Mémoires embarquées non volatiles à grille flottante : challenges technologiques et physiques pour l'augmentation des performances vers le noeud 28nm

Adam Dobri

▶ To cite this version:

Adam Dobri. Mémoires embarquées non volatiles à grille flottante : challenges technologiques et physiques pour l'augmentation des performances vers le noeud 28nm. Micro et nanotechnologies/Microélectronique. Université Grenoble Alpes, 2017. Français. NNT : 2017GREAT030 . tel-01693008

HAL Id: tel-01693008
https://theses.hal.science/tel-01693008

Submitted on 25 Jan 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

Spécialité : **Nano électronique et nano technologies**

Arrêté ministériel : 25 mai 2016

Présentée par

Adam DOBRI

Thèse dirigée par **Prof. Francis BALESTRA**

préparée au sein du **Laboratoire Institut de Microélectronique, Electromagnétisme et Photonique - Laboratoire d'hyperfréquences et de caractérisation** dans **l'École Doctorale d'Électronique, Électrotechnique, Automatique et Traitement du Signal**

Embedded Non-volatile 1T floating-gate memories: technological and physical challenges for augmenting performance towards the 28 nm node

Mémoires embarquées non-volatiles à grille flottante: challenges technologiques et physiques pour l'augmentation des performances vers le nœud 28 nm

Thèse soutenue publiquement le **13 Juillet 2017**,
devant le jury composé de :

Monsieur Gérard GHIBAUDO
Directeur de recherche CNRS Alpes, Président
Monsieur Albdelkader SOUIFI
Professeur, INL/INSA de Lyon, Rapporteur
Monsieur Pascal MASSON
Professeur, Université de Nice Sophia Antipolis, Rapporteur
Monsieur Francis BALESTRA
Directeur de recherche CNRS Alpes, Directeur de thèse

Acknowledgements

The old proverb that "it takes a village to raise a child" can easily be scaled to become "it takes an entire team to train a PhD student" as there is no way that I could have succeeded in this work without the support of my co-workers and peers in many different places. In keeping with the spirit of my time in Grenoble the acknowledgements will be written in a mixture of French and English; I will try my best to avoid ST-isms that do not exist in either language.

Je voudrais d'abord remercier mon directeur de thèse, Francis Balestra, et mon manager à ST, Fausto Piazza, d'avoir accepté d'encadrer cette thèse sur les mémoires flash. Francis a une bonne vue globale sur la microélectronique et Fausto est vraiment fort en mémoires embaquées, j'ai toujours bien apprécié mes échanges avec eux. Je suis convaincu qu'un quart d'heure avec les experts vaut bien plus qu'une heure passés avec la biblio.

Dans mon travaille quotidienne j'ai beaucoup travaillé avec Simon Jeannot qui m'a appris le travail d'un ingénieur en « Process Integration » et les informations sûr les diélectriques. Il mérite aussi un grand « Thank you ! » pour tout son aide avec la rédaction de ce manuscrit, merci de m'avoir guidé ces dernières années tout en me laissant choisir le chemin de mon projet. L'équipe entière de M40 représentait un milieu excellent pour l'apprentissage du métier, c'a été bien d'avoir des experts de la flash à seulement quelques mètres chaque fois que j'avais une question. Je m'en souviendrai aussi des repas et cafés (et bien sûr les gâteaux) partagés avec l'équipe. Du côté CEA-Leti j'aimerais bien remercier mon encadrante, Carine Jahan, qui m'a guidé et aidé avec la partie intégration en salle blanche au CEA-Leti ainsi que la rédaction.

Je voudrais aussi remercier l'équipe de traitement thermique à ST qui m'a accueilli pendant mon stage de master en 2011 et qui m'a présenté le monde de la microélectronique industrielle. C'est grâce à eux que j'ai rapidement amélioré mon français et pris la connaissance des thèses CIFRE. Une thèse en Process Integration nécessite une forte interaction avec plein d'autre personnes, des ingénieur(e)s et technicien(ne)s de Process Development aux gens en salle blanche et de la métrologie. Il n'est pas possible de lister tous les noms mais je voudrais remercier Clément, Mickaël, Nathalie, Nelly, Nils, Olivier, Patrick et Vincent de m'avoir aidé lors de mon temps à Crolles.

Vingt-cinq kilomètres à l’ouest de Crolles il y avait encore une grande équipe à remercier au CEA-Leti. Alain nous a beaucoup aidé avec la gestion et avancement des lots en salle, surtout lors de la dernière année quand Carine et moi étions plus souvent à Crolles qu’au CEA-Leti. Nos lots n’auraient jamais vu les testeurs sans la soutien de Christian qui nous a développé une gravure wordline. Pour tout ce qui est caractérisation électrique je dois un grand « merci » à Jean et Alex pour leur collaboration lors du développement de la technique « OSS » ainsi qu’Alain et Fabienne pour les analyses paramétriques.

Il ne faut pas oublier non plus les amis et autres thésards qui ont été là pour rendre les trois dernières années un véritable plaisir. Ils ont été là pour m’aider à naviguer la vie d’une thésard et l’administration française. Je tiendrai toujours près du cœur tous les sorties en ski/randonnée/vélo dans cette région magnifique et lies bières/raclettes qu’il fallait prendre pour les planifier. Pour ceux qui sont déjà partis j’espère qu’on se croisera dans l’avenir et pour ceux qui n’ont pas encore soutenu leurs thèses, courage !

I would also like to acknowledge my family back home in Canada, and thank them for putting up with me being away so often. I know my parents would have preferred that I find a PhD position a closer to home. A big “thank you” also goes out to my future wife, Sholpan, who was always there for me during the difficult times of the PhD and who always found a way to motivate me to keep going.

Table of Contents

List of Figures

List of Tables

List of Acronyms

ALD	Atomic Layer Deposition
BT	Break Through etch step
CEA	*Centre d'Energie Atomique*, the French Alternative Energies and Atomic Energy Commission
CEA-Leti	*Laboratoire d'électronique des technologies de l'information*, microelectronics institute in Grenoble, France
CHE	Channel Hot Electrons
CMP	Chemical Mechanical Planarization
CPU	Central Processing Unit
CV	Capacitance-Voltage curves or relationship
CVD	Chemical Vapour Deposition
DRAM	Dynamic Random Access Memory
EDX (or EDS)	Energy Dispersive X-Ray Spectroscopy
EEPROM	Electrically Erasable Programmable Read Only Memory
EOT	Equivalent Oxide Thickness
HCI	Hot Carrier Injection
HF	Hydrofluoric Acid
HKMG	High-k Metal Gate structure
HV	High Voltage devices on an embedded flash product
IGD	Inter Gate Dielectric (see IPD)
IPD	Inter Poly Dielectric (see IGD)
ITRS	International Technology Roadmap for Semiconductors
IV	Current-Voltage curves or relationship
LAE	Leti Alumina Etch chemistry
LPCVD	Low Pressure Chemical Vapour Deposition
LV	Low Voltage, or logic CMOS device on an embedded flash product
ME	Main Etch step
MEMS	Micro Electrical Mechanical Systems
MIM	Metal-Insulator-Metal device
MOCA	Name of Maskset for damascene-type capacitors at the CEA-Leti
MOCVD	Metal Organic Chemical Vapour Deposition

MOS	Metal Oxide Semiconductor
MOSCAP	Metal Oxide Semiconductor capacitor
MOSFET	Metal Oxide Semiconducting Field-Effect Transistor
MRAM	Magneto-resistive Random Access Memory
NAND	Type of flash memory, named for the logic gate its configuration resembles
NLDD	(N-type) Lightly Doped Drain
NOR	Type of flash memory, named for the logic gate its configuration resembles
NVM	Non-Volatile Memory
ONO	Classical IGD, $SiO_2/Si_3N_4/SiO_2$
OSS	Oxide Stress Separation
PCM	Phase Change Memory
PLDD	(P-type) Lightly Doped Drain
PVD	Physical Vapour Deposition
RRAM	Resistive Random Access Memory
SIMS	Secondary Ion Mass Spectrometry
SNOW	Side Nucleation On Wall deposition of SiO_2
SONOS	Silicon-Oxide-Nitride-Oxide-Silicon structure for charge trapping memory
STI	Shallow Trench Isolation
TCAD	Technology Computer Aided Design
TDDB	Time Dependent Device Breakdown
TEM	Transmission Electron Microscope or Micrography
TEOS	Tetraethyl orthosilicate, precursor for SiO_2 deposition
TREQ	Devices like a flash cell with the floating and control gates shorted
USB	Universal Serial Bus
VARIOT	Variable Oxide Thickness
WKB	Wentzel–Kramers–Brillouin, approximation used in calculating tunneling probabilities

Résumé en Français – Summary in French

En croissant de zéro vers un marché de 400 milliards de dollars, l'industrie de la microélectronique est devenue un marché aux multiples facettes avec des technologies qui se sont différenciées pour desservir une grande variété de marchés. La cellule flash NOR à grille flottante est utilisée depuis longtemps sur le marché de la mémoire et joue toujours un rôle majeur dans les mémoires flash embarquées au nœud technologique 40 nm. La multi-couche dioxyde de silicium / nitrure de silicium / dioxyde de silicium trilayer, "ONO", est presente tant que diélectrique entre grilles depuis des décennies. Un bon diélectrique entre grilles est électriquement mince pour contrôler le couplage entre la grille flottante et la grille de contrôle tout en restant suffisamment épais pour bloquer les fuites (perte de données); la mise à l'échelle agressive améliore les performances mais augmente le risque de perte de données. L'utilisation continue de l'ONO soulève plusieurs questions. Comment saurons-nous quand la mise à l'échelle de l’ONO est devenue trop agressive? Peut-il être remplacé? L'ONO serait-il compatible avec les processus logiques des portes métalliques aux futurs nœuds? Et les matériaux à haute permittivité représentent-ils une bonne option pour le remplacement de l’ONO? Les recherches menées au cours des deux dernières décennies sur les dispositifs NAND planaires pour des applications à haute densité peuvent aider à orienter les réponses concernant le flash NOR incorporé en identifiant des matériaux candidats tels que l'alumine et les silicates de hafnium.

Comme la perte de charge de la porte flottante conduit à la perte de données, il est donc très intéressant de savoir d'où viennent ces pertes. L'approche logique serait de mesurer directement le courant de fuite à travers un diélectrique entre grilles sur un condensateur de test; ceci est impossible car les courants aux champs électriques bas sont en dessous des limites de mesure. Les méthodes de portes flottantes ont été inventées afin d'étendre les limites de mesure, mais elles ne sont toujours pas assez sensibles pour atteindre tels champs. Nous avons développé et breveté la technique de séparation des contraintes d'oxydes (OSS pour « Oxide Stress Separation » en anglais) pour mesurer les courants de fuite à travers le diélectrique entre grilles dans une cellule flash nominale, décrite au chapitre 2. La prémisse de la technique OSS est qu'en choisissant soigneusement la tension de seuil et les conditions de polarisation il est possible de ne pas avoir de champ électrique dans l'oxyde du tunnel. Avec l'oxyde de tunnel à l'état de bande plate, toute la chute de potentiel se produit dans le diélectrique entre grilles et tout courant de fuite peut être

considéré comme passant via le diélectrique entre grilles. Les courants de fuite de l'ordre de 10^{-22} à 10^{-23} A sont déduits des variations de la tension de seuil de la cellule flash testée et peuvent être comparés aux résultats des tests de rétention des données faits aux températures élevées. Au nœud technologique 40 nm, l'ONO reste de bonne qualité et ne contribue qu'une faible partie des courants de fuite de la grille flottante. La technique OSS sera utile pour évaluer les propriétés de rétention des données dans les futurs nœuds, qu'ils contiennent ou non du diélectrique entre grilles ONO ou matériaux haute permittivité.

L'un des candidats potentiels pour le remplacement du nitrure de silicium dans l'ONO est l'alumine car elle offre un constant diélectrique accru pour un meilleur couplage tout en conservant une bande interdite relativement élevée pour réduire les courants de fuite. Pour la première fois, au chapitre 3, des puces flash embarquées à 40 nm ont été fabriquées en utilisant des diélectriques entre grilles à base d'alumine soit sous forme de couches d'alumine sur 4 nm de dioxyde de silicium, avec ou sans couche finale de dioxyde de silicium au-dessus de l'alumine. Les modifications du flux de processus standard pour le produit flash embarqué ont été conséquentes, avec de nombreuses étapes au CEA-Leti. La couche d'alumine est facilement retirée des grilles de contrôles des dispositifs MOS haute tension et des ouvertures du condensateur diélectriques entre grilles en utilisant le procédé de gravure humide développé au CEA-Leti. L'étape de gravure qui définit les “word lines” des cellules flash a été plus compliquée et a abouti à un profil sous-optimal, bien que l'alumine exposée ait été complètement retirée, ce qui a réduit les risques de contamination dans les étapes de la procédure.

Les dispositifs d'essai élémentaires ont montré que les couches d'OA et d'OAO présentent de bonnes performances intrinsèques, avec des fuites et des tensions de claquage qui semblent permettre la réduction de l'épaisseur d’oxyde équivalent du diélectrique entre grilles. En utilisant ces tests électriques, nous voyons que la bicouche semble être plus prometteuse que les trilouches et que l'alumine déposée en utilisant l'ozone comme précurseur d'oxygène est de meilleure qualité. La caractérisation complète de ces empilements n'a pas été effectuée au cours de ce travail, en raison de problèmes de calendrier. Néanmoins, les matériaux seront évalués par une caractérisation fine (conservation des données, rendement et mesures OSS) pour donner un aperçu complet du potentiel offert par les piles d'alumine pour les diélectriques entre grilles. Sur nos premiers échantillons, les améliorations attendues de la performance d'effacement des cellules flash ne sont pas observées avec les diélectriques entre grilles les plus minces. La perte de couplage due à une diminution de la surface de la grille flottante, causée par un profil de ligne de mot sous-optimal, est censée éliminer les gains de couplage liés à des diminutions de l'épaisseurs électriques. Après cette première démonstration de cellules entièrement intégrées à base d'alumine, la nouvelle amélioration des procédés de gravure “word line” devrait faire ressortir les avantages d'une épaisseur électiruqe réduite sur les performances cellulaires.

Au niveau du nœud suivant, 28 nm, les dispositifs logiques CMOS sont formés en utilisant des empilements de portes métalliques avec un diélectrique de haute permittivité (HKMG) qui nécessiteront des modifications du diélectrique entre grilles. La première question que l'on peut se poser est de savoir si l’empilement HKMG dégraderait l'ONO si un schéma d'intégration à deux portes similaire était conservé. Grâce aux analyses de CV, IV et de claquage dans le chapitre 4, nous avons montré que le HKMG ne dégrade pas les propriétés de l'ONO. En fait, la résistance de l’empilement HKGM à la dégradation a augmenté dans les tests de claquage dépendants du temps. La question suivante est de savoir si un empilement de portes à haute permittivité permettrait ou non un meilleur couplage sans augmentation des fuites. Ceci a également été testé au chapitre 4 avec des IGD constitués de 4 nm de dioxyde de silicium suivis par 7 à 22 nm de silicate d'hafnium ou d'oxyde d'hafnium. Les silicates de hafnium ont montré une permittivité plus élevée, des courants de fuite plus faibles, un temps au claquage plus long et de meilleures durées de vie que les échantillons d'oxyde d'hafnium. Par conséquent, les silicates de hafnium sont plus susceptibles de représenter une amélioration pour les applications diélectrique entre grilles. L'intégration d'un diélectrique entre grilles à base de silicate d'hafnium épais s'est révélée compliquée car la gravure sèche n'était pas suffisante pour retirer la matière de tous les emplacements souhaités pendant la première étape de gravure. Dans tous les cas, au niveau du nœud de 28 nm et au-delà, la méthode de dépôt du métal de grille deviendra critique en termes de conformité afin de prendre en compte les grilles flottantes, les motifs non présentes dans le traitement logique normal.

En conclusion, l'ONO actuel est de très haute qualité et peut donc ne pas avoir besoin d'être immédiatement remplacé dans le nœud de 28 nm. Il est impératif de tester une intégration flash embarquée de 40 nm avec l'ONO et l’empilement de grille métallique haute k qui a été testée sur les condensateurs au chapitre 4. Si l'ONO est remplacé par un IGD à permittivité élevé, les silicates d'alumine et d'hafnium se sont révélés être de bons candidats pour l'utilisation dans les cellules flash. Comment ils sont intégrés dans un produit flash intégré est tout aussi important que leurs propriétés matérielles; surtout en ce qui concerne la gravure “word line”. Cette étape de gravure définit les cellules flash et est critique car les différences de largeur de grille de commande peuvent entraîner une perte de couplage capacitif qui ne peut pas être récupérée via des diélectriqued entre grilles plus minces. Les futures intégrations des diélectriques entre grilles d'alumine ou de silicate de hafnium nécessiteront une collaboration très étroite avec des équipes de gravure sèche / humide afin de trouver des conditions de gravure qui marchent. Dans ces optimisations, une caractérisation précise du diélectrique entre grilles lui-même telle que la technique OSS sera très utile pour comprendre les dispositifs afin de guider le développement du processus.

General Introduction

While “may our ones stay ones and our zeros stay zeros” is rarely made as a toast, the desire to record and store information has been a motivator of humans’ technological progress for several millennia. From early markings on cave walls to magnetic hard-drives, the decreasing cost and increasing density of memories have opened up new markets and applications for them. During the search for a *universal memory* that is fast, reliable, non-volatile, durable and most importantly economical, the electronics industry has been forced to make a series of trade-offs for its different types of memories. Dynamic Random Access Memory (DRAM) is cheap, fast and durable; however as it requires constant refreshing it is not used for long-term storage. Alternatively, mechanical hard drives can store large amounts of data for a long time, but they are slow and therefore could never be expected to replace the Central Processing Unit (CPU) cache.

Embedded flash memories represent a trade-off between multiple individual systems and are used by the microelectronics industry in a variety of applications where information must be treated quickly and securely over a wide range of temperatures. These Non-Volatile Memories (NVMs) are used to secure smartcard data, run microcontrollers and manage automotive sensors. One of the workhorses of the NVM market, the floating gate flash cell, involves controlling the quantity of charge stored on an isolated gate. As embedded flash memories advance through the 40 and 28 nm technology nodes, the choice of material for the Inter Gate Dielectric (IGD) may become critical; a leaky IGD would render the chip unsuitable for use.

What are Non-Volatile Memories?

As the name implies, Non-Volatile Memory is the type of memory which can retain its stored information specifically it does so without a power supply [Cap99]. This ability to store information without expending energy for constant refreshing is incredibly important in terms of the cost (power consumption) and reliability (no loss of data in the case of power failure). These memory chips are all around us; the average consumer will certainly be familiar with “flash drives” in the form of USB keys and the SecureDigital series of memory cards shown in Figure 1. Less familiar applications include the embedded memory chips in smart cards or sensors, also shown in Figure 1.

Transcend
SD HC
16GB
4974
SPC58 C Line Automotive MCUs
for car body and security applications
SPC58 C

Figure 1: Flash memory card, smartcard and microcontrollers containing flash memory [Stm16]

Solid-state, semiconductor-based NVM have advanced steadily since they were introduced to the marketplace. As the consumer appetite for memory has grown along with capacity of the chips, the price per unit decreases are somewhat hidden and often pass-by underappreciated. For example, in 2003 a 128MB USB drive was about 33 USD [Mcc17] (44 USD in 2017) while in 2017 128GB drives are offered at a similar price point. These advances in performance and cost are due to the ever-shrinking size of memory cells and the economies of scale that come with packing more and more cells onto larger silicon wafers. The rate at which the microelectronics industry has shrunk the size of the transistors and memory cells followed the International Technology Roadmap for Semiconductors (ITRS), laid out by the semiconductor industry over several decades. The approximate rate of doubling the number of transistors per chip every two (later adjusted to three) years is referred to as Moore's Law, after Gordon Moore, an eventual co-founder of intel, who noted the trend in his 1965 essay [Gor65].

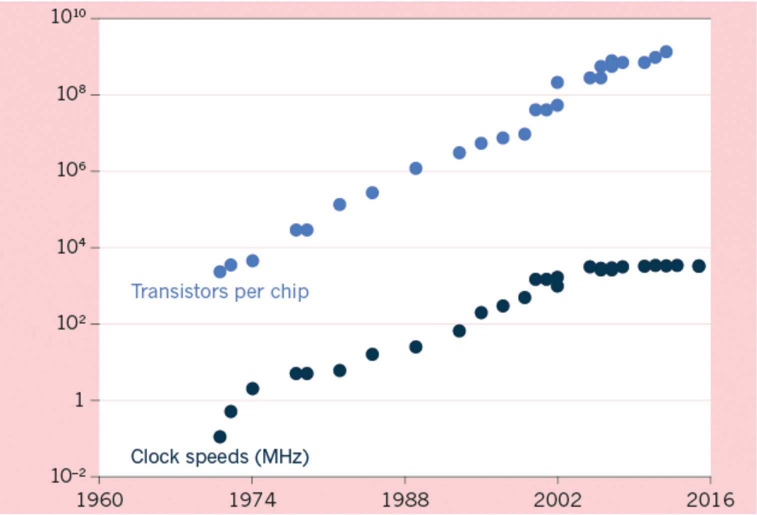


Figure 2: The number of transistors per chip and clock speeds have increased by orders of magnitudes since the 1970s, with clock speeds leveling off in the mid-2000s to avoid overheating [Wal16]

As the scaling limits of the ITRS are approached, the number of companies working on the smallest technology node has fallen in recent years, and many companies have focused more on finding new applications or products instead of simply chasing smaller and faster transistors. The market also faces greater segmentation as each technology may have its own scaling trajectory as “thinner layers” and “smaller features” no longer have the same cost/benefit ratios as they did during the late 1900s. For example, design rules and voltages optimized for logic devices at a new node can be re-optimized for specific applications such as Micro Electromechanical Systems (MEMS), 3D integration or mixed-signal integration [Whi16]. Embedded flash memories are no exception as different technologies have been developed for different applications.

Presentation of this thesis

Embedded floating gate flash memories require relatively high performance logic circuits to be fabricated on the same silicon wafers as high performance memory cells. This co-integration requires trade-offs to be made and leads to a different scaling trajectory than either CMOS (complementary metal-oxide-semiconductor) only technologies or stand-alone memory. This thesis describes the comprehension and amelioration of 40 nm embedded flash memory cells and their extension towards the 28nm node. Chapter 1 of this thesis describes the global microelectronics and NVM market, as well as the technical background. Chapter 2 outlines the development of the Oxide Stress Separation technique which can be used to advance the characterization of flash cells; it can be used to detect if the scaling has become too aggressive. If this point of overly aggressive scaling has been reached, then new design or materials’ solutions will be required to keep meeting the markets’ memory demand. Chapter 3 presents an overview of the up-to-date flash memory process flow and the critical steps. This embedded flash product integration is explained within the context of the integration of alumina in the flash cell as a method of improving the performance of current embedded flash cells; these represent the world’s first 40 nm embedded flash based on alumina. Chapter 4 lays the foundation for the application of the current floating gate flash memory integration scheme at the next technology node using high-k metal gate technologies. Chapter 5 provides general conclusion and perspectives for future work.

Chapter 1: Non-Volatile Memory Market and Technological Developments

Table of Contents

1.1 Introduction to the global microelectronics market

From the humble beginnings of a simple transistor in a research lab to the forecasted market value of 372 billion USD in 2020, integrated circuits have shaped our world in ways that would have been difficult to imagine. Over the years, the markets have been driven by a variety of different applications. The proliferation of personal computers driving annual growth in the 1980s, while 1990s will be remembered for the Intel vs AMD microprocessor battle and ever-increasing DRAM requirements. The next decade included a technical slow-down as processor clock speeds leveled off at the same time that the economic slow-downs, such as the dot.com bust and the 08-09 global recession, hit. The current decade has included an explosion in smartphone growth, offsetting the decline of personal computer sales, while the Internet of Things and mobile devices are expected to carry the market on through 2020 [Ici17]. Along with processing power, these devices, whose respective markets are presented in Figure 3, need to have access to some form of memory.

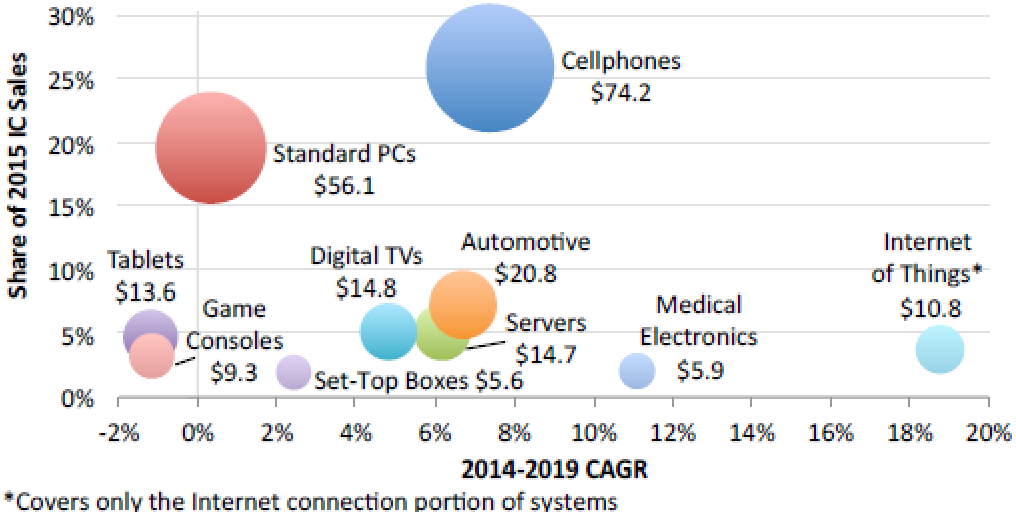
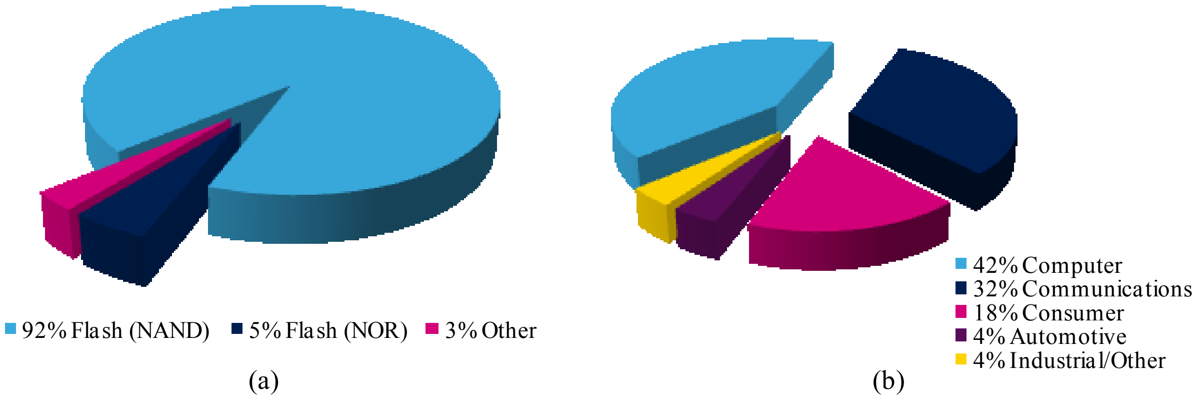


Figure 3: IC market value and current grown rates for different market segments [Ici17]

1.1.1 Memory Market

The majority of the 85 billion USD memory market is still dedicated to DRAM, exemplified by Samsung's 2015 ground-breaking of a $14 billion fab, expected to mainly produce these volatile memories [Joh15]. However, flash memory unit shipments recently surpassed DRAM and represent a significant fraction of the market, 40% [Ici17]. Even if the discussion of NVM in conferences and journals world-wide is dominated by more exotic concepts such as Phase Change Memory (PCM), Magneto-Resistive Random Access Memory (MRAM) and Resistive Random Access Memory (RRAM), the market remains dominated by Flash, at 97%, as shown in Figure 4(a). While these emerging memories have a chance at surpassing flash memories, the industry is loath to make anything larger than incremental changes as long as flash memories can continue to meet their needs.

The NVM market is incredibly diverse and this leads to many different technologies vying for different market segments featuring varying *mission profiles*, or operating conditions. The diversity in products reflects the lack of a universal memory, that is, a memory that is fast, dense, reliable, durable and economical. This leads to the fragmented breakdown in flash memory device types shown in Figure 4(b). Companies are therefore obligated to make trade-offs when selecting a technology for use in each application. The first criterion is the difference in temperature range to be faced by the devices and the second is the speed/performance requirements. For example, a gas sensor in a car's engine has different requirements than a smartcard's chip that will never be used far outside of room temperature. The reliability is also very important for certain applications, for example, corrupted memory in an a camera's SD card may lead to the loss of an image, while in a vehicle's crash-detection sensor the same error could lead to an airbag failure, causing serious injury or death to the occupants. It is therefore not difficult to understand why car manufacturers will therefore pay a premium for higher quality chips.



(a)

(b)

Figure 4: Breakdown of the (a) 39.9 billion USD 2017 (forecasted) NVM Market by device type and (b) Flash memory market breakdown by system type [Ici17]

Within the flash memory market there are two principle categories, NAND and NOR, named for the logic gates that their connections resemble. NAND flash chips are typically higher density for storing large amounts of data in blocks that are read periodically. NOR flash chips are typically lower density for data that is read, written and erased more often, with the former two operations being random access.

The smartphones, solid-state drives and other mobile devices generally incorporate the high density NAND flash that dominates with 92% share of the flash market. Applications where reliability is of upmost importance such as smart cards, automotive and code storage tend to incorporate NOR flash. The trade-off between density and reliability/speed is evident in Figure 5. The majority of NOR devices are 16MB or less while in the case of NAND the vast majority are at least 64GB. It should be noted that NOR flash units are not appropriate solely for embedded applications, nor are NAND flash units only applicable to standalone applications.

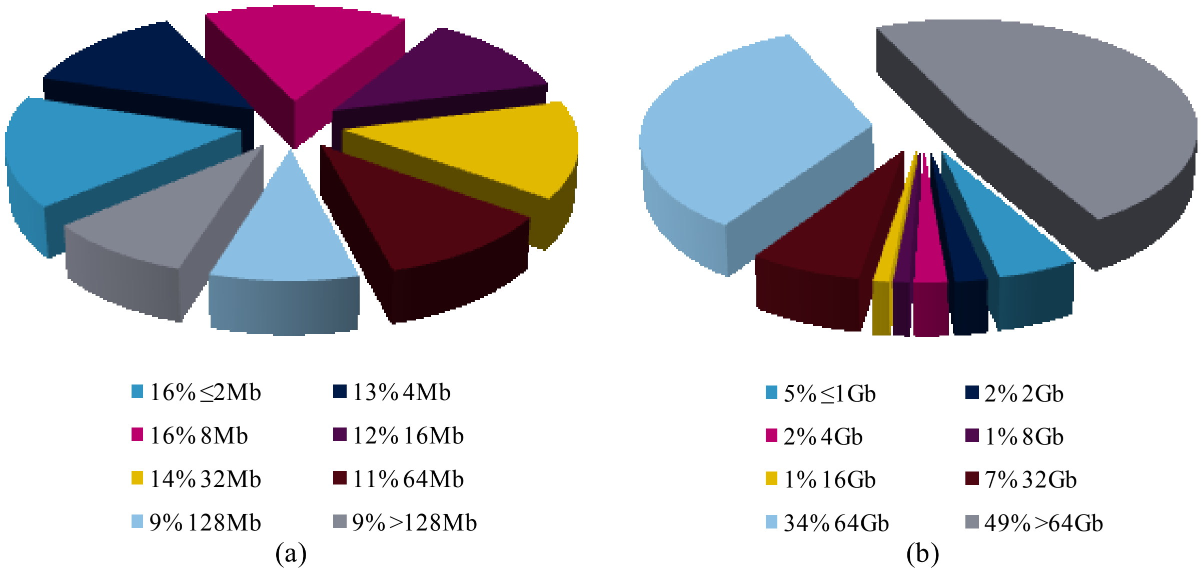


Figure 5: Breakdown by unit size for the (a) 5.5 billion units of NOR flash (b) 12.1 billion units of NAND flash that are forecast to be sold in 2017 [Ici17]

1.2 Working Principles of Flash Memory

One approach to making an NVM is to design a transistor with a threshold voltage that can be modulated. Thus, the fact that a drain current is measured (or not measured) at a given gate voltage is indicative of the state of the bit. By convention, the high and low-threshold states are referred to as programmed and erased, respectively. The read conditions of the cell use a gate voltage in between the two states, as a result a current is only measured in the low-threshold state. The threshold voltage (V_T) of a metal oxide semiconducting field effect transistor (MOSFET) can be written as Equation 1.1 where V_{T_0} is the threshold voltage for a cell with no oxide charge and the second factor is related to the charge distribution in the oxide and the oxide capacitance [Pav97].

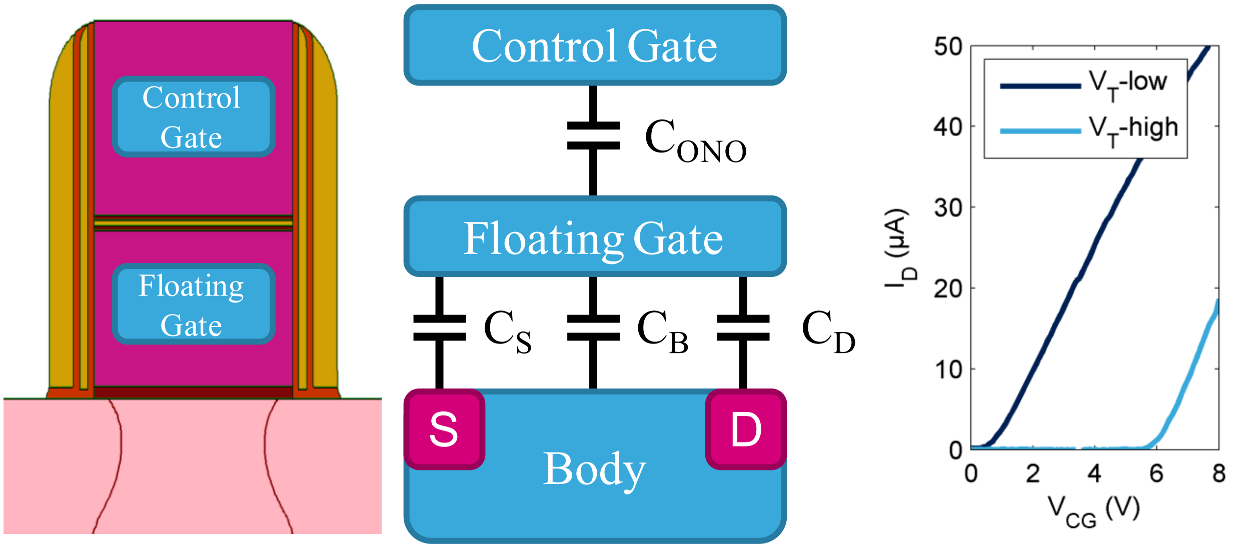
$$V_T = V_{T_0} - \frac{Q}{C_{OX}} \qquad Equation\ 1.1$$

In introductory electronics courses this equation is usually used within the context of the V_T changes introduced by trapped charges in the oxide. However, if engineers can control Q and C_{OX}, then they can make a memory cell with different V_T states.

Two different ways of obtaining a charge, Q, lead to two common types of NVM: charge-trapping and Floating Gate (FG) devices. In the former, charges are trapped in a charge-trapping layer such as Si_3N_4 in SONOS (Silicon Oxide Nitride Oxide Silicon) memories. In the latter, charges are injected into the *floating gate*, a conductor that is insulated from the control gate and the channel. These trapped charges can be expelled from the charge-trapping layer or floating gate, making the memories EEPROMs (Electrically Erasable Programable Memories). Flash memory is a subset of EEPROM that has cells arranged in such a way that many cells are erased in parallel "in a flash" to save time.

1.2.1 Floating Gate Devices

Floating gate devices modulate the threshold voltage by changing the number of electrons stored on a gate that is disconnected from all other electrode, it is "floating". Figure 6(a) shows a diagram of a cross-section of a FG device, while Figure 6(b) shows a simplified electrical model. One can imagine that increasing the number of electrons stored on the FG increases the electrostatic repulsions with electrons in the channel. Thus, a larger gate bias is required to reach inversion and the cell arrives at a higher threshold voltage state (for an NMOS device) in Figure 6(c).



(a) (b) (c)

Figure 6: (a) TCAD cross section of a flash cell (b) capacitive model of the same cell (c) IV curve for a device in both the programmed and erased states

The potential of the floating gate (V_{FG}) is analogous to the control gate of a classical MOSFET in that it determines the conductivity of the channel, and thus, the logic state of the cell. Given that the floating gate is isloated, it cannot be accessed directly; however, it is influenced by capacitive coupling with the terminals of the device. By considering the case where the charge on the floating gate is zero, it is possible to write Equation 2.3 in which Q is the charge on the floating gate, C_{ONO}, C_S, C_B and C_D are the capacitances between the floating gate and the control gate, source, body and drain regions respectively. V_{FG}, V_C, V_S, V_B, and V_D are the floating gate, control gate, source, body, and drain potentials respectively [Pav97].

$$Q = 0 = C_{ONO}\,(V_{FG} - V_C\,) + C_S\,(V_{FG} - V_S\,) + C_B\,(V_{FG} - V_B\,) + C_D\,(V_{FG} - V_D\,)\quad \mathit{Equation}\ 1.2$$

By defining C_T as the sum of the above capacitances and α_j=C_j/C_T as the coupling coefficient for the contact "*j*", one can rewrite Equation 1.2 in the form of Equation 1.3.

$$V_{FG} = \alpha_G V_G + \alpha_S V_S + \alpha_B V_B + \alpha_D V_D \quad \mathit{Equation}\ 1.3$$

For grounded source and bulk contacts, Equation 1.3 becomes Equation 1.4.

$$V_{FG} = \alpha_G \left(V_G + \frac{\alpha_D}{\alpha_G}\, V_D \right) \quad \mathit{Equation}\ 1.4$$

Traditional MOS equations are defined relative to the control gate of the MOS transistor, and for a grounded drain, the threshold voltage of the floating gate is related to the control gate by Equation 1.5.

$$V_T^{FG} = \alpha_G V_T^{CG} \quad Equation\ 1.5$$

For the case where Q is non-zero, Equation 1.5 becomes Equation 1.6

$$V_{FG} = \alpha_G V_G + \alpha_D V_D + \frac{Q}{C_T} \quad Equation\ 1.6$$

Rewriting in terms of the control gate bias, for very small V$_D$ leads to Equation 1.7.

$$V_T^{CG} = \frac{V_T^{FG}}{\alpha_G} - \frac{Q}{C_T \alpha_G} = \frac{1}{\alpha_G}\ V_T^{FG} - \frac{Q}{C_{ONO}} \quad Equation\ 1.7$$

By combining Equation 1.5 into Equation 1.7, one can see that the threshold voltage shift is determined by the charge on the floating gate and the capacitive coupling between the floating gate and the control gate in Equation 1.8.

$$\Delta V_T = V_T - V_{T_O} = -\frac{Q}{C_{ONO}} \quad Equation\ 1.8$$

This relation quickly leads to the question of how the charges are added and/or removed. Electrons can be added to the floating gate by hot-carrier injection through which electrons gain energy in the electric field between the source and drain before undergoing random collisions with a fraction of the electrons entering the floating gate. Electrons can also be added or removed by Fowler-Nordheim tunneling, which uses a high electric field to increase the tunneling probability of electrons to tunnel into the FG [Pav97].

1.2.2 Write Information: Hot Electron Injection

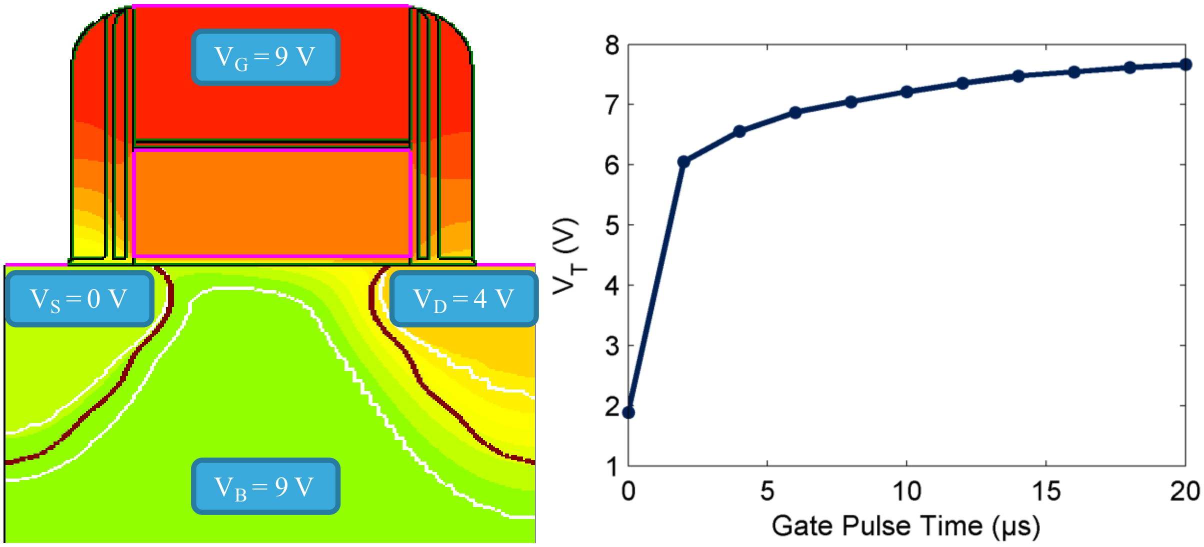
With the channel in inversion and a bias applied to the drain, electrons will travel from the source towards the drain, gaining energy as they move in the lateral electric field. If the bias is sufficiently high, the electrons' kinetic energy gains are no longer kept in equilibrium by energy losses to lattice vibrations and the electron energies relative to the conduction band can increase. These electrons are referred to as "hot" electrons, which can be injected into the floating gate under the following conditions:

1) Their kinetic energy must be higher than the potential barrier of the tunnel oxide.

2) They must be travelling in the direction towards the barrier.

3) The electric field in the tunnel oxide must be collecting the electrons.

An example of these conditions and the resulting change in V_T are shown below in Figure 7.



(a) (b)

Figure 7: (a) Bias conditions and electrostatic potentials and (b) V_T (t) curve for programming a flash cell

The quantitative calculation of the electron injection across the tunnel oxide is very non-trivial as it requires the knowledge of the electron energy and momentum distributions across the channel, the shape and height of the tunnel barrier and the probability that an electron can cross the barrier with a given wave vector, energy and distance from the interface. It is further complicated when impact-ionization becomes an additional energy-loss mechanism [Pav97].

Two different approaches can facilitate the calculation of the injection probabilities. The first is the "lucky electron" model, which relies on the electron being able to travel far enough between collisions that it gains enough kinetic energy to cross the tunnel barrier, provided that the next collision directs it towards the barrier. Finally, the electron must follow a collision-free trajectory towards the interface. Each of the events has an associated probability and they combine to yield a lumped probability, which can be used to simulate the injection current [Pav97].

A second, more rigorous approach relies on the assumption that the electrons can be treated as a gas in quasi-thermal equilibrium with the electric field, characterized by an effective temperature. This appraoch uses the non-local relation between the effective electron temperature and the drift-field to calculate the probabilities of injection [Pav97].

Hot Carrier Injection (HCI) is fast, with cells being programmed in a few microseconds. However, it is also requires relatively high currents because only a fraction of the electrons are injected into the floating gate. In the example in Figure 7, the integrated drain current of approximately 80 μA over 3 μs is ≈2.5 nC, while the V_T-shift of 6 V is obtained by an increase of less than 1 fC; only about 1 in 10^{-6} electrons are injected into the floating gate. This inefficiency makes this type of programming more appropriate for applications where write speeds are more important than power consumption.

1.2.3 Erase Information: Fowler-Nordheim Tunneling

The second method of charge transfer to the floating gate is by quantum mechanical tunneling through the tunnel oxide. By applying a large bias across the oxide, the tunneling distance across the triangular barrier from the channel into the conduction band of the oxide decreases, allowing electrons to flow. An example of this situation is shown in Figure 8, as an electron tunnels from the n^+ poly-Si floating gate on the left side across the tunnel oxide towards the p-type silicon substrate. Programming by this mechanism is also possible, it is often used in NAND flash.

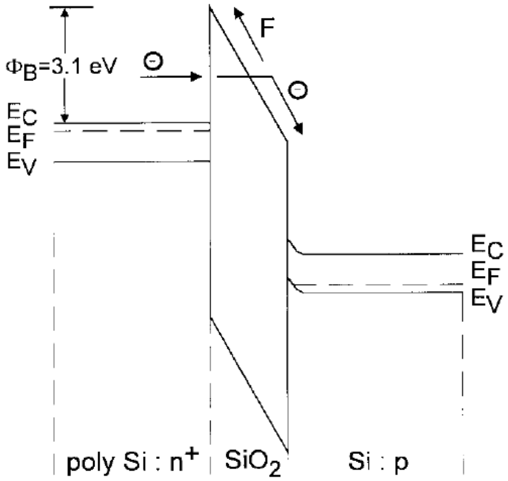
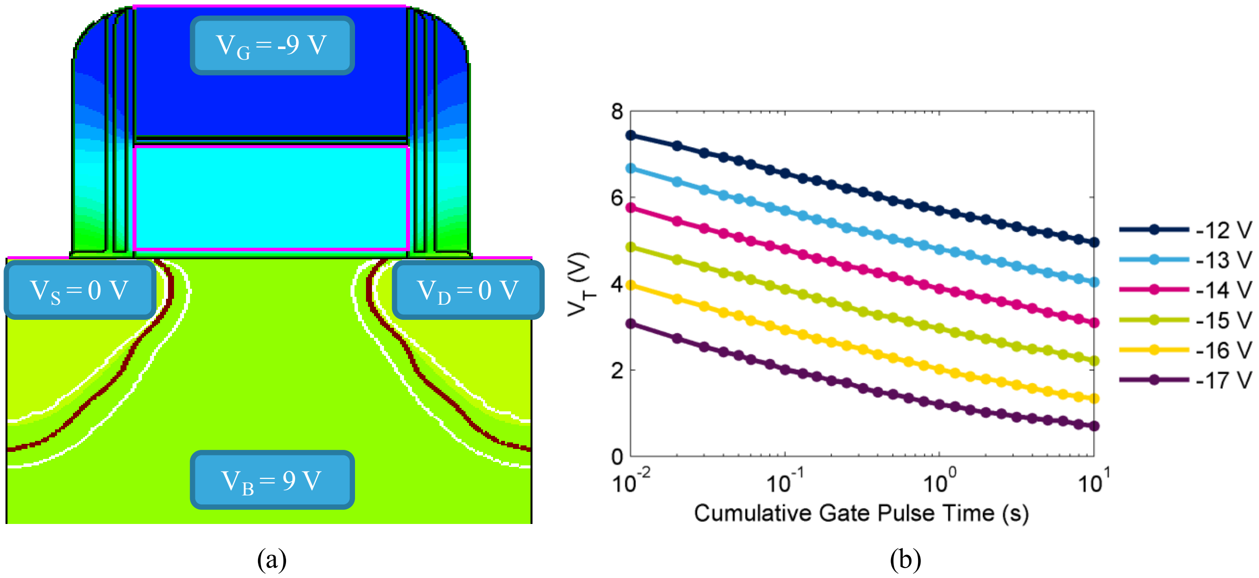


Figure 8: Fowler-Nordheim tunneling in a MOS structure [Pav97]

In practice, the erase operation is completed by applying a large bias between the body and the control gate of the flash cell, which slowly empties the stored electrons from the floating gate as shown in Figure 9. The erase speed is much slower than the writing speed, as a result this is one of the reasons that flash cells are erased in large blocks, or pages. Increasing the biases used for the erase operation can increase the speed, at the cost of being obligated to generate high voltage pulses.



(a) (b)

Figure 9: (a) Bias conditions and electrostatic potentials and (b) V_T (t) curve for erasing a flash cell at different gate biases for a grounded body

Using the Wentzel-Kramers-Brillouin (WKB) approximation for quantum mechanical tunneling and treating the poly-Si gate like a metal, the current density can be described by Equation 1.9 where q is the charge on an electron, F is the electric field, h is Planck's constant, Φ_B is the barrier height, m_{ox} is the effective electron mass in the oxide and J is the current density.

$$J = \frac{q^3F^2}{16\pi^2h^2\Phi_B}e^{-\frac{4(2m_{ox}^*)^{\frac{1}{2}}\Phi_B^{\frac{3}{2}}}{3\hbar qF}} \quad \mathit{Equation}\ 1.9$$

This equation is often written in the simpler form of Equation 1.10, that takes into account the fact that barrier height depends on the voltage and that the oxide field is reduced due to the voltage drop in the substrate; A and B are functions of the electric field [Pav97].

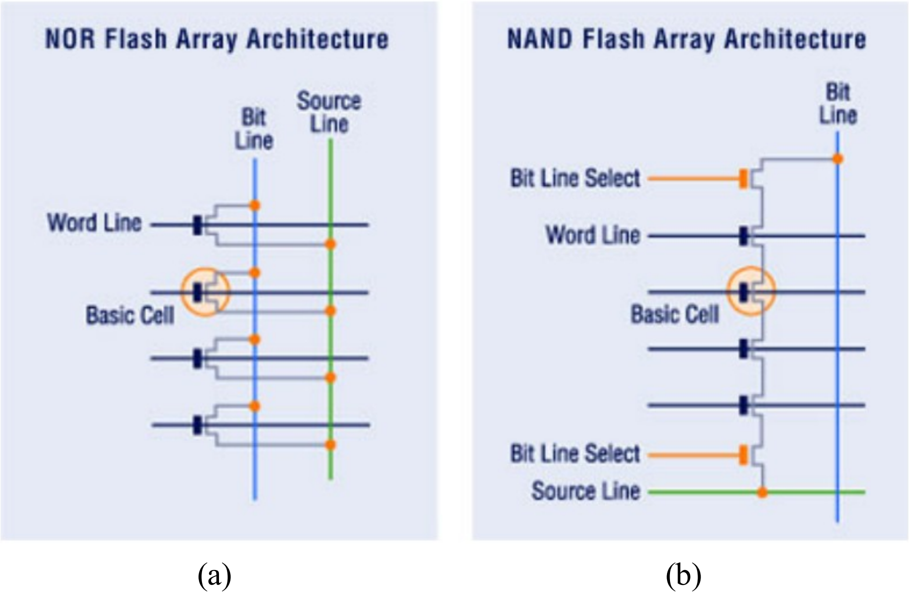
$$J = A\cdot F^2\cdot \mathrm{e}^{-\frac{\mathrm{B}}{\mathrm{F}}} \quad \mathit{Equation}\ 1.10$$

As the current scales with the electric field, the oxide thickness and applied bias must be chosen carefully to balance two important but opposing characteristics. These two characteristics are the high current levels required for programming/erasing and the low current levels required for data retention. For example, a cell having a 5 nm thick tunnel oxide would be programmed and erased easily; however the leakage currents would not meet the data retention standard. The lower limit for tunnel oxide scaling is often cited as 8 nm because a leakage patch could be provided by a single defect in a thinner oxide [Deg04].

Fowler-Nordheim Tunneling is much slower than HCI, the program operation takes milliseconds. The mechanism is also used to erase the cells, often a few kbits at the same time in order to overcome the inherent sluggishness of Fowler-Nordheim tunneling. For applications where low-power operation is more important than high speeds, the Fowler-Nordheim tunnaling programming method is well suited as the write currents are lower than those required for HCI. This is because the programming efficiency is ≈100% as electrons are injected directly from the channel into the floating gate.

1.2.4 NAND, NOR and Embedded Flash

There are many different types of memory architectures that are considered to be Flash Memory, which vary in the way that the cells are programmed/erased and how they are accessed. For example, cells can be altered by Fowler-Nordheim tunneling, channel hot electrons, hot holes or source-side hot electrons in either a series or parallel configuration. Two of these have come to be considered "standard" in the microelectronics industry, NOR and NAND, shown in Figure 10.



(a) (b)

Figure 10: (a) NOR and (b) NAND flash array architectures [Tal02]

In general, the first, NOR, is more likely to be used for code storage while the second type, NAND, dominates the mass-storage market [Bez03]. NOR is more adapted to higher performance applications as each cell has three contacts, allowing for faster and random access. NAND cells are organized in a chain, without contacts between the word lines, as shown in

Figure 11. This increase in density is offset by slower access times as individual cells cannot be randomly accessed.

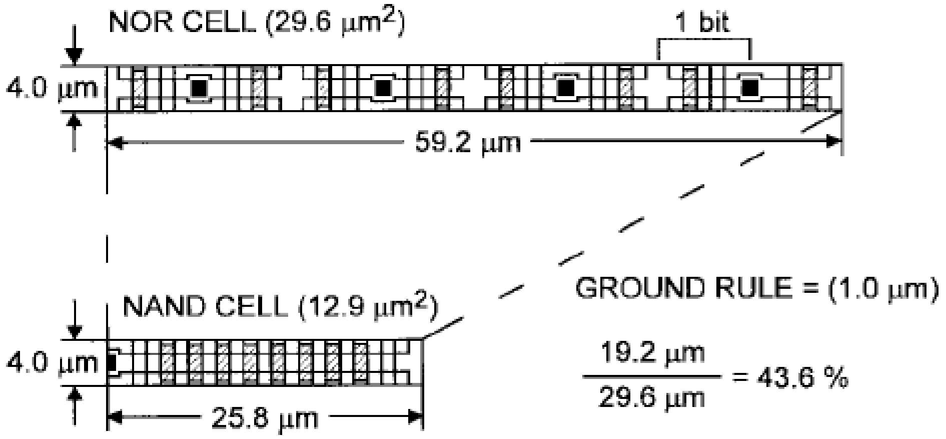


Figure 11: 8 bits arranged in NOR and NAND configurations showing the density differences [Pav97]

The density of NAND chips can also be increased by storing multiple bits on a single cell by way of multimodal V_T distributions. The current leaders in standalone flash memory have recently succeeded in turning the word line so that it is vertical, therefore stacking the memory layers. These 3D-NAND or V-NAND chips achieve densities of >3 GB/mm^2 to form up to 256 GB chips by stacking 36-48 layers, sometimes with multiple bits per cell [Mea16].

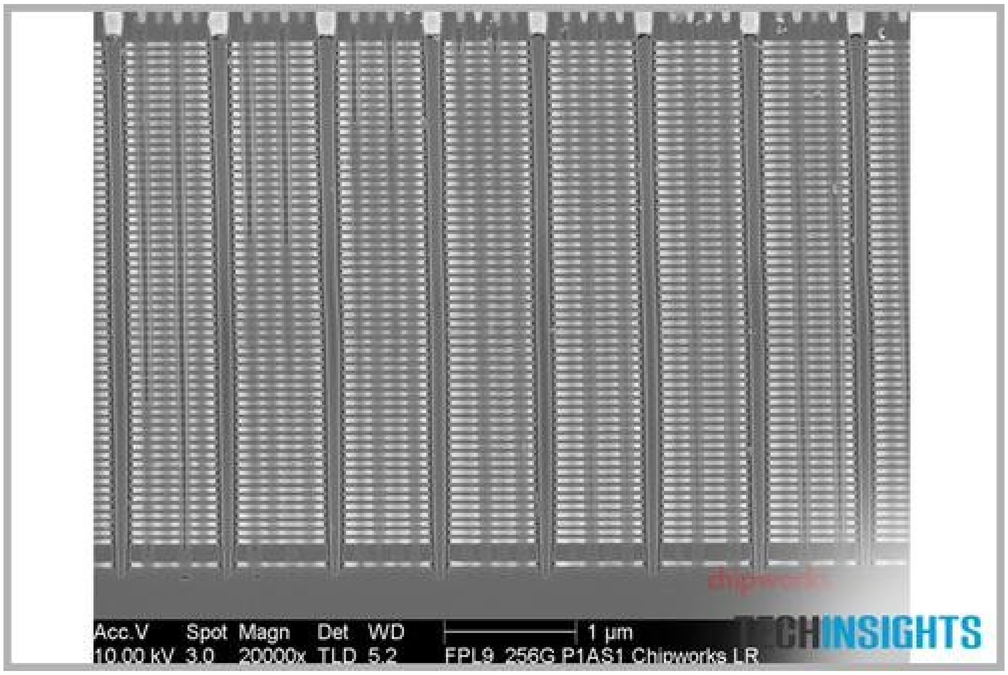

Acc.V Spot Magn Det WD 1 μm
10.00 kV 3.0 20000x TLD 5.2 FPL9 256G P1AS1 Chipworks LR
CHINSIGHTS

Figure 12: Cross section of 48-layer Vertical NAND by Sandisk/Toshiba for the iPhone 7 [Cho16]

As one moves away from high density chips for Solid-State Drives and flash cards, the ability to randomly access data drives the decision to use NOR flash even though it cannot be made as dense as NAND flash. For System on Chip or embedded applications, multiple types of devices are integrated on the same wafer during fabrication, making the co-existence of the flash devices and logic circuits more important than solely their density. The inclusion of multiple devices on one chip avoids the need to package multiple chips together and allows manufacturers to offer a more complete product. Access times are also reduced as the devices are located on the same chip. For these embedded memories, the density is less important as the memory acts to support the chip's function instead of being the sole function. Embedded memories can be found in products like microcontrollers, smart cards and automotive chips.

Each application has its own mission profile to describe the conditions under which the product must work. For example an engine sensor in a vehicle must work at start-up in the winter and while the engine is running in the summer. Therefore a standalone flash chip with an operating temperature from 0°C to 70°C is not sufficient for the automotive industry but well-suited for a personal computer's storage. Reliability is another area where embedded flash outshines standalone memories; a failing crash detection sensor in a car presents a much greater risk than a corrupted memory card in a camera and therefore demands a more reliable chip. Standard mission profiles for different applications are summarized in Table 1.

Table 1: Industry-standard mission profiles for embedded NOR flash

	Automotive	Smart Card	Microcontrollers
Data Retention	30 years at 80 °C	10 years at 80 °C	10 years at 80 °C
Erase Time	500 ms	1 ms	500 ms
Operating Temp.	150-160 °C	80 °C	125 °C
Endurance (in cycles)	500k	500k – 1M	100k

The co-integration of flash and CMOS logic presents a very challenging engineering problem due to the sequential nature of the semiconductor fabrication process; either the CMOS or the flash will end up being subjected to a higher than normal thermal budget and extra chemical deposition/removal steps. In addition to the flash cells, high voltage (HV) MOS circuits are also required to generate the voltage pulses that are used during program and erase. Each of these three types of devices (HV, logic and flash) would ideally have their own optimized oxides and gates.

1.3 Motivation to use high-k in Flash

From the 1960s until 2007 when Intel demonstrated 45 nm-node high-k metal gate transistor based processors, the microelectronics industry was dominated by silicon/silicon dioxide based technologies. Silicon dioxide worked very well because it was a high quality, native oxide of the readily-available silicon wafers that could be grown or deposited and then etched with good selectivity. However, as the gate oxides became thinner and thinner the gate leakage currents increased until the leakage of the "OFF" transistors became a large problem [Rob15]. These very thin gate oxides were used to attain the high gate capacitances required for high-speed switching. The gate capacitance can be described by Equation 1.11 where *A* and *t* are the area and thickness of the gate, respectively while ε_o is the vacuum permittivity and *k* (often written as the Greek, κ) is relative permittivity or the dielectric constant.

–

High-k materials, having a dielectric constant higher than the 3.9 of $SiO_{2,}$ were introduced because they could provide higher gate capacitance, without resorting to thinner oxides, thus avoiding the high off-state currents. This change in material led to the use of the Effective Oxide Thickness (EOT) as a way to describe a layer. The EOT is the thickness of SiO_2 that would give the same areal capacitance and is described in Equation 1.12 where t is the thickness and k is the dielectric constant.

———

The introduction of high-k materials into transistors faced many technical hurdles such as finding stable materials and dealing with fermi-level pinning. After realizing stable high-k materials, metal gates were introduced at the same time to avoid Poly-Si/high-k reactions and to better control the potential in the channels of the transistors [Rob15]. The motivations for using high-k materials in flash cells are similar for the transistor, as will be explained in the following subsections. Each of these motivations is related to the common goals shared between embedded NOR flash chips:

- Good coupling between the floating gate and the control gate for improved erase performance (faster and/or at lower voltages).
- Ultra-low leakage currents to maintain >10 years of data retention.
- Economically feasible process (limiting the number of masks and Poly-Si depositions).
- Good co-integration with core CMOS logic processes and their design libraries.

1.3.1 Increased coupling without increasing leakage

A critical parameter of a flash cell is the coupling coefficient between the control gate and the floating gate, ($\alpha_{CG} = C_{ONO}/C_T$) as described in Section 1.2.1. As the erase by Fowler-Nordheim tunneling is driven by the electric field in the tunnel oxide and V_{FG} is proportional to $\alpha_{CG}V_{CG}$, an increase in the coupling factor can have a positive impact on erase performance. Figure 13 shows the TCAD simulated V_T-erase values as a function of control gate bias for three different ONO EOT values. The simulations show that if the EOT is decreased by about 10%, the same erase speed can be reached with a control gate bias that is 1 V less. By decreasing the maximum gate bias required during cell operation, the requirements of the charge pumping circuitry used to generate the program/erase pulses are reduced. In practice, the chipmaker would need to decide if reducing the control gate erase pulses and keeping the same performance is more desirable than maintaining the same gate pulses and increasing erase speed.

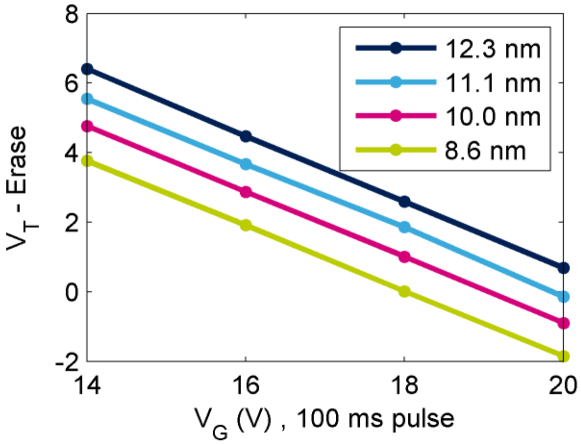


Figure 13: TCAD simulated V_T-erase values after a 100 ms gate pulse with different ONO EOTs. Initial $V_T = 8$ V

Based on Equation 1.11, there are three ways to increase the capacitance: increase the area, decrease the dielectric thickness or increase the dielectric constant. Increasing the area of the cell is not an option as this would run counter to the trend of device scaling in the microelectronics industry. Decreasing the thickness of the ONO too much will lead to leakage current problems. It is therefore of interest attain a higher capacitance by changing the dielectric to another material having a higher k-value.

In general, higher-k materials have smaller band gaps as shown in Figure 14, which complicates the material selection. For example, titanium dioxide has a very high k value, around 80, however it has almost no conduction band offset with silicon, which makes it a very poor barrier to electron conduction [Cas02]. In addition, the defects in high-k layers can contribute to leakage currents that can be consequential enough to lead to data-retention failures [Pad08,Lar11].

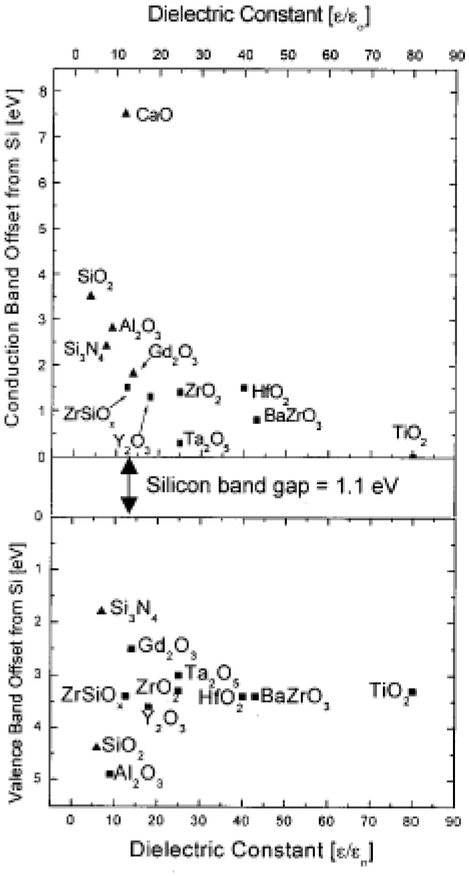


Figure 14: Conduction and Valence band offsets vs dielectric constant for different materials [Cas02]

1.3.2 Co-integration with high-k metal gate CMOS logic processes

For embedded flash memories, the composition of the logic devices must also be taken into account, as process steps are often shared to reduce thermal budgets and/or the number of lithography masks. For example, in STMicroelectronics 65nm embedded flash process; the same poly-Si layer that forms the gate of the HV transistors also forms the floating gate of the flash cells while a second poly-Si simultaneously forms the control gate of the flash and the gate of the logic transistors shown in Figure 15 [Pia10]. It would be possible to make each of these gates individually, but having a process flow with two poly-Si depositions is more economical than having one with four.

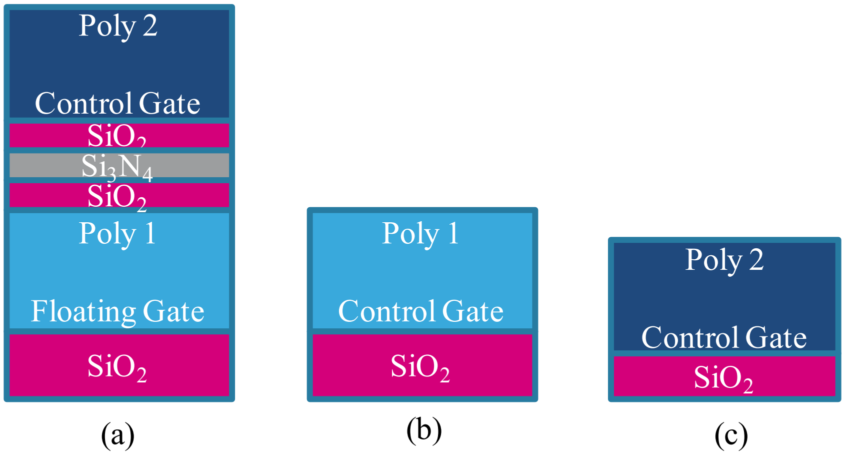


Figure 15: (a) Flash, (b) High Voltage and (c) Logic devices in a two-poly scheme [Pia10]

In the case of Figure 15, the ONO is deposited over the whole wafer and then removed in the HV and logic zones before the gate oxide is grown by oxidation and the second poly-Si is deposited. Another way to think about this co-integration is to see that the flash memory stack is terminated by the gate stack of the logic transistors. When the embedded flash moves towards a technology node whose logic transistors normally include high-k dielectrics and metal gates, it becomes necessary to think about the use of high-k materials in flash devices. Conversely, while the logic gates are poly-Si, there is a barrier to the introduction of high-k materials in the flash devices due to worries about contamination in the cleanroom.

1.4 The use of high-k materials in flash

In the early 1970s the first reprogrammable floating gate NVMs were produced with a basic structure similar to that of Figure 16, with tunnel oxides and IGDs on the order of 85 and 400 nm of SiO_2 respectively [Tar72]. Although other configurations – such as SuperFlash [Mic17a] and SplitGate

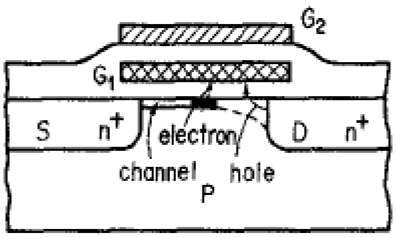


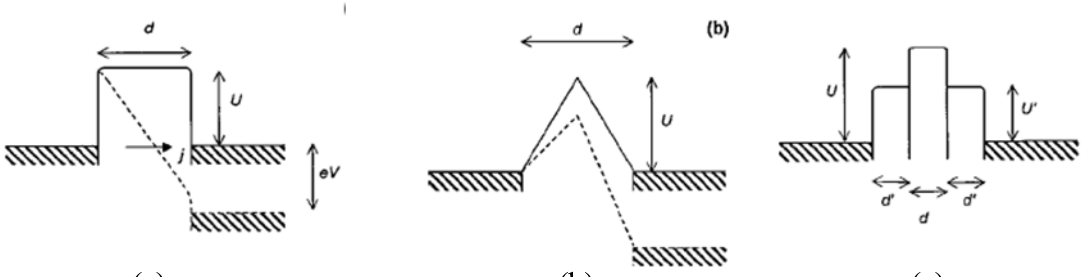
Figure 16: Early Flash cell from 1972 [Tar72]

[Shu12] – have been used in industry, this configuration is still relevant forty years later. By the early 1980s, researchers discovered that $SiO_2/Si_3N_4/SiO_2$ tri-layers were more robust than SiO_2 single-layers in time zero breakdown tests [Wat84], in addition to offering thinner EOTs for a stack with the same physical thickness and reduced thermal budgets [Ger85]. Silicon nitride has a higher-k value than silicon dioxide, 7.6 vs 3.9, so this where flash devices started to include high-k materials. However, in microelectronics parlance, silicon nitride is a very commonly used material in the cleanroom and did not represent the same sort of revolutionary jump that the move to high-k/metal gate structures that did.

Although this thesis focuses on the integration of high-k materials for the IGD, the concepts introduced during high-k tunnel oxide research are interesting and will also be discussed Section 1.4.1. Section 1.4.2 introduces the V_T instability problem that comes with high-k materials integration before section 1.4.3 summarizes some of the research into high-k materials for flash.

1.4.1 Multi-layers for bias dependent tunnel barriers

While working on the tunnel barrier in 1998, Likharev proposed a barrier structure that is dependent on the applied bias. This structure is attractive because the barrier can decrease with applied bias, making program and erase easier while keeping good retention characteristics at low biases. This is clearly more applicable to a tunnel oxide than an IGD as the latter should always insulate. The barrier structure is explained below in Figure 17, as the height of a (a) square SiO_2 barrier does not significantly change with applied bias while a (b) triangular barrier's height is strongly affected. As it would be difficult to implement a truly triangular barrier, (c) the step-wise barrier is a more realistic implementation [Lik98].



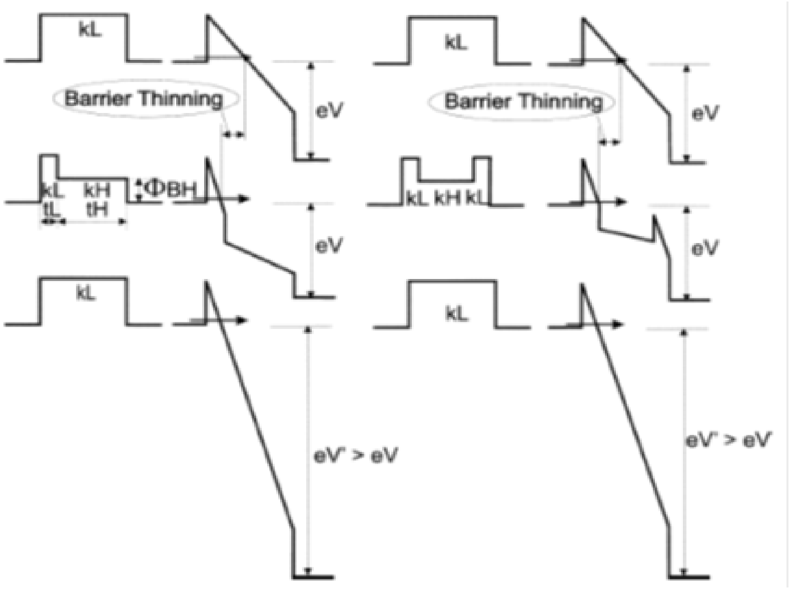
(a) (b) (c)

Figure 17: (a) Square, (b) Triangular and (c) Step-Wise barriers proposed by Likharev in 1998

The tri-layer barrier structure is not only much easier to fabricate than a triangular barrier, but it was shown, via simulation by Casperson in 2002, to have similar tunneling properties to an ideal triangular barrier. They suggested $Si_3N_4/Al_2O_3/Si_3N_4$ and $ZrSiO_x/Al_2O_3/ZrSiO_x$ as likely candidates [Cas02], neither of which has entered current use.

By the early 2000s, Govoreanu et al. presented a model for tunnel oxide currents in multilayer dielectric stacks and concluded that the most promising two-layer stacks (three-layer stacks were not yet considered) were a combination of SiO2 and a material with a moderate barrier height and significantly larger relative permittivity [Gov03a]. Their modeling work was extended to trilayer stacks and denoted "VARIOT" for VARIable Oxide Thickness, as the effective tunneling thickness is variable.

Based on Gauss' Law, which states that the net electric flux through any closed surface is equal to 1⁄ε times the net electric charge enclosed within that surface, one can see that electric-field strengths (and band structures) are dependent on the materials. This leads to weaker fields (more shallow band slopes) in high-k materials as the product of εE (the dielectric permittivity and the electric field) is continuous at an interface with no trapped charge. As shown Figure 18, low-k/high-k or low-k/high-k/low-k stacks can be used to yield barriers that are small under large biases (program/erase) but present long tunneling distances at low biases (retention) [Gov03b].



(a) (b)

Figure 18: (a) Two and (b) three layered tunneling barriers showing the redistribution of the potential drop in stacks with different relative permittivity values (kL = low-k, kH = high-k values) [Gov03b]

Later that year, Govoreanu et al. fabricated Si/SiO_2/ZrO_2/TiN capacitors which confirmed their predictions about low-k/high-k stacks. Under large applied biases the tunneling current through the stack is mainly controlled by the thin SiO2 layer, whereas at low bias the thick two-layer tunneling barrier significantly reduces the current, shown in Figure 19. Once Govoreanu's hypothesis had been validated, it was then possible to propose a wide variety of barriers.

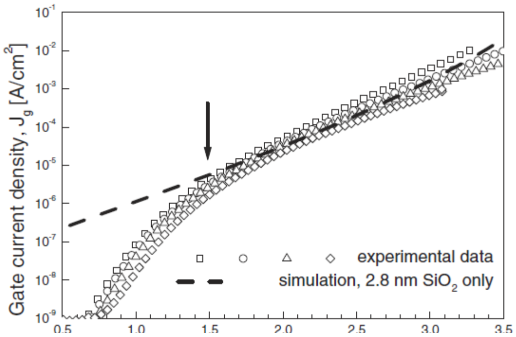
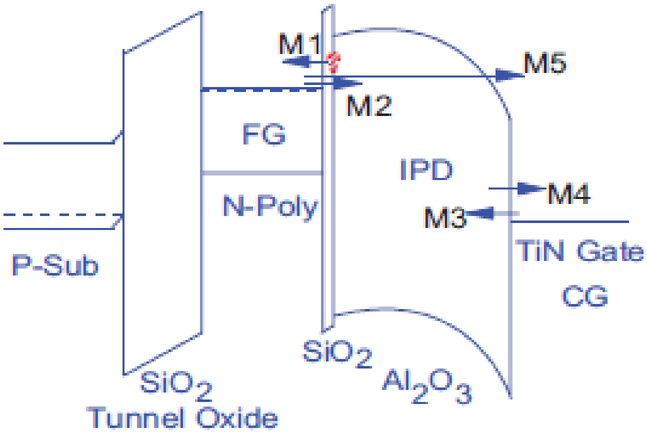


Figure 19:IV for SiO_2/ZrO_2 bi-layers [Gov03c]

1.4.2 Flash threshold voltage instabilities

As the interfaces between the ONO and the gates were of good quality the V_T-instability related to trapping/detrapping at these interfaces was not problem. These interfaces can become a problem when the ONO is replaced by high-k materials. Errors in cell verification associated with electrons redistributing as in Figure 20, after programming, erasing or reading.



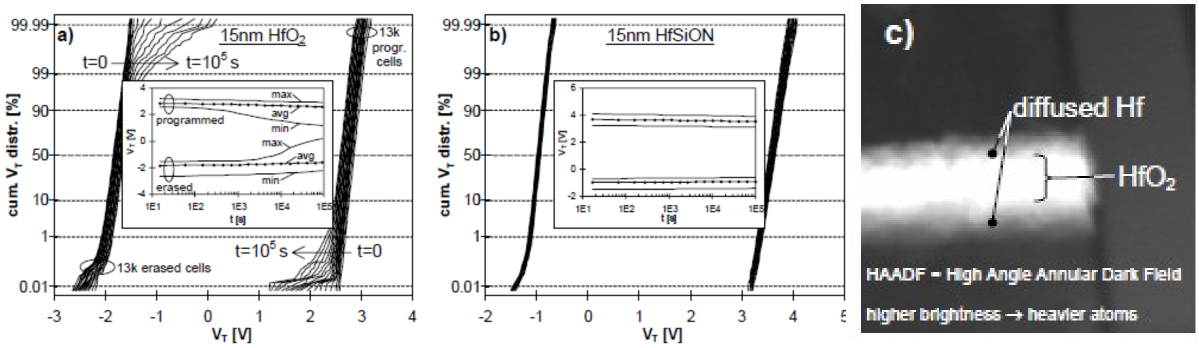
M1) detrapping from IGD to FG
M2) trapping from FG to IGD
M3) trapping from CG to IGD
M4) detrapping from IGD to CG
M5) charge loss across the IGD

Figure 20: Flash cell band structure with five possible mechanisms that would modify the V_T, electron movement towards the control gate leads to a decrease in V_T and towards the channel it is an increase in V_T [Tan13]

The relative importance of the possible mechanisms is governed by the trap characteristics. For example, with an Al_2O_3 IGD, there are many shallow traps near the interface (with SiO_2), while there are few such at the interface of $HfAlO_x$ so electron detrapping from the IPD to the FG is important in the former, but not the latter case. On the other hand, the smaller band-offset and numerous shallow traps of $HfAlO_x$ lead to discharging from the FG to the CG via the IGD. Therefore it was of interest to try to combine the two properties as a possible solution such as an $HfAlO_x/Al_2O_3/$ $HfAlO_x$ trilayer [Tan13].

1.4.3 State of the Art for high-k materials in flash

The idea of using high-k materials in flash started in the late 90s – a 10 nm Al_2O_3 IGD was reported by Lee et al. from Bell Laboratories in 1997. Lee predicted that Al_2O_3 could have similar leakage currents as a 20 nm ONO while erasing three orders of magnitudes faster than a 15 nm ONO [Lee97]. The first reported use of a high-k material for the IGD was by van Duuren et al. in 2006. They fabricated 26k arrays of 2T-FN/FN cells at the 0.18 µm technology node using both HfO_2 or HfSiON (Hf/Si=47/53) deposited by MOCVD with EOT values between 5 and 10 nm along with two reference ONO splits (5-5-5 and 6-6-6 nm). Both high-k materials were able to reduce the Program/Erase voltages by more than 3V (10 ms pulses for a V_T window of 4.2V) with good endurance. HfSiON performed better than HfO_2, due to less of a reaction with the polysilicon gate, shown in Figure 21, and the lifetime was extrapolated to be 10 years at 80 °C [Van06].



(a) (b) (c)

Figure 21: Cumulative V_T distributions from time = 0 to 10^5 s showing (a) significant changes in the HfO_2 sample (b) no changes in the HfSiON sample and (c) a TEM micrograph showing intermixing between the poly-Si control gate and the HfO_2 that is believed to be the origin of the poor performance [Van06]

At roughly the same time, Blomme et al. presented “the first stacked gate memory cells with engineered tunnel barrier and alumina interpoly dielectric”. These were single stacks of tri-layer tunnel barrier, poly-Si floating gate, $SiO_2/$ Al_2O_3 (10/80 Å) IGD and a poly-Si control gate. They achieved program/erase voltages of 9V, which, in comparison to industry standard cells, requiring 16-19V, represented a huge improvement [Blo06].

Based on the work of van Duuren and Blomme, Miranda et al. produced a comparison between Al_2O_3 and HfSiON using the same flash array that van Duuren and Blomme used. The motivation to use Al_2O_3 the IGD was the higher conduction band offset, in the hope that it would provide a greater tunneling barrier for high temperature data retention and reduce the reaction with the polysilicon. Again, the program/erase voltages were decreased by 2-3 V (to get a 4.2V window in 10 ms) with HfSiON giving a bigger improvement compared to Al_2O_3, shown in Figure 22, which is expected given that the EOT is smaller. V_T spread increases with Al_2O_3 thickness, which had a wider distribution than HfSiON because the Al_2O_3 deposition was not very uniform. The motivation for using Al_2O_3 was validated as 15 nm Al_2O_3 met retention standards of 10 years at 125 °C (V_T window loss of 1V) [Mir06].

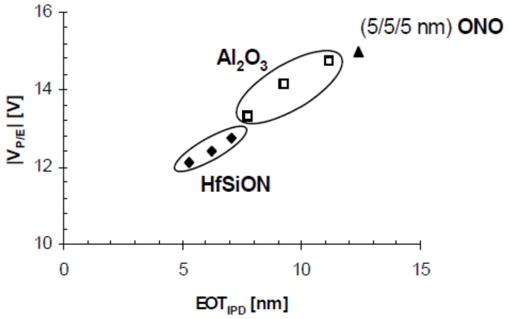
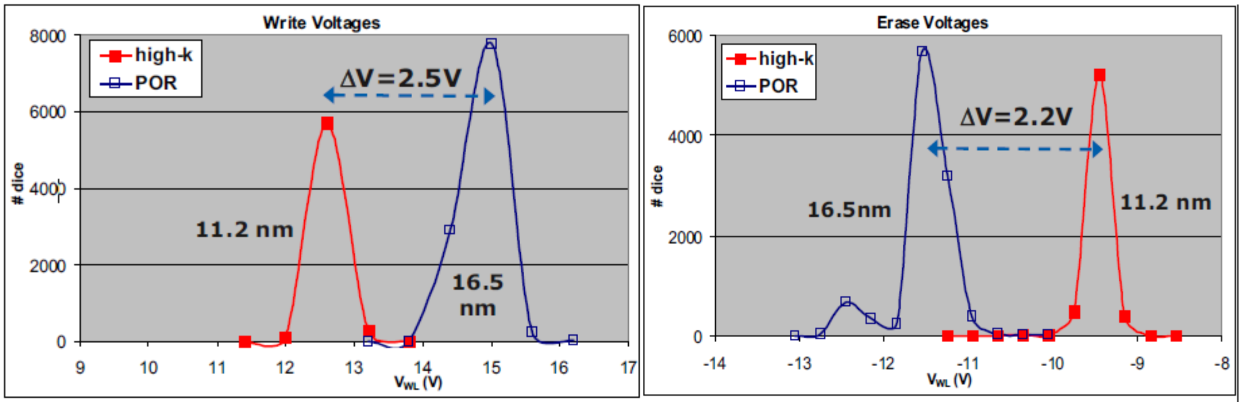


Figure 22: V_P and V_E vs EOT for HfSiON and Al_2O_3 layers from 15 to 25 nm thick [Mir06]

Wellekens et al. demonstrated fully-planarized flash cells with IGDs of SiO_2/ Al_2O_3 (SiO_2 ranged from HF-last to 5 nm / and Al_2O_3 ranged from 7 to 12 nm) and determined that data retention is determined by room temperature charge loss, with the bottom oxide thickness being a key parameter, 5 nm blocks charge loss even at high temperature. Poly-gate and post-deposition anneal both increase retention. Erase saturation (charges injected from the control gate to the floating gate that compensate the loss of charges traversing the tunnel oxide during erase) can be reduced by moving to a TiN gate, at a cost of worse retention, possibly due to adsorbed water that doesn't degas before TiN deposition or damage from the PVD [Wel06].

Power et al. reported program/erase and reliability data on a 2Mb demonstrator using 0.130 µm technology with $SiO_2/Al_2O_3/SiO_2$ (5/7/2 nm) IGDs. Retention was strongly affected by the top-oxide, as it serves to protect the underlying alumina layer [Pow07]. However only a year later at the same conference they present that the top oxide is not necessary. In the original work the alumina was not sufficiently crystallized and therefore was susceptible to chemical attack. The bottom oxide (4 nm in this work) is required to suppress electron injection [Pow08].

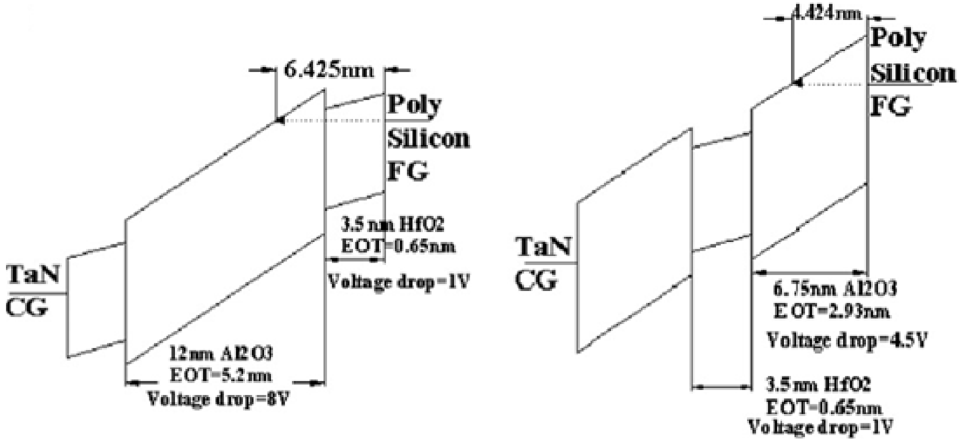
Kakoschke et al. produced 2Mb arrays using 130 nm eFlash technology with SiO_2/Al_2O_3 (1 to 5.5 nm of SiO_2, overall EOTs from 6.0 to 8.5 nm) IGDs. They were able to decrease program/erase voltages by about 3V. Careful IGD thickness optimization was required to avoid early write/erase saturation, minimize tail bits and have good retention, with 4 nm SiO_2 / 10.4 nm of Al_2O_3 being a strong candidate. They showed room for process optimization as there was significant oxide regrowth at the edge of the IGD; they believed that their post-deposition anneal was too short to fully crystallize the alumina [Kak08]. Shum et al. also produced a 400Kbyte demonstrator at 0.13µm node with the program and erase distributions shown in Figure 23. Both distributions were shifted by at least 2 V for a programming window almost 5 V wider. They also tested Al_2O_3 thin films from ozone and water precursors, sometimes with additional Si. High deposition temperature and low silicon content give the best films. The study claimed that water give the best interface as there is less continued oxide growth [Shu09].



(a) (b)

Figure 23: The 0.13 µm technology product 400 Kbyte product demonstrator's write and erase distributions for the POR (16 nm EOT ONO) and high0k (12 nm EOT Al_2O_3) [Shu09]

The trilayers became more appealing for IGD applications in 2008 when Zhang et al. broke from the idea that one should sandwich the high-k material by two high band-gap (low-k) materials such as ONO, then modeled and fabricated these and their opposite, the high-low-high barrier. They showed that $HfO_2/Al_2O3/HfO_2$ and $Al2O_3/HfO_2/Al2O_3$ were both better than single-layers, and that the high-low-high (HAH) barrier performed better in terms of leakage currents than the low-high-low (AHA) barrier, likely due to increased direct-tunneling distances, as described in Figure 24 [Zha08].

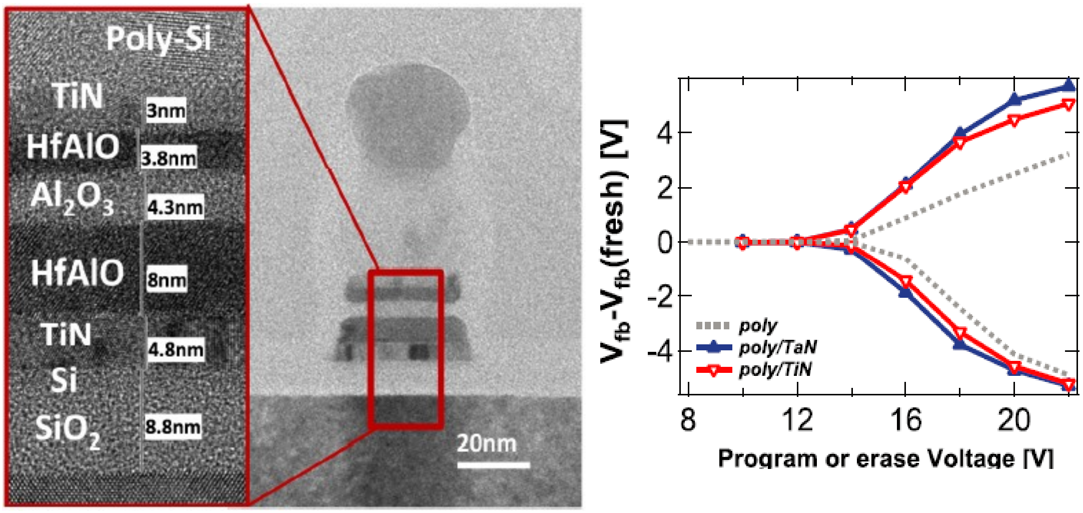


(a) (b)

Figure 24: Simulated band diagram for a 10 V bias on a (a) $HfO_2/Al_2O3/HfO_2$ and a (b) $Al2O_3/HfO_2/Al2O_3$ IGD

In the mid-early 2010s the focus for NAND flash shifted to multi-layers with very low (single-digit) EOTs in order to meet the decreasing C_{IGD} constraints for planar cells. The move to planar cells requires a jump in EOT decreases to make up for the lost capacitance from the sidewalls as the control gate wraps around the floating gate in non-planar geometries. Multi-level cells also require the programming window to be as large as possible in order to separate multiple distributions. This series of tests was often done using a capacitor vehicle outlined by Wellekens et al. that greatly simplifies the fabrication process [Wel11]. The capacitor test structure is a p-type substrate / tunnel oxide / floating gate / IGD / control gate structure where a single, non-selective etch step can pattern the whole gate stack. This greatly reduces the complexity of the fabrication of the cells.

After testing stacks such as HfO_2 [Lin11], Al_2O_3, HfGdO [Bre14a], ScAlO, GdScO [Lis13], LaAlO [Jay10], YAlO [Hua10,Hua11], $Al_2O_3/La_2O_3/Al_2O_3$ [Kim10], $HfAlO/SiO_2/HfAlO$, and a leading multilayer is $HfAlO/Al_2O_3/HfAlO$ [Bre14b] as shown in Figure 25(a). The hybrid poly-Si/metal floating gate is used to control program saturation [Blo10] while the metal control gate is used to reduce erase saturation [Bre15], exhibited in Figure 25(b).



(a) (b)

Figure 25: (a) TEM cross section of 20nm planar NAND cell cross having a high-k multilayer IGD and hybrid Poly-Si/TiN floating and control gates [Bre14b] (b) Example of the program and erase saturation that the hybrid floating gate was introduced to avoid [Blo10]

Although interesting, the cell described in Figure 25 would not be particularly useful for the 40 nm embedded flash discussed in this thesis for two reasons. The first reason is that the current cells have a control gate that wraps around the floating gate, and they are single level cells; an EOT of 5-6 nm is not necessary. The second reason is that program saturation at large gate biases during FN programming are does not occur as the cells are programmed by HCI.

As described in Section 1.2.4, the leading chipmakers in standalone flash are not making use of these advances for planar NAND and have instead moved to vertical, or 3D-NAND as their way to increase density/reduce cost. SanDisk/Toshiba used an ONO-based charge trapping [Mel17] while Samsung uses TANOS-based charge trapping. Although Micron has made 16nm planar NAND with a high-k IGD [Keu15], they have also adopted a vertical configuration of floating gate memory for their latest products [Mic17b]. At the moment, the major players have all announced 64 layer vertical NAND [Mea16] so it remains to be seen if the manufacturers move towards the 2D NAND technical advances such as hybrid floating gates and high-k multilayers in their 3D products.

1.5 Conclusions on NVM market and technological developments

The microelectronics market has developed from only a handful of transistors into a soon-to-be 400 billion dollar market with implications in nearly every aspect of modern life. Non-volatile memories themselves have taken up a considerable share of this market through a series of changes and improvements to their basic technologies. One mainstay of the microelectronics market has been the floating gate flash cell comprised of two poly-silicon layers separated by an ONO trilayer. The ever-increasing demands to improve the technology have started to call into question the continued use of this trilayer.

In the quest for higher density NAND chips, designers first looked to planarizing the flash cells and replacing the ONO with high-k multilayers in order to attain very aggressive EOTs in the single-digit range. Although high-k IGDs allowed more densely packed cells, they cannot compete with the gains offered by turning the word lines in the vertical direction and stacking 48-64 layers. In the future the industry may transition to vertical cells, with high-k IGDs, but for now the ONO remains dominant.

In the high performance applications targeted by embedded flash memories, the uptake of high-k materials has been much slower, with single-level ONO cells still being indispensible. The point at which high-k materials in the IGD become necessary, either to increase coupling between the control gate and floating gate or to co-integrate with high-k metal gate based logic devices, is nonetheless approaching. Although the motivations to change embedded flash products are different from those of standalone memories, much of the research driven by the latter can be used for embedded flash advancements. The VARIOT concept and organization of high-k multilayers will all be just as important for embedded flash, as will any potential threshold voltage instabilities.

Based on the literature, the most promising materials for use as an IGD for embedded flash at the 40 nm and 28 nm nodes are based on alumina, hafnium oxide and hafnium silicates; these are the materials that will be further tested in this thesis. Alumia has the advantage of the highest bandgap of the three which can decrease high temperature leakage currents. However, the k-value is not as large as the others, so the improvements in erase performance are expected to be smaller. 40 nm embedded flash products countaing alumina have been fabricated and are outlined in Chapter 3. Hafnium oxide is expected to have a highest k-value of the three, offering the best increase in erase performance. On the contrary the lower bandgap, or a reaction with the control gate could cause problems with data retention. Hafnium silicate has the advantage of having a higher band gap and being more stable than HfO_2, it is also already used in production of logic devices at the 28 nm node. Although simple HfSiON and HfO_2 samples were tested, we were not able to fabricate working flash devices from these materials as explained in Chapter 4. During the lengthy processing time in the fab, we developed a new electrical characterization method for low field measurements in flash cells as outlined in Chapter 2.

1.6 Bibliography

[Blo06] P. Blomme, et al., "Scalable Floating Gate Flash Memory CellWith Engineered Tunnel Dielectric and High-K (Al2O3) Interpoly Dielectric" in *IEEE Non-Volatile Semiconductor Memory Workshop*, Monterey, CA, USA, Feb., 2006

[Blo10] P. Blomme et al., "Novel dual layer floating gate structure as enabler of fully planar flash memory," *Symposium on VLSI Technology*, Honolulu, 2010, pp. 129-130.

[Bre13] L. Breuil, et al., "A novel multilayer Inter-Gate Dielectric enabling up to 18V Program / Erase window for planar NAND flash" in *IEEE International Memory Workshop*, Monterey, CA, USA, May, 2013

[Bre14a] L. Breuil. et al., “HfO2 Based High-k Inter-Gate Dielectrics for Planar NAND Flash Memory”, *IEEE Electron Device Letters*, vol. 35, no. 1, 2014

[Bre14b] L. Breuil. et al., “Integration of a Multi-layer Inter-Gate Dielectric with Hybrid Floating Gate Towards 10nm Planar NAND Flash” in *IEEE International Memory Workshop*, Taipei, Taiwan, 2014

[Bre15a] L. Breuil. et al., “Intergate Dielectric Engineering Toward Large P/E Window Planar NAND Flash” *IEEE Transactions on Electron Devices*, Vol. 62, No. 5, May, 2015

[Bre15b] L. Breuil. et al., “Optimization of Ru based Hybrid Floating Gate For Planar NAND Flash” in *IEEE International Memory Workshop*. Monterey, CA, USA, 2015

[Cas02] J.D. Casperson, et al., “Materials issues for layered tunnel barrier structures” *Journal of Applied Physics 92*, 2002

[Cho16] J. Choe. (2016, Oct. 6). *Inside iPhone7's SanDisk/Toshiba 48L 3D NAND* [Online]. Available: http://www.eetimes.com/author.asp?section_id=36&doc_id=1330584

[Deg04] R. Degraeve, et al., “Analytical Percolation Model for Predicting Anomalous Charge Loss in Flash Memories”. *IEEE Transactions on Electron Devices*, vol 51, no. 9, Sep 2004

[Ger85] G. Gerosa, et al., "A high performance CMOS technology for 256K/1MB EPROMs" in *IEEE International Electron Devices Meeting*, Washington, DC, USA, Dec. 1985

[Gov03a] B. Govoreanu, et al., “A model for tunneling current in multi-layer tunnel dielectrics”, *Solid-State Electronics*, vol. 47, 2003, pp 1045-1053

[Gov03b] B. Govoreanu, et al., “VARIOT: A Novel Multilayer Tunnel Barrier Concept for Low-Voltage Nonvolatile Memory Devices”, *IEEE Electron Device Letters*, vol. 24, no 2, Feb. 2003

[Gov03c] B. Govoreanu, et al., “Enhanced Tunneling Current Effect for Nonvolatile Memory Applications”, *Japan Journal of Applied Physics*, vol. 42, 2003, pp. 2020-2024

[Hua10] X. D. Huang and P. T. Lai, "YAlOx as inter-poly dielectric for improved performance of flash-memory application," *2010 IEEE International Conference of Electron Devices and Solid-State Circuits*, Hong Kong, 2010, pp. 1-4.

[Hua11] X. D. Huang, et al., "Improved Performance of Yttrium-Doped Al_2O_3 as Inter-Poly Dielectric for Flash-Memory Applications," in *IEEE Transactions on Device and Materials Reliability*, vol. 11, no. 3, pp. 490-494, Sept. 2011.

[Ici17] I.C. Insights, “The McCLEAN REPORT 2017: A Complete Analysis and Forecast of the Integrated Circuit Industry” Scottsdale AZ, USA, 2017

[Jay10] S. Jayanti, et al., “Technique to improve performance of Al_2O_3 interpoly dielectric using a La_2O_3 interface scavenging layer for floating gate memory structures”, *Applied Physics Letters*. Vol. 96, no. 9, 2010

[Joh15] R.C. Johnson. (2015, May 8). *Samsung Breaks Ground on $14 Billion Fab: World's most expensive semi fab.* [Online] Available: http://www.eetimes.com/document.asp?doc_id=1326565

[Kak08] R. Kakoschke, et al., “Use of Al2O3 as inter-poly dielectric in a production proven 130 nm embedded Flash technology”, *Solid-State Electronics*, vol. 52, 2008, pp. 550-556

[Kim10] H.J. Kim, et al., “Memory characteristics of $Al_2O_3/La_2O_3/Al_2O_3$ multi-layer films with various blocking and tunnel oxide thicknesses”, *Materials Science in Semiconductor Processing*, vol. 13, no. 1, Feb. 2010, pp 9-12

[Kue15] W. Kueber, et al., “A Highly Reliable and Cost Effective 16nm Planar NAND Cell Technology” in *IEEE International Memory Workshop*, Monterey, CA, USA, May 2015

[Lar11] L. Larcher and A. Padovani, “High-k related reliability issues in advanced non-volatile memories” *Microelectronics Reliability*, vol. 50, 2010, pp. 1251–1258

[Lee97] W.H. Lee, et al., “A Novel High K Inter-poly Dielectric(IPD), Al_2O_3 For Low Voltage/high Speed Flash memories: erasing in msecs at 3.3V” in *Symposium on VLSI Technology*, Kyoto, Japan, Jun., 1997

[Lik98] K. Likharev, “Layered tunnel barriers for nonvolatile memory devices”, *Applied Physics Letters*, vol. 73, 1998

[Lin11] C-L. Lin, et al., “Electrical characteristics and TDDB breakdown mechanism of N_2-RTA-treated Hf-based high-k gate dielectrics”, *Microelectronic Engineering*, Vol. 88, 2011, pp 950–958

[Lis13] J. Lisonoi, et al., “High-k gadolinium and aluminum scandates for hybrid floating gate NAND flash” *Microelectronic Engineering*, vol. 109, 2013, pp. 220-222

[Mic17a] Microchip Technology Inc. (2017) *What is SuperFlash® Technology?* [Online] Available: http://www.microchip.com/design-centers/memory/serial-parallel-flash/what-is-superflash-technology

[Mic17b] Micron Technology Inc (2017) *3D NAND* [Online] Available: https://www.micron.com/products/nand-flash/3d-nand

[Mir06] A.H. Miranda, et al., "Reliability Comparison of Al2O3 and HfSiON for use as Interpoly Dielectric in Flash Arrays" in *IEEE European Solid-State Device Research Conference*, Montreux, Switzerland, Sep. 2006

[Mcc17] J.C. McCallum. (2017, Jan. 17). *Flash Memory and SSD Prices (2003-2017)* [Online]. Available: http://www.jcmit.net/flashprice.htm

[Mel17] C. Mellor. (2017, Feb. 23). *Tosh doubles 64-layer 3D flash chip capacity with a bit of TLC: 1TB chip incoming* [Online]. Availible: https://www.theregister.co.uk/2017/02/23/tosh_doubles_64layer_3d_flash_chip_capacity/

[Moo65] G. E. Moore, “Cramming More Components onto Integrated Circuits”, *Electronics*, Apr. 1965, pp. 114–117

[Mea16] L. Mearain. (2016, Dec. 21). *3D NAND set to dominate SSDs, kill off traditional flash: Manufacturers are ramping up production of new 64-layer 3D NAND* [Online]. Availible: http://www.computerworld.com/article/3152808/data-storage/3d-nand-set-to-dominate-ssds-kill-off-traditional-flash.html

[Orb16] Z. Or-bach. (2016, Aug. 29). *28nm Was Las Node of Moore's Law* [Online]. Available: http://www.eetimes.com/author.asp?section_id=36&doc_id=1330366

[Pad08] A. Padovani, “Statistical Modeling of Leakage Currents Through SiO2/High-k Dielectrics Stacks for Non-Volatile Memory Applications” in *IEEE InternationalReliability Physics Symposium*, Phoenix, AZ, USA, Jul. 2008

[Pav97] P. Pavan, et al., "Flash memory cells-an overview," in *Proceedings of the IEEE*, vol. 85, no. 8, pp. 1248-1271, Aug 1997.

[Pow07] J.R. Power, et al., "Improved Reliability of a High-k IPD Flash Cell through use of a Top-oxide" in *IEEE Non-Volatile Semiconductor Memory Workshop*, Monterey, CA, USA, Aug. 2007

[Pow08] J.R. Power, et al., "Improved Retention for a Al2O3 IPD Embedded Flash Cell without Top-Oxide" in *IEEE Non-Volatile Semiconductor Memory Workshop*, Opio, France, May, 2008

[Rob15] J. Robertson and R. Wallace, “High-K materials and metal gates for CMOS applications” *Materials Science and Engineering: R: Reports*, Feb. 2015

[Shu09] D. Shum, et al., “ALD-Al2O3 as an Inter-Poly Dielectric for a Product Demonstrator in a proven eFlash Technology” in *IEEE International Memory Workshop*, Monterey, CA, USA, May, 2009

[Shu12] D. Shum, et al., “Highly Reliable Flash Memory with Self-aligned Split-gate Cell Embedded into High Performance 65nm CMOS for Automotive & Smartcard Applications” in *IEEE International Memory Workshop*, Milan, Italy, May 2012

[Stm16] STMicroelectronics. (2016, May 11). *New Automotive Microcontrollers from STMicroelectronics Pave the Way to Smart Driving via More Secure and Connected Cars*. Available : http://www.st.com/content/st_com/en/about/media-center/press-item.html/p3826.html

[Tal02] A. Tal (2002, Oct.). *Two Flash Technologies Compared: NOR vs NAND* [Online] Available: https://focus.ti.com/pdfs/omap/diskonchipvsnor.pdf

[Tan13] B. Tang et al., “Read and Pass Disturbance in the Programmed States of Floating Gate Flash Memory Cells With High-κ Interpoly Gate Dielectric Stacks” *IEEE Transactions on Electron Devices*, Jul. 2013

[Tan14] B.J. Tang, et al., “Optimization of inter-gate-dielectrics in hybrid float gate devices to reduce window instability during memory operations” *Microelectronics Reliability*, vol. 54, no. 9-10, 2014, pp 2258-2261

[Tar72] Y. Tarui, et al., "Electrically reprogrammable nonvolatile semiconductor memory", *IEEE Journal of Solid-State Circuits*, vol. 7, no. 5, Oct., 1972

[Van06] M. van Duuren, et al., "Performance and Reliability of 2-Transistor FN/FN Flash Arrays with Hafnium Based High-K Inter-Poly Dielectrics for Embedded NVM" in *IEEE Non-Volatile Semiconductor Memory Workshop*, Monterey, CA, USA, Feb. 2006

[Wal16] M.W. Waldorp. (2016, Feb. 9). *The chips are down for Moore’s law: The semiconductor industry will soon abandon its pursuit of Moore's law. Now things* *could get a lot more interesting*. Available: http://www.nature.com/news/the-chips-are-down-for-moore-s-law-1.19338

[Wat84] T. Watanabe, et al., "Stacked SiO2/Si3N4/SiO2dielectric layer for reliable memory capacitor" in *IEEE International Electron Devices Meeting*, San Francisco, CA, USA, Dec. 1984

[Wel06] D. Wellekens, et al., "Al_2O_3 Based Flash Interpoly Dielectrics: a Comparative Retention Study" in *IEEE European Solid-State Device Research Conference*, Montreux, Switzerland, Sep. 2006

[Wel11] D. Wellekens, et al., "An Ultra-thin Hybrid Floating Gate Concept for Sub- 20nm NAND Flash Technologies" in IEEE Memory Workshop, Monterey, CA, USA, May, 2011

[Whi16] M. White. (2016, Aug. 24). *Established Technology Nodes: The Most Popular Kid at the Dance* [Online]. Availible:http://semimd.com/favre/2016/08/24/established-technology-nodes-the-most-popular-kid-at-the-dance/

[Zha08] L. Zhang, et al., "Multi-layer high-j interpoly dielectric for floating gate flash memory devices", *Solid-State Electronics*, vol. 52, 2008, pp 564-570

Chapter 2: Ultra Low Field Current Measurements for Data Retention

Table of Contents

2.1 Introduction to flash cell data retention

With the requirement of ten-years of data retention for non-volatile memories it is clear that the threshold voltage must not shift significantly over time. If it does, then ones will be misread as zeros, or zeros will be misread as ones. As a ten-year-long test would not be very practical, the industrial standard is to perform accelerated tests at high temperatures on millions of cells and then extrapolated. Data retention can be impacted by both extrinsic (oxide defects and contamination) and intrinsic (electron detrapping, thermionic emission, field-assisted emission) phenomena [Pav97, Bez03].

As described in Section 1.2, the distinction between cell representing a zero or a one bit is whether or not the drain current is above a certain threshold when the read bias (V_{read}) is applied to the control gate. After program or erase operations, the tails of the V_T distributions are approximately the program-verify or erase-verify levels, respectively, as in Figure 26. The relationship between these three values defines the level of V_T-shift that results in data-retention failures; they are carefully chosen to optimize program/erase speed, program/erase pulses, endurance and data-retention.

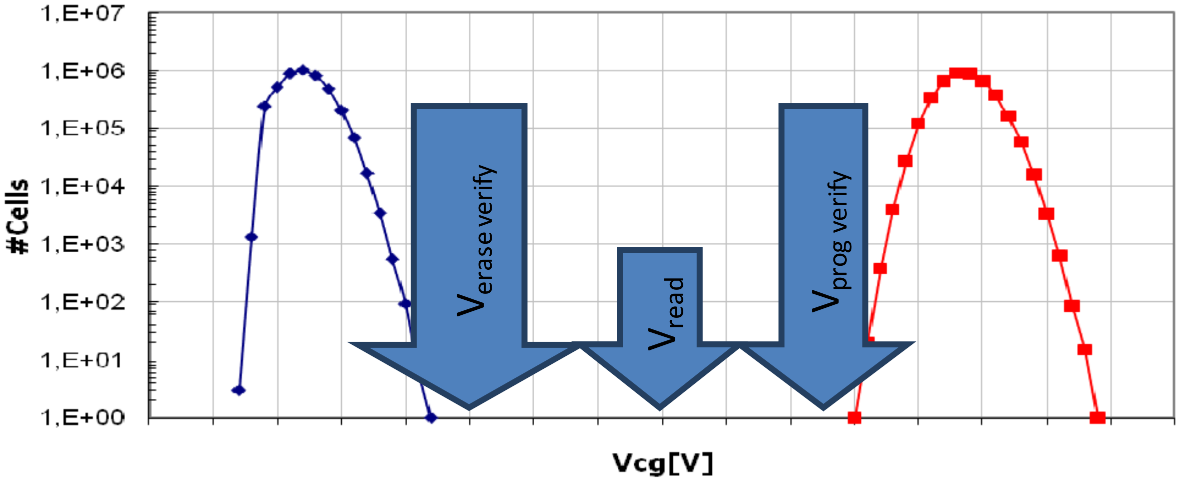


Figure 26: Typical V_T distribution on a flash product for program and erase after compacting algorithm [Pia10]

In the case of the 40 nm embedded flash discussed in this thesis, the programmed state corresponds to electrons stored on the floating gate, and it is the state most likely to be impacted by leakage currents. For a fixed program verify level and read bias, the V_T-shift at which failures will begin to occur is given by Equation 2.1.

$$\Delta V_{T\ at\ failure} = V_{Program\ Verify} - V_{Read} \quad Equation\ 2.1$$

The industrial standard for estimating data retention times is to measure time to failure at high temperatures (200-300 °C) and extrapolate to the temperature specified in the product's mission profile using Equation 2.2, where t_0 and E_a are pre-exponential constant and activation energy extracted from the accelerated test, k is the Boltzmann constant and T is the temperature in K [Des99a].

$$t_{retention} = t_0 e^{\frac{E_a}{kT}} \quad Equation\ 2.2$$

The reported activation energy for the charge loss varies widely and is temperature dependent, so both De Salvo et al. and Kim et al. disagree with the simple model in Equation 2.2 [Des99a,Kim04]. Both referred to an initial transient regime followed by a second, steady-state regime that dominates retention. De Salvo et al. considered the first regime to be electron movement within the nitride and described the second regime using the Fowler-Northeim transport across the ONO [Des99a]. Kim et al. considered the initial phase to be detrapping of electrons from the bottom oxide/nitride interface, which were trapped during programming and the second regime to be that of Figure 27 [Kim04]. The industrial standard of simply using Equation 2.2 to extrapolate to operating temperatures remains in current use. However the model may not be applicable to different IGDs; Lee et al. showed that there was an incoherence between the data retention model for ONO and OAO samples [Lee10]. The replacement of the ONO by another IGD demands a considerable effort to update the data retention models. Not to be constrained by leakage models, there is a motivation to follow the charge on the floating gate directly.

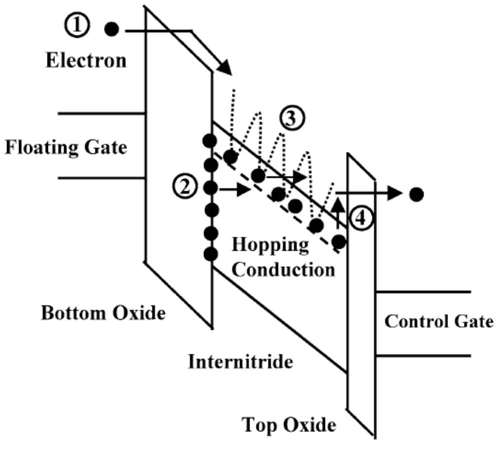


Figure 27: ONO conduction mechanism proposed by Kim et al. in 2004 [Kim04]

Another way to approach the data retention problem is from the point of view of the change in the number of electrons stored on the floating gate as a function of time in the form of leakage currents. Recalling Equation 1.8 from Chapter 1, the amount of charge that corresponds to this V_T – shift is calculated in Equation 2.3 and divided by the time in order to yield the allowed average leakage current.

$$\Delta Q = -(V_{Program\,Verify} - V_{Read}) \times C_{ONO} \quad Equation\ 2.3$$

Taking typical $V_{prog\text{-}verify}$ and V_{read} values of 6 V and 4 V, respectively, along with C_{ONO} as 0.14 fF leads to 0.28 fC (1750 electrons) being the amount of charge loss that leads to data-retention failures. As the standard for NVM is ten years of data retention the average leakage current out of the floating gate over ten years needs to be blow 10^{-24} A (10^{-15}A/cm²), or fewer than one lost electron every two days.

Figure 28 shows the TCAD representation of a flash cell along with the energy band and electric field diagrams in the vertical direction during data retention in the programmed state. There are non-zero electric fields in the tunnel oxide and ONO that can direct electrons towards either the channel or control gate, respectively. It is therefore of great interest to be able to separate the IGD and tunnel oxide current components during testing in order to identify the each layer's contribution.

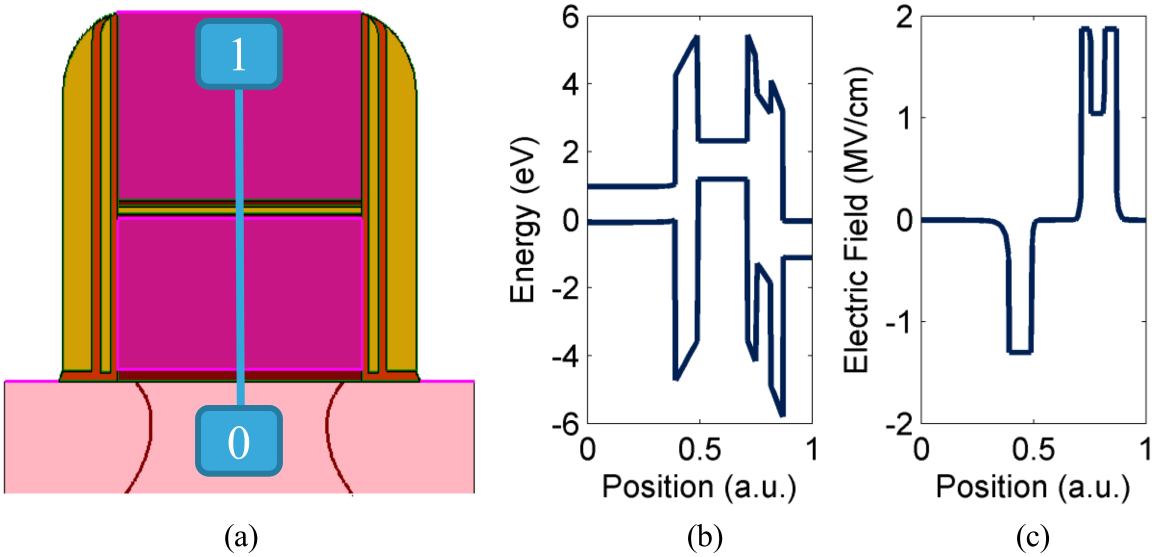


Figure 28: (a) TCAD representation of flash cell and TCAD-simulated (b) energy band (c) electric field diagrams for a programmed cell at the data retention condition

The obvious approach to decorrelating the two leakage currents is to measure the two layers independently at the relevant biases. Figure 29, the IV curve of a 10,000 µm² Poly1-ONO-

Poly2 test structure, illustrates the problem with this approach; the average electric field across the ONO during data retention E_{ret} yields a current well below the resolution limit of the testing equipment. In practice, the detection limit of the Agilent B1500 semiconductor device analyzer is on the order of 10^{-14}A (10^{-10}A/cm² for these planar devices); this is well over the 10^{-15}A/cm² threshold required to for ten years data retention. If one wanted to make a test structure large enough to detect retention-level leakage currents of roughly ten times the noise level, its area would need to be on the order of ten cm², considerably larger than many die sizes.

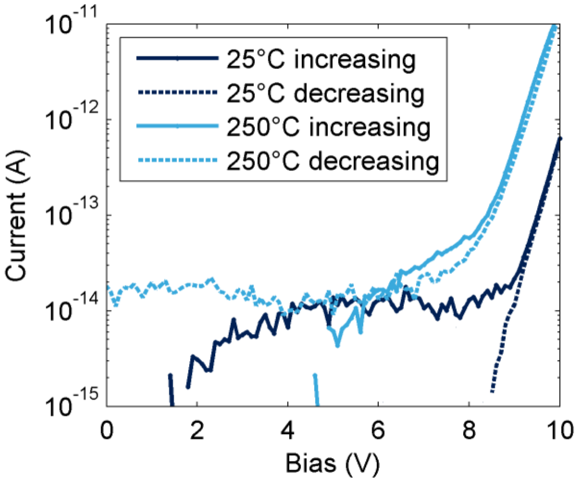


Figure 29: IV measurement from 0 V to +10 V to 0 V on a 10,000µm² test structure at 25 and 250°C

As the E_{ret} is so far below the field where currents can be measured, extrapolations are unlikely to be valid due to transitions in the conduction mechanism between high and low-fields. The difference in area between the flash cell and the planar test structure (a factor of over 100,000) increases the likelihood that the direct IV measurement will be impacted by defects in the ONO instead of its intrinsic conduction properties. The planar device also does not take into account the 3D effects in the inherently 3D flash cell [Zak11].

These factors provide the motivation for the Oxide Stress Separation (OSS) technique as a way to measure ultra-low currents across the IGD of a nominal flash cell.

2.2 Historical Floating Gate Techniques

By the end of the 1970s, microelectronics engineers studying the reliability of gate oxides were interested in measuring currents at levels below the background noise levels of their equipment. They realized that the sensitivity of a MOSFET's transfer characteristics to its properties could be exploited. This was done in 1980 by Gaensslen and Aitkenn order to measure hot electron currents. The gate of a MOSFET was charged to an initial bias then left floating while the drain current (under constant drain bias) was recorded in time. The initial discharge rate of the gate capacitor could be calculated to determine channel hot electron currents on the order of 10^{-13}A [Gae80]. A few years later in 1986, Nissan-Cohen presented the Floating Gate Method for gate currents down to 10^{-18}A. The method works by using two identical transistors sharing the same poly-Si floating gate that is capacitively coupled to metal control gates, as shown in Figure 31. Only one of the transistors is stressed and the threshold voltages of both transistors are measured in time. As only one of the two transistors is stressed, the difference in V_T – shift between the two can be used to quantify both the transistors degradation and gate leakage current [Nis86].

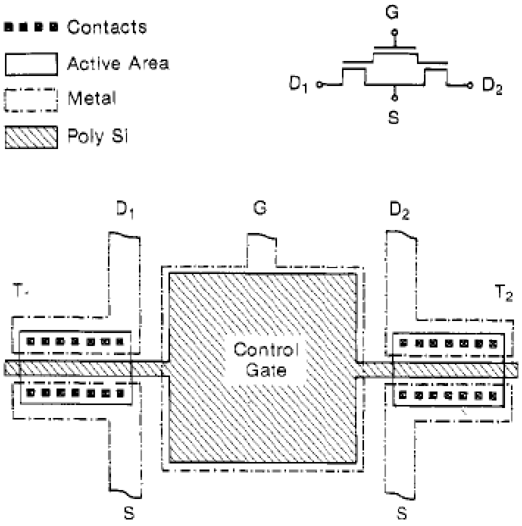


Figure 30: Two transistors share a common poly-Si floating gate in Nissan-Cohen's gate current/degradation measurements [Nis86]

Unfortunately, this method only accounted for the MOSFET degradation due to Hot Carrier Injection, and not the degradation by Fowler-Nordheim tunneling, which can also change the I_dV_g characteristics. A workaround was found in 1991 by Fishbein in the form of a test structure as shown in Figure 31, where the source of the tunneling capacitor (S1) is grounded, but the source (S2) and drain (D) can be biased to only a few volts below the common gate (G). The configuration avoids the damage caused by tunneling currents in the monitoring transsitor and was able to measure currents as low as 2×10^{-17} A (2×10^{-13} A/cm²) [Fis91]. This method has also been applied to the ONO to detect current densities around 10^{-13} A/cm² at high temperature [Des99b]

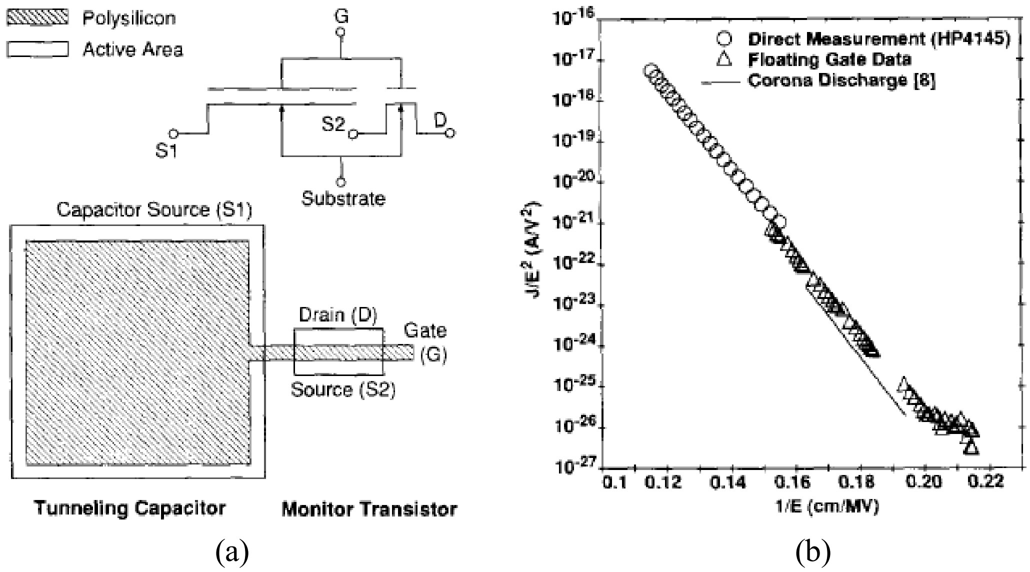


Figure 31: (a) Fishbein's test structure separated the tunneling capacitor and monitoring transistor which shared a common [Fis91] (b) Floating Gate measurements align well with direct measurements on Fowler-Nordheim axes down to about 5 MV/cm [Des99b]

The V_T-shift of a nominal flash cell under stress was later used by Manabe and related to the leakage current by $I_g = - C_{ONO}\times\Delta\ V_T /\Delta t$. They measured currents on the order of 10^{-20}A and assumed that all of it was across the tunnel oxide; the only mention of the interpoly was to describe its capacitance [Man99]. Later works by Tao showed an interest in the ONO, however their interpretation of the V_T-shift was the redistribution of electrons in the ONO; again, leakage currents were something that only occurred across the tunnel oxide [Tao07].

2.3 The "Oxide Stress Separation" technique

When a potential bias is applied to the control gate of a flash cell, the potential drop occurs in the body , tunnel oxide and the ONO. For example, the TCAD simulation in Figure 32(a) shows the situation for a +6 V control gate bias applied to a virgin (no charge on the floating gate) flash cell. This division of the potential drop throughout the stack makes it difficult to identify the origin of any V_T – shifts occurring under the gate stress. Starting from the fact that changing the number of electrons on the floating gate will change its potential, a configuration in which the potential drop occurs entirely across the ONO can be imagined, as in Figure 32(b). In this configuration, the electric fields in the channel and the tunnel oxide are zero, so any V_T – shifts can be attributed to electrons traversing the ONO. The effective separation of the potential drops in the tunnel oxide and ONO leads to the name of the technique, *Oxide Stress Separation*.

This technique was first presented in the poster session at the IEEE Semiconductor Interface Specialists Conference In Washington DC, USA in December 2015 before being refined and then published in Microelectronics Reliability [Dob17]. With the CEA-Leti we have submitted an application for a patent for "Method for determining a leakage current through an inter-gate dielectric structure of a flash memory cell" in December 2016.

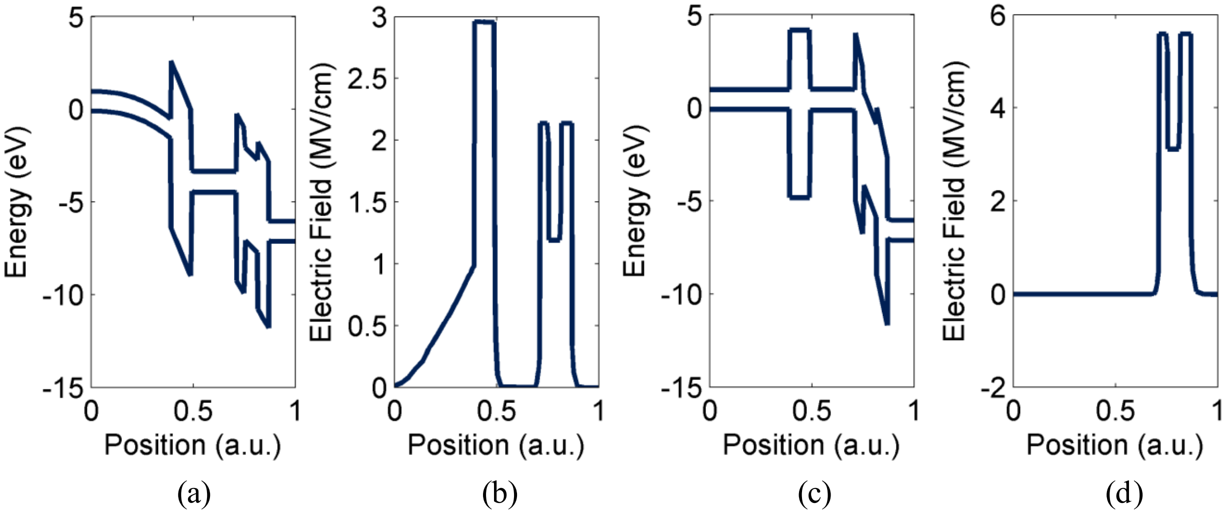


Figure 32: TCAD-simulated energy band structure and Electric field diagrams for Vg = +6 V applied to the control gate of a cell (a-b) having no charge on the floating gate and (c-d) at the OSS condition

2.3.1 Determining the OSS condition

In practice, this OSS condition occurs when the potential of the floating gate is equal to the flat-band voltage of the MOSCAP formed by the body/tunnel oxide/Poly1. The cell can be programmed (or erased) to raise (or lower) the floating gate potential energy in order to arrive at the OSS condition. Equation 2.4 describes the floating gate potential while the potential drop in tunnel oxide and the ONO are then given by Equations 2.5 and 2.6, respectively.

$$V_{FG} = \alpha_G(V_G + V_{T_0} - V_T) + \alpha_D V_D + \alpha_B V_B + \alpha_S V_S \quad Equation\ 2.4$$

$$\Delta V_{tunnel} = V_{FG} - V_{FB} \quad Equation\ 2.5$$

$$\Delta V_{ONO} = V_C - V_{FG} \quad Equation\ 2.6$$

With the body, drain and source at ground, the V_T and V_G pair that minimizes the potential drop in the tunnel oxide ($V_{FG} = V_{FB}$) are related by Equation 2.7. The voltage drop in the ONO at this condition is given by Equation 2.8.

$$V_T^{OSS} = V_{CG}^{OSS} + V_{T_0} - \frac{V_{FB}}{\alpha_{CG}} \quad Equation\ 2.7$$

$$\Delta V_{ONO} = V_{CG} - V_{FB} \quad Equation\ 2.8$$

There are an infinite number of simultaneous solutions to Equations 2.7 and 2.8 so the targeted voltage drop in the ONO must be fixed to the desired value before calculating the V_G^{OSS} and V_T^{OSS} required for the OSS measurement. The cell is programmed to the V_T^{OSS} and V_G^{OSS} is applied to the control gate. As electrons leave the floating gate, the V_T decreases according Equation 2.9 and this change in time is used in Equation 2.10 to calculate a current.

$$V_T = V_{T_0} - \frac{Q}{C_{ONO}} \quad Equation\ 2.9$$

$$I = -C_{ONO}\frac{dV_t}{dt} \quad Equation\ 2.10$$

2.3.2 How to apply the OSS Technique

The ability to determine the OSS condition requires the knowledge of the V_{FB} of the tunnel oxide, as well as α_G and V_{T_0} of the flash cell. These can all be determined from four common test structures that are normally embedded in a flash chip, pictured in Figure 33. The flash cell is the nominal device under test and the TREQ (from the French: *transistor équivalent*) that has the same width, length, tunnel oxide thickness and doping as a flash cell except the floating and control gates are shorted. The CV curves of MOS capacitor structure and the ONO capacitor are used to determine the V_{FB} of the tunnel oxide and the EOT of the ONO, respectively.

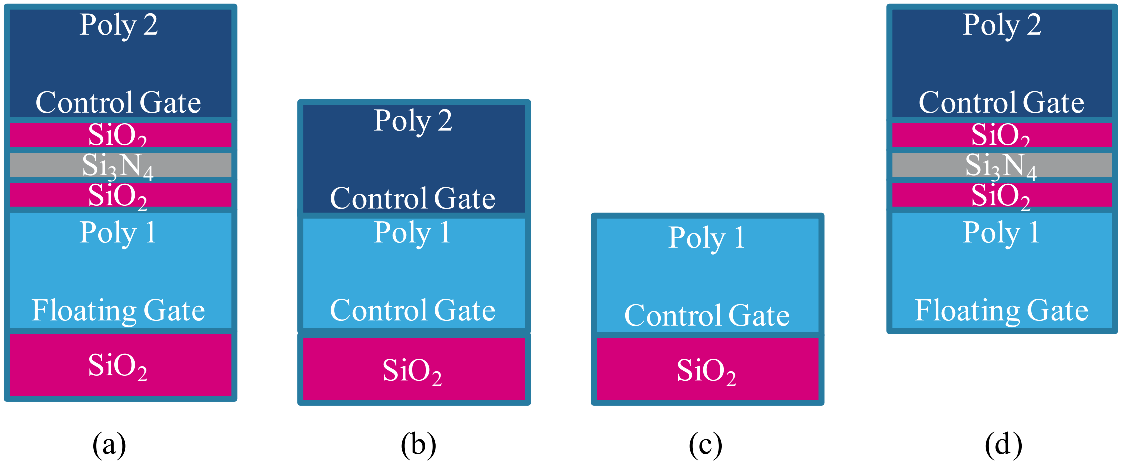


Figure 33: (a) nominal flash cell (b) TREQ or flash-equivalent transistor (c) Tunnel oxide MOS capacitor (d) Poly-ONO-Poly capacitor

2.3.3 Calculating α_G and V_{T0}

As shown in Equation 2.4, the coupling coefficients and threshold voltages can be used to calculate the V_{FG}. Intuitively, if the relation between the V_G and V_{FG} of a flash cell is known then the α_G and V_{T0} can also be extracted. The TREQs are have the same channel width, length, oxide thickness and doping as their respective flash cells, however they allow the floating gate poly-Si to be directly addressed. Therefore, if the $I_DV_{G/FG}$ characteristics are known, then the V_{FG} of a flash cell can be calculated from its measured I_D. The I_DV_G curve of a flash cell is compared with the $I_DV_{G/FG}$ of two TREQs to obtain the V_{FG} vs V_G relation in Figure 34(a) and (b). Using the V_T of the flash cell and the linear fit of the V_{FG} vs V_G relation to 2.4, a pair of α_G and V_{T0} is extracted.

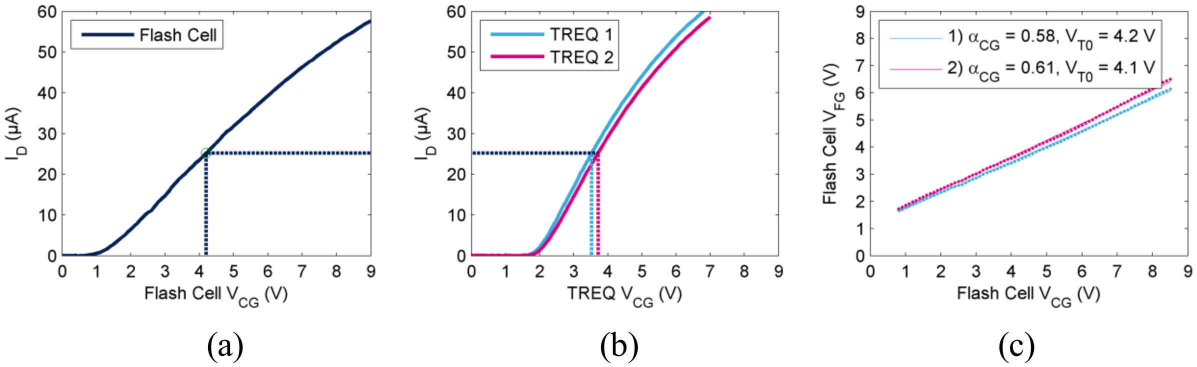


Figure 34: (a) I_DV_G curve of flash cell (b) $I_DV_{FG/G}$ curve of TREQ (c) V_{FG} and V_G relation for flash cell

In the previous example with one flash cell and two TREQs, two coupling coefficients and two V_{T0} values were extracted because minor variations in the fabrication process lead to a distribution of I_DV_G characteristics. Therefore multiple flash cells are compared against multiple TREQs in order to obtain α_G and V_{T0} distributions. An example of these distributions is shown in Figure 35 for 17 flash cells and 31 TREQs, which yield 527 α_G and V_{T0} values, centred at 0.64 and 4.1 V, respectively.

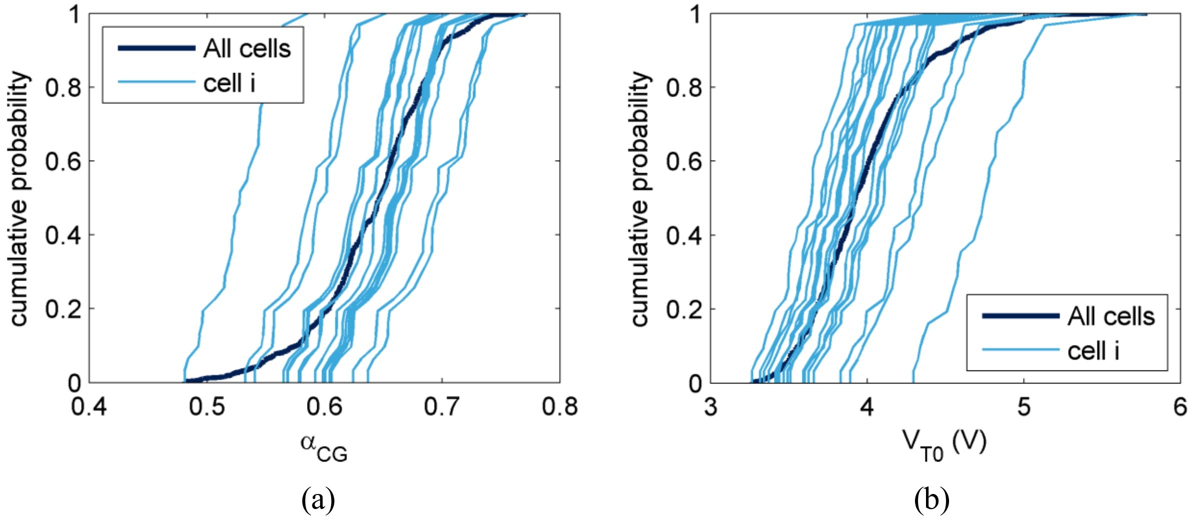


Figure 35: α_G and V_{T0} cumulative distribution functions for 17 flash cells and 31 TREQs

2.3.4 Deviations from the ideal case

In the event that the nominal α_{CG} and V_{t_0} values used to calculate the OSS conditions in Equation 2.7 are not precisely known for the flash cell under test, Equation 2.4 becomes Equation 2.11. Small fluctuations in the values lead to the cell not being in the exact OSS state (E_{tunnel}=0 MV/cm) but remaining in a state where the vast majority of the potential drop continues to occur in the ONO (E_{ONO}>>E_{tunnel}).

$$V_{FG} = V_{FB}\left(\frac{\alpha_G^{true}}{\alpha_G^{nominal}}\right) + \alpha_G^{true}\left(V_{T0}^{true} - V_{T0}^{nominal}\right) \quad Equation\ 2.11$$

During an OSS measurement, any leakage current will lead to a change in V_T, and therefore a change in V_{FG} that moves the cell out of the OSS condition. The impact of using a test set-up that cannot alter the V_G during the measurement is simulated in Figure 36, showing the impact of a 500 mV V_T-shift during an OSS measurement for a test where V_{CG} remains a constant +6 V. Although E_{tunnel} becomes non-zero, E_{ONO}>>E_{tunnel} and the principle of the OSS measurement remains; the potential drop in the tunnel oxide is negligible and ONO leakage currents are dominant.

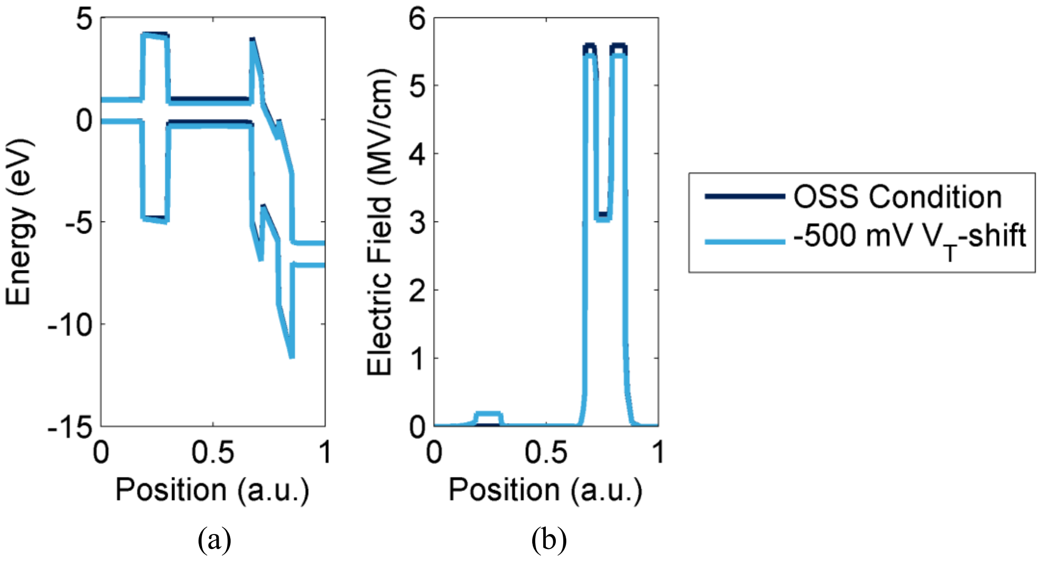


Figure 36: TCAD-simulated (a) band structure and (b) electric field strength for a flash cell in the OSS condition and the same cell after a 500mV decrease in V_T under the same +6 V V_G

2.4 OSS Application

The V_T as a function of time for a first OSS measurement for 60 hours at 250°C is shown below in Figure 37. The relationship is clearly not linear as would be expected if there were only one current component, there is also a fast initial shift in V_T.

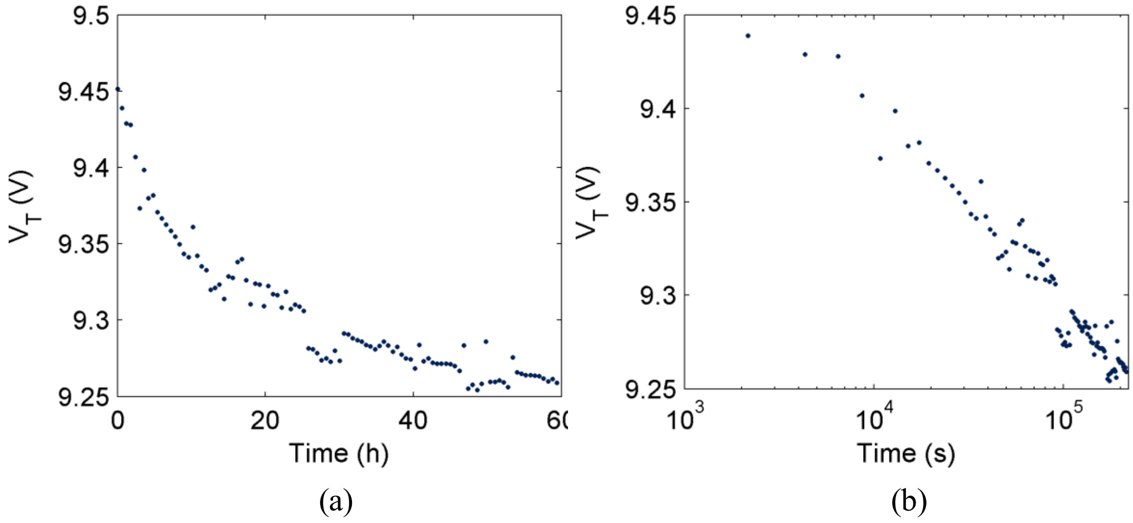


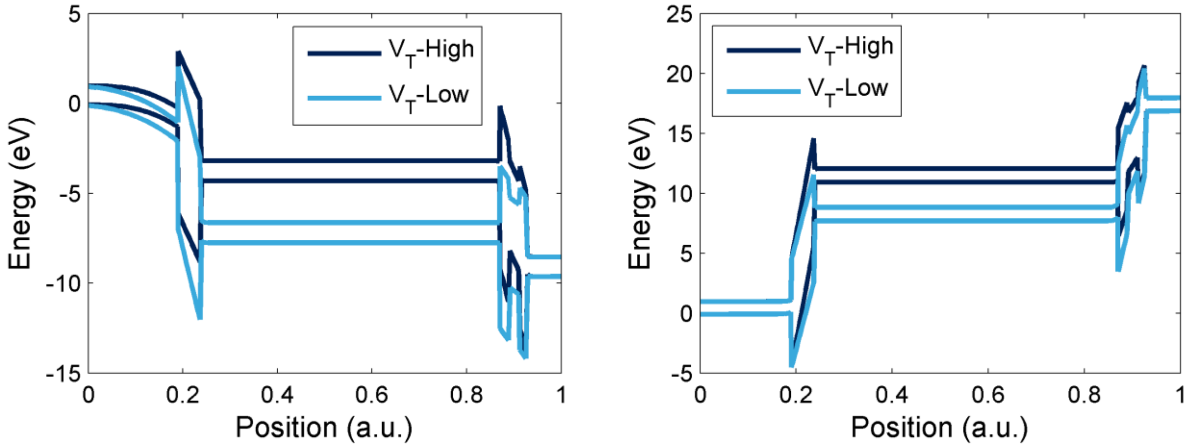
Figure 37: V_T in time for a first OSS measurement at 250°C in (a) linear and (b) semi log scales with a ΔV_{ONO} of 5.4 V (E_{ONO} = 4.2 MV/cm)

In looking at the structure of a flash cell it is natural think of the other phenomena that can impact the threshold voltage. If the V_T is modified by changing the number of electrons on the floating gate ~10 nm away from the channel, then electrons anywhere else in the structure should also affect the electrostatics of the channel and therefore the V_T. It is a well-known phenomenon that charges trapped in the gate oxide of a MOSFET impact the V_T; the closer the charge to the channel, the stronger the effect [Hu10].

This section is therefore divided into two parts: the first to study the impact of the tunnel oxide of flash-equivalent transistors and the second to measure the leakage currents across the ONO.

2.4.1 Influence of Program/Erase operations on tunnel oxide

During the program/erase cycle of a flash cell, the potential drop in the tunnel oxide will vary due to the different biases used in the operations. During program (erase) the floating gate's energy level moves up (down) during the operations, reducing the magnitude of the potential drop, as shown using TCAD simulations in Figure 38. This provides the motivation to reproduce these conditions on a tunnel oxide using a TREQ device so that any V_T changes are not hidden within programming/erasing a flash cell.



(a)

(b)

Figure 38: TCAD simulated of band diagrams (a) programming from the low V_T to the high V_T state and (b) erasing from the high V_T state to the low V_T state

Using Equation 2.5 and the program/erase conditions, the largest potential drops in the tunnel oxide are roughly -12 V at the beginning of the erase pulse and +6 V at the onset of programming. The actual amount of time spent in each configuration is shorter because the band structure changes during the operation, but for practical reasons 3.2 µs and 50 ms were used for the length of stress during program and ease, respectively. The pulse characteristics are summarized in Table 2, with a bias applied to the drain as well during the program simulation in order to generate the hot carriers that would be present during normal operation.

Table 2: Simulated Program and Erase pulses for dummy cells

	V_{CG} (V)	V_D (V)	Time	Type of Stress
Program	+6	+4	3.2 µs	Hot Carrier Injection
Erase	-12	0	50 ms	Fowler-Nordheim

The simulated program and erase pulses were applied to sets of three TREQs with a read measurement occurring after each pulse. The TREQs were exposed to either both program and erase, erase only, or program only pulses. The V_T – shift values throughout 100 sets of pulses are shown in Figure 39, with the erase pulses having a much larger impact on the V_T. In all cases, the V_T increased, which is consistent with negative charges being trapped in the tunnel oxide. As these negative charges would be close to the channel interface they would inhibit the inversion of the channel, increasing the V_T. It should be noted that these represent worst-case scenarios as during the program and erase operations the electric field strength decreases during the operation as the floating gate potential changes. Prior to an OSS measurement the cell is programmed and erased only two or three times to verify that it is functional; the ≈0.5-0.6 V V_T increase that comes with 100 program/erase operations is therefore not expected.

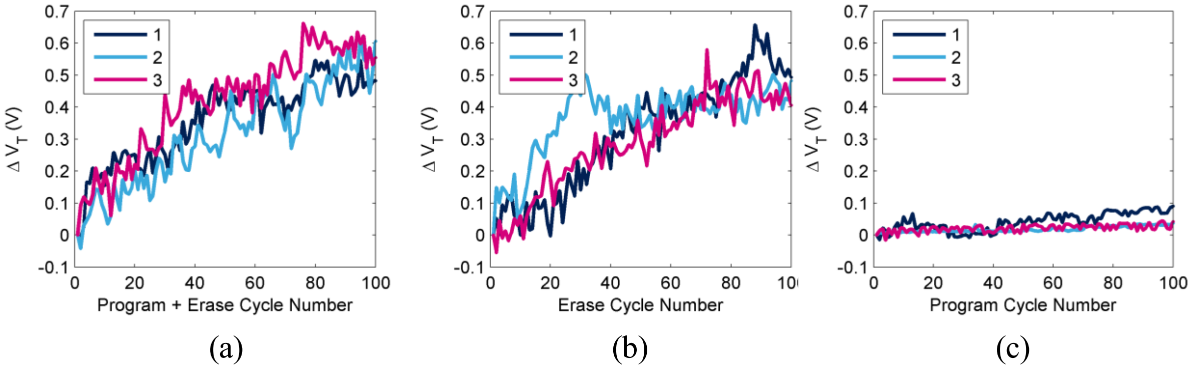


Figure 39: TREQ V_T – shift during simulated (a) program and erase (b) erase only (c) program only pulses on sets of three devices at 250 °C

2.4.2 Tunnel oxide relaxation at the OSS condition

After a few simulated program and erase cycles, a -1 V bias was applied to the control gate for 30 hours, with the V_T read at regular intervals. The -1 V bias corresponds to the flat-band voltage of the MOSCAP formed by the body/tunnel oxide/poly-Si floating gate as measured on a test capacitor of the same wafer. This is referred to as the TREQ OSS measurement because the tunnel oxide's flat-band condition is targeted, like in a flash OSS measurement. The evolution of the V_T in time in Figure 40 fits well with an exponential decay, such as in Equation 2.12 where A_0 is a pre-exponential constant related to the number of trapped charges, A_1 is a constant related to the initial threshold voltage of the TREQ and τ is a time constant related to the relaxation of the trapped charges.

$$V_t(t) = A_0 e^{-\frac{t}{\tau}} + A_1 \quad Equation\ 2.12$$

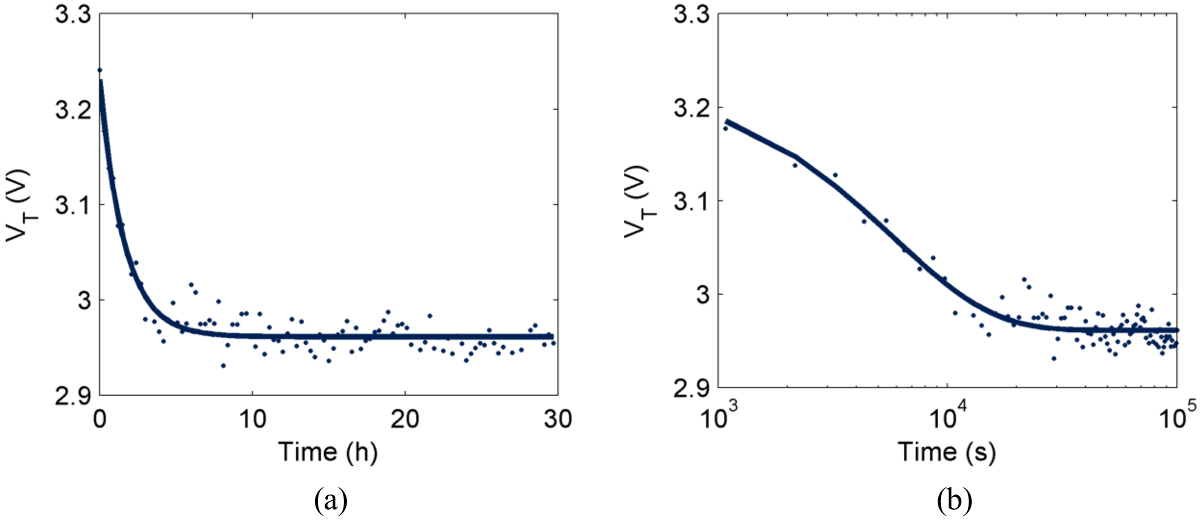
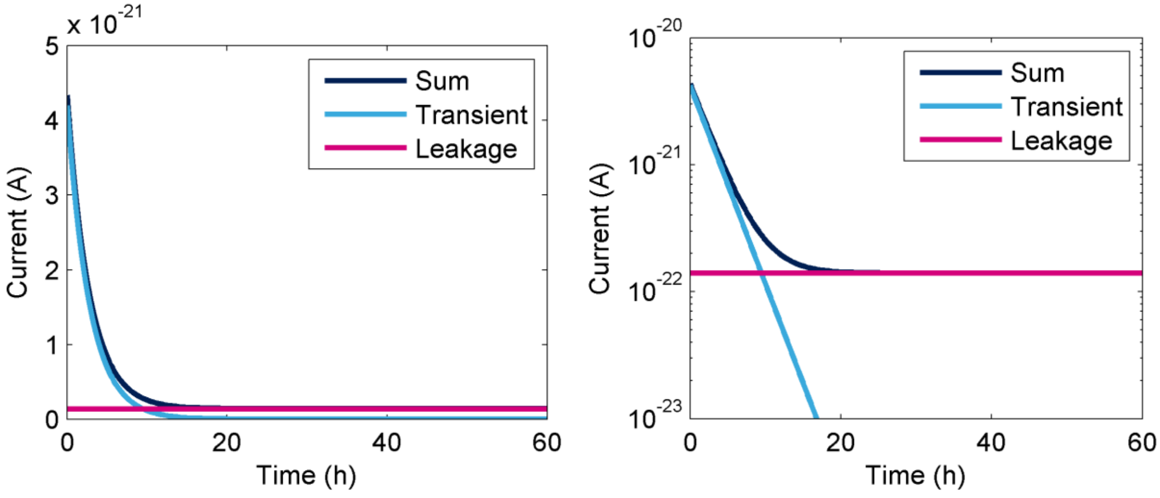


Figure 40: TREQ OSS measurement at 250°C in (a) linear and (b) semi log timescales with a fit of Equation 2.12 (solid line)

The interpretation of the V_T's behaviour is that electrons, trapped in the tunnel oxide during program/erase cycles, slowly de-trap to the body, causing a decrease in V_T. It is clear that Equation 2.12 would have an impact on the measured current as its derivative would be included in Equation 2.10, which would then include two components with the first component being transient as shown in Equation 2.13

$$I = C_{ONO}\frac{A_0}{\tau}e^{-\frac{t}{\tau}} + I_{ONO} \quad Equation\ 2.13$$

We will see in Section 2.4.3 that τ is typically around 10,000 s and A_0 is a few hundred mV. These values and an ONO leakage current of about 10^{-22} A are used to illustrate the two components in Figure 41. The first, detrapping component decays in time to zero, while the second component is a constant for a given OSS test. An OSS measurement should have a long enough duration that the detrapping component does not impact the extraction of the leakage component. The key result of the measurements on TREQs is that the tunnel cannot be seen as having no influence on the V_T. Therefore, any OSS measurement needs to take into account this initial change in the V_T, in practice this obliges the use of long (60 hour) measurements.



(a) (b)

Figure 41: Example of the impact of tunnel oxide detrapping on an OSS measurement assuming τ = 10,000s, $A_0 = 0.3$ V and a leakge current of about 10^{-22} A. These values are typical in OSS measurements.

2.4.3 Flash Cell OSS

As the example of an OSS measurement in Figure 37 showed, two features of the data are immediately clear; the V_T does not level-off as in the case of the TREQs and the read measurements are extremely noisy. At 250°C, 100 sequential read measurements typically have a standard deviation of about 10-15 mV. This level of noise makes it impossible for one to use numerical differentiation to directly determine the current based on the $\frac{dV_T}{dt}$ as in Equation 2.10, even if the transient component discussed in Section 2.4.2 has decayed to zero. This provides the motivation to determine the leakage currents via a model fit to the experimental data.

As the complete flash cell structure contains the same width, length, tunnel oxide thickness and doping has the dimensions as the TREQs it is natural to propose a V_T (t) model for a flash cell at the OSS condition that includes the components seen in Equation 2.12. For a given electric field in the ONO, the current is assumed to be constant, which leads to the addition of a linear component in the V_T (t) model, to become Equation 2.14. The constants A_0 and τ are analogous to those described in Equation 2.12 for the TREQs; A_0 represents how much charge is trapped in the tunnel oxide and τ represents a detrapping time constant. A_1 is related to the geometry of the cell and the intial amount of charge on the floating gate, while A_2 is the component related to the change in V_T induced by charge loss.

$$V_t(t) = A_0 e^{-\frac{t}{\tau}} + A_1 + A_2 t \quad Equation\ 2.14$$

The model provides a good fit to the experimental data, as shown in Figure 42.

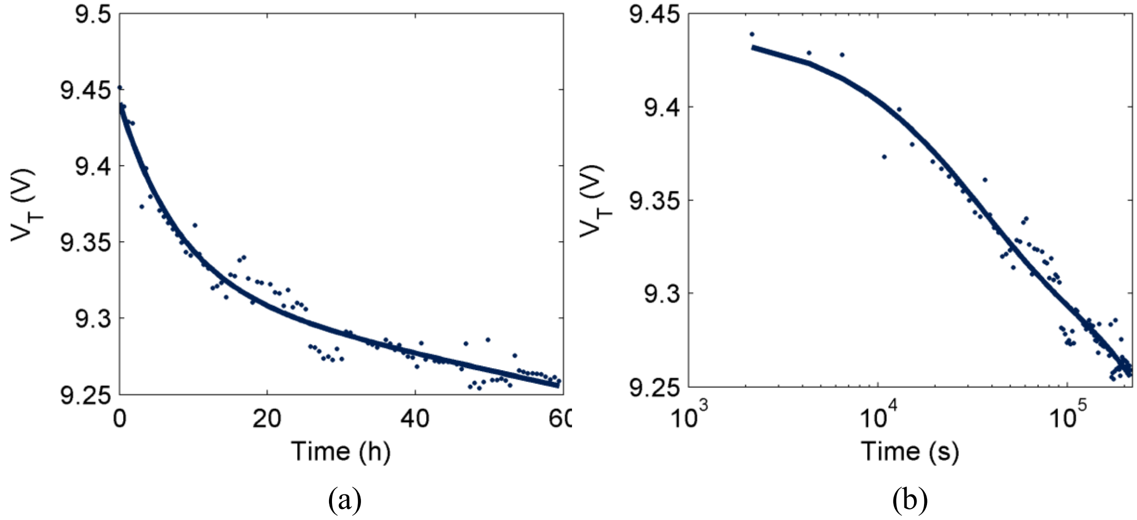


Figure 42: OSS measurement at 250°C in (a) linear and (b) semi log timescales with a fit of Equation 2.14 (solid line) at ΔV_{ONO} of 5.4 V (E_{ONO} = 4.2 MV/cm)

As each OSS measurement has a different initial V_T -value and the interest is in the $\frac{dV_T}{dt}$ instead of the V_T itself the data from multiple OSS measurements are shown together in the form of ΔV_T (t) in Figure 43. One can see that the total V_T – shift does not necessarily scale with the E_{ONO}, this can be explained by the variability of the tunnel oxide trapping (A_0 and τ) from one test to another. This is not a problem as only the $\frac{dV_T}{dt}$ behaviour is of interest, so long as the OSS measurement is long enough that the tunnel oxide detrapping does not convolute the extraction of the leakage current.

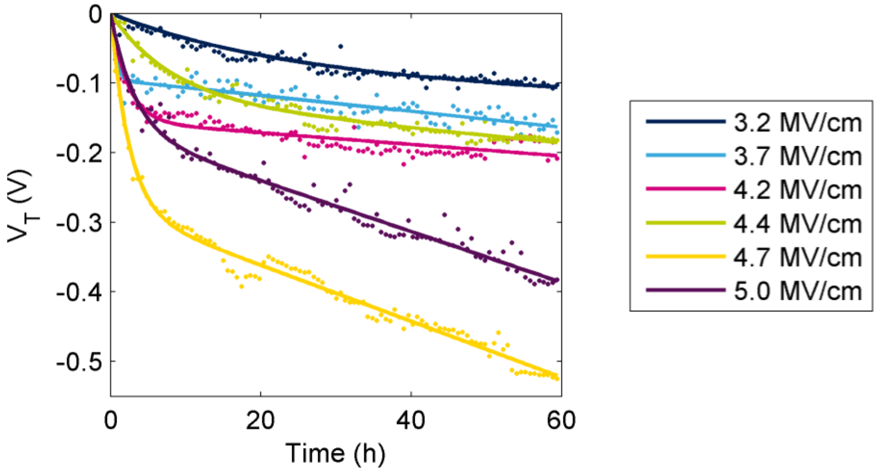


Figure 43: $V_T(t)$ shift curves for multiple OSS measurements (points) and their fits to the model in Equation 2.14 (solid curves) at 250°C

The constant component, A_2, is considered to be due to the leakage current and is used to build the IV curve in Figure 44. For the sake of comparison, a 1V V_T – shift over ten years is approximately a 5×10^{-25}A current or about one electron every four days. The IV data extracted from OSS measurements is fitted to a model for trap-assisted tunnel currents, given in Equation 2.1.5 [Bel02].

$$ln(I) = A_3(E_{ONO}) + A_4 \quad Equation\ 2.1.5$$

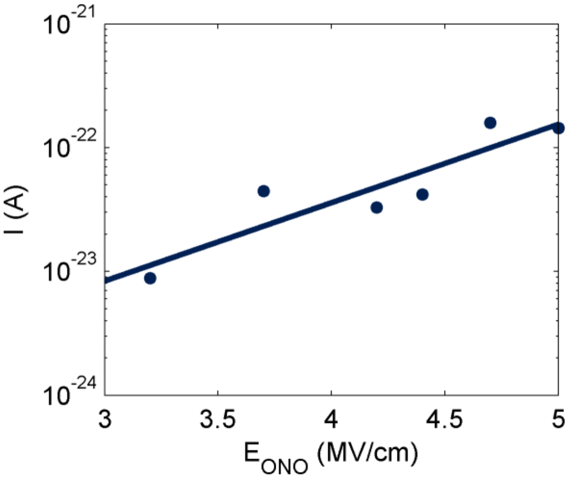


Figure 44: IV curve constructed from OSS measurements at 250°C

The direct IV measurement of the large area (10,000μm²) ONO test structure at 250°C is compared with the currents from the OSS measurements in Figure 45. Despite having an area over 100,000 times smaller than that of the poly-Si/ONO/poly-Si test structure, the OSS technique is able to measure current densities several orders of magnitude below direct measurements. The low-field current density on the order of 10^{-10} A/cm² represents the noise level of the direct measurement at 250°C. As expected, the Fowler-Nordheim extrapolation from high-field regime considerably underestimates the low-field currents as it is not considered to be the dominant mechanism at low fields. The current density at approximately 3 MV/cm is coherent with the predicted values of a few tens of fA/cm² [Moo13].

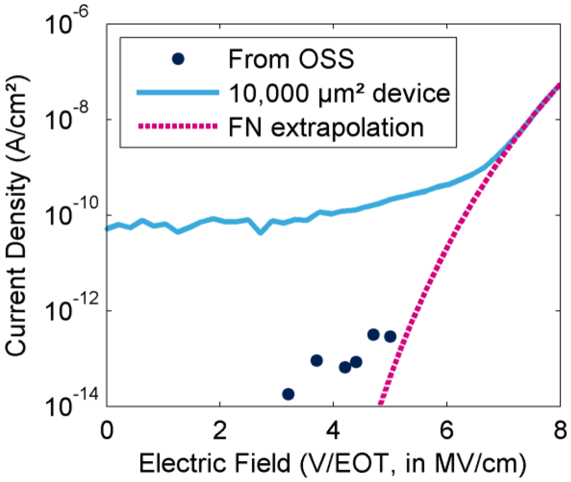


Figure 45: Current density vs electric field for nominal devices by OSS and large area test structures by direct measurement with Fowler-Nordheim extrapolation at 250 °C

2.4.4 Comparison to Data Retention

By extrapolating the ONO IV curve in Figure 44 to lower electric fields, the leakage currents at the data retention condition were estimated. From these currents, the component of the data-retention V_T – shift related to the charge loss from the floating gate to the control gate was deduced using Equation 2.10. The calculated V_T – shifts (solid lines) are compared, in Figure 46, with the average measured V_T – shifts from 20 individual flash cells (open circles with standard deviation error bars) from data retention tests starting from three different initial V_T values.

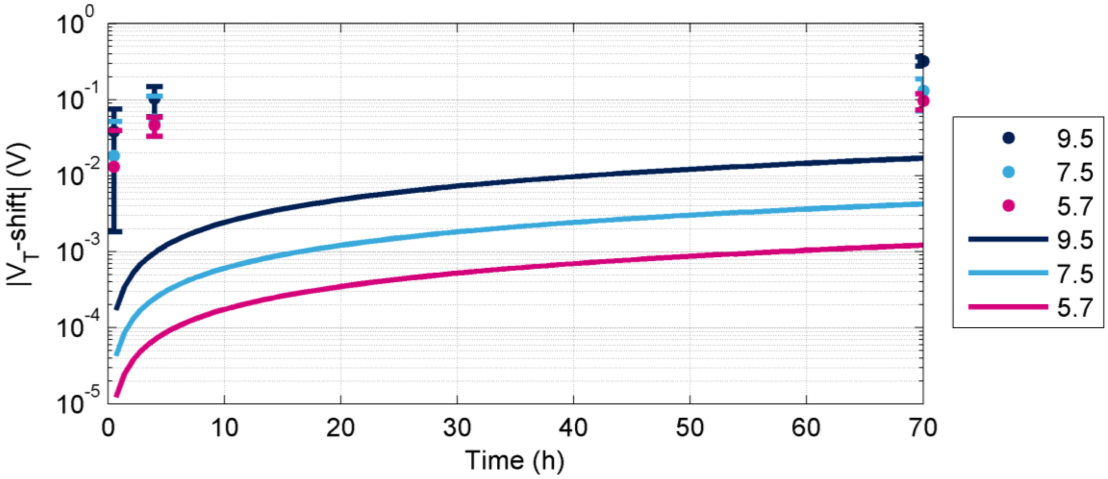


Figure 46: Measured V_T – shift (open circles) for three different initial V_T values (in V) during a retention bake and the predicted contributions from ONO leakage currents (solid lines)

One can see that the ONO leakage currents represent only a small contribution to the V_T – shift measured in standard bake tests. This is expected for the high quality, sub-13nm EOT ONO of a 40 nm embedded flash product. We can therefore see for the first time that intrinsic charge loss in the data retention tests comes from the tunnel oxide and not the ONO.

2.5 Conclusions on ultra low field current measurements

The low-field leakage currents that define whether or not a flash cell fails in terms of data retention have long been beyond the resolution of direct measurements. Floating gate measurements have been used to push the limits of low-field ONO characterization however the test structures are different from nominal flash devices, and the lowest electric fields are still not low enough. In place of direct measurements, models to describe charge loss through the ONO have been proposed by conducting long term data retention bake tests on programmed cells. The development of these models required several different sets of ONOs having a range of thicknesses; this requirement will be difficult to meet as the first sets of new IGDs are integrated and tested.

We have proposed a new technique, Oxide Stress Separation (OSS) for the evaluation of low and medium-field leakage currents across the ONO inter-gate dielectric in floating gate flash memories. The OSS technique measures the charge loss from a nominal flash cell, improving the resolution over standard test devices 100,000 times larger. The charge loss is confined to the pathway across the ONO by the careful selection of the biasing conditions and threshold voltage of the cell. Currents on the order of 10^{-22} to 10^{-23} A can be extracted from OSS measurements based on the sensitivity of a flash cell's threshold voltage to the charge on its floating gate.

The technique allows for the evaluation of low-field leakage currents on nominal flash cells and has been applied to determine the contribution of ONO leakage to the data retention properties of 40 nm embedded flash cells. Most of the intrinsic data retention losses are considered to be via the tunnel oxide. Leakage currents through the ONO layer to the control gate are currently a very small part of the data-retention shift; this may change considerably as high-k materials are integrated into embedded flash memories.

2.6 Bibliography

[Bel02] H.P. Belgal, et al., “A new reliability model for post-cycling charge retention of flash memories” in *IEEE Reliability Physics Symposium Proceedings*, Dallas, TX, USA, Apr 2002

[Bez03] R. Bez, E. Camerlenghi, A. Modelli and A. Visconti, "Introduction to flash memory," in *Proceedings of the IEEE*, vol. 91, no. 4, pp. 489-502, April 2003.

[Des99a] B. De Salvo, et al., “Experimental and Theoretical Investigation of Nonvolatile Memory Data-Retention”, *IEEE Transactions on Electron Devices*, vol. 46, no. 7, 1999

[Des99b] B. De Salvo, et al., “Investigation of low Field and high temperature SiO_2 and ONO leakage currents using the floating gate technique”, *Journal of Non-Crystalline Solids*, vol. 245, 1999, pp. 104-109

[Dob17] A. Dobri, et al., “Development and application of the Oxide Stress Separation technique for the measurement of ONO leakage currents at low electric fields in 40 nm floating gate embedded-flash memory”, *Microelectronics Reliability*, vol. 69, Feb 2017

[Fis91] B. Fishbein, D. Krakauer and B. Doyle, "Measurement of very low tunneling current density in SiO_2 using the floating-gate technique," in *IEEE Electron Device Letters*, vol. 12, no. 12, pp. 713-715, Dec. 1991.

[Gae80] F. H. Gaensslen and J. M. Aitken, "Sensitive technique for measuring small MOS gate currents," in *IEEE Electron Device Letters*, vol. 1, no. 11, pp. 231-233, Nov 1980.

[Hu10] C. Hu, “Chapter 5: MOS Capacitor” in *Modern Semiconductor Devices for Integrated Circuits,* 1st Edition. Pearson. 2010

[Kim04] J.H. Kim and J.B. Choi, "Long-term electron leakage mechanisms through ONO interpoly dielectric in stacked-gate EEPROM cells," in *IEEE Transactions on Electron Devices*, vol. 51, no. 12, pp. 2048-2053, Dec. 2004.

[Lee10] S.H. Lee, et al., “Investigation on the Retention Reliability of Scaled $SiO_2/Al_xO_y/SiO_2$ Inter-Poly Dielectrics for NAND Flash Cell Arrays” *IEEE Electron Device Letters*, vol. 31, no. 4, Apr., 2010

[Man99] Y. Manabe et al., "Detailed observation of small leak current in flash memories with thin tunnel oxides," in *IEEE Transactions on Semiconductor Manufacturing*, vol. 12, no. 2, pp. 170-174, May 1999. [Nis84] Y. Nissan-Cohen, "A novel floating-gate method for measurement of ultra-low hole and electron gate currents in MOS transistors," in *IEEE Electron Device Letters*, vol. 7, no. 10, pp. 561-563, Oct 1986.

[Moo13] P. Moon et al, “Methodology for improvement of data retention in floating gate flash memory using leakage current estimation”, *Microelectronics Reliability*, Volume 53, Issues 9–11, September–November 2013, Pages 1338-1341

[Nis86] Y. Nissan-Cohen, “A Novel Floating-Gate Method for Measurement of Ultra-Low Hole and Electron Gate Currents in MOS Transistors”, *IEEE International Electron Device Letters*, vol. 7, no. 10, 1986

[Pav97] P. Pavan, R. Bez, P. Olivo and E. Zanoni, "Flash memory cells-an overview," in *Proceedings of the IEEE*, vol. 85, no. 8, pp. 1248-1271, Aug 1997.

[Pia10] F. Piazza, et al., “High Performance Flash Memory for 65 nm Embedded Automotive Application” in *IEEE International Memory Workshop*, Seoul, Korea, May, 2010

[Tao07] G. Tao, C. Ouvrard, H. Chauveau and S. Nath, “Experimental study of carrier transport in multi-layered structures”, *Microelectronics Reliability*, Volume 47, Issues 4–5, April–May 2007, Pages 610-614

[Zak11] A. Zaka, et al., “Characterization and 3D TCAD simulation of NOR-type flash non-volatile memories with emphasis on corner effects”, *Solid-State Electronics*, vol. 63, 2011, pp. 158-162

Chapter 3: High-k Materials Integration in 40 nm Embedded Flash

Table of contents

3.1 Introduction

As one must walk before they can run, one should also make small changes towards high-k materials in the current node before attempting to implement more complicated gate stacks into products that do not yet exist. In this context, the goal of this chapter is to describe the integration of a high-k material in a 40 nm embedded flash product from STMicroelectronics. This is the current technology node in development for embedded non-volatile memories. This node utilizes a $SiO_2/Si_3N_4/SiO_2$, or "ONO" integrate dielectric and a poly-silicon control gate, with silicon based gate oxides. The product flow includes LV (logic), flash (NVM), and High Voltage (HV) devices in a process flow outlined in the following pages. The process uses a *two-poly* integration where the logic devices' gate is also the control gate of the flash. The process is very modular and the performance of the LV devices is not negatively impacted by the flash cells. The flash cells are affected by the logic processing, as they are exposed to the logic gate processing [Pia10]. As the 40 nm CMOS logic processes in STMicroelectronics were already well defined, any suggestions for improving the flash cell at the 40 nm node needed to be compatible with these processes.

The choice of $SiO_2/Al_2O_3/SiO_{2,}$ or "OAO" to replace the ONO at 40 nm was made because alumina presents a good compromise in k-value vs bandgap (see Figure 14) at 9 and 8.6 eV, respectively [Cas02]. Alumina is regularly used in the CEA-leti cleanroom as well, so there is readily accessible process knowhow. The OAO trilayer is used because switching the nitride with alumina represents a move to high-k material while maintaining a similar OXO structure, which allows for a similar integration approach. The inclusion of OA samples, those without a top-oxide was a logical extension from OAO as alumina is considered to be stable and to offer a large enough band-gap that the top SiO_2 layer isn't necessary [Pow08]. The literature is not completely consistent in the recommendation for or against a top-oxide above an alumina [Pow07,Pow08]. Both O_3 and H_2O precursors are tested as oxygen sources during Al_2O_3 growth as they were shown to impact the flash cell operation [Shu09].

In order to obtain results at multiple EOT levels, three different thicknesses of alumina were deposited, with and without top oxides as shown in Figure 47 and outlined in more detail in Table 3. ONO reference splits were also included to isolate the effects of changing the IGD from the effects related to the other process changes, for example different spacers or ion implantations. Having three thicknesses should provide a good enough range of EOT values that trends can be established without increasing the experimental complexity to unmanageable levels. One IGD consisting only of a 20 nm alumina layer was also fabricated.

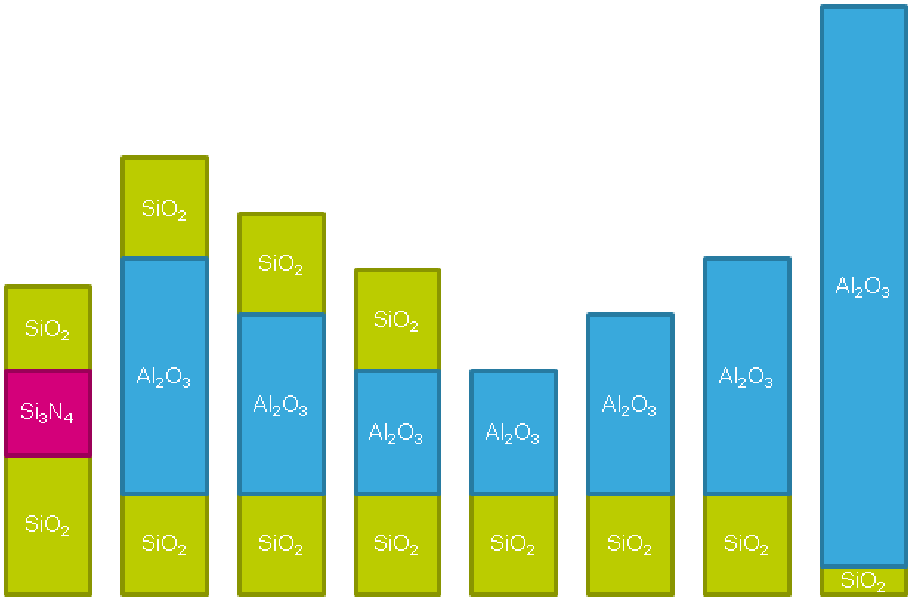


Figure 47: Overview of splits for 40 embedded flash batches integrating alumina into the IGD

Table 3: Table of splits for 40 embedded flash batches with alumina

Sample	ONO	OAO 8.5 nm	OAO 6.5 nm	OAO 4.5 nm	OA 4.5 nm	OA 6.5 nm	OA 8.5 nm	A200 200 nm
Top SiO_2	Yes	Yes	Yes	Yes	-	-	-	-
Nitride	Yes	-	-	-	-	-	-	-
Alumina (precursor)	-	Both H_2O and O_3						O_3
Anneal	Yes	5 minutes at 950 °C or 30 seconds at 1030 °C						
Bottom SiO_2	4 nm SiO_2							

In a modern 300 mm microelectronics fab this change of one layer requires many changes and updates to the rest of the process flow in order to avoid the risk of alumina contamination or simply to compensate for other changes. The purpose of this chapter is to outline the necessary changes.

3.2 Overview of 40 nm embedded flash process flow and modifications for alumina integration

A typical 40 nm process flow includes hundred of processing steps (thousands if measurements are included) to make embedded flash devices; they are greatly simplified to make the flow-chart shown in Figure 48. As alumina is not commonly used in the “Front End” (before metallization) in the production fab, a modified version is proposed, also in Figure 48, in order to limit potential contamination of cleanroom equipment. In the standard process flow, the ONO is deposited at roughly the 250th step; more than one thousand steps are potentially impacted by the change from the standard ONO to the OAO. The impact on the rest of the process steps ranges from no change required for a standard decontamination of the wafer backside to several months of process development to develop a new etching protocol. This section describes the process flow used at STMicroelectronics to produce embedded flash memories at the most recent technology nodes. It focuses on the adaptations and development required to incorporate alumina as part of the IGD. This general process flow is patented under the “Method of manufacturing an electrically programmable, non-volatile memory and high-performance logic circuitry in the same semiconductor chip” [Pes01].

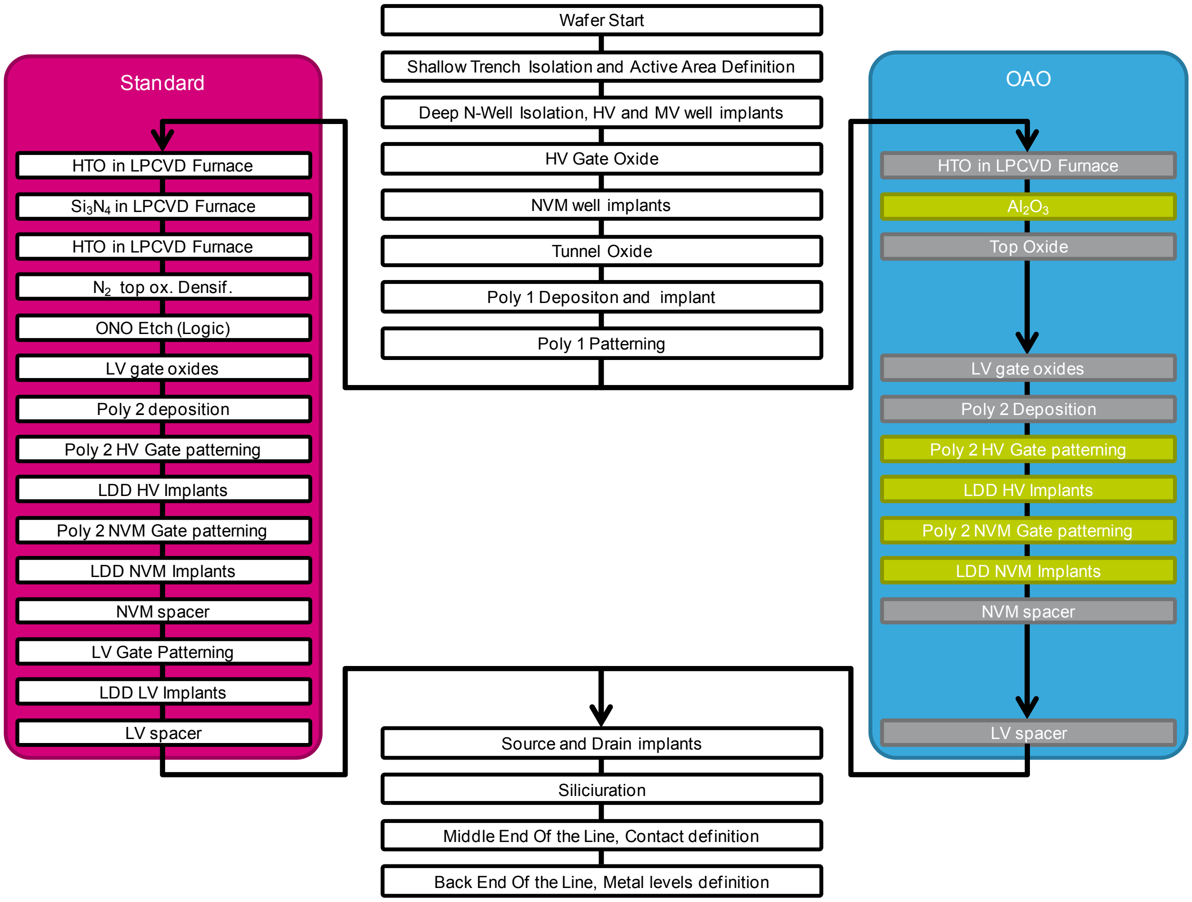


Figure 48: Process flow overview for standard and OAO 40 embedded flash processes carried out at the CEA-Leti highlighted in green while processes modified at STMicroelectronics are highlighted in grey

3.2.1 STI and Active definition

From the bare Si wafer, the Shallow Trench Isolation (STI) is completed in order to define the active regions of the die. Several ion implantation steps are completed to make the deep n-well isolation and to define the N and P wells of the HV devices. The result is shown in Figure 49. Note that the drawings are not necessarily to scale and are intended only to show general configurations.

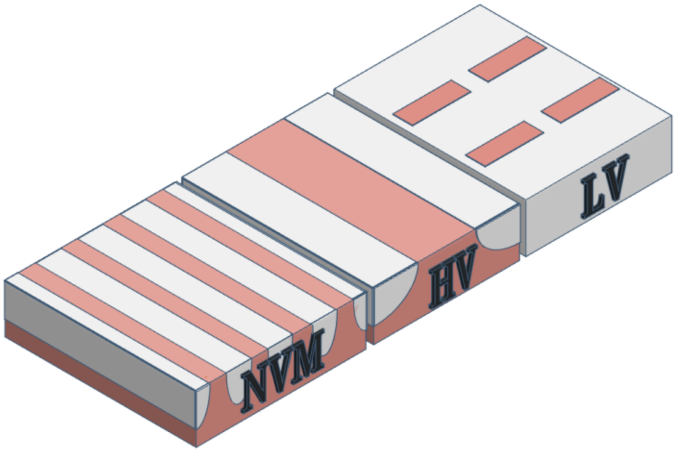


Figure 49: NVM, HV and LV regions post STI and active region definition

3.2.2 HV and Tunnel Oxide definition

The HV oxide is deposited and acts as the screen oxide for the P-well implant in the NVM zone. This HV oxide is then removed in the NVM zone before the tunnel oxide is deposited. The end result is an oxide layer of different thicknesses covering the wafer as shown in Figure 50.

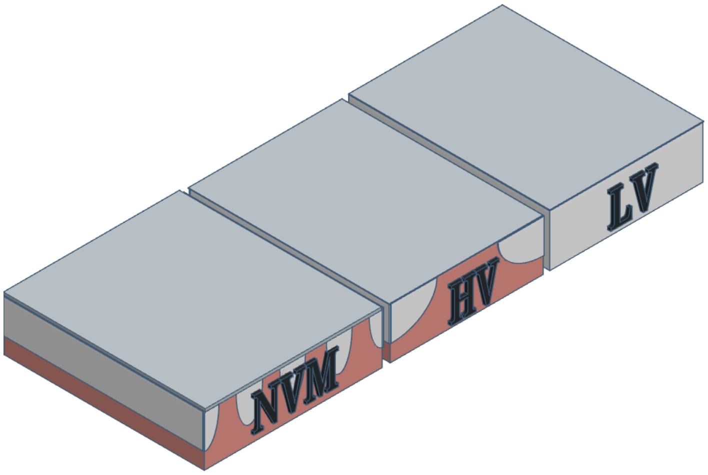


Figure 50: NVM, HV and LV regions post HV and tunnel oxide definition

3.2.3 Poly1 Deposition and patterning

The "Poly1" layer of poly-Si is deposited and then patterned to remove the Poly1 from LV zone, define the gates in the HV zone and define lines in the bit line direction in the NVM zone, as shown in Figure 51. This is the last step where the OAO batches followed the exact same steps as a standard batch.

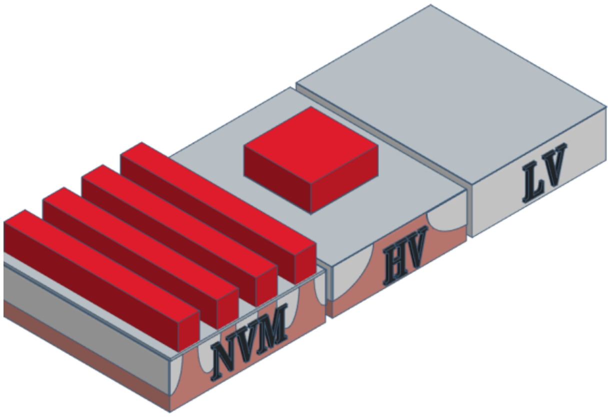


Figure 51: NVM, HV and LV regions post Poly1 deposition and patterning, note that the bars of Poly1 in the NVM region will later form the floating gates

3.2.4 ONO Deposition and patterning

The ONO needs to be of good quality and very conformal as the Poly1 is already patterned. Therefore, the ONO is deposited by Low Pressure Chemical Vapour Deposition (LPCVD), and then removed from the LV zone. For the OAO batches the same requirement exists. The bottom oxide is deposited by LPCVD and then the wafers were sent to the CEA-Leti for the alumina deposition by Atomic Layer Deposition (ALD) in an ASM Polygon using TMA as the Al-precursor and either H_2O or O_3 as the oxygen precursor. The ALD meets the standards for quality and conformality required for this deposition. The alumina from the backside of the wafer is removed to insure that no alumina is present on the backside when they return to the production fab. After the cleaning of the wafer backside the wafers were annealed for five minutes at 950°C, generally under a nitrogen environment.

For the top oxide a good quality SiO_2 should be used, however the exposed alumina was not allowed into the LPCVD tool. Therefore a different method was used; a SNOW (Side Nucleation On Wall) type oxide was deposited. This is a cycled Plasma Enhanced Chemical Vapour Deposition (PECVD) at 350 °C where organic liquid containing silicon in a helium carrier gas coats the wafer before being oxidized by oxygen in an argon carrier gas. A single cycle deposits approximately 0.4 nm and can be repeated multiple times, thirteen times in the case of the top-oxide of the OAO. This differs from an ALD process because in ALD the steps are self-limiting whereas the first organic seed-layer is not truly self-limiting at the mono-layer.

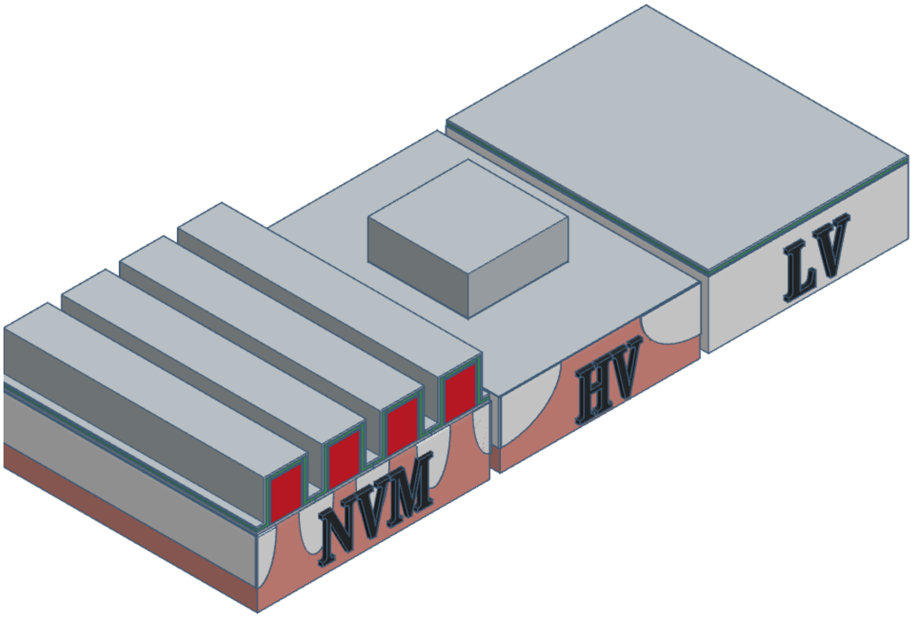


Figure 52: NVM, HV and LV regions post OAO deposition (note: the NVM region is shown in cross section)

At this point the OAO is not removed from the LV zones for two reasons: functioning LV devices are not necessary for the testing of the flash chip and leaving the OAO in the LV zones means one less etching step, saving process development resources.

3.2.5 Gate Oxide growth and patterning

As the LV devices are not patterned in these batches, the gate oxide processes are effectively just additional anneals for the exposed OAO. The tool in which these oxidations take place heats the wafer from the backside during a rapid thermal oxidation and presents the context for an interesting problem.

In the normal process flow, the HV oxides, Poly1 and the ONO are still present on the backside of the wafer as there is no reason to remove them. In the modified flow the alumina is removed after deposition so the stack on the backside of the wafer includes the Poly1, covered by the 4 nm bottom oxide (instead of the ONO). This 4 nm oxide is quickly removed during the backside decontamination steps involving HF, leaving the Poly1 exposed on the backside of the wafer. During the rapid heating under an inert atmosphere some of this Poly-Si sublimed and deposited itself onto a detector. This deposit interfered with the temperature measurement of the wafer and the tool increased the lamp's power output to compensate to a point where the lamp burnt out. This rapid heating had never caused problems in the standard configuration because the Poly1 was covered by the ONO on the backside of the wafer. This example shows some of the potential complications that can arise during the testing of a new integration; even the changes that appear completely innocuous may involve repercussions at a later processing step.

3.2.6 Poly2 Deposition and Poly2HV Etch

The 80 nm Poly2 is usually deposited using a production-scale furnace, so the OAO lots were processed on a single-wafer tool that was better suited to a small batch of wafers containing alumina. The first critical OAO etching step is the Poly2HV etch which does not directly impact the NVM or LV (they remain covered by the photo resist). This etching step patterns the gates of the HV logic, as shown in Figure 51, that are used in the charge-pumping circuits; these circuits are required for the operation of the 4 MB test-chip. This etch also forms the OAO capacitors, large area devices with contacts on the Poly1 and Poly2 that are used as charge capacitors on commercial products and to test the reliability and medium/high-field conduction on our samples (not shown in the figure).

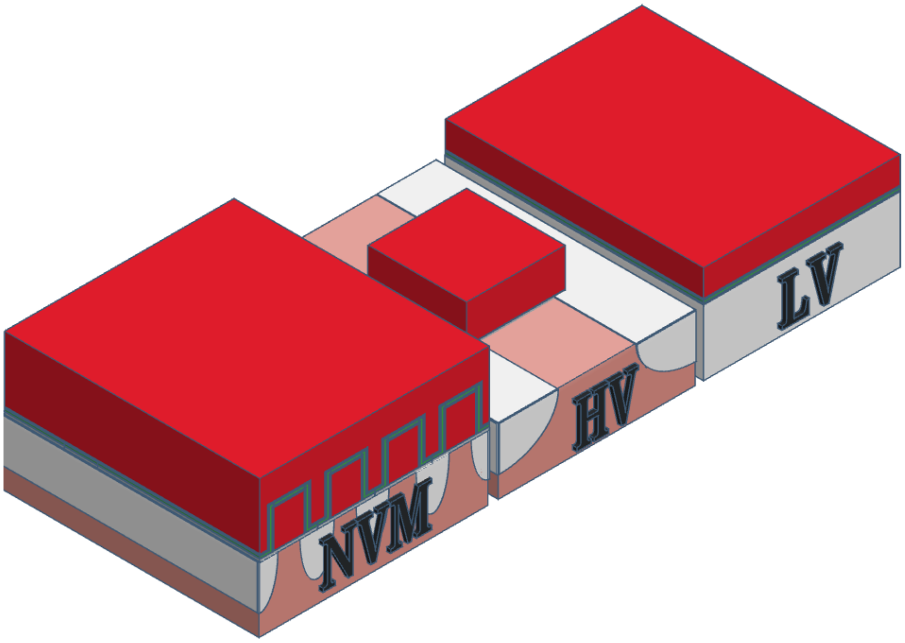


Figure 53: NVM, HV and LV regions post Poly2HV etch (note: the NVM region is shown in cross section)

3.2.6.1 Poly2HV Etch: 1st trials

As the Poly2HV etch must remove the Poly2 and the IGD, different options are available. The first step in developing the etching process at the CEA-Leti was to calibrate the etch rate and end-point detection for the 80 nm layer of Poly2; this was done using full-sheet (or non-patterned) wafers of Poly2 deposited on an oxide layer. A first set of three trial wafers was then completed in the CEA-Leti cleanroom where the entire poly-Si/SiO_2/Al_2O_3 was removed in a dry etch, summarized in Table 4. The first three etches were chosen to show the results of only the dry etch (HV 1A), with the standard stripping protocol added (HV 1B) and with a modified stripping to improve alumina removal (HV 1C). The modification was the addition of a wet etch step to remove alumina. This step is referred to as LAE (Leti Alumina Etch) in lieu of its chemical composition for reasons of confidentiality. The SC1 refers to a standard clean involving hydrogen peroxide and sodium hydroxide to remove organic and metallic contaminants while [Plu00].

Table 4: Poly2HV Etch 1st trials

Wafer	Dry Etch	Stripping and Wet Etch
HV 1A	Poly2 + SiO_2 + Al_2O_3	HF (0.1%, 30s)
HV 1B		O_2 plasma (40 s) + HF (1%, 60s) + SC1
HV 1C		O_2 plasma (40 s) + HF (1%, 60s) + LAE + SC1

This dry etch approach (HV 1A) successfully removes the poly-Si/SiO_2/Al_2O_3 in larger, open areas, arriving on the bottom SiO_2 layer, as shown in Figure 54. When the Poly1 is patterned, such as in the HV MOS devices, shown in Figure 55, the dry etch was insufficient for the removal of the OAO from the vertical sidewalls. The dry etch also induces “trenching” where the tunnel oxide is degraded close the sidewall. The decrease in alumina on the sidewalls after the LAE treatment (HV 1C) was promising because it indicated that the wet etch can remove alumina.

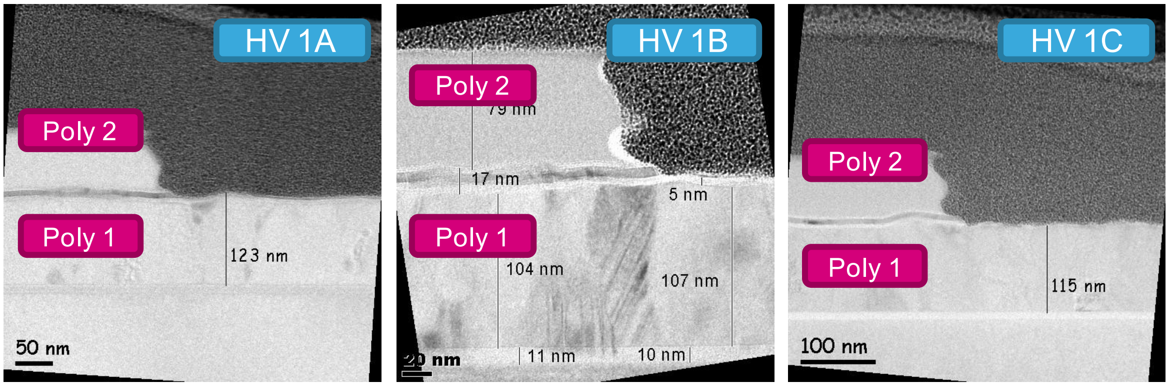


Figure 54: TEM images of ONO capacitor opening after first Poly2HV etch trials

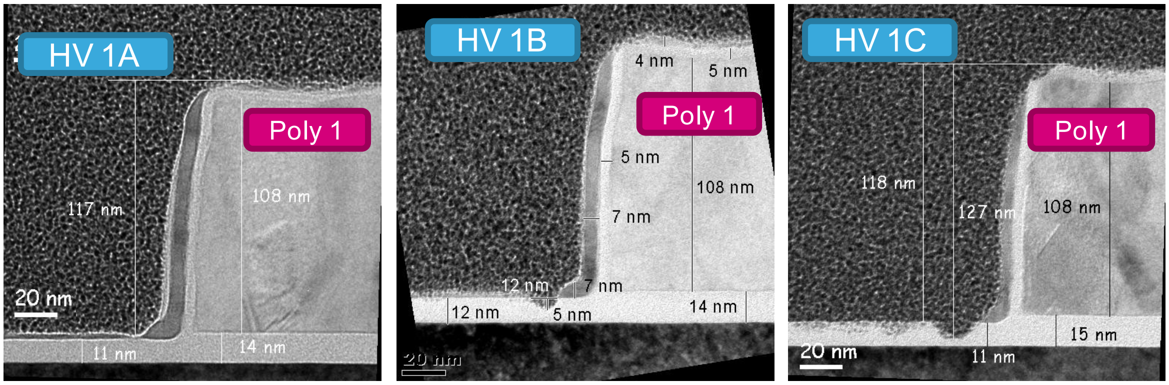


Figure 55: TEM images of HV MOS devices after first Poly2HV etch trials

3.2.6.2 Poly2HV Etch: 2nd trials

In order to improve the removal of the alumina, the alumina dry etch time was increased. The bias was also reduced in order to avoid the trenching problem during the dry etch. In addition, a “dry plus wet” approach was also used for the second and third wafer of the set of etch trials. In these cases, only the Poly2 is removed by a dry etch before the top oxide and alumina removed by wet etching as summarized in Table 5. Two different levels of over etch (etch time longer than the ideal time based on etch rates) were tested, 15 and 100%.

Table 5: Poly2HV Etch 2nd trials

Wafer	Dry Etch	Stripping and Wet Etch
HV 2A	Poly2 + SiO_2 + Al_2O_3	O_2 plasma (40 s) + HF (1%, 60s) + SC1
HV 2B	Poly2	O_2 plasma (50 s) HF (1%, 60s) + LAE (15% over etch) + SC1
HV 2C	Poly2	O_2 plasma (50 s) HF (1%, 60s) + LAE (100% over etch) + SC1

The TEM images in Figure 56 confirm that the full dry etch of the poly-Si/SiO_2/Al_2O_3 is not feasible. Although the trenching into the tunnel oxide is reduced, most of the alumina remains on the sidewalls. Wafers HV 2B and 2C were much more promising as the alumina was removed from the sidewalls. There appears to be a small amount of alumina remaining in the lower corner of wafer HV 2B, having only a 15% LAE over etch. In order to err on the side of caution in terms of alumina removal, the 100% LAE over etch processes is considered to be the best process as it does not pose a danger to the underlying SiO_2.

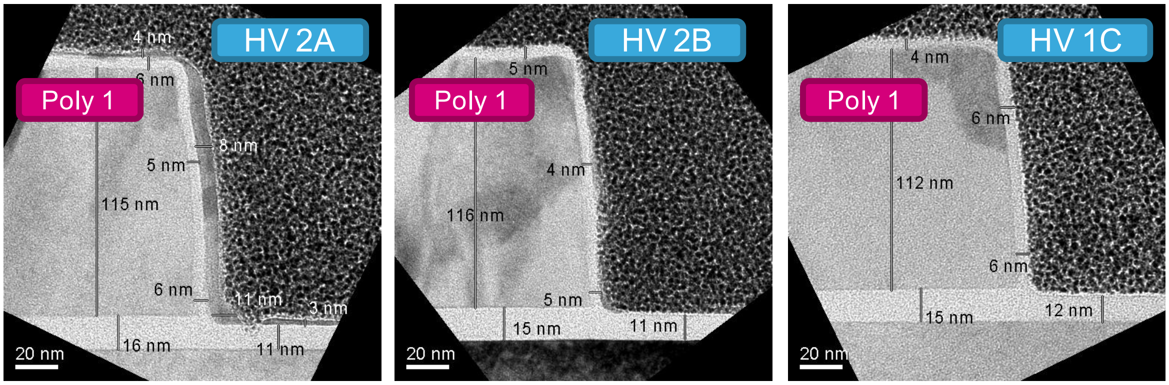


Figure 56: TEM images of HV MOS devices after second Poly2HV etch trials

One potential danger of over etching the alumina is that the LAE could undercut the Poly2 in the areas where the OAO remains exposed, such as in the lower corners of the OAO capacitor openings. Another TEM analysis was completed in order to either confirm or eliminate the fears of undercutting. Figure 57 shows no evidence of alumina undercutting so the etch process for HV 1C was used as the reference process for the rest of the wafers.

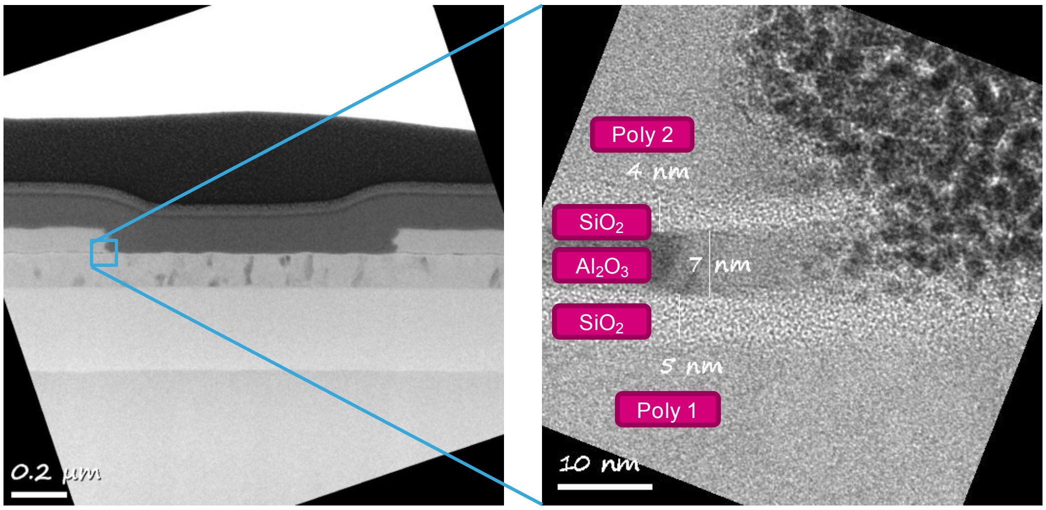


Figure 57: TEM images of ONO capacitor opening of wafer HV 2C having 100% LAE over etch to confirm that there is no risk of undercutting the Poly2

3.2.7 Poly2NVM Etch

The Poly2NVM etch is the most critical step in the process flow as it defines the flash cells. The Poly2, ONO and Poly1 are all etched in this step. At the end of the Poly2NVM etch, a few nm of oxide remains, acting as a screen oxide during the NVM n-type Lightly Doped Drain (NLDD) implantation. The wet strip of the photoresist following the NVM NLDD includes HF and removes the screen oxide, leading to the configuration in Figure 58, and the corresponding TEM image in Figure 59(a) and (c).

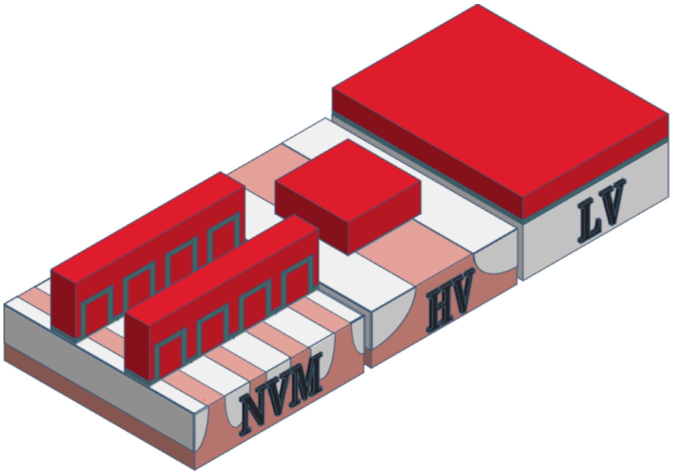


Figure 58: NVM, HV and LV regions post Poly2NVM etch

The Poly2NVM etch does not perfectly remove the vertical part of the ONO and nitride residues remain as “fences” across the trenches between the word lines, also shown in Figure 59. In the standard process flow the fences do not represent a potential contaminant while in the OAO lots, fences of alumina would greatly complicate the fabrication because they would often be exposed during later steps.

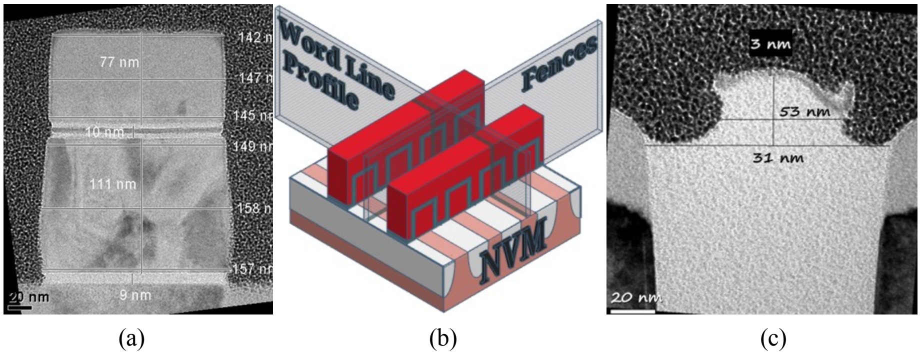


Figure 59: (a) Flash cell profile in the word line direction (b) model showing TEM cut directions (c) nitride “fences” perpendicular to the word lines (note: TEM images are from the reference process flow with an ONO IGD)

3.2.7.1 Poly2NVM Etch: 1st trials

These etch steps were developed at the CEA-Leti , with the general process being the following steps, which are summarized in Figure 60.

- BT1: Non-selective dry etch removes the first ~20 nm of the Poly2
- ME1: etch until top-oxide is exposed, then uses different etch chemistries to etch the top oxide, alumina and bottom oxide.
- BT2: non-selective dry etch
- ME2: second main-etch to remove most of the poly-Si
- OE: selective etch to etch the Poly-Si

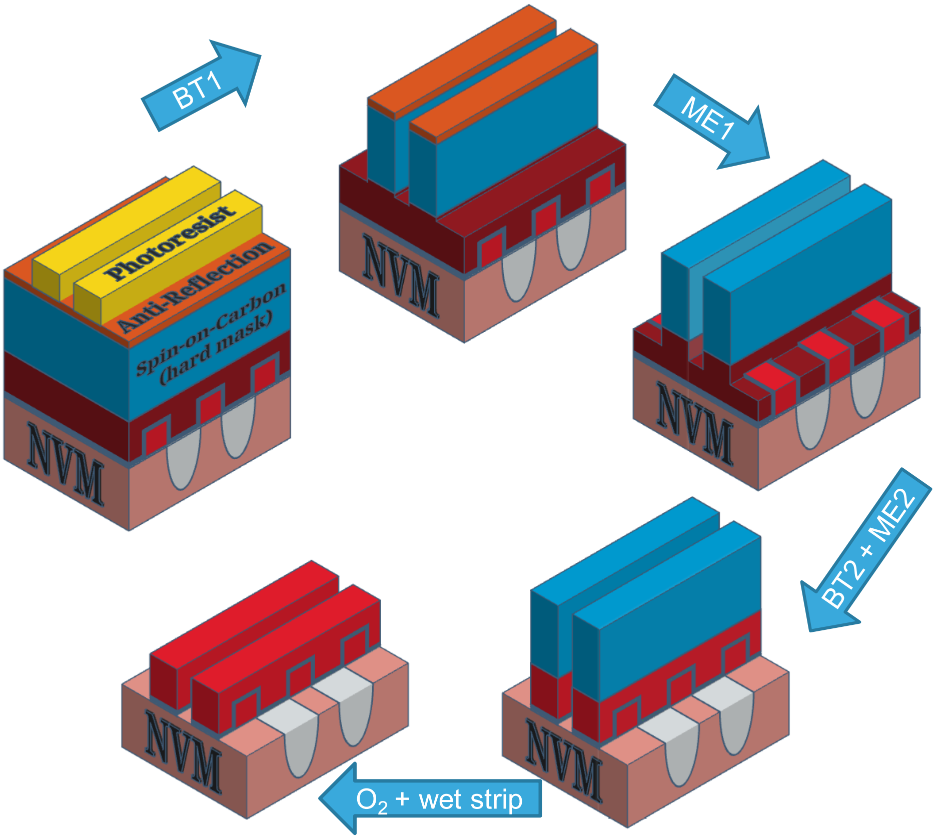


Figure 60: Graphical representation of the Poly2NVM etch process for OAO wafers

The first set of trial etches was three wafers, summarized in Table 6 to provide an intermediate snapshot part way through the dry etch and to confirm the removal efficiency of the "fences" of OAO between adjacent word lines.

Table 6: Poly2NVM Etch 1st trials

Wafer	Dry Etch	Stripping and Wet Etch
NVM 1A	Partial (stop after BT2)	No stripping + HF (0.1%, 30 s)
NVM 1B	Full	O_2 plasma + HF (1%, 45 s)
NVM 1C	Full	O_2 plasma + HF (1%, 45 s) + LAE

After the second breakthrough step the dry etch was stopped to provide an intermediate snapshot of the etching process in Figure 61. The lightly coloured material on the sidewalls of the Poly2 is polymer residues, considered to be from the CH_2F_2 that is used during the breakthrough etch. These residues block the reactants in the second Main Etch, leading to a non-vertical profile where the width of the Poly2 is about 35 nm smaller than width of the Poly1 in wafers NVM 1B and 1C. The profile is unaffected by the LEA treatment. At the end of these etch trials all of the tunnel oxide has been consumed and there is no oxide left to act as a screen oxide during NLDD NVM ion implantation, this is not ideal.

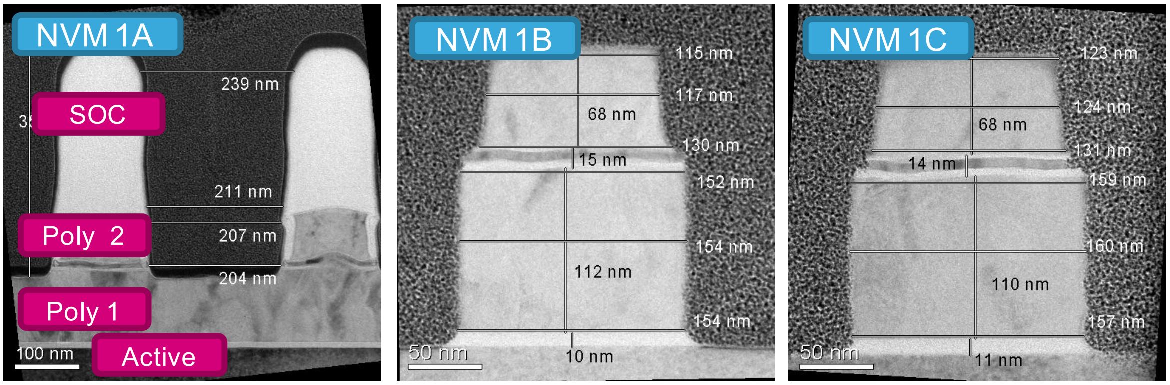


Figure 61: TEM images of the word line profiles after the first Poly2NVM etch trials

Although wafers NVM 1B and 1C showed that the LAE treatment does not impact the word line profile, the LAE is indispensable in the removal of the alumina fences, without it a massive fence of alumina would be present, as shown in Figure 62(b-c).

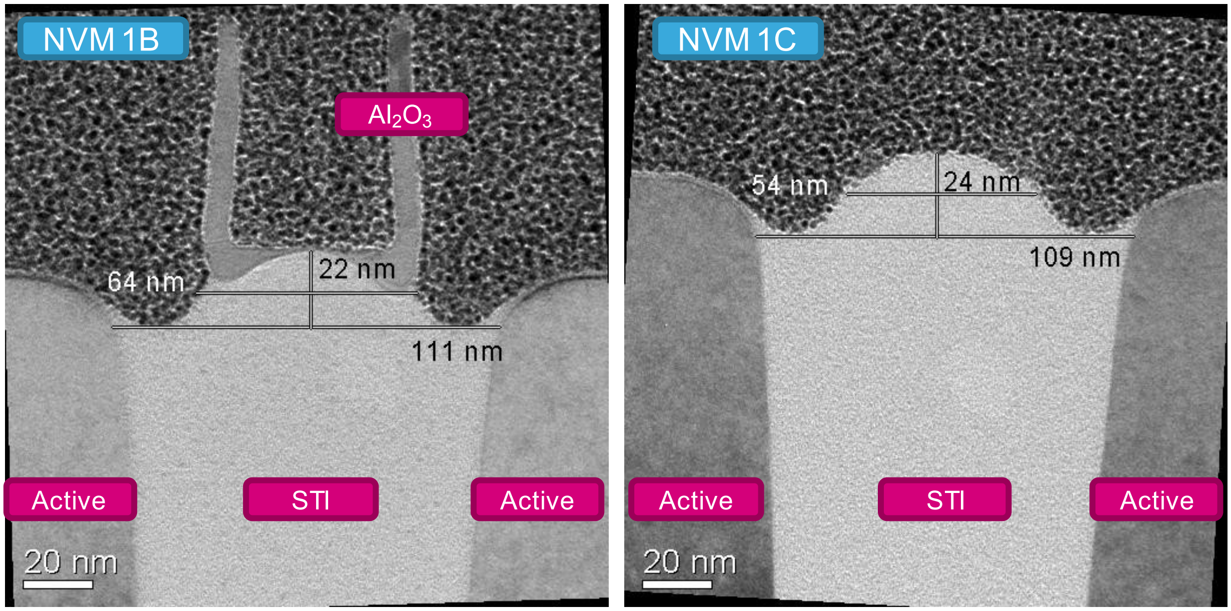


Figure 62: TEM images between the word lines to confirm alumina fence removal by LAE process after the first Poly2NVM etch trials

3.2.7.2 Poly2NVM Etch trails – 2nd loop

The differences in width at the top and bottom of the word line were concerning enough to warrant a second etch optimization loop. In the second optimization loop four more wafers were etched with a modified dry etch and different HF budgets. The dry etch used less CH_2F_2 in the hopes of reducing the thickness of the polymer sidewalls that mask the second part of the etching process. The ME2 time was also reduced in order to consume less of the tunnel oxide. All four wafers had the same optimized dry etch and different HF budgets according to Table 7. The HF budget was reduced in order to avoid the complete removal of the oxide on the active Si, which should behave as a screen oxide during ion implantation.

Table 7: Poly2NVM Etch 2nd trials

Wafer	Dry Etch	Stripping and Wet Etch
NVM 1A	Full (reduced CH_2F_2 in BT, shorter ME2)	O_2 plasma + HF (1%, 45s) + LAE
NVM 1B		O_2 plasma + HF (1%, 30s) + LAE
NVM 1C		O_2 plasma + HF (1%, 20s) + LAE
NVM 1D		O_2 plasma + HF (1%, 10s) + LAE

Unfortunately, this second process, compared with the first in Figure 63 did not yield an improved profile. However the changes in the dry etch did improve the result with respect to the amount of tunnel oxide remaining as a screen oxide.

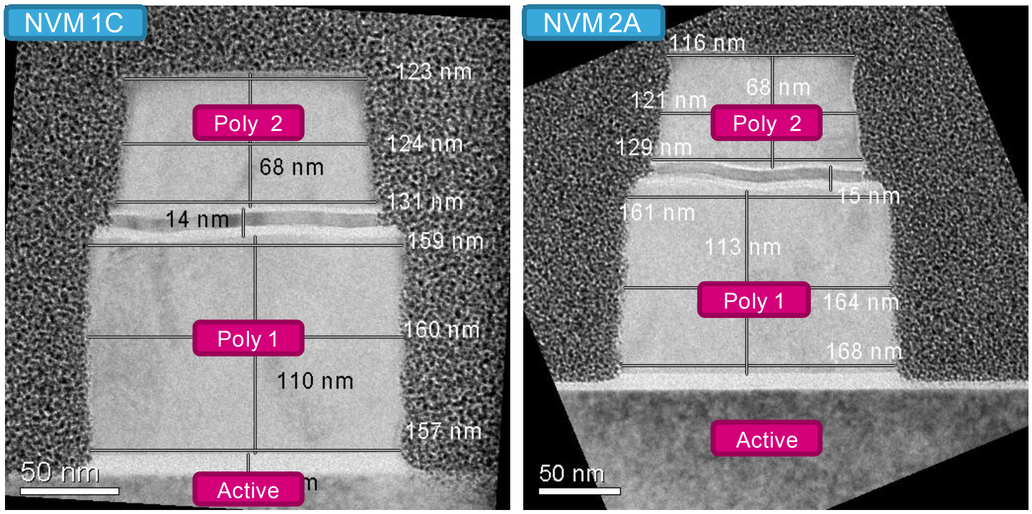


Figure 63: TEM images of the word line profiles after the (NVM 1C) first and (NVM 2A) second Poly2NVM etch trials

Although the word line profile is considered to be acceptable, the updated dry etch does not fully remove the fences as seen in Figure 64. Even at the highest of the tested HF budgets of 45 s, the fences remain. The benefits of having the screen oxide during the implant are far outweighed by the cost of potential contamination by the alumina fences in later processing steps. We have therefore chosen the process of wafer NVM 1C as the process of record for the remaining samples.

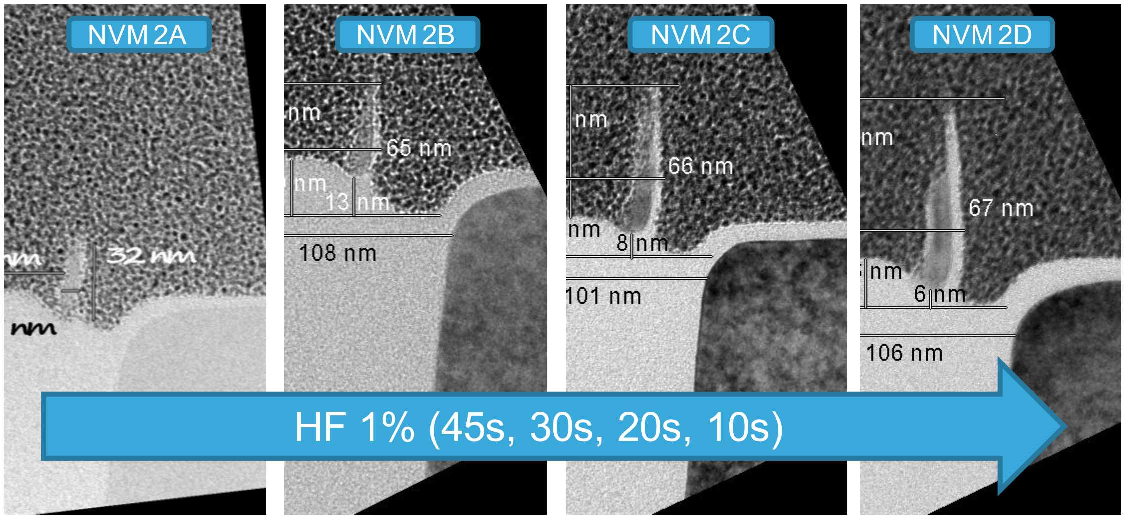


Figure 64: TEM images between the word lines to confirm the effect of HF budget on the alumina "fences" in the second Poly2NVM etch trials

3.2.8 NVM Spacer and Spacer A

For the OAO lots the re-oxidation and 10 nm nitride layer deposition by ALD to form the NVM spacer were completed before the NLDD NVM implant for two reasons. Firstly, the word line profile could lead to ion implantation into the IDG, and increased risk of implant chamber contamination if no spacer is present. Finally, the Poly2NVM etch described in the previous section does not leave a screen oxide for the implant. By moving the NVM spacer formation and etching steps before the NLDD NVM implantation, contamination risk is reduced and the implant can be made across the screen oxide that remains after NVM spacer etch, the profile of which is shown in Figure 65.

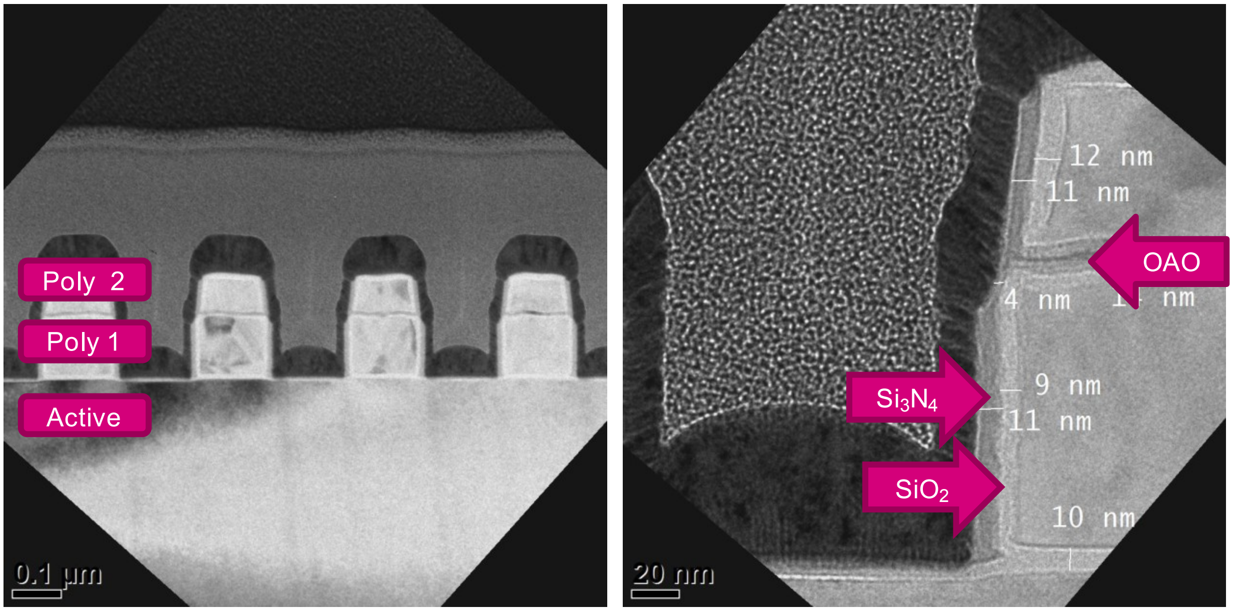


Figure 65: TEM images of the word line profile post NVM spacer etch showing the spacer comprised of ≈10 nm SiO_2 by the reoxidation and the deposited Si_3N_4

After the first NVM spacer is etched, a second spacer of 4 nm of SiO_2 and 35 nm of Si_3N_4 is added to fully encapsulate and protect the flash cells. This strategy of covering the high-k material with a spacer to reduce the risk of high-k materials' contamination is a common technique in the microelectronic industry.

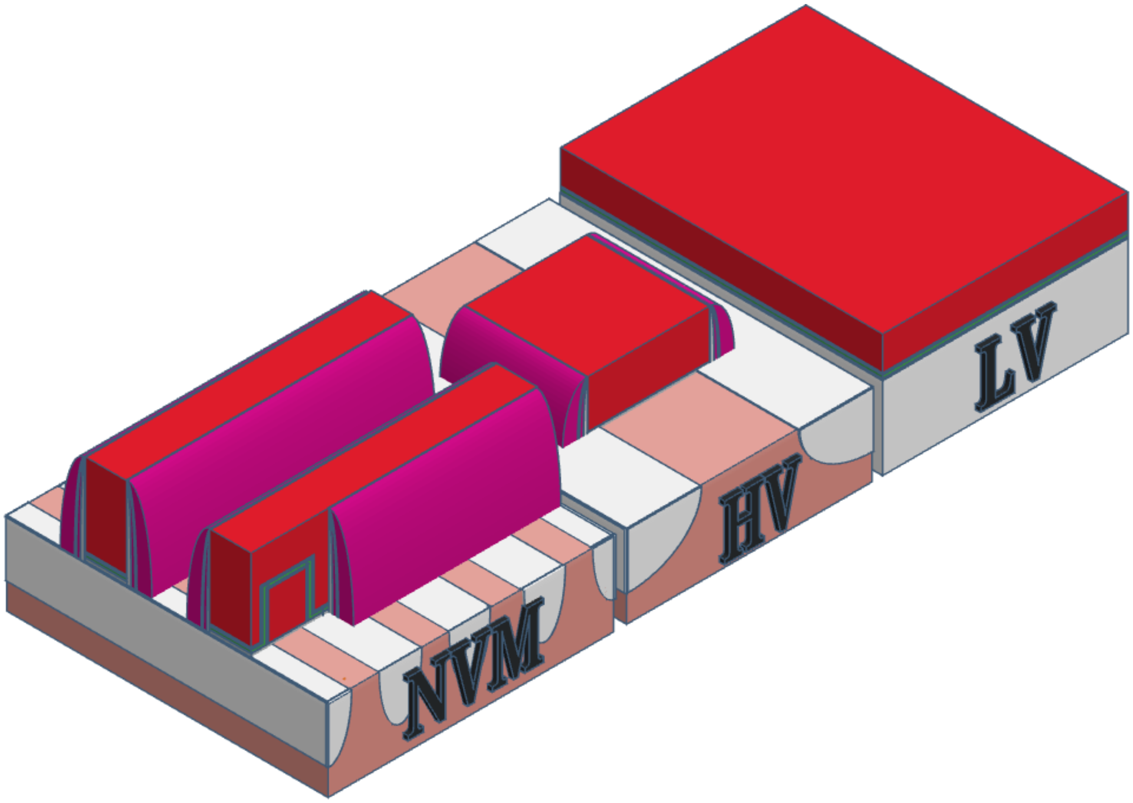


Figure 66: NVM, HV and LV regions post 35 nm spacer (note: the spacers fully surround the HV MOS and NVM devices, the cut-away is to show the spacer structure)

3.2.9 Logic Gate Etch and Spacer

As the LV devices are not necessary for the testing of the flash cells, this step was skipped for the first three pilot of wafers. However at the electrical testing stage the 4 MB test chip did not work; there were short circuits between the inputs and outputs of the chip. Analysis of the computer aided design layers and the masks showed that the Poly2 resistances defining the input and output contacts are patterned using the same mask as the LV gates. Although functional LV devices are not required, the patterning step is therefore necessary for the testing of the 4 MB test chip. The LV gate patterning was added for the remaining wafers. The addition of a last-minute gate etch represented a significant deployment of resources.

In the standard flow, a silicon dioxide hard mask is deposited after the NVM spacer, the LV gates are patterned and then the oxide is removed, using the spacer's nitride as an etch-stop layer to protect the flash. The LV gates were then etched using high-k compatible tools to remove the Poly2, top-oxide and alumina. A 10 nm nitride spacer was then deposited to complete the flash/HV spacers and to cover the alumina that was exposed during the gate etch, leading to the configuration in Figure 67.

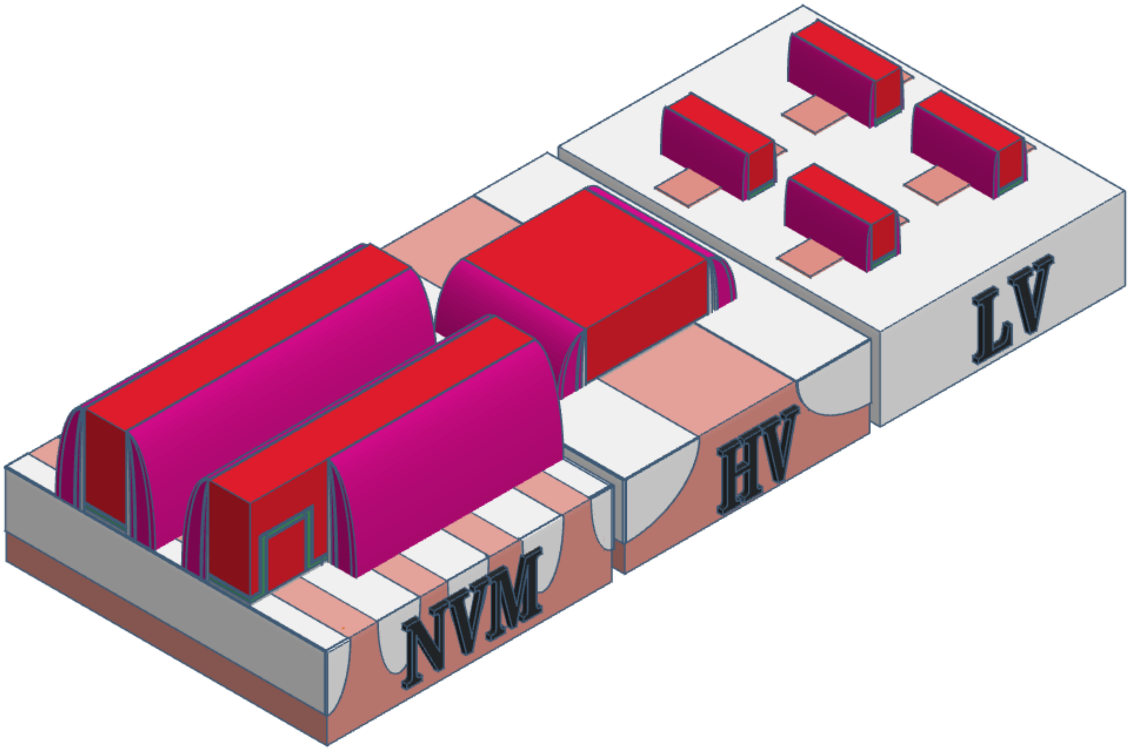


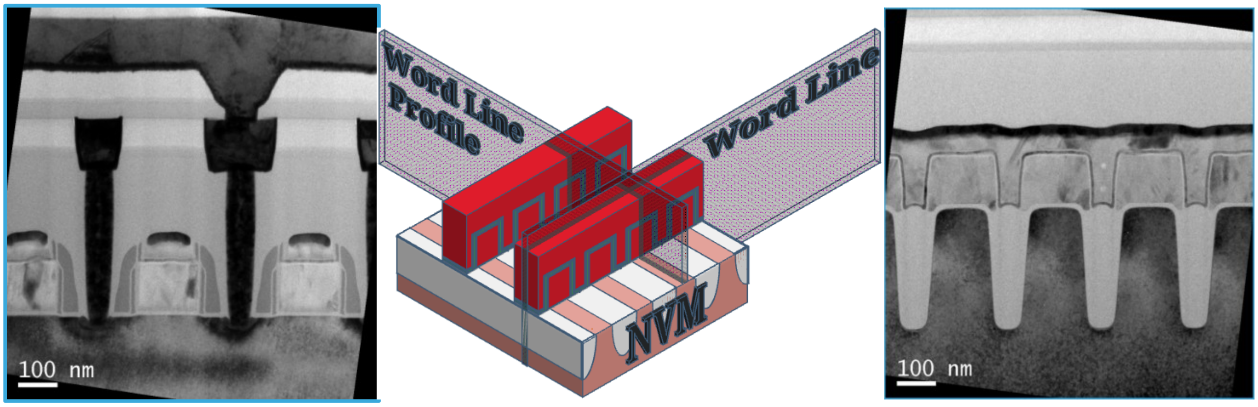
Figure 67: NVM, HV and LV regions post gate etch and final spacer nm spacer (note: the spacers fully surround the MOS devices and NVM word line, the cut-away is to show the spacer structure)

3.2.10 Salicide and Contacts

After the final spacer, the poly-Si is salicided, where appropriate, to improve the contact resistance. The contacts are formed by depositing a thick layer of SiO_2 over the wafer which is then planarized in CMP. Contact holes in this oxide layer are etched and then filled with TiN/W to form the contacts. Excess W is removed in an additional CMP step before the next dielectric layer is deposited and patterned to form the first Cu metal layer. This is repeated for the rest of the metallization layers.

3.3 Alumina-based embedded flash cells at the 40 nm node

The cross-sections of the first OAO flash cells are shown up to the first metal layer in Figure 68, the difference in Poly1 and Poly2 widths is visible in (a).

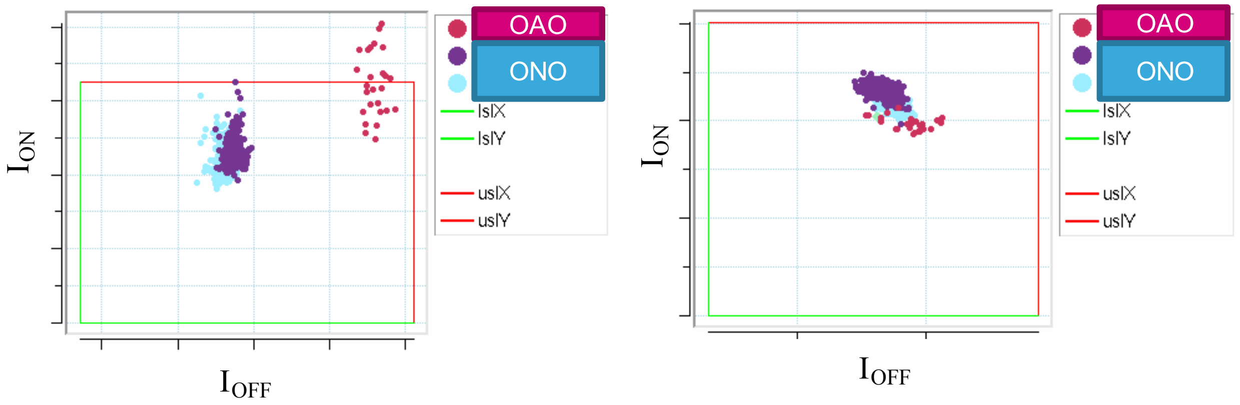


(a) (b) (c)

Figure 68: Final TEM images of functional 40 nm embedded flash cells with alumina in the IGD (a) in the direction of the word line and (c) in the word line as described in the (b)

3.3.1 Flash and HV device validation

The I_{ON}/I_{OFF} currents for the HV devices are shown below in Figure 69. Both the NMOS and PMOS devices are functional, with the latter being well-centered. The NMOS devices have slightly higher I_{ON} values, however they can easily be re-centered by adjusting the ion implantation step.



(a) (b)

Figure 69: I_{ON}/I_{OFF} for the (a) NMOS and (b) PMOS HV devices showing three wafers having an OAO IGD and two other batches with the standard ONO for comparison

The flash cells of the three remaining wafers from the batch dedicated to debugging the process flow were functional, shown in Figure 70, representing the first time that alumina has been integrated into the IGD of a 40 nm embedded flash product. In testing these wafers, it was discovered that the 4 Mb test chip did not work due to short circuits between the inputs and outputs. This necessitated the addition of the LV logic gate etch step which also defines the input and output resistances.

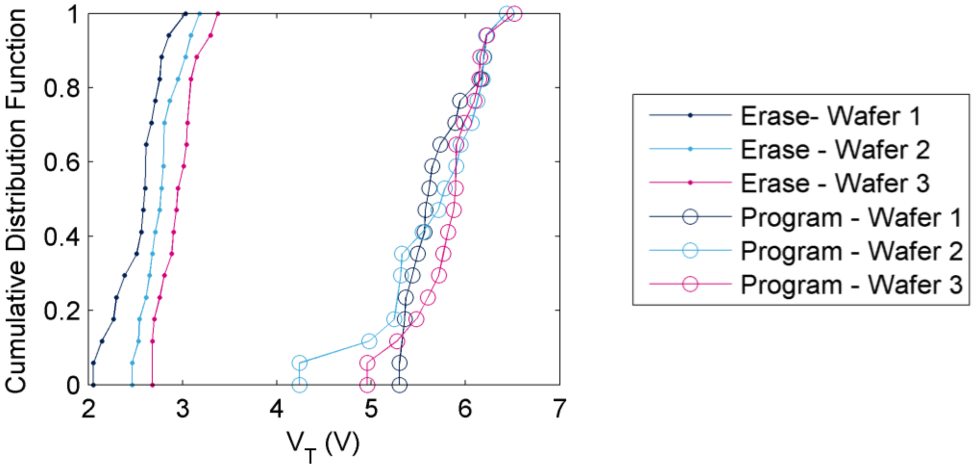


Figure 70: CDF for the programmed (V_T-high) and erased (V_T-low) states of the first three 40 nm embedded flash devices having an OAO structure

The gate etch outlined in Section 3.2.9 was added to the remaining batches in order to render the 4 Mb matrix testable. This additional step further increased the processing time in the cleanroom and the final batches passed the first stages of parametric testing on May 28th, 2017. We hope that the 4 Mb matrices can be tested in the summer of 2017.

3.3.2 Leakage and breakdown characteristics of alumina based IGDs

In parametric testing the 10,000 µm² plate of Poly1/IGD/Poly2 is tested to determine the EOT, leakage, and breakdown properties of the IGDs. The first leakage characteristics are shown in Figure 71, where the y-axis represents the bias applied to Poly2 that will induce à 0.1 mA/cm² through the IGD. Each point represents a single device as multiple devices per wafer were tested. In both directions, the ONO is the highest performing IGD. In the negative direction (a), the OA samples seem to align along a trend that is more favourable than the OAO samples; an extrapolation of both populations to an EOT of 10 nm would lead to a larger negative bias being required for the OA samples to conduct the same amount of current as an OAO sample. This corresponds to electron injection from the control gate (Poly2) towards the floating gate (Poly1). In the positive direction (b) all the OAO and OA samples seem to follow the same general trend.

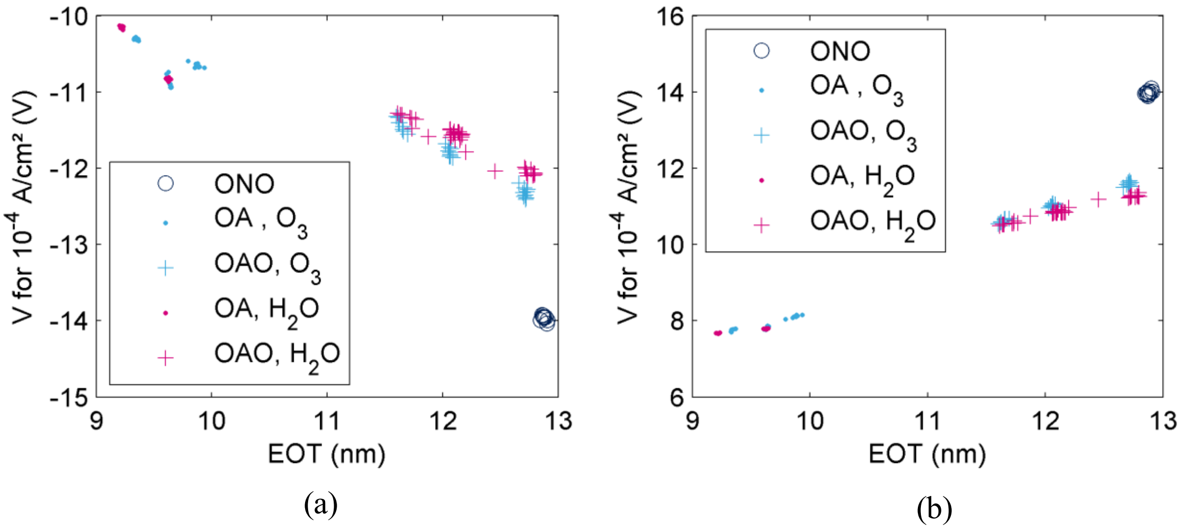
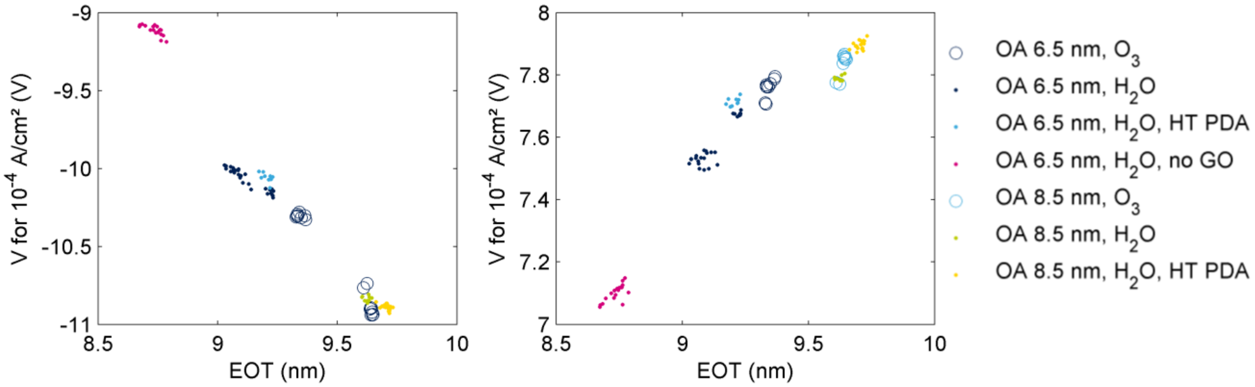


Figure 71: The Poly2 bias on a large area capacitor required for a 0.1 mA/cm² leakage current as a function of EOT in the (a) negative and (b) positive direction (bias applied to Poly2) for the standard ONO and OAO/OA IGDs with either ozone or water as the oxygen precursor

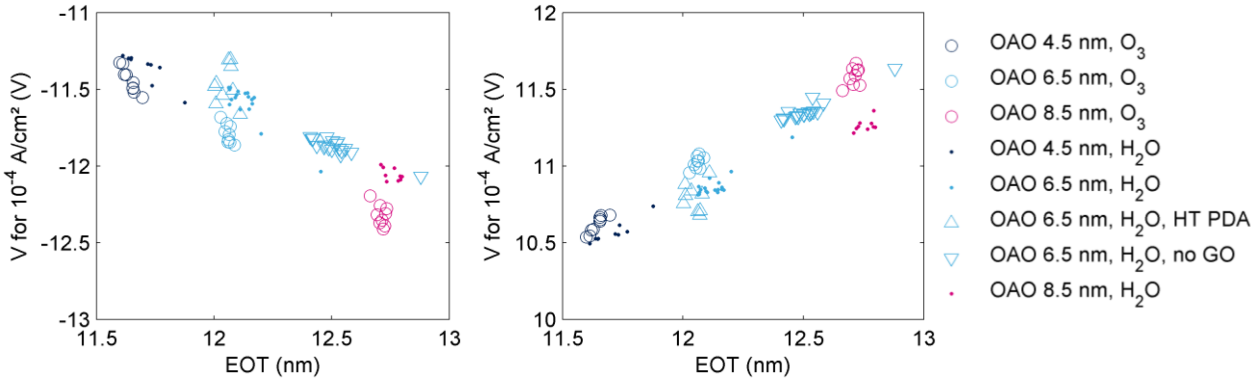
If we look more closely at the OA populations in Figure 72 we can see that the differences related to the oxygen precursor (H_2O vs O_3) during ALD growth are small. In the negative direction there is no impact, and in the positive direction there is a slight improvement (increase) in the voltage required for 0.1 mA/cm² leakage currents for the 8.5 nm thick alumina sample. An increase in the thermal budget via a higher temperature post deposition anneal, “HT PDA” increases the leakage currents in the negative direction while decreasing them in the positive direction. The omission of the gate oxide growth steps , “no GO”, has a significant impact on the IGD, mostly in the fact that the EOT is approximately 0.5 nm thinner; the “no GO” samples remain in the same bias/EOT trend as the other OA samples. For thick OA samples, the addition of the HT PDA reduces leakage in the positive direction, while not degrading in the negative direction. We can conclude the HT PDA is beneficial to OA samples.



(a) (b)

Figure 72: The Poly2 bias on a large area capacitor required for a 0.1 mA/cm² leakage current as a function of EOT in the (a) negative and (b) positive direction (bias applied to Poly2) for OA samples with either ozone or water as the oxygen precursor and different thermal treatments

The differences between the precursors are more readily observed for the OAO samples in Figure 73; the use of ozone consistently increases the magnitude of the bias required to reach the same leakage current in a water-precursor-based sample. The HT PDA treatment increases dispersion in the values but does not seem to significantly impact the leakage currents in either direction. Unlike the OA samples, when a top oxide is included, the omission of the gate oxide grown leads to an increase in EOT by about 0.5 nm.



(a) (b)

Figure 73: The Poly2 bias on a large area capacitor required for a 0.1 mA/cm² leakage current as a function of EOT in the (a) negative and (b) positive direction (bias applied to Poly2) for OAO samples with either ozone or water as the O-precursor and different thermal treatments

The breakdown voltages of the IGDs are shown for the negative direction in Figure 74, they are not shown in the positive direction as they all fall on the same trend line. For the negative direction the use of the ozone precursor improves the resistance to breakdown for OAO

IGDs, with no impact on the OA populations. The division of some IGDs into two populations implies that there is an integration problem somewhere that introduces extrinsic degradation. Although the resistance to breakdown is improved by not including the gate oxide growth, these gate oxidations are necessary in the fabrication of an embedded flash product. The HT PDA does not degrade the IGDs in terms of breakdown.

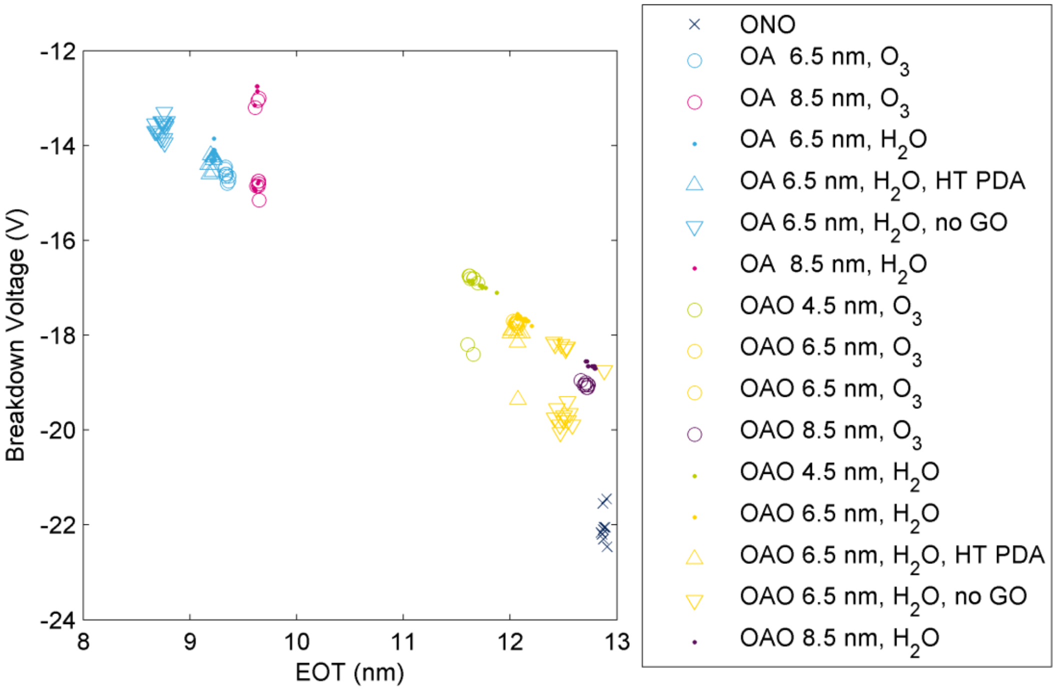


Figure 74: Negative (bias applied to Poly2) breakdown voltage vs EOT for a series of OAO/OA IGDs and the standard ONO

Based on the batch of alumina-based IGDs, the alumina that is grown using ozone as the oxygen precursor is intrinsically better than that grown using water as a precursor. The HT PDA is beneficial for the alumina in terms of leakage and breakdown.

3.3.3 Alumina-based IGD Flash Cells

Each of the IGDs yielded functional flash cells, with some examples shown in Figure 75. The reduction in V_T in the programmed state is understandable due to the altered NLDD implantations for the OAO and OA samples with respect to the standard ONO. As the NLDD implant was completed after the spacer for the OAO and OA samples, there are fewer n-type dopants added to the p-type canal, leading to a higher threshold voltage. The program and erase pulses, being the same for all devices, lead to a V_G-V_T during programming is lower, lower drain current and a less efficient program operation for the OAO and OA samples. The erase state is not strongly correlated with the IGD, an unexpected result as a thinner EOT for the IGD increases the control gate-floating gate coupling, α_G, and the efficiency the erase operation.

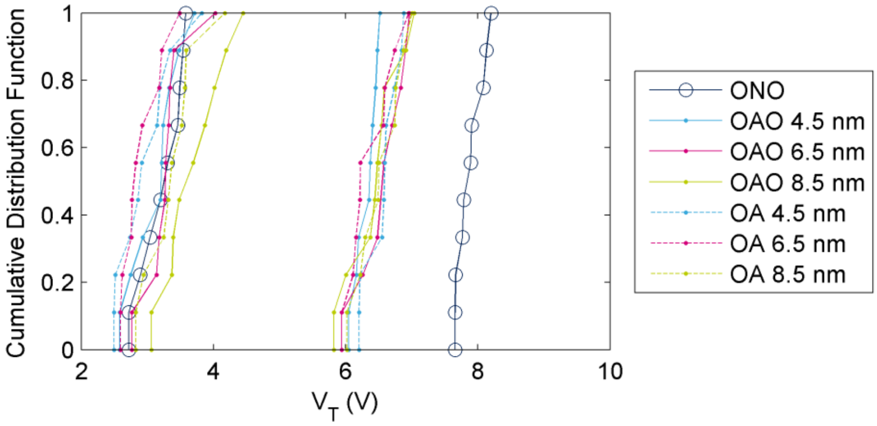


Figure 75: CDF for the programmed (V_T-high) and erased (V_T-low) states of the several OAO and OA IGDs along with the standard ONO showing little improvement in erase performance for lower EOTs

In order to see investigate wether or not the coupling coefficient is affected by the change in IGD, we plot an extracted α_G vs the EOT of each IGD in Figure 76. The coefficient does not follow the expected increase at lower EOTs. We believe that that this is due to the word line profile after the Poly2NVM etch; decreases in the width of the word line decrease the area of the IGD capacitor between the floating gate and control gate. For example, in Figure 61 the top of the Poly2 layer is 123 nm wide, nearly 25% lower than the width of the Poly1. This loss of coupling via area decrease can offset any gains in coupling that were made by thinner IGDs. Therefore, the Poly2NVM etch must be improved before gains in coupling can be made using alumina-based IGDs.

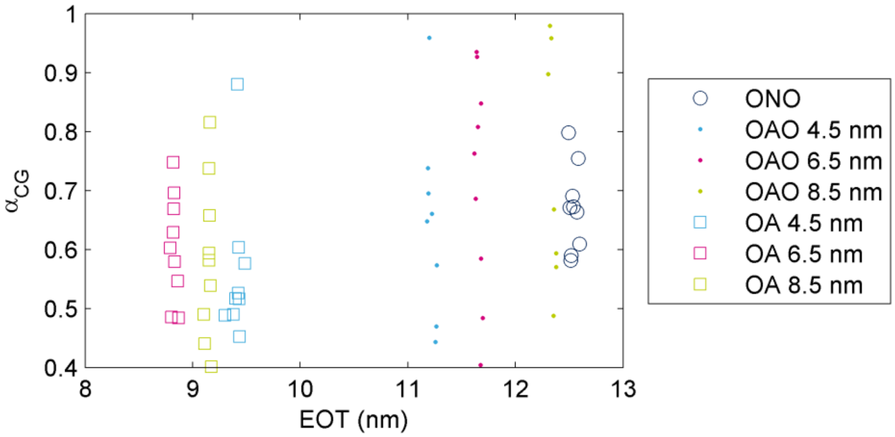


Figure 76: α_G vs EOT for a range of OAO and OA samples having alumina layers from 4.5 to 8.5 nm

3.4 Conclusion

This chapter has explained a basic overview of the production flow that can be used to fabricate embedded flash products at the 40 nm node. The modular process flow fabricates the flash cells and CMOS logic devices in such a way that the former does not have significant impact on the latter. Although we chose an integration scheme, outlined in Section 3.2, that did not form functional CMOS logic devices, the results indicate that completely functional embedded flash products are feasible. The modifications were not detrimental to the functionality of the HV devices.

When the ONO is replaced by an OAO or even an OA the first processing step after IGD deposition would be the ONO patterning step where it is removed from the LV CMOS regions. As described in Section 3.2 this step was not completed for the first lots, however given the success of the LAE etch developed at the CEA-Leti this patterning step should be easily reproduced. The next step is the growth of the gate oxides which simply act as additional annealing steps for the IGD and poses no problem. The results of the Poly2HV etch in Section 3.2.6 showed that the approach of a dry etch plus a wet etch was highly effective in opening the ONO plate capacitors and completely removing the Al_2O_3 from the HV MOS devices.

The Poly2NVM etch that defines the word lines of the flash matrices proved to be the most complicated processing step. Although the dry and wet etch were able to define the word lines, the non-vertical profile is not ideal. With more wafers and time to dedicate to this etching process it would surely be possible to improve the side-wall profile for future batches. The worries of “fences” containing alumina between the word lines causing problems in later processing steps proved to be unwarranted due to their complete removal by the LAE etch. In fact, the LAE etch of OAO was more effective than the standard ONO etch in clearing the space between the word lines. Without exposed Al_2O_3 after the spacer depositions the wafers should be able to undergo the standard LV CMOS processing without worrying about alumina contamination of the equipment.

We have shown that the use of OAO or OA to replace the ONO in the 40 nm embedded flash process at STMicroelectronics is feasible from a process integration standpoint. Both OAO and OA IGDs are of good quality in terms of leakage currents and breakdown voltages. These results, obtained with electrical testing performed at rather higher electric fields, are promising; they will have to be completed with characterization at much lower fields and with longer times to more deeply evaluate the alumina based IGDs. Future characterization (data retention bakes, OSS measurements) should give a complete overview of the potential of these layers. The top oxide does not appear to impact leakage from the floating gate to the control gate and in the opposite direction it has a negative impact; OA IGDs show a more promising trend. The alumina grown using ozone as the oxygen precursor outperforms the alumina grown using water as the precursor and the shorter, higher temperature post deposition anneal further improves the alumina leakage properties.

Functional flash devices can be made with the variety of OA and OAO IGDs, however they do not present an improvement in erase performance with respect to the standard ONO. We believe that the degraded word line profile after the Poly2NVM etching step leads to a decrease in IGD area that reduces control gate-floating gate coupling. This reduction in coupling therefore offsets any gains in coupling made by using IGDs with lower EOTs, eliminating the expected erase performance gains. With more etch development resources, the profile could be improved and the increases in coupling could be regained.

3.5 Bibliography

[Cas02] J.D. Casperson, et al., “Materials issues for layered tunnel barrier structures” *Journal of Applied Physics 92*, 2002

[Plu00] J.D. Plummer, et al., “Chapter 4: Semiconductor Manufacturing – Clean Rooms, Wafer Cleaning and Gettering” in *Silicon VLSI Technology: Fundamentals, Practice and Modelling, 1st ed.* Upper Saddle River, NJ, Prentice Hall, 2000

[Pes01] D. Peschiaroli, A. Maurelli, E. Palumbo, F., Piazza, “Method of manufacturing an electrically programmable, non-volatile memory and high-performance logic circuitry in the same semiconductor chip” US Patent 6482698 B2, Nov. 19, 2002

[Pow07] J.R. Power, et al., "Improved Reliability of a High-k IPD Flash Cell through use of a Top-oxide" in *IEEE Non-Volatile Semiconductor Memory Workshop*, Monterey, CA, USA, Aug. 2007

[Pow08] J.R. Power, et al., "Improved Retention for a Al2O3 IPD Embedded Flash Cell without Top-Oxide" in *IEEE Non-Volatile Semiconductor Memory Workshop*, Opio, France, May, 2008

[Pia10] F. Piazza, et al., “High Performance Flash Memory for 65 nm Embedded Automotive Application” in *IEEE International Memory Workshop*, Seoul, Korea, May, 2010

[Shu09] D. Shum, et al., “ALD-Al_2O_3 as an Inter-Poly Dielectric for a Product Demonstrator in a proven eFlash Technology” in *IEEE International Memory Workshop*, Monterey, CA, USA, May, 2009

Chapter 4: Embedded Flash Potential in 28 nm Node

Table of contents

4.1 Introduction

The processes changes outlined in Chapter 3 were dedicated to the integration of a high-k material into a 40 nm embedded flash product. The next question that one must ask is “what happens at 28 nm”? Until this point, CMOS logic used silicon dioxide gate oxides and poly-silicon control gates but at the 28 nm node high-k metal gate stacks are used. Just as the 40 nm embedded flash cell involved an IGD that was exposed to the LV gate oxide and capped-off by the LV poly-Si gate, the 28 nm node embedded flash cell could utilise the same approach. A reminder of the 40 nm stacks and the potential 28 nm stacks are shown below in Figure 77.

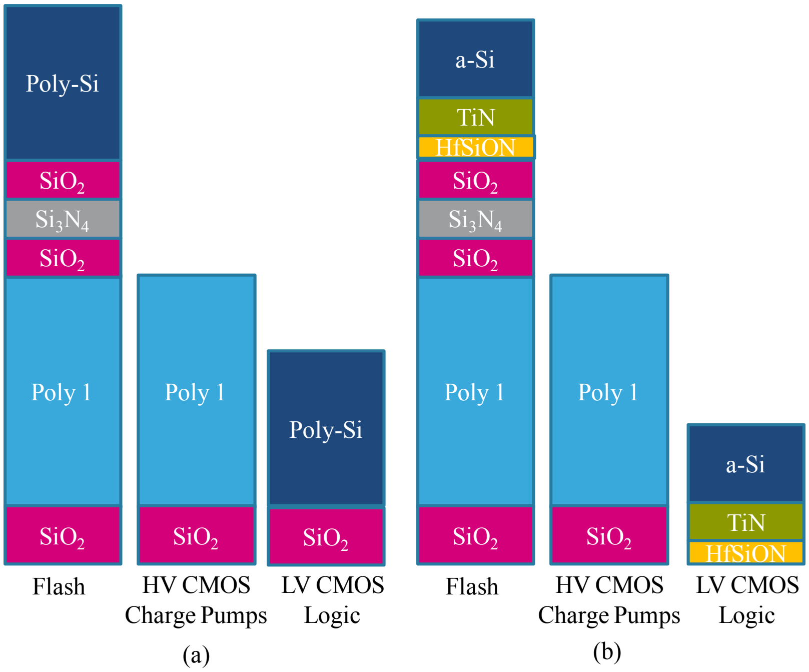


Figure 77: Flash, HV CMOS and LV CMOS stacks for (at) 40 nm and (b) 28 nm processes

This chapter describes some of the potential IGD stacks for a 28 nm embedded flash product. These are tested on a simplified IGD capacitor, described in Section 4.2, in order to avoid to process complexities described in Chapter 3. The first series of samples are described in Section 4.3 and seeks to answer the question of whether or not the HKMG stack from the logic process would degrade the ONO. Section 4.4 follows up on the incremental change in order completely replace the ONO by a high-k stack. This set of samples tests whether or not a fully high-k IGD can go futher and improve performance relative to an ONO-based solution. The question of whether or not we can integrate a high-k IGD into a 40 nm embedded flash product is answered in Section 4.5.

4.2 Capacitor-based IGD test vehicle

Due to the process complexity, as described in Chapter 3, an embedded Flash product is a very inefficient way to do basic screening of IGD stacks. This is mainly related to the difficulties associated with the development of the etching processes required to fabricate the cells. In place of a functioning flash matrix, there is a motivation to find other test vehicles that can provide the basic information required to screen IGDs before attempting their integration in a more complete product. One such test vehicle is the inlaid or damascene capacitor, which uses a planarization step to define the devices, eliminating the need for etching step development. The "MOCA" lithography mask for 300 mm wafers from CEA-Leti is used to define the cavities in a 550 nm SiO_2 layer, giving the type of sample the moniker "MOCA wafers". Figure 78 shows an outline of the process flow from bare Si wafers to W planarization, which are more deeply explained in Section 4.2.1. An n-type substrate is used because in a flash cell the floating gate is n-doped.

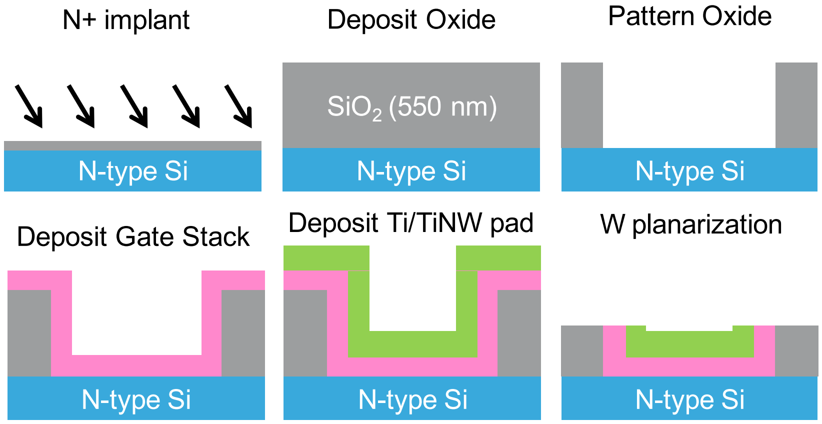


Figure 78: MOCA process flow for damascene capacitors

Electrical characterization of the MOCA wafers are done via two contacts: the capacitor gate via a prober's point and the backside of the wafer via the chuck of the tester. The MOSCAP structure is characterized by capacitance measurements to obtain the EOT and to help understand charge trapping in the stack. Current measurements are also used to characterize the quality of the stacks in terms of leakage currents and breakdown. Metal-Insulator-Metal, or "MIM" style capacitors would have been better for leakage current and breakdown characterizations as the potential drop and charge in the silicon substrate do not need to be taken into account. However, this type of sample was not possible to fabricate using the simplified MOCA maskset.

4.2.1 MOCA Capacitor Process Flow

The process flow for the MOCA wafers is outlined in Figure 79. The steps in blue take place in the 300 mm profuction fab of STMicroelectronics while the steps in purple take place in the cleanroom at the CEA-Leti. The wafers were provided by the CEA-Leti however the first processing step occurs at STMicroelectronics as the CEA-Leti cleanroom does not have the equipment for the LPCVD deposition of SiO_2 onto 300 mm wafers. The patterning is then completed at the CEA-Leti because the maskset belongs to the CEA-Leti. The wafers then return to STMicroelectronics for the deposition of the gate stacks. As the wafers were to include layers and treatments that were already in regular use in the cleanroom they used parts of two different production flows:

- 40 nm embedded flash products already have the well anneal, ONO deposition and ONO densification anneal recipes, so for these steps the wafers were considered to be a flash product.
- 28 nm logic devices contain the gate-stack of interest so for these steps the wafers were considered to be a CMOS product

Before switching from the 40 nm flash process flow towards the 28 nm logic flow, the ONO was removed from the backsides of the wafer to allow for electrical contact via the bulk silicon.

After the completion of the stacks, the standard TiN/Ti/W plug is deposited before the wafers are sent to the CEA-Leti for the final CMP and then electrical testing.

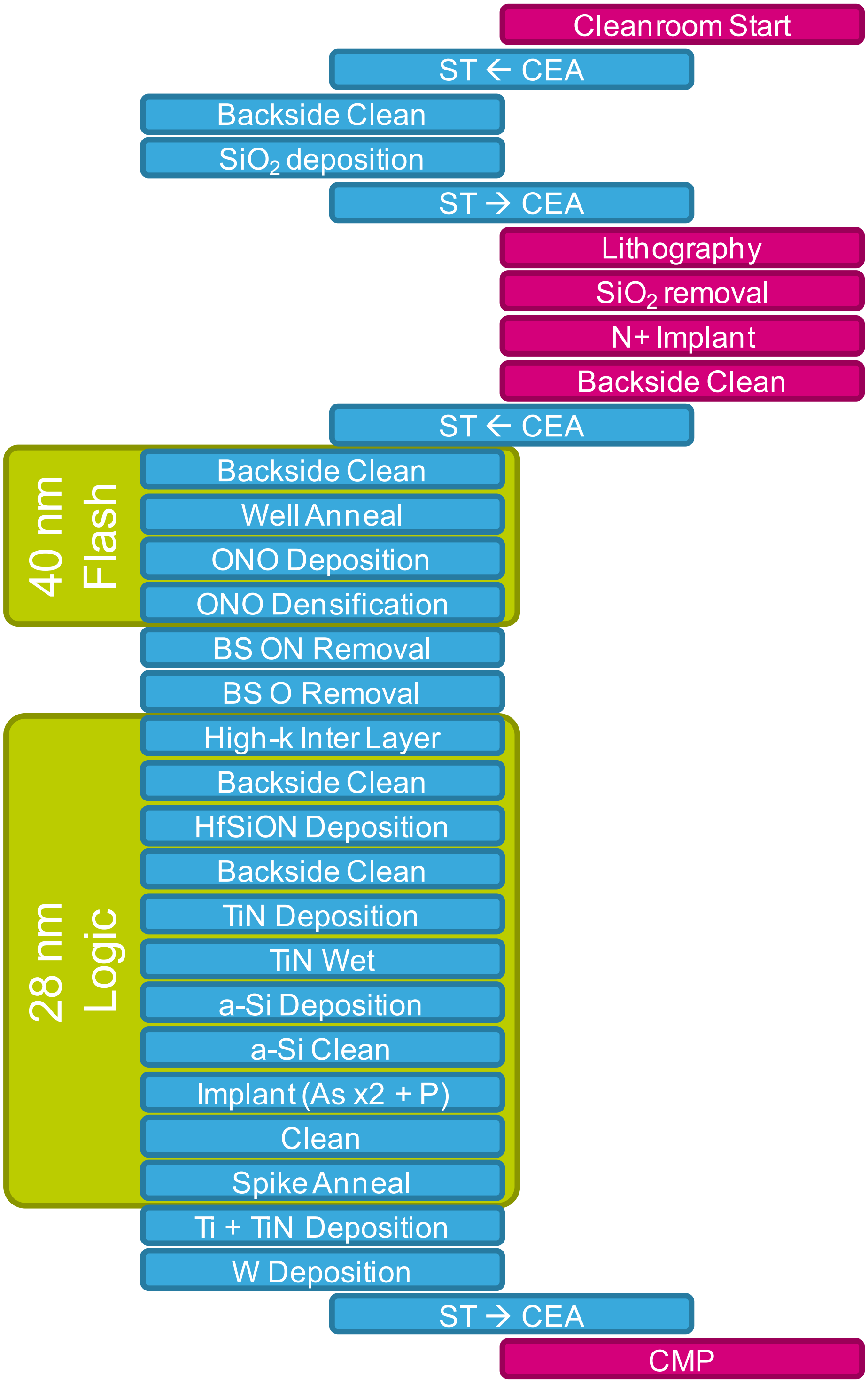


Figure 79: Process flow for MOCA wafers between the CEA (purple) and STMicroelectronics (blue) cleanrooms using processing steps from the 40 nm flash and 28 m logic processes

4.3 ONO and High-k Metal-Gate Logic compatibility evaluation

In the evolution towards embedded Flash comprising high-k metal gate based logic devices, a logical step is to continue using a standard ONO as the IGD. One might call this the "if it isn't broken, don't fix it" approach. This provides the context for the testing of the compatibility of the ONO with the high-k metal gate stack. The gate stack itself is 1.8 nm of HfSiON, deposited by MOCVD, 5.5 nm of TiN, deposited by PVD and 43 nm of a-Si, deposited by CVD, therefore it is interesting to test the impact of each of the components. Figure 80 shows six of the stacks that were tested, including two where a SiO_2 layer is used in place of an ONO in order to eliminate the effect of charge trapping in the nitride.

This work was presented at the Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon in Athens, Greece in April 2017 [Dob17].

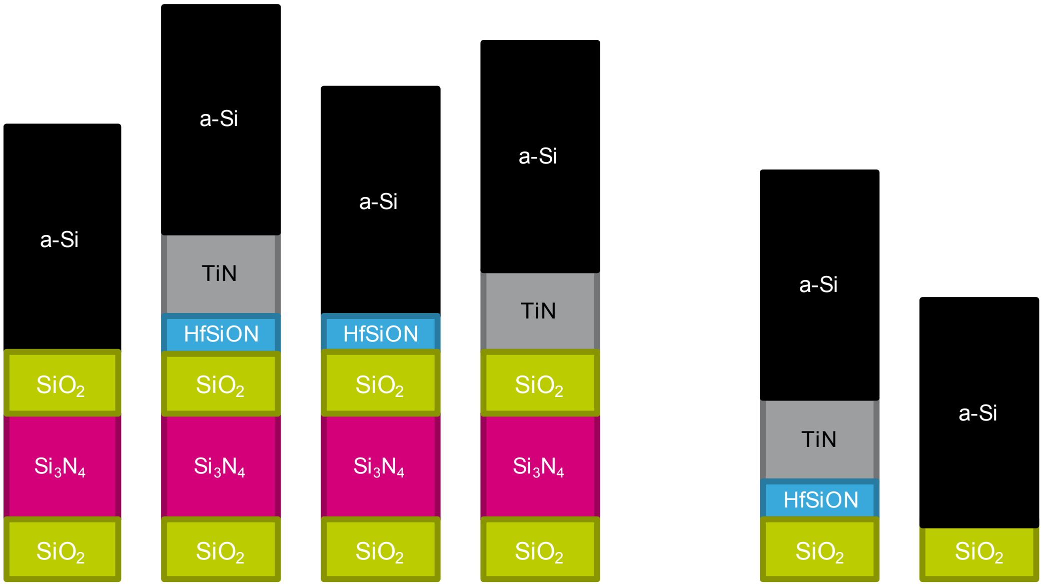
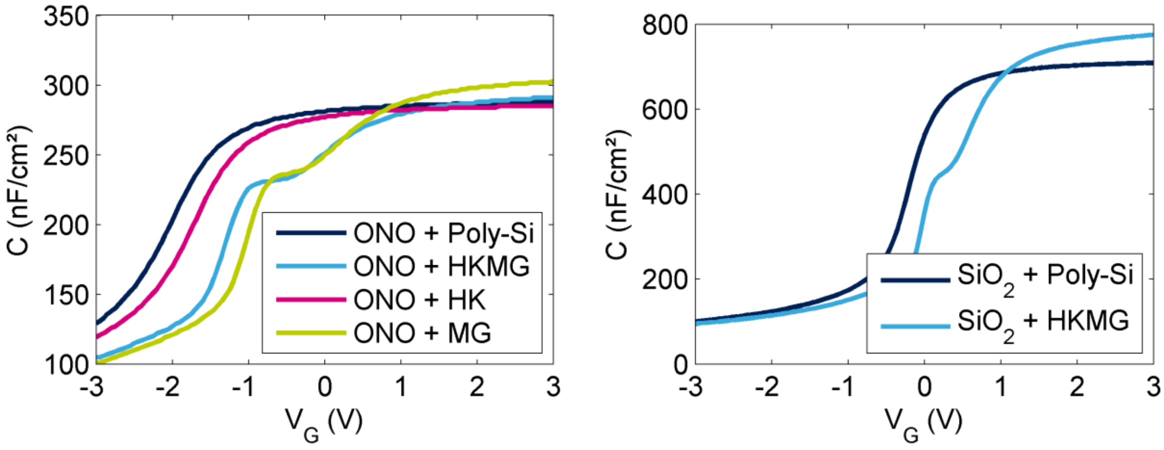


Figure 80: ONO and SiO_2 based MOCA capacitor samples for HKMG screening

4.3.1 ONO and HKMG Characterization

The results from the capacitance voltage characterization of 20 devices per wafer at 100 kHz are shown in Figure 81 and their characteristics are summarized in Table 8. The most noticeable features are that the EOT decreases, the flat-band (V_{FB}) increases and a shoulder peak appears when the poly-Si is replaced by the HKMG. These occur whether or not the dielectric is the ONO, or simply a SiO_2 layer



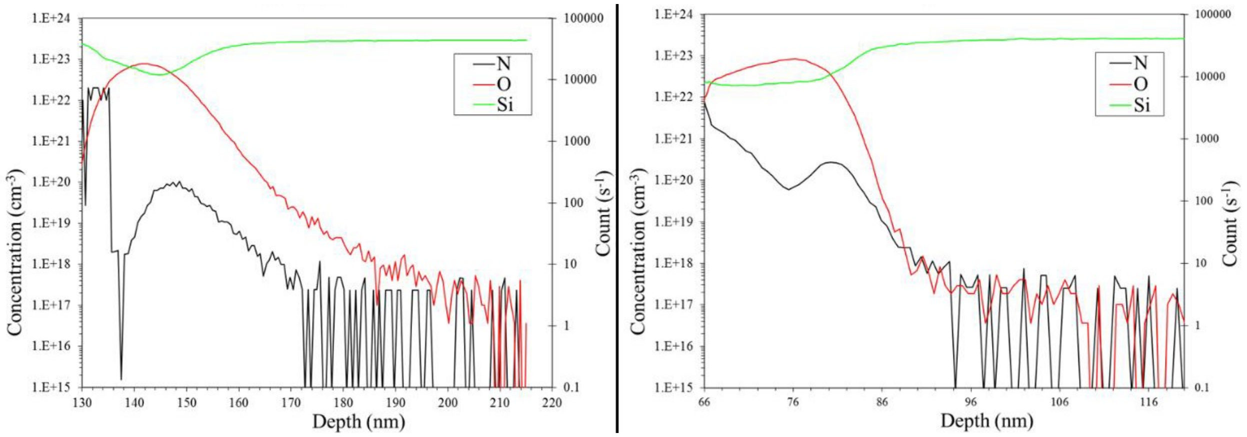
(a) (b)

Figure 81: CV curves at 100 kHz for (a) ONO and (b) SiO_2 based samples, median values for 20 devices

Table 8: Summary of EOT and V_{FB} characteristics' for ONO and HKMG capacitors

Name	Si	HKMG	HK	MG	SiO_2/Si	SiO_2/HKMG
Dielectric	ONO	ONO	ONO	ONO	SiO_2	SiO_2
1.8 nm HfSiON	-	Yes	Yes	-	-	Yes
TiN	-	Yes	-	Yes	-	Yes
a-Si	Yes	Yes	Yes	Yes	Yes	Yes
EOT (nm)	11.9	11.5	12.0	11.2	4.7	4.2
V_{FB} (V)	-1.5	0.1	-1.2	0.0	-0.1	0.7

The decrease in EOT from 11.9 to 11.5 nm when switching from a Poly-Si to a HKMG is counterintuitive because the thin dielectric layer is added so the EOT would be expected to increase. This is the case with the HK sample, which shows a slightly higher EOT of 12.0 nm. However, the EOT is impacted by both the physical thickness and the dielectric constant of the stack for a fixed area. Nitrided SiO_2, or SiON is known to have a higher k-value than pure SiO_2 [Fis04]. Therefore the decrease in EOT for the MG sample is attributed to an increase in the average k-value of the stack caused by the diffusion of nitrogen from the TiN. In the HKMG sample the net effect is a decrease in EOT. This theory is supported by SIMS analysis that shows in Figure 82 the increased nitrogen concentration when the HKMG stack is used [Mor17].



(a) (b)

Figure 82: SIMS of N, O and Si for (a) poly-Si and (b) HKMG stack [Mor17]

The second notable feature, the shoulder peak, is also present only when the metal gate is used. The prominence of the shoulder is frequency dependent and CV analysis determines that it is related to traps at 0.65 and 0.97 eV below the Si conduction band [Mor17]. This trap could potentially impact the V_T of a flash cell by charging/discharging during cell operation, similar to the V_T-instabilities discussed in Section 1.4.2. Although in a working flash cell, the bottom oxide of the ONO is in contact with the highly doped poly-Si floating gate, not bulk silicon, so the trap level may not be probed during operation.

The change in V_{FB} can be related to the difference in the work function of the gates and/or the amount of charge trapped in the dielectrics. The flat-band and threshold voltage are much less of an issue for IGD applications than they were during the development of HKMG stacks for CMOS logic. For CMOS applications, one of the problems with high-k materials was that their defects, via Fermi-level pinning, negatively impacted the ability to control the transistors' threshold voltages [Rob15]. In a flash cell, the IGD is in contact with highly doped poly-Si floating gate that is more than 100 nm away from the channel, so there is no Fermi-level pinning.

The current voltage characteristics of the ONO samples up to breakdown around 20 V are shown in Figure 83. The medium and high-field currents are very similar so the conduction across the ONO is considered to be the limiting factor. The voltage drops of interest across the ONO for data retention properties are on the order of 2-3 V so these measurements are insufficient for comparing the samples as described in Chapter 2.

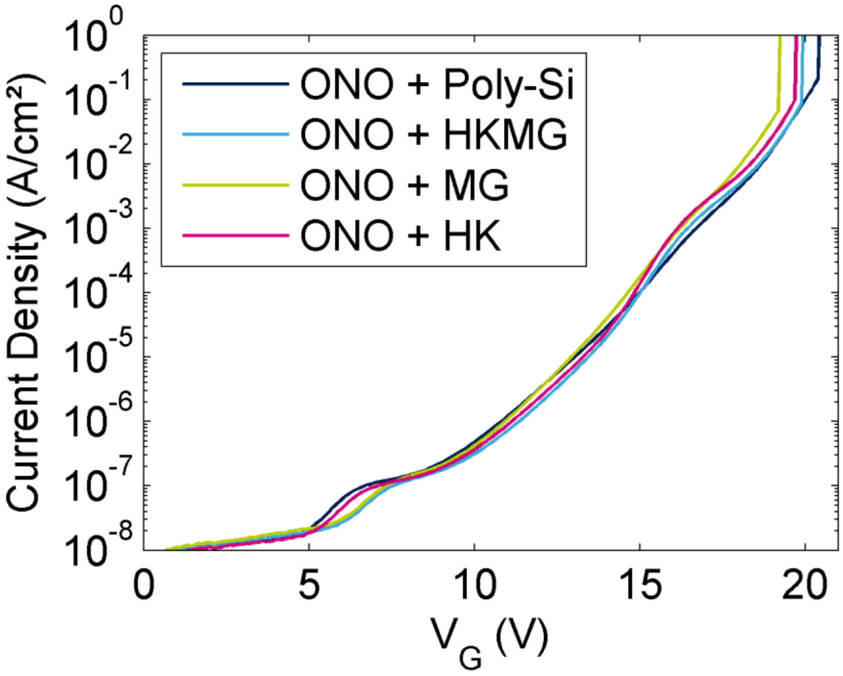


Figure 83: IV characteristics at 25°C, median values for 20 devices

Higher resolution measurements using an Agilent B1500 Semiconductor Parameter analyser were completed to measure the current-voltage properties between gate voltages of 0 and +7 V and are shown in Figure 84. The large hysteresis in the both stacks is immediately apparent, as the currents during the return trace from +7 to 0 V are orders of magnitudes lower than the currents during the initial trace from 0 to +7 V. This is explained by charge trapping in the nitride layer of the ONO; some of the electrons injected across the bottom oxide are trapped in the nitride during the increasing sweep. For the return sweep, as the bias decreases, the electrons trapped in the nitride remain trapped, so the current does not follow the same trace because the electric field in the ONO is modified due to the trapped charges. In addition, the trapped charges cause a shift of the flat-band voltage in the positive direction when the CV curve after the IV sweep is compared with the virgin CV curve in Figure 84. This phenomenon of a V_{FB} (and therefore V_T) that is dependent on the charge density in the ONO is the basic idea behind SONOS-style charge trap memories.

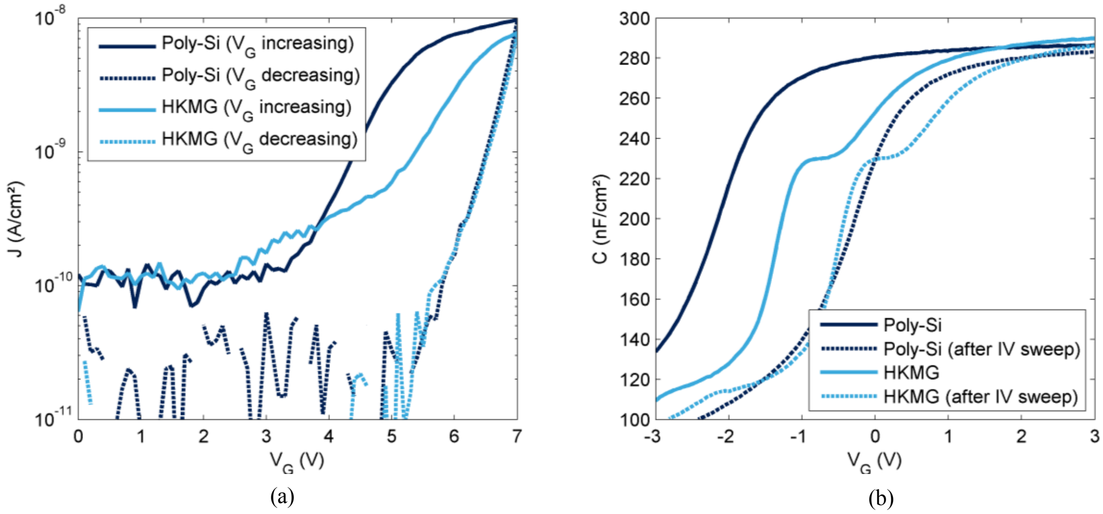


Figure 84: (a) IV measurements from 0 to +7 V and back to 0 V and (b) CV measurements before and after the IV measurement

The trapped charge in the ONO's effect on the capacitor can be modeled by solving Gauss' law in the ONO. Appendix A describes the model used to calculate the ONO band structure when charges (interfacial or bulk) are taken into account. The change in band structure due to electrons trapped in nitride of the ONO under a gate bias of +7 V is shown for charge densities up to 2x10^{13}/cm² in Figure 85. As the number of electrons trapped in the nitride increase, the conduction and valence bands in the nitride bend upwards, resulting in an enhancement of the electric field in the top oxide and a reduction in the bottom oxide. As any electron traversing the ONO needs to be supplied by injection from the silicon substrate, the decreasing electric field in the bottom oxide drastically reduces the current because the injection at the Si/SiO_2 barrier is highly dependent on the electric field [Pav97].

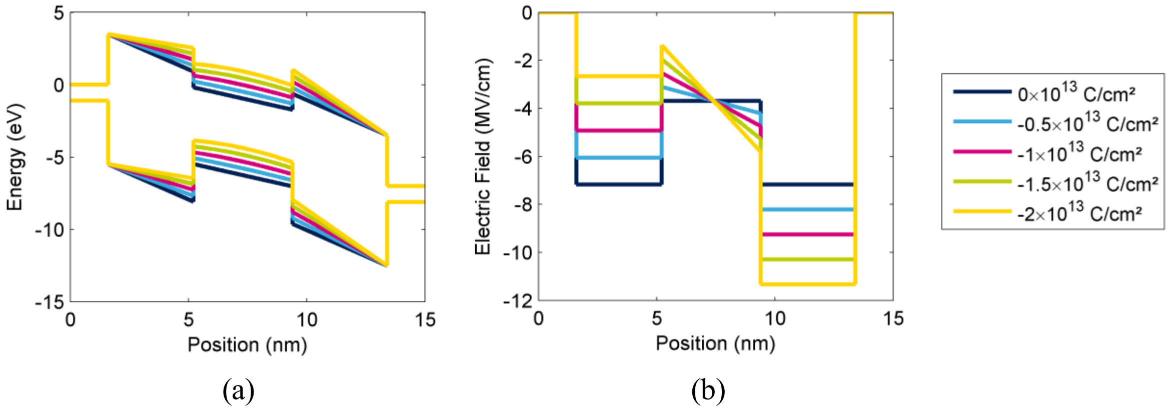


Figure 85: Modeled effect of trapped electrons in the silicon nitride of the ONO on the (a) energy band structure and (b) electric fields for silicon nitride charge densities of 0,5,10,15,20 x10^{12}/cm² (V_G = 7 V)

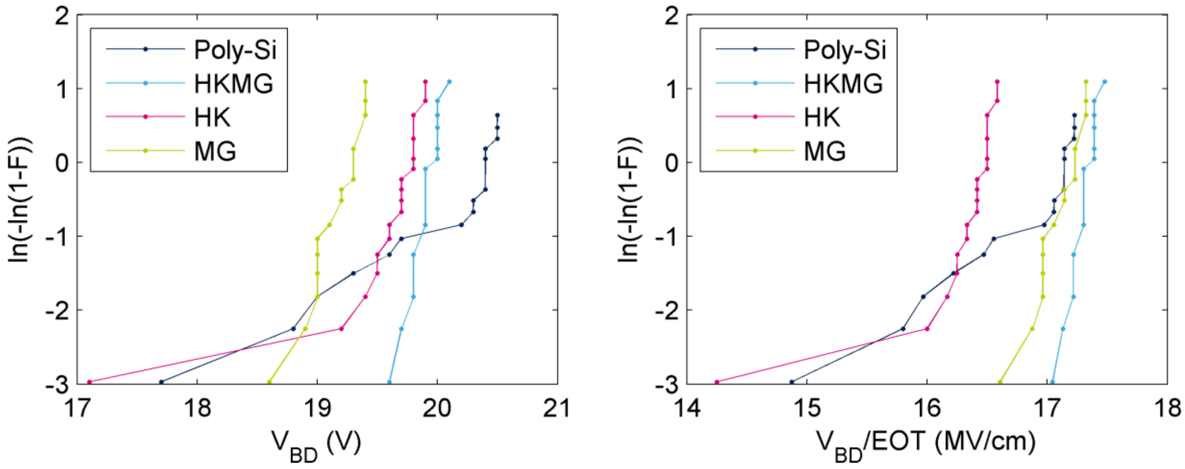
For an n-type MOSCAP in accumulation the electric field in the oxide is considered to be given by Equation 4.1 [Hu10].

$$E_{OX} = \frac{V_G - V_{FB}}{EOT} \qquad Equation\ 4.1$$

This illustrates the difficulty associated with using MOSCAP devices such as these (n-type substrate measured in accumulation) for current voltage and breakdown analysis. As a gate bias is applied, the current tends to decrease due to both the trapped charges' influence on the ONO band structure and the reduction in the average electric field due to the changing V_{FB} which increases with the number of electrons trapped in the ONO. In a MIM-style device, there is no bulk semiconductor flat band to take into account and the decreases in current can be attributed to trapped charges [Moo14].

4.3.2 ONO and HKMG Breakdown Characterization

The IV curves shown in Figure 83 are the median values for 20 devices, so one other way to compare the four samples is to look at the distribution of breakdown voltages, shown below in Figure 86(a). Weibull statistics are often useful in characterizing oxide reliability, so *ln(−ln(1−F(t))* is plotted instead of *F(t)* itself, which is the failure rate at time, *t* [Ghe14].The breakdown voltages were also normalized by the EOT in Figure 86(b) in order to use a metric more closely related to the average electric field in the ONO. The switch from poly-Si to HKMG does not show any degradation and in fact reduces the dispersion in V_{BD}.



(a) (b)

Figure 86: Weibull plots for ONO-based devices (a) without and (b) with normalization by the EOT

Time dependent device breakdown (TDDB) characterization was carried out on 20 devices at 125°C at two different gate biases (16 and 17 V) with the results shown in Figure 87. In general, the high-k layer increases the time to breakdown, while the metal gate decreases it. The net effect of the two is a slight increase in time to breakdown.

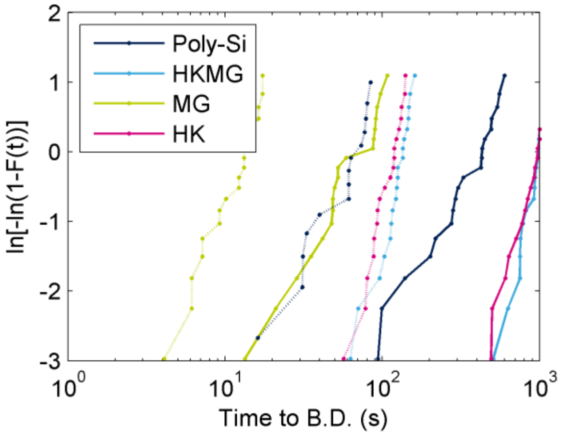


Figure 87: Weibull plot for TDDB measurements of ONO samples at 125°C under +16 V (thick lines) and +17 V (thin lines) gate biases

Using the time to breakdown for 63% of the devices (W = 0 in a Weibull plot), the time to breakdown can be plotted against the V_G/EOT in Figure 88 in order to better compare the different samples. Adding the HfSiON layer to the ONO increases the lifetime of the devices while the switch to a TiN gate on the ONO does not appear to have a significant impact. A combination of the two phenomena, in the switch to HKMG from poly-Si control gates results in an overall increase in the time to breakdown. This is a very promising result as it shows that depositing the HKMG stack directly onto the ONO does not induce any intrinsic degradation.

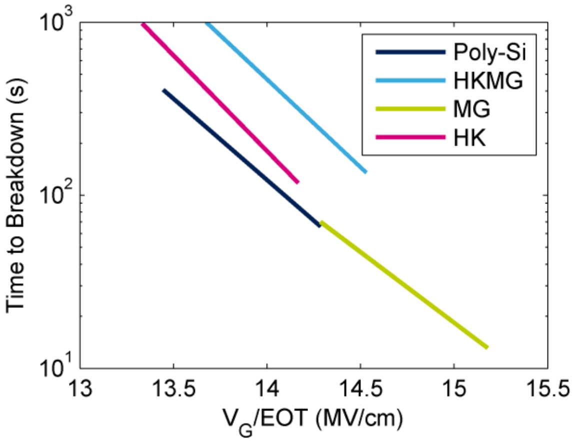


Figure 88: Time to breakdown for 63% of devices at 125°C vs V_G/EOT for ONO samples

4.3.3 Conclusions and perspectives on ONO and HKMG

The results presented in Sections 4.3.1 and 4.3.2 show that the switch from a Poly-Si control gate to the gate stack from a HKMG logic process does not degrade the properties of the ONO dielectric. The electric field at breakdown and the time to breakdown are both improved by the switch to the HKMG stack. It is therefore of interest to attempt a 40 nm embedded flash product integration having the standard ONO dielectric and the HfSiON/TiN/a-Si control gate from the 28 nm logic process. Additional electrical and physical characterization of the final flash product should be completed in order to evaluate the impact of the traps that appear to be induced by the use of the metal gate.

In the context of this ONO and HKMG flash cell, the deposition method of the TiN may have to be modified because the PVD is not compatible with the topography of a flash cell. Recall from Section 3.2.4 that the ONO and gate stack are deposited onto the first Poly-Si layer, it has a finger-like structure with a much higher aspect ratio than the substrate in a typical 28 nm logic gate deposition. Figure 89 shows TEM cross sections and EDX elemental analysis for HKMG stacks deposited onto patterned Poly1 by either PVD or ALD. The ALD deposition leads to a much more conformal TiN layer, especially on the sidewalls where no TiN was deposited by PVD. These differences in conformality were expected as they are well-known [Plu00]. The impact of switching from PVD to ALD TiN deposition on the CMOS logic devices would need to be investigated.

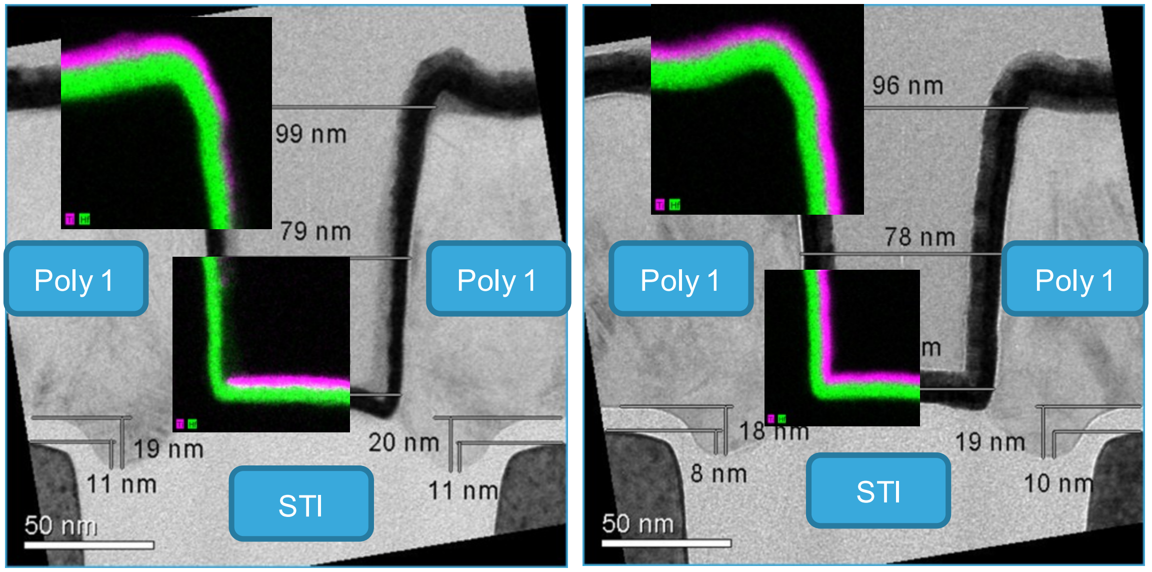


Figure 89: TEM cross section comparison of TiN deposited by (left) RF-PVD and (right) ALD. The insets are elemental mappings by EDX for Hf (in green) and Ti (in purple).

4.4 Fully high-k IGDs for co-integration with HKMG logic

Now that we have shown that the HKMG stack does not degrade the ONO IGD we can investigate whether or not we can become more aggressive in the IGD selection by completely replacing the ONO. In this section, HfSiON deposited by Metal Organic Chemical Vapour Deposition and HfO_2 deposited by Atomic Layer Deposition are considered to be the contenders. The first question is whether or not we can move to lower IGDs in order to improve performance without leakage or reliability problems. The second question is whether or not an IGD that is promising at the materials level can be successfully integrated in a 40 nm embedded flash product.

4.4.1 Hafnium Aluminates IGD selection and fabrication

A series of samples based on $HfAlO_x$, Al_2O_3 and multilayers of the two was planned, based on the advanced IGDs described in Chapter 1. The planned gate stacks for the batch of 24 wafers included comparisons for the effects of the:

- Al_2O_3 thickness
- $HfAlO_x$ thickness
- Presence of a SiO_2 layer at the top interface
- Presence of the SiO_2 layer at the bottom interface
- Poly-Si vs HKMG gate stacks
- $HfAlO_x/Al_2O_3/HfAlO_x$ multilayers to increase direct tunneling distance (see Figure 24)

Unfortunately, due to a series of technical problems [Gra15], this selection of IGDs was abandoned and IGD based on hafnium oxide and hafnium silicates were targeted in replacement.

4.4.2 Hafnium Oxide and Hafnium Silicates IGD Fabrication and Characterization

The remainder of the batch of wafers was divided into samples based on HfO_2 and on HfSiON, summarized in Figure 90, to study the impact of EOT, and the top and bottom interfaces. A series of three high-k thicknesses on a 4 nm SiO_2 bottom oxide and the same HfSiON/TiN/a-Si gate were made for both HfSiON and HfO_2 to see the trend in properties with thickness. For the thickest HfSiON sample, two addition splits were defined, without bottom oxide and without TiN in the gate. The splits to study the top and bottom interfaces were not repeated for the HfO_2 series as the HfSiON samples were predicted to be more promising as described in state of the art in Section 1.4.3.

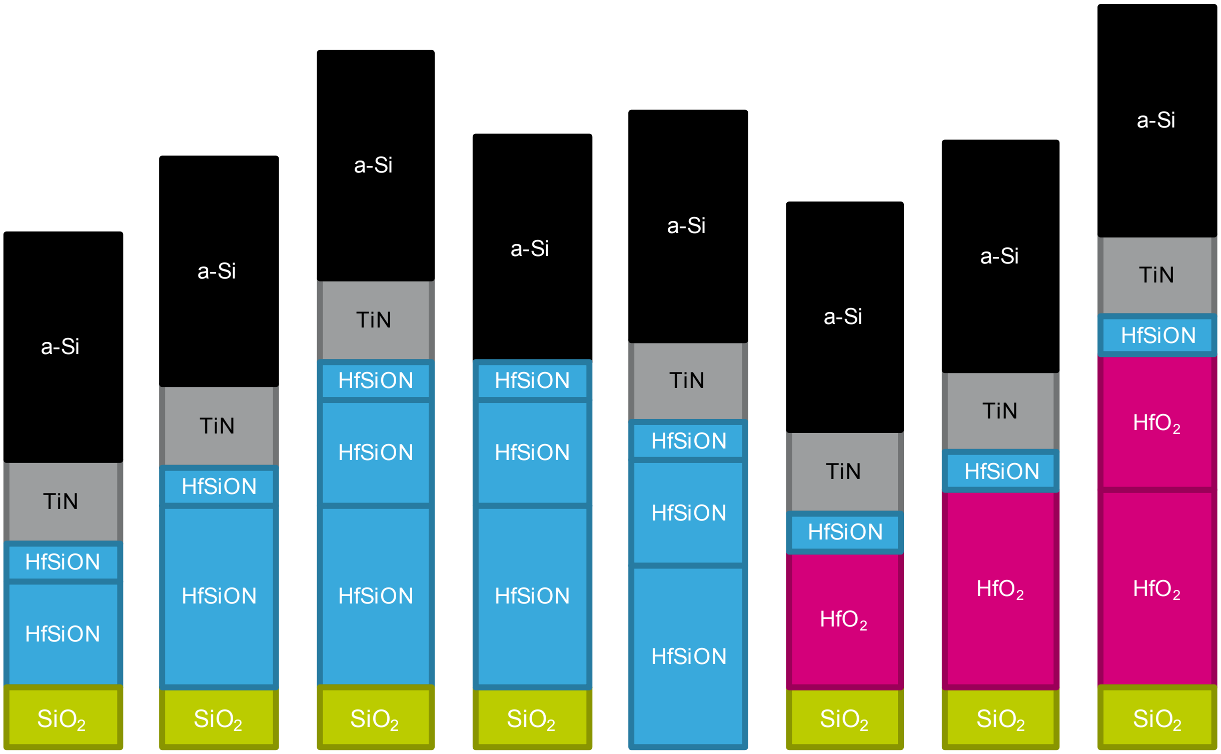


Figure 90: IGD stacks for MOCA capacitors based on hafnium silicates

The fabrication of these samples followed the same MOCA capacitor process flow steps outlined in Section 4.2.1 and Figure 79. The change was that the ONO deposition and removal was replaced by bottom oxide then HfSiON or HfO_2 deposition and removal. In order to limit deposition recipe development efforts, only two deposition recipes for each material were developed:

- HfSiON 7 nm and 12 nm by Metal Organic Chemical Vapour Deposition in an AMAT Centura tool, followed by a plasma nitridation and an anneal at 850 °C
- HfO_2 9 nm and 13 nm by Atomic Layer Deposition in an ASM Polygon, followed by an anneal at 650 °C

Backside of wafer cleaning processes were developed for the worst cases of 12 nm of HfSiON and of 13 nm of HfO_2. Only having two thicknesses available meant that the third thickness was obtained by a double deposition, for example 12 nm of HfSiON was deposited and then 7 nm of HfSiON was deposited to obtain the 19 nm sample. It is important to note that the thick high-k layers were capped by the HKMG stack, so the 19 nm HfSiON sample actually involves three HfSiON deposition steps: 12 nm and 7 nm to form the 19 nm IGD plus 1.8 nm as part of the HKMG stack.

4.4.3 Characterization of Hafnium oxide and Hafnium Silicate

The effect of the high-k layer's thickness on the CV characteristics for the six splits having a bottom oxide and thick high-k layer topped by the HKMG are shown in Figure 91. A slight shoulder peak is observed in all splits, similar to the shoulder peak observed with the ONO plus HKMG samples in Section 4.3.1.

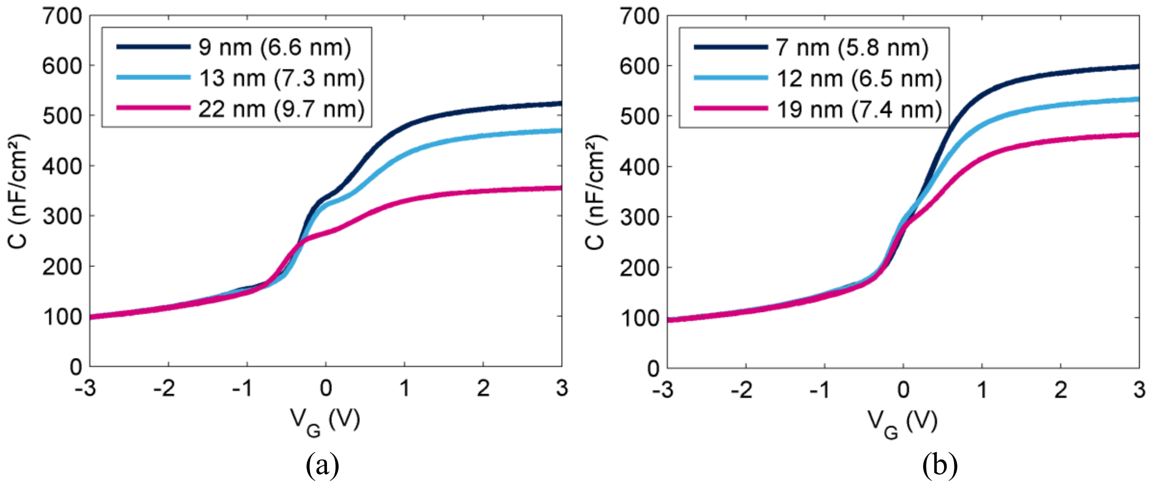


Figure 91: CV characteristics at 100 kHz for three different nominal thicknesses (EOT in parentheses) for IGDs based on (a) SiO_2 + HfO_2 and (b) SiO_2 + HfSiON

The equivalent oxide thickness of the IGD is given by the sum of the bottom oxide, the IGD itself and the HKMG stack as given by Equation 4.2

$$EOT_{total} = EOT_{Bottom\ Oxide} + EOT_{IGD} + EOT_{HfSiON\ from\ HKMG} \quad Equation\ 4.2$$

Based on the EOT of the 4 nm SiO_2 with HKMG sample (Figure 80), the first and last components are considered to sum to 4.5 nm, reducing Equation 4.2 to Equation 4.3.

$$EOT_{total} = 4.5\ nm + thickness_{IGD} \times \left(\frac{k_{SiO_2}}{k_{IGD}}\right) \quad Equation\ 4.3$$

The total EOT is plotted as a function of the nominal high-k thickness to yield a straight line in Figure 92 where the slope is inversely proportional to the k-value of the high-k. The dielectric constant of HfSiON in these samples is approximately 26, which falls close to the reported range of 15-25 for $HfSiO_x$ [Cas02]. For HfO_2 it is 16, lower than expected as the silicate typically has a lower k-value [Cas02], although the difference could be explained by the crystalline phase as the monoclinic phase has a lower k-value, as low as 14 [Bre14]. As the relation between EOT and nominal thickness is linear we believe that we have a consistent material.

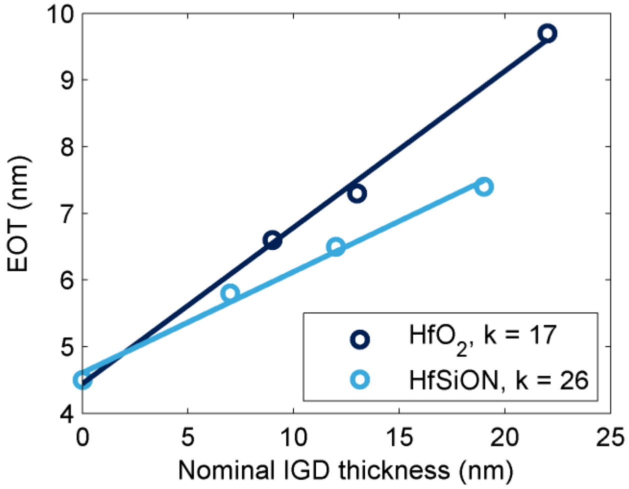


Figure 92: k-value determination for SiO_2 + HfO_2 and SiO_2 + HfSiON samples by EOT vs nominal thickness analysis

The current voltage characterizations of the six samples, shown in Figure 93, show that increasing the thickness of the high-k layer leads to considerably reduced leakage currents. Due to the wide range of EOT values, the comparison of leakage current vs V_G/EOT is made in Figure 94.

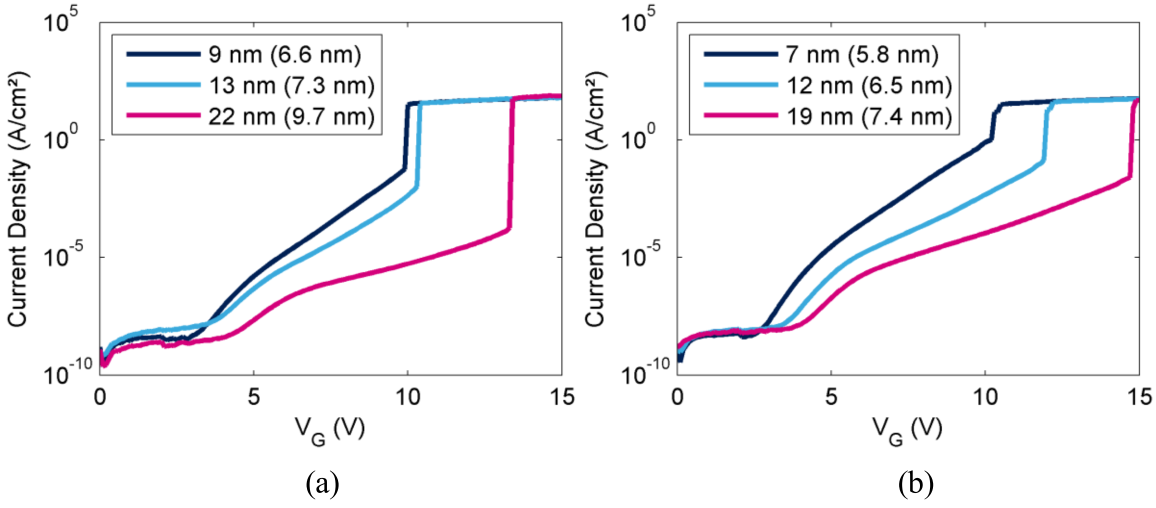


Figure 93: IV characteristics for (a) SiO_2 + HfO_2 and (b) SiO_2 + HfSiON IGD series for different nominal high-k thicknesses (EOT in parentheses)

When the gate bias is normalized by the EOT, the difference between the thick high-k samples is smaller, for example the 9 and 13 nm HfO_2 samples follow almost the same curve in Figure 94(a). The ONO IGD (with HKMG stack) sample is included as a reference for comparison and we can see that it generally performs better than the high-k samples, although the high-k IGDs have EOTs 2-4 nm thinner than the ONO sample. At higher fields, the 19 nm HfSiON sample exhibits lower currents than the ONO in Figure 94(b).

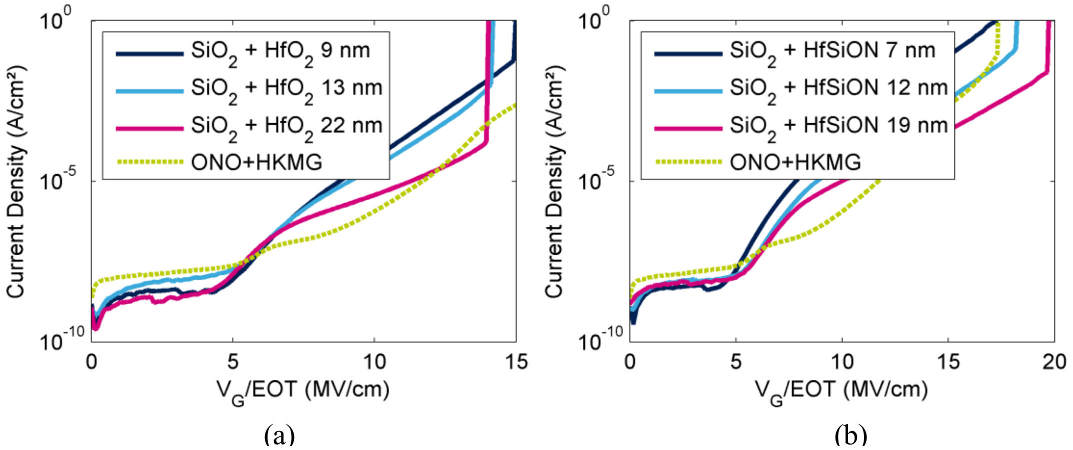


Figure 94: Current density vs average electric field for the ONO and (a) HfO_2 and (b) HfSiON IGDs

In order to better compare samples with a wide range of EOT values, the current density as a function of EOT is plotted in Figure 95 where each series is a different gate bias. For a given gate bias the HfSiON samples exhibit lower leakage currents than the HfO_2 samples in a similar EOT range (6.5 to 7.5 nm). The ONO with HKMG sample is also included as a reference, however the EOT is considerably larger than those of the high-k IGDs so a relevant comparison is difficult.

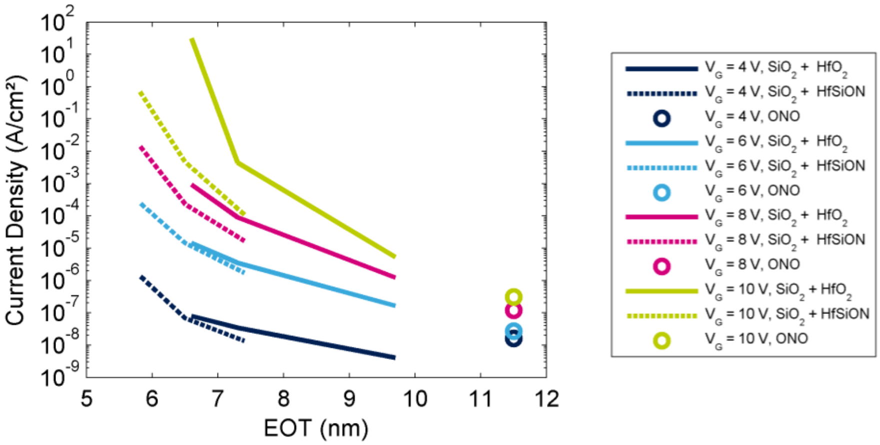


Figure 95: Current density as a function of EOT at a given gate bias for the HfO_2 and HfSiON IGDs, an example of the ONO IGD is included for comparison

4.4.3.1 Top and Bottom interface effects for 19 nm HfSiON

The samples having either no bottom oxide or no TiN allow for the study of the effects of the bottom and top interfaces, respectfully and their CV curves are shown in Figure 96. Just as was observed in the ONO plus HKMG trails in Section 4.3.1, the shoulder peak is only observed when the TiN metal gate is present; the nitridation of the HfSiON alone does not appear to contribute as much to the nitrogen diffusion in the stack as the TiN gate does. Trapping in the high-k is readily observable in the sample without a bottom oxide. As the CV measurement sweeps from depletion (V_G = -3 V) to accumulation (V_G = +3 V) the traps close to the Si/HfSiON interface are filled by electrons from the substrate. During the sweep back from accumulation to depletion, the trapped charges shift the CV curve in the positive direction. When the 4 nm of SiO_2 are present, the charge transfer to the high-k is blocked at such low biases (+/- 3V).

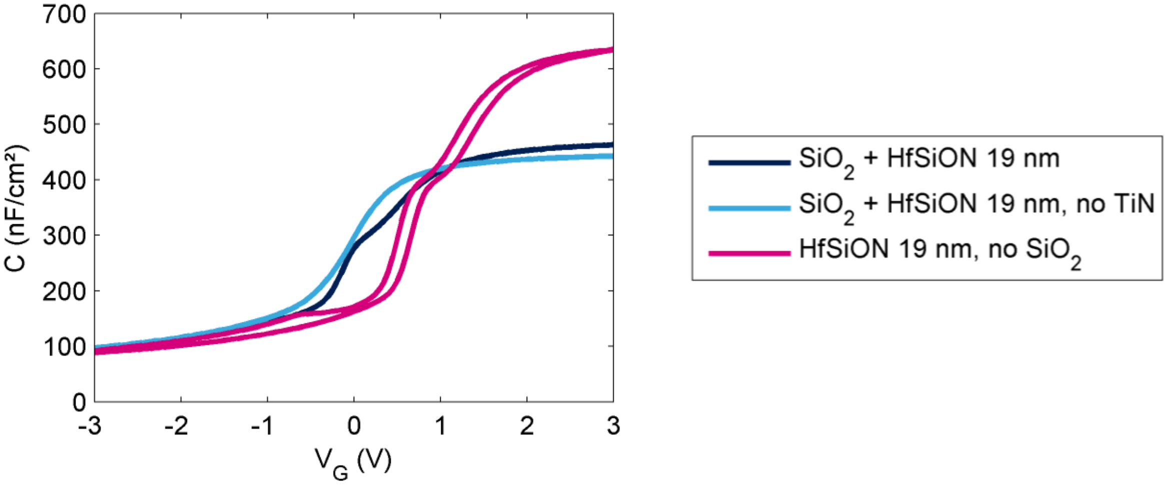


Figure 96: CV characterization of top and bottom interfacial effects for SiO_2 + 19 nm HfSiON IGD (25°C, 100 kHz)

The current voltage characterization for the three splits is shown in Figure 97 vs both the gate bias and the average electric field. The TiN in the control gate does not at all impact the current density, so the conduction is considered to be controlled only by the IGD and not the metal gate, however the TiN may still play a role in the reliability of the devices. Without the bottom oxide, the injection barrier into the IGD changes and the low-field currents are increased, as expected because the HfSiON bandgap is smaller than that of SiO_2, about 6 vs 9 eV [Cas02].

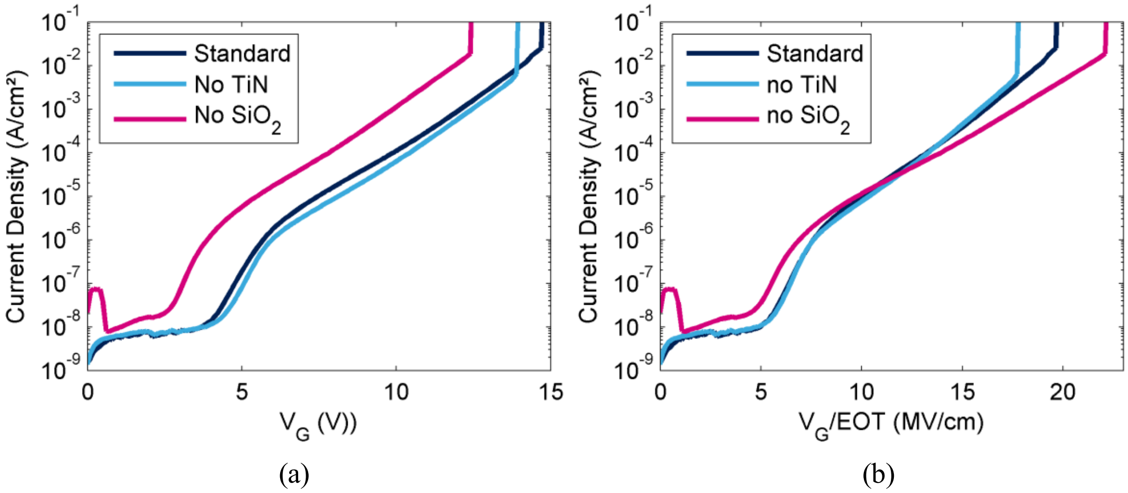


Figure 97: IV characterization of top and bottom interfacial effects for SiO_2 + 19 nm HfSiON IGD in (a) gate bias and (b) average electric field

4.4.4 Breakdown Characterization

The breakdown voltages from the IV curves for 20 devices of each of the thick high-k samples are plotted in Figure 98. As expected, the thicker layers break down at higher gate biases, with the 19 nm HfSiON sample showing the highest breakdown voltage. The two smaller HfO_2-based samples do not differ significantly and the HfSiON samples appear to have breakdown voltages more dependent on the thickness.

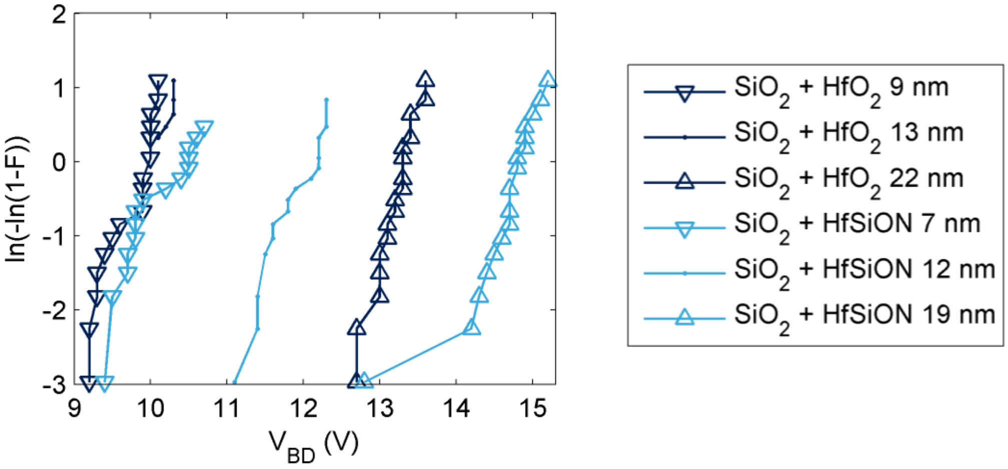


Figure 98: Failure rates against gate bias for the HfO_2 and HfSiON IGDs for 20 devices at 25 °C

Note that 14 out of 20 devices in the 13 nm HfO_2 series broke down at only a few volts. This is considered to be a processing error, most likely during the CMP step. As there are many splits and few wafers per split, the parameters of the CMP process were not able to be optimized for each split. This resulted in some wafers having uneven planarization. For example, if the CMP leaves 200 nm or 400 nm of SiO_2 at different radii then a ring will appear due to the change in color based on the change in thin-film thickness. The important thing is that the capacitor devices are well defined; similar rings are observed on several wafers without impacting electrical measurements. In the case of this wafer, only the 6 center devices were well-defined and testable.

In order to better compare the range of samples, the failure rate is also plotted against the average electric field at breakdown in Figure 99, with the ONO HKMG sample for comparison. The HfSiON based samples break down at higher electric fields, while the HfO_2 based samples are grouped together showing break down at lower fields. This is a promising result for the use of HfSiON as an IGD.

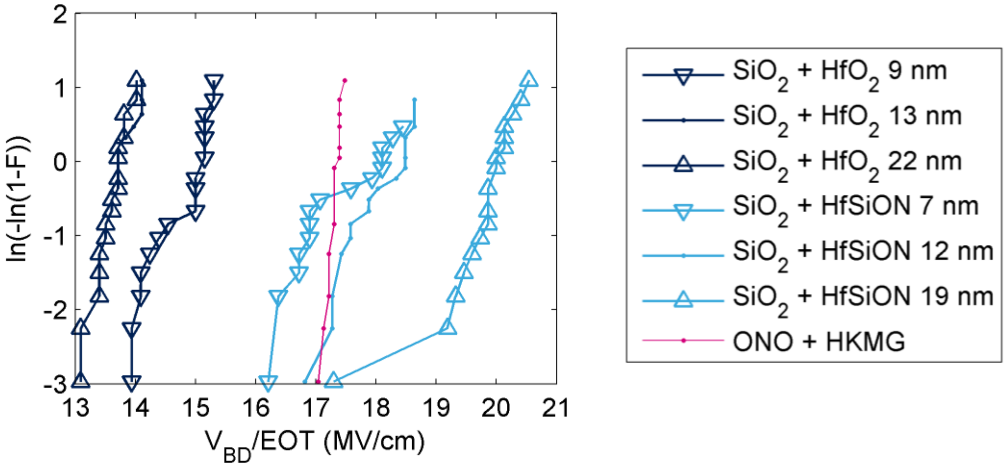
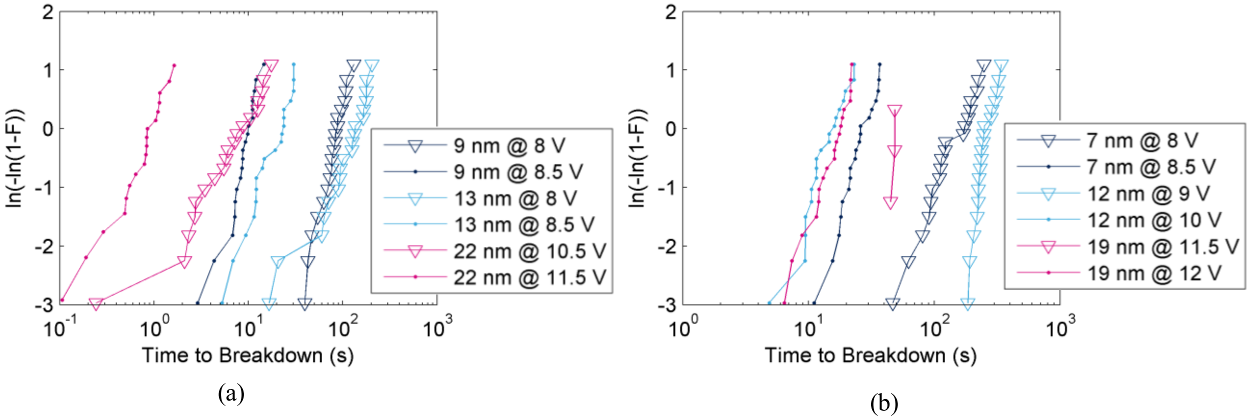


Figure 99: Failure rates against electric field for the HfO_2 and HfSiON IGDs for 20 devices at 25 °C

The Weibull plots from the TDDB tests of the HfO_2 and HfSiON samples at biases from 8 to 12 V are shown below in Figure 100. As there is a variety of different biases used, the differences are not immediately apparent without further treatment.



(a)

(b)

Figure 100: Weibull plots for TDDB measurements of (a) SiO_2 + HfO_2-based IGDs and (b) SiO_2 + HfSiON IGDs at 125°C at gate biases from 8 to 12 V

The time to breakdown for 63% of the devices, W = 0 in the Weibull plot, was determined by the linear fit in the Weibull plot and this characteristic time is plotted against the gate bias in Figure 101. The ONO with HKMG sample is included as a reference. The general trend is that the thicker high-k layer leads to a more robust device. To better compare the samples of different EOTs, the time to breakdown for 63% for the devices, is plotted against the average electric field in Figure 102.

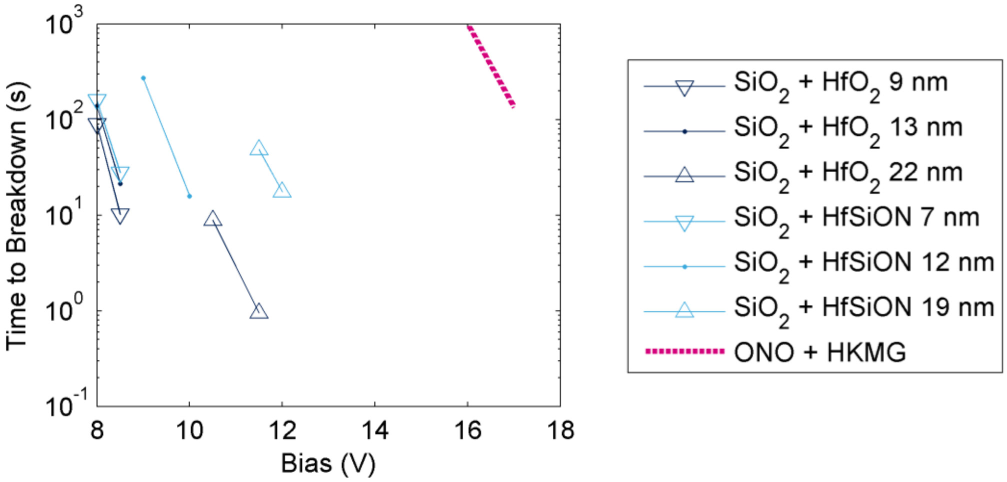


Figure 101: Time-to-breakdown for 63% of devices vs gate bias at 125°C

Based on the TDDB results, the HfSiON is more robust than the HfO_2 at 125 °C. Within the HfO_2 series, an increase in HfO_2 thickness does not increase the time to breakdown; the 9 nm sample is the most robust. Within the HfSiON series, there is little difference between the 7 and 12 nm samples; however the 19 nm sample offers an increased resistance to breakdown. The linear extrapolation of the thickest HfSiON is on the order or magnitude of the ONO and HKMG sample. This is another promising result supporting the use of a thick HfSiON IGD

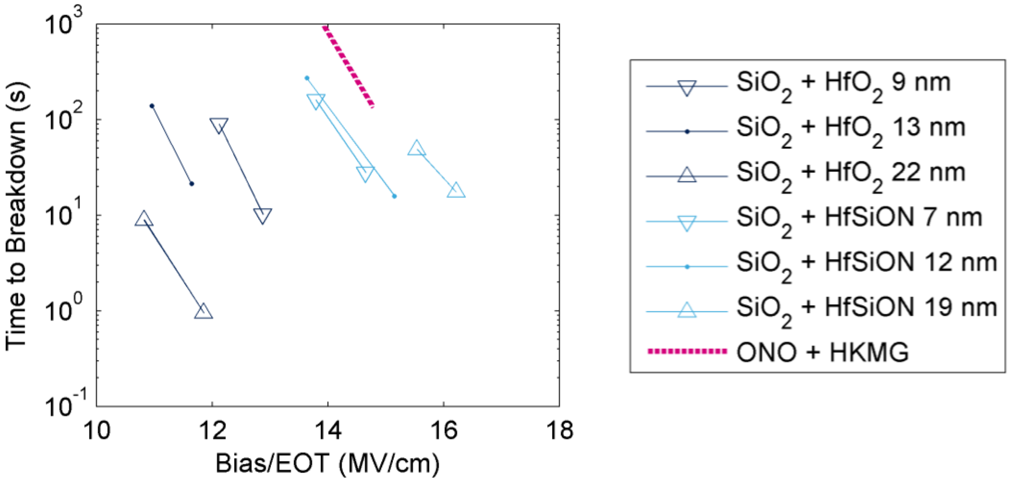
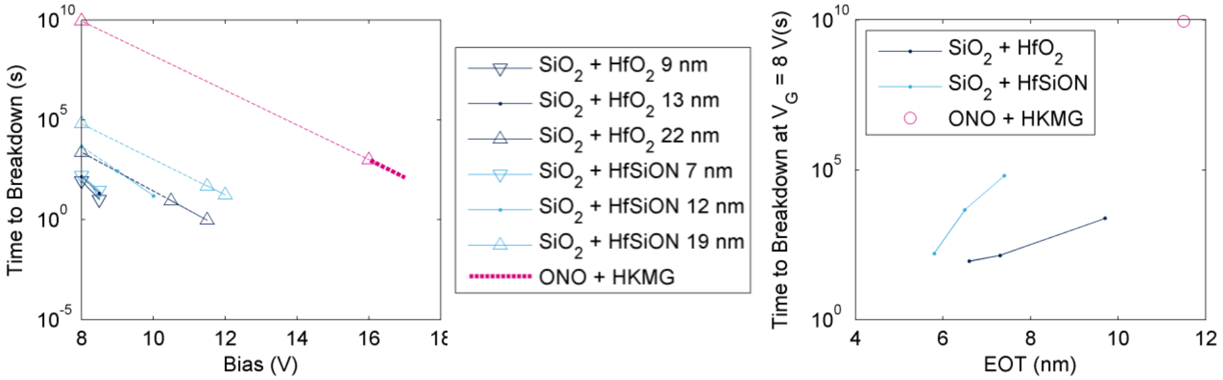


Figure 102: Time-to-breakdown for 63% of devices vs average electric field at 125°C

At this point, we believe that the HfSiON is intrinsically more robust than the HfO_2. However, before it can be integrated into a product we need to be sure that it would still yield a sufficient lifetime at normal operating conditions. Normally, an extrapolation would be made to lifetime at 5 V but given the small number of samples we make this extrapolation to 8 V in Figure 103(a) as this is closer to the regime where the devices were tested. This lifetime at 8 V is then plotted against the EOT Figure 103(b), showing that the HfSiON samples have a longer lifetime at 8 V. From this figure it is difficult to compare with the ONO plus HKMG sample, although the slope of the HfSiON curve is promising as it implies that thicker IGDs would have much longer lifetimes.



(a)

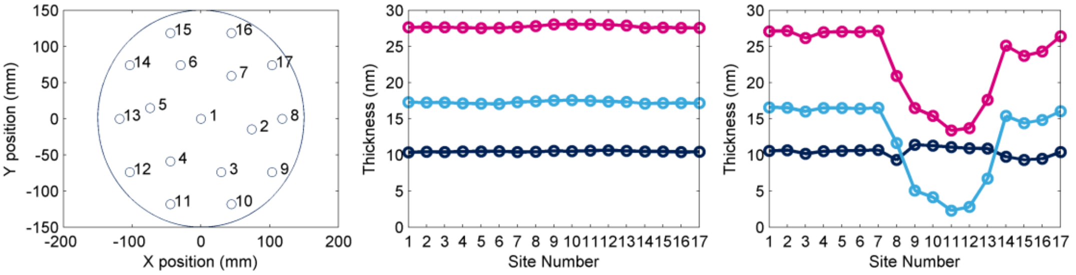
(b)

Figure 103: (a) IGD lifetimes are extrapolated to 8 V in order to compare , in (b) the IGD lifetime vs the EOT

4.4.5 Hafnium silicate based IGD integration with 40 nm embedded flash product

One batch of embedded flash wafers was dedicated to the fabrication of devices having the IGDs based on those in Figure 90: a 4 nm SiO_2, plus a thick high-k layer and a TiN/a-Si control gate. The TiN was deposited using ALD in order to avoid the problems with conformality as seen in Section 4.3.3. The etching of these materials is very common in the cleanroom, however the layers are typically very thin. These wafers present high-k layers an order of magnitude thicker than those normally etched and they are crystalline. Therefore, the first step was to determine the etch rate of thick HfSiON using blanket wafers; the HfSiON etch rate was approximately 5.4 nm per minute.

The approach taken for the Poly2HV etch (outlined in Section 3.2.6) was to use the standard poly-Si and TiN etch steps to remove the gates and then etch the HfSiON based on time, with an over-etch to ensure that the HfSiON is completely removed. The first pilot wafer exhibited an observable endpoint for the poly-Si, but the endpoint of the TiN etch was unclear. The ellipsometry results measured at 17 locations on the wafer in Figure 104 showed that the TiN was not homogenously removed over the entire wafer. Sites 8 through 13 showed a decrease in HfSiON thickness whereas the other points did not show a reduced HfSiON thickness. This is consistent with TiN residues on parts of the wafer blocking the HfSiON etch.

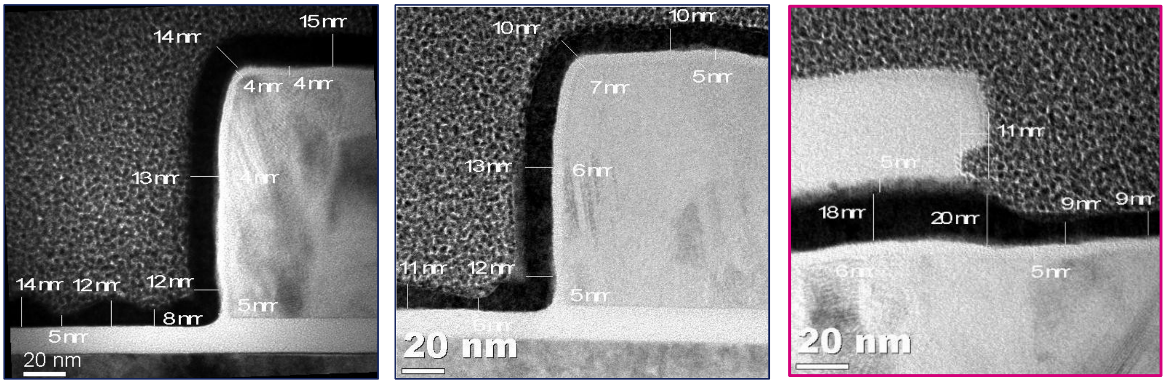


(a) (b) (c)

Figure 104: (a) mapping of ellipsometry measurements (b) measured IGD thicknesses before gate depsotion (c) measured IGD thicknesses after first etch

The second pilot wafer was etched with longer breakthrough and overetches for the TiN before applying a 25% longer HfSiON etch. TEM analysis of the first two wafers showed, in Figure 105, that a significant amount of HfSiON remains after the etching, 16 and 10 nm for the first and second pilot wafers, respectively. The HfSiON also remains on the sidewalls of the HV

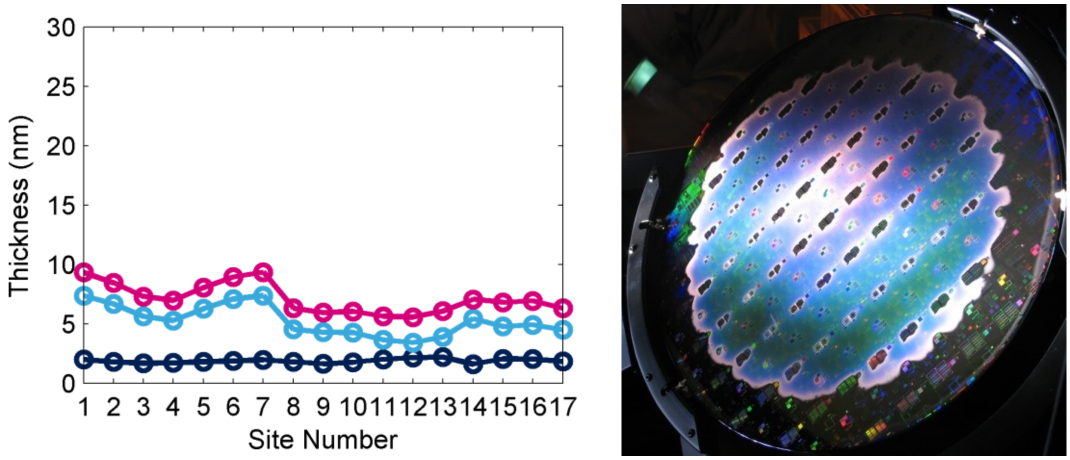
MOS devices. The rightmost image shows the removal of the poly-Si and TiN at the ONO capacitor opening.



(a) (b) (c)

Figure 105: TEM images of (a) HV MOS in first pilot wafer (b) HV MOS in second pilot wafer (c) ONO capacitor opening in second pilot wafer

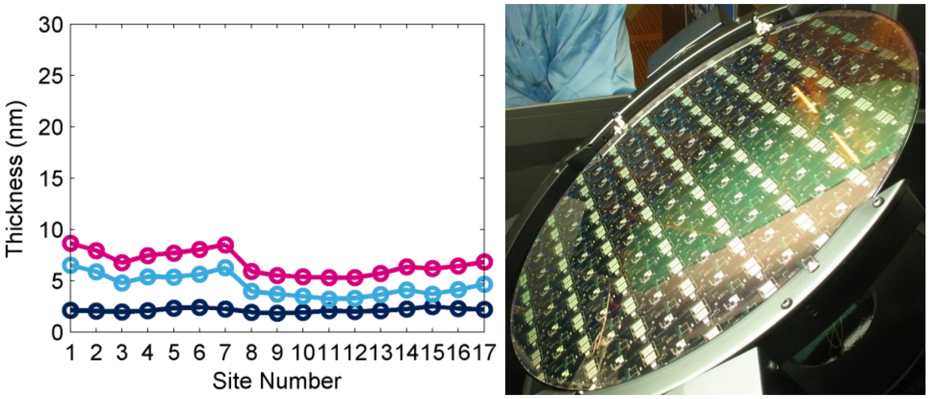
The third pilot wafer was then etched using a very long HfSiON etch step to compensate for the slower-than expected high-k etch rate. The ellipsometry results in Figure 106 show total 5-6 nm of HfSiON on top of 2 nm of SiO_2. This likely corresponds to the complete removal of the HfSiON (leaving only ≈8 nm of SiO_2) while the ellipsometry continues to report a SiO_2 and an HfSiON layer as the model used is defined as having two layers [Jaw17]. Visual inspection of the wafer, Figure 106, clearly showed a problem as there was a residue present on most of the wafer.



(a) (b)

Figure 106: (a) measured IGD thicknesses of the third pilot wafer (b) visual inspection of the wafer showing resist residues on the wafer

The remaining residue on the wafer is considered to the photo resist that is hardened by the bombardment of ions during the very long HfSiON over-etch. The over-etch was then reduced for a fourth pilot wafer to avoid the hardening of the photo resist. The results, shown in Figure 107, were positive as the HfSiON appears to have been removed and no resist is visible on the wafer after the stripping.

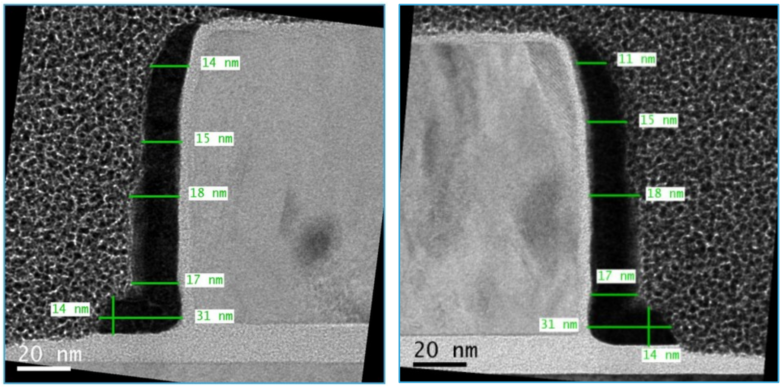


(a)

(b)

Figure 107: (a) measure IGD thicknesses of the fourth pilot wafer (b) visual inspection of the wafer showing the complete removal of the photo resist

As the ellipsometry measurements are made on a relatively large flat surface, TEM analysis of the HV MOS devices is required to fully evaluate the etch processes. The TEM analysis in Figure 108 shows that the third pilot etch completely removes the HfSiON from horizontal surfaces and that reduction in over etch for the fourth pilot was reasonable as the HfSiON is removed without the photo resist hardening. The thick HfSiON appears to remain on the sidewalls.



(a)

(b)

Figure 108: TEM analysis of the HV MOS transistors showing removal of the HfSiON on the horizontal surfaces but leaving HfSiON and possibly TiN on the sidewalls of the (left) third and (right) fourth pilot wafers

The composition of the sidewall seen in Figure 108 was analyzed using EDS and summarized below in Figure 109 for the fourth pilot wafer. The thick dark layer is confirmed to be HfSiON, which remains in same rounded form that is commonly seen after dry etching of spacers. In the bottom corner, the least accessible place during a dry etch, some TiN from the gate remains as well. We can therefore say that the modified dry etch for the second set of pilot wafers did effectively remove the entire HfSiON/TiN/a-Si stack, but only from the horizontal surfaces.

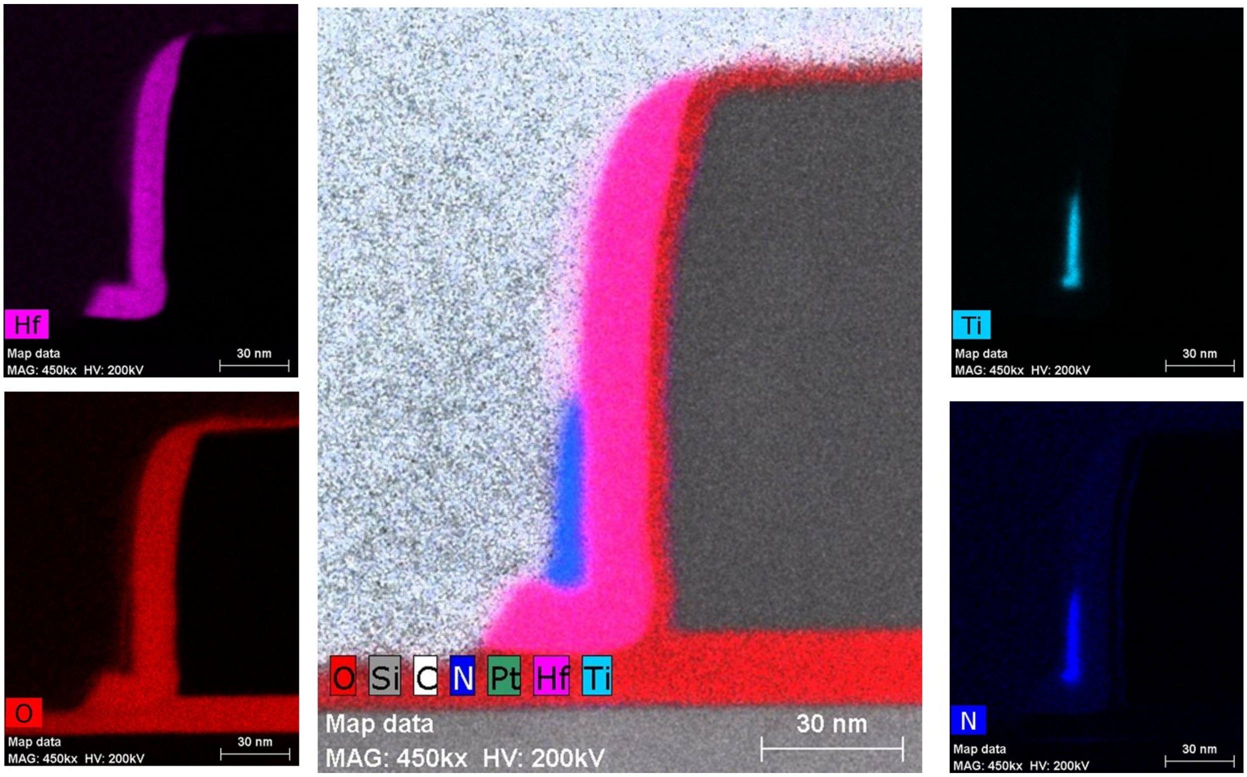


Figure 109: EDS mappings for Hf, O, Ti and N in a HV MOS device from the fourth pilot wafer

The remaining HfSiON on the sidewalls would not necessarily be an insurmountable hurdle for the HV device characteristics. However the Ion Implantation step that follows the Poly2HV etch would contaminate implantation chambers, so the HfSiON needs to be completely removed. Alternatively, the HfSiON could be covered by a spacer before the ion implantation, however this would complicate the integration. The difficulties associated with this etching step led to this series of samples being abandoned as it was clear that it would not be possible to develop a cleaner Poly2HV etch, and then the more complicated Poly2NVM etch with a limited number of wafers in a short amount of time.

4.4.5 Conclusions and Perspectives on thick HfO_2 and HfSiON IGDs

Once the constraint of having to use an ONO has been lifted, the options for IGDs are nearly endless. We first defined a series of samples based on $HfAlO_x$, Al_2O_3 and multi-layers of the two based on the most advanced IGDs outlined in Section 1.4.3. These were later changed to a series based on HfSiON and HfO_2 due to a series of technical challenges that left the first set impossible to fabricate.

Through CV, IV and TDDB characterization we have tested the use of IGDs consisting of 4 nm of SiO_2 and 7-22 nm high-k layers of HfO_2 or HfSiON for their applicability to future use as IGDs. Overall, the HfSiON is better suited to use as an IGD than the HfO_2. The CV characterization showed that in both cases the TiN metal gate contributes to an electrically active trap, but without electrical measurements on flash cells the impact of the trap is difficult to evaluate. The dielectric constant of the HfSiON, at 26 is higher than that of $HfO_{2,}$ at 16, an unexpected trend as HfO_2 was expected to have the higher k-value. The higher-k value can lead to more capacitive coupling with IGDs that are physically larger.

At a given gate bias, the HfSiON samples showed lower leakage currents than their HfO_2 counterparts in a given EOT range. Although they did not show lower leakage currents than the ONO-based IGDs, the EOT of the thickest HfSiON sample was 4 nm less than that of the ONO. It is possible that with a slightly thicker HfSiON, the currents can be reduced to below those of the ONO while still maintaining a smaller EOT to increase coupling. In terms of the electric field at breakdown, these HfSiON layers actually performed better than the ONO based IGD of a larger EOT. The HfO_2 based IGDs broke down at lower fields, and in TDDB measurements they were less reliable and showed a degradation in robustness when the thickness of the HfO_2 was increased. The HfSiON based IGDs broke down later, and at higher fields than the HfO_2 based IGDs while also exhibiting the trend that a thicker HfSiON layer improves resistance to breakdown. In fact, the 19 nm HfSiON sample showed TDDB characteristics on the same order of magnitude as the ONO sample, a very promising result.

We have attempted to integrate HfSiON into a 40 nm embedded flash product in the same process as outlined in Chapter 3. Being an order of magnitude thicker than normal, and crystalline instead of amorphous, the HfSiON dry etching proved to be a big challenge. An etch chemistry was found that could remove the metal gate and the HfSiON from planar surfaces. However, where the IGD was deposited on a patterned poly-Si layer, such as the flash floating gates and HV transistors' gates, the etch was ineffective at removing the IGD from the sidewall. The HfSiON and TiN that remained on the sidewall would have posed problems in terms of contamination and/or integration and the series was abandoned. We believe that with time and resources, a more appropriate etching approach could be found in the future.

4.5 Conclusions and Perspectives

At the 40 nm node, a business-as-usual approach was possible with regards to using an ONO and a common polysilicon control gate that is shared with the CMOS logic devices. This is no longer the case at the 28 nm and later nodes as the CMOS logic changes to a high-k metal gate stack. Although it would be possible to add a polysilicon layer that would function only as the flash control gate, this addition polysilicon layer and its patterning would increase the cost of the process. In trying to avoid the increased expense of the extra layer, different IGDs are proposed that would share the control gate with the CMOS logic devices. Due to the more exploratory nature of these IGDs, they were tested using damascene capacitor test vehicle, outlined in Section 4.2 in lieu of an embedded flash product.

The first set of IGDs, proposed in Section 4.3 answered the question of whether or not the HKGM would degrade the ONO in a"bridge" solution. This solution combines aspects of the current 40 nm embedded flash technology and the 28 nm CMOS logic process. They are based on an ONO IGD combined with the HKMG stack. The replacement of the poly-silicon control gate by the HKMG stack does not degrade the IGDs current voltage characteristics which remain dominated by the conduction of the ONO itself. The leakage currents at the low-fields, relevant for data-retention, were not accessible using the simple capacitors. The switch to a HKMG stack also appears to increase the reliability of the IGDs as they were more robust in time dependent breakdown tests, a very promising result. Nonetheless, capacitance voltage measurements indicate that the TiN in the HKMG stack induces an electronically active trap above the midgap silicon; at this time the impact on flash operation is unknown.

None of the basic characterizations carried out provided any reason to avoid an ONO with a HKMG stack. Therefore a test of embedded flash products having an ONO and a HKMG stack is highly recommended, however it will pose integration challenges such as those outlined in Chapter 3. The biggest potential issue with the fabrication of the IGD and gate stack will be the deposition of TiN, as the standard method will not provide a conformal layer over the patterned Poly1 layer that forms the floating gates of the flash.

The second set of IGDs, in Section 4.4.3 focused on answering the question of whether or not we can further improve performance with more aggressive EOTs. These IGDs were a 4 nm SiO_2 layer covered by a thick 7-22 nm high-k layer of either HfSiON or HfO_2 and capped off by the HKMG stack from the 28 nm CMOS logic process. CV characterizations showed that the HfSiON, with its nitridation and annealing actually has a higher k-value than the HfO_2 after its nitridation and annealing. This is beneficial because a thicker HfSiON layer could be used to provide the same capacitive coupling between the control gate and floating gate. The HfSiON based samples also exhibited lower leakage currents and were more robust than the HfO_2 based samples. The basic characterization of the damascene capacitors indicates that an HfSiON based IGD is more likely to represent an improvement than an HfO_2 based IGD. The thick HfSiON samples showed higher resistance to breakdown, longer lifetimes and time dependant breakdown characteristics on the order of those observed with the standard ONO.

The integration of a flash product having a thick HfSiON proved to be complicated as the wafers were abandoned at the Poly2HV etch step. In this case the dry etch was not sufficient to remove the HfSiON from the sidewalls of the HV MOS transistors. In the future it would likely be possible to find a solution comprising a dry and a wet etch, or a protective spacer. The impact of the change in the TiN deposition method, to something more conformal, on the CMOS logic devices will have to be evaluated as well.

4.6 Bibliography

[Bre14] L. Breuil. et al., “HfO2 Based High-k Inter-Gate Dielectrics for Planar NAND Flash Memory”, *IEEE Electron Device Letters*, vol. 35, no. 1, 2014

[Cas02] J.D. Casperson, et al., “Materials issues for layered tunnel barrier structures” *Journal of Applied Physics 92*, 2002

[Dob17] A. Dobri, et al., “Evaluation of ONO compatibility with high-k metal gate stacks for future embedded flash products” in *Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon*, Athens, Greece, Apr. 2017

[Fis04] D. Fischer, et al. “Effects of Nitridation on the Characteristics of Silicon Dioxide: Dielectric and Structural Properties from ab initio Calculations” *Phyical Review Letters.* Jun., 2004

[Ghe14] A. Ghetti, “Gate Oxide Reliability: Physical and Computational Models” in *Predictive Simulation of Semiconductor Processing*. Springer Berlin Heidelberg. 2014

[Gra15] F. Grassaud. (2015, Apr. 25). *Incendie dans une salle blanche au CEA de Grenoble* [Online] Available: http://france3-regions.francetvinfo.fr/auvergne-rhone-alpes/incendie-salle-blanche-au-cea-grenoble-712213.html

[Hu10] C. Hu, “Chapter 5: MOS Capacitor” in *Modern Semiconductor Devices for Integrated Circuits,* 1st Edition. Pearson. 2010

[Jaw17] J.A.Woollam. (2017). *Ellipsometry Tutorial* [Online]. Availible: https://www.jawoollam.com/resources/ellipsometry-tutorial

[Moo14] P. Moon et al., “Field-dependent charge trapping analysis of ONO inter-poly dielectrics for NAND flash memory applications”, *Solid-State Electronics*, Apr. 2014

[Mor17] D. Morillon et al. , “High voltage MOSFETs integration on advanced CMOS technology: characterization of thick gate oxides incorporating High K Metal Gate stack from logic core process” in *IEEE International Conference of Microelectronic Test Structures*, Grenoble, France, 2017

[Plu00] J.D. Plummer, et al., “Chapter 9: Thin Film Deposition” in *Silicon VLSI Technology: Fundamentals, Practice and Modelling, 1st ed.* Upper Saddle River, NJ, Prentice Hall, 2000

[Rob15] J. Robertson and R. Wallace, “High-K materials and metal gates for CMOS applications” *Materials Science and Engineering: R: Reports*, Feb. 2015

Chapter 5: General Conclusion

Chapter 5

In growing from nothing, to a soon-to-be 400 billion dollar market, the microelectronics industry has developed into a multifaceted market with technologies that have differentiated themselves in order to serve a huge variety of markets. The floating gate NOR flash cell has long been used in the memory market and still plays a major role in embedded flash memories at the 40 nm technology node. The silicon dioxide/silicon nitride/silicon dioxide trilayer, "ONO", has been the integrate dielectric (IGD) for decades. A good IGD is electrically thin to provide good floating gate to control gate coupling while remaining thick enough to block leakage (data loss); aggressive scaling improves performance but increase risk of data loss. The continued use of the ONO brings up several questions. How will we know when the ONO scaling has become too aggressive? Can it be replaced? Would the ONO be compatible with the high-k metal gate logic processes at future nodes? And can high-k materials represent a good option for ONO replacement? The research conducted in the last two decades on planar NAND devices for high density applications can help guide the responses concerning embedded NOR flash by identifying candidate materials such as alumina and hafnium silicates.

As charge loss from the floating gate leads to the loss of data, it is therefore of great interest to know from where these loses are originating. The logical approach would be to directly measure the leakage current through an IGD on a test capacitor; this is impossible because the currents at low electric fields are below the measurement limits. Floating gate methods were invented in order to extend the measurement limits, but they are still not sensitive enough to reach such low fields. We have developed and patented the Oxide Stress Separation (OSS) technique for measuring leakage currents across the IGD in a nominal flash cell, described in Chapter 2. The premise of the OSS technique is that by carefully selecting the threshold voltage and biasing conditions it is possible to have no electric field in the tunnel oxide. With the tunnel oxide at the flat band condition, the entire potential drop occurs in the IGD and any leakage currents can be considered to occur via the IGD. The leakage currents on the order of 10-22 to 10-23 A are deduced from the changes in the threshold voltage of the flash cell under test and can be compared with the results from data retention bake tests. At the 40 nm technology node, the ONO remains of good quality and only contributes a small fraction of the leakage currents from the floating gate. The OSS technique will be useful in evaluating the data retention properties in future nodes, whether not they contain ONO or high-k IGDs.

One of the potential candidates for the replacement of the silicon nitride in the ONO is alumina as it offers an increased dielectric constant for better coupling while still having a relatively high band gap to reduce leakage currents. For the first time, in Chapter 3, 40 nm embedded flash chips have been fabricated using alumina-based IGDs either in the form of alumina layers on 4 nm of silicon dioxide, with or without a final silicon dioxide layer on top of the alumina. The modifications to the standard process flow for the embedded flash product were consequential, with many steps taking place at the CEA-Leti. The alumina layer is easily removed from the control gates of the High Voltage MOS devices and the IGD capacitor openings using the wet etch process developed at the CEA-Leti. The etching step that defines the word lines of the flash cells was more complicated and resulted in sub-optimal profile, although the exposed alumina was completely removed, reducing worries of contamination in the proceeding steps.

Elementary test patterns showed that OA and OAO layers exhibit good intrinsic performances, with leakages and breakdown voltages that seem to allow for the reduction of the IGD's EOT. Using theses electrical test we see that the bilayer seems to be more promising than trilayers and that alumina deposited using ozone as the oxygen precursor is of better quality. The complete characterization of those stacks was not completed during this work, due to timing issues. Nevertheless, the materials will be evaluated by fine characterization (data retention, yield and OSS measurements) to give a complete overview of potential offered by Alumina stacks for IGD. On our first samples, the expected improvements in flash cell erase performance are not observed with the IGDs of lower EOT. The loss of coupling due to a decrease in the surface area of the floating gate, caused by the sub-optimal word line profile, is believed to eliminate the gains in coupling linked to decreases in EOT. After this first demonstration of fully integrated alumina-based based cells, new Poly2NVM etch process improvement should bring back the benefits of reduced EOT on cell performances.

At the next node, 28 nm, the CMOS logic devices are formed using high-k metal gate (HKMG) stacks which will necessitate changes to the IGD. The first question one can ask is whether or not the HKMG stack would degrade the ONO if a similar two-gate integration scheme is retained. Through CV, IV and breakdown analysis in Chapter 4 we have shown that the HKMG does not degrade the properties of the ONO. In fact, the HKGM stack resistance to degradation in time dependent breakdown tests. The next question is whether or not a fully high-k gate stack would allow for improved coupling without increases in leakage. This was also tested in Chapter 4 with IGDs consisting of 4 nm of silicon dioxide followed by 7 to 22 nm of hafnium silicate or hafnium oxide. Hafnium silicates showed a higher-k value, lower leakage currents, longer time to breakdown and better lifetimes that hafnium oxide samples. Therefore hafnium silicates are more likely to represent an improvement for IGD applications. The integration of a thick hafnium silicate-base IGD proved complicated as the dry etch was not sufficient to remove the IGD from all the desired locations during the first etching step. In all cases at the 28 nm node and beyond, the gate metal deposition method will become critical in terms of conformality in order to accommodate for the patterned floating gates, not present in normal logic processing.

In conclusion, the current ONO is of very high quality and may therefore not need to be immediately replaced in the 28 nm node. It is imperative to test a 40 nm embedded flash integration with the ONO and the high-k metal gate stack that was tested on capacitor devices in Chapter 4. Should the ONO be replaced by a high-k IGD, both alumina and hafnium silicates have shown themselves to be good candidates for the use in flash cells. How they are integrated into an embedded flash product is just as important as their material properties; especially with regards to the word line etch. This etch step defines the flash cells and is critical because differences in control gate width may lead to a loss in capacitive coupling that cannot be regained via thinner IGDs. Future integrations of alumina or hafnium silicate IGDs will require very close collaboration with dry/wet etch teams in order to find suitable etching conditions. In these optimizations, precise characterization of the IGD *itself* such as OSS technique will be very useful for the understanding the devices in order to guide the process development.

Appendix A: Charge trapping in the ONO

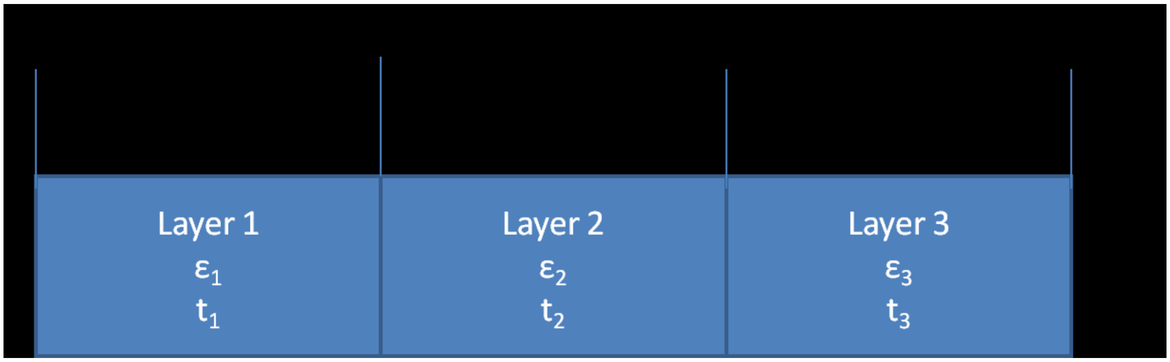
One can estimate the electric potential in a multilayer dielectric structure by solving Poisson's Equation in each layer, where E is the electric field, ρ is the volume charge density and ε is the dielectric constant of the layer.

$$\nabla \cdot \mathrm{E} = \frac{\rho_{\mathrm{free}}}{\varepsilon}$$

The boundary conditions between the layers are subject to Gauss' Law where Φ_D is the surface integral of the displacement field, D (equal to εE) over the area A and Q_{free} is the charge at the interface.

$$\Phi_D = \oiint \boldsymbol{D} \cdot d\boldsymbol{A} = Q_{free}$$

The general case for a trilayer is three layers having thickness, t_i, and dielectric constant ε_i. The gate is at the right hand side. σ is the areal charge density at the interface.



At the bottom interface:

$$\oiint \boldsymbol{D} \cdot d\boldsymbol{A} = Q_{free}$$

$$(\pi r^2 \varepsilon_2 E_2 - \pi r^2 \varepsilon_1 E_1) = \pi r^2 \sigma_{b(ottom\ interface)}$$

$$\varepsilon_2 E_2 - \varepsilon_1 E_1 = \sigma_b$$

$$E_2(x = x_b) = \frac{\sigma_b + \varepsilon_1 E_1}{\varepsilon_2}$$

At the top interface

$$\varepsilon_3 E_3 - \varepsilon_2 E_2 = \sigma_{t(op\ interface)}$$

$$E_2(x = x_t) = \frac{\varepsilon_3 E_3 - \sigma_t}{\varepsilon_2}$$

In the nitride

$$\frac{dE_2}{dx} = \frac{\rho}{\varepsilon_2}$$

$$E_2(x) = \frac{\rho x}{\varepsilon_2} + C_1$$

$$-\frac{dV_2}{dx} = \frac{\rho x}{\varepsilon_2} + C_1$$

$$V_2(x) = -\int \frac{\rho x}{\varepsilon_2} + C_1 dx$$

$$V_2(x) = -\frac{\rho x^2}{2\varepsilon_2} - C_1 x + C_2$$

With V(x=0) = 0V and V(x=x_G)=V_G

$$\Delta V_1 + \Delta V_2 + \Delta V_3 = V_G$$

$$-E_1 t_1 + \left[-\frac{\rho x^2}{2\varepsilon_2} - C_1 x + C_2\right]_{x\_b}^{x_t} - E_3 t_3 = V_G$$

$$-E_1 t_1 + \left[-\frac{\rho {x_t}^2}{2\varepsilon_2} - C_1 x_t + C_2\right] - \left[-\frac{\rho {x_b}^2}{2\varepsilon_2} - C_1 x_b + C_2\right] - E_3 t_3 = V_G$$

$$-E_1 t_1 - \frac{\rho {x_t}^2}{2\varepsilon_2} - C_1 x_t + \frac{\rho {x_b}^2}{2\varepsilon_2} + C_1 x_b - E_3 t_3 = V_G$$

$$-E_1 t_1 + \frac{\rho}{2\varepsilon_2}({x_b}^2 - {x_t}^2) + C_1(x_b - x_t) - E_3 t_3 = V_G$$

Note : $x_b - x_t = -t_2$ and ${x_b}^2 - {x_t}^2 = t_1^2 - (t_1^2 + 2t_1 t_2 + t_2^2) = -2t_1 t_2 - t_2^2$

$$-E_1 t_1 - \frac{\rho}{2\varepsilon_2}(2t_1 t_2 + t_2^2) - C_1 t_2 - E_3 t_3 = V_G$$

$$C_1 = \frac{1}{t_2}\left(-E_1 t_1 - \frac{\rho}{2\varepsilon_2}(2t_1 t_2 + t_2^2) - E_3 t_3 - V_G\right)$$

At the bottom interface	At the top interface
$\frac{\sigma_b + \varepsilon_1 E_1}{\varepsilon_2} = \frac{\rho x_b}{\varepsilon_2} + C_1$ $\frac{\sigma_b + \varepsilon_1 E_1}{\varepsilon_2} = \frac{\rho x_b}{\varepsilon_2} + \frac{1}{t_2}\left(-E_1 t_1 - \frac{\rho}{2\varepsilon_2}(2t_1 t_2 + t_2^2) - E_3 t_3 - V_G\right)$	$\frac{\varepsilon_3 E_3 - \sigma_t}{\varepsilon_2} = \frac{\rho x_t}{\varepsilon_2} + C_1$ $\frac{\varepsilon_3 E_3 - \sigma_t}{\varepsilon_2} = \frac{\rho x_t}{\varepsilon_2} + \frac{1}{t_2}\left(-E_1 t_1 - \frac{\rho}{2\varepsilon_2}(2t_1 t_2 + t_2^2) - E_3 t_3 - V_G\right)$

Bottom interface

$$\frac{\sigma_b + \varepsilon_1 E_1}{\varepsilon_2} = \frac{\rho x_b}{\varepsilon_2} + \frac{1}{t_2}\left(-E_1 t_1 - \frac{\rho}{2\varepsilon_2}(2t_1 t_2 + t_2^2) - E_3 t_3 - V_G\right)$$

$$t_2\sigma_b + t_2\varepsilon_1 E_1 = \cancel{\rho t_2 t_1} - \varepsilon_2 E_1 t_1 \cancel{- \rho t_1 t_2} - \frac{1}{2}\rho t_2^2 - \varepsilon_2 E_3 t_3 - \varepsilon_2 V_G$$

$$E_1[t_2\varepsilon_1 + \varepsilon_2 t_1] + E_3[\varepsilon_2 t_3] = -t_2\sigma_b - \varepsilon_2 V_G \ - \frac{1}{2}\rho t_2^2 \qquad Equation\ 1$$

Top interface

$$\frac{\varepsilon_3 E_3 - \sigma_t}{\varepsilon_2} = \frac{\rho x_t}{\varepsilon_2} + \frac{1}{t_2}\left(-E_1 t_1 - \frac{\rho}{2\varepsilon_2}(2t_1 t_2 + t_2^2) - E_3 t_3 - V_G\right)$$

$$t_2\varepsilon_3 E_3 - t_2\sigma_t = \rho t_2(t_1 + t_2) - E_1 t_1 \varepsilon_2 - \frac{\rho\varepsilon_2}{2\varepsilon_2}(2t_1 t_2 + t_2^2) - \varepsilon_2 E_3 t_3 - \varepsilon_2 V_G$$

$$t_2\varepsilon_3 E_3 - t_2\sigma_t = \cancel{\rho t_2 t_1} + \rho t_2^2 - E_1 t_1 \varepsilon_2 \cancel{- \rho t_2 t_1} - \frac{1}{2}\rho t_2^2 - \varepsilon_2 E_3 t_3 - \varepsilon_2 V_G$$

$$E_1[t_1\varepsilon_2] + E_3[t_2\varepsilon_3 + \varepsilon_2 t_3] = t_2\sigma_t + \frac{1}{2}\rho t_2^2 - \varepsilon_2 V_G \quad Equation\ 2$$

Equations 1 and 2 are summarized below, to be solved numerically

$$E_1[t_2\varepsilon_1 + t_1\varepsilon_2] + E_3[t_3\varepsilon_2] = -t_2\sigma_b - \varepsilon_2 V_G \ - \frac{1}{2}\rho t_2^2$$

$$E_1[t_1\varepsilon_2] + E_3[t_2\varepsilon_3 + \varepsilon_2 t_3] = t_2\sigma_t - \varepsilon_2 V_G + \frac{1}{2}\rho t_2^2$$

When E_1 and E_2 are known, the constants C_1 and C_2 can be calculated:

$$C_1 = \frac{1}{t_2}\left(-E_1 t_1 - \frac{\rho}{2\varepsilon_2}(2t_1 t_2 + t_2^2) - E_3 t_3 - V_G\right)$$

$$C_2 = \frac{\rho {t_1}^2}{2\varepsilon_2} + C_1 t_1 - E_1 t_1$$

This allows one to calculate the potential in the second layer

$$V_2(x) = -\frac{\rho x^2}{2\varepsilon_2} - C_1 x + C_2$$

The values for the trapped charges in the nitride (a few 10^{13}/cm^2) trapped were used as they represent the order of magnitude of charge than can be stored in the nitride of the ONO according to experiments on ONO MOSCAP devices.

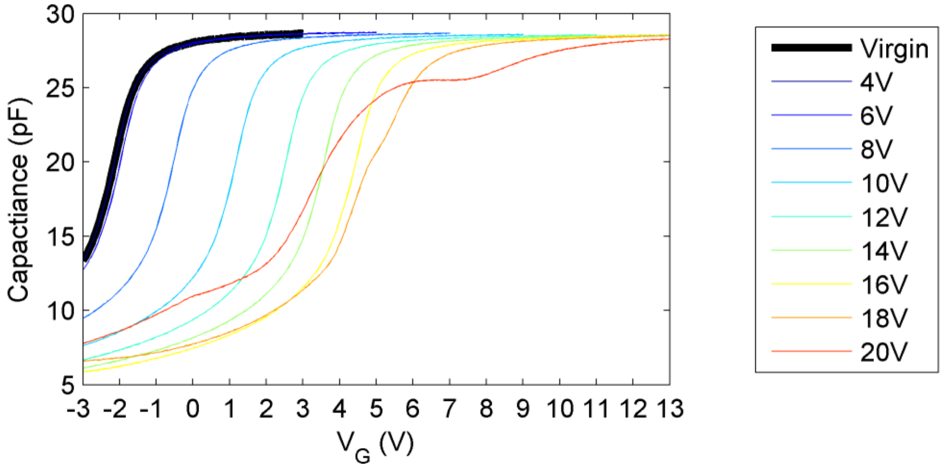


Figure A1: CV curves after 100 ms V_G pulses of increasing magnitude to show the effect of charge trapping.

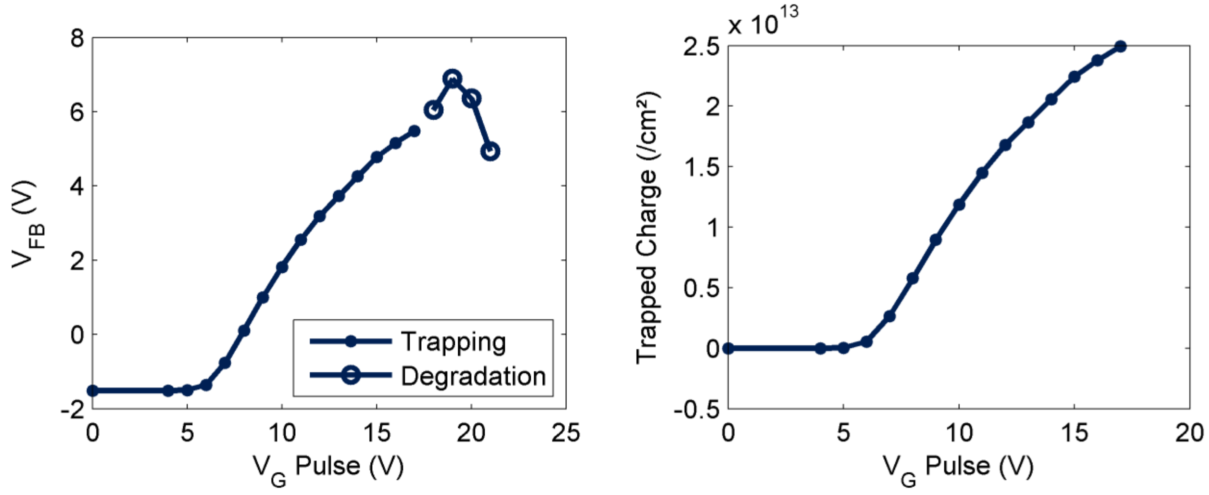


Figure A2: (left) flatband voltage as a function of gate pulse strength. A simple shift in the characteristics is considered to be trapping in the nitride, while the change in the C was considered to be degradation/breakdown (right) the concentration of trapped charge (assuming that it is centered in the middle of the ONO) from the relation that $\Delta V_{FB} = \frac{Q_{ONO}}{2\times C_{ONO}}$