition of TiN by ALD offers superior purity and stoichiometry without plasma damage. ALD-TiN enables the application of high- k dielectrics at EOT values reduced by several Ångströms, for the same level of capacitor leakage.

Finally, the MIM architecture allows to limit capacitor height giving less bitline coupling and parasitic capacitance. Its low electrode resistance minimizes RC delay and offers fast read/write times required for the high-speed designs of today.

On a different note, capacitor-less (1T/0C) DRAM cells have recently been reported [9], in order to completely bypass the scaling limits of capacitor development in the Gigabyte era.

5. Conclusions

Capacitor technology for embedded-DRAM has flourished over many process generations due to evolutionary innovations like 3D-capacitors, hemispherical polysilicon grains and (oxy)nitride dielectrics. Below the 100nm node MIM capacitors are adopted with novel high- k dielectrics. ALD-deposition appears an enabling technology for both dielectrics and metal electrodes.

Acknowledgements

Part of this work was realized under European Eureka program MEDEA+ T-126 on embedded memories. ALD development was supported by CEA-LETI.

References

- [1] J.Lutzen et al, VLSI Tech.Dig., p.178, 2002.
- [2] H. Sunami, IEDM Tech.Dig., p. 694, 1985.
- [3] Y.Takemae et al.,ISSCC Tech.Dig, p.250, 1985
- [4] M.Sakao et al., IEDM Tech.Dig, p.655,1990
- [5] H.S Jeong et al., IEDM Tech.Dig., p.353, 2000
- [6] S.J Won et al., VLSI Tech.Dig., p.97, 1997
- [7] S.J. Won et al., IEDM Tech.Dig, p.789, 2000
- [8] P. Mazoyer et al. IITC Techn. Dig, p.117, 2003
- [9] R.Ranica et al., VLSI Tech.Dig., p.128, 2004

Embedded Memory Design: Challenge or Nightmare

Johan Hendrickx^a, Patrick van de Steeg

^a Philips Semiconductors, High Tech Campus 46, 5656 AE Eindhoven, the Netherlands, johan.hendrickx@philips.com

Abstract

Design of embedded memories in the most advanced technologies (90nm, 65nm, ...) is a very complex task with many challenging aspects. This paper highlights some of these, mainly focusing on the design of embedded SRAM and ROM compilers.

1. Introduction

Key elements in the design of embedded SRAM and ROM memories are definition of the design space and specification, choice of the most appropriate architecture, schematic design of sense amplifier and self timed loop, layout, floorplanning and extraction.

2. Key aspects

The definition of the design space and more generally the functional specification of an embedded memory compiler is the first decision, which has to be made, focusing on the IP reuse aspects. The correct trade-off between speed, power and area determines the suitability of a memory compiler in the customer application.

Next to that, the most appropriate architecture has to be defined, either focusing on area, speed or power, including leakage reduction techniques. This also includes choosing the optimal combination of threshold voltages available within the technology (in 90nm and 65nm technologies, low, standard and high VT flavours are available).

The design of the sense amplifier should take the statistical variations of the transistors into account to determine the required minimum differential voltage. The self timed loop (or dummy loop) should be set as fast as possible to achieve the best access time, but not too fast to avoid yield loss. Also here statistical analysis has to be performed on both active and dummy circuits.

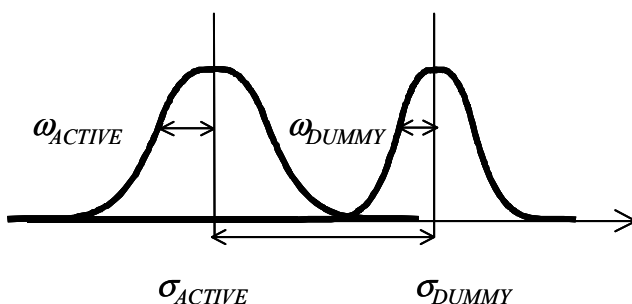


Fig. 1: The margin within the self timed loop is determined by the difference between the active access path and the dummy access path.

A dense layout requires a well-thought floorplanning to avoid wasted area. When complemented with a dense (full-custom) layout of the different leafcells, a very high efficiency can be achieved.

During the validation of an embedded memory compiler, extraction is one of the most important issues. As shown in figure 2, the back-end contains the most dominant layers (including contact and VIA) and therefore a full 3-D extraction of the SRAM bitcell is essential to include all its capacitive parasitics.

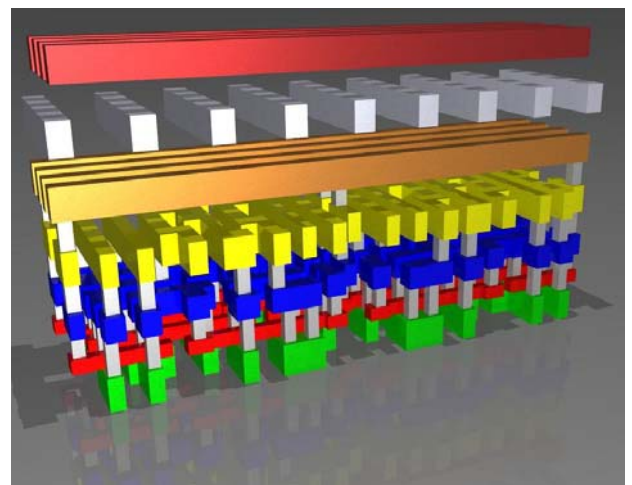


Fig. 2: 3-D representation of an SRAM bitcell

3. Conclusion

Many issues need to be tackled when designing embedded memories in advanced technologies. Powerful tools (simulation, extraction, ...) offer a great help in supporting the design of embedded memories and its verification. However, experienced architects and designers are indispensable for the design of very competitive and state-of-the-art embedded memories.

A small granular controlled leakage reduction system for SRAMs

Peter Geens, Wim Dehaene
K.U. Leuven, ESAT -MICAS
Kasteelpark Arenberg 10, B-3001 Heverlee, Belgium
Tel. +32 16 32 11 04
 {Peter.Geens,Wim.Dehaene}@esat.kuleuven.ac.be

Abstract

In current and future technologies the leakage of the transistors is a major contributor to the total power dissipation. That effect is even reinforced in SRAM circuits because the matrix contains a high proportion of cells that are not accessed for extended periods of time. This paper presents a memory that by use of small granular control of the supply voltages can succeed in gaining at least a factor 10 in total power reduction. Through the use of a distributed last stage of the SRAM decoder the stand-by and active cycle of the SRAM matrix can be controlled per word. This allows to selectively wake up only the needed word and peripheral circuitry. When combined with monitoring and DC-DC conversion circuitry for the stand-by voltage, a closed loop system is attained that can minimise power consumption in the SRAM matrix.

1. Introduction

Due to the seemingly unstoppable increase in demand for more functionality in mobile applications, power consumption has become the key specification in 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 deep-submicron technologies needed to meet the functionality requirements, stand-by power consumption is dominated by leakage currents.

The equation for subthreshold leakage as already defined in BSIM2, see also formula (1), shows which parameters contribute to the current. The dependency of the subthreshold current on bulk bias and drain induced barrier lowering (DIBL) is clearly shown. The other small channel effects can be taken into account by the effective width (W_{eff}) and length (L_{eff}) of the device.

$$I_{\text{leak}} = I_0 \cdot \exp\left(\frac{-V_{T0} - \gamma \cdot V_{BS} + \eta \cdot V_{DS}}{n \cdot V_{th}}\right) \cdot (1 - \exp(\frac{-V_{DS}}{V_{th}})) \quad (1)$$

$$I_0 = \mu_0 C_{ox} \frac{W_{\text{eff}}}{L_{\text{eff}}} (V_{th})^2 e^{1.8} \quad (2)$$

where

μ_0 : mobility
 C_{ox} : gate oxide capacitance
 W_{eff} : effective width
 L_{eff} : effective length
 I_{leak} : total subthreshold leakage
 V_{T0} : transistor threshold voltage
 γ : linearised body coefficient
 V_{BS} : bulk-source voltage
 η : DIBL coefficient
 V_{DS} : drain source voltage
 V_{th} : thermal voltage

With the advent of deep-submicron devices also came the reduction of the gate-oxide thickness. This reduction leads to a higher electric field across the oxide. As a result the tunnelling of electrons through the gate oxide into the substrate and from substrate to the gate becomes possible. This tunnelling current is referred to as gate leakage. [1] Being quantum mechanical in nature the gate leakage current is virtually temperature-independent. Simplified equations from [5] yield the following formula (3).

$$I_{\text{gate}} = K \cdot W \cdot \left(\frac{V}{t_{\text{ox}}}\right)^2 \cdot \exp(-\alpha T_{\text{ox}}/V) \quad (3)$$

where

I_{gate} : total gate oxide leakage
 t_{ox} : oxide thickness
 α and K : fit parameters
 W : gate width
 V : gate-bulk voltage

These formulas, (1) and (3), also show which degrees of freedom a designer has and their respective impact on the current. From a circuit design perspective, the width and length of the device are given by the operational constraints of the circuit. The technological parameters are considered to be a given, although channel engineering to reduce leakage in transistors is a viable option [3]. There are two choices left for a designer to reduce the leakage current. The first one is a choice on the threshold voltage.

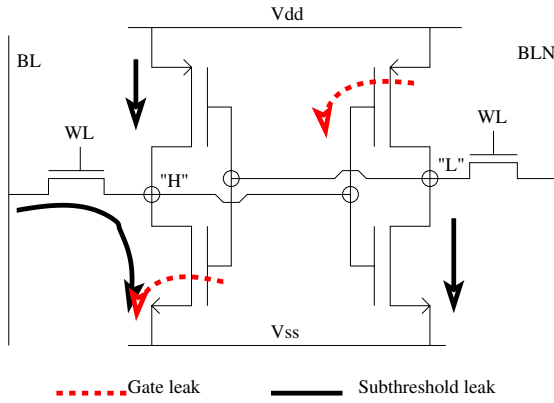


Figure 1. Leakage currents in a 6T-SRAM cell

This choice can be made by either using differently processed transistors or modulating it by varying bulk bias. Secondly the drain-source voltage can be modified to suppress the leakage current. For digital system the drain-source voltage is closely related to the supply voltage. This dependency will be actively exploited here both to reduce the subthreshold and gate leakage.

Various techniques have been developed to reduce both leakage currents, subthreshold and gate-tunnelling, in the core logic. These techniques like Multi-Threshold CMOS [9], Super Cut-off CMOS [8] and Zig-Zag Super Cut-off CMOS [2], have however the disadvantage of losing the information stored on the internal nodes when put into sleep-mode. This makes them not suitable for memory circuits, and certainly not in the data matrix. Although there are several designs that use high V_T transistors in the memory matrix to reduce the leakage. These designs suffer from a delay penalty which will only increase due to the evolution towards sub 1-V supplies.

First this paper will look into the influence of the leakage currents specifically for SRAMs. Then a solution will be proposed that reduces the leakage currents. The conclusions will be drawn and the future work will be announced.

2. SRAM specific

The biggest source of leakage power in SRAMs is the cell matrix. In this matrix generally only one word is accessed at a given moment. The rest of the cells are standing by and drawing leakage currents.

In the non-accessed cell, the pass-transistors are off. Like depicted in figure 1, there are several leakage mechanisms at work in a non-accessed cell. When the cell is in a stable written state, one of the internal nodes high or near the supply voltage, the other one low or near the ground, one PMOS and one NMOS are in the off-state. In this state they will draw a subthreshold leakage current. This current will be drained or compensated by their saturated counterparts, but represents an unuseful consumption of power. On the saturated transistors another leakage mechanism comes into play. Due to having drain and gate at a different voltage, gate leakage is present in those devices.

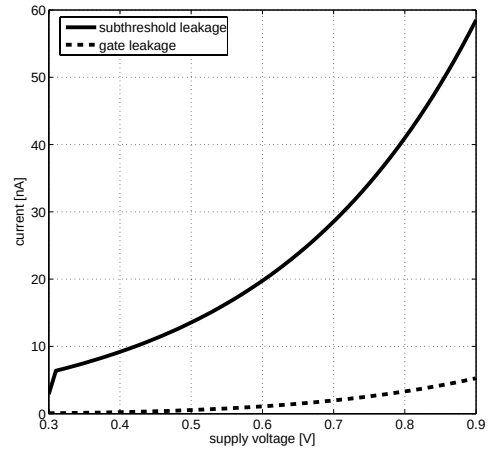


Figure 2. leakage in a 6T-SRAM cell afo the supply voltage

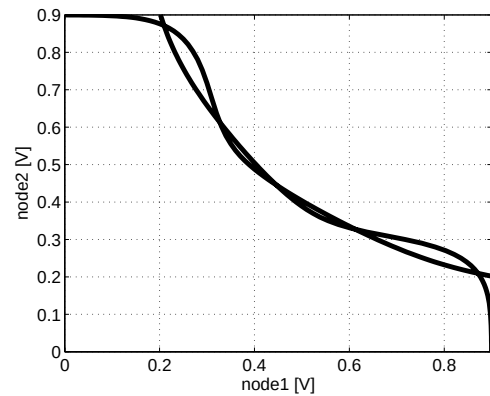


Figure 3. butterfly diagram under read conditions

The contribution of both leakage currents is shown in the figure 2 as is their dependency on the supply voltage.

The leakage currents flowing in a six transistor cell, as depicted in figure 1, have a common variable. The supply voltage plays an important role both in the subthreshold and gate leakage of the cells, see also formulas (1) and (3).

The cells are designed to guarantee a safe operation during access. The static noise margin (SNM), as defined by Seevinck et. al in [7], under halve select conditions is traditionally the determining factor in this.

SRAM cells are designed to be stable under the most stringent of conditions : half-access read. In non-accessed cells however the effective SNM is much higher due to the off state of the pass transistors as illustrated in fig 3 and 4. This will limit the possible disturbances on the internal nodes of the cell severely. The extra margin that is present due to this gives a degree of freedom in the system design of the SRAM. While a certain SNM has to be guaranteed, the supply voltage on the non-accessed cells can be lowered to reduce both subthreshold and gate leakage.

Recent publications like [4] and [6] have already proved the feasibility of this technique. However they

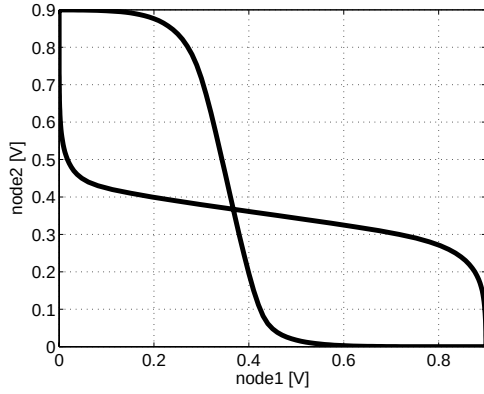


Figure 4. butterfly diagram under non-accessed conditions

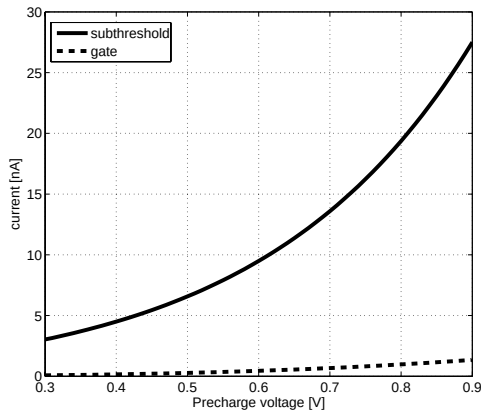


Figure 5. leakage currents in a 6T-SRAM pass transistor after the precharge voltage

still use large parts of the memory as their granularity of control.

While in the above mentioned approaches the focus lies in reducing the leakage generated by the “core” four transistors, bitline leakage also comes into play. As both pass transistors are in cut-off, they also have a subthreshold component that has to be taken into account. The wordline is kept low by the buffers in the decoders, which have to drain the gate leakage current from the pass transistors too. Both effects can be greatly reduced by lowering the precharge voltage on the bitlines as shown in figure 5. Some adaptation of the sense-amplifiers to accommodate for the lowered precharge voltage is of course necessary when applying this technique.

3. Proposed solution

Different approaches to reduce static power consumption in SRAMs using varying supply voltages already have been published, like [4] and [6]. They have a fairly large granularity of control and in the case of [6] even take the cache-access history into account. In [6] the conclusion is drawn that any control algorithm more compli-

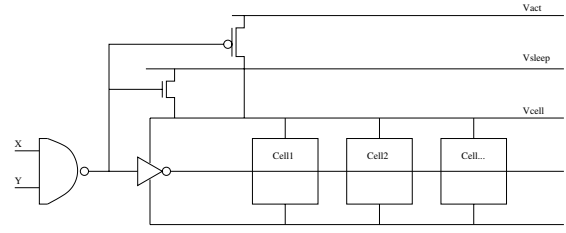


Figure 6. Word structure with distributed decoding

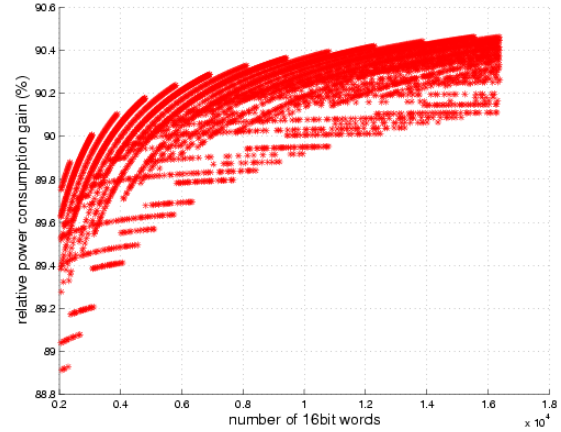


Figure 7. Relative reduction of power consumption after the number of words

cated than putting all lines periodically in sleep-mode, or “drowsy” mode is not worth the extra complexity.

The solution presented here in figure 6 however combines a small granular control with a low overhead in complexity and area. By distributing the last stage of the decoders into the matrix a form of hierarchical wordline decoding is present. This distributed last stage enables an activation control on word level, the smallest grain of control useful.

By using the local wordline also as a signal to switch between active and sleep supply a very simple but effective control system is obtained. This also implies a cost on the delay of the memory. Some time is needed to reactivate the word and some active power is sacrificed for greatly reduced leakage current.

Simulation of this concept in a 65nm technology suggest that this approach is entirely feasible. The integrity of the data during wake-up is guaranteed by firstly reducing the precharge voltage on the bitlines and secondly by having the last wordline buffer stage in the word wake-up at the same rate as the cells. The extra delay will still allow for an access time of 3ns. This is fast enough for the ultra-low power applications this is aimed at.

The figures 7 to 8 show high level simulation results for the power gain that is possible when this solution is applied. The different plots shown, include gain figures in function of size of the matrix or power supply for a fixed size. All take an 200Msample/s access activity and an 80% supply conversion efficiency into account.

One important issue [4] nor [6] have an answer for is the choice of the sleep-voltage. Both have a heuristic ap-

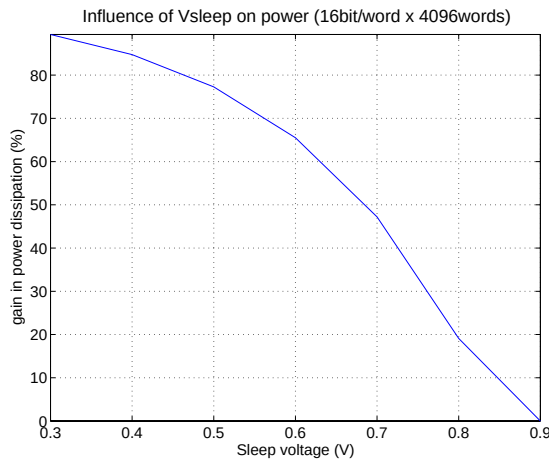


Figure 8. power consumption afo vsleep

proach to selecting that voltage. It is obvious from 1 and 3 that the sleep voltage should be chosen as low as possible. That boundary condition is given by the retention of the data in the cells and by the noise margin on top of that. In an environment where variability is a serious issue, this will require some real-time on die monitoring.

One type of monitoring has already been presented in another context by [10]. In [10] replicated flipflops are modified to fail at higher retention voltages than the flipflops used in the core of the system. This could easily be adapted for SRAM circuits. Another approach is building a control loop based on actual inline measurement of the cells. This requires some more analogue components but would provide better accuracy.

A secondary effect of the choice of the sleep voltage is the increase in wake-up delay for the circuit with lower sleep-voltages. While [4] chooses a single cycle wake-up and access, [6] takes an extra clock-cycle to wake up the circuit into account. The latter approach is more suited for high speed cached systems, where the wake-up delay is considered as a “cache-miss”.

For low power circuits where clock speeds tend to be lower anyhow, the first approach as proposed both here and in [4] are better suited. This is due to a lower penalty in delay and still maximising saving on the power by keeping as many cells as possible in their sleep state.

4. Conclusion

By manipulating the supply voltage in the SRAM cell matrix a factor 10 less power is consumed when this proposed technique is used. This can be achieved because the supply reduction works on three effects: DIBL, bulk-effect and electric field reduction. While the technique has its drawbacks in increased delay, this can be fine tuned on a higher level by modifying the sleep voltage to have less wake up latency. That comes of course at a cost of increased leakage.

Through the use of the lowest level distributed decoder it is guaranteed that a maximum of cells is kept in its sleepy state. The area overhead that comes with it is

around 10% for 90% less power dissipation and as such is not considered a high price.

5. Future work

The techniques described above are being used in the design of a 8kbyte SRAM in a 65nm technology. On top of that this design will include real-time monitoring for the SNM and form a platform to close the loop for the voltage regulation. The system is designed to give a higher than 90% saving on total power at an operating frequency of 100MHz.

6. Acknowledgements

This research was sponsored by the SLI project at IMEC Leuven.

7. References

- [1] 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
- [2] M. Drazdziulis, P. Larson-Edefors, D. Eckerbert and H. Eriksson, “A power Cut-Off Technique for Gate Leakage Suppression”, *IEEE proc. of ESSCIRC 2004*, p. 171-174
- [3] R.W. Mann et. al, “Ultralow-power SRAM technology”, *IBM Journal on Research & Development*, VOL.47, NO. 5/6, September/November 2003
- [4] 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
- [5] A. Chandrakasan, W. Bowhill and F. Fox, *Design of High-Performance Microprocessor Circuits*, IEEE Press, 2001
- [6] N. Kim et al., “Circuit and Microarchitectural Techniques for Reducing Cache Leakage Power”, *IEEE Trans. on VLSI*, vol. 12, no. 2, feb. 2004
- [7] 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
- [8] H. Kawagushi, K. Nose, T. Sakurai, “A Super Cut-Off CMOS (SCCMOS) Scheme for 0.5 V Supply Voltage With Picoampere Stand-By Current”, *IEEE Journal of Solid-State Circuits*, vol. 35, no. 10, october 2000
- [9] S. Mutoh, T. Douseki, Y. Matsuya, T. Aoki, S. Shigematsu, J. Yamada, “1-V Power Supply High-Speed Digital Circuit Technology with Multithreshold-Voltage CMOS”, *IEEE Journal of Solid-State Circuits*, vol. 30, no. 8, augustus 1995
- [10] 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

Impact of process variations on the alpha-particle-induced SER of embedded SRAMs

Tino Heijmen and Bram Kruseman

Philips Research Laboratories, Prof. Holstlaan 4 (WAY41), 5656 AA Eindhoven, The Netherlands,
email: {tino.heijmen, bram.kruseman}@philips.com

Abstract

We investigate the spread in soft-error rate (SER) owing to variations in the process parameters. The alpha-particle-induced SER of embedded SRAMs processed in 0.18 and 0.13 μm technologies was determined experimentally using accelerated testing. The SER in both the 0.18 and the 0.13 μm processes show design-to-design and batch-to-batch variations. In addition, the 0.13 μm SRAMs show a variation in SER between individual samples from the same batch. Also, we performed circuit simulations to study the statistical variations in the critical charge of the SRAM cell. The total spread in SER for 0.18 μm SRAMs obtained from our calculations is as large as a factor of 3. This result is in agreement with the experimental data. We show that the use of the high- V_T process option can reduce SER, because of a decrease in the collection of induced charges. Our results illustrate the importance of accurate simulation methods and stress the need to test several samples, batches, and designs in order to characterize the SER of an SRAM compiler.

1. Introduction

Radiation-induced soft-errors are an increasingly important reliability issue in state-of-the-art CMOS integrated circuit (IC) technologies [1]. The number of circuit nodes that are sensitive to upsets tends to grow, while the upset probability per bit stays constant or increases. Another effect of IC technology scaling is that the relative variations in the process parameters become larger. Also, the number of process options is growing with every new technology generation. For example, different values of the threshold voltage (V_T) are being applied.

In general, the SER of an IC is dominated by the SER of its embedded SRAM instances, because of the large number of sensitive bits and the relatively high upset probability. SRAM modules can be effectively protected from soft errors with error-correction coding (ECC) or cell hardening. However, these solutions are costly and/or deteriorate the performance, area, or power consumption of the design. Therefore, a reliable analysis of the expected SER for embedded SRAMs is needed to determine if protection should be applied.

In the current paper we study the impact of (unwanted) process variations on the alpha-particle-induced SER of embedded SRAMs in 0.18 and 0.13 μm technologies. Furthermore, the difference in SRAM SER for the standard- V_T (SVT) and high- V_T (HVT) process options is investigated.

2. Experiment

The chips under study are listed in Table 1. All investigated modules are single-port 6T SRAMs. Standard- V_T transistors are used in the whole memory, except if indicated otherwise. Chips A and B contain instances of the same 0.18 μm high-speed (HS) SRAM compiler. Similarly, chips C and D contain instances of the same 0.13 μm high-speed compiler. Chip E contains two instances of a 0.13 μm low-power (LP) SRAM compiler, which are identical except that one instance uses standard- V_T and the other one high- V_T transistors in the whole memory.

Chip	SRAM compiler	V_{DD} (V)	Remarks
A	0.18 μm HS	1.8	
B	0.18 μm HS	1.8	3 batches
C	0.13 μm HS	1.2	
D	0.13 μm HS	1.2	2 batches
E	0.13 μm LP	1.2	contains SVT and HVT modules

Table 1 : Devices studied in this work.

Alpha-accelerated SER (alpha-ASER) tests were performed using an Am241 alpha-emitting source, which was placed over the decapped package of the IC sample under test. The SRAMs were tested at 90, 100, and 110% of the nominal V_{DD} . All tests were done in compliance with the JESD89 standard [2].

3. Results

Alpha-ASER test results are shown in Figure 1 for the 0.18 μm high-speed SRAMs on chips A and B. The SER per Mbit for chip A is on average 40% lower than for chip B. Furthermore, the SER of batches 1 and 3 of chip B are approximately 6% higher and lower compared with batch 2 of chip B, respectively. The SER of individual samples from the same batch agree within the experimental accuracy. Figure 2 shows the results for the 0.13 μm high-speed SRAMs. The SER for chips C and D differs with 35% on average. The average difference in SER between the two batches of chip D is 16%. Different from chip B (0.18 μm), the individual samples from the same batch of chip D (0.13 μm) show a significant spread, typically of about 5%. These results show that statistical variations form a potential problem

for the accurate SER characterization of an SRAM compiler. Therefore, a more detailed statistical analysis is highly relevant.

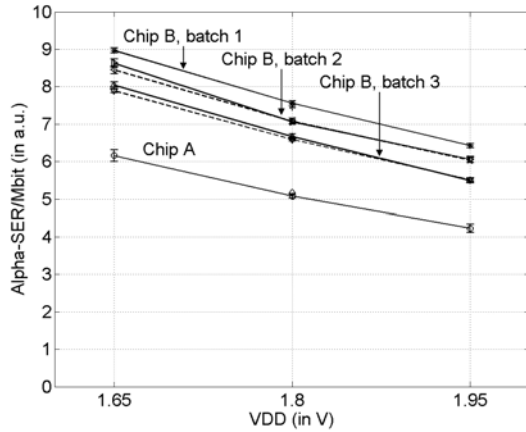


Figure 1 : Alpha-SER of 0.18 μm high-speed SRAM.

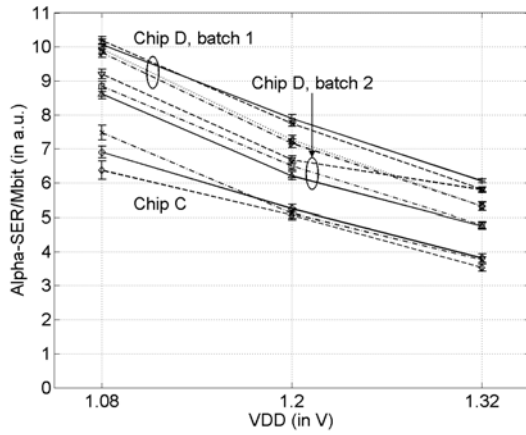


Figure 2 : Alpha-SER of 0.13 μm high-speed SRAMs.

The modelled spread in SER for the high-speed SRAM compiler in 0.18 μm is shown in Figure 3. First, the statistical variation in the critical charge (Q_{crit}) for the SRAM cell was calculated using SPICE-like circuit simulations [3]. The effect of variations in the processing parameters was evaluated with a Monte-Carlo analysis, which was repeated for the slow, nominal, and fast process corners. Then, the corresponding SER value was computed using an analytical model. This model had been constructed by combining experimental alpha-ASER data with simulated Q_{crit} results [4]. Figure 3 shows that the difference between the worst-case slow and best-case fast SER values equals a factor of 3. Fast transistors have a larger pull-up or pull-down strength than slow transistors. Therefore, the voltages levels of the storage nodes are better stabilized and Q_{crit} is larger. This explains the lower SER for the fast process corner and the higher SER for the slow corner. For convenience, the spread in the experimentally observed alpha-ASER of chips A and B have been indicated. The simulation results are in agreement with the experimentally observed variations.

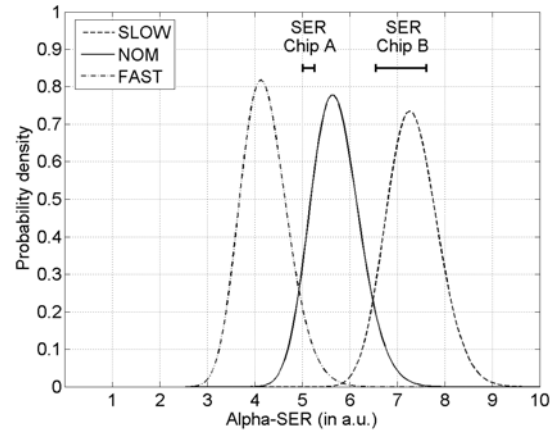


Figure 3 : Statistical variation in SER of 0.18 μm high-speed SRAMs.

Next, the sensitivity of Q_{crit} to $+4\sigma$ and -4σ variations of individual process parameters were calculated for the nominal process corner [5]. The Q_{crit} data were translated into SER variations using the analytical SER model. The results for the most sensitive parameters are shown in Figure 4.

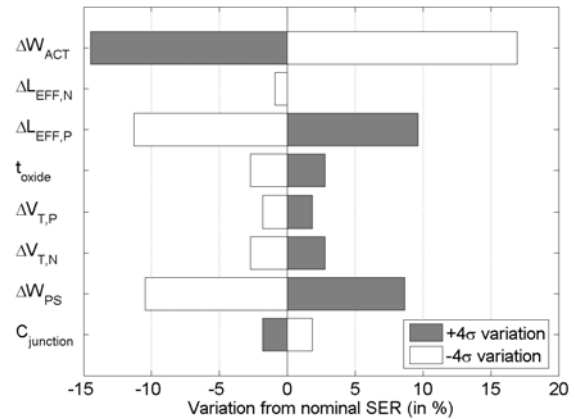


Figure 4 : Effect of processing variations on SER.

The variation of the width of the ACTIVE area (ΔW_{ACT}) is the parameter with the largest influence on SER. This parameter is directly related to the effective width of both NMOS and PMOS transistors. Also the poly-silicon width (ΔW_{PS}) and the effective PMOS transistor length ($\Delta L_{\text{EFF,P}}$) affect SER, because they are both related to the transistor length. The aspect ratio W/L of a transistor in the feedback loop directly influences Q_{crit} and thus the SER sensitivity. This is because a transistor with a larger aspect ratio has a larger gain and therefore contributes more to the active stabilization of the cell. SRAM bit cells are especially sensitive to induced electrons that are collected by storage nodes. Such charge collection results in a lowering of the node voltage, in which case the pull-up PMOS transistor tries to stabilize the node level. Therefore, the influence of the aspect ratio is larger for PMOS than for NMOS transistors.

Figure 5 shows the effect of applying the HVT instead of the SVT process option for the transistors in the SRAM modules on chip E. On average, a 43% higher

SER was measured for the SVT module compared to the HVT module. Since both modules are located on the same device and are identical except for the transistor V_T , the difference must be related to the use of the SVT and HVT process options.

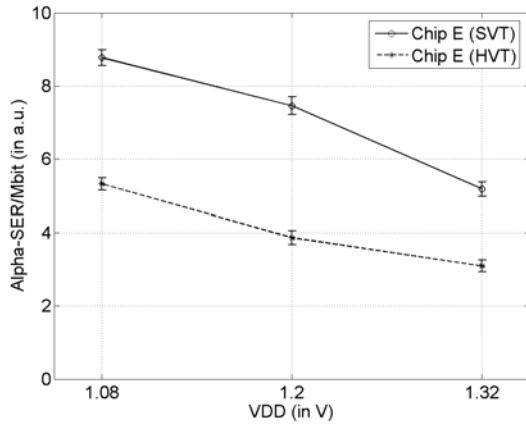


Figure 5 : Alpha-SER of 0.13 μ m low-power SRAM.

The SER sensitivity of an SRAM cell is determined by three parameters: the diffusion area of the sensitive transistor drains, the critical charge Q_{crit} , and the efficiency with which charges are collected by the circuit nodes after a particle hit. The drain diffusion area is the same for SVT and HVT, since the cells share the same layout. The difference in SER cannot be explained by a change in Q_{crit} either. The lower SER for HVT suggests a larger value of Q_{crit} . However, Figure 6 shows that actually Q_{crit} is slightly lower for HVT than for SVT. The cause of this difference is that for HVT the pull-up and pull-down transistors in the feedback loop have a smaller gain.

Consequently, the difference in SER has to be related with different charge collection efficiencies for the two process options. An HVT transistor differs from an SVT transistor in that an extra channel doping has been added to the former to increase V_T . This additional doping increases the channel resistance and thus decreases the collection efficiency for charges induced in the channel by incident alpha particles.

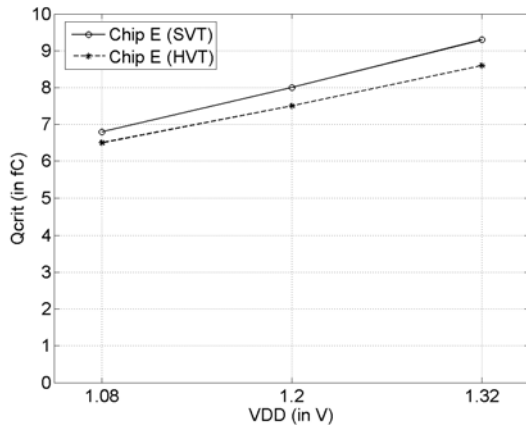


Figure 6 : Q_{crit} of 0.13 μ m low-power SRAM.

4. Conclusions and Future Work

The experimental alpha-ASER results presented in this paper show that for embedded SRAMs processed in 0.18 and 0.13 μ m technologies a substantial design-to-design and batch-to-batch variation is observed. The SER of individual samples from a single batch of 0.18 μ m SRAMs agree within the experimental accuracy. However, for a comparable 0.13 μ m SRAM compiler also a significant sample-to-sample variation is found. This result shows that the impact of processing variations on the spread in SER of embedded SRAM is growing. As a consequence, several designs, multiple batches per design, and multiple samples per batch have to be tested in order to obtain an accurate prediction of the nominal and worst-case SER for an SRAM compiler. An alternative would be to calculate SER data by simulation. However, current state-of-the-art simulation approaches are not (yet) capable of providing highly accurate SER data.

The presented simulation results show that the spread in SER for 0.18 μ m SRAMs is as large as a factor of 3. With the growing impact of process variations, the SER variation will grow for newer technologies. A clear difference in SER exists for devices manufactured at the nominal, slow, or fast process corners. An even larger spread in SER would be found if also different operational conditions were considered, i.e., when the process corners were combined with variations in temperature and supply voltage. Especially V_{DD} affects the SER, as can be seen from Figures 1, 2 and 5.

The present work considers the SER induced by alpha particles only. Also neutrons produced by cosmic rays are capable of generating soft errors. However, number statistics for neutron-accelerated SER (neutron-ASER) testing is usually not as good as for alpha-ASER testing. This results in a reduced accuracy of the experimental neutron-SER data. Accurate test data are essential to statistically confirm differences in SER between designs, batches, or samples. In general, a smaller impact of process variations on neutron-SER than on alpha-SER is expected, because neutron-SER typically shows a weaker dependency on Q_{crit} .

The application of the high- V_T process option results in a lower SER compared to standard V_T SRAMs, due to a change in the charge collection efficiency. The detailed explanation of this effect is a subject of further research. This example shows that the use of process options can influence the SER of a design, not only via Q_{crit} , but also via the charge collection efficiency.

Thus far, all presented SER data were normalized to 1 Mbit. Alternatively, the SRAM SER normalized to 1 mm^2 could be more relevant to a designer. Figure 7 shows the trend for the average SER per area as a function of technology generation. The result for the 90 nm generation is an estimation based on test data not discussed in this paper. The impact of area occupied by the memory periphery is not included. It is clear that as the SER/Mbit stays approximately equal, the SER/ mm^2 doubles with every new process technology.

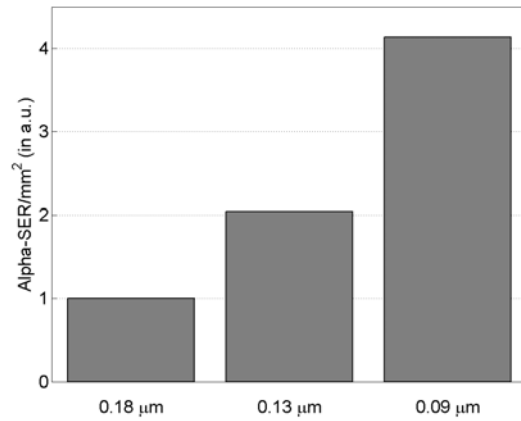


Figure 7 : Trend of average SER per SRAM area.

In summary, we have shown that wanted and unwanted process variations have a major effect on the statistical spread in the SER of embedded SRAMs. This result stresses the need for more advanced simulation techniques than currently available. Furthermore, it implies that various samples, batches, and designs have to be tested in order to accurately characterize the SER for an SRAM compiler. This is becoming increasingly important as the SRAM SER per mm² of memory is approximately doubling with every new technology generation.

Acknowledgments

The authors gratefully acknowledge the help from Rutger van Veen in the alpha-accelerated SER testing. The authors are indebted to Roelof Salters and Koos van Winkelhoff for critically reading the manuscript.

References

- [1] R. Baumann, in IEDM Tech. Dig., 329 (2002).
- [2] JEDEC standard JESD89 (2001).
- [3] T. Heijmen, B. Kruseman, R. van Veen, and M. Meijer, in Proc. IRPS, 675 (2004).
- [4] T. Heijmen, submitted to IOLTS (2005).
- [5] L.B. Freeman, IBM J. Res. Dev. **40**, 119 (1996).

High Density SRAM robust to radiation-induced Soft Errors in 90nm CMOS technology

P. Roche, F. Jacquet, G. Gasiot, C. Caillat, B. Borot, J.P. Schoellkopf

STMicroelectronics, 850 rue Jean Monnet 38926 Crolles, France, Ph. : +33 4 76 92 50 64, e-mail : philippe.roche@st.com

Abstract

A quasi alpha and neutron SER immune High Density SRAM is fabricated in a standard 90nm CMOS technology. Comparisons with ECC protected SRAM and regular SRAM are discussed. Applications of this hardening technique to TCAM and latches are illustrated. A testchip embedding various 90nm SRAM memory cells demonstrates effectiveness throughout SER tests performed in France, USA and Canada.

Keywords : SRAM, SER, TCAM, 90nm, MIM and DRAM.

Introduction

It is today averred that radiation-induced soft errors constitute a reliability issue for CMOS technologies below the 130nm node [1,2]. To meet the increasing need for both robust and highly integrated memory devices, an original 3D memory device has been developed mixing SRAM and eDRAM capacitors. This hardening technique, named rSRAMTM, aims at increasing the cell capacitance for strongly reducing the Soft Error Rate (SER) while not impacting the device area and minimizing the timing and power penalties.

In this paper, the rSRAM device fabrication and its electrical performances are first presented. To assess the rSRAM SER performances at silicon level, several testchips have been manufactured in both 120 nm and 90nm standard CMOS technologies. This paper is focused on the latest achievements in 90nm. The impact of the extra load on the timing performances is discussed at device and circuit levels by comparing 90nm robust and standard SRAM compilers. Comparisons are then presented between a standard SRAM, an rSRAM and its counterpart protected with a classical Error Code Correction (ECC). Applications to protect latch-based devices, such as Flip-flop or TCAM, are also illustrated. Finally, experimental evidences of the robustness gain are disclosed for two 90nm SRAM bit cells, 1.15 and 0.99 μm^2 , tested with alpha particles and neutrons using radiation facilities located in France, USA and Canada.

rSRAM Device Fabrication

The rSRAM cell is built in the third dimension (3D) of a standard 6 transistor High Density SRAM cell above which two stacked capacitors are symmetrically added (cf. Fig. 1). They are inserted between the polysilicon and the Metal 1 levels and manufactured with a standard embedded DRAM (eDRAM) process flow (see SEM cross-sections in Fig. 2 and 3). This architecture is generically labeled as Metal-Insulator-Metal (MIM) capacitor. The SRAM cell layout and footprint are rigorously identical to the non-hardened version with only the 3 additional layers for the capacitor definition. As another

attractive feature, this device can be indifferently processed with and without the capacitor steps resulting respectively in a hardened or non-hardened SRAM.

With our 90nm stacked MIM capacitor solution, the internal node capacitance can be increased without any area penalty up to 12fF/unit for an additional leakage current as low as 0.1 fA/unit (at 25°C and with the common electrode biased at VDD/2, cf. Fig.1). The minimum amount of charges required to flip the cell logic state, or critical charge, is accordingly highly increased ($Q=C \cdot V$) leading a much lower SER. Note that other types of capacitors have been recently proposed to reduce the SRAM SER [3,4,5], but with only few added fF per cell (implying a lower robustness gain) or using a non standard process. Table 1 synthesizes the main differences between other published “6T+2C” robust SRAM solutions.

rSRAM Design Performances

As previously pointed out, the rSRAM cell symmetry and the transistor sizing remain strictly identical to the standard version. Both the read current and the Static Noise Margin (SNM) are unchanged, whereas the intrinsic write operation is slowed down proportionally to the extra loads on the two internal nodes. For a target capacitor value of 12fF per node, the memory cell write time is increased by a factor 3 in typical conditions (see Fig. 4). However, this intrinsic timing penalty only results for the entire range of a robust single port compiler in a +30% maximum degradation of the write cycle time, compared to its equivalent non robust version (see Fig.5). Concurrently, the access time and read cycle time are not significantly impacted i.e. by less than 1%. As shown in Fig. 6 and 7, other trade-offs between robustness and write-timing performances can be achieved by acting on either the capacitor shape or dielectric.

For a Single Port SRAM, Table 2 summarizes the comparison between a standard, an rSRAM and an ECC solution (pipeline architecture). The parameters for this comparison on a 256 Kb cut in normal operation conditions are: cycle time, access time, area, power and leakage. The strong penalties measured for the ECC solution are induced by the additional parity bits (+7 bits for a 32 bit-word using a single error correction and double error detection algorithm) and by the extra combinatorial logic. The High Density of the rSRAM is highlighted in Table 2 by comparing the costs at wafer and die levels for the three solutions. Despite of a higher wafer cost induced by the additional eDRAM process steps (+3 masks), the die cost of the rSRAM is lower in this example than the ECC counterpart. Note that there is no yield loss for the rSRAM.

Other Applications

Beyond the attractive alternative to ECC, other potential interests of the rSRAM arise for flip-chip package and logic hardening. First, this solution avoids using the classic and costly “keep out areas” for SRAM and latches beneath and around the lead solder bumps [1]. When there is a large amount of bumps as explained in [6], it would also allow minimizing design complexity and chip area. Second, the rSRAM concept can simultaneously protect the memory and logic parts of an ASIC without respectively using ECC and Triple Module Redundancy [1]. As an illustration, Figure 8 shows a Flip-flop hardened with the rSRAM technique applied to both the master and the slave latches. As for the SRAM memory cell, the addition of the capacitors degrades the timing performances. Slight modifications (not detailed in this paper) to the design have been applied to minimize this timing degradation with an area cost of +5%.

Such a robust solution is also applicable to multiple port SRAM, register file, high speed cache, CAM, and all kinds of memory structure for which usual Error Code Correction (ECC) techniques cannot be easily implemented. As an illustration, a 90nm High Density Ternary CAM memory cell has been hardened without any area penalty using the rSRAM technique applied on both the data and mask storage nodes. As previously discussed for the rSRAM, the write cycle time of this rTCAM is increased. The access and read cycle times are again not significantly impacted (less than 1%). The comparison mode remains unchanged with this hardening technique compared to a standard design. This feature combined with the high integration, constitutes a key advantage over an ECC implementation.

Testchips and experimental SER characterizations

The first experimental evidence of the hardening efficiency provided by the rSRAM technique was presented last year [7], with a testchip fabricated in a 120nm CMOS technology. Accelerated SER tests exhibited a complete alpha immunity, even at a very low supply voltage. Full neutron insensitivity was measured at Los Alamos at 1.32V with a residual Failure-In-Time (FIT) rate at 1.2V i.e. below 10 FIT/Mb.

New validation testchips have been fabricated with a 90nm CMOS technology. We present in this paper experimental SER results on one of these circuits embedding different robust and standard SRAMs. As shown on the circuit layout in Fig.9, four memory cells are used: 1.15 and 0.99 μm^2 , standard and robust. The circuit has been successfully tested encapsulated in a PBGA package.

An Americium 241 alpha radioactive source was used with normal chip operating conditions and power supplies ranging from 0.8 to 1.5V. Other test parameters (test pattern, frequency, error bitmaps, etc) not detailed here, have been used to verify the overall consistency of the results. The characterization procedure was strictly JESD89 [8] compliant. Fig 10 plots the measured alpha-SER vs. VDD at room temperature for each

robust and standard memory cells. These experimental data again prove the strong robustness enhancement: the two rSRAM cells are quasi invulnerable (<0.2 FIT/Mb is equivalent to a Mean-Time-Between-Failure $>1\text{E}6$ years) whereas the regular 1Mbit SRAMs experienced hundreds of fails/sec under the same intense alpha bombing.

The neutron tests were performed at the Canadian TRIUMF facility [9] with again a full JESD89 compliance. The experimental results, still under analysis at the time of the paper writing, have shown that the rSRAM cell is here again more robust than the conventional one. The neutron SER is reduced below 100 events per Mb instead of hundreds in a worst case configuration i.e. at 0.9V. It is noteworthy that no neutron-induced latch-up was recorded even at extreme conditions (1.65V and 125degC) [10].

Conclusion

The 90nm rSRAM devices behave under radiations like in the CMOS 120nm node: they are quasi alpha immune and very robust against neutrons in nominal operating conditions.

As the rSRAM is using a standard process option relying on the CMOS eDRAM process flow, it is directly scalable to the next CMOS generations. The 65nm rSRAM cells, ranging from 0.499 μm^2 to 0.615 μm^2 , have been drawn without any area penalty and are currently manufactured on testchips.

Reference

- [1] R. Baumann, “SER tutorial”, Proceeding IEDM 2001.
- [2] P. Roche *et al.*, “Comparisons of Soft Error Rate for SRAMs in Commercial SOI and Bulk below the 130 nm Technology Node”, *IEEE Trans. Nucl. Sci.*, vol. 40, Dec. 2003.
- [3] Soon-Moon Jung *et al.*, “Soft error immune 0.46 μm^2 SRAM cell with MIM node capacitor by 65 nm CMOS technology for ultra high speed SRAM,” *IEEE International Electron Devices Meeting IEDM 2003, Technical Digest*, Pages:11.4.1 - 11.4.4, 8-10 Dec. 2003.
- [4] Y. Z. Xu *et al.*, “Process impact on SRAM alpha-particle SEU performance,” 2004 IEEE 42nd Annual International Reliability Physics Symposium Proceedings, Pages:294 - 299, 25-29 April 2004.
- [5] Y. Fujii *et al.*, “Soft error free, low power and low cost superSRAM with 0.98 μm^2 cell by utilizing existing 0.15 μm -DRAM process,” 2004 Symposium on VLSI Technology, Digest of Technical Papers, Pages:232 - 233, 15-17 June 2004.
- [6] J. Stinson *et al.*, “A 1.5GHz Third Generation Itanium Processor”, Proceeding ISSCC 2003, February 2003.
- [7] P. Roche, F. Jacquet, C. Caillat, J. P. Schoellkopf, “An Alpha Immune and Ultra Low Neutron SER High Density SRAM”, proceedings of *IRPS 2004*, pp671-672, April 2004.
- [8] JEDEC standard n° JESD89, www.jedec.org.
- [9] E. Blackmore, 2000 IEEE Radiation Effects Data Workshop, pp.1-5, Reno, Nevada.
- [10] P. E. Dodd *et al.*, “Neutron-induced soft errors, latchup, and comparison of SER test methods for SRAM technologies,” International Electron Devices Meeting 2002, IEDM 2002. Digest., pp. 333-336, 8-11 Dec. 2002.

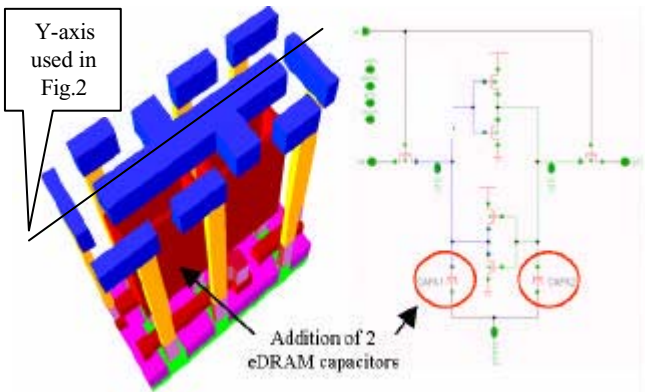


Fig.1: Overview of the rSRAM cell (6T+2C). The rSRAM cell is built from a standard 6 transistor High HD SRAM cell above which two stacked capacitors are symmetrically added. The common electrode is biased at VDD/2.

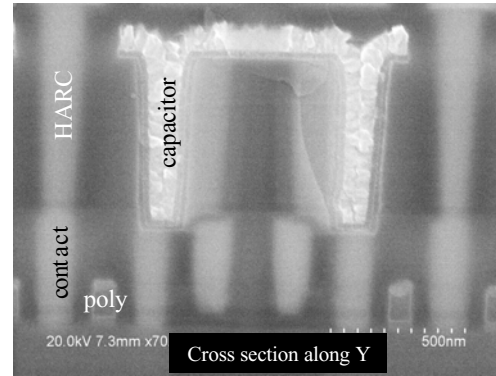


Fig.2: SEM cross-section of the 1.15μm² rSRAM cell. The capacitors are located between polysilicon and Metal1 Levels. Three additional mask levels are required to build the capacitor: one for the capacitor's trench, one for the top electrode definition and a third one for the High Aspect Ratio Contact (HARC).

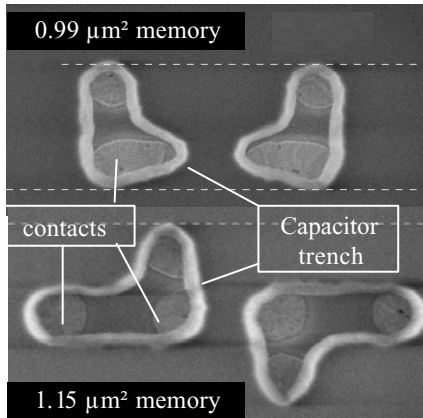


Fig.3: SEM top views of the 0.99μm² and 1.15μm² rSRAM cells at the process step of trenches' formation.

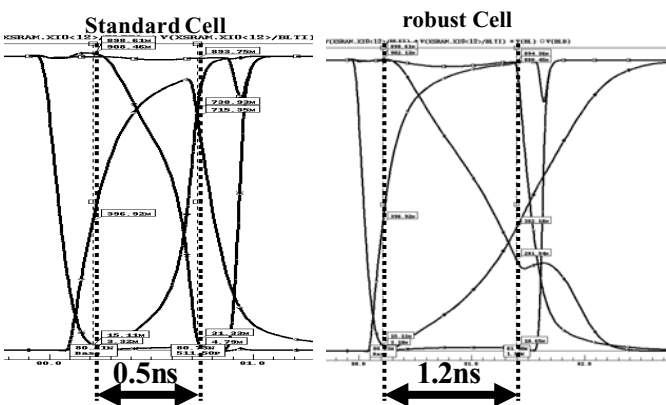


Fig. 4: Intrinsic write time of the 1.15μm² rSRAM memory cell with an added capacitor of 12fF. Simulation conditions are: nSlow pFast - 0.9v - 125°C.

	Ref. [5]	Ref. [3]	Ref. [4]	This paper
Type of capacitor	MIM capacitor	MIM capacitor	MIM capacitor	eDRAM capacitor
Technology	Thin Film Transistor technology	Standard option	Standard option	Standard option
Capacitor added per node / technology	+20fF fF/cell /150nm	+2 fF/node /65nm	+2fF/node /150nm	+36fF/node /130nm +12fF/node /90nm
Cell area	0.5 (TFT PMOS)	1.0	1.0	1.0
SER reduction	1000x (only measured with alpha)	6x (only simulated with neutrons)	6x (only measured with alpha)	250x [6] (measured with neutron and alpha)

Table 1: Comparison of various published "6T+2C" robust SRAM solution recently presented in international conferences.

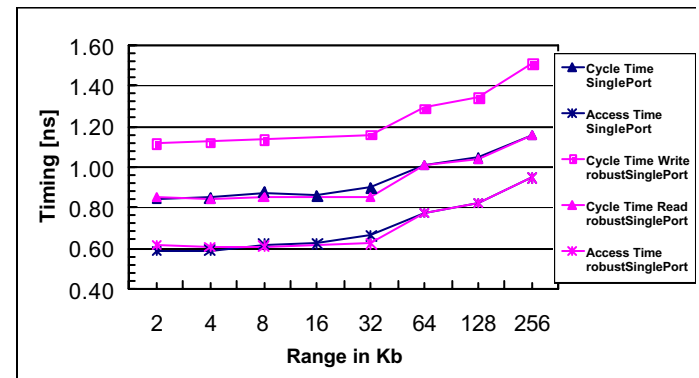


Fig.5: Comparisons of cycle and access times between two 90nm Single Port SRAM compilers: a standard and a robust one. Simulation conditions are: typical process corner - nominal voltage - room temperature.

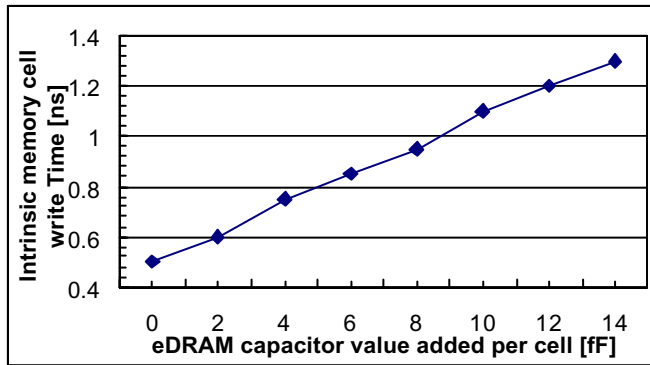


Fig.6: Memory cell write time for a 90nm standard single rSRAM as a function of the added capacitor value. Simulation conditions are: typical process corner - nominal voltage – room temperature.

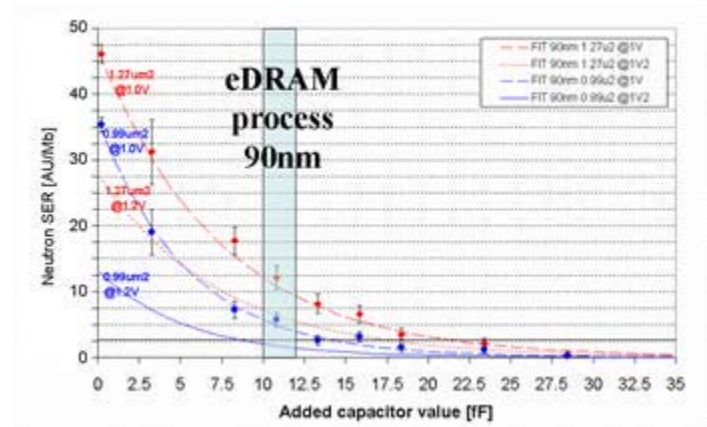


Fig.7: Normalized SER rates for the rSRAM as a function of the added capacitor value. Simulations performed with proprietary Monte Carlo codes for various memory cells and power supplies [2].

For a same cut 8Kx32 ST CMOS 90nm GP Typ case , 1.0V, 25°C	SRAM Specs	rSRAM ? in %	SRAM ECC (pipeline) ? in %
tcycle read (ns)	1.16	+1 %	0 %
tcycle write (ns)	1.50	+30 %	0 %
taa (ns)	0.95	+1 %	one cycle later
area (mm2)	0.4	+0	+25%
power (uW/MHz)	50	+5 %	+40 %
leakage (uA, WC)	400	+0	+25 %
Process	CMOS	CMOS eDRAM	CMOS
Wafer Cost	1	+10%	+0
Die Cost (10Mb/ 32-bit word)	1	+10%	+25%

Table 2: Comparisons in arbitrary units of Standard, Robust and ECC protected SRAMs for the same 256Kb cut in nominal conditions for a standard 90nm CMOS technology.

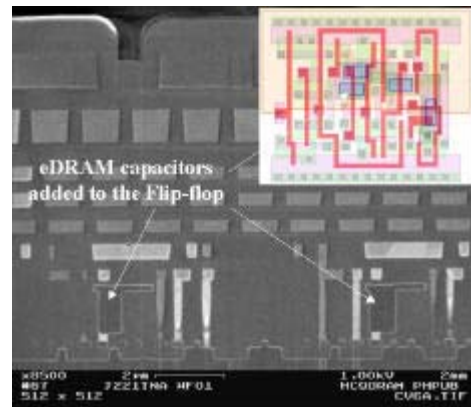


Fig.8: SEM cross-section and layout of a robust flip-flop, protected by 2 pairs of stacked capacitors.

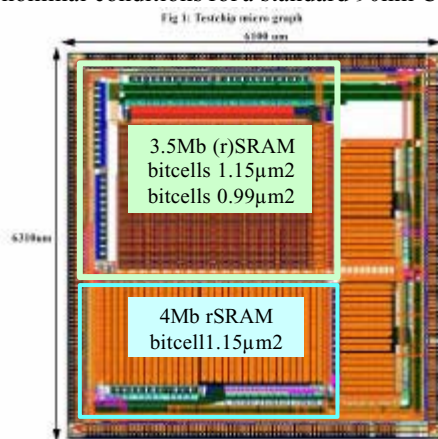


Fig.9: Layout of the validation circuit fabricated in a 90nm CMOS technology. It contains several standard and robust SRAMs with two different memory cells 1.15μm² and 0.99μm².

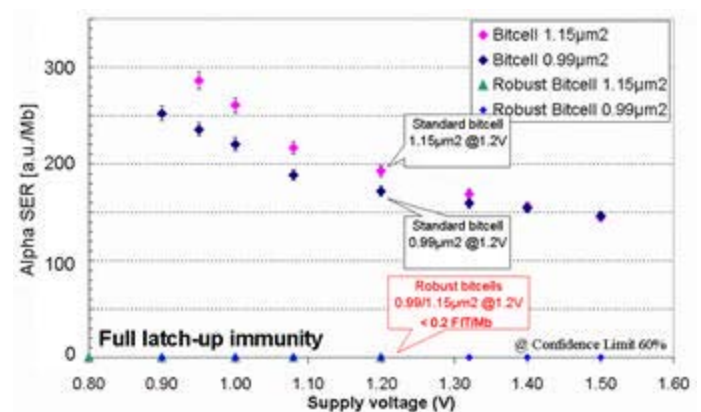


Fig.10: Normalized alpha SER results for the 90nm SRAM and rSRAMs memory cells. Variations with the power supply at room temperature. Extrapolations use a natural flux of 1E-3 alpha/cm².h.

A model to explain the C-V curves of DRAM capacitors with silicon electrodes and trapping dielectrics

L. Lopez^{a, b}, P. Masson^a, D. Nee^b and R. Bouchakour^a

^a L2MP-UMR CNRS 6137, IMT-Technopôle de Château Gombert, 13451 Marseille cedex 20, France.

^b STMicroelectronics, ZI de Rousset BP2, 13106 Rousset cedex, France.

pascal.masson@polytech.univ-mrs.fr

Abstract

The DRAM capacitor that we have studied in this paper is composed of two n-type-Si electrodes and of an ON (Oxide-Nitride) dielectric. The C-V curve of this device has a U shape however the simulated curve with a classical modeling has a bell shape. We propose a new model that can explain this anomalous curve shape. This model is based on a dynamic charge trapping in the Oxide/Nitride interface. This model has been validated by C-V measurement using three different devices.

1. Introduction

DRAM capacitors are designed to get an expected capacitance value with a small capacitance variation versus polarization. However, the shape of C-V curves of DRAM capacitors is rarely investigated. The aim of this paper is to explain for the first time the U shape of the C-V curves of the DRAM capacitors made of n-type-Si electrodes and trapping dielectric like Si_3N_4 , Ta_2O_5 or Al_2O_3 . This problem has been rarely shown [1] and has been never clearly explained. As shown by figure 1(a), the classical modeling of this kind of device leads to a bell shape. The classical modeling takes into account the depletion effect of the both silicon electrodes.

2. Experimental details

The schematic structure of the devices that we used is shown in figure 2(a). We have used three devices, Device A, Device B and Device C, that come from a planar test vehicle that is described elsewhere [2]. The details of the physical parameters are summarized in table I. The main difference between the devices is the bottom electrode (electrode 1) that is made of polysilicon for Device A and Hemispherical Silicon Grain (HSG) [3] for the other ones. Moreover, the HSG layer is n-doped by implantation whereas the polysilicon layer is n-doped in situ. Actually, the electrode 1 of device A has been performed using the same process operation than the one of the electrode 2. The implantation dose of the electrode 1 of Device C is equal to the one of Device B divided by 10. The ON deposition operation is the same for all the devices. This operation is composed of three steps: oxidation of electrode 1; nitridation of the oxide; nitride deposition. The use of different materials to perform electrode 1 has induced a variation of the thicknesses of the grown oxide and the deposited nitride that compose the ON. As a consequence, the thickness of the ON

deposited on polysilicon is higher than the one on HSG. These layers have been characterized by Transmission Electron Microscopy (TEM). The Area Enhancement Factor (AEF) induced by HSG has been empirically evaluated, using the results of TEM and capacitance characterizations. C-V curves are performed by an Agilent 4156C using the Quasi-static measurement with leakage compensation. The gate potential is applied on electrode 2 and electrode 1 is grounded.

3. Modeling

Our C-V modeling is based on a parasitic dynamic charge that induced a parasitic capacitance that distorts the C-V curve. We have taken the assumption that a parasitic charge Q_{int} is found in the Oxide/Nitride interface. This charge is expected to be equal to zero when $V_G = 0$.

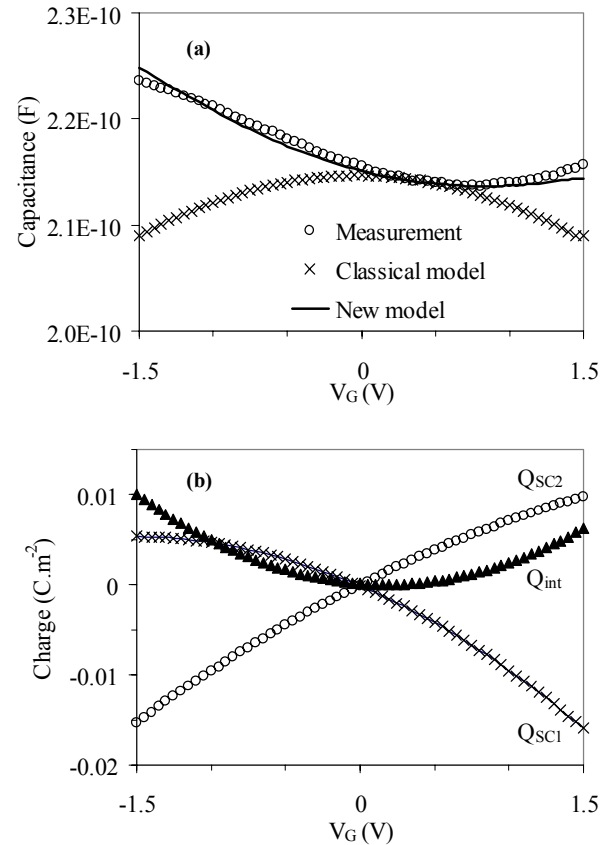


Fig. 1: Simulation and measurement of Device A: (a) Comparison simulation/measurement of the C-V curves; (b) Simulation of the Q-V curves.

As said by S. D. Khanin [4], the gate potential of the structure is given by:

$$V_G = \Phi_{S2S1} + 4\psi_{S2} + 2\psi_{S1} + 2V_{ON} \quad (1)$$

Where Φ_{S2S1} is the contact potential difference between electrode 2 and electrode 1; ψ_{S1} and ψ_{S2} are respectively the surface potentials of electrode 1 and electrode 2; V_{ON} is the voltage drop across the ON.

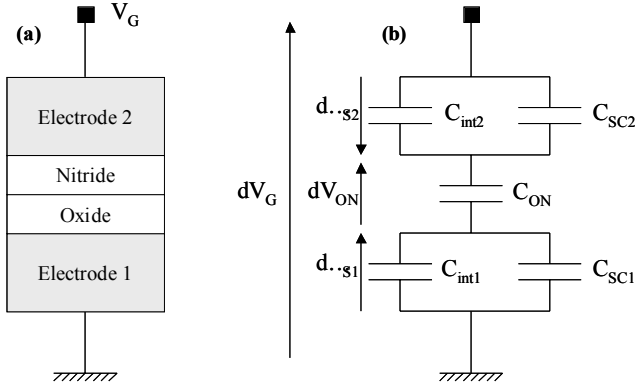


Fig. 2: Schemes of the capacitor: (a) Scheme of the structure (b) Equivalent circuit of the model.

	Device A	Device B	Device C
Material of electrode 1	Polysilicon	HSG	HSG
Material of electrode 2	Polysilicon	Polysilicon	Polysilicon
Mask Area (μm^2)	27250	27250	27250
Area Enhancement Factor	1	1.9	1.9
Oxide thickness (nm)	1.3	1.0	1.0
Nitride thickness (nm)	5.5	3.8	3.8
Doping concentration of electrode 1 (cm^{-3})	$7 \cdot 10^{19}$	$2.5 \cdot 10^{19}$	$2.5 \cdot 10^{18}$
Doping concentration of electrode 2 (cm^{-3})	$7 \cdot 10^{19}$	$7 \cdot 10^{19}$	$7 \cdot 10^{19}$

Table 1: Physical characteristics of the devices.

By developing the expression of V_{ON} , we obtain:

$$V_G = \Phi_{S2S1} + 4\psi_{S2} + 2\psi_{S1} + 2 \frac{Q_{SC2}/\psi_{S2} + Q_{int2}}{C_{ON}} \quad (2)$$

Where C_{ON} is the capacitance of the ON dielectric layer; Q_{SC2} is the charge in the semiconductor given by Sze [5] of electrode 2; Q_{int2} is the equivalent charge in electrode 2 induced by the parasitic charge Q_{int} and its centroid. The calculation of Q_{int2} is given by:

$$Q_{int2} = Q_{int} \frac{T_{SiO2}}{T_{SiO2} + 2T_{Si3N4} \frac{\epsilon_{SiO2}}{\epsilon_{Si3N4}}} \quad (3)$$

Where T_{Si3N4} and T_{SiO2} are respectively the thickness of the nitride layer and the thickness of the oxide layer; ϵ_{Si3N4} and ϵ_{SiO2} are respectively the dielectric constants of the nitride layer and of the oxide layer.

The neutrality of charges is given by (4). Using (2) and (4), we can calculate the surface potentials for a given V_G and a given Q_{int} .

$$Q_{SC1}/\psi_{S1} + 2Q_{SC2}/\psi_{S2} + 2Q_{int} = 0 \quad (4)$$

The calculation of the total capacitance of the structure is given by:

$$C/V_G = \frac{dQ_{SC2}/\psi_{S2} + 2dQ_{int2}}{dV_G} + 4 \frac{dQ_{SC1}/\psi_{S1} + 2dQ_{int1}}{dV_G} \quad (5)$$

Where Q_{SC1} and Q_{int1} are respectively the charge of the semiconductor of electrode 1 and the equivalent charge in electrode 1, induced by the parasitic charge Q_{int} and its centroid.

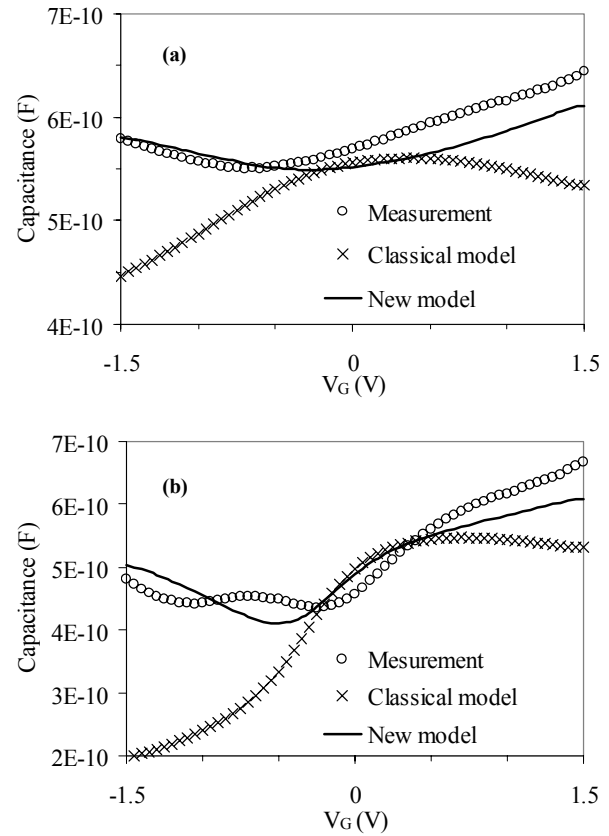


Fig. 3: Comparison simulation/measurement of the C-V curves: (a) Device B; (b) Device C.

Considering $C_{int1} = -dQ_{int1}/d\psi_{S1}$ and $C_{int2} = -dQ_{int2}/d\psi_{S2}$, as the equivalent parasitic capacitance respectively of electrode 1 and of electrode 2 and considering $C_{SC1} = -dQ_{SC1}/d\psi_{S1}$ and $C_{SC2} = -dQ_{SC2}/d\psi_{S2}$, as the depletion capacitances respectively of electrode 1 and of electrode 2, the inverse capacitance expression of (6) can be replaced by (7). Thus, we obtain the equivalent circuit in figure 2(b).

$$\frac{1}{C} = 4 \frac{d\psi_{S1}}{dQ_{SC1}} + 4 \frac{d\psi_{S2}}{dQ_{SC2}} + 2 \frac{dV_{ON}}{dQ_{SC2} + 2dQ_{int2}} \quad (6)$$

$$\frac{1}{C} = \frac{1}{C_{SC1}} + \frac{1}{2C_{int1}} + \frac{1}{C_{SC2}} + \frac{1}{2C_{int2}} + \frac{1}{C_{ON}} \quad (7)$$

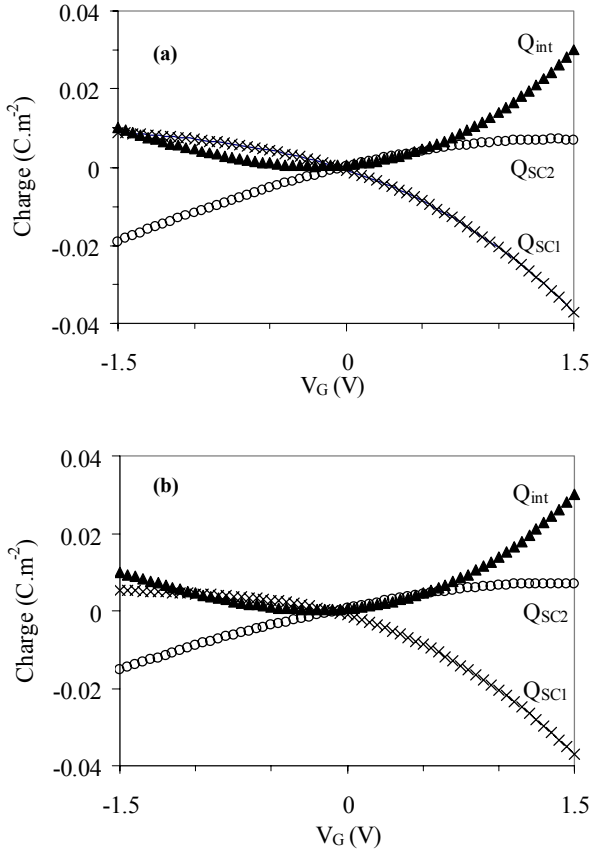


Fig. 4: Simulation of the Q-V curves: (a) Device B; (b) Device C.

4. Results and discussion

As when can see in figure 1, the measured C-V curve of Device A has a U shape. Thus, the HSG layer is not the cause of the distortion of the C-V curve. The assumption of a parasitic charge Q_{int} in the Oxide/Nitride interface has been chosen for several reasons. Firstly, the distortion is seeable for positive and negative polarization and for quasi-static or high frequency [2] capacitance measurement. Thus, the parasitic charge can't be situated in one dielectric/electrode interface. Secondly, some models [6] of MNOS memories use the assumption of a static charge trapped in the oxide/nitride interface. In our case, the trapped charge is not static because the dielectric layers are too thin: there is no charge retention. Moreover, the electrical characteristics of our devices are not modified by UV exposition. That is why we have taken the assumption that $Q_{int} = 0$ when $V_G = 0$.

We have used a polynomial expression to model the variation of Q_{int} with V_G in order to fit the simulated C-V curves with the measured ones. We have used the same data of Q_{int} for Device B and Device C. The Q_{int} values are lower for Device A. This is explained by the higher ON thickness of Device A because the mechanism of trapping of Q_{int} is expected to be linked to the leakage of the structure. The calculated values of charges

corresponding to the C-V curves of Device A in figure 1(a) are plotted in figure 1(b). The calculated values of charges corresponding to the C-V curves of Device B and Device C in figure 3 are plotted in figure 4.

In figure 1(a), the comparison between measurement and simulation of the C-V curve of Device A with our new model is very good. In figure 3, the comparison between measurement and simulation of the C-V curves of Device B and Device C with our new model is good but not perfect. The presence of HSG in theses two last devices induces complex effects (doping concentration variation, dielectric thickness variation...) that are not taken into account in our simple 1D modeling. However, the shape and the order of magnitude of the simulation with our new model is quite good comparing to the classical modeling that doesn't use Q_{int} . One can note that the only parameter that is modified between the simulation of Device B and Device C is the doping concentration of electrode 1. As we can see in figure 4, the simulated charges of Devices B and Devices C are very similar. However, the curves of the surface potentials versus V_G of Device B and Device C are not the same, as shown by figure 5. In spite of the U shape of the C-V curves, figure 5 shows that the regimes of the both electrodes are classical: with a positive V_G , the electrode 2 is depleted ($\Phi_{S2} < 0$) and the electrode 1 is accumulated ($\Phi_{S1} > 0$); with a negative V_G , the electrode 2 is accumulated ($\Phi_{S2} > 0$) and the electrode 1 is depleted ($\Phi_{S1} < 0$).

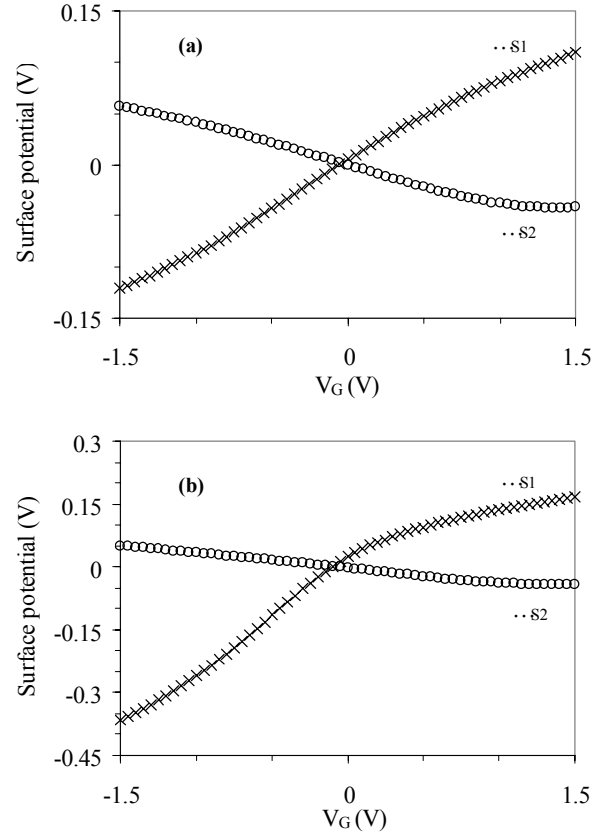


Fig. 5: Simulation of the surface potential curves: (a) Device B; (b) Device C.

5. Conclusion

For the first time, we have proposed and validated a model that can describe the anomalous C-V curves of DRAM capacitors made of n-type-Si electrodes and ON dielectrics. We have shown that the U-shape C-V curve is induced by a parasitic capacitance coming from a dynamic charge trapping in the oxide/nitride interface.

References

- [1] S. Bruyère, D. Roy, D. Jacques, C. Boccacio, IEEE IRPS, 298 (2003).
- [2] L. Lopez, D. Nee, P. Masson, R. Bouchakour, IEEE IRPS, 498 (2004).
- [3] H. Watanabe, N. Aoto, S. Adachi, T. Ishijima, R. Ikawa, K. Terada, Appl. Phys. Lett. **58**, 251 (1991).
- [4] S. D. Khanin, I. A. Uritskaya, V. Y. Uritsky, , IEEE Int. Symp. on Electrets, 211 (1996).
- [5] S. M. Sze, *Physics of Semiconductor Devices*, Wiley interscience (1981).
- [6] D. Frohman-Bentchkowsky, M. Lenzlinger, J. of appl. Phys. **40**, 3307 (1969).

Timed verification of the SPSMALL memory*

Manuel BACLET¹ and Rémy CHEVALLIER²

¹ LSV - CNRS UMR 8643 - ENS de Cachan

61, Av. du Président Wilson - 94235 Cachan Cedex - France

baclet@lsv.ens-cachan.fr

² ST Microelectronics - D.A.I.S - Central R&D

850, rue Jean Monnet - 38921 Crolles cedex - FRANCE

remy.chevallier@st.com

Abstract

The aim of the paper is to verify a small synchronous memory component with the real-time model checker Uppaal, taking into account the electrical propagation delays through gates and along wires.

1 Introduction

Timed automata Timed automata have been proposed by R. Alur and D. L. Dill in [AD90]. They are a suitable model for describing and verifying real-time systems. They have been widely studied and efficient algorithms have been designed and implemented in real-time model-checker such as HyTech¹, Kronos² or Uppaal³ and the verification of many real cases has been successfully carried out.

System on Chips The micro-electronics firms have two main focuses on the digital products: firstly the improvement of the product capabilities, and secondly the decrease of the product cost. In order to conciliate these two antagonistic objectives, the only way is the increase of the system integration capability. One chip includes today more functionality than few chips of yesterday plugged together: a complete system, processors, peripherals and memories are gathered now in the same component. This component is called a System on Chip (SoC). The other big issue is that the improvement of the system capability induces the explosion of the data flows in the SoC and then, the explosion

of the memory size needed in the component. The memories integrated in a SoC take then more and more surface on the silicon and the time spent by the system to read and write data in the memories is directly impacting the performances of the SoC. The timings of the memories must be then carefully computed and checked. The verification by timed automata is then a good solution in order to be sure that the right behavior of the memory is preserved even if the timings of the memories needed by the SoC are very aggressive.

The Spsmall memory component In this paper, we propose to model and verify the Spsmall component, which is small SRAM module manufactured by ST Microelectronics, using the real-time model checker Uppaal. The main point is that we want to take into account the propagation delay of electrical signals through transistors and along wires. The Spsmall module is a small memory with a maximum total capacity of 64 kbits (3 to 512 words of 2 to 256 bits). We chose to model the smallest memory consisting in three words of two bits, which leads to a netlist of 305 transistors.

The electrical network consists in the connection of several *logical gates* and *registers*. Each component has a characteristic *propagation delay* which corresponds to the time needed to obtain the right output value if its inputs are kept constant. Those

*Partially supported by project MEDEA+ Blueberries

1. <http://www-cad.eecs.berkeley.edu/~tah/HyTech/>

2. <http://www-verimag.imag.fr/TEMPORISE/kronos/>

3. <http://www.uppaal.com/>

delays leads to the definition of *setup* and *hold* time for the input signals of the memory.

Related works Timed verification of digital circuits has been studied in [BJMY02]. The authors succeeded the checking of various circuits of sizes up to 22 gates. If we suppose that 6 transistors are needed for each gate, this leads to more or less 132 transistors. Hence, the SPSMALL memory, with its 305 transistors, is beyond this threshold. In our approach, we take into account several characteristics that allows the verification of much larger circuits:

- the module we want to model is a memory. Hence, we use the fact that the module is naturally divided into several subentities, namely the input latches, the address decoder, the memory array and the output stage.
- we take advantage of the *symetry* of the design, since the memory array consists in 3 banks of 2 bits memories.

Finally, we think that our approach can be adapted to deal with other memory design.

Outline The memory component and the specifications it is intended to have are detailed in section 2. In section 3, we describe how the modeling of the memory was carried out. Finally, in section 4, we detail our results.

2 Description of the memory

2.1 Functional description

The memory is a synchronous device: all operations (read, write) are synchronized with the rising edge of a unique clock. We now describe briefly the different signals involved in these operations. Address (A), input data (D), write enable not (WEN) and chip select not (CSN) signals are latched in at the rising edge of the clock (CK). If CSN is set high at the rising edge of CK, it inhibits the access to the memory for the current clock cycle, and the memory is held in stand-by mode. WEN precises the direction of a memory operation (read when set to 1 and write when set to 0). Output data are internally latched at the completion of each access, and

are valid until the next positive edge of CK. During write operations, data written into the memory are copied back to the output data pins (write-through). The OEN signal is used to disable the output pins (high impedance state).

2.2 Timing description

Two timings are defined for each input signals: the setup timing, *setup* and the hold timing, *hold*. These timings are the delays before and after the rising edge of the clock for the signals to be stable to be correctly interpreted by the system. This means that if a rising edge occurs at date t , the input signal must keep constant value in the interval $[t - \text{setup}, t + \text{hold}]$. Those definitions are motivated by the fact that the electrical signals are not instantaneous and hence delays of propagation along wires and through gates must be considered to obtain right behaviors.

2.3 Architectural description

The real implementation of the memory is described in Fig. 1.

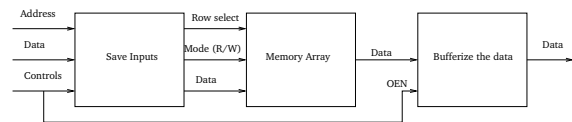


Figure 1: General design

The «save inputs» part is connected to the input signal and saves in registers the value of the data, the address, the chip select signal and the write enable signal. The «memory array» part includes the memory array and the select logic associated to it. The «bufferize the data» part manages the outputs with the output enable signal and the data saved in the memory or the data set at the input of the design (write mode).

2.4 Specifications

The properties we want to check are the following:

- When the environment performs a write operation, the new data is effectively stored in the memory (if the address is valid) and available

on the output pin *taaw* ps after the rising edge of the clock. (At a rising edge of the clock, if *WEN* is set to 0, *CSN* to 0,

- if A_0A_1 is not 11, the value of the input data D_0D_1 is saved into the memory array at address A_0A_1
 - the output Q_0Q_1 is set to the value of D_0D_1 after *taaw* ps.)
- ii. When the environment performs a read operation, the right value stored in the memory array is available on the output pins *taa* ps after the rising edge of the clock. (At a rising edge of the clock, if *WEN* is set to 1 and *CSN* to 0, Q_0Q_1 is set to the value kept into memory at address A_0A_1 after *taa* ps. If the address is invalid, the output values remains unchanged)

3 Modeling

We propose to model this memory by a network of timed automata. We were provided a RTL netlist described in VHDL which was obtained from the transistor netlist with the tool TLL [BDPG02] from the TNI-Valiosys company which is involved in the MEDEA+ Bluberries project. This tool was also used to compute the values of the propagation delays along wires and through gates.

The main complexity issue we faced was the potentially high number of clocks if we model each gate separately, leading to untractable verification. Hence we decided to carry out several abstractions and we now detail them.

3.1 Abstracting the model

On our first tries, the model-checking was not tractable. Hence, we proposed several abstractions that have led to a quite smaller model and that can be used to model memories of any capacity.

The OEN (output enable not) signal is only used in the output stage circuit and we can suppose that it is always set to 0 since verifying timings for this signal can be carried out by modeling only the output stage.

The CSN (chip select not) signal was also abstracted and in the sequel, we suppose that it is always set to 0.

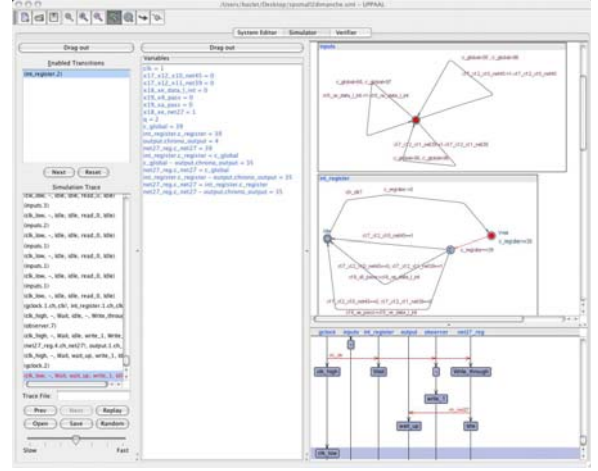


Figure 2: A snapshot of Uppaal

The following idea was considered: since we are only interested in checking timings, we can suppose that the component has correct behaviors when using non aggressive cycle, setup and hold timings. Hence, we can check the memory behaviors using only one data bit and only one address bit. Hence, we set A_1 and D_1 to 0.

3.2 Modeling with Uppaal

Using these ideas, we propose an abstracted modeling which is of much smaller size and it was possible to produce easily an Uppaal model since Uppaal provides debugging facilities and a graphical interface (see Fig. 2). Descriptions of models handled by Uppaal can be found in many papers [BL96, LPY97]. We now give a short description of the automata associated to the different sub-entities of the memory module.

3.2.1 The clock modeling

The clock signal was modeled with a 2-states automaton: *clk_low* and *clk_high*. The rising edge of the clock event is modeled by the message *ch_clk* that the automaton sends periodically.

3.2.2 Inputs modeling

The three input signals (WEN, A, D) were modeled with a 1-state automaton. Those signals behave in the following way: they can change their value at

any time except within intervals of the form $[t - \text{setup}, t + \text{hold}]$ where t is the date of a rising edge of the clock and *setup* and *hold* are the characteristic timings of the considered input signal.

3.2.3 The memory array modeling

It is modeled by a 2-states automaton which has the following behavior: when it receives a rising edge of the clock signal (*ch_clk*), it enters a wait state. When it is able to leave this state, it comes back to the idle state while updating (or not) the values stored in the memory array depending on the current input values.

3.2.4 The output latch modeling

This automaton has three states: *Write_through*, *Read* and *Idle*. When the automaton receives a rising edge of the clock signal (*ch_clk?*), it goes to one of the states named *Write_Through* or *Read*, depending on the operation performed during the cycle and then wait in this state for a duration that depends on the operation. Then it comes back to the *Idle* state while updating, if needed, its output (and if it is the case, it sends a message *ch_output_latch*).

3.2.5 The output stage modeling

This automaton models the combinatorial circuit that manages the output of the output latch. When the automaton detects that the signal of the output latch changes its value (by receiving a message *ch_output_latch*), it sets the output signal Q to the undetermined value (usually denoted by the letter X), it enters a wait state (*Wait_u*). After some delay, it comes back to the *Idle* state while updating the value of signal Q.

4 Model-checking and results

Thanks to Uppaal, we were able to show that our modeling has the nice property of being deadlock free and to ensure that the access time in read mode is 0.77 ns and in write-through mode 0.57 ns. More precisely, we showed that those timings are optimal since the memory model has incorrect behaviors when choosing smaller timings.

Conclusion

In this paper, we describe a modeling in Uppaal of a small SRAM component. We detail how the modeling was abstracted from the memory description (so that model-checking remains tractable) and we succeeded in verifying formally the modeling. More precisely, we were able to ensure that the read and write access timings are optimal. Future works will consist in finding a more precise modeling (e.g. by taking into account the CSN signal that we abstracted) and maybe seeing whether the timings of the memory can be computed from the transistors netlist and few reference gate timings.

References

- [AD90] R. Alur and D. L. Dill. Automata for modeling real-time systems. In *Proc. 17th International Colloquium on Automata, Languages, and Programming (ICALP'90)*, volume 443 of *Lecture Notes in Computer Science*, pages 322–335. Springer, 1990.
- [BDPG02] P. Bricaud, T. Delaye, F. Poirot, and A. Gorun. Do standardized embedded IP transistor view exists for SoC IP integration? In *Proc. 4th Sophia Antipolis forum on MicroElectronics (SAME'2001)*, 2002.
- [BJMY02] M. Bozga, H. Jianmin, O. Maler, and S. Yovine. Verification of asynchronous circuits using timed automata. In *Proc. 1st Workshop on the Theory and Practice of Timed Systems (TPTS'2002)*, volume 65 of *Electronic Notes in Theoretical Computer Science*. Elsevier, 2002.
- [BL96] J. Bengtsson and F. Larsson. UPPAAL, a tool for automatic verification of real-time systems. Master's thesis, Department of Computer Science, Uppsala University (Sweden), 1996. ISSN 0283-0574.
- [LPY97] K.G. Larsen, P. Pettersson, and W. Yi. UPPAAL in a Nutshell. *Journal of Software Tools for Technology Transfer*, 1(1–2):134–152, 1997.

Device Simulation Study of SEU in SRAMs based on Double-Gate MOSFETs

K. Castellani-Coulié^a, D. Munteanu^a, J.L. Autran^{a,b},
V. Ferlet-Cavrois^c, P. Paillet^c, P. Masson^d

^a L2MP / Bât. IRPHE, 49 rue Joliot Curie-BP 146, 13384 Marseille Cedex 13, France

^b also with Institut Universitaire de France

^c CEA/DAM/DIF, 91680 Bruyères-le-Châtel, France

^d L2MP / Polytech' Marseille, Technopôle de Château – Gombert 13451 Marseille Cedex 20, France

Abstract

The sensitivity to heavy ions of static random-access memories (SRAM) based on Double-Gate transistors is analyzed by 2-D numerical simulation. Two different Double-Gate structures are used in the construction of the memory cells, which are then exposed to an ionizing particle strike. The behavior under irradiation of the two different cells is analyzed in terms of threshold LET and critical charge, and compared with results obtained previously in SRAM based on Single-Gate SOI and bulk CMOS devices.

1. Introduction

Because of its excellent control of the body potential and the elimination of floating body effects, Double-Gate (DG) transistors can be particularly interesting structures for radiation hardened applications. Double-Gate MOS transistors incite presently very much interest mainly because it offers extremely reduced short channel effects (due to a reinforced electrostatic coupling between the silicon film and the gate electrode), which allows the use of low-doped channels. Therefore Double-Gate devices present an enhanced mobility and the elimination of doping level fluctuation effects. These features make Double-Gate devices one of the most promising solutions for meeting the roadmap requirements for the end-of-roadmap integration [1].

The aim of the present study is precisely to evaluate the sensitivity to heavy ions of static random-access memories (SRAM) based on Double-Gate transistors. The Single Event Upset (SEU) corresponds to the logical change of state of a memory point following an ionizing particle strike. The effects of ionizing particles on electronic devices result from the electrical perturbation associated with the conductive electron-hole pair column created along the track. Previous papers concerning the analysis of Single Event Upset mechanisms in SRAMs [2-5] are related to bulk or partially-depleted/fully-depleted Single-Gate SOI SRAM cells. Only few papers present studies of the total dose impact on the operation of Double-Gate transistors [6]. At the best of our knowledge, the sensitivity to SEU of SRAM cells based on Double-Gate devices has not yet been addressed in the literature.

In this paper, we will evaluate the sensitivity, in terms of threshold LET and critical charge, of a CMOS SRAM cell based on 0.25 μ m or 0.2 μ m gate length Double-Gate transistors.

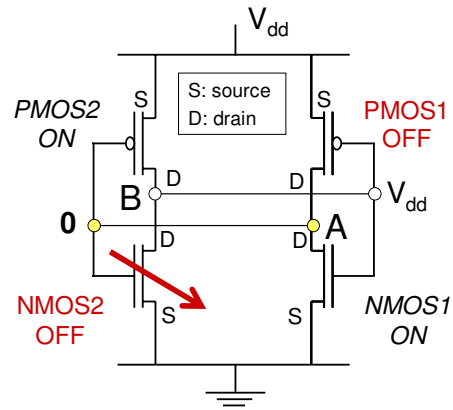


Fig. 1. Initial state of the simulated SRAM cell with the impacted transistor.

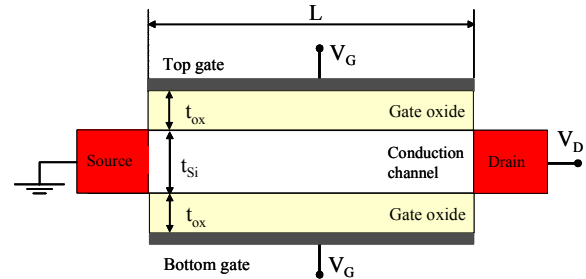


Fig. 2. Schematic of the simulated Double-Gate transistors.

2. Simulation conditions

A. Simulated device

The CMOS SRAM memory cells (Fig. 1) were modeled in the device domain using the ISE (Integrated System Engineering) 2D device simulator [7] in mixed mode. Fig. 1 shows a schematic of the simulated SRAM cell. The Double-Gate transistors used to build the cell have been designed for low-power operation. Note that, in both structures, the two access transistors (corresponding to a 6-transistor cell design) were not simulated (Fig. 1). A schematic cross-section of the Double-Gate device architecture (both NMOS and PMOS transistors) used in this work is shown in Fig. 2.

The first structure is based on a 0.25 μ m gate length Double-Gate transistor having a supply bias of 1.8V. Lightly Doped Drain (LDD) extensions have been

considered with a silicon film thickness (t_{si}) of 20nm and a gate oxide (t_{ox}) of 5.4nm. The body doping is $8.5 \times 10^{17} \text{ cm}^{-3}$ and the off current $10^{-10} \text{ A}/\mu\text{m}$.

The second technology modeled in this work is based on a 0.2- μm gate length Double-Gate transistor. The main difference compared with the first technology is a much thinner gate oxide and a lower Silicon film doping. The devices have a silicon film thickness of 20nm, the gate oxide is 1.5nm and the body doping is $1 \times 10^{16} \text{ cm}^{-3}$. The supply bias is 1V and the off current $10^{-10} \text{ A}/\mu\text{m}$.

B. Simulation models

In all simulations, impact ionization and hydrodynamic models were turned on. Note that impact ionization rate depends on the carrier energy and no longer on the local electric field, as it is the case in Drift-Diffusion simulations. The effect of ion strike has been simulated using the DESSIS HeavyIon module, considering an electron-hole pair column centered on the ion track axis to model the ion strike. The electron-hole pair generation model uses a Gaussian radial carrier distribution with a fixed characteristic radius of 0.02 μm (very thin track) and a Gaussian time distribution centered on 50ps, with a characteristic time of 0.2ps. The LET value is fixed along the track and expressed in $\text{pC}/\mu\text{m}$. In this work, the track was simulated at normal incidence in the middle of the body.

The case of tracks crossing the drain of one off-state transistor of the memory cell is generally considered as the most efficient to trigger an SEU. For this reason, we have studied here the case of an ion striking the off-NMOS transistor (NMOS2 on Fig. 1).

C. Terminology

When crossing the device, the ionizing particle loses its energy. This energy lost by unit of length represents the Linear Energy Transfer (LET), which is expressed in $\text{pC}/\mu\text{m}$ ($1 \text{ pC}/\mu\text{m} \approx 100 \text{ MeV} \cdot \text{cm}^2/\text{mg}$).

The bipolar gain for heavy-ion irradiations, β_{HI} , corresponds to the amplification of the deposited charge. β_{HI} is obtained from the ratio between the total collected charge at the drain electrode and the deposited charge [4]:

$$\beta_{HI} = Q_{\text{collect}} / Q_{\text{dep}} \quad (1)$$

The total collected charge at the drain electrode is given by:

$$Q_{\text{collect}} = \int_0^t I_{\text{drain}} dt \quad (2)$$

where Q_{collect} corresponds to the charge collected at $t=10^{-8} \text{ s}$ whereas Q_{dep} is the charge deposited in the sensitive volume.

At threshold LET, β_{HI} becomes:

$$\beta_{HI} = Q_{\text{crit}} / Q_{\text{dep}} \quad (3)$$

The critical charge Q_{crit} is defined as the minimum electrical charge collected at the drain electrode and able to upset a SRAM cell.

3. Study of the SEU sensitivity

A. Current and voltage considerations in DG SRAM

When interacting with the matter constituting the device, the ionizing particle induces the creation of an electron-hole pairs column. These electron-hole pairs move by diffusion and are collected by the electrodes of the component by drift, inducing a parasitic current in the device. The delay and the magnitude of this current are determining parameters for the SEU occurrence. The higher the LET is, the higher the current magnitude and the delay.

The drain current induced by the particle strike in the impacted n-channel Double-Gate transistor with $L=0.2\mu\text{m}$ is presented in Fig. 3 for various LET values. The current peak is very high for extremely important LET (8 and $4 \text{ pC}/\mu\text{m}$). In this case, the current can be decomposed into two parts. The first one corresponds to the capacitive effect induced by the ion [7-8]. It appears simultaneously with the particle strike (which has a Gaussian time distribution as presented on Fig.3.b on the right scale). The second part of the curve corresponds to the ambipolar diffusion of charges partly generated by the particle and partly by impact ionization. For lower LET (0.5 and $0.2 \text{ pC}/\mu\text{m}$) the particle energy is not high enough to induce a visible capacitive effect.

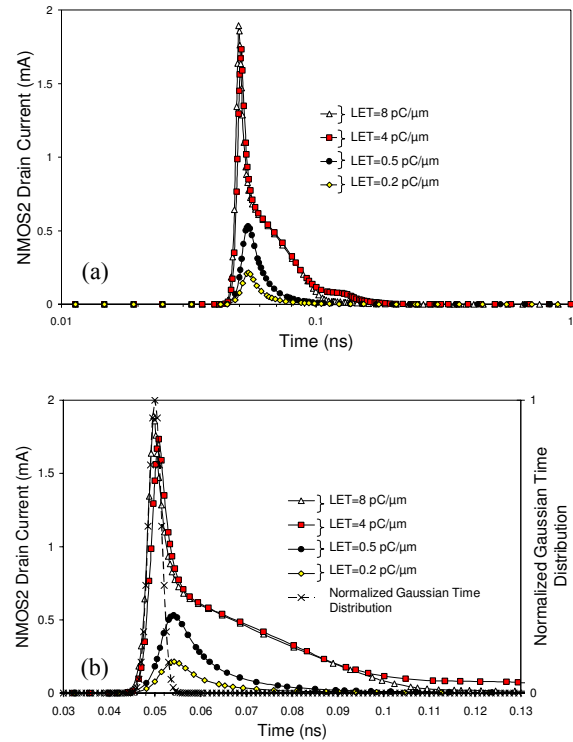


Fig. 3. a) Drain current generated by the ionizing particle at the NMOS2 drain electrode (semi-logarithmic scale). b) Drain current generated by the ionizing particle at the NMOS2 drain electrode: linear scale (left scale) and normalized Gaussian time distribution of the charge generated by the ionizing particle (right scale).

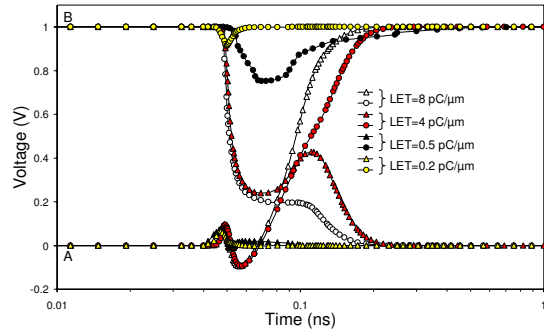


Fig. 4. Time variations of voltage versus LET at both nodes (A and B in Fig. 1) of the SRAM.

Then the current corresponds only to the diffusion process and it does not appear simultaneously with the charge generation. There are fewer charges generated by impact ionization and the second part of the current is not as visible as for a LET equal to 8 or 4 pC/μm.

Fig. 4 presents the variation of voltages at nodes A and B corresponding to the drain currents presented in Fig. 3. When the memory cell flips, its logical state changes. The flipping of the cell is quicker when the LET is higher (as for LET=8 pC/μm and 4 pC/μm compared to LET=0.5 pC/μm and 0.2 pC/μm).

B. Comparison of the sensitivity of the two Double-Gate-based SRAM cells

The main parameters which characterize the sensitivity to heavy ions are summarized in Table I for both simulated SRAM cells. The threshold LET of the 0.2 μm SRAM is about twice higher than that of the 0.25 μm SRAM cell. We can state that the 0.2 μm is more hardened against SEU. The main difference between the two structures is the body doping which can modify the bipolar amplification.

To analyze these differences, we have studied the currents and the deposited charges in the two structures for LET=3 pC/μm (Fig. 5). For the same LET, the 0.2 μm SRAM drain current and the total collected charge are higher than those of the 0.25 μm SRAM. The difference in currents can be partly explained by the surface gate area. It was shown [8] that when the surface area is reduced, the prompt component of the current increases. The reason is that small surface capacitors have a higher bandwidth, and then can respond to shorter signals. Our

results are then in concordance with ref. [9]. However, other parameters have to be taken into account in the analysis, such as the body doping and then the impact ionization, which are different between the two structures. For example, at $t=50$ ps (where the charge generation is maximum) the impact ionization is $4 \times 10^{19} \text{ cm}^{-3}$ for the 0.2 μm SRAM cell and $7 \times 10^{26} \text{ cm}^{-3}$ in the 0.25 μm SRAM. Therefore, the 0.2 μm SRAM has a lower bipolar amplification than the 0.25 μm SRAM (Table I).

The voltage variation at nodes A and B has been also studied for LET=3 pC/μm: as shown in Fig. 6, the 0.25 μm cell flips whereas the 0.2 μm does not.

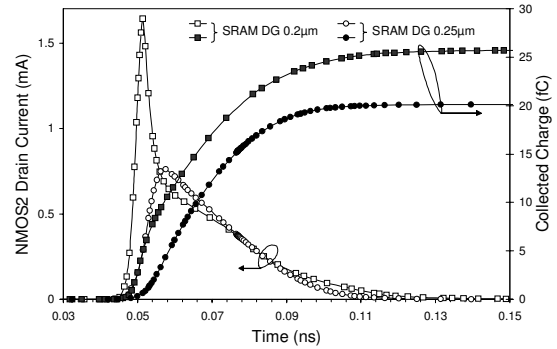


Fig. 5. Current (left scale) and collected charge versus time at the NMOS2 drain electrode (impacted electrode), for LET=3 pC/μm.

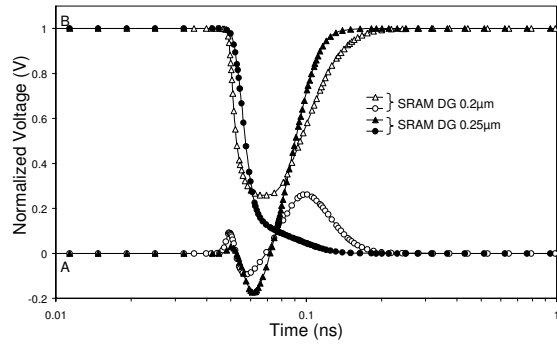


Fig. 6. Time variations of voltage at nodes A and B for LET=3 pC/μm.

Technology	Gate length (μm)	Supply Voltage (V)	t_{si} (nm)	LET _{th} (pC/μm)	Q _{dep} (fC)	Q _{crit} (fC)	β_{HI}
Bulk	0.25	2.5	---	0.02	---	6.64	---
SOI PD	0.25	1.8	---	0.009	---	16.1	16
DG	0.25	1.8	20	2	40	23	0.58
DG	0.2	1	20	4	80	29	0.36

Table I: Sensitivity to heavy ions – Comparison between different technologies.

Fig.7 presents the flipping for both cells corresponding to the threshold LET ($4\text{pC}/\mu\text{m}$ for the $0.2\mu\text{m}$ DG SRAM and $2\text{pC}/\mu\text{m}$ for the $0.25\mu\text{m}$ DG SRAM). At threshold LET the difference between the two technologies is even more visible in terms of current and total collected charge (Fig. 8). As expected, the critical charge is higher in $0.2\mu\text{m}$ SRAM (Table I) than in $0.25\mu\text{m}$ SRAM.

C. Double-Gate SRAM versus other technologies

Results concerning bulk $0.25\mu\text{m}$ SRAM [9-10] and SOI Partially-Depleted (PD) $0.25\mu\text{m}$ SRAM [10] are also reported in Table I, for comparison with the Double-Gate SRAM. The most important point concerns the critical charge and the threshold LET: for the same gate length ($0.25\mu\text{m}$), the critical charge and the threshold LET increase from bulk to Double-Gate SRAM. In Table I, the threshold LET of the PD SOI is lower than that of the bulk SRAM due to floating body effect, which implies that SOI PD SRAM does not exhibit the hardness against SEU that could have been expected from the reduced charge collection volume of SOI technologies. This phenomenon has been already thoroughly studied in literature [10-11]. Table I shows that the bipolar amplification is also considerably reduced from PD SOI to DG SRAM. These results are very encouraging for the DG SRAM cell, which seems to be significantly less sensitive than PD SOI or bulk technologies. DG transistors, and consequently DG-based SRAM cells, exhibit the perfect configuration for SEU hardness: an extremely reduced charge collection volume ($t_{\text{si}}=20\text{nm}$) and a very low bipolar amplification. Finally, we note that the threshold LET values found here for DG-based SRAM are very high (higher than the required value for hardened components which is $100\text{MeV.cm}^2/\text{mg}$), but there are only qualitative results, since we have considered here long channel lengths and very low off-state current. For obtaining more precise values of the threshold LET, structures calibrated in real technologies have to be simulated. The threshold LET is expected to be lower in real scaled devices having short channel lengths and higher off-state current.

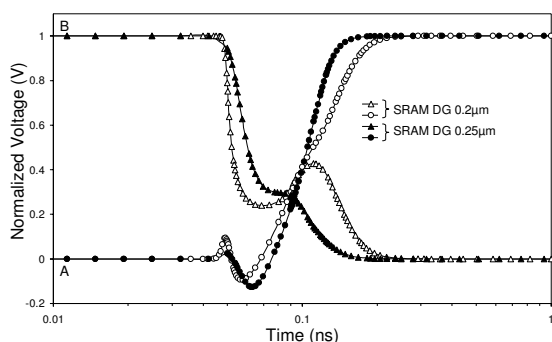


Fig. 7. Time variation of voltage at both nodes of the SRAM at threshold LET ($4\text{pC}/\mu\text{m}$ for the $0.2\mu\text{m}$ DG SRAM and $2\text{pC}/\mu\text{m}$ for the $0.25\mu\text{m}$ DG SRAM).

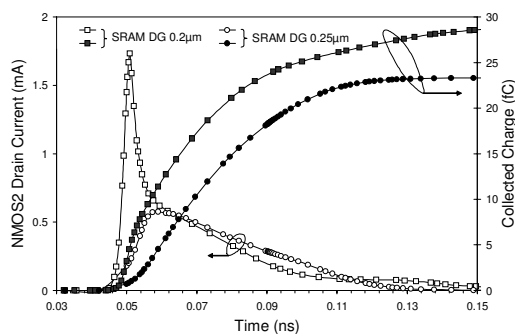


Fig. 8. Current and collected charge versus time at the NMOS2 drain electrode (impacted electrode) at threshold LET.

4. Conclusion

In this paper, we have analyzed in terms of threshold LET and critical charge, the sensitivity of two different SRAM cells based on Double-Gate devices. The $0.2\mu\text{m}$ DG SRAM cell appears to be more hardened against SEU than the $0.25\mu\text{m}$ DG one. This can be mainly explained by the body doping of these cells which are really different. We show also that compared with PD SOI and bulk technologies, the DG SRAM cell is sensibly less sensitive to heavy ion irradiation and will be an excellent candidate for hardened applications.

References

- [1] For a recent review, see D. Hisamoto, Short course IEDM 2003 and references therein.
- [2] P. Dodd et al., "SEU-Sensitive Volumes in Bulk and SOI SRAMs From First-Principles Calculations and Experiments", *IEEE Trans. Nucl. Sci.*, vol.48, n°6, pp. 1393-1903, 2001.
- [3] K. Castellani-Coulié et al. "Various SEU conditions in SRAM studied by 3-D device simulation", *IEEE Trans. Nucl. Sci.*, Vol. 48, n°6, pp. 1931-1936, 2001.
- [4] V. Ferlet-Cavrois et al. "Insights on the Transient Response of Fully and Partially Depleted SOI Technologies Under Heavy-Ion and Dose-Rate Irradiations", *IEEE Trans. Nucl. Sci.*, vol.49, no.6, p. 2948, 2002.
- [5] K. Castellani-Coulié et al. "Study of a SOI SRAM Sensitivity to SEU by 3-D Device Simulation", *IEEE Trans. Nucl. Sci.*, vol.51, no 5, pp 2799-2804, 2004.
- [6] J-P. Colinge et al, "Effects of total-dose irradiation on gate-all-around (GAA) devices", *IEEE Trans. Nucl. Sci.*, vol.40, no.2, p. 78, 1993.
- [7] G. Hubert et al., "Study of basic mechanisms induced by an ionizing particle on simple structures," *IEEE Trans. Nucl. Sci.*, vol. 47, pp. 519-526, 1999.
- [8] V. Ferlet-Cavrois et al. "Charge Collection by Capacitive Influence Through Isolation Oxides", *IEEE Trans. Nucl. Sci.*, vol.50, no.6, p. 2208, December 2003.
- [9] J.M. Palau et al., "Device Simulation Study of the SEU Sensitivity of SRAMs to Internal Ion Tracks Generated by Nuclear Reactions", *IEEE Trans. on Nucl. Sci.*, vol. 48, No 2, pp 225-231, 2001.
- [10] G. Gasiot et al., "Comparison of the Sensitivity to Heavy Ions of $0.25\mu\text{m}$ Bulk and SOI technologies", *IEEE Trans. Nucl. Sci.*, vol. 49, No 3, pp 1450-1455, 2002.
- [11] Y.Tosaka, K.Suzuki, T.Sugii, "Alpha-particle induced soft errors in submicron SOI SRAM", in *Proc. 1995 Symp. VLSI Technology Dig. Tech. Papers*, Kyoto, Japan, June 1995, pp.39-41.

A System-Level Approach for Memory Robustness

Riccardo Mariani^a, Gabriele Boschi, Anna Ricca

Yogitech SpA, via Lenin 132/p, 56017 San Martino a Ulmiano, Pisa, Italy

^a Contacting author: riccardo.mariani@yogitech.com

Abstract

New ultra-deep submicron technologies are bringing not only new advantages such extraordinary transistor densities or unforeseen performances, but also new uncertainties such modelling complexity, coupling effects, leakage contribution and increased sensitivity to internal and external disturbs. Soft Errors are not anymore a pure radiation effect important for aerospace application only: a new frontier is opening concerning reliability of modern System-on-Chip, and as in the past memory systems should be the first to be worried by such escalation. In this paper, classical gate-level protection techniques for soft-error protection are enriched with a system-aware approach, where the memory system is analyzed based on its role in the application. A protection block (called F-MEM) is presented, together with the results of its application to a proof-of-concept memory system. This work has been developed in the framework of MEDEA+ T126 project called BLUEBERRIES.

1. Introduction

As the technological evolution of the semiconductor industry marches on to the deep submicron territory and beyond, extreme system complexity, shrinking feature size, the accompanying reduction in supply voltage and increase in operating speed results in an increased susceptibility to uncertainties and faults. Modern system on a chip experiment failures caused not only by radiation induced particles but also by cross talk, model and system unforeseen behaviors. If we define fault robustness as the ability of a system to continue to function reliably despite the existence of faults, we can conclude that modern systems are less and less robust.

Nowadays, deep-sub micron technologies and complex systems are used in high volume applications where robustness is a key factor for system safety: for instance, in automotive industry fault tolerance is a must in critical systems involving airbags, brakes, engine control and future x-by-wire cars. But it is not only a matter of safety: also reliability and availability become to be relevant for high-demanding and very high-volume applications such modern communication systems. Could you accept an abrupt reset of your mobile phone each hour while trying to transmit Mbits of data?

As robustness requirement increases in such applications, technology hardening or pure circuit redundancy cannot be anymore a solution by themselves: costs, overhead and performance degradation must be efficiently tuned based on the system and on the application. This paper presents an innovative approach that starts from system-level consideration, and then it applies a mix of protection techniques to modern memory system.

2. Background

It is generally accepted that “failure” is the event that occurs when the delivered service of a system deviates from the specified service, or “mission”. An “error” is the part of the system state, which is liable to lead to failure. “Fault” is the phenomenological cause of the error. System dependability is defined as the quality of the delivered service such that reliance can justifiably be placed on this service [1, 2].

Based on previous definitions, we can define “soft” error the end effect of a transient fault causing a failure of a circuit where only data are destroyed (e.g. a bit-flip): we distinguish it from “hard” error where also the internal structure of the semiconductor material is damaged, and the fault is therefore permanent or at least intermittent. Transient faults can be caused by external (such alpha-particles hitting the silicon) or internal (timing faults due to cross talk, ground bounce) disruptive events, and generally manifest itself as a transient pulse on the output of a logic cell.

There is a great amount of data and of philosophical discussions concerning effects of soft-errors in modern technologies [4, 7]. Concerning embedded RAMs, some Texas Instrument data report that elimination of BPSG in new 10Mb generations contribute to sharply decrease SER per bit, but the big growth of embedded RAMs brings to a global SER increase per chip [15]. Therefore, FIT/bit is decreasing, but FIT/chip is increasing. The same appears in other references [15-18].

Roughly speaking, memory protection techniques can be divided into three main groups, depending where the technique itself is applied: at manufacturing level, at software-level or at hardware-level. Manufacturing techniques include for example Silicon On Insulator or wafer thinning. The advantages of these solutions are the low impact on performance and die area, while on the other hand they are expensive since they impact the technology process. Software-level techniques include specific coding approaches to make the system robust to soft errors. For example the ABFT [5] are checksum schemes to validate results by a numerical subroutine operating in a fault-prone environment. These techniques often require heavy modifications of the SW that can be unacceptable in some cases. Several well-known techniques exist at hardware-level. The simplest is the parity, which is able to detect errors only. Error correction is typically achieved by using Error Correction Codes (ECC) [3]. Many solutions with different ECCs and calculation mechanisms exist: ECC can be applied on different word sizes, using different codes (Hamming, APC, Reed Salomon etc. [8]) and different implementations that differ in terms of area, speed and performance overheads.

ECC is in fact one of the most effective techniques in terms of on-line fault-tolerance: however, its use is still confined to some applications, since overheads are important and so system architects are often reluctant to widely use such approach. Some advanced techniques exist, such as the “ECC for small-word memories and/or for maskable operations” [9]. They consist in code sharing between several words. The big advantage is the overhead reduction area. The performance overhead that is introduced cannot be negligible with standard memories, but can be reduced if bit-writable memories are used. In another interesting technique, the “ECC without speed penalty” [9], a speed optimization of the Error Detection mechanism and a fault-decontamination technique are both used. While the system/processor is stalled, the error is corrected and the fault-free data is provided to the core, modifying the flip flops containing the faulty data. Decontamination implies a modification of the core that in some cases can be unacceptable.

Concerning metrics to assess the robustness level, in general reliability can be measured in different ways. Mean Time To Failure (MTTF) is one of this. FIT (Failure in Time) is also commonly used, and it is inversely related to MTTF. Both MTTF and FIT are related to the definition of what a “failure” is, that arises once again the debate on the mission and the “acceptable” degradation of the mission of a given system. Mean Time to Repair (MTTR) is particularly interesting when fault decontamination techniques are used. Vulnerability factors are also very useful to complete the analysis [12, 13]: for instance, the Architectural Vulnerability Factor (AVF) is the probability that an error will occur given a fault in a cell, and it can be associated to a given circuit architecture, flip-flop schemes, etc... In safety-critical application a more system-level oriented metric is used, such for instance the IEC61508 norms [20].

3. A system-level approach

YOGITECH’s vision on these aspects is that a system-level approach is mandatory to achieve the best compromise between reliability and costs; moreover, it is only a *combined* utilization of all the techniques that can lead to a dependable system. The attention should be moved to failures instead of faults: the mission of a given system should be analyzed (e.g. with failure mode and effect analysis, FMEA) and errors should be classified depending on their failure leading. This general concept should be traduced in a methodology including SW reliability analysis, use of fault-injection at different abstraction-level, selection of a mix of HW protecting techniques, etc... This is what we call the “Fault-Robust” approach, where “robustness” is defined as the ability of a system to continue to function despite the existence of faults in its component subsystems or parts, even if system performance may be diminished or otherwise altered – but always in a *reliable* way - until the faults are corrected. The word “reliable” can be substituted by “safe” or “available” based on the application requirements (e.g. automotive, telecom, etc..), but it is always related to the system’s mission.

This approach is applicable to memory systems as well. This paper gives an example of how system-level

concepts could be applied together with a mix of protection techniques (parity, ECC, ECC with scrubbing) in order to provide the best compromise between reliability and related overheads.

Metrics are needed to specify the level of robustness at system-level. In [14], we defined a Mission Vulnerability Factor (MVF), where a system or a part of a system or of an application is measured in terms of the effects of an error on the given mission. In other words, we relate MVF with the probability that an error generates a failure. In case it is applied to a sub-system, the MVF is weighted (WMVF, weighted MVF) according to the criticality of the sub-system mission and to the overall mission. For instance, an application can have procedures with a very high WMVF (like system routines, schedulers, variables affecting sensors such as ABS) and others with a low WMVF, like data representing digital images, etc...The WMVF can be also correlated with IEC61508.

Based on this definition, the approach presented in this paper starts from a high-level analysis of the application using a given memory system. Then, vulnerability factors are computed, FMEA is carried out and the data to be stored in the memory system are ranked based on the effects of a fault toward the overall mission of the system, i.e. based on WMVF. Then, based on WMVF, the right mixture of protection techniques is applied. The following paragraph gives an overview of the outcome of such decision process.

4. The F-MEM

Let’s consider as the system to be protected a microcontroller unit based on an ARM7TDMI, with the AMBA AHB as the backbone bus to connect the CPU to internal embedded memory (a RAM). The embedded RAM is connected to the AHB through an AHB slave called RAM controller. The proposed protection system is called F-MEM and it is schematically represented in Figure 1. It is composed by three main blocks: the FMEM core (which controls all the FMEM features), the CPU IF (which provides a way for the processor to control the FMEM) and the RAM controller wrapper (to give a DMA to the FMEM core and for other features) including the ECC coder and decoder, to detect/correct the errors.

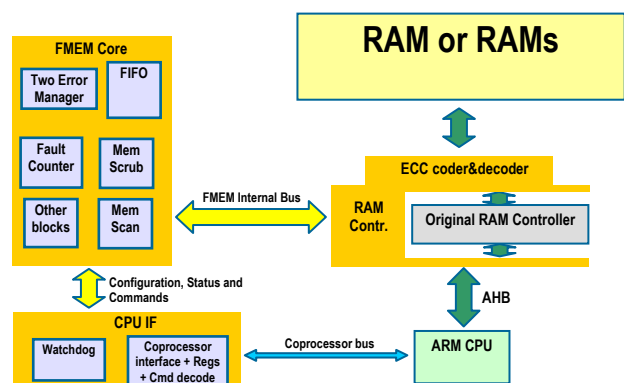


Figure 1: the protected system with F-MEM

F-MEM block can apply parity or SEC-DED or no protections at all to the memory partitions, in order to

optimize the various overheads to the target application depending on the WMVF defined before: an example of such partitioning is shown in the next paragraph. A “scrubbing” technique is also used to periodically inspect the memory in order to find and correct errors. The parts of memory to be scrubbed are configurable by SW, and the scanning is typically done during RAM idle periods or on SW/HW demand.

As detailed in [14], the aim of the scrubbing is to discover single errors as soon as possible:

- to avoid double errors ;
- to reduce the performance and speed overheads (errors are detected and corrected before the processor attempts an access) enabling the removal of the ECC decoder from the read path

For what concerns error managing and fault logging, unrecoverable errors on same word are signalled to the core/system in various ways, providing the information on the address/data of the location and enable system-level fault-decontamination techniques. F-MEM also includes fault logging: single and double faults counters are available to the core. This is useful for statistics and for dynamic management of the protection.

The SW control over the whole protection mechanism is another important feature of the F-MEM: using basic coprocessor instructions, SW has full control over protection mechanisms, scrubbing, configuration, logging etc. For example, SW can start a scrubbing when it is waiting for an interrupt, or no processes are active at a certain moment. Interrupts or other events will suspend the scrubbing, thus obtaining a negligible performance overhead.

5. Results

A first configuration of the F-MEM consists in the application of SEC-DED codes added to the whole RAM. In the architecture of figure 1, we considered a 6-bit ECC added to each 16-bit word, with a memory overhead of 37.5%. Scrubbing could be present or not, depending on the desired target MTTF. The following graph represents how MTTF is changing depending on the memory size, with a fixed FIT = 0.01 errors/bit per 10^9 hours.

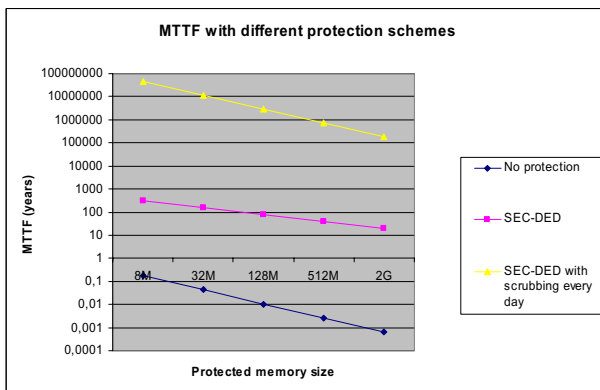


Figure 2 : MTTF vs memory size for a given FIT

The presence of the ECC coder-decoder in the memory data path affects the access timing: 3-4 extra ns are needed with a 0.18um technology, where the worst is during read operations, since both ECC decoding and

code comparing should be performed. Some schemes exist to reduce the speed overhead, mainly taking the ECC decoder out of the memory data path and performing the ECC decoding and code check in parallel. If a data is not containing errors, the processor can continue its operation as in a system without ECC protection, while a “fault decontamination” procedure should start in case of errors (e.g. stopping the processor, flushing the pipeline, repeating the operation, etc...).

The second configuration of the F-MEM (currently under development) is fully applying the system-level approach previously described, as also includes the possibility to remove the ECC decoder from the data path to reduce the speed overhead. For a memory of 16Mb, the table 1 reports the resulting partition after the FMEA analysis, together with the memory sizes, the estimated WMVF, the selected protection technique, and the correspondent MTTF for each sub section.

Partition content	Estimated WMVF	Size	Protection technique	Calculated MTTF
Not critical data	-	6Mb	None	- (not important)
Sub-Critical data	2%	1.5Mb	Parity	45 years
Medium-Critical data (e.g. stack)	10%	512Kb	Parity	27.2 years
Critical data, but used in not timing critical functions	40%	≈8Mb	SECDED in the data path (16bit data + 6bit code)	792 years
Critical data, used in timing critical functions	50%	64Kb	SECDED outside the data path, scrubbing	108.85 years

Table 1: a possible FMEM configuration based on WMVF

It is worth to note that last partitions have an estimated WMVF of 40% and 50%: this is a worst case estimation since, as reported in [13] for the AVF, the most critical parts of a microprocessor (such instruction queue and executions units) have AVF upper-bounds between 2% and 40%.

Summarizing the results, the final MTTF of the 16Mb memory system is around 14 years. Without any protection and with the same WMVF settings, the MTTF is 0.17 years, i.e. 62 days. Considering the whole 16Mb protected with SECDED, the MTTF is 224 years. Concerning the memory area penalty, as already said with a 6-bit SECDED on a 16-bit wide word on the whole memory the overhead is 37.5%. With the memory system partitioned as in the previous table, the total area overhead decreases to 19.5%. The achieved MTTF and overhead are fully acceptable for the application.

Concerning overheads, preliminary results on a subset of the previous described techniques shows that the memory controller area overhead (including Core, CPU IF and RAM Contr. Extension) is between 1.5 and 3K equivalent gates. Performance overhead (for scrubbing and scanning) is negligible. Some additional techniques should be applied if the application program shall also be executed from the protected memory: in

such case the performance degradation is less than 5%. Speed overhead (data for 0.18um low leakage technology) shown negligible overheads for the SECCDED outside the data path.

The following table summarizes the tradeoffs of the possible F-MEM configurations. For example, it is possible to change the coding scheme to decrease the area overhead (from 16 to 32-bit ECC coders). In this case, the speed overhead will increase due to the relatively slow ECC calculation, and also performance overhead will increase due to the halfword and byte write managing. Leveraging on the ECC parallel processing, it is possible to introduce a parametric write buffer which decreases the speed penalty. The drawback is a small area increase. Similar considerations can be done for the other parameters.

The F-MEM approach can be used with some minor differences both for embedded RAMs and DRAMs.

Configuration parameter	Area overhead increases Speed/Performance overhead increases		
	32-bit and more	16-bit	8-bit
ECC coding scheme	32-bit and more	16-bit	8-bit
ECC Parallel processing	NO	Partial	YES
Wrapped Ram Controller	YES	YES	NO
Pipelined FMEM	NO	NO	YES
Shadow variables	NO	NO	YES
SW information to FMEM	Overheads depend on several factors.		
SW error management	Overheads depend on several factors.		
Protection and memory partitioning	Overheads depend on several factors.		

Table 2: F-MEM tradeoffs

6. Conclusions: the validation process

We have seen as the main aspects to be considered for the protection of a memory system are the *robustness level* (including the choice between fault detection & fault correction, and the use of techniques for fault forecasting and fault decontamination), *overheads* (including the area penalties, the speed penalties, the performance penalties and power penalties) and *flexibility* (including the need to manage many types of memories and memory controllers, with different characteristics). The proposed approach takes into account all these aspects starting from system considerations, such as a correct estimation and validation of the WMVF.

We strongly believe this system-aware robustness process would be the mandatory approach for the very complex devices of the near future. Of course, a very strict validation process is needed in order to be confident with such high-level approach: functional validation with *fault injection* is therefore mandatory to complete the specification and design stage of a memory system using such high-level techniques. In particular, fault injection must be done both at a lower circuit level

(e.g. with the approaches described in [10] or in [11]) and at a higher level including errors and failures injection both in HW and in SW running in the system. YOGITECH is using this combined fault injection approach to validate these techniques [19].

7. Acknowledgments

We would like to thank Leonardo Boschi for his support during F-MEM implementation and all the BLUEBERRIES collaboration for the very useful discussions carried-on during the project.

References

- [1] J.C Laprie, "Dependable Computing and Fault Tolerance Concepts and Terminology", IEEE Computer, 1985
- [2] H. Tahne, "Safe and Reliable Computer Control: Systems Concepts and Methods", Mech. Lab, Univ. Stock, 1996
- [3] F.J. MacWilliams and N.J.A. Sloane, The Theory of Error-Correcting Codes, North-Holland: New York, NY, 1977
- [4] Tezzaron semiconductor, "Soft Errors in Electronic Memory – A White Paper", 2004
- [5] M. Turmon and R. Granat, "Algorithm-Based Fault Tolerance for REE Numerical Applications: Basis and Implementations", 1998
- [6] M. Spica, "Do we need anything more than single bit error correction (ECC)?", 2004
- [7] R. Mariani, "Soft errors on digital components" chapter in "Fault Injection Techniques and Tools for VLSI Reliability Evaluation", Kluwer Academic Publisher.
- [8] K. Chakraborty, P. Mazumder, "Fault-Tolerance and Reliability Techniques for High-Density Random-Access Memories", 2002
- [9] M. Nicolaidis, "Design of robust circuits with respect to SEU/SET and fault simulation" Proceedings of European Workshop on Radiation Hardened Electronics, 2004
- [10] R. Mariani et al, "A SEU Injection Tool to evaluate DSP-based Architecture for Space Applications", ICCD 2000
- [11] "ROBAN", Datasheet v2.2, IROC Technologies, 2002
- [12] S. Mukherjee et al. "Cache scrubbing in Microprocessors: Mith or Necessaity?", 2004
- [13] S. Mukherjee et al. "A Systematic Methodology to Compute the Architectural Vulnerability Factors for a High-Performance Microprocessor", 2003
- [14] R. Mariani, G. Boschi, "Scrubbing and partitioning for Protection of Memory System", submitted to IOLTS 2005.
- [15] R. Baumann, "Silicon Amnesia: Radiation Induced Soft Errors", RADECS 2001 Short Course, September 2001
- [16] J.M. Palau, G. Hubert, K. Coulie, B. Sagnes, M.-C. Calvet, S. Fournier "Device Simulation Study of the SEU Sensitivity of SRAMs to Internal Ion Tracks Generated by Nuclear Reactions" IEEE Trans. Nucl. Sci., Vol.48, pp. 225-230, Dec.2001.
- [17] L. Z. Scheick et al. "SEU Evaluation of SRAM Memories for Space Applications", IEEE Trans. Nucl. Sci., 47 (6), 2001
- [18] F. Wrobel et al. "Simulation of Nucleon-Induced Nuclear Reactions in a Simplified SRAM Structure: Scaling Effects on SEU and MBU Cross Sections", IEEE Trans. On Nuclear Science, vol. 48, no.6, Dec 2001
- [19] A. Benso, R. Mariani, P. Prinetto, L. Tagliaferri, "e-Fault: a Fault injection Flow Built on Functional Verification Software", submitted to ITC 2005.
- [20] CEI International Standard IEC 61508, 1998-2000

SESSION C

MRAM 1/2

Chairmen: **Didier SAVE**, GEMPLUS, France (Gémenos)
 Jean-Pierre NOZIERES, SPINTEC, France (Grenoble)

Recent Advances in MRAM Technology

W.J. Gallagher

IBM Watson Research Center, Yorktown Heights, NY 10598, USA,

Abstract

MRAM provides a good combination of all attractive memory attributes – high performance, high density, low power, non-volatility, and high write endurance. In this paper, we review our work on a 16Mbit MRAM demonstrator chip. The chip, design on a 180nm CMOS technology base, utilized a $1.42\ \mu\text{m}^2$ cell and achieved a read/write access time performance of 30ns. We describe aspects of the technology development that led to this chip, indicating both some of the ultimate potential for MRAM technology and some of the power-performance compromises involved in a practical chip demonstration. In addition, we also discuss recent advances in MRAM technology. These include the use of an advanced tunnel barrier material, MgO, which promises much larger signals for MRAM read operations.

Toggle-MRAM Performance and Reliability

S. Tehrani

Freescale Semiconductor

Abstract

Magnetoresistive Random Access Memory (MRAM) is an emerging technology that extends beyond the space traditionally covered by existing nonvolatile memories. It employs a tunneling magnetoresistive device integrated with standard silicon-based microelectronics, resulting in a unique combination of nonvolatility, high-speed operation and unlimited read/write endurance. The development of MRAM technology has faced many challenges impacting the manufacturability of high-density devices. One of the most difficult hurdles has been the control of the bit micromagnetic behavior to achieve adequate switching and disturb margins for reliable programming of large arrays. The "conventional" astroid bit select and writing approach has not been able to achieve the necessary operating margin. We have developed a novel bit structure and "toggle" switching mode that solves this bit select problem, greatly improving the switching margin, reliability and device scalability. The new bit uses a

balanced synthetic-antiferromagnet trilayer structure and a novel write pulse sequence to provide robust switching performance with immunity from $\frac{1}{2}$ -select disturbs. Using this approach, we have developed a 4Mb MRAM based on a 1 transistor-1 magnetic tunnel junction (1T1MTJ), $1.55\mu\text{m}^2$ memory cell built in a $0.18\mu\text{m}$ CMOS dual gate process. The properties of our toggle-switching MRAM bit, and large arrays of such bits, is discussed and compared to the conventional switching approach. The performance of the 4Mb Toggle-MRAM circuit is described and the requirements for magnetic and electrical properties of the magnetic materials are discussed. Also, the intrinsic reliability of the magnetoresistive tunnel junction (MTJ) and the metal interconnect system of MRAM will be presented. Time dependent dielectric breakdown (TDDB) and resistance drift were the two main failure mechanisms identified for intrinsic memory bit reliability. Results indicated that a lifetime over 10 years is achievable under the operating conditions.

Low power MRAM design based on new switching mechanism

Hiroaki.Yoda^a, Tadashi Kai^a, Masahiko Nakayama^a, Sumio Ikegawa^a, Yoshiaki Asao^a,
Hiromitsu Hada^b, Nobuo Ishiwata^b, and Shuichi Tahara^b

^a Corporate Research & Development Centre, Toshiba Corporation, 1 Komukai Toshiba-cho, Saiwai-ku, Kawasaki 212-8582 Japan, hk.yoda@toshiba.co.jp

^b System Devices Laboratories, NEC Corporation, 1120 Shimokuzawa, Sagami-hara, Kanagawa 229-1198, Japan

Abstract

MRAM (Magnetoresistive Random Access Memory) is believed to have potential to become an ultimate non-volatile random access memory. Lots of efforts have been made in order to solve its issues and to prove the potential [1], [2],[3]. Among them, Savtchenko switching improved the critical issue for the production, i.e. small writing margin. However the programming current is thought not to be small but may increase with reducing MTJ(Magnetic Tunnelling Junction) width. Therefore, the scalability is not clear, especially in the range of 100nm width MTJ. The new writing mechanism was proposed which expanded the writing margin[4]. MRAM array is designed based on the new writing mechanism. The free layer takes S figure-like spin moment configuration throughout the switching process while takes double C figure-like spin moment configuration in half selection by bit lines. This results in reasonable programming current as well as reasonable write margin. To reduce the programming current further, yoke wires are closely placed to MTJs, which has large field generation efficiency.

Finally, the scalability of MRAM is estimated. We believe MRAM has the scalability beyond 100nm MTJ width.

References

- [1] W. J. Gallagher, *et al.*, *J. Appl. Phys.*, 81 (8), pp.3741-3746.
- [2] M. Durlam, *et al.*, *Symposium on VLSI Circuit Digest of Technical Papers*, pp.158-161, 2002
- [3] K.Ounadjela, *et al.*, *9th Joint MMM/Intermag Conference*, GE 01,2004
- [4] T. Kai, *et al.*, *IEDM Technical Digest*, pp583-586,2004

Development strategy for writing margin of high density MRAM

J. E. Lee, S. C. Oh, H.-J. Kim, Y. K. Ha, J. S. Bae, K. T. Nam, S. O. Park, H. S. Kim,
U-In Chung and J. T. Moon

Process Development Team, Samsung Electronics Co., Ltd.
San #24, Nongseo-Ri, Giheung-Eup, Yongin-City, Gyeonggi-Do, Korea, 449-711
Tel: 82-31-209-8833, E-mail: jangeun2002.lee@samsung.com

Magnetic random access memory (MRAM) has been developed in the many major semiconductor companies for several years. However, the mass product of high density MRAM has been delayed due to the issues of writing operation. Writing margin issues are associated with switching field distribution caused by interlayer couplings, cell shape variation and etch damage and disturbances caused by operations to nearby cells. Three switching methods such as conventional switching, C-S switching and toggle switching are developed and analyzed to apply to high density MRAM with reliable operating margin.

We demonstrated that roughness control of magnetic tunnel junction (MTJ) film, choice of free layer materials with small M_s and pinned layer with small stray field can effectively suppress switching deviation of the conventional type cells. But, these cells still presented narrow writing margin in the asteroid curve due to the incoherent switching and switching field distribution. We suggest a new type of goggle cell shape and discuss its switching mechanism in comparison with conventional shapes such as rectangular and elliptical shapes. For the goggle shapes, the dependence of the switching characteristics on the saturation magnetization, the free layer thickness, and the aspect ratio was studied. We found that the L-like asteroid curve with an enlarged

writing margin could be obtained from a goggle shape, especially when the aspect ratio is smaller than 2.0. This seems to be closely related with the kink formation. We also found that the writing margin could be improved further as the saturation magnetization and thickness of the free layer were optimized. It is also revealed that a goggle shaped MTJ with a SAF free layer has a remarkably large writing margin.

Toggle switching mode MRAM is tested and characterized with respect to the free and pinned layer material and thickness. The crystal anisotropy field (H_k) and the exchange coupling field (H_{ex}) of the free layer are the most governing factors in toggle MRAM switching characteristics. With magnetization curves, we were able to find the optimum thickness combinations of free layer and spacer materials. For CoFeB where H_k is about 20 ~ 30 Oe, one needs to have a reasonably high H_{ex} to enlarge the switching margin. In case of NiFe, H_k is usually much smaller than H_{ex} , so that further decrease of H_{ex} between two magnetic layers are needed. The pinned layer roughness and thickness are another factors to be optimized to reduce the toggle switching field. Cell aspect ratio 3.0 is needed to minimize the switching distribution and enhance the switching stability.

STATE OF THE ART IN MAGNETIC RANDOM ACCESS MEMORY

K. Ounadjela

Cypress Semiconductor, 3901 N. First Street, San Jose, CA 95134

Sophisticated storage technologies based on magnetoelectronics are already at an advanced stage. In 2005, Magnetic Random Access Memories (MRAM), a new type of computer memory will go on the market. MRAMs use magnetic hysteresis to store data and magnetoresistance to read data. These devices are made of confined magnetic structures integrated on a circuit chip and function like a static semiconductor RAM chip

with the added feature that the data are retained when the power is turned off. Unlike present forms of nonvolatile memory, they would have switching rates and rewritability challenging those of conventional RAM. I will be presenting the new development in this area as well as the technological challenges which have been overcome to make this memory manufacturable.

Thermally Assisted MRAM for low power applications

O. Redon^{a,b}, M. Kerekes^b, R. Sousa^b, L. Prejbeanu^b, H. Sibuet^a, F. Ponthennier^a,
A. Persico^a, J.P. Nozières^a

^a Cea/Léti, CEA Grenoble, 17 Rue des Martyrs, 38054 Grenoble Cedex, France

^b SPINTEC, CEA/CNRS, URA 2512, CEA Grenoble, 17 Rue des Martyrs, 38054 Grenoble Cedex, France

Conventional Magnetic Random Access Memory (MRAM) relies on the superposition of two orthogonal magnetic fields to create the selectivity. The use of the simple Stoner-Wolffarth astroid as a write selection scheme was first proposed, but MRAM developers had to face poor write margin as well as endurance and reliability issues due to the thermal activation of half-selected bits [1]. The Savtchenko approach (“Toggle switch”) was then proposed by Freescale to solve these major issues for industrialisation [2]. To date, many MRAM developers have adopted this scheme despite the penalty of a much higher switching energy. In addition, none of these approaches allows for an appropriate scaling along the ITRS roadmap scheme, mostly due to the increase of the switching current upon scaling and the poor thermal stability for small feature sizes. The Toggle approach may therefore be suitable for the two next technological nodes but alternative write schemes are necessary beyond the 65nm node.

The thermally assisted switching (TAS) combines a local heating of the selected bit and a single low amplitude magnetic field [3]. The magnetic tunnel junction (MTJ) stack used in the cell was slightly modified to be adapted for the TAS operation, wherein the storage layer now comprises a ferromagnetic layer exchange biased by a low blocking temperature (T_B) antiferromagnetic material. The antiferromagnetic materials of the storage and reference layers are chosen to have the largest difference in their T_B . During the write operation, a voltage is forced across the junction. The hot electrons injected through the tunnel barrier relax by emission of phonons and magnons which in turn heats up the storage layer. When the storage layer temperature exceeds T_B , the ferromagnetic layer is freed and can be reversed under the application of a small unique magnetic field. The magnetic field is sustained beyond the heating voltage pulse to cool the MTJ under magnetic field and insure a correct pinning of the storage layer.

This alternative write approach offers superior advantages over the conventional schemes. First, the selectivity is perfect as only the selected heated junctions can be written regardless of the external field amplitude (non heated cells are field insensitive). Note that as the selectivity is temperature dependent, the T_B of the storage layer has to be adjusted with respect to the chip operation temperature range. Second, the exchange bias storage layer has an increased intrinsic anisotropy due to exchange bias which allows i) dots with a low aspect ratio down to 1:1 which minimises the switching field and makes it scaling-independent ii) thermally stable cells even for small feature sizes and without the need to increase the thickness of the storage layer. Finally, the

exchange biasing of the storage layer guarantees a high protection against magnetic erasure: This memory can only be erased by the superposition of a high temperature and a magnetic field.

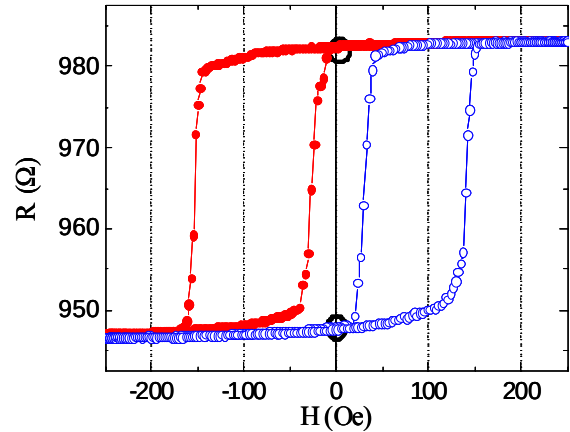


Fig. 1: Magnetic loops of the exchange biased storage layer before (red) and after write procedure (blue). The MTJ was written using 20ns voltage pulse of 2V and an applied magnetic field of 200Oe.

Figure 1 shows the magnetic loops of the storage layer before and after write operation. The loops are clearly shifted from zero field due to the exchange bias on the storage layer. In zero magnetic field, we have two distinct states corresponding to the digital information “0” or “1”. In this graph, the write operation was performed using 20ns heating voltage pulse of 2V and a small external magnetic field of 20Oe. The minimisation of the switching field implies the use of a low AR dot but also an optimisation of the crystallographic texture of the MTJ stack. Indeed a compromise should be found to achieve simultaneously a high ratio of the exchange field over the coercivity (good (111) texture) and minimum of H_c at T_B (weak (111) texture).

The second term to minimise to achieve low power operation is the heating power density. Figure 2 shows the measured heating power versus the dot size for 100ns heating pulse width. Although the absolute power is continuously decreasing with the dot size, the required power density is increasing. This effect is related to the power dissipated into the surrounding leads. When the leads are heated up (case of large junctions) the thermal loss is decreased which reduces the required power density. For the smallest junctions, the power is too small to generate a sizeable heating of the leads leading to a larger required power density. Similarly, when the pulse width is decreased below 10ns, the heat transfer into the leads is not triggered, hence a required power density which is size-independent. Note that using the

heating of the leads (by reducing their width relative to the MTJ dimension) to reduce the power density is risky business as thermal cross-talk may occur, as well as current densities closer to the electromigration limit.

The best option to minimize the required power density while keeping a constant value upon scaling consists in adding thermal barriers on each size of the MTJ. These thermal barriers made from low thermal conductivity material with reasonably low resistivity are very efficient to decrease the heating power density as they help to confine the generated heat into the MTJ. The simulation in Figure 2 shows the corresponding reduction of power (yellow triangles). It is important to point out that in this scheme, the power density becomes size-independent – a very important factor for future scaling.

To conclude, the thermally assisted approach offers a promising solution for the next generation of MRAM as it can solve most of the current issues (write selectivity, power consumption, thermal stability) whilst offering full scalability to 65 nm node and beyond. The integration of thermal barriers is the last technological hurdle before a real implementation on large CMOS arrays.

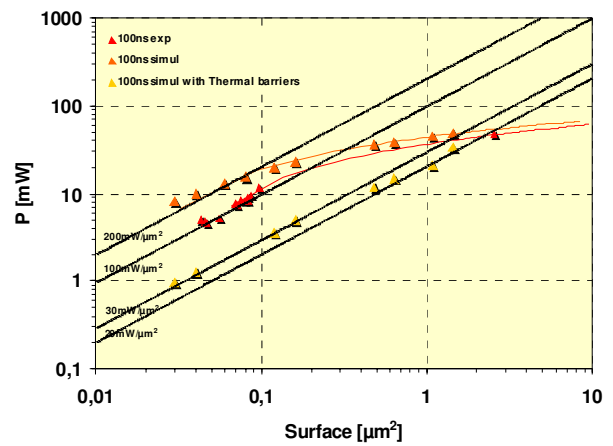


Fig. 2: Power vs surface for 100ns voltage pulsewidth. The red triangles are the experimental data, the orange triangles are thermal simulations and the yellow triangles are simulation with two thermal barriers added on the top and bottom surfaces of the MTJ.

References

- [1] B.F. Cockburn, records of MTDT04, 2004, 46
- [2] B.N. Engel et al, IEEE trans. Mag., 41, 2005, 132
- [3] I.L. Prejbeanu, IEEE Trans. Mag., 40, 2004, 2625
- [4] J. Van Driel et al, JAP, 88, 2000, 975

SESSION D

MRAM 2/2

[illegible]

Switching speed in high density MRAM based on precessional or spin-transfer switching

T. Devolder^a, C. Mauroury^a, H.W. Schumacher^a, C. Chappert^a, C. Maufront^b, J.-V. Kim^a, R. Fournel^b, Y. Suzuki^c, B. C. K. Yagami^{d, f, c}, Tulapurkar^{d, e}

^a Institut d'Electronique Fondamentale, UMR CNRS 8622, Université Paris Sud, 91405 Orsay cedex (France), devolder@ief.u-psud.fr

^b ST Microelectronics, Central R&D, 850 rue Jean Monnet, 38921 Crolles cedex

^c Osaka Univ., Graduate School of Engineering Science, Toyonaka, Osaka 560-8531, Japan

^d NanoElectronics Institute, AIST, Tsukuba 305-8568, Japan

^e CREST-JST, 4-1-8 Honcho, Kawaguchi 332-0012, Japan.

^f SSNC, Semiconductor Technology Development Group, SONY Corp., Atsugi, Kanagawa 243-0014, Japan.

Abstract

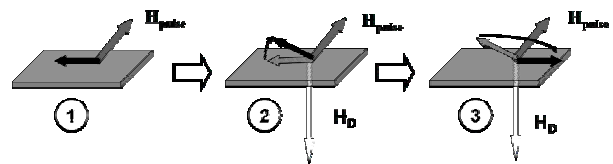
We discuss the speed issues of the writing step in Magnetic Random Access Memories. At present, the writing is performed by slow-rising magnetic fields. The magnetization vector follows the most decreasing slope in the energy landscape. The switching path is relaxation-dominated and can be described by the famous Stoner-Wohlfarth model to a satisfactory extent. However, the switching trajectory will drastically change when the operating frequency of MRAM will reach circa 500 MHz. Indeed when the fields are applied in a short and fast-rising manner, the switching path is then dominated by the precession torque, leading to a so-called "precessional dynamics". The associated trajectories can be ballistic and the switching times can be compressed far below the nanosecond regime. I will show our related experimental results and discuss them with simple models.

Despite its advantages, the energy cost of the precessional reversal does not scale well for cells below 200 nm and above 5 GHz. An alternative strategy is to transfer spin angular momentum from a reference layer directly into the magnetic cell to switch its magnetization, as originally proposed by J. Slonczewski. I will show experiments indicating that spin-transfer switching (i.e. without any applied field) have the relevant properties to relay today's technology both for frequency scaling above 5 GHz and for size scaling below 200nm. We will report the peculiar influence of the thermal fluctuations on the switching speed above 1 GHz and discuss its implications for the macrospin or non-macrospin character of sub-200nm magnetic cells.

Paper

In recent papers [1, 2] we have shown that precessional switching of magnetization is a convenient way to ensure, fast (sub ns), reliable, energy cost efficient writing of a M-RAM cell. In that technique, magnetization switching is triggered by a controlled pulse of magnetic field (H_{pulse}) applied perpendicular to

the easy magnetization axis of the M-RAM cell free layer. As schematized in figure 1, this drives the cell magnetization out of the plane, which in turn induces a fast precession around the large resulting demagnetizing field (H_D) perpendicular to the layer. This precession is a very efficient way to switch the magnetization from one direction to its reverse. It can for instance be obtained for H_{pulse} amplitudes down to half the cell magnetic



anisotropy field H_K [3], contrary to quasi-static mechanisms based on Stoner-Wohlfarth (SW) astroid.

Fig. 1: Schematic illustration of a precessional switching event

Bit selection and direct overwrite can then be obtained by superimposing a pulse of magnetic field H_{easy} applied along the free layer easy magnetization axis. No tight synchronization is needed between the two pulses, provided the longitudinal field is stable when the perpendicular field pulse is applied. In this scheme, the switching/no switching limit is given by the "dynamical astroid" [3] represented in Fig.3 for realistic addressing conditions. These include finite rise-time (178 ps) and imbricated or synchronous field pulses. An analytical model was derived to calculate the dynamical astroid in the limit of matched pulse duration. In this limit the hard axis field ($h_x < 0$) and the easy axis field ($h_y > 0$) are linked by the following expression [6]:

$$\frac{h_y}{h_k} = \frac{\left(1 + \frac{h_x}{2h_k}\right) \left(1 + \frac{h_x}{h_k}\right)}{2 - \frac{h_x}{h_k}}$$

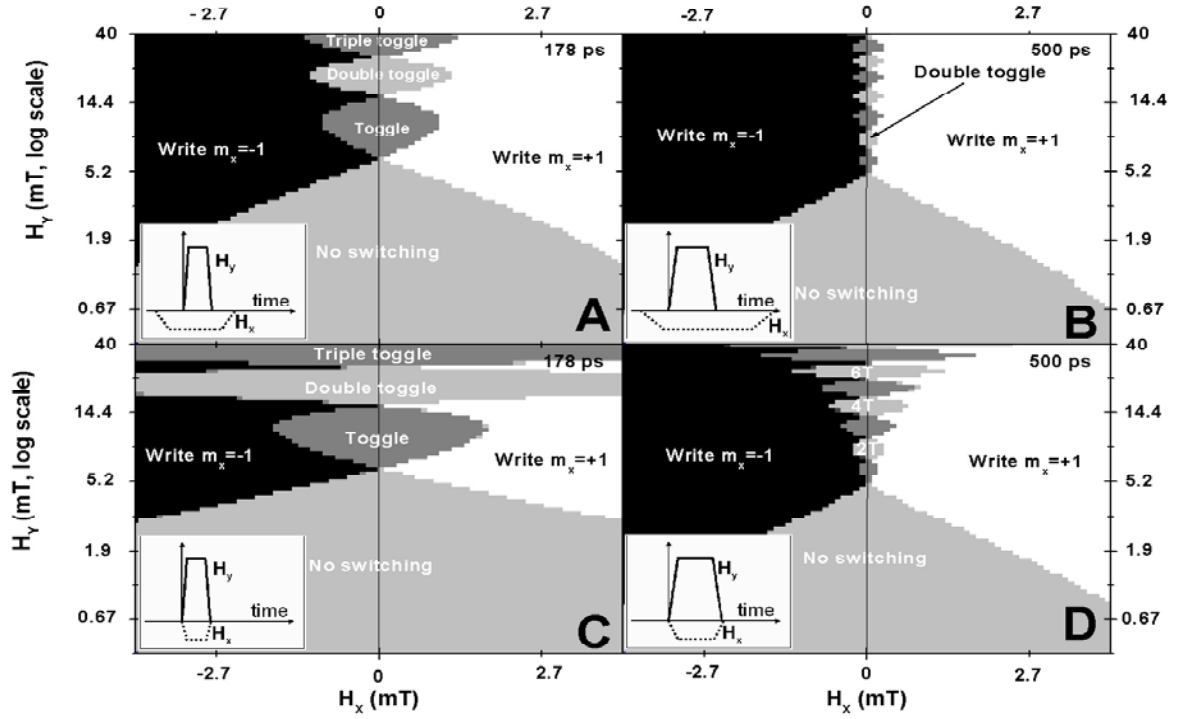


Fig.2: Simulated switching diagrams for an MRAM cell with CoFe free layer of demagnetizing factors $N_x = 0.0036$, $N_y = 0.064$ and $N_z = 0.99$. The damping parameter is $\alpha = 0.03$. The Hard axis pulse duration is (A), (C): 178 ps and (B), (D): 500 ps. The pulses H_y and H_x are synchronous in (C), (D) the relaxation is done in constant H_x for frames (A) and (B). The different regions of no switching, toggle, multiple toggle and direct-write are represented. The symbols “2T” stands for double toggle.

Based on the Landau-Lifshitz-Gilbert equation we have developed an approximate analytical theory of this precessional switching process, valid for any damping parameter α . The minimal transverse switching field is [5]:

$$H_{bif} = \frac{H_k}{2} + 0.59 \alpha \sqrt{H_k N_z M_s}$$

whereas the precession frequency is:

$$\omega \approx 0.847 \gamma_0 \sqrt{M_s (H_y - H_k/2)}$$

where H_k is the anisotropy field, N_z the vertical demagnetizing factor and M_s the saturation magnetization.

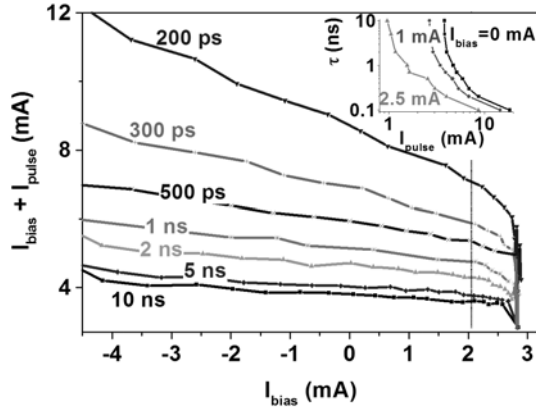
This allowed us to study in detail the influence of α in the bit writing scheme. We could also optimize the process for reducing cell size, versus the onset of super-paramagnetism. Moreover, we have studied how the writing parameter window is affected by a distribution of cell parameters in a large array. Finally, we have studied the influence of intercell dipolar coupling, when the cell size and array density are varied according to the characteristic linewidth of the semiconductor technology nodes.

Despite its advantages, the energy cost of the precessional reversal does not scale well for cells below 200 nm and above 5 GHz. An alternative strategy is to transfer spin angular momentum from a reference layer directly into the magnetic cell to switch its magnetization, as originally proposed by J. Slonczewski.

The spin-transfer magnetization switching may restore the scalability in high density Magnetic Random Access Memories (MRAM) beyond several Gbit/chip. The proof of concept has been done as early as year 1999 but many important issues for real applications are waiting for answers. Among the major concerns are the operating frequency limitations and the correlated magnitude of the needed write current I_C . In a practical memory circuit, the peak value of the write current I_C determines the minimal size of the write transistor, which sets a lower limit on the memory areal density.

In year 2000, J. Sun predicted that the switching speed $1/\tau$ in spin-transfer induced magnetization switching should scale with $|I - I_C| \ln \theta_0$, i.e. both with the overdrive current $I - I_C$ and with the initial misalignment θ_0 between the transported spin polarization and the macrospin to be reversed. However in experiments so far θ_0 was the misalignment of the magnetization of the free layer from its easy axis, mostly arising from finite temperature fluctuations. Increasing the switching speed can thus be done by increasing either the current pulse I , which is not technologically desirable or by preparing a more favourable initial condition with $\theta_0 \neq 0$ as suggested by Li et al.. A straightforward strategy is to change θ_0 by a

magnetic field pulse transverse to the easy axis, as classically done in magnetic field switching. However in practical memory architectures, this strategy would require additional addressing lines and large transistors to provide enough current, which would significantly increase the technological complexity. Another possible strategy is to exchange-bias the pinned layer at an angle θ_0 with the free layer's easy axis. However in this case a substantial part of the magneto-resistance signal is wasted at the expense of the SNR at the reading step.



Here we present a novel strategy to gain in switching speed while keeping the full magneto-resistance ratio and not requiring applying any magnetic field. We precharge the device with a dc bias current to excite a steady state precession with a finite cone angle,

so that the magnetization is very unlikely to be and stay near $\theta_0=0$ when the write current pulse is then sent to the device. The so-prepared initial precession increases the torque efficiency of the pulsed current and significantly accelerates the reversal event for given current amplitude. Equivalently, it reduces the total current needed to reverse in a certain duration.

References

- [1] H.W. Schumacher, C. Chappert, P. Crozat, R. C. Sousa, P. P. Freitas, J. Miltat, J. Fassbender, and B. Hillebrands, *Phys. Rev. Lett.* **90**(1), 017201 (2003).
- [2] H.W. Schumacher, C. Chappert, R. C. Sousa, P. P. Freitas, and J. Miltat, *Phys. Rev. Lett.* **90**(1) 017204 (2003).
- [3] J. Miltat, G. Aburquerque and A. Thiaville in *Spin Dynamics in Confined Magnetic Structures, Volume I*, edited by B. Hillebrands and K. Ounadjela (Springer, Berlin, 2001).
- [4] T. Devolder and C. Chappert, "Spectral analysis of the precessional switching of the magnetization in an isotropic thin film" *Sol. St. Com.* **129**(2), 97 (2004).
- [5] T. Devolder and C. Chappert, "Precessional switching of thin nanomagnets: analytical study", *Eur. Phys. J. B.* **36**, 57 (2003).
- [6] T. Devolder and C. Chappert, "Cell writing selection when using precessional switching in a Magnetic Random Access Memory", *J. Appl. Phys.*, **95**(4) 1933, (2004).
- [7] A. A. Tulapurkar, T. Devolder, K. Yagami, P. Crozat, C. Chappert, A. Fukushima, Y. Suzuki, *Appl. Phys. Lett.* **85**(22) 5358 (2004) ; T. Devolder , C. Chappert, P. Crozat, A. Tulapurkar, Y. Suzuki, J. Miltat, K. Yagami, *Appl. Phys. Lett.* **86**, 062505 (2005) .

Giant Tunneling Magnetoresistance and Tunneling Spin Polarization in Magnetic Tunnel Junction Memory Cells using MgO Tunnel Barriers

Stuart Parkin

IBM Almaden Research Center, San Jose, California, USA, Email: parkin@almaden.ibm.com

The relationship of the spin polarization of current tunneling from a ferromagnetic material through a tunnel barrier to its magnetization is of considerable current interest, particularly with regard to the development of magnetic tunnel junctions with high tunneling magnetoresistance suitable for magnetic memory and sensor applications¹. The tunneling spin polarization (TSP) can be directly measured using superconducting tunneling spectroscopy (STS) or inferred from measurements of tunneling magnetoresistance (TMR). The TSP depends not only on the ferromagnetic material but also on the tunneling barrier. Using crystalline (100) oriented MgO tunnel barriers and *bcc* CoFe electrodes, TSP values exceeding 85% are found from STS studies at 0.25 K with TMR values greater than 300% at 5K². The same ferromagnetic electrodes using conventional amorphous alumina tunnel barriers exhibit TSP values of only ~50%.

TMR values exceeding 220% are observed at room temperature using MgO tunnel barriers and ferromagnetic electrodes formed from alloys of Co-Fe and boron, a glass forming element. Cross-section transmission electron microscopy show that the CoFeB layers are amorphous but there is the likely possibility that the layer is crystallized at the MgO interface perhaps, in part, due to migration of B away from the MgO interface.

The read performance of magnetic tunnel junction (MTJ) magnetic memory cells in magnetic random access memories (MRAM) is strongly dependent on the magnitude of the tunneling magnetoresistance (TMR). The giant TMR values we find with MgO barriers will give rise to read times significantly shorter than those

possible with alumina tunnel barriers. MTJs with MgO tunnel barriers have other important advantages. In particular the structures are highly thermally stable with thermal stability to temperatures well above 400 C.

For both alumina and MgO tunnel barriers the sign of TSP is positive corresponding to the preferential tunneling of majority spin electrons. However, using magnetic electrodes formed from the ferrimagnetic rare-earth transition-metal alloys, such as Co-Gd, we show that the sign of the TSP and TMR can be either positive or negative depending on the composition of the alloy and temperature. For alloys of Co and Fe with Pt and Pd we show that the TSP is strongly influenced by bonding at the tunnel barrier/ ferromagnetic electrode interface. In particular, we compare results for CoPt alloys with AlN and Al₂O₃ tunnel barriers. For AlN barriers the TSP decreases linearly with increasing Pt content whereas for Al₂O₃ tunnel barriers the TSP has only a weak dependence on Pt content for up to ~50 atomic % Pt. These results can be understood as a consequence of stronger Co - oxygen bonds as compared to Pt - oxygen bonds and demonstrate the significant role of tunneling matrix elements in determining the magnitude of the tunneling spin polarization and tunneling magnetoresistance

References

- [1] S. S. P. Parkin, X. Jiang, C. Kaiser, et al., Proc. IEEE **91**, 661 (2003).
- [2] S. S. P. Parkin, C. Kaiser, A. Panchula, et al., Nature Mater. **3**, 862 (2004).

Preparation of TMR multilayer for MRAM production

W. Maass^a, B. Ocker^a, J. Langer^a, Y. Huai^b

^aSingulus Technologies AG, Hanauer Landstr. 103, D-63796 Kahl, Germany, wolfram.maass@singulus.de

^bGrandis Inc, 1266 Cadillac Court, Milpitas CA 95035, USA

Abstract

The preparation of MTJ multilayer under semiconductor fabrication conditions is crucial to achieve a high-yield and cost-effective production of MRAM chips, one of the most promising candidates for a next generation non-volatile solid-state memory. Employing Linear Dynamic Deposition (LDD) technology as key component of a semiconductor production PVD tool provides superior TMR process performance with outstanding uniformity across wafers up to Ø300mm. In this paper these results are reviewed in comparison with other available deposition technologies.

1. Introduction

Magnetic Tunnel Junctions (MTJ) are the key element to store information in the non-volatile Magnetic Random Access Memory (MRAM), and for other spintronics applications as ultra-high density magnetic recording heads and spin-logic devices [1] due to its high tunnel magneto-resistance (TMR, up to 220% at room temperature [2]) and large and tunable ranges of resistance-area product (RA), from below $1\Omega\cdot\mu\text{m}^2$ to $\text{k}\Omega\cdot\mu\text{m}^2$. MRAM has tremendous promise as a universal memory, which could replace Flash, SRAM and DRAM [3]. A big market potential is seen for the new MRAM technology, in particular as embedded memory in combination with CMOS logic for all kinds of so-called handheld devices.

A basic MTJ stack consists of two ferro-magnetic thin film electrodes separated by a thin tunnel barrier such as Al_2O_3 or MgO . Fabrication of high-performance MTJ stacks (which are usually composed of more than 6 thin film layers) is one of most crucial steps in MRAM manufacturing. The special requirements for MTJ stack deposition are very different from those for multilevel metallization in silicon semiconductor devices. The most important requirements, among others, include precise thin layer controllability down to 0.01 nm, excellent material uniformity across 200 to 300 mm wafers, smooth surface (interface) morphology, fast multilayer deposition capability, controllable oxidation process for the thin insulating barrier, and low costs of ownership of the system. These challenges cannot be met with traditional PVD systems using simple parallel target-substrate configuration. Currently there are two types of specially designed PVD cluster systems that are capable of meeting these demanding requirements: (a) linear dynamic

deposition (LDD) design, and (b) non-coaxial oblique substrate-target design [4]. The present paper highlights the superior features of the LDD technology along with some comparisons between the two novel PVD designs.

2. Thickness uniformity and repeatability

A film deposited by a PVD cathode (usually DC magnetron sputtering) on a fixed wafer substrate underneath will not have a particularly good thickness uniformity because the erosion profile of the cathode will be imaged into the substrate. There exist some techniques for overcoming this non-uniformity issue, e.g. by rotating the non-coaxial wafer underneath the cathode or rotating a off-centered system of permanent magnets behind the target. For the case of the very thin films under discussion here, this method requires several tens of rotations to be accomplished in order to obtain the desired uniformity effect. Since the deposition time for the very thin film is short (e.g. few seconds for about 1 nm and below), a very high rotation speed of the wafer is needed with implications of drawbacks: compatibility with ultrahigh vacuum chamber, long-term mechanical reliability, and particle generation. In addition, magnetic thin films usually require magnetic anisotropy induced by applying a magnetic field at substrate during deposition. Unless the target-wafer separation is large, this rotation of aligning field at substrate level will interfere with the magnetron field from the target, resulting in plasmas fluctuation that, in turn, influences the thickness uniformity and precise film thickness control. Large target-wafer separation, however, results in a very low collection efficiency of the deposition system, i.e. most of the very expensive materials are deposited onto shields and not on the wafer.

On the other hand, the so-called linear dynamic deposition (LDD) technology that moves wafers linearly underneath the target can achieve high thickness uniformity in large-scale deposition (up to 300mm) without rotating the wafer. Thickness uniformity of $1\sigma < 0.3\%$ has been achieved for Al and Ru, which are the two most critical layers used in MTJ stacks. All other materials as CoFe, NiFe, NiFeCr, Ta, and PtMn exhibit thickness uniformity $1\sigma < 0.6\%$.

The linear dynamic deposition technology also bears other superior deposition characteristics such as good coating efficiency by using short target-wafer distance (TWD, approx. 2 inch), precise thickness controllability (by varying precisely the

pass-speed of the wafer underneath the target) down to 0.01 nm and mechanical reliability because no high speed wafer rotation and movement are involved and no mechanical shutter is needed.

200 mm EA-skew α :

Range: $-1.5^\circ \leq \alpha \leq +1.4^\circ$

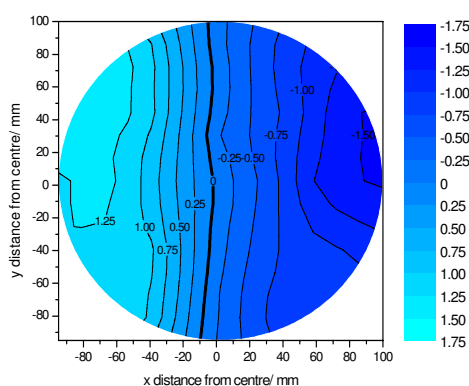


Fig. 1: Magnetic uniformity across the wafer for a film of Ta 3/NiFe 10 (in nm)

A specially designed electro-magnet as part of the LDD-PVD system generates a uniform magnetic field on the substrate, achieving exceptional magnetic uniformity. Fig.1 shows the uniformity of the magnetic skew of a 10 nm thick NiFe film across 200mm and 300 mm wafers, respectively.

Additionally, by LDD wedge-shaped layers over a large substrate (up to 300mm) along the wafer motion direction can be realized by simply accelerating or de-accelerating wafer pass-speed underneath the target. This technique is very useful to evaluate the thickness dependence of MTJ properties (e.g. TMR or RA) over a large thickness range [5] on a single wafer.

3. Surface and interface smoothness

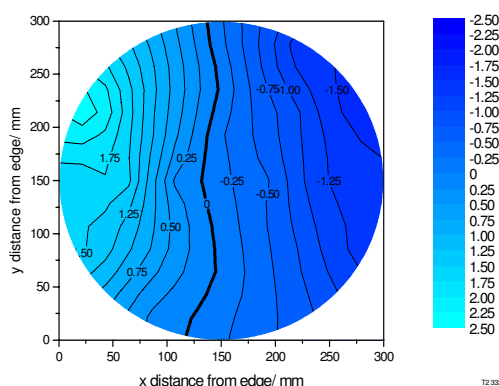
The surface and interface smoothness is extremely important in MTJ stacks to achieve high TMR and breakdown voltage, and a low interlayer coupling field (H_{in}) between the storage layer and the reference layer. It is in particular important since a smooth barrier topology is prerequisite for a high RA uniformity across a MRAM die, which is crucial for the reliable performance of the chip. Long mean free path of ad-atoms at low deposition pressure promotes high surface mobility for ad-atoms at the substrate surface, resulting in smooth surface morphology.

In conventional PVD systems with longer TWD ad-atoms arriving at the substrate are fairly collimated with angles of incident close to normal. The resulting grain growth will in tendency lead to rougher films. The LDD technique combining short target-wafer distance and low deposition pressure

(< 1μbar) enables smooth film growth due to high surface mobility of ad-atoms. Moreover, since the substrate is moving in and out of the ad-atom flux, a multiple-angle deposition is achieved with more

300 mm EA-skew α :

Range: $-1.6^\circ \leq \alpha \leq +2.6^\circ$



oblique angles of incidence. Since the TWD is small, the ad-atoms arrive at substrate with virtually no gas phase collisions. Both these factors result in increased surface mobility along the plane of the substrate, leading to highly smooth interfaces that some specialists believed so far can only be

Roughness Analysis

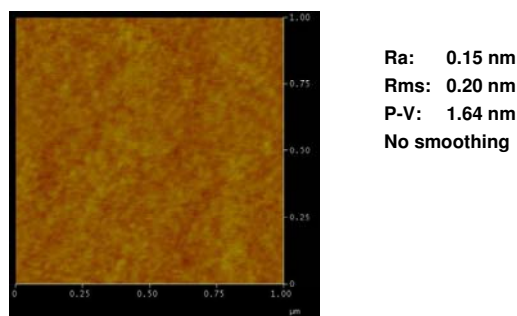


Fig. 2: Surface roughness of a typical bottom stack of Ta5/PtMn20/CoFe2.2/Ru0.8/CoFe2.2 (in nm).

achieved by ion beam deposition technology. Fig 2 shows the AFM measured surface roughness of a typical MTJ bottom electrode (Si/Ta5/PtMn20/CoFe2.2/Ru0.8/CoFe2.2, thickness in nm), exhibiting a RMS of 0.20 nm without using any additional smoothing techniques, on which the thin tunnel barrier layer (Al_2O_3 or MgO) directly grows. Fig. 3 shows a TEM cross section image of a low-RA - MTJ ($\text{RA} \sim 10 \Omega \cdot \mu\text{m}^2$) with a deposited thickness of the barrier Al of 0.5 nm, using CoFeB as amorphous pinned layer. As can be seen, the Al_2O_3 barrier is smooth and still homogeneous and continuous, proving that an ultra-thin high-quality

barrier is typically produced using LDD technology.

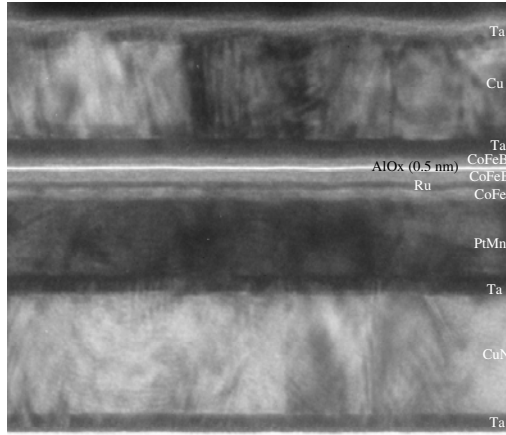


Fig. 3: TEM cross-section of a MTJ stack of Ta/CuN/Ta/PtMn/CoFe/Ru/CoFeB/Al₂O₃(0.5)/CoFeB/Ta/Cu/Ta

4. High TMR performance by Al₂O₃ and by MgO barriers

The discussed characteristics of LDD (thickness uniformity, precise thin film controllability and surface and interface quality) lead to a high MTJ performance. Fig. 4 shows TMR values with RA ranging from $< 1 \Omega \cdot \mu\text{m}^2$ to $\text{k}\Omega \cdot \mu\text{m}^2$ for a

thickness t , and subsequent either low-energy remote plasma oxidation or natural oxidation. TMR values were measured at room temperature using both lithography patterned junctions and the current-in-plane tunneling technique (CIPT) [6] on un-patterned films. A TMR ratio as high as 73% at $\sim 3 \text{ k}\Omega \cdot \mu\text{m}^2$ for an Al₂O₃ barrier, measured by CIPT, has been a benchmark for this barrier type. In the low RA range high TMR has been demonstrated for barriers prepared by natural as well as by plasma oxidation, where the plasma oxidation seems to have advantages, by resulting in slightly higher TMR. First investigations of LDD - TMR stacks using MgO as barrier show ultra-high TMR ratios up to 200% at RA below $100 \Omega \cdot \mu\text{m}^2$.

LDD technology also demonstrated excellent process repeatability and thin film controllability. Fig. 5 shows wafer-to-wafer variations for TMR and RA values obtained from patterned wafers. One sigma of 1.27% and 5.0% for TMR and RA has been achieved, respectively, where 5% RA variation corresponds approximately to 0.01 nm variation of barrier Al thickness.

5. Mass production of MTJs

The results discussed above relate to a 10-cathode/300mm chamber in a cluster tool arrangement with additionally one oxidation and one pre-clean module. It is an additional advantage of the LDD that processes developed for one wafer

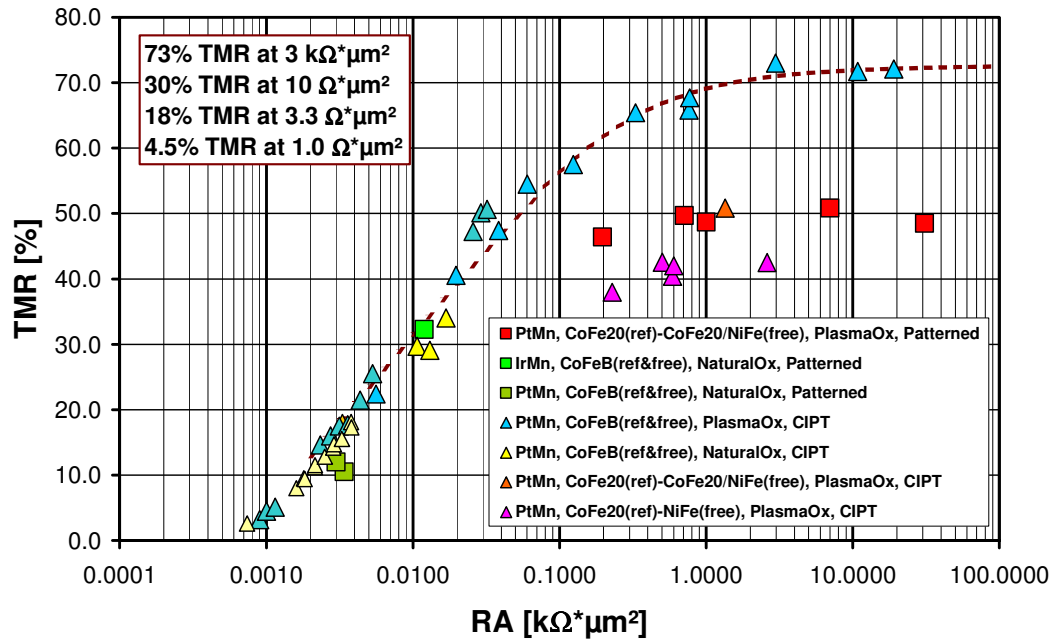


Fig. 4: TMR values vs. RA for different MTJ stacks (triangles correspond to CIPT, squares to patterned MTJs).

representative MTJ stack of Ta₁₀/PtMn₂₀/CoFe_{2.2}/Ru_{0.8}/CoFe_{2.2}/Al₂O₃(t)/CoFe_{1.5}/NiFe₄/Ta₅. The aluminum oxide barrier was formed by first depositing a thin Al layer of

size can directly be used for other types and sizes of substrates. This important feature will ease the transfer of results in the chain from R&D to large scale production.

The high collection efficiency discussed above in combination with high rate deposition, the low

latency time between depositions of individual stack layers and the effective handling of wafers in the cluster (only one deposition chamber needed to deposit all materials) results in a high throughput of the LDD system. Approximately 9 wafer/hour (based on a typical generic MTJ stack) per deposition module and 18 wafer/hour for a fully equipped system can be achieved, important for mass production.

existing designs, the LDD - PVD cluster system incorporating short TWD offers significant cost of ownership advantages for MRAM stack deposition. The superior thickness and material uniformity, as well as the high intrinsic surface smoothness make LDD also an attractive technology for other critical thin film applications in the semiconductor industry.

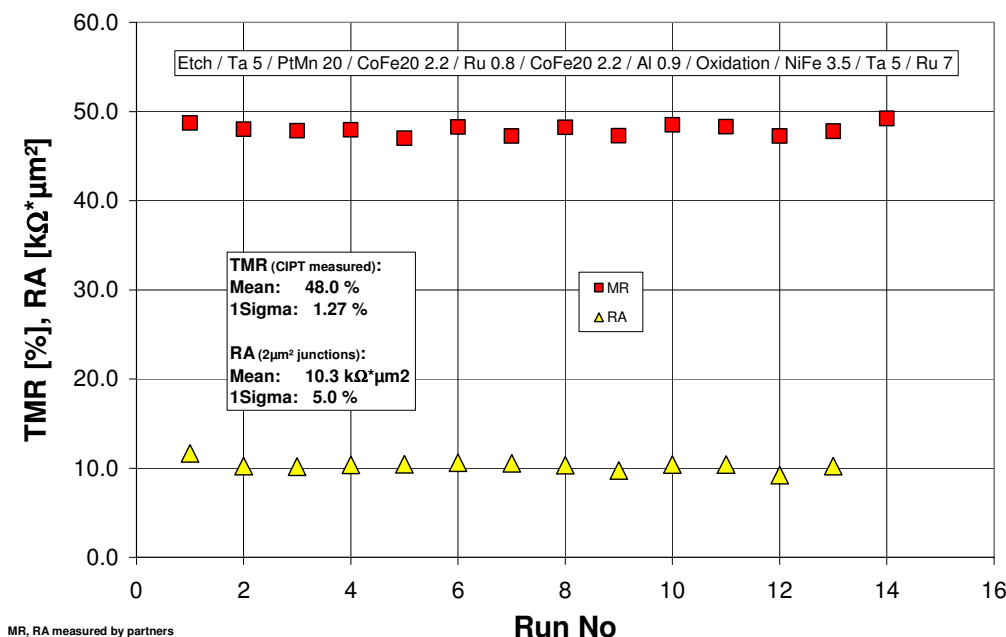


Fig. 5: TMR and RA repeatability

Another benefit of the high collecting efficiency of the LDD are significant lower costs of deposition in comparison to long target-wafer separation designs, including the case of non-axial oblique deposition designs [4]. In particular relevant are expensive materials like PtMn, IrMn or Ru, which will reduce the added costs/wafer by a significant factor for all sizes of wafers by using LDD.

Since LDD allows the use of adequately large targets without negative impact on tool footprint, the increased lifetime of targets enables to run a LDD machine about four weeks (7day/24hour production), until the next target-change and strip-and-clean downtime. In combination with a fast vacuum recovery after a scheduled downtime (16 hours until restart of tool qualification) this feature results in a high effective uptime.

6. Summary

Linear Dynamic Deposition (LDD) technology employed in a novel 10-cathode PVD module and incorporated in a cluster tool has been discussed in detail to be the technology of choice for the manufacturing of TMR multi-layers for MRAM production. Superior process performance including ultra-high TMR by using MgO barriers proves the maturity of this technology. Compared with

Acknowledgments

The authors would like to thank K.-H. Schuller, (Singulus Technologies), G. Proudfoot (Singulus Consultant) as well as M. Pakala (Grandis) for providing process results and for many helpful discussions and suggestions.

References

- [1] S.A. Wolf, D.D. Awschalom, R.A. Buhrman, J.M. Daughton, S. von Molnár, M.L. Roukes, A.Y. Chtchelkanova, and D.M. Treger, *Science* **294**, 1488 (2001).
- [2] S.S.P. Parkin, Ch. Kaiser, A. Panchula, P. Rice, B. Hughes, M. Samant and S.-H. Yang, *Nature Online*, Oct. 31 (2004); S. Yuasa, T. Nagahama, A. Fukushima, Y. Suzuki, K. Ando, *Nature Online*, Oct. 31 (2004).
- [3] R.P. Cowburn, *Materials today*, **July/August**, 32 (2003).
- [4] T. Tsunoda, M. Kosuda, K. Tsunekawa and N. Watanabe, *Semiconductor Manufacturing* **4/9**, 145 (2003).
- [5] W. Kula, J. Wolfman, K. Ounadjela, E. Chen, and W. Koutny, *J. of Applied Physics* **93**, 8373 (2003).
- [6] D.C. Worledge, P.L. Trouillions, *Appl. Phys. Lett.* **83**, 84 (2003).

Current Status and Issues for the Patterning of MRAM Devices

Thomas J. Kropewnicki^a, Ryan Patz, Jeffrey Stokes, Alex Paterson,

Theodoros Panagopoulos, and John Holland

Etch Product Business Group, Applied Materials, Inc., 974 E Arques Ave., Sunnyvale, CA 94085-4520 USA

^a Tom.Kropewnicki@amat.com

Abstract

The pervasiveness of mobile electronics has resulted in a rapidly growing market for nonvolatile memories. Flash memory is currently dominant in this area, but its extendibility to future technology nodes may be limited. Therefore, several alternative technologies, such as magnetic RAM (MRAM), phase change memory, polymer memory, and ferroelectric RAM, are actively being pursued. Each of these technologies has certain advantages over flash memory, including long retention times, low read or write times, or low power consumption. Our applications lab has recently seen a sharp increase in the amount of activity for nonvolatile memory development, with MRAM being the most common application requested.

A major challenge of these new memory technologies is the new materials that they introduce into the process flow. Especially challenging from the perspective of patterning MRAM devices is the etching of the nonvolatile materials they contain. The patterning of these materials is therefore going to require a thorough understanding of the etch mechanism of these materials and the process / device interactions to ensure high device yield. This talk will begin with an update on our recent efforts to fundamentally characterize the nonvolatile material etching process. In particular, we will discuss the impact of fundamental plasma parameters on the etch process, as well as plasma chemistry effects.

The second half of the paper will cover our view on the issues that need to be addressed for reliable MRAM production. MRAM development from an equipment perspective has benefited from needs closely matched to those of etching high κ materials, specifically high temperature etching in high density plasmas. Additionally, the complexities of the MRAM patterning process appear to challenge the single universal process chamber mentality, leaning instead toward an approach for MRAM systems that employ multiple chambers for etching, rinsing, and stripping. Will these trends continue in the future? The talk will conclude with a vision for the future of MRAM production.

Non-Contact Measurement of the Magnetic Properties of MRAM Wafers

W. Van Drent and F.Z. Vajda

ADE Technologies, 80 Wilson Way, Westwood, MA 02090, U.S.A.
email: fvajda@adetech.com, tel: +1-781-467-3513

Using the well known longitudinal Kerr-effect technique, a high resolution, high field instrument was designed specifically to measure MRAM wafers.

Although Kerr-effect is a well known measurement technique, customary implementations appear to indicate that it lacks the sensitivity necessary to measure magnetic materials with a low packing ratio or buried magnetic layers. The design principles presented in this paper allow us to measure MRAM arrays, where the packing ratio is in the order of 10-20%, as well as materials buried hundreds of Angstroms deep. Furthermore, utilizing a unique vector field control technique and a patented vector Kerr detection scheme, it is possible to measure materials with sub Oe coercivities, and also apply maximum fields above 1

Tesla. Thus, a single instrument is capable of serving a wide range of magnetic measurement needs for MRAM development and production, such as: MRAM arrays, including asteroid-type vector field measurements, full wafer mapping at low and/or high fields.

In the full paper, measurement results will be presented indicating the field control accuracy of our technique, as well as the sensitivity of the detection scheme on TMR sheet film stacks. It will be shown that these types of measurements are critical for MRAM development, as well as production deposition process control.

Spin-wave dynamics of a square magnetic bit in remanence

S.O. Demokritov^a, V.E. Demidov^b, B. Hillebrands^b, Ch. Back^c

^a Institut für Angewandte Physik, Westfälische Wilhelms-Universität Münster, D-48149 Münster, Germany,
e-mail: demokrit@uni-muenster.de

^b Fachbereich Physik, Technische Universität Kaiserslautern, D-67663 Kaiserslautern, Germany

^c Experimentelle und Angewandte Physik, Universität Regensburg, D-93040 Regensburg, Germany

Abstract

Quantized spin-wave eigenmodes in single, 16 nm thick and 0.75 to 4 μm wide square permalloy islands with a remanent four-fold closure domain structure have been investigated by micro-focus Brillouin light scattering spectroscopy. Up to five eigenmodes were detected. Spatially resolved measurements allowed the determination of the mode profiles and their classification. The main direction of the spin-wave quantization was found to be perpendicular to the local static magnetization. An additional less pronounced quantization along the direction parallel to the static magnetization was also observed.

1. Introduction

Substantial efforts have been put recently into the understanding of the excitation spectrum of small ferromagnetic elements (see, e.g., [1-7] and references therein). These excitations determine the response of small magnetic elements subject to ultrashort magnetic field pulses. This is of great importance for advanced magnetic recording technology, as switching time is further reduced and pushed well into the precessional regime. The same excitations cause the high-frequency magnetic thermal noise in the elements which limits the figure of merit of magnetic sensors and read heads. In the last several years high-frequency confined spin-wave eigenmodes of micrometer-sized magnetic elements have been systematically studied for the straightforward case of elements possessing an almost monodomain state [1]. Significantly less is known about the magnetization dynamics in mesoscopic systems with remanent inhomogeneous distribution of the static magnetization. The simplest of these systems is a magnetic disk in the vortex state characterized by an axial symmetry. The corresponding confined spin-wave modes are now intensively studied both theoretically and experimentally [2-5,7] and seem to be well understood. Another, more complex magnetic structure with reduced symmetry is a square in the flux closure Landau state with four-fold symmetry in the magnetization distribution [8]. Regarding the dynamic mode spectrum, no reliable theoretical prediction and only few experimental findings have been reported up to now, although such a state has been known for almost 70 years. The only publications reporting experimental results are Ref. [4, 6] where two low-frequency quantized modes were observed.

Due to the finite size of each of the four triangular domains in the given Landau state, confinement effects for the spin wave excitations are likely. It is the subject

of this paper to report on the discovery and the properties of these eigenmodes. In single Permalloy squares of 16 nm thickness with a lateral width L between 0.75 and 4 μm up to five eigenmodes are identified and their mode profiles are determined. [9]

2. Sample preparation

Single Permalloy ($\text{Ni}_{80}\text{Fe}_{20}$) squares were produced by e-beam evaporation on a Si substrate and capped with 2 nm Al for corrosion protection. Each square was placed inside a single turn Au loop of 300 nm thickness connected to a 50 Ω microstrip transmission line. The loop has inner and outer diameters of 8 and 12 μm , respectively. For a sketch of the sample see Fig. 1(a). Using non-patterned films evaporated in the same batch as the samples under investigation, the sample thickness, d and its saturation magnetization, M_s were measured by means of X-ray fluorescence and a superconducting quantum interference device magnetometer. The values obtained are 16 nm and 800 G, respectively. Before each measurement the samples were demagnetized using a gradually decreasing AC magnetic field applied perpendicularly to the plane of the samples. The presence of a four-fold closure domain structure was checked for each sample by means of magnetic force microscopy. A typical MFM image is shown in Fig. 1(b).

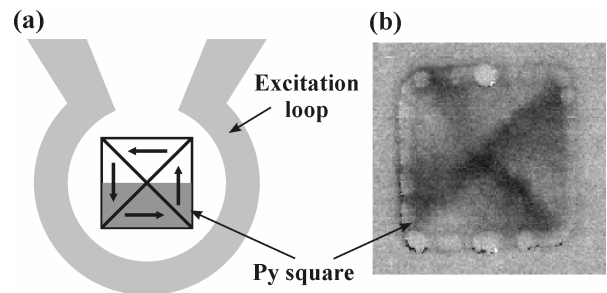


Fig. 1: (a) Schematic layout of a sample: a magnetic permalloy square with a closure domain four-fold Landau structure is surrounded by a gold excitation loop. The solid lines show the domain walls, whereas the arrows indicate the directions of the static magnetization in each domain. The shadowed area has been imaged using micro-focus BLS. (b) A typical MFM image for the square with a lateral width of 1 μm .

3. Experiment

First, spectra of thermally excited quantized spin waves were studied using the micro-focus Brillouin light scattering setup described elsewhere [10]. The setup is characterized by a diameter of the probing laser spot less than 300 nm and allows for simultaneous detection of

spin-wave excitations in a wide range of wavevectors of up to $2 \cdot 10^5 \text{ cm}^{-1}$. The results of these measurements are presented in Fig. 2. For the measurements the probing laser spot was positioned in the middle of one of the four triangular domains of the closure domain structure. The accumulation time of each spectrum was 2 hours. The lower frequency limit (2 GHz) was determined by the influence of the elastically scattered light, non-filtered by the Brillouin spectrometer due to its finite finesse for the given Free Spectral Range of 72 GHz.

As seen from Fig. 2, up to six peaks can be clearly distinguished in the spectrum of the square with $L = 0.75 \text{ }\mu\text{m}$. With increasing size the frequencies corresponding to the peaks decrease and the spectrum becomes denser. The broad spectrum for $L = 4 \text{ }\mu\text{m}$ appears to be continuous. Such an evolution can be associated with the fact that the finite decay length of spin waves becomes smaller than the size of the square. This length can be estimated for our samples as 1-2 μm . As a result in larger squares ($L = 2 - 4 \text{ }\mu\text{m}$) spin-wave modes are only

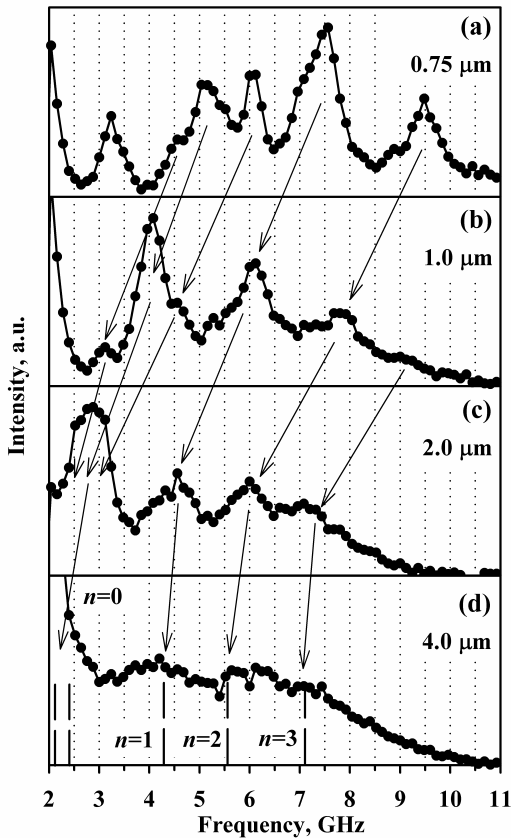


Fig. 2: Spectra of thermal spin waves for the squares with a lateral width L of 0.75 μm (a), 1 μm (b), 2 μm (c), and 4 μm (d). The arrows follow to the shift of the peaks with increasing size of the square. Dashed lines in part (d) indicate the frequencies corresponding to the distributions of the dynamic magnetization presented in Fig. 3.

weakly resonant and the spectrum comprises several broad, strongly overlapping peaks.

In order to classify the detected eigenmodes, space resolved measurements were performed allowing for an imaging of the lateral distributions of the dynamic magnetization. For this purpose the largest ($L = 4 \text{ }\mu\text{m}$) square was chosen for the reason of maximum spatial

resolution. Since the frequencies of the eigenmodes of such a square were not clearly resolved, micro-focus BLS imaging was performed for frequencies in the range from 2 to 8 GHz with a frequency step size of 100 MHz. In order to improve the sensitivity and to filter modes with fast spatial variations, an external excitation field at a fixed frequency was applied to the sample by applying a monochromatic microwave current through the excitation loop. The probing laser spot was scanned in two dimensions with a step size of 200 nm. Since BLS imaging with high spatial resolution is very time consuming, and taking into account the four-fold symmetry of the system being investigated, only half of the square was scanned as shown in Fig. 1. Several different mode profiles were observed at different frequencies of the external field. It was found, that transitions between profiles of adjacent modes demonstrating qualitatively different distributions of the dynamic magnetization are smooth and take place in a certain frequency interval. Therefore the frequencies of the eigenmodes were assigned to the center of the frequency interval, at which a given spatial mode distribution has been observed. These values are indicated in Fig. 2(d) by dashed lines. They are in agreement with the overall dependence of the frequencies of the eigenmodes on the size of the square obtained for smaller squares.

The measured distributions of the dynamic magnetization are presented in Fig. 3. Figures 3(a-e) were measured at 2.1, 2.3, 4.3, 5.6, and 7.1 GHz, respectively. As seen from the figure, all modes are

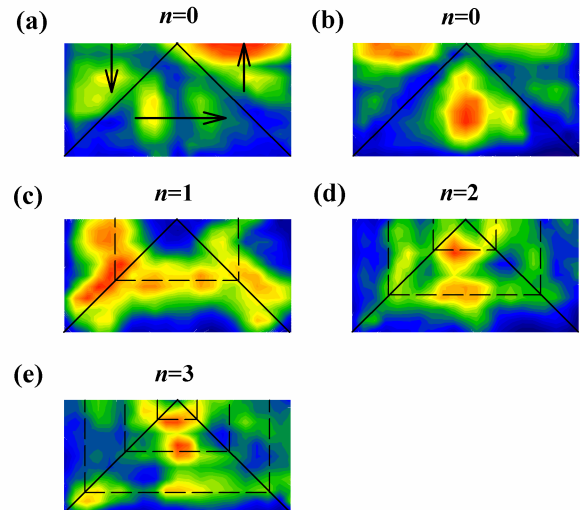


Fig. 3: Distributions of the amplitude of dynamic magnetization corresponding to the eigenmodes. Diagonal solid lines indicate the domain walls. Arrows in part (a) indicate directions of the static magnetization. The corresponding transversal quantization numbers n are indicated near the distribution.

spread over the entire domain and clearly show a quantization behavior. They can unambiguously be labeled by the number n of antinodes of the amplitude of the dynamic magnetization along the direction from the center of the square to its edges, i.e. along the direction perpendicular to the static magnetization in the domain (below we address this quantization as “transversal quantization”). The mode shown in Fig. 3(c) has one

antinode along this direction, 3(d) - two antinodes, and 3(e) - three antinodes. Consequently, we label these modes with the transversal quantization number n equal to 1, 2 and 3, respectively. Similar to the spin-wave modes observed in magnetically saturated stripes [1] the frequencies of the modes increase with increasing quantization number. This fact will be discussed below.

In addition to the modes with $n = 1, 2, 3$, two modes with practically uniform profile along the direction perpendicular to the static magnetization can be seen in Figs. 3(a) and 3(b). In agreement with the above notation we label them with $n = 0$. However, these two modes clearly differ from each other as the distribution of the dynamic magnetization along the direction parallel to the static magnetization is concerned: the mode in Fig. 3(a) has an antinode at the center of the domain, whereas the mode in Fig. 3(b) shows there a node. Thus, they demonstrate a second quantization along the direction of the static magnetization in the domain, which will be referred to as “longitudinal quantization”.

As seen from Fig. 2, with decreasing size L of the squares, the wide low-frequency peak corresponding to closely lying modes with $n = 0$ and different longitudinal profiles splits into several individual peaks, which can be clearly distinguished for the smallest square. Since the frequency splitting caused by the longitudinal quantization is very small, the presence of the longitudinal quantization is unambiguously confirmed experimentally for the modes with the transverse quantization number $n = 0$ only. However, one should expect a similar longitudinal quantization for the modes with $n > 0$. Most probably, the frequency splitting due to the longitudinal quantization is smaller for the higher-order transversal modes, which does not allow for the observation of individual peaks. However, as seen from Fig. 2, the peaks corresponding to the eigenmodes with $n > 0$ are clearly asymmetrical. They demonstrate shoulders at their low-frequency slopes, which can result from small-amplitude peaks located very close to the main ones. This fact might be considered as a hint to the longitudinal quantization for higher-order transversal modes.

4. Discussion

To our knowledge, a rigorous theory of quantized spin waves in the Landau domain structure is not existing yet. However, the observed quantized modes can be qualitatively understood based on the dispersion law for plane spin waves of a continuous film, magnetically saturated in its plane [11]. In fact, the frequency of a spin wave with the wavevector oriented perpendicularly to the static magnetization strongly increases with increasing wavevector. Consequently, a standing spin wave with larger number of antinodes (larger effective wavevector) along the direction perpendicular to the static magnetization should have a higher frequency. This is in agreement with our experimental findings. As far as longitudinal

quantization is concerned, the estimations show that for the used thickness of the samples and the value of the saturation magnetization of $\text{Ni}_{80}\text{Fe}_{20}$ the same increment in the longitudinal component of the wavevector results in rather weak changes in the spin-wave frequency. This might be the reason why the observed longitudinal quantization is less pronounced than the transversal one.

5. Conclusion

In conclusion, we have determined the spin-wave eigenmode spectrum of $\text{Ni}_{80}\text{Fe}_{20}$ squares with a four-fold closure domain structure. It is shown that the spectrum comprises a series of modes spread over the entire domain. The latter modes are quantized in both in-plane directions. The quantization along the direction perpendicular to the static magnetization is characterized by a significant frequency separation between the modes and can be clearly observed in micrometer-sized squares. The frequencies of eigenmodes quantized in this direction increase with increasing number of antinodes. The frequency splitting due to the longitudinal quantization is much smaller. As a result, it is only pronounced in the squares with submicrometer lateral dimensions.

This work was supported in part by the priority programme SPP1133 “Ultrafast magnetization processes” of the Deutsche Forschungsgemeinschaft and by the European Communities Human Potential programme under contract number HRPN-CT-2002-00318 ULTRASWITCH.

References

- [1] S.O. Demokritov, B. Hillebrands, and A.N. Slavin, *Phys. Rep.* **348**, 441 (2001).
- [2] B.A. Ivanov and C.E. Zaspel, *Appl. Phys. Lett.* **81**, 1261 (2002).
- [3] V. Novosad, M. Grimsditch, K.Yu. Guslienko, P. Vavassori, Y. Otani, and S.D. Bader, *Phys. Rev. B* **66**, 052407 (2002).
- [4] J.P. Park, P. Eames, D.M. Engebretson, J. Berezovsky, and P.A. Crowell, *Phys. Rev. B* **67** R 020403 (2003).
- [5] M. Buess, R. Höllinger, T. Haug, K. Perzlmaier, U. Krey, D. Pescia, M.R. Scheinfein, D. Weiss, and C.H. Back, *Phys. Rev. Lett.* **93**, 077207 (2004).
- [6] H. Stoll, A. Puzic, B. van Waeyenberge, P. Fischer, J. Raabe, M. Buess, T. Haug, R. Höllinger, C. Back, D. Weiss, and G. Denbeaux, *Appl. Phys. Lett.* **84**, 3328 (2004).
- [7] L. Giovannini, F. Montoncello, F. Nizzoli, G. Gubbiotti, G. Carlotti, T. Okuno, T. Shinjo, M. Grimsditch, *Phys. Rev. B* **70**, 172404 (2004).
- [8] L.D. Landau, E.M. Lifshitz, *Phys. Z. d. Sowjetunion* **8**, 153 (1935).
- [9] K. Perzlmaier, M. Buess, C.H. Back, V.E. Demidov, S.O. Demokritov, B. Hillebrands, *Phys. Rev. Lett.*, **94**, 057202 (2005).
- [10] V.E. Demidov, S.O. Demokritov, B. Hillebrands, M. Laufenberg, and P.P. Freitas, *Appl. Phys. Lett.*, **85**, 2866 (2004).
- [11] R.W. Damon and J.R. Eshbach, *J. Phys. Chem. Solids* **19**, 308 (1961).

Magnetic tunnel junction device model for MRAM and hybrid magnetic-CMOS components design

V. Javerliac, F. Bernard-Granger, Li. Prejbeanu

Spintec, URA 2512 CEA-CNRS, 17 Rue des Martyrs, 38054 Grenoble Cedex, France
email: javerlia@drfmc.ceng.cea.fr

The development of hybrid magnetic-CMOS components such as MRAM, Thermal-MRAM, MFPGA or RF require efficient models for magnetic nano-structures. Model parameter extraction like BSIM3v3 transistors are very accurate but need the realisation of many wafers and the use of specific scan routines. Major issue is to develop a physical model of the magnetic tunnel junction based on theoretical-empirical laws and

simulations (like finite element method). This model has been developed in Verilog-A and validate on Cadence Virtuoso Platform with Spectre 5.1.41 simulator. So, we can introduce it on hybrid system and perform magnetic and CMOS co-simulations, the first step to a “Magnetic Design Kit”.

Brillouin Light Scattering Observation of the Damon-Eshbach Mode in a Nano-Granular (SiO₂)Co/GaAs Film

A.A. Stashkevich^a, Y. Roussigné^a, G. Viau^c, N.A. Sergeeva^{ad}, M.P. Kostylev^d, L. V. Lutsev^b, A. I. Stognij^e, and N.N. Novitskii^f

^a LPMTM CNRS (UPR 9001), Université Paris 13, 93430 Villetaneuse, France, as@lpmtm.univ-paris13.fr

^b Research Institute 'Ferrite-Domen', 196084 Saint-Petersburg, Russia

^c ITODYS CNRS (UMR 7086), Université Paris 7, 75251 Paris, France

^d Electrotechnical University, 197376 Saint-Petersburg, Russia

^e Minsk Research Institute of Radiomaterials, 220024 Minsk, Belarus

^f Institute of Solid State and Semiconductor Physics, National Academy of Sciences of Belarus, 220072 Minsk, Belarus

Abstract

Brillouin light scattering measurements of dispersion characteristics of Damon-Eshbach surface spin wave (SW) mode were carried out on a granular film containing cobalt nanoparticles in amorphous silicon dioxide matrix. The measurements showed a pronounced spin wave signal at 28-29 GHz.

1. Introduction

The numerous applications of granular films, consisting of magnetic nanoparticles in an insulating matrix have continuously driven technological and academic research [1-3]. Their magnetic and electrical properties can be varied over a very wide range by independently modifying the particle sizes, their separation and composition [3]. Such structures are considered as the most important candidates for the devices based on spin-dependent tunnelling. Previous papers mostly addressed either the static characteristics of such nano-composites or investigated low-frequency, typically in the range 0.1 – 1000 Hz, behaviour of magnetic susceptibility χ . In this paper we report the results of a Brillouin Light Scattering experiment in the Damon-Eshbach geometry carried out in 25 – 30 GHz frequency range. It was backed with MFM imaging of the domain structure and SQUID measurements of magnetisation curves at ambient and low temperature.

2. Sample preparation and experimental set-up

The experiments were performed on a 600 nm thick film of amorphous silicon dioxide containing cobalt nanoparticles SiO₂(Co) grown on a 0.4 mm thick substrate of gallium arsenide (n-GaAs) of (100)-orientation. The (SiO₂)_{100-x}Co_x/GaAs film was grown by ion beam co-sputtering of composite quartz and cobalt targets. The concentration of cobalt nanoparticles in the silicon dioxide deposit, amounting to $x = 80$, was adjusted by changing the ratio of cobalt and quartz target areas. The film composition was checked by X-ray analysis. Prior to deposition the substrate was polished by a low-energy oxygen ion beam, so that the maximum roughness height did not exceed 0.5 nm.

Brillouin light scattering (BLS) spectroscopy has proved to be one of the most powerful methods providing access to the parameters which monitor the physical properties of magnetic films. It relies upon the inelastic scattering of light by thermally excited magnons and allows to directly measure dispersion characteristics of spin waves from which one can determine the pertinent magnetic parameters such as saturation magnetisation, anisotropy and exchange constants.

The BLS measurements (p-s scattering) were carried out in the usual Damon-Eshbach geometry. Light of a single-mode Ar⁺ laser using a power of 350 mW at wavelength λ of 514 nm was focused onto the sample and the frequency spectrum of the backscattered light was analyzed using a computer controlled Sandercock-type (2 × 3)-pass tandem Fabry-Pérot interferometer. The value of the wavenumber $q_{\parallel}$ was changed by varying the angle of light incidence θ measured against the surface normal of the sample: $q_{\parallel} = (4\pi/\lambda) \sin \theta$. Cross polarisations between the incident and scattered

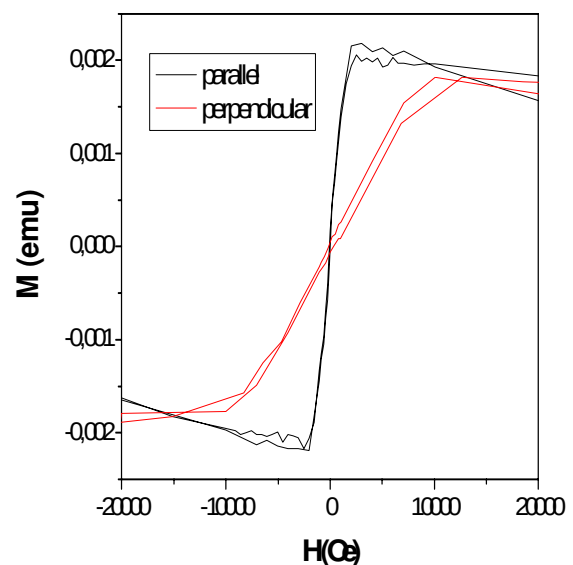


Fig. 1. Sample magnetisation curve

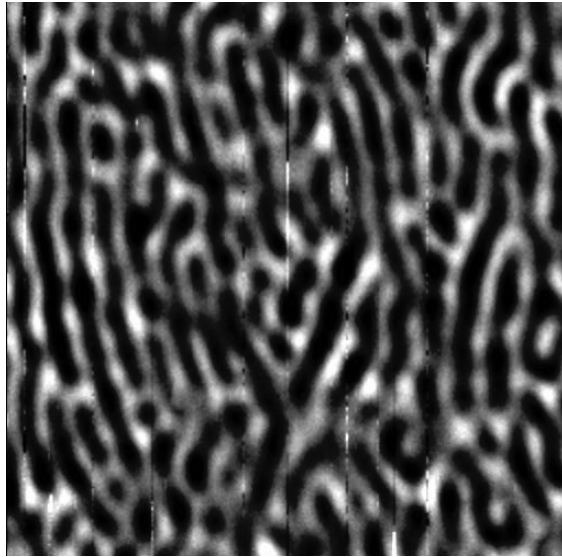


Fig.2 MFM image of the domain structure.

beams were used in order to practically suppress the light scattered by phonons.

3. Results and discussion

Before BLS measuring SW spectra the film was characterised in a static regime. Magnetisation curves of the film were measured with a Quantum Design MPMS-

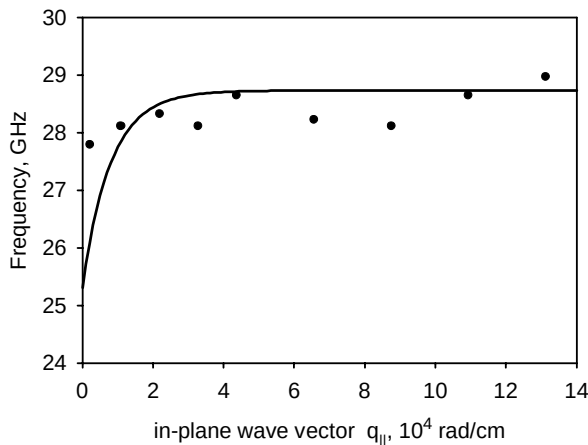


Fig. 3 Spin-wave spectrum measured by BLS technique. Dots are measured data. The solid curve represents numerical simulations of the Damon-Eshbach mode.

5S SQUID magnetometer. Magnetisation curves of the Co-SiO₂/AsGa film at room temperature with the field parallel and perpendicular to the film are given in Fig.1. At 300 K the coercivity and remanence are very small either in the parallel or in the perpendicular orientation. With the field perpendicular to the film the value of 9000 G was necessary to reach the saturation, this value is consistent with the $4\pi M_s$ value of the film obtained from the Co concentration in the film. In the parallel configuration the sample was fully saturated with an external field of 2000 Oe. On the whole, the sample exhibits ferromagnetic rather than superparamagnetic behaviour. This was partially confirmed by the presence of a domain structure (see Fig.2) obtained with a Digital Instruments Dimension 3100 AFM/MFM.

In Fig.3 is given a typical experimental spectrum measured with a relatively high external magnetic field $H_e = 5000$ Oe, thus ensuring full saturation of the sample (filled dots). The position of each experimental point was obtained

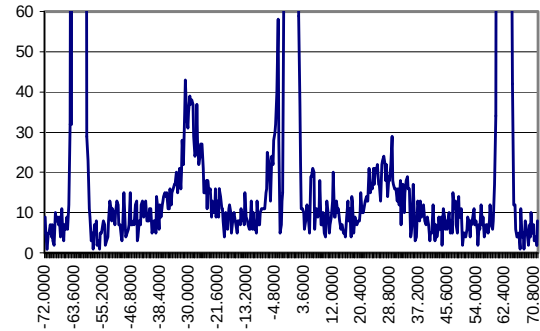


Fig. 4 BLS intensity spectrum for the angle of incidence of 40°

after averaging the positions of the Stokes and anti-Stokes peaks. The solid line is a square least fit with Damon-Eshbach mode calculated according to the classical formula [5]. Input fitting parameters are as follows : applied field $H=5000$ Oe, saturation magnetisation $4\pi M_s=9000$ Oe (the value obtained from the magnetisation curve in the perpendicular film), film thickness $d=600$ nm. For the fitted parameter, the gyromagnetic factor γ , we obtained a reasonable value of 3.02 GHz/Oe. The peculiarity of shapes of BLS spectral lines (see Fig.4) suggests a superposition of a wide, less intense and symmetric contribution of SWR resonances and an additional relatively narrow and intense Damon-Eshbach line on the anti-Stokes side.

4. Conclusion

The BLS measurements carried out at 26-29 GHz in a granular (SiO₂)₂₀Co₈₀/GaAs film, have detected the existence of Damon-Eshbach SW modes. The peculiarity of shapes of BLS spectral lines suggests the presence of SWR resonances.

Acknowledgements

This work was supported by the French Ministry of National Education and Research (ACI NR0095 "NANODYNE"). The authors are grateful to D.Billet for the MFM image.

References

- [1] J.Inoue and S.Maekawa, Phys. Rev. B 53, R11927 (1996)
- [2] K.Yakushi, S.Mitani, K.Takanashi, J.-G. Ha, and Fujimori, JMMM **212**, 75 (2000).
- [3] J.S.Micha, B.Dieny, J.R.Régnard, J.F.Jacquot, and J.Sort, JMMM **272-276**, 967 (2004).
- [4] S.Sankar, D.Dender, J.A.Borchers, D.J.Smith, R.W.Erwin, S.R.Kline, and A.E.Berkowits JMMM **221**, 1 (2000).
- [5] R.W.Damon and J.R.Eshbach J.Phys.Chem. Solids **19**, 308 b(1961).

Dynamic heating studies in Thermo-magnetic MRAM junctions

M. Kerekes^a, R. Sousa^a, O. Redon^a, L. Prejbeanu^a, U. Ebels^a, P. Xavier^b, B. Dieny^a,
J.P. Nozières^a

^a Spintec, URA 2512 CEA-CNRS, 17 Rue des Martyrs, 38054 Grenoble Cedex, France

^b UMR 5130 IMEP CNRS-UJF-INPG, Grenoble, France, 23 Av des Martyrs, 38016 Grenoble, France

Email: kerekes@drfmc.ceng.cea.fr

In order to overcome some limitations of the conventional MRAM architecture (write selectivity, power consumption, and thermal stability) a new thermo-magnetic write approach has been proposed. The magnetic tunnel junction (MTJ) stack is modified to include a second antiferromagnetic (AF) layer to exchange-bias the storage layer. During the write procedure a pulse of current is sent through the stack, increasing its temperature above the storage layer blocking temperature and below that of the reference layer. The system is then cooled under a magnetic field, larger than the high temperature coercive field of the storage layer. In this work the dynamic heating of sub-micron size MTJs developed for the thermo-

magnetic write scheme will be investigated using nanosecond and subnanosecond heating pulses. The influence of the voltage pulse width and junction area on the required power density will be discussed. The thermo-magnetic writing scheme is demonstrated through cycling of the cell between the two low and high resistance levels, and important features such as the selectivity, the protection from stray field erasure and the write time were confirmed. The submicron MTJs can be written using 500ps heating pulses proving that the thermo-magnetic approach is competitive with the conventional scheme in terms of write speed.

Thermal and non-thermal ultrafast magnetization control: New avenues in magnetic memory?

F. Hansteen^{a,b}, A. V. Kimel^a, A. Kirilyuk^a, D. Stanciu^a, R. V. Pisarev^c and Th. Rasing^a

^a IMM, Radboud University Nijmegen, 6525 ED Nijmegen, The Netherlands

^b The Norwegian University of Science and Technology, 7491 Trondheim, Norway

^c Ioffe Physico-Technical Institute, 194021 St.-Petersburg, Russia

Abstract

Ultrafast (10^{-12} sec and faster) magnetization dynamics and the search for mechanisms of high speed magnetic switching have become topics of ultimate interest in both condensed matter physics and information storage technology. In search of mechanisms for ultrafast optical control of magnetic ordering, we have studied laser induced spin dynamics in ferromagnetic and antiferromagnetic compounds. We find that ultrafast laser pulses can, via both thermal and non-thermal mechanisms, change the magnetic anisotropy of a material and cause spin reorientation or precession of the magnetization. Additionally, an ultrafast non-thermal effect of optically generated magnetic fields has been discovered. We demonstrate that this effect can be used to directly control the precessional dynamics of the magnetization in both magnetic garnet films and rare-earth orthoferrites. The existence of ultrafast non-thermal mechanisms, as well as ultrafast thermal processes open exciting new possibilities for applications in magnetic data storage and spin electronics.

1. Introduction

The fastest known way to alter the direction of the magnetization in a material is by precession [1, 2]. In ferromagnets the precessional frequency ν depends on the strength of the magnetic field H as $\nu = \gamma H$, with the gyroscopic constant $\gamma = 28 \text{ GHz/T}$. In antiferromagnets, which are becoming increasingly important for technological applications, no angular momentum is associated with the order parameter, and therefore spin dynamics in these materials is intrinsically much faster than in ferromagnets [3, 4].

Ultrafast manipulation and control of the magnetic order in ferromagnets and antiferromagnets requires means to influence the magnetic system on an ultrafast timescale. Modern femtosecond laser sources are able to excite a medium on timescales much shorter than that on which fundamental processes such as spin precession or spin-lattice relaxation take place. Pulsed laser sources of this type can therefore be used in optical pump-probe experiments to study ultrafast processes in solids in general, and spin dynamics in particular, with subpicosecond temporal resolution [5, 6].

Laser-induced switching in magnetic materials has until now been achieved only through thermal mechanisms that are limited in speed by the phonon-magnon interaction time. In metals this time is typically

of the order of a few picoseconds and in dielectrics typically several hundreds of picoseconds. Another problem is that the repetition rate is severely limited by the time it takes for heat to be transported away [7]. The existence of *nonthermal* mechanisms of interaction between light and spins could lead to a solution of these problems.

In this paper we report our recent results on ultrafast optical control of magnetization. We demonstrate that the heating caused by ultrafast laser pulses can change the magnetic anisotropy of an antiferromagnet within 1ps and cause spin reorientation. Furthermore, we demonstrate how two *non-thermal* photomagnetic effects allow for complete control of magnetization on the femtosecond timescale. This could open new avenues in the field of magnetic memory.

2. Experimental

An all-optical pump and probe technique was used to trigger and to study ultrafast spin dynamics in ferromagnetic (FM) and antiferromagnetic (AFM) materials. Regeneratively amplified 100fs laser pulses of wavelength $\lambda = 805 \text{ nm}$ from a Ti:Sapphire laser system were incident on the sample at a repetition rate of 500Hz. The resulting spin dynamics was probed by measuring either the Faraday rotation (in garnet samples with a net magnetic moment) or the linear magnetic birefringence (in TmFeO_3 samples) as a function of the time delay between the pump and probe pulses.

Two compounds belonging to the group of rare-earth orthoferrites were studied. TmFeO_3 is a canted antiferromagnet which due to strongly temperature dependent magnetic anisotropy exhibits a continuous 90° spin reorientation phase transition in the temperature range from 80K to 90K. Such canting of spins gives rise to a weak spontaneous magnetization $4\pi M_s = 100 \text{ G}$. Despite the small magnetization this material exhibits a giant Faraday rotation of about $3000^\circ/\text{cm}$ due to strong spin-orbit interaction. Also ferromagnetic garnet films of composition $\text{Lu}_{1.6}\text{Y}_{0.7}\text{Bi}_{0.7}\text{Fe}_{3.9}\text{Ga}_{1.1}\text{O}_{12}$ with in-plane magnetization $4\pi M_s = 550 \text{ G}$ were studied. These are highly transparent at the laser wavelength and also exhibit strong magneto-optical effects.

3. Thermally induced spin reorientation

The spin-reorientation phase transition taking place between the temperatures of 80K and 90K in TmFeO_3 can be initiated by laser pulses [4]. When the sample is

held at temperatures close to this transition, spin reorientation is triggered by the laser-induced temperature increase. The spin reorientation can be probed by measuring the change in magnetic linear birefringence, as shown in Fig. 1. The relaxation process can be divided into three distinct sections. First, the excitation decays via phonon cascades and the phonon system thermalizes rapidly with a time constant of about 0.3ps. A new lattice temperature is reached and the equilibrium anisotropy axis is changed. In a ferromagnet this would cause the magnetization vector to precess around the new equilibrium. In an antiferromagnet the resulting motion of the spins takes place in the plane spanned by the anisotropy field and the AFM vector. In Fig. 1 this process is denoted 2, and it has a characteristic time of about 5ps. After the initial relaxation we observe oscillations of the antiferromagnetic moments around their new equilibrium (process 3).

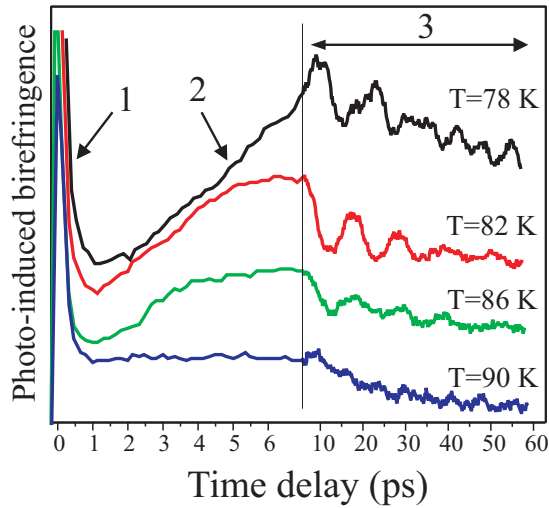


Fig. 1: Dynamics of the laser-induced spin reorientation as it is seen in birefringence measurements. Process 1 corresponds to the period when the new orientation of the easy axis of magnetization is set, process 2 is the reorientation of the spins to the new equilibrium and process 3 is the oscillations of the spins around the new equilibrium with the frequency of the antiferromagnetic resonance.

The experiment shows that the AFM spins in TmFeO_3 are reoriented by several tens of degrees within only a few picoseconds, much faster than the laser-induced destruction of the AFM order in similar materials. This demonstrates that spin reorientation triggered by changes in the magnetocrystalline anisotropy is a much faster process than phonon-magnon scattering.

In Fig.2 it is illustrated how this principle of laser-induced spin reorientation could be used to reverse the magnetization in a ferromagnetic layer. In an exchange-coupled FM/AFM bilayer laser-induced reorientation of the AFM vector by 90° will trigger, via the exchange coupling, a precession of the FM moment into the opposite state. The reversal speed of the ferromagnet is determined by the precession of the ferromagnetic spins in the strong interfacial exchange field and does not

require any additional applied field. This exchange field can be strong and spin reorientation times in the FM layer may reach the picosecond timescale. Note that in contrast to hybrid recording, the magnetic anisotropy is not simply reduced by heating, but modulated. This results in orders of magnitude faster spin reorientation in ferromagnetic materials.

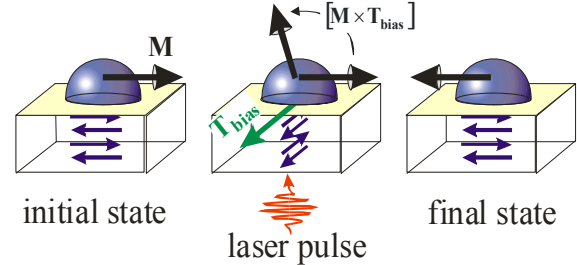


Fig. 2: Ultrafast laser-induced switching of a ferromagnet / antiferromagnet bilayer. The magnetization of the ferromagnetic layer precesses in the exchange field of the antiferromagnet into the new state [4].

Despite the fact that thermally induced spin reorientation can be fast, thermal processes inherently suffer from low repetition rates as excess heat must be transported away. This poses a serious limitation for many technical applications. The existence of non-thermal mechanisms of interaction between light and spins could lead to a solution of this problem.

4. Non-thermal optical spin manipulation

From studies of the laser induced magnetization dynamics in ferrimagnetic garnet films two distinct *non-thermal* and *ultrafast* photomagnetic effects have been identified:

1. Photo-induced magnetic anisotropy.
2. Optically generated magnetic field pulses.

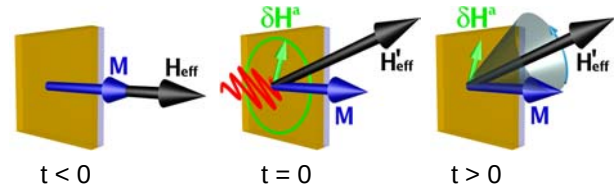


Fig. 3: A graphical illustration of the excitation process and subsequent precessional dynamics triggered by linearly polarized laser pulses. The photo-induced magnetic anisotropy δH^a changes the equilibrium ($H_{\text{eff}} \rightarrow H'_{\text{eff}}$) and M immediately starts precessing.

Linearly polarized laser pulses modify the magnetocrystalline anisotropy and create a new equilibrium state for the magnetization. This state is long-lived ($\tau \gg 3\text{ns}$) and is believed to be caused by optically induced electron transfer between ions on non-equivalent sites in the crystal. Effectively this causes a redistribution of ions, with an accompanying change in the magnetocrystalline anisotropy [8-10]. The direction of the photo-induced magnetic anisotropy δH^a depends on both the plane of polarization of the laser pulses and the orientation of the magnetization with respect to the crystallographic axes of the garnet film. The process of

laser-induced magnetocrystalline anisotropy and the subsequent precessional dynamics is illustrated in Fig.3. Immediately after the photo excitation the magnetization starts precessing around the new equilibrium $\mathbf{H}_{\text{eff}}$.

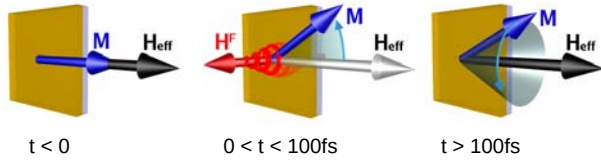


Fig. 4: A graphical illustration of the excitation by circularly polarized light. During its presence in the sample the laser pulse acts as a magnetic field of about $\mathbf{H}^F=0.6\text{T}$ along the $\mathbf{k}$ -vector of the optical pulse. The magnetization precesses in this strong field for the 100fs duration of the pulse. Subsequent precession takes place in the effective field $\mathbf{H}_{\text{eff}}$.

A different mechanism exists for circularly polarized laser pulses (see Fig. 4). They were found to act as strong transient magnetic field pulses of about $\mathbf{H}^F=0.6\text{T}$ during their presence in the sample. This is known as the inverse Faraday effect [11-13]. Pump pulses of opposite helicity give rise to precession with opposite phase, indicating that the direction of $\mathbf{H}^F$ is determined by the helicity of light. As for linearly polarized pulses, circularly polarized pulses also modify the magnetic anisotropy. However, the direction of $\delta\mathbf{H}^a$ now only depends on the initial orientation of $\mathbf{M}$ with respect to the crystal axes, which is determined by the balance between the externally applied field and the anisotropy fields of the sample. It can thus be tuned by varying the strength of the applied field.

These two photomagnetic effects can be combined to achieve complete optical control of the magnetization on the femtosecond timescale. We have demonstrated an approximately 1° switching of the magnetization within 100fs and using only a single laser pulse, see Fig. 5. The applied field is tuned so that after precessing for 100fs in the strong axial field $\mathbf{H}^F$, the magnetization ends up exactly in the new equilibrium state $\mathbf{H}'_{\text{eff}} = \mathbf{H}_{\text{eff}} + \delta\mathbf{H}^a$ simultaneously created by the optically induced anisotropy field. No further precession will then take place, as the magnetization is oriented along the effective field direction. It remains stable in this new state for several nanoseconds until the photo-induced anisotropy relaxes.

Direct optical control of magnetization can also be achieved by use of multiple circularly polarized pump pulses. This approach works equally well in garnets and orthoferrites. In Fig. 6 is shown how two pump pulses, separated in time by about 17-18ps, can be used to trigger and control the precession of $\mathbf{M}$. By carefully timing the arrival of the second pump pulse amplification as well as complete stopping of the precession can be achieved. This is the so-far best demonstration of the unsurpassed speed with which magnetization control can be achieved using laser techniques. Strikingly, a change in the arrival time of the second pump pulse by only about a picosecond decides whether the system will be left in a stable state (lower curve) or with a large precession amplitude (upper

curves). No existing electronics even remotely approaches these timescales.

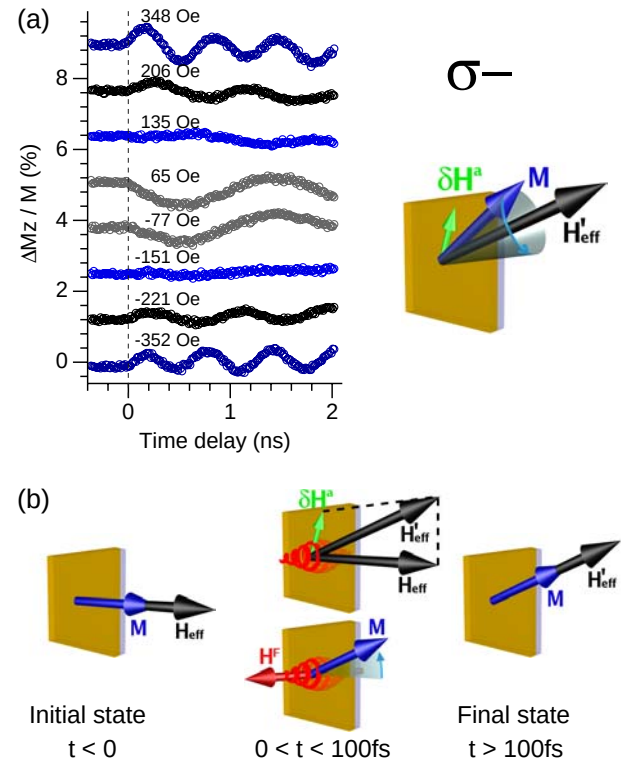


Fig. 5: All-optical switching of the magnetization within 100fs. (a) Precession of the magnetization triggered by circularly polarized laser pulses of helicity σ at different values of the in-plane applied magnetic field. At an applied field of $\pm 150\text{Oe}$, no precession is observed due to a perfect balance of the two photomagnetic effects. (b) Graphical illustration of the switching process. Initially at $t < 0$ the magnetization is along $\mathbf{H}_{\text{eff}}$. During the presence of the laser pulse $0 < t < 100\text{fs}$ photo induced modification of the anisotropy fields leads to a new long-lived equilibrium along $\mathbf{H}'_{\text{eff}}$. Simultaneously, the strong optically generated magnetic field $\mathbf{H}^F$ causes the magnetization to precess into the new state. After $t > 100\text{fs}$ the optical pulse is gone and the $\sim 1^\circ$ switching of $\mathbf{M}$ is complete.

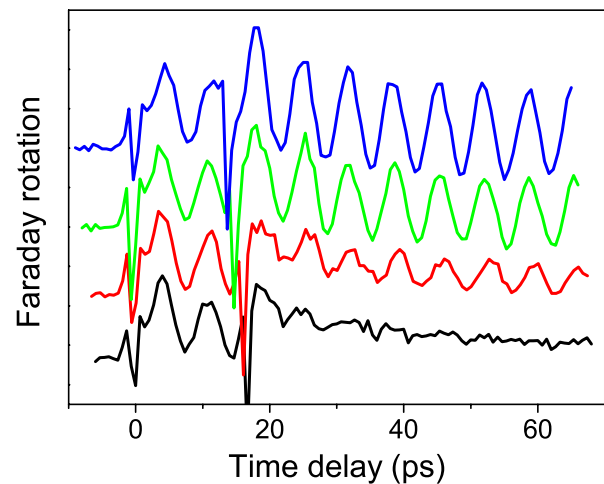


Fig. 6: Direct optical control of the magnetization dynamics in the orthoferrite sample by two circularly polarized pump pulses of the same helicity. Depending on the time of arrival of the second pump pulse the precession can be amplified or stopped completely.

5. Conclusions

We have demonstrated picosecond thermally induced spin reorientation in the antiferromagnetic orthoferrite TmFeO_3 and *sub-picosecond non-thermal* optical control of the magnetization in both orthoferrite and garnet samples. In view of the great variety of magnetic materials, the direct effect of light on the spontaneous magnetization in other materials is foreseen. Our findings open new insight into the understanding of ultrafast magnetic excitations and may provide new prospects for applications of ultrafast photomagnetic phenomena.

References

- [1] Th. Gerrits, H. A. M. van den Berg, J. Hohlfeld, L. Bär and Th. Rasing, *Nature* **418**, 509-512 (2002).
- [2] H. W. Schumacher et al., *Phys. Rev. Lett.* **90**, 017204 (2003).
- [3] A. V. Kimel et al., *Nature* **429**, 850-853 (2004).
- [4] C. Kittel, *Phys. Rev.* **82**, 565 (1951).
- [5] E. Beaurepaire, J. C. Merle, A. Daunois and J. Y. Bigot, *Phys. Rev. Lett.* **76** 4250-4253 (1996).
- [6] J. Hohlfeld, E. Matthias, R. Knorren and K. H. Bennemann, *Phys. Rev. Lett.* **78**, 4861-4864.
- [7] J. Hohlfeld, Th. Gerrits, M. Bilderbeek, Th. Rasing, H. Awano and N. Ohta, *Phys. Rev. B* **65**, 012413 (2002).
- [8] R. W. Teale and D. W. Temple, *Phys. Rev. Lett.* **19**, 904-905 (1967).
- [9] V. G. Veselago, R. A. Doroshenko, S. G. Rudov, *Zh. Eksp. Teor. Fiz.* **105**, 638-647 (1994). [*JETP* **78**, 341-345 (1994)].
- [10] A. Stupakiewicz, A. Maziewski, I. Davidenko, V. Zablotskii, *Phys. Rev. B* **64**, 064405 (2001).
- [11] L. P. Pitaevskii, *Sov. Phys. JETP* **12**, 1008-1013 (1961).
- [12] J. P. van der Ziel, P. S. Pershan and L. D. Malmstrom, *Phys. Rev. Lett.* **15**, 190-193 (1965).
- [13] P. S. Pershan, J. P. van der Ziel and L. D. Malmstrom, *Phys. Rev.* **143**, 574-583 (1966).

Magnetic Random Access Memory Integrated on Silicon (the MEMIS project)

C. Duluard^a, A. Bsiesy^{a,b}, V. Safarov^c, A. Filipe^d, A. Francinelli^{c,d}, H. Achard^e

^a SPINTEC, CEA/Grenoble and CNRS(URA 2512), , 38054 Grenoble Cedex 9, France

^b Université Joseph Fourier-Grenoble I, France

^c CRMCN, CNRS (UMR 6631) and Université de la Méditerranée, 13288 Marseille, France

^d SPINTRON, Hôtel Technologique - BP100 - 13382 Marseille Cedex 13, France

^e LETI, CEA/Grenoble, 17 av. des Martyrs, 38054 Grenoble Cedex 9, France

The extension of spintronics to heterostructures combining ferromagnets and semiconductors is a very promising research field that can lead to the development of innovating spintronic devices. In the field of MRAMs, a further development would lie in the integration of magnetic material at the CMOS level i.e. directly on silicon. Indeed, the recent demonstration of spin-polarized electrons injection from ferromagnetic metal into semiconductors shows the possibility, at least from the fundamental point of view, of integrating embedded MRAMs [1], [2]. However, the main difference with standard CMOS process would be the deposition over the silicon substrate of spin-polarized injector and collector. Consequently, the integration of ferromagnet material into a CMOS process seems to be a major challenge towards enabling embedded magnetic devices. In this work, we have designed a single cell MRAM demonstrator integrating ferromagnet electrodes directly onto silicon. The studied device is planar where spin injection and collection are obtained by two ferromagnet/tunnel insulator barrier electrodes, labelled “injector” and “collector” respectively. By a proper design based on shape-dependent magnetic anisotropy, the ferromagnets of these electrodes can show different coercive fields. In these conditions, the electron spin-dependent injection and collection towards and from silicon can be analysed through magnetoresistance characterization. The ferromagnet electrode magnetization switching under an external static magnetic field will be used to write “0” or “1” logic values. This field will be adjusted to a value that can switch the magnetization of the injector while keeping that of the collector unchanged. This “write” process is used for the purpose of demonstrating the MEMIS concept viability. In the future, the integrated memory cells addressing will use the traditional Field Induced Magnetic Switching produced by current pulses in metal lines running over the ferromagnet electrode.

The first part of this paper will be devoted to electrical characterization of the injector and collector electrodes to ensure that the current through the ferromagnet/tunnel barrier/silicon (F/I/S) structure involves direct tunnelling, a spin-conservative transport mechanism. Capacitance-voltage as well as current-voltage characteristics at ambient and low temperatures have been analysed. These results show that, above flat band potential, a direct electron tunnelling can take place in the F/I/S structure. However, at smaller voltages a defect-related current takes place. The concentration of these defects turned out to be largely determined by the ferromagnet deposition process. The figure 1 shows the density of current versus the voltage for the NiFe (10 nm)/ SiO₂ (2 nm)/Si MOS structure.

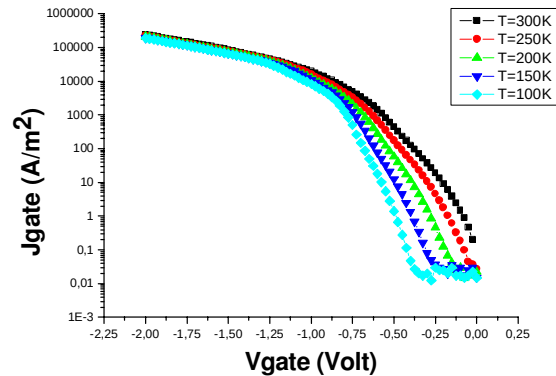


Fig. 1: Current density vs. gate voltage obtained on a p-type MOS structure of 4 μm^2 surface and at different temperatures. The gate is 10 nm of NiFe. The oxide is a 2 nm thick SiO₂.

The figure 2 shows the capacitor at several frequencies versus the voltage of a MOS structure with a gate of Ni, an aluminum oxide and a P – doped silicon substrate.

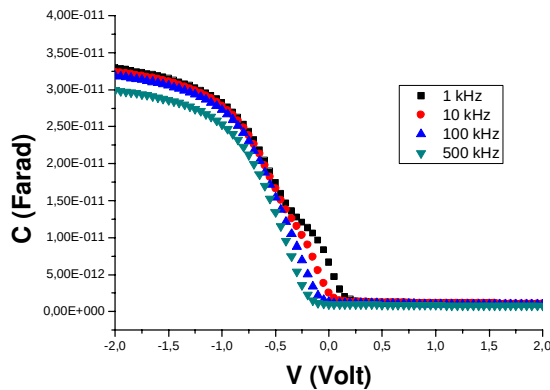


Fig. 2: P-MOS capacitor C-V characteristic with Ni gate. The surface of the capacitor is $2400 \mu\text{m}^2$. The oxide is a 2,5 nm thick Al_2O_3 .

In the second part of this paper, magnetic characterization of the electrodes in relation with their thickness and aspect ratios (length/width) will be presented. In particular, it will be shown that only a very thin magnetic “dead-layer” is observed on the ferromagnet/insulator interface. This rules out the undesirable spin depolarisation at the interface electronic states prior to injection into silicon. The figure 3 shows the magnetization versus the thickness of a layer of Co onto an oxide of silicon covered by a layer of Pt. Moreover, no significant dead layer is built up even after a 30mn thermal annealing at 450°C in the N_2/H_2 forming gas ambient. The same results are obtained when a Al_2O_3 is used as a dielectric layer.

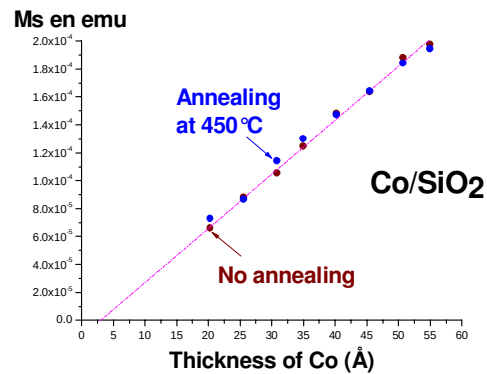


Fig. 3 : Magnetization of a Co layer versus the thickness of Co. The linear extrapolation gives us no significant dead – layer between the Co layer and the oxide of silicon.

Finally, some preliminary results on the magnetoresistance measurements will be presented.

References

- [1] A. Fert and H. Jaffrès, Physical Review B, 64, 184420, 2001
- [2] V.F. Motsnyi, Applied Physics Letters, 81, 265, 2002

SESSION E

NV Innovative Architectures

Chairmen: **Agostino PIROVANO**, STMicroelectronics, Italy (Agrate)
 Thierry YAO, AMI Semiconductor, Belgium (Oudenaarde)

NVM based on FinFET device structures

F. Hofmann, M. Specht, U. Dorda, R. Kömmling, L. Dreeskornfeld, J. Kretz, M. Staedele,

W. Rösner, L. Risch

Infineon Technologies AG, Corporate Research, D-81739 Munich, Germany

franz.hofmann@infineon.com

Abstract

High density data flash memories are essentially used in mobile applications. For logic applications FinFET type devices are known to have good scalability down to 10 nm gate length. This device architecture combined with a trapping layer enables memory cells with feature sizes well below 50 nm. To show the scaling potential of SONOS FinFET memories, devices are processed on SOI wafers with fin widths varying from 8 nm to 30 nm and gate lengths scaled down to 20 nm. We discuss three different storage modes of FinFET trapping layer NVM devices: a) single bit SONOS cell, b) multi level SONOS cell and c) NROM dual bit device. For a) and b) program and erase is done with Fowler-Nordheim tunneling and for c) channel hot electrons are used for programming and hot holes are injected for the erasing of the localized charges. SONOS FinFET memory devices show excellent functionality down to 20 nm channel length.

1. Introduction

The non volatile memory market for mobile applications with small form factor is dominated by high density floating gate NAND Flash memories for data storage; these devices have experienced explosive growth in the past years. Key drivers are the mobile digital consumer applications like digital cameras, handsets, mobile phones, MP3 players, PDA,... Future driver for highest density flash memories will be camcorders and mobile video applications. Thus portable memory with small form factor in the 10 GByte range at low cost for the consumer market will be needed. For instance video compression codec like MPEG4 can store 45 minutes of a movie in one GByte. Mechanical drives like DVD recorder or hard disks will be outperformed by flash due to weight, size, power consumption and shock constraints. In order to get this high integrated NVM up to 10 GB memory density feature sizes below 50 nm are required. Currently highest density NVMs are realized by floating gate NAND architectures. Here typical 32 cells are arranged in one string. The serial memory transistors are between two select transistors and there are only two contacts for the string, so small cell size of $4.5 F^2$ can be achieved, which is close to the lithographic limit. Operating in multi level mode with 4 levels, an effective cell size of $2.25 F^2$ has been demonstrated [1]. NAND cells are programmed and erased by Fowler-Nordheim tunneling with $t_p=200 \mu\text{sec}$, $V_p=18\text{-}20\text{V}$;

$t_e=2 \text{ msec}$, $V_e=20\text{V}$ [1]. The standard industrial specification for NVM is a data retention of 10 years at $T=85\text{-}125^\circ\text{C}$ after 10^5 cycles. This spec results in a very challenging leakage current of $I_L < 10^{-24} \text{ A/cell}$, so the tunnel oxide has to be thicker than 6-8 nm. For sub 60 nm feature size some of the most challenging issues are tunnel oxide thickness and capacitive coupling of floating gate to floating gate [2]. Trapping layer memories do not suffer from these drawbacks. There are a lot of investigations for an ideal cell device. In CMOS logic research multi gate FinFETs turn out to have the best scalability down to sub 10 nm due to improved electrostatic gate control of the channel [3]. Therefore, by combining a trapping layer with a tri gate FinFET device architecture a highly scalable NVM device is created [4]. Further, since the thickness of the trapping layer gate stack is smaller than the thickness of a floating gate control gate stack, this offers higher three dimensional integration density for SONOS FinFET architecture.

2. SONOS single bit FinFET with 20 nm fins

The FinFETs are fabricated on SOI wafers with a silicon thickness of 50 nm and a buried oxide of 100 nm [5]. The Si-fins are orientated in $\langle 110 \rangle$ with fin thickness of 20 nm and e-beam lithography was used for defining the small fins. Fin etch was done with a hard mask by dry etch. After a sacrificial oxidation a tunnel oxide of 3 nm is grown. The nitride trapping layer is deposited by LPCVD and afterwards a thermal blocking oxide of 4.8 nm is grown on top of the nitride (the final thickness of the nitride is 4 nm).

Table 1: Process flow

- FIN PATTERNING IN $\langle 110 \rangle$ DIRECTION
ELECTRON BEAM LITHOGRAPHY
RESIST TRIMMING, DRY ETCHING
- SACRIFICIAL OXIDE
- THERMAL TUNNEL OXIDE IN DILUTED O_2
- LPCVD NITRIDE
- THERMAL OXIDATION OF NITRIDE (TOP OXIDE)
- IN-SITU PHOSPHOROUS DOPED 100 NM POLY-Si
- GATE PATTERNING : EBEAM LITHOGRAPHY,
TEOS HARD MASK, DRY ETCHING
- ULTRA THIN SPACERS OF 12 NM
- AS-IMPLANTATION, RTP ANNEAL

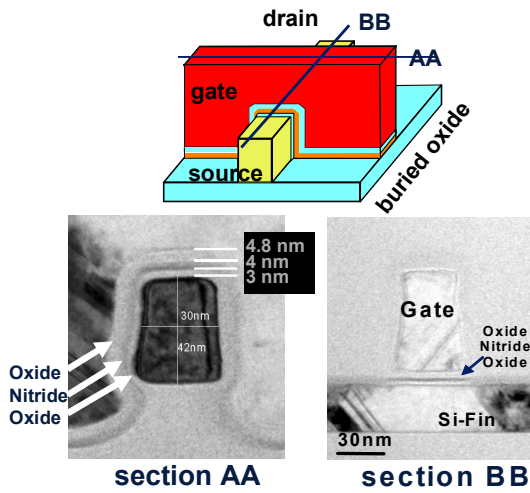


Fig. 1: Schematic FinFet ONO memory cell and two TEM cross sections, $w_{fin}=30$ nm.

Finally, gate poly silicon is defined by a hard mask with e-beam lithography and dry etching. A TEOS spacer of 12 nm is deposited and source and drain regions are As-doped with a dose of $5 \times 10^{15} \text{ cm}^{-2}$ followed by a RTP anneal (Table 1). Fig. 1 shows a schematic FinFET memory with ONO trapping layer and two TEM cross sections.

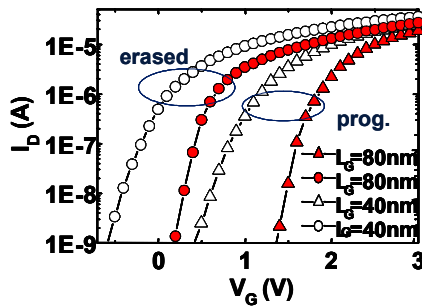


Fig. 2: Transfer characteristics of FinFET ONO transistors with $L_g=80$ and $L_g=40$ nm $w_{fin}=20$ nm after program and erase.

Threshold voltage shifts of $\Delta V_{th} \sim 1$ V are measured after programming with $V_p=11.5$ V at $t_p=3$ msec (Fig. 2). Typical erase parameters are $V_e=-11.5$ V and $t_e=100$ msec. Due to the thick oxides excellent retention properties are obtained (Fig. 3a). The endurance characteristic is given in Fig. 3b, where no degradation is seen up to 10^4 cycles.

These tri gate SONOS cells could be organized in a NAND architecture. The programming disturb is successfully reduced due to the thick tunnel oxides (Fig. 4a). As the memory cell is fabricated on SOI, the erasing procedure in an array is done by successive erasing one word line after the other. For read disturb, the cells are stressed with $V_g=2.5$ V, $V_d=1$ V and no window closure was seen after 10^4 sec (Fig. 4b). Arranging this cell in a NAND architecture 4-5 F^2 cell size can be obtained.

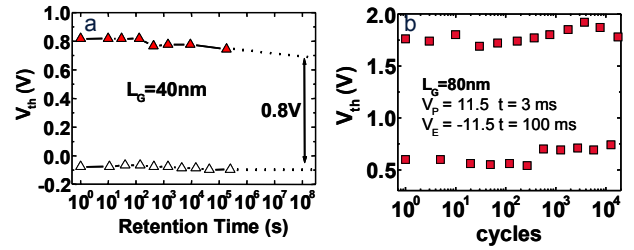


Fig. 3a: Retention characteristic of a $L_g=40$ nm FinFET.
Fig. 3b: Cyclability of $L_g=80$ memory device.

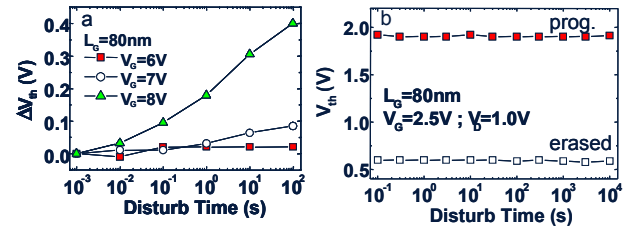


Fig. 4a: Program disturb for a NAND type array. Parasitic shifts up to $V_g=7$ V are acceptably small.
Fig. 4b: Read disturb at $V_g=2.5$ V and $V_d=1$ V.

3. SONOS FinFET with multi level scheme

The FinFETs with extremely scaled fin thickness are fabricated in a similar way as above [6]. The ONO dimensions are $t_{ox} \approx 3.2$ nm, $t_{nitride} = 6.5$ nm $t_{top} = 5.5$ nm. The poly gate is in situ phosphorus doped. The TEM cross section in Fig. 5 shows an ultra thin fin, with parts below 5 nm. Note, the fin is nearly completely surrounded by the gate, so this device works as a gate all around memory transistor. With these very small fins the gate length can be scaled down to 20 nm without degrading the transistor characteristics like subthreshold slope (Fig. 6).

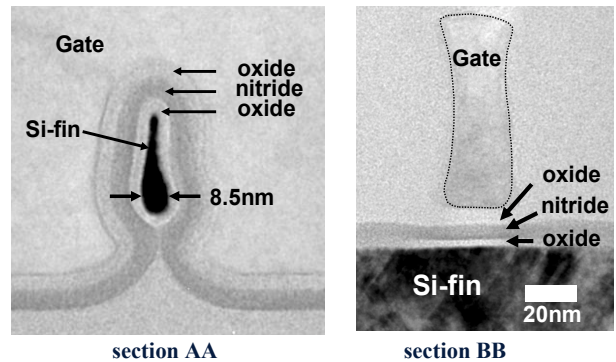


Fig. 5: TEM cross sections for FinFETs with ultra thin fin width $w_{fin} < 8.5$ nm, $L_g=20$ nm. The sections are defined in Fig. 1.

For these ONO-transistors a threshold shift up to $\Delta V_{th} \sim 4$ V is observed (Fig. 6). So this FinFET memory transistor can be operated in a multi level scheme.

Programming parameters are $V_p=12.5V$, $t_p=2$ msec and for erasing $V_e=-11.5V$, $t_e=2$ msec. In Fig. 6 the transfer characteristics for a 4 level storage device are shown. A separation of $\Delta V_{th} \sim 4V$ for the highest and lowest level is achieved. Intrinsically these small devices give only a small saturation current, at a maximum of $1\mu A/\text{device}$. The retention time of the four levels is given in Fig. 7a. Excellent retention properties with an extrapolated $\Delta V_{th}=1V$ between the levels at 10 years are observed.

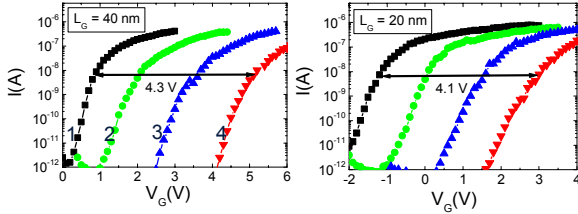


Fig. 6: Multi-level transfer characteristics for a FinFET memory transistor with $L_g=40$ nm (left) and $L_g=20$ nm (right) at $V_D=1V$.

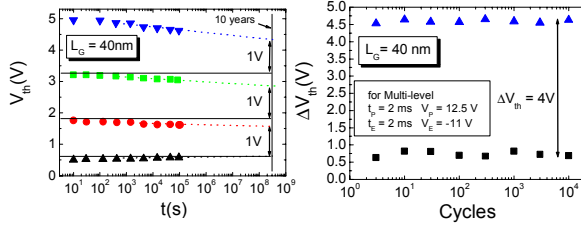


Fig. 7a (left): Retention for a multi level ONO device with $L_g=40$ nm.

Fig. 7b (right): Cycle characteristics of the device programmed into the highest and erased into the lowest level.

The cyclability measurements are given in Fig. 7b. Here the device is programmed into the highest level and erased into the lowest level. A large shift of $\Delta V_{th} = 4V$ is achieved and no window closure up to 10^4 cycles is observed. The number of electrons stored in the trapping layer could be calculated (Fig. 8). These devices are ultra small and the number of electrons stored is in the order of a few 100. In Fig. 9 detrapping of a single electron is visible. If one electron is tunneling back to the silicon channel a V_{th} shift of about 5 mV is measured. Fig. 9 shows the voltage shift in the sub threshold region of a FinFET transistor with $L_g=80nm$. A multi level scheme in a NAND architecture gives an effective cell size of $2 \times 2.5 F^2$ and the bit storage can be doubled. However for these large shift devices good control of read and write disturb requires careful optimization of the ONO dielectric.

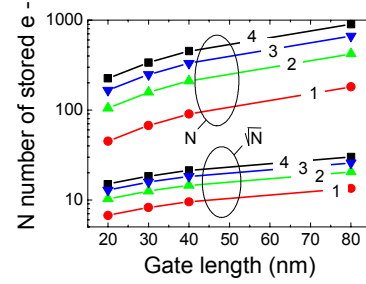


Fig. 8: Number of electrons stored in the ONO trapping layer and the statistical spread $\sqrt{N}$. The levels are labeled according to Fig. 6.

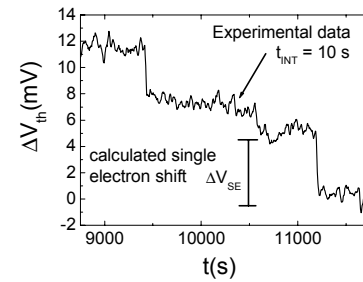


Fig. 9: Measured single electron discharge at room temperature in a device with $L_g=80$ nm. The calculated and measured shift is 5 mV.

4. Dual bit type FinFET memory

The FinFET devices with an ONO stack can also be operated in the dual bit mode with a two bit storage in a portion of the trapping layer at source and drain [7]. The fin is surrounded on three sides by an ONO trapping layer and by the word line, so also the position of the trapped charges are on three sides of the fin at source and drain (Fig. 10). The dimension of the trapping layer are $t_{ox}=3$ nm, $t_{Nitride}=4$ nm, $t_{top}=5$ nm. The tri gate FinFET transistors are programmed, erased and read in analogy to planar NROM (Fig. 11). The programming is done by channel hot electron injection at $V_g=8V$, $V_d=3.7V$ and $t_p=10\mu sec$. Some hot electrons are injected at the drain side near the pn junction from the three sides of the silicon fin into the adjacent nitride layer. A threshold shift of $\Delta V_{th} \sim 1.3V$ is achieved for FinFETs with a moderate channel doping of $3 \times 10^{17} \text{ cm}^{-3}$. At this doping concentration, the devices with a fin width of $w_{fin} \approx 25$ nm are fully depleted. The gate is in-situ doped with boron during the deposition process. Reading of bit one requires a drain voltage of about 1.5V and is done in the reverse mode, where source and drain definitions are interchanged.

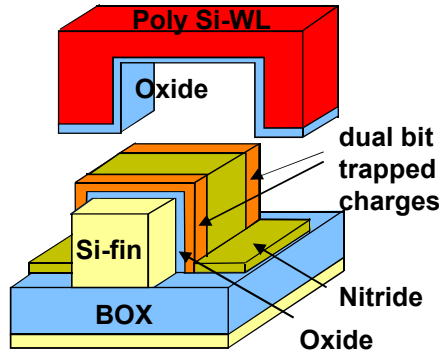


Fig. 10: Schematic FinFET memory device in the dual bit mode.

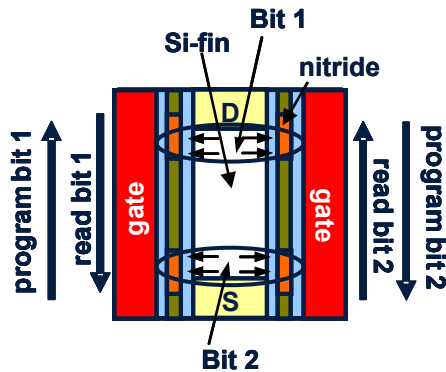


Fig. 11: Top view of a FinFET memory device in the dual bit mode.

Fig. 12 shows the input characteristics of a device with $L_g=100$ nm, $w_{fin}=30$ nm in the erased and programmed state. The parasitic shift of bit 2 is smaller than 20% (Fig.13). Bit 1 is erased by band to band tunnel enhanced hot hole injection at $V_g=-5.5$ V, $V_d=3.25$ V, $t_e=1$ msec. Bit 2 could be erased separately in a reverse mode. The retention characteristic is shown in Fig. 14. NROM like two bit storage is another scheme for high density memories with FinFETs.

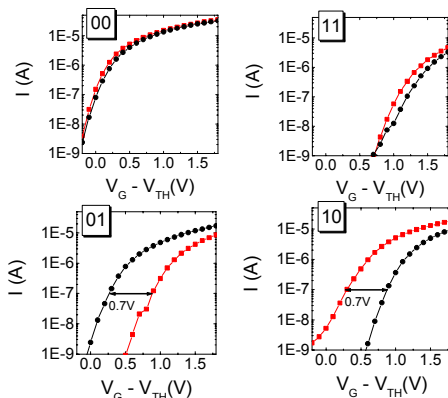


Fig. 12: Transfer characteristics of the two bits for a transistor with $L_g=100$ nm; $V_{SD}=1.5$ V

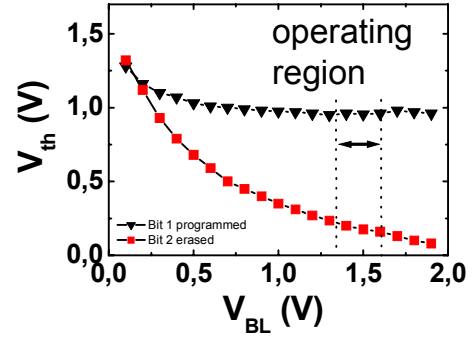


Fig. 13: Dependence of the threshold voltage as a function of the read voltage (V_{SD}) for a programmed bit 1. The two read conditions are defined in Fig. 11.

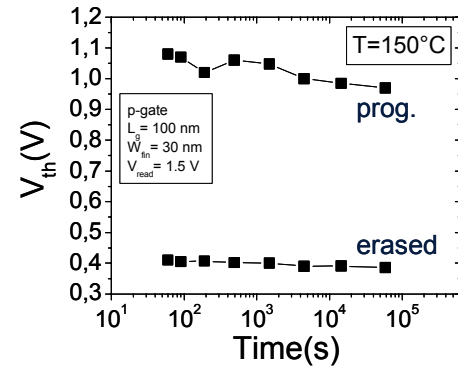


Fig. 14: Retention time for bit 1 (prog.) and bit 2 (erased) at $T=150^\circ\text{C}$, $L_g=100$ nm.

Conclusion

We have demonstrated experimentally that FinFET SONOS memory devices can be scaled down to sub 10 nm fin width and 20 nm gate length. SONOS devices can be reliably operated in single or multi level modes. In a dual bit approach, the FinFET memory can be scaled down to about 100 nm gate length. So SONOS FinFET memory cell are proposed for ultra high density data storage.

References

- [1] J. Park, S. Hur, J. Lee, J. Sel, J. Kim, S. Song, S. Son, Y. Kim, S. Cha, J. Choi, U. Chung, J. Moon, B. Ryu, IEDM Tech. Dig. (2004).
- [2] S. Lai, MIT Standford-UC Berkeley Nanotechnology Forum (2004).
- [3] D. Hisamoto, L. Wen, J. Kedzierski, H. Takeuchi, K. Asano, C. Kuo, E. Anderson, T. King, J. Bokor, C. Hu, Transaction on Electron Devices vol. 47, p2320 (2000).
- [4] P. Xuan, M. She, B. Harteneck, A. Liddle, J. Bokor, T. King, IEDM Tech. Dig. (2003).
- [5] M. Specht, R. Kömmling, L. Dreeskornfeld, W. Weber, F. Hofmann, J. Kretz, E. Landgraf, T. Schulz, J. Hartwich, W. Rösner, M. Städele, R. Luyken, H. Reisinger, V. Klandzievski, E. Hartmann, L. Risch, VLSI Tech. Dig. p.244, (2004).
- [6] M. Specht, U. Dorda, L. Dreeskornfeld, F. Hofmann, J. Kretz, E. Landgraf, T. Schulz, J. Hartwich, W. Rösner, M. Städele, R. Luyken, H. Reisinger, A. Graham, E. Hartmann, L. Risch, Electron Device Letters, Vol. 25, p810, (2004).
- [7] M. Specht, R. Kömmling, F. Hofmann, V. Klandzievski, W. Weber, J. Kretz, E. Landgraf, T. Schulz, J. Hartwich, W. Rösner, M. Städele, R. Luyken, H. Reisinger, A. Graham, E. Hartmann, L. Risch, Electron Device Letters, Vol. 25, p810, (2004).

Modeling of a Double-Gate FinFlash memory

S. Jacob, B. De Salvo, G. Le Carval and S. Deleonibus

CEA-LETI, 17 rue des Martyrs 38054 Grenoble Cedex 9, France stephanie.jacob@cea.fr

Abstract

In this paper, we propose a thorough theoretical investigation of a memory structure based on the double-gate FinFET architecture, named FinFlash. Three-dimensional simulations of the FinFlash memory are shown. Analytical models which allow us to explain the simulation results are also proposed.

1. Introduction

Conventional floating gate Flash memory devices are going through several limitations [1] and it is widely believed that the standard planar Flash cannot be scaled beyond the 50 nm node [2]. In particular, the scaling of tunnel and control dielectric thickness is limited by concerns for data-retention, yielding low coupling ratios and consequently high operating voltages. Moreover, in NOR cells, the drain turn-on phenomenon limits the channel length, and consequently the cell area scaling. These two factors make future scaled cells suffering of short channel effects and punch-through phenomenon, as well as of small I_{on}/I_{off} ratio, thus reducing the sensing margin.

Today, the use of the FinFET architecture [3] appears as one of the most promising solutions to push further the scaling of Flash memories. Several double-gate memory devices have been proposed in the literature, using as storage node either doped poly-Si [4, 5] or nitride layers [6-10]. These devices show an improved electrostatic control of the channel region, thus reducing the drain-induced barrier lowering and the short channel effects, and also a large drive current, thus improving the memory access time.

In this paper we present three-dimensional simulation results of a double gate poly-Si FinFlash structure that shows improved scaling down capabilities as compared to a traditional planar structure. Moreover, the simulated electrical behaviors will be explained through simple analytical models.

2. Memory structure

The structure used for simulations is described in Fig.1. The proposed device is implemented on a SOI substrate. The buried oxide (BOX) thickness is 50 nm. The channel is a Silicon fin, with a 30nm-thick spacer-shaped poly-Si floating gate on both sides of the channel. The 100 nm thick silicon oxide, on the top of the fin, called “blocking oxide”, suppresses the top-channel parasitic conduction. The tunnel / top dielectric is SiO₂ of thickness $t_{tunnel\ ox} / t_{top\ ox}$.

3. Simulation results

We performed an exhaustive three-dimensional simulation study of the double-gate memory architecture using SILVACO tools (ATHENA, ATLAS 3D) [11].

3.a Threshold voltage simulations

Firstly, a simple analytical formula of the neutral threshold voltage for the double gate memory structure is proposed.

For a double-gate MOSFET (Fig.2), the threshold voltage can be expressed as follows [12, 13, 14]:

$$V_{th} = V_{fb} + 2\phi_f + \frac{qN_A W_{Si} t_{ox}}{2\epsilon_{ox}} \quad (1)$$

with V_{fb} : flatband voltage, ϕ_f : electrons Fermi level, N_A : channel doping, W_{Si} : fin width, ϵ_{ox} : SiO₂ permittivity, t_{ox} : tunnel oxide thickness.

Assuming that the potential drop across the floating gates in the FinFlash device is negligible, it follows that the double gate memory structure, from a threshold voltage point of view, is equivalent to a double gate MOSFET structure with the following gate oxide thickness: $t_{ox} = t_{tunnel\ ox} + t_{top\ ox}$.

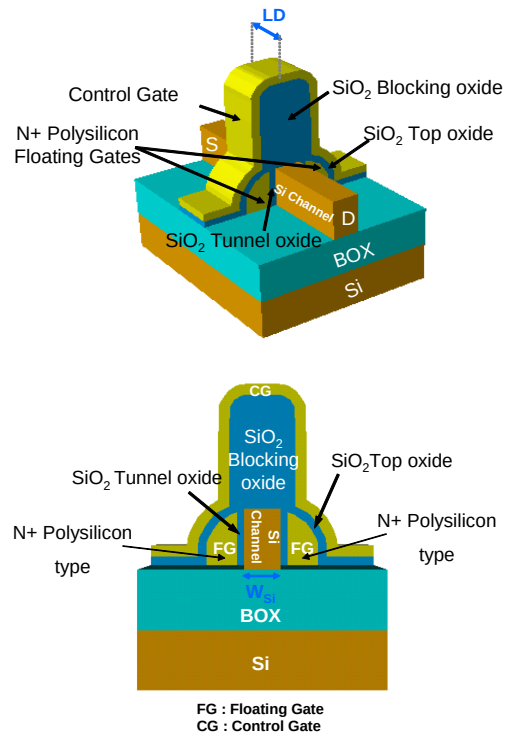


Fig. 1: Views of the simulated FinFlash structure.

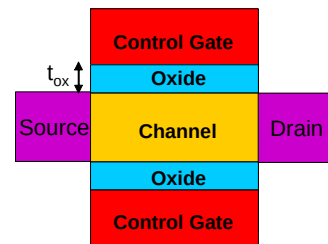


Fig. 2: Cross-section of a double gate MOSFET structure.

Fig.3 shows a comparison between the neutral threshold voltage of a planar conventional memory device with single control gate (SG) and the one of the double gate memory device (DG). It appears that the channel control is really improved with a double gate structure. The “Vt-roll off” phenomenon appears for gate lengths below 70 nm instead of 100 nm as in the case of conventional planar structure.

We studied also the effects of several technological parameters on the neutral threshold voltage of the double gate FinFlash structure.

We can observe in Fig.4 the influence of the channel doping on the FinFlash threshold voltage. Two doping values were simulated (10^{15} cm^{-3} and 10^{18} cm^{-3}). For gate lengths above 70 nm, the neutral threshold voltage saturates to a value which increases with the doping, in agreement with Equation (1).

Fig.5 shows the influence of the control gate material on the FinFlash neutral threshold voltage. We simulated four types of material through their work functions. We see that the threshold voltage is proportional to the gate work function. We can explain this, by writing Equation (1) as follows:

$$V_{th} = \phi_{MC} - \phi_S + 2\phi_f + \frac{qN_A W_{Si} t_{ox}}{2\epsilon_{ox}} \quad (2)$$

being $V_{fb} = \phi_{MC} - \phi_S$, where ϕ_{MC} is the work function of the control gate and ϕ_S is the silicon work function.

We also studied the impact of some geometrical parameters on the neutral threshold voltage of the double gate FinFlash memory.

Fig.6 shows the FinFlash neutral threshold voltage versus the top and tunnel oxide thickness. A linear behavior appears. Once again this result is in good agreement with Equation (1).

Finally, we observe in Fig.7 that the FinFlash neutral threshold voltage is also dependent on the fin thickness (T_{Si}). It appears that the threshold voltage decreases with T_{Si} when the fin is thinner than 70 nm.

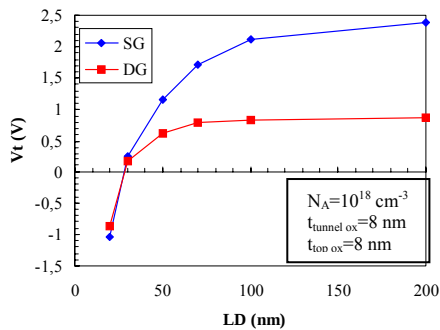


Fig. 3: Neutral threshold voltage (V_t) of a conventional planar memory device with single gate (SG) and of a double gate FinFlash memory (DG) versus the device gate length (LD).

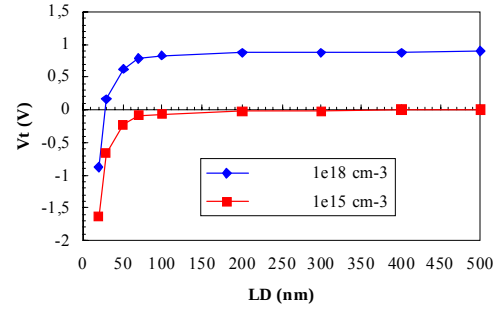


Fig. 4: FinFlash neutral threshold voltage (V_t) versus gate length (LD) for different channel doping.

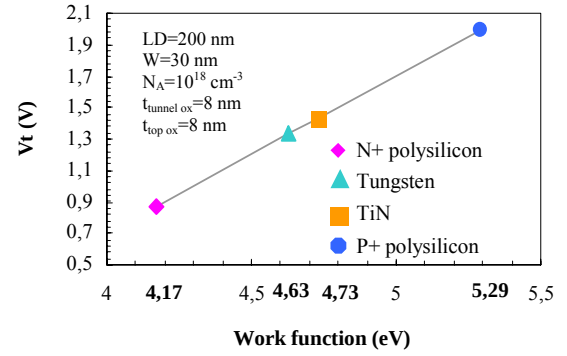


Fig. 5: FinFlash neutral threshold voltage (V_t) versus the work function of the control gate material.

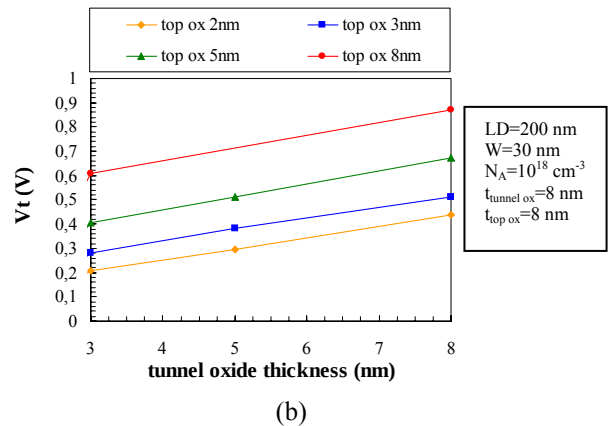
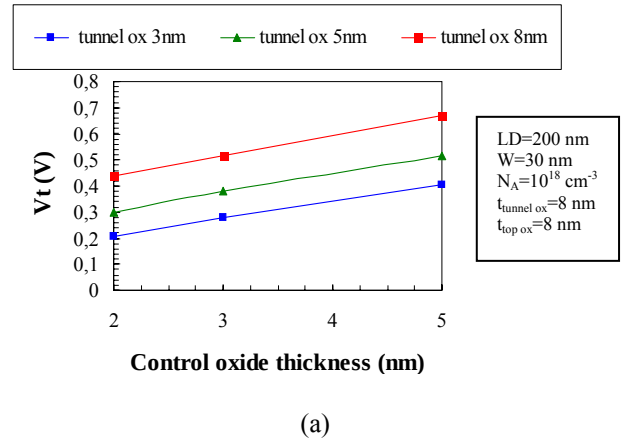


Fig.6: FinFlash neutral threshold voltage (V_t) versus (a) the top oxide and (b) the tunnel oxide thickness.

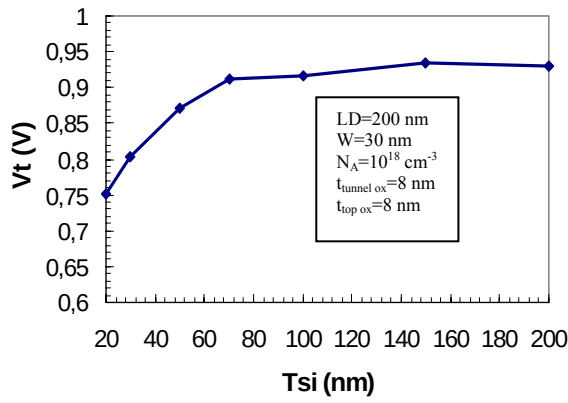


Fig.7: FinFlash neutral threshold voltage (V_t) versus the fin thickness (T_{Si}).

3.b Memory effect simulations

Next simulations concern more particularly the FinFlash memory effects. We observed the impact of a negative charge, stored in the FinFlash floating gates, on the device threshold voltage. We performed these simulations for double gate FinFlash devices with the following characteristics: $L_D=200\text{nm}$, $W_{Si}=30\text{nm}$, $T_{Si}=50\text{nm}$, channel doping $=10^{18}\text{ cm}^{-3}$, $t_{\text{tunnel ox}}=8\text{nm}$, $t_{\text{top ox}}=8\text{nm}$. Both the control gate and the floating gates are n^+ poly-Si.

We clearly see on Fig.8 that a negative charge stored in one or both the floating gates causes the channel potential decrease. Moreover, when only one floating gate is charged, we observe that the minimum of the channel potential is closer to the charged floating gate, whereas the minimum is centered when both floating gates are symmetrically charged.

Fig.9 shows the I_d - V_g characteristics for different values of stored charge density located in one floating gate or in both the floating gates.

The corresponding V_t shift values are reported in Table 1. It clearly appears that the threshold voltage shift is proportional to the stored charge (as shown in Fig.10) and, in the case of both floating gates symmetrically charged, the resulting V_t shift is double.

Finally, our simulations have allowed us also to put in evidence that the two floating gates are not independent. In other words, in Fig.11 we can observe that the presence of a charge in one floating gate has an influence on the potential of the other floating gate. So, we can deduce that there is a coupling factor between the two floating gates.

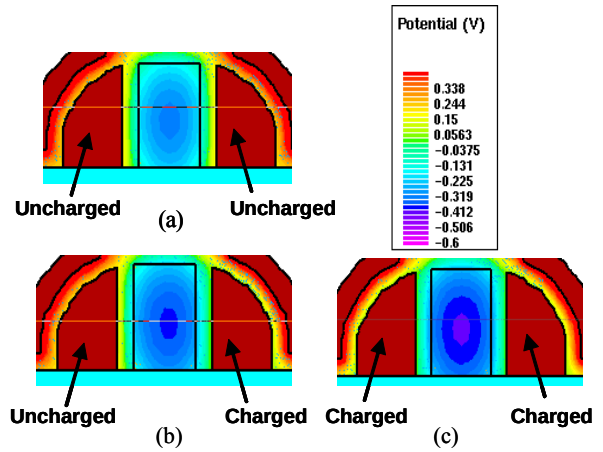


Fig 8: Cross-sectional views of the double gate structure with (a) no charge on both the floating gates, (b) one floating gate charged and (c) both floating gates charged.

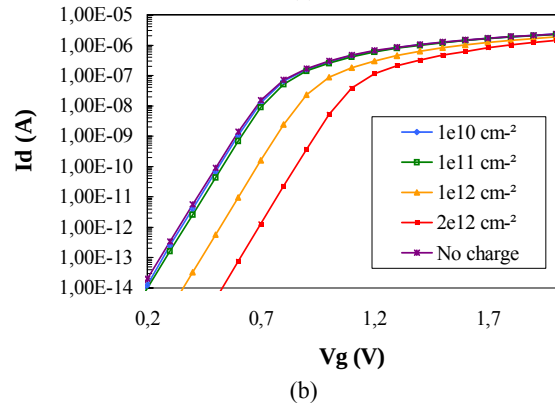
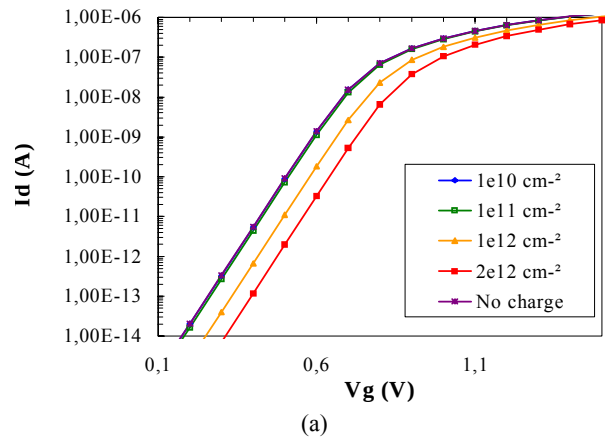


Fig.9: FinFlash I_d - V_g curves when (a) only one floating gate is charged or (b) when both the floating gates are symmetrically charged. The effects of different charge densities are illustrated.

Density (cm^{-2})	Vt shift (mV)	
	One floating gate charged	Both floating gate charged
10^{10}	0.9	8
10^{11}	10	27
10^{12}	100	197
2.10^{12}	197	418

Table 1: FinFlash threshold voltage shift (= charge state V_t – uncharged state V_t) for different values of the charge density.

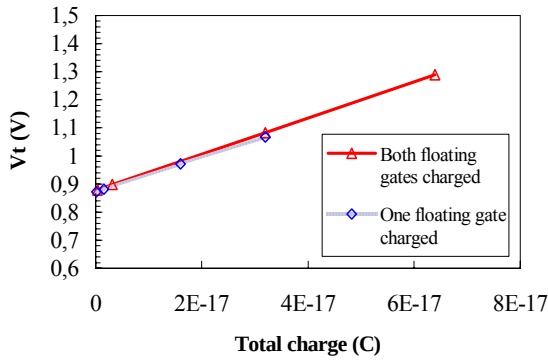


Fig.10: FinFlash threshold voltage versus the total charge trapped in one or symmetrically in both the floating gates.

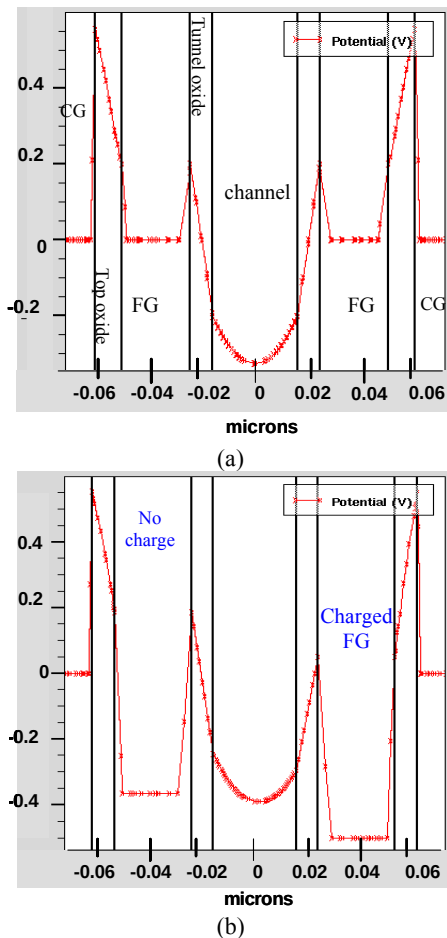


Fig 11: Potential profile (cross-section) of the FinFlash structure with (a) no charge trapped in the floating gates and (b) with one floating gate charged.

4. Conclusion

Double-gate FinFlash memories are good candidates for future generations of NVM devices which can challenge the down scaling problem thanks to the improved channel control. In this paper, we performed three-dimensional simulations of a double gate poly-Si FinFlash memory. Firstly, the effects of several

technological and geometrical parameters on the neutral threshold voltage of the double gate structure have been thoroughly investigated. Secondly, the memory effects (threshold voltage shifts) due to the charging of the floating gates have been clearly studied. Finally, simple analytical equations which can explain the structure behaviors have been also introduced.

Acknowledgements

The author would like to thank Alexandre Ferron from SILVACO-Grenoble for his helpful advices on simulations.

References

- [1] International Technology Roadmap for Semiconductors, 2003 Edition: <http://public.itrs.net/Files/2003ITRS/Home2003.htm>
- [2] "Non Volatile Memories: a look into the future", S. Lai, MIT-Stanford-UC Berkeley Nanotechnology Forum, 26/02/2004.
- [3] "FinFET- A Self-Aligned Double-Gate MOSFET Scalable to 20 nm", D. Hisamoto et al., IEEE Transactions on Electron Devices, Vol. 47, No. 12, 2000.
- [4] "New Silicon-on-Insulator (SOI) Flash memory with side channel and side floating gate", H.Choi et al., Jpn. Journ. Appl. Phys., Vol.42, pp. 3361-3363, 2003.
- [5] "Optimized Cell Structure for FinFET Array Flash Memory", E.S. Cho et al., IEEE ESSDERC 2004, p.289.
- [6] "Nano-Scale SONOS Memory with a Double-Gate MOSFET Structure", I.H. Cho et al., Journal of the Korean Physical Society, Vol. 42, No. 2, pp. 233-236, 2003.
- [7] "FinFET SONOS Flash memory for embedded applications", Peiqi Xuan et al., IEDM 2003, p.609.
- [8] M.Specth et al., VLSI Tech.Dig. 2004.
- [9] « 20 nm tri-gate SONOS memory cells with multi-level operation », M.Specth et al., IEDM 2004, p.1083.
- [10] « Damascene Gate FinFET SONOS Memory Implemented on Bulk Silicon Wafer », Chang Woo Oh et al., IEDM 2004, p.893.
- [11] <http://www.silvaco.com>.
- [12] "Semiconductor Thickness Effects in the Double-Gate SOI MOSFET", B. Majkusiak et al., IEEE Transaction on Electron Devices, vol. 45, No. 5, Mai 1998.
- [13] "Analytical surface potential expression for thin-film double-gate soi MOSFET", K. Suzuki et al., Solid-State Electronics, vol. 37, No. 2, pp. 327-332, 1994.
- [14] "A continuous, Analytic Drain-Current Model for DG MOSFETs", Y. Taur et al., IEEE Electron Device Letters, vol. 25, No. 2, 2004.

3D Simulation Study of Gate and Noise Coupling in Advanced Floating Gate Non Volatile Memories

Andrea Ghetti, Luca Bortesi, Aurelio Mauri, Loris Vendrame

STMicroelectronics, FTM – NVM Technologies Development, Via Olivetti 2, 20041 Agrate Brianza, Italy
E-mail: andrea.ghetti@st.com

Abstract

In this paper we present a 3D simulation study of gate and noise coupling in advanced floating gate non volatile memories. First, the simulation methodology is introduced and validated against experimental data. Then, physical effects that significantly impact the coupling calculation are pointed out. Finally, the method is applied to a sensitivity study of both gate and noise coupling in different non volatile memory architectures showing the increasing importance of this kind of analysis in designing more advanced technologies.

1. Introduction

Floating gate (FG) non volatile memory (NVM) operation is based on the capacitive coupling between FG and all other electrodes (Fig. 1). In this way, by applying an external voltage to the control gate (CG) it is possible to control the FG potential in order to capture electrons during cell writing or emit them during cell erasing. Using a simple capacitive model, the FG potential can be computed as [1]:

$$V_{FG} = \sum_i \alpha_i V_i \quad (1)$$

where V_i and α_i are the i -th electrode potential and coupling coefficient, respectively. The coupling coefficient is computed as [1]:

$$\alpha_i = \frac{C_i}{\sum_j C_j} \quad (2)$$

where C_i is the coupling capacitance between FG and the i -th electrode. The sum in Eq. (1) and (2) runs over all electrodes. Of particular importance is the gate coupling coefficient α_G that controls the fraction of the CG potential transferred to the FG: the higher the better.

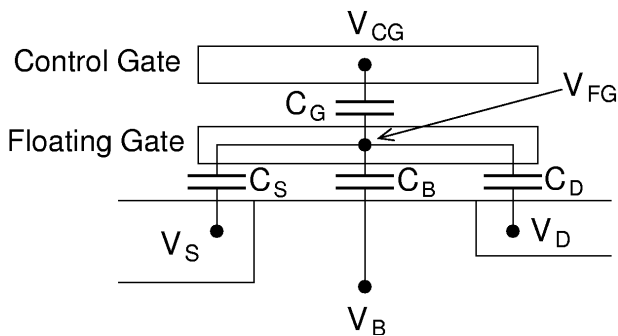


Fig. 1: Schematic cross section of a floating gate transistor showing the intrinsic capacitances.

This capacitive model is valid for a single cell. However, NVM cells are organized in large matrix of cells closely packed, so that each FG can couple also with adjacent FG and, if present, with the drain contact of nearby cells. The coupling with nearby cells has two detrimental effects: 1) it adds additional terms at the denominator of Eq. (2) thus reducing α_G ; 2) it introduces a dependence of the FG potential of a cell on the FG potential of the nearby cells. Clearly, this “interference” depends on the electrical status of nearby cells, i.e. on the programmed pattern. Therefore it is not predictable and represents a kind of disturb or “noise” that must be minimized. For practical purposes, only the noise due to the first neighbour cells has to be considered. Thus, in NAND Flash there are eight disturbing cells (Fig. 2), and the total noise expressed as coupling coefficient is $\alpha_n = 2\alpha_x + 2\alpha_y + 4\alpha_{xy}$. In NOR-type array, there is a contact plug every two cells along the bit line direction that effectively decouples cells on one side from cells on the other side of the contact reducing to five the number of disturbing cells (Fig. 3), thus $\alpha_n = 2\alpha_x + \alpha_y + 2\alpha_{xy}$.

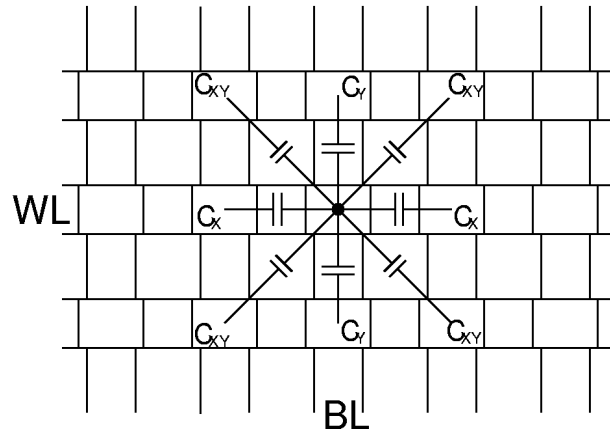


Fig. 2: Schematic drawing of parasitic capacitances among nearby cells in a NAND Flash array.

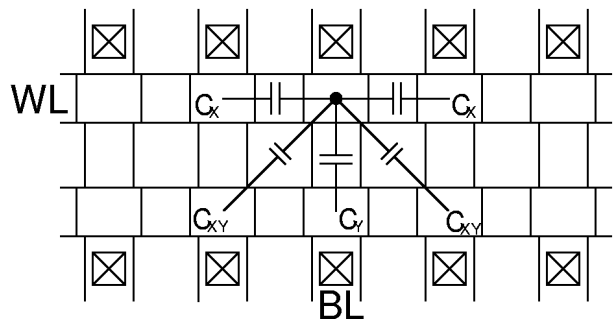


Fig. 3: Schematic drawing of parasitic capacitance among nearby cells in a NOR Flash array.

The coupling noise shows up macroscopically as a shift of the cell threshold voltage (V_{th}) distribution even of hundreds of millivolt only because the programmed pattern has changed [2]. This shift is of great concern, especially for multi level applications where the immunity margins between the different levels are very tight, and must be properly addressed.

The aim of this work is to investigate this kind of problem. Since coupling among nearby cells in a memory array is intrinsically a 3D problem very sensitive to the geometry we first developed a methodology to build the 3D cell structure. Then coupling coefficients are simulated with different methodologies and validated against experimental data. The so validated methodology has been applied to investigate the scaling limit of different cell architectures from the coupling noise point of view.

This paper is arranged as follows. Section 2 describes the simulation methodology and how it has been validated against experimental data. Section 3 would like to point out important physical effects that must be taken into account for this kind of analysis. The proposed methodology is then extensively exploited in Section 3 to perform an analysis of gate and noise coupling scaling in different NVM technologies.

2. Methodology description and validation

A. Experimental

Although coupling coefficients are defined by Eq. (2), this expression is not of practical usage since the single capacitances between FG and all other electrodes are not directly measurable simply because the FG is not accessible. Therefore, it is more convenient to exploit the linear relationship of Eq. (1). For example, concerning α_G , Eq. (1) implies that ratio of the cell gain (β) and the “equivalent transistor” gain (where the FG is shorted to the CG, so it can be biased and act as the gate of a MOS transistor) coincides with α_G [1]. For the same reason, also the ratio of the sub-threshold slopes provides α_G .

A similar strategy based on the NVM cell I-V characteristics can be adopted to experimentally measure the different coupling noise coefficients. As a matter of fact, from Eq. (1) it also derives that

$$\alpha_i = \frac{\Delta V_{th}^{vict}}{\Delta V_{th}^i} \quad (3)$$

where ΔV_{th}^{vict} is the threshold voltage variation of the “victim” cell due to a ΔV_{th}^i threshold voltage variation of the “aggressor” cell produced by a write/erase operation. Practically, V_{th} of the “victim” and “aggressor” cells are measured before and after a write operation on the “aggressor” cell.

B. Simulation methodology

Since simulation tools allow computing the single capacitances, Eq. (2) can be exploited to define the simulated coupling coefficients. One kind of such tools are commercial programs devoted to the extraction of the interconnect R-C parameters [3]. They allow to handle

large piece of the cell matrix, but have limited capability in describing the complex geometry of a NVM cell. Moreover, usually they do not consider transport in the silicon (i.e. silicon is treated as a contact) and do not allow floating contact, preventing to use Eq. (3) that is exploited in reality to measure coupling noise.

To overcome these restrictions, one can use also TCAD tools. They include by construction transport in the silicon substrate, allow floating contact and a better definition of the cell geometry, but at the expenses of a limited number of cells that is possible to simulate because of a larger use of computational resources. With TCAD tools, in particular device simulators, it is possible to perform an AC analysis which provides all the capacitances, and hence to compute all coupling coefficients using Eq. (2). In the following this method will be referred to as simulation Method A. On the other hand, they also allow to compute V_{th} leaving FG really floating, hence allow to use also Eq. (3). In the following this method will be referred to as simulation Method B.

We verified that Method A provides the same capacitances as R-C extractors when applied to the same structure with the same approximations. However, because of a better description of the geometry and the possibility to use floating contacts and Eq. (3) we adopted in this work TCAD tools [4]. This is also needed because, as we will see in the following (cfr. Sec. 3), there are special cases where Method A and Method B do not provide the same results because of an additional effect on the channel transport that can be accounted for only by Method B.

C. Validation

The simulation methodology was validated against experimental data coming from a NOR Flash cell. For best possible accuracy in the geometry description active area and source line cross sections have been imported from TEM pictures (Fig. 4) and extruded in the channel direction according the nominal layout. Simulated 2D boron channel profile of the active area has been extended in the channel direction, while simulated 1D S/D arsenic profile has been extended in the channel direction with a gaussian profile to reproduce the real channel length. The structure has been meshed and then reflected twice to get a 2x2 array of cells (Fig. 5).

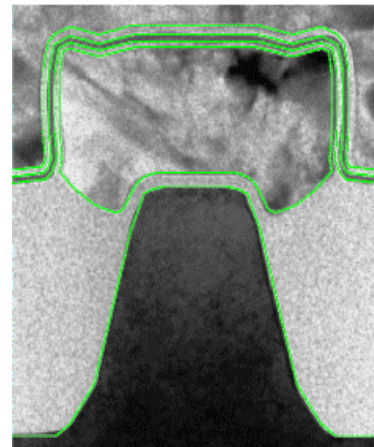


Fig. 4: Cell cross section and imported profile (green line).

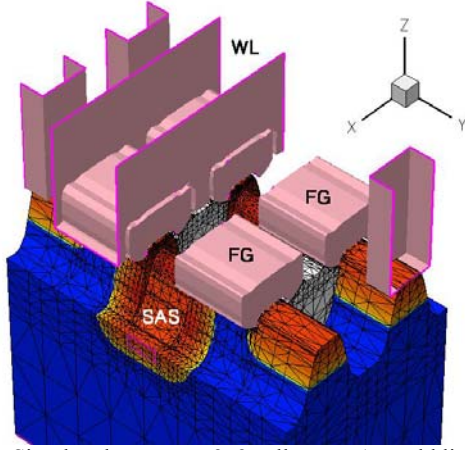


Fig. 5: Simulated structure: 2x2 cell array. A world line, a drain contact and most of the dielectrics are not shown.

Notice the presence of the source line between the two lines of cells. Particular attention has been paid to the inclusion of the correct silicon over etch due to the self-aligned source etch (SAS) since it significantly affects α_y . Without this silicon consumption the source line would be closer to the FG shielding more effectively the cell on the other word line resulting in a lower α_y . Notice also that we tried to generate a finer mesh in the dielectric region between cells for a more accurate calculation of the noise coupling coefficients.

The mini array has been simulated with Method A and B and the results are compared to experimental data in Table 1. The agreement is very good especially because of the good accuracy with which the geometry has been described. Notice that, for this particular geometry, the two simulation methods give the same results. However, this is not always the case as discussed in the next Section.

	Exp.	Method A	Method B
α_G	0.65	0.662	
α_x	0.003	0.0031	0.0032
α_y	0.0078	0.0078	0.0078
α_{xy}	0.00082	0.0007	0.0011
Total noise : $\alpha_n = 2\alpha_x + \alpha_y + 2\alpha_{xy}$	0.01544	0.0154	0.0164

Table 1: Comparison of experimental and simulated values of coupling coefficients.

3. Direct FG-to-Channel coupling

Figure 6 shows a schematic cross section of the self-aligned STI (SASTI) architecture [5]. First tunnel oxide and FG are formed. Then, the entire gate stack and STI trenches are etched in order to get a FG self aligned to the active area. However, in this scheme, the top corners of the active area are not shielded by the FG and, therefore, the FG of the nearby cell in the word line direction (x direction) can directly affect the transport in the channel. As a matter of fact, when the “aggressor” cell is programmed its FG potential goes below the “victim” FG potential. In this way, the “victim” FG is pulled down because of the standard capacitive coupling. Moreover, the surface potential of the top corner can be additionally reduced by the negative potential of the

nearby cell, thus causing a reduction of the current on the side of the channel facing the “aggressor” cell.

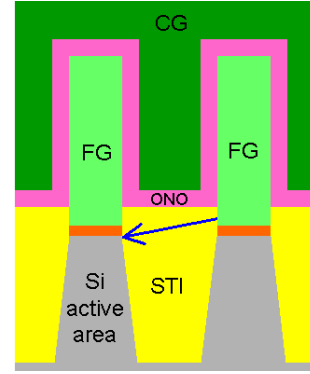


Fig. 6: Schematic cross section along the word line of the self aligned STI architecture.

This is demonstrated in Fig. 7 showing the reduction of the surface concentration of electrons due to the standard capacitive coupling and the direct FG-to-channel interference. The standard capacitive coupling reduces the surface concentration symmetrically, while the direct FG-to-channel interference reduces electron concentration only on the side of the “aggressor” cell.

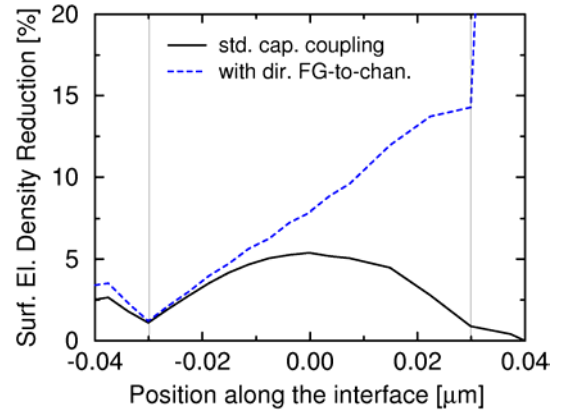


Fig. 7: Relative reduction of the surface electron concentration (proportional to the current) along the silicon/STI interface. The two vertical lines mark the top part of the silicon.

Since V_{th} is usually defined at a particular current level, the reduction of the current due to the direct FG-to-channel interference translates in an increase of V_{th} of the “victim” cell, hence in an increased α according to Eq. (3). This effect is quantified in Table 2 and can have a significant impact on the total noise coupling. Notice that Method A and Method B (the only accounting for this direct FG-to-channel coupling) differ only when the “aggressor” is on the nearby bit line (x direction), i.e. the “aggressor” faces the side of the channel.

	Method A	Method B
α_x	0.0036	0.0128
α_y	0.0080	0.0078
α_{xy}	0.0005	0.0007
Total noise : $\alpha_n = 2\alpha_x + \alpha_y + 2\alpha_{xy}$	0.0162	0.0348

Table 2: Comparison of coupling noise provided by the two methods on the SASTI architecture pointing out the significant effect of the direct FG-to-channel interference.

4. Sensitivity analysis

After having demonstrated that this simulation methodology based on 3D device simulation is accurate enough, we applied it to the study of the coupling coefficient scaling in different types of NVM cells (NAND and many variants of NOR) belonging to next generation technology node.

As an example, we report here the simulation results relative to the NOR cell architecture schematically shown in Fig. 8 [6]. In this scheme, the FG is made of two parts: the first is self aligned with the STI (P1) while the second (P2) is defined lithographically. The structure is defined by many geometrical parameters. A complete DOE would require too long time. Therefore we report here a kind of sensitivity analysis where parameters are changed one at the time around a nominal point.

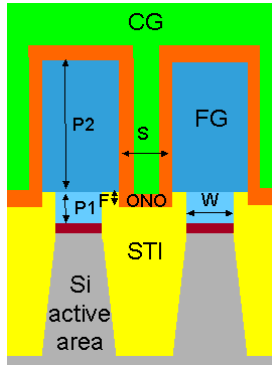


Fig. 8: Schematic cross section along the word line of the analyzed structure. Some of the geometrical parameters defining the structures are also shown.

Since silicon samples are not yet available for the advanced technologies this architecture is aimed to (the value of this work is precisely that it allows to reliably get fundamental parameters for NVM cell such as the coupling coefficients well in advance of costly silicon samples), we generated the simulated structure using a new 3D process emulator tool [7]. This tool allows to build a 3D structure exploiting solid modeling operations and simple process emulation steps. It also has a scripting interface that allows batch runs to automate the entire simulation flow. Once the geometry is generated in this way, it is meshed and then reflected twice in order to get the 2x2 cell array similar to Fig. 5, but with the cross section of Fig. 8, which is then used in the simulation according to both Method A and B.

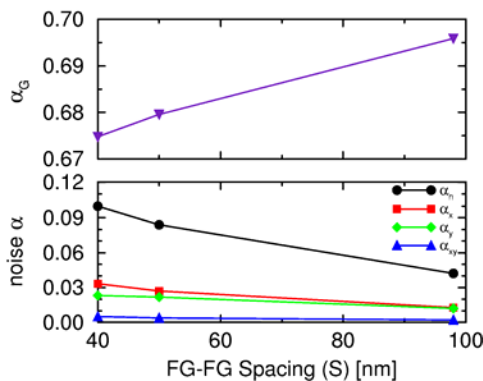


Fig. 9: Simulated coupling coefficients as a function of the FG-FG spacing S.

Figure 9 reports the simulation results as a function of the FG-FG spacing (S). As it can be seen, S has a strong impact on the coupling coefficients. In particular the total noise α_n is strongly reduced because so is α_x that counts twice. α_G increases because the denominator of Eq. (2) decreases.

Figure 10 reports the simulation results as a function of the height of the self aligned part of the FG (P1). Increasing P1 α_x and α_y increase because of the increasing facing area, while α_G decreases for the increase of the total capacitance (not shown). Moreover, the difference between Method B and Method A, that is the contribution of the only direct FG-to-channel coupling, saturates since this effect is completely activated as soon as the top corner of the silicon active area is “visible” from the other FG (Fig. 10 top frame).

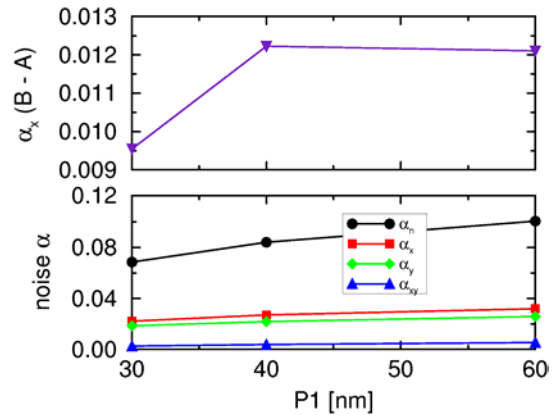


Fig. 10: Bottom frame: simulated noise coupling coefficients as a function P1. Top frame: contribution to α_x of only the direct FG-to-channel coupling (i.e. Method B – Method A)

5. Conclusions

The reduction of cell physical dimensions implies an increase of the coupling noise that is a major concern for NVM scaling, especially for multi level applications. We introduced and validated a simulation methodology based on 3D TCAD tools that accurately predicts coupling coefficients. This methodology is then a valuable instrument to investigate NVM cell architectures for future generation technologies from the coupling coefficients point of view well before silicon samples are available.

Acknowledgment

The authors would like to thank T. Ghilardi, M. Bonanomi and L. Chiavarone of STMicroelectronics for providing experimental data.

References

- [1] P. Pavan and R. Bez, in “Flash Memories”, P. Cappelletti, C. Golla, P. Olivo and E. Zanon Eds., Kluwer, Norwell, 1999.
- [2] C. Lee, H. Lee, S. Park, C. Park, K. Kim and K. Kim, *Proc. of Non Volatile Semiconductor Memory Workshop*, 2004.
- [3] Silvaco, Santa Clara, *CLEVER manual*, ver. 2.0, 1999.
- [4] ISE, Zurich, *DESSIS manual*, rel. 9.5, 2004.
- [5] M. Ichige *et al.*, VLSI Tech. Symp., 2003.
- [6] Y. Song *et al.*, VLSI Tech. Symp., 2003.
- [7] ISE, Zurich, *DEVISE manual*, rel. 9.5, 2004.

Cost-competitive embedded non-volatile technology for Flash and EEPROM applications

Do Dormans^a, Dick Boter^a, Jerome Dubois^a and Jochen Garbe^b

^a Philips Semiconductors, ICN, Gerstweg 2, 6534 AE Nijmegen, The Netherlands

^b Philips Semiconductors, BL-Identification, Georg Heyken Strasse 1, 21147 Hamburg, Germany
Email: Do.Dormans @ philips.com

Abstract

We have developed a next generation embedded flash/EEPROM technology that is cost-competitive with the industry 0.13 μm technology. Cell sizes scale well with the technology. Improved endurance performance has been achieved by introducing nitrided tunnel oxide. The high performance is shown both for EEPROM and flash products that are manufactured in the same generic process.

1. Introduction

Philips Semiconductors has introduced an embedded non-volatile memory technology based on a 2-transistor flash cell (Fig. 1) operated by uniform FN tunneling both for program and erase [1]. Though the cell size is relatively large, module sizes of memories with up to 16Mb flash are smaller than modules based on a 1-transistor flash cell. The 2T-FNFN concept enables operation at low read (1.2V) and write voltages (1.5V) [2]. The same basic cell and technology are used to build a byte-erasable EEPROM memory [3] thus enabling the unique combination of the two non-volatile options on one chip. The 2T-FNFN concept was introduced in a 0.18 μm CMOS technology.

With the continuous pressure from the market to reduce cost we have ported this technology to a next generation that is cost-competitive with the industry 0.13 μm technology node. Both flash and EEPROM scale well with the technology and the process has been optimised such that two of the non-volatile specific masking steps can be skipped.

2. Technology

The flash/EEPROM option is embedded in a high performance 5 or 6 metal, dual gate oxide, dual Vt CMOS process. The process architecture (Table 1) is chosen to minimize the effect of the additional flash/EEPROM processing steps on logic. Thus, CMOS baseline performance is maintained and all logic and analogue IP developed for the CMOS process can be used as such in flash/EEPROM products.

In the industry, the 0.13 μm node is a 1.2V technology using 193nm lithography and a dual damascene copper back-end. In our next generation embedded flash/EEPROM technology design rules have been scaled to the minimum without needing 193 nm lithography. A conventional aluminium / low-k back-end is used and processing is done on 8" wafers. Logic oxide thicknesses

have been chosen to comply with external supply voltages up to 1.8V. All these measures avoid an increase of wafer cost.

Both flash and EEPROM cell dimensions have been scaled significantly (Table 2). Drain engineering of the 2T cell was done to allow a reduction of the channel length without increasing the off-current of the cell or punch-through during write. The inter-poly dielectric (ONO) thickness has been scaled to an electric oxide thickness of 13.5nm to maintain the coupling coefficient of the cells. A cross section of the cell is shown in figure 2. For the tunnel oxide we have introduced a post-nitridation to extend the endurance performance to typically 100k for flash and 1M for EEPROM.

3. Results

Program and erase times for flash are 1ms and 100ms, respectively. For EEPROM a program cycle (including erase) takes 2.7ms. Operating voltages are $\pm 15\text{V}$ for both devices. Programming and erasing is done with single pulses, i.e. without any verification. Threshold voltage distributions (Fig. 3) are comparable to 0.18 μm cell distributions. Endurance characteristics for conventional and nitrided tunnel oxide are shown in figure 4. The nitrided tunnel oxide suppresses interface state generation thus extending the endurance by almost one order of magnitude to well over 1M cycles.

This is further exemplified in figure 5 showing the margin threshold voltage in a 72kB EEPROM product at 95°C demonstrating an endurance at product level >500k cycles.

The high quality of the process is also demonstrated in figure 6 showing the bit fail fraction of flash cells (measured on 256 16Mb Flash modules at 85°C) as function of the number of program/erase cycles. Programming is done in checkerboard mode causing a small fraction of erratic bits (fast programming cells). This bit-fail fraction, however, does not increase with the number of program/erase cycles and is well below the level that can be corrected for by the error correction, even in the margin modes.

References

- [1] T. Ditewig et al., ISSCC 2001, p. 2.4.
- [2] D. Dormans et al., SSDM 2001, p. 540.
- [3] D. Dormans et al., NVSMW 2003, p. 21.

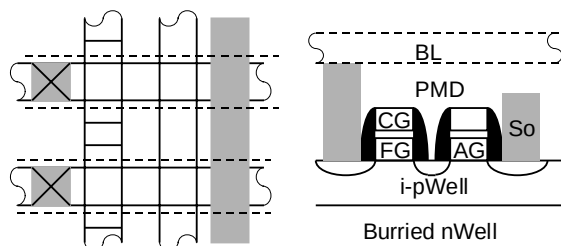


Figure 1: Layout of 2T-Flash/EEPROM cell.

• 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
• Pre-metal-dielectric & local interconnect formation
• 5- or 6-layer aluminium / low-k back-end

Table 1: Process architecture.

Floating gate transistor	W/L = 0.192 μm / 0.176 μm
Access transistor	W/L = 0.192 μm / 0.208 μm
Flash cell size*	0.50 μm^2
EEPROM cell size*	0.86 μm^2
Tunnel oxide	8.5 nm (nitrided)
Inter-poly dielectric	ONO 6/5/5 nm

* Cell size without overhead.

Table 2: Cell dimensions.

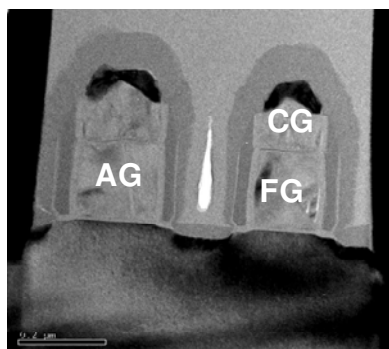


Figure 2: TEM cross-section of the 2T cell.

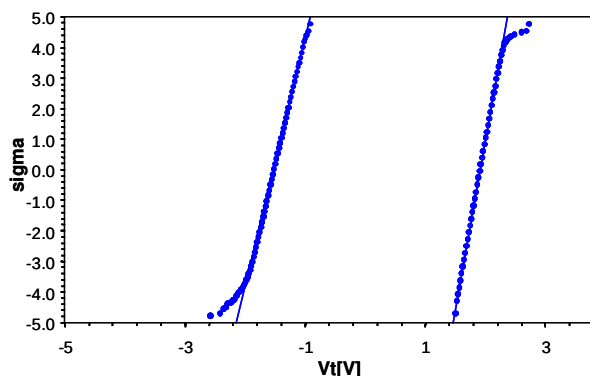


Figure 3 Cumulative threshold voltage distribution of flash cells in a 2.7Mb test array after a single pulse write and erase.

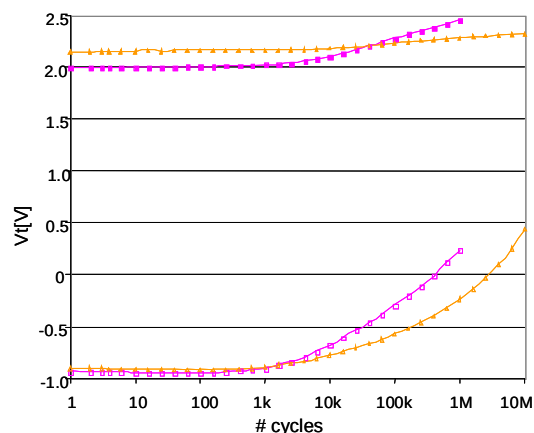


Figure 4: Endurance of EEPROM cells with conventional (pink symbols) and nitrided (orange symbols) tunnel oxide.

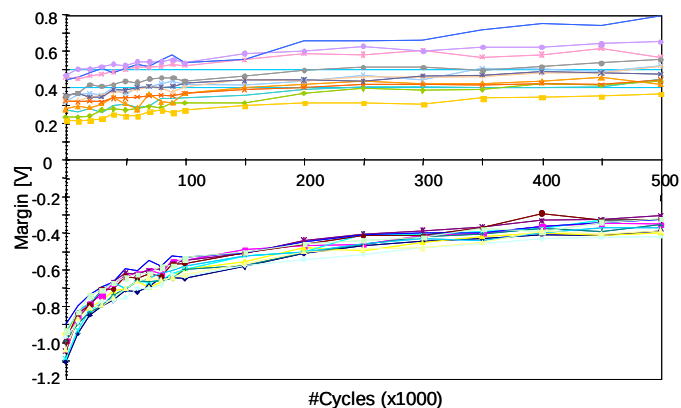


Figure 5: Margin voltage of programmed and erased cells as measured on a 72kB EEPROM product at 95°C. Different lines refer to different dies.

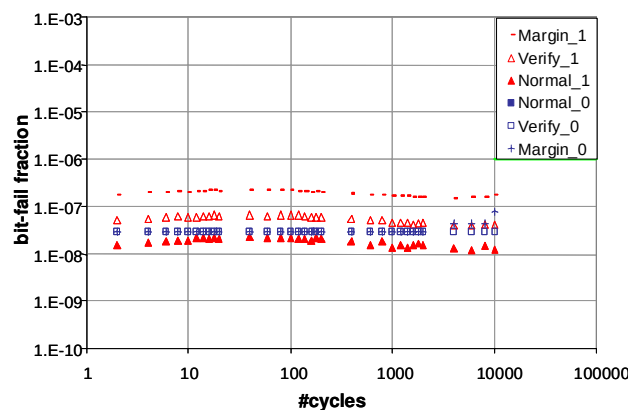


Figure 6: Bit-fail fraction of programmed (blue) and erased (red) cells measured at 85°C on 256 16Mb Flash products. The bit fail fraction is measured in application mode ("normal") and two margin modes ("verify" and "margin").

Composition control and ferroelectric properties of sidewall $\text{Sr}_{0.8}\text{Bi}_{2.2}\text{Ta}_2\text{O}_9$ in integrated 3-Dimensional ferroelectric capacitors

L. Goux^a, J.G. Lisoni^a, M. Schwitters^a, V. Paraschiv^a, D. Maes^a, L. Haspeslagh^a, D.J. Wouters^a, N. Menou^b, Ch. Turquat^b, V. Madigou^b, Ch. Muller^b and R. Zambrano^c

^a IMEC, Kapeldreef 75, B-3001 Leuven, Belgium, gouxl@imec.be

^b L2MP, UMR CNRS 6137, Université de Toulon et du Var, BP 132, F-83957 La Garde Cedex, France

^c STMICROelectronics, MPG Devt Center Catania, Stradale Primosole, 95121 Catania-Italy

Abstract

The development of 3-Dimensional Ferroelectric Capacitors (3-D FeCAP's) is required for scaling FeRAM along with the CMOS technology roadmap. A critical issue in this 3-D is the control of the material quality of the ferroelectric layer at the capacitor sidewall. In this work, we studied the specific electrical behavior of 3-D FeCAP's with $\text{Sr}_{0.8}\text{Bi}_{2.2}\text{Ta}_2\text{O}_9$ (SBT) ferroelectric thin film deposited by MOCVD. Poor ferroelectric behavior of the material at the sidewalls could be related to bismuth segregation during the MOCVD deposition. However, after process optimization, bismuth segregation could be minimized and the ferroelectric polarization value of the sidewall SBT is almost equal to that of the planar SBT film.

1. Introduction

Ferroelectric Random Access Memory (FeRAM) has for long time been considered as promising candidate to become the ideal memory because it combines very attractive properties such as low-voltage and low-power operation, fast write, non-volatility, and high cycling endurance. However, FeRAM is difficult to scale along with the CMOS technology roadmap, and current state-of-the-art 1-64 Mb FeRAM developments [1-4] in 0.25 μm down to 0.13 μm technologies are reaching signal window limitations resulting in large cell size [5]. One solution to bridge this gap would be to evolve from 2-Dimensional (2-D) to 3-Dimensional (3-D) Ferroelectric CAPacitors (FeCAP's), in order to gain polarization signal from the contribution of the FeCAP sidewalls (SW's) [5]. For such 3-D configurations, one of the main concerns is related to the electrical characteristics of the FeCAP SW's, as reports on good control of both electrode and ferroelectric material properties in the SW are still missing in the literature.

In this work, we have integrated 2-D and 3-D $\text{Sr}_{0.8}\text{Bi}_{2.2}\text{Ta}_2\text{O}_9$ (SBT)-based FeCAP structures in a 0.35 μm CMOS technology up to metal interconnection. We show that the ferroelectric contribution of the sidewalls depends on the MOCVD (Metal Organic Chemical Vapor Deposition) deposition temperature of SBT, and we relate this contribution to temperature-dependent Bi-segregation process in the SW's.

2. Experimental

The integrated FeCAP's were realized as follows. We deposited a 350 nm thick bottom electrode (BE) stack of Pt/IrO₂/Ir/Ti(Al)N on silicon. The BE stack was

patterned in arrays of stripes. Then 130 nm amorphous SBT was deposited by MOCVD at low temperature to ensure good step coverage. The SBT layer was then crystallized at 700°C, 60min. in oxygen. Pt TE was deposited and patterned in stripes, leading to the realization of 2-D configurations (Fig. 1a) and 3-D configurations (Fig. 1b). Further integration steps included H-sealing, isolation and W-via formation to metal interconnect level.

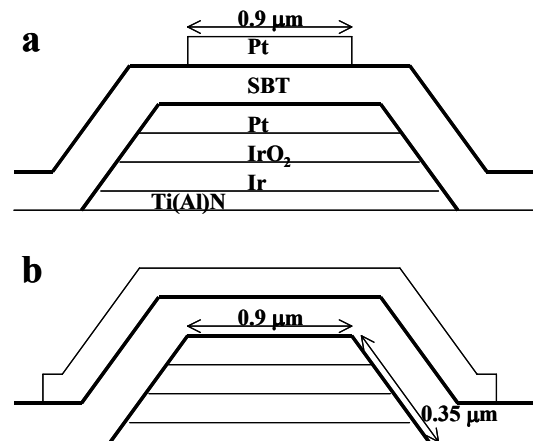


Fig. 1: Schematic of the cross-section taken along the width of $0.9 \times 100 \mu\text{m}^2$ stripe-type FeCAP's, for 2-D (a) and 3-D configuration (b).

The design area of a stripe is $1 \times 100 \mu\text{m}^2$. From SEM cross-section images of the actual integrated structures, we estimated that the effective area of the 2-D FeCAP stripe is $\sim 0.9 \times 100 \mu\text{m}^2$ whereas the effective area of the 3-D stripe is $\sim 1.6 \times 100 \mu\text{m}^2$ because of the contribution of SW's (Fig. 1b). From hysteresis loops measured at 1 kHz after division by the design area of $1000 \mu\text{m}^2$ (corrections using real effective area are given where appropriate), we extracted the remnant polarization (P_r) of 2-D and 3-D FeCAP stripe arrays.

3. Influence of the MOCVD deposition temperature on the ferroelectric properties of SBT sidewalls

The liquid precursors used for SBT MOCVD were $\text{Bi}(\text{thd})_3$, $\text{Sr}(\text{thd})_2\text{pmdeta}$, and $\text{Ta}(\text{O}-i\text{Pr})_4(\text{thd})$. The growth rate of SBT was observed to increase linearly with the deposition temperature (T_{dep}) from 405°C to 430°C, suggesting a reaction-rate limited deposition regime [6]. However the growth rate saturates from 430 to 440°C, which in turn suggests a change to a mass-transport limited regime [6]. Hence, the influence of T_{dep}

on the Pr value measured after integration for both 2-D and 3-D stripe arrays was investigated for $T_{\text{dep}} = 405^\circ\text{C}$ and $T_{\text{dep}} = 440^\circ\text{C}$.

Table 1 shows Pr values obtained for 3 different samples, namely SBT440 ($T_{\text{dep}} = 440^\circ\text{C}$), sample SBT405 ($T_{\text{dep}} = 405^\circ\text{C}$), and sample SBT405opt ($T_{\text{dep}} = 405^\circ\text{C}$, using optimized process conditions). Table 1 also reports the parameter ' $SW\_efficiency$ ', corresponding to the efficiency of the SW SBT in contributing to the Pr of the 3-D FeCAP. This parameter is defined as follows:

$$SW\_efficiency = 100 \times \frac{Pr_{3-D} - Pr_{2-D}}{(1.6/0.9)Pr_{2-D} - Pr_{2-D}} \quad (1)$$

The value of $(1.6/0.9)Pr_{2-D}$ corresponds to the expected Pr_{3-D} considering the increase in effective area from the 2-D to the 3-D configuration. $SW\_efficiency$ is 0% for $T_{\text{dep}} = 440^\circ\text{C}$, which means that the sidewall region is not ferroelectric and thus that the nature of the SBT material significantly varies from the planar region to the sidewall region. However, $SW\_efficiency$ increases up to 50% for $T_{\text{dep}} = 405^\circ\text{C}$, and up to 88% after further optimization at $T_{\text{dep}} = 405^\circ\text{C}$ (Table 1).

	Pr_{2-D} ($\mu\text{C}/\text{cm}^2$)	Pr_{3-D} ($\mu\text{C}/\text{cm}^2$)	$SW\_efficiency$
SBT440	2.6	2.6	0%
SBT405	4.5	6.3	50%
SBT405opt	4.5	7.5	88%

Table 1: Characteristics of 2-D and 3-D FeCAP's for samples SBT440 ($T_{\text{dep}} = 440^\circ\text{C}$), SBT405 ($T_{\text{dep}} = 405^\circ\text{C}$), and SBT405opt (optimized MOCVD at $T_{\text{dep}} = 405^\circ\text{C}$).

4. Composition analysis of SBT in the sidewalls

A. As-grown defects at the sidewalls for $T_{\text{dep}} = 440^\circ\text{C}$

To have more insight into the ferroelectric behavior of SBT sidewalls, SEM investigation was made after MOCVD deposition of SBT. Fig. 2 shows the morphology of as-deposited SBT grown at 440°C . Mushroom-like defects are observed, especially in the sidewall region of the patterned BE substrate. These defects are not observed if SBT is grown at 405°C .

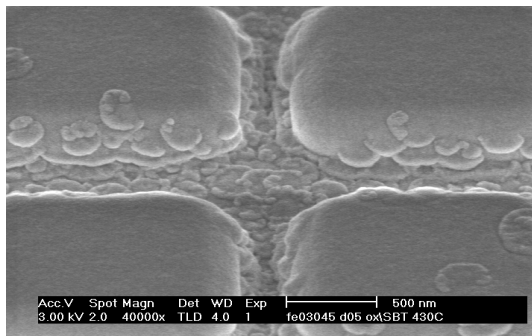


Fig. 2: SEM picture after SBT MOCVD at 440°C on patterned Pt/IrO₂/Ir/Ti(Al)N.

As we also observed these defects with the same large quantity in the sidewalls when SBT is deposited on patterned SiO₂ substrates (not shown), they cannot be

attributed to the interaction of SBT with the different layers constituting the BE in the sidewall region (i.e. IrO₂, Ir and Ti(Al)N). Hence the growth of mushroom-like defects is related to the surface topography, in other words to the influence of the use of a patterned surface on the MOCVD growth mechanism of SBT.

After crystallization at 700°C , 60min., these defects are disappeared, only slightly bumped SBT material remains at the sidewalls. Still, they may be at the origin of the non-ferroelectric behavior of sidewall SBT.

B. Local composition analysis

Samples for TEM observations and EDS analyses were prepared by the Focus Ion Beam (FIB) technique, allowing composition analysis along the width of stripe-type FeCAP's. From the weight percentage of the Sr, Bi, and Ta elements probed in the SBT layer, the average quantities of these elements (namely x , y , and z in $\text{Sr}_x\text{Bi}_y\text{Ta}_z\text{O}_9$) have been calculated in an SBT unit cell.

For sample SBT440, which had MOCVD deposition at $T_{\text{dep}} = 440^\circ\text{C}$, large variations in Bi content were observed: y varied from 2.2 in the planar region up to 3.5 in the sidewall region. The location where the concentration of Bi drastically increases matches the location of the mushroom-like defects (Fig. 2).

For sample SBT405, the increase in Bi concentration in the sidewalls was observed to be lower: $y = 2.6$ at most in the sidewalls. It is further minimized for sample SBT405opt, for which $y = 2.3$ at most in the sidewalls.

C. Discussion

These results indicate that the MOCVD growth of SBT at 440°C on patterned substrates goes along with the diffusion of Bi from the planar areas into the sidewall areas. The excess of Bi in the sidewall areas leads to the segregation of Bi apart from the SBT phase, and to the growth of mushroom-like defects. X-ray diffraction diagrams (not shown) obtained after SBT deposition at 440°C on patterned substrates exhibit a sharp peak at 28.4° in 2θ (for $\lambda_{\text{Cu}} = 1.5418 \text{ \AA}$), most probably corresponding to crystalline Bi₂O₃ phase. Such Bi-rich and crystalline mushroom-like defects have already been observed after MOCVD deposition of SBT [7].

After anneal at 700°C of SBT layers grown at 440°C on patterned substrates, the mushrooms have disappeared (according to SEM imaging), leaving some slightly-bumped SBT material in the sidewalls together with small voids observed underneath (revealed by means of SEM and TEM cross-sections). This suggests a back-diffusion of SBT elements at 700°C , which is driven by the 'Aurivillius' phase formation. However, according to the EDS analysis, this back-diffusion is incomplete and leaves a higher amount of Bi in the sidewall than in the planar region. It is probable that the Bi-rich sidewalls are difficult to convert to the 'Aurivillius' phase at 700°C [7]. As the SBT sidewall is not ferroelectric for $T_{\text{dep}} = 440^\circ\text{C}$ (Table 1), it is indeed probable that no 'Aurivillius' phase is present in the sidewalls after anneal at 700°C .

For sample SBT405 that had MOCVD deposition at $T_{\text{dep}} = 405^\circ\text{C}$, the Bi variation along the width of the FeCAP's is much less pronounced than for SBT440: this is consistent with the absence of mushrooms observed after MOCVD at $T_{\text{dep}} = 405^\circ\text{C}$. Moreover, the contribution of SBT sidewall to the Pr value of the array for SBT405 proves the formation of the 'Aurivillius' phase at 700°C if $T_{\text{dep}} = 405^\circ\text{C}$ (Table 1). Still, as the parameter *SW efficiency* is only 50%, this contribution is not maximal. This can again be attributed to the difficult conversion of Bi-rich SBT to the Aurivillius phase, and thus to a worse crystallization in the SW's as compared to in the planar region. This can also simply be related to the critical influence of the Aurivillius composition on the ferroelectric properties, as widely reported in the literature [8].

Sample SBT405opt was processed with optimized MOCVD deposition conditions of SBT at $T_{\text{dep}} = 405^\circ\text{C}$. As a consequence, Bi variations are observed to be further reduced and the Pr extracted from the 3-D hysteresis loop is $7.5 \mu\text{C}/\text{cm}^2$ (Table 1), and the comparison with the Pr extracted from the 2-D loop yields a *SW efficiency* increased up to 88%.

5. Conclusion

In this work we have compared electrical properties of 2-D and 3-D SBT-based ferroelectric capacitors after integration. We evidenced that the ferroelectric contribution of the SBT sidewalls in 3-D configuration is strongly dependent on the Bi segregation occurring in the sidewall region during MOCVD deposition of SBT layer. This process was reduced by lowering the

deposition temperature from 440°C to 405°C , and minimized after further optimization of the MOCVD conditions at 405°C . As a result of this improvement, the ferroelectric properties of the SBT layer in the sidewalls are almost equal to those of the SBT layer in the planar region, resulting in an increase of $2Pr$ from $9 \mu\text{C}/\text{cm}^2$ (for planar capacitors) to $15 \mu\text{C}/\text{cm}^2$ for 3-D capacitors integrated up to metal interconnect.

The present work of the control of segregation processes in the sidewall region paves the way to further FeRAM development in sub-100 nm CMOS technologies where 3-D capacitors with even higher aspect ratio will be required.

Acknowledgment

The authors acknowledge H. Vander Meeren and M. Willegems of IMEC for technical assistance, and the partial financial support through the IST project 2000-30153 FLEUR.

References

- [1] S. Oh et al., IEDM Tech. Dig. 835-838 (2003).
- [2] Y.J. Song et al., VLSI Tech. Dig. 169-170 (2003).
- [3] Y. Nagano et al., VLSI Tech. Dig. 171-172 (2003).
- [4] H. Mc.Adams et al., VLSI Circuits Dig. 175-176 (2003).
- [5] R. Zambrano, Integrated Ferroelectrics **53**, 247-255 (2003).
- [6] G.G. Condorelli et al., Materials Science in Semiconductor Processing **5**, 167 (2003).
- [7] F.S. Hintermaier et al., United States Patent, nb 6,133,051 (2000).
- [8] N. Nukaga et al., Jpn. J. Appl. Phys. **38**, 5428 (1999).

Technological Advances for Memory Applications: Crystalline High-K Gate Dielectrics and Alternatively Doped Gates

Y. Stefanov, R. Komaragiri, U. Schwalke

Institute of Semiconductor Technology, Darmstadt University of Technology
Schlossgartenstr. 8, 64289 Darmstadt, Germany, stefanov@iht.tu-darmstadt.de

Abstract

We report on current advances in CMOS processing that can be successfully implemented for advanced sub-one nanometer gate oxide memory fabrication. Crystalline rare-earth high-k materials hold big promise as ultra-low leakage replacements of silicon dioxide as gate dielectrics. Alternatively doped gates are an easy and efficient solution for avoiding gate depletion problems.

1. Introduction

For the past four decades microelectronic circuits have become denser [1], faster, cheaper and more reliable. This unparalleled development has been fuelled by the continuous scaling down of the building blocks of integrated circuits: the MOSFET and the MOS capacitor. Probably the most critical area in scaling is the gate dielectric. For its thickness to continue decreasing in the sub-nanometer region, several serious issues have to be addressed.

First of all, tunneling currents through the gate dielectric increase exponentially and will soon exceed the on-currents thus greatly decreasing power efficiency. In order to lower leakages, gate dielectrics with higher dielectric permittivity (high-k) than that of SiO_2 need to be introduced in CMOS processing. Some of the most suitable candidates are crystalline rare-earth dielectrics, which possess sufficient processing stability, good interfacial properties with silicon and do not require SiO_2 buffer layer. With their help the equivalent oxide thickness (EOT) can be scaled down to fractions of a nanometer.

Furthermore, when EOT is approaching 2 nm and below, polysilicon gate capacitance (C_{pd}) cannot be neglected when compared to C_{ox} , as the total gate capacitance (C_g) is reduced in the presence of gate depletion (Fig. 1a) when the devices are turned on [2], resulting in a reduced drain current (I_d). The reduction in C_g will become even more severe when crystalline high-k gate dielectrics such as Pr_2O_3 are used, since in this case EOT can be scaled down to fraction of a nanometer. Extremely increased gate doping reduces gate depletion to a certain extent, but a highly doped polysilicon gate on the other hand introduces a high-density parasitic charge layer (due to the applied gate bias, a depletion region is still present at the polysilicon gate-oxide interface and the ionized dopant atoms form a high dense parasitic charge layer at the interface). These parasitic charges interact with the carriers in the channel, an effect known as remote coulomb scattering (RCS) [3], and degrade carrier mobility.

In order to eliminate both gate depletion and RCS without the need for introducing metal gate electrodes, we propose an alternative gate doping scheme in which the polysilicon gates are inversely doped, i.e. the n+ polysilicon gate in the nMOSFET is replaced by a p-type polysilicon gate and the p+ polysilicon gate in the pMOSFET with a n-type polysilicon gate (Fig. 1b). For inversely doped gates, gate depletion effect and RCS are completely eliminated since the gate is driven into accumulation when the device is switched on.

In this work we present results on fully functional Pr_2O_3 nMOSFETs with ultra-low gate leakages as well as both measurement and simulation results on pMOSFETs with alternatively doped gates that eliminate gate depletion and RCS.

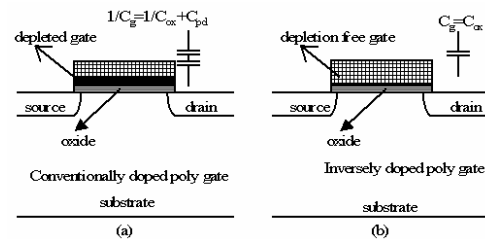


Fig. 1 Schematic diagram of polysilicon gate-oxide interfaces of (a) conventionally doped polysilicon gate structures and (b) inversely doped polysilicon gate structures illustrating the total gate capacitance (C_g).

As DRAM memory is traditionally the process driver for technology development, we can expect these concepts to be applied first in DRAM production. Both the capacitor and the transistor in a DRAM cell (Fig. 2) can benefit from improved gate capacitance and decreased leakages, allowing further downscaling and improved performance.

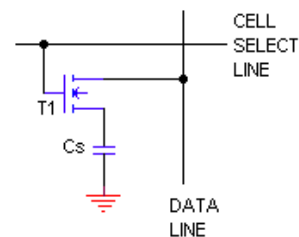


Fig. 2: Dynamic RAM cell schematic

2. Device Fabrication and Simulations

Crystalline Pr_2O_3 high-k gate dielectric was integrated in a modified n+ polysilicon CMOS process in order to fabricate fully functional Pr_2O_3 MOSFETs.

Praseodymium oxide layers of 17nm (EOT=1.8nm) were grown on pre-implanted p-type Si(100) 3-inch wafers in a molecular beam epitaxy (MBE) system [4] using a commercially available ceramic Pr_6O_{11} source. The layers were covered in-situ with 50nm of undoped polysilicon as unprotected Pr_2O_3 is not stable against air. Next, a 250nm of in-situ phosphorous doped polysilicon layer was deposited at approx 750°C. The n+ polysilicon gate stack was structured by optical photolithography and reactive ion etching (RIE) with SF_6/Cl_2 . No polysilicon residues, surface damage, or thickness changes of the gate dielectric were observed after the RIE step. Subsequently n+ source/drain implantation was performed, followed by a brief rapid thermal anneal (RTA) at 900°C for 10sec. in an inert atmosphere for activation of the implanted doping. A low temperature oxide was deposited and contact holes were etched. Standard Al/Ti back-end metallization and forming gas anneal at 425°C for 10min. completed the fabrication. Further details on the process integration of crystalline Pr_2O_3 have been published previously [5].

In order to prove the concept of inversely doped gates, pMOSFETs were fabricated with highly doped n⁺-polysilicon gates with a phosphorus concentration of 10^{20} cm^{-3} and low-doped n-polysilicon gates with a doping concentration 10^{18} cm^{-3} . Conventional processing was used, except for removing the resist for gate lithography only after source and drain implantation in order to protect the oppositely doped gates.

pMOSFETs and nMOSFETs with various gate doping profiles of 10^{15} - 10^{20} cm^{-3} and channel lengths ranging from 1 μm down to 150nm with an EOT of 2nm were process-simulated using TSUPREM4. The main aim of these process simulations is to obtain devices with identical threshold voltage (V_{th}). The simulated structures were then imported into the device simulator MEDICI to obtain device I-V characteristics.

3. Results and Discussion

3.1. Crystalline Pr_2O_3 MOSFETs

The Pr_2O_3 MOSFETs are fully functional as seen from their sub-threshold (Fig. 3) and output characteristics (Fig. 4). Gate leakages of Pr_2O_3 transistors with gate length $L < 10\mu\text{m}$ are well below 10pA (detection limit of the measurement system), corresponding to less than 10^{-6} A/cm^2 . Due to a high density of extrinsic defects, however, larger gates ($L > 10\mu\text{m}$) are often leaky (Fig. 5). When the output characteristics are compared to what is expected from simulation using SiO_2 of the same EOT, a reduction in the effective mobility by approximately 20% needs to be taken into account to obtain the measured drain currents. Part of this performance degradation is related to a high density of interface charges and oxide charges.

Other evidence for the susceptibility of Pr_2O_3 MOSFETs to charge trapping effects is the increased sub-threshold swing ($S=145\text{mV/dec}$) and the V_{th} -shifts to more negative biases during measurements with increasing sweep number (Fig. 6). The V_{th} -shifts may be due to trapping of positive charges or de-

trapping of negative oxide charges. The magnitude of the shift decreases with increasing number of sweeps, as shown in the inset of Fig. 6. The total shift is 250mV and corresponds to a trap density of approximately $2 \times 10^{12} \text{ q/cm}^2$. Full recovery and even overcompensation is observed after applying negative gate biases. Furthermore, charge trapping is found to be partially recoverable with time at room temperature.

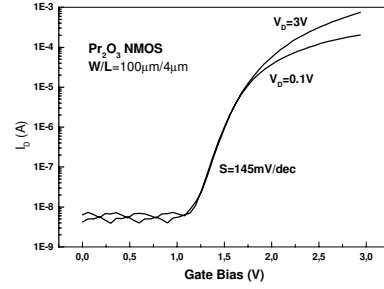


Fig. 3: Sub-threshold characteristics of Pr_2O_3 MOSFETs, $W/L=100/4\mu\text{m}$.

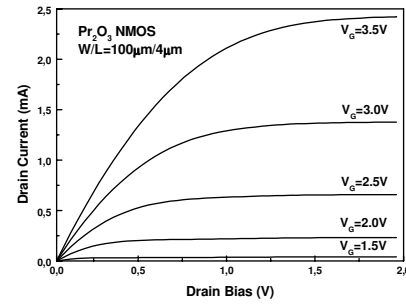


Fig. 4: Output characteristics of Pr_2O_3 MOSFETs, $W/L=100/4\mu\text{m}$.

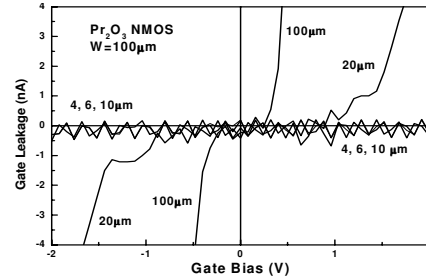


Fig. 5: Size dependence of gate leakage of Pr_2O_3 nMOSFETs. Gate lengths: 100, 20, 10, 6 and $4\mu\text{m}$.

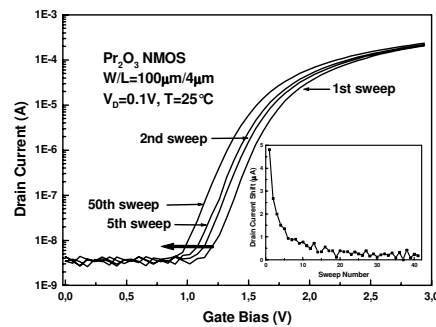


Fig. 6: V_{th} -shift produced during measurements. Inset shows the drain current shift as a function of the number of sweeps.

Tunnelling current atomic force microscopy [6] was used to investigate further the Pr_2O_3 gate dielectrics at various stages of processing. While scanning, a certain bias was applied to the samples and the resulting currents between the scanning tip and the sample were recorded. Since topography and current images are measured simultaneously and independently from each other, direct correlation of the sample location and its electrical properties is possible within nanometer resolution (Fig. 7). Clearly, the quality of the Pr_2O_3 dielectric at this point is inferior to state-of-the-art SiO_2 films. The AFM scans indicate both structural and electrical non-uniformities on the high-k sample, whereas no such non-uniformities are visible on thermally grown SiO_2 . Besides the structural defects already seen in the topographical AFM image of the new material which affect the local electrical properties, pure electronic defects can also be identified which are only visible in tunneling current images. Local tunnelling AFM I-V sweeps (Fig. 8) of structural and electronic defects in point contact mode are currently performed to obtain further information on local electron transport mechanisms in epitaxial Pr_2O_3 gate oxides.

In Fig. 9 the build-up of enhanced leakage sites during consecutive scans of the same surface area due to local charge trapping is shown. Note that no change in topography by AFM is observed. To which extent these trapping sites are related to grain boundaries of the two orthogonal Pr_2O_3 domains which form during growth on $\text{Si}(100)$ surfaces is subject to further investigation.

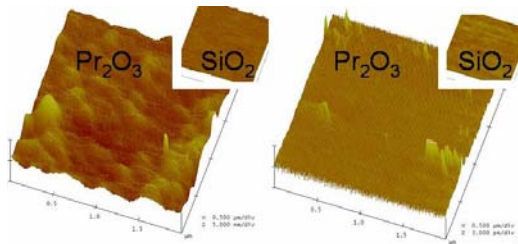


Fig. 7: AFM topography (left) and tunneling current (right) images measured simultaneously on crystalline Pr_2O_3 film. Inset shows SiO_2 reference for comparison.

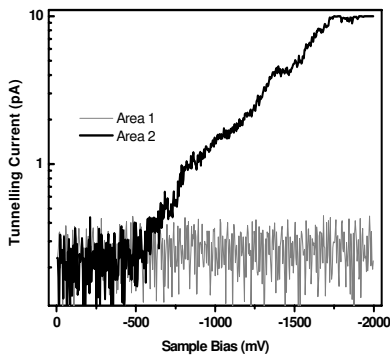


Fig. 8: Local Current-voltage characteristics of two different sites on the Pr_2O_3 surface measured by TUNA-AFM in the point-contact mode. Area 1: non-leaky site, Area 2: leaky site.

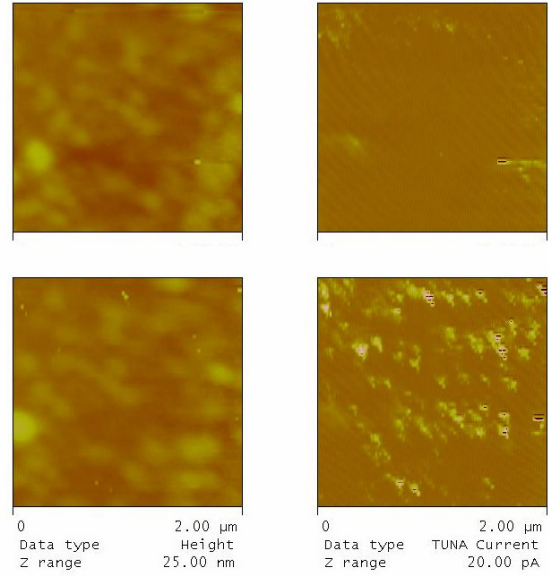


Fig. 9: AFM topography scans (left) and tunneling scans (right). Top row: fresh crystalline Pr_2O_3 film; bottom row: build-up of trapping sites after 3 consecutive scans.

3.2. Alternatively Doped Gates

The simulated capacitance-voltage (C-V) characteristics of a pMOSFET are shown in Fig. 10. Even for the low-doping case (10^{17}cm^{-3}), the inversely doped (n-type) gate of the pMOSFET is not depleted when the device is turned on, and retains the maximum gate capacitance value. The elimination of gate depletion in inversely doped gates was experimentally verified using pMOSFETs with a channel length $1.2\mu\text{m}$. Although the gate doping is reduced by two orders of magnitude, from 10^{20}cm^{-3} down to 10^{18}cm^{-3} , the drain current increases substantially as shown in Fig. 11. In the absence of gate depletion one can take full advantage of the reduced gate work function of the low-doped gate to increase the gate overdrive ($V_{\text{GS}} - V_{\text{D}}$). Although the experimental devices are buried channel devices rather than surface channel devices, the principle still holds. Furthermore, our simulation studies show that, if one down scales the lateral device dimensions to sub-200 nanometer, a surface channel device can be achieved with inversely doped gates [7]. The simulated device drain characteristics, which are shown in Fig. 12, are obtained for devices with a channel length of 150 nm. For inversely doped gate devices, a substantial performance improvement in terms of drain current is observed. This performance improvement of inversely doped CMOS devices is due to (1) the elimination of gate depletion and (2) the enhancement of mobility since RCS is removed in the absence of parasitic gate charges. In addition, due to the reduction of V_{t} adjust implant dose, the doping concentration in the channel is reduced [7], thus resulting in an additional mobility improvement due to reduced impurity scattering.

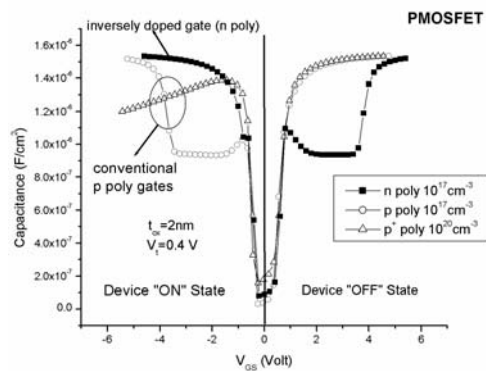


Fig. 10: C-V characteristics of a pMOSFET (or nMOS capacitor) with conventional polysilicon gates and inversely doped gates.

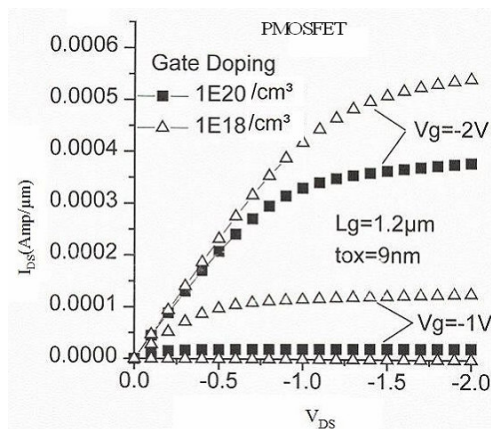


Fig. 11: Measured I-V characteristics of a pMOSFET with inversely doped gates

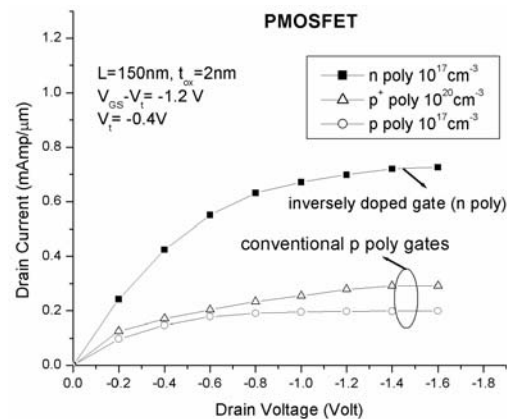


Fig. 12: Simulated I-V characteristics of a pMOSFET with an inversely doped n-polysilicon gate (10^{17} cm^{-3}) and with conventionally doped p-polysilicon gates (10^{17} and 10^{20} cm^{-3}) clearly showing the improvement in drain current.

4. Conclusions

We have presented results on functional nMOSFETs with crystalline Pr_2O_3 gate dielectrics. Although device performance is not ideal at this early stage of development, one can clearly see the significant decrease in leakage currents compared to thermal Si_2O and other possible high-k replacements of the same EOT.

Additionally, unlike most other high-k dielectrics, crystalline rare-earth metal oxides possess sufficient CMOS processing stability and do not require an interfacial Si_2O layer which makes them very suitable for advanced ICs, allowing both to scale down transistor size in logic and decrease the aspect ratio of the trench for capacitors in DRAM memory cells substantially. We have extended our work to other lanthanide oxides such as Nd_2O_3 . Integration and electrical characterization results will follow in the near future.

Furthermore, we have demonstrated the advantages of using inversely doped gates both in device measurements and simulations clearly indicating the superior performance over the conventionally doped counterparts. The elimination of polysilicon gate depletion will significantly improve performance in future generation microelectronic devices.

The use of crystalline high-k gate materials and inversely doped gates will be more advantageous when combined together, which can lead to scaling down devices into the sub-50 nanometer regime without significant process integration and device stability problems.

Acknowledgements

This work was partly funded by the German Federal Ministry of Education and Research (BMBF) under the KrisMOS project (01M3142C) and by the German Research Foundation (Deutsche Forschungsgemeinschaft) under grant GRK 749/1-02.

References

- [1] ITRS, <http://public.itrs.net/Files/2003ITRS/Home2003.htm>
- [2] R.Rios and N.D.Arora, 1994 IEDM Techn. Digest, p. 613.
- [3] F Gámiz, et al., Semicond. Sci. Technol., 18 (2003), p 927.
- [4] H. J. Osten et al., IEDM (2001).
- [5] U. Schwalke, et al., J. Osten, et al., H.-J. Müssig, Proc. ESSDERC, p. 407, Sept. 2002
- [6] A. Olbrich et al., Appl. Phys. Lett. 73, 3114 (1998).
- [7] R.Komaragiri, F.Zaunert and U.Schwalke, Proc. SAFE 2004, Nov. 25– 26, 2004 – Veldhoven, the Netherlands.

A Two-Terminal Semiconductor Memory Made in a Single Nanolithography Step

A. M. Song^a, M. Missous^a, and W. Seifert^b

^a School of Electrical and Electronic Engineering, University of Manchester, P. O. Box 88, Manchester M60 1QD, United Kingdom. Email: A.Song@manchester.ac.uk

^b Solid State Physics and Nanometer Consortium, Lund University, S-221 00 Lund, Sweden

Abstract

Conventional semiconductor memory devices have at least three terminals. In this work, we used a nanometre-scale channel with a broken geometric symmetry and realised a new type of memory devices with only two terminals. The charge retention time at room temperature is in the order of minutes, which is about four orders of magnitude longer than a typical DRAM. Furthermore, the self-switching memory has a very simple planar structure, which was made by fabricating insulating trenches on the substrate and required only a single nanolithography step. The device therefore also allows for low-cost and high-density integrations.

1. Introduction

It has long been realised that nanotechnology is not only a key enabler of the next generation electronics. This is because new effects associated with nanometre dimensions allow inventions of completely new ideas. Despite these great promises, very few new concepts have made it into practical applications, in strong contrast to the enormous investments and efforts. This is because most novel nanoelectronic devices only function at cryogenic temperatures, which is in contrast to optical devices in which transitions between conduction and valence bands are less affected by temperature. Moreover, they virtually always require extremely critical lithography resolutions, for instance single electron transistors require a dimension of a few nanometres, resulting in a yield too low for mass production. Recently, we have demonstrated a novel type of nano-diode, called self-switching diode (SSD) and made by simply etching narrow trenches with asymmetric layout on a semiconductor substrate [1]. Without using any doping junction or tunnelling barrier, the working principle of the SSD is completely different from that of a normal diode.

In this work, we employ the SSD structure and demonstrate a new type of two-terminal memory device using two different material systems, InGaAs-InP and InGaAs-InAlAs heterostructures. The charge retention time at room temperature is in the order of minutes, which is orders of magnitude longer than a typical DRAM. Furthermore, the self-switching memory (SSM) has a very simple planar structure, which was made by fabricating insulating trenches on the substrate and required only a single nanolithography step. The device

therefore also allows for low-cost and high-density integrations. From the working principle, the device is expected to function with silicon material and is also compatible to CMOS technology.

2. Results and discussions

Figure 1 shows a scanning electron micrograph of a typical self-switching memory device. The channel length is about 1.2 μm . The dark areas were etching down through the two dimension electron gas in an InGaAs-InP heterostructure, resulting in two insulating trenches. If a voltage is applied to the left and right terminals of the device, the current can only flow through the narrow channel between the two trenches. We reported in Ref. [1] that the device functioned as a nano-diode at room temperature, because of the asymmetric geometric layout of the device.

The memory effect was most clearly observed at low temperature. Figure 2 (a) shows the I-V characteristics and hysteresis effect of a typical device at 24 K. By using a memory detecting voltage of -0.5V, and memory switching voltages of +1V and -1V, we performed real-time demonstration of the memory effect, as shown in Fig. 2(b). We found that the memory holding time was at least 10 hours in separate experiments. Note that here the long switching pulses of 1V and -1V were used for clarity. The hysteresis effect was also observed by applying triangular waves up to 10 kHz (the limit of our setup) to the device, meaning that the memory switching time is below or at least in the order of ms [2].

The memory effect can be explained by the charging and discharging the surface states on the etched sidewalls of the trenches. For a nanometer-scale channel, the large surface to bulk ratio makes the devices extremely sensitive to charging and discharging of the surface states. Figure 3(a) is a sketch of an SSM. The conduction band profile at zero bias along the dashed line in (a) is shown in Fig. 3(b). The device is in equilibrium and has a unified Fermi energy. Since the device is not conducting at zero bias, shown in Fig. 2(a), the potential bottom of the channel is above the Fermi energy, meaning that there are no free electrons in the channel. The surface states at both inner and outer sidewalls of the trenches have a narrow energy distribution (the percentage of the filling of the surface states here is only for illustration purpose, and at this stage, no exact value is known but this does not affect the analysis in the following). When the applied bias is

beyond (or lower than) the negative threshold voltage, V_{th} (-1V for the example of Fig. 2(a)), the surface states of the outer sidewalls will be charged with more electrons and becomes more negative (or less positive). However, the inner sidewalls will be discharged by charge transfer into the channel and become more positive (or less negative), which is more important to the channel conductance because the inner sidewalls are more close to the channel. Overall, the channel effectively becomes more open. If the applied bias starts to sweep upward from -1V, the more opened channel will lead to the blue I-V curve in Fig. 2(a). The opposite process happens if the applied bias starts to sweep downward from +1V, the more closed channel will lead to the red I-V curve in Fig. 2(a).

We have also demonstrated the device at room temperature. At 300K, the memory “0” state no longer corresponded to a zero current, but rather a low, finite current, which was nevertheless sufficiently distinguishable from the memory “1” state with a much higher current. This is shown in Fig. 4. The memory holding time can be estimated to be at least a few minutes, which is about four orders of magnitude longer than a typical DRAM. Yet, the device lithography is much simpler, allowing potentially much lower cost to mass produce. Also, the devices can be made using silicon material, and are compatible with the CMOS technology.

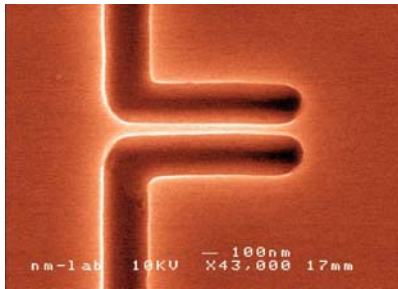


Fig. 1: A scanning electron micrograph of a typical self-switching memory device. The channel length is about 1.2 μm , and the narrow gap between the two trenches is about 50nm wide.

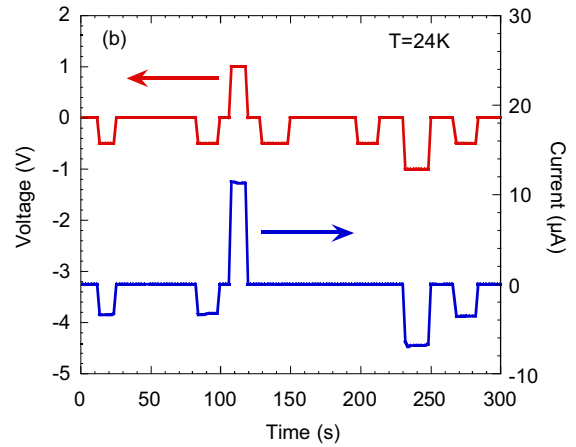
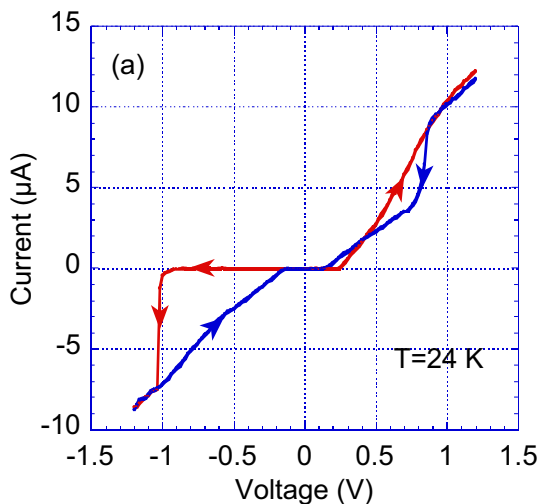


Fig. 2: (a) I-V characteristics and hysteresis effect of a typical device at 24 K. (b) Real-time demonstration of the memory effect, with a memory detecting voltage of -0.5V, and memory switching voltages of +1V and -1V.

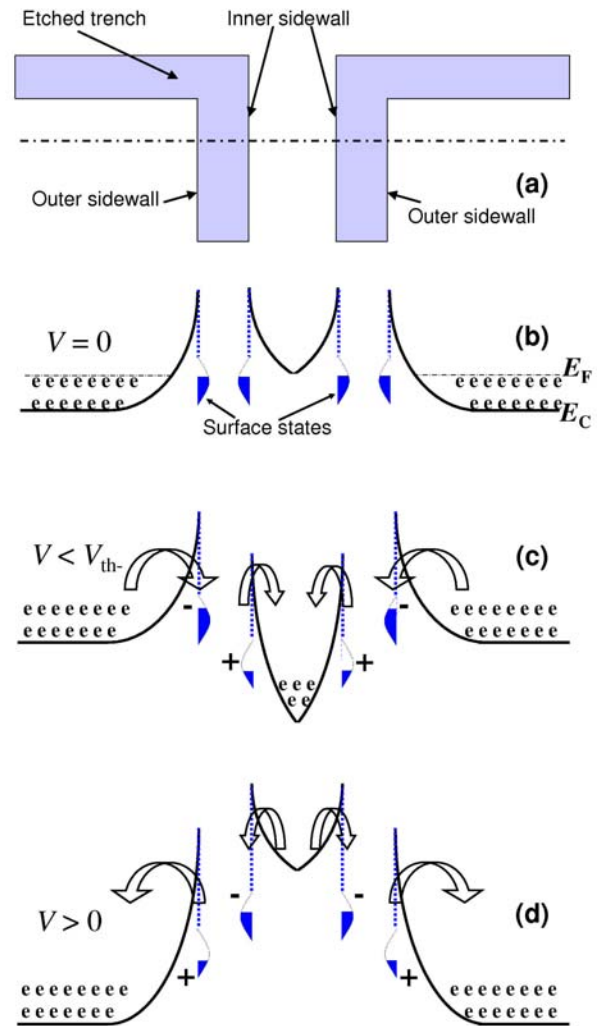


Fig. 3: (a) A sketch of an SSM. (b) The conduction band profile along the dashed line in (a). (c) When the applied bias is beyond (or lower than) the negative threshold voltage, V_{th} , the surface states of the outer sidewall will be charged with more electrons and becomes more negative. However, the inner

sidewall will be discharged by charge transfer into the channel and become positive, which is more important because the sidewall is more close to the channel. Overall, the channel effectively becomes more open. (d) The opposite process happens if the applied bias is positive.

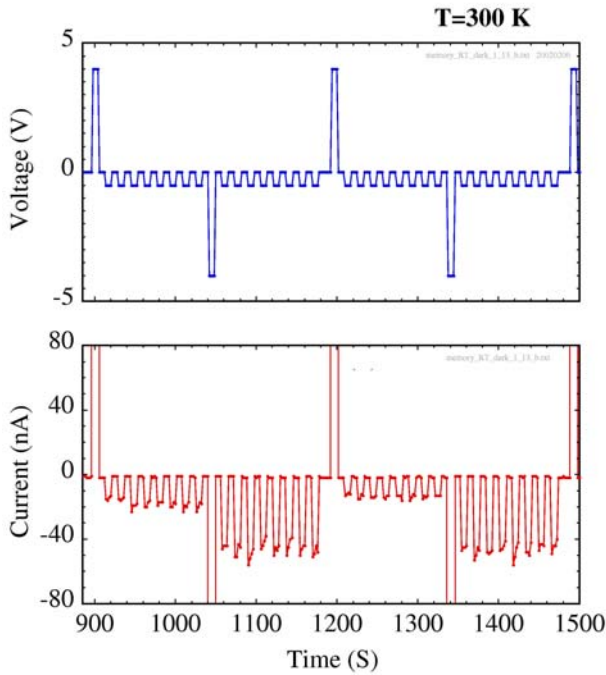


Fig. 4: Memory effect observed at room temperature. The real-time experiments were performed with a memory detecting voltage of -0.5V, and memory switching voltages of +4V and -4V.

3. Conclusions

We have fabricated a new type of room-temperature, two-terminal memory with a charge retention time reaching the order of minutes, which is about four orders of magnitude longer than a typical DRAM. From the working principle, the device is expected to function with silicon material with much lower cost and is also compatible to CMOS technology.

4. Acknowledgements

This work was supported by the Engineering and Physical Science Research Council (EPSRC) GR/S17741/01, and EU Framework Five project NEAR.

References

- [1] C. Balocco, A. M. Song, and M. Missous, Appl. Phys. Lett. **85**, 5911 (2004).
- [2] A. M. Song, P. Omling, I. Maximov, M. Missous, L. Samuelson, W. Seifert, Appl. Phys. Lett. **86**, 042106 (2005).

Development of Atomic Layer Deposition Processes for Multicomponent Oxide Dielectrics

Marko Vehkamäki,^a Mikko Ritala^a and Markku Leskelä^a

^a Laboratory of Inorganic Chemistry, Department of Chemistry, University of Helsinki, Finland

email: marko.vehkamaki@helsinki.fi

Abstract

The different chemical approaches that have been reported in Atomic Layer Deposition (ALD) of multi-component oxide materials such as SrTiO₃ and SrBi₂Ta₂O₉ is reviewed. Development of complex ALD processes requires new ALD-specific precursor chemistry or the use of highly reactive oxygen sources such as oxygen radicals or oxygen plasma to activate otherwise unsuitable metal precursors.

1. Introduction

Atomic Layer Deposition (ALD) [1] thin film processes are entering IC mass production lines due to the excellent conformality and large area uniformity of ALD films. Commonly studied high-K ALD oxides, are Al₂O₃, ZrO₂, HfO₂ and metal silicates. Although these materials already have higher permittivities than SiO₂ and SiO_xN_y dielectrics, there is an ever-increasing need for higher charge densities. This suggests that ALD processes for very high permittivity materials such as SrTiO₃, BaTiO₃ and BST could prove useful in future memory devices, for example in future generations of both stack and trench DRAM architectures. Also, there is considerable interest in the development of non-volatile memories such as ferroelectric random access memories. ALD is an interesting candidate for depositing very thin films mixed oxides with difficult compositions such as SrBi₂Ta₂O₉ over large wafer surfaces.

Atomic Layer deposition of multi-component materials is more challenging than that of binary oxides, especially if the material contains metals with difficult to fabricate binary oxides, such as SrO, BaO and Bi₂O₃. Well known stable binary oxides such as TiO₂ or Ta₂O₅ may help self-limited incorporation of more difficult oxides into oxide mixtures. ALD conditions for multi-component oxides are generally found in fairly low temperatures (150-350°C) so post-deposition annealing will likely be required in order to obtain desired high permittivity or ferroelectric phases.

2. ALD growth studies of multicomponent oxide materials

Multi-component oxide dielectric ALD studies include SrTiO₃ [2-5] BaTiO₃ [2,6] SrTa₂O₆ [7-8], SrBi₂Ta₂O₉ [8-9], Bi₂Ti₂O₇ [10], Bi₄Ti₃O₁₂ [11] and recently also Bi_{1-x}Y_xTi₃O₁₂ [9]. These materials share a common processing issue of a strong influence of material composition on their dielectric permittivities, leakage current densities or ferroelectric remanent polarizations. This makes ALD a lucrative option especially if films need to be processed into deep trenches.

3. Chemical approaches used in ALD of Multicomponent oxides

For binary oxides such as TiO₂ the ALD process consists of four steps which are repeated until a sufficient film thickness is achieved. First, the metal precursor, typically a metal alkoxide, metal halide or organometallic compound, is pulsed into the reactor in a sufficient dose to consume all available reaction sites on the growth surface, thus producing a saturated molecular layer. Second, the reactor chamber is purged with inert gas, typically N₂, in order to remove all excess reactant prior to the next precursor pulse. Once the reactor is properly purged, the oxygen source, typically H₂O, O₃ or oxygen plasma is led to the surface in a third step. The fourth and final step is purging of the excess reactant, thus ending the deposition cycle with a surface on which the cyclic deposition can continue. Two main approaches have been used in metal precursor delivery in multi-component oxide processing. The first is mixing different metal precursor/oxygen source reactions with one another. The second is the use of single-cocktail mixtures, which are reacted with a surface to add multiple metals in a single pulse.

Using a separate metal precursor/oxygen source reaction for each metal needed in the multi-component film provides the best control for incorporation of each separate metal into the film. The film composition is controlled by changing the ratio in which the individual reactions are performed. This is a more time-consuming way, as each metal precursor pulse requires a separate purging time, an oxygen source pulse and a oxygen source purging time. If films with uniform compositions are to be applied to very deep trenches (aspect ratios >10:1) this may be the only viable approach.

Reactions of Sr and Ba cyclopentadienyl/H₂O reactions mixed with Titanium tetraisopropoxide/H₂O reactions were used to deposit amorphous or polycrystalline SrTiO₃ [2,3] at 250-325 °C and BaTiO₃ [2,6] at 275 °C. The use of cyclopentadienyl precursors produces films with very low carbon concentrations, at 0.1-0.3 at%, which is probably due to the stability of cyclopentadienyl ring. Kosola et al. reported that the use of Sr(thd)₂/O₃ reactions mixed with titanium isopropoxide/water reactions were found to result in a mixture of SrCO₃ and TiO₂ which could be annealed to form SrTiO₃ [4]. The use of bimetallic Sr-Ta alkoxides for deposition of SrTa₂O₆ has been reported using either oxygen plasma [5] or H₂O [6] as the oxygen source. In both cases some deviation from the precursor 1:2 Sr/Ta ratio was observed, thus further study on the reliability of bimetallic alkoxides in ALD use is needed.

Mixing the bimetallic alkoxide/H₂O reactions with Bi(N(Si(Me)₃)₂)₃/H₂O reactions could be used for deposition of SrBi₂Ta₂O₉. [8] The process temperature was a very low 190 °C, which means that H₂O evacuation times need to be long. The as-deposited Sr-Bi-Ta-O films were amorphous and post-deposition annealing was used to crystallize the films. Schuisky et al. studied mixing of Bi(Ph)₃/H₂O reactions with titanium tetraisopropoxide/H₂O water reactions in deposition of Bi-Ti-O films [10] It was seen that the use of Bi(Ph)₃ resulted in the incorporation of metallic bismuth into the films, and that it was difficult to exceed a Bi/Ti ratio of 0.6 using this precursor combination. Clearly, if Bi(Ph)₃ is to be used, more reactive oxygen sources such as oxygen plasma are needed.

The single-cocktail liquid injection method is often used in MOCVD as it is an easy way to introduce several chemicals into the reactor at once. This has also led to several reports about its use in ALD experiments as well. It can be difficult to acquire uniform cation distributions in deep and narrow trenches this way, as differences in precursor reactivities may lead to composition profiles in trenches. Nevertheless, successful deposition of several materials has been reported on planar substrates using alternating reactions of cocktail mixtures and an oxygen source.

Deposition of SrTiO₃ was achieved at 250-350 °C by injecting a Sr(thd)₂/Ti(OⁱPr)₄-mixture and subsequently treating the surface with an O₂ plasma discharge [5] A mixture of Bi(Ph)₃ and bimetallic SrTa₂(OEt)₁₀dmae₂ was used by Shin et al. for deposition of SrBi₂Ta₂O₉. Oxygen plasma was used as the oxygen source. Uniform coverage of a 5" wafer was reported. The SBT films were amorphous and were annealed to form the SrBi₂Ta₂O₉ layered perovskite phase. Cho et al. reported ALD of Bi₄Ti₃O₁₂ using a Bi(mmp)₃/Ti(mmp)₄ mixture for the metal pulse and an O₂/O₃ mixture as the oxygen source. Self-limiting growth was reported at 250-350°C and the deposited amorphous films were post-deposition annealed to form Bi₄Ti₃O₁₂. An injected mixture of Bi(mmp)₃, Ti(mmp)₄ and Si(OEt)₄ was used with O₃ at 325 °C to deposit Bi_{1-x-y}Ti_xSi_yO_z [12] which is an amorphous dielectric candidate for use DRAM applications.

4. Conclusions

Promising results have been reported for ALD of several key technological multi-component oxide materials in the recent years. Increased activity in ALD multi-component oxide dielectric research is needed in order to provide a wider selection of process chemistries and in order to evaluate the feasibility of currently known processes for use in memory applications.

References

[1] M. Ritala and M. Leskelä in Handbook of Thin Film Materials, vol. 1, chapter 2, ed. H. S. Nalwa, Academic Press, 2002

[2] M. Vehkamäki, T. Hatanpää, T. Hänninen, M. Ritala and M. Leskelä. *Electrochem. Solid-State Lett.* **2**, 504 (1999)
 [3] M. Vehkamäki, T. Hänninen, M. Ritala, M. Leskelä, T. Sajavaara, E. Rauhala and J. Keinonen. *Chem. Vap. Deposition* **7**, 75 (2001)
 [4] A. Kosola, M. Putkonen, L-S Johansson and L. Niinistö., *Applied Surface Science*, **211**, 102 (2003)
 [5] J.H. Lee, Y.J.Cho, Y.S.Min and D. Kim and S.W.Rhee., *J. Vac. Sci. Technol. A* **20**, 1828 (2002)
 [6] T. Hatanpää, M. Vehkamäki, I. Mutikainen, J. Kansikas, M. Ritala and M. Leskelä, *Dalton Transactions* **8**, 1181 (2004)
 [6] W.J Lee, I.K. You, S.O. Ryu, B.G. Yu, K.I.Cho, S.G.Yoon and C.S.Lee., *Jpn. J. Appl. Phys., Part 1*, **40**, 6941 (2001)
 [7] M.Vehkamäki, M. Ritala, M. Leskelä, A.C. Jones, H.O.Davies, T. Sajavaara and E. Rauhala., *J. Electrochem. Soc.* **151**, F69 (2004)
 [8] M. Vehkamäki, T. Hatanpää, M. Ritala and M. Leskelä., *J. Mater. Chem.*, **14**, 3191 (2004)
 [9] W.C. Shin, S.O. Ryu, I.K. You, S.M Yoon, S.M. Cho, N.Y. Lee, K. D. Kim, B.G. Yu, W.J.Lee, K.J.Choi and S.G.Yoon., *Electrochem. Solid-State Lett.* **7**, F31 (2004)
 [10] M. Schuisky, K.Kukli, M.Ritala, A Härsta and M. Leskelä, *Chem. Vap. Deposition*, 2000, **6**, 139
 [11] Y.J. Cho, Y-S Min, J-H Lee, B-S Seo, J.K. Lee, Y.S. Park, and J-H Choi., *Integr. Ferroelectr.* **59**,
 [12] Y-S Min, Y.J.Cho, I.P Asanov, J.H. Han, W.D. Kim and C.S.Hwang, *Chem. Vap. Deposition* **11**, 38 (2005)

A Macro Model of Programmable Metallization Cell Devices

Nad E. Gilbert, Chakravarthy Gopalan, and Michael N. Kozicki

Center for Solid State Electronics Research, Arizona State University, Tempe, AZ 85287-6206, USA
michael.kozicki@asu.edu

Abstract

This paper presents a new macro model for Programmable Metallization Cell (PMC) devices developed using PSPICE and incorporating only common circuit elements used in all SPICE-based programs. The model comprises an equivalent circuit which represents the dynamic operation of the PMC switching element and encompasses both pseudo DC and transient conditions.

1. Introduction

Continued scaling of non-volatile memory (NVM) will require the use of novel materials and storage mechanisms. One highly promising new NVM technology, Programmable Metallization Cell (PMC) memory, uses electrochemical control of nanoscale quantities of metal in solid electrolyte films [1]. The electrolyte is formed by dissolving silver or copper in a chalcogenide or transition metal oxide glass. A silver- or copper-containing layer and an inert electrode formed in contact with the electrolyte film creates a device in which information is stored via large resistance changes caused by the oxidation of the silver or copper metal and reduction of the mobile metal ions in the electrolyte. Polarity-dependent switching between the resistance states occurs within a few tens of nanoseconds for voltages of a few hundred mV and currents in the μA range. PMC possesses the endurance, retention, and CMOS compatibility required of advanced memory devices and has excellent scaling prospects due to the low operational power and physical scalability to the nanoscale regime [2].

Even though the technology has been under investigation by the semiconductor industry for several years, there has been no publication as yet of a comprehensive device model. This paper is therefore the first to present a macro model for PMC and discusses the equivalent circuit required to represent dynamic device operation. The performance of the model is compared with data from Ag-Ge-Se electrolyte devices.

2. PMC Device Operation

The details of PMC device operation have been reported elsewhere [1] but critical aspects are summarized here. The device is essentially a thin film (<100 nm thick) of solid electrolyte sandwiched between two electrodes as shown in Fig. 1. Solid electrolytes based on chalcogenide or transition metal oxide glasses can be made to contain large amounts of highly mobile metal ions. An example of such an electrolyte is 33 atomic % Ag dissolved in $\text{Ge}_{30}\text{Se}_{70}$, i.e., $\text{Ag}_{33}\text{Ge}_{20}\text{Se}_{47}$, which actually comprises a continuous high resistivity Ge_2Se_3 backbone and a dispersed nanoscale Ag_2Se

superionic phase [3]. This nanostructure, which also exists in the sulfide variant [4], leads to a high overall film resistivity even though mobile ion content is high. An anode consisting of oxidizable silver and an electrochemically inert cathode (e.g., tungsten) formed on an Ag-based electrolyte will allow ions to be reduced at the cathode and ions to enter the electrolyte via oxidation at the anode at an applied bias of a few hundred mV. Once initiated, the voltage required to sustain electrodeposition drops to around half the initial value. Charge neutrality is maintained by a balance between oxidation and reduction. Non-uniformities in the ion distribution and in the electric field at the cathode will promote localized Ag electrodeposition and a metal-rich filament will extend out from the electrode. The growing filament has a resistivity many orders of magnitude lower than that of the surrounding electrolyte and will supply electrons for subsequent ion reduction. In this way, the advancing electrodeposit harvests ions from the electrolyte, plating them onto its surface to extend itself forward. The average distance each ion has to travel in the ion-rich electrolyte to be reduced is around a few nm and so the growth process is rapid (around 1 nm/nsec). Once the electrodeposit has grown from the cathode to the anode, the resistance of the structure drops by several orders of magnitude. The decreasing resistance due to the electrodeposition effect increases the current flowing through the device until the current limit of the source is reached. At this point, the voltage drop falls to the threshold for electrodeposition and the process stops to yield the final resistance of the structure. The electrodeposition process is reversible by reversing the polarity of the applied bias; if the electrodeposit is made positive with respect to the Ag electrode, it becomes the anode and will dissolve via oxidation, the ions so created being reduced on the Ag electrode until the conducting link is completely removed. It is important to note that it is the asymmetry of the structure that allows cycling of the devices between a high resistance off state and a low resistance on state and thereby allows them to operate as switching/resistance change elements in memory circuits.

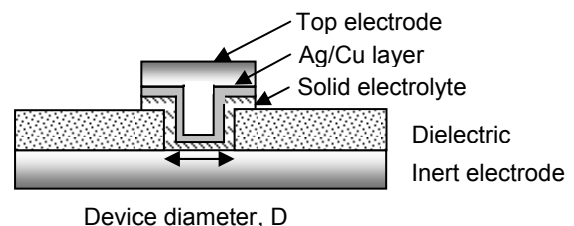


Fig. 1: Schematic of a PMC memory device. The solid electrolyte contacts the lower inert electrode through a via of diameter D in the dielectric. A thin silver or copper layer and a top electrode cap the structure.

3. PMC Spice Macro Model

There are essentially three ways to model the electrical characteristics of circuit elements in SPICE-based circuit simulators: C-code, Hardware Description Language (HDL), and SPICE Macro Models (SMM). Our method of choice was SMM. Although this is not as fast as the other approaches, we could use our understanding of physical device operation to create an equivalent circuit on which to base the model. This allowed us to use basic SPICE circuit elements - resistors, capacitors, voltage controlled sources, etc.

Until now, the only equivalent circuit of PMC was for the off-state resistance [5]. This circuit was developed via impedance spectroscopy on devices with electrode spacings of several tens of microns. These large structures behaved more like “traditional” electrochemical systems and required a Warburg impedance in the equivalent circuit, as well as solution resistance and geometric and double layer capacitance components. Note that for smaller devices, in which the dimensions involved in ion transport are small (a few tens of nm or less), the diffusion-dependent Warburg term can be neglected. In this previous work, a quasi-dynamic model was developed in HSPICE for neural net applications, in which the solution resistance was decreased according to the bias across the device using a voltage controlled resistor element in HSPICE [6].

In order to design circuits, including memory arrays, using PMC devices, a more accurate SMM of the dynamic operation of smaller devices was developed. Fig. 2 shows a schematic of our new SMM. The concept is based on the reduction and oxidation of Ag in the electrolyte. In the simplest sense, reduction of Ag in the electrolyte during the write can be thought of as electrons being stored on the Ag ions in the electrolyte, thereby forming the resistance-lowering electrodeposit, while the oxidation of the Ag constitutes a release of electrons. In the model, the electrons are stored on

capacitor CAG and the consequent electrodeposition is represented by the decrease in resistance of GON, which acts as a voltage controlled resistor. To erase the device, the electrons are pulled off the capacitor and the resistance of GON is increased, thereby simulating the Ag being oxidized and taken out of the electrolyte. In the following section, the physical details of the dynamics of the model will be discussed.

4. Physical Description of the SMM

The SMM can be broken into four sections: physical parasitics, threshold comparators, write dynamics, and erase dynamics. All sections with the exception of the threshold comparators are included in Fig. 2. A schematic of each individual section will be given from this point on in the paper for convenience.

The physical parasitics of the devices were adapted from the equivalent circuit in ref. [5] but now for devices of smaller geometry. These parasitics are qualitative and no attempt has been made to directly measure them. Fig. 3 shows a detailed schematic of the equivalent circuit. Capacitors CA and CC represent the parasitic capacitances of the anode and cathode to the substrate. Resistors RA and RB represent the contact resistance to the anode and cathode respectively. The voltage source VS is used to measure the current flowing through the device. Capacitor CED models the capacitance of the dielectric around the electrolyte. The voltage controlled voltage source EVC and capacitor CG model the variable capacitance of the electrolyte which decreases when the device is written and increases when it is erased. Device GON is a voltage controlled current source that is governed by eqn. 1 below.

$$I_{Gon} = V_{an,ac} * V_{con} + \frac{V_{an,ac}}{R_{off}} \quad (1)$$

By solving for the effective resistance between nodes AN and CA, the on resistance is determined by eqn. 2.

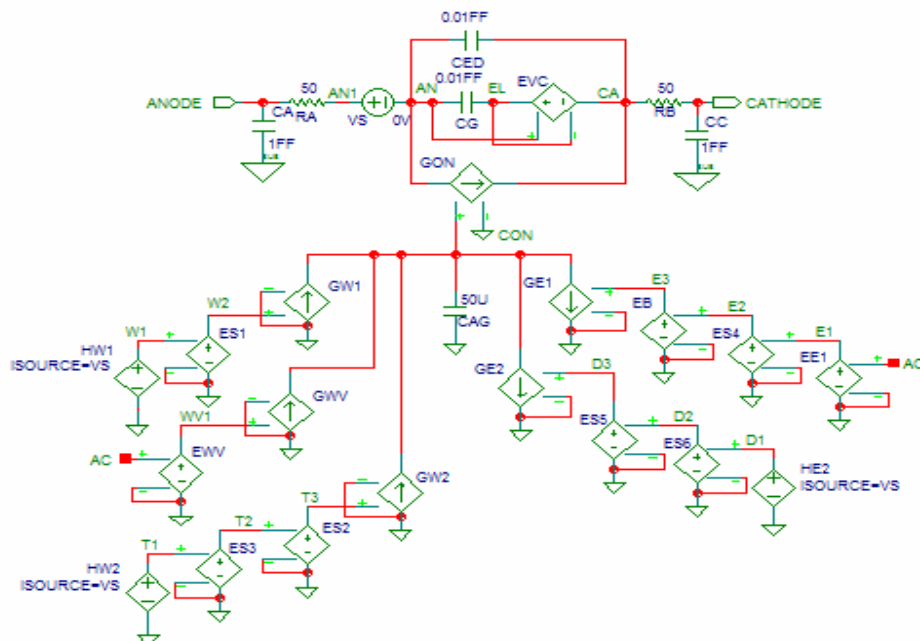


Fig. 2: The core section of the SMM for PMC, which operates on the principle of stored charge being converted to Ag for write and converted back to charge for the erase.

$$R_{on} = \frac{R_{off}}{R_{off} * V_{con} + 1} \quad (2)$$

The control voltage is designed to never go below 0 V so that a negative resistance and singularities are avoided.

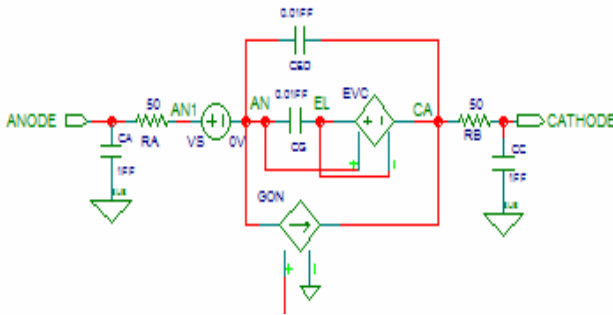


Fig. 3: Equivalent circuit for PMC parasitics.

There are several thresholds that govern the switching dynamics of PMC devices. Fig. 4 is a schematic of elements EC1 to EC10 which function as comparators. The two voltages of greatest importance are the bias from AN to CA and the voltage on node CON or the storage capacitor. The difference between nodes AN and CA is measured by E1 and then compared to a number of references. References 1, 2 and 3 and E elements EC1 EC2, and EC3 are for the write operation of the device. References 4, 5, and 6 and E elements EC4, EC5, EC6 are for the erase. EC7 and RF7 activate the “second” (lower electrodeposition threshold) write in the write cycle and EC8 and RF8 stop the erase cycle when the resistance of the device has returned to the off value. EC9 and EC10 along with their corresponding references are used to switch between the voltage and current modes in the erase.

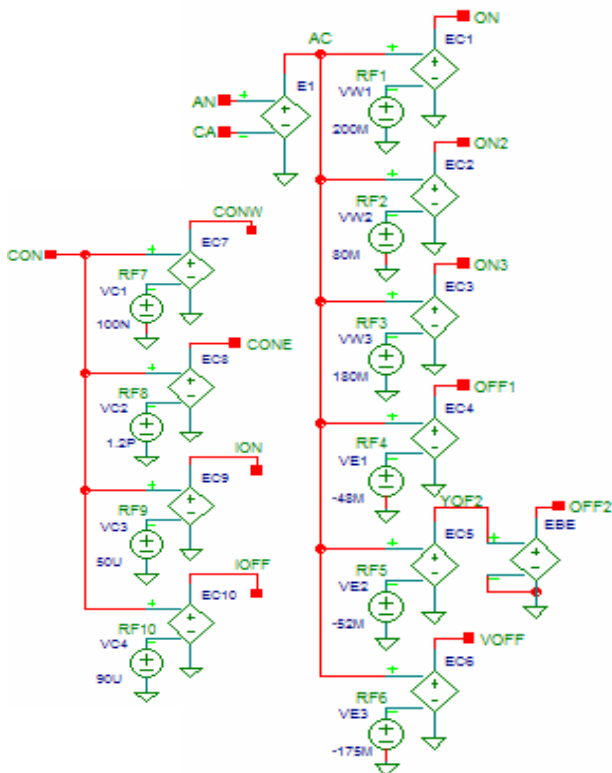


Fig. 4: Comparators and threshold for PMC switching dynamics.

The write cycle is both voltage (threshold) driven and current driven. The amount of electrons present is a major contributing factor in the amount of Ag stored in the electrolyte, so a combination of both current and voltage determinations in the intrinsic device must be made in order to set the subsequent resistance of the element. The write section of Fig. 2 is given again in Fig. 5. The three contributing voltage controlled current sources, GW1, GWV, and GW2 represent the three write mechanisms in the device. The initial write mechanism is controlled by the voltage across the device being large enough to initiate electrodeposition, determined by the comparator EC3 in Fig. 4. Element EWV, in Fig. 5, takes the voltage across the intrinsic device multiplied by a gain, which is determined by curve fitting, and then converts it into a current, given by GWV. The current is integrated over time as a voltage on the large capacitor CAG. When the initial voltage threshold has been developed across the device, determined by the comparator EC1 in Fig. 4, the current flowing through the device is measured by the current controlled voltage source HW1 and converted to a voltage and allowed to pass to GW1. The same integration process begins and is varied by the gains chosen for each element. Once a substantial charge has been stored on CAG, (or Ag stored in the electrolyte through GON), the second potential will be initiated if the minimum potential to support the electrodeposition appears across the device, determined by EC2 in Fig. 4. The gains of this leg are much less than the others because this process is much slower than the other two.

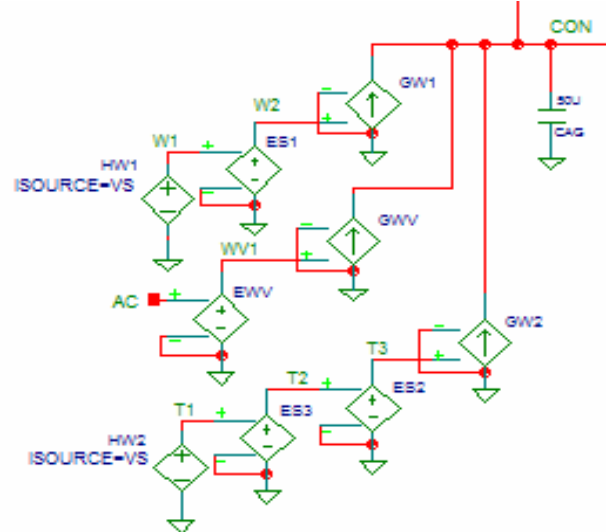


Fig. 5: Dynamic write section of SMM for PMC.

The final section is the erase circuitry, given in Fig. 6. It is also governed by the voltage and current in the device but in the reverse direction. If there is not a substantial Ag deposit in the glass, then the erase is dominated by the electrochemistry and the bias across the device is the dominant factor. The transition is determined by comparator EC9 and EC10. The voltage controlled voltage sources E1 to E3 perform this function. When the minimum erase potential is achieved, determined by VE2 and EC5 in Fig. 4, the voltage across the device is measured by E1 and allowed

to pass through E2 when the device is in voltage (electrochemical) erase mode. E3 in Fig. 6 and EBE in Fig. 4 ensure that node CON does not fall below zero volts. If the voltage across the device exceeds a large negative potential, determined by comparator EC6 in Fig. 4, the erase process is accelerated by initiating the voltage mode regardless of the current mode operation. The voltages created by E3 to E1 are then converted to a current by GE1 and pull charge off of the capacitor, which translates to an increase in the resistance (or silver being taken out of the electrolyte). The current mode of the erase works exactly opposite to the write current mode. When the appropriate potential across the device is reached, HE2 converts the negative current in the device into a positive voltage that is converted to a current by GE2 to pull charge off the capacitor. Element ES5 enables or disables the voltage mode dependent on the voltage on node CON.

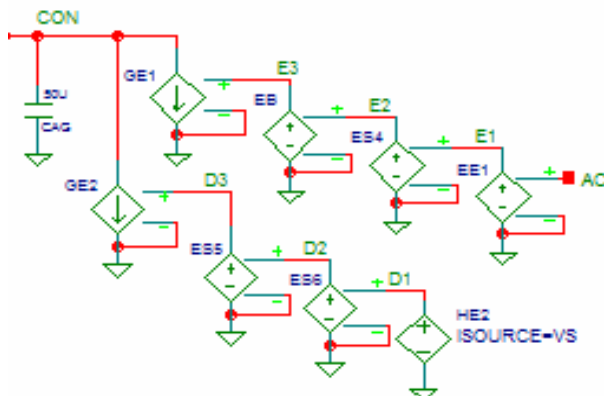


Fig. 6: Dynamic erase section of SMM for PMC

5. Experimental Results

After a number of transient simulations were run to ensure that the device model was operating similar to pulse and pseudo-DC test data, a curve fitting exercise was done. Due to the transient nature of the device, the model was designed to only be run in transient simulations. Most all of our simulations were performed in PSPICE, but due to the fact that only basic SPICE macro model elements and equations were used, the SMM can easily be run in any SPICE-based simulator with simple syntax changes for the given simulator.

In order to accurately simulate the device, the circuit used to measure the pseudo-DC IV characteristics of the device also required modelling. A model was designed to simulate the Agilent 4155 SPA to fit the data found in Fig. 7. The resistance of the same device as a function of voltage was also analysed at the same time and is presented in Fig. 8. The large deviation in resistance for the subthreshold portion in Fig. 8 is due to inaccuracies in measurement and has not been considered at this stage due to the relatively small error in simulated current.

6. Conclusions

In this paper, a SMM that more accurately describes the physical dynamic operation of the PMC device element was presented. Each section of the SMM

representing a mechanism relating to write, erase and parasitics of the device were described. Adequate curve fitting of experimental data to the simulated output was also demonstrated.

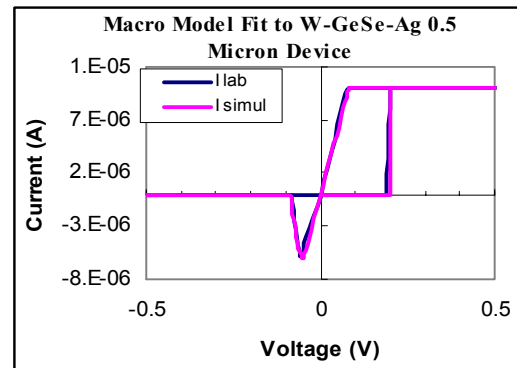


Fig. 7: Results of curve fitting simulation data with experimental IV data (D = 0.5 micron).

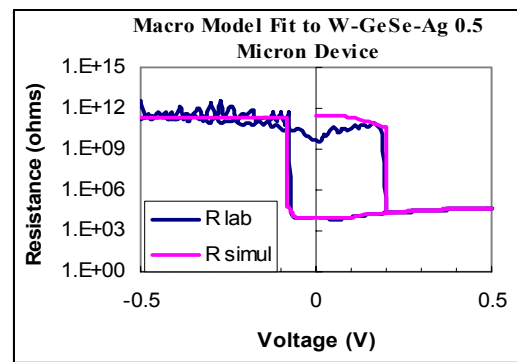


Fig. 8: Results of curve fitting simulation data with experimental RV data (D = 0.5 micron).

Acknowledgements

This work was supported by Axon Technologies Corporation. Special thanks to Desert Microtechnology Associates, Inc. for the use of their design tools.

References

- [1] M.N. Kozicki, C. Gopalan, M. Balakrishnan, M. Park, and M. Mitkova, "Non-Volatile Memory Based on Solid Electrolytes," Proceedings of the 2004 Non-Volatile Memory Technology Symposium, 10-17 (2004).
- [2] M.N. Kozicki, M. Park, and M. Mitkova, "Nanoscale Memory Elements Based on Solid State Electrolytes," IEEE Trans. Nanotechnology, in press.
- [3] M. Mitkova, M. N. Kozicki, H. Kim, T. Alford, "Local Structure Resulting From Photo- and Thermal Diffusion of Ag in Ge-Se Thin Films," J. Non-Cryst. Sol., vol. 338-340C, 552-556 (2004).
- [4] M. Mitkova, M.N. Kozicki, H.C. Kim, and T.L. Alford, "Thermal and Photodiffusion of Ag in S-Rich Ge-S amorphous films," Thin Solid Films, vol. 449, 248-253 (2004).
- [5] W.C. West, K. Sieradzki, B. Kardynal, and M.N. Kozicki, "Equivalent Circuit Modeling of the Ag/As_{0.24}S_{0.36}Ag_{0.40}/Ag System Prepared by Photodissolution of Ag", J. Electrochem. Soc., vol. 145, 2971-2974 (1998).
- [6] B. Swaroop, W.C. West, G. Martinez, M. N. Kozicki, L.A. Akers, "Programmable Current Mode Hebbian Learning Neural Network Using Programmable Metallization Cell, Proceedings of the International Symposium on Circuits and Systems, vol. 3, 33-36 (1998).

possible the particle size, such as to enable downscaling of the size of the memory element to the size of a contact hole. Simultaneously, downscaling of the size of the crystallites enhances the ON/OFF current ratio and the reliability of the devices. Preparation of the CuTCNQ layer by spontaneous electrolysis of Cu in a saturated TCNQ solution in ACN at room temperature resulted in inhomogeneous surface coverage with large crystalline aggregates (Fig. 2a). Much smoother CuTCNQ films with submicron particle size were obtained by carrying out the reaction in BuCN at 110°C (Fig. 2b). Memories prepared according this procedure were initially in OFF state, and could be switched to a low impedance ON-state, and a high impedance OFF-state by applying respectively a sufficient negative or positive voltage to the Al line. This observation agrees well with previously published results for CuTCNQ films [9] reporting that the contacts between the Al electrode and CuTCNQ likely play in essential role in the polarized memory effect. Furthermore all I-V curves crossed exactly at 0 V and 0 A, indicating that the switching behaviour does likely not rely on ionic movement inside the memory material. Good switching properties with typical ON/OFF current ratios between 10^2 and 10^3 were observed for all memory sizes (Fig. 3).

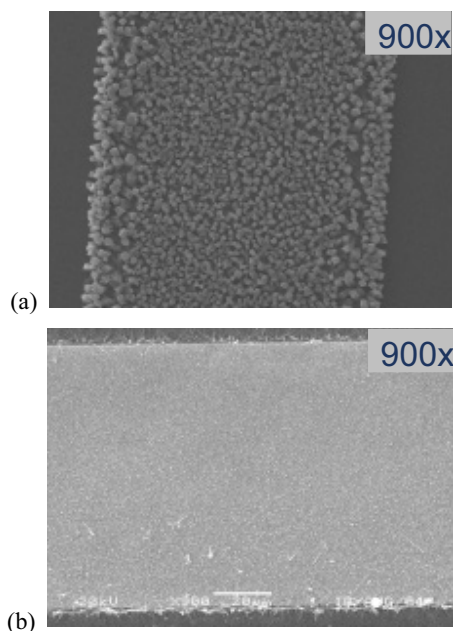


Fig. 2: Scanning electron microscopy images of Cu microlines covered by CuTCNQ prepared from (a) ACN at room temperature (10 min, 50 μm width line) and (b) BuCN at 110°C (10s, 100 μm width line).

Number of write/erase cycles

Devices were submitted to extensive ON/OFF cycling. They typically can be written and erased more than a thousand times (Fig. 4 shows ON and OFF currents measured during an experiment of 1100 switching cycles). The most observed failure mode is that a memory can no longer be turned OFF after extensive cycling. We have observed that sometimes in the course of the cycling, this failure mode sets in for a number of write/erase cycles before normal ON/OFF

switching is observed again. As a result, the histogram of the OFF-currents is too broad, and its upper tail partially overlaps with the ON-state histogram (histograms of experiment including 1100 cycles shown in Fig 4) in the current state of research. Further work is underway to improve reliable cycling to the OFF state.

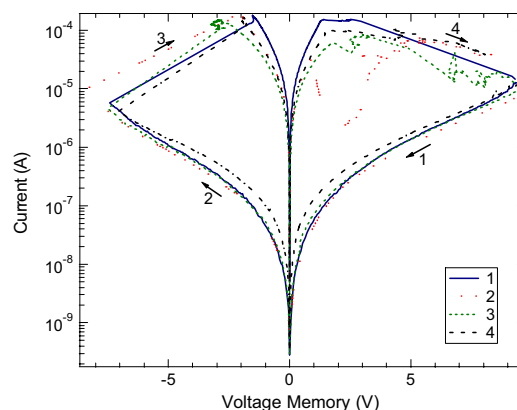


Fig. 3: Log|J|-V curve of a CuTCNQ memory 50 μm x 200 μm prepared from hot BuCN at 110°C. Load resistor $R = 47 \text{ k}\Omega$.

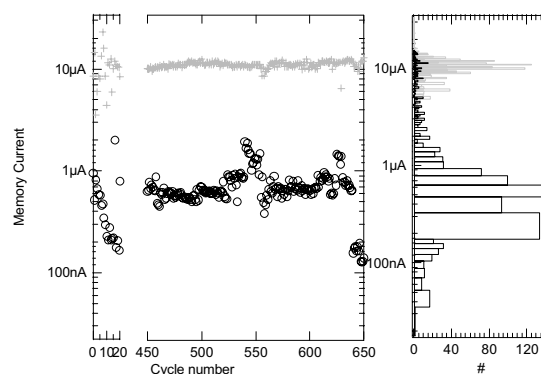


Fig. 4: On-state currents (grey) and OFF-state currents (black) measured at $V = -1\text{V}$ at the memory element, for 200 μm x 200 μm CuTCNQ memory (prepared from BuCN at 110°C), cycled repetitively over $R = 100\text{k}\Omega$ between -7V and $+7\text{V}$. Left panel: 650 memory cycles; Right panel: histograms of ON-state and OFF-state currents for 1100 cycles.

Temperature stability

Examination of the switching properties of CuTCNQ memories at various working temperatures revealed that the switching process is virtually not affected by temperatures up to 80°C (Fig. 5a,b). Reproducibility of the cycles gradually decreased at higher temperatures; leading to highly irreproducible ON and OFF switching at 120°C (Fig. 5c).

Additional measurements at room temperature with devices heated for one hour in air to 120 and 170°C showed still reproducible I-V curves (Fig. 6). This indicates that the device itself is stable to temperatures of at least 170°C and that the loss of reproducibility observed during measurements at high temperature must be related to the current passing through the device.

4. Conclusions

We are investigating the use of CuTCNQ for cross-bar memory cells exhibiting conductivity switching. Results so far show better than 1000 write/erase cycles, and good thermal stability. Work is underway to scale the device dimensions to contact-hole sizes, and to further improve the reproducibility of the switching to the OFF state. We are simultaneously also elucidating the mechanisms of the conductivity switching.

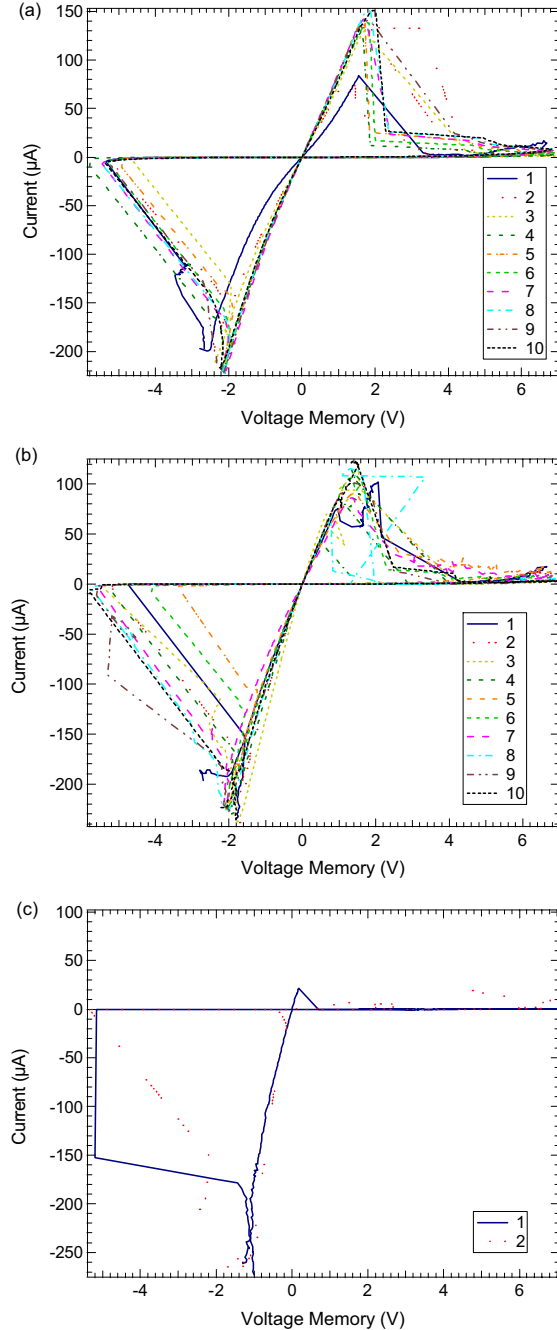


Fig. 5: I-V curves of CuTCNQ memory elements ($200\ \mu\text{m} \times 200\ \mu\text{m}$, prepared BuCN at 110°C) recorded at temperatures of (a) 40°C , (b) 80°C , and (c) 120°C ($R = 22\ \text{k}\Omega$).

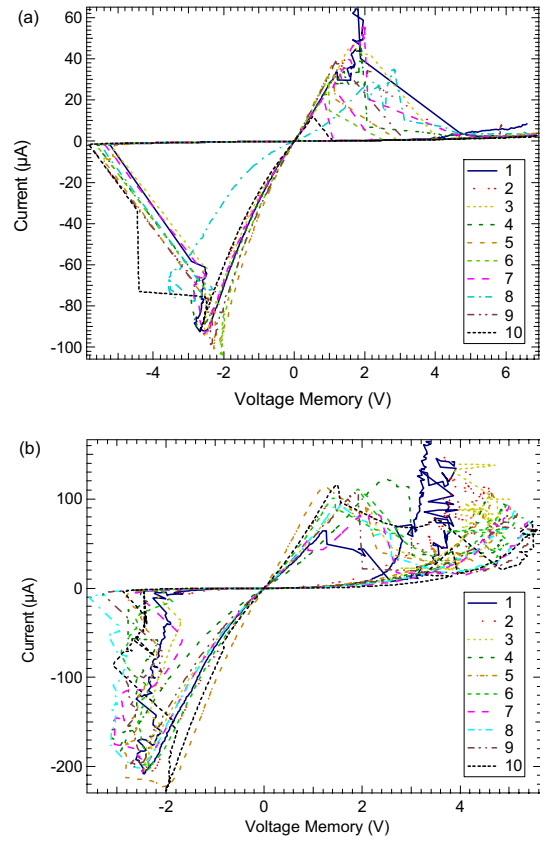


Fig. 6: I-V curves of CuTCNQ memory elements ($200\ \mu\text{m} \times 70\ \mu\text{m}$, prepared by spontaneous electrolysis from hot BuCN): (a) memory heated to 125°C for 1 hour ($R = 47\ \text{k}\Omega$) and (b) 170°C ($R = 22\ \text{k}\Omega$) before measurement at room temperature.

Acknowledgements

S. De Jonge, K. Myny and D. Aerts are respectively acknowledged for preparation of Si/SiO₂ wafers, photolithographic processing, and electrical measurements. This research was performed within the framework of the NOSCE MEMORIAS project of the European Union (FP6-507934).

References

- [1] R.S. Potember, T.O. Poehler, and D.O. Cowan, Appl. Phys. Lett. **34**, 405 (1979).
- [2] A. Bandyopadhyay and A.J. Pal, Appl. Phys. Lett. **82**, 1215 (2003).
- [3] R.M.Q. Mello, E.C. Azevedo, A. Meneguzzi, M. Aguiar, L. Akcelrud, and I.A. Hümmelgen, Macromol. Mater. Eng. **287**, 466 (2002).
- [4] D.R. Stewart, D.A.A. Ohlberg, P.A. Beck, Y. Chen, R.S. Williams, J.O. Jeppesen, K.A. Nielsen, and J.F. Stoddart, Nanolett. **4**, 133 (2004).
- [5] J. Chen, J. Su, W. Wang, and M.A. Reed, Physica E **16**, 17 (2003).
- [6] D. Ma, M. Aguiar, J.A. Freire, and I.A. Hümmelgen, Adv. Mat. **12**, 1063 (2000).
- [7] L. Ma, Q. Xu and Y. Yang, Appl. Phys. Lett. **84**, 4908 (2004).
- [8] R. Sezi, A. Walter, R. Engl, A. Maltenberger, J. Schumann, M. Kund and C. Dehm, IEDM Technical Digest 2003
- [9] C. Sato, S. Wakamatsu, K. Tadokoro, and K. Ishii, J. Appl. Phys. **68**, 6535 (1990).

Organic Electrically Bistable Materials for Non-Volatile Memory Applications

A. Pirovano^a, S. Conoci^b, R. Sotgiu^a, S. Petralia^b, and F. Buonocore^b

^a Central R&D, NVMTD, STMicroelectronics, via C. Olivetti 2,
20041 Agrate Brianza (Mi), Italy. E-mail: sabrina.conoci@st.com

^bPost Silicon Technology, FTM, STMicroelectronics, Stradale Primosole 50, Catania, Italy

Abstract

Organic electrically bistable materials exhibiting conductance switching have been investigated and their potential exploitation in non-volatile memory applications has been tested. Theoretical modelling of both geometrical and electronic structure of such materials has been carried out and tentatively correlated to the mechanisms responsible of the bi-stable switching property.

1. Introduction

Several inorganic and organic compounds have been proposed for non-volatile applications that should allow better performance or lower costs (or both) with respect to Flash technology. Among these emerging concepts, conductance switching devices are the most attractive due to their large reading signal, fast programming operations, and expected good scalability [1,2]. Recently, the investigation of voltage-induced conductance-switching phenomena in organic compounds have attracted growing interest, and novel efforts have been dedicated to the exploitation of their intrinsic low-cost capabilities and the possibility to tailor their properties with selective changes of the molecular structure. In this paper, we report a study of some Organic Electrically Bistable (OEB) compounds (having similar molecular structures) exhibiting conductance switching as potential materials for use in high density memories applications.

2. Experimental

All samples have been prepared by depositing films of the active OEB materials (thicknesses of several hundred nanometers) through spincoating on ITO substrates, followed by evaporation of the top electrode (1mm x 1mm size and 30-40 nm thickness). Annealing processes of the samples have been also performed. Absorption (UV-Vis) spectra have been recorded by a Perkin Elmer Lambda 900 instrument. Melting Point has been obtained by BÜCHI B-545 instrument.

Atomic Force Microscopy (AFM) images have been acquired in air by using a Digital 3100 in tapping mode. Commercially available tapping etched silicon probes (Digital) with a pyramidal shape tip having a nominal curvature of 10 nm and a nominal internal angle of 35° were used. Conductive Probe (CP)-AFM measurements have been carried out by Digital 3100 AFM equipped by

CP modulus. A gold tip (radius nominally 20 nm) has been used as conductive probe.

An Agilent 4155C semiconductor parameter analyser has been employed to measure the device current-voltage (IV) characteristics and to evaluate the data retention capabilities. An Hp81110A pulse generator has been instead used to investigate the programming speed. Electrical measurements have been performed on both annealed and not annealed samples, showing better results for the first type of samples. Moreover, samples stability has been assessed by comparing measurements performed in air and in vacuum (10^{-5} Torr), the last one showing a better reproducibility.

3. Physical Characterizations

Absorption spectra of these OEB compounds, recorded in methanol solution, present absorption maxima ranging between 559 and 520 nm. The energy of these bands corresponds to energy difference between HOMO (High Occupied Molecular Orbital) and LUMO (Low Unoccupied Molecular Orbital) orbitals and represents the band gap of these molecules. Thus, the values of band gaps calculated from the absorption spectra range between 2.19 and 2.53 eV.

Thermal stability measurements show that the Melting Points (MP) of these compounds range between 185 and 360°C depending of the chemical structures.

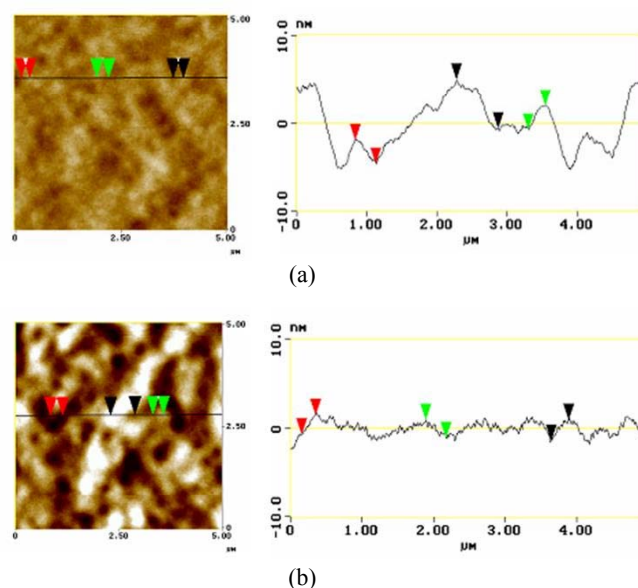


Fig. 1: AFM images of OEB materials films before (a) and after (b) the annealing processes.

Thus, the possibility of integration in a standard CMOS back-end process can be allowed, especially for the compounds exhibiting highest values of MP. The morphology of the deposited films on ITO substrate has been inspected by Atomic Force Microscopy (AFM) before and after the annealing processes. Fig.1 shows the representative results obtained. As it can be noticed, the annealing processes provoke a relevant flattening of the films. Indeed, the related RMS values decrease from 2.50 nm for the not annealed films to 1.0 nm for the annealed ones.

4. Electrical Results and Scalability Tests

Fig. 2 reports the typical current-voltage characteristic of the OBD compounds, showing the well-known bistable characteristic. As shown in the inset of the same figure, the programmed resistance ratio evaluated at 0.3 V is about 10^3 . This value is comparable to the set/reset ratio of phase-change memory devices and order of magnitude better than the reading signal ratio in other emerging technology like MRAM and FeRAM, demonstrating the good reading window that characterizes this technology. Moreover, a switching voltage lower than 1.5V for programming is a definitive advantage for low-power applications, system management issues, and scalability.

Repetitive programming with squared pulses 1 ms long were performed, showing a reproducible transition of the memory switching for several tenths of cycles. Despite these preliminary results are very promising, some issues related to the development of a suitable cap layer to avoid ambient contamination and to improve the device uniformity must be solved before assessing the real endurance capabilities of this organic molecule. Finally, data retention capabilities of the programmed state were evaluated. In these materials the stable state is constituted by the OFF conditions, while the ON state could show a progressive data lose. At room temperature, a 10% degradation is observed after 20000 seconds, suggesting a good stability of the programmed

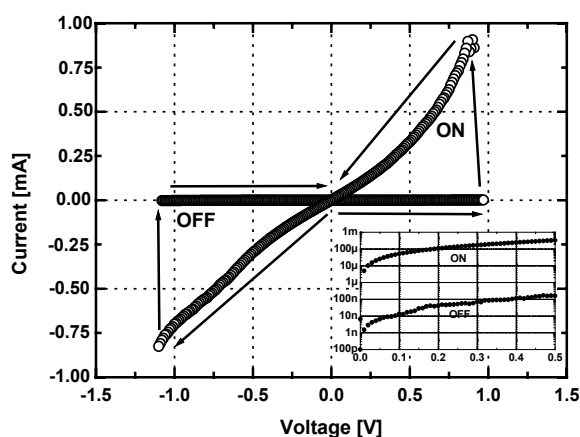


Fig. 2: Typical I-V curve of the organic electrically bistable compounds, showing the hysteretic characteristic. The inset report an enlarged view of the curve to appreciate the reading signal.

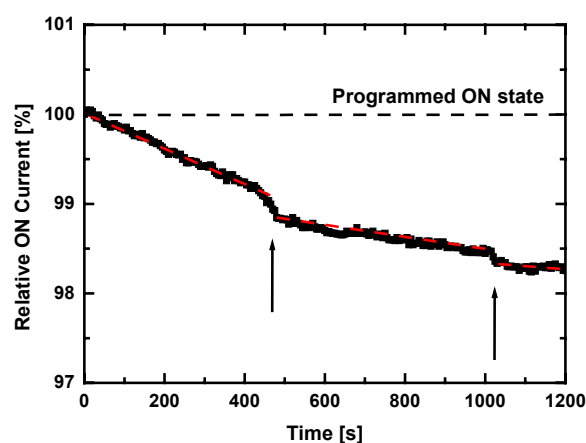


Fig. 3: Data retention characteristic showing a stair-case charge loss.

state. Also in this case measurements in air showed worse results, confirming the role of ambient contaminants in degrading the electrical properties of these materials. Finally, in Fig. 3 an enlarged view of the data retention measurement showing the ON state current normalized to the programmed value is reported. It is interesting to note a stair-case behaviour of the data loss, probably related to a mechanism of charge-loss of a single molecule that switches back to the high resistive state.

Finally, in order to perform a first assessment on the scalability of these materials, Conductive Probe Atomic Force Microscope (CP-AFM) measurements have been also carried out on OEB films. In all cases, bistable IV characteristics are observed, featured by different values of ON/OFF voltages and cycles. Fig. 4 depicts the typical IV characteristics, for both sweep directions, on both not annealed (a) and annealed (b) films.

5. Theoretical Modelling

To clarify the mechanism responsible for the bistable memory switching in organic molecules, numerical simulations employing Functional Density Theory (FDT) for single molecule calculation have been performed. In our calculations we used the DMol³ code [3,4]. As it has been observed in the electrical characterization, OEB materials undergo a relevant

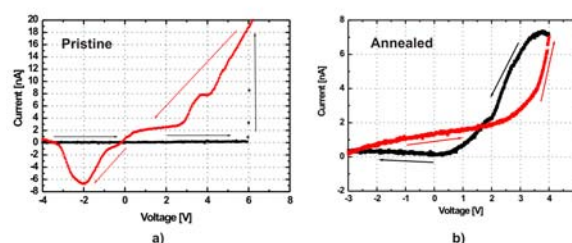


Fig. 4: Typical I-V curves obtained by CP-AFM inspection: (a) pristine films; (b) annealed films.

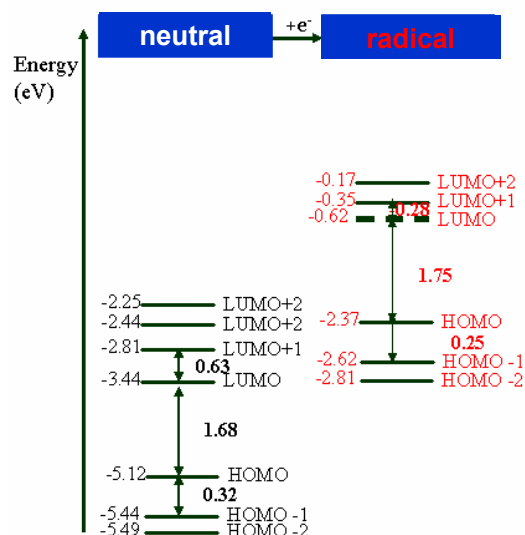


Fig.5: Simulated energy levels of a typical OEB compound.

increase of conduction when at a given threshold voltage is reached. This switch is supposed to be connected to an electron transfer from the electrode to the single molecule. Therefore it can be useful to investigate the electron structure of these organic compounds in the neutral state and when is charged with an electron (radical state).

Structural simulations of both neutral and radical states have been developed in vacuum assumption. The energy levels of a typical OEB compound in both neutral and radical state are reported in Fig. 5. As it can be noticed in the neutral state the HOMO-LUMO gap is 1.65 eV, while for radical state this is 1.5 eV. By looking at these values, not so much modification of HOMO-LUMO gap has been observed. On the other hand, when

an electron is injected in the compound (radical state) a decrease of the separation in energy of the LUMO levels has been obtained. This shrinking of the LUMO levels separation should favor the transport of electrons, corroborating the experimental results of higher conduction when electrons are injected into these molecules (ON states).

6. Conclusions

The present study has shown that the OEB compounds here investigated are promising materials to be employed as active layers in non-volatile memory applications. The integration of these materials in a standard CMOS back-end process is also possible. Finally, theoretical simulations suggest that the injection of electrons in the single molecule of these materials leads to an increasing of the conductance, corroborating in this way the experimental results.

7. Acknowledgments

This work has been performed in the framework of the NOSCE MEMORIAS project funded by the EC (FP6-507934). The authors thanks to IMM CNR (Catania, Italy) for the technical support in the Scanning Probe Microscopies measurements and the Politecnico di Milano for the electrical characterization in vacuum.

References

- [1] R. Sezi *et al.*, IEDM 2003 Tech. Dig., 259 (2003).
- [2] A. Bandyopadhyay and A. J. Pal, Appl. Phys. Letters **82** (8), 1215 (2003).
- [3] B. Delley, J. Chem. Phys. **92**, 508 (1990);
- [4] B. Delley, J. Chem. Phys. **113**, 7756 (2000).

Growth at low temperature of very high k material on silicon.

M. Detalle^a, E. Fribourg-Blanc^b, E. Cattan^a, D. Remiens^a

^a IEMN – DOAE – MIMM Team, Bâtiment P3, cité scientifique Université des Sciences et Techniques de Lille, 59652, Villeneuve d'Ascq Cedex France, detamic@wanadoo.fr

^b Present address: LETI - CEA/GRENOBLE, 17, rue des Martyrs, 38054 Grenoble Cedex 9, France

Abstract

We present the synthesis of thin films in the solid solution lead magnesium niobate with lead titanate (PMN-PT with 30 % lead titanate) by sputter deposition. The study was aimed at characterizing post deposition annealed films at various temperatures. We proved full perovskite crystallization from 400 °C to 750 °C. From dielectric measurements, it is shown that quality of material improves gradually in this temperature range to reach a maximum of dielectric characteristics at 700 °C annealing. Dielectric constant reaches 2300 at dielectric maximum for this last film. Low temperature annealing was specially studied for possible application in above IC.

1) Introduction

With the development of the mobile phone, electronic functions, ie. camera, photography, internet, are rapidly expanding. In parallel the dimensions of the telephone setup decrease. To be able to fit all these electronic functions in a very small space, new materials with very high permittivity should necessarily be integrated, in the aim to replace the decoupling capacitors. In this context, we have developped a new material, well known in the bulk form to reach very high dielectric permittivity: the PMN-PT.

2) PMN-PT potentialities

In terms of materials, PMNT ($x \text{ Pb}[\text{Mg}^{1/3}\text{Nb}^{2/3}\text{O} - (1-x) \text{ PbTiO}]$) provide a versatile choice for decoupling capacitors. In the range of $x < 0.35$, PMNT show relaxor ferroelectric property [1]. This gives rise to a high dielectric constant (high-k material) and a diffuse phase transition for low lead titanate concentrations. Usually, normal ferroelectrics materials are characterized by sharp dielectric peak at

the transition where dielectric permittivity is the highest. This transition characterized by its Curie temperature exhibits high dielectric constant greater than 10,000 for bulk material. Relaxor ferroelectrics show a smoother dielectric variation with temperature because of their diffuse phase transition and can still be used in this range if properly engineered [1 – 6]. However, relaxor ferroelectrics exhibit a frequency-dependent dielectric maximum temperature.

3) Experimental method

The PMN-PT films were deposited by rf magnetron sputtering on silicon substrate. A 3-inches sputtering target made of mixed PbO , MgNb_2O_6 and TiO_2 powders was used [7]; a composition of $x = 0,30$ was chosen for being close to morphotropic phase boundary but still in the relaxor range. The deposited films are amorphous and a post annealing treatment is necessary to crystallise the films in the perovskite phase. We demonstrate that with a conventional annealing the pure perovskite phase without second phase is obtained at very low temperature: 450°C. This temperature is largely compatible with the silicon technologies and in particular the above IC.

We have systematically measured the relative permittivity and the loss factor of the films with the help of an HP4192A impedance analyser.

4) Results and discussion

For films post annealed at 400°C the relative permittivity reaches 500, while a maximum room temperature permittivity is obtained (2300) for a post annealing treatment at 700°C (Fig. 1).

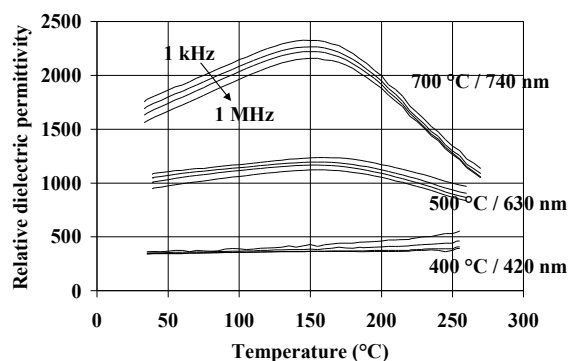


Fig. 1 : Dielectric constant versus temperature for films at various annealing temperatures.

Then, we chose to study low temperature annealing (450°C), to be compatible with above IC technology. As we can see on figure 2, relative dielectric constant reaches 400 at room temperature for a film thickness of 200 nm. Dielectric constant raises with thickness up to 800 for a 700 nm film.

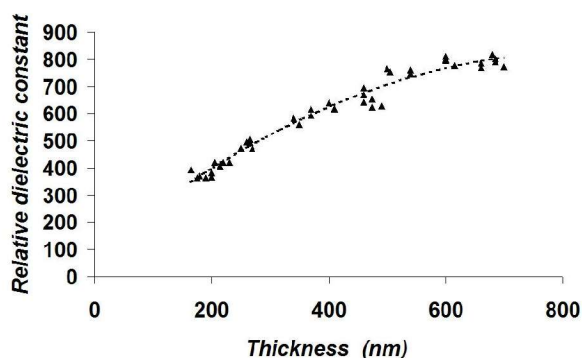


Fig. 2 : Dielectric constant measured at room temperature versus thickness for films annealed at 450°C

5) Conclusion

In this study we have shown that it was possible to obtain PMN-PT thin films on platinum with good dielectric and ferroelectric properties with an annealing temperature above 450°C. It is now possible to consider the integration of PMNT thin films in applications like IC components which require low process temperature.

A mechanism which could explain low temperature crystallization has been investigated. Our study focused on interfaces phenomenon between PMNT thin films and platinum bottom electrode.

All these results will be presented at the conference and a model will be proposed to explain the

perovskite phase formation of PMN-PT at these low temperatures.

References

- [1] Zhang Q M, Zhao J, Cross L E. Aging of the dielectric and piezoelectric properties of relaxor ferroelectric lead magnesium niobate-lead titanate in the electric field biased state. *J Appl Phys* 1996;79(6):3181-7.
- [2] Tavernor A W, Li H-P S, Stevens R. Production and characterisation of composite relaxor ferroelectric multi-layer structures. *J Eur Ceram Soc* 1999;19:1859-63.
- [3] Park S-E, Shrout T R. Ultrahigh strain and piezoelectric behavior in relaxor based ferroelectric single crystals. *J Appl Phys* 1997;82(4):1804-11.
- [4] Kim Y-M, Lee W-J, Kim H-G. Deposition of PZT films by MOCVD at low temperature and their change in properties with annealing temperature and Zr/Ti ratio. *Thin Solid Films* 1996;279:140-4.
- [5] Li X-S, Tanaka T, Suzuki Y. Characterization of lead zirconate titanate thin films deposited at low temperature by reactive facing target sputtering. *Thin Solid Films* 2000;375:267-70.
- [6] Baek J-G, Gomi K, Isobe T, Senna M. Flat dielectric properties of capacitor $(1-x)\text{Pb}(\text{Mg}_{1/3}\text{Nb}_{2/3})\text{O}_3\text{-xPbTiO}_3$ by controlled concentration gradient. *Mat Sci Eng B* 1997;49:46-52.
- [7] Fribourg-Blanc E, Cattani E, Remiens D, Dupont M, Osmont D. Rf-sputtering of PMNT thin films. *J Phys IV* 2001;11:145-9.

Analysis of US patents (1966-November 2004) in phase change memory

Semyon D. Savransky^a and Tz-Chin Wei^b

^a The TRIZ Experts, 6015 Pepper Tree Court, Newark, CA 94560, USA E-mail: chalcogenide_glasses@yahoo.com

^b The TRIZ Experts, 11F, #1, Lane 228, JungAn St., JungHe City, Taipei 235, Taiwan E-mail: peterwei@ms9.hinet.net

Abstract

The initial results of US patents analysis in phase change memory field in the framework of TRIZ are presented. The statistics shows that the development of non-volatile semiconductor (electrical) phase change memory occurs now at the second stage of technique evolution.

1. Introduction

The famous paper "Reversible Electrical Switching Phenomena in Disordered Structures" [1] by Stanford R. Ovshinsky was published in November 1968 and it created new field of phase change memory (PCM) devices. At the first time we analyze development activities in PCM in the framework of TRIZ methodology [2-5]. We believe that TRIZ can be useful for progress in memory technology. The prime goal of this research is to apply TRIZ to non-volatile PCM development. In order to achieve it we should:

- 1) Figure out patents levels and heuristics already used in the field;
- 2) Study trends of non-volatile PCM evolution;
- 3) Identify problems important for PCM development and predict difficulties that this technique might face during implementation;
- 4) Select useful generic heuristics from TRIZ database for further memory developments;
- 5) Find possible solutions to these problems using TRIZ methodology.

2. Short TRIZ Overview

TRIZ¹ is a structured problem solving methodology that help apply known generic heuristics to new problems [2-5]. TRIZ author, Genrich S. Altshuller, used to study patents in order to show that the same generic heuristics can be re-applied in various technical fields. Altshuller distinguished the five levels of patents between trivial legal documents (with tiny inventiveness) and truly breakthrough discoveries:

Level 1. Routine design problems solved by methods well known within the same technical field or inside a company. (About 32% of the solutions occurred at this level).

Level 2. Minor corrections (~45%) to an existing technique, by methods known within the industry.

Level 3. Fundamental improvements (~18%) to an existing technique which resolve contradictions, by methods known outside the industry.

Level 4. New generations (~4%) using a new scientific (rather than technological) principle to perform the primary functions of the technique.

Level 5. Rare scientific discoveries or pioneering inventions (less than 1%) of essentially a new technique.

The original Ovshinsky's patent US3,271,591 corresponds to the 5th level, ECD's patent US5,335,219 fits to the 4th level, and recent Nanochip's US2004/0145848 patent application matches the 3rd level, while some of recent (Ovonyx, Intel, Micron) patents belong to the levels 1 and 2. It is worth to note that these low levels solutions are necessary for a technique implementation. TRIZ maintains that a technical problem can have few solutions at high (3-4) levels. The TRIZ methodology offers several tools for finding these solutions based on the generic heuristics usage [2-5].

After examinations of numerous high-level patents, G. S. Altshuller and other TRIZ experts identified and codified a common set of inventive heuristics used across numerous technical fields from mechanical to software engineering. More than 100 such heuristics are recognized now [5], while inventors in a particular field (including PCM) use often intuitively less than 10% of them. Many PCM inventors continuously employed the same approach for solution of the problem they faced. It points to the existence of strong psychological bias of thinking inside the box for most of these developers. TRIZ helps to overcome such limits [2,3].

Improvement of one of the technique parameters often leads to deterioration of others or the same in different mode. The basic concept of TRIZ is the recognition and resolution of a **contradiction**. A contradiction arises from mutually exclusive demands that may be placed on the same or related (sub)system of a technique. The contradictions are the hidden root(s) of the technical problems and their resolution leads the technique toward an **ideal system** along different trends of the **evolution**.

¹ Created in USSR in the middle of the XX century and unknown to Western world until recently (~1995), TRIZ is now a popular problem solving methodology in many countries. Currently, many Fortune 500 companies (including Procter & Gamble, Samsung, Mitsubishi and BMW) use TRIZ to design better products and less complicated processes, to understand disruptive technique evolution trends and to improve their handling of intellectual property.

As a technique evolves during a life cycle of birth, childhood, growth, maturity, and decline it should become better, i.e. its performance follows so-called S-curve (see the figure 1) [2,4]. Usually during a technique evolution its ability to satisfy human needs (e.g., faster and bigger memory) increases while its cost and power consumption decrease, i.e., the technique becomes more ideal. The ideal system, according to the TRIZ philosophy, is a non-existent technique with all of its functions still being executed. This ideal system, analogous to the definition of limit in mathematics, is unrealizable in practice. For example, any memory chip contains many subsystems that help but do not perform the primary functions (e.g., information retention) and keep a chip away from ideal.

To understand the stage where a technique is on its evolution, TRIZ suggests to use two metrics from patent analysis, i.e. number of patents and level of invention in addition to conventional metrics of performance and profit. Figure 1 shows the observed trends in these metrics and the correlations among them common for various techniques [2].

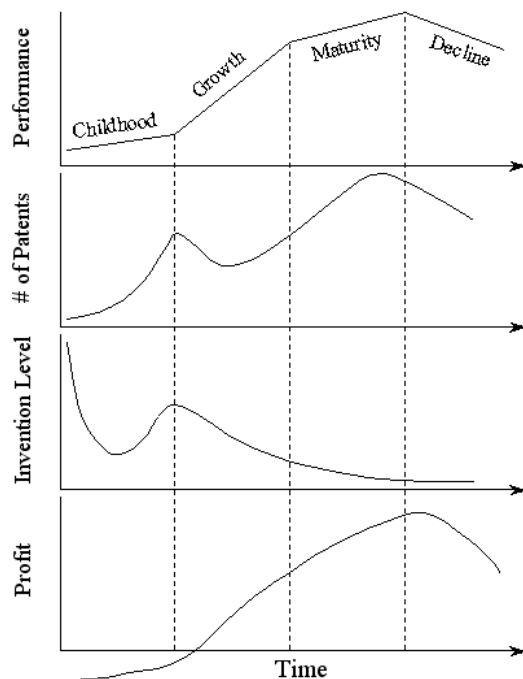


Figure 1. Correlations of four metrics for determining the stage of technique evolution.

3. US Patents in semiconductor phase change memory field

We have analyzed about 500 US patents issued from 1966 up to November 2004² for chalcogenide PCM. In this paper we report results for more than 300 patents

important for non-volatile electrical memory³. Table 1 shows that these patents legally issued only in one country are quite representative because most of inventors and assignees have USA jurisdiction.

The level of invention is determined by analyzing the patents in the framework of TRIZ [3]. Figure 2 shows the results of the preliminary analysis. The first level patents issued recently are omitted for clarity.

Country	# patents	# assignee
United States	302	27
Japan	14	7
Netherlands	4	1
Italy	3	1
Taiwan	3	1
Israel	2	1
France	1	1
United Kingdom	2	2
Republic of Korea	1	1

Table 1. PCM patents geography. Top assignees are Energy Conversion Devices / Ovonyx, Inc.; Micron Technology, Inc. and Intel Corporation – all USA. Only 30 patents (less than 10%) were issued for 15 non-US assignees.

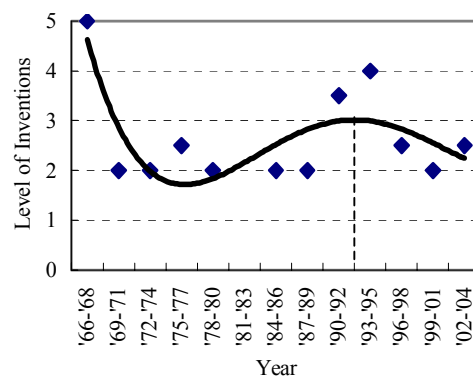


Figure 2. Estimated highest level of inventions issued in 3-year period in chalcogenide PCM field.

Compare the figure 2 and the third graph in the figure 1 we figured out that semiconductor non-volatile PCM that was born in 1962-1963 just moved into the second (growth) stage of technique evolution [2,4] due to many supportive patents on the low levels. Figure 2 shows that this transition occurred in around 1991-1994 after quite long childhood (1966-1990). The heuristics used for innovative developments of PCM are recovered⁴ along with various statistical results about players, problems and proposed solutions in the field. Only most important statistical results are reported in this paper.

² It was arbitrary chosen and to coincide with the Conference deadlines. We are going to add newest patents and expand database to European, Japanese, etc. patents in the future.

³ Our optical phase change memory patents analysis will be reported in www.triz-journal.com in 2005.

⁴ The already applied and perspective heuristics will be reported elsewhere.

The simple analysis of US patent numbers shows (the figure 3) that the speed for issue of PCM patents has recently increases in about 100 times that also indicates the transition to the second (growth) stage of the technique evolution [2,4].

The table 2 confirms this conclusion for number of patent applications. Number of assignees (mostly companies participated in PCM development) also grew up from one in 1964-1974 to three in 1984-1989 to about ten at the turn of the XX century and to more than twenty in 2004.

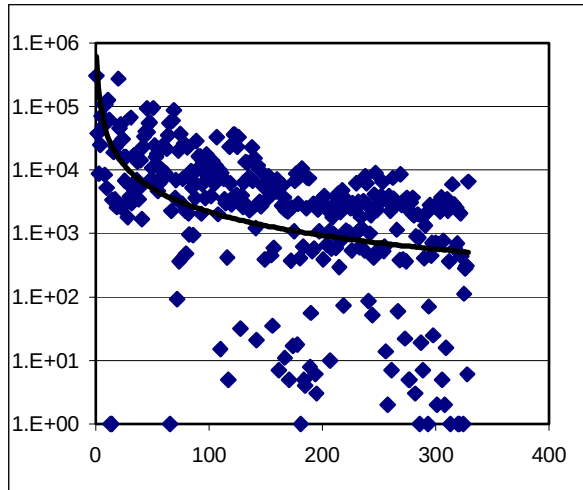


Figure 3. Frequency of US patents issued in chalcogenide PCM field. The Y-axis shows the difference between 2 closest patent numbers in logarithmic scale (e.g., the difference between US patents 3,619,732 and 3,644,741 - which are closest to each other in the field - is about 25000). The X-axis presents linear patent order analyzed in this work from first US 3,271,591 up to the latest US 6,822,903.

Years	# of Applications	# of Assignees
1963-1965	1	1
1966-1968	2	1
1969-1971	5	1
1972-1974	1	1
1975-1977	8	2
1978-1980	3	1
1981-1983	3	1
1984-1986	15	3
1987-1989	8	3
1990-1992	13	4
1993-1995	22	5
1996-1998	41	9
1999-2001	128	19
2002-2004	82	22

Table 2. Time frame for successful patent applications and assignees in chalcogenide phase change memory field.

The table 3 shows patents distribution according to the top (most frequent) classes of the International Patent Classification (IPC) for major assignees. The five most active IPC classes listed in table 3 represents about a

quarter of the all subclasses in this study. These 5 classes cover about 94% of all PCM patents issued in USA for the period of time under the investigation. Top ten assignees hold about 84% of patents in the field. ECD's key patents reflect wide diversity of technical knowledge of this company, while Micron's, Intel's and Ovonyx's patents show their high activities in non-volatile semiconductor PCM.

The figure 4 in comparison with the second graph of the figure 1 also confirms recent explosion of activities in the field and transfer of chalcogenide phase change non-volatile memory technology into the second growth stage of technique evolution.

IPC Subclasses	H01L	G11C	G11B	G06F	H03K
Assignee					
BAE	4	0	0	0	0
ECD	35	18	0	2	4
HP	8	15	2	0	0
Hitachi	4	0	0	0	0
Intel	24	8	0	0	0
IBM	0	1	2	1	0
Matrix	4	17	0	1	0
Micron	72	10	0	0	0
Ovonyx	23	18	0	0	0
SST	2	0	0	0	0

Table 3. Activities of ten major players in the field of semiconductor non-volatile PCM.

(Full names of the companies as mentioned in the assignee fields of the patents are: BAE Systems Information and Electronic Systems Integration, Inc.; Energy Conversion Devices, Inc.; Hewlett-Packard and Hewlett-Packard Development Companies.; Hitachi, Ltd.; Intel Corporation; International Business Machines Corporation; Matrix Semiconductor, Inc.; Micron Technology, Inc.; Ovonyx, Inc.; Silicon Storage Technology, Inc. in the order given in the table.)

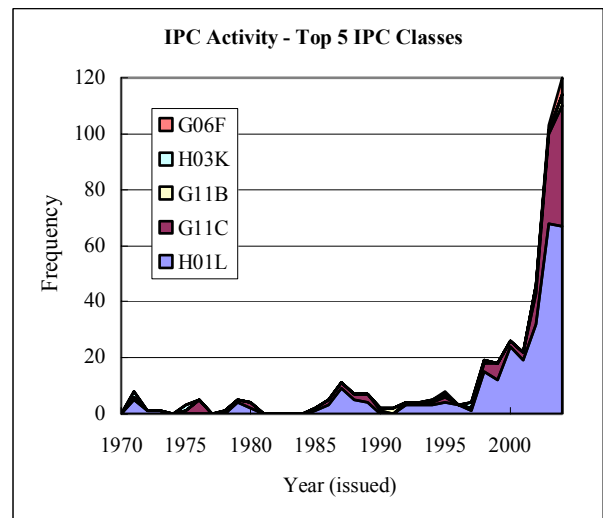


Figure 4. Top IPC classes distributions in time. Most activities are related to "Other Semiconductor Devices" (H01L) and "Static Information Storage" (G11C) classes.

4. Conclusion

We have analyzed few hundreds US patents selected in the field of optical and semiconductor PCM in the framework of TRIZ. Such patent analysis helps to figure out used heuristics, to prioritize technical problems and to foresee possible solutions, to identify evolution trends for PCM technique and to envisage future non-volatile memory roadmaps. The statistical results reported in this paper indicate that the semiconductor non-volatile PCM grows up presently at the second stage of the technique evolution. We figure out that only a few heuristics are used by inventors in the PCM field while several other generic TRIZ heuristics seems promising for this technique development. These results will be reported elsewhere.

References

- [1] S. R. Ovshinsky, Phys. Rev. Lett., **21**, 1450 (1968).
- [2] G. S. Altshuller, *Creativity as an Exact Science: The Theory of the Solution of Inventive Problems*, New York: Gordon and Breach, 1984.
- [3] S. D. Savransky, *Engineering of Creativity: Introduction to TRIZ Methodology of Inventive Problem Solving*, CRC Press, 2000, 408 pp.
- [4] G. G. Doncean, Y.P. Salamatov, S.D. Savransky, *Theoria Evolutiei Sistemelor Tehnice*, RO-INI, 2003, 237 pp. (in Romanian).
- [5] M. A. de Carvalho, S. D. Savransky, Tz-Chin Wei, *121 Heuristics for Solving Problems*, LuLu, 2004, 346 pp. (ISBN: 1-4116-1689-8).

Effects of threshold switching and parasitic capacitance in the programming transient of chalcogenide phase-change memories

D. Ielmini^a, D. Mantegazza^a, A. L. Lacaita^a, A. Pirovano^b and F. Pellizzer^b

^a DEI, Politecnico di Milano, piazza L. da Vinci 32, Milano, Italy (e-mail: ielmini@elet.polimi.it)

^b STMicroelectronics, Central R&D, via C. Olivetti 4, Agrate Brianza, Italy

Abstract

We investigated the effects of threshold switching in the programming dynamics of phase change memory (PCM) cells. We show that the parasitic capacitance of a stand-alone memory cell causes a large current overshoot in correspondence of the amorphous chalcogenide threshold switching. Experimental evidence for the parasitic current pulse is provided, and it is shown that the current peak can cause a sudden melting and quenching of the chalcogenide material, as in the reset operation. The impact on the programming and phase-change dynamics of the PCM is discussed, referring to both stand-alone cells and integrated array cells.

1. Introduction

The scaling down of conventional Flash cell beyond the 45 nm node is today considered a difficult task, due to physical limitations of the tunnel oxide thickness which must fulfil the 10 year data retention criterion [1]. As a result, alternative memory concepts suitable for non-volatile storage are being investigated. One of the most promising memory technology is the phase-change memory (PCM), where the memory switching of a chalcogenide material (typically $\text{Ge}_2\text{Sb}_2\text{Te}_5$, or GST) is used to store the logic bit [2,3]. A current pulse (reset) leads to melting and quenching in the active layer, resulting in a high-resistance amorphous phase (see Fig. 1a). A lower current pulse (set) causes crystallization in the amorphous GST, yielding a low resistance (Fig. 1b-c). Programming with few tens of ns have been demonstrated, together with a large cycling endurance (10^{12} set/reset cycles) and a good data-retention below 100°C [2,4]. In addition, due to the relatively simple design, the PCM is largely scalable, and no limitations have been pointed out down to at least 22 nm [2].

Despite the large potential of the PCM concept, many conceptual and practical challenges have to be addressed. In particular, large efforts are being dedicated to the analysis of the programming dynamics [5-7]. Experimental studies have shown that the set operation on the amorphous GST (necessary to switch the memory from a high to a low resistance) results in a stacked distribution of crystalline and (residual) amorphous phases (Fig. 1b) [5,6]. On the other hand, numerical simulations of the PCM programming based on a self-consistent electro-thermal and phase change model indicates that threshold switching results in localized current paths within the amorphous volume, leaving crystalline filaments after set (Fig. 1c) [7].

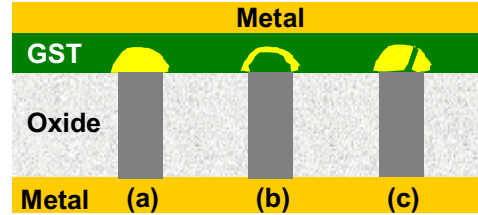


Fig. 1: Schematic of three PCM cells, after reset (a), and set operation (b,c). Amorphous and crystalline GST phases are shown as light and dark colours, respectively.

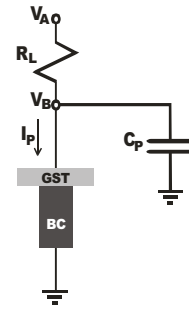


Fig. 2: Schematic view of the standard set-up for electrical characterization of stand-alone PCM cells. The parasitic capacitance C_P at the contact node is also shown.

In this work, we study the programming dynamics of the PCM, pointing out the effects of threshold switching in presence of a parasitic capacitance. It is shown that switching leads to a current overshoot, pushing the GST temperature above the melting point. We provide evidence that the current overshoot leads to a parasitic reset, significantly affecting the programming dynamics and phase distribution. The occurrence of the parasitic reset in the integrated array cell is discussed, referring to the PCM capacitance in that case.

2. Experimental set-up

The experimental set-up for characterizing current-voltage (I-V) and programming curves of PCM cells is shown in Fig. 2 [4,6]. The voltage pulse V_A is applied to the stand-alone PCM cell with a series resistance R_L . An electrometer is connected to node B to probe V_B on top of the memory cell, then the current is estimated from:

$$I_P = \frac{V_A - V_B}{R_L}, \quad (1)$$

where a load resistance $R_L = 2.5\text{k}\Omega$ was used.

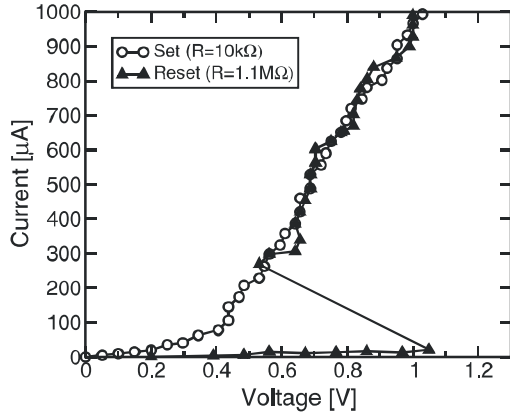


Fig. 3: Measured I-V characteristics for PCM cells with different resistance R . The cell was prepared in two different resistance states by the application of set and reset programming pulses, resulting in $R=10\text{k}\Omega$ (set state) and $1.1\text{M}\Omega$ (reset state).

Fig. 2 shows the I_P - V_B characteristics, that is the current-voltage characteristics of a PCM cell, for two different values of the low-field resistance ($R=10\text{k}\Omega$ and $1.1\text{M}\Omega$, measured at a read voltage of 0.2V). The different resistance values were obtained with the application of either a set or a reset pulse, leading to the low- R set and the high- R reset states respectively. The curve for the reset state display a very high resistance for small currents (OFF state), then threshold switching occurs once a specific voltage (the threshold voltage V_T) is reached. Threshold switching, which characterizes most amorphous chalcogenide materials, consists in an abrupt switch of the conduction behaviour to a dynamic ON state, characterized by a relatively small resistance and large current. As a consequence of the particular set-up with the series load resistance, threshold switching also results in a V_B snap-back. The I-V curve for the set state does not feature any threshold switching process, although a largely non-linear conduction characteristic can be observed. For cell states, the linear characteristic in the ON region is limited by the bottom contact (BC, see Figs. 1,2), with a resistance $R_{ON}=0.5\text{k}\Omega$.

2. Effects of parasitic capacitance

Fig. 2 also shows the parasitic capacitance C_P in parallel with the PCM cell. To estimate the capacitance value, we used the dynamic characterization procedure shown in Fig. 4. The figure shows the falling edge of the measured V_B waveform, in response to a sub-threshold voltage pulse with amplitude V_A ($V_A=0.5\text{V}<V_T$) applied to a PCM cell in the reset state. The exponential falling-edge of the V_B pulse features a time constant which is given by:

$$\tau = R \parallel R_L C_P \approx R_L C_P \quad (2)$$

where R is the OFF-state resistance of the PCM cell. The approximation in Eq. (2) is possible due to the very large OFF resistance of the reset PCM ($R=1.2\text{M}\Omega$ in the figure). An average $C_P=12\text{pF}$ was obtained averaging the results over about 100 measurements.

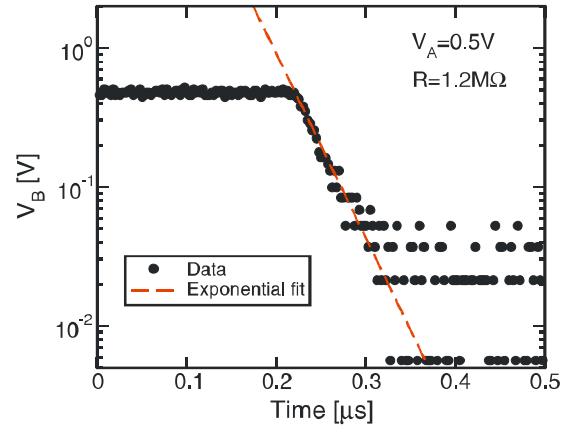


Fig. 4: Measured V_B waveform obtained with the application of a sub-threshold voltage pulse with amplitude $V_A=0.5\text{V}$, to a PCM cell in the reset state with a resistance $R=1.2\text{M}\Omega$.

The parasitic capacitance is due to several contributions at node B, including the contact tip for electrical measurement, the contact pad on the wafer, the active probe used for sensing V_B and the capacitance of the chalcogenide cell itself. Due to the relatively large capacitance value, the impact of C_P on the programming transient cannot be neglected. In particular, the current I_P flowing in the PCM must be obtained according to the improved formula:

$$I_P = \frac{V_A - V_B}{R_L} - C_P \frac{dV_B}{dt} \quad (3)$$

which results from a simple analysis of the electrical network. The capacitive component of I_P in Eq. (3) can contribute significantly to the cell current, especially for abrupt changes of V_B as in the threshold switching transient. This can result in strong effects during the set operation, i.e. the programming pulse applied to a reset cell to provide crystallization and resistance drop to a low value.

3. Programming transient

We studied the effects of the capacitive component of I_P in the programming transient during the set operation. The PCM cell was first prepared in a high-resistance reset state ($R=1.5\text{M}\Omega$) by applying a reset current pulse of 1mA for 50ns . Then we applied a programming square pulse V_A of amplitude 1.3V and width $1\mu\text{s}$ to the cell. Fig. 5 shows the measured V_B and calculated I_P during the programming pulse. The measured V_B is initially close to the applied V_A , consistently with the high-resistance behaviour of OFF-state amorphous chalcogenide. After a delay time of about 300ns from the start of the pulse, V_B suddenly drops to about 0.55V , as a result of threshold switching snap-back. Also the current I_P , calculated according to (3), reflects the change from high to low resistance of the active material. Most important, in correspondence of the switching event a large current overshoot of about 1.25mA is detected: this is the capacitive component of Eq. (3), resulting from the discharge of C_P through the PCM when V_B suddenly snaps back.

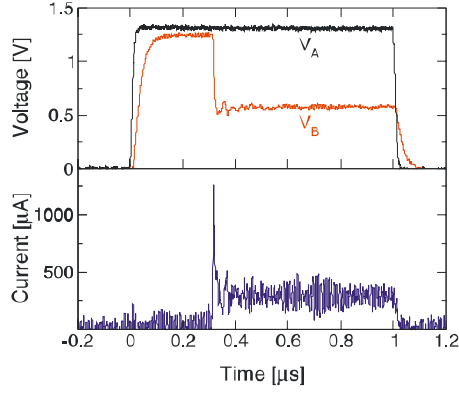


Fig. 5: Top: measured voltage V_B as a function of time, in response to an applied squared voltage pulse with amplitude $V_A=1.2V$ and width $1\mu s$. Bottom: corresponding current I_P calculated by (3), showing the current overshoot in correspondence of the threshold switching at $t=250ns$.

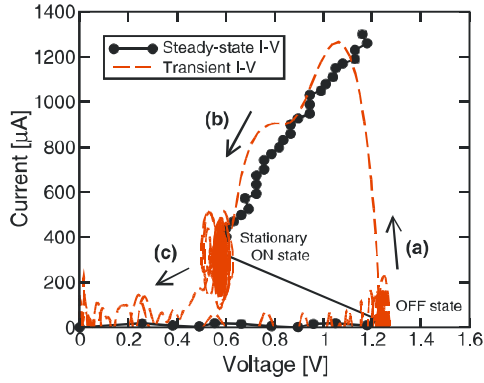


Fig. 6: Steady-state and transient I-V characteristics of the PCM device. The transient curve is obtained plotting I_P - V_B values from Fig. 5.

Fig. 6 shows the I_P - V_B characteristics, comparing the transient behaviour obtained from the V_B and I_P waveform of Fig. 5 to the “stationary” curve, evaluated from Eq. (1) in analogy with data in Fig. 3. In particular, the current I_P in the stationary characteristics was evaluated after sufficient time from the switching event, thus allowing for a sufficient decay of the capacitive contribution of Eq. (3). The device is initially biased in the high-resistance OFF state, then at the switching time the current suddenly raises to the ON-state I-V (arrow a), while the voltage remains almost constant as a result of the large parasitic capacitance. The stationary condition is then reached along the ON-state characteristics (arrow b), where the cell remains for about $750\mu s$. Finally, as the voltage pulse is switched off the current vanishes according to the high-resistive characteristic (arrow c). Note that the current overshoot during the switching transient largely exceeds the reset current of $1mA$ which initially prepared the memory in the high-resistance state with $R=1.5M\Omega$. The effects of such current overshoot are thus to be carefully evaluated for the understanding of the programming dynamics.

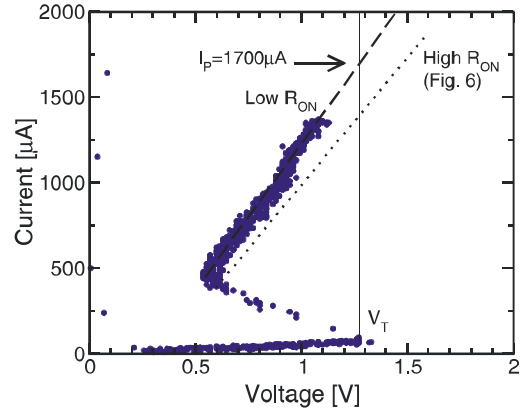


Fig. 7: Measured I-V characteristic for a PCM with relatively low ON-state dynamic resistance.

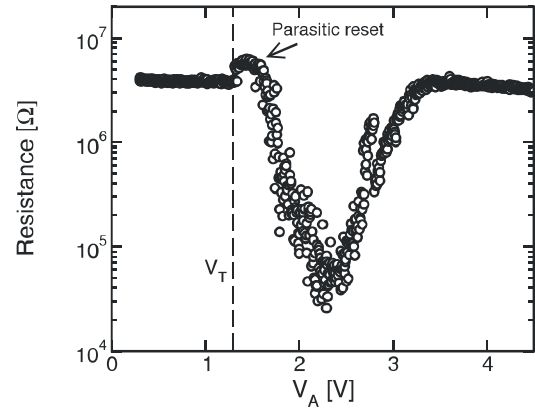


Fig. 8: R measured on the cell in Fig. 7 after programming, as a function of V_A . Note the peak revealing the parasitic reset.

4. Evidence for parasitic reset

According to Fig. 5, the estimated time-width of the current overshoot is about $20ns$ at most. Although the peak current is comparable to that used in the reset operation, the effective rise of the GST temperature above the melting point should be demonstrated. To provide direct evidence for GST melting due to the current overshoot, we characterized the programming curve for cells with relatively low contact resistance R_{ON} , in order to maximize the peak current at switching. Fig. 7 shows the I-V curve of one selected cell: at V_T , the ON current is about $1.7mA$, far larger than the reset current ($1.35mA$) used to prepare the cell prior to the application of the program pulse. Fig. 8 shows R measured on the cell in Fig. 7 after a programming pulse with voltage V_A . The three typical regions of the programming curve can be clearly identified [5]: for $V_A < V_T$, the OFF-state programming current does not affect the original phase, hence resistance of the reset cell. Above V_T , a relatively large current flows in the ON-state cell, leading to substantial crystallization and resistance drop (set operation). Finally, above $V_A=2.2V$, the resistance increases as the current is enough to melt the active volume and form the amorphous phase (reset operation). Most interestingly, programming with V_A just above V_T causes a resistance *increase* with respect to the initial reset state, which was never reported previously.

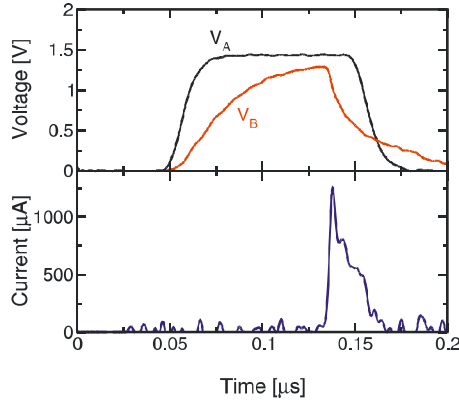


Fig. 9: Top: measured V_B waveform as a result of the application of a squared voltage pulse with amplitude $V_A=1.45V$ and width 100ns. Bottom: corresponding current I_P calculated by (3). Threshold switching and current overshoot occur about 15ns before the pulse ends.

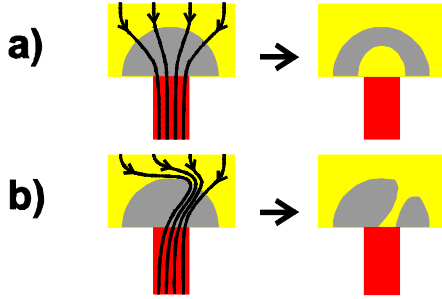


Fig. 10: Schematic of the current distribution (left) and the final phase distribution after the set operation (right). The cases of a stand-alone cell with relatively high C_P (a) and the integrated cell with negligible C_P (b) are shown.

To investigate the resistance increase in Fig. 8, we measured the V_B and I_P waveform for V_A pulses just above threshold voltage. Fig. 9 shows the transient behaviour of V_B and I_P when a square voltage pulse with $V_A=1.45V$ and width 100ns is applied to a cell in the reset state. It is shown that, since the threshold switching occurs close to the end of the pulse width, the current I_P vanishes just after the switching overshoot, and no stationary ON current flows for sufficient time to provide crystallization as in the set operation. The only phase-change effects thus result from the current overshoot, and the resistance increase in Fig. 8 demonstrates that this is able to provide GST melting and quenching as in the standard reset operation.

5. Impact on set operation

The parasitic reset, which is inherent to any set pulse applied to a stand-alone cell, leads to the formation of a new amorphous-GST volume, where the current flows uniformly in the programmable volume as shown in Fig. 10a. The final phase distribution will thus reflect the uniform current distribution and heating, resulting in a hemispherical crystalline volume within an amorphous shell, in agreement with the already reported series distribution of phases after the set operation [5,6].

Note that this type of phase change is a consequence of the parasitic reset, and occurs in stand-alone PCM cells previously prepared in the high-resistive state and in presence of a high parasitic capacitance. On a more quantitative basis, the parasitic reset given by the capacitive term in (3) is significant as $R_L C_P \gg \tau_s$, where τ_s is the switching time needed for completing the chalcogenide transition from the OFF- to the ON-state (τ_s is about 100ps [8]). In a real PCM cell array, the cell resistor is connected to an integrated select-MOSFET [9], thus parasitic C_P and series resistance R_L are far smaller than the stand-alone case. As a result, the time constant $R_L C_P$ be estimated in the ps range for integrated array cells, which is sufficiently small compared to the switching time. As a result, no parasitic reset is expected in the set operation on real-device cell structure. Instead, threshold switching will preferentially occur around corners of the amorphous GST where the electric field is maximum (see Fig. 10b), resulting in a strongly localized current and in a filament-like crystalline shape [7].

6. Conclusions

We have studied the programming transient in stand-alone PCM cells, pointing out the role of parasitic capacitance at switching. A detailed electrical cell description shows that a parasitic reset occurs during the set operation. The parasitic current is due to the typically large capacitance in parallel with the cell resistor, and is not present when the cell is connected to the select transistor, as in the integrated memory array.

7. Acknowledgments

The authors gratefully acknowledge A. Redaelli (Politecnico di Milano) and R. Bez (STMicroelectronics) for fruitful discussions and suggestions.

References

- [1] S. Lai, "Future trends of nonvolatile memory technology", Intel press release, December 2001.
- [2] S. Lai, "Current status of the phase change memory and its future," IEDM Tech. Dig., 255-258, 2003.
- [3] A. Pirovano, *et al.*, "Scaling analysis of phase-change memory technology", IEDM Tech. Dig., 699-702, 2003.
- [4] A. Pirovano, *et al.*, "Reliability study of phase-change non-volatile memories," IEEE Trans. Device Mater. Reliab., 422-427, 2004.
- [5] D. Ielmini, *et al.*, "Analysis of phase distribution in phase-change non-volatile memories," IEEE Electron Device Lett. 25, 507-509, 2004.
- [6] A. Itri, *et al.*, "Analysis of phase-transformation dynamics and estimation of amorphous-chalcogenide fraction in phase-change memories," Proc. IRPS, 209-215, 2004.
- [7] A. L. Lacaita, *et al.*, "Electrothermal and phase-change dynamics in chalcogenide-based memories," IEDM Tech. Dig., 911-914, 2004.
- [8] D. Adler, M. S. Shur, M. Silver and S. R. Ovshinsky, "Threshold switching in chalcogenide-glass thin films", J. Appl. Phys. 51, 3289-3309, 1980.
- [9] F. Pellizzer, *et al.*, "Novel μ trench phase-change memory cell for embedded and stand-alone non-volatile memory applications", Symp. VLSI Tech. Dig., 2004.

Economics of Alternative Non-Volatile Memories

Radu G. Andrei

Web-Feet Research, 769 Madison, Monterey, CA 93940, USA

1. Introduction

The mature \$17 billion nonvolatile memory (NVM) market created by the floating gate technologies is increasingly challenged by a series of advanced technologies, which have not been productized yet. Although it is quite difficult to compare mass manufactured products with technologies in the development phase, there is a compelling need to do it.

The development of a market for new mobile communication/computing/consumer terminals drives the need not only for higher processing power, but also for higher storage efficiency. This newly created market opportunity has lead to an over competitive situation in which more than thirty nonvolatile memory technologies and technology variations compete for a piece of the \$17 billion market, which is expected to exhibit a double digit compound growth rate over the next five years. Because of the prohibitive cost of a sustainable technology development policy, the natural question is to find out which of these technologies are likely to be commercial successes and which ones failures.

This paper, along with the presentation material, is an attempt to competitively assess the advanced nonvolatile memory technologies and technology variations in comparison with the floating gate technologies, which are used as reference.

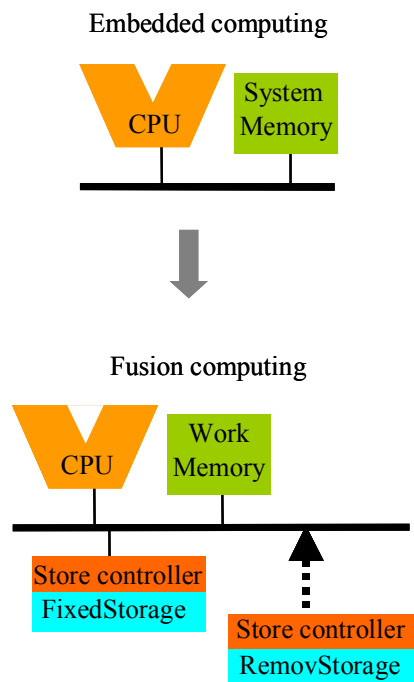


Fig. 1: Evolution of the mobile terminal system architecture

2. Background

The new mobile terminals will require, in addition to more processing power and more storage capability, a new system architecture. The main drivers for this new architecture will be the need to store a disproportionately high amount of digital data and to allow for a cost-efficient expansion of system functionality. One way to improve system efficiency is by splitting the system memory into work memory and storage. System storage can be either fixed or removable. Figure 1 illustrates this development.

This new architecture allows increasing system functionality without major changes in the hardware configuration. It can be done by increasing the capacity of fixed storage, of removable storage, or of both. When removable storage capacity is increased, just minor or no hardware and software changes are needed. In order to allow for increased functionality in successive product generations, the work memory capacity will need to increase only marginally, while the storage capacity is expected to increase exponentially. Figure 2 illustrates this trend.

Work memory and storage have different technical and commercial requirements. Figure 3 summarizes these requirements.

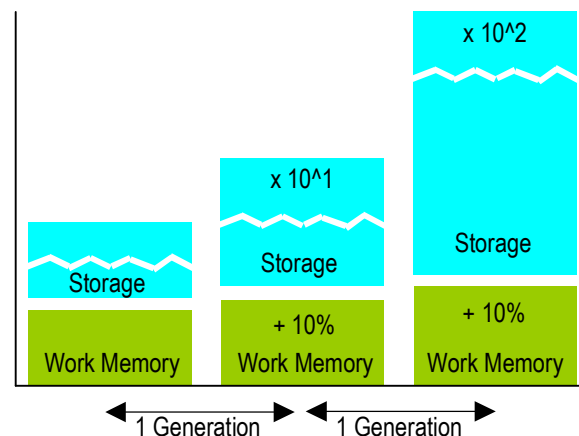


Fig. 2: Forecast capacity evolution for work memory and storage

	Work memory	Storage
Random access	Yes	No
Non volatility	Yes/No	Yes
Access speed	High	Low
Power consumption	Low	Low
Cost per bit	Medium	Very Low

Fig. 3: Technical and commercial requirements

3. Assessment Criteria.

Three different criteria need to be used to efficiently assess the commercial viability of the NVM technologies: market size, “system friendliness” and manufacturing economics. The first criterion has been traditionally used to assess economic viability. However, because of the very competitive landscape and the different degree of maturity of competing technologies, the other two criteria have to be considered as well.

“System friendliness” is defined as a factor that is indirectly proportional with the amount of resources a memory consumes in order to be embedded into a system. The ideal system memory will not need any non-system related resources (additional hardware, software or power). System resources, used exclusively to make system memory work, like error detection/correction and wear leveling, will increase system complexity, will reduce system performance and will increase system costs. “System friendliness” measures the hidden system costs introduced by a NVM technology.

Manufacturing economics gauge the cost per bit (normalized cost) for individual technologies. Normalized costs are directly related with bit density, manufacturing complexity and manufacturing learning curve. While bit density and manufacturing complexity are technology specific, the manufacturing learning curve is directly proportional with the manufactured volumes, which in turn depend on market acceptance. Therefore, one can safely conclude that technologies, which will not be able to capture a certain market share, have a reduced potential to become a significant factor in the market.

4. Mature NVM Technologies / Variations

The mature NVM technologies are:

1. Floating gate technologies (NOR, ACT, NAND, AND, AG-AND) come in SLC and MLC variations
2. Trapped charge technologies (SONOS, NROM, MirrorBit) come mainly in a MBC variation

The first technology currently dominates the market with a share of approximately 95%. The trapped charge technology has entered the market in a meaningful way last year. It can be considered to be a mature technology as well.

5. Advanced NVM Technologies / Variations

Most advanced NVM technologies, those which have not been productized yet, use a cell selection mechanism implemented in semiconductors and a non-semiconductor mechanism to store the data.

1. Ferroelectric principle memories (FRAM – 2T/2C, 1T/1C, 1T)
2. Magnetic principle memories (MRAM – AMR, GMR, TMR, CMR)
3. Phase change memories (CRAM, PMC)
4. Dipole change memories (PFRAM)
5. Polymer ion memories (PIRAM)
6. Nano-tube memories (NRAM)
7. Probe memories (Millipede)

A second group of advanced NVM technologies uses only semiconductors to implement both the data storage and cell selection functions.

1. N-FG (Nano Floating Gate)
2. McRAM (NV/SRAM monocrystalline)
3. RITD (Resonant Interband Tunneling Diode)
4. SESO (Single Electron Shut Off)
5. PLEDM (Phase-state Low Electron-number Drive Memory)

A third group of advanced NVM technologies uses no semiconductors at all to implement both functions.

1. NanoMem (storage and selection on plastic)
2. SpinRAM (all-metal storage and selection)

6. Conclusion

This assessment method has the capability to show the relative merits of the individual technologies, which can be used to calculate their degree of competitiveness. The more competitive technologies are expected to win a higher share of the market, which in turn will ensure the resources to further develop the respective technology. The less competitive technologies are expected to be obsolete quickly by the market.

SESSION F

Memories and high- κ dielectrics

Chairmen: **Franz HOFMANN**, Infineon Technologies, Germany (Munich)
 Frans LIST, Philips Semiconductors, Netherlands (Eindhoven)

Use of W Damascene as True Metal1 in a 130 nm Flash Technology.

Stephan Louwers^a

^a Atmel Fab7, Zone Industrielle, 13106 Rousset cedex, France. e-mail stephan.louwers@rfo.atmel.com

Abstract

W damascene has been integrated as a true metal1 in our 130 nm Flash technology. This shows the possibility to extend the lifetime of AlCu/W metallisation to this technology node. The dielectric in which the metal1 trenches are etched is essentially PECVD TEOS with SiN bottom and top layers, that serve as stop layers for the metal1 trench etch and the W-CMP respectively. The trench etch uses a C₅F₈ etch chemistry, and has been optimised for resist budget and capability to stop in a SiN stop layer. To improve the gap fill and the ability of the metal1 W-CMP process to stop in the top SiN layer the barrier deposition has been optimised. In the W deposition the thickness of the nucleation layer has been minimised. This thickness has a large impact on the metal1 sheet resistance. Optimisation has been done by implementing an atomic layer deposition. Finally a low-pressure W-CMP process has been developed, to minimize dishing and erosion to reasonable levels.

E-test data are within expectations, and a comparison has been made between our W damascene process and a Cu technology. Also process functionality has been demonstrated, and a successful integration of the W-damascene technique has been obtained.

1. Introduction

With shrinking dimensions the metal patterning with a subtractive AlCu process is becoming increasingly difficult, and its manufacturability is more and more problematic. Since the smallest dimensions typically occur at metal1, especially at this level difficulties arise. On the other hand, switching from AlCu to Cu involves important investments, both in terms of equipment and process development effort. Therefore, in the development of our 130 nm Flash technology, we decided to process the metal1 in a W damascene technique. It was considered that the manufacturability of a damascene technology is better than for subtractive AlCu processing. Also, implementing W as metal1 allowed us to make do with our current equipment set, and to push a switch to Cu at least one process generation further away.

Many have already presented work on a W damascene local interconnect [1,2], but in our case the metal1 is a true metal1, with many different lines/spaces. This paper presents the integration of this interconnect technology. Also a comparison with a Cu metal1 will be made.

2. Process architecture

The metal1 dielectric is a stack of SiN/PECVD TEOS/SiN. The Bottom SiN (300 Å) serves as etch stop during the metal1 trench etch, while the top SiN (450 Å) is a stop layer during the metal1 W-CMP. The thickness of the PECVD TEOS layer essentially determines the resistance of metal1 lines, as well as the interline capacitance of those lines. It can be tuned between 4000 and 7000 Å without introducing too much difficulties in the processing. Simulations showed that a thickness of 5000 Å is a fair compromise between a reasonable sheet resistance (typically 350 mΩ/□) and a not too high interline capacitance. Also, by choosing 5000 Å of PECVD TEOS resistance and capacitance values are such that the use of our existing design library, based on a subtractive AlCu technology, was facilitated.

3. Metal trench etch

For the metal1 etch a C₅F₈ etch chemistry is used. With this chemistry a good selectivity to the bottom SiN layer (>10:1) can be obtained [3], which is necessary to stop in the 300 Å thin SiN layer. In addition, by optimizing the etch recipe microloading effects have been minimized, and the dependence of the trench depth on line width is negligible. Differences in trench depth between various metal1 lines are limited to ~100 Å.

In terms of resist budget it was considered necessary to leave at least a 1000 Å of resist at the shoulders of the metal1 lines after etch, to have enough process latitude for any further process modifications. As Fig. 1 shows, this is not an issue, a minimum of 1500 Å is present on top of the BARC.

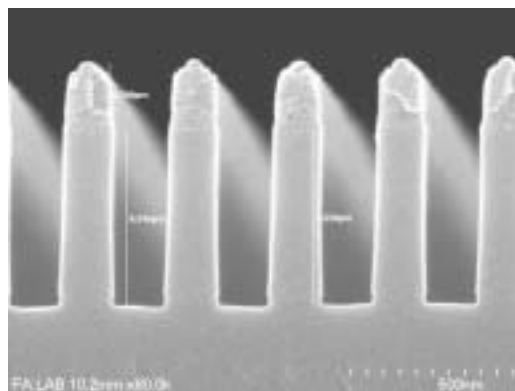


Fig. 1. Cross sections through 0.16/0.16 μm/μm lines spaces after metal1 etch.

4. Trench Fill

During the subsequent barrier deposition a Soft Sputter Etch (SSE) is necessary to clean the top of the contact plugs underneath the metall lines. A 'spin-off' of the SSE is a slight tapering of the dielectric sidewalls at the top of the metall trenches [4], improving W-fill. A too extensive SSE, however, reduces the thickness of the SiN top layer and leads to extensive tapering. Both these effects negatively impact the control of the W-CMP process, since they make it more difficult to stop in the top-SiN layer. We found that a 50Å SSE is sufficient for good contact resistance, and leaves a maximum SiN available as stop layer for W-CMP. Increasing the sputter etch to 100 or 150 Å seriously increases the tapering (see Fig. 2a and 2b).

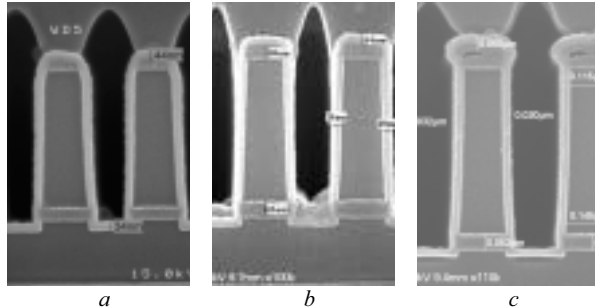


Fig. 2. Cross sections after W nucleation layer with various barrier deposition processes. a: 150 Å SSE/100 Å IMP-Ti/1x50 Å CVD-TiN, b: 50 Å SSE/100 Å IMP-Ti/1x50 Å CVD-TiN, c: 50 Å SSE/200 Å IMP-Ti/1x50 Å CVD-TiN.

For gap fill purposes the Ti layer should not be too thick. As Fig. 2c shows, a 200 Å IMP-Ti layer leads to an extensive overhang, with detrimental results for gap fill. A 100 Å IMP-Ti layer does not have this disadvantage. For these reasons a choice was made for a 50 Å SSE, followed by a 100 Å IMP-Ti deposition. The barrier deposition was finalized by a 1x50 Å CVD-TiN layer, which has not been optimised.

The subsequent W deposition consists of a nucleation layer and W-fill layer. Gap fill is improved by thinning up the nucleation layer, but making it too thin will make the process susceptible to Ti attack by WF₆ in case of insufficient TiN coverage. In addition the process control will deteriorate (reproducibility and uniformity of the nucleation layer). It was considered that 300 Å is a reasonable minimum for conventional CVD nucleation layers. With an Atomic Layer Deposition technique (ALD) even lower values can be implemented.

Since the resistivity of the nucleation layer is significantly higher than for the W-fill layer, also the sheet resistance is influenced by the nucleation layer thickness. It's impact depends on the linewidth. In Fig. 3 a comparison is given between a CVD nucleation layer (300 or 600 Å) and an ALD nucleation layer (40 Å). For 0.16 µm metall lines an decrease in nucleation layer from 600 Å to 300 Å will lead to an associated sheet resistance decrease from 680 mΩ/□ to 370 mΩ/□. A 40 Å ALD nucleation layer will give even lower sheet resistance values of 270 mΩ/□.

Also the dependence of the sheet resistance on CD variations is dramatically better for thin nucleation layers. For a 600 Å nucleation layer not only will the

sheet resistance depend on the designed metal line width, but also natural CD variations from die to die will lead to a large spread in sheet resistance. The latter variations will be largely eliminated with a 300 Å nucleation layer, while a 40 Å ALD deposition will give a nearly flat sheet resistance for metal line widths up to 0.1 µm. This means that not only for the stability of the sheet resistance there are many advantages to ALD deposition, but also that this technique is scalable at least one more process generation. This is not the case for CVD type depositions.

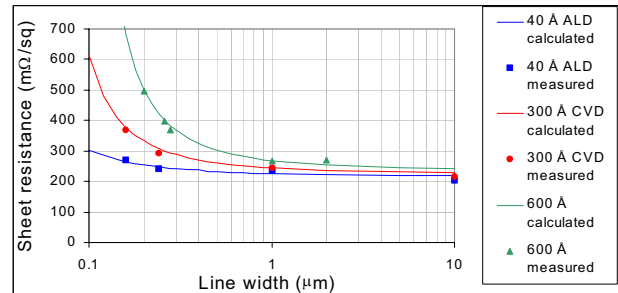


Fig. 3. Impact of the nucleation layer thickness on the metall sheet resistance.

Finally, in Fig. 4 a comparison is given between the gapfill performance with a 300 Å CVD nucleation layer and a 40 Å ALD nucleation layer. It can be seen that the latter process gives superior trench fill. For these reasons we are currently qualifying the ALD nucleation layer in our 130 nm process.

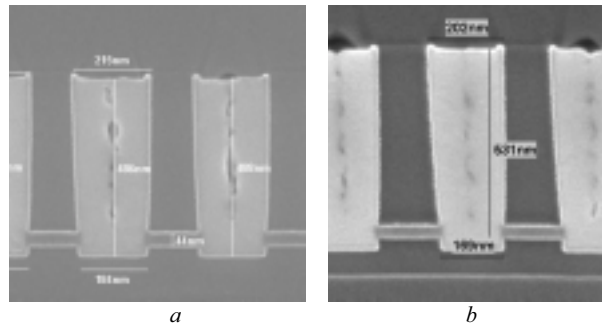


Fig. 4. Trench fill comparison between W processes incorporating a 300 Å CVD nucleation layer (a) and a 40 Å ALD nucleation layer.

5. W CMP

One of the most critical steps is W-CMP, since dishing and erosion will immediately impact the metall sheet resistance and lead to inadmissible variations over the die, the wafer and the lot. We have implemented design rule restrictions to minimize these effects (maximum linewidth, minimum spacing for large metall lines). In

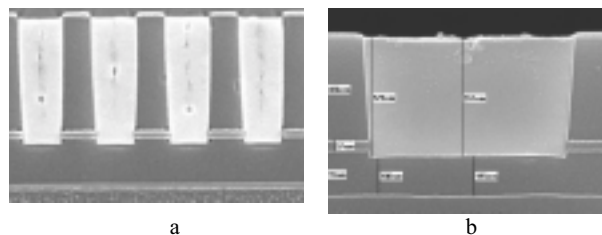


Fig. 5. Cross sections through test structures after W-CMP with various lines/spaces. a: 0.16/0.16 µm/µm. b: 2/0.26 µm/µm.

a low-pressure W-CMP process has been developed. The low pressure will effectively minimize the dishing and erosion, although care must be taken that the uniformity does not get out of control [5]. Fig. 5 shows some representative cross sections. For metal lines up to 2 μm wide the W-loss is minimized, and the W-CMP typically stops in the SiN layer or just below. For wider lines a more considerable W-loss is seen; up to 25% of the total W thickness, depending on surface density. In Fig. 6 the W-loss is indicated for various other test structures. Although the numbers in this figure might seem large, it must be realized that these tests have been done on very large test structures with a surface of 2.5 mm by 1.6 mm. In real products there might be structures of this size for 0.16/0.16 $\mu\text{m}/\mu\text{m}$ lines spaces (memory), but not for 5/1 $\mu\text{m}/\mu\text{m}$ lines spaces for instance, and in reality the W-loss can be expected to be much less.

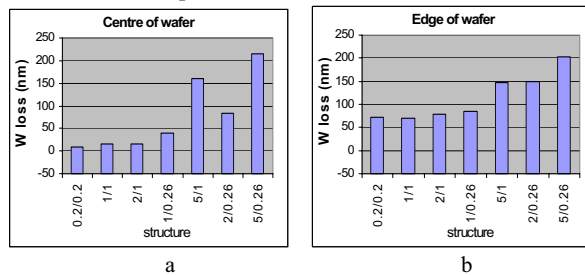


Fig. 6. W loss due to W-CMP for various lines/spaces at the center and the edge of the wafer. The 5/0.26 structures are a violation of the design rules, but have been added for completeness

To demonstrate the integration of W damascene metall on contacts and under via1, in Table 1 various electrical data are shown. As can be seen, string yield is 100%, and the corresponding contact resistance is reasonable. Also no large spread in contact resistance exists, indicating no particular problems. In addition to these data, product functionality on a number of Flash memories up to a 256 Mb parallel Flash memory has been demonstrated, using the presented technology as metall1.

	Contact chains		
	string yield	Resistance	st. dev.
silicided p+	100 %	15.2	0.87
silicided n+	100 %	14.7	0.70
silicided poly	100 %	11.3	0.67
unsilicided poly	100 %	503	18.56
via1	100 %	5.45	0.19

Table 1. Contact resistance and string yield in $\Omega/\text{contact}$ for various types of contact.

6. Impact of the metall sheet resistance

When choosing W as metall1, obviously considerations need to be made on the sheet resistance, which is high compared to an AlCu process, let alone a Cu process.

When looking at the logic part of the product, we must distinguish several situations. The logic part consist essentially of a number of rows of standard cells, that within one row are laid out side to side. The metall interconnects inside standard cells are short, and their resistance is limited to approximately 10 Ω . Also between adjacent cells the metall interconnects will be short, and equivalent to a maximum of ca. 20 Ω . These

values are negligible compared to the resistances of poly lines or transistors. However, long interconnects between different rows of standard cells can constitute much higher resistances of several hundreds of Ω , which can introduce additional RC delay. To avoid these problems, these interconnects can be transferred to the less resistive higher metal layers.

For the memory part the higher sheet resistance of the W lines can introduce a certain voltage drop. Taking into account the sheet resistance of the W metall1, and the maximum distance from a memory cell to the nearest by via1, a maximum voltage drop of ca. 30 mV can be calculated. This value is negligible compared to standard process variations.

Also, the sheet resistance of the W lines will impact the speed with which the cells can be read. This is difficult to evaluate up front, but to compare a W metall1 with a Cu metall1 a number of wafers has been manufactured with a Cu backend. In Table 2 some typical backend e-test parameters are compared. For completeness also the metall2 and metall3 parameters are given. Although the metall1 sheet resistance is three times lower for the Cu process than for the W process, speed measurements show that only a 14 % lower read time has been realised. Although this is a significant difference, and implicates that for high speed applications a Cu backend will be advantageous, it is not large enough to disqualify the W process for many standard applications.

Process	Metall sheet resistance ($\text{m}\Omega/\square$)			Via1 resistance (Ω/via)
	width 0.16 μm	width 1.0 μm	width 10 μm	
W/Al	370	234	205	5.45
Cu	97	91	83	1.1

Process	Metall2 sheet resistance ($\text{m}\Omega/\square$)	Via2 resistance (Ω/via)	Metall3 sheet resistance ($\text{m}\Omega/\square$)
W/Al	86	4.34	50
Cu	66	0.8	25

Table 2. Comparison of backend e-test parameters for a W/Al process and a Cu process.

7. Conclusions

We have developed a W damascene technique as a true metall1, allowing many different combinations of lines/spaces. This has been successfully implemented in our 130 nm Flash technology. After optimisation of trench etch, barrier deposition, W deposition and W-CMP (all described herein), good process control is obtained, while electrical parameters are within expectations. Furthermore, product functionality has been demonstrated.

A comparison between our W/AlCu process and a Cu process learns that the W damascene process is a viable alternative for many applications where high speed is not absolutely necessary.

References

- [1] 70 nm NAND Flash Technology with 0.252 μm^2 cell size for 4 Gb Flash Memory, Yong-Sik Yim et al., IEDM 2003.

- [2] A highly manufacturable 110 nm DRAM Technology with $8F^2$ vertical transistor cell for 1 Gb and beyond, H. Akatsu et al., 2002 Symposium on VLSI Technology.
- [3] Incident angular dependence of SiO_2 and SiN etching with mass-analyzed CF_x^+ ion beam irradi
- [4] DamaClean reactive preclean for Cu Damascene applications, R. Rozbicki et al., The 3rd technical and scientific meetings of CREMSI, 2000.
- [5] Low-Pressure CMP for 300-mm Ultra Low-k ($k=1.6-1.8$)/Cu Integration, S. Kondo et al., IEDM 2003.

Investigation of SiO₂/HfO₂ gate stacks for application to non-volatile memory devices

J. Buckley^a, B. De Salvo^a, G. Ghibaudo^b, M. Gely^a,
J.F. Damlencourt^a, F. Martin^a, G. Nicotra^c and S. Deleonibus^a

^a CEA-LETI, 17 rue des Martyrs 38054 Grenoble Cedex 9, France julien.buckley@cea.fr

^b IMEP-CNRS/INPG, Grenoble, France

^c Microsystems and Microelectronics Institute (CNR -IMM), Stradale Primosole, 50, 95121 Catania, Italy

Abstract

Engineering of the tunnel barrier of non-volatile memories is addressed in this paper. The strong potential of multilayer stacks for reducing the programming times of these devices, without altering their retention characteristics, is studied. To this aim, experimental results showing the improved field sensitivity of the gate current of MOS devices with SiO₂/HfO₂ gate stacks compared to SiO₂ of identical electrical thickness (EOT) are presented. Simulations of the corresponding tunnelling currents are also reported. Our electrical measurements suggest that it is of great interest to investigate the use of SiO₂/HfO₂ stacks as tunnel barriers for NVMs.

1. Introduction

In current flash non-volatile memory (NVM) devices, the tunnel oxide thickness is of about 7-8 nm and would be difficult to downscale in the future if 10 year data retention is still wanted [1]. This requirement limits further improvement of device performance in terms of programming speed and operating voltages. It is therefore necessary to find innovative solutions in order to bypass this restriction. A promising one is the use of multilayer dielectric stacks to practice tunnel barrier engineering. The main potential of these stacks would be to allow shorter writing/erasing times and/or smaller operating voltages than the ones obtained using conventional SiO₂ without altering the 10 year data retention constraint.

Multi-layer tunnel barriers for flash memory applications were first introduced by K.K. Likharev [2] in 1998 with the concept of “Crested Barriers” (CBs). In ref. [3] he demonstrated through tunnelling current simulations that the use of a triple-layer stack (using a Si₃N₄(4nm)/AlN(5nm)/Si₃N₄(4nm) combination) would enable far better performance than today’s single layer SiO₂ barriers, in terms of programming speed and data retention. These improvements were due to the high slope of the simulated current-voltage characteristic of such a stack compared to one layer of silicon oxide.

Other authors have shown the possibility of achieving high performance by using barriers consisting of low-k/high-k/low-k combinations. In these cases, the low-k oxide generally presented was SiO₂, which has the important advantage of guaranteeing a good quality interface with the Si channel. Experimental data for SiO₂/ZrO₂ stacks was shown in Ref. [4]. It was also shown experimentally [5] that the use of a SiO₂/HfO₂

tunnel barrier in a NVM memory cell favoured low programming voltages.

In Ref.[6], we presented a theoretical analysis of both “Conduction Band - Crested Barriers” and “Dielectric Constant – Crested Barriers” in view of their application to floating-gate non-volatile memory devices. The main conclusion was that low-k/high-k/low-k dielectric stack combinations are very good candidates as a tunnel dielectric for NVMs.

In this paper we present experimental results that clearly show the improved field sensitivity of the gate current of MOS devices with SiO₂/HfO₂ gate stacks compared to SiO₂ of identical EOT. These measurements suggest the possibility of using similar structures for replacing the tunnelling dielectric of NVMs. Simulations of the corresponding tunnelling currents are also reported.

2. Device description and experimental results

Very large (250 μm x 160 μm gate surfaces) single and double layer high-k n-MOS transistors have been fabricated, using HfO₂ and thermal SiO₂ (See Fig. 1). Interfacial 1.65 nm, 2.2 nm and 2.6 nm thick SiO₂ layers were thermally grown on p-doped (10¹⁷/cm³) silicon substrates. Three nanometre thick hafnium oxide was deposited by Atomic Layer Deposition (ALD) in a PulsarTM 2000 tool from ASM Microchemistry Ltd, using HfCl₄ as a precursor. All transistor gates were fabricated with TiN coated by n+ degenerated poly-Si. High resolution TEM pictures of all fabricated samples are available in Fig. 2. The extracted EOTs of samples 1 and 2 are respectively 2.3 nm and 2.7 nm, allowing them to be comparable with references A and B.

Experimental transport curves are shown in Fig. 3-a for both samples and the reference. We can observe that samples 1 and 2 have higher levels of current than the references for electric fields F_{ox}>7MV/cm, therefore allowing a higher writing speed if integrated in a device. For smaller fields (2-3MV/cm), devices 1 and 2 have smaller leakage currents than the reference, allowing better retention. The reason for this is illustrated in Fig. 4. For low fields the direct-tunnelling current through the double-layer stacks is smaller than the one through the Reference SiO₂ layer, due to its larger physical thickness, as illustrated by the conduction band diagrams in Fig. 4-a and -c. For higher fields, the voltage drop across the barrier is mainly applied to the interfacial SiO₂, because of its low dielectric constant compared to

a high-k. This leads to direct tunnelling through the interfacial layer (Fig. 4-b), which is shorter to cross than in the case of the single layer sample (Fig. 4-d).

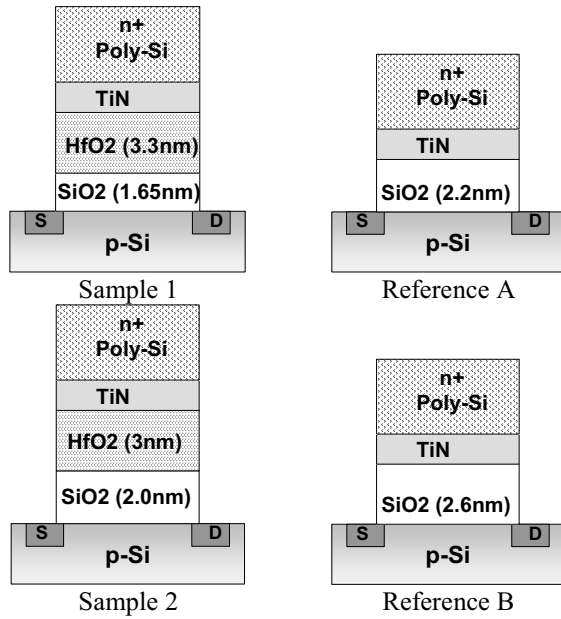


Fig. 1: Samples developed in this work, i.e. single- and double-layer transistors with thermal SiO₂ and ALD HfO₂.

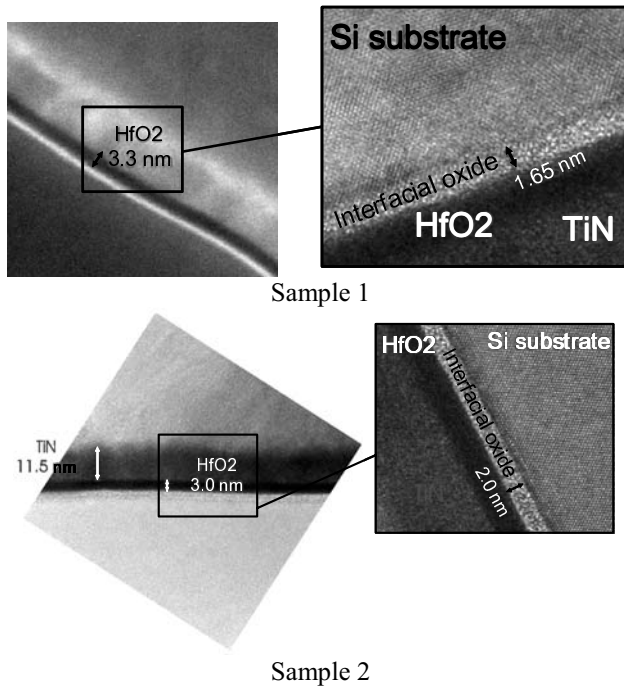


Fig. 2: High Resolution TEM analyses of samples 1 and 2 (corresponding to Fig. 5).

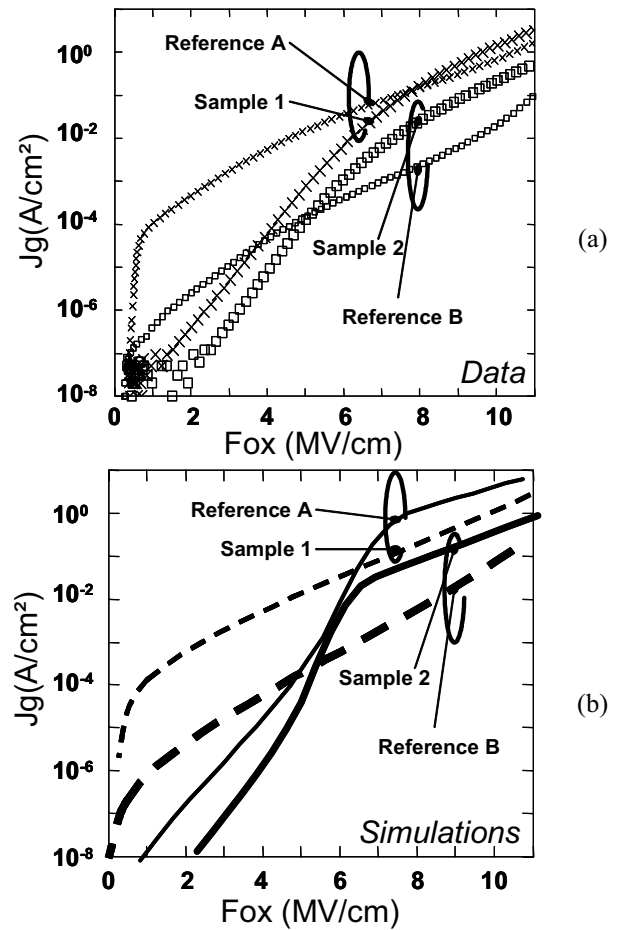


Fig. 3 - (a) Experimental transport curves (current versus equivalent oxide electric field) for the stacks drawn in Fig. 1 (b) Corresponding simulated transport curves using the current model described in paragraph II-B (parameters: $\Phi_1=1.5\text{eV}$, $m_{\text{eff}}=0.17m_0$, $k=20$ for HfO₂, and $\Phi_2=3.15\text{eV}$, $m_{\text{eff}}=0.5m_0$, $k=3.9$ for SiO₂).

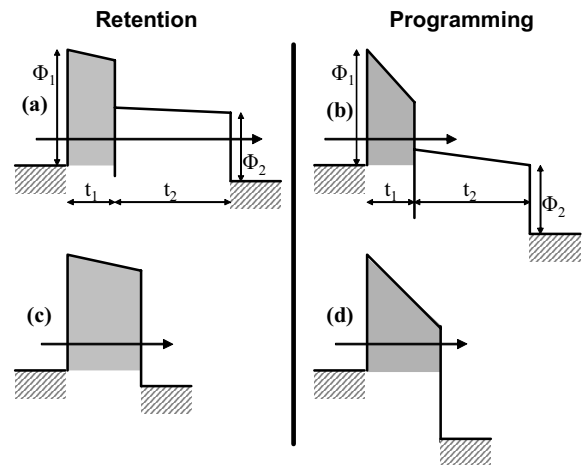


Fig. 4 - (a) Conduction band diagram of electrons tunneling through a SiO₂/high-k stack in the case of a small electric field. (b) Conduction band diagram of electrons direct tunneling through the SiO₂ layer of a SiO₂/high-k stack and entering directly into the conduction band of the high-k in the case of large electric fields. (c) Conduction band diagram of electrons tunneling through SiO₂ in the case of a small electric field. (d) Conduction band diagram of electrons direct tunneling through SiO₂ in the case of large electric fields.

3. Gate tunnelling current simulations

Tunnelling current calculations are developed here for triple layer stacks for which simulation examples are shown. The model is then applied to the cases of the double-layer devices presented in part 2.

In this paper, the tunnelling probability of electrons is determined by using a Wentzel-Kramers-Brillouin (WKB) formalism for the transparency of the involved dielectric stacks, which does not take into account electron reflections. In our case we used it for 2 and 3 layer stacks. A method that includes electron reflections is the Transfer Matrix formalism [7]. A problem arises though, which is the introduction of oscillations in the current-voltage simulations. Such oscillations aren't observed experimentally because of electron scattering. In Fig. 5-a, b we show simulations for two different stacks: SiO₂ (5nm) and SiO₂ (1.5nm)/HfO₂(10nm)/SiO₂(1.5nm). They indicate that the WKB current simulations give results that are very close to the ones that are obtained with Transfer Matrix, but without any oscillations. Details on these simulations, when a WKB formalism is used, are explained hereafter.

In the case of a triple-layer symmetrical dielectric stack (see conduction band diagram in the inset of Fig. 5-a) for which the middle layer has parameters m_2 , ϵ_2 , Φ_2 and t_2 , and both side layers parameters m_1 , ϵ_1 , Φ_1 and t_1 (with: m_1 , m_2 : electron effective masses – ϵ_1 , ϵ_2 : permittivities – Φ_1 , Φ_2 : electron barrier heights compared to Si – t_1 , t_2 : layer thicknesses), the corresponding WKB transparency can be expressed analytically as follows:

$$T_{WKB}(m_{1,2}, \Phi_{1,2}, t_{1,2}, \epsilon_{1,2}, E, V_{TB}) = \exp\left(\frac{-2\sqrt{2}}{\hbar} [I_1(1, E, V_{TB}) + I_2(1, 2, E, V_{TB}) + I_3(1, 2, E, V_{TB})]\right) \quad (1)$$

where: $\hbar$: reduced Planck constant – E : electron energy – V_{TB} : voltage drop across the dielectric stack tunnel barrier and where:

$$I_1(1, E, V_{TB}) = I_1(m_1, \Phi_1, t_1, \epsilon_1, E, V_{TB}) = \int_0^{t_1} \sqrt{m_1 \left(\Phi_1 - qV_{TB} C_{TB} \frac{x}{\epsilon_1} - E \right) H \left(\Phi_1 - qV_{TB} C_{TB} \frac{x}{\epsilon_1} - E \right)} dx \quad (2)$$

$$I_2(1, 2, E, V_{TB}) = I_2(m_2, \Phi_2, t_2, \epsilon_2, E, V_{TB}) = \int_{t_1}^{t_1+t_2} \sqrt{m_2 \left(\Phi_2 - qV_{TB} C_{TB} \left(\frac{t_1}{\epsilon_1} + \frac{x-t_1}{\epsilon_2} \right) - E \right) H \left(\Phi_2 - qV_{TB} C_{TB} \left(\frac{t_1}{\epsilon_1} + \frac{x-t_1}{\epsilon_2} \right) - E \right)} dx \quad (3)$$

$$I_3(m_1, \Phi_1, t_1, \epsilon_1, E, V_{TB}) = \int_{t_1+t_2}^{2t_1+t_2} \sqrt{m_1 \left(\Phi_1 - qV_{TB} C_{TB} \left(\frac{t_2}{\epsilon_2} + \frac{x-t_2}{\epsilon_1} \right) - E \right) H \left(\Phi_1 - qV_{TB} C_{TB} \left(\frac{t_2}{\epsilon_2} + \frac{x-t_2}{\epsilon_1} \right) - E \right)} dx \quad (4)$$

where: H : Heaviside function – C_{TB} : tunnel barrier capacitor per surface unit

In the case of electron injection from n-doped or p-doped silicon, corresponding respectively to an

accumulation or inversion layer, and if tunneling through SiO₂ is considered, the current can be expressed as being [8]:

$$J(V_{ox}) = f_{esc} Q T_{WKB}(m_{SiO_2}, \Phi_{SiO_2}, t_{SiO_2}, \epsilon_{SiO_2}, E_{fund}(Q), V_{ox}) \quad (5)$$

Where:

f_{esc} : escape frequency – Q inversion (electron injection from p-type Si case) or accumulation (electron injection from n-type Si case) charge – V_{ox} : voltage drop across the oxide – m_{SiO_2} : electron effective mass in SiO₂ – ϵ_{SiO_2} : SiO₂ permittivity – Φ_{SiO_2} : SiO₂ electron barrier height compared to Si – t_{SiO_2} : oxide thickness – V_{ox} : oxide voltage drop – E_{fund} : electron fundamental energy in the silicon accumulation/inversion layer

The same formula can also be used for triple-layer stacks:

$$J_{STACK}(V_{STACK}) = f_{esc} Q T_{WKB}(m_{1,2}, \Phi_{1,2}, t_{1,2}, \epsilon_{1,2}, E_{fund}(Q), V_{STACK}) \quad (6)$$

Where:

V_{STACK} : voltage drop across the stack

According to Ref. [9], the escape frequency for electrons can be expressed as being:

$$f_{esc} = \frac{2}{3} \left(\frac{2E_{fund}}{\hbar} \right) \quad (7)$$

Where: $\hbar$: Planck constant

As described in Ref. [9-10], if a variational approach is used to model the energy well at a Si/dielectric interface, E_{fund} has the following expression:

$$E_{fund} = \frac{3\hbar}{2} \left[\frac{q^2 m_{Si}}{4\hbar^2 \epsilon_{Si}} Q^2 \right] \quad (8)$$

Where: $\hbar$: reduced Planck constant – m_{Si} : electron mass in Si – ϵ_{Si} : Si permittivity

Current simulations of our samples based on the here presented model are shown in Fig. 3-b. They were carried out by setting:

$$I_3 = 0 \quad (9)$$

in order to adapt the calculations for dielectric stacks that contain only two layers. We can observe that they are very close to the measurements in the case of the SiO₂ layers.

For the SiO₂/HfO₂ stacks (samples 1 and 2), the transition of the current characteristics between direct tunnelling through both layers of the stack (Fox<5MV/cm) and direct tunnelling through the SiO₂ layer only (Fox>7MV/cm) is a little sharper in the simulations than in the experiments. This can be explained by the fact that the SiO₂/HfO₂ interface is in practice not as good quality as considered in the model. As a result, the electronic potential transition at this interface isn't as sharp for our devices as in the simulation.

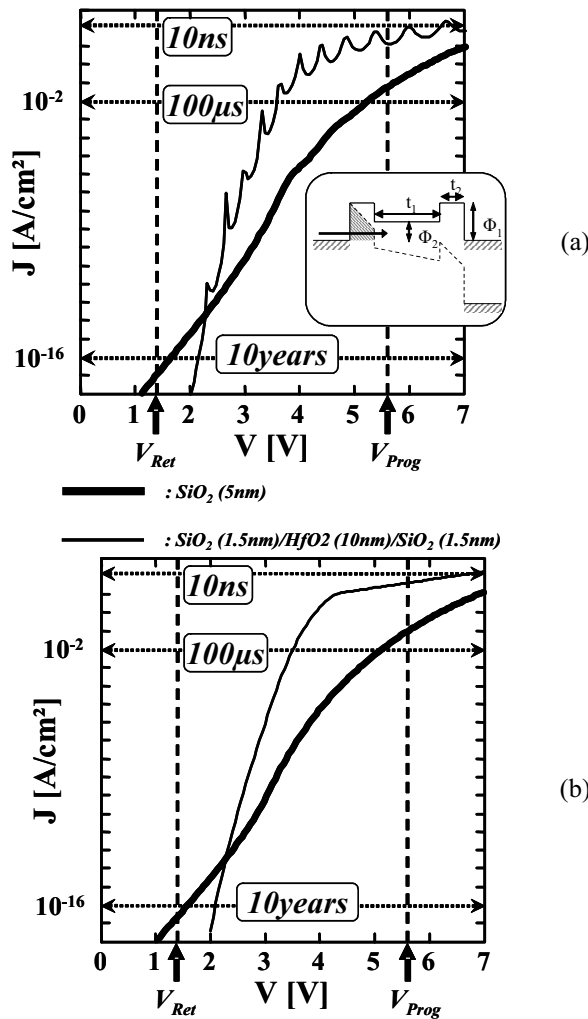


Fig. 5. Simulation of a current-voltage (J-V) curve of a triple-layer barrier compared to SiO_2 , (of same EOT) made by using the escape frequency model described previously. In graph (a), simulations have been done with a Transfer Matrix formalism for the transparency (inset: band diagram of the $\text{SiO}_2/\text{HfO}_2/\text{SiO}_2$ stack) and in graph (b) the Wentzel Kramers Brillouin approximation is used.

4. Conclusion

We have shown in this paper experimental results that are promising for improving the field-sensitivity of the tunnel dielectric of NVMs, by using stacks made of SiO_2 and HfO_2 rather than only SiO_2 .

Indeed, these results are encouraging and will have to be carried out in a near future for larger high-k

thicknesses and symmetrical stacks for which the conduction for gate injection as well as substrate injection could be studied, before being integrated in NVM devices.

Acknowledgements

This work has been partly carried out with the support of Freescale. The author would like to thank Ramachandran Muralidhar, Bob Steimle, Rajesh Rao and Bruce White from Freescale for fruitful discussions.

References

- [1] International Technology Roadmap for Semi-conductors, 2003 Edition: <http://public.itrs.net/Files/2003ITRS/Home2003.htm>
- [2] K.K. Likharev, "Layered Tunnel Barriers for Nonvolatile Memory Devices", Applied Physics Letters, Vol. 73, No. 15, p.2137 (1998)
- [3] K.K. Likharev, "Riding the Crest of a New Wave in Memory", IEEE Circuits and Devices Magazine, Vol. 16, No. 4, p.16 (2000)
- [4] P.Blomme, B. Govoreanu, M. Rosmeulen, J. Van Houdt, K. De Meyer, "Multilayer tunneling barriers for nonvolatile memory applications", Proc. of Dev. Research Confer. , p.153 (2002)
- [5] 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 Electron Device Letters, Vol. 24, No. 2, p. 99 (2003)
- [6] J.Buckley, B.DeSalvo, G.Ghibaudo*, M. Gely, J.F.Damlencourt, A.M. Papon, X. Garros and S.Deleonibus, "Engineering of "Conduction Band - Crested Barriers" or "Dielectric Constant - Crested Barriers" in view of their application to floating-gate non-volatile memory devices", SNW'04
- [7] Y.Ando, T. Itoh, "Calculation of transmission tunneling current across arbitrary potential barriers", Journal of Applied Physics, Vol. 61, No. 4, p. 1497 (1987).
- [8] B. De Salvo, G. Ghibaudo, G. Pananakakis, P. Masson, T. Barron, N. Buffet, A. Fernandes, B. Guillaumot, "Experimental and Theoretical Investigation of Nano-Crystal and Nitride-Trap Memory Devices", IEEE Transactions on Electron Devices, Vol.48, No.8, p.1789 (2001)
- [9] R. Clerc, P. O'Sullivan, K.G. McCarthy, G. Ghibaudo, G. Pananakakis, A. Mathewson, "A physical compact model for direct tunnelling from NMOS inversion layers", Solid State Electronics, Vol. 45, Iss. 10, p. 1705 (2001)
- [10] T. Ando, A.B. Fowler, F. Stern, "Electronic properties of two-dimensional systems", Review of Modern Physics, Vol. 54, p. 437 (1982)

Scaling Down the Interpoly Dielectric for Next Generation Flash Memory: Challenges and Opportunities

B. Govoreanu^a, D. Brunco^b, J. Van Houdt^a

^{a)} IMEC Leuven, Kapeldreef 75, B-3001 Leuven, Belgium

^{b)} Intel Corp. c/o IMEC, Kapeldreef 75, B-3001 Leuven, Belgium

Phone: +32-16-281 337, Fax: +32-16-281 844, E-mail: govorean@imec.be

Abstract

In this work, we review the main issues in scaling down the interpoly dielectrics for future Flash memory technology generations. To reach the sub-10 nm EOT range, new solutions are proposed, exploiting the opportunities offered by the high- κ dielectrics. Their effectiveness increases when midgap and p-type metal gates are considered, instead of the conventional polysilicon control gate.

1. Introduction

Scaling of the conventional tunnel oxide for Flash memory technology is close to its limits. Research is needed to circumvent this lack of scalability, and thus to allow the voltages to scale down and keep pace with mainstream CMOS technology. On a longer term, *revolutionary research* [1] may eventually lead to entirely new NVM devices. In a nearer future, *evolutionary research* will presumably maintain the “template” of the existing memory devices, e.g. floating gate device structures in the present case, and seek to improve the elements responsible for charge transport on both a short time scale (for programming and erasing), as well as on a long time scale (for retention). A possible route is to research novel tunnel barriers and interpoly dielectrics (IPD's) that are able to meet the performance requirements of Flash memory. Although the successful implementation of a memory concept will ultimately be dictated by cost, a thorough investigation is needed to explore their capabilities and limits. In this work, we focus on the scalability of the IPD's in the sub-10 nm range.

2. Interpoly dielectric requirements

Good IPD's must fulfill several conditions. First, to allow fast enough operation of the device they must ensure good electrical coupling to one external electrode, commonly referred to as the control gate, which implies thin electrical thickness. This allows for a significant part of the voltage drop over the entire stack to be transferred to the tunnel dielectric. However, too strong a coupling will actually decrease the V_T shift corresponding to a given amount of transferred charge, hence the coupling ratio is to be defined by this compromise. Secondly, to allow for good information storage, they must provide excellent isolation at any operating condition, involving very low leakage currents, under retention and read-out (disturb) conditions. On top of these requirements, a growing concern is the so-called erase saturation problem, which slows down or even stops the erase operation, due to the

increasing high-field conduction through the IPD, which tends to compensate for the electrons removed from the floating gate towards the channel. All these conflicting requirements lead to practical difficulties in scaling down the IPD.

The state-of-the-art interpoly dielectrics usually consist of an oxide-nitride-oxide (ONO) stack and received quite a lot of research attention [2,3]. Notably, the nitride layer not only reduces the electrical thickness, but it also acts as charge capture layer, thus decoupling the fields in the top and bottom oxide. However, it is widely accepted that the conventional ONO stack has little room left for scaling, due to the individual layers reaching the physical limits beyond which they will not be able to perform as expected. Particularly, the bottom oxide can no longer be scaled down due to intrinsic or generated defects, which may cause V_T instabilities (on a short time scale) or increase the charge loss (on a long time scale). The top oxide, the primary role of which is to block hole injection from the control gate has also reached its limits, at around 4 nm. Furthermore, the nitride layer cannot be made very thin, since its ability to retain charge leaking from the floating gate decreases and also the time-dependent dielectric breakdown limit may be reached.

The lack of scalability of the tunnel oxide is one of the most striking features of the Flash memory forecasts, as discussed in the ITRS [4]. To overcome this issue, we proposed a novel tunnel barrier concept, namely VARIOT [5, 6], which has a strong potential for scaling towards 4 nm EOT and below, while maintaining the retention leakage at acceptably low levels. But even in the absence of such evolutionary concepts, that provide some work-around for maintaining the backbone of the mature floating-gate industry-standard cell [7], the IPD has to be scaled down anyway. The main reasons for this are given below:

- The first one is the need for lower operating voltages, to reduce the gap between NVM and CMOS, which then can be combined in embedded memory applications without consuming a lot of peripheral area. This is shown in **Fig. 1**, from which one can estimate the voltage lowering for a conventional tunnel oxide to be about 0.5 V/nm. The programming times were calculated by numerical time integration, with a tunneling current model discussed in [8].
- The second one, and even more important in the stand-alone memory case, comes from the architectural change of the stacked gate cell. With

the technology downscaling, the control gate can no longer be wrapped around the floating gate. This causes loss of coupling capacitance from the sidewalls (**Fig. 2**). To maintain the same coupling, additional capacitance must be provided to compensate for the coupling area loss. This may be achieved by reducing the electrical thickness of the IPD. Given the typical horizontal and vertical dimensions of the floating gate for present and future technologies, the decrease of the coupling capacitance can be of up to 60 %, which requires a dramatic decrease of the IPD thickness in order to compensate for.

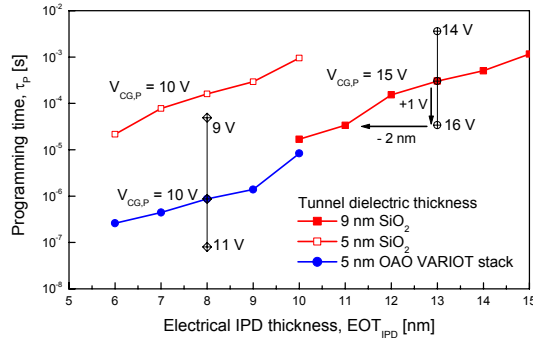


Fig. 1: Programming time as a function of the electrical thickness of the IPD. The symmetric $\text{SiO}_2/\text{Al}_2\text{O}_3/\text{SiO}_2$ (OAO) VARIOT stack has 4.6 nm of Al_2O_3 . A full numerical model was used for the tunneling current calculations.

At first sight, this leads to the use of very high- κ materials to combine a thick physical thickness with a low electrical thickness. In this case however, other constraints may be activated. One of them is the electrostatic interference between neighbouring cells, which has been discussed both for NAND [9] and NOR [10] arrays. In fact, this is just parasitic coupling, which could require low- κ materials as inter-cell isolation, in order for it to be reduced. Another constraint is given by the pitch size, i.e. the physical thickness of the IPD should remain small as compared to the pitch size.

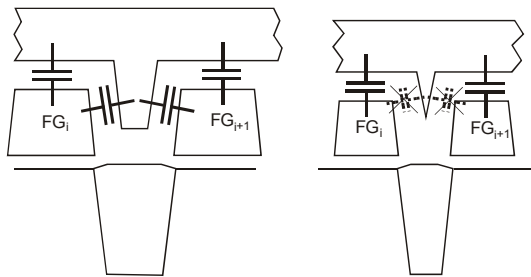


Fig. 2: Schematic drawing illustrating the effect of the downsizing on the coupling capacitance loss. The small space in between two adjacent cells can no longer accommodate both IPD and control gate word line.

With these arguments, we have provided a strong motivation to direct research efforts towards alternative IPD's. For practical reasons, we divided these into two categories, the first one being sub-10 nm EOT IPD, which can not be achieved with ONO and which should serve to reduce operation voltage, hence addressing the first requirement mentioned before. A second category

of IPD's targets EOT's of 5 nm and below and should serve the purpose of fabricating cells in a high density stand-alone memory technology.

3. Sub-10 nm interpoly dielectrics

3.1. Triple-layer OHO stacks

The most straightforward, yet naïve approach, to achieve sub-10 nm EOT IPD's is to replace the nitride layer in an ONO stack by a high- κ layer. This is more effective for thicker nitride layers, but given the relative gain in EOT, summarized in Table I, the total EOT improvement is rather small, of up to 2 nm at best. If the replacement of the nitride is combined with a thinning of the middle layer, then the EOT improvement is somewhat better. We fabricated such stacks, with either Al_2O_3 or HfO_2 , which showed no directly measurable leakage on large area capacitors for equivalent field oxides of up to 3 – 4 MV/cm.

Material	Al_2O_3	HfO_2	HfAlO_x	Hf-Silicates
EOT gain [el.nm/ph.nm]	0.12	0.3	~0.25	< 0.2

Table I: Relative EOT gain with respect to the reference Si_3N_4 for different high- κ materials.

3.2. The Reverse-VARIOT concept

A more elegant approach is to redesign the IPD stack, exploiting the opportunities offered by high- κ materials. To this end, one possibility is to reconsider the role of the top oxide in the ONO stack, i.e. it blocks hole injection from the top poly, which would otherwise take place through nitride, with a hole barrier of only approximately 2.0 eV, and a better hole conductivity as compared to that of electrons. As high- κ materials have generally large bandgaps and band offsets with respect to those of the Si conduction and valence bands, a suitable high- κ material with a large bandgap, like e.g. Al_2O_3 may eventually allow skipping the top oxide, and thus reducing EOT by a few nm at once. Combined with an additional EOT reduction from the high- κ material itself, a decrease of the EOT down to 6-7 nm seems feasible. **Fig. 3** shows leakage currents through such dual-layer stacks with Al_2O_3 . Splits with EOT's of intermediate values show currents in between both sets of curves, as expected. Since the bottom SiO_2 layer is relatively thick, the VARIOT effect is reduced, in the positive polarity range. This is important for retention, at low positive biases, of up to 1.5 – 2V, as well as for gate disturb conditions.

The negative polarity shows even smaller currents, at similar $|V|$ conditions. This is due to two reasons. First, the **Reverse-VARIOT effect** [11] makes the potential barrier for control gate electrons more persistent to the applied bias, as opposite to the VARIOT case, where the barrier is more sensitive (**Fig. 4**). This is attributed to the redistribution of the electric field in the individual layers of the stack, as follows from basic electrostatics. Secondly, the benefit comes from the gate material

itself, which has a workfunction of around 5 eV (p-type), and thus the IPD stack appears to have a higher potential barrier as well.

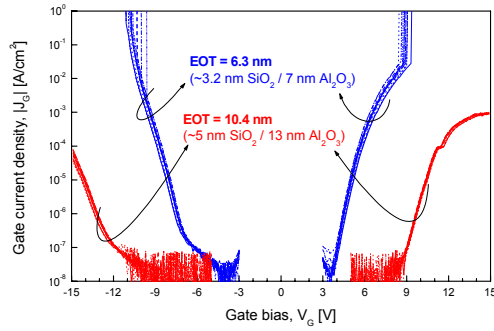


Fig. 3: Leakage currents through double-layer $\text{SiO}_2/\text{Al}_2\text{O}_3$ stacks with TiN gate electrode. Data were measured on large area capacitors.

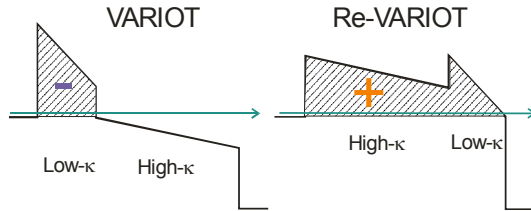


Fig. 4: Illustration of the Reverse-VARIOT effect. For comparison, the VARIOT effect is also shown. At similar voltage drops over the two stacks, the tunneling barrier is much larger in the Re-VARIOT case (“+”), as compared to the VARIOT case (“-”).

The erase saturation problem appears in the negative polarity range, when the IPD current during erase operation is comparable to the actual erase current. This condition is application dependent. However, a range of current values of $10^{-4} - 10^{-3} \text{ A/cm}^2$ seems to cover well the most typical situations. For the 10.4 nm EOT stack this only happens at -15 V , which will never be reached over the IPD stack during the regular device operation. For the 6.3 nm stack, this value reduces to only around -8 V , which is still rather good. A limitation of this approach is given by the fact that the SiO_2 layer has to be sufficiently thick (more than 3 nm), for sufficiently suppressing the inherent VARIOT effect. Furthermore, if there is sufficient EOT budget, a symmetric high- κ /low- κ /high- κ Reverse-VARIOT stack could be considered, as well. One important challenge here is to control the interfacial layer that may eventually form underneath the high- κ material.

4. Aggressively scaled interpoly dielectrics

If an electrical thickness below 5 nm is required, e.g. 4.5 nm, then the situation is much more critical. This is shown in **Fig. 5**, for several $\text{SiO}_2/\text{Al}_2\text{O}_3$ IPD stacks, sandwiched between two poly electrodes. To maximize retention performance, the most convenient way is to actually use a single layer of Al_2O_3 as IPD, since it provides the largest voltage window. When considering the erase saturation issue, the optimal stack is realized for a bottom SiO_2 of about 3 nm. However, the obtained

optimum may not be sufficient for preventing the erase saturation sufficiently, since a larger voltage may drop over the IPD.

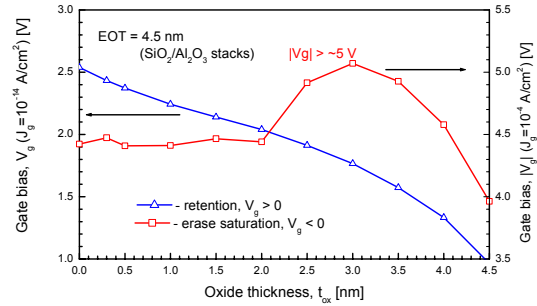


Fig. 5: Maximum gate bias at which a fixed-current retention condition is fulfilled (left) and maximum gate bias at which the constant-current erase saturation condition is still not an issue (right).

An improvement is possible if the potential barrier corresponding to the erase saturation case is increased, as already discussed in the previous section. Indeed, the use of metal gates reduces the current corresponding to gate electron injection, as shown in **Fig. 6**. On the positive side, the effect of the metal gate is mostly a shift to the right, as expected, due to the modification of the flatband voltage.

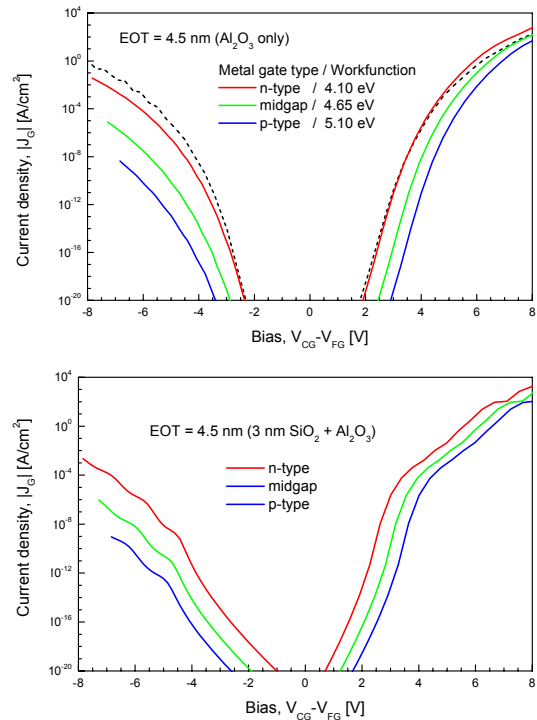


Fig. 6: Effect of the metal workfunction on the IV curves of Al_2O_3 only (top) and $\text{SiO}_2/\text{Al}_2\text{O}_3$ (bottom) stacks. The dotted lines (top plot) show the current for highly doped n+ poly top gate, for comparison purposes.

On the negative side, the main effect is a downward shift; hence using metal gate effectively reduces the erase saturation concern. It should be mentioned that these results were obtained assuming ideal materials, with no defect-assisted conduction, thus giving the largest possible window.

Fig. 7 shows the sensitivity of the current density for IPD voltage drops in the erase saturation region of interest versus the metal workfunction. According to this plot, 3-4 decades of current decrease are possible within a 1 eV window, corresponding to the transition from n-type to p-type metal gate.

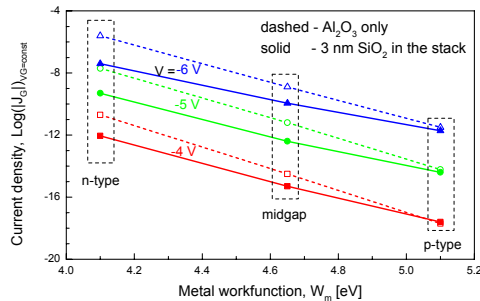


Fig. 7: Current sensitivity to metal workfunction, for an EOT stack of 4.5 nm, at different voltages.

From the previous results, it should be obvious that the best IPD stack depends upon current requirements, gate material, and whatever may be the most critical issue (retention, erase saturation, gate disturb). If the control gate remains n⁺ poly, then a single layer of Al₂O₃ remains the best choice if retention and disturb are the prime issues. By contrast, a stack with ~3 nm bottom oxide is the best solution if erase saturation is the main concern. When using midgap or p-type metal gates, the gate injection current decreases drastically, removing the erase saturation as a concern. This allows to focus on other issues, like retention, in which case a pure Al₂O₃ layer would perform best, or on disturb, in which case an Al₂O₃/SiO₂ with about 3 nm of SiO₂ would be the best choice. On top of this, the defects in the high-κ materials will presumably add more constraints, especially as far as charge retention is concerned. Another issue that must be carefully considered is the breakdown voltage [12], since the field in such electrically thin IPD stacks is very high.

We have mainly discussed here Al₂O₃-based alternative IPD's. However, we carried out a similar study for HfO₂ and concluded that, for the sub-5 nm EOT, its potential is below that of Al₂O₃. On the one hand, a somewhat thicker SiO₂ layer (around 0.5 nm) is needed to inhibit the VARIOT effect for the positive polarity, whereas on the other hand the smaller barrier height partly cancels out the improvement associated with the higher κ value, in the negative voltage range. Hf Silicates and Aluminates however, have material parameters closer to that of Al₂O₃ and also offer the possibility of "tuning" the material parameters to some extent, thus being interesting candidates, as well.

5. Conclusion

In summary, we discussed several options for scaling down the interpoly dielectrics to below 10 nm and 5 nm EOT, respectively. Sub-10 nm EOT IPD's are doable with dual (or triple) layer stacks and poly control gate.

Highly scaled sub-5 nm EOT IPD's, which would be required for the 45 nm technology node, can hardly be achieved with poly CG. Using midgap or p-type metal gate appears to be a must, whereas the design of the optimal IPD stack is subject to the available (high-κ) materials and particular cell specs. Finally, with increasing electric field through ultrathin EOT stacks, the IPD breakdown becomes an additional issue.

Acknowledgments – The authors would like to acknowledge useful discussions with P. Blomme, IMEC and A. Akheyar, Infineon Technologies assignee at IMEC. Parts of this work have been supported by the MEDEA+ Crescendo/Nemesys projects as well as by the IMEC Industrial Affiliation Program on Advanced Memory Technology together with Intel Corp. and Infineon Technologies.

References

- [1] H. Maes, J. Van Houdt: "Silicon scaling and its consequences for memory technology", Proc. NVSM Workshop, pp. 7-11, 2003 (keynote presentation).
- [2] S. Mori, E. Sakagami, H. Araki, Y. Kaneko, K. Narita, Y. Ohshima, N. Arai, K. Yoshikawa, "ONO inter-poly dielectric scaling for nonvolatile memory applications", IEEE Trans. El. Dev., **38**(2): 386-390, 1991.
- [3] S. Mori, Y.Y. Araki, M. Sato, H. Meguro, H. Tsunoda, E. Kamyia, K. Yoshikawa, N. Arai, E. Sakagami: "Thickness scaling limitation factors for ONO interpoly dielectric for nonvolatile memory devices", IEEE Trans. El. Dev **43**(1): 47-53, 1996.
- [4] Semiconductor Industry Association (SIA): International Technology Roadmap for Semiconductors (ITRS), online at <http://public.itrs.net>, ed. 2003.
- [5] 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.
- [6] P. Blomme, B. Govoreanu, M. Rosmeulen, US Patent 6,784,484, 2004.
- [7] P. Cappelletti, R. Bez, A. Modelli, A. Visconti: "What we have learned on Flash memory reliability in the last ten years", IEDM Tech. Dig, pp. 489-492, 2004.
- [8] B. Govoreanu, P. Blomme, K. Henson, J. Van Houdt, K. De Meyer: "An Effective Model for Tunneling through Ultrathin Oxides/High-k Gate Stacks from Silicon Inversion Layers and the Impact of the Stack Parameters on the Leakage Current", Solid-St. Electron., **48**(4): 617-625, 2004.
- [9] J.-D. Lee, S.-H. Hur, J.-D. Choi: "Effects of floating-gate interference on NAND Flash memory cell operation", IEEE El. Dev. Lett. **23**(5): 264-266, 2002.
- [10] C. Lee, H. Lee, S. Park, C. Park, K. Kim, K. Kim: "Physical scaling limit of NOR Flash memory cell based on floating-gate interference effect", Proc. NVSMW, pp. 65-66, 2004.
- [11] B. Govoreanu, Ph.D. Dissertation, KU Leuven, 2004.
- [12] J.W. McPherson, J. Kim, A. Shanware, H. Mogul, J. Rodriguez: "Trends in the ultimate breakdown strength of high dielectric-constant materials", IEEE Trans. El. Dev. **50**(8): 1771-1778, 2003.

ISSG RTO (In-situ Steam Generation Rapid Thermal Oxidation) Grown Tunnel Oxide on Advanced FLOTOX E2PROM Memories

Gilles Festes, Stéphanie Angle, Jean-Marc Bédrine, Mikael Berrebi, Magali Cerisier,
Eleonore Daemen, André Dumas, Michel Grégoire, Elsa Hugonnard,
Pascal Rohr, Eva Serres and Alexis Solere

ATMEL Rousset, ZI Rousset - avenue Olivier Perroy - 13106 ROUSSET Cedex - FRANCE

1. Abstract

This paper reports the benefits of an ISSG-RTO oxide as tunnel oxide in an Advanced FLOTOX E2PROM memory. Our results indicate that the uniformity and control of the tunnel oxide is considerably improved by ISSG-RTO compared to conventional furnace oxide : less sensitive to substrate doping variations, better corner rounding. Quality in terms of charge to breakdown (Qbd) is improved whatever current injection direction. Such enhanced reliability of the ISSG oxide is demonstrated on a 256kb E2PROM memory from the view point of data retention and cycling tests.

2. Introduction

The quality of the tunnel oxide in embedded E2PROM technologies is important to guarantee product reliability over time. Tunnel oxide must be well controlled in thickness, specially on heavily doped N⁺ substrate (FLOTOX cell) and close to active corners (STI technology) to prevent parasitic tunneling current which leads to data retention issues. Oxide quality (intrinsic Qbd) must also be high enough to improve memory endurance (cycling).

3. Experimental

ISSG RTO is an oxidation process where H₂ and O₂ react at the wafer surface. ISSG RTO recipe was optimized considering oxidation temperature and H₂ percentage during oxidation.

Our results indicate that the reliability of ISSG oxide is considerably improved as the process temperature increases with respect to charge-to-breakdown (Fig. 1 – Furnace oxide : 850C / wet oxidation). This improvement is explained by the reduction of the dangling bonds as the oxidation temperature increases¹.

No positive effect of ISSG H₂ content was seen on Qbd within the range 5%-30%, which is still consistent with previously reported results^{1,2}.

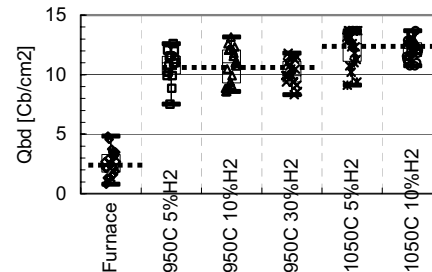


Fig. 1 : Variation of Qbd of Furnace oxide and ISSG oxides with H₂ percentages and process temperature (80000 tunnel window array 400nm x 100nm)

The optimized ISSG RTO recipe is defined as : 5% H₂ percentage during oxidation, 1050C oxidation temperature.

4. Results

The evaluation of ISSG-RTO oxide was developed on a floating gate 0.18μm 256kb E2PROM memory. Fig 2. shows the cross section of the FLOTOX cell : a 80nm tunnel oxide window is formed by placing dummy nitride spacers on either side of a nitride implant mask³. By construction, the tunnel window is self-aligned with BN⁺ drain. In order to prevent parasitic phenomena during write operation (like hot hole injection in Floating Gate and Band to Band Tunneling current in the substrate), BN⁺ implant was optimized in terms of dose, energy, twist and tilt. Resulting dose was increased from 2.3e14cm⁻² to 7.2e14cm⁻². Since the tunnel oxidation is done after BN⁺ implantation, the tunnel oxide thickness target must be adjusted on bare silicon to reach 75Å on product.

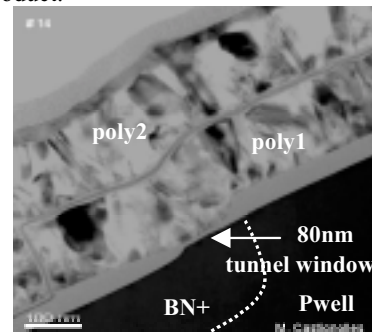


Fig. 2 : TEM cross-sections in P2/P1 storage device of an E2PROM cell (0.18μm embedded technology)

As expected, for such high doping concentrations, furnace tunnel oxide growth is very dependent on BN+ arsenic implant dose (Fig. 3a). The target must be reduced to 68Å on bare silicon to reach 75Å on product.

ISSG rapid thermal oxidation is less sensitive to substrate doping : only 2Å thicker oxide on a high BN+ doped region is found (Fig. 3b).

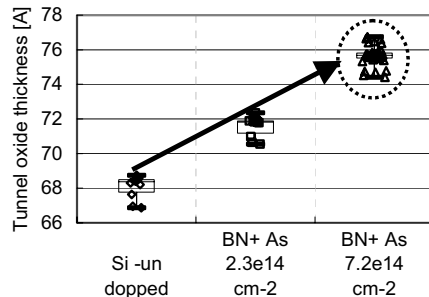


Fig. 3a : Furnace Tunnel oxide thickness variation with substrate doping.

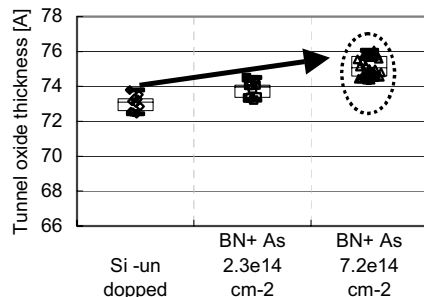


Fig. 3b : ISSG-RTO Tunnel oxide thickness variation with substrate doping.

- TEM cross-sections

TEM cross-sections through tunnel window area perpendicular to Bit Lines in the 256kb E2PROM array have been done on both furnace oxide and ISSG RTO splits (Fig. 4). A better uniformity of oxide thickness is shown on ISSG RTO : the furnace oxide results in some thinning on active corners, which is explained by a non-uniform As doping close to STI edges. Since the ISSG oxidation is much less dependent on substrate doping, ISSG RTO thickness is more uniform, even on active corners.

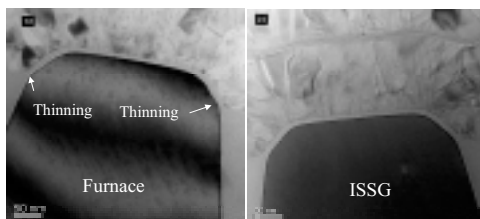


Fig. 4 : TEM cross-sections in tunnel window (ISSG RTO and Furnace oxide)

- Intrinsic Qbd (Charge-To-Breakdown)

Intrinsic charge to breakdown of a tunnel oxide is an indicator of oxide quality. Qbd tests have been provided by injecting electrons in both directions : from BN+ area to Poly1 (as in clear operation) and from Poly1 to BN+ (as in write operation). Qbd for ISSG-RTO is a factor of 3.5 higher than for furnace oxide when electrons are injected from BN+ interface to P1. On the other direction, the gain is only a factor of 2 (Fig. 4a). In that case, Qbd performance is restricted by top oxide / Poly1 interface which is not a smooth interface due to polysilicon grains⁴.

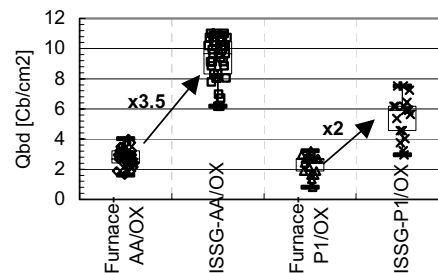


Fig. 4a : Variation of Qbd of Furnace oxide and ISSG oxides with e-injection from P1 to substrate (P1/OX) or from substrate to P1 (AA/OX).

During Qbd test, charge trapping between 1 and 10s has been measured on capacitors with and without STI edges (Fig. 4b). There is a negative influence of STI edges on both Furnace and ISSG-RTO splits. Nevertheless, ISSG-RTO intrinsic oxide quality on STI edges is nearly as good as furnace oxide quality on planar capacitor (trapping~0.04V/s)³.

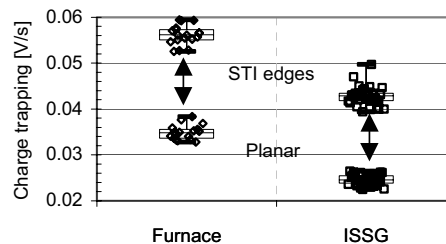


Fig. 4b : Variation of Charge Trapping of Furnace and ISSG oxides on capacitors with and without STI edges

- Data Retention on 256kb E2PROM

Data retention tests at Room temperature have been performed on both Furnace and ISSG-RTO tunnel oxides. Sensibility towards oxide thickness is a good indicator of process latitude : two thickness around 75Å typical target have been tested. Fig. 5 and Fig. 6 show retention results respectively for Furnace and ISSG oxides : only Margin High distributions have been reported since Margin Low distributions

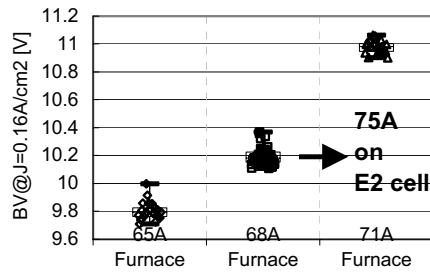


Fig. 5a : Voltage across Tunnel Oxide @J=0.16A/cm2 vs Furnace oxide thickness target.

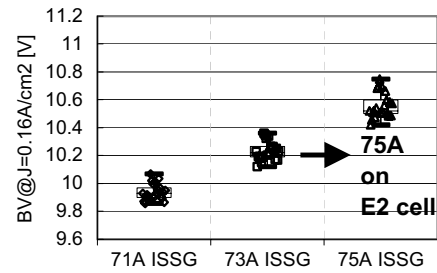
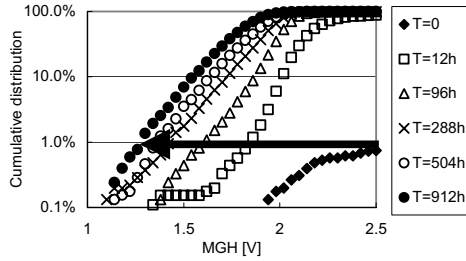
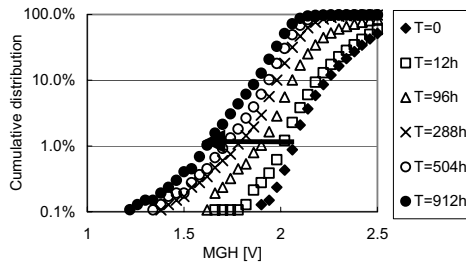


Fig. 6a : Voltage across Tunnel Oxide @J=0.16A/cm2 vs ISSG-RTO oxide thickness target.

65Å Furnace oxide (~72Å oxide on BN+)



68Å Furnace oxide (~75Å oxide on BN+)



71Å Furnace oxide (~78Å oxide on BN+)

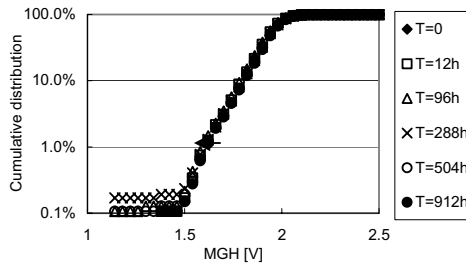
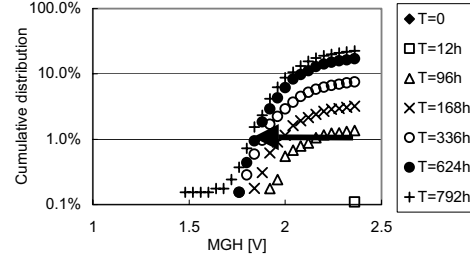


Fig. 5b : Margin High Distribution vs Retention time On 65Å-68Å-71Å Furnace tunnel oxide wafers

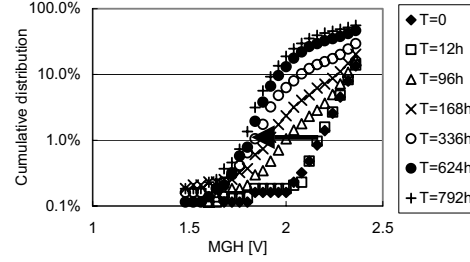
haven't shown any variation over retention time (due to the fact that margin value is clamped at -600mV).

When tunnel oxide is upper specification limit (77Å), cell programming is more difficult, resulting in lower margins. Nevertheless, retention over time is very good. When tunnel oxide is on target (75Å), initial margin is better but there is a slight shift of the distribution after 1000hours. This behavior is worse for furnace oxide.

71Å ISSG-RTO oxide (~72-73Å oxide on BN+)



73Å ISSG-RTO oxide (~75Å oxide on BN+)



75Å ISSG-RTO oxide (~77Å oxide on BN+)

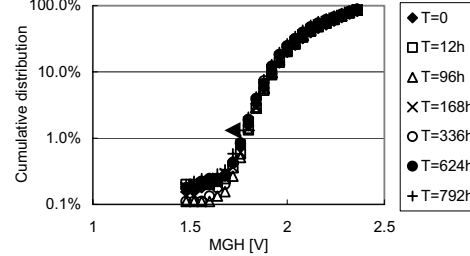


Fig. 6b : Margin High Distribution vs Retention time On 71-73-75Å ISSG tunnel oxide wafers

At tunnel oxide around 72-73Å on BN+ doped region (which is 2Å lower specification limit), initial margins are very good in both cases (close to 2.4V clamp value), but the margin loss over retention time is not acceptable on a Furnace tunnel oxide and still good on a ISSG tunnel oxide. Looking to I-V characteristics on tunnel oxide capacitors, both splits have nearly the same curve, resulting in the same voltage (9.8-9.9V) for a given current density (Fig. 5a and 5b). Based on these results we can conclude that the process latitude is significant larger for the ISSG-RTO process.

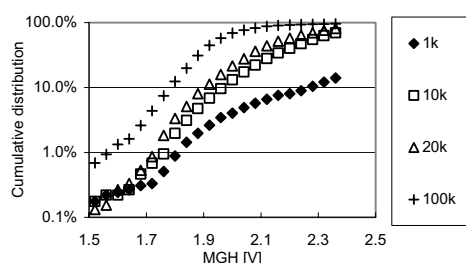
Another important point, as illustrated on the TEM cross section (Fig.4), is that the Furnace grown Tunnel oxide on heavily N⁺ doped region is not uniform close to the STI edges. This local thinning of the Tunnel oxide results in lower BV values (200-300mV). As shown in figure 5a and 5b, reducing the Furnace grown Tunnel oxide by 200-300mV can become catastrophic for data retention. The process latitude using Furnace tunnel oxide is so extremely reduced.

By switching to an ISSG-RTO tunnel oxide, there is no thinning at the STI edges and the data retention is less sensitive towards the Tunnel oxide thickness, we can conclude that the process latitude is significantly improved.

- Endurance on 256kb E2PROM

Cycling tests have been performed in nominal test conditions (25C / Programming Time = 2ms – Fig .7). There is no difference in margin loss over number of cycles. Based on better intrinsic Qbd (by a factor 2 to 3.5), a gain in endurance by at least a factor 2 was expected. This is not obvious on product cycling, which can be controlled by other mechanisms (charge pump degradation, junction walk-out, ...). Margin High cumulative distributions are sharper on ISSG-RTO tunnel oxide, even after 100kcycles. This could be related to the more uniform tunnel oxide over the whole memory array.

68Å Furnace oxide (~75Å oxide on BN+)



73Å ISSG-RTO oxide (~75Å oxide on BN+)

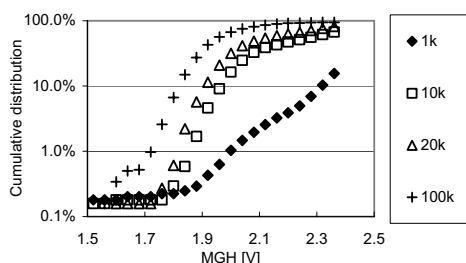


Fig. 7 : Margin High Distribution vs Cycles
On Furnace and ISSG tunnel oxide wafers

- Data Retention after Cycling

Data retention tests have been done on 100k cycled parts. Compared to initial margins, about 500mV loss is due to cycling ; 100mV loss are related to retention loss after cycling. There are slightly better results on retention after cycling on ISSG-RTO oxide than on furnace oxide (50mV compared to 100mV loss).

5. Conclusions

ISSG RTO tunnel oxide results in a better intrinsic quality compared to a conventional furnace oxide. The ISSG RTO process also leads to more uniform tunnel oxide on heavily doped substrate (growing is less doping dependant) : this property is a key-point to guarantee a well-controlled tunnel oxide in an Advanced FLOTOX E2PROM. Data retention (and cycling) tests on a 256kb E2PROM show that a more reliable product can be achieved using ISSG tunnel oxide.

Therefore, single-wafer ISSG RTO processing for tunnel oxide is demonstrated to be a more robust alternative to the conventional batch-type furnace processing.

References :

- [1] K.C. Chen, "Cycle time and Process Improvement by Single Wafer Thermal Processing in Production Environment", 10th IEEE International conference on Advanced Thermal Processing of Semiconductors – RTP2002, pp171-175.
- [2] T.Y. Luo, "Effect of H₂ content on Reliability of Ultrathin In-Situ Steam Generated (ISSG) SiO₂", IEEE Electron Device Letters, Vol.21, No9, pp430-432.
- [3] Atmel US Patent 6624027, "Ultra small thin windows in floating gate transistors defined by lost nitride spacer"
- [4] Tuung Luoh, "Single-Wafer Polysilicon Engineering for the Improvement of Over Erase in a 0.18um Floating-Gate Flash Memory", IEEE Trans. On Semiconductor Manufacturing, Vol.16, No2, pp155-159.
- [5] C. Dijkstra, "ISSG RTO grown Tunnel oxide for Improved Reliability of Flash Memories", NVSMW2003, pp85-86.

SONOS flash memories with HfO₂ or HfSiON

Rob van Schaijk, Michiel van Duuren, François Neuilly, Wilko Baks,
Almudena H. Miranda, Michiel Slotboom, Nader Akil and Pablo G. Tello

Philips Research Leuven, Kapeldreef 75, B-3001 Leuven, Belgium
Phone: +32-16-281945, Fax: +32-16-281706, Email: Rob.van.Schaijk@philips.com

Abstract

SONOS embedded flash memories, with either the top layer or trapping layer made of the high-K materials HfO₂ or HfSiON, have been characterized. The memories with a high-K top layer are indicated by SHINOS (Silicon-High-K-Nitride-Oxide-Silicon). This paper shows the operation and reliability aspects of N- and P-type SHINOS memories: endurance, retention and read disturb.

A strong gain in program/erase efficiency can be obtained for SHINOS (high-K top layer) compared with SONOS memories. If HfO₂ or HfSiON replaces the nitride layer, the trapping efficiency is significantly reduced. By using P-type SHINOS it is possible to operate the devices with a symmetric V_t window around 0V. In this case, read disturb is avoided and the advantages of the 2T NOR architecture are fully used, meaning fast random access and real low power operation.

The P-type SHINOS devices show good endurance and retention. The retention of the SHINOS devices can be considerably improved by trading-off the improved program/erase efficiency with a thicker bottom oxide.

Keywords: SONOS, flash, non-volatile memory, high-K materials, HfO₂, HfSiON, reliability

Introduction

The present industry standard for non-volatile semiconductor memories is based on charge storage on a floating gate. These types of devices suffer from scaling issues for the coming generations [1][2], which make charge storage in the nitride layer of a silicon-oxide-nitride-oxide-silicon (SONOS) device an attractive option [3][4]. The nitride based memory devices offer better CMOS compatibility, a higher radiation tolerance, no erratic bits and better endurance, compared to their floating gate based counterparts.

The SONOS devices operate with a relatively thin bottom oxide (2.0-2.5nm), with a nitride layer sandwiched between this bottom oxide and a thicker top oxide. The cell is programmed (erased) by direct tunnelling of electrons (holes) uniformly through the bottom oxide into the nitride, driven by a gate bias. The trapping properties of the nitride make downscaling of the tunnel oxide possible, which results in a reduction of the high voltage for program and erase and therefore in the reduction of overall memory area. The revival of interest for this old concept is enabled by the greatly improved semiconductor fabrication methods and is

mainly driven by the better scaling perspectives and the ease of integration, which is essentially important for embedded memories. The main reliability concerns for SONOS memories are data retention and read disturb and both need improvement to obtain reliable SONOS memories [5][6]. In this paper solutions for these concerns are proposed and investigated. The data retention issue is mainly caused by charge loss through the thin bottom oxide of ~2nm. A thicker bottom oxide is no solution because erasing by means of tunnelling becomes impossible due to erase saturation [6][7]. A high-K top dielectric (HfO₂ and nitrided Hafnium Silicate; HfSiON) allows for a thicker bottom oxide without compromising the program/erase efficiency [8][9]. Besides the replacement of the top oxide layer by a high-K dielectric layer, we also tried replacing the nitride layer by these high-K materials.

The origin of the read disturb is the positive gate voltage during read, which is needed because the V_t window is not symmetric around 0V, causing an accelerated V_t shift of the erased cells. The underlying cause for the non-negative erased V_t is the so-called erase saturation. It is shown that P-type SHINOS has a symmetrical V_t window around 0V and reduced erase saturation.

The SHINOS memories described in this paper use 2 transistor (2T) NOR architecture, which is well suited for low power applications demanding random accessibility like for instance direct code execution for smart cards [10].

SHINOS fabrication and cell structure

The device performance is determined by the properties of the layer stack consisting of bottom oxide, trapping layer and top dielectric. The bottom oxide thickness mainly determines the data retention, but the need for direct tunnelling of electrons and holes excludes oxide thicknesses above ~3nm. Bottom oxide thicknesses of 2.2nm, 2.4nm and 3.0nm (diluted wet oxide) were investigated.

The trapping layer of 6nm was in most devices stoichiometric nitride, but also HfO₂ and nitrided hafnium silicate (HfSiON) were investigated as trapping layers. The HfO₂ and HfSiON layer were deposited using MOCVD technique. The HfSiON layer is a HfSiO₄ layer nitrided in NH₃ for 30' at 800°C.

The top layer splits were oxide, HfO₂ and HfSiON. The top oxide was deposited 8nm high temperature oxide (HTO). Both HfO₂ and HfSiON have a thickness of 10nm or 14nm. After complete processing of the devices, both the HfO₂ and HfSiON layer are poly-

crystalline. From CV measurements, the K value was derived and for both layers $K \approx 15$. On top of the HfO_2 layer a HfSiO_4 cap of 1nm was deposited to prevent reaction with the poly gate. In figure 1, a cross section TEM graph of the SHINOS device is shown.

The NOR array architecture of the 2T SONOS cell consists of an access transistor and low V_t SHINOS transistor in series. The access transistor makes a symmetric V_t window around 0V possible. The poly wordlines consist of access and control gates (AG, CG), whereas shared drain contacts connect the cells to metal bitlines. In this paper results of capacitors, 256 bit and 26kbit arrays with either N-type or P-type SONOS/SHINOS devices will be presented.

Results and Discussion

SHINOS with high-K top layer:

In fig. 2, program and erase curves ($\pm 10\text{V}$) are shown for SONOS with 8nm HTO top layer and SHINOS with 10nm HfO_2 top layer. In both cases the bottom oxide is 2.2nm with a 6nm nitride layer. Like all V_t 's in this paper, the V_t 's in this graph are measured with a $5\mu\text{A}$ current criterion at $V_d = 0.5\text{V}$. The increased program/erase efficiency for the SHINOS device is clearly visible. Already after $100\mu\text{s}$ a large difference between SHINOS and SONOS in both the programmed and erased V_t can be observed.

Erasing does not result in a negative V_t due to erase saturation: at high erase voltages and/or long erase times an equilibrium is reached between electrons tunnelling through the top layer and holes through the bottom oxide [6]. Also programming saturation exists for the SHINOS device. This is not due to equilibrium between electron and hole tunnelling during programming, but caused by an electron current (leakage) to the gate. This effect occurs in SHINOS because there is no effective energy barrier between nitride and HfO_2 [11].

These leakage currents in SHINOS devices result in much lower Q_{bd} values compared with SONOS. In figure 3, Weibull plots of Q_{bd} measurements are shown for SHINOS (14nm HfO_2) and SONOS (8nm HTO).

For N-type SHINOS, good endurance up to 10^6 cycles with a standard deviation for the programmed/erased states of $\sim 100\text{mV}$ is shown in fig. 4. The V_t window is between 3.0V for the programmed state and 0.5V for the erased state. This implies that a read voltage of approximately 2V is needed for this device; meaning boosting is necessary and excluding low power use.

This high read voltage results in a considerable read disturb on the erased state. Read disturb measurements at a gate stress of 2.5V on a 10^5 cycled SONOS array are shown in fig. 5. The erased V_t distributions during stress show that all bits behave in the same way; meaning no 'tail' bits can be observed in contrast to what is seen with floating gate devices. The significant V_t shift during 10^5 seconds proves the necessity for a low read voltage.

This problem can be solved largely with P-type SHINOS devices: first the saturation effect is reduced, because the P-type gate of a P-type transistor injects fewer electrons into the nitride than the N-type gate of an N-type transistor [7], and secondly, thanks to the

complementary behaviour of the P-type device, the V_t window is symmetric around 0V, see fig. 6. The definition of the V_t for the enhancement P-type devices is chosen negative. The program saturation of the P-type SHINOS is no issue, because when this happens, the V_t is already very negative.

Also for P-type SHINOS a huge gain in program/erase efficiency can be obtained when a high-K material is used as top layer. For 8nm oxide, 10nm and 14nm HfO_2 and 10nm HfSiON the V_t window is shown in fig. 7. A reduction in program/erase time of a factor of 100 can be gained changing 8nm oxide into 10nm HfSiON .

In fig. 8, the endurance for P-type SHINOS is shown with a symmetric V_t window of 4V around 0V. This means that 0V can be used and no extra disturb during read is present. In the P-type SHINOS case both the programmed and erased V_t become lower after 10^5 cycles. This is opposite to N-type SHINOS, where the V_t of both states increases after cycling. It seems that in P-type SHINOS positive fixed charge is present after cycling in contrast to negative fixed charge in N-type SHINOS. The reason for this difference is not yet understood.

All data presented above were measured for 2.2nm thick bottom oxides. By trading-off the lower program/erase voltages for SHINOS against a thicker bottom oxide, the data retention can be improved. Although a strong reduction of the V_t window was observed upon going from 2.2nm to 3.0nm, SHINOS still works for 3.0nm whereas SONOS simply does not work above 2.5nm [5] (see fig. 9).

For P-type SHINOS with 3.0nm bottom oxide, the endurance is shown in fig. 10. For this 'thick' bottom oxide, it is more difficult to obtain a symmetric V_t window, due to the strongly decreased hole tunnelling.

The difference in data retention between 2.2nm and 3.0nm bottom oxide for P-type SHINOS is shown in fig. 11. More than 3 orders in retention time can be gained if 3.0nm bottom oxide is used.

SHINOS with high-K trapping layer:

If the nitride trapping layer is changed for either HfO_2 or HfSiON , the program/erase efficiency is decreased significantly. In fig. 12, P/E curves for nitride and HfSiON are shown. Nitride is superior compared with these high-K materials, most likely due to the higher trap density. Also the data retention of these three materials (sandwiched between oxides) is best for nitride.

Conclusions

We demonstrated SHINOS memories, with the top dielectric layer or the trapping layer made of high-K material (HfO_2 or HfSiON). The replacement of the top oxide significantly improves the program/erase efficiency. The devices operate at write voltages lower than 10V with P/E times of around 10ms. For N-type SHINOS the V_t window is not symmetric around zero due to erase saturation, resulting in a considerable read disturb. This problem can be solved largely with P-type SHINOS devices, which reduces erase saturation and exploits low voltage read, potentially improving on

reliability performance. As trapping material, nitride is superior compared with the high-K materials HfO_2 and HfSiON . For SONOS a main reliability issue is the data retention. By trading-off the lower program/erase voltages for SHINOS against a thicker bottom oxide, the data retention time can be improved by three orders of magnitude.

References

- [1] H. Maes and J.F. Van Houdt, *proc. IEEE NVSMW*, 2003, pp. 7
- [2] M. van Duuren *et al.*, *proc. IEEE NVSMW*, 2003, pp. 73

- [3] F.R. Libsch and M.H. White, W.D. Brown and J.E. Brewer, Ed. New York, IEEE press, 1998, pp. 309
- [4] Y. Kamigaki and S. Minami, *IEICE Trans. Electron*, vol. E84C, 2001, pp. 713
- [5] R. van Schaijk *et al.*, *Microelectronic Engineering* 72, 2004 pp. 395
- [6] R. van Schaijk *et al.*, *proc. ESSDERC*, 2004, pp. 270
- [7] H. Bachhofer *et al.*, *Journal of Appl. Phys.* 89, 2001, pp. 2791
- [8] S. Choi *et al.*, *Journal of Appl. Phys.* 94, 2003, pp. 5408
- [9] C.H. Lee *et al.*, *proc. IEDM*, 2003, pp. 613
- [10] G. Dormans *et al.*, *proc. SSDM*, 2001, pp. 540
- [11] J. Robertson, *Journal Vac. Sci. Technol. B18*, 2000, pp. 1785

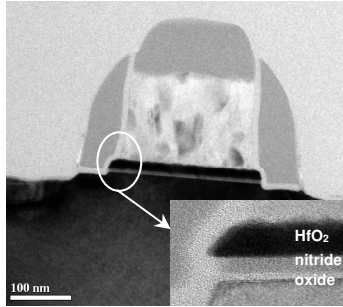


Figure 1. TEM graph of a SHINOS device with 2.2nm bottom oxide, 6nm nitride and 10nm HfO_2 top layer.

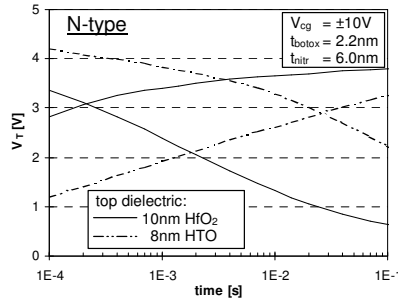


Figure 2. P/E curves for N-type SONOS and SHINOS device (average V_t of 256 bit arrays).

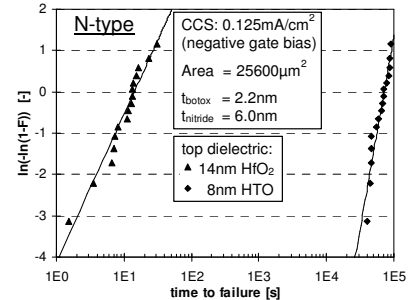


Figure 3. Time to failure for constant current stress (CCS) of SONOS and SHINOS capacitors.

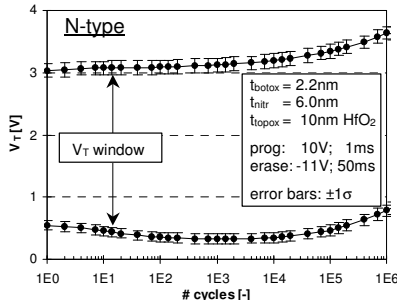


Figure 4. Endurance of a 256 bit array of N-type SHINOS.

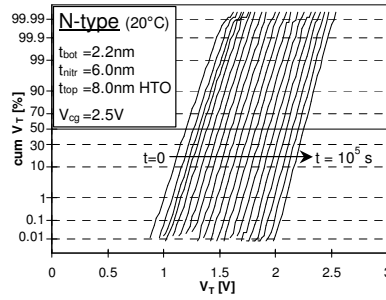


Figure 5. Cumulative V_t distribution of the erased V_t during 2.5V gate stress on a 26kbit N-type 2T SONOS array.

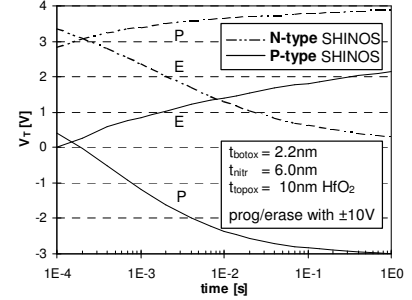


Figure 6. P/E curves of N-type and P-type SHINOS.

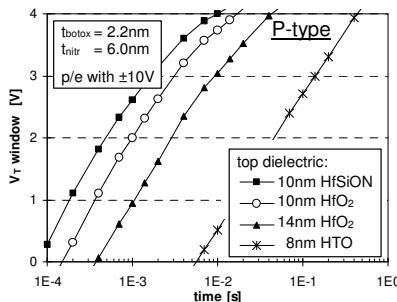


Figure 7. V_t window (difference between programmed and erased V_t) for P-type SONOS and SHINOS devices.

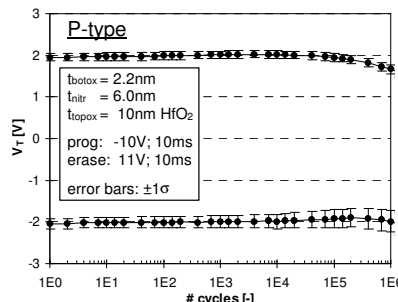


Figure 8. Endurance of a 256 bit array of P-type SHINOS.

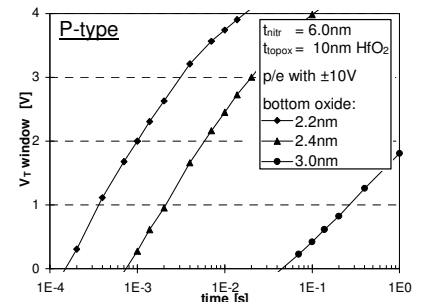


Figure 9. V_t window for P-type SHINOS with different bottom oxide thickness.

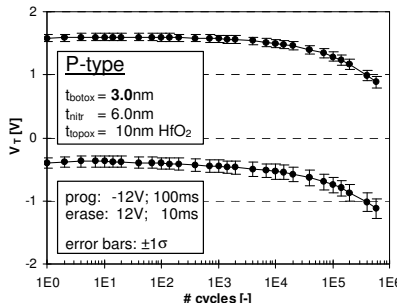


Figure 10. Endurance of a 256 bit array of P-type SHINOS with 'thick' bottom oxide.

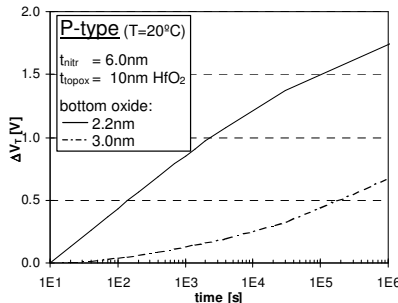


Figure 11. Data retention of P-type SHINOS with 2.2nm and 3.0nm bottom oxide thickness.

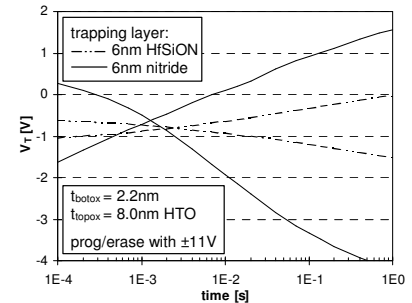


Figure 12. P/E curves of P-type SHINOS with nitride or HfSiON as trapping layer.

Fully Compatible Novel SNONOS Structure for Improved Electrical Performance in NAND Flash Memories

Jeong Hee Han, Ju Hyung Kim, Seung Hyun Lee, Sanghun Jeon and Chungwoo Kim

Materials & Devices Research Center, Samsung Advanced Institute of Technology,
Mt. 14-1, Nongseo-Ri, Giheung-Eup, Yongin-Si, Gyeonggi-Do, Korea 449-712,
Email: jeonghee.han@samsung.com

Abstract

A fully compatible gate stack structure with an additional stoichiometric silicon nitride layer (referred to as “SNONOS”) is used to improve the electrical performance of nitride-based memories. By employing the new structure, the detrimental “back-tunneling” effect is suppressed without sacrificing device capacitance. Thus, a threshold voltage improvement of approximately 1.2V is observed for the erase state. Retention for devices with similar erase performance is also improved with the SNONOS structure. Further mitigation of the “back-tunneling” effect is observed for devices with high work function gate material.

1. Introduction

With increasing demand for high-density, high-performance semiconductor memories, NAND Flash memories are capturing more and more memory market share. Research on non-volatile NAND Flash memories has steadily garnered attention for the past several years. However, as conventional floating-gate type NAND memory is continually scaled down to the nanometer regime, many integration obstacles become more evident. One of the main problems is the capacitance coupling effect which occurs between adjacent devices [1]. In addition, traditional floating-gate type memories are also vulnerable to critical pin-hole defects in the tunnel oxide [2]. Thus, more interest is being given to nitride-based alternatives.

Nitride-based memories, in particularly, silicon-oxide-nitride-oxide-silicon (SONOS) illustrated in Fig. 1a, differ from floating-gate type memories in that a trap-filled nitride layer stores the charge, rather than a conductive polysilicon layer. Since the nitride is a dielectric with unique charge traps, there is less capacitive coupling concern and hardly any fatal pin-hole defects in the tunnel oxide [2]. However, SONOS is not without its difficulties. Due to the nitride layer being a dielectric, a thinner tunnel oxide is required. Moreover, a phenomenon widely referred to as the “back-tunneling” effect limits the electrical performance [3].

Due to the simplistic structure of the NAND Flash SONOS device structure, there are very few options in improving the electrical performance and reducing the back-tunneling phenomenon. Previous attempts to use alternative (i.e. high-k) materials have been studied [4], but integration problems and a need for a metal gate limits its potential. However, by employing a silicon nitride-silicon oxide bi-layer in the blocking oxide, a drastic improvement in electrical performance can be

achieved. In other words, the new gate stack is a silicon-nitride-oxide-nitride-silicon (SNONOS) structure and is illustrated in Fig. 1b. Additionally, the structure can be fabricated by conventional silicon processing and can be completely integrated with present-day equipment and technology.

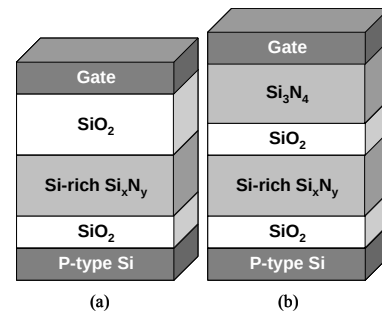


Fig. 1: Illustration showing the gate stacks of (a) SONOS and (b) SNONOS memory devices.

2. Device Fabrication and Experiment

Samples were fabricated on p-type (100) wafers. The gate stack structure was fabricated using only conventional silicon processes. The tunnel oxide was thermally grown at 850°C in an oxygen atmosphere and annealed in N₂O at the same temperature. A silicon-rich nitride layer was then deposited using LPCVD at 730°C with a dichlorosilane (DCS) to NH₃ ratio of 4. The higher ratio was used to deliberately increase the number of traps in the charge storage layer [5]. The blocking oxide was also deposited using LPCVD at 450°C with a SiH₄ to O₂ ratio of 5. For SNONOS samples, an additional stoichiometric silicon nitride was deposited with a DCS to NH₃ ratio of 0.13. This DCS:NH₃ ratio resulted in a film with much less trap density than the silicon-rich nitride. E-beam evaporation at 2×10⁻⁶ Torr was used to deposit either aluminium or gold as the gate material. Aluminium and gold were used since they have similar electrical properties (i.e. work function) to n⁺ or p⁺ polysilicon, respectively [6]. Film thicknesses were measured by spectral ellipsometry and confirmed using transmission electron micrographs (TEM). Capacitance-voltage measurements were performed to ensure that the equivalent oxide thickness (EOT) was similar for both SONOS and SNONOS samples.

3. Results and Discussion

The back-tunneling effect is predominantly observed during erase operation of the device, as shown in Fig. 2. In order to remove trapped electrons from the nitride layer, a negative bias is applied to the gate. However, for devices with thinner blocking oxides, electrons from

the gate can tunnel through the blocking oxide layer and back into the nitride. As a result, the threshold voltage either saturates or increases, as shown in Fig. 3. Thus, the device may not reach a full “erased” state. It is important to note that applying a higher erase bias on the gate amplifies the undesirable effect. Yet, reducing the erase bias will increase the time needed for erase operation.

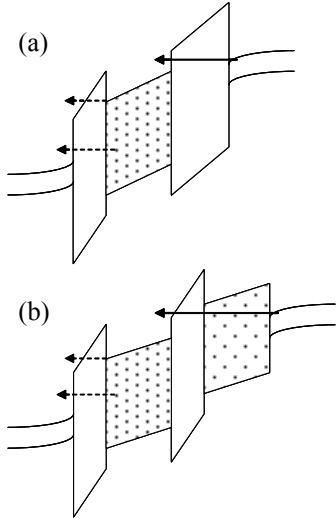


Fig. 2: Energy band diagrams showing the charge paths during the erase operation of (a) SONOS and (b) SNONOS gate stack structures. The dashed arrows denote the removal of charge from the nitride layer. The solid arrows denote the “back-tunneling” effect.

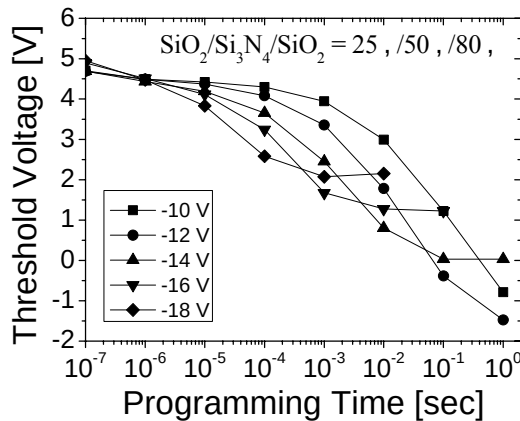


Fig. 3: Using a more negative voltage initially results in faster erase operation. However, the saturation threshold voltage increases.

With the SNONOS gate stack structure, a physically thicker blocking oxide is integrated without sacrificing capacitance. Fig. 4 shows the electrical performance comparison between optimized SONOS and SNONOS devices with the same capacitance. Both structures have a tunnel oxide thickness and charge storing nitride thickness of 30Å and 70Å, respectively. The blocking oxide for the SONOS samples is 65Å. To maintain the capacitance, the blocking layer of the SNONOS samples consists of a 30Å oxide layer and a 70Å nitride layer. Both devices have aluminium metal gates.

With a write voltage of +17V, both SONOS and SNONOS behave similarly. For write programming pulses longer than 2ms, a slight decrease in threshold voltage is observed for SNONOS devices. This may be attributed to interface affects in the blocking layer of the device. However, the decrease is not significant because most NAND memory devices are written in several microseconds. Moreover, the threshold voltage never falls below 1V, a critical voltage for NAND-type memory specifications.

As for the erase operation characteristics of the two devices, there is a more drastic difference. For erase programming pulses greater than 20μs, the SNONOS structure provides better performance. The threshold voltage of the new structure is reduced much faster than the SONOS counterpart with the same pulse length. For a 2ms pulse, the SNONOS threshold voltage is approximately 1.2V lower than the SONOS threshold voltage. Because the thickness of the tunnel oxide and the charge storing nitride layer are the same, it is believed that the reduction in the back-tunneling effect for the SNONOS device is the cause for the improved performance.

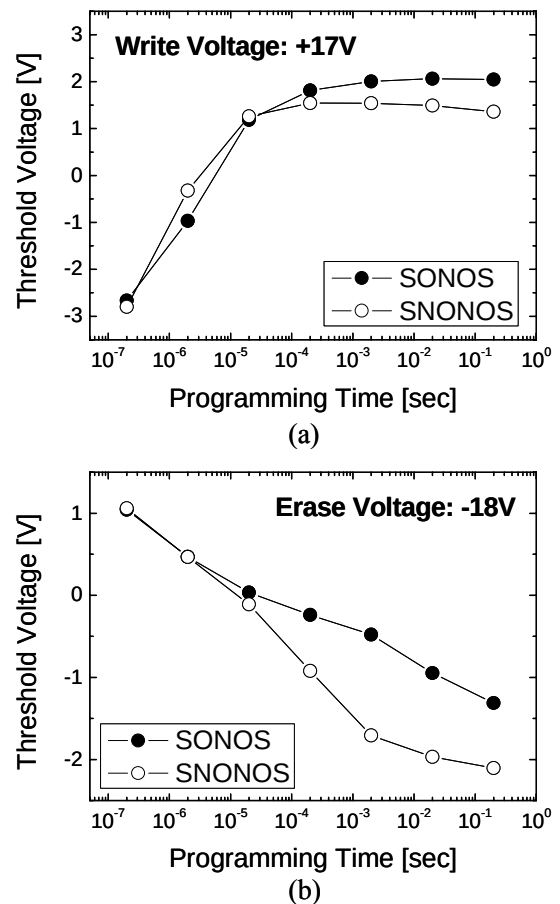


Fig. 4: (a) Write and (b) erase characteristics for SONOS and SNONOS devices. Although the write performance is similar for the two devices, SNONOS provides better erase performance due to the suppression of “back-tunneling” effect.

It is commonly known that there is a trade-off between charge retention and erase operation characteristics in nitride based memories [7]. For

devices with thicker tunnel oxides (and thinner blocking oxides to compensate for the capacitance), charge retention may be better. However, the back-tunneling effect is more severe and erase performance worsens. Conversely, devices with thinner tunnel oxides have faster erase operation characteristics but poor charge retention. In this study, retention was characterized by the threshold voltage shift (electron charge loss) of a programmed device after a two hour 250°C oven bake. Fig. 5 illustrates the trade-off for SONOS and SNONOS devices fabricated for optimization purposes. It was observed that the SNONOS devices had improved retention characteristics with similar erase properties. Once again, this typifies the improvement of electrical performance with the new gate stack structure. Additionally, electron loss stemming from charge paths across the blocking oxide may also be reduced with the physically thicker blocking layer of SNONOS, resulting in a smaller threshold voltage shift.

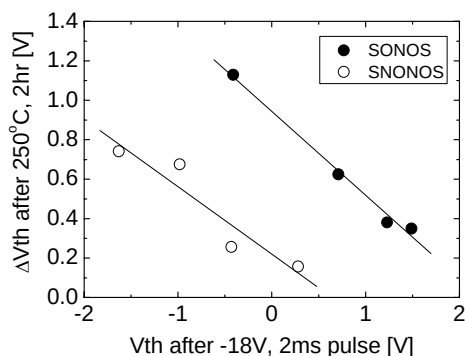


Fig. 5: Charge retention vs. erase characteristics. SNONOS gate stack structures show dramatic improvement of the trade-off.

In an effort to further improve the erase performance by reducing the back-tunneling effect, a metal gate with a higher function was used. Devices with gold electrodes having a work function of approximately 5.1eV were compared to devices with aluminium electrodes having a work function of approximately 4.1eV. The write and erase characteristics are shown in Fig. 6. For the write operation, the device performance is similar, regardless of the electrode metal. However, for the devices with gold electrodes, a more noticeable decrease in threshold voltage occurs after 2ms. This may be a result from hole “back-tunneling” effect. The hole barrier height for electrodes with higher work functions is reduced. Thus, during the write operation, holes from the gate tunnel through the blocking layer and enter the nitride, thus decreasing the threshold voltage.

However, when employing a high work function gate, the difference in erase performance is much more drastic. For the devices presented in Fig. 6b, the erase performance is similar for pulse widths up to 200μs. However, the NONO dielectric stack in combination with the gold electrodes can achieve threshold voltages of 1.5V and 2.0V lower than ONO dielectric stack with gold electrodes and NONO with aluminum electrodes,

respectively. By increasing the electron barrier height, the length of the blocking layers which the electron must tunnel through is widened.

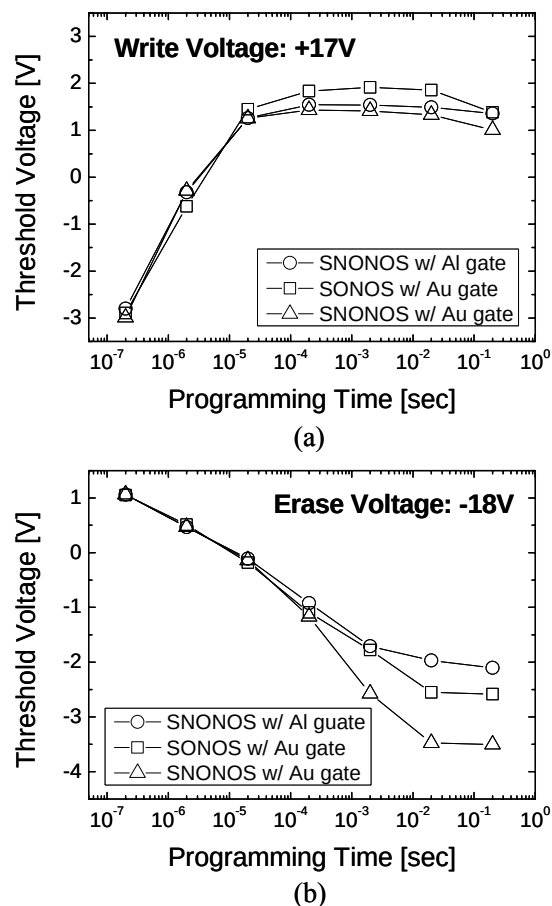


Fig. 6: (a) Write and (b) erase characteristics showing the effect of high work-function gates. Write operation is similar for all devices shown, whereas erase operation shows a dependence on work function.

5. Conclusions

A new gate stack structure was investigated to mitigate the “back-tunneling” effect which prevents nitride-based memories from achieving good erase performance. With an additional stoichiometric silicon nitride layer between the blocking oxide and gate, the threshold voltage can be reduced approximately 1V for the same erase pulse length. The additional layer also results in the improvement of the retention-erase trade-off commonly observed for nitride-based memories. With the incorporation of a gate material with higher work function, the back-tunneling effect can be further suppressed. By this means, the charge transport through the blocking layers is lengthened and results in further improved erase performance. Because it uses no high-k materials, there are no integration issues and the novel structure can be completely fabricated by conventional silicon processes.

Acknowledgement

This work was supported by the National Program for Tera-Level Nano-Devices through the Ministry of Science and Technology, Korea.

References

- [1] J. D. Lee *et al*, IEDM Tech. Dig., 264 (2002)
- [2] M. H. White *et al*, Circuits and Devices, **16**, 22 (2000)
- [3] H. Reisinger *et al*, VLSI Symp. Tech. Dig., 113 (1997)
- [4] C. H. Lee *et al*, IEDM Tech. Dig., 613 (2003)
- [5] Y. Yang and M. H. White, Solid-State Electron. **43**, 2025 (1999)
- [6] M. Koyama *et al*, IEDM Tech. Dig., 299 (2004)
- [7] L. Lundkvist and C. Svensson, Solid-State Electron. **16**, 811 (1973)

Scaling effects in dual-bit split-gate nitride memory devices.

L. Breuil, L. Haspeslagh, M. Lorenzini, J. De Vos, J. Van Houdt

IMEC Kapeldreef 75, 3001 Leuven, Belgium

Breuil@imec.be

Abstract

This paper investigates the aggressive scaling of a dual-bit split-gate memory device based on nitride storage. Devices operating with very short storage area lengths of less than 10 nm are demonstrated. Using a optimized erase scheme, the erase performance is drastically enhanced as the storage area is scaled down. It is shown that this device allows for very aggressive scaling without having the main problems associated with the two bits interference as seen in other concepts.

1. Introduction

Devices based on nitride storage have shown a high potential for scaled-down Non Volatile Memory. E.g., the NROMTM concept can achieve very high density by physically storing two bits in the same cell [1]. However, the channel length scaling is still difficult for two reasons. Firstly because of the mutual influence of both bits, known as the second bit effect [2]. Secondly because programming with Channel Hot Electrons Injection (CHEI) requires a high drain voltage which limits the scaling of the channel length due to punchthrough. In other studies, the physical isolation of the bits has been addressed on single gate devices by complicated sidewall process [4, 5]. Also, spacers have been used as additional gates allowing Source Side Injection programming (SSI) [6]. The main drawback of these solutions is the difficult control of the storage areas lengths defined by spacers. Therefore, we have developed a split-gate device where two bits are stored under different Program Gates (PG) while sharing the same Control Gate which serves as the wordline in the memory array (fig. 1) [3]. This structure, allows physical isolation of the bits, removes the second bit effect, and allows the use of the SSI programming mode. The storage areas are not defined by spacers, and yet, we show in this paper that they can be made even shorter. The structure of the cell is fully self-aligned, making it highly scalable. Indeed, because of the punchthrough mentioned above, the lower limit of the channel length in the NROMTM case is much higher than the actual feature size as made available from advanced lithography. Therefore our three-gate-device will not lead to a longer channel than a single gate device. Moreover, scaling of the PG transistors is expected to improve the erase operation because of narrower storage regions for electrons and holes. In this paper, we investigate the aggressive scaling of these PG transistors, and the associated effects on erase performance.

2. Device structure and processing

The device structure (fig. 1) and basic processing steps were presented in [3]. In this study, the device is processed in a 90nm technology node. Very short effective PG length (L_{PGeff}) are obtained by combining two methods. Firstly, an isotropic hardmask trimming technique was applied for the first poly lines definitions, allowing to obtain PG lengths (L_{PG}) reduced by about 45 nm as compared to the PG length on mask (L_{PGmask}). Fig. 2 shows that a good uniformity of L_{PG} is obtained. L_{PGmask} can be defined as small as 100nm, leading to L_{PG} of about 55 nm. Secondly, The LDD junctions are diffused under the PG's, to around 55nm from the PG edge as estimated by TCAD. Therefore, L_{PGeff} can be scaled down to a few nm's ; and different L_{PGeff} are obtained by varying (L_{PGmask}) in steps of 10nm. The different PG lengths used in this study are summarized in table 1.

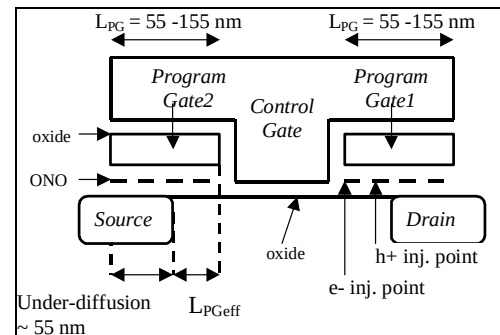


Fig. 1: Device structure.

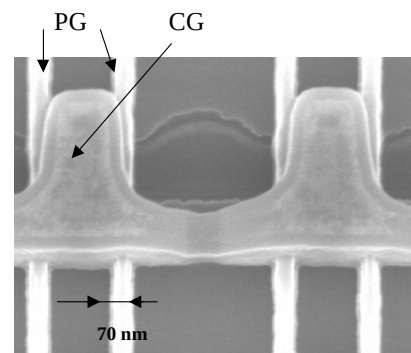


Fig. 2: SEM picture of two adjacent devices with $L_{PG} = 70$ nm.

L_{PGmask}	200	140	130	120	110	100
$L_{PG} \sim$	155	95	85	75	65	55
$L_{PGeff} \sim$	100	40	30	20	10	0

Table 1: Different available PG length (in nm), and estimated effective PG lengths.

3. Short PG sub-channel characteristics

The operating principles of this device can be found in [3,7]. As L_{PG} is scaled down, important short channel effects are observed for the PG transistor. Drain Induced Barrier Lowering can be substantially reduced by reading the device from the source [7] (fig.3).

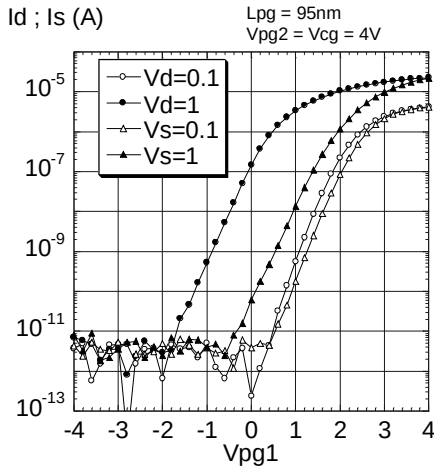


Fig. 3: Important DIBL when reading a short PG device from the drain, can be strongly reduced by reading from the source.

However, another barrier lowering component is caused by the voltage V_{cg} applied to the adjacent CG. When the PG sub-channel becomes short enough, the CG potential is able to decrease the V_t of the PG transistor (fig. 4). This effect is more difficult to avoid since V_{cg} has to be high enough for generating a high enough read-out current. Fig. 5 shows typical $I_s V_{pg1}$ transconductance characteristics as a function of the PG length, under read-out conditions. The PG transistor behaviour gets more and more degraded as the PG sub-channel is scaled down.

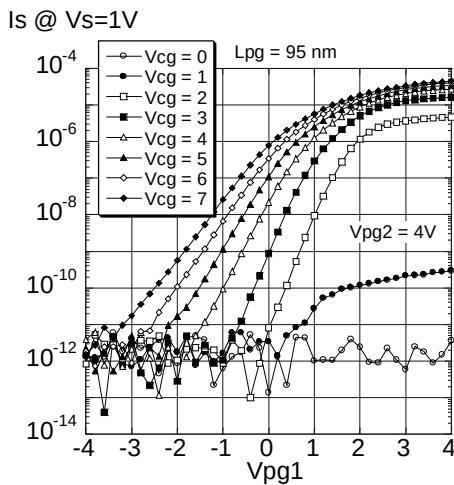


Fig. 4 : $I_s V_{pg1}$ transconductance characteristics for different CG voltages. PG barrier lowering induced by the CG voltage decreases the V_t of the PG transistor.

Is @ Vs=1

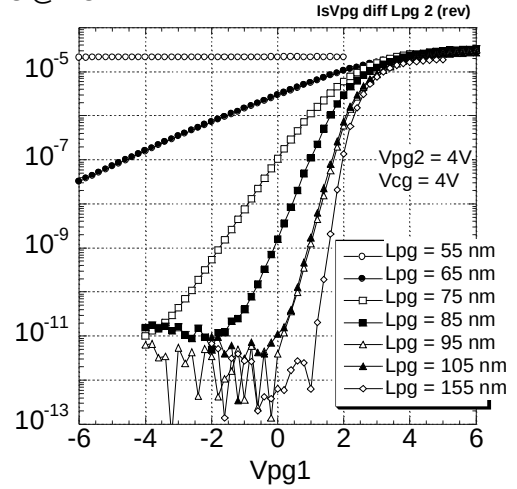


Fig. 5 : PG transistor transconductance characteristics under read-out conditions as a function of the PG length.

It is important to note that this effect is not detrimental for the device operation since the current difference between selected and unselected cells is managed by the CG transistor, which does not show any SCE (Fig. 6). The PG scaling limitation is reached when the PG voltage has no more influence on the device channel current. Referring to fig. 5, the case $L_{PG} = 55$ nm corresponds to this limit, where the junctions are probably entirely diffused under the PG's. For $L_{PG} = 65$ nm, however, the PG voltage can still modulate the current, and it is possible to sense a charge stored under the PG. Since the subthreshold characteristic is very bad, however, a practical limitation could be an insufficient read-out current difference between programmed and erased states. Comparing these two cases ($L_{PG} = 55$ nm and $L_{PG} = 65$ nm) one can estimate experimentally the under-diffusion to be between 55 and 65 nm, in good agreement with the simulation. Thus, the effective PG channel length for the shortest operational device is below 10 nm.

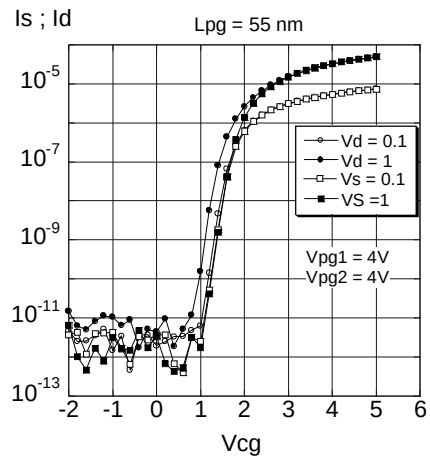


Fig. 6 : CG transistor transconductance characteristic.

4. Scaling effects on erase operation

For the devices operation, we consider here only the V_t of PG1, knowing that the situation is identical for PG2 because the cell is symmetrical [3]. The V_t is measured from the source side (reverse read), at $V_s=1V$, with the following other gates biases : $V_{pg2} = 4V$ (for screening the bit stored under PG2, and $V_{cg} = 4V$ (giving a read-out current of about $25\mu A$ - see fig. 6 -). Since the intrinsic V_t (V_{t0}) is dependent on the PG length because of SCE (see fig. 5), the V_t shift $V_t - V_{t0}$ is considered for comparison. Prior any erase operation, the PG1 transistor is programmed to $V_{t0} + 1V$. Erase is performed by Band to Band Tunneling induced hot holes by applying $+5V$ on the drain and $-5V$ on PG1. It has been shown previously that the erase operation is faster with a shorter PG length down to 140 nm , because the holes are injected closer to the stored electrons [7]. Fig. 7 shows the erase characteristics of further scaled PG length devices.

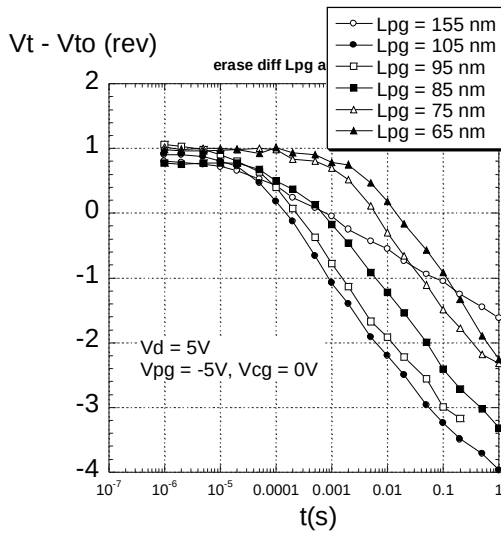


Fig. 7 : Erase characteristics as a function of the PG length.

A maximum erase performance is found for $L_{pg} \sim 95 - 105\text{ nm}$ (or $L_{pg\text{eff}} \sim 40 - 50\text{ nm}$). For shorter PG lengths, erase is slower, which is not expected from the above hypothesis [7]. It has been checked that no programming occurred under the Control Gate ($V_t(\text{CG})$ unchanged), when programming the shorter PG1 devices. This slower erase is rather explained by the influence of the CG (biased at $0V$) on the drain deep depletion during erase, in those cases where the PG is short enough. Fig. 8a and 8b show simulations of the band to band tunneling generation rate for the CG biased to 0 or $-5V$, in case of short and long PG's. For a shorter PG, the generation rate is decreased when the CG is grounded, and this can compensate for the advantage of having holes injected closer to the electrons.

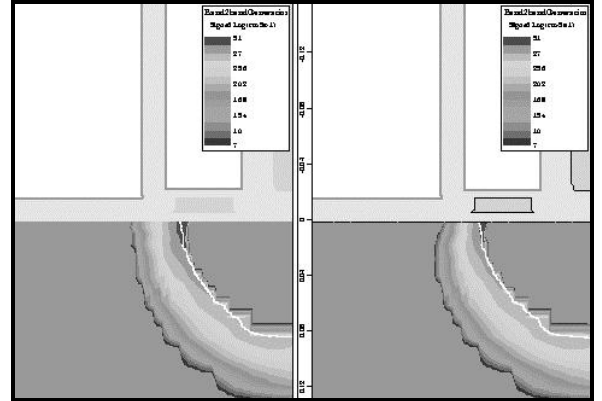


Fig. 8a : Band to band tunneling generation rate of a short PG device during erase with $V_d=5V$, $V_{pg1}=-5V$, and $V_{cg}=-5V$ (left), $V_{cg}=0V$ (right). The generation rate is higher when adding the negative CG bias.

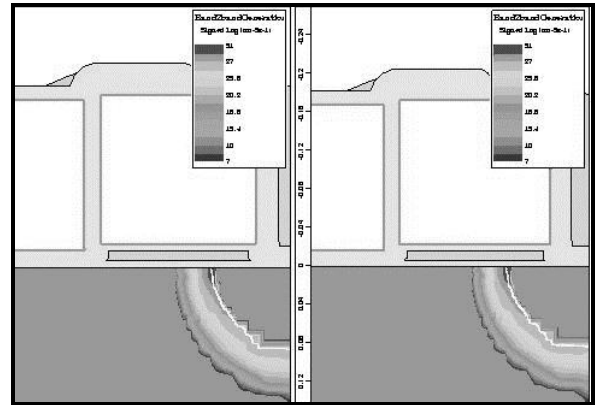


Fig. 8b : Band to band tunneling generation rate of a long PG device during erase with $V_d=5V$, $V_{pg1}=-5V$, and $V_{cg}=-5V$ (left), $V_{cg}=0V$ (right). The generation rate is similar when adding the negative CG bias.

The effect of adding a negative CG bias during erase is shown in fig. 9. As expected, the improvement of the erase characteristic is all the more important when the PG is shorter. Under this erase scheme, the benefit of injecting the holes closer to the electrons is maintained without reducing the hole generation rate (filled curves in fig. 9), and very fast erase operation can be obtained on very short PG lengths.

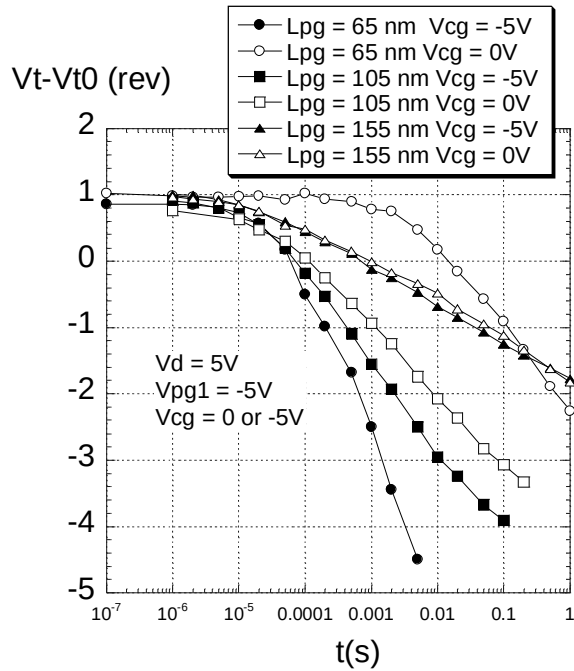


Fig. 9 : Erase characteristics as a function of the PG length with CG biased to 0 or -5V, respectively.

6. Conclusion

Aggressive scaling of a nitride storage split-gate device has been investigated. It has been shown that the storage sub-channels can be scaled to less than 10nm length, while a good device transistor behaviour is preserved by the CG sub-channel. Further, it has been shown that by using an optimized erase scheme, the erase operation improves as the storage sub-channel length is reduced, leading to very fast erase operation. Future investigation will concern the programming characteristics and reliability of these highly scaled devices.

References

- [1] Eitan, B. et al, IEEE El. Dev. Let. Vol. 21, No 11 (2000).
- [2] Chang, Y-W., et al., IEEE El. Dev. Let. Vol. 25, No 2 (2004)
- [3] Breuil, L. et al. IEEE NVSM Workshop (2003) , p. 46
- [4] Hsieh, C-S. et al., IEEE Trans. Elec. Dev., Vol. 51, No. 11 (2004)
- [5] Lee, Y.K., et al. IEEE El. Dev. Let. Vol. 25, No 5 (2004)
- [6] Ogura, T. et al., VLSI Symp. Tech. Dig. (2003)
- [7] Breuil, L. et al. IEEE NVSM Workshop (2004) , p. 83

Low voltage and low power embedded 2T-SONOS flash memories

M. Slotboom*, D. Dormans*, M. van Duuren, R. van Schaijk, N. Akil and R. Beurze*

Philips Research Leuven, Kapeldreef 75, B-3001 Leuven, Belgium, email: michiel.slotboom@philips.com

*Philips Semiconductors, Gerstweg 2, 6534 AE Nijmegen, The Netherlands

Abstract

16Mb 2-transistor (2T) SONOS flash memories have been processed in a 0.18 μ m CMOS technology. The 2T NOR architecture allows fast random access and low power program and erase by tunneling. Also, negative erased threshold voltages enable low voltage read without boosting. This paper shows the operation and reliability aspects of 2T N-type SONOS memories: endurance, retention and read disturb. A further reduction in read voltage is demonstrated by using P-type SONOS.

1. Introduction

Compared to floating gate flash, the SONOS flash cell has the potential of improved scalability, better CMOS compatibility and reduction of the high voltage for program/erase (P/E) and therefore reduction of overall System-on-Chip area [1]. The embedded 2T SONOS technology is well suited for low voltage / low power applications demanding random accessibility like for direct code execution for smart cards.

Main reliability concerns for SONOS memories, *viz.* retention, erase saturation and read disturb, are investigated in this paper. It is shown that P-type SONOS has a symmetrical threshold voltage (V_T) window around 0V and a reduced erase saturation.

2. Cell structure and device fabrication

The NOR array architecture of the 2T cell is given in fig. 1. The 2T SONOS cell consists of an access MOST and low V_T SONOS transistor in series. The access MOST not only allows low-voltage read because the SONOS transistor can be erased to a negative V_T , but also allows programming by uniform tunneling to and from the nitride layer [1]. Poly wordlines consist of access and control gates (AG, CG), whereas shared drain contacts connect the cells to metal bitlines (BL) and the sources are connected by tungsten local interconnect lines (LIL) [2].

Test arrays up to 2.7Mb as well as 16Mb circuits including error correction and charge pumps, have been fabricated in a 0.18 μ m CMOS process [2]. The cell size of 0.78 μ m² is proven scalable to 0.21 μ m² in the 90nm node.

After STI formation, an ONO layer is deposited (19 or 22 $\text{\AA}$ thermal oxide, 53 $\text{\AA}$ LPCVD nitride, 80 $\text{\AA}$ HTO), see fig. 2. An ONO-protect mask is used to remove the ONO from the AG area. This mask is also used to adjust the

V_T of the AG transistor and SONOS device independently. Next, the AG oxide is grown and the poly gate is deposited and patterned. The ONO is etched in a self-aligned way with respect to the CG gate (see fig. 2). An MDD implant is followed by high voltage (HV) and logic device formation ending with silicidation, tungsten contacts and LIL and a 6-layer aluminum back-end metallization.

3. Results and Discussion

A V_T window of 2.7V is obtained by P/E for 10ms at ± 11 V on the CG as is shown in fig. 3. Lowering the V_T of the SONOS cell by using the ONO-protect mask reduces the erased V_T by ~ 0.6 V and, by that, the required gate voltage during read. Also shown in fig. 3, erasing results in a negative V_T , but is limited due to erase saturation: at high erase voltages a dynamic equilibrium is reached between electrons tunneling through the top oxide and holes through the bottom oxide [3].

Fig. 4 shows that long cumulative erase times at a high erase voltage cause significant degradation of the endurance of the memory due to erase saturation. A one-order longer erase saturation time results in a one-order worse endurance, without yielding a larger V_T window.

Fig. 5 shows the endurance of a 2.7Mb array up to 10^6 cycles avoiding erase saturation conditions. The standard deviation of the threshold voltage for both the programmed and erased state is $\sigma = 75$ mV.

The data retention of uncycled cells is shown in fig. 6. No tail bits appear, which supports the observation of the deterministic and predictable degradation behavior of SONOS memories [4].

P/E cycling shows only a minor effect on the retention, see fig. 7. In case of a 22 $\text{\AA}$ thick bottom oxide, the initial V_T window is moved upward. In that case only the programmed V_T shifts significantly in time allowing a more simple design with constant read voltage and access time. However, for thicker bottom oxides, higher P/E voltages are needed and the sensitivity to erase saturation increases. Extrapolation to 10 years lifetime results in a typical V_T window of 0.6V and 0.9V for a bottom oxide thickness of 19 $\text{\AA}$ resp. 22 $\text{\AA}$ and allows reading at $V_{CG} = 1$ V. Not shown here, the retention of cycled cells at temperatures up to 150°C shows only a small extra charge loss compared to 20°C [4].

Fig. 8 shows that an initial larger V_T window of SONOS remains significantly larger during retention, which is not the case for floating gate memories. This might be explained by the SONOS detrapping behavior of fast shallow traps and slow deep traps. Only the magnitude

of the V_T window is important for retention, not the program conditions that were used to reach this window. The feasibility of 2T NOR SONOS has been demonstrated with fully functional 16Mb arrays with dedicated periphery (*e.g.*, sufficient read margins in checkerboard mode, an access time of 55ns and a total current dissipation of 2.2mA during parallel write of all cells). The 16Mb array shows good yield over the wafer, see fig. 9. All logic and HV device parameters remain within specifications proving the CMOS compatibility of the embedded SONOS memory module.

Read disturb measurements at a gate stress of 2.5V on a 10^5 times cycled array are shown in fig. 10. The significant V_T shift during 10^5 seconds proves the necessity for a low read voltage. P-type SONOS provides a symmetric V_T window around 0V, see fig. 11. Besides, it largely reduces erase saturation, because the P-gate injects fewer electrons into the top oxide than an N-gate [5].

The V_T window and endurance of a P-SONOS array are good up to more than 10^6 cycles while reading at V_{CG} voltage of 0V is feasible, see fig. 12.

4. Conclusions

Fully functional, well yielding and reliable 16Mb 2T SONOS NOR memories embedded in a state of the art 0.18 μ m process are demonstrated. Operation voltages are lower than 11V with P/E times of around 10ms. No tail bits are observed in retention, gate stress and after cycling. A P-SONOS memory largely reduces erase saturation and exploits low voltage read, potentially improving on reliability performance. The 2T SONOS NOR technology provides a scalable flash cell in a highly manufacturable CMOS compatible process module.

References

- [1] "Nonvolatile Semiconductor Memory technology", W.D. Brown and J.E. Brewer (ed.), IEEE press, New York, p308-357 (1998)
- [2] G. Dormans *et.al.*, proc. SSDM, p540 (2001)
- [3] R. van Schaijk *et.al.*, Microel. Eng. 72, p395 (2004)
- [4] R. van Schaijk *et.al.*, proc. ESSDERC, p270 (2004)
- [5] H. Bachhofer *et.al.*, J. Appl. Phys. 89, p2791 (2001)

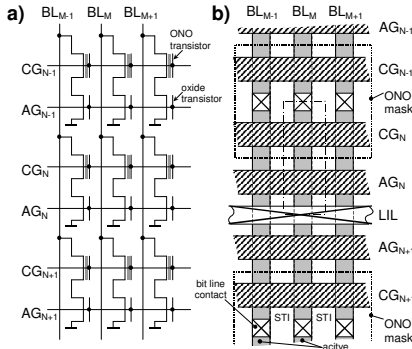


Fig. 1. Schematic electrical (a) and topview array layout (b) of a 2T-SONOS NOR memory.

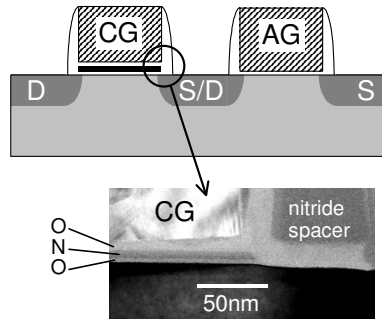


Fig. 2. Schematic cross-section of 2T SONOS memory cell with zoom-in on TEM graph of the ONO stack at the gate edge.

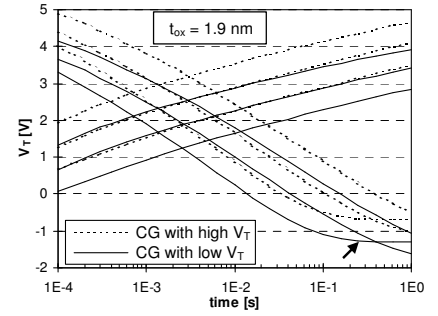


Fig. 3. P/E curves of 2.7Mb array for write voltages of ± 9 , 10 and 11V (average V_T). The erase saturation is marked by the arrow.

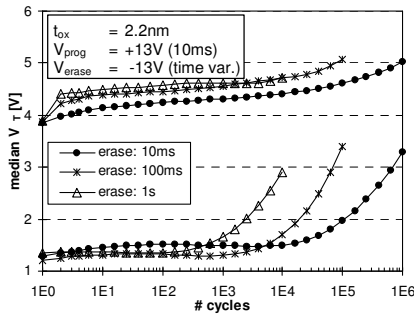


Fig. 4. Endurance of 0.5Mb arrays for the same erase voltage, but various erase times (all in the saturation regime).

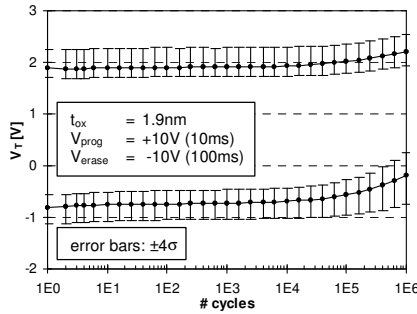


Fig. 5. Endurance of a 2.7Mb SONOS array.

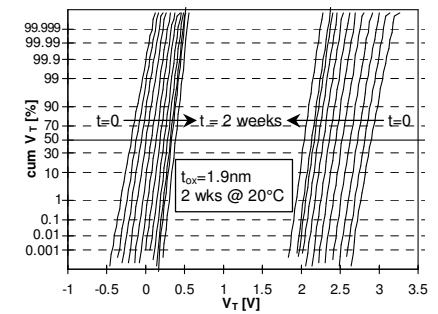


Fig. 6. Cumulative V_T distributions of two 1.3Mb arrays during a 2-weeks retention test.

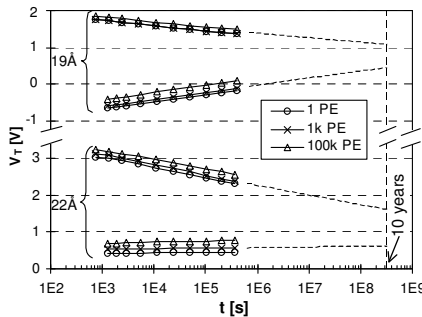


Fig. 7. Data retention of 0.5Mb arrays at $T=20^\circ\text{C}$ for 19Å and 22Å bottom oxide.

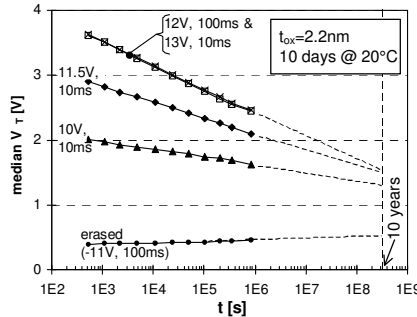


Fig. 8. Retention of several 0.5Mb arrays for different initial programming conditions.

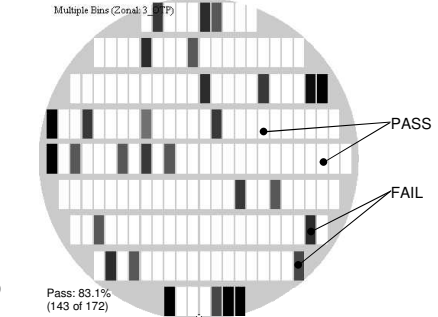


Fig. 9. Wafer map of overall yield of 16Mb SONOS array with fully featured periphery.

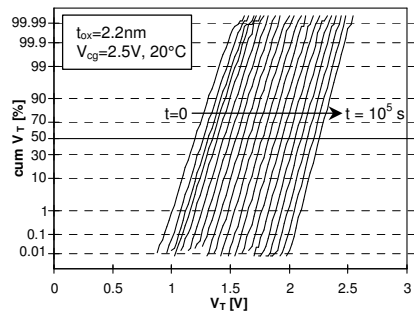


Fig. 10. Cumulative V_T distribution during gate stress on a 26kb SONOS array.

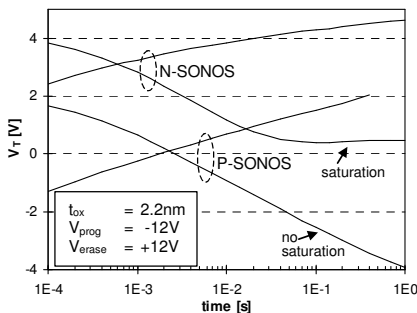


Fig. 11. P/E curves of N- and P-SONOS (average V_T in 256 bit arrays).

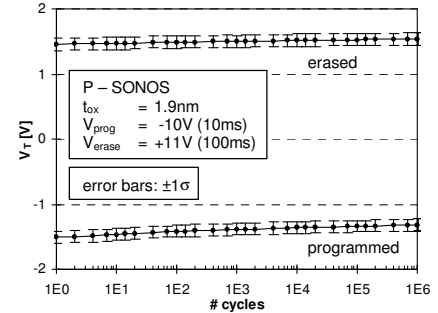


Fig. 12. Endurance of a 256 bit P-SONOS array.

Retention Reliability Improvement Using Nitride with Varying Trap Density in SONOS Type Non-volatile Memory

Ju Hyung Kim, Jeong Hee Han, Yo Sep Min, Moon Kyung Kim, Yeon Seok Jeong, Sanghun Jeon, Jae Woong Hyun, Jung Hoon Lee, Hee Soon Chae, Soo Doo Chae and Chungwoo Kim

Material and Devices Research Center, Samsung Advanced Institute of Technology, Mt.14-1, Nongseo-Ri, Giheung-Eup, Yongin-Si, Gyeonggi-Do, Korea 449-712, E-mail:juh.kim@samsung.com

Abstract

A new SONOS type structure employing bi-layered nitride with varying trap density and high-k blocking oxide is proposed for NAND Flash memory applications. The bi-layered silicon nitride with varying trap density is composed of a $\text{SiN}_x/\text{Si}_3\text{N}_4$ stack and Al_2O_3 was used for the high-k blocking oxide. Compared with conventional SONOS structures, the proposed structure shows remarkable improvement of charge retention reliability while maintaining good programming/erase performance.

1. Introduction

Recently, there has been an increasing need of high density memory for mass storage applications such as memory cards, digital cameras, USB drives, etc. NAND Flash memory technology has been applied to these data storage application fields. With the continual scaling down of devices, however, conventional floating-gate based technology has limitations such as cell-to-cell coupling interference and disturbance in NAND Flash memory [1, 2]. Charge trapping memories, such as SONOS, are a promising solution to overcome the aforementioned floating-gate limitations. However, SONOS memory also has some problems for NAND Flash memory applications regarding high performance and reliability. One problem is that there is a strong trade-off between the programming/erase speed and charge retention in the conventional SONOS memory structure [3, 4].

In this paper, we propose a new SONOS type structure, where conventional silicon nitride and blocking oxide (SiO_2) are replaced with a bi-layered silicon nitride with varying trap density and a high-k blocking oxide (Al_2O_3), respectively, in order to mitigate the trade-off relationship between them. We fabricate both conventional and proposed devices and compare the electrical characteristics of these devices in terms of programming/erase speed and retention property.

2. Experimental

We fabricated MONOS and MHNOS type capacitor samples which are described in detail in Table 1. Both structures have different types of trapping nitride. One employs the uni-layered silicon-rich nitride (SiN_x). The other employs a bi-layered silicon nitride which consists of a $\text{SiN}_x/\text{Si}_3\text{N}_4$ stack. The stoichiometry of the Si_3N_4 layer is almost stoichiometric, as shown Table 2.

All samples were fabricated on p-type (100) wafers. The tunnel oxide was thermally grown at 850°C in an oxygen atmosphere and annealed in N_2O at the same

temperature. Next, the uni-layered silicon nitride (SiN_x) was formed using LPCVD at 730°C with a dichlorosilane (DCS)/ NH_3 gas ratio of 2. In the case of bi-layered silicon nitride ($\text{SiN}_x/\text{Si}_3\text{N}_4$), the first nitride (Si_3N_4) was deposited using LPCVD at 730°C with a DCS/ NH_3 gas ratio of 0.13 and the second nitride (SiN_x) was deposited using LPCVD at 730°C with a DCS/ NH_3 gas ratio of 2.

For the bi-layered nitride, the first silicon nitride (Si_3N_4) layer on the tunnel oxide has a lower trap density while the second silicon nitride (SiN_x) has a relatively higher trap density, due to the higher DCS/ NH_3 gas ratio [5]. Namely, the trap density varies within the bi-layered nitride structure.

Next, a blocking oxide (SiO_2) was also deposited using LPCVD at 450°C . In the case of high-k blocking oxide, Al_2O_3 was formed by atomic layer deposition (ALD) at 400°C using tri-methyl-aluminum (TMA) and ozone gas (O_3), followed by an O_2 anneal at 700°C for 1 min. Finally, after each dielectric stack was formed, Al or Pt electrodes were formed using conventional photolithography and evaporation processes.

MONOS type	MHNOS type
Tunnel oxide (SiO_2) 25 Å	Tunnel oxide (SiO_2) 30 Å
Trapping nitride 50 Å - SiN_x (50Å) - SiN_x (30Å) / Si_3N_4 (20Å)	Trapping nitride 60 Å - SiN_x (60Å) - SiN_x (40Å) / Si_3N_4 (20Å)
Blocking oxide (SiO_2) 70 Å	High-k Blocking oxide (Al_2O_3) 120 Å
Metal electrode (Al)	Metal electrode (Pt)

Table 1: Sample description of MONOS and MHNOS type devices.

	DCS/ NH_3 ratio	Si (at %)	N (at %)	Stoichiometry
SiN_x	2/1	55.3	44.7	Si (3.9) N (3.1)
Si_3N_4	1/7.5	41.9	58.1	Si (2.95) N (4.05)

Table 2: Stoichiometry of Silicon nitride as DCS/ NH_3 gas ratio results from TEM EDS.

3. Result and Discussion

Generally, in SONOS type memories, Si-rich nitride (SiN_x) with high trap densities results in improvement of programming speed with worse charge retention. Conversely, stoichiometric nitride (Si_3N_4) with low trap density leads to improvement in the charge retention, but has poor programming speed [5]. Trap density in the silicon nitride, as well as the thickness of tunnel oxide, is the main contributor to the trade-off

relationship between the programming/erase speed and the charge retention in SONOS devices. Therefore, we propose a SONOS type structure where the bi-layered nitride has varying trap density. The purpose of utilizing the bi-layered nitride is to improve the charge retention reliability while maintaining the programming/erase performance.

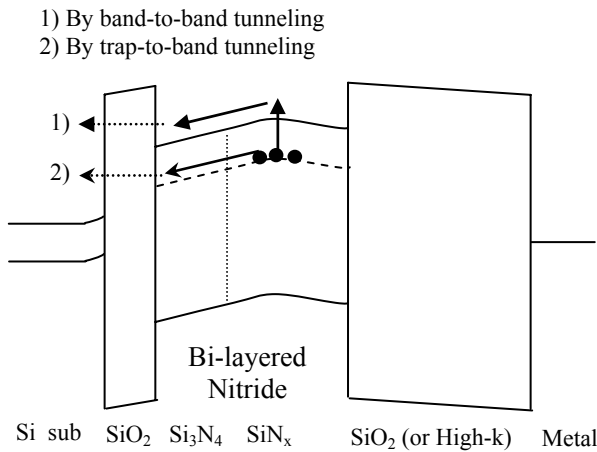


Fig. 1: Schematic band diagram illustrating the charge loss path in programmed state of MONOS or MHNOS structure with bi-layered nitride with varying trap density. Si₃N₄ has a low trap density and SiN_x has a high trap density

As shown in Fig.1, the charge loss paths in the programmed state energy band diagram of MONOS (or MHNOS) structure with bi-layered nitride are represented. A slight internal field exists in the structure due to the trapped charge distribution. With the two paths, the trapped charge can escape through the tunnel oxide to the substrate, whereas the blocking oxide is too physically thick for the trapped charge to leak through it. One path is leakage by band-to-band tunneling after Pool-Frenkel emission of the trapped charge. The other path is leakage by trap-to-band tunneling after transport of the trapped charge toward the trapping nitride-tunnel oxide interface. Leakage by trap-to-trap tunneling via traps in the tunnel oxide could also be included in the latter.

In the SONOS type structure with bi-layered nitride, leakage by trap-to-band tunneling after transport of the trapped charge toward the trapping nitride-tunnel oxide interface can be reduced, because transport of the trapped charge using a nitride trap as a stepping stone is reduced in the stoichiometric nitride (Si₃N₄) with low trap density. Lundkvist et al. also reported that the trapped charge tunnels from the nitride back to the silicon substrate due to the internal field in the retention mode and the charge loss rate is proportional to the trap density of the nitride [6].

The program/erase speed and programmed state retention characteristics of MONOS structure with uni-layered and bi-layered nitride are shown in Fig. 2 and Fig. 3, respectively. The charge retention characteristics are represented by the flat-band voltage shift due to the charge loss at the programmed state.

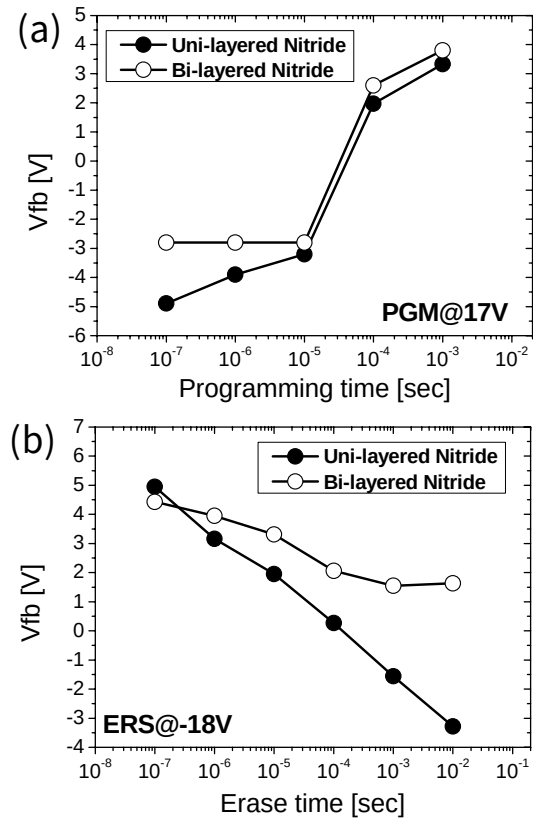


Fig. 2: (a) Programming and (b) erase characteristics of the MONOS device with uni-layered and bi-layered silicon nitride structure. Program and erase voltage are 17V and -18V, respectively.

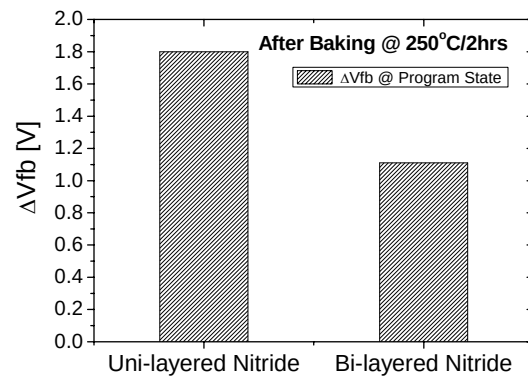


Fig. 3: Retention characteristics in the programmed state of the MONOS device with uni-layered and bi-layered silicon nitride structure. Flat-band voltage shift is measured after baking at 250 °C for 2 hours in the oven.

As expected, the retention characteristics of MONOS devices with bi-layered nitride are better than those with uni-layered nitride. Similar behavior is shown in the programming characteristics. However, in the case of the erase characteristics, MONOS devices with bi-layered nitride show worse performance. It could be thought that the stoichiometric silicon nitride with low trap density plays a role in reducing the amount of trapped charge tunneling back to the silicon substrate. The back-tunneling current from the gate also plays a

role for erase times over 100 μ s. The trade-off relationship may still remain although it can be slightly improved with further thickness and process optimization of the structure.

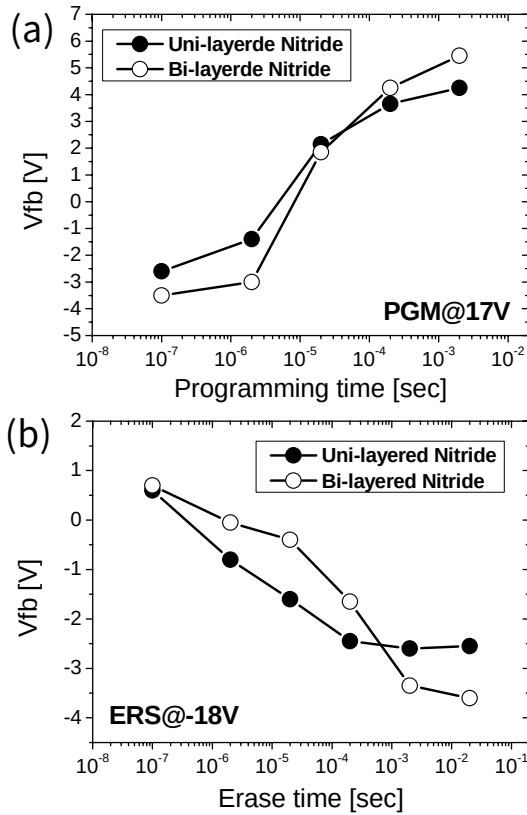


Fig. 4: (a) Programming and (b) erase characteristics of the MHNOS device with uni-layered and bi-layered silicon nitride structure. Program and erase voltage are 17V and -18V, respectively.

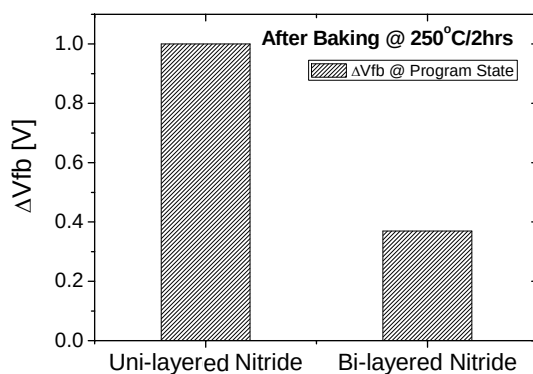


Fig. 5: Retention characteristics in the programmed state of the MHNOS device with uni-layered and bi-layered silicon nitride structure. Flat-band voltage shift is measured after baking at $250^{\circ}C$ for 2 hours in the oven.

For this reason, high-k dielectric (Al_2O_3) was used for the blocking oxide in the MONOS structure with bi-layered nitride. By utilizing the high-k blocking oxide, the physical thickness can be increased without

increasing EOT. Electric field in the tunnel oxide increases and the electric field in the blocking oxide decreases when the blocking oxide has a higher dielectric constant [7]. As a result, the back-tunneling current from the gate at the erase mode can be suppressed. In addition, the back-tunneling current from the gate can be further suppressed by utilizing high work-function metal gate such as Pt ($\approx 5.3eV$).

The programming/erase speed and the retention characteristics of MHNOS structure with uni-layered and bi-layered nitride are shown in Fig. 4 and Fig. 5, respectively. In the MHNOS device, the programming/erase characteristics are not deteriorated by utilizing bi-layered nitride in the structure. The programming/erase characteristics of the MHNOS device with bi-layered nitride is superior to those of the MHNOS device with uni-layered nitride for programming times longer than 30 μ s and erase times longer than 500 μ s. It is thought that the employment of the high-k blocking oxide and high work-function metal gate leads to the enhanced electric field in the tunnel oxide and the suppression of back-tunneling current from the gate. Thus, the negative effect of bi-layered nitride for the erase performance is reduced. For the retention characteristics, the effect of utilizing bi-layered nitride in the structure is remarkable. That is to say, the flat-band voltage shift by the charge loss is reduced to about 0.37V in the MHNOS device with bi-layered nitride.

4. Conclusion

We propose a new SONOS type structure with a bi-layered nitride layer with varying trap density and high-k blocking oxide. The results demonstrate that the charge retention reliability is remarkably improved while maintaining good program/erase performance for the proposed structure, compared with the conventional structure.

Acknowledgement

This work was supported by the National Program for Tera-Level Nano-Devices through the Ministry of Science and Technology, Korea.

References

- [1] Y.S. Yim, IEDM, 819 (2003)
- [2] J.D. Lee, IEEE EDL, 264 (2002)
- [3] M.L. French and M.H. White, IEEE Comp Pack and Manu Tech part A, **17**, 390 (1994).
- [4] J. Bu and M.H. White, Solid-State Electron. **45**, 113 (2001)
- [5] Y. Yang and M.H. White, Solid-State Electron. **43**, 2025 (1999)
- [6] L. Lundkvist and C. Svensson, Solid-State Electron. **16**, 811 (1973)
- [7] C.H. Lee and K. Kim, IEDM, 613 (2003)
- [8] M.K. Kim and S. Tiwari, IEEE Nanotech. **3**, 417 (2004)

Flash memory with high-k tunnel dielectrics: comparative retention study.

Pieter Blomme^{a,b}, Amal Akheyar^c, Jan Van Houdt^a, Kristin De Meyer^{a,b}

a) Imec, Kapeldreef 75, 3001 Leuven, Belgium; blommep@imec.be

b) KULeuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium

c) Infineon Technologies affiliated to Imec

Abstract

The tunnel and interpoly dielectric thickness of floating gate Flash memories can hardly be scaled any more. Therefore, electron tunneling through these dielectrics for programming or erasing the memory cells requires a very high voltage. A possibility for drastically reducing the necessary voltage for electron tunneling without requiring a reduction of the physical thickness of the dielectrics is the use of engineered multilayer oxide/high-k stacks. In this work, we study the data retention that can be achieved using these stacks in memory cells. We show that HfO₂ with TiN metal gate provides 10 years of data retention up to 105°C but is very leaky in combination with a polysilicon gate due to interaction between the poly and the HfO₂. Al-based materials such as Al₂O₃ or Al/Hf mixed oxides do not suffer from this interaction and can be used in order to obtain memory cells with 10 years of data retention up to 90°C even when a polysilicon floating and control gate are used.

Introduction

Floating gate Flash nonvolatile memory technology imposes very strict requirements on leakage through the tunnel gate dielectric and interpoly dielectric: a charge of a few thousands of electrons is stored in the floating gate and this charge has to remain there for at least 10 years, which leads to a maximum allowable leakage current at or below the level of 1 electron per day. Due to this strict requirement, both tunnel oxide thickness and interpoly thickness can only be reduced very slightly [1], leading to the excessively high voltages to program and erase the memory cells.

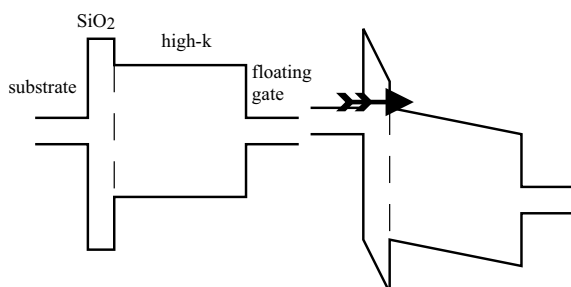


Fig. 1. Band diagram of an SiO₂/high-k stack when no bias is applied (left-hand side) showing a large tunnel barrier and when a moderate bias is applied (right-hand side) showing a strongly reduced barrier width. The high dielectric constant of the second layer limits the electric field in this layer so that a high field in the SiO₂ layer can be obtained for a moderate applied bias.

Reduction of the voltage that is necessary for electron tunneling through the dielectric can be achieved by the use of engineered dielectric stacks such as the crested barrier [2] or the Variot stack [3], the latter being easier to integrate in a practical memory cell.

A Variot stack is a stack consisting of two dielectrics with different dielectric constant, typically a SiO₂/high-k dielectric stack [4] using for example Al₂O₃ or HfO₂ as high-k material (fig. 1). Strong voltage reduction and good endurance have been demonstrated by using these stacks [5][6] for programming or erasing memory cells by electron tunneling. In this work, we show that not only an important gain in programming voltage but also 10 years of data retention can be achieved using an SiO₂/high-k stack.

Experimental

We have combined results from two types of memory devices for this study, being large single poly (or metal gate) memory cells [5] and dual poly split gate 0.18μm HIMOSTM [7] memory cells; We have also used a pseudocell setup [8] for short time data retention measurements on capacitors. We have not seen strong impact of the cell type on the data retention and the validity of the pseudocell approach is confirmed by comparing pseudocell data retention with the retention measured on HIMOSTM memory cells (fig. 2). The studied dielectric stacks are two-layer stacks consisting of 2nm-3nm SiO₂ and 8nm-10nm of high-k material being either AL-CVD HfO₂, Al₂O₃ or Al/Hf mixed oxide or MOCVD

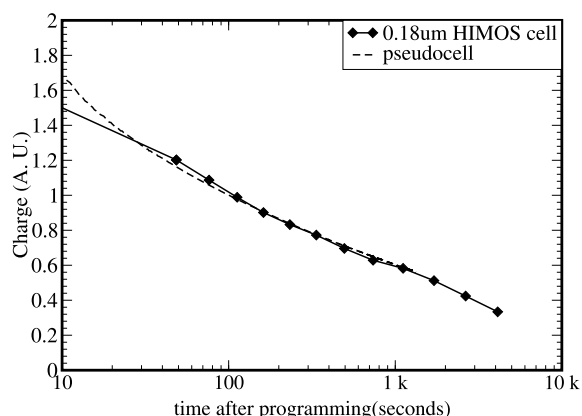


Fig. 2. Comparison of a pseudocell retention measurement with a retention measurement performed on a 180nm HIMOSTM memory cell with an identical dielectric stack. In order to see a clear charge loss within the maximum allowable pseudocell measurement, both measurements are performed at 125°C.

HfSiON.

Results and discussion

When using a HfO₂ tunnel dielectric with TiN metal gate, 10 years of data retention can be achieved up to 105°C. As can be seen in fig. 3, The charge loss can be modeled by an exponential decay function:

$$V_{th}(t) = V_{th}(\infty) + (V_{th}(0) - V_{th}(\infty)) \cdot \exp\left(-\frac{t}{\tau}\right) \quad (1)$$

and the temperature dependence of the charge loss follows an $1/T$ law (fig. 4). However, when the TiN metal gate is replaced by a polysilicon gate, the charge leakage behavior of the HfO₂ layer is drastically changed; the decay of the threshold voltage is no longer exponential but follows a logarithmic law. Furthermore, the observed leakage is much stronger. As a result, the SiO₂/HfO₂ dielectric stack combined with a polysilicon gate can not be used in order to obtain memory cells with 10 years of data retention.

Two different issues could cause this difference in retention behavior: the memory cells with TiN metal floating gate shown in [5] are manufactured using a low thermal budget process so that the HfO₂ layer is not fully crystalline. Also, the interaction between the polysilicon and the bulk of the HfO₂ material can not be excluded. In order to separate these two effects, we have manufactured memory cells with a TiN gate and the thermal budget corresponding to the polysilicon process (labeled *TiN+anneal* in fig. 4). The data retention as measured on these cells is very close to the retention reported on a cell with low temperature budget. Therefore we can conclude that crystallization of the HfO₂ layer is not a problem with respect to retention, but the retention problems are caused by interaction with the top poly layer (or with the polysilicon deposition process). This hypothesis has been confirmed by depositing a nitride capping layer on the HfO₂ before polysilicon deposition, demonstrating improved retention behavior as can be seen in fig. 5

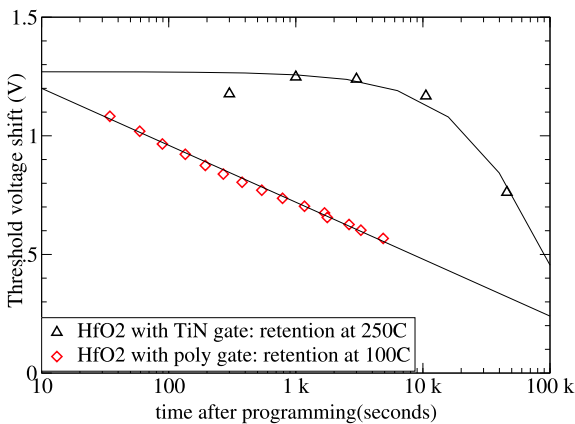


Fig. 3. Threshold voltage shift of memory cells with SiO₂/HfO₂ stacked tunnel dielectric. The use of a polysilicon gate leads to a drastic increase in the leakage current and a change in the shape of the V_{th} decay.

although the data retention is not as good as when using a metal gate.

The use of a SiO₂ / hafnium silicate (MOCVD Hf-SiON) tunnel dielectric with polysilicon gate leads to a slight improvement in data retention as compared to HfO₂ with polysilicon gate, but significant charge loss is already observed at 125°C. However, the retention properties of the HfSiON stack are strongly dependent of the preparation recipe. With further material optimization, memory cells with HfSiON could therefore possibly be manufactured as well with a good data retention.

TABLE I
OVERVIEW OF THE DIFFERENT STUDIED DIELECTRIC STACKS.

SiO ₂ layer	1.5nm	3nm	2nm	2nm
high-k layer	8nm HfO ₂	10nm HfO ₂	10nm Al ₂ O ₃	8nm Al/Hf
Gate electrode	TiN	N-type poly	N-type poly	N-type poly
equivalent oxide thickness	2.89nm	5.16nm	4.93nm	4.0nm
10 year data retention up to	105°C	38°C	91°C	91°C
activation energy	0.93eV	-	1.2eV	1.36eV

When using a SiO₂/Al₂O₃ dielectric stack with a polysilicon gate, the data retention is clearly better than for HfO₂ with polysilicon gate. The charge loss of memory cells with a SiO₂/Al₂O₃ tunnel dielectric stack follows an exponential decay as for the HfO₂ with TiN gate. 10 years of data retention is extrapolated down to an operating temperature of 91°C (fig. 6) assuming a $1/T$ temperature dependence of the leakage. As the combination of Al₂O₃ with a polysilicon gate does not lead to strong charge loss, Al₂O₃ is suitable for floating gate NVM applications as tunnel and interpoly dielectric material.

The data retention of a memory cell using a Hf/Al mixed oxide deposited by alternating HfO₂ and Al₂O₃ ALCVD deposition cycles is similar to that of cells with pure Al₂O₃ (fig. 6). The strong similarity between the data retention on Al₂O₃ and on the studied

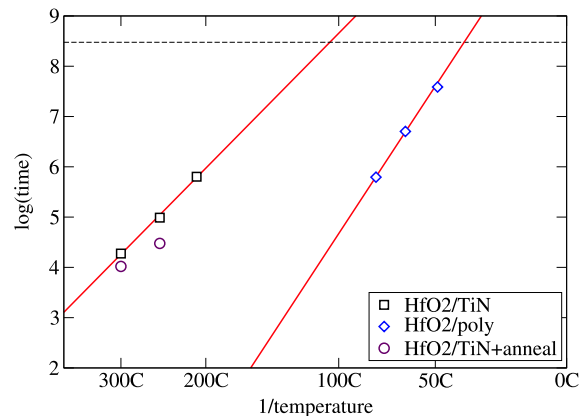


Fig. 4. Temperature dependence of the data retention with an SiO₂/HfO₂ tunnel dielectric. The samples labeled *TiN + anneal* have been subjected to the thermal budget of the poly process.

mixed oxide suggests that the same leakage mechanism might be dominant both in the pure Al_2O_3 as well as in the mixed oxide layer.

Conclusions

We have investigated the data retention of memory cells with different high-k materials. Without further optimizations, HfSiON is not useable as tunnel dielectric for nonvolatile memory applications. We have shown that good data retention is obtained in floating gate Flash cells with a $\text{SiO}_2/\text{HfO}_2$ tunnel dielectric stack when interaction with the top polysilicon can be avoided, indicating the need for either an adequate capping layer on top of the HfO_2 or the use of a metal gate. Aluminum containing oxides, being pure Al_2O_3 or Al-Hf mixed oxide are suitable for making memory cells providing 10 years of data retention and low voltage tunnel programming, also for the case of a polysilicon gate.

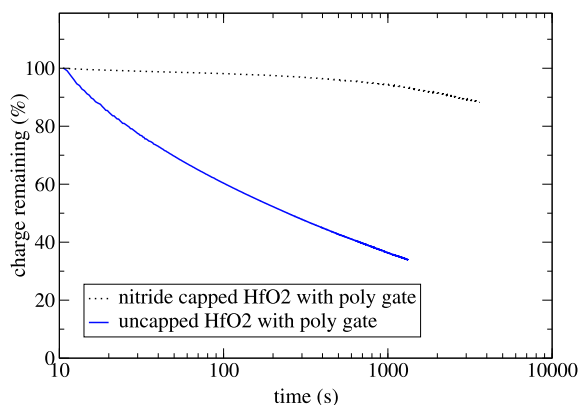


Fig. 5. Charge loss at 125C with a gate stack consisting of 2nm SiO_2 /10nm HfO_2 /N-type poly or 2nm SiO_2 /10nm HfO_2 /3nm SiN /N-type poly. The charge retention is strongly improved by the nitride capping but the leakage is still higher than when using a TiN metal gate.

Acknowledgments

P. Blomme thanks the IWT-Vlaanderen for financial support. This work was performed under IMEC's memory affiliation program with Intel and Infineon Technologies.

References

- [1] International technology roadmap for semiconductors, 2003 edition.
- [2] K. Likharev, "NOVORAM: CMOS Compatible Implementation Options", in *Proc. NVSMW*, 2001
- [3] P. Blomme, B. Govoreanu, M. Rosmeulen, US patent 6784484 and EP 1253646A1
- [4] B. Govoreanu, P. Blomme, M. Rosmeulen, J. Van Houdt, K. De Meyer, *IEEE El. Device Letters*, vol. 24 pp. 99-101, 2003
- [5] P. Blomme, B. Govoreanu, J. Van Houdt and K. De Meyer, in *Proc. IEEE NVSMW*, pp. 93-94, 2003
- [6] P. Blomme, L. Haspeslagh, J. De Vos, A. Akheyar, M. Lorenzini, J. Van Houdt and K. De Meyer, in *Proc. IEEE NVSMW*, pp. 63-64, 2004
- [7] J. Van Houdt, L. Haspeslagh, D. Wellekens, L. Deferm, G. Groeseneken, *IEEE Transactions on Electron Devices*, 40(12):2255-2263, 1993.
- [8] P. Blomme, B. Govoreanu, M. Rosmeulen, J. Van Houdt and K. De Meyer, "Multilayer tunnel barriers for nonvolatile memory applications", in *Proc. DRC* 2002

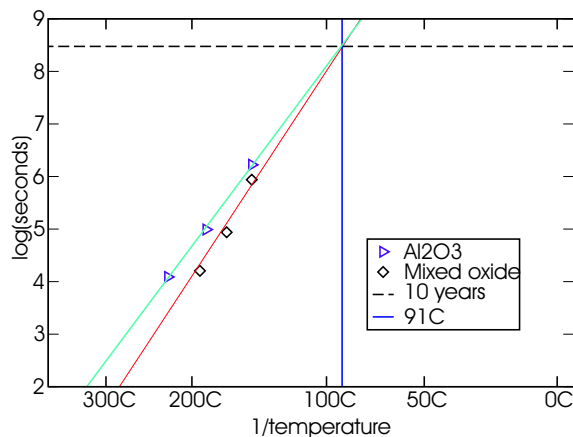


Fig. 6. Temperature dependence of the data retention with $\text{SiO}_2/\text{Al}_2\text{O}_3$ or $\text{SiO}_2/\text{Al-Hf}$ mixed oxide tunnel dielectric.

A 4 MegaBit, 128-bit data embedded flash memory allowing high-speed in place code execution.

Marylène Combe, Jean-Michel Daga

Embedded Non Volatile Memory group,
ATMEL ROUSSET, Zone Industrielle 13106 Rousset Cedex FRANCE.
marylene.combe@rfo.atmel.com , jean-michel.daga@rfo.atmel.com

Abstract

A 128-bit data operation, 4Mb flash has been designed in a 0.18um embedded flash process. This memory has two banks of equivalent size, allowing read while writing operation for improved application updates, and data management during code execution. Using the pre-fetch buffer concept, in place execution of the code at up to 100MHz can be obtained on a 32-bit MCU.

1-Introduction

Historically, 32-bit MCU based solutions have been driven by the need of increased performance in high-end applications, targeting the automotive market for example. More recently, 32-bit MCU solutions have been expanding in various applications, including secure products. The resulting competition to deliver more and more integrated and low cost family of products will certainly accelerate the move from 8,16-bit solutions to 32-bit ones [1]. According to analyst estimations [2], it is clear that the growing part of the market is on Flash based micro-controller units, which are expanding dramatically, while ROM based market is almost flat. The success of flash solutions is easy to understand. Flash offers many advantages due to the in system programmability allowing on the fly application updates, removing inventory issues associated with ROM, but also improving code protection [3]. In addition, cost of the flash has been decreasing due to its improved density, and embedded flash is now available almost concurrently with the latest CMOS logic process.

In order to take benefit of the MCU instruction set, embedded flash have been moving to 32-bit operation. Anyway, it is well known that flash accesses are slow compared to other types of memories. A lot of work has been done recently to improve flash random access, using optimized boosting techniques [4], improved

sensing techniques [5] [6], or dual supply concept [7]. Using these techniques, random access to flash at 30-40 MHz can be obtained. To further improve performance, high-speed (up to 66MHz) sequential burst access, with first read latency has been implemented on advanced stand-alone flash memories. Anyway, in order to take benefit of high-speed sequential access, code linearity is mandatory. Due to the lack of flexibility of sequential burst, this mode is typically used to fill a cache memory, and code is executed from the cache. In order not to limit system overall performance, maintain a good level of code flexibility, and decrease the need for expensive cache memories, new solutions must be implemented. The objective is to provide in place execution of the code, at system speed. That means about 100MHz flash access speed on a 0.18um technology, 200MHz at the 0.13um node. In this paper, we present a 128-bit 4Mb flash allowing random access to four sequential 32-bit instructions at 40MHz. Using a pre-fetch buffer concept, 32-bit instructions are accessed at up to 160MHz, after an initial read at 40MHz. In section 2 of the paper, the flash specification and the pre-fetch buffer concept are presented. More details about the flash architecture are provided in section 3. Simulation and first silicon results are given on section 4, while section 5 is dedicated to the conclusion.

2- Flash specification and pre-fetch buffer concept

The main objective was to provide in place execution of the code at up to 100MHz, on a 32-bit ARM based solution implemented in 0.18um. Using the previously reported design techniques, memory random read access was targeted to 25ns. In order to further improve performance, widening flash words and using a pre-fetch buffer concept was decided. The concept of pre-fetch buffer is illustrated on Fig. 1, based on 128-bit data words.

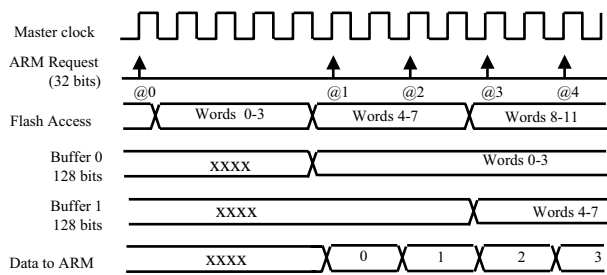


Fig. 1: Pre-fetch buffer concept

The embedded flash controller executes successive memory accesses. 25ns after the first access, data is valid on the memory data output bus, latched on an external 128-bit buffer and a second access is launched. Each clock cycle, 32 bits of data coming from buffer 1 are provided to the CPU. 4 clock cycles later (25ns), the second 128-bit of data that have been read and latched on buffer 2, are ready to be provided to the CPU at the clock frequency, and so on. That way, except for the first access, code can be executed each clock cycle, without a large SRAM buffer. Each 128-bit memory access can be executed randomly in the memory array. Compared to a sequential burst access, this approach provides better flexibility, and minimizes code constraints. As shown in Fig. 2, the optimal word width will be determined according to the targeted MIPS and expected random access speed. Considering 25ns random access, 128-bit wide words were necessary to allow up to 100 MIPS, on a ARM7 architecture. Actually this was even better than what was really feasible in 0.18um, due to logic speed limitations, and 128-bit appeared as a very safe option.

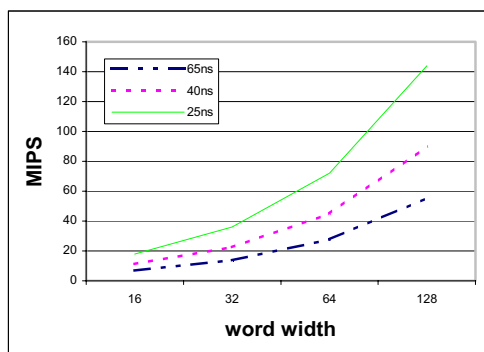


Fig. 2: MIPS vs. word width and random access speed (ARM7)

In addition to the word widening for speed improvement, the memory has been divided in two 2Mb planes, allowing read while writing operation. This feature offers improved flexibility for applications updates, or concurrent code execution and data management using the same device.

3- Memory architecture

The design uses a 2 transistors FLOTOX memory cell. This technology has the advantage to overcome over-erase and disturb problems, and provides short erase time compared to other technologies such as ETOX. The array architecture provides page erase granularity to minimize area compared to byte or word erase like memories. To reach the targeted access time, dual supply concept had been used [7].

The memory contains two planes: A and B, as described on Figure 3. Each of them has 512 pages of 128 words each. In most of the modes, the two banks can be seen like two different memories.

Data loading of the page content into an internal cache is performed prior to start programming the entire page in parallel. In the same plane, memory read is allowed during the data loading, forbidden once the programming has been started. When a programming is applied on one plane, reading is allowed on the other one.

The column register block separates the two planes. It is used as column decoding in write mode.

A whole page can be programmed simultaneously, so it is also used as page data register. Because of the two planes, architecture and layout have been modified to allow the circuitry to provide high voltage either to a plane or the other one. Special care has been taken on the parasitic devices and leakage currents, in order not to disturb the read access in a plane if a programming is performed on the other one. Input data are provided to column latches on a dedicated input bus. So, loading is not disturbing for a read operation on both planes. Figure 4 shows a column register.

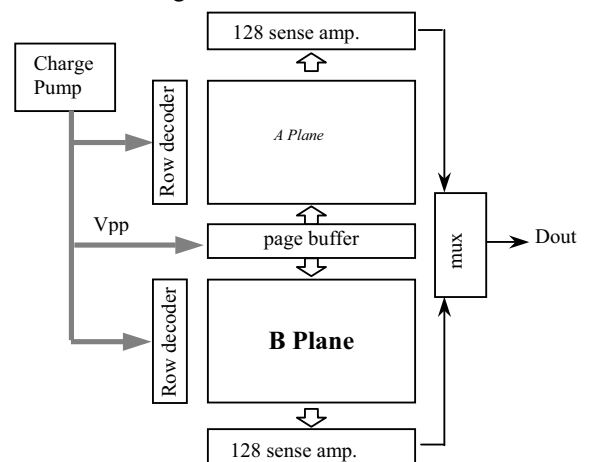


Fig. 3. Memory architecture

The column register is composed of three parts: the low voltage part including a path transistor (*Nld*) and a latch, a level shifter isolated from the low voltage latch by the

high voltage path transistor controlled by *PathHV* signal, and a high voltage multiplexer allowing to connect the bit line of plane A (B) to V_{pp} if programming occurs on this plane. To set-up a write operation, 1 data is latched, 0 otherwise. When 1 is latched, the output of the level shifter *outls* switches to V_{pp} , turning $NvppA(B)$ devices ON. If write is performed on plane A *WriteA* signal is driven to V_{pp} , *WriteB* to 0. NwA and NwB transistors goal is to avoid read disturbance when writing on plane B and A respectively.

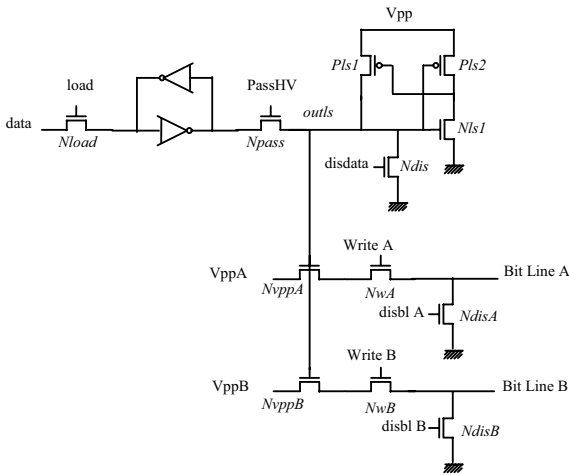


Fig. 4: Column register

The read path can be divided into three main parts: the row decoder (WLdecod), the Bit Line decoder (BLsel) and the sense amplifiers.

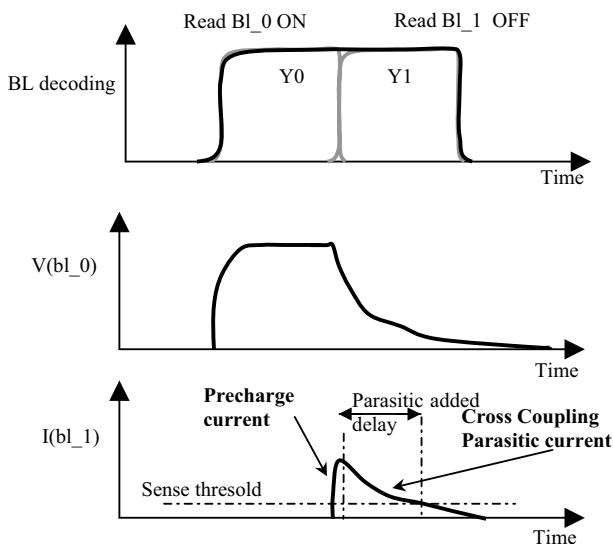


Fig. 5: Illustration of the BL parasitic current due to BL to BL coupling capacitance

The row decoder provides high voltage when programming and a medium voltage (about 3.3v) when reading. Providing 3.3v on the word line minimizes the bit selection transistor resistance. The use of Supply33 instead of an internal boosted voltage speeds up the word line selection and decreases power consumption [7].

The read column circuitry has been adapted to the memory capacitance and targeted performances. Critical dimensions becoming smaller and smaller, spacing between bit lines decreases drastically. So the bit line to bit line coupling capacitance contribution increases, becoming an important part of the total parasitic capacitance. In some cases, this coupling capacitance can have a dominant contribution to the read behaviour, slowing down the read operation. The uncontrolled discharge of the previously read column induces parasitic current on the surrounding bit line. In case a read is being performed on that bit line, it will be disturbed during the uncontrolled discharge, resulting in 20ns of additional delay in worst case. Figure 5 shows Bit line current when reading an OFF cell (no current), surrounded by ON cells (driving current). To avoid that problem, a new Bit line selection circuitry has been designed, in order to control the discharge of the first read Bit line, while reading the second Bit line.

Each side of the memory, 128 sense amplifiers are available to read the memory data value. Their outputs cross a long path (third of the memory periphery for the worst case), before being latched by the external circuitry. Because of the path length on the dout bus, buffer size, bus width and layout have been optimized to minimize cross talk effects and IR drops.

4- Simulations and silicon results

The 4 Mb memory has been implemented on a 0.18um embedded technology using the dual supply concept. Both 3.3V and 1.8V are used for consumption and speed optimization. The layout of the block is given in Fig. 6. The layout of different sub-blocks and the whole placement and routing have been optimised to minimize the read access time.

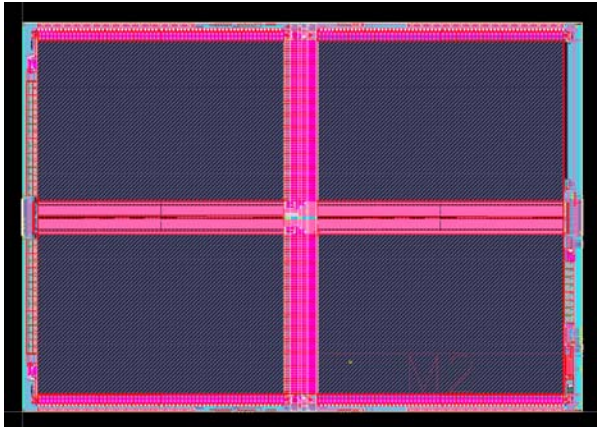


Fig. 6: Layout of the 4Mb memory

Read post layout simulations were performed considering various asynchronous accesses on different patterns such as checkerboard and so on. On such asynchronous memory, the access time is very dependant on the read pattern and the way the memory cell is accessed. So, the PVT analysis is performed using various read sequences and patterns.

Simulated results were in between 12ns and 20ns assuming different patterns and PVT corners. The worst case was obtained at high TEMP (125°C junction), low voltage supply (1.6v/3.0v) and worst process option. Figure 7 shows access time for 3 different supply voltages, across the full temperature range.

First silicon measurements have been performed and show values in line with SPICE simulations.

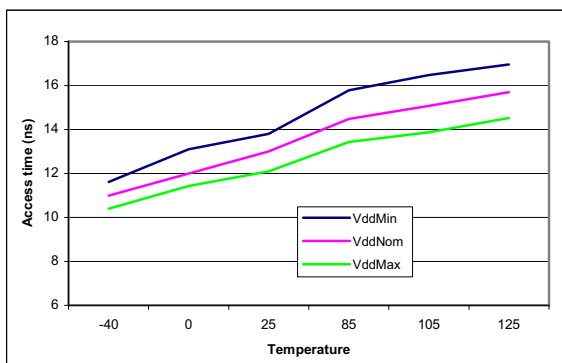


Fig. 7: Simulated Access time, at 3 different Vdd, across temperature

5- Conclusion

Flash MCU are expending very fast, due to the added value of In System Programmability. Flash memory performance optimisation is a hot topic, impacting the overall system resources allocation: improving flash read access speed to achieve in place execution of the code minimizes the need for expensive cache memory. Designing multi bank memories allows simultaneous code execution and data updates, removing the need for additional EEPROM memory. We showed in this paper that the only way to achieve up to 100MIPS code execution was to improve as much as possible random access time, and use wide words associated to a pre-fetch buffer system operation. This technique allows at system speed code execution, while maintaining a good level of flexibility. Based on the 128-bit 4Mb designed flash, more than 100MHz access can be obtained on a 0.18um technology. With such level of performance, flash is no more limiting the system overall performance. At the 0.13um node, near 250MHz is the targeted system speed. So the challenge is open for optimised flash at this node, based on the proposed solutions of this paper.

References

- [1] P. Bishop, "Atmel's ARM-based Flash Microcontrollers Bridge the Gap between 8- and 32-bit Applications", *Electronik*, September 2004
- [2] SEMICO report, June 2003
- [3] J.M.Daga, "Embedded Flash : an overview", SAME, October 2004
- [4] T. Tanzawa and S. Atsumi, J., "Optimization of Word Line Booster Circuits for low-Voltage flash Memories", *J. Solid-State Circuits*, vol. 34, no. 8, pp. 1091-1098, Aug. 1999
- [5] T.Tanzawa et al, "Design of a Sense Circuit for Low Voltage Flash memories" *J. Solid-State Circuits*, vol. 35, no. 10, pp. 1415-1421, Oct. 2000
- [6] N. Otsuka and M. Horowitz, "Circuit techniques for 1.5V Power Supply Flash Memory", *J. Solid-State Circuits*, vol. 32, no. 8, pp. 1217-1230, Aug. 1997
- [7] J.M. Daga et al, "Embedded EEPROM Speed Optimization Using System Power Supply Resources", *PATMOS*, September 2004
- [8] M. Combe et al, "EEPROM architecture and programming protocol " USSN 10.737,676., 2003

Lifetime Model and Procedure for Failure Rate Prediction of Moving Bits in case of Room Temperature Retention

T. Yao, D. Medjahed, P. Gassot, A. Lowe, P. Cacharelis

AMI-Semiconductor Belgium BVBA – Westerring 15, B-9700 Oudenaarde, Belgium
thierry_yao@amis.com

Abstract

We propose a model and a procedure for the prediction of the failure rate of moving bits in Floating-Gate based Non-Volatile Memories. The model and the procedure are based on a limited number of assumptions to enhance the confidence of the prediction. We show that the erratic moving bits are less critical and less numerous than the non-erratic ones. Therefore, the procedure requires no additional assumption such as the Equivalent Cell principle. Finally, we introduce the notion of lifetime predictability as part of the procedure.

1. Introduction

Moving Bits (MB), equally called Fast Bits or Tail Bits, are of the most critical reliability issues of Floating-Gate (FG) based Non-Volatile Memories (NVM). Several studies have been published to demonstrate their physical origin and/or to model them for failure rate prediction. The main conclusion is that MBs suffer from anomalous charge loss due to anomalous Stress-Induced Leakage Current (SILC) through the dielectric isolating the FG of the NVM cell. One of the most well known models is the percolation model published by R. Degraeve [1]. That model has several advantages: it is simple to use and it characterises the quality of the Tunnel Oxide (TunOX) by giving the Density of Oxide Traps (D_{OT}). Finally, it is able to predict the failure rate of the NVM device. Nevertheless, we found that model diverges from our data in long term MB experiments. Starting from that observation, we worked on building a more accurate model and an efficient procedure for MB failure rate prediction.

This paper summarizes the work done on modelling the variation in time of the threshold voltage of an FG based NVM and its application to the MB issue in case of Room Temperature (RT) retention. We present, for the first time, a general model of lifetime independent of the type of leakage current based on five assumptions. Then we explain the choice made for the model of leakage current, finalizing the analytical model. The lifetime model is then used in a procedure for the prediction of the failure rate in case of RT retention of the HIMOSTM Flash EEPROM [2]. We show, for the first time, that the procedure for failure rate prediction requires no additional assumption such as the Equivalent Cell (EC) principle. Moreover, we introduce in the procedure a new concept that is the lifetime predictability. This concept is a series of exercises required to verify the quality of the predictions.

2. The general lifetime model

The general lifetime model shown in this section is applicable to many type of FG NVM cell in several conditions of FG charge leakage. The model is built in two steps: Firstly a model of the FG voltage, and secondly a general leakage law.

The model of the FG voltage V_{FG} is based on the calculation of the FG charge Q_{FG} which assumes that the FG is coupled to each relevant terminal of the cell via linear capacitors (1). Although that approach is standard [3], one must understand that the linearity of the capacitors is an assumption which might not be acceptable for some NVM cells such as single-poly EEPROM having strongly non-linear capacitors (MOS cap).

$$Q_{FG} = \sum_{i \in \{\text{terminals}\}} C_i \times (V_{FG} - V_i) \quad (1)$$

where V_{FG} is the FG voltage, V_i is the voltage applied to the i^{th} terminal and C_i is the capacitance linking the FG to the i^{th} terminal.

One of the terminals is the Control-Gate (CG) of the cell. The threshold voltage V_T of the cell can be defined as the CG voltage needed to have the required voltage at the FG in the read conditions. One can show from (1) that Q_{FG} is a linear function of the threshold voltage V_T .

$$Q_{FG} = \gamma \times C_{FG} \times (V_{Tneu} - V_T) \quad (2)$$

Where V_{Tneu} is the threshold voltage of the cell when Q_{FG} is null and γ is the coupling factor of the FG to the CG (3).

$$\gamma = \frac{C_{CG}}{\sum_{i \in \{\text{terminals}\}} C_i} = \frac{C_{CG}}{C_{FG}} \quad (3)$$

Where C_{FG} is the FG capacitance which is equals to the sum of all coupling capacitors C_i .

Finally, it is possible to calculate the FG voltage as a function of V_T . from (1) and (2),

$$V_{FG} = \gamma \times (V_{Tneu} - V_T) + \frac{\sum_{i \in \{\text{terminals}\}} C_i \times V_i}{C_{FG}} \quad (4)$$

The general lifetime model is based on the five assumptions listed below:

1. There is a leakage current I_{leak} between the FG and one or more terminals all being at the same potential V_{term} .
2. I_{leak} is a function of V_{FG} and V_{term} only as shown in **Eq. 5**.
3. I_{leak} is null if $V_{FG} = V_{term}$.
4. I_{leak} is monotonically increasing with $V_{FG} - V_{term}$.
5. $(1/I_{leak})$ is integrable except at $V_{FG} = V_{term}$.

The 1st assumption is valid while there is a main leakage path. The 2nd assumption excludes erratic behaviours that are time dependent. We will see later how this point is managed. The 3rd and 4th assumptions are usual rules for resistive-like current laws. Finally, the 5th assumption is a requirement to build the model.

$$I_{leak} = I_0 \times \eta \left(\frac{V_{FG} - V_{term}}{V_0} \right) \quad (5)$$

Where, I_0 and V_0 are the parameters of the model, and η is a function.

According to 3rd assumption, **Eq. 4** and **Eq. 5**, there exists a value of V_T where there is no leakage current. That value is the 'equilibrium threshold voltage' V_{Teq} (**6**).

$$V_{Teq} = V_{Tneu} - \frac{V_{term}}{\gamma} + \frac{\sum_{i \in \{terminals\}} C_i \times V_i}{\gamma \times C_{FG}} \quad (6)$$

Combining (4), (5) and (6) we obtain:

$$I_{leak} = I_0 \times \eta \left(\frac{V_{Teq} - V_T}{V_0 / \gamma} \right) \quad (7)$$

The leakage current modifies the FG charge as shown in (8).

$$I_{leak} = - \frac{dQ_{FG}}{dt} \quad (8)$$

Equations 7 and 8 give a differential equation that can be integrated according to the 5th assumption.

$$\begin{cases} t = \tau \cdot \left[\kappa \left(\frac{V_{Teq} - V_T}{\nu} \right) - \kappa \left(\frac{V_{Teq} - V_{Tini}}{\nu} \right) \right] \\ \nu = \frac{V_0}{\gamma} \\ \tau = C_{FG} \times \frac{V_0}{I_0} \end{cases} \quad (9)$$

Where t is the time for the memory cell to reach V_T knowing that its initial state is V_{Tini} . Function κ is the

integral of $1/\eta$. ν and τ are the basic parameters of the model.

Equation 9 is the general model which gives important information on the behaviour of a FG cell as a function of time and voltage without fixing the leakage current. Moreover, the introduction of physical variables such as V_{Tini} , $V_{Tneutral}$ and V_{Teq} simplifies the understanding of the equation. One can see that (Fig. 1):

- a. κ is a function of $V_{Teq} - V_T$. It implies that modifying the stress conditions has exactly the same effect as changing the initial state.
- b. In general, κ is infinity at $V_T = V_{Teq}$, so that the time offset defined by the initial state becomes negligible. In other words, the threshold voltage V_T becomes weakly dependent on the initial V_T after a while.
- c. The threshold voltage V_T converges toward the equilibrium voltage V_{Teq} .
- d. The threshold voltage V_T at a given time t can be approached by a linear function of V_{Teq} and so a linear function of the stress voltages (cf. **Eq. 6**) since points **a** and **b**.

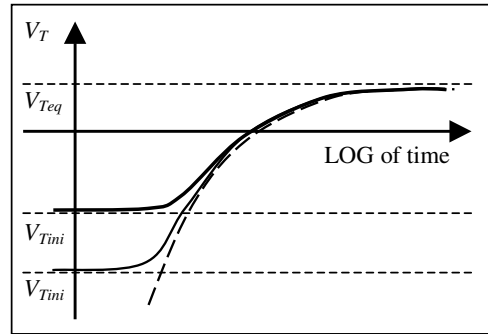


Fig. 1: Example of V_T -time characteristic deduced from the general model. The hashed-line shows the asymptote towards which all characteristics converge irrespective to V_{Tini} . The thin and thick lines show two examples of characteristics corresponding to two initial V_T 's. Shifting V_{Teq} translates the asymptote and therefore all the converging curves.

3. The analytical lifetime model

The calculation of the lifetime requires a model for the leakage current I_{leak} . Several models have been proposed: Hopping Conduction Model (HCM) [4], Fowler-Nordheim (FN) [5], exponential [1] and SILC based model [6]. Physical models such as SILC are not practical because of their complexity which disallows short calculations. We used Hoefler's model [7] for the leakage current (10) because it is capable of modelling a large range of current laws such as FN ($\chi = 1$), Frenkel-Poole ($\chi = -0.5$), Ohm ($\chi = 0$) or any arbitrary law. Moreover, it can be integrated (according to the 5th assumption).

$$\eta(x) = x^{\chi+1} \times \exp \left(- \frac{\chi}{|\chi|} \times x^{-\chi} \right) \quad (10)$$

$$t = \tau \times \frac{\chi}{|\chi|} \times \left\{ \exp \left[\frac{\chi}{|\chi|} \times \left(\frac{\nu}{|V_{Teq} - V_T|} \right)^\chi \right] - \exp \left[\frac{\chi}{|\chi|} \times \left(\frac{\nu}{|V_{Teq} - V_{Tini}|} \right)^\chi \right] \right\} \quad (11)$$

Where x is the variable and χ is a parameter to be added to V_0 and I_0 . By definition, $\chi/|\chi|$ is 1 if $\chi=0$. Parameter χ must be strictly lower than 1 to be in line with the 3rd assumption. The analytical model (11) is calculated by using the definition of κ (9) and η (10).

4. The failure rate procedure

The lifetime model shown above (11) has been used to predict the failure rate of MB in case of RT retention. However, that model can be used for more engineering and reliability purposes such as the extraction of cell parameters, erratic behaviour [3], disturbs [8], or retention [5].

The MB failure rate prediction is a reliability work that requires a well defined procedure. One problem is always encountered when trying to define the procedure: MB often have erratic behaviour. Indeed, the MB, that loose quickly their charge, are generally not numerous and often starting and/or stopping to move in an unpredictable way [5]. The so called start and stop MBs show that the leakage current I_{leak} can also be time dependent, in contradiction to the 2nd assumption of the general lifetime model. Because of the erratic MB, it is impossible to determine which bit will fail in the field. So, the MB problem can only be treated statistically. Different statistical methods have been proposed to overcome the erratic behaviours. Thus, A. Scarpa [5] worked on the tail of the Cumulative distribution Function (CdF) of cell threshold voltage assuming that the tail has a constant slope in a Gaussian distribution. D. Ielmini [9] introduced the Equivalent Cell (EC) principle on the CdF of cell threshold voltage V_T assuming that those ECs can replace real cells. R. Degraeve [1] re-used, for the percolation model, the EC principle but on the CdF of the threshold voltage shift $V_T - V_{Tini}$ making an additional assumption on the initial state. Moreover, he assumed a distribution law given by the percolation principle. We propose a new procedure for the failure rate prediction which requires no such assumption.

The new procedure is based on two important observations that will be shown in the following section: Firstly, most of MB display no erratic behaviour over a long period. Secondly, erratic MB behave like non-erratic MB when they move. So, it is possible to apply the lifetime model to each MB when they display no erratic behaviour, knowing that the erratic behaviour can only delay the time they fail. So, the procedure used to predict the lifetime is:

1. Locate the MB based on the CdF of the threshold voltage of all cell.
2. Extract the lifetime parameters for all MB.
3. Isolate the erratic MB from the non-erratic ones based on the extracted parameters.

4. Calculate the failure rate excluding the erratic MB.

The extraction of the parameters ν and τ is done on each MB. The value of parameter χ is shared by all MBs. It has been extracted on the most moving bits only so that we optimised the leakage current I_{leak} in the largest possible voltage range. So, the proposed procedure has two advantages compare to Hoefler's one [7]: Firstly, the number of parameters extracted per cell is higher, and secondly, it does not use the EC principle.

5. Results

Figure 2 shows, for a few cells, that there exist MBs having no erratic behaviour over a long period (9 month). The figure also proves the good fit we can achieve with the lifetime model. The parameters extracted on erratic MB are located in the extrinsic part of their CdF (fig. 3) because the lifetime model is limited by the 2nd assumption of the general model. However, one can see that 80% of the MB population has displayed no erratic behaviour during a 9 month Room Temperature retention test. Moreover, the erratic MBs behave exactly like the non-erratic ones when they move. Indeed, figure 4 reveals that if the lifetime model is fitted in the periods of their movement, their ν - τ parameters completely overlap the ones of the non-erratic MB. Therefore, the erratic behaviour is not dominating the MB, and the failure of the erratic MBs is delayed by the fact that they are not moving constantly. It implies that the failure rate can be calculated excluding the erratic MB.

Before calculating the failure rate we have to check its predictability. This is a new notion we introduced which consists of verifying that the lifetime model plus the collected data are good enough to give reliable predictions. In a first time, the capability of the model to make a prediction is checked and validated (Fig. 5) by extracting the parameters ν - τ of the lifetime model using the data from the beginning of the RT retention test (1st month). Then we verify that the data at the end of the test (9 month) is correctly predicted. The consistency of the predictions is verified (Fig. 6) by comparing predictions done using the parameters ν - τ extracted on longer and longer periods of the RT retention test. Figure 6 shows that predictions based on short test (< 1 month) are significantly different than the ones based on longer tests (≥ 2 month). However, the predictions made for the fastest MB using the data from the 1st month of the RT retention test are rather good. More generally, we found that a good parameter extraction requires a minimum shift of the MB threshold voltage. Below that minimum the imprecision in the data can dominate. Once the lifetime predictability is validated, we just have to calculate the failure rate (Fig. 7) using all the extracted parameters.

References

- [1] R. Degraeve et al., IEDM, (2001)
- [2] D. Wellekens et al., IEEE Trans. On Elec. Dev. **47**, 2153 (2000)
- [3] A. Chimenton et al., IEEE Trans. on Dev. & Mat. Rel. **1**, 179 (2001).
- [4] H. Kameyama et al., IRPS **38**, 194 (2000).
- [5] A. Scarpa et al., IRW, 24 (2000).
- [6] D. Ielmini et al., Solid-State Elec. **46**, 1749 (2002)
- [7] A. Hoefler et al., IRPS **40**, 21 (2002).
- [8] J. De Blauwe et al., IEDM, 343 (1996).
- [9] D. Ielmini et al., IRPS **39**, 73 (2001)

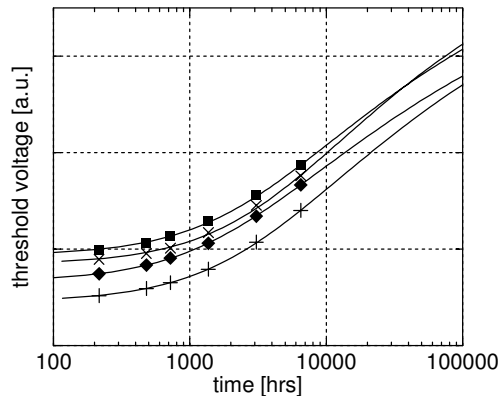


Fig. 2: Examples of V_T -time characteristics of non-erratic MBs – Measured data (symbols) versus lifetime model – Parameter extraction done using all data points – Extrapolation over 10 years of V_T -time characteristic.

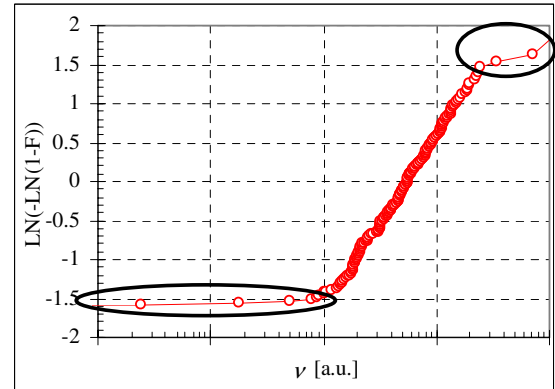


Fig. 3: CdF of parameter ν extracted on all MBs using the all data from the 9 month RT retention test - erratic MBs have their parameter ν in the extrinsic parts of the distribution (circled areas).

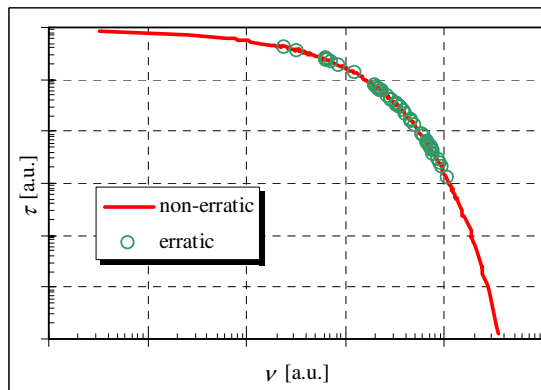


Fig. 4: Correlation between parameters ν and τ of the non-erratic (thick line) and the erratic (circles) MBs – the extraction of the ν - τ parameters is done on the periods of their movement only.

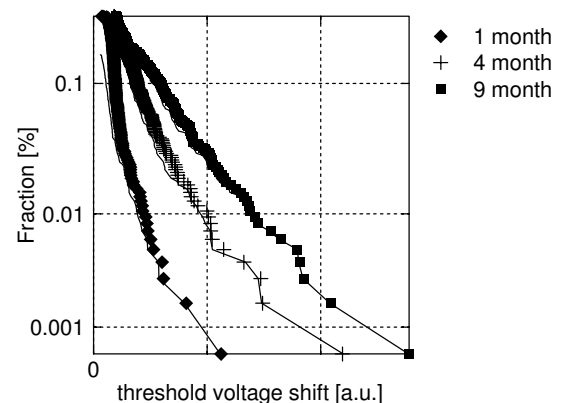


Fig. 5: CdF of the threshold voltage shift – Measured data (symbols) versus lifetime model (lines) – Parameter extraction done using data points from the 1st month only – Extrapolation over 9 month.

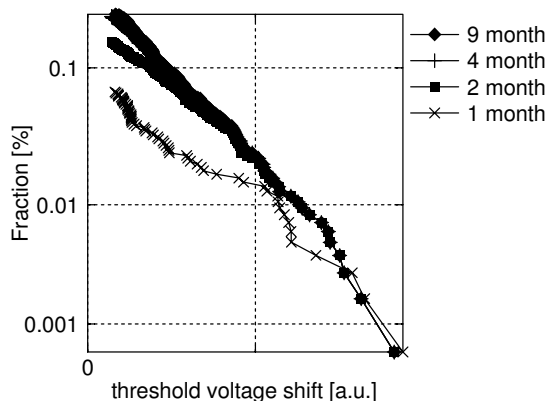


Fig. 6: CdF of the predicted threshold voltage shift assuming 10 years of RT retention conditions – Prediction using the parameters extracted with the measured data from the first 1, 2, 4 and 9 month of RT retention test.

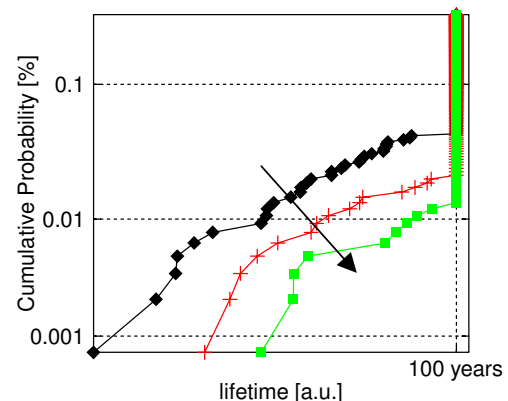


Fig. 7: Lifetime CdF for increasing threshold voltage limits (arrow) – Compliance at 100 years.

EEPROM retention time extrapolation from floating gate SILC measurements

S. Burignat¹, N. Baboux¹, C. Plossu¹ and P. Boivin²

¹ Laboratoire de Physique de la Matière, UMR CNRS 5511,
Bât. Blaise Pascal, INSA de Lyon, 7 Av. Jean CAPELLE, 69621 Villeurbanne Cedex, France
² ST Microelectronics, ZI Rousset, BP2, 13106 Rousset Cedex, France

Abstract

This work presents new experimental data on ultra-low level Stress Induced Leakage Currents measured in 7 – 8 nm EEPROM tunnel oxides at room temperature, by using the floating gate technique. Current-voltage characteristics are successfully modeled by a one-step tunneling model taking into account the spatial and energetical profiles of defects in SiO₂. New phenomena are pointed out, a negative differential resistance and an abrupt decrease of leakage current below low-level threshold gate potentials. From our model, these behaviors can be well simulated considering discrete defects energy levels in the SiO₂ bandgap. Moreover threshold voltages measured around 2 to 3 V appear to be very close to retention conditions floating gate potentials in EEPROM cells. According to a simple model, EEPROM retention times are extrapolated and it is shown that the particular behavior of SILC currents could explain the fact that retention times remain in the product specifications even for hardly cycled cells.

1. Introduction

Charge retention remains the main reliability issue preventing further SiO₂ tunnel oxide thickness reduction in Floating Thin Oxide (FLOTOX) Electrically Erasable Programmable Read Only Memory (EEPROM) devices. Because direct evaluation of retention times at room temperature require unacceptably long measurement times in industrial environment, these are usually predicted from extrapolation of high temperature measurements according to a “1/T dependent model”. Nevertheless, this model has been recently controversial and a new “T dependent model” has been proposed [1,2], which remains to be validated. Floating gate charge loss is supposed to be mainly associated to Stress Induced Leakage Currents (SILC) as conduction path through the tunnel oxide. The modelization of EEPROM retention kinetics taking into account this charge loss requires the knowledge of oxide current-voltage characteristics at very low potentials (less than 3 V). These data are experimentally inaccessible on small-size memory cells but also on large area plane capacitors since their level remains below the resolution of conventional current measurement apparatus.

In this work, we have used the so-called “floating gate technique” to obtain low-level current-voltage I-V characteristics of virgin and stressed tunnel oxides. Despite the fact that the floating gate technique requires very long measurement times, it has been shown to be a powerful technique to reach low-level (inferior to 10⁻¹⁵ A) leakage currents. Specific test structures have been

implemented on EEPROM product wafers, which allow both direct and floating gate tunnel oxide current measurements. We first present this EEPROM-like test structure, then we give a brief view of the floating gate technique principle. A physical modelization of experimental I-V measurements is proposed in paragraph 3. Finally EEPROM cell retention times are extrapolated from a simple electrical model.

2. Experimental

Description of test structures

Specific EEPROM-like test structures shown in Fig. 1 were fabricated¹ in the scribelines of memory device wafers from a standard double-poly EEPROM technology.

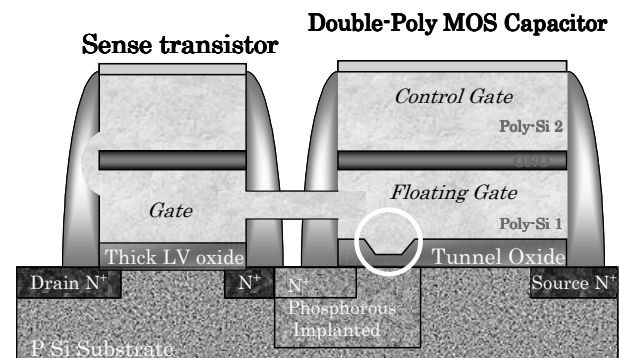


Fig. 1: EEPROM-like floating gate test structure.

These structures are macroscopic double-poly MOS capacitors that reproduce the different stacked layers of the active area of a memory cell state transistor: the control gate (Poly2), the interpoly oxide-nitride-oxide (ONO) dielectric layer, the floating gate (Poly1), the SiO₂ tunnel oxide and the N⁺⁺ implanted drain.

<i>ONO dielectric</i>	
Area	64 000 μm ²
Equivalent oxide thickness	20 nm
<i>Thick LV oxide</i>	
Area	36 640 μm ²
Electrical thickness	27 nm
<i>Nitrided tunnel oxide</i>	
Area S	19 360 μm ²
Electrical thickness t _{ox}	7-8 nm
<i>Sense transistor</i>	
Width	10 μm
Length	0.7 μm

Table 1: Main geometrical parameters of the double-poly MOS test structures.

¹ Samples were supplied by ST Microelectronics Company, Rousset, France.

The other specificity of this test structure is that the floating gate (Poly1) is directly connected to the gate of a NMOS sense transistor as illustrated in Fig.1. The Poly1 and Poly2 gates are polycrystalline silicon layers deposited by chemical vapour deposition and phosphorus N^{++} doped. The main geometrical parameters of the double-poly MOS test structures are reported in Table 1. The thermally grown tunnel oxide and the deposited ONO layer are 7-8 and 20 nm thick respectively. The tunnel oxide area is $19\,360\,\mu m^2$. External pads are accessible for the control gate, floating gate, source, drain, sense transistor gate and substrate.

Current measurements

Direct current measurements were classically performed using an optimized probe station protected by a double shielding Faraday's cage, allowing to reach a current resolution of 10^{-15} A. The floating gate technique was used to measure lower level leakage currents. It is based on potential time monitoring of the Poly1 floating gate once it has been externally charged and then left floating. The sense transistor gate which is connected to the floating gate is used to monitor the time variations of the floating gate potential V_{FD} via the measurement of the source current I_{DS} , by considering the transconductance g_m of the sense transistor.

$$g_m = \frac{dI_{DS}}{dV_{FD}} = \frac{dI_{DS}/dt}{dV_{FD}/dt} \quad (1)$$

The leakage current I_{FD} through the tunnel oxide is linked to the floating gate charge Q_{FG} time variation by :

$$I_{FD} = \frac{dQ_{FG}}{dt} \quad (2)$$

Q_{FG} is related to the floating gate potential by the tunnel oxide capacitance C_{FD} , the influence of the sense transistor gate capacitance being negligible.

The leakage current I_{FD} can finally be expressed by relation (3).

$$I_{FD} = C_{FD} \frac{dV_{FD}}{dt} + V_{FD} \frac{dC_{FD}}{dt} \quad (3)$$

The experimental characteristics $I_{DS}(V_{FD})$ and $C_{FD}(V_{FD})$ are used to extract the leakage current I_{FD} . The acquisition of the main characteristic $I_{DS}(t)$ is performed for several weeks to reach a sufficient resolution. Current levels as low as $2 \cdot 10^{-17}$ A were reached despite the fact that the measurement time to reach this current level is more than 50 days.

Stress conditions

In the aim to compare the retention time for a virgin and a stressed structure, stresses have been performed using the constant current stress (CCS) technique. The injected charge Q_{inj} through the tunnel oxide is calculated by simple integration of the experimental injected gate current with time.

Experimental results

Typical experimental I-V curves are given in Fig.2 for a virgin test structure compared to a stressed structure. For the virgin structure, the floating gate technique has been used. It can be seen a negative

differential behavior between 4 V and 3 V, followed by an abrupt current decrease below a threshold voltage around 3 V. This behavior has been observed on many samples from different wafers. The degraded structure has been CCS stressed up to a total injected charge Q_{inj} of $3.2\,C/cm^2$ that corresponds approximately to the injected charge after 3×10^5 write/erase cycles in a memory cell. The I-V curve has been obtained by direct current measurements. A typical SILC behavior between 3 and 7 V is observed but it must be noted that current values below 3 V are not significative because their level reaches the 10^{-15} A current resolution limit in direct measurement mode.

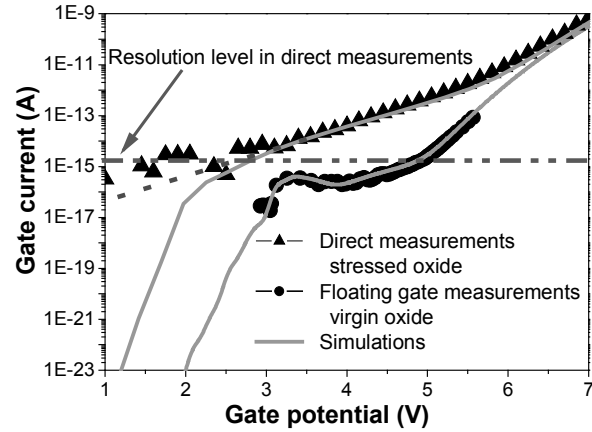


Fig.2 : Experimental I-V curves of a virgin structure (floating gate measurements) and a stressed one (direct measurements). Resolution level in direct measurement mode is visualized (dotted line). Comparison with simulated I-V curves.

3. Conduction model

A general consensus in the literature associates SILC currents to tunneling conduction mechanisms assisted by defects in the SiO_2 bulk. We have developed a one-step trap-assisted tunneling model in the WKB approximation and considering a 1D (perpendicular to the Si/ SiO_2 interface) mechanism. As the floating gate is N^{++} polysilicon and the silicon drain highly phosphorus doped, we have considered a FERMI-DIRAC distribution for the states densities in the two electrodes of the tunnel oxide capacitance. BEDNARCZYK approximation for the FERMI-DIRAC integral was used [4] for calculation simplification. The two interface charges Q_{FG} and Q_D in the floating gate and in the drain respectively were calculated classically by (4), taking into account the trapped charge Q_{ot} in the oxide :

$$Q_{FG} = \frac{\epsilon_{ox}}{t_{ox}} V_{ox} - Q_{ot} \left(1 - \frac{\bar{X}}{t_{ox}}\right) \quad (4a)$$

$$Q_D = + \frac{\epsilon_{ox}}{t_{ox}} V_{ox} - Q_{ot} \frac{\bar{X}}{t_{ox}} \quad (4b)$$

where ϵ_{ox} is the tunnel oxide dielectric constant, t_{ox} the tunnel oxide thickness, V_{ox} the tunnel oxide potential, $\bar{X}$ the oxide charge centroid measured from the cathode. As the presence of a charge in the bulk oxide can modify the tunneling barrier shape (as illustrated in Fig. 3), we have considered two average electric fields F_{ox1} and F_{ox2} delimited by the centroid $\bar{X}$.

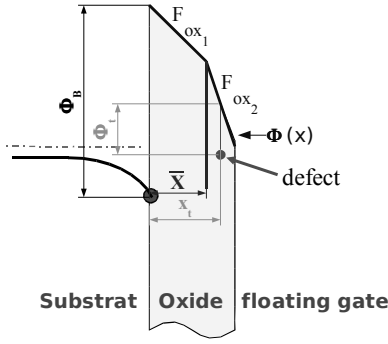


Fig.3: Tunnel oxide potential barrier and average oxide electric fields F_{ox1} and F_{ox2} .

$$F_{ox1} = -\frac{V_{ox}}{t_{ox}} + \frac{Q_{ot}}{\epsilon_{ox}} \left(1 - \frac{\bar{X}}{t_{ox}}\right) \quad (5a)$$

$$F_{ox2} = -\frac{V_{ox}}{t_{ox}} - \frac{Q_{ot}}{\epsilon_{ox}} \frac{\bar{X}}{t_{ox}} \quad (5b)$$

Considering that each defect is active on an area corresponding to its capture cross-section σ_t , only a proportion ρ of the total tunnel area S is concerned by the SILC current J_{SILC} , the remaining area being considered as virgin without defect (J_{virgin}). The general expression of the total current density J is then expressed by :

$$J = \rho J_{SILC} + (1 - \rho) J_{virgin} \quad (6)$$

where N_t is the total number of defects involved in SILC mode conduction, S the total tunnel area and ρ given by :

$$\rho = \frac{N_t \sigma_t}{S} \quad (7)$$

J_{virgin} is modeled by a Fowler-Nordheim injection through a perfect oxide without defects. The SILC current density J_{SILC} is calculated from the SILC current I_{SILC} ($J_{SILC} = I_{SILC}/S$) expressed by (8) :

$$I_{SILC} = \frac{4 \pi m_e q k T}{h^3} \int_0^{t_{ox}} \sigma_t fct(x) \int_0^{+\infty} \Xi(E, x) \ln \left(\frac{1 + \exp\left(\frac{E_{F1} - E}{kT}\right)}{1 + \exp\left(\frac{E_{F2} - E}{kT}\right)} \right) fct_E(E) dE dx \quad (8)$$

where x is the distance from the injecting interface, E the energy level of an incident electron relatively to the conduction band level of the cathode electrode at the interface, $\Xi(E)$ the generalized tunnel barrier transparency, E_{F1} and E_{F2} the Fermi levels in the cathode and the anode respectively, $m_e = 0.5 \times m$ the electron effective mass in SiO_2 and m the electron mass, q the electron charge, k the Boltzmann constant and T the temperature. $fct_t(x)$ and $fct_E(E)$ are the spatial and energetical distribution functions of defects in the oxide layer.

Simulation results

Both spatial and energetical defect profiles in the tunnel oxide can be tailored to fit the experimental I-V

curves using the previous model. The extracted profiles corresponding to experimental data of Fig.2 are given in Fig.4 and Fig.5 respectively.

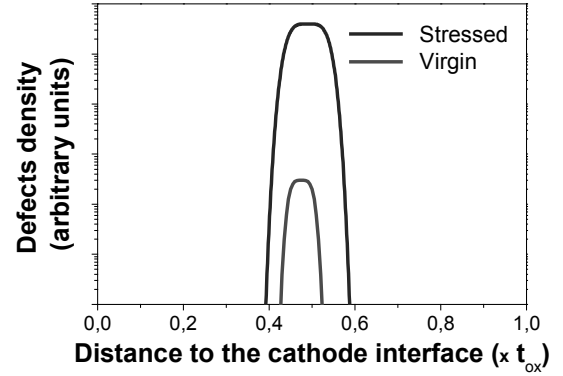


Fig.4 : Spatial defect profiles for both virgin and stressed structures.

Defects appear to be located very close to the middle of the oxide layer. Their spatial profiles are symmetrical with respect to the middle oxide that is coherent with the gate polarity symmetry of experimental SILC characteristics. The proportion ρ of the tunneling area that is affected by SILC is about 0.2 % for the virgin structure, which is very small, but increases up to 25.4 % for the stressed structure.

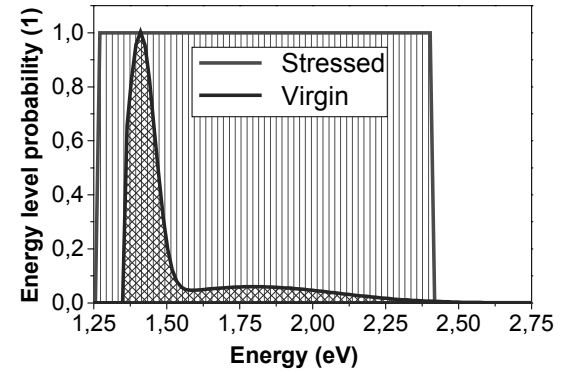


Fig.5 : Energetical defects profiles for both virgin and stressed structures. The energy level is measured from the bottom of the cathode electrode conduction band at the oxide interface.

In the case of virgin structures, the energetical defect profile can be accurately extracted from floating gate measurements, as shown in Fig.5. A main defect energy level in the SiO_2 bandgap appear around 1.4 eV up from the injecting electrode conduction band at the interface, corresponding to an energy level located at 2.8 eV from the bottom of the SiO_2 conduction band. In the case of stressed structures, the energetical defect profile is enlarged and can be modeled by a constant energy band as shown in Fig.5. This can be explained by the generation of stress-induced defects in SiO_2 resulting in an increase of the proportion ρ of tunneling area affected by SILC and consequently a non-negligible coupling between defects.

Simulated I-V curves are given in Fig. 2 for both virgin and stressed structures. A very good agreement with experimental data is obtained. Both negative differential resistance and current threshold voltage

phenomena of the virgin structure are well modeled. For the stressed sample, the simulation results in a current decrease threshold voltage around 2 V, which is fixed by the bottom level of the defects band energy (around 1.25 eV up from the injecting electrode conduction band at the interface in Fig.5). This current threshold voltage phenomenon appearing for low potentials is not accessible by direct measurements.

4. Retention time extrapolation

In EEPROM cells, the floating gate potential in both write and erase states is usually between 1 V and 3 V in absolute values. The current decrease threshold voltages previously observed on tunnel oxide I-V characteristics are in the same potential domain. So they appear to be very close to floating gate potentials in retention conditions and could greatly influence retention times. From a simple electrical model [5], we have extrapolated retention times of an EEPROM cell by considering that the floating gate charge loss kinetics is governed by the tunnel oxide I-V characteristic measured on the test structure. The retention time $t_{1,2}$ for a variation of the floating gate potential from V_{FG1} to V_{FG2} can be expressed as :

$$t_{1,2} = -\frac{1}{A_{D,tun}^{tun}} \int_{V_{FG1}}^{V_{FG2}} \frac{C_{tun}^S}{J_{tun}(V)} dV \quad (11)$$

where C_{tun}^S is the tunnel oxide capacitance per unit area, J_{tun} the current density through the tunnel oxide and $A_{D,tun}$ the capacitive coupling coefficient between the floating gate and the drain given by :

$$A_{D,tun} = \frac{C_{tun}}{C_T} \quad (12)$$

C_{tun} is the the tunnel oxide capacitance and C_T the total capacitance of the structure. Equation (13) shows that the retention time depends on the geometry of the structure through the coefficient $A_{D,tun}$ and on the tunnel oxide conduction law through the integral.

We have compared EEPROM cell retention times for virgin and degraded structures calculated from simulated I-V curves of Fig.2, for a 20% charge loss and for different initial floating gate potentials V_{FDinit} . Results are presented in Fig.6 and Tab.2. As an example, for $V_{FDinit} = 3$ V, the retention time should be 233 years for a virgin cell and should decrease down to 4.7 days for a degraded cell which becomes obviously outside the product specifications. With $V_{FDinit} = 2$ V, the retention time for the degraded cell increases to 33 years and returns into the specifications. Nevertheless it is important to note that if the significative part (between 3 V and 5 V) of the stressed oxide experimental I-V curve in Fig.2 is considered instead of the simulated one, and is extrapolated towards lower potentials without any inflexion and abrupt decrease of the leakage current (dotted line in Fig. 2), the extrapolated time retention for $V_{FDinit} = 2$ V would be greatly inferior. Yet presently produced EEPROM devices remain reliable even after cycling and tunnel oxide degradation. This could support

the fact that there is actually a current decrease phenomenon at low potentials, which would be associated to the presence of specific defects energy levels in SiO₂, allowing to keep retention times into the product specifications.

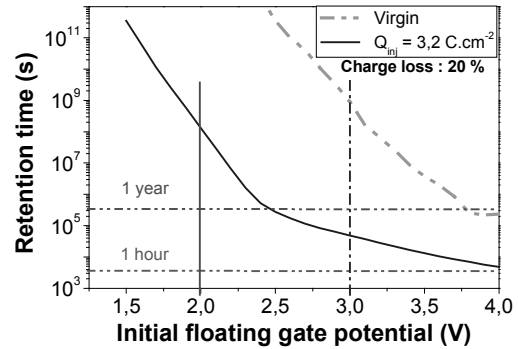


Fig.6 : Extrapolated EEPROM cell retention time at 20% charge loss as a function of initial floating gate potential, for both virgin and stressed tunnel oxides.

V_{FDinit}	4 V	3 V	2 V	1.5 V
Virgin	22.5 days	233 years	$4.4 \cdot 10^{-7}$ years	$4.7 \cdot 10^{10}$ years
Stressed	8.33 hours	4.7 days	33 years	91 600 years

Table 2 : EEPROM cell retention times at 20% charge loss for different initial floating gate potentials between 1.5 V and 4 V, for both virgin and stressed tunnel oxides.

4. Conclusion

A new phenomenon reflected by a brutal decrease of the leakage current below low threshold voltages has been observed in EEPROM 7-8 nm virgin tunnel oxides by using the floating gate technique. This behavior has been modeled by a physical one-step trap-assisted tunneling conduction model and related to the energetical profile of involved defects in SiO₂. These defects may be responsible or precursor to SILC currents. On the other hand, threshold voltage values are close to EEPROM floating gate potentials in retention conditions. From simulation results obtained on stressed oxides, retention times of EEPROM cells have been extrapolated and it has been shown that the abrupt decrease of leakage currents at low potentials could explain that product retention specifications remain satisfied even by cycled degraded memory cells. The link between SILC currents and EEPROM retention properties remains to be further investigated thanks to new floating gate measurements and in particular as a function of temperature.

References

- [1] De Salvo B, IEEE Trans.El.Dev. **46**(7) (1999), 1518-1524
- [2] De Salvo B, IEEE Elec.Dev.Lett. **20**(5) (1999), 197-199
- [3] Nissan-Cohen Y, IEEE El.Dev.Lett. **7**(10) (1986) 561-563
- [4] Vasileska D., Journal of Modeling and Simulation of Microsystem **1**(1) (1999) 49-56
- [5] De Blauwe J., IEEE Trans.El.Dev. **45**(8) (1998), 1751-176

A New Reliable P/E Methodology for Fully Fowler-Nordheim Operating Non Volatile Memories

Fernanda Irrera, Giuseppina Puzzilli
Electronics Department, University of Rome "La Sapienza"
irrera@die.uniroma1.it

1. Introduction

In this work we propose an innovative methodology for programming non volatile memories by Fowler-Nordheim tunnel using short voltage pulses. The key feature of the pulsed methodology is in preserving endurance and data retention. For this reason, use of such methodology in principle allows for a reduction of the cell tunnel oxide, implying a scaling of the operation voltages. Speed and supply voltages can be traded off through the optimization of the P/E waveform, being the reliability no longer a major concern.

The applications of the pulsed P/E methodology are mainly of two kinds: on one side, in memory architectures requiring fully Fowler-Nordheim operation, as in the case of NAND arrays; on another side, in all those cases in which reliability is a main target, as for discrete storage nodes cells with extremely thin tunnel oxides.

The underlying physics is presented in Sec.2. Results on Flash cells and arrays with 7 nm and 3.5 nm thick tunnel oxides are reported in Sec. 3.

2. Basic theory

It has been recently demonstrated that the creation of a stable trap in SiO₂ during electrical stress is a two step process and requires a finite time [1]. First a trap precursor is created which physically consists in an unstable state of the SiO₂ amorphous network. Then precursors can be converted in stable traps if the external perturbation (i.e. electrical stress) is longer than a critical time (trap creation time, τ_{CR}). Nevertheless, defect precursors can also revert to stable bonds keeping off the electrical stress for a time longer than a characteristic time τ_{ANN} . In these conditions a spontaneous annealing of trap precursors occurs.

In order to verify this physical picture, experiments of electrical stress were performed using square waveforms with the following features: pulse time duration t_{ON} , amplitude V_{STRESS} and time between two pulses t_{OFF} . Stress waveform is schematically sketched in fig. 1. We refer to this stress condition as pulsed stress technique. Samples under test were MOS capacitors with 7nm (set A) and 3.5 nm (set B) oxide thickness [2, 3].

Self annealing of precursors can be quantitatively evaluated comparing values of SILC obtained under pulsed condition with those under a constant applied voltage (DC stress). In fact, spontaneous annealing of

precursors occurring during pulsed stress causes a correspondent decrease of stable defect generated during stress and, as a consequence, of SILC respect to the DC case. In this context, the characteristic times play important roles: τ_{CR} is a threshold time such that if $t_{ON} < \tau_{CR}$ no significant number of traps is created during a voltage pulse; τ_{ANN} is a characteristic time such that if $t_{OFF} \gg \tau_{ANN}$ almost all trap precursors generated during a pulse are annealed.

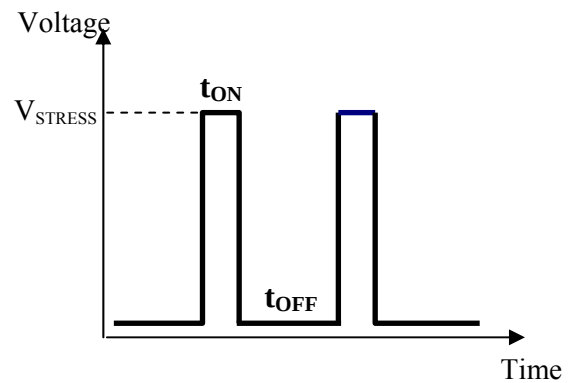


Fig.1. Qualitative picture of the waveform used in the pulsed stress experiment.

Fig. 2 shows the reduction of SILC measured on set A as a function of t_{OFF} . Samples A were stressed in different conditions, but the same charge (0.6 C/cm^2) was injected.

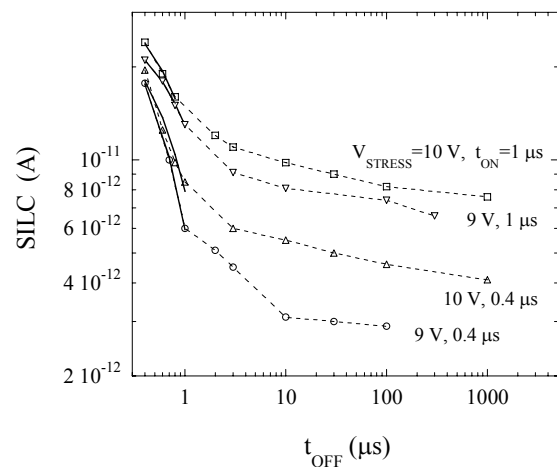


Fig.2. Measured SILC (at 4.5V) on set A vs. t_{OFF} after pulsed stress ($t_{ON} = 1 \text{ μsec}$ and 0.4 μsec) $V_{STRESS} = 9 \text{ V}$ and 10 V . Exponential interpolation of the short time behavior is sketched with bold solid line.

At very short t_{OFF} values, only a few precursors can be annealed during a period, the most being converted into stable traps during the subsequent pulse. Increasing t_{OFF} SILC decreases rapidly and for $t_{\text{OFF}} > 10 \mu\text{s}$ it saturates becoming independent on t_{OFF} , since all precursors created during a pulse are annealed before the next pulse. Precursor annealing can be modeled by the following rate equation for the density of precursors (N_{PREC}) [2]:

$$\frac{dN_{\text{PREC}}}{dt} = -\frac{N_{\text{PREC}}}{\tau_{\text{ANN}}} \quad (1)$$

The values of τ_{ANN} interpolating experimental data are in the range of $0.6 \div 1 \mu\text{sec}$. τ_{ANN} is almost independent on the stress parameters, since the precursor recovery takes place when no external voltage is applied. We observed a weak dependence on temperature in the range $20\text{-}150^\circ\text{C}$, in the sense that the higher the temperature the faster the annealing. We suppose that τ_{ANN} is correlated to the physical and chemical characteristics of the particular oxide film. In fact, for set B we observed the complete saturation of SILC for $t_{\text{OFF}} > 1 \mu\text{s}$, in opposition to $10 \mu\text{s}$ found for thicker films. This discrepancy can be ascribed to different bond energies, perhaps due to the different physical distance from the interface.

Figs. 3 and 4 show SILC data after stress for set A and B respectively, as a function of t_{ON} . The values of V_{STRESS} were 8 V, 9 V and 10 V for set A and 5.2 V for set B. τ_{CR} can be associated to the abrupt increase of SILC and is in the range of several tens up to hundreds of nanoseconds in the cases investigated here. As a first remark, the timescale for creation of traps depends mainly on the stress conditions rather than on the oxide thickness. Furthermore, SILC increases with V_{STRESS} and t_{ON} . This behavior was modeled both for thin (set A) [4] and ultrathin (set B) oxides [5].

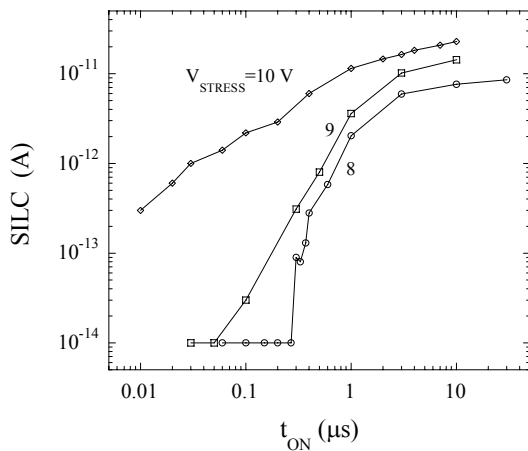


Fig.3 Measured SILC (at 4.5V) on set A vs. t_{ON} after stress performed with $t_{\text{OFF}} = 10 \mu\text{sec}$, $Q_{\text{INJ}} = 0.3 \text{ C/cm}^2$.

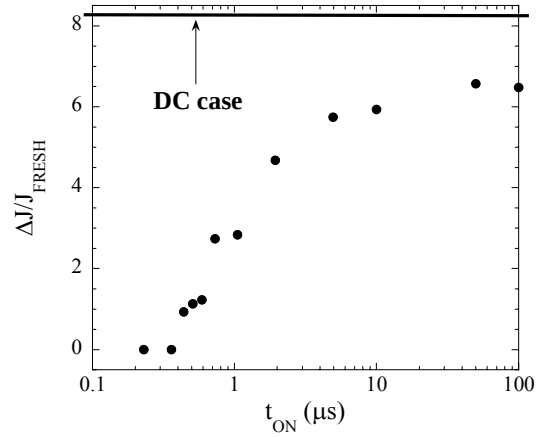


Fig.4. Percentile variation of SILC ($J_{\text{SILC}} - J_{\text{FRESH}} / J_{\text{FRESH}}$) measured (at 2.6V) on set B vs. t_{ON} after stress performed with $V_{\text{STRESS}} = 5.2\text{V}$, $t_{\text{OFF}} = 1 \mu\text{sec}$, $Q_{\text{INJ}} = 3 \text{ C/cm}^2$. Solid line is the value due to a DC stress with the same V_{STRESS} and Q_{INJ} .

3. Pulsed programming and erase of Flash Memories

3.1 Standard Flash cells with 7nm thick tunnel oxide

Flash cells investigated here were fabricated by STMicroelectronics with 7 nm-thick tunnel oxides and conventional process. The cell threshold voltage V_T after UV-erase was 2.3 V, with an intrinsic spread of about 0.35 V given by the wafer-fab. The target threshold voltage shift ΔV_T was 1.5 V (corresponding to a target injected charge $Q_{\text{AVT}} = 10^{-6} \text{ C/cm}^2$).

The target of the work was the optimization of the waveform with respect to the oxide degradation [6]. Several different waveforms were taken into account, keeping constant t_{OFF} (10 μs), but varying the program voltage V_{PR} (up to 20 V), t_{ON} (up to 500 ns, according to results obtained on the creation times) and the number of pulses N_p to give the target charge Q_{AVT} (10^{-6} C/cm^2). On the (V_{PR} , t_{ON}) plane, Q_{AVT} identifies a line representing all possible combinations of voltage and duration achieving the target ΔV_T . The crucial point is that along such a line the density of new traps as a function of V_{PR} and t_{ON} exhibits a (broad) minimum, clearly corresponding to minimum SILC degradation. Such a behavior is essentially due to the fact that, with respect to the minimum, if V_{PR} is increased the increase of the term representing trap creation dominates on the decrease of that describing precursor annealing, while the opposite is true if V_{PR} becomes smaller. In order to develop a methodology aimed at minimizing oxide damage under pulsed tunneling conditions, an analytical description accounting for all main physical mechanisms of interest (creation of stable traps and precursors, precursors annealing, and conversion of residual ones into stable traps) represents a useful starting point. To this purpose, the following equation describes the density of traps created by a pulsed and DC stress ($N_{\text{TRAP}}^{\text{PULSED}}$ and $N_{\text{TRAP}}^{\text{DC}}$, respectively) for the same fixed

value of $Q_{\Delta V_T}$:

$$N_{\text{TRAP}}^{\text{PULSED}}(V_{\text{PR}}, t_{\text{ON}}) = N_P \cdot [1 - \alpha_{\text{ANN}}(t_{\text{ON}})] \cdot N_{\text{TRAP}}^{\text{DC}}(V_{\text{PR}}, t_{\text{ON}}) \quad (2)$$

where α_{ANN} represents the fraction of trap precursors annealed at the end of the pulse, and $(1 - \alpha_{\text{ANN}})$ is the fraction of residual precursors which do not anneal and finally become stable traps. In this work, α_{ANN} as a function of t_{ON} was experimentally determined measuring SILC variations with respect to the DC case on the MOS capacitors described in Sect. 2. $N_{\text{TRAP}}^{\text{DC}}$ can be extracted using the trap creation model proposed in Ref.[4] and, solving eq. 2, $N_{\text{TRAP}}^{\text{PULSED}}$ can be finally calculated. Results are shown in fig. 5, in the case of $N_P = 2, 3$, and 4 ($\Delta V_T = 1.5$ V, hence $Q_{\Delta V_T} = 10^{-6}$ C/cm²).

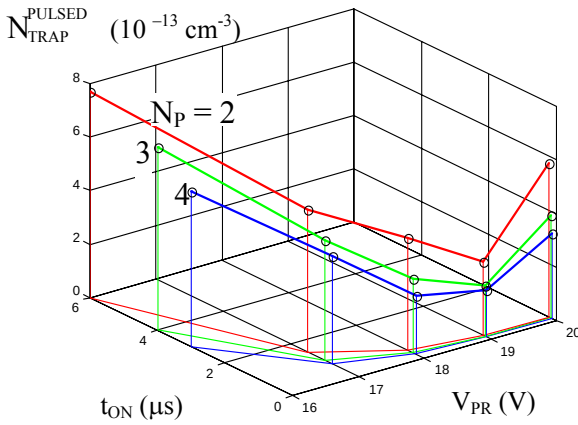


Fig.5. Density of created traps during FN tunnel programming vs. program voltage and pulse duration for $N_P = 2, 3$ and 4, for a target $\Delta V_T = 1.5$ V and $Q_{\Delta V_T} = 10^{-6}$ C/cm².

With the aid of fig.5 and according to operation constraints we chose $N_P = 3$, $V_{\text{PR}} = 18$ V, $t_{\text{ON}} = 333$ ns, $t_{\text{OFF}} = 10$ μ s. The correspondent program time, given by the following equation:

$$t_{\text{PR}} = N_P t_{\text{ON}} + (N_P - 1) t_{\text{OFF}} \quad (3)$$

is equal to 21 μ sec.

Some tens of Flash cells were programmed and erased with a few cycles using the optimized pulsed waveform and the obtained medium value of ΔV_T is 1.5 V with distribution of the programmed V_T which reproduces the 0.35 V spread in the UV- erased values. Furthermore, write/erase were also performed varying only the t_{OFF} value, and it was verified that $\Delta V_T = 1.5 \pm 0.175$ V. This demonstrates the accuracy of the new programming methodology (to the extent of a reasonable set of samples).

The impact of precursors annealing on memory reliability was investigated performing 10^5 cycles with waveforms different from the optimized one only for the value of t_{OFF} (which has been varied in the range

0.4-20 μ s). The obtained data of ΔV_T vs. t_{OFF} reported in fig.6 with closed symbols clearly show that, apart from the known intrinsic spread of values, only when $t_{\text{OFF}} \geq 10$ μ s cycling has no effect on ΔV_T . For shorter t_{OFF} the time between one pulse and the next one is not sufficient for the precursors created during the former to completely anneal out before the beginning of the latter pulse, thus confirming the results obtained with the capacitors. In fig. 6, values of the program voltage shift of the same samples before cycling are also shown (open symbols), in order to demonstrate that the behavior found after cycling is not due to characteristics of the fresh set, but only to the duration of t_{OFF} . The reproducibility of this result has been verified on a wide set of samples, all subject to 10^5 cycling with the optimized waveform ($t_{\text{OFF}} = 10$ μ s), confirming that the medium value of the distribution of the measured ΔV_T was 1.5 V with a spread of ± 0.175 V.

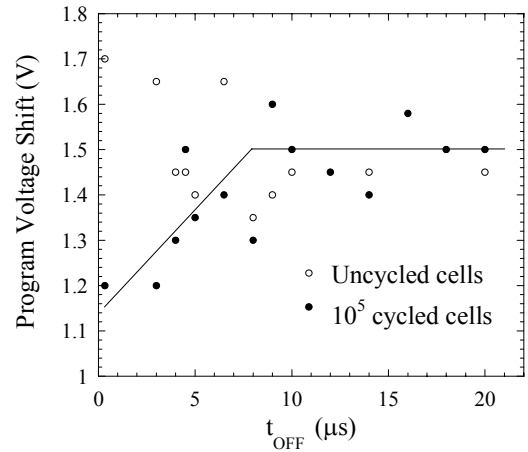


Fig. 6. Program voltage shift measured on several set as a function of t_{OFF} after 10^5 cycles (closed circles) and just a few cycles (open circles).

3.2 Silicon Nanocrystal Flash memory array with 4 nm thick tunnel oxide

We studied memory arrays made of 256Kbit cells ($L=0.3\mu\text{m}$, $W=0.2\mu\text{m}$) in a complete parallel connection (hereafter CAST, acronym of cell array stress test). Cells featured silicon nanocrystals in the place of a uniform floating gate [7] and a ultrathin (4 nm) tunnel oxide. The arrays were fabricated by STMicroelectronics. A 7nm thick control dielectric layer was deposited after nanocrystals formation, covered by a n^+ -doped polysilicon gate.

A total stress dose of 10^5 P/E cycles was imposed to the arrays, with P/E pulses of amplitude $V_G = \pm 11$ V and duration equal to 10 ms (standard cycling). We also performed 10^5 P/E cycles with a pulsed programming waveform (pulsed cycling) featuring amplitude $V_G = +11$ V, $t_{\text{ON}} = 2\mu\text{s}$ and $t_{\text{OFF}} = 1\mu\text{s}$, and a single erase pulse with duration $\tau_E = 10$ ms and amplitude $V_G = -11$ V. The number of programming pulses was fixed to give the same threshold voltage shift as in the standard waveform stress experiment. In

both cases cycling was periodically interrupted to monitor the window program [8]. Set investigated here were not optimized from the point of view of the program window.

Fig. 7 shows the threshold voltage window between the programmed and the erased array state as a function of P/E cycling using the standard and the pulsed technique. V_T has been extracted from the array I_D - V_G characteristics as the gate voltage corresponding to a drain current level $I_D=10^{-5}$ A. The V_T window has been reported normalized to its initial value ΔV_{T0} (0.5V) to allow for a more immediate comparison of the results. A more severe window closure is encountered for standard cycling with respect to pulsed tunnel cycling, resulting in a residual V_T window equal to 0.58 times ΔV_{T0} for standard cycling and 0.85 times ΔV_{T0} for pulsed cycling, after 10^5 P/E cycles.

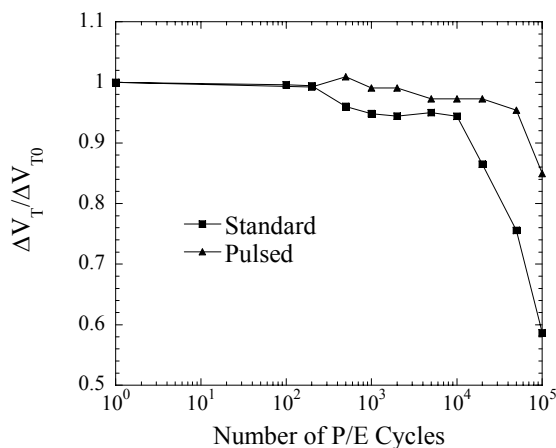


Fig. 7. Normalized threshold voltage window for nanocrystals Flash arrays as a function of the number of P/E cycles, for standard (squares) and pulsed (triangles) cycling.

After cycling the same arrays were programmed by substrate hot electrons opening a window $\Delta V_T = 1.5$ V, and data retention was investigated. In order to accelerate the data retention test, we applied a voltage of -2.5V for 10^5 s to the control gate (gate stress technique).

Fig.8 shows the normalized program windows as a function of gate stress time after 10^5 P/E standard and pulsed cycles. In that figure, very different short time behaviors can be distinguished depending on the programming methodology, whereas the long time dependence is similar. In particular, the standard programmed cells immediately start to lose charge, whereas the pulsed programmed cells retain the initial charge within 0.5 % for approximately one thousand seconds. This can be explained by the fact that standard cycling induces the creation of a greater trap density than in pulsed condition, and these traps assist conduction by trap assisted tunnel. On the contrary, at long-times the two curves in fig.8 tend to become parallel, as if the leakage mechanism is the same. In this case, leakage can be ascribed to direct tunnel from the floating gate into the substrate.

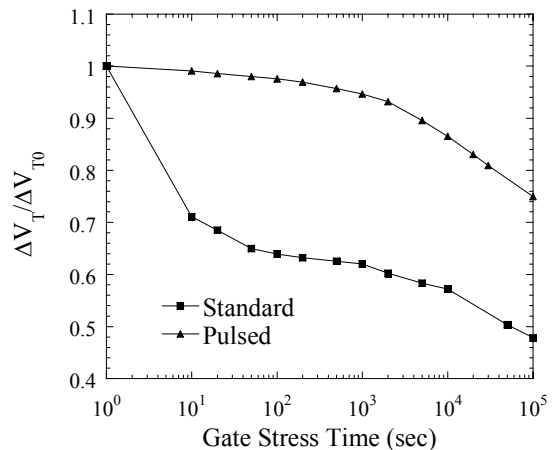


Fig. 8. Normalized program window $\Delta V_T/\Delta V_{T0}$ vs. the gate stress time performed after 10^5 standard and pulsed cycles.

Conclusions

In this work we demonstrate Flash memory tunnel programming from the channel using short voltage pulses. With the pulsed programming methodology noticeable improvement of tunnel oxide reliability can be achieved, reflecting in better cycling and data retention features. The reliability improvement can be traded-off with slightly larger voltages, to obtain faster cell programming. In conclusions, it is possible to choose the best compromise between speed and supply voltages, being the reliability no longer a major concern.

References

- [1] R. Feruglio, F. Irrera, B. Riccò, Microscopic Aspects of Defect Generation in SiO₂, *Microelectr. Reliab.*, 42 (2002), p. 1427.
- [2] F. Irrera, B. Riccò "SILC dynamics in MOS structures subject to periodic stress", *IEEE Trans. Electr. Dev.*, 49 (2002), 1729.
- [3] G. Puzzilli, D. Caputo, F. Irrera "Fast and reliable tunnel programming of nanocrystal nonvolatile memories", *IEEE Trans. Electr. Dev.*, 51 (2004), 1205.
- [4] F. Irrera "Degradation kinetics of thermal oxides" *App. Phys. Lett.* 79 (2001), 182.
- [5] F. Irrera, G. Puzzilli "Degradation of ultra-thin oxides", *IEEE Trans. on Dev. and Mat. Reliab.*, 4 (2004), 530.
- [6] F. Irrera, B. Riccò "Pulsed tunnel programming of nonvolatile memories", *IEEE Trans. Electr. Dev.*, 50 (2003), 2474.
- [7] B. De Salvo, G. Ghibaudo, G. Pananakakis, P. Masson, T. Baron, N. Buffet, A. Fernandes, B. Guillaumont "Experimental and theoretical investigation of nanocrystal and nitride-trap memory devices" *IEEE Trans. on Electr. Dev.*, 48 (2001), 1789.
- [8] G. Puzzilli, D. Caputo, F. Irrera, C. Monzio Compagnoni, D. Ielmini, A. Spinelli, A.L. Lacaita "Improving Floating-Gate Memory Reliability by Nanocrystal Storage and Pulsed Tunnel Programming" *IEEE Trans. on Dev and Mat. Reliab.*, 4, (2004), 390.

STI engineering for high program & erase performance of STI-bounded nitride-storage flash memory cells

M. Isler^a, J.-M. Schley^a, S. Riedel^a, T. Mikolajick^a, C. Ludwig^a, K.-H. Kuesters^a, G. Tempel^a, J.-U. Sachse^a, P. Deconinck^a, R. Mikalo^a, R. Reichelt^a, N. Schulze^a, E. Stein v. Kamienski^a, M. Strassburg^a, J. Willer^b, F. Lau^b, R. Hagenbeck^b, P. Haibach^b

^a Infineon Technologies Dresden, Koenigsbruecker Str. 180, D-01099 Dresden, Germany

^b Infineon Technologies Munich, Otto-Hahn-Ring 6, D-81739 Munich, Germany

Abstract

The impact of effects due to shallow trench isolation (STI) on the performance of narrow-channel nitride-storage flash memory cells is investigated. By engineering a negative STI step height the gate-control is improved, short-channel effects are reduced and the erase speed is significantly increased. One order of magnitude improvement in erase performance is achieved by increasing the hole generation rate by band-to-band tunnelling due to pronounced fringing fields. This translates into an endurance improvement of one order of magnitude and is particularly important for shrinking technology nodes since the erase performance deteriorates as the channel length is downscaled.

interface strongly depends on the STI step height. For negative values of the step height (level of active area above that of the STI) short channel effects can be significantly reduced. This can be recognized by comparing the subthreshold transfer characteristic for $V_{ds} = 0.1$ V and 1.3 V in Fig. 4a: The DIBL (drain-induced barrier lowering) is clearly reduced for negative step height compared to memory cells with positive step height. In addition, the subthreshold slope and the transconductance (Fig. 4b) is larger in case of negative step height. The reason for that and for the reduced DIBL is the improved gate control due to more pronounced fringing fields in case of negative step height. Another benefit of a negative step height is the increased effective active area by utilizing the vertical walls similar to the FinFET concept [4].

1. Introduction

Nitride-storage non-volatile flash memory cells using a nitride trapping layer to store 2 bits into one memory cell are promising for code and data storage applications due to their high-density and fast speed [1]. Programming is achieved by injection of channel hot electrons and erase by injection of hot holes generated by band-to-band-tunneling. Recently, a novel STI-bounded architecture was proposed for such memory cells in order to achieve even higher densities [2]. STI effects on the device characteristics are known to some extent from CMOS technology [3], but so far there is no experience how the STI corner affects the program and erase performance of narrow-channel nitride-storage memory cells. The subject of this work is to investigate and to optimize these effects.

2. STI effects on transfer characteristics

The principal device structure is shown in Fig. 1. The memory cell is very similar to a state-of-the-art STI-bounded NMOS transistor with the gate oxide replaced by an oxide(O)-nitride(N)-oxide(O) layer stack. The channel width is approximately 90 nm. The shallow trench isolation affects the electrical characteristics mainly in two ways: (i) boron segregation into the STI oxide reduces the boron concentration in the channel particularly at the STI edge (Fig. 2), (ii) the electric field originating from the polysilicon gate penetrates into the STI oxide and leads to fringing fields which act on the carriers in the channel (Fig. 3). The second effect which could be enhanced by positive charges at the STI/Si

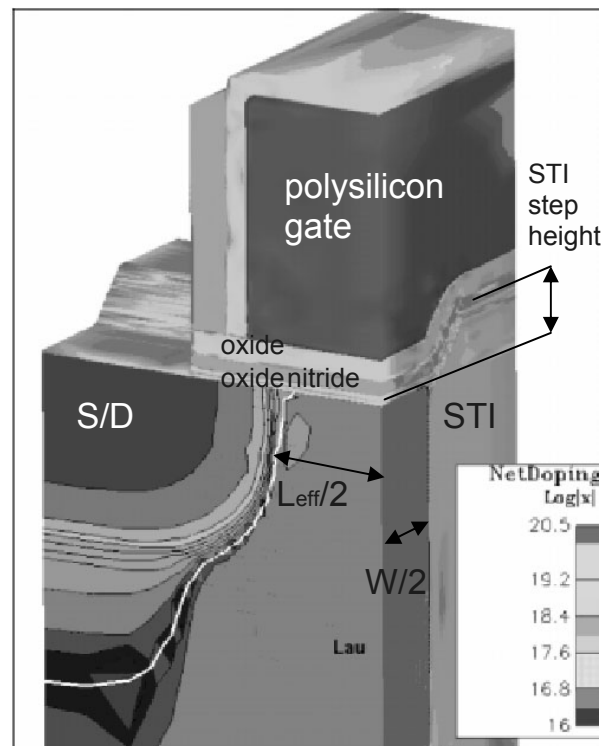


Fig. 1: STI-bounded nitride-storage memory cell (quarter device structure). Cells with different values of the STI step height (here positive) are investigated.

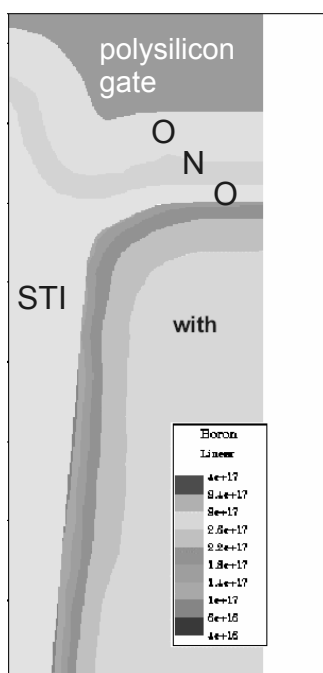


Fig. 2: Boron concentration (in cm^{-3}) in the channel (along width direction) according to 3D process simulation.

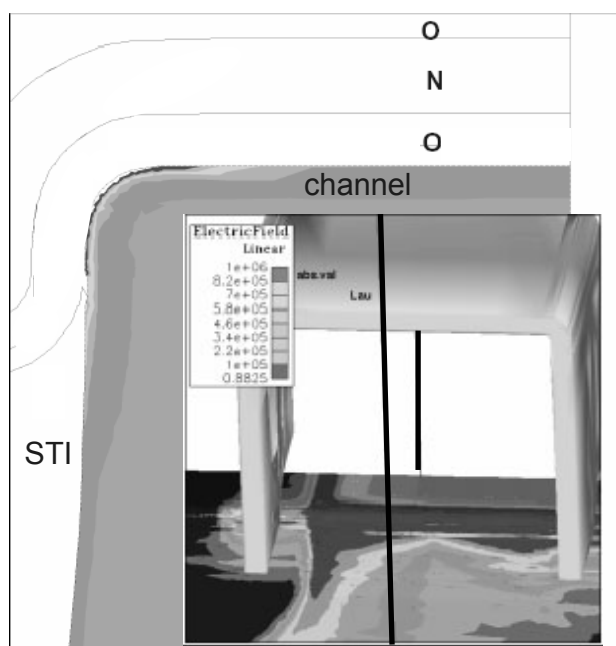


Fig. 3: Electric field (in V/cm) in the channel (along width direction) for open-channel conditions according to 3D device simulations. The inset shows the position of the cut plane.

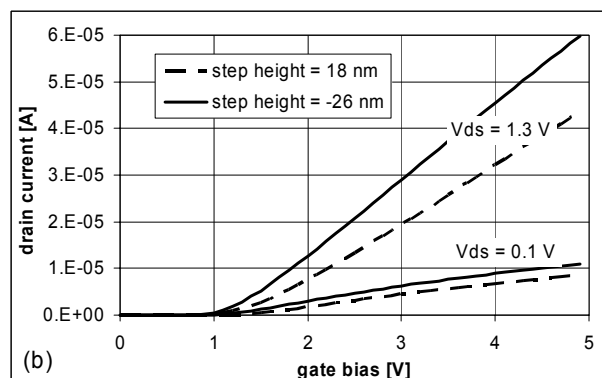
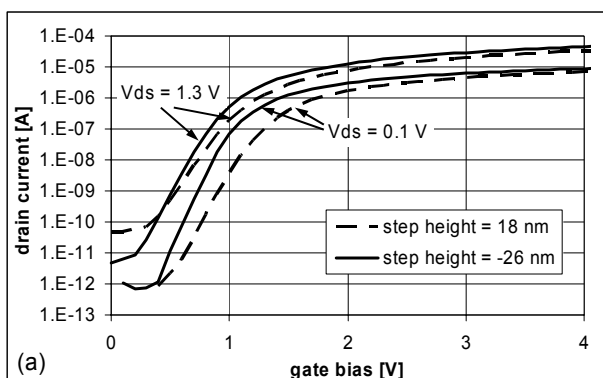


Fig. 4: Measured transfer characteristics for $V_{ds} = 0.1 \text{ V}$ and 1.3 V for a positive and negative STI step height in logarithmic scale (a) and linear scale (b).

3. Program performance

The program speed is not significantly affected by the STI step height (Fig. 5) even though the number of electrons in the channel is significantly reduced by increased step height due to degraded transconductance. 3D device simulations have been performed in order to quantify the sensitivity in terms of threshold voltage shift per injected charge in the nitride. (Fig. 6) Clearly, for a positive step height the threshold voltage shift is larger than for a negative step height for the same number of electrons in the nitride. Therefore, for the same threshold voltage shift more electrons have to be injected for a negative step height in order to suppress the channel current particularly at the STI edge. In this way, the larger channel and injection current for negative step height is balanced against the larger amount of necessary charge in the nitride. Therefore the program speed is approximately the same for negative and positive step heights.

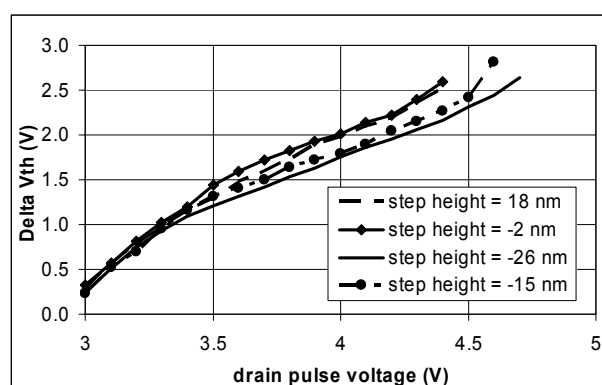


Fig. 5: Measured program characteristics for different values of STI step height. The gate bias is 9 V and the pulse length is 250 ns .

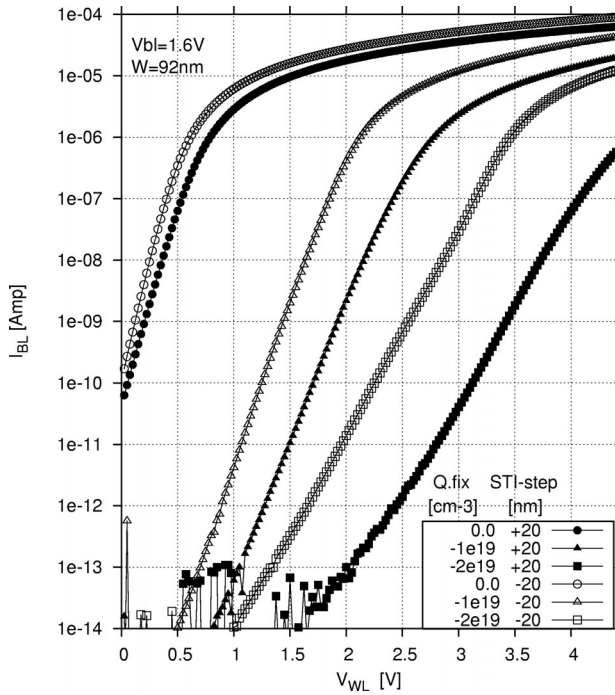


Fig. 6: Simulated transfer characteristics for negative and positive step heights with charge packets of ~ 250 electrons ($-1 \times 10^{19} \text{ cm}^{-3}$) and ~ 500 electrons ($-2 \times 10^{19} \text{ cm}^{-3}$) in the nitride.

4. Erase performance

In contrast to the programming behavior, the erase performance depends strongly on the STI step height (Fig. 7). For negative step height the erase speed is significantly improved. This is attributed to the influence of step height on the hole generation rate by band-to-band-tunneling (BBT). A common measure for this generation mechanism is the GIDL (gate-induced drain leakage) current. The measured GIDL curves as a function of negative gate bias are shown in Fig. 8 for different values of the STI step height. For a negative step height of -25 nm the initial GIDL current is approximately one order of magnitude higher (at a gate bias of -7 V) than for a positive step height of 18 nm . Fringing fields, which are pronounced for a negative step height, enhance the band-to-band tunneling rate. In order to explain the difference in the erase speed, the GIDL curves for the programmed rather than for the initial state must be considered, because the erase starts from a high- V_{th} -state and the injected electron charge has a large influence on the BBT mechanism. This can be recognized by the shift of the GIDL curves for the programmed state with respect to the initial state along the gate axis of approximately 6 V . The difference between the GIDL currents for negative and positive step heights is also maintained for the programmed state and therefore explains the difference in erase speed. In addition, for negative step height a higher hole injection probability due to the enhanced field at the STI corner could help to improve the erase speed.

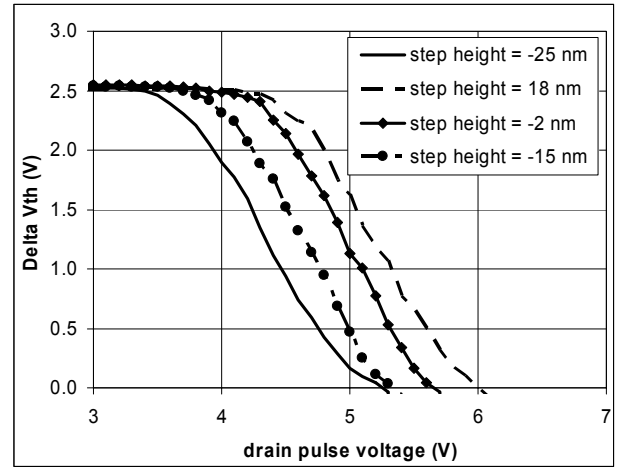


Fig. 7: Measured erase speed for different values of the STI step height. The gate bias is -7 V and the erase pulse length is $100 \mu\text{s}$.

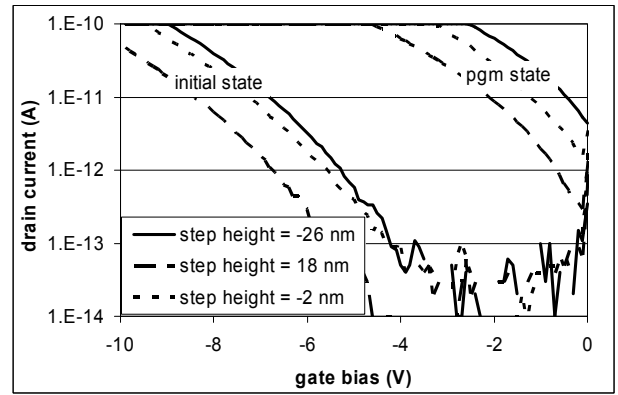


Fig. 8: Measured GIDL curves for initial and programmed state for different values of STI step height. The drain bias is 1.6 V and the source is floating.

5. Endurance

The improvement of erase performance for negative step height translates into improved endurance, because the erase performance degrades with increasing number of program and erase cycles (Fig. 9), whereas the program performance improves (not shown here). As shown in Fig. 9, the number of erase pulses increases with increasing number of cycles. Comparing the number of erase steps for negative and positive step height, the number of erase pulses after 10k cycles for a step height of -25 nm is approximately the same as for a step height of 18 nm after 5 cycles. Extrapolating this result, at least one order of magnitude improvement in the maximum number of cycles (i.e. the endurance) can be expected for a negative step height compared to positive step height.

6. Conclusions

For STI-bounded nitride-storage flash memory cells a negative STI step height significantly reduces short-channel effects, improves the gate-control and considerably increases the erase speed due to stronger fringing fields. This improvement of erase performance translates into improved endurance. Erase performance is also critical with regard to future technology nodes since by downscaling of channel length the erase performance degrades significantly.

Acknowledgements

The authors would like to thank all project members of the Infineon & Saifun development team for support and valuable discussions. This work was financially supported by the Federal Ministry of Education and Research of the Federal Republic of Germany (Project No. 01M 3160).

References

- [1] B. Eitan et al, IEEE Electron Device Lett., vol 21, pp. 543-545, Nov. 2000.
- [2] J. Willer et al, Proc. of 2004 Symposium of VLSI Technology, pp 76.
- [3] C. H. Wang et al, IEEE Trans. on Electron Devices, vol 46, pp. 139-144, Jan. 1999.
- [4] M. Specht et al, IEEE Electron Device Lett., vol 25, pp. 810-812, Dec. 2004.

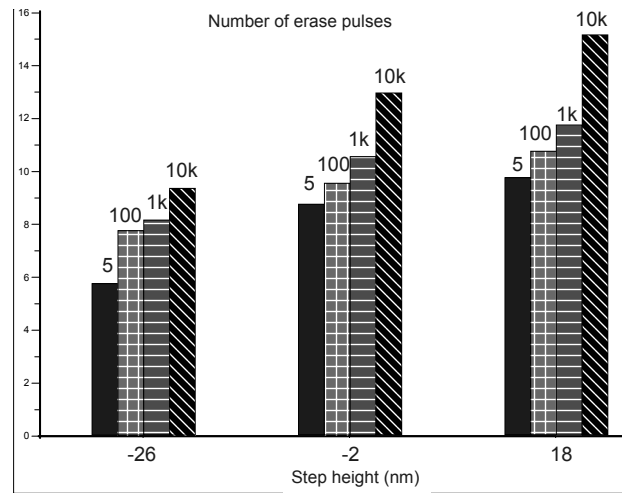


Fig. 9: Measured erase performance of a memory unit of 2000 bits vs number of cycles for different STI step heights.

Low Temperature Postmetallisation Annealing of High-k Dielectrics

Y.F.Loo^a, S.Taylor^b, P.Taechakumput^b, A.C.Jones^a, P.R.Chalker^c and L.M.Smith^d

^a Department of Chemistry, University of Liverpool, UK, L69 7ZD, y.f.loo@liverpool.ac.uk

^b Department of Electrical Engineering, University of Liverpool, UK, L69 3GJ

^c Department of Materials Science and Engineering, University of Liverpool, UK, L69 3BX

^d Epichem Oxides and Nitrides Division, Power Road, Bromborough, Wirral, UK, CH62 3QF

Abstract

The objective of this work is to compare the postmetallisation annealing properties of Hafnium dioxide (HfO₂) deposited from (I) Hf(OBu^t)₂(dmop)₂ and (II) Hf(OBu^t)₂(mmp)₂; and zirconium dioxide (ZrO₂) from (III) Zr(OBu^t)₂(dmop)₂ and (IV) Zr(OBu^t)₂(mmp)₂. Capacitance-Voltage (C-V) measurements were conducted on thin films deposited from (I), (II), (III) and (IV), before and after postmetallisation annealing in forming gas at 350°C. The annealing process has improved the properties of (I) and (II) but deteriorate the properties of (III) and (IV). Accurate permittivity values of the oxides are also obtained, which are very close to ideal k-value. This is supported by the more accurate thickness measurements using the Transmission Electron Microscopy (TEM).

1. Introduction

SiO₂ technology has reached its limit as the drastic scaling with the dimension of the gate dielectrics down to sub-0.1µm. This has motivated researchers to investigate high-k materials as a substitute for SiO₂ [1].

The metal oxide chemical vapour deposition (MOCVD) of HfO₂ and ZrO₂ thin films by using the mononuclear and novel complexes (I) Hf(OBu^t)₂(dmop)₂, (II) Hf(OBu^t)₂(mmp)₂, (III) Zr(OBu^t)₂(dmop)₂ and (IV) Zr(OBu^t)₂(mmp)₂ has been proven successful over a wide range of temperatures and film thicknesses (5 – 300nm) [2,3]. HfO₂ and ZrO₂ are both potential high-k dielectrics with k-value around 25 [4]. Therefore, we now report the electronic properties of these thin films.

2. Theory

Definition of permittivity:

$$\varepsilon = k\varepsilon_0 \quad (1)$$

The product $k\varepsilon_0$ is called the permittivity of the dielectric, ε_0 is the permittivity of free space and k is the dielectric constant [5]. The capacitance in the presence of a dielectric is given by:

$$C = k\varepsilon_0 \frac{A}{d} \quad (2)$$

where A is the contact area and d is the oxide thickness.

Flatband capacitance for silicon:

$$C_{FBSi} = \frac{\varepsilon_0 \varepsilon_s}{\lambda_p} \quad (3)$$

where ε_s is the dielectric constant of silicon and λ_p is the extrinsic Debye length at room temperature given by:

$$\lambda_p = \sqrt{\frac{(\varepsilon_0 \varepsilon_s 0.025)}{qN_D}} \quad (4)$$

where q is the electronic charge, 1.60×10^{-19} C and N_D is the doping density for silicon, 1×10^{15} cm⁻³.

Flatband capacitance of the metal oxide semiconductor (MOSC):

$$C_{FBMOSC} = \frac{1}{\frac{1}{C_{OX}} + \frac{1}{C_{FBSi}}} \quad (5)$$

where C_{ox} is the thin film oxide capacitance. The flatband voltage shift (δV_{FB}) value corresponds to the value of C_{FBMOSC} .

Oxide fixed charge, Q_f may then be calculated by the following:

$$\delta V_{FB} = \phi_{ms} - \frac{Q_f}{C_{ox}} \quad (6)$$

where ϕ_{ms} is the aluminium – silicon (Al – Si) work function difference is -0.35 [6]. This relationship assumes that the work function for Al deposited on HfO₂ and ZrO₂ is the same in both cases. For the conduction band offset to Si (ΔEC), we used the value calculated by Robertson, i.e. 1.5eV for HfO₂ and 1.4eV for ZrO₂ [7].

CV hysteresis, $\delta|V_H|$ is quantified as follows:

$$\delta|V_H| = |V_{FB}^- - V_{FB}^+| \quad (7)$$

where V_{FB}^- represent flatband voltage shift from negative to positive bias and V_{FB}^+ represent flatband voltage shift from positive to negative bias.

3. Procedure

C-V measurements were conducted by using HP 4192 impedance analyser. The measurement frequency used is 10 kHz and the probe signal is 30mV RMS. All the samples were deposited by liquid injection MOCVD technique at 450°C. Samples (I) and (III) were cut into 4. This produces samples Hf(a), Hf(b), Hf(c) and Hf(d) from (I); and samples Zr(a), Zr(b), Zr(c) and Zr(d) from (III). This enables the annealing of 6 minutes on Hf(a) and Zr(a); 12 minutes on Hf(b) and Zr(b); 18 minutes on Hf(c) and Zr(c) and 24 minutes on Hf(d) and Zr(d). Annealing is performed in a furnace with forming gas (10% hydrogen and 90% nitrogen mixture gases) flowing at a rate of 1.5 litres/minutes at a temperature of 350°C.

Cross section TEM samples were prepared by hand-grinding to $\approx 40\mu\text{m}$ and final thinning by using Gatan Precision Ion Polishing System (PIPS). TEM is conducted by using JEOL 2000FX.

4. Results and discussion

The TEM results show that the thickness of HfO_2 (deposited from (I)) thin film range from 276 – 317nm and the thickness of ZrO_2 (deposited from (III)) thin film range from 222 – 244nm. Fig. 1 and 2 show one of the cross section TEM micrographs of (I) and (III) respectively. Columnar growth can be observed from these micrographs, which corresponds to the SEM results obtained previously [2]. Fig. 2 also shows that the ZrO_2 thin film is slightly more crystalline compared to HfO_2 in Fig. 1. This can be confirmed by the XRD measurement of both (I) and (III) as shown in Fig. 3. The full width half maximum of (I) and (III) is $\sim 0.96^\circ$ and $\sim 0.51^\circ$ respectively. From the C-V measurements obtained previously [2] and by using equation (2) along with the TEM results, k-values could be calculated. The k-value for HfO_2 range from 22 to 25 and k-value for ZrO_2 range from 21 to 23. These values are very close to the ideal k-value of 25 for these oxides.

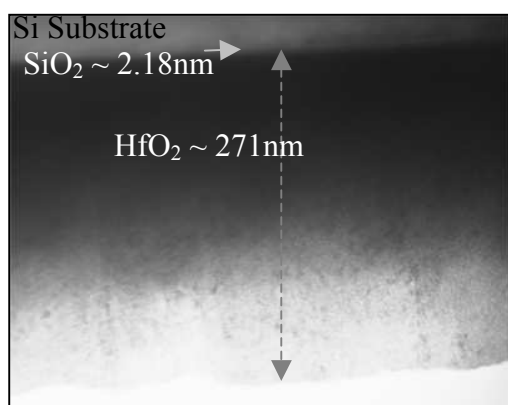


Fig. 1: Cross section TEM of HfO_2 thin film.

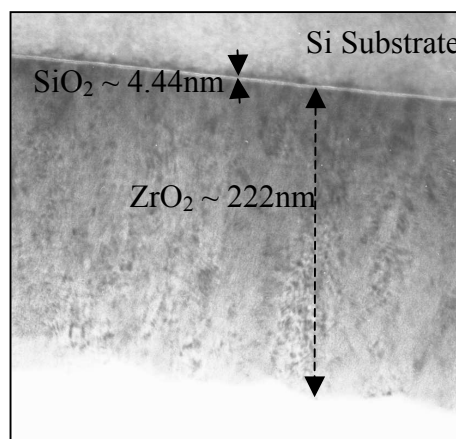


Fig. 2: Cross section TEM of ZrO_2 thin film.

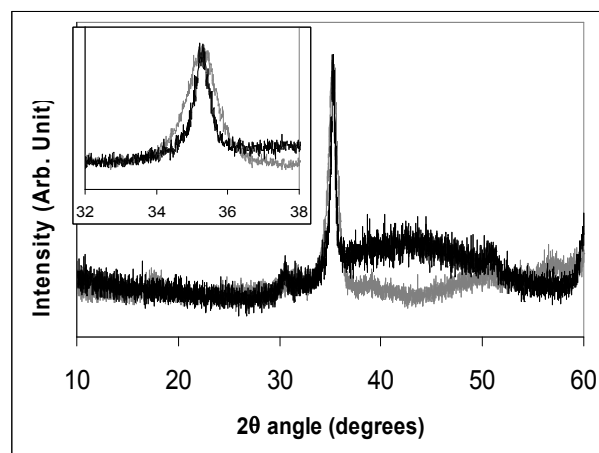


Fig. 3: XRD of (I) indicated by gray colour trace and (III) by black colour trace. Insert shows a magnified highly crystalline peak.

C-V measurements before annealing are shown in Fig. 4 for HfO_2 and in Fig. 5 for ZrO_2 . Hysteresis is in the clockwise sense for both films indicating electron trapping in the bulk of the dielectric over the timescale of the CV measurement. In general the level of hysteresis for the HfO_2 is lower than for the ZrO_2 film indicating a lower level of electron traps in the as deposited films. In Fig. 5 the inversion region is at the same level of capacitance as the accumulation region during negative to positive sweep due to the response of the minority carriers in the silicon substrate. This may be due to some Zr atom diffusion into the underlying Si substrate creating generation centres in the silicon and causing enhanced generation of minority carriers. This may be due to the affinity between Zr and Si or due to the greater mobility of Zr in Si compared to Hf.

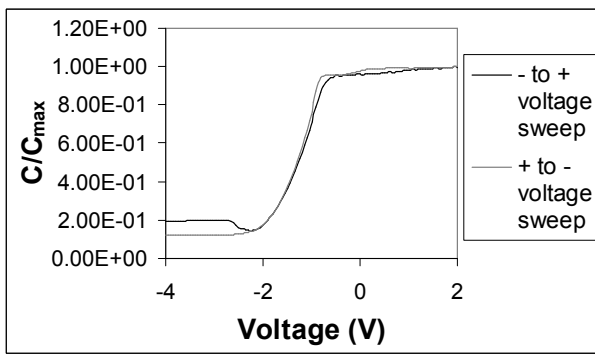


Fig. 4: C-V measurements for Hf(c) thin film.

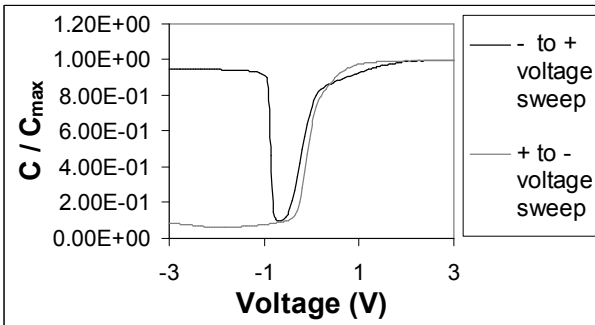


Fig. 5: C-V measurements for Zr(a) thin film.

By using eq. (3), (4) and (5) with C-V measurements before and after annealing, δV_{FB} could be found and shown in graphical format in Fig. 6. Annealing in forming gas for HfO₂ (I) always reduces oxide fixed charge from a positive value to a less positive value (confirmed using eq.(6)), with the optimum result achieved between 18 to 24 mins annealing time. Thin film HfO₂ (II) also shows the same behaviour up to 18mins annealing time. Therefore, annealing has improved the δV_{FB} closer to ideal for (I) and (II). However, annealing in forming gas for ZrO₂ (III) always increases oxide fixed charge from a negative value to a positive value (again confirmed by using eq. (6)) (see Fig. 6). Thin film ZrO₂ (IV) shows improvement up to 12mins annealing time, then deteriorates at 18mins annealing time and then shows improvement again at longer annealing time of 24mins. The consistent improvement for HfO₂ and deterioration for ZrO₂ could be due to the higher crystallinity in the ZrO₂ sample as described previously. It was also found that the leakage current for ZrO₂ was significantly higher than for HfO₂ films.

Annealing in forming gas generally increases δV_H for HfO₂ (I), ZrO₂ (III) and ZrO₂ (IV) except for HfO₂ (II). By using eq. (7), δV_H could be calculated and shown in graphical format in Fig. 7. The hysteresis is in the clockwise direction indicating electron trapping in electronic states at the Si-dielectric interface and / or in the bulk of the dielectric film. Annealing decreases hysteresis for HfO₂ (II) between 6mins to 18mins annealing time. This shows a decrease in electron trapping for that annealing time. After 18minutes, further annealing causes increased δV_H . This could be due to generation of electron states by prolonged exposure to the active annealing species. The active annealing

species in a post-metallisation forming gas anneal are generally considered to be H⁺ ions (protons).

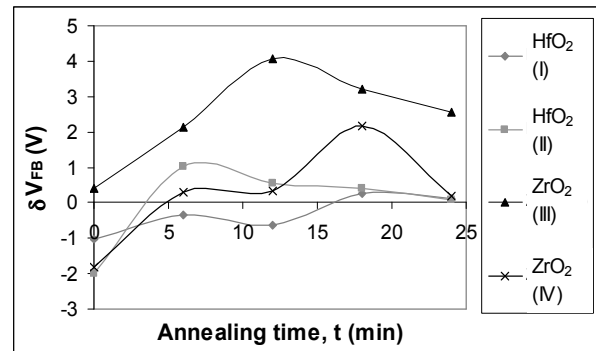


Fig. 6: δV_{FB} versus annealing time for thin films (I), (II), (III) and (IV). (II) and (IV) extracted from Zoofakar [8].

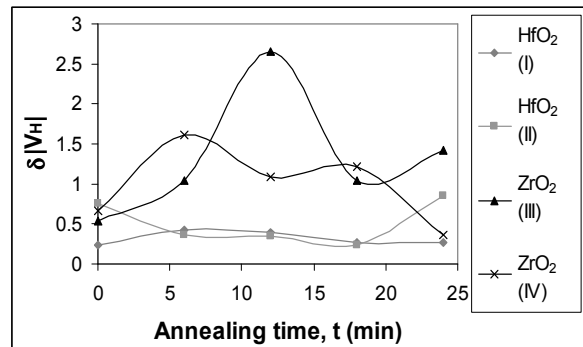


Fig. 7: δV_H versus annealing time for thin films (I), (II), (III) and (IV). (II) and (IV) extracted from Zoofakar [8].

5. Conclusion

Annealing in forming gas (H₂/N₂) always reduces oxide fixed charge from a positive value to a less positive value for both HfO₂ thin films, (I) and (II) deposited from Hf(OBu^t)₂(dmop)₂ and Hf(OBu^t)₂(mmp)₂ precursors, respectively. Hence, δV_{FB} values in both cases are closer to ideal with annealing than without.

Annealing in forming gas (H₂/N₂) always increases oxide fixed charge from a negative value to a positive value for ZrO₂ thin film (III) deposited from Zr(OBu^t)₂(dmop)₂ precursor. Therefore, δV_{FB} values are further away from ideal with annealing than without. δV_{FB} values for ZrO₂ (IV) deposited from Zr(OBu^t)₂(mmp)₂ precursor are closer to ideal between 6 to 12 mins annealing and move further from ideal at 18mins annealing and move closer to the ideal again at 24mins annealing.

Accurate k-values were obtained for both HfO₂ and ZrO₂ with the accurate thickness obtained from X-TEM. HfO₂ has k-value range from 22-25 and ZrO₂ has k-value range from 21-23. These values are very close to ideal k-value of 25. TEM micrographs also show a more crystalline film in the case of ZrO₂.

Acknowledgements

YFL and ST would like to thank A. Zoofakar for his M.Sc work and R. Murray for his assistance with the TEM microscope and R. Potter for XRD discussion.

References

- [1] G.D. Wilk, R.M. Wallace and J.M. Anthony, J. Appl. Phys., **89**, 5243 (2001).
- [2] Y.F. Loo, R. O’Kane, A.C. Jones, H.C. Aspinall, R.J. Potter, P.R. Chalker, J.F. Bickley, S. Taylor and L. M. Smith, J. Mater. Chem, submitted.
- [3] P.A. Williams, J.L. Roberts, A.C.Jones, P.R. Chalker, N.L. Tobin, J.F. Bickley, H.O. Davies, L.M. Smith and T.J. Leedham, Chem. Vap. Dep., **8**, 163 (2002).
- [4] “Identifying the Most promising High-k Gate Dielectrics” by R.M. Wallace and G.D. Wilk, Semiconductor International, Reed Elsevier Inc., July 2001.
- [5] “University Physics” by H.D. Young and R.A. Freedman, 10th Edition, Addison Wesley Longman, Inc., USA, ISBN 0-201-60322-5, 2000, pg. 785.
- [6] “MOS Physics and Technology” by E.H. Nicollian and J.R. Brews, John Wiley, New York, ISBN 0471085006, 1982, pg. 466.
- [7] J. Robertson, J. Vac. Sci. Technol. B, **18**, 1785 (2000).
- [8] “Investigation of Alternative Gate Dielectric Of Sub Micrometer MOSFET” by A.Zoolfakar, M.Sc thesis, University of Liverpool, 2002, Appendix 2 and 4.

EEPROM failure analysis methodology: can programmed charges be measured directly?

C. De Nardi^a, R. Desplats^a, P. Perdu^a, F. Beaudoin^a and J-L. Gauffier^b

^aLaboratoire d'Expertise commun CNES et TES, 18 avenue Edouard Belin, 31401 Toulouse cedex 9, France.

^bLNMO, INSA Génie Physique, 135 av de Rangueil, 31055 Toulouse cedex 4, France.

Abstract

A method to measure “on site” programmed charges in EEPROM devices is presented. Electrical AFM based techniques (EFM and SKPM) are used to probe directly floating gate potentials. Both preparation and probing methods are discussed. Sample preparation to access floating gate/oxide interfaces at a few nanometers distance without discharging the data reveals to be the key point, more than the probing technique itself.

1. Introduction

For space applications, memory devices are used to perform many applications, from storing the mission parameters to the communication algorithms. These components must ensure harsh qualification tests as reliability is a question of utmost importance.

In the case of Electrically Erasable Programmable Read Only Memory (EEPROM) or flash-EEPROM several tests are performed: i) endurance after a high number of write/erase cycles, ii) data retention after temperature or field ageing and iii) internal (mobile ions coming from device top layers) or external (irradiations, atmospheric particles) contamination sensibility [1]. All of these tests lead to a reduction of programmed charge level in EEPROM which may alter memory data reliability.

Usually, the current leakage trough the cell bit line and the storage transistor threshold voltage are monitored. Such tests highlight EEPROM charge fluctuation without providing a direct measurement of the charge density.

The objective of our paper is to determine if it is possible to directly measure the programmed charges (electrons) in such non volatile memories in order to differentiate two fully functional nominal bits programmed at opposite values “1b” and “0b”. The final goal is to differentiate a nominal “0b” binary state from a degraded “0b” one that suffered a previous known reliability test.

While an accurate calculation of the exact programmed charge number is complex, we target an order of magnitude of few hundreds of charges. The mobile nature of the charges is of course a problem. While the programming phenomena by Fowler-Nordheim tunneling takes advantage of it, it is also known that sources of energy may make the charges go away thus erasing the programming value. For example, temperature annealing (120°C) and UV illumination are two known techniques to reduce charge trapping. Sample preparation can also degrade the charges integrity if special care is not taken.

After a quick review of EEPROM programming principles, we will describe the descrambling technique that we used to locate physically programmed data in the memory table. Then we will present two sample preparation methodology

approaches: front side and backside. Finally, using a combination of Atomic Force Microscopy (AFM) based techniques we will tackle the task to reveal “0b” state from “1b” state.

2. Devices under test (DUT)

For this study, we have selected two similar NOR EEPROM with standard cell structures (see Fig. 1). The first one is a 128 kbits memory from ST Microelectronics. It is a 0.35µm CMOS technology organized in 16k×8bits which operates from 5.5V down to 2.5V. Its SEM cross-section view is presented in Fig. 2. The second one is a 64 kbits from Atmel, a 0.35µm CMOS technology organized in 8k×8bits with 5V power supply.

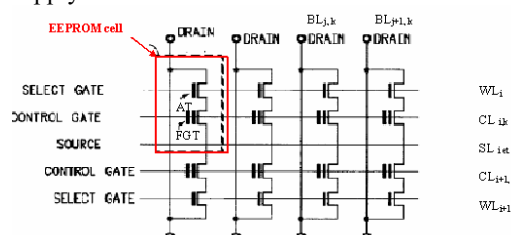


Fig. 1: EEPROM cell structure. It is composed of two NMOS transistors: the access transistor (AT) and the floating gate transistor (FGT). The FGT has two gates: a control gate (CG, poly 2) and a floating gate (FG, poly 1).

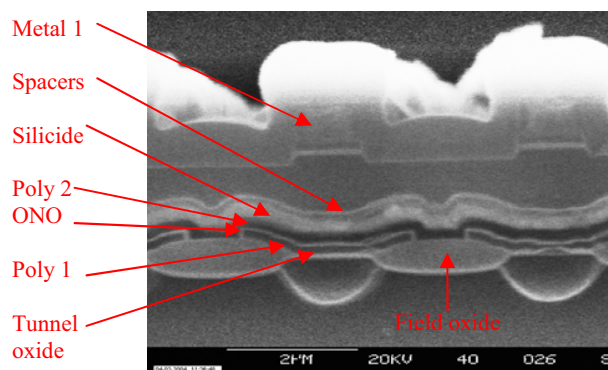


Fig. 2: SEM cross section parallel to word lines (poly 2) presenting the physical structure of the 128 kbits EEPROM. The floating gate is located at poly 1 level and isolated by an Oxide Nitride Oxide (ONO), nitride spacers and a 10-15 nm thin gate oxide.

In such nonvolatile device, the programmed data correspond to charge density accumulated in the floating gate (FG)/tunnel oxide interface. In this paper, EEPROM conventions proposed in [2] are used:

- An erased cell has electrons stored in its FG and presents a positive threshold voltage VT_1 . A programmed cell has

no electrons on its FG and presents a negative threshold voltage VT_2 .

- When a read voltage between VT_1 and VT_2 is applied to the CG, the erased cell remains OFF, which corresponds by convention to the logic binary state “0b” whereas the programmed cell is conductive and corresponds to the logic state “1b”.

3. Electrical descrambling

AFM based techniques are meant to be used on small areas (the maximum scan size is $100\mu\text{m}$ by $100\mu\text{m}$ because of limited piezoelectric x-y scanner displacements). Therefore, the area of interest must be located precisely before hand to match the electrical address with its physical location.

For EEPROM, we have chosen emission microscopy (EMMI) to locate physical address. The EMMI principle [3] is to collect photons in the near infrared spectrum, emitted by transistor during read or write access. Since the bulk silicon is transparent to these wavelengths, light emission can be collected through the backside, making this approach suited for high integrated technologies with a high number of metal levels. By looping a write “FOh” (“11110000b”) cycle on specific address, a high V_{pp} (12-20V) is applied to [2]:

- the FGT drain through the AT while programming a “1b” (programmed conductive state), the FGT control gate being grounded and its source being floating.
- the FGT control gate, while programming a “0b” (erased blocked state), the FGT drain being floating and its source being grounded.

In time, photons are collected and the address becomes visible (see Fig. 3). Programming in loop only one byte each time, the whole memory array organization is deduced. Since photons are detected only during “1b” programming, localization of bits in the same address itself is also achievable.

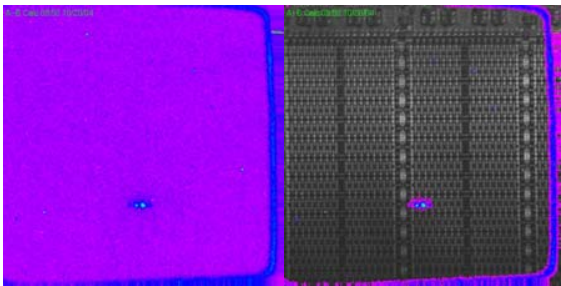


Fig. 3: EMMI image (left), LSM and EMMI superimpose image (right) obtained from the back side with HgCdTe camera with 10 min integration time. The byte which address is 1025d (401h) is programmed at “FOh” in loop. Only the four most significant bits programmed at “1b” emit photons. Views at 100X.

4. Sample preparation

Programmed charges in EEPROM are stored at the poly1/tunnel oxide interface. To make near-field electrical measurements, samples have to be deprocessed less than 100 nm distance from the interface. Indeed, metal-coated AFM tips have to probe near the floating gates to be sensitive to electrical charges without discharging them. Thus, two preparation approaches were investigated:

- Traditional front side layer by layer deprocessing, down to the ONO layer. Floating gates potentials will be measured through ONO.
- Relative new and promising backside approach, down to the tunnel oxide layer. Floating gates potentials will be measured through tunnel oxide.

It must be underlined that during whole sample preparation process, Reactive Ion Etching (RIE) is not recommended to avoid charging oxides and so disturbing real FG potentials. SEM inspections are performed on different samples from AFM ones for the same reasons.

Before hand, samples are dropped into HNO_3 fuming acid at 90°C during 4-5min to remove plastic packages [4] and neutralized with a standard cleaning procedure.

4.1 Front side deprocessing

The front side approach objective is to prepare samples down to ONO layer. It is achieved in two steps:

- 1- Buffer Oxide Etch (BOE = 1ml HF + 5ml NH_4F) to expose poly 2 totally free of silicide,
- 2- Choline hydroxide [5] [6] or Tetra Methyl Ammonium Hydroxide (TMAH) [7] [8] to remove poly 2 and expose ONO layer.

In our experiment, front side deprocessing was found to be destructive with respect to programmed charge. Indeed it is difficult to obtain a perfect ONO level exposed with poly 1 gates still isolated. The issue comes from the BOE first etch step. On one hand it must be long enough (10-12 min) to completely remove WSi_2 silicide and to expose poly 2. On the other hand, the more the BOE etching time is increased, the more the spacers are weakened and do not protect poly 1 anymore. As a consequence, some floating gates were over-etched by the following choline (or TMAH) treatment, and charges vanished.

4.2 Back side preparation

The backside approach is particularly promising for detection of gate oxide and tunnel oxide holes and defects like micro-cracks and pinholes [9]. Our back side approach objective is to prepare the sample up to the tunnel oxide layer. We used a procedure similar to ones developed in [7] and [9], which can be resumed as follows:

- 1- Epoxy gluing,
- 2- Bulk silicon removal by parallel lapping,
- 3- Selective chemical etch of remaining silicon by choline hydroxide or TMAH,
- 4- SEM inspections or electrical AFM measurements.

Step 1) First of all, we had to glue the sample front side on a conductive holder in order to manipulate it during the whole process. Indeed, the chip thickness after preparation will be less than $4\mu\text{m}$. Thus, it is glued on a metal AFM disk with EP110 epoxy from Allied, which is found to be resistant to TMAH and choline hydroxide treatments. We cured the epoxy during half an hour at 120°C .

Step 2) The sample, composed of the chip on the AFM disk, is waxed on the polishing holder. It is now ready to be thinned by parallel lapping from the die backside. A $15\mu\text{m}$ diamond film is used to thin the sample until its thickness is around

30 μ m. Then, 6 μ m diamond slurry and finally 1 μ m diamond slurry are used to make a mirror polish. Indeed, cracks remaining on backside surface after polishing step will lead to chemical etching rate non-uniformity. As shown in Fig. 4, choline hydroxide (or TMAH) etches preferentially silicon scratches. After the final polishing step, the sample is cleaned with Q-tip soaked into a mixture of DI water and soap. Its thickness is around 25 μ m and can be measured by laser reflectometry. Then the sample is removed from the polishing puck by melting away the wax.

Step 3) To remove the remaining silicon and leave the tunnel oxide layer untouched, a solution with a strong silicon to oxide etching selectivity and a low etching rate is needed. Two products commonly used in Failure Analysis were tried: Choline hydroxide solution in water [5] [6] and TMAH solution in methanol [7] [8]. For the first choice, we used 50% choline hydroxide solution in water around 90°C. According to [5], in these conditions, etch rate for bulk silicon is around 400 nm/min and 0.08 nm/min for oxide, so silicon to oxide selectivity is around 5000. For the second choice, we used 80ml of 25% TMAH solution in methanol mixed with 20ml of isopropanol at 85°C. According to [7], etch rate of bulk silicon is around 350nm and 0.1nm/min for oxide, so a selectivity of 3500.

We found that both choline hydroxide and TMAH are suitable for the backside approach and leave the tunnel oxide and field oxide intact on 95% of the chip surface. Before choline or TMAH etch, we recommend a 5s BOE dip to remove the native oxide. The sample is inspected after a first 40min choline or TMAH treatment. Subsequent etches of 10min each are used if the silicon is not completely removed. Usually it takes between 60min and 90min to have a surface free of silicon for both solutions.



Fig. 4: Choline hydroxide etching progress at t=0min (left) and t=40min (right). Optical views at 2.5X.

Step 4) The high viscosity of choline hydroxide, and on a less extent TMAH, created rinsing and cleaning problems. After a long DI water cleaning, we recommended successive 2min ultrasonic baths with PRS1000, acetone and DI water. A key point must be underlined: silicon or choline/TMAH residues obstruct charge measurements by AFM electrical modes at floating gate interfaces. Indeed, it is known that all electrical techniques are very dependant of the sample surface. The tunnel oxide surface must be free of screen charges, free of contaminations and perfectly clean. In addition, the smoother the surface is, the better the measurements are. Fig. 5 presents the surface state after backside preparation.

To conclude, the presented backside methodology allows exposing the tunnel oxide level perfectly intact with floating gates still isolated.

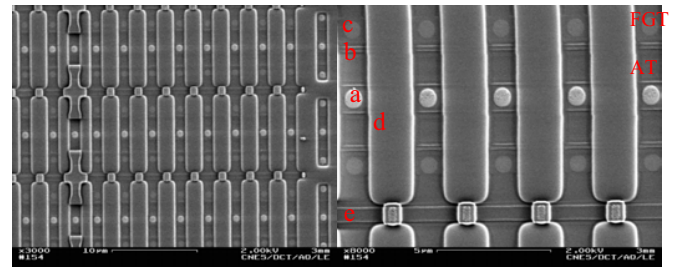


Fig. 5: SEM view of tunnel oxide level from backside, showing drain contacts (a), spacers (b), tunnel oxides (c), field oxides (d) and source lines (e) of the 128 kbits EEPROM sample.

5. EEPROM charge measurement

The objective is to directly measure EEPROM programmed charges at their location, i.e. at the floating gate /tunnel oxide interface. Electrical modes of AFM [10] seem to be promising because they are very sensitive to electrostatic interactions and provide a nanoscale spatial resolution, especially:

- Electrostatic Force Microscopy (EFM) which measures variations of electric field gradient between tip and sample with 45mV resolution on 100nm thin metal lines [11], [12].
- Scanning Kelvin Probe Microscopy (SKPM) which is dedicated to probe sample surface potential down to 10mV resolution [13].

Both techniques use two scans [14]. The first scan records topography in AFM tapping mode, whereas the second scan is achieved at a 10nm to 100nm lift distance from the surface in order to be sensitive to electric field. During the EFM second scan, the cantilever experiments mechanical oscillations imposed by piezoelectric materials and a DC voltage is applied to the tip and/or the sample. The SKPM second scan is performed with no imposed mechanical oscillations. Applying an AC voltage to the tip, it experiments smaller oscillations only due to Coulomb force interactions.

5.1 Front side measurements through ONO layer

As explained in the previous section, ONO, spacers and field between cells are deteriorated by front side deprocessing. As a consequence, floating gates are no longer isolated and partially etched: the programmed charges vanished during front side deprocessing. Therefore the backside approach is chosen.

5.2 Backside measurements through tunnel oxide layer

Tapping AFM, EFM and SKPM scans are performed on tunnel oxide level from the sample backside on 64 kbits and 128 kbits samples. AFM tapping measurements demonstrated the perfect FGT tunnel oxide integrity; no holes are detected (see Fig. 6 left). SKPM image on tunnel oxide surface (see Fig. 6 right) shows a uniform potential proving that its surface is free of electrical contaminations (screen charges...) which could interfere with the measurement of programmed charged stored deeper. Probing 40-50nm deep, SKPM image (not shown) demonstrates the perfect integrity of floating gates, spacers and field oxides.

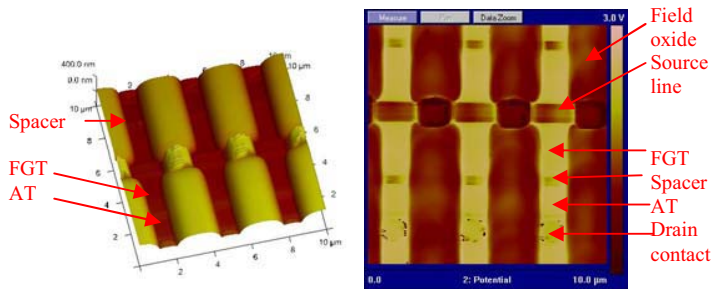


Fig. 6: (left) Tapping AFM scan on tunnel oxide level showing topography information obtained from backside. Thick field oxide has bump bevel shape. No holes are detected on gate oxides. (right) SKPM scan (less than 5 nm deep) showing uniform potential of oxide surface and spacer integrity. No electric contaminations such as screen charges are detected (128 kbits memory).

To image charges in tunnel oxide/FGT floating gate interfaces, we have to find optimal SKPM parameters (VDC tip, VAC tip, VDC sample and lift distance) which allow to probe at the right depth with maximum electrical sensitivity. In addition, working on floating structure directly through oxide is a complex problem to stabilize SKPM signals. It is different from measuring stable passive voltage contrast on biased metal lines as presented in [11] and [12] or biased PN junctions as shown in [13]. Indeed, the working point, where tip and floating sample are in equilibrium, has to be found.

Fig. 7 shows topographic amplitude AFM (left) and electrical SKPM (right) scans showing charged floating gate and uncharged ones on the 64 kbits memory sample. We have to notice that topography is exactly the same for both types of gate, proving that SKPM contrast only comes from electrostatic charges stored in tunnel oxide/FG interface. The tip being at positive DC voltage, attractive electrostatic forces between tip and electrons stored on floating gates occur, resulting in a high potential signal. Thus, charged gates appear bright on SKPM images. Inversely, while scanning gates empty of electrons, the positive tip experiments low repulsive forces resulting in low dark contrast in SKPM images.

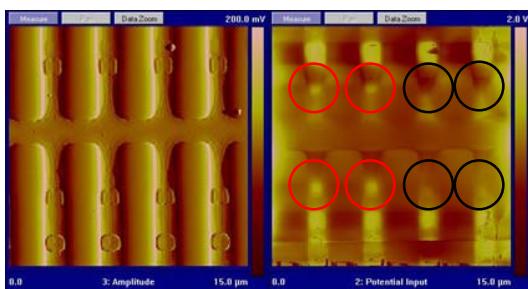


Fig. 7: Amplitude AFM (left) and SKPM (right) scans on "FOh" ("11110000b") programmed byte area. SKPM detected two charged FGT gates (bright in red circles) and two uncharged ones (dark in black circles) as programmed in each byte center (64 kbits memory)

Knowing on one hand the data we have programmed into the table memory ("FOh") and on the other hand the physical location of each byte address obtained by descrambling techniques, we confirmed our SKPM recordings.

EFM measurements did not yield conclusive results possibly because it offers less sensitivity and probes deeper than SKPM. Further investigations are underway.

5.3 Discussion

The SKPM results were very encouraging. It validated that our backside sample preparation methodology did not remove all programmed charges. It also validated the AFM electrical mode ability to measure charges "on site". But as SKPM scans are conducted in open loop, we have only qualitative and relative data (see Fig. 8). Indeed, it is only by comparison of two different levels of charges appearing in the same SKPM scan (100µm by 100µm maximum size) that we can conclude to orient the failure analysis process.

The measured difference between nominal "0b" and "1b" EEPROM programmed cells is around 450mV. Knowing that we can achieve a resolution up to 10mV, we can hope to distinguish nominal "0b" from degraded "0b".

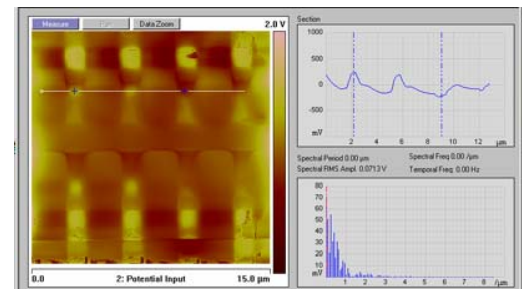


Fig. 8: Relative SKPM measurements between charged FGT gates (bright) and uncharged ones (dark). SKPM difference detected between them $\approx$ 450 mV as shown (64 kbits memory)

6. Conclusion

With the goal of measuring "on site" programmed charges of EEPROM devices, we have presented two sample preparation methodologies: front side and backside. Only the backside approach has been shown to maintain ONO, spacers and tunnel oxide integrity and floating gates isolation. Despite the physical integrity of the structure, the possibility that charges may have vanished during the preparation process was still quite strong.

Using SKPM, an AFM electrical mode, charged gates and uncharged gates were differentiated, thus validating both sample backside methodology and probing technique. The charges did not vanish and were successfully detected.

In the case of two different charge levels, SKPM appears limited to nearby gates that could appear in the same image with a sufficient charge density difference to be detected.

References

- [1] Jan Van Houdt, IMEC, ESREF Tutorial (2000)
- [2] S.Zink et al, ST Microelectronics, Brevet FR2803080 (1999)
- [3] P.Sardin et al, ST Microelectronics, ANADEF proceedings (2004).
- [4] P.Perdu et al, CNES, ESREF proceedings (2000)
- [5] V. Korchnoy, Intel, ISTFA proceedings (2002)
- [6] J. Higgins, WaferTech L.L.C., ISTFA proceedings (2001)
- [7] F. Zachariasse, Philips, ISTFA proceedings (2001)
- [8] M. Morris et al., ON Semiconductor, ISTFA proceedings (2000)
- [9] K. Upadhyaya et al., Intel, ISTFA proceedings (2003)
- [10] A.N.Erickson, Multiprobe Inc. ISTFA tutorial (2001)
- [11] U.Behnke et al, Duisburg University, ESREF Proceedings (1999)
- [12] W.Martin et al, Duisburg University, ISTFA proceedings (2000)
- [13] A.Doukkali et al, Laboratoire de Microélectronique ISMRA Philips (LAMIP), Applied Surface Sciences, (March 2004)
- [14] F.M.Sherry et al, Digital Instruments Veeco Metrology Group, Digital Instruments application notes AN 27, www.veeco.com.

Flash Memory Cell: An Automated Diagnosis Tool for Geometric Failures

B.Saillet^b, J.M. Portal^a, D. Née^b

^a L2MP-Polytech, IMT Technopôle de Château Gombert, 13451 Marseille, France, portal@polytech.univ-mrs.fr

^b ST-Microelectronics, Z.I de Rousset, BP2, 13106 Rousset, France

Abstract

The objective of this paper is to present an automated geometric failure diagnosis tool for Flash cell. This tool focuses on speeding up the diagnosis process of geometric failures. It is based on a mathematical model generated with a "Design Of Simulation" (DOS) technique. The DOS technique takes as input, simulations results of a Flash cell with different geometries and produces, as output, a polynomial equation of the threshold voltages. The diagnosis process is performed in generating all cell geometries that match measured threshold voltages. From this set of cell geometries, the distribution of each geometric parameter is generated and compared with the specification window to diagnose any process variations.

1. Introduction

The evolution of the market of portable devices such as smart card or mobile communication media has a straight impact on the use of Flash memories. Indeed, Flash have become over the last few years very relevant choices for any application requiring non-volatile semiconductor memory and represent 50% of the NVM market [1]. The core element of the Flash memory cell is a Floating Gate Transistor (FGT). The threshold voltage of a FGT can be shifted repetitively from a high to a low state corresponding to the two logical values of the memory cell. The V_T of a FGT is shifted when the cell is programmed by a write or an erase operation. When a negative charge is stored on the floating gate node, the V_T increases, the cell is then defined as written. Respectively, when the charge on the floating gate node is positive, the V_T decreases, the cell is then defined as erased. In most of the Flash actually on the market, the charge transfer from and into the floating gate is based on two main mechanisms. The hot-electron injection (HEI) mechanism is used for the write operation. The HEI mechanism involves a lateral electric field to "heat" the electron and a transversal one to inject the carriers through the tunnel oxide [1]. The Fowler-Nordheim (FN) tunneling mechanism is used for the erase operation. The FN mechanism injects carriers in the floating gate when a high electric field is applied through a thin oxide [2]. These both tunneling mechanisms are completely dependant of the geometry of the cell [1], inducing critical problems of process control.

Failures in the NVM have been well investigated. In [3] Cappelletti and al. present a degradation mechanism due to drain stress during write/erase

cycling with emphasis on the critical impact of the device dimensions on the degradation mechanism. In [4] an accelerated current test method is proposed to evaluate the tunnel oxide. Charge trapped in the tunnel oxide is also a major concern for the reliability of the Flash, in [5] a test structure is proposed to quantify the trap density. In [6] a systematic test methodology is introduced to identify oxide defect related failure mechanism. From all this study, it appears that the geometry of the cell is a major concern to check the reliability of the Flash memory cell. Indeed, any small variations of any geometric parameters among the cells in a memory array leads directly to a spread threshold voltage distribution, limiting the overall performance of the memory. Moreover, the sensitivity of Flash memories to process variations has a straight impact on yield improvement. Knowing that, yield is a key factor to determine the success of a fabrication unit, product specific as well as general speed up techniques for process monitoring are absolutely necessary.

The fundamental discussion of this paper is to present a tool developed to speed up the diagnosis process of geometric failures in the Flash memory cell. From the measurement of the threshold voltages during the Parametric Test (PT), this tool generates the potential distribution of each geometric parameter and compares them to the specification window in order to diagnose any process drift. The paper is organized as follows. Section 2 presents electrical and polynomial models of the floating gate transistor used to simulate the Flash cell. The section 3 is dedicated to the diagnosis process based on the use of the presented model. The efficiency of the tool is studied with simulations and silicon example. Finally, Section 4 gives some concluding remarks and work perspectives.

2. Flash cell model

2.1. Electrical model

A Flash model is developed based on a Mos Model 9 description [8]. This MM9 based Flash model is a compact charge sheet model written in HDLA [7]. This model takes into account static and dynamic behaviours. The approach used to develop this model is based on a charge neutrality formulation coupled with a MOS Model 9, as presented in [9]. In this approach, the charge neutrality, including the charge stored in the floating-gate, is applied to determine the floating gate potential from which all the variables can be calculated in the MM9 model formulation as well as in the tunneling current equations. The MM9 is the intrinsic transistor that appears in the model between the drain node, the source node and the floating gate node, with

all the potentials referenced to the bulk. The MM9 model card is given by ST-Microelectronics and results from characterization performed on dummy cell (Flash cell with floating gate connected to the control gate). The Flash specific geometric parameters are extracted with physical characterizations using scanning electron microscopy (SEM) and electrical characterization on specific test structures. This electrical model is used to extract the three threshold voltages (V_{TUV} , V_{TER} and V_{TWR}). To do so, five simulations have to be performed. The details of each one of these simulations are given below:

- a static simulation of the cell is performed with no charge on the floating gate ($Q_{FG} = 0$ C) to extract the V_{TUV} ,
- a dynamic simulation of the cell is performed using the write signals to extract the injected charge on the floating gate ($Q_{FG} = Q_{FG-WR}$),
- a static simulation of the cell is performed with Q_{FG-WR} charge on the floating gate to extract V_{TWR} ,
- a dynamic simulation of the cell is performed using the erase signals to extract the injected charge on the floating gate ($Q_{FG} = Q_{FG-ER}$),
- a static simulation of the cell is performed with Q_{FG-ER} charge on the floating gate to extract V_{TER} .

2.2. Polynomial model

The aim of a mathematical model is to link the three threshold voltages of the cell (V_{TUV} , V_{TER} and V_{TWR}) to the geometric parameters. To do so, a given number of electrical simulations using the Flash cell electrical model are necessary. First of all, the geometric parameters of the Flash cell have to be identified. The Fig.1 presents the Flash cell structure with the location of the geometric parameters included in our polynomial model. W_{pp} is the inter-polysilicon oxide width, W is the channel width, L is the channel length, L_{pp} is the inter-polysilicon oxide length, T_{ONO} is the inter-polysilicon oxide thickness and T_{OX} is the tunnel oxide thickness.

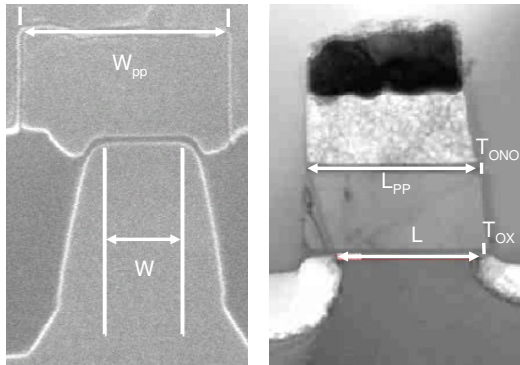


Fig. 1: Flash cell structure overview with location of the geometric parameters.

To generate the polynomial model of Flash cell, a “Design Of Experiment - DOE” technique is used. Our experiments being a set of simulations, this technique

is here called “Design Of Simulation – DOS”. The main idea, when using such a technique, is to minimize the number of cell geometries needed to obtain the mathematical model. In order to establish the needed cell geometries, a Doehlert matrix [10] is used. More concretely, each geometric parameter is affected to a factor with a given number of variation levels. This affectation is summarized in the *Table 1*.

Parameters name	Factor	Variation levels
W_{pp}	X5	3
T_{ONO}	X1	5
L	X3	7
W	X4	7
T_{OX}	X2	7

Table 1: List of parameters with their respective factors and variation levels.

According to the Doehlert matrix with the above choice of factors, the map of simulations is build around 31 different Flash cell geometries’ values. Knowing that 5 simulations are necessary to extract the three threshold voltages of a given cell, a total of 155 simulations are necessary to build the mathematical model. From the result of these simulations, a multi-regression algorithm extracts, for the three threshold voltages, the coefficient “b0”, “bi”, “bii” and “bij” of each polynomial model. The coefficient “b0” represents the mean value of each threshold voltage. The coefficients “bi”, “bii” and “bij” represents, respectively the influence, of each single factor, of each square factor and of each interaction of two factors. The mathematical model of each threshold voltage (V_{TN}) is given in a general manner by the equation 2.

$$V_{TN} = b_{N0} + \sum_i b_{Ni} \cdot X_i + \sum_i b_{Nii} \cdot X_i^2 + \sum_{ij} b_{Nij} \cdot X_i \cdot X_j \quad (2)$$

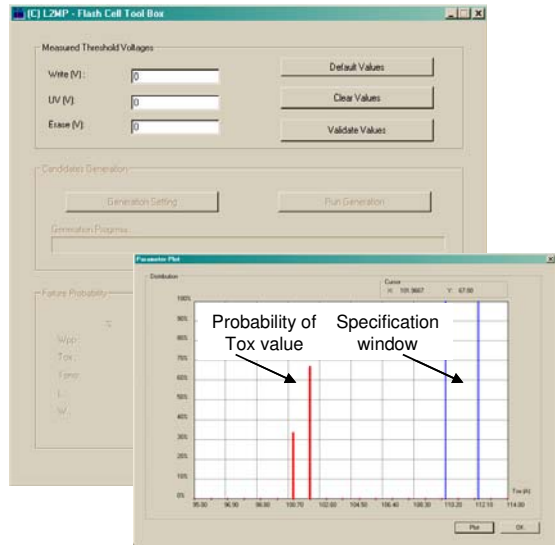


Fig. 2: Flash cell Diagnosis tool overview (main and distribution windows).

To validate the mathematical model, the difference between the values obtained from the equations and the ones obtained from the simulations are calculated for the 31 cells. The maximal differences are 10mV for V_{TUV} , 30mV for V_{TER} and 10mV for V_{TWR} . These differences represent, respectively less than, 0.3%, 1% and 0.2% of the threshold voltages values.

3. Diagnosis Process

The diagnosis tool is able to give the potential distribution of each geometric parameters of a cell and thus allows diagnosing if these distributions are out of the specifications windows. The diagnosis tool takes as inputs, the 3 threshold voltages (V_{TUV} , V_{TER} and V_{TWR}) measured during the Parametric Test. From this set of measurement and using the polynomial model, the tool is able to generate all the cell geometries that match the 3 threshold voltages. All the cell geometries are then used to build the distribution of each geometric parameter. These distributions are then compared with the specification windows in order to establish the probability of each parameter to be out of specification. An overview of the tool, programmed in C++, is given Fig.2 with the main windows and the distribution plot windows.

3.1. Method validation by simulation

A first validation step of the method is done by simulation. A population of 50 cells, with different geometries is simulated and the thresholds voltages V_{TUV} , V_{TER} , V_{TWR} are extracted. This population of 50 cells is defined as follow:

- 10 cells with a Single Parameter Failure (SPF), i.e. one geometric parameter is out of specification (once under the specification and once over the specification),
- 40 cells with a Double Parameters Failure (DPF), i.e. two geometric parameters are out of specification (all possible couples of parameters are studied with under and over specification value).

The simulations of these 50 defective cells exhibit 38 cells (5 with a SPF and 33 with a DPF) with threshold voltages out of the specifications. The diagnosis tool is run for each of these 38 cells, and the results obtained are summarized Table 2.

	SPF	DPF
Number of defective cells	5	33
Number of successfully diagnosed cells	5	32
Efficiency	100%	96.97%

Table 2: Summary of the diagnosis results obtained with simulated defective cells.

The results obtained with our diagnosis tool exhibit a very high degree of efficiency. The Table 2 results

are calculated from the capability of the tool to diagnose at least the parameter with the largest failure. However some details have to be given concerning the DPF cases, because two faulty parameters are involved. It appears that the tool is able to diagnose successfully, for 20 cells, the both faulty parameters. For 12 cells, the parameter with the largest failure is diagnosed. So globally the efficiency of the tool in the case of two faulty parameters is 96.97%. This set of diagnosis based on the simulation of defective cells validates successfully our tool.

3.2. Method validation on silicon

The next step in term of validation is to run our tool on real cell processed on a 0.15 μ m embedded Flash technology from ST-Microelectronics (Rousset-France). One example of real case is detailed here after. The threshold voltages of the cells, in the Test Element Group included in the scribe lines, are extracted during the parametric test. The cells with threshold voltages out of the specification are candidate for our diagnosis tool. From the diagnosis, it appears that this cell has two defective parameters the channel length and the tunnel oxide thickness. In order, to validate these results a SEM analysis is performed as illustrated Fig.3.



Fig.3: Diagnosis validation with physical analysis

For confidentiality reason, the metric values are not given, however the SEM analysis has shown the two following problems:

- A reduction of the flash cell length around 40nm compare to the target of the geometry.
- A problem of salicide thickness on the control gate node.

The results of the physical analysis match in a satisfactory way the results obtained with our diagnosis tool. Indeed, the channel length reduction is given by our tool with a probability of 100% to be defective. The second problem of the cell is a salicide thickness reduction, which is electrically equivalent to an extra resistor-capacitor dipole on the control gate. From a geometric parameter point of view, this extra capacitor

is equivalent to a reduction of T_{ox} or an increase of T_{no} as established by our tool. So, our diagnosis tool results are fully coherent with the experimental results. Studied cases based on simulation as well as experimental results on silicon, fully validate our tool. It is important to note that the parameters with the largest impact on threshold voltages (W and L) are all the time fully diagnosed. The very few cases, that have been not or partially diagnosed belongs to small variations on parameters with a low impact on the cell. This last remark sets the limitation and the confidence that users must have on this diagnosis tool.

4. Conclusion

Flash Memories are based on analog cells. That leads to critical problems of process control and design. In this paper, a polynomial model of the cell threshold voltages versus the cell geometry is proposed. This model is the core element of a diagnosis tool. From the threshold voltages measured during the parametric test, this tool generates for each geometric parameter its distribution and its probability to be defective. The proposed tool is successfully validated with simulations and real silicon data. The on going perspective of this work is to evaluate this tool in a systematic way during the parametric test.

References

- [1] P.Pavan et al., "Flash Memory Cells – An Overview", Proc. IEEE, vol. 85, n° 8, pp.1248-1271, August 1997.
- [2] M. Lenzlinger et al., "Fowler-Nordheim tunneling into thermally grown SiO_2 ", Journal of Applied Physics, vol. 40 (1969) 278.
- [3] P.Cappelletti et al., "Failure Mechanism of Flash Cell in Program/Erase Cycling", IEDM Tech. Dig., 1994, pp.291-294.
- [4] P.Cappelletti et al., "Accelerated Current Test for Fast Tunnel Oxide Evaluation", Proc. IEEE Int. Conf. On Microelectronic Test Structures, 1991, Vol.4, N°1, pp.81-85.
- [5] V.Verma et al., "Proposed On-Chip Test Structure to Quantify Trap Densities within Flash Memories", Proc. IEEE Int. Workshop on Memory Tech. Design and Testing, pp.22-26, 1996
- [6] D.M.Hoffstetter et al., "A Systematic Test Methodology for Identifying Defect Related Failure Mechanisms in EEPROM Technology", Proc. IEEE Int. Conf. On Microelectronic Test Structures, 1994, Vol.7, pp.114-118.
- [7] "ELDO User's Manual", Mentor Graphics Corp., 1998.
- [8] R. Velghe, D. Klaassen and F. Klaassen, "Compact MOS Modeling for Analog Circuit Simulation", IEDM Tech. Dig., pp.485-488 (1993).
- [9] L. Larcher et al., "A New Compact DC Model of Floating Gate Memory Cells Without Capacitive Coupling Coefficients," IEEE Trans. on Electron Devices, Vol.49, N.2, pp. 301-307, 2002.
- [10] D.H. Doehlert, "Uniform Shell Designs", Applied Statistics, 19, pp.231-239.(1978).

A Novel Embedded OTP NVM Using Standard Foundry CMOS Logic Technology

M. Fliesler, G. Rosendale, J. Wang, C. Ng, ZS Liu, J. Peng
Kilopass Technology, Inc. Santa Clara, CA USA
m.fliesler@kilopass.com, www.kilopass.com

Abstract—A novel embedded one-time programmable (OTP) nonvolatile memory (NVM), using only standard Foundry CMOS Logic technology, is described. IP modules with densities from 1Kb to 1Mb were constructed and tested. Reliability data is presented for 1Mb memory modules fabricated in 0.13 μ technology.

Keywords—OTP, NVM, antifuse, gate oxide breakdown

I. INTRODUCTION

High-volume consumer electronics products feature short product life cycles and demand product differentiation with a low manufacturing cost structure. Design cycle time may be reduced and product features enhanced by using embedded programmable memory or logic. Unfortunately, current factory or field-programmable embedded nonvolatile memory (NVM) technology either has an inflexible manufacturing cycle (ROM) or adds significant cost per wafer (Flash). In addition, NVM technology typically lags standard logic technology by a generation or more, and NVM technology scaling is becoming progressively more difficult.

We have developed XPM, or Super Permanent Memory, a novel one-time-programmable (OTP) embedded NVM, which is implemented in standard Foundry CMOS Logic process. No extra mask or processing steps are required. The memory features high density, high performance, excellent scalability and low cost, compared with other NVM embedded memories (Table 1).

II. IMPROVEMENT OVER PRIOR ART

The XPM memory cell utilizes an antifuse principle. The programming mechanism is gate oxide breakdown [1]. A standard dual-oxide CMOS technology is used to create a compact 1.5 transistor memory cell. The memory cell is constructed using the thin-oxide transistors, while the programming circuits are constructed using the thick oxide (I/O) transistors. The memory may be programmed with either an internal charge pump or an external programming voltage (V_{pp}).

The memory achieves high reliability and programming yield by circuit design techniques which limit internal gate oxide stress [2]. Data retention is estimated to be > 10 years.

Table 1: Comparison of XPM with Other NVM

	ROM	OTP	Flash	XPM
Overview				
Process	CMOS	EPROM	Flash	CMOS
Process Generation	Leading	2-3 Gen Lag	1-2 Gen Lag	Leading
Scalable Technology	Yes	No	No	Yes
Benefits Comparison				
Programming	Mask	Field	Field	Field
IC Performance	High	Low	Low	High
Read speed	Fast	Slow	Medium	Medium
Wafer cost	100%	130%	150%	100%
Power Consumption	Low	Medium	Medium	Medium
Scalability, Reliability	High	Medium	Medium	High

The XPM memory cell is an improvement over prior art [3], [4] which used gate-oxide breakdown in a single-gate-oxide CMOS technology, but was unsuitable for high-density memory due to large memory cell area. Other prior art [5] implemented a fuse principle for OTP NVM, but utilized a different fuse material, which required additional processing steps, and was not available in standard Foundry CMOS.

III. MEMORY CELL OPERATION

The basic memory array schematic is shown in Fig. 1. The memory cell has three terminals: a "read" wordline (wr), a "program" wordline (wp), and a bitline (bl). Voltages applied to these terminals are labeled V_{wr} , V_{wp} , and V_b , respectively, in Fig. 1.

Terminal voltages for programming and read operations are listed in Table 2.

Both the core (V_{dd}) and I/O (V_{ddIO}) power supplies are used. For the 0.13 μ general-purpose (013G) Foundry process, $V_{dd} = 1.2V \pm 10\%$, and $V_{ddIO} = 3.3V \pm 10\%$.

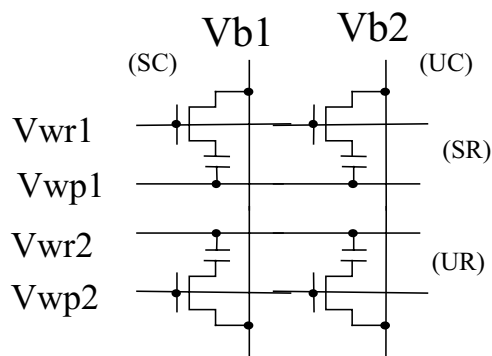


Fig. 1: XPM Memory Array Schematic

During programming, VddIO is applied to the selected row (SR) read wr, the selected column (SC) bl is pulled low, and a short Vpp pulse is applied to the selected program wordline wp. This causes the wp gate oxide to break down, thus programming the memory cell.

During programming, unselected columns (UC) are held at VddIO, which is approximately Vpp/2, and unselected rows (UR) are held at ground. In this way, stress voltages across unselected memory elements are limited to a maximum of Vpp/2.

Reliability is achieved by operating the device in the safe area of the time-dependent dielectric breakdown (TDDB) graphs. Memory cell reliability is ensured by the use of a short (50us) program pulse with a low duty cycle.

In read mode, the cell is selected by applying Vdd to WLr, and VddIO to WLp. A programmed bit will conduct current on the bitline, which will be sensed by the sense amplifier and output a "1". An unprogrammed bit will not conduct current, and will output a "0".

Table 2: NVM Terminal Voltages

		Vbl	Vwp	Vwr	Program
Program	SC/SR	0	VPP	VPP/2	Yes
	SC/UR	0	0	0	No
	UC/SR	VPP/2	VPP	VPP/2	No
	UC/UR	VPP/2	0	0	No
Read	SC/SR	0	VddIO	Vdd	Yes
	SC/UR	0	0	0	No
	UC/SR	Vdd	VddIO	Vdd	No
	UC/UR	Vdd	0	0	No

IV. MEASUREMENTS

Memory arrays of various densities, from 1Kb to 1Mb, were fabricated in 0.13μ CMOS Logic technology. Performance is very good; in packaged test chip form, the 1Mb memory has < 1uA typical standby current, and achieves a typical access time of < 60ns, with < 8mA active current at 5MHz operating frequency [6]. The 1Mb memory is byte-programmable and byte read (128Kx8). Program time is 50us/byte, giving a total memory program time of <7s [7].

The 1Mb array (Fig. 2) was used as the reliability qualification vehicle. It passed the standard Foundry NVM reliability tests, including 1000 hrs HTOL at 125°C (Figs. 3,4), and 168 hrs retention bake at 250°C (Fig. 5). No failures were observed.

Memory distributions, measured in a test mode, were use to detect any changes in the memory cell characteristics. After 1000 hrs HTOL, there was <200mV shift in the memory distribution. The memory had 800mV of power supply margin to the lower end of the datasheet operating spec.

Unbiased data retention bake (DRB) had similarly good performance. After 168 hrs of 250°C DRB, the memory read margin actually improved. The distribution shifted by ~600mV. Most of this shift occurred within the first 48 hrs bake, and is due to a detrapping effect. In contrast, charge-storage type NVM's typically lose read margin after DRB.

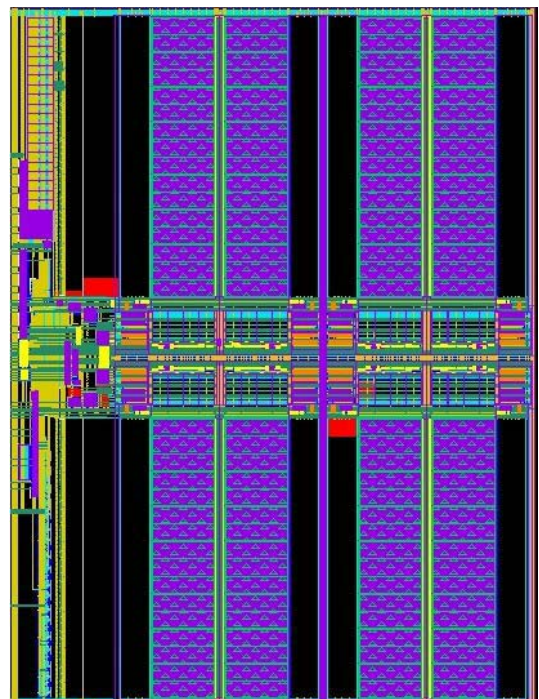


Fig. 2: 1Mb XPM Memory in 0.13u CMOS

V. CONCLUSIONS

A novel OTP embedded NVM structure has been developed, which is implemented in standard CMOS Logic technology. It achieves high performance, high density, scalability, and reliability at low cost. Memory modules have been fabricated at standard commercial foundries in 0.18 μ m, 0.15 μ m, 0.13 μ m, and 90nm, and have passed standard NVM reliability tests.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the contributions of A. Xiao and J.Zhang for testing and reliability data.

REFERENCES

- [1] Peng, Jack, et al, internal communication
- [2] Peng, Jack, et al., internal communication
- [3] Candelier et al, US Patent #6,421,293
- [4] Smith, et al., U.S. Patent #6,525,955
- [5] Anthony, et al, U.S. Patent #6,580,144
- [6] Kilopass XPM Memory KP-13T Series data sheet.
- [7] Kilopass XPM Memory Testing Methodology Guidelines

Fig. 3: 1Mb Pre-HTOL Memory Distribution

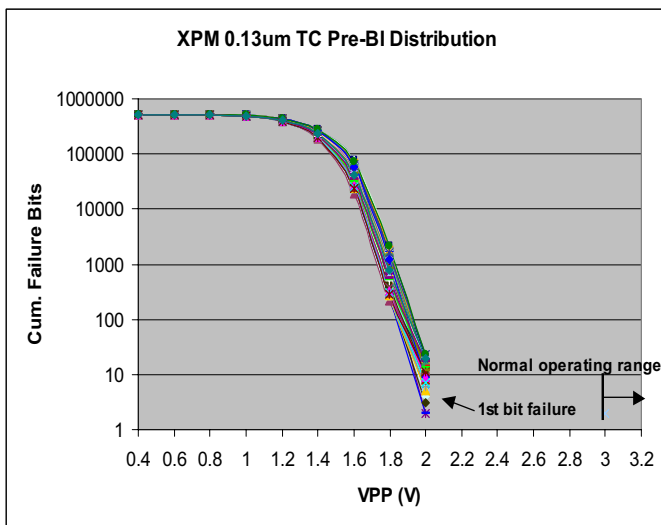


Fig. 4: 1Mb Post 1000 hrs HTOL Memory Distribution (125C)

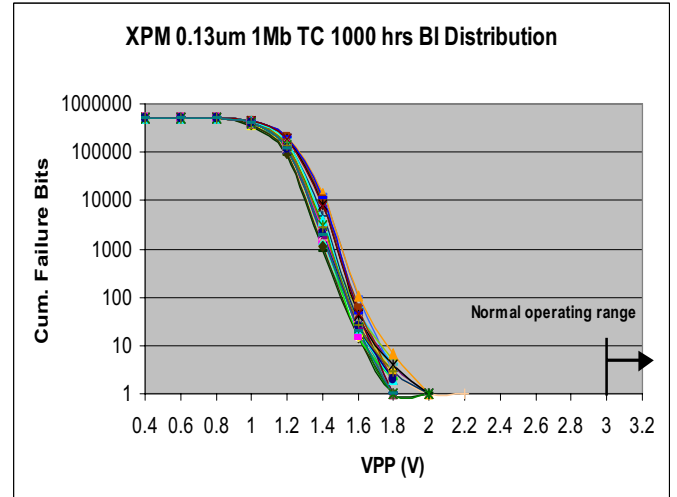
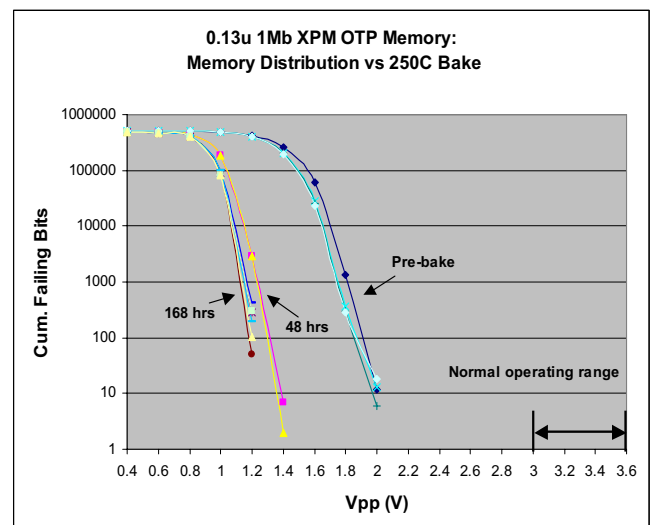


Figure 5: 1Mb Memory Distribution vs. Data Retention Bake (250°C: 0 hrs, 48 hrs, 168 hrs)



LIST OF AUTHORS

A

H. Achard.....145
 B. Acred 21
 A. Akheyar.....239
 N. Akil.....219, 231
 S.G.H. Anderson 21
 S.A. Anderson 37

R. Andrei 199
 S. Angle.....215
 Y. Asao 107
 G.B. Assayag 29
 J.-L. Autran 93

B

N. Baboux 251
 Ch. Back 131
 M. Baclet 89
 J.S. Bae 109
 W. Baks 219
 C. Balocco 49
 T. Baron 45
 F. Beaudoin 267
 J.-M. Bédrine.....215
 F. Bernard-Granger.....135
 M. Berrebi 215
 A. Berthelot.....67
 R. Beurze 231
 R. Bez 17
 P.W.M. Blom 9
 P. Blomme.....239

B. Boeck 67
 P. Boivin.....251
 C. Bonafos 29
 B. Borot 81
 L. Bortesi 157
 G. Boschi 97
 D. Boter 161
 R. Bouchakour 59, 85
 L. Breuil 227
 D. Brunco.....211
 A. Bsiesy.....145
 J. Buckley 207
 F. Buonocore 185
 S. Burignat.....251
 C. Busseret.....41, 45

C

P. Cacharelis.....247
 C. Caillat.....67, 81
 A. Campera 33
 P. Candelier 59
 M. Carrada.....29
 K. Castellani-Coulié 93
 E. Cattani.....189
 M. Cerisier.....215
 H.S. Chae 235
 S.D. Chae.....235
 P.R. Chalker.....263

K.M. Chang.....21
 C. Chappert.....117
 M. Charleux.....67
 H.-H. Chen.....63
 N. Cherkashin.....29
 R. Chevallier 89
 U-In Chung.....109
 Claverie 29
 H. Coffin 29
 M. Combe 243
 S. Conoci 185

D

E. Daemen 215
 J.-M. Daga.....243
 J.F. Damlencourt.....207
 D.M. De Leeuw 9
 K. De Meyer 239
 C. De Nardi.....267
 B. De Salvo.....25, 153, 207

J. De Vos 227
 P. Deconinck 259
 W. Dehaene.....73
 S. Deleonibus.....153, 207
 D. Deleruyelle 25
 V.E. Demidov.....131
 S.O. Demokritov.....131

D

R. Desplats.....267
 M. Detalle.....189
 T. Devolder.....117
 B. Dieny 139
 P. Dimitrakis 29
 U. Dorda 149

E

U. Ebels 139

F

M. Fanciulli 29
 S. Favier N.A.
 P. Fazan.....55, 63
 V. Ferlet-Cavrois.....93
 S. Ferraton.....41
 G. Festes 215

G

L. Gabette.....67
 B. Gallagher.....103
 J. Garbe 161
 G. Gasiot.....81
 H.P. Gasquet 21
 P. Gassot 247
 J.-L. Gauffier.....267
 P. Geens 73
 G.H. Gelinck.....9
 M. Gely 25, 207

H

Y.K. Ha 109
 H. Hada.....107
 R. Hagenbeck.....259
 P. Haibach 259
 J. Hamilton 21
 J.H. Han 223, 235
 F. Hansteen.....141
 L. Haspeslagh 163, 227
 T. Heijmen.....77
 J. Hendrickx.....71

I

G. Iannaccone 33
 D. Ielmini.....195
 S. Ikegawa.....107

J

S. Jacob 153
 F. Jacquet 81
 V. Javerliac 135
 S. Jeon.....223, 235

D. Dormans.....161, 231
 L. Dreeskornfeld 149
 J. Dubois.....161
 C. Duluard 145
 A. Dumas 215
 D. Dutratre 67

N. Emonet.....67

A. Filipe.....145
 BM. Fliesler 275
 R. Fournel 117
 D. Fraboulet 67
 A. Francinelli 145
 E. Fribourg-Blanc 189

J. Genoe.....181
 E. Gerritsen.....67
 A. Ghetti 157
 G. Ghibaud.....207
 N.E. Gilbert.....177
 C. Gopalan.....177
 L. Goux 163
 B. Govoreanu.....211
 M. Grégoire 215

P. Heremans 181
 B. Hillebrands 131
 F. Hofmann.....149
 J. Holland.....127
 B. Hradsky 21, 37
 Y. Huai 123
 V. Huard 67
 E. Hugonnard.....215
 J.W. Hyun 235

F. Irrera 255
 N. Ishiwata.....107
 M. Isler 259

Y.S. Jeong.....235
 A.C. Jones 263
 N. Jourdan.....67

K

T.	Kai	107
v.	Kamienski	259
M.	Kanoun.....	45
M.	Kerekes	113, 139
C.	Kim.....	223, 235
H.J.	Kim.....	109
H.S.	Kim.....	109
J.H.	Kim.....	235
J.K.	Kim.....	223
J-V.	Kim.....	117
M.K.	Kim.....	235

L

A.L.	Lacaita.....	195
D.	Lafond.....	25
J.	Langer.....	123
F.	Lau.....	259
G.	Le Carval.....	153
J.C.	Ledezma.....	21
J.E.	Lee	109
S.H.	Lee	223, 235
M.	Leskelä.....	175
S.W.	Lin	49

M

W.	Maass.....	123
V.	Madigou.....	163
D.	Maes.....	163
P.	Malinge	59
D.	Mantegazza	195
R.	Mariani.....	97
A.W.	Marsman	9
F.	Martin	207
P.	Masson.....	59, 85, 93
L.	Mathew	63
C.	Maufront	117
A.	Mauri	157
C.	Mauroury.....	117
P.	Mazoyer	59, 67
D.	Medjahed.....	247

N

R.C.G.	Naber	9
M.	Nagoga.....	63
M.	Nakayama.....	107
K.T.	Nam	109
D.	Née.....	85, 271
F.	Neuilly	219
C.	Ng.....	275

T.	Kim.....	N.A.
A.V.	Kimel	141
A.	Kirilyuk	141
R.	Komaragiri	167
R.	Kömmeling	149
M.P.	Kostylev	137
M.N.	Kozicki.....	177
J.	Kretz	149
T.J.	Kropenwnicki	127
B.	Kruseman	77
K.-H.	Kuesters.....	259

V.	Lipponen	N.A.
J.G.	Lisoni	163
Z.S.	Liu.....	275
Y.F.	Loo.....	263
L.	Lopez	85
M.	Lorenzini.....	227
S.	Louwers.....	203
A.	Lowe	247
C.	Ludwig.....	259
L.V.	Lutsev.....	137

N.	Menou	163
R.	Mikalo.....	259
T.	Mikolajick	259
Y.S.	Min.....	235
A.H.	Miranda.....	219
M.	Missous	49, 171
G.	Molas	25
F.	Monsieur	67
L.	Montès	41
J.T.	Moon.....	109
C.	Muller.....	163
R.	Müller.....	181
D.	Munteanu	93
P.	Mur	N.A.
R.	Muralidhar.....	21, 37

B.-Y.	Nguen.....	63
G.	Nicotra	207
E.	Nolot	N.A.
P.	Normand.....	29
N.N.	Novitskii.....	137
J.P.	Nozière.....	113, 139

O

B.	Ocker	123
S.C.	Oh	109

P

P.	Paillet	93
T.	Panagopoulos	127
V.	Paraschiv	163
S.O.	Park	109
S.	Parkin	121
A.	Paterson	127
R.	Patz	127
F.	Pellizzer	195
J.	Peng	275
P.	Perdu	267
M.	Perego	29
A.	Persico	113

R

R.	Ranica	59
R.A.	Rao	21
R.	Rao	37
Th.	Rasing	141
O.	Redon	113, 139
J.	Regolini	67
R.	Reichelt	259
D.	Remiens	189
A.	Ricca	97

S

J.-U.	Sachse	259
M.	Sadd	21, 37
V.	Safarov	145
B.	Saillet	271
S.D.	Savransky	191
S.	Schamm	29
J.-M.	Schley	259
J.P.	Schoellkopf	81
N.	Schulze	259
H.W.	Schumacher	117
U.	Schwalke	167
M.	Schwitters	163
W.	Seifert	171
N.A.	Sergeeva	137
E.	Serres	215
S.	Setayesh	9
H.	Sibuet	113
T.	Skotnicki	59
M.	Slotboom	219, 231
L.M.	Smith	263

S.	Okhonin	63
K.	Ounadjela	111

J.	Peschke	21
S.	Petralia	185
M.	Piazza	67
A.	Pirovano	17, 185, 195
R.V.	Pisarev	141
C.	Plossu	251
F.	Ponthennier	113
J.-M.	Portal	271
Li.	Prejbeanu	135
L.	Prejbeanu	113, 139
E.J.	Prinz	21, 37
G.	Puzzilli	255

S.	Riedel	259
G.	Rinkenberger	21
L.	Risch	149
M.	Ritala	175
P.	Roche	81
P.	Rohr	215
G.	Rosendale	275
W.	Rösner	149
Y.	Roussigné	137

A.	Solere	215
A.M.	Song	49, 171
R.	Sotgiu	185
A.	Souifi	45
B.	Soumireu	13
R.	Sousa	113, 139
M.	Specht	149
M.	Staedele	149
D.	Stanciu	141
A.A.	Stashkevich	137
Y.	Stefanov	167
R.F.	Steimle	21, 37
E.	Stein	259
T.	Stephens	63
A.I.	Stognij	137
J.	Stokes	127
M.	Strassburg	259
S.	Straub	21, 37
Y.	Suzuki	117
C.T.	Swift	21, 37

T

P.	Taechakumput	263
S.	Tahara	107
C.	Tanase	9
S.	Taylor	263
S.	Tehrani	105

V

F.	Vajda	129
M.	Van Duuren	219, 231
J.	Van Houdt	211, 227, 239
R.	Van Schaijk	219, 231
W.	VanDrent	129

W

J.	Wang	275
T.C.	Wei	191
B.E.	White Jr	21, 37

X

P.	Xavier	139

Y

B.C.K.	Yagami	117
T.	Yao	247

Z

H.	Yoda	107
R.	Zambrano	163

P.G.	Tello	219
G.	Tempel	259
F.J.	Touwslager	9
K.	Tulapurkar	117
Ch.	Turquat	163

M.	Vehkamäki	175
L.	Vendrame	157
G.	Viau	137
M.L.	Villani	N.A.
A.	Villaret	59

J.	Willer	259
D.J.	Wouters	163

J.A.	Yater	21, 37

J.	Zimmermann	41

NOTES

NOTES

NOTES

NOTES

NOTES

NOTES

NOTES

NOTES

NOTES

NOTES

**The ICMTD-2005 is organised by L2MP
and associated institutes:**



ORGANISATION

Rachid Bouchakour
L2MP
Marseille, France

Eric Gerritsen
Philips Semiconductors
Crolles, France

Pascal Masson
L2MP
Marseille, France

Pascale Mazoyer
STMicroelectronics
Crolles, France

Christophe Muller
L2MP
Toulon, France

COMMUNICATION

Corinne Joachim
CREMSI
Rousset, France

Cathy Paitel
L2MP
Marseille, France

**The ICMTD-2005 is organised with financial and logistics
support from the following organisations:**



SPONSORS