Development and Characterisation of a Process Technology for a 0.25µm SiGe:C RF-BiCMOS embedded Flash Memory

Dissertation
Zur Erlangung des akademischen Grades
Doktor der Ingenieurwissenschaften
(Dr.-Ing.)
der Technischen Fakultät
der Christian-Albrechts-Universität zu Kiel

Alexander Fox

Kiel
2006

1. Gutachter	Prof. Dr.-Ing. Peter Seegebrecht
2. Gutachter	Prof. Dr. rer. nat. Helmut Föll
3. Gutachter	Prof. Dr.-Ing. Reinhard Knöchel
Datum der mündlichen Prüfung	11.7.2006

Acknowledgements

During the time of this dissertation project I worked as part of the scientific staff at the IHP microelectronics research centre, Frankfurt (Oder), Germany. The dissertation has been supervised by Professor Dr.-Ing. Peter Seegebrecht of the “Technische Fakultät der Christian-Albrechts-Universität“, Kiel, Germany.

First of all I want to thank Professor Dr.-Ing. Peter Seegebrecht for the supervision of this dissertation, for the confidence he put in me and especially for the time and effort that this required to spend. I enjoyed all the discussions, which gave this dissertation-project a clear structure.

Next I want to thank Professor Dr. rer. nat. Helmut Föll and Professor Dr.-Ing. Reinhard Knöchel to agree to take the responsibility and effort of being the second revisers.

An important support for my work was the positive influence of my colleagues at the IHP.

I want to thank Karl-Ernst Ehwald for the countless technical discussions. I have been in the unique position to work close to a scientist of his experience and deep understanding of physical mechanisms. I also thank him for giving me confidence in my work throughout the project.

I want to thank Dr. Dieter Knoll for technical discussions regarding the baseline process, the HBT and also general technical and project-related topics. I am also thankful for the helpful discussions about the language and structure of the dissertation text.

I want to thank Dr. Bernd Heinemann and Dr. Holger Rücker for being always open for technical and general discussions that supported the development of this the project.

I want to thank Dr. Bernd Tillack for his support as department head of the technology department of the IHP. I am very glad that I have been able to do this work at the IHP in such a straight way and with all the support I needed.

I want to thank Felix Fürnhammer for his support of the project during his time as department head of the technology department of the IHP. I also thank him for being one of the main initiators of the embedded-flash-project at the IHP and of the cooperation with the NTU Kiev.

For the great support of preparing the silicon wafers I want to thank Reiner Barth, (clean room), Andre Wolf (maintenance), Sigrid Orlowski, Martina Glante, Klaus Glowatzki, Renate Gericke and Angelika Gregror (clean-room teams), and by this everybody of IHP’s clean room staff.

I want to thank the members of the process research department for their support in discussing, developing and adjusting all the single process steps: Dr. Steffen Marschmeyer, Dr. Thomas Grabolla, Dr. Harald Richter, Ulrich Haak, Dr. Achim Bauer, Dr. Beate Kuck, Dr. Klaus-Detlef Bolze, Katrin Blum and Thomas Morgenstern.

For the countless electrical measurements I want to thank Dr. Peter Schley, also for the many technical discussions about measurement related topics, Detlef Schmidt and Dr. R. Sorge.

For producing the numerous SEM pictures I want to thank Dr. Wolfgang Höppner, Heike Pfeiffer, Renate Naumann, Gabriele Morgenstern and Monika Döppner.

For doing the preparation of the TEM images I want to thank Dr. Petr Formanek and Dr. Günther Weidner.

I want to thank Christoph Wolf for the time he invested in preparing the complex test programs for the functional testing of the 1-Mbit memory and for doing the functional testing itself.

I want to thank Prof. Dr. Rolf Krämer and Dr. Michael Methfessel for the discussions regarding the role of embedded flash memories in a SOC environment and the required specifications.

I want to thank Dr. Biswanath Senapati for the work related to the electrical modeling of the flash cells and of the high-voltage MOS transistors.

I want to thank Dr. Valeriy Stikanov, Alex Gromovyy, Andriy Hudyryev from the NTU Kiev for the very nice cooperation and their work on the circuit design of the 1-Mbit memory.

Finally I want to thank Dr. Dag Behammer, who stands at the beginning of this work, as he has encouraged me in my decision to join the IHP for working on my dissertation, and who has over all been of great influence in evoking my interest in the field of microelectronics.

Chapter 1

Introduction

Today's SiGe:C BiCMOS technologies are key enablers for single chip wireless and broadband communication systems. They combine high performance RF-bipolar transistors required for analogue functions, e.g. wireless data transmission, with VLSI CMOS for digital data processing. A cost effective process technology makes this functionality feasible for the consumer market. We see today, for example, a rapidly growing spectrum of wireless applications, starting from mobile phones, Bluetooth or WLAN up to wireless sensors or sensor networks, which are in return drivers for technology development. This dissertation describes for the first time in detail the integration of a suitable embedded flash memory module into such a SiGe:C BiCMOS process technology.
The SiGe and subsequently the SiGe:C HBT pushed the silicon technology towards high frequency applications, which were earlier covered by costly III-V-semiconductor based processes [1][2], thus making RF-applications available for a mass market which were only niche products before. Additionally, the compatibility with VLSI CMOS allowed the combination of dense digital with analogue functions and paved the way for complete single chip solutions [3]. Research directions are on the one hand improving the SiGe-HBT itself towards higher frequencies, thus principally demonstrating the potential of this material, and on the other hand developing less complex process technologies compatible with standard CMOS. Besides SiGe BiCMOS for highest frequencies also Si-only BiCMOS or RF-CMOS technologies are being developed for similar applications.
Many research and development activities are going on to further combine different components on the same chip, e.g. high quality passives (inductors, capacitors, resistors or transmission lines), high voltage power MOS-devices (LDMOS), different kinds of memories (SRAM, DRAM, NVM), up to micromechanical components (MEMS) and sensors. The main reasons for this system-on-chip integration (SOC) [4], away from multi-chip solutions, are saving costs, reducing size, saving power (less output drivers), enhancing reliability, optimizing the design of clock, bus and control signals, reducing electromagnetic interferences, at board level, and, especially for embedded memories, enhancing system flexibility (by reconfigurable non-volatile memories), and achieving faster access times (reduced capacitive coupling) [5]. This has to be bought by more complex process technologies, issues like interference of the integrated components (e.g. noise problems) and complex functional testing requirements.
In the special case of mass market wireless communication systems, which are often battery-driven portable applications (e.g. mobile phones), cheap process technologies in terms of number of required mask levels and number of process steps are needed, which offer the required RF-performance and allow designing circuits with low power consumption. These boundary conditions are also valid for all additionally integrated components, and have thus also a big influence on the choice of a concept for an embedded memory.

For embedded non-volatile memories the most mature technology today is the floating-gate flash memory. The information is stored by charging an isolated piece of silicon in a MOS-

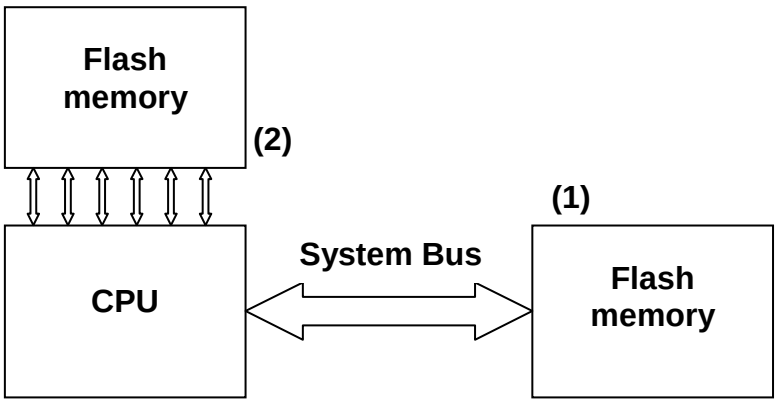


Figure 1: Flash memory incorporation in a System-on-Chip, (1) communication via system bus, (2) direct memory access by the CPU

transistor-like memory cell. The advantage of this concept is the CMOS compatibility, as no new materials or new kinds of process steps are needed. Much work was done understanding the reliability issues, scaling the cell size, reducing the operation voltage (which is significantly higher than CMOS operation voltage due to the high electric fields needed for cell-programming), or enhancing the memory's speed. Since the further development in all of these areas appears to reach some limits, more research was done in the last years to develop alternative memory concepts (SONOS, FeRAM, MRAM, PCM, etc.), which means in most cases investigating new materials and their process integration. These technologies are emerging and will in future take over the mainstream position of the floating gate embedded flash memory.
Floating-gate embedded flash memories are offered today in different forms by most Si-foundries as an optional module for their CMOS and also RF-CMOS processes, and semiconductor companies offer ICs with integrated NVMs. The combination of a Si-only BiCMOS process with an embedded flash memory is reported [6][7]. The integration of an embedded flash memory into SiGe:C BiCMOS is a consequent development, but not yet established. No literature is available on the process integration issues, and, to my knowledge, only one company has announced a SiGe process with NVM option to be available in future [8], but yet without further specification.

As SiGe:C applications evolve from mainly RF products to real BiCMOS products, which make use of the VLSI CMOS possibilities, memories ranging from low density (a few byte, e.g. non-volatile registers) up to medium density (~Mbit, e.g. for operation system storage) are required. The embedded flash process technology needs to produce cells that fulfil the size and performance requirements for such a memory, allow memories operating at low supply voltages and with low power consumption, and add only low additional cost to the process. An important requirement is the read access time. Fig. 1 shows two ways of how a flash memory can be incorporated into a system. The CPU can communicate with the flash memory via the system bus (1) or be directly connected with the flash memory (2). In the first case the speed requirements are more relaxed, while in the second case the flash memory should be readable at the clock frequency of the CPU. This makes, for embedded applications, the read access time more critical than other memory parameters. A modular integration scheme with low impact on the original CMOS and HBT devices would be favourable, e.g. for reusability of libraries.

This dissertation describes the development and characterisation of a process technology for an embedded floating-gate flash memory within a SiGe:C technology platform. The baseline

process is a fully featured SOC SiGe:C RF-BiCMOS process developed at the IHP microelectronics research institute [9]. It consists of an industry standard 0.25µm CMOS part and different integrated high performance SiGe:C HBTs. It is optimized for cost effective processing, e.g. different HBTs are integrated with only one mask level on top of the CMOS part. The flash memory integration is done with the main targets of achieving a process with low added cost and allowing memories with low power consumption, while cell size and performance need to fulfil the boundary condition of realizing up to Mbit memories.
The low cost issue is addressed by a process integration scheme with a low number of additional mask steps. This is achieved by the choice of the flash cell concept and by sharing masks and process steps for different purposes.
The target of low power consumption is addressed by the memory concept. Cells with a Fowler-Nordheim (FN) programming scheme were chosen, having intrinsically low power consumption during write and erase operations.
An additional task is the realization of devices capable of handling the required voltages for cell programming. These are well above the standard CMOS supply voltages, especially for FN-writing. A concept for integrating high voltage MOS transistors is presented, again targeting low added processing cost.
One more goal is to reduce the impact on the original CMOS and HBT devices as much as possible, which results in a modular process, to be able to offer an optional embedded flash memory only on demand.

The primary result of this dissertation project is the successful realization of a process technology for integrating an FN-programmed, floating gate embedded flash memory into the 0.25µm SiGe:C BiCMOS process of the IHP. The integration of flash cells and high voltage devices is done with 4 mask levels on top of the baseline BiCMOS flow, which is among the lowest numbers for embedded FN-programmed cells reported so far. The process implementation has been investigated geometrically by means of SEM and TEM. The different newly developed devices have been characterised electrically. The results indicate that the process is capable of fabricating devices with the required performance. To demonstrate the feasibility of the process for fabricating medium density memories, a 1-Mbit circuit has been demonstrated in cooperation with the Technical University of Kiev.
The dissertation was done as a research project at the IHP microelectronics research institute, using the process technology and characterisation facilities of the IHP. I acted as project leader and was working on myself on all of the process integration issues. The IHP process research department, where the single process steps were developed, supported my work, as well as the characterisation department, where the measurements were performed, the clean-room staff, who prepared the wafers, and the CAD department of the Technical University of Kiev, who worked on designing the 1Mbit memory chip based on the developed technology.
Results of this work have been published at the ICM 2004 [10]. The complete SOC BiCMOS technology, including embedded Flash, has been presented at the BCTM 2005 [11].

The thesis is structured as follows:

Chapter 1: Introduction.
Chapter 2 briefly introduces the field of non-volatile memories, then gives an overview of existing embedded non-volatile technologies and finally describes the fundamentals of the chosen memory concept, from single cell functionality to memory-array operation.
Chapter 3 explains the process integration scheme. Starting with a presentation of the baseline BiCMOS process, the newly developed additional process modules for integrating the flash cells and the peripheral high-voltage devices are described and discussed in detail.

Chapter 4 presents the results of implementing the proposed technological process flow in the 0.25µm pilot-line of the IHP. A geometrical analysis of the different devices is presented, followed by a discussion of important process steps, as well as the impact of these steps on the original BiCMOS devices.
Chapter 5 presents the results of electrical characterization of the individual devices. DC and transient characteristics of the flash cells are shown and discussed, also with respect to their dependence on important technological parameters. The results of basic reliability investigations done at single memory cells are shown. A DC characterisation of the high voltage MOS transistors is done. Finally the impact o the flash memory integration on the DC characteristics of the CMOS transistors and the HBT is discussed.
Chapter 6 shortly introduces the 1Mb demonstrator circuit and shows the most important results of functional testing.
Chapter 7 closes this work giving a summary and conclusions.

Chapter 2

The chosen embedded NVM concept

This chapter gives a brief introduction into the field of non-volatile memories, describes examples of technologies for embedded flash memories, before explaining more in detail the chosen memory concept. The aim is to range the chosen memory concept within the field of existing embedded non-volatile memories. The principles of different approaches are introduced and discussed. A more detailed description of the different memory concepts can be found in textbooks [5][12][13], review papers [14][15][16][17][18][19][20][21], and the respective specific publications given below.
For easier understanding of layouts and schematic cross section views please refer to the respective legends.

2.1. Non - volatile memories

Non-volatile memories are memories that keep their information without regularly refreshing their contents (unlike DRAM) and without a power supply connected (unlike SRAM). Two important performance parameters are the retention (-time) and the endurance. The retention states how long the information can be stored. This refers to the fact that the information is lost after some time, which is in floating gate memories due to electrons leaking off the floating gate. Typically a retention time of 10 years is specified. The endurance of a memory cell says how often the contents of a memory cell can be changed. This refers to the fact that each programming process of a cell leads to some degradation of its characteristics and can be done only a limited number of times. For floating-gate memories the limitation is given by the degradation of the tunnel-oxide caused by electrons passing through it. Typical values are 10^5 – 10^6 maximum allowed write / erase cycles.

2.1.1. Floating gate memories

These are today's mainstream non-volatile memories. A schematic cross-section and characteristics are shown in Figure 2. The information is remembered by the charging state of an isolated piece of silicon, the floating gate (FG), which is placed between a control gate (CG) and a channel in the silicon substrate or well area. This structure forms a MOS-like transistor (usually n-Type) whose threshold voltage V_T is determined by the amount of charge on the floating gate. The V_T-shift is detected by measuring the change in the drain-current at a control-gate voltage $V_{CG,r}$. The voltage applied to the control gate is capacitively coupled to the floating gate via the interpoly-dielectric. This coupling is determined by the control-gate coupling ratio k_{Cg}, being the ratio of the CG/FG capacitance and the FG/"rest of the world" capacitance. A high coupling ratio is usually favourable, as high FG-potentials are needed to get the required electric field for cell programming.
The floating gate's history goes back to UV-erasable EPROMs [22], then the invention of the electrically erasable EEPROM [23], and subsequently the flash-EEPROM [24], which allowed larger memories due to a reduced cell size, and introduced this technology to a broad market in the late '80s, after its reliability was demonstrated [25].

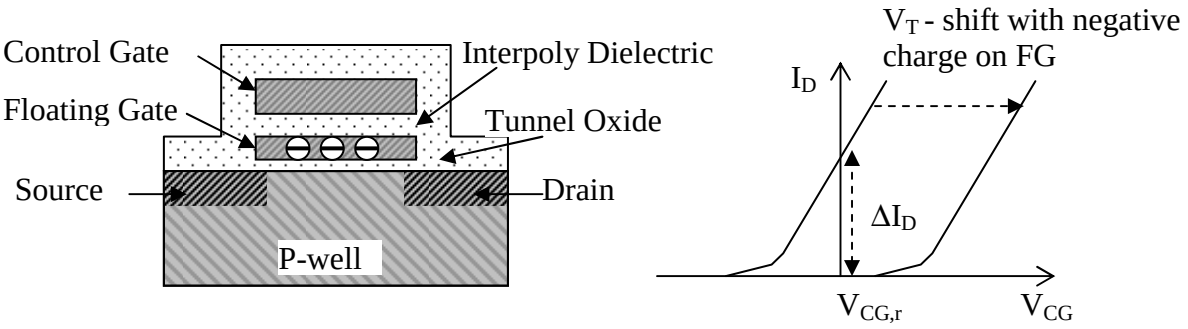


Figure 2: Schematic cross-section and characteristics of a floating gate memory transistor

The first differentiation that can be made is between (full feature-) EEPROM and Flash (-EEPROM) memories. In EEPROMs every single byte can be written and erased separately. In Flash memories only the writing is on byte level, while the erase operation is performed for all cells of a defined memory sector simultaneously. For selecting a single cell in EEPROMs, a select transistor is required in series with the actual memory transistor in every cell, leading to the two-transistor cell structure. Due to their larger cell size, EEPROMs are mainly used for smaller memory sizes and functions like parameter storage. The smaller cell size of the 1-transistor Flash memory cell, which in principle does not need a select transistor, leads to the feasibility of this technology for code and mass storage applications [14].
Programming of the cells, which means transferring electrons on or off the floating gate, is in most cases done by one of two different mechanisms, either Fowler-Nordheim (FN) tunnelling or hot-electron injection. At sufficiently high electric fields electrons overcome the oxide barrier by FN tunnelling. This mechanism can be used in both directions, getting electrons on and off the floating gate. FN tunnelling can be realized between the floating gate and the silicon substrate, either in a separate tunnelling area or in the channel area, from the floating gate to the source or drain, or between two poly-silicon layers. In the latter case it can be locally enhanced by a specially designed geometry. Hot-electron injection uses the possibility for electrons to gather enough energy in the lateral electric field at the drain side of the transistor to get over the oxide barrier, if the vertical electric field is also favourable for them to do so. Only a fraction of the total drain current is injected to the floating gate this way. The injection efficiency has been raised by optimized drain engineering, means like substrate enhanced injection (CHISEL [26]), and optimized cell architectures, e.g. split-gate cells for the so-called source side injection (see also below, section “HIMOS”), which divides the formation of the lateral and vertical electric fields to two separate gates, thus allowing optimization of the conditions for both separately. Hot-electron injection can only be used to get electrons onto the floating gate. Discharging is in any case done by FN-tunnelling. FN programming requires higher voltages, while hot-electron programming consumes more current. A comprehensive comparison, also with respect to the reliability issues of the different mechanisms, is for example given in [20].
Another difference between types of flash memories is the memory array’s architecture. The most important are NOR and NAND arrays. The NOR array allows fast random access to each byte during writing and reading. The wordline connects all control gates of one row, while in the crossing bitline each drain is contacted by ½ contact (two cells share one contact). The sources are all connected together via (silicided) active areas. This is the widely used, so-called common ground NOR array (Figure 5, in the description of the ETOX memory below). In the NAND array a number of cells belonging to one bitline are connected in series [5]. This leads to very compact cells (no separate bit-line drain contacts are necessary), at the expense of a slow random access time, as the reading current has to be passed through all cells connected in series. Mass data storage, stand-alone Flash memories up to 8 Gb have been

realized in NAND architecture [27]. NOR type memories are used for code storage applications with medium density requirements. Due to the specific requirements (medium density and high speed), NOR is the main architecture for embedded memories. Other concepts like AND [28], AG-AND [29] or DiNOR [30] play in general only minor roles.
A big issue in floating gate memory devices is the reliability in terms of endurance and retention. This is also one main limiting factor for progress in cell scaling, operation speed and operation voltage scaling. For array functionality always the behaviour of the worst cell is important, making statistical measures necessary when investigating reliability. Much work has been undertaken to understand the physical background and optimize the memory cell structure, the operating conditions and the process technology in this respect. Major roles play the tunnel-oxide and the interpoly-dielectric. The tunnel-oxide has to be produced without defects. Then, its characteristics change with continued cell writing and erasing. Stress-induced leakage current (SILC) occurs with the generation of electron traps within the oxide, thus affecting the retention. The transient programming characteristics change due to oxide charge build-up within the oxide. The reading current decreases due to degradation of the Silicon/Oxide interface. Other, more recently investigated effects are statistical fluctuations in the tunnel-oxide conduction behaviour, which have especially an impact on cell-array functionality. All these effects limit the minimum allowed tunnel-oxide thickness, which in return has an impact on the other cell properties [31]. Typical tunnel-oxide thickness is 8-10nm. The interpoly-dielectric is in most cases an ONO layer-stack. One reason for this is a reduction of enhanced electric fields at corners of the floating gate. In such corner situations, which occur as the control-gate is formed over an already patterned floating-gate, an intentionally introduced charge-build-up at the oxide/nitride interface shields the geometry related field enhancement. This leads to a reduced poly-poly leakage and thus enhanced reliability. The higher dielectric constant of the ONO leads additionally to a reduced EOT, which has a positive influence on the coupling ratio. Typically the EOT is around 20nm; minimum values of 11nm have been reported. Cappelletti has presented comprehensive summaries on flash memory reliability at the IEDM in 1994 and 2004 [32][33]. Other general summaries are given in [34][35][36]. SILC and related effects are investigated in [37][38][39][40]. Investigation of interpoly-dielectric related issues can be found in [41][42][43][44].

2.1.2. Nitride trap memory

This kind of memory distinguishes itself from the floating gate memory by storing electrons not in a single piece of silicon, but in discrete traps in a silicon-nitride layer or at the interface between silicon nitride and silicon oxide. It has the same transistor-like structure, where the charge-trapping layer or layer stack, placed between the control-gate and the channel, replaces the floating-gate. The concept is actually even older than that of the floating-gate cell, but it suffered for a long time from reliability problems and worse performance compared to floating gate cells. During the last years, as the floating-gate cell reaches its scaling limits, more research was done again in this area to overcome the obstacles. Advantages are, besides scalability, the relatively simple process integration (e.g. no second poly level is required) and the possibility to store 2 bits per cell by localized charge injection at source and drain. Advanced scaled layer stacks (SONOS – Silicon / Oxide / Nitride / Oxide / Silicon, and recently the introduction of high-k materials) enhance the performance of this technology[45][46]. Nanocrystal memories, where the charge is stored in small silicon dots embedded in an oxide layer, are a similar concept [47].

2.1.3. Advanced memory concepts

Driven by the known limitations of the floating-gate memory cell and by the idea of developing a “unified memory”, being a memory that combines the advantages of the

different existing kinds of memories (short access times, small cell size, non-volatile, high/unlimited endurance, low power consumption, CMOS compatibility), research activities are in various directions today to explore new memory concepts. The main directions are FeRAM, MRAM and PCM. Some of these concepts have reached a level of first products or demonstrators [48]. For CMOS integration many open questions exist, mainly because of the new materials involved. For embedded memory application those concepts will be especially interesting, which are integrated in the backend-of-line part (metallization) of the fabrication process. This reduces the interference with the fabrication of the other active devices.

2.2. Embedding flash memories

Embedding of Flash memories into standard logic took place almost from the beginning, as required materials and process steps were compatible. First embedded memory processes, derived from the stand-alone EEPROM processes, were complex and thus costly. For the integration of memory cells and the required high-voltage devices 7-9 extra mask levels were needed on top of the baseline CMOS process [21]. For applications with low-density CMOS parts also modified Flash-memory processes with embedded CMOS capabilities have been developed [21]. With the upcoming demand for small to medium density NVM within complex CMOS devices, specific processes for embedded Flash memories have finally been developed with the goal to reduce the added process complexity and cost, while achieving the required performance. The number of additionally required mask levels could be reduced to 2-5 extra mask levels, depending on the properties of the baseline flow [49]. The single-poly cell approach for low-density embedded memories works with even less.
Some selected examples of important embedded Flash memory technologies will be described in following. The aim is to give an overview of the topic and what is going on in the world. More detailed information can be found in the respective references.

2.2.1. FLOTOX

The FLOTOX cell is the original EEPROM cell with a 2-transistor structure [5][23]. It has been thoroughly investigated and modelled, e.g. [50][51], and is also used as embedded memory. It is both, written and erased by FN-tunnelling. A separate tunnelling-area (tunnel-window) outside the actual transistor part of the cell, located within the drain diffusion region, is formed for this purpose. The drawbacks of this concept are the relatively big cell area due to the select-transistor and the separate diffusion area, as well as the high number of mask steps and complex process technology for realizing an embedded EEPROM including high-voltage devices. The implementation in a modern CMOS technology, where the CMOS gate oxide is thinner than the minimum allowed tunnel oxide thickness, also needed a re-thinking

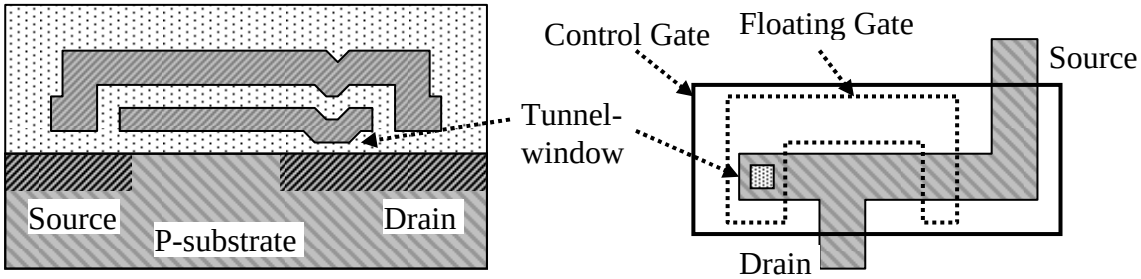
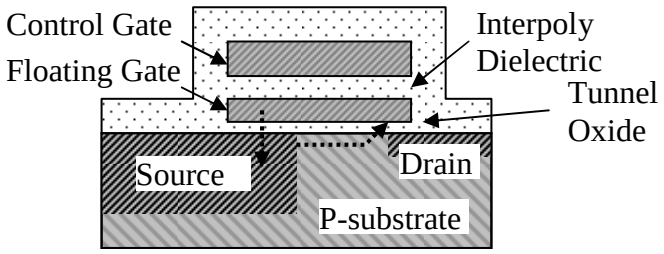


Figure 3: Schematic cross-section and layout of the FLOTOX memory cell (without select transistor)



	Source	Control-Gate	Drain
Read	GND	5V	1V
Program	GND	12V	5 – 7V
Erase	5V	-8V	Float

Typical voltage values after [5]

Figure 4: Schematic cross-section of the ETOX memory cell and operation conditions

of the concept. An example of a scaled process technology for a 0.35µm embedded EEPROM has been presented in [52]

2.2.2. ETOX™

This approach, presented for the first time by Intel in 1988 [53], is called the "industry standard" flash memory cell [5][19]. It realizes a NOR, common ground, stacked gate, one transistor flash memory. It is programmed by hot-electron injection at the drain and erased by FN tunnelling to the source. A schematic cross section is shown in Figure 4, together with the operating conditions. Typical tunnel-oxide thickness is 8-10nm. An ONO stack forms the interpoly dielectric. Due to the different operation modes for write and erase, the source and drain junctions need to be optimized separately. The source is formed with a smooth profile, having a large overlap with the floating gate. This is to allow applying high erase-voltages, avoiding junction breakdown and high BTBT currents, which are the main reliability problem in this technology. The drain junction is steep and optimized for hot electron injection efficiency. The channel implant has to be optimized to get a compromise between the diverging requirements of the source and drain junctions. The structure is critical for short-channel effects (due to the smooth and deep source junction), leading to scalability limitations. In modern technologies the erasing scheme has been changed to a channel erase scheme [54], where source, drain and substrate have the same potential and the current is distributed over the whole tunnel oxide area. High voltage transistors are formed separately by introducing an additional HV-gate-oxide and dedicated well implants. The arrangement of one-transistor cells in the common ground NOR array leads to a relatively compact cell layout as shown in Figure 5. Control gate and floating gate are formed self-aligned to each other. The technology has reached the 90nm technology-node in 2004 [55], and is thus one of the most

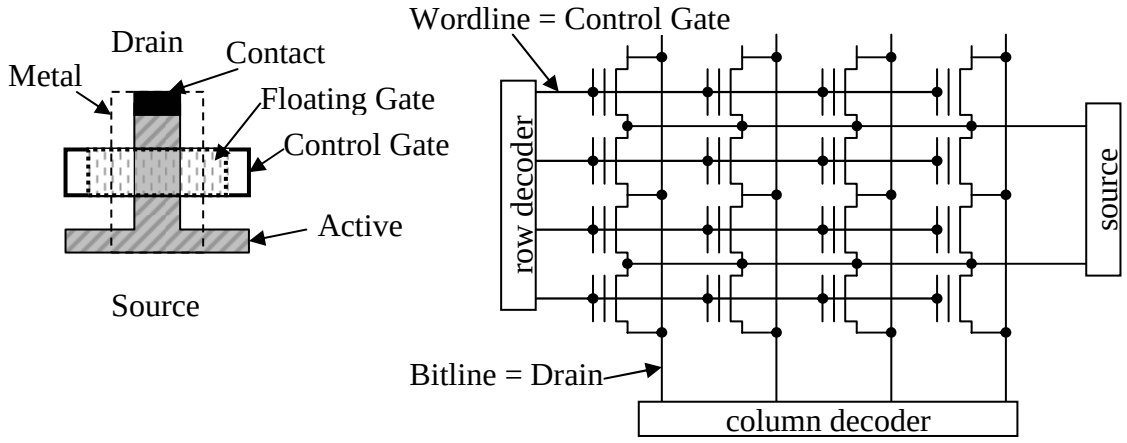
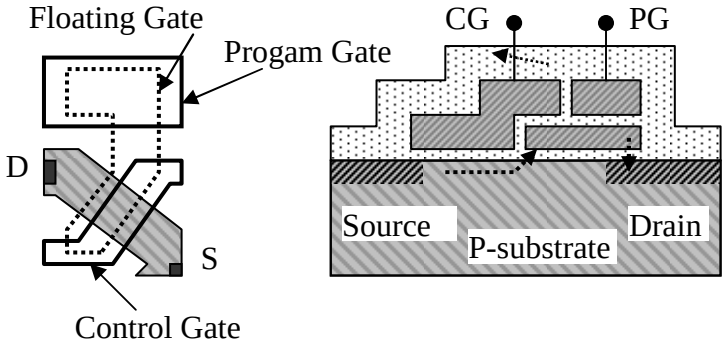


Figure 5: ETOX cell layout and cell arrangement in the common ground array



	S	D	CG	PG
W	gnd	3.3V	0.9V	9V
DE	-	3.3V	-6V	-6V
PE	-	gnd	5.5V	-5V
R	gnd	1V	1.7V	gnd

W: Write, DE: Drain Erase,
PE: Poly erase, R: Read

Figure 6: The HIMOS cell: layout, schematic cross section and operating conditions [49]

advanced NVM technologies. To use it in combination with CMOS for wireless applications, in most cases the CMOS is embedded in the flash process [54][56], but also the other way, an ETOX-like embedded flash in CMOS, has been presented [57]. Issues for embedding are the complex technology (number of mask steps) and the current consumption at the drain during programming, with drain voltages above the usual supply voltage in modern technologies (which is a problem for low-voltage and low-power applications).

2.2.3. HIMOS™

This is a real dedicated embedded flash technology. It has been invented at the IMEC in the early 90's, and was further developed and optimized until today [58][59][49][60]. The goal was to reduce the process complexity and the operating voltages, thus reducing both, the entry cost of the technology and the chip costs by reduced complexity of the periphery. This has been achieved by a new cell concept, which is a split-gate cell combined with an additional program gate. This structure significantly enhances the efficiency of the hot-electron injection, as it allows controlling the lateral and vertical electric field separately, and thus achieving the optimum injection conditions. The drain voltage could be reduced to a minimum value of 3.3V. The drain current during writing is low, because of the low voltage at the control gate (1V), leading to low power consumption. The source-side injection relieves the requirements to the drain junction as the injection point is now shifted to the source-side of the floating gate. The split-gate structure overcomes the overerase problems of 1-transistor cells (see below in this chapter, section 2.3.2. "Array operation"), and gives the possibility to have an erased V_T well below 0V, which results in high reading currents and fast read access times. Also, a low excess charge in the written state has a positive impact on reliability. Erasing can be done by FN tunnelling to the drain or by FN poly-poly erase. As the drain-erase leads to enhanced short-channel effects, similar to the ETOX source-erase, scaled technologies use the poly-poly erase option. The cell size is bigger than for ETOX, because of

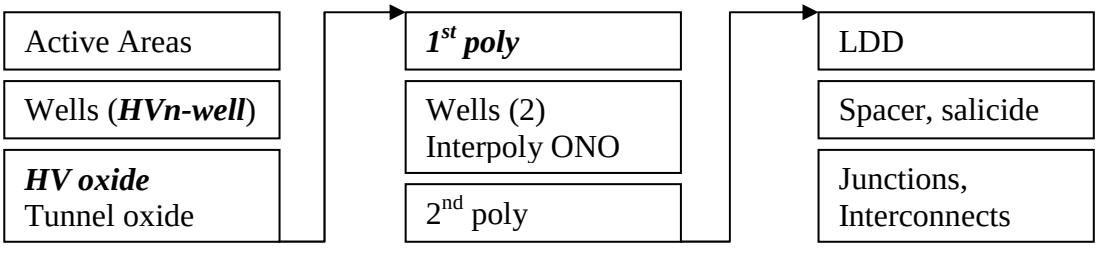
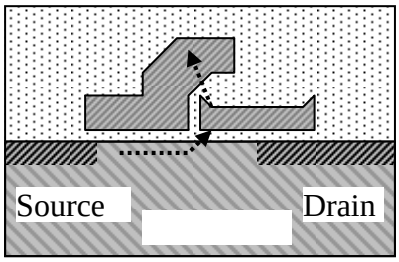


Figure 7: Process flow for HIMOS integration in CMOS (after [60]). ***Bold/italic*** are additional masks needed



	S	D	CG
W	-5µA	9V	1.7V
E	gnd	gnd	12V
R	1V	gnd	1.7V

W: Write, E: Erase, R: Read

Figure 8: The Superflash cell: schematic cross section and operating conditions [64]

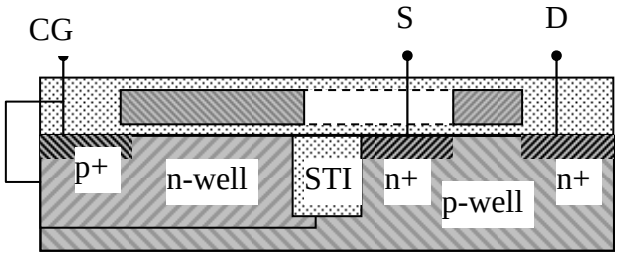
the additional program gate. The control gate and floating gate are not self-aligned. The cell-layout, together with a schematic cross section and a table with operating conditions are shown in Figure 6. The arrangement in a "quasi virtual ground" array is described in [61], together with 1Mb circuit results. The CMOS integration only needs 2-5 masks, depending on the baseline process. An example is given in [60]. The process flow is shown in Figure 7. Here, 3 additional masks are needed, one for the HV-n-well, one for the tunnel oxide/high-voltage-oxide dual oxide process, and one for patterning the floating gate. The required medium voltage transistors for the 3.3V operation are already available in the process and are not counted. The first poly layer is here used for the floating gate and the high voltage transistors. The second poly forms the control gate, the program gate, the medium voltage transistors and the CMOS transistors. The process has been presented for the 90nm technology node [60], and concepts for 45nm have already been discussed.

2.2.4. Superflash

This is also a pure embedded flash concept. It was developed by SST [62], and is used by different companies (including TSMC) as CMOS embedded flash option. It is a NOR, common ground, split-gate, double poly silicon technology. It is written by source-side hot electron injection and erased by poly-poly FN tunnelling [63][64]. For erasing, an electric field enhancement is achieved geometrically by forming a tunnel-injector tip. This is achieved by using effects of corner shaping at oxidation of patterned poly silicon. A high drain/floating gate coupling is achieved by a dedicated implant forming a big drain overlap. This is needed for controlling the floating gate potential during erasing. Not much is published about the CMOS integration, or the high-voltage concept. The concept provides the advantages of source-side injection and the split-gate structure described above.

2.2.5. Single poly cells

In this technology a separate active area forms the control-gate. The integration is very simple (e.g. no second poly silicon level is required), at the cost of a bigger cell size, which makes it the technology of choice if only small memory density is needed. Fig. 9 shows the schematic cross section of one possible realization presented by IBM in 1994 [65], other solutions are presented in [66][67]. In the IBM cell, one part of the floating gate overlaps an n-well area, which acts like the control gate of a stacked gate cell. It can only be positively biased, because of the pn-junction to the substrate. Single poly cells can be programmed either by hot electron injection or by FN tunnelling. The process integration of the flash cell itself only needs one additional process module, the formation of an additional oxide layer, forming the tunnel oxide. This usually requires one lithographic mask level. Additional processing is usually necessary to make the required on-chip high-voltage handling possible.



	S	D	CG
W	0V	5.5V	6.5V
E	float	7.0V	0V
R	0V	1.5V	3V

W: Write, E: Erase, R: Read

Figure 9: Example of a single poly cell. Schematic cross section and typical operation conditions

2.3. The chosen memory concept

The embedded flash memory concept that has been chosen for integration into the RF-BiCMOS process is a uniform channel FN programmed, channel FN erased, stacked gate, double poly silicon, triple well, dual voltage, NOR flash memory.
The idea to use FN tunnelling for both, writing and erasing, in an embedded flash memory, especially for portable applications in the mobile communication field, has already been proposed in 1993 [68]. It describes a split gate cell structure, written by FN tunnelling to the drain and erased by FN tunnelling from the channel (states are switched here compared to e.g. hot-electron-written ETOX). In 1997 a similar concept has been proposed for a 1-transistor stacked gate cell [69], also applying drain-side FN tunnelling for writing and channel FN erasing. Uniform channel FN writing and erasing, thus achieving a completely symmetrical cell, has been published in 1999 [70], proposing a 2-transistor cell, arranged in a common ground NOR array, embedded in an 0.25µm CMOS process. In 2000 two publications, [71] and [72], at the same time proposed a uniform channel FN writing and erasing, 0.25µm, 1-transistor embedded flash technology using a triple well that allows splitting the high voltage between control-gate and well. In [72] the STI is used to separate the wells of each bitline, to allow bit-by-bit control of writing and erasing. In 2002 an uniform channel FN written, channel FN erased, stacked gate, double poly silicon, 2-transistor, triple well, dual voltage, common ground NOR technology as low-voltage, low power embedded flash technology, implemented in a 0.18µm CMOS process has been proposed [73]. The publication describes it as the embedded flash technology of choice due to its robustness, reliability and cost effectiveness. It should be noted that the select transistor in the 2-transistor cell does not allow a full feature EEPROM usage of the cell (= bit-wise erasing) in the case of uniform channel programming. It is merely used during reading, where it leads to immunity against the overerase problem and to more freedom in the choice of the V_T-window used for memory operation.
The advantages of using uniform channel FN tunnelling for both, writing and erasing are the following:

- Uniform channel FN-tunnelling has the highest programming efficiency. Virtually all of the current flowing in the selected cell is the FG charging current. Only junction leakage and, in array operation, the leakage of the unselected cells reduce the efficiency. This results in a low power consumption during programming.
- FN tunnelling requires high voltages, but these can easily be generated on-chip by charge-pumps, also from low supply voltages, due to the low power consumption required for programming. This also results in a small area consumption of the charge-pump, compared to other flash memory solutions. The low voltage operation capability is an advantage compared to concepts using hot electron injection, especially for battery-driven systems.

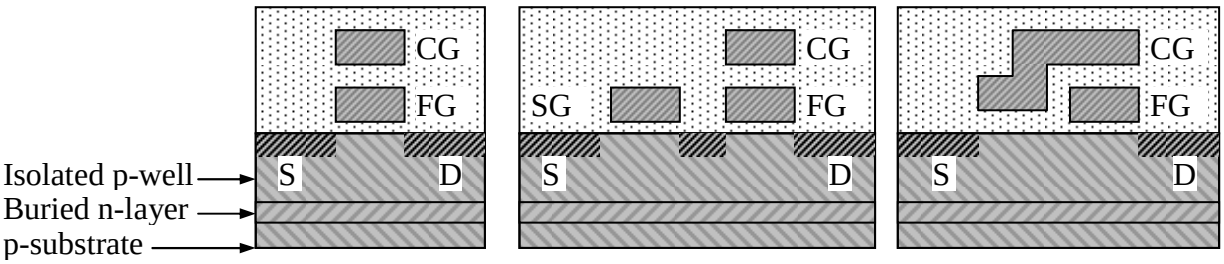


Figure 10: Schematic cross sections of the 1-transistor cell, the 2-transistor cell and the split-gate cells

- The low power consumption makes highly parallel writing of cells possible, which in principle allows reducing the average writing time to a fraction of the actual cell writing time.
- For uniform channel FN programming no separate optimization of the source or drain junction is necessary. This results in a reduced process complexity and thus easier and cheaper process integration. It also favours a modular integration scheme, which has an advantage because of reusability of libraries, thus saving technology entry costs.
- No enhanced short-channel effects, as observed in source or drain erasing schemes, are present.
- Channel erasing has good reliability properties, as one of the main issues, the BTBT at biased junctions, is not present
- The implementation of a triple well technology relieves the high voltage requirements on the devices involved, as the total programming voltage can be split between the control gate and the well/channel.

The low power consumption, low supply voltage operation and the low cost process technology match the requirements of portable, mobile communication applications.
The realization of a 1-transistor memory cell leads to a small cell size, while a 2-transistor cell or split-gate cell have the advantages of

- overerase immunity
- the possibility of low voltage reading without word line boosting, as required in other technologies. This is because the erased V_T can be low, even below 0V in this case, leading to high reading currents at low control gate voltages.
- a reduced complexity, and thus reduced area of the peripheral circuitry [73].

The developed process technology allows the formation of all three types of memory cells, 1-transistor, 2-transistor and split-gate cells. Schematic cross sections are shown in Figure 10.
A NOR configuration of stacked gate cells allows realizing medium density memories (several Mbit) with fast random access times, as required by the targeted applications.

2.3.1. Cell operation

The applied voltages in write, erase and read mode are shown in Figure 11. The principal functionality is the same for all of the three cell types, thus only the 1-transistor case is shown. All voltages are given with respect to the substrate. For **writing**, a high positive voltage $V_{CG,w}$ is applied to the control gate and a high negative voltage $-V_{SDW,w}$ is applied to the source, drain and isolated p-well simultaneously. Alternatively either source or drain can be floating. The buried n-layer is kept at 0V in this case. Under such bias conditions, a high positive voltage is coupled to the floating gate, depending on its charging state even higher than $V_{CG,w}$ for erased cells (V_T<0V). An inversion layer is formed under the oxide and electrons start tunnelling from the channel to the floating gate in the whole tunnel oxide area. For **erasing**, a high negative voltage $-V_{CG,e}$ is applied to the control gate, and a high positive voltage $V_{SDW,e}$ is applied to the source, drain and isolated p-well simultaneously. The buried n-layer is on the

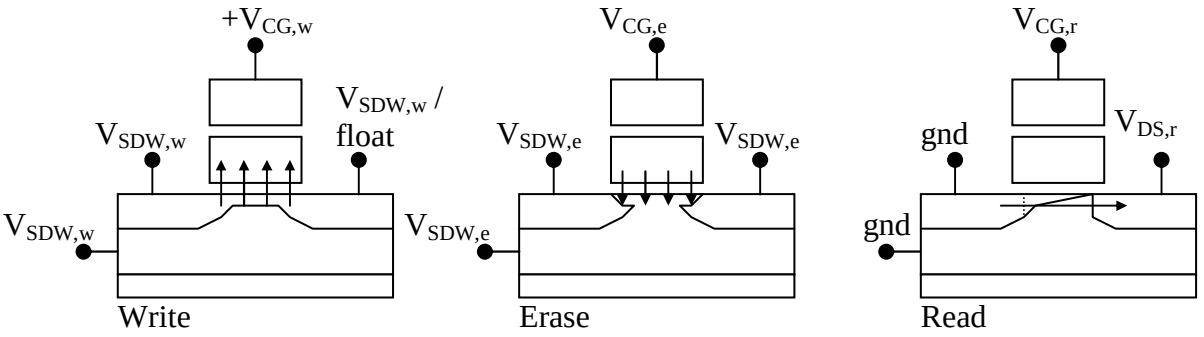
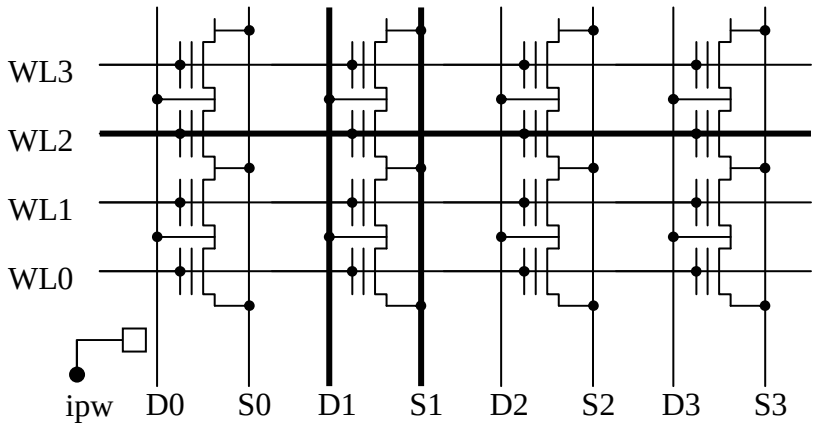


Figure 11: Operating conditions for channel FN programming; typical values: $V_{CG,w} = -V_{CG,e} = -V_{SDW,w} = V_{SDW,e} = 6V$; $V_{CG,r} = 0V$; $V_{DS,r} = 1.5V$

same potential as the isolated well in this case. A high negative voltage is coupled to the floating gate, again possibly exceeding $-V_{CG,e}$, now in the case of written cells. The channel area is in accumulation, while the overlap areas of the source and drain with the floating gate are inverted. This again leads to a uniform tunnelling current through the tunnel oxide. A more detailed analysis of the programming kinetics is given in chapter 5 and appendix A. For **reading**, the control gate is biased with a voltage $V_{CG,r}$ between the written and the erased V_T state, see also Figure 2. It must be as much as possible above the erased $V_{T,e}$ to get a high reading current, but keep a safe distance to the written $V_{T,w}$, so that the current of a written cell is still low and its state can be correctly determined. The drain voltage should be as high as possible for high reading currents. The limit is given by the read disturb effect, which is an unintentional change of the floating-gate charging state during reading (soft-writing). It is typically specified that the cell's state should be safely kept for 10 years of continuous reading.

2.3.2. Array operation

The arrangement of the cells in an array and the operation conditions for writing and erasing are shown in Figure 12. Unlike usual common ground NOR, for each bitline the sources and drains have to be connected separately, leading to 2 metal lines per bitline. The reason for this becomes clear after a closer look to the operating conditions.



Write conditions:	
WL, sel.:	$V_{CG,w}$
D, sel.:	$V_{SDW,w}$
S, sel.:	$V_{SDW,w}$
WL, non-sel.:	gnd
D, non-sel.:	gnd
S, non-sel.:	gnd
ipw:	$V_{SDW,w}$

Erase conditions:	all WL:	$V_{CG,e}$
	all D, all S, ipw.:	$V_{SDW,e}$

Figure 12: NOR array configuration and operating conditions for FN programmed memory cells. Thick lines mark selected lines.

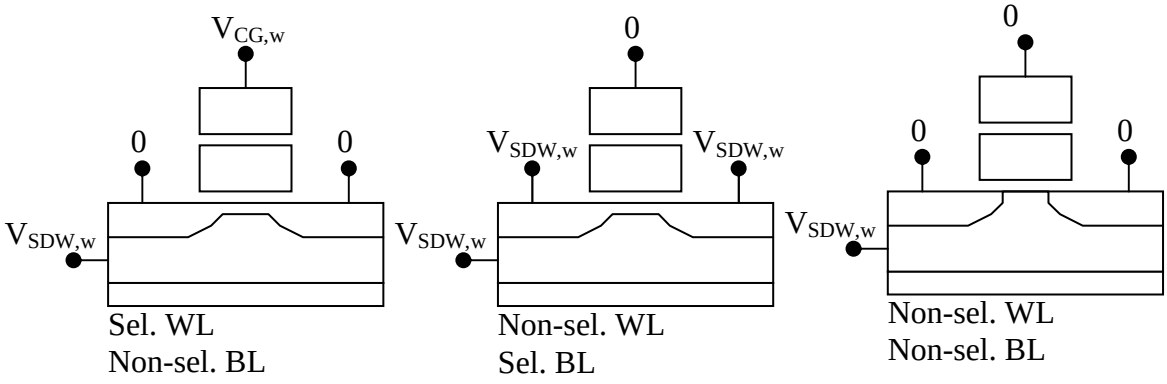


Figure 13: Inhibit conditions for different cells in the array.

For **writing**, the wordline of the selected bit (here WL2) is set to a high potential and the drain and source lines of the corresponding bitline (here D1 and S1) are set, together with the isolated p-well of the whole array, to a high negative potential. This leads to the writing conditions described above for the selected cell. To prevent the other cells of WL2 from being written, their source and drain lines are set to 0V, and to prevent the other cells of bitline 1 from being written, their wordlines are set to 0V. These inhibit conditions leave also the unselected cells biased with some voltages. Three cases can be distinguished as shown in Figure 13, (1) cells in the selected wordline, in a non-selected bitline, (2) cells in a non-selected wordline, but in the selected bitline, and (3) cells in a non-selected wordline and a non-selected bitline. These voltages lead to the so-called write disturb effect, which is discussed in chapter 5. The different potentials of the selected and non-selected source lines do not allow the usual common-ground configuration. One exception is the 2-transistor cell, if the possibility is used to leave the source floating (which is possible as there is an inverted channel under the gate). If the select-transistor is able to keep the voltage difference between a selected and unselected bitline without too much leakage, a common ground configuration is theoretically possible (and used for example in [70]).
For **erasing,** simply all wordlines are biased with a high positive potential and all bitlines and the isolated p-well are on a high negative potential.
For **reading,** the selected wordline is biased with $V_{CG,r}$, the selected drain line with $V_{DS,r}$ and the selected source line is grounded. The unselected source and drain lines are all grounded, as well as the isolated p-well. For the unselected wordlines the conditions are different for the different cell types. (1) For one transistor cells the unselected wordlines must be set to a voltage V_{off} , which must be well below the $V_{T,e}$ of the erased cell, to prevent a current in non-selected cells of the selected bitline. Consequently, for the case of a $V_{T,e}< 0V$ a negative voltage must be provided within the memory circuit for reading. At this point also the overerase problem of one-transistor cell-arrays becomes clear. If only one cell of a bitline has a $V_{T,e}$ below V_{off}, its drain current is sensed when reading any other cell in the same bitline, leading to wrong reading results. (2) For the split-gate cell the V_T of erased cells can't be below the V_T of the select-transistor-part of the cell, which is above 0V, so the unselected wordlines are switched off by the select-transistor part of the cell when applying 0V to the control gate. (3) In a 2-transistor cell the current in unselected cells is switched off by the select transistor, also in the case of a negative $V_{T,e}$ of the cell. The control-gates of the unselected cells as well as the select-gates are set to 0V.
Overerased cells do not disturb reading in split-gate or two transistor memories, making these solutions more robust. In general, the 2-transistor cell gives most freedom in choosing the V_T-window of all three cell-types.

Chapter 3

The Flash / BiCMOS Process Integration Scheme

This chapter presents the developed process integration concept. First, the baseline BiCMOS process is introduced followed by the description of the integration scheme for the flash memory devices, including the different options for memory cells themselves as well as the high-voltage MOS transistors. After a presentation of the overall process flow, the processing of the individual devices is explained more in detail.
For easier understanding of layouts and schematic cross section views please refer to the respective legends given in the appendix of this dissertation (p.111).

3.1 The baseline BiCMOS process

The baseline process is a 0.25µm BiCMOS process, consisting of an industry standard CMOS part, combined with a low-cost, high performance SiGe:C HBT module and different passive devices for RF circuit design [9]. Fig. 14 shows a schematic cross section. The main features are: STI isolation; a triple well approach (n-well, p-well and an isolated p-well formed by an implanted buried n-layer); MOS transistors for digital applications (V_{DD} = 2.5V), including the possibility of an isolated NMOS for improved signal isolation; a Co salicide process for reduced poly silicon resistance; low-cost integrated SiGe:C HBTs with epi-free, implanted collectors, allowing on the same chip 3 different HBTs (f_T / BV_{CEO} of 28GHz / 7.5V, 52GHz / 3.8V, 75GHz / 2.4V) by layout variation; poly silicon resistors of different sheet resistances; high-Q MOS varactors; a 4 level Al-interconnect system, including a 2µm thick top metal layer for high-Q inductor fabrication; and a 1fF / μm^2 MIM capacitor.

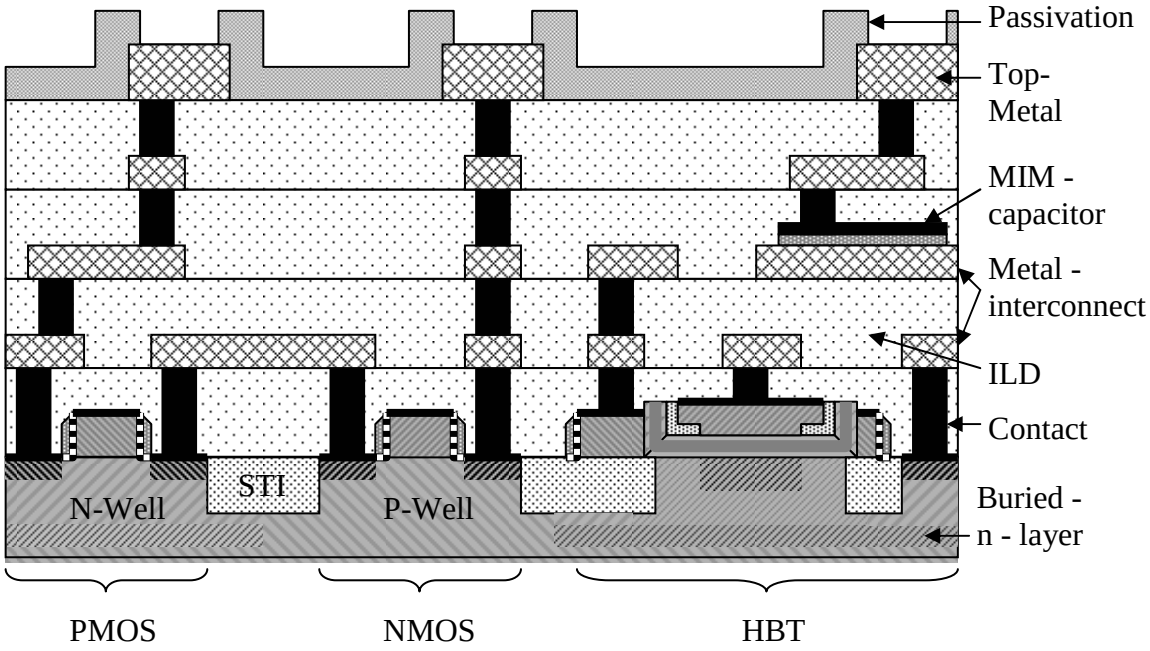


Figure 14: Cross section view of different devices of the BiCMOS technology

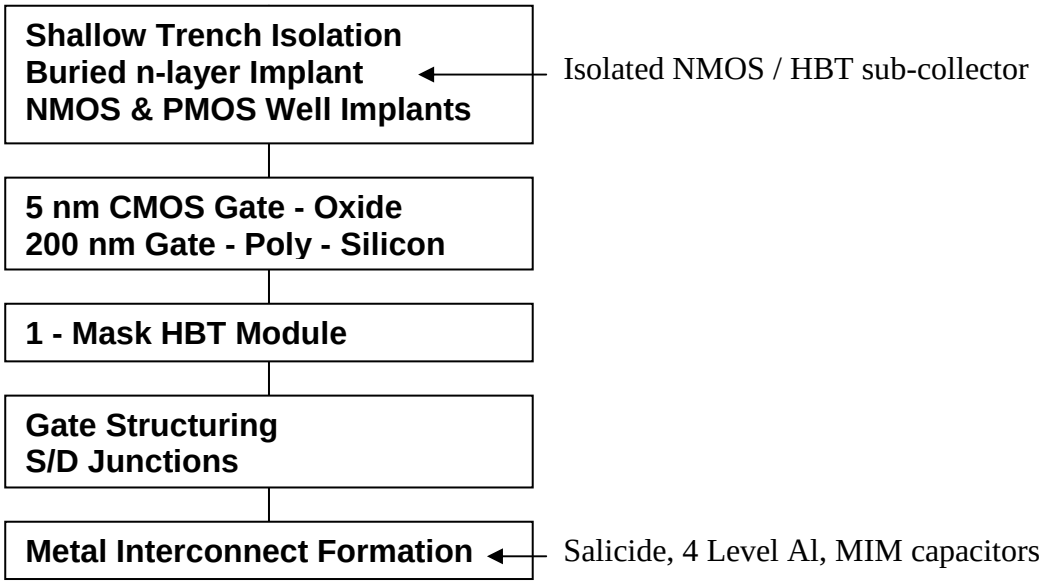
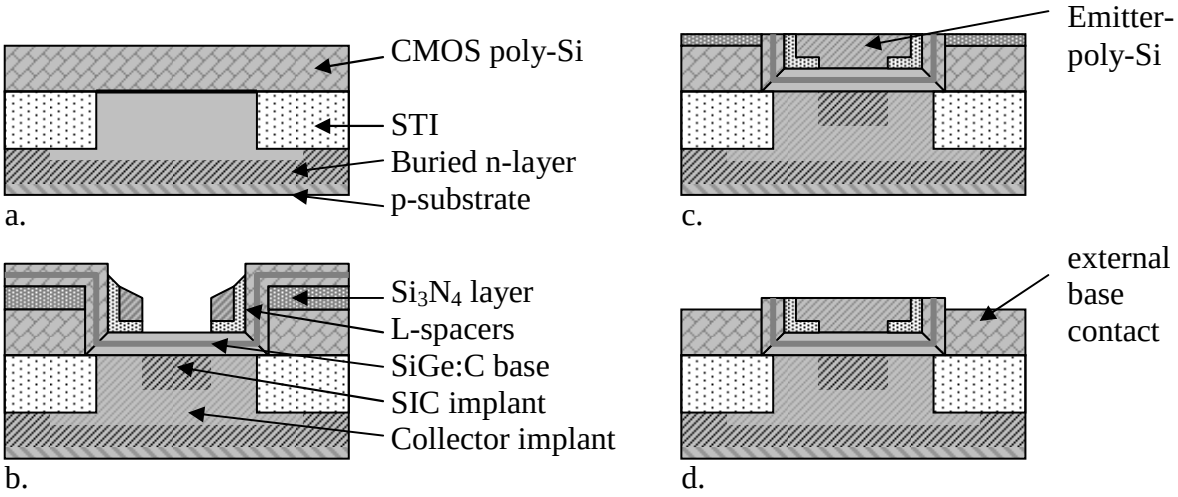


Figure 15: Schematic BiCMOS process flow without embedded flash memory

The principle process flow can be seen in Fig. 15. It starts with the formation of the shallow trench isolation by trench-RIE, oxide filling and CMP planarization. In a next step the different wells are implanted in the respective areas. A 5nm gate oxide is thermally grown, and a 200nm poly silicon layer is deposited, which form the gate stack of the MOS transistors. At this stage of the process the HBT is fabricated, requiring only one lithographical step. The HBT process is described below more in detail. After the HBT module the CMOS gates are formed. First the poly silicon is structured by RIE. After a poly silicon re-oxidation and the formation of sidewall spacers, the source and drain junctions are implanted. The silicon surfaces are silicided in a Co-salicidation process. Finally the interconnect system is produced by subsequent deposition of the interlayer dielectrics, CMP planarization, contact hole etching, contact filling, metal deposition and metal structuring. On top of the last metal layer a passivation layer stack is deposited and opened again by RIE on the contact pads. The complete process flow requires 19 lithographic mask steps. The CMOS part is a typical process sequence for industrial fabrication in this technology node, while the integration of the HBT is an advanced process for low cost, high-performance BiCMOS applications.
The description of the HBT module of [9] will be shortly repeated in following. The basic steps of the 1-mask HBT module are presented in Fig. 16. Fig. 16a shows the situation after the CMOS poly silicon deposition. The poly silicon layer fully covers the HBT area; STI defines the base area of the HBT and separates it from the collector contact area (not drawn in Fig. 16), which is connected with the intrinsic collector area under the base via the buried n-layer. The HBT fabrication starts with the deposition of an Si_3N_4 protection layer. The HBT mask is used to remove the Si_3N_4 and the gate poly silicon from the HBT regions by RIE, and to carry out a chain of P-implants, which form and adjust the collector doping. After removing the gate oxide from the base area by wet etching, the SiGe:C base and a low-doped poly-Si emitter layer are successively deposited by CVD. L-shaped inside spacers are formed and a SIC implantation is carried out (Fig. 16b). An in-situ n-doped poly-Si emitter layer is deposited. All silicon material is then removed from the Si_3N_4 layer by CMP, thus isolating the emitter from the external base (Fig. 16c). After removing the Si_3N_4 from the CMOS poly-Si by wet etching, the CMOS device fabrication is continued as described above. The state at the end of the HBT module is shown in Fig.16d. CMOS gate etching and PMOS S/D implants are used to complete the HBTs external base contact area later in the process.



a. After CMOS gate oxide growth and CMOS poly deposition
b. After Si_3N_4 deposition, bipolar window etching, collector implant, SiGe:C epitaxy, inside L-spacer formation and SIC implant
c. After emitter poly-Si deposition and CMP
d. After residual Si_3N_4 removal

Figure 16: 1-mask HBT module

The HBT fabrication leaves the regions outside the HBT in the same state as they have been before the module and adds only little topography, so the following CMOS steps can be performed without modification. The position of the HBT-module in the process flow ("HBT before gate" – integration scheme) minimizes the impact of the thermal steps during HBT formation on the MOS devices. This integration scheme is possible due to the C incorporation in the base layer, which leads to a reduced B out-diffusion and thus prevents degradation of HBT parameters by CMOS thermal steps, e.g. poly re-oxidation and S/D RTA.

3.2 Flash Memory Integration

In following, the process steps for realizing the 4-mask-layer, modular integration of the flash memory will be described. After presenting the integration scheme, the process flow for the different possible flash cells and for the high-voltage MOS transistors is explained in detail.

3.2.1. Integration scheme

Fig. 17 shows the integration scheme for the flash memory module. The additional processing can be separated into 4 blocks, corresponding to the 4 additional lithographic mask levels needed to include the flash memory fabrication into the BiCMOS process.
The first block is carried out before the CMOS gate oxide growth, and is the formation of the tunnel oxide by a standard dual oxide process. This process consists of a masked removal of the thick oxide (tunnel oxide) outside of the areas of the later tunnel oxide by wet etching, before the thin oxide (CMOS gate oxide) is grown. Please note that, if the baseline process already has a dual gate oxide option with a suitable thick oxide thickness, the integration of the flash memory only requires 3 additional mask steps, as this block is not needed.
The second block is a masked chain of ion implantations for doping the floating gate and the flash p-well, which is used for both, the flash-cells and the high-voltage transistors. This is done before the HBT module is carried out, thus using the HBTs thermal steps for annealing the implantation damage.
The main structuring steps are carried out after the HBT module and before the CMOS gate formation. In the third block a masked RIE of the first poly silicon layer is done. This etching separates the floating gates belonging to one wordline from each other. The same etching step

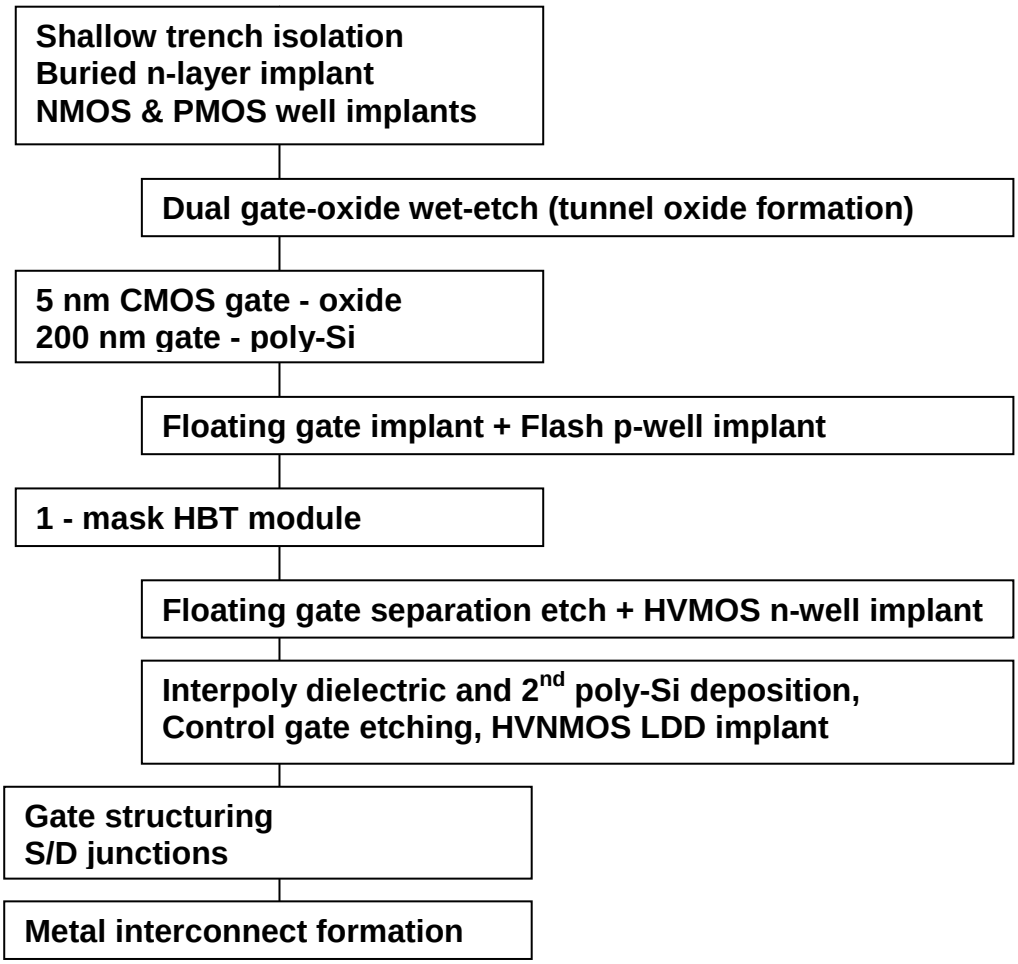


Figure 17: Embedded flash memory integration scheme

is used to remove the first poly silicon layer from the HVMOS areas. Before removing the resist, the ion implantation of the n-well for the HVPMOS is carried out.
In the last block the interpoly dielectric layer and the control gate poly silicon layer are deposited and structured. Using the same mask, an LDD implant for the HVNMOS is done.

The process continues with the CMOS processing steps, which also do the self-aligned floating gate etching and the formation of the source and drain junctions of the memory cells and the high-voltage transistors, thus saving extra processing.
By the position of flash memory module in the overall process flow the impact on the CMOS devices is minimized, as the integration is completed before the CMOS gate formation. The integration after the HBT module minimizes the interference of HBT processing with flash memory processing, which could be a problem because of the flash cells' relatively high topology. This integration scheme takes benefit from of the above mentioned stability of the SiGe:C HBT to degradation to added thermal budget. A low cost process is achieved by using the same mask layers and process steps for different purposes, and by using CMOS process steps also for flash memory fabrication. Especially the fabrication of the HVMOS transistors does not require any additional mask layer.

3.2.2. Fabrication of the floating gate memory transistor

Fig. 18 shows the schematic layout fragment of a 1-transistor memory array. The following Figures 19 - 22 show cross section views at different stages of the process. The cuts are along line A in bitline direction, crossing the wordline; line B in wordline direction in the area of the control gate; and line C in wordline direction, but outside the CG area. In both directions (x

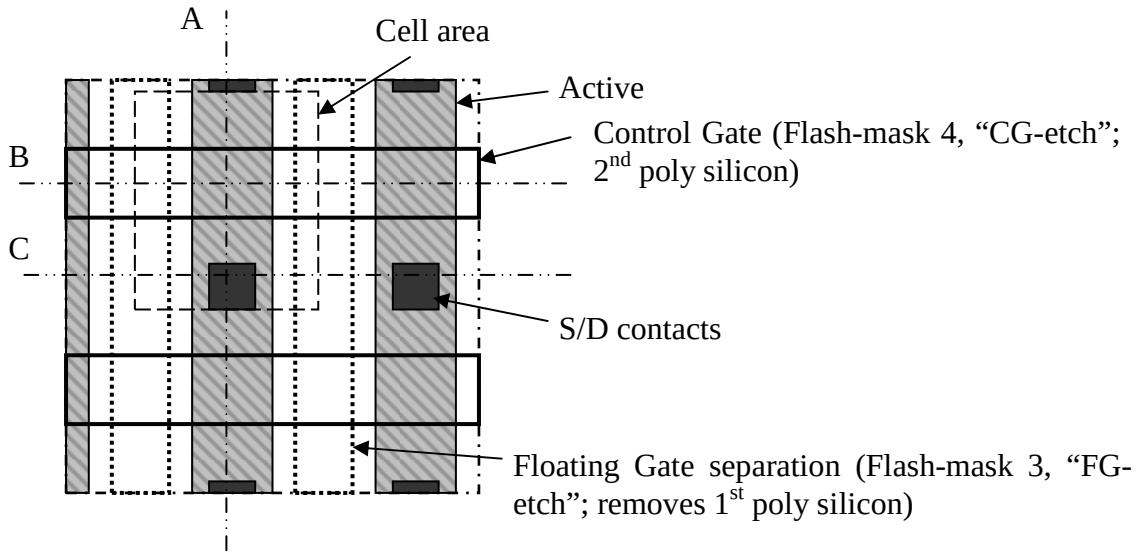
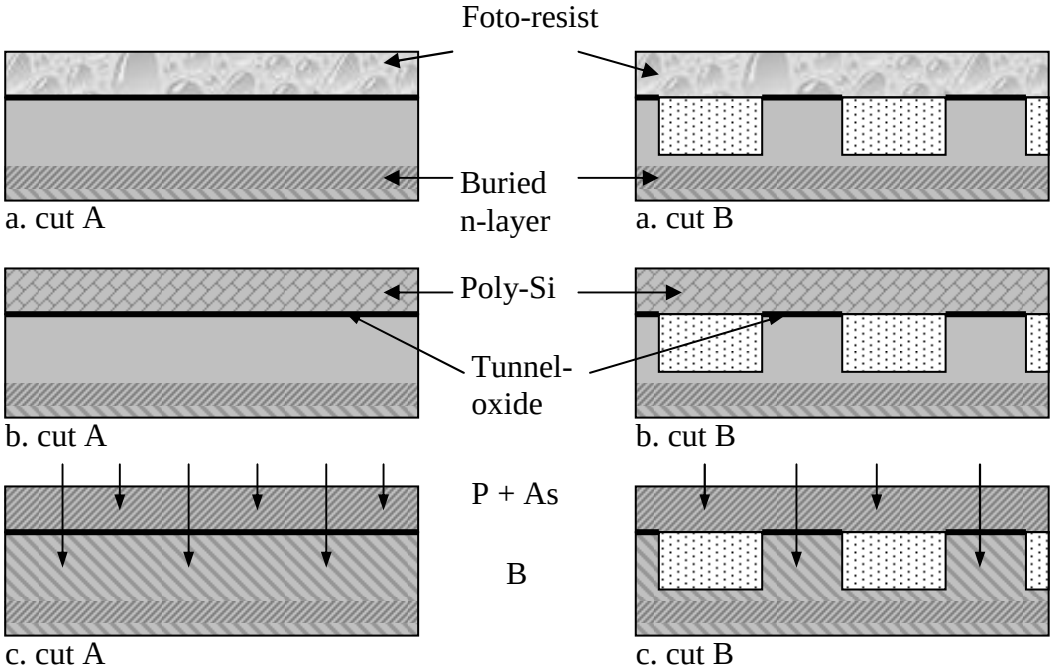


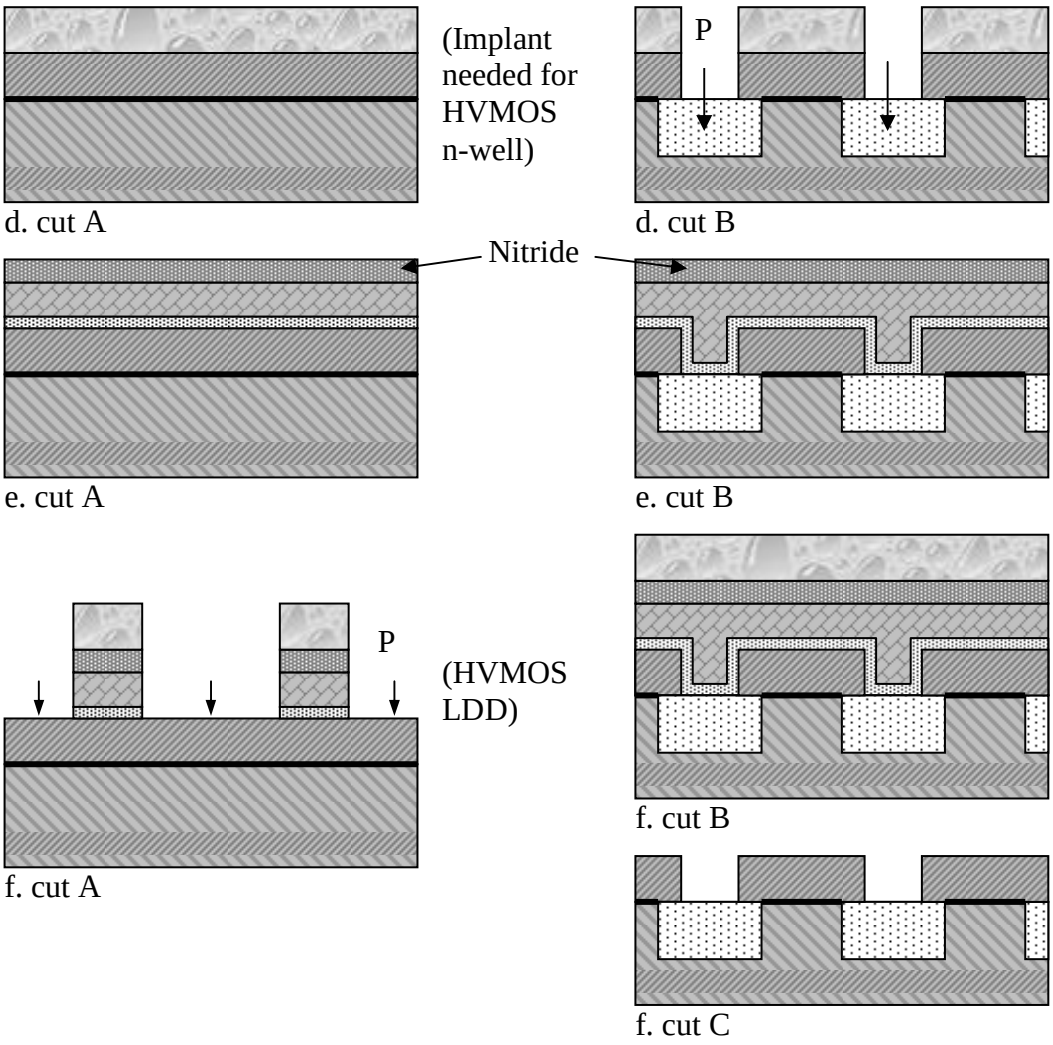
Figure 18: Schematic layout fragment of a 1-transistor cell array

and y) the chosen fragment of the array covers a distance of about 2 memory cells. Not shown are the metal interconnects, only the placement of the contacts is drawn in Fig. 18. The metal connection can, for example, be done in a way that a metal-1 line and a metal-2 line are drawn in parallel, one for connecting the sources, the other connecting the drains of the memory cells belonging to one bitline. The first two masks of the flash module are not drawn in Fig. 18, as the first mask (DGT) completely covers the flash cells, while the second mask (FImp) is open in the whole area of the flash cells.
Figure 19 shows the application of the first two masks of the embedded flash process for cut



a. Flash-mask 1 (DGT): dual gate-oxide wet-etch (resist open outside tunnel oxide areas)
b. CMOS gate-oxide growth and gate poly-Si deposition (1st poly silicon layer)
c. Flash-mask 2 (FImp): Floating gate and flash-p-well implant (resist covers areas outside flash cell)

Figure 19: Process flow for flash memory cell: flash-mask 1 (DGT) and flash-mask 2 (FImp); cut lines A and B according to Fig. 18

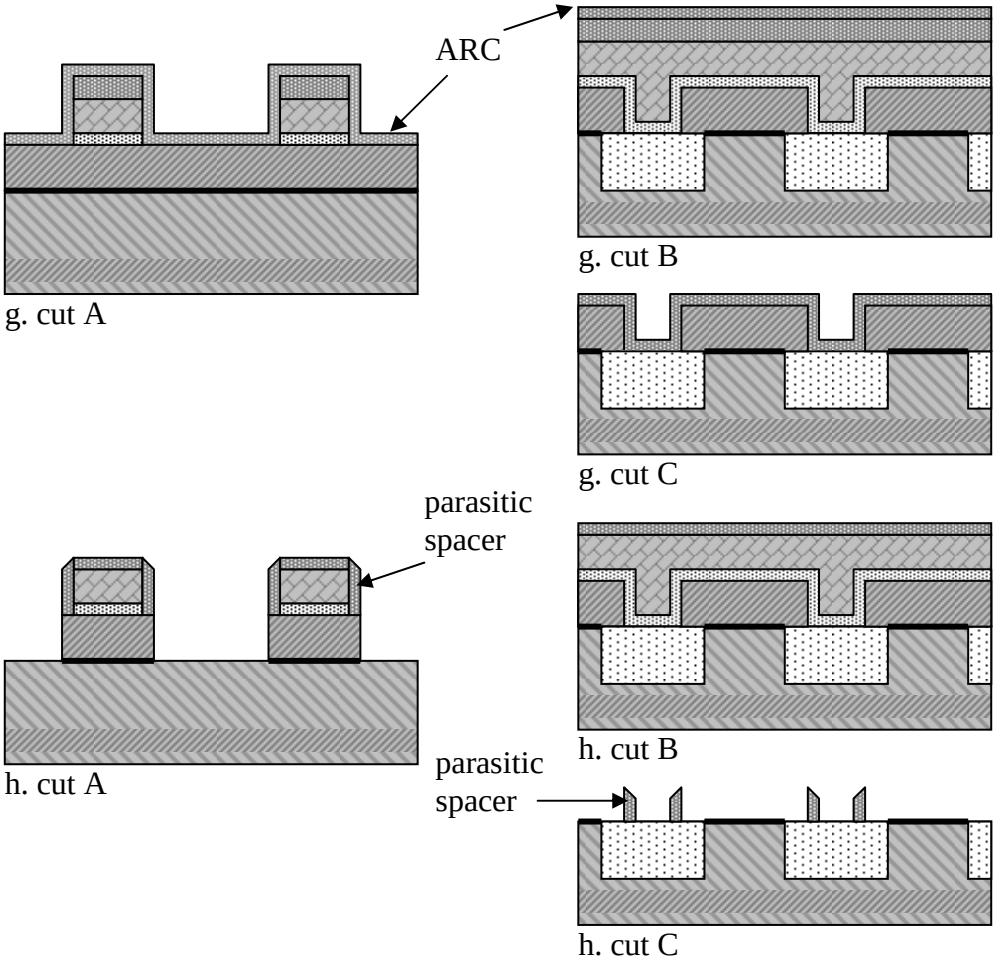


d. Flash-mask 3 (FG-etch): floating gate separation etching
e. Interpoly dielectric, second poly silicon and nitride deposition
f. Flash-mask 4 (CG-etch): control gate etching

Figure 20: Process flow for flash memory cell: flash-mask 3 (FG-etch) and flash-mask 4 (CG-etch); cut lines A, B and C according to Fig. 18

lines A and B (C is the same as B at this stage). The STI has been formed and the well implants have been carried out. On the active areas an oxide has been grown that will become the tunnel oxide later on. The first resist mask (DGT) covers the active areas at the places where the tunnel oxide is formed, while all oxide is removed from the silicon surfaces of the other active regions in an HF-wet etching step. After resist removal and a cleaning step the CMOS gate oxide is grown. The tunnel oxide looses a bit of thickness during the cleaning and reaches its final thickness during the thermal gate oxidation. The first (CMOS) poly silicon layer is subsequently deposited.

The second resist mask of the flash module (FImp) covers all areas outside the flash cell array and protects them from the following ion implantations. A first ion implantation is done for highly n-doping the first poly silicon layer in the areas of the floating gate. As in this technology the floating gate is the same poly silicon level as the MOS transistor gates, this has to be a masked process, unlike in many other embedded flash technologies, where the MOS transistors are made of the second poly silicon layer. The same resist mask is used to do the well and V_T doping for the flash cells (flash p-well). The doping is done by a chain of B implantations of different doses and energies, resulting in a retrograde profile that provides a



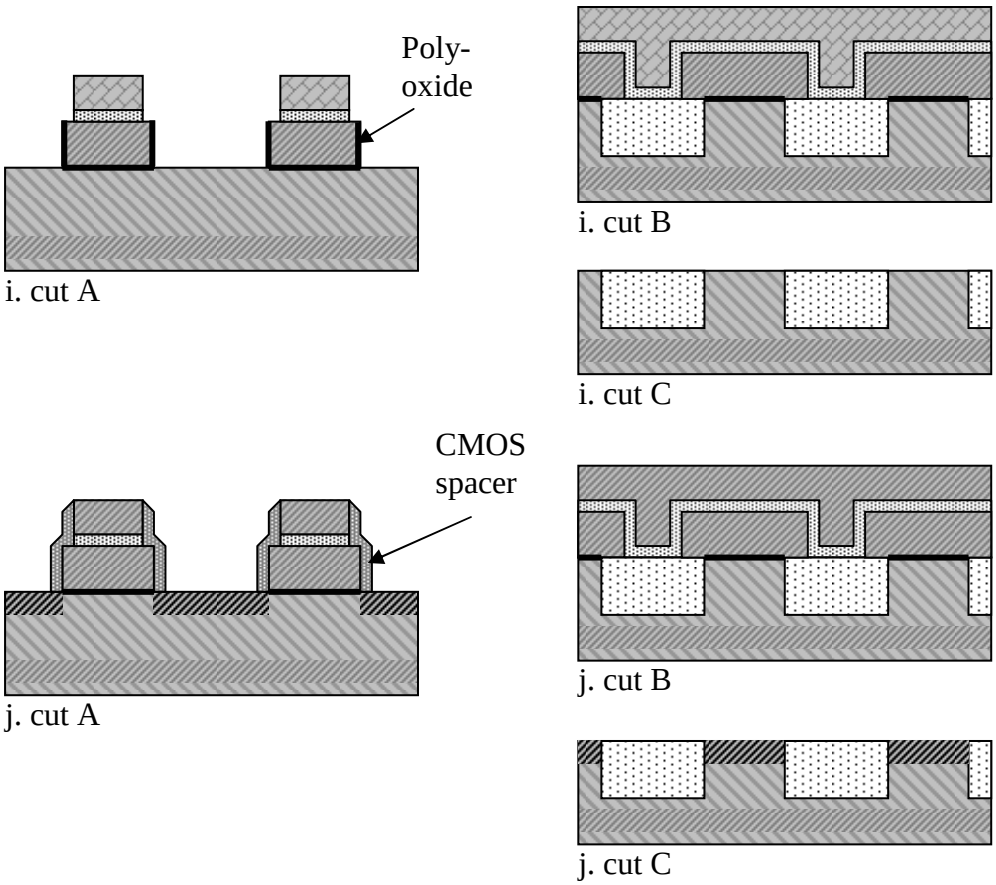
g. CMOS ARC deposition
h. CMOS gate etch, self-aligned floating gate structuring

Figure 21: Process flow for flash memory cell: CMOS steps, gate etching; cut lines A, B and C according to Fig. 18

compromise between the requirements of a low V_T and punch-through immunity. The B implants are carried out through the first poly silicon layer, with energies ranging from 90keV to 360keV. This process block ends with the removal of the resist mask, leaving the wafer ready for the HBT module.

After completing the HBT module, which has no impact on the areas of the flash memory, besides thermal implant annealing, the flash fabrication is continued with applying the third flash mask (FG-etch). This mask is used to etch narrow slits in the first poly silicon layer to separate the floating gates that will later be covered by the same control gate (Fig. 20d, cutB). The etching can not be done using the CMOS gate etching mask, which is also used for structuring the 1st poly silicon layer, as the deposition of the interpoly dielectric and the second poly silicon is still to come. Some conflicts would appear later in the process (parasitic spacers etc.) if the MOS gates were already structured then. The distance of the slits defines the width of the floating gate (the length is defined later by self-aligned etching with the control gate). As indicated by the arrows in Fig.20d, cut B, an implantation is carried out after the etching. This is needed for the formation of the HVMOS transistors, and has no impact on the flash cells, as here it only goes into the oxide of the STI.

As a next step the interpoly dielectric is formed, using a thin thermal oxidation step, followed by a thermal, low pressure CVD oxide deposition and a thermal oxide-densification step. On top of the dielectric the second poly silicon layer is deposited. By using a conformal



i. Nitride wet-etch (original step from BiCMOS flow)
j. CMOS spacer formation; source and drain junction implants

Figure 22: Process flow for flash memory cell: CMOS steps, S/D junctions; cut lines A, B and C according to Fig. 18

deposition, the narrow slits between the floating gates are filled with poly silicon during this step, leaving a more or less planar surface. Finally, a silicon nitride layer is deposited. This layer is needed later in the process as hardmask for self-aligned floating gate etching during the CMOS gate RIE (Fig. 20e).
The fourth flash mask (CG-etch) is used to structure the second poly silicon layer. The RIE process consists of different steps for etching the silicon nitride hardmask and the poly silicon layer. By a long overetch, the poly silicon is removed out of the slits outside the control gate areas (Fig. 20f, cut C). Moreover the CMOS and HBT areas are cleared from the 2nd poly silicon during this step. The interpoly oxide is etched partly by RIE and partly by a subsequent HF cleaning step, in order to remove it also from the sidewalls in the FG slits. At this point also a shallow P ion implantation is carried out, as indicated by the arrows in Fig.20f, cut A. This is required for the HVMOS fabrication to form LDD extensions. The last block of the flash module ends with removing the resists mask. The regions outside the flash memory are now in the same state again as after the gate poly silicon deposition, which was before the second block of the flash memory integration (not counting the influence of the thermal steps and the HVMOS LDD implant; for a detailed discussion of the impact of the flash integration on the CMOS process see chapter 4.3.). The flash memory transistor is not finished yet. This happens during the following processing steps of the CMOS core.
The CMOS gate processing starts with the deposition of an antireflective coating (ARC) (Fig.21g). The ARC is a silicon rich silicon nitride layer with adjusted optical properties. In the following gate RIE step, the flash cell is not covered by any resist. The formerly deposited

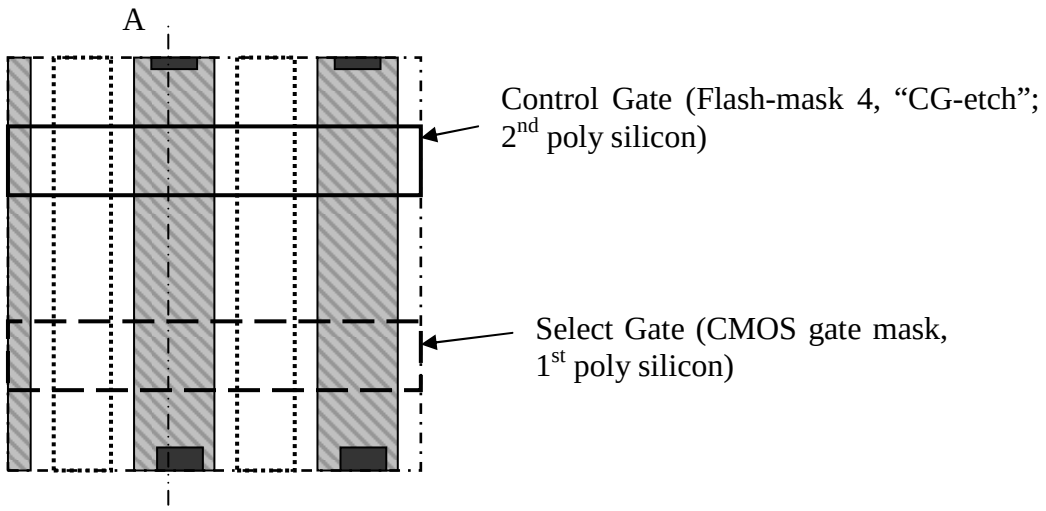
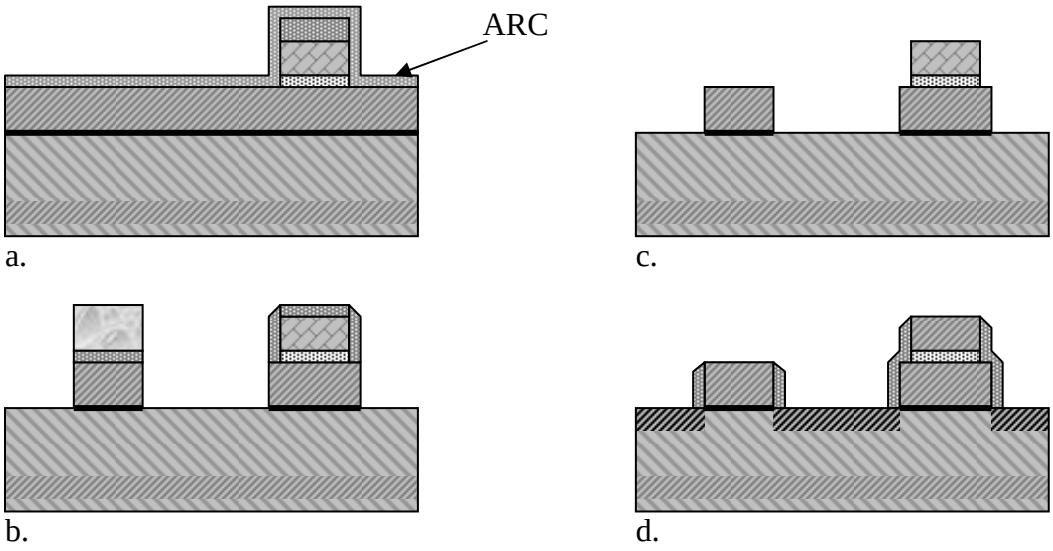


Figure 23: Schematic layout fragment of a 2-transistor cell array

silicon nitride hardmask protects the control gate during this step, and leads to a self-aligned etching of the floating gate. The gate length of the floating gate is larger than that of the control gate, due to spacers formed by the CMOS ARC layer (Fig.21h, cut A). These spacers also appear in the floating gate separation slits outside the control gate (Fig.21h, cut C). They will be removed later in the process (Fig.22i).
At this point a poly silicon re-oxidation is carried out in the CMOS flow, also affecting the memory transistor. All nitride layers, including the ARC on the CMOS gates, the parasitic nitride spacers and the hardmask on the floating gate are then removed by a wet etching step in phosphoric acid (Fig.22i).
The source and drain junctions are implanted after the formation of sidewall spacers (Fig. 22j). The spacers are produced by conformal oxide and nitride deposition, followed by an anisotropic RIE. The source and drain ion implantation is a combined As and P implant for the NMOS transistors and a B implant for the PMOS transistors. After an RTA step for dopand activation a salicide process forms a low resistance CoSi layer on the silicon surfaces. Finally the formation of the metal interconnect-system is carried out, forming the contacts, metal layers, MIM capacitors and the passivation.

3.2.3. The 2-transistor cell

The schematic layout of the 2-transistor cell is shown in Fig. 23. The difference compared to the 1-transistor layout is the placement of the select gate (=CMOS gate mask) parallel to the control gate (=CG-etch mask), between the control gate and the source contact, resulting in a bigger cell size.
The fabrication of the 2-transistor cell is done with the same technological process flow as the 1-transistor cell; only the cell layout is changed. Fig. 24 shows the fabrication of the select transistor after the flash memory module has been completed. The control gate of the memory transistor has been structured at this stage of the process. The ARC for CMOS gate lithography has been deposited (Fig. 24a). In following the gate lithography is done and the select transistor's gate is etched together with the CMOS gates and the floating gate (Fig. 24b). The next steps are poly silicon re-oxidation, nitride removal (Fig. 24c), spacer formation and source / drain junction ion implantation (Fig. 24d).
The select transistor consists, like the CMOS transistors, of the first poly silicon layer only. It has the tunnel oxide as gate oxide and the same p-well and V_T implants as the flash cell. The latter makes the optimization of these ion implantation processes different compared to the 1-transistor or split-gate cells. This will be discussed more in detail in chapter 5.1.



a. After all flash memory process modules and CMOS ARC deposition
b. After CMOS gate etching
c. After nitride wet-etching
d. After spacer formation and S/D junction implantation

Figure 24: Process flow for 2-transistor cell: building the select transistor; cross section along line A (Fig. 23)

3.2.4. The split-gate cell

The schematic layout for fabricating a split gate cell structure with this process flow is presented in Fig. 25. By producing an overlap of the FG-etch and CG-etch masks in the memory transistor's channel area, the split gate structure can be achieved. The CMOS gate mask is needed in this case to protect the source area from the CMOS gate RIE. Cross

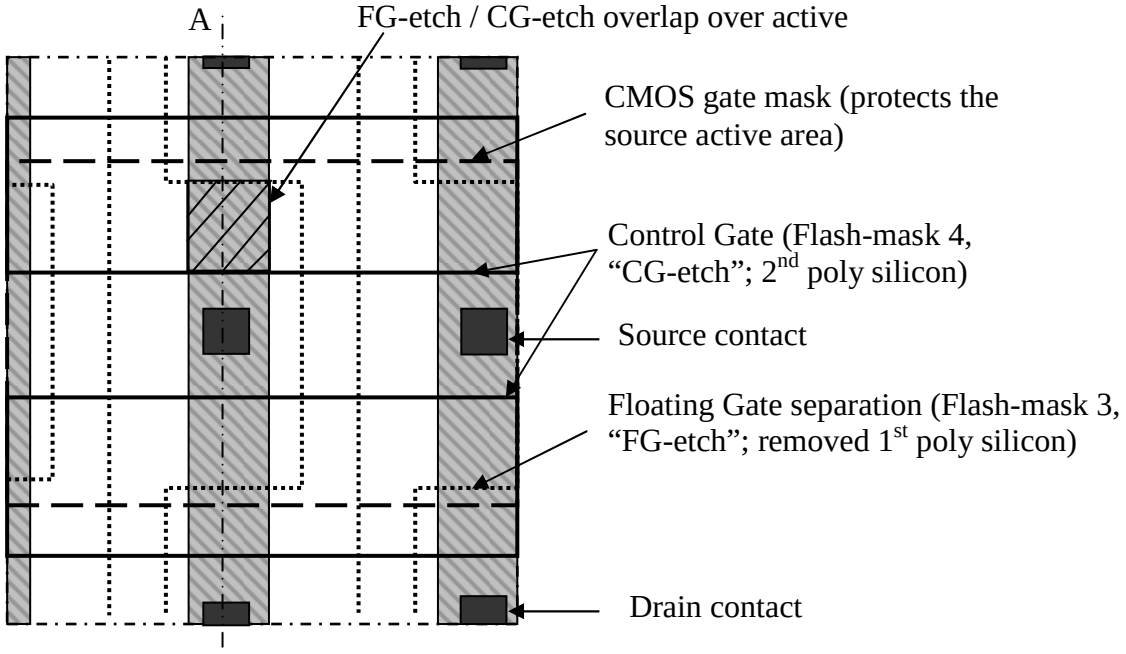


Figure 25: Schematic layout fragment of a split gate cell array

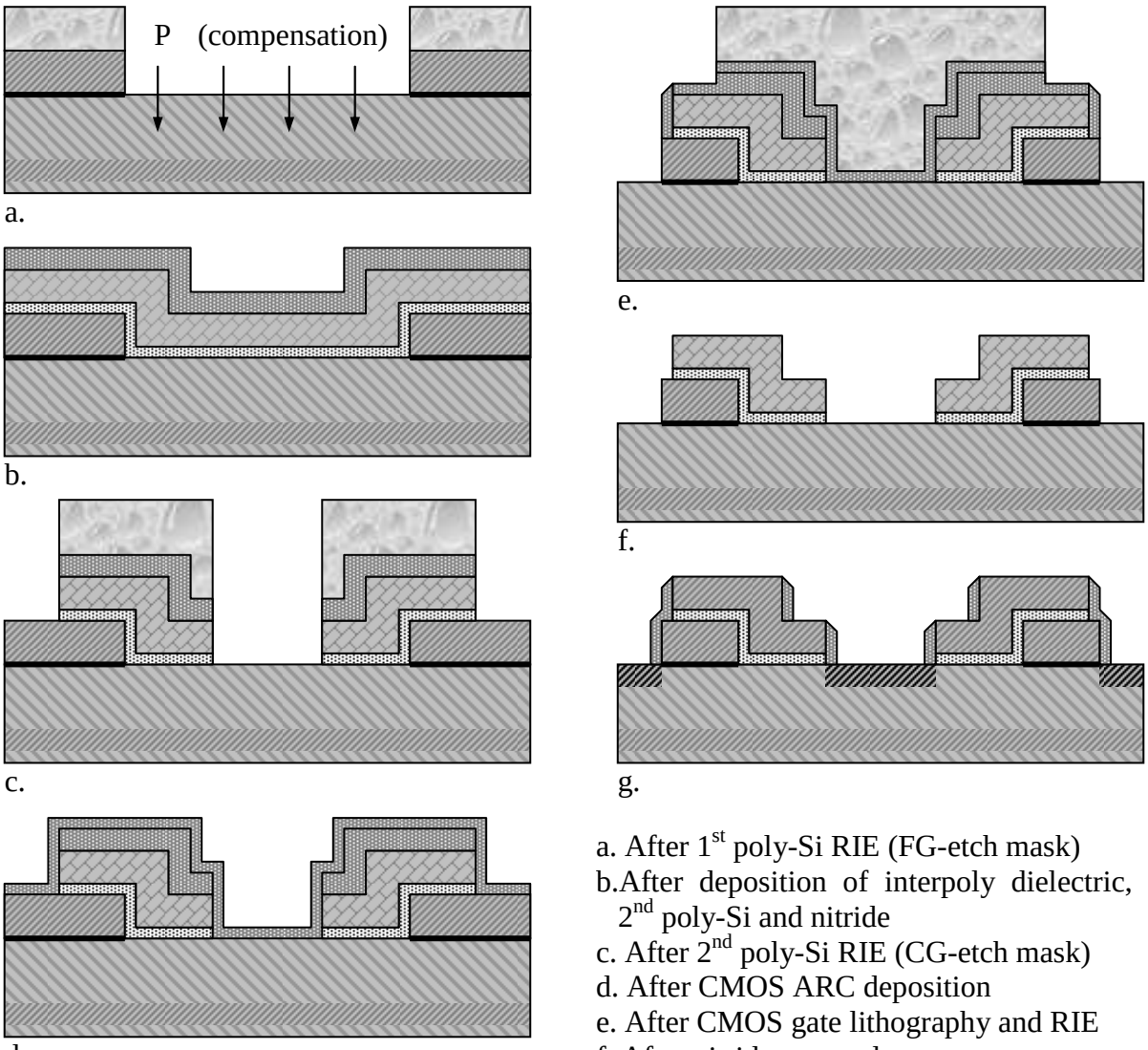


Figure 26: Split-gate cell process flow; cross sections along cut line A (Fig. 25)

sections of a cut along line A of Fig. 25, perpendicular to the wordline direction, are shown at different stages of the process in Fig. 26.

The tunnel oxide formation, floating gate ion implantation and flash-p-well ion implantation are done in the same way as for the 1-transistor cell. The floating gate etching mask (3rd mask of the flash module, FG-etch) is opened here also in the active transistor area. The floating gate and the tunnel oxide are removed in this place (Fig. 26a). With the same mask a P-implant is carried out as indicated by the arrows, which is needed for fabrication of the HVMOS transistors. This implant also adjusts the V_T of the select transistor part of the split-gate cell, as it reduces the net doping (the P compensates the B partly). Fig. 26b shows the state after deposition of the interpoly dielectric layer, the 2nd poly silicon layer and the silicon nitride hardmask layer. The 4th mask of the flash module (CG-etch) is now applied, placed over the former edge of the FG-etch mask (Fig. 26c). The overlap of the CG-etch mask over the remaining floating gate produces the memory transistor part of the cell, while in the already etched area the select-transistor part is formed. The process continues with the CMOS process steps. The CMOS ARC layer is deposited (Fig. 26d) before the gate lithography and etching are performed. Here, the CMOS gate mask has to be placed over the cell's source

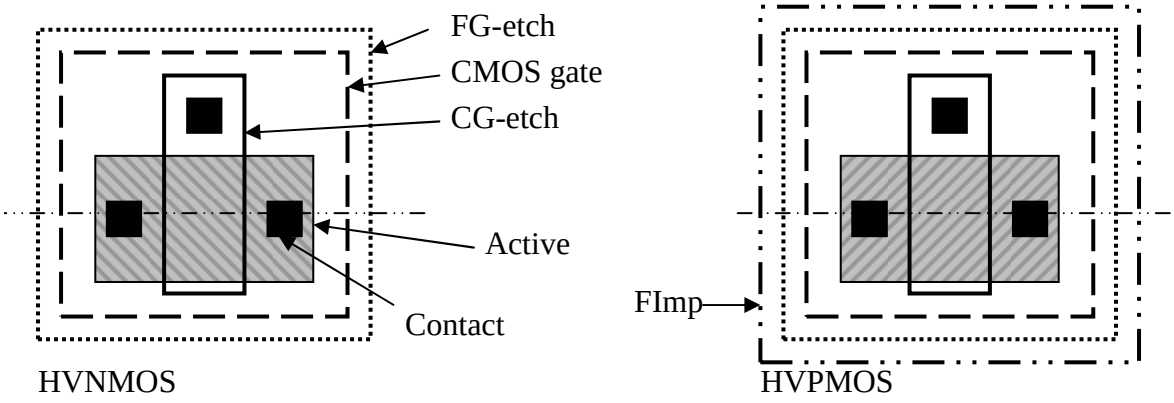


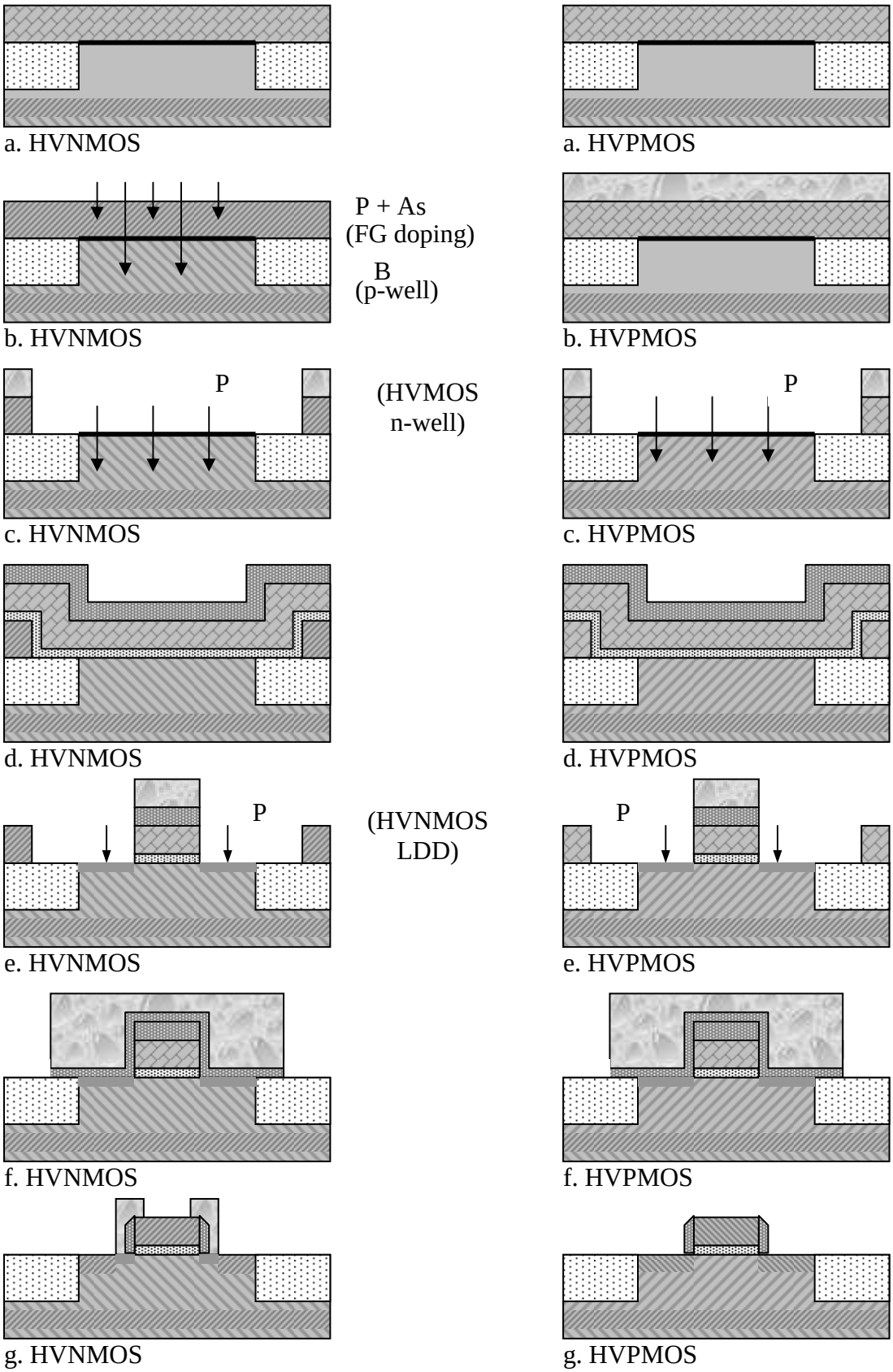
Figure 27: Schematic layout of the HVMOS transistors

area, where the 1st poly silicon layer has already been removed by the FG-etch, to protect the active layer from being unintentionally etched during the CMOS gate RIE (Fig. 26e). The next steps are silicon nitride removal by wet etching (Fig. 26f), the spacer formation, and the ion implantation to produce the source and drain junctions (Fig. 26g).
It should be noted that the CG-etch mask in this concept defines the total gate-length of floating gate + select gate. The adjustment of the CG-etch mask and the FG-etch mask with respect to each other defines the actual length (or the proportion) of both, which makes this a critical process parameter here, and special care must be taken to minimize the misalignment during the respective lithography steps. It should also be noted that the select transistor part is similarly constructed as the HVNMOS transistor described below, with the interpoly oxide layer as gate dielectric and a partly compensated flash p-well.

3.2.5. High voltage MOS transistor integration

The Integration of the high voltage transistors is in general critical with respect to the cost of the embedded memory process. By the way it is done here it does not need an additional lithographic step, which is achieved by reusing process steps of the memory transistor fabrication and of the CMOS fabrication. At different steps the process is adjusted with respect to the HVMOS fabrication. Firstly, the adjustment of the doping profile in the flash-p-well is also done with respect to the HVMOS. Secondly, an additional P-implantation for adjusting the wells of the HVMOS transistors, which is not required for the memory cell itself, is done after FG-etching. Thirdly, the interpoly oxide fabrication is optimized with respect to the HVMOS transistor. The fourth point is another additional shallow P - implantation carried out after the control gate etching for forming LDD extensions.
Fig. 27 shows the layout of the HVMOS transistors. The only difference between HVNMOS and HVPMOS is that the FImp mask (2nd flash mask) blocks the flash p-well implantation at the HVPMOS, and the S/D implants are done with the respective masks of the CMOS process (not drawn in Fig. 27). The gate is drawn with the CG-etch mask in each case. The FG-etch mask is used to clear the HVMOS area from the first poly silicon layer. The CMOS gate mask protects the active areas during the CMOS gate RIE, which is necessary as the first poly silicon is already removed there at this step.
Fig. 28 shows cross section views along the cut lines indicated in Fig. 27 for the HVNMOS and the HVPMOS at different stages of the process. The first cross section shows the state after the first poly silicon deposition (Fig. 28a). The second mask of the flash module (FImp) is then opened only at the HVNMOS (Fig. 28b). The flash p-well is implanted in this area (and the poly silicon is doped, which is not important for the HVMOS), while nothing changes for the HVPMOS. Defined by the FG-etch mask, the first poly silicon layer is now removed from the HVMOS areas (Fig. 28c). After the RIE, before removing the resist, the n-well for the HVPMOS is implanted. This P ion implantation is done in both devices, as the

FG-etch mask is open in the HVNMOS and in the HVPMOS. Due to this, the net doping of the HVNMOS's p-well is reduced to its final level here. Altogether the flash module leads to three different wells: the p-well of the flash cells, the lower doped p-well of the HVNMOS and the n-well of the HVPMOS. The next process steps are the deposition of the interpoly dielectric, the second poly silicon layer and the silicon nitride hardmask layer (Fig. 28d). The control gate etching with the CG-etch mask structures the second poly silicon layer and forms the gates of the HVMOS transistors. Fig 28e shows the cross section after this RIE step. At this point also a shallow P ion implantation is carried out. This is needed to form LDD extensions for the HVNMOS transistor, which is required to reach the necessary breakdown voltage, see chapter 5.
Fig. 28f shows the application of the CMOS gate mask as protection from the CMOS gate RIE. The source and drain areas must be protected. The first poly silicon is already removed here, and the RIE would attack the silicon substrate. The following steps are the formation of the spacers, and the formation source and drain junctions by ion implantation (Fig. 28g). Note that the source and drain junction implants are not carried out self-aligned for the HVNMOS, which is again to from LDD regions at this device. The HVPMOS does not need such a special treatment. The dose of the HVNMOS LDD is so low that is has no impact on the HVPMOS transistor.



a. After CMOS poly-Si deposition
b. 2nd flash mask (FImp)
c. 3rd flash mask (FG-etch)
d. Interpoly oxide, 2nd poly-Si and nitride
e. 4th flash mask (CG-etch)
f. CMOS gate mask
g. Spacer formation,
S/D implantations

Figure 28: Process flow for the high-voltage MOS transistors

Chapter 4

Process Implementation

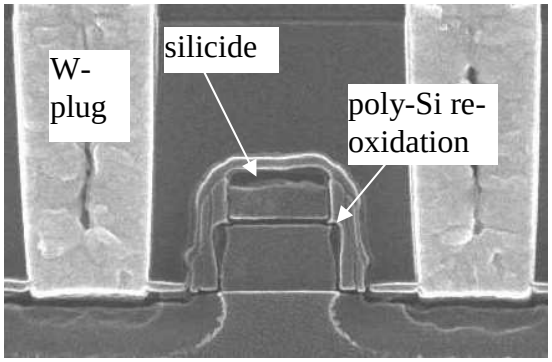
The process described in chapter 3 has been implemented in the pilot line of the IHP. The toolset for the 0.25µm BiCMOS technology fabrication has been used. It consists basically of (1) lithography tools for DUV (wavelength 248nm) and i-line (wavelength 365nm) resist patterning; (2) PECVD tools for deposition of isolators (silicon nitride, silicon oxide, HDP deposition option for dense oxides, e.g. at STI); (3) CVD tools for thermal deposition of silicon nitride, silicon oxide and poly silicon (also in-situ doped); (4) Ovens for thermal silicon oxide growth and thermal implant anneal; (5) RIE tools for dry etching of the different materials; (6) ion implantation tools; (7) metal deposition tools; (8) CMP tools for oxide, metal and poly silicon planarization; (9) CVD tool for epitaxial deposition of SiGe:C; (10) wet-etching and cleaning tools for isotropic layer removal and surface cleaning; (11) RTA tools for implant annealing.
In this chapter, first, geometrical results obtained by SEM and TEM are presented to illustrate the device formation during the process. Then important process steps and process parameters are discussed more in detail, showing the interaction with the overall process flow and the impact on device properties. Finally, the impact of the flash memory integration on the original CMOS and HBT devices will be discussed.

4.1. Geometrical Results

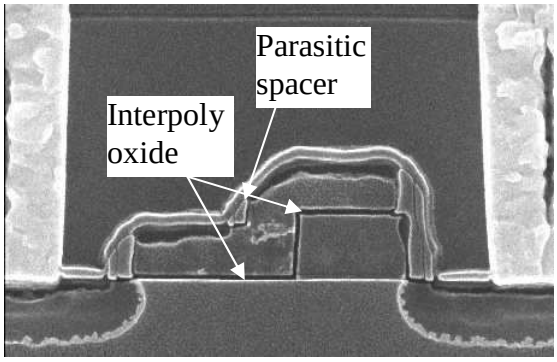
To characterize the process geometrically, SEM and TEM investigations were done. Off-line cross section pictures, as well as in-line top views on the wafer during processing have been performed. This was done throughout the whole process development, to assess the results of the different deposition and structuring steps. Following, typical views of the present state of the process will be shown.

4.1.1. Flash Cells

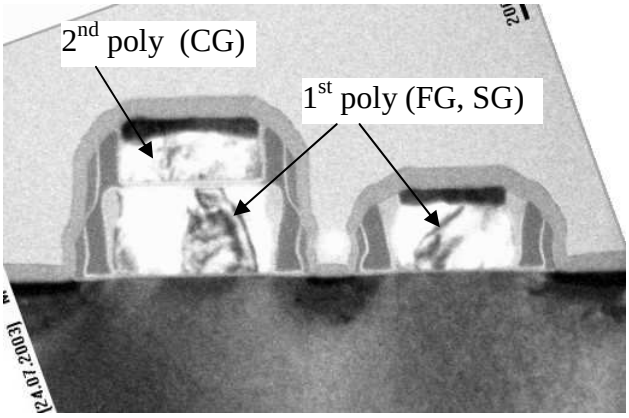
Fig. 29 shows cross sections of the 1-transistor cell, the 2-transistor cell and the split-gate cell, taken after full processing, corresponding to cut-lines A in Fig. 18, Fig. 23 and Fig. 25, respectively. The double poly silicon structure can clearly be seen. The poly silicon thickness is 200nm and 160nm for the FG and CG, respectively. On top of the control gate, the silicide layer for lower sheet resistance can be seen. A 20nm interpoly oxide layer separates the two poly silicon layers. The tunnel oxide under the first poly silicon layer with a thickness of around 8nm is too thin to be seen in SEM. The floating gate length is larger than that of the control gate, which is because of the parasitic spacers formed out of the ARC during CMOS gate etching (see chapter 3.2.2.). The thicker oxide that can be seen at the upper part of the floating gate sidewalls stems from the poly silicon re-oxidation step. The floating gate is doped at this step already by ion implantation, and the doping gradient in the poly silicon from top to bottom leads to an enhanced oxidation of the upper part (a changed oxidation rate of doped silicon is a well-known effect). The sidewall spacers that do the separation of the source and drain junctions from the floating gate can be seen. They consist of a silicon oxide and a silicon nitride layer. Especially in the TEM picture of the 2-transistor cell a second pair



a.



b.

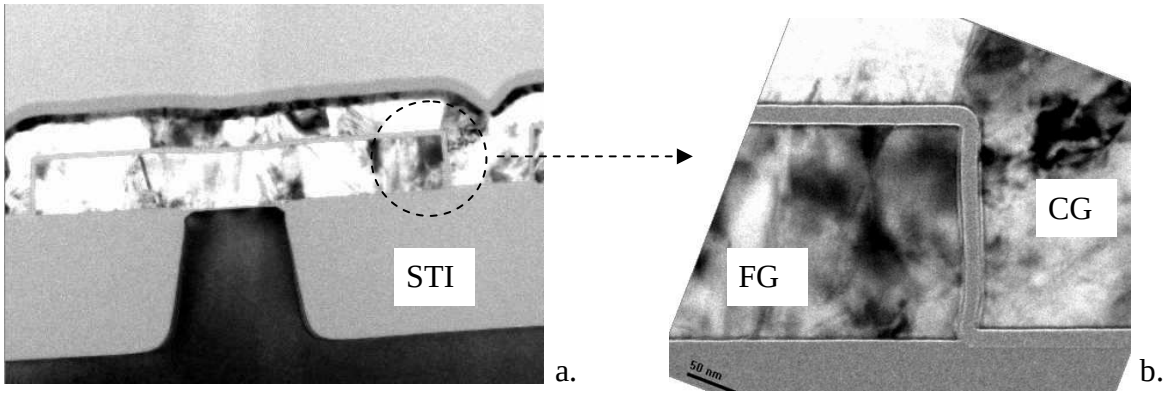


c.

a. 1-transistor cell (SEM)
b. Split-gate cell (SEM)
c. 2-transistor cell (TEM)

Figure 29: Cross section views of the different cell types after full processing; cuts are along bitline direction, crossing the wordline

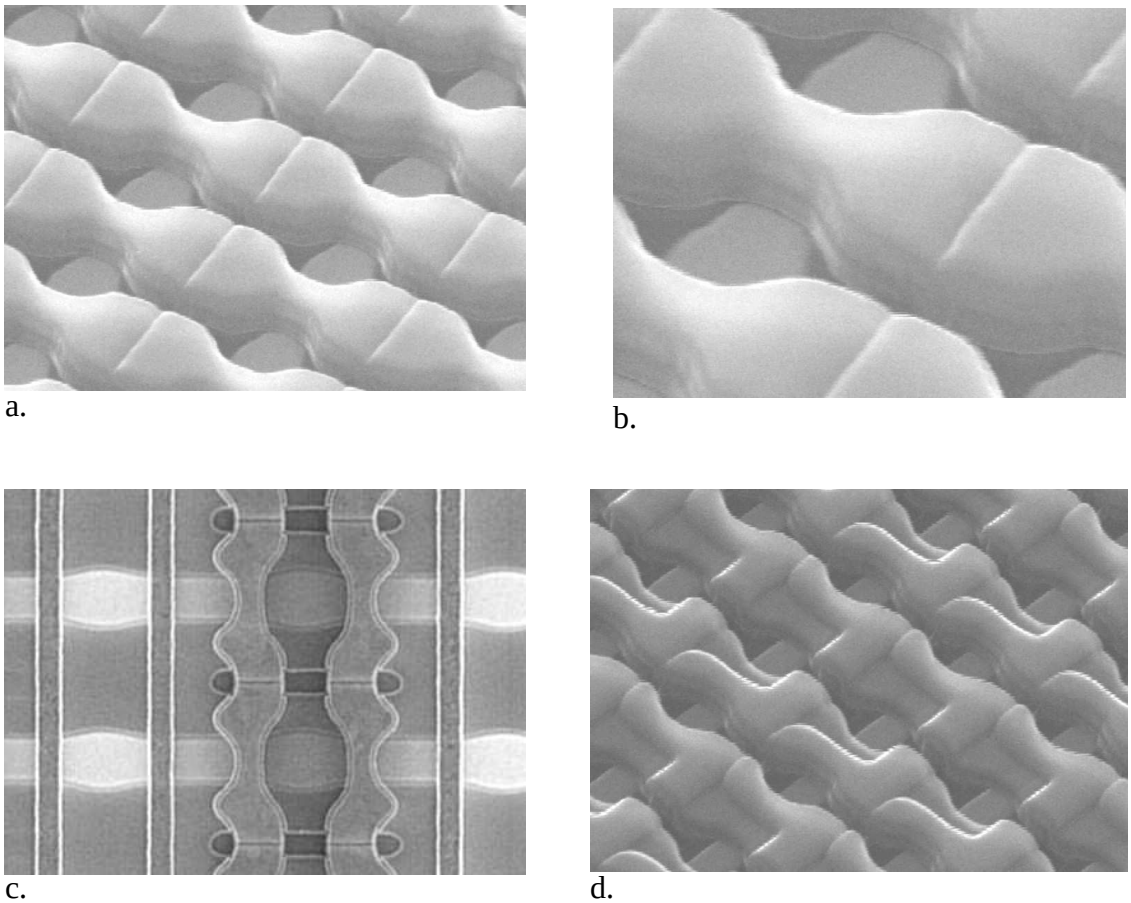
of spacers on top of the first ones can be seen, which stem from the silicidation process, where parts of the wafer are protected from silicidation by a protection layer to prevent silicide formation. This is done to produce the different (non-silicided) poly silicon resistors offered by the process. During the RIE of this layer the spacers are formed. The highly n-doped source and drain junctions can be seen in Fig. 29a and Fig. 29b due to a decorative etching. This etching solution has a doping dependent etch rate, and makes these areas visible this way. The silicon surface in the source and drain areas is also silicided. Especially at the 1-transistor cell, the metal contact plugs can be seen, which connect the cells to the bitline metal. In the cross section view of the split-gate cell its basic structure is visible (Fig. 29b), e.g. how the interpoly oxide forms the gate dielectric for the select transistor part of the cell. Due to the step height, parasitic spacers are formed at the point where the control gate steps off the floating gate. In the 2-transistor cell the select transistor gate is formed of the first poly silicon layer. As the FImp mask is open in the whole cell area, it is also already doped before the poly re-oxidation is done and shows the same enhanced oxidation in the upper part of the poly silicon sidewalls as seen at the floating gate.
Fig. 30 shows TEM cross section views, along wordline direction, corresponding to cut line B in Fig. 18. The images have been taken after full processing. Fig. 30a covers a distance of about one memory cell, showing how the interpoly dielectric and the 2nd poly silicon layer, that forms the control gate, cover the floating gate. The slits separating two neighbouring floating gates, which were etched with the 3rd flash mask (FG-etch) can be seen to the left and to the right of the floating gate. The 2nd poly silicon layer fills these slits. A small notch can be seen in the topology of the control gate at this place. The silicide on top of the control gate is again visible. The patterned substrate can be seen. The silicon (dark area) forming the active



a. b.

Figure 30: TEM cross-section views after full processing; cuts are along wordline direction

transistor area is isolated by the STI to the left and to the right. A zoomed view on the floating gate edge can be seen in Fig. 30b. The interpoly dielectric layer separating the two poly silicon layers conformally covers the floating gate. No thinning of the layer at the edge is seen, and the sidewall layer has the same thickness as the top layer. The floating gate etching process produces a steep profile of the slits.
Fig. 31 shows SEM top views on cell arrays before the formation of the metal interconnects. The views on the 2-transistor array are taken with a 90° top view, while the 1-transistor and



a. b.

c. d.

a. 1-transistor cell array
b. Zoomed 1-transistor cell array
c. 2-transistor cell array
d. Split gate cell array

Figure 31: SEM top views on the different kinds of cell arrays before metallization

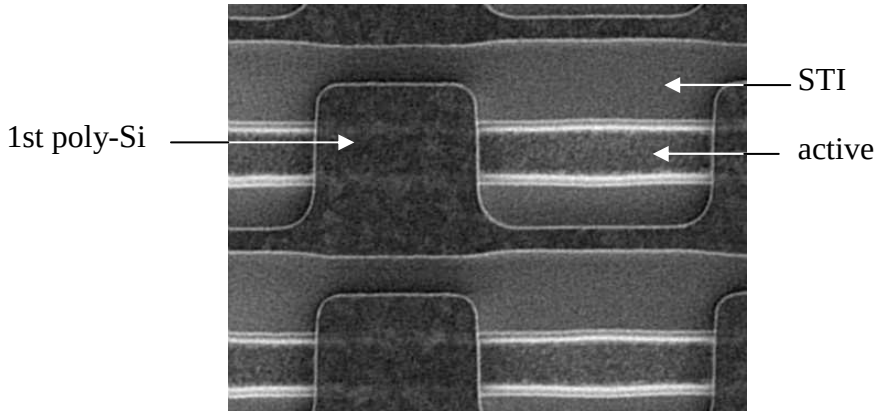
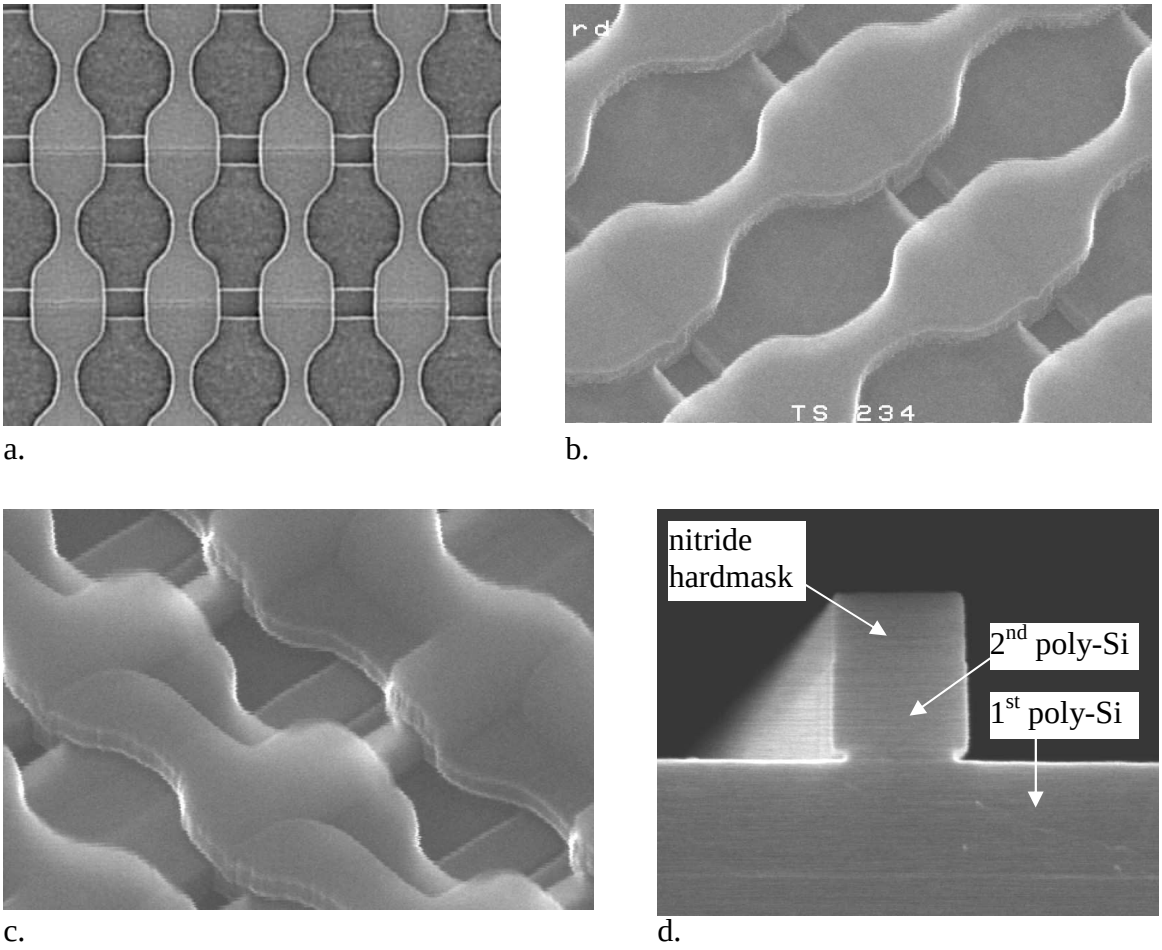


Figure 32: SEM top view on the split gate cell array after floating gate dry etching

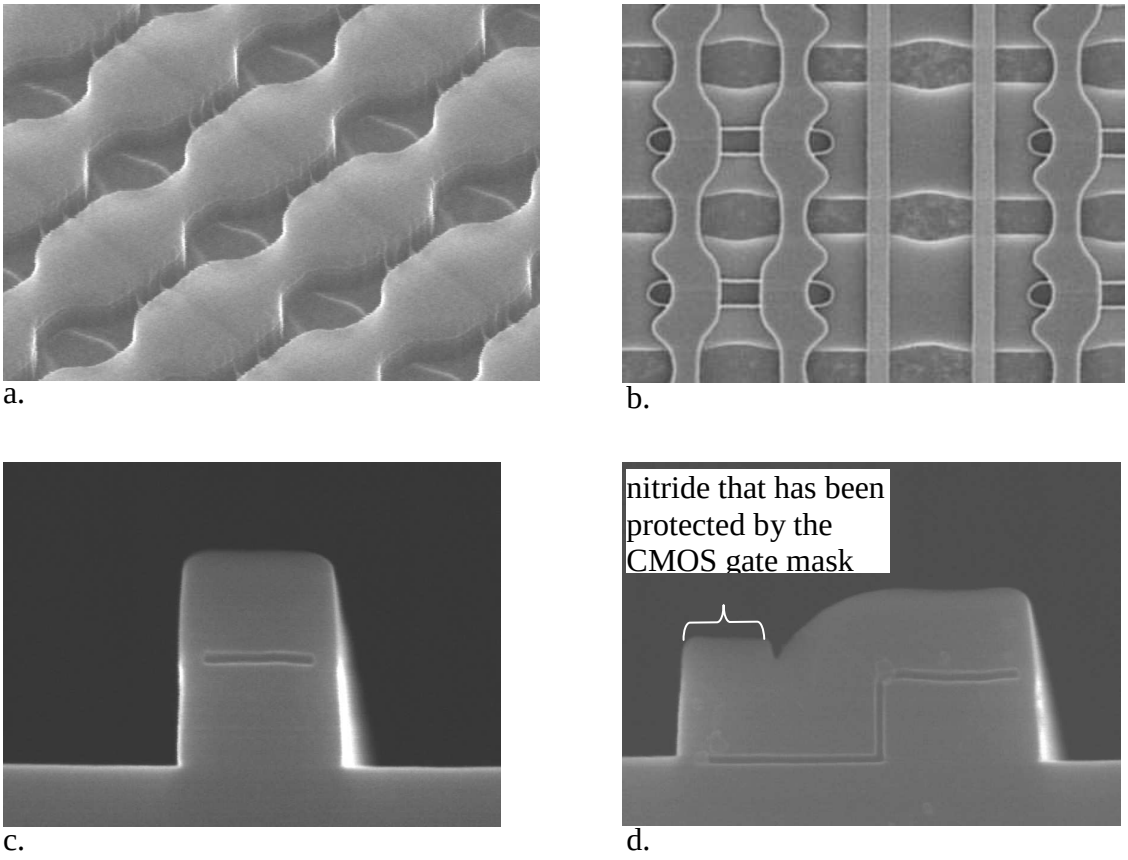
split gate cell arrays are taken with a tilted and rotated wafer. The layout, especially the shape of the control gate, differs from the schematic layouts presented in chapter 3. This is because the layout was optimized to get a high coupling factor at minimum cell size with the given design rules of the technology. In the case of the 1-transistor cell, for example, the control gate has been drawn wider outside the active area, to raise the interpoly capacitance while keeping the minimal gate length on the active area, and keeping the minimum cell size in bitline direction, which is defined by the gate pitch with a contact between the gates. This all leads to the bone-like shape that can be seen in the SEM image. Fig. 31b shows a zoomed view on one single cell in the array. The double poly structure can be seen, as well as the notch in the control gate at the place where two floating gates are separated by the slit in the 1^{st} poly silicon layer. The picture is taken after spacer formation, which is the reason why no sharp poly silicon corners are visible. At the bottom of the double gate structure the active area (bright), crossing the control gate in bitline direction, can be seen.
For the 2-transistor cell array, similar considerations as for the 1-transistor cell lead to the shape of the control gate that can be seen in Fig. 31c. Again, the gate length on the active area is minimal, while otherwise empty area on the STI is used to enhance the CG-FG-coupling. The FG-etch mask does not form a continuous slit here, but only covers 2 control gates at once, to allow an unbroken select gate, as this is also formed of the first poly silicon layer.
Finally, a realization of a split gate cell array can be seen in Fig. 31d. To make it more understandable, an image after floating gate etching (FG-etch mask) is shown in Fig. 32. At the source side, the first poly silicon layer is already removed from the active area and the STI. As also shown in the schematic layout (Fig. 25), the FG-etch mask does not form a simple slit as in the 1-transistor or 2-transistor cell. The FG-mask is extended over the source active area, so that the split-gate structure is realized when the interpoly-oxide and the second poly silicon are deposited. All these are only examples of realizations of the cell layouts; one might find other optimized variants.
The situation in a more early stage of the process is shown in Fig. 33. These images are taken after the control gate etching, at the end of the actual embedded flash process module, as schematically presented in Fig. 20f (without resist here). Cross section views across the wordline are shown (Fig. 33d and Fig. 33e). The result of etching the silicon nitride hardmask and the control gate can be seen. The dry etch process uses the interpoly oxide as stopping layer, so that the first poly silicon layer is untouched. The hardmask dry etching step is optimized to produce a steep profile, which is needed for controlling the gate length during the following etching steps. The result can be seen here. The Fig. 33a and Fig. 33b show top views on a 1-transistor cell array, first a 90° top view, then a view on a rotated and tilted wafer. In the 90° view the layout becomes visible. The rotated and tilted view gives a good impression of the control gate covering the floating gates (which are not yet etched between



a. 1-transistor cell array SEM top view, 90°
b. 1-transistor cell array SEM top view
c. Split gate cell array SEM top view
d. 1-transistor cell SEM cross-section view
e. Split gate cell SEM cross section view

Figure 33: SEM views of the different cells and cell arrays after control gate dry etching

the CGs), climbing from one cell to the next and filling the slits produced with the FG-etch mask. The First poly silicon layer is etched only in the separation slits at this stage of the process. The remaining material between the control gates will be removed later during the CMOS gate RIE. The source and drain active areas are thus still covered by poly silicon and cannot be seen. A tilted view shows the split gate array at this process step (Fig. 33c). It should be compared to the view after FG-etch in Fig. 32 to be understandable. In the tilted view, the small bar of the first poly silicon, which was added to enhance the coupling factor, cannot be seen sharply, because it is partly transparent for the electrons of the SEM.
The next set of pictures illustrates the situation after CMOS gate etching (Fig. 34). In the top view on the 1-transistor array now the stacked gate structure of the cell becomes visible (Fig.



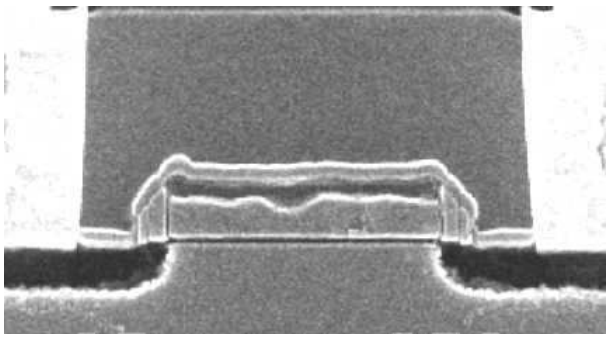
a. 1-transistor cell array top view
b. 2-transistor cell array top view
c. 1-transistor cell cross section view
d. Split gate cell cross section view

Figure 34: SEM views of the different cells and cell arrays after CMOS gate dry etching

34a). The active area of the source and drain junctions can be seen now, crossing the wordline in bitline direction, as the first poly silicon layer has been removed. Between the control gates, at the edges of the former floating gate separation slits, the residual nitride spacers formed by the ARC layer can be seen, which will be removed by a wet etching step later. In the top view on the 2-transistor array the parallel control gate and select gate lines can be seen, crossed by the active area running in bitline direction (Fig. 34b).
The floating gate is etched self-aligned with the control gate. Cross section views of the different cell types are shown in the Fig. 34c and Fig. 34d. The silicon nitride hardmask is thinned, but not completely removed (unfortunately there is only a low contrast between silicon nitride and silicon in the images). The attack of the CMOS gate etching on the silicon nitride hardmask can be observed at the split gate cell, where the CMOS gate mask has covered the source area to protect the silicon substrate. The mask overlaps the select gate part of the cell, so a step in the nitride can be seen at the place of the former resist edge (Fig. 34d).

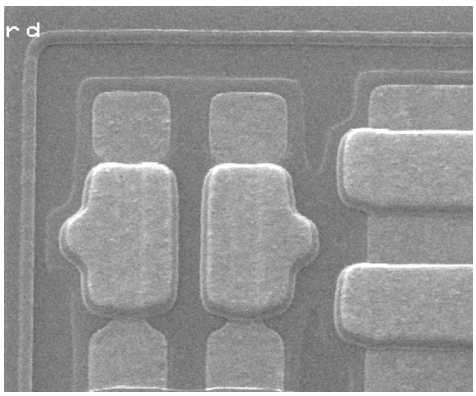
4.1.2. High voltage transistors

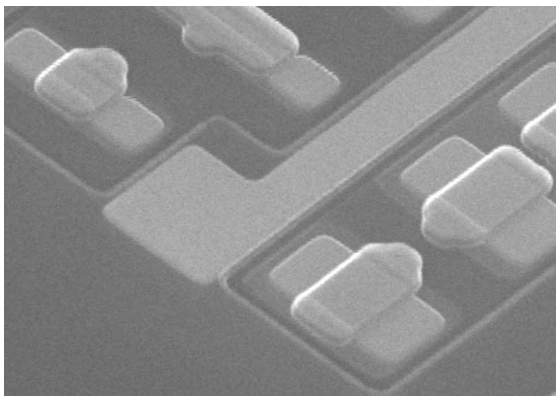
Examples of SEM investigations of the high voltage transistors will be presented now. Fig. 35 shows a cross section view after full processing (Fig. 35a), together with top views after silicide formation (Fig. 35b, immediately before metal interconnect formation) and after CMOS spacer formation (Fig. 35c). In the top views, details of the peripheral circuitry of a complete flash memory are shown. HVMOS transistors of different gate length and width are visible. Fingerprints of the formerly applied FG-etch and CMOS gate mask can still be seen surrounding the HVMOS areas. Fig. 36 shows the situation after control gate lithography and



a. Cross section after full processing
b. Top view after silicide formation
c. Top view after CMOS spacer formation

a.

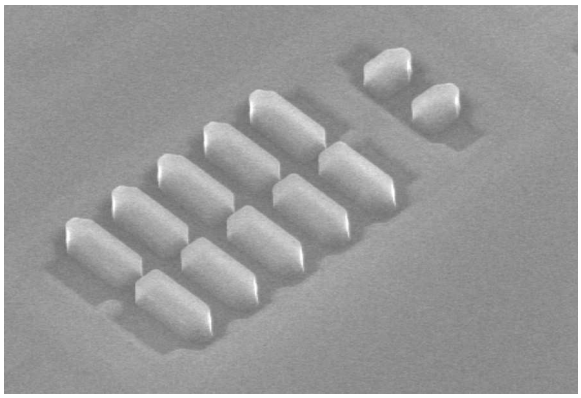


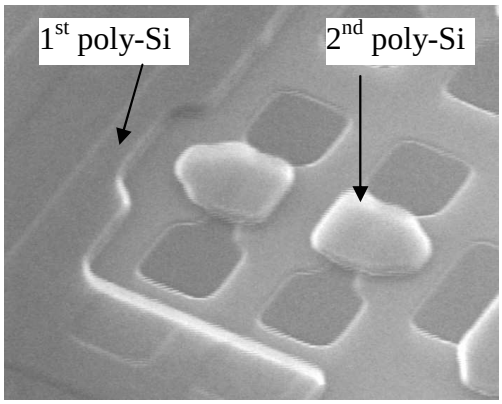


b. c.

Figure 35: SEM images of the HVMOS transistor

after control gate etching. The resist covers the later high voltage gates in an area where the first silicon layer has been removed during the floating gate etching with the FG-etch mask. The second poly silicon layer is the only poly silicon layer in that area, with the interpoly oxide under it. The edge of this area, where the second poly silicon climbs off the first poly silicon layer, can clearly be seen. After etching, the source and drain active areas become visible. Outside the area that was covered by the control gate mask, the first poly silicon layer is cleared of the second poly silicon layer again. The first poly silicon will be removed from the regions where it is not needed by the CMOS gate RIE step.
Fig. 37 shows cross section views after control gate etching and after CMOS gate etching.

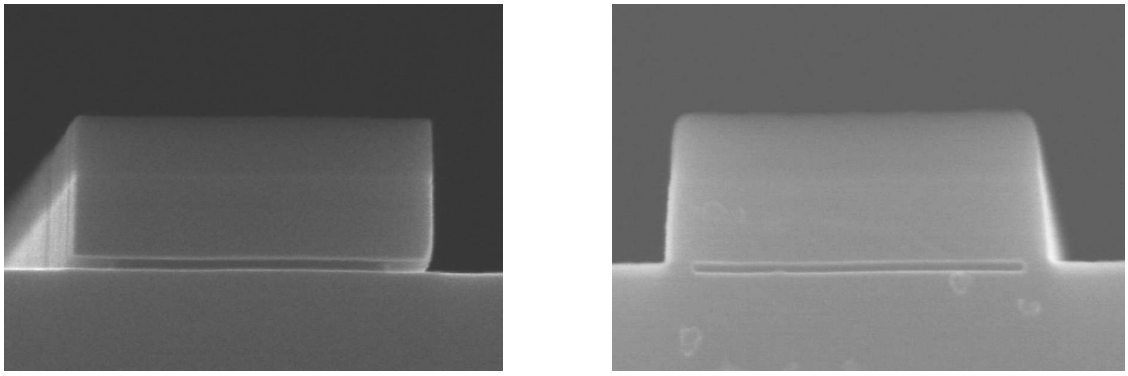




a. b.

a. After control gate lithography b. After control gate dry etching

Figure 36: SEM views of the HVMOS transistor at control gate patterning



a. b.

Figure 37: SEM cross sections of HVMOS transistors after control gate etching (a) and after CMOS gate etching (b)

Note that in Fig. 37b the CMOS ARC layer, which will be removed later by wet etching, fully covers the HVMOS.

Fig. 38 finally shows a top view on a detail of a complete flash memory. It shows the combination of the different devices in one chip: the memory cell array, low voltage CMOS transistors and HVMOS transistors.

4.2. Important process steps and process parameters

4.2.1. The tunnel oxide

First, the tunnel oxide fabrication will be reviewed. The tunnel oxide is one of the key elements of the floating gate cell. The electrons enter and leave the floating gate through the tunnel oxide at high electric fields, and its potential barrier keeps the electrons on the floating gate at low electric fields. Thus its electrical properties are very important for both, the memory performance and reliability. Results of electrical characterization are presented in chapter 5. In general, the quality in terms of low density of process induced defects in the bulk

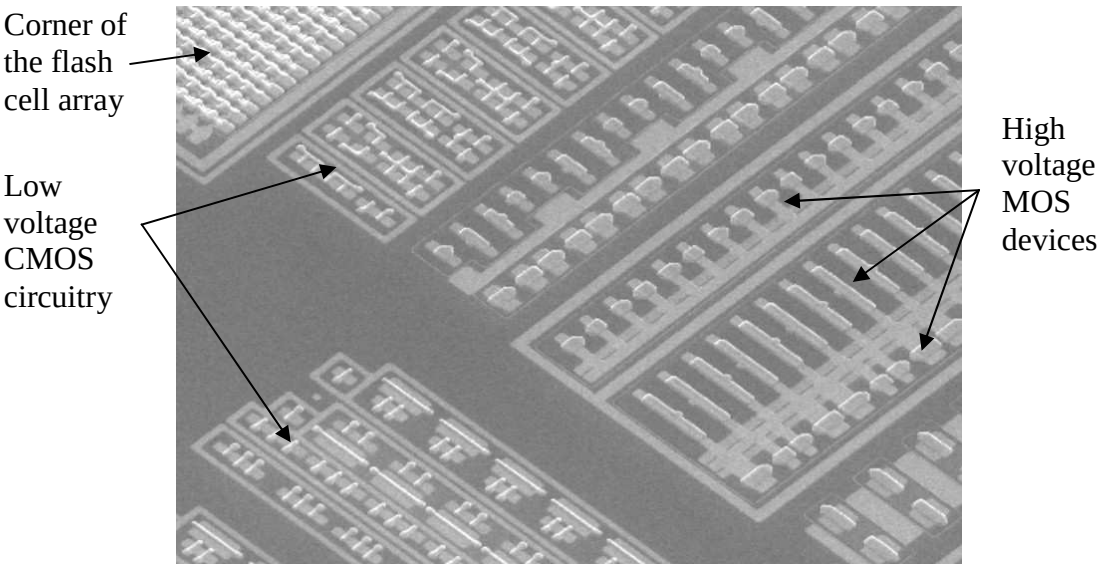


Figure 38: SEM top view on a detail of a flash memory before metal interconnect formation

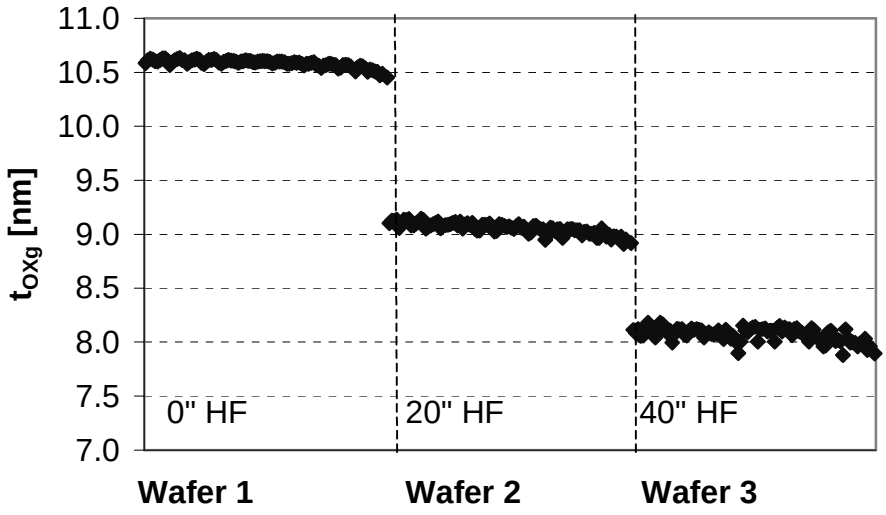


Figure 39: Tunnel oxide thickness extracted from electrical CV measurement of MOS capacitors (see appendix E); 100 sites measured on each wafer

and at the interface must be high to guaranty memory yield and reliability. Defect generation (traps) during memory operation defines a lower limit for the allowed tunnel oxide thickness. The upper limit is given by the upper limit of the acceptable programming time. This time depends very sensitively on the oxide thickness, as the tunnelling current has a strong dependence on the tunnel oxide thickness (exponentially). A difference of 1nm in the tunnel oxide thickness leads approximately to a factor of 10 in writing time (the higher CG/FG coupling factor at thicker tunnel oxides can hardly compensate this trend). Controlling the oxide's thickness, with respect to the homogeneity over the wafer, from wafer to wafer and from production lot to production lot, is thus important. The absolute thickness should be chosen as thick as possible, with respect to the required programming performance, to achieve the maximum reliability.

As the baseline process is a single-oxide process with only 5nm oxide thickness, which is too thin from the reliability point of view to be used as tunnel oxide, a second thermal oxide needs to be produced. This is done with the described dual gate oxide processing (chapter 3.2.2.). In this process, the final tunnel oxide thickness is influenced by three separate process steps: 1) the thermal oxide growth itself, 2) the pre-cleaning performed before the thin (CMOS) oxide growth and 3) the thermal oxidation of the thin oxide. The thermal oxidation steps are industry standard steps, which have a tight control of oxide thickness and homogeneity. The second oxidation step adds only a fraction of the thin oxide thickness to the tunnel oxide, as the process is diffusion controlled and has a square-root transient characteristic at this oxide thickness. The influence of the CMOS gate oxide pre-cleaning step can be seen in Fig. 39, showing the tunnel oxide thickness (from electrical CV-measurements) of 3 wafers (100 sites measured on each wafer). The main influencing component of the cleaning procedure is a wet etching in diluted HF. The wet etching time has been varied to investigate the influence (with the same thermal oxidation and other processing conditions). A reference wafer without HF wet etching (0") has been processed. It can be seen that for longer HF times the scattering of the oxide thickness becomes slightly higher. For an etching time of 20" the scattering is almost the same as for the reference without HF etching, while for 40" a significant difference can be seen. The bigger difference in thickness between 0" and 20" compared to 20" and 40" can be explained by an additional "handling time" of the wet etch tool, that has to be added in the case of 20" and 40" etching time. The given etching times are here only the nominal

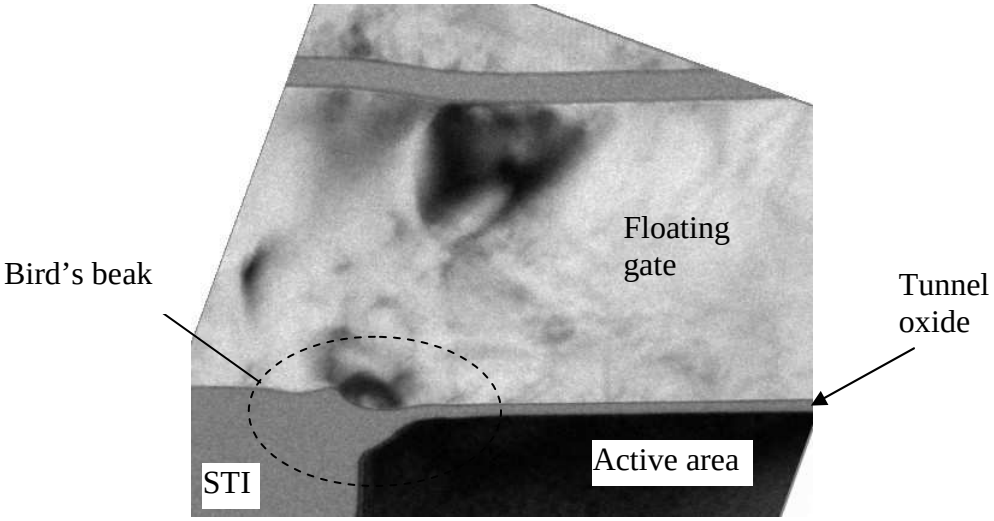


Figure 40: TEM cross-section of the STI corner covered by the tunnel oxide

values entered in the tool's program. The real etching time is thus prolonged by a fixed amount of time.
A critical point for the tunnel oxide is the shallow trench edge. It must be made sure that it is well covered by oxide to prevent any leakage currents here, especially as the edge is also critical with respect to mechanical stress that can impact the processing and the electrical properties. Fig. 40 shows a TEM view of the STI corner. The tunnel oxide thickness is even thicker than on the main active area, due to a "bird's beak" that was formed during STI formation, and that was kept by not too long HF cleaning before the tunnel oxide growth and before the CMOS oxide growth.
The influence of the tunnel oxide thickness on the programming time is presented in chapter 5. The lower limit of the tunnel oxide thickness for guaranteed reliability can only be determined by extensive reliability tests, which are not part of this thesis. Some basic investigations are presented in chapter 5.

4.2.2. The interpoly oxide / HVMOS gate oxide

In this technology, the interpoly oxide and the gate dielectric of the high voltage transistors are formed simultaneously, which is different in most other embedded flash memory technologies. The oxide must fulfil the requirements of both applications. A layer stack consisting of a thin (2 nm) thermal oxide, grown at 700°C in a dry oxidation process, followed by an LPCVD oxide deposition at 638°C from a TEOS source, which is finally densified at 700°C in O_2 ambient, is used here. Densified stacked thermal / LPCVD oxide has been investigated in the early 90's for use in submicron CMOS processes and its applicability and low defect density has been demonstrated [78][79][80][81]. The first thin thermal oxide provides a good interface quality for the HVMOS devices. The use of an LPCVD deposited oxide for formation of the main part of the layer has some advantages for the flash cells and the integration scheme (compared to a pure thermal oxide): 1) a conformal deposition compared to thermal oxidation of doped poly silicon, especially at the floating gate edge (Fig. 30), which prevents unwanted formation of a tip at the poly edge that could result in an oxide thinning and worse cell reliability; 2) the added thermal budget is lower for an LPCVD process compared to thermal oxidation; and 3) the consumption of (the CMOS-) poly silicon during thermal oxidation would enhance the impact of the flash integration on the CMOS flow and disturb the modular character of the integration scheme. The significant effect of the initial thermal oxidation and final densification on the HVMOS transistor's V_T distribution can be seen in Fig. 41. The impact is on both, HVNMOS and HVPMOS transistors. The

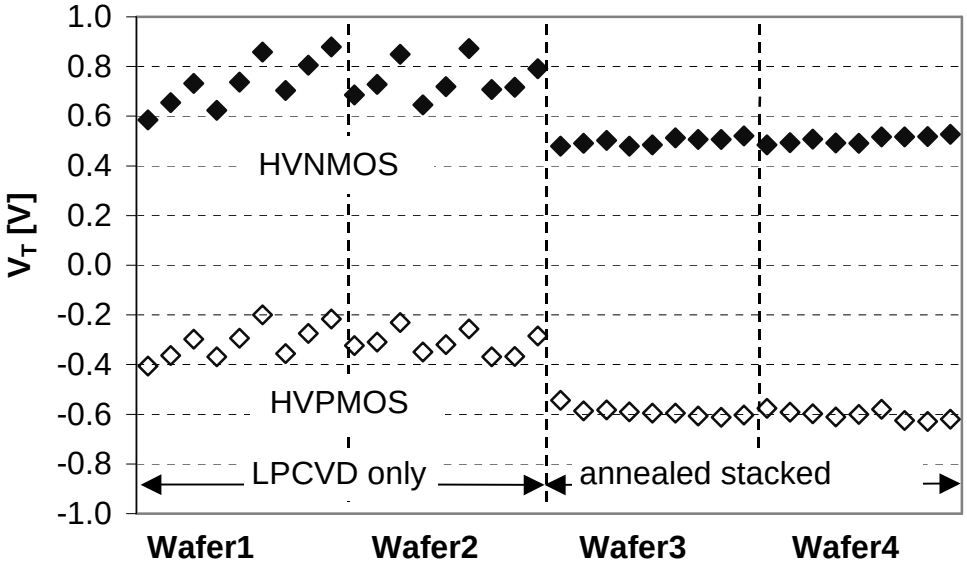


Figure 41: HVMOS V_T values of differently processed wafers: wafers 1 and 2 with LPCVD gate oxide, wafers 3 and 4 with stacked thermal oxide and LPCVD oxide + annealing

scattering of the V_T-values is reduced and the mean value is lower (or higher in absolute values for the HVPMOS). This is thought to be because of a reduced number of interface and oxide electron traps. The chosen total thickness of around 23 nm results in acceptable absolute V_T values for the HVMOS transistors (in combination with the HVMOS wells) and a gate breakdown voltage high enough for the targeted operating voltages. For the flash cells, the reliability of a 1 Mbit circuit must be demonstrated in an extensive reliability investigation, to judge the interpoly oxide quality. This is, again, not part of this thesis. Basic investigations of the reliability of flash memory cells are presented in chapter 5.1.5.
It should be mentioned that usually an ONO stack forms the interpoly dielectric layer for reliability reasons in most floating gate process technologies. The use of such a layer stack is also possible in this approach, but would need an adjusted (and maybe more complex) control gate etching and cleaning, and the impact on the HVMOS would have to be reviewed, where the traps at the silicon oxide/silicon nitride interface could cause problems (see also 5.1.2).

4.2.3. The Flash p-well and the HVMOS wells

The wells of the different MOS devices must have a surface doping concentration that adjusts the intrinsic threshold voltage and a concentration profile in depth that prevents punch through and other short channel effects. For the flash cells, a chain of B ion implantations of different energies and doses forms the p-well. Different optimum V_{Ti} values exist for the different cell types. For the 1-transistor cell and the split gate cell the V_{T0} of a cell without floating gate charge should be in the middle between the written and erased V_T - state. This configuration leads to minimized excess charge on the floating gate of the programmed cells and thus low electric fields and better retention characteristics. Furthermore, if the V_{T0} would be too far from this centre, either the writing or the erasing time would rise unacceptably. For the memory transistor in the 2-transistor cell, the same considerations are valid, but it has to be taken into account that the select transistor is placed within the same well. Here, the V_T should not be too low in order to achieve small enough leakage currents at the unselected cells of one bitline. Depending on the subthreshold slope, the $V_{T,w}$ must be well above 0V. It must be kept in mind that the leakage currents of all cells in one bitline add together.
In the presented integration scheme, the wells of the high voltage transistors are not independently produced from the flash cells. One additional P-ion implantation compensates

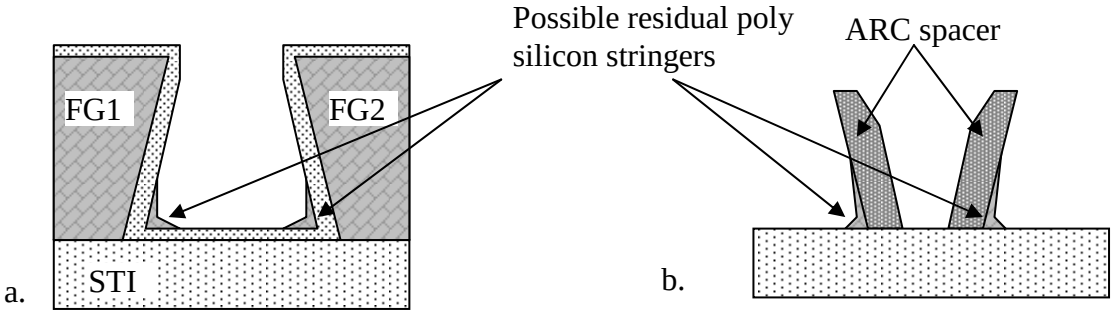
the flash p-well partly to form the p-well for the HVNMOS, while this P-ion implantation alone forms the well for the HVPMOS. This implantation should be adjusted in a way that the V_T values of the HVNMOS and HVPMOS are symmetrical to 0V, which is beneficially for the circuit design. As the gate dielectric of the HVMOS is thicker than that of the flash cells, the resulting V_T has an acceptable absolute value with the reduced concentration of the compensated p-well. Here, it can be adjusted to around + / - 0.6 V in combination with the 1-transistor memory cell.

4.2.4. Floating gate etching

The purpose of the floating gate etching with the 3rd mask of the embedded flash integration flow (FG-etch) is to separate the floating gates that are covered by the same control gate from each other. The self-aligned floating gate etching with respect to the control gate during the CMOS gate etching later in the process defines the length of the floating gate, the width has to be defined before the deposition of the second poly silicon. The etched slits can be very narrow, which is beneficial for the cell size and for the control gate resistance. Narrow slits lead to an almost planar surface after the deposition of the second poly silicon layer, thus keeping a low wordline resistance after silicidation. In case of a too deep notch in the CG, a formation of parasitic spacers stemming from the deposited and structured silicide blocking layers (which are needed for poly silicon resistor fabrication) could disturb the silicide formation on the wordline.

The FG etching itself is done in a plasma etching process consisting of different process steps. First, the gross part of the poly silicon is etched in an anisotropic process, until the underlying oxide is reached. Then, a very selective over-etching is done, which etches silicon significantly faster than silicon oxide, to ensure that on the whole wafer no silicon is left at the bottom of the slits. This is necessary to compensate layer thickness variations and etch rate variations across the wafer. This two-step procedure is necessary as the optimizations are done with different targets. The first step has to produce the required steep profile of the slits, while the second one needs the high selectivity. This is not likely to be achieved with only one RIE process.

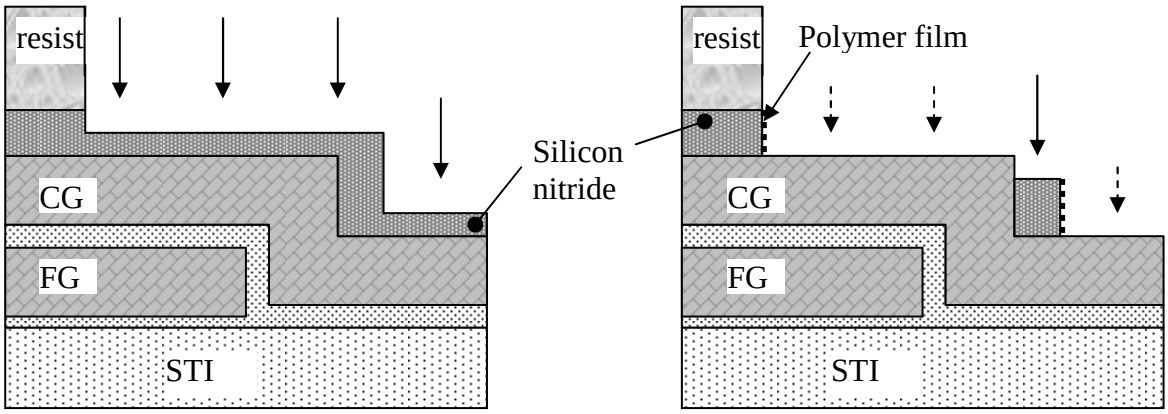
The target profile is a rectangular slit, with steep (90°) poly silicon sidewalls. The reason for this is to prevent shorts between the cells of neighbouring wordlines. The critical points are at the process steps shown in Fig. 20f, cut C and Fig. 21h, cut C. Here, all poly silicon has to be removed out of the area between two control gates during the control gate etching and during the CMOS gate etching. A negative slope of the sidewalls could lead to residual poly silicon at the bottom corners of the slit (Fig. 42a), which would form stringers and cause a short



Negative FG sidewall slope: Control gate etching (= Fig. 20f, cut C, before oxide removal)

Positive FG sidewall slope: CMOS gate etching (= Fig. 21h, cut C)

Figure 42: Unwanted residual poly silicon at non - 90° slope of floating gate slit sidewall



Hardmask main-etch: isotropic silicon nitride dry etching

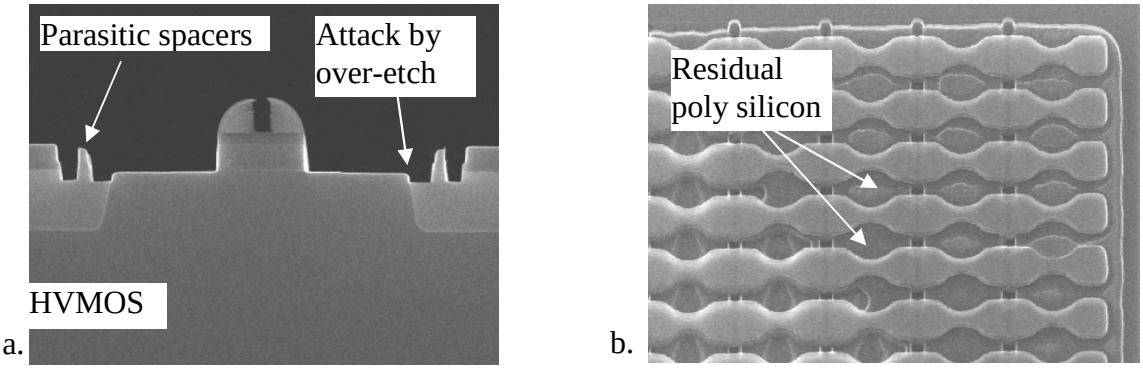
Hardmask over-etch: selective removal of residual silicon nitride

Figure 43: Control gate etching: main-etch and over-etch of the first part of the etching process, which is the silicon nitride hardmask patterning

between the wordlines of two neighbouring cells. A positive slope of the sidewalls would lead to the same problem during the CMOS gate etching (Fig. 42b). Residual silicon at the bottom corner could lead here to a short between neighbouring floating gates. The preferred profile is thus a 90° sidewall slope, tending more towards a negative slope (or a slight notch at the bottom of the slit as can be seen in Fig. 30b), as it is easier to optimize the control gate etching in this respect than the CMOS gate etching, which has the most strict profile and final dimension requirements in the process, and on top of this the modular character of the embedded flash process can only be kept if no additional requirements are added to the CMOS gate RIE.

4.2.5. Control gate etching

The control gate etching is a rather complex dry etching process that has to be carefully optimized. Layers of different materials have to be structured on a topology resulting from the floating gate etching. As it has been the case for the floating gate etching, the process has to be split into different parts with different properties. The main differentiation is done with respect to the different materials that have to be etched. The three layers that need to be etched outside the area covered by the resist mask are the silicon nitride hardmask layer, the control gate poly silicon layer and the interpoly oxide layer. The resulting three main parts of the etching process are further subdivided. The hardmask etching needs to be anisotropic to produce steep sidewalls, in order to transfer the resist pattern without losing any gate length to the underlying layers. Furthermore, the formation of parasitic spacers at edges of the first poly silicon layer (at the edges of the former floating gate mask) must be prevented. To do this, the etching process must have a high selectivity between the etched silicon nitride and the underlying poly silicon, to allow a long over-etch that removes all nitride material from these steps. In this case, the over-etch must even remove more material (in height) than the actual main hardmask etching: the step height that has to be cleared is higher than the hardmask thickness, and, as the etching must be anisotropic, the material is mostly removed from top down. The hardmask etching process thus consists of a main-etch, patterning the silicon nitride until the poly silicon is reached, followed by a highly selective over-etch, which clears the steps in the first poly silicon from nitride spacers. To keep the control gate length constant during this over-etch, which is not as anisotropic as the main etch, the nitride sidewalls are protected by a layer of polymer material, which was build up during the main etch by the chemical reactions of the chemical species that are present during etching (etch gases and the



Effect of too low selectivity of the hardmask etching process during control gate patterning

Effect of non-optimized cleaning after control gate etching: masking of the following floating gate self-aligned structuring

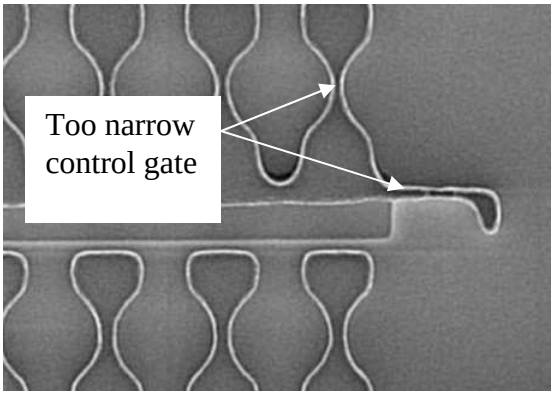
Figure 44: Control gate etching: problems at early stages of the process development

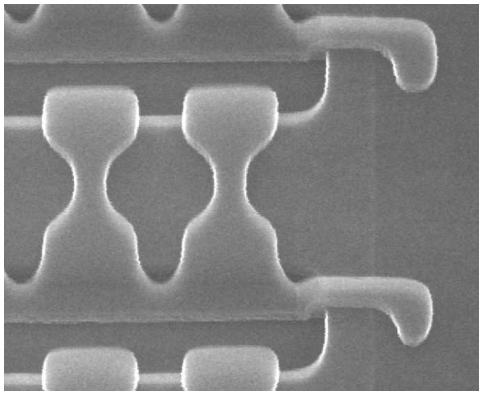
etched material). Fig. 43 demonstrates the two-step etching procedure in a schematic cross section view.

In general, the more selective the over etch is, the more relaxed are the requirements for the following poly silicon etching (as explained below), and the better is the gate-length transfer from the mask to the poly-silicon (as the hardmask over-etch is not etching the underlying poly silicon in an anisotropic way).

The following poly silicon etching process has to do a similar job as the hardmask etching: firstly produce steep poly silicon sidewalls and then remove the poly-silicon from the steps formed by the first poly silicon layer, especially out of the slits that were etched with the floating gate mask. This part of the etching process is thus similarly divided in a main-etch part and an over-etch part, with the same considerations as for the hardmask etching. The higher the selectivity of the hardmask etching in the first part has been, the more poly-silicon is now left on the interpoly oxide, and the better (higher) is the ratio of main-etch versus over-etch in this case. This relaxes the selectivity requirements (poly silicon versus interpoly oxide) of the over-etch. The last part of the etching, the patterning of the interpoly oxide, is done partly by dry etching and partly by wet etching in HF acid. This combined etching has been found to be the best solution to clear the wafer not only from the silicon oxide, but also from the polymer films formed during the former parts of the etching procedure.

Fig. 44 shows SEM images taken during the development of the etching process. Fig. 44a illustrates a non-optimized over-etch of the hardmask etching. The selectivity of the over-etch was too low. The underlying poly silicon layer is etched too much, which transfers to the following etching steps and the total etching goes far too deep (the silicon surface is attacked and the STI oxide partly removed). In addition, parasitic silicon nitride spacers at edges of the first silicon (see Fig. 43, e.g. at the place where the first poly silicon is removed to form the HVMOS transistors) are not yet prevented. The spacers masked the following etching steps, leading to the parasitic spacers seen in the image. The development of a prolonged over-etch with significantly higher selectivity solved these problems. Fig. 44b shows the result of a non-optimized cleaning after the control gate etching. Residual polymer films were still present at the next etching step, the CMOS gate etching, and thus partly prevented the proper self-aligned etching of the floating gate. It can be seen that the first silicon layer is not completely removed between the control gates.





Without ARC Adjusted silicon nitride ARC

Figure 45: Effect of ARC on control gate patterning

4.2.6. Control gate lithography: anti reflective coating

It has been found that for a proper lithographical photo resist patterning the application of an antireflective coating (ARC) is necessary for the lithography process of the control gate layer. At the present lithographical resolution level this technique is a standard measure. Here the situation is even more critical, as the resist has to be patterned on a non-planar surface. The first poly silicon is already patterned and the second poly silicon, covered by the silicon nitride hardmask, follows this topology. In places where the second poly silicon layer steps off the first poly silicon layer, the exposing light is scattered and reflected back into the resist, partly focused. The resist is exposed in an uncontrolled way, depending on slight thickness and topology variations. To prevent this, an antireflective coating is used. This is a layer placed under the resist that prevents the backscattering of light from the substrate by subtractive interference and / or absorption. Due to the present topology, the absorbing factor should in this case be dominant.

The approach was to change the optical properties of the nitride hardmask itself so that it acts as an ARC, instead of introducing a separate layer, which would have lead to a change in the control gate-etching regime again. The proposed solution does not impact the process flow. In general, for every lithographic layer, the applied ARC has to be individually adjusted with respect to the reflecting properties of the underlying layers. The optical properties of silicon

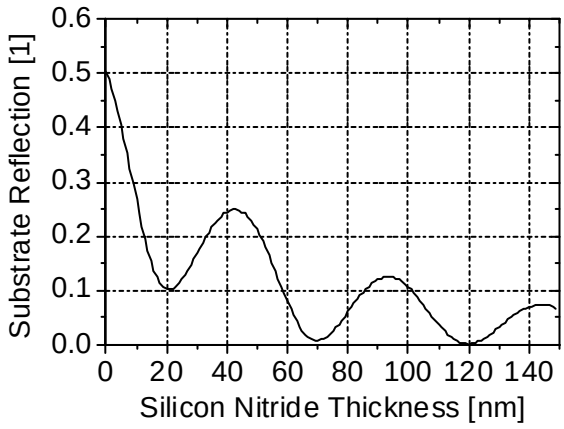


Figure 46: Swing-curve: Calculated substrate reflection versus silicon nitride thickness for the silicon rich silicon nitride on top of the control gate layer stack

nitride can be adjusted by the amount of silicon in the material. Fig. 46 shows a calculated reflection versus thickness curve (so called „swing-curve“) for the chosen silicon-rich silicon nitride. At a thickness of 120nm the reflection is completely suppressed. This thickness is also an acceptable choice for the hardmask functionality of the layer. Fig. 45 shows a comparison of a detail of the embedded flash memory (a line of reference cells) processed without any ARC and with the adjusted silicon nitride layer that is used as ARC here. The better control of the shape of the pattern can be clearly seen.

4.3. Process impact on CMOS and HBT

The impact of the added flash memory process steps on the original HBT and CMOS will be discussed now. Electrical measurements done with this respect are presented in chapter 5.3.
Five points can be defined in which the embedded flash module impacts the MOS devices and their processing.
(1) The formation of the tunnel oxide. This can hardly be prevented when embedding a floating gate memory device, as there are strict limitations to the thickness of this oxide, which do not match the typical gate oxides of scaled CMOS technologies. The applied dual gate oxide process usually needs an additional thermal oxidation step to form the tunnel oxide before the thin oxide is grown. It was tried to prevent this here by re-using an already present so called “sacrificial oxide”, which has been grown earlier in the process and can be adjusted to the required thickness. The wet etch and cleaning regime is changed, which might lead to a different STI oxide / active step at the STI edge. The pre-cleaning done before the thin oxide growth has been be reviewed and adjusted, as it has an immediate influence on the tunnel oxide. It has been tried to keep the total HF wet etch time applied to the CMOS devices almost unchanged, to minimize the impact here. The decision if the influence of the changes is acceptable can strictly speaking only be done after reliability and yield testing of the CMOS devices.
(2) The added thermal budget. Three process steps must be noted when counting the thermal budget of the process: the interpoly dielectric formation, the poly silicon deposition and the silicon nitride deposition. Because of the process integration scheme, the thermal steps have no influence on the critical source and drain junctions, which are formed later in the process. Only the well doping profile might be affected, but the damage of the implantation is already annealed and the dopands are activated at this stage, so that the impact can be neglected, which is also seen in the electrical measurement results presented in chapter 5.
(3) A longer nitride wet etching is required. The nitride wet etching step used in the original CMOS flow to remove the ARC layer, which was used at the gate lithography, has to be prolonged. The silicon nitride hardmask on top of the control gate has to be safely removed at this place. A longer nitride wet etching can be done with low risk, as the selectivity of the phosphoric acid towards the other materials is very high.
(4) A thin thermal oxide is formed on the first poly silicon layer. The formation of the interpoly oxide in three steps, a thermal oxide growth, followed by an LPCVD oxide deposition and densification, leads to a slight oxidation of the surface of the first poly silicon layer. This leads to a loss in poly silicon thickness and maybe a slightly enhanced surface roughness, due to the grain structure of the poly silicon. As the grown oxide is very thin, only about 2 nm, these effects can be neglected.
(5) The additional HVMOS LDD ion implantation, which was introduced to raise the breakdown voltage of the HVNMOS transistors, also slightly dopes the first poly silicon layer in the area where the later CMOS transistor gates are formed. The dose of the implanted ions is very low compared to the ion implantation of the source and drain junctions (by orders of magnitude), which later in the process dopes the gate poly silicon finally. Thus is does not significantly change the final doping of the CMOS transistor gates, which is also indicated by electrical measurement results.

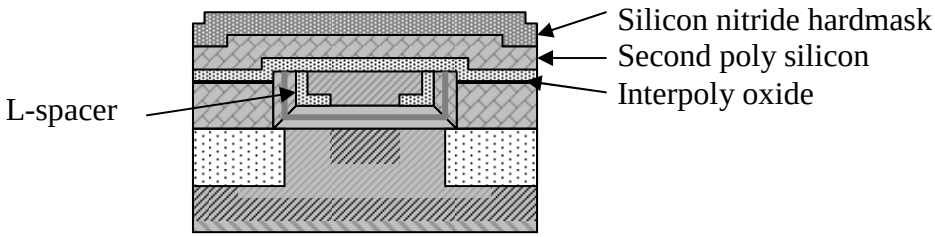


Figure 47: Schematic cross section of the HBT covered by the additional layers deposited during flash memory fabrication

The other CMOS process steps are unchanged. The etching of the contact holes was reviewed, as the stacked gate structure has a higher topology than the other devices, but it was found that no change in this process step was necessary.

Two points can be identified where the embedded flash process interferes with the HBT processing:
(1) Layer deposition and removal. The major part of the flash memory integration is done after the HBT has been build. The additional layers that are deposited to form the flash cells are also deposited and removed again from the HBT. Deposited are the interpoly oxide, the second poly silicon layer and the silicon nitride hardmask, see Fig. 47. They are removed again during the control gate etching. The last part of this etching procedure, the removal of the interpoly oxide, immediately touches the HBT. Due to reasons explained above this step is performed as a combination of dry etching and HF wet etching. The wet etching is the last part. This step is highly selective towards silicon and silicon nitride, only the oxide is significantly removed. The only oxide present at the HBT surface is that of the L-shaped spacers that separate the emitter from the external base. These spacers are indeed partly removed by the added HF cleaning, as can be seen in Fig. 48. Measurements of the device characteristics and yield did not show a negative influence of this fact, which might be because these notches are filled again with oxide later in the process.
(2) The additional thermal budget described above also affects the HBT. The possible impact is on the doping profiles in the intrinsic HBT layer stack, especially the B doped base layer. It must be mentioned at this place, that the carbon doping of the SiGe base layer has been

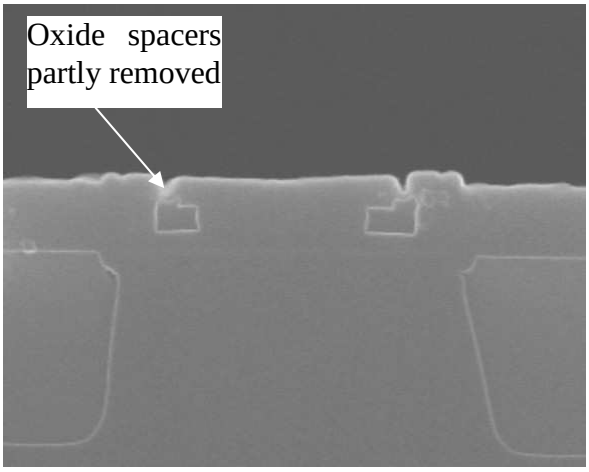


Figure 48: HBT after flash memory processing and CMOS ARC deposition

originally introduced to reduce the boron out-diffusion out of the base layer. This fact makes the proposed integration scheme (flash after HBT) possible. An investigation of the electrical parameters is presented in chapter 5, also with this respect. It could be shown that the flash memory integration has no negative impact, neither on the electrical parameters of single devices nor on the yield of HBT arrays.
In summary, as the discussed process changes do not lead to a critical change in the BiCMOS device parameters, the presented process is here regarded as a modular integration.

Chapter 5

Device Characterization

In this chapter the electrical properties of the individual devices will be presented and discussed. The flash memory cells themselves will be investigated, including separate characterisations of the tunnel oxide and of the interpoly oxide. Static, dynamic and reliability behaviour will be discussed. DC-measurement results of the HVMOS transistors will be presented. Finally, the characteristic of the CMOS transistors and the HBTs prepared with the BiCMOS/embedded flash process will be compared to devices prepared with the BiCMOS process only, to demonstrate the modular character of the technology.
A summary of the geometrical dimensions of the investigated test structures is given in appendix D. Parameters used for simulations are listed in appendix C, unless not differently stated in the respective figures.
The I(V) measurements of all devices, except flash memory cells, have been performed using an Agilent 4156c Parameter Analyzer. Parametric measurements and all measurements performed at flash memory cells have been done using a Keithley S600 parametric tester, combined with a programmable pulse generator for cell writing and erasing.

5.1. Flash memory cells

The different flash memory cells have been investigated, starting with separate characterisations of the tunnel oxide and the interpoly oxide at capacitor structures, followed by the presentation of static and transient characteristics of the complete flash memory cells and finally basic tests with respect to the flash memory cells' reliability.
A tunnel oxide thickness t_{oxg} given in a figure is the average of values measured electrically after appendix E on different sites of one wafer.

5.1.1. The tunnel oxide

The tunnel oxide is the key element of the floating gate cell. It has to fulfill the compromise of being thin enough to let electrons enter the floating gate at reasonable programming voltages, while preventing the electrons to leak off again at low voltages. The latter must also be guaranteed after the FN-stress by repeated write/erase cycles. The compromise is possible due to the extremely steep current-voltage characteristics, described by the Fowler-Nordheim tunneling equation (see appendix A). The electron current that passes the oxide barrier at high electric fields drops rapidly towards lower electric fields. The FN-current at high field strengths is determined by the electric field at the injecting electrode side of the oxide (the electric field can vary within the oxide from substrate to gate at the presence of oxide charges). This field determines the shape of the barrier, especially its width, and thus the tunneling probability [5][84]. The leakage current at low fields, which determines the flash cells retention, is an electron transport via oxide defects (Fig. 53). The most important is the so-called SILC [40], a trap-assisted tunneling current via electron traps. The traps are generated by the FN current passing through the oxide during the memory operation. SILC has weaker electric field dependence than FN tunneling and is thus dominating at low electric fields, while FN tunneling is observed at high electric fields.

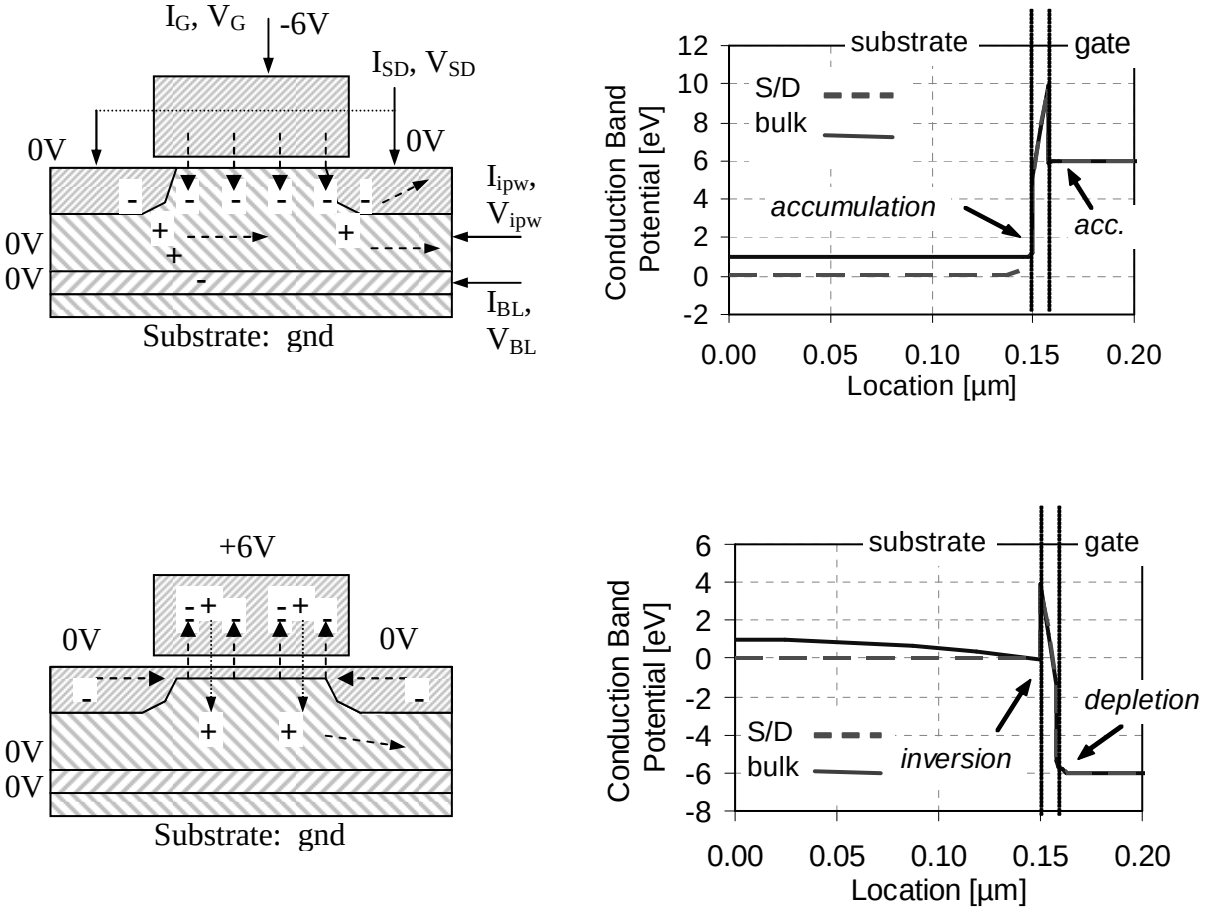


Figure 49: Current components and band diagrams of a MOS structure

In order to have a direct access to the tunnelling current, a MOS-transistor like test-structure has been build (app. D, "large area"-device), with the first poly-silicon layer as gate electrode, with the tunnel-oxide as dielectric layer, and with the source, drain and wells made with the same processing steps as the flash cells. It must be noted that the doping concentration of the gate poly silicon is higher than that of the floating gate in flash cells, as in this test-structure it is doped twice: firstly together with the floating gate, secondly together with the source and drain junctions. The tunnel oxide area is significantly bigger than the tunnel oxide area in the flash cells, to allow the observation of low currents at low voltages, and to produce a sufficiently high capacitance for CV-characterisation. The poly-Si gate is completely surrounded by the S/D junctions here, so that there is no separate contact for source and drain. Fig. 49 shows a schematic cross section of the structure, together with the energy band diagrams showing the conduction band bending in the bulk region and in the source/drain-gate overlap region of the device for positive and negative gate-voltage. The calculation of the oxide electric field in every point of the tunnel oxide under different bias conditions is a complex task, and only completely possible by 2D device simulation. In a general view, at high positive gate voltages versus source, drain and p-well (= cell writing), the silicon surface under the oxide is inverted in the p-well area, and the accumulated S/D areas determine its potential. At V_G=0V (not shown in the figure), the oxide field in the p-well area is slightly higher than in the S/D areas due to the higher flat-band voltage. When the gate voltage becomes negative, the oxide field in the p-well area first becomes 0V/cm at flat-band conditions, and then catches up with the electric field of the S/D areas. The field strength rises faster in the p-well area, because the silicon surface is accumulated here, while in the S/D areas a depletion region builds up. Finally, at very high negative gate voltages (= cell erasing),

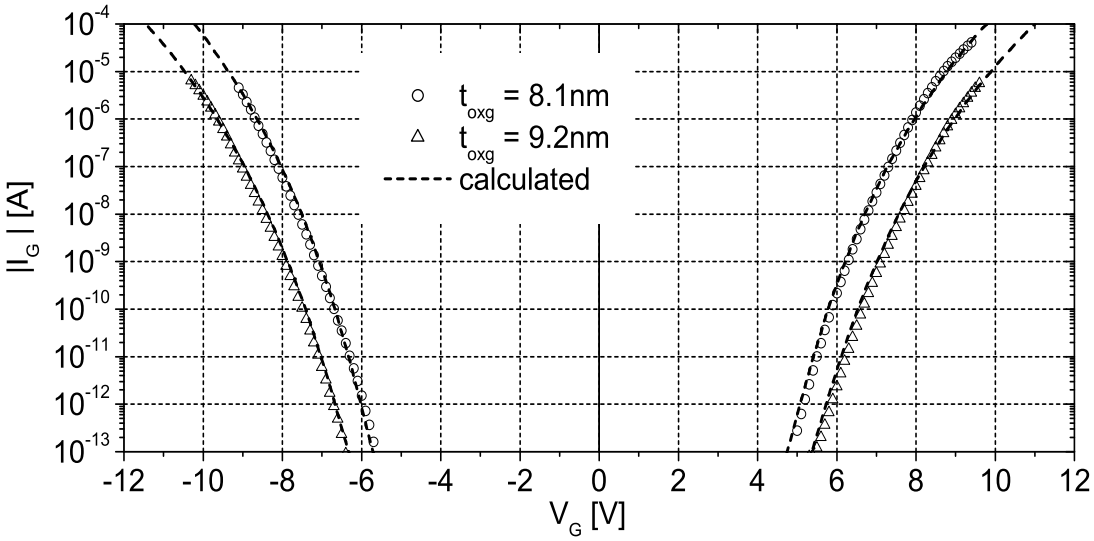


Figure 50: Tunnelling current through a MOS structure for different oxide thicknesses; parameters for the calculated curve see appendix C

an inversion layer forms in the S/D areas, and the whole surface potential is controlled by the accumulated p-well area. An efficient method to calculate the oxide electric field has been developed and is presented in appendix A.

In Fig. 50 the measured gate current is shown in dependence of the applied gate voltage. The measured curves of devices with two different oxide thicknesses are shown. The strong dependence of the current on the oxide thickness can be seen: for about 1nm difference in thickness the current differs more than one decade. The measurements are compared to calculations after appendix A. Over the whole presented voltage range an FN-like behaviour is observed in good agreement to the model. As the oxide is not stressed, no SILC component is detected in the observed current range of $I_G > 10^{-13}$A.

In Fig. 51 the current components measured at the gate contact, the S/D contact and the (isolated) p-well contact are shown for different V_{SD} for positive gate voltages V_G. For highly positive gate voltages the tunnelling current is mainly supplied from the S/D junctions, which supply electrons to the inversion layer. A p-well current is measured, which is proportional to

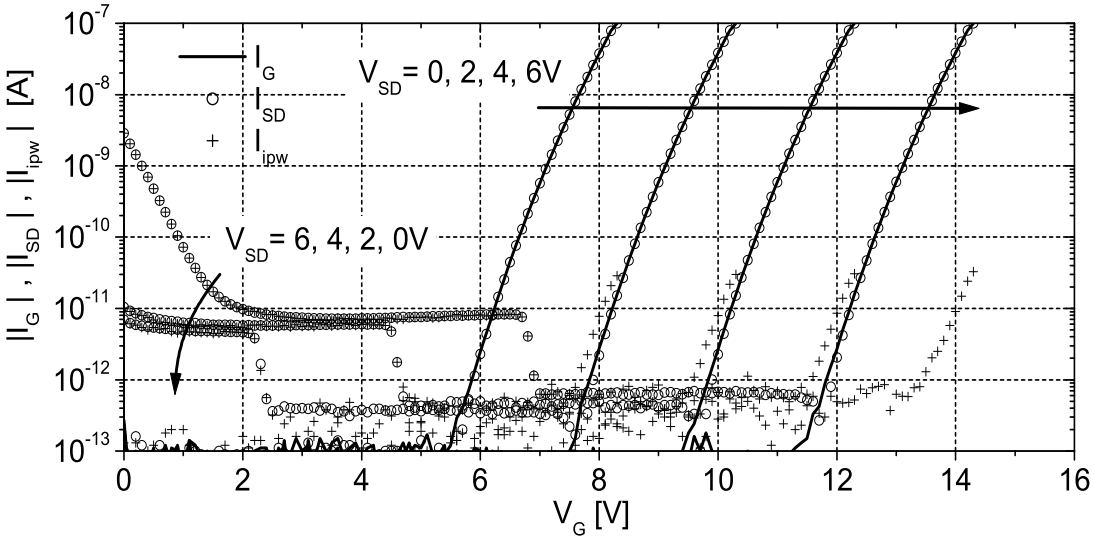


Figure 51: Measured gate, p-Well and S/D current components for different values of V_{SD}; t_{oxg} = 9.2nm

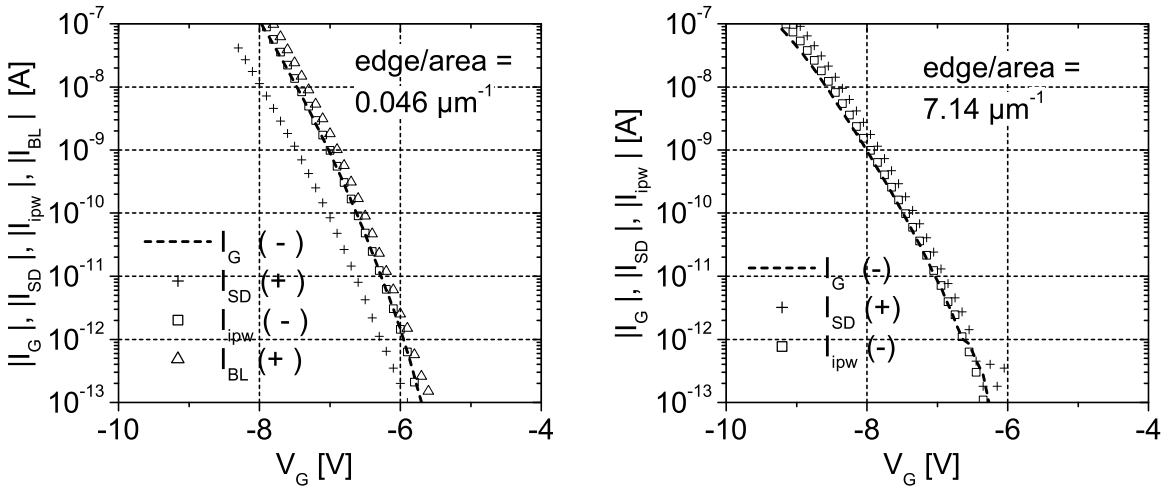


Figure 52: Gate, S/D, p-well and buried layer current components at negative gate bias for MOS-structures with different gate-edge/gate-area ratios; the signs in the legend indicate the direction of the current, "+" means a current flowing into the contact (Fig. 49); t_{oxg} = 8.1nm

the S/D current. This can be explained by hot holes tunnelling from the gate into the p-well and by photons leading to an electron-hole pair generation in the p-well [85]. The hot holes stem from electron/hole pairs generated in the gate poly-Si by the high-energy electrons of the primary tunnelling component (Fig. 49), which also lead to the photon generation. For high V_G (well above V_{SD}) applying an S/D voltage V_{SD} merely leads to a parallel shift of the curves, showing how the whole oxide field is controlled by the potential of the inversion layer. This voltage regime is present during cell writing. For the selected cell V_{SD} is on the same potential as the p-well, for the unselected cell the tunnelling is suppressed by the applied V_{SD} (chapter 2). For lower V_G a significant increase of I_{ipw} and I_{SD} can be seen with increased V_{SD}. This is due to the onset of BTBT. The BTBT is a major reliability concern for source erasing or drain erasing schemes, as it is regarded as an origin for oxide degradation during programming. In the presented array operation scheme using uniform channel erasing, this regime is avoided, leading potentially to a high reliability of the flash cells.

The different current components at negative gate bias are presented in Fig. 52. Measurement results from two different MOS structures with different total oxide areas are shown (app. D). An important difference between both test structures is the ratio of the total gate oxide area and the length of the gate poly-Si edge (= length of the SD junctions, L_{SD}). The standard structure of Fig 52a has a relatively large poly-Si electrode surrounded by an S/D junction. Thus, the fraction of S/D junction area of the total gate area is relatively small. The structure of Fig. 52b consists of 1000 MOS-transistor-like structures connected in parallel, each with dimensions similar to the floating gate of one flash cell. Due to the small gate length, the ratio of the S/D junction area and the total gate area is significantly higher here and comparable to a flash cell. The edge/area ratio given in Fig. 52 is the length of the poly-Si gate edge divided by the poly-Si gate area. While the different oxide area leads to a different absolute value of the current, the different edge/area ratio results in a different ratio of the current components measured at the different terminals of the device.

The gate current is in both cases determined by the FN-injection. In Fig. 52a it can be seen that the highest current is I_{BL}. This current component stems from electrons that tunnel through the oxide and generate electron/hole pairs in the p-well due to their high energy. Both, the tunneling electrons and the generated electrons are mainly collected by the underlying (n-doped) buried layer and add together to the current measured at the buried layer contact. A fraction of the electrons also goes to the S/D junctions, but I_{SD} is much lower than I_{BL} due to the low edge/area ratio in this test structure. The generated holes are detected at the p-well contact and lead to I_{ipw}.

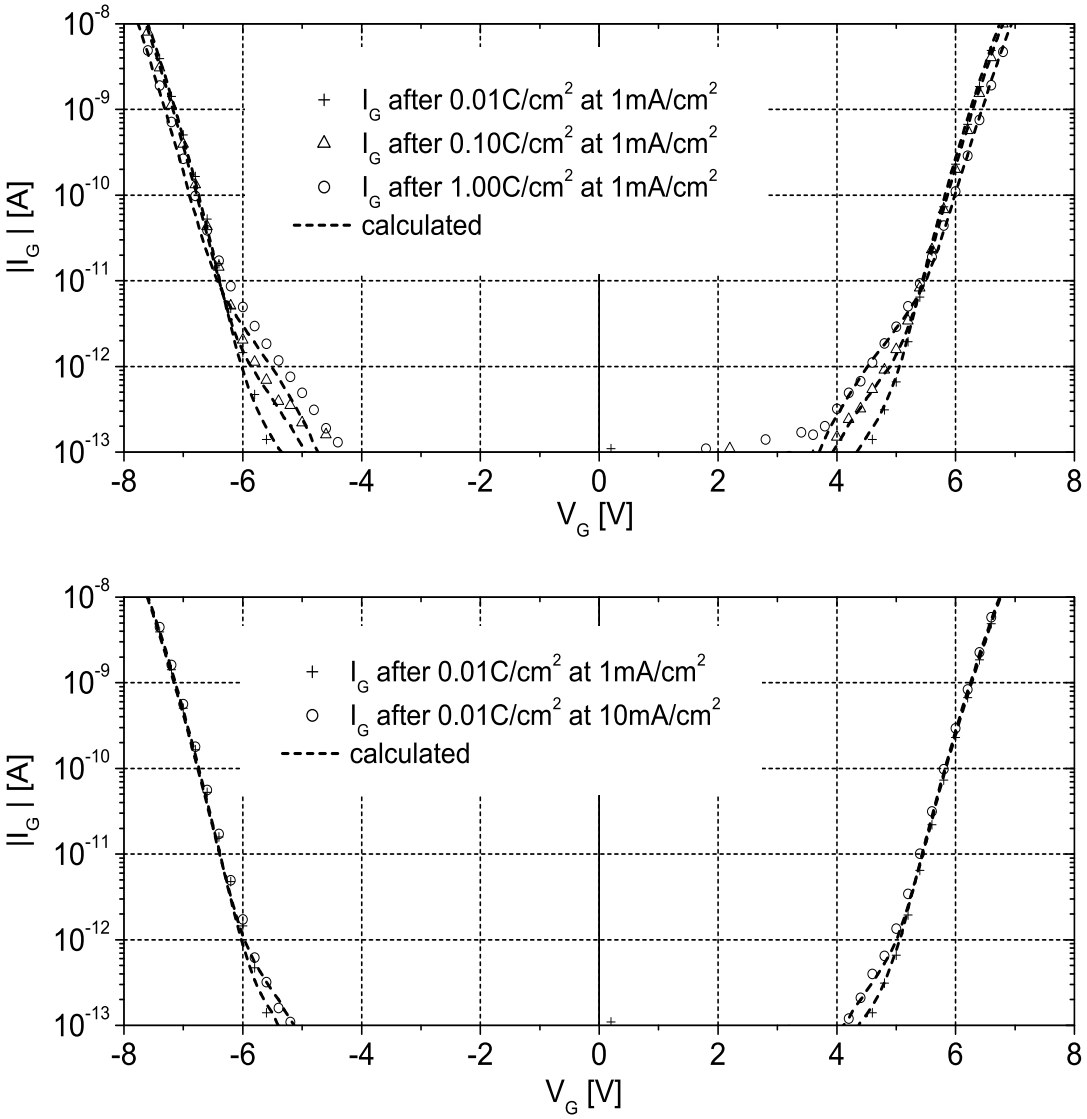


Figure 53: Gate current measured after constant current stressing; t_{oxg} = 8.1nm

For a different edge/area ratio the distribution of the current is different. Due to the significantly larger edge/area ration of the device of Fig. 52b, the PN-junction of the S/D implantation plays a major role here. The electrons are collected at the S/D junctions rather than in the buried layer. So in this case I_{SD} has the highest absolute value here, while I_{BL} could not even be measured ($I_{BL} < 10^{-13}$A). I_{ipw} has a higher value than I_G in this case, so the generated current is even higher than the injected tunneling current.

An important issue for the reliability of a flash memory is the degradation of the tunnel oxide with repeated programming. Fig. 53 shows the change in the current/voltage characteristics after a constant current stress. The FN current leads to a generation of electron traps within the oxide. This has two important consequences: (1) the traps lead to an additional current component by trap-assisted tunneling (SILC), and (2) electrons that are captured in the traps charge the oxide and lead to a reduction of the electric field at the emitting interface. The consequence of (2) is a reduced tunneling current at high electric fields. For the flash cells this means that, after repeated programming, (1) the retention decreases and (2) the programming becomes slower. The trap generation has in literature found to be dependent on both, the amount of charge that has been injected through the oxide and the current density of the FN stress [40]. Fig. 53a demonstrates the dependence of FN current and SILC on the injected charge for a constant stress current density, while in Fig.53b the influence of the current

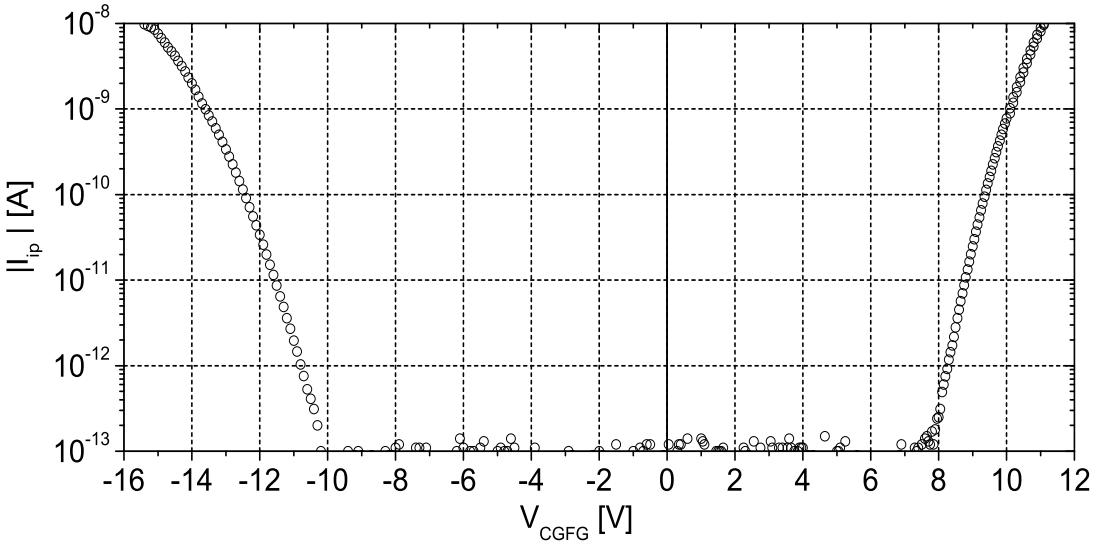


Figure 54: Interpoly oxide current measured at a patterned poly-poly capacitor

density is shown for constant current stressing with the same total injected charge. A comparison with calculations shows a good agreement with an empirical model for SILC presented in [40] (appendix A).
It must be mentioned that the oxide degradation effects seen at this MOS test-structure show the large-scale behaviour. As trap generation is a small-scale, statistical effect, in a memory matrix of millions of cells it thus different for each cell, while the MOS test structure shows the average behaviour. On top of a usual Gaussian distribution of the regular SILC current in flash cells, which is thought to be trap assisted tunnelling with help of one trap in the path of an electron, there is a statistical tail of cells that show anomalous SILC [33], which is thought to be caused by more than one trap involved per leaking path. Different kinds of such leaky cells exist [82], and also the programming and erasing distribution has a statistical tail due to fast cells. To observe such effects, array-like test-structures and statistical methods are necessary.

5.1.2. The interpoly oxide

The interpoly dielectric has to fulfil the compromise of producing a high capacitive coupling between the control gate and the floating gate, while preventing a leakage current between both. As explained in chapter 3, the interpoly dielectric is here an oxide layer that is produced by combined thermal and LPCVD oxide formation. In most flash memory technologies, an LPCVD ONO stack is used for this purpose. The reason for this is the better ratio of capacitance and leakage compared to a thermally grown poly-oxide that was used earlier. It has especially an advantage due to the topology of the floating gate. Electrons residing in traps at the oxide/nitride interface shield the enhanced electric fields at the corner of the floating gate [41]. Here, the use of ONO is in principle also possible, but would need some technological adjustment. The use of an oxide layer only results in a simpler process, at the cost of programming speed due to reduced scalability.
From statistical reliability investigations of flash cells (which are not part of this thesis) the minimum allowed interpoly oxide thickness must be determined. Here, a value of around 23nm has been chosen as a first guess, which is also a reasonable value for the gate oxide of the HVMOS transistors.
Fig. 54 shows the measurements of the current through the interpoly oxide. The test structure is a large CG/FG capacitor (app. D). It can be seen that the current is lower at negative gate bias compared to the current measured at the same absolut positive gate voltage. This is due to the enhanced electric field at the FG edge produced by the FG-etch slits. The current is for both, positive and negative gate voltages higher than calculated by the regular FN equation for

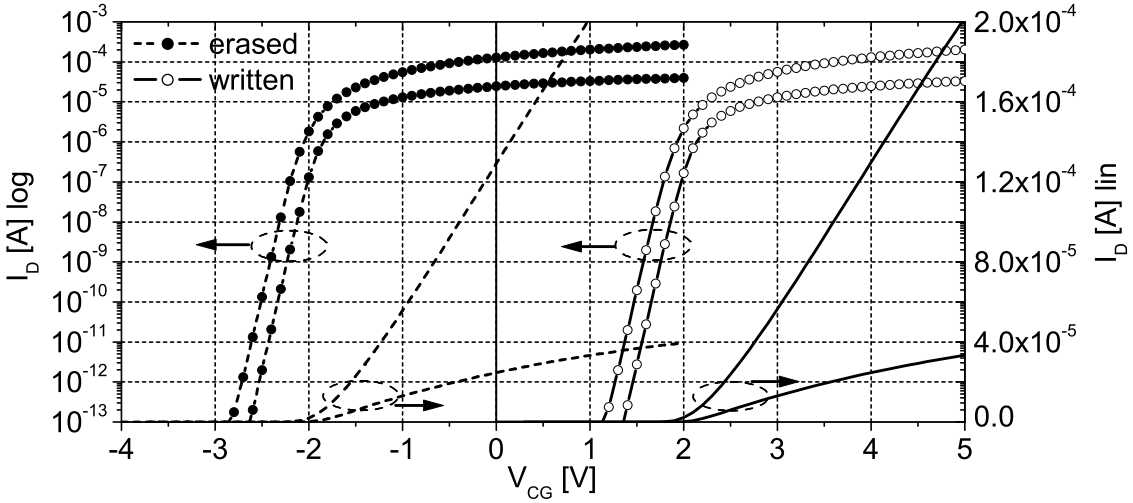


Figure 55: Transfer characteristics of the written and erased 1-transistor cell; curves for V_{DS}=0.1V and V_{DS}=1.5V in logarithmic and linear scale; t_{oxg} = 8.1 nm

the present oxide area and thickness (appendix A). If from the measured current the leakage of one memory cell is calculated and compared to the leakage of the tunnel oxide, then it can be concluded that the interpoly oxide does not impact the reliability. It must be mentioned that this has to be verified by reliability measurements of flash cells, as neither statistical effects nor low-field behaviour are measured here.

5.1.3. Static characteristics of the flash cells

The static behaviour of a flash memory cell determines the reading operation of the memory. Two main parameters can be identified. (1) The current of an erased cell (on-current) that is supplied to the sense amplifier during the read operation. The value of this current should be as high as possible to allow a fast random read access, this means a fast decision if the cell is written or erased. The higher the current the faster the parasitic capacitances are loaded and the faster this decision can be made by the sense amplifiers. (2) The leakage current of a written cell (off-current). This current must be low enough to ensure that the total leakage of all unselected cells in a bitline is significantly lower than the on-current of the selected cell, so that for the selected cell the written and erased state can be distinguished.
Fig. 55 shows the measured transfer characteristics of the written and of the erased 1-

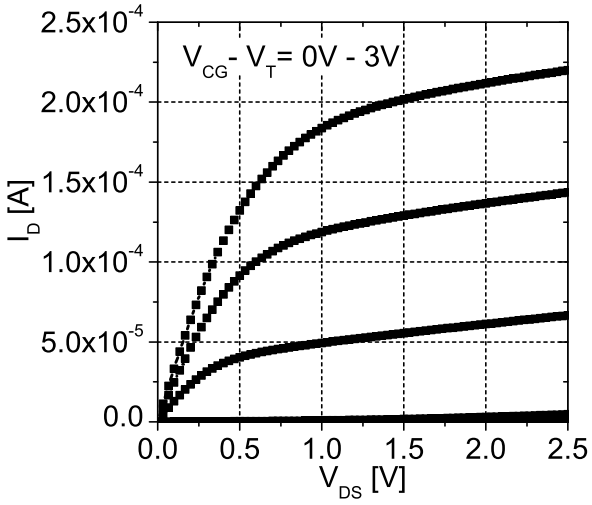


Figure 56: Output-characteristics of the 1-transistor flash cell; V_{CG} changed in 1V steps; t_{oxg} = 8.1 nm

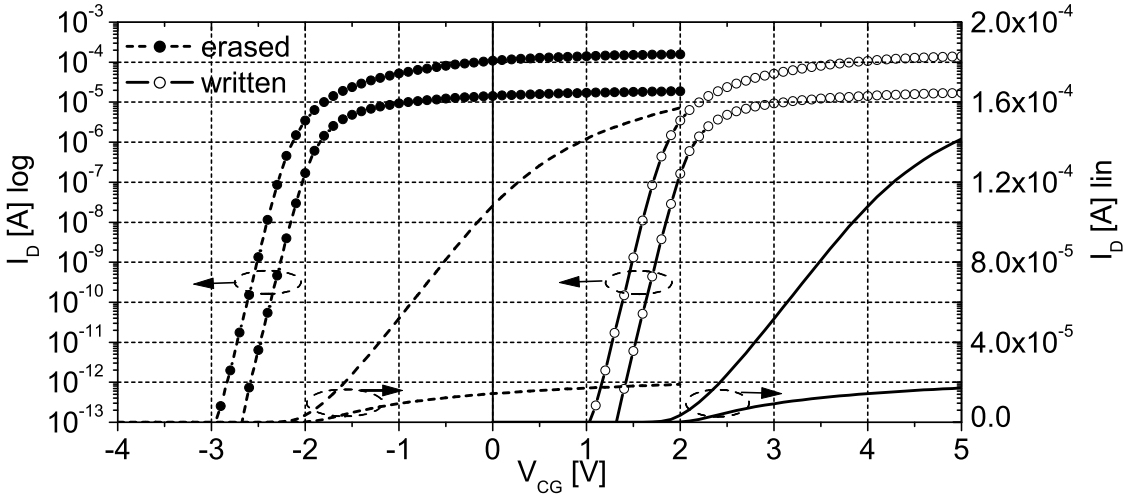


Figure 57: Transfer characteristics of the written and erased 2-transistor cell; curves for V_{DS}=0.1V and V_{DS}=1.5V in logarithmic and linear scale; V_{SG}=2.5V; t_{oxg}=8.1nm

transistor cell. Curves for V_{DS}=0.1V and V_{DS}=1.5V are shown. At V_{DS}=0.1V the V_T of the cell is determined, while V_{DS}=1.5V is a typical value for the reading operation. For reading at V_{CG}=0V and V_{DS}=1.5V a drain current of about I_D=130µA is measured at the erased cell, while the off-current of the written cell is less than I_D=0.1pA. It can be seen that the reading current could be raised if the cell would be erased to a more negative V_T (I_D=160µA for 0.5V lower V_T). The subthreshold behaviour shows no parallel leakage path, e.g. stemming from the trench corner or similar effects. The difference between the curves for V_{DS}=0.1V and V_{DS}=1.5V has two origins, firstly the capacitive coupling between the drain and the floating gate, and secondly short channel effects. Due to the drain/FG coupling the FG potential is raised together with the applied drain voltage and the drain current becomes higher as well.
The limitations of $V_{CG,r}$ and V_{off} for the memory operation become visible here. $V_{CG,r}$ must be at least about 1V below $V_{T,w}$, to be sure to have no current when reading a written cell. This margin must even be higher due to the statistical distribution of $V_{T,w}$ within one memory. On the other hand, $V_{CG,r}$ must be well above $V_{T,e}$, to have a high reading current. V_{off} must be at least about 1V below $V_{T,e}$, or more depending on the statistical $V_{T,e}$-distribution. This makes the need for a sensitive adjustment of the operation conditions of the 1-transistor cell clear. If

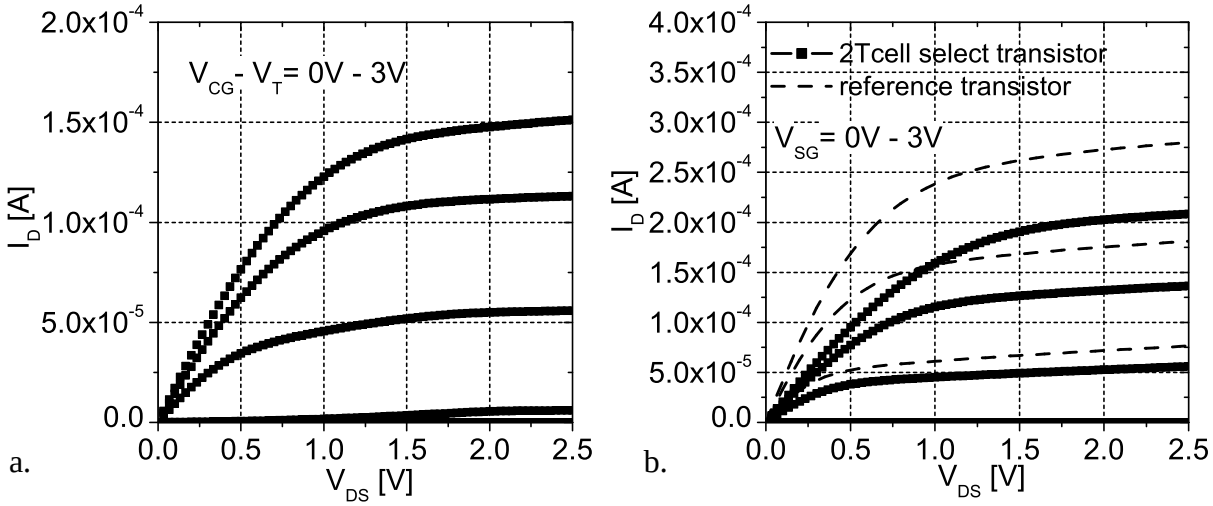


Figure 58: Output-characteristics of the 2-transistor cell; a. memory cell, with V_{SG}=2.5V; b. select-transistor, with V_{CG}=V_T+4.5V, and reference transistor, with normalized I_D (multiplied by W_{SG}/W_{ref}) ; t_{oxg} = 8.1 nm

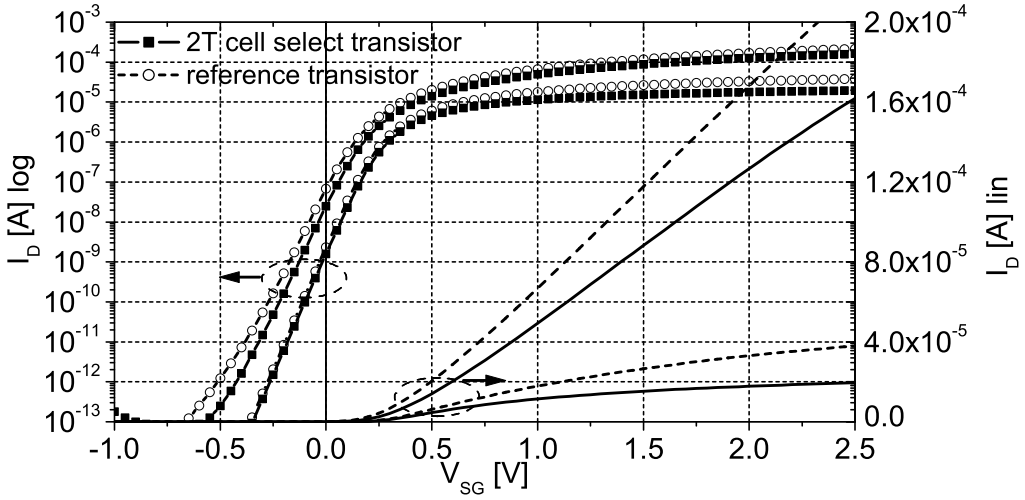


Figure 59: Transfer characteristics of the select transistor of a 2T cell, compared to a reference transistor without a flash cell in series; curves for V_{DS}=0.1V and V_{DS}=1.5V in logarithmic and linear scale; flash-cell: V_{CG} =V_T+4.5V; I_D of reference transistor normalized by ($W_{SG}/W_{ref.}$) ; t_{oxg} = 8.1 nm

no negative voltages are generated during reading, then the maximum difference between V_{off} and $V_{CG,r}$ can only be V_{DD}. $V_{T,e}$ must be >1V above V_{off}, so the resulting distance between $V_{T,e}$ and V_{read} limits the reading current. The p-well of the flash cells would have to be adjusted to produce an intrinsic V_T close to $V_{CG,r}$ to have acceptable programming times, but which would then not be the optimum choice for reliability. The introduction of a negative voltage (<-3V) for V_{off} relaxes all these constraints and allows a V_T-window as presented in Fig. 55, at the cost of a higher power consumption during reading and more complexity in the peripheral circuitry. This measure has been chosen at the memory chip presented in chapter 6.
Fig. 56 shows the measured output characteristics of the 1-transistor cell. In the presented normalized view (V_{CG}-V_T) the characteristics is the same for both, the written and the erased cell. The increase of I_D with raising V_{DS} in the saturation region can be ascribed to the capacitive coupling between the drain and the floating gate [12].
The transfer characteristics of the 2-transistor cell are presented in Fig. 57 and Fig. 59, showing the behaviour of the floating gate transistor and the select-transistor, respectively. For the written cell, the same considerations as for the 1-transistor cell are valid. Here also a minimum of 1V difference between $V_{T,w}$ and $V_{CG,r}$ is necessary. The difference of the currents in the subthreshold region between V_{DS}=0.1V and V_{DS}=1.5V is slightly higher than for the 1-transistor cell, which is due to a slightly shorter L_{FG} in this test device, leading to enhanced short-channel effects. For the erased cell, only the reading current is important. Turning off the unselected cells is done by the select-transistor. Fig. 58 shows the output characteristics of the flash cell and the select-transistor of the 2-transistor cell. The reading current I_D of the flash cell at V_{DS}=1.5V is limited at high V_{CG} due to the select-transistor that is connected in series. The select-transistor (measured with V_{CG}=V_T+4.5V applied to the flash cell) is compared to the reference transistor without flash cell in series. The reference transistor has a larger W_G, but for comparison the current is normalized to the width of the transistor combined with the flash cell. The curves first show the slope determined by the drain/FG coupling, before they saturate due to the limitation by the select transistor. In the characteristics of the select-transistor no such limitation is seen, which is due to the relatively high voltage that is applied to the control gate of the flash cell.
The restrictions for the 1-transistor cell with respect to the choice of these voltages are not valid here. As a consequence, if no additional circuitry is used in a 1-transistor flash memory

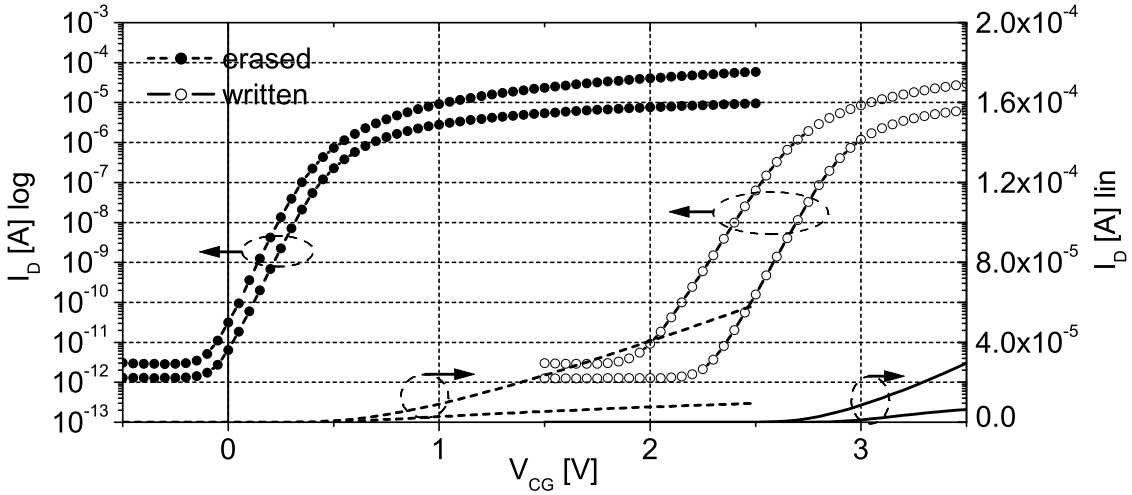


Figure 60: Transfer characteristics of the written and erased split-gate cell; curves for V_{DS}=0.1V and V_{DS}=1.5V in logarithmic and linear scale; t_{oxg} = 8.3 nm

to generate a low V_{off}, the reading current of the 2-transistor cell would be higher. Here for V_{CG}=0V it is about I_D=105µA.
The transfer characteristics of the select-transistor can be seen in Fig. 59. Again, a comparison of a select-transistor within a flash cell and a separate reference transistor without floating gate transistor in series is presented. It can be seen that the current of the separate device is higher for $V_{SG}>V_T$. In the subthreshold region a higher sensitivity to short-channel effects is visible for the reference transistor. The off-current at V_D=1.5V at V_{SG}=0V is about $2x10^{-8}$A. This value is quite high, and shows the requirement of further process optimization if a 2-transistor flash memory should be produced. If 5,000 cells are connected in one bitline, the off-current of the unselected cells is as high as the reading current of the selected cell, so that the state of the selected cell cannot be determined. The maximum allowed number of cells in one bitline depends on the required difference in reading current to fast and safely determine the state of the selected cell and is thus in this case well below 5,000 cells. The doping concentration in the p-well of the flash cell needs to be raised, in order to achieve a higher V_T of the select-transistor and a lower off-current.

The next figures show the static characteristics of the split-gate cell. In Fig. 60 the transfer

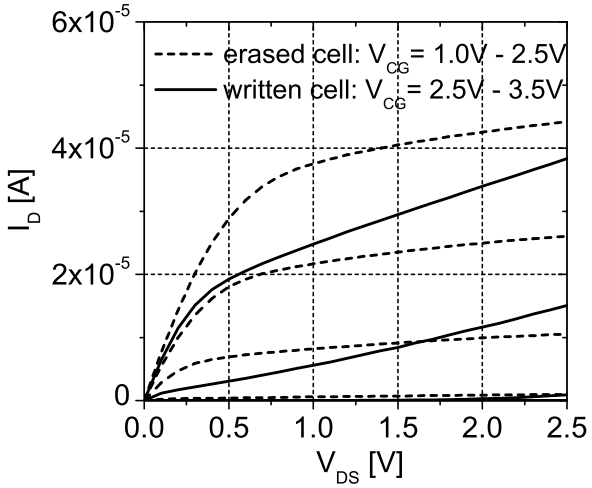


Figure 61: Output-characteristics of the split-gate cell; V_{CG} changed in 0.5V steps; t_{oxg} = 8.3 nm

characteristics is shown for a written and an erased split-gate cell. The reading current is strongly limited by the select-transistor part of the cell. The scale of the I_D-axis in Fig. 60 is the same as in Fig. 57 and Fig. 55 for an easier comparison. It should be noted that the transfer characteristics looks different for the written and the erased cell, which is not the case for the other cell types. This is because for the erased cell, if the floating gate part of the cell has a V_T below that of the select-gate part, the characteristics of the select-gate part of the cell are measured, while in the case of a written cell the floating gate part of the cell determines the behaviour. The lower limit of the $V_{T,e}$ for a split gate cell is fixed by the select-gate part, and no over-erase can take place. Stronger short-channel effects can be seen for the written cell compared to the 1-T or 2-T cells. This can be explained by a reduced p-well doping compared to the other cells, as here the compensation implant for producing the HVMOS transistors has an influence. The implantation is carried out only in the select-gate part of the cell, but the dopands can diffuse also into the region under the floating gate, thus also leading to a compensation of the p-well here. Another reason could be a slightly shorter length of the floating gate due to misalignment of the CG-etch and the FG-etch masks. The problem could be solved by an adjusted (higher) p-well doping concentration, which would also reduce the writing time and allow higher $V_{T,w}$ values.
The relatively high off-current that is seen in the figure must be attributed to a layout error in the test-structure. This error leads to a V_D-dependent leakage current flowing parallel to the drain current of the cell. Due to this, the off-current of the split-gate cell could not be determined within this work.
The output characteristics of the written and erased split-gate cell are shown in Fig. 61. The effect of capacitive coupling between the drain and the floating gate can be seen in the characteristics of the written cell. This coupling has no effect on the select-transistor part of the cell and is therefore not observed in erased cells.

Comparing the static characteristics of the different cell types it can be concluded that the different cell types need a separate adjustment of the process. Especially the doping level of the p-well needs to be optimized with respect to the chosen cell-type, and with respect to the way of memory operation. The process in its present state can be used for a 1-transistor memory that operates with a generated negative V_{off} or for a 2-transistor memory with a limited number of cells combined in the same bitline. For the other possible options a re-optimization of the process parameters is necessary.

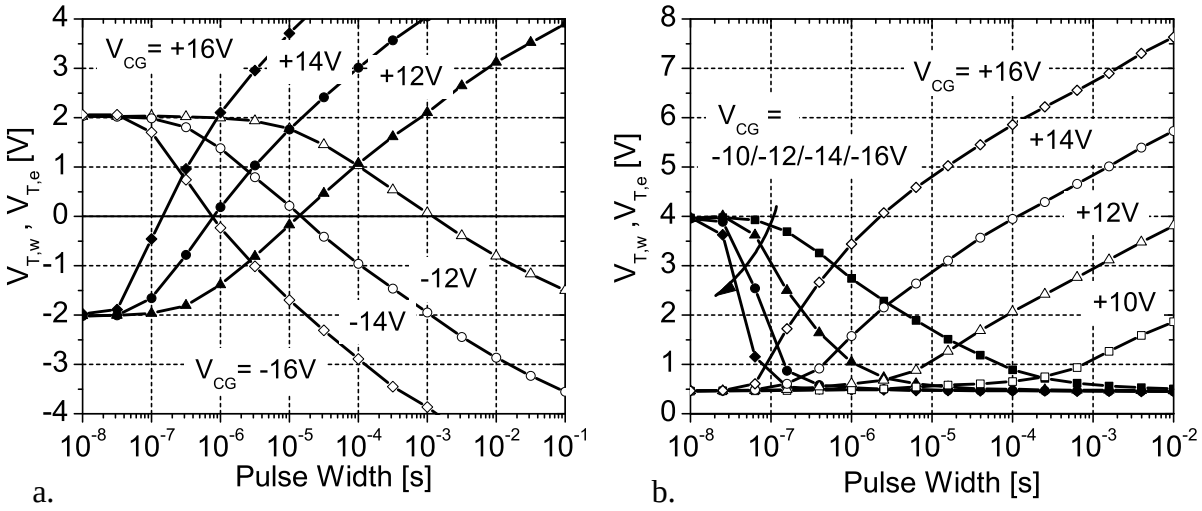


Figure 62: Transient characteristics of the 1-transistor cell (a.) and the split-gate cell (b.); a. t_{oxg}=7.9nm; b. t_{oxg} = 8.8 nm

5.1.4. Transient behaviour of the flash cells

In following, the transient behaviour of the different types of flash cells will be presented. The overall behaviour will be discussed and the influence of different process parameters on the cell programming will be shown.

The transient behaviour describes the change of the flash cell's characteristics with time when the programming voltages are applied to the different terminals. The effect of programming is a change in the amount of charge that resides on the floating gate (the degradation of the oxide due to the programming current is an issue discussed under reliability). Changing the charge on the floating gate leads to a shift of the transfer curves along the V_{CG}-axis, which is described by a change in V_T. The transient behaviour is analyzed by measuring the $V_{T,w}$ or $V_{T,e}$, respectively, after applying programming pulses of different pulse width. The results of the individual write or erase processes for the same value of V_{CG} are combined to one curve showing the $V_{T,w}$ or $V_{T,e}$ versus the pulse width. In a curve showing the write behaviour, the cell is erased again between each point of the curve, and in a curve showing the erase behaviour the cell is written again between each point of the curve (re-set pulses). To solve the problem that at for short pulse width the applied programming pulse does not necessarily change the state of the cell and the cell's V_T would be changed by the re-set pulse between the measured programming pulses, an additional pulse is applied before the actual re-set that in case of a writing curve first definitely writes the cell before the re-set, and that in case of an erasing curve first definitely erases the cell before the re-set. Thus, as the final V_T after programming does not depend significantly on the initial state of the cell for long enough pulses (Fig. 64), the cell is in the same state again before each measurement point belonging to one programming pulse.

An important parameter for programming is the programming voltage, being the voltage between the control gate and the source, drain and p-well (S, D and p-well are all on the same potential at selected cells). During the measurements, S, D and the isolated p-well have been on the same potential as the substrate, so that V_{CG} is the programming voltage.

Comparing the different cell types it must be noted that there is no principal difference between the 1-transistor cell and the 2-transistor in the programming behaviour. Only in the present device realizations the cell layouts are different, leading to different coupling ratios and thus different programming speeds. In the following investigations only the behaviour of one type is shown, e.g. when showing the influence of process parameters, as it is qualitatively the same for both kinds of cells. The split-gate cell behaves in principal also in

the same way, with the only difference that there is a lower limit for the $V_{T,e}$, which is given by the V_T of the select-gate part of the cell.

Fig. 62 shows the programming curves of the 1-transistor cell and the split-gate cell within a V_T-range that is reasonable for memory operation. The strong dependence of the required programming time to reach a specified $V_{T,w}$ or $V_{T,e}$ on the programming voltage (=V_{CG}) can be seen. To write the cell to a $V_{T,w}$=2V, a programming pulse of 1ms is required for V_{CG}=12V, while it becomes as short as 1µs for V_{CG}=16V. Erasing the 1-transistor cell to $V_{T,e}$= -2V requires a longer pulse than reaching $V_{T,w}$= +2V at the same absolute programming voltage. It can be seen how the chosen V_T-window determines the required programming times. As demonstrated below, changing V_{Ti} with an adjusted p-well doping has an influence on this behaviour.

One important characteristic in the transient behaviour of the split gate cell is the limited $V_{T,e}$. The split-gate cell cannot be erased below the V_T of the select-gate part of the cell. If the V_T of the floating-gate part of the cell is below this value, then only the characteristics of the select-gate part are measured. The minimum $V_{T,e}$ is 0.45V for the measured cell.
As the control gate covers also the poly-Si sidewall of the floating gate, the capacitive CG/FG coupling is higher than in the 1-transistor cell, leading to higher electric fields in the tunnel oxide and thus faster programming. It can be seen that at V_{CG}=16V the cell is already written to $V_{T,w}$=3.5V, compared to $V_{T,w}$=2V for the same conditions at the 2-transistor cell. This is achieved despite the fact that the p-well doping concentration is lower in the split-gate cell, and the tunnel oxide is thicker in the investigated device. Erasing is due to the high k_{CG} and low $V_{T,i}$ significantly faster than in the 1-transistor cell. The narrowing of the erasing-curves for high absolute V_{CG} is due to the shape of the erase pulse. It has been found that the rising front of the programming pulses has a time constant of about $\tau_{prg}=1.5\text{x}10^{-8}$s, so that V_{CG} is not at its final level in this time range. The erase curves for V_{CG}=-14V and V_{CG}=-16V are thus showing a slower behaviour as it would be the case for rectangular pulses.
An adjustment of the p-well doping would lead to a more reasonable ratio of write time/erase time and to an enhanced reliability. This underlines the necessary adjustment of the technology with respect to the chosen kind of flash cell.

Fig. 63 shows the transient characteristics of the 1-transistor cell for a wide range of V_T-values and programming voltages. The results of the measurements are compared to calculated curves after appendix A. The excellent agreement of measurements and calculations shows that the mechanisms behind the transient behaviour are well understood.
Some properties of the cell could be investigated with help of the calculations. E.g. the saturation of the $V_{T,w}$ at around 9V for programming with V_{CG}=18V could only be modelled by taking the current through the interpoly oxide into account. When a programming voltage is applied then this current rises with rising V_T of the cell, as the potential of the floating gate drops and thus the electric field in the interpoly oxide rises (appendix A). At the same time the current through the tunnel oxide drops. When both currents have the same value, the charging of the floating gate stops, as the same amount of charge that enters the floating gate on one side leaves it on the other side. Thus the level of the saturated $V_{T,w}$ gives an information about the CG/FG leakage current.
The deviation of the measured $V_{T,e}$ for long programming pulses at V_{CG}=-18V cannot be explained by the current through the interpoly oxide. This is rather due to the onset of oxide charging, which leads to a slower cell erasing (appendix B). The curve at V_{CG}=-18V was the last curve that was measured in this test, so the tunnel oxide is already stressed by the repeated set and re-set pulses.

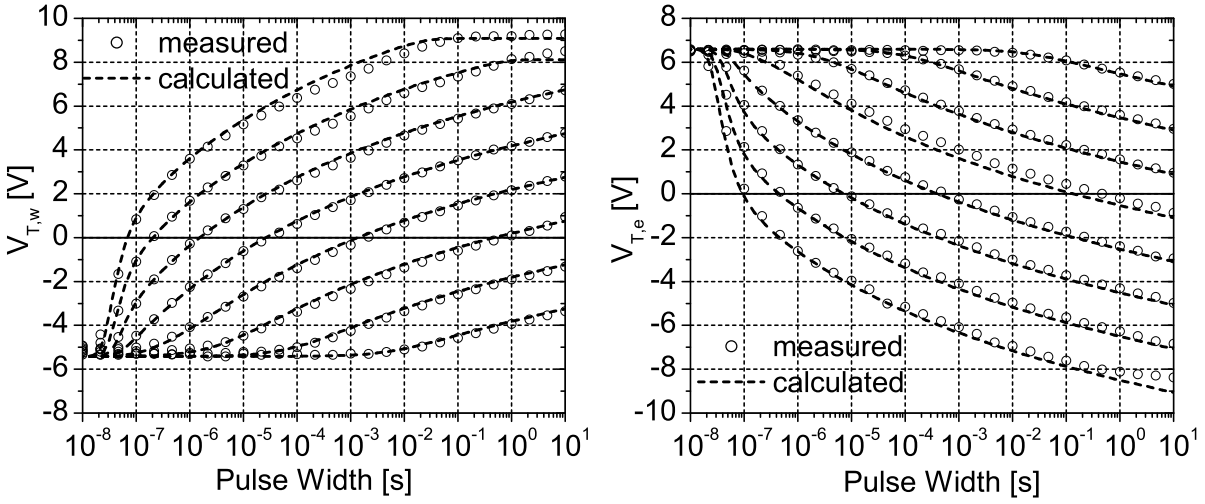


Figure 63: Comparison of measured and calculated programming curves of the 1-transistor cell over a wide range of programming voltages and V_T-values; a. cell writing, V_{CG} is changed from 4V to 18V in 2V-steps; b. cell erasing, V_{CG} is changed from -4V to -18V in -2V-steps; cell parameters: appendix D; simulation parameters: appendix C; t_{oxg}=8.2nm

The steep edge of the curves at short pulse widths can be explained by the shape of the programming pulses. The measured behaviour could be modelled correctly in this region by calculating with a time constant of $\tau_{prg}=1.5x10^{-8}$s at the rising edge of the pulse. It should be mentioned that for matching the writing curves over the whole range of programming voltages and for modelling the writing curves as well as the erasing curves with the same set of parameters it was necessary to take the depletion and inversion of the highly doped poly-Si into account.

An interesting feature of channel FN programming is the insensitivity of the final $V_{T,w}$ and $V_{T,e}$ on the cell's initial V_T before applying the programming pulse. This behaviour is demonstrated in Fig. 64. By adjusting the re-set pulses, the initial V_T before programming has been set to -1V, -2V and -3V, and to 1V, 2V and 3V for writing and erasing, respectively. It can be seen that the curves quickly approach each other. This effect can be used for example to handle the overerase problem in a 1-transistor memory: after erasing a small write pulse

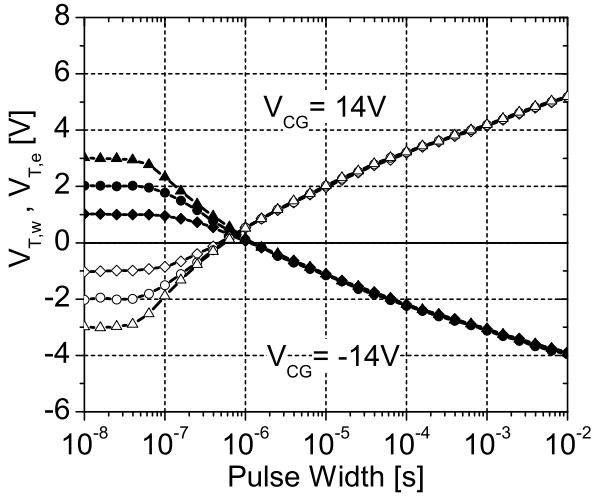


Figure 64: Programming curves of the 1-transistor cell for different initial V_T values; measured at a 1-transistor cell; t_{oxg} = 7.9nm

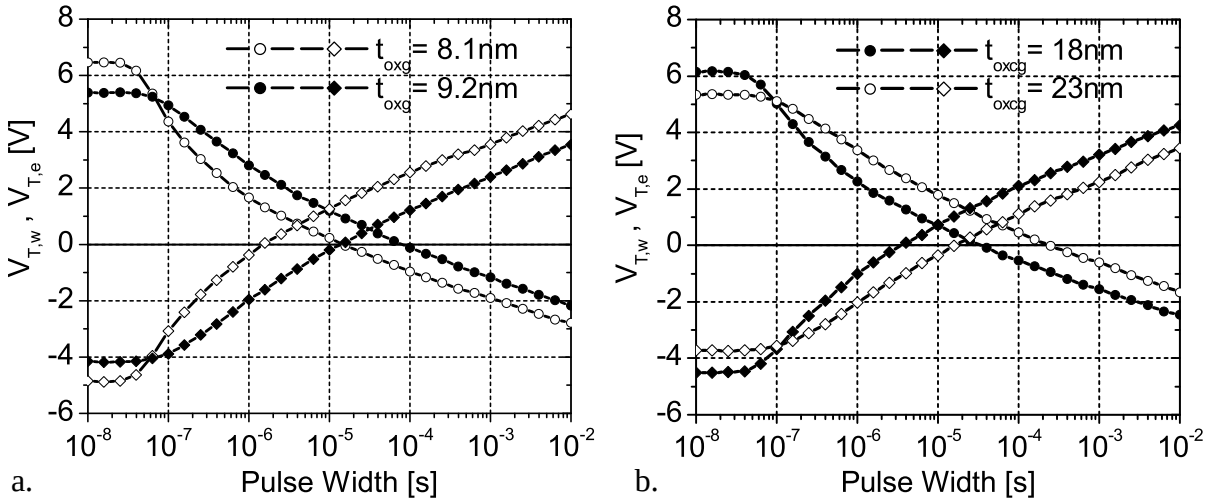


Figure 65: Influence of the tunnel oxide thickness (a.) and the interpoly oxide thickness (b.) on the transient behaviour; programming voltage V_{CG}=+/-14V; 1-transistor cell;
a. t_{oxcg}=23nm; b. t_{oxg}=9.4nm

can be applied to all cells, bringing them to the same minimum $V_{T,e}$ level.

In following the influence of selected process and layout parameters on the transient cell behaviour will be presented and discussed.
Fig. 65 shows the influence of the tunnel oxide thickness and the interpoly oxide thickness. In both cases, the programming is faster for a thinner oxide layer. A thinner tunnel oxide leads to an enhanced electric field within the oxide, as the floating gate potential drops at a much slower rate than the oxide thickness (the potential drops due to the reduced control gate coupling). The electric field depends proportionally on both, the inverse oxide thickness and the FG potential, and in total it becomes higher and with it the FG-charging current. A thinner interpoly oxide raises the capacitive coupling between CG and FG and thus raises the floating gate potential. This leads to an enhanced electric field and an enhanced FG-charging current. For a fast programming both oxide layers should be as thin as possible. The lower limits are

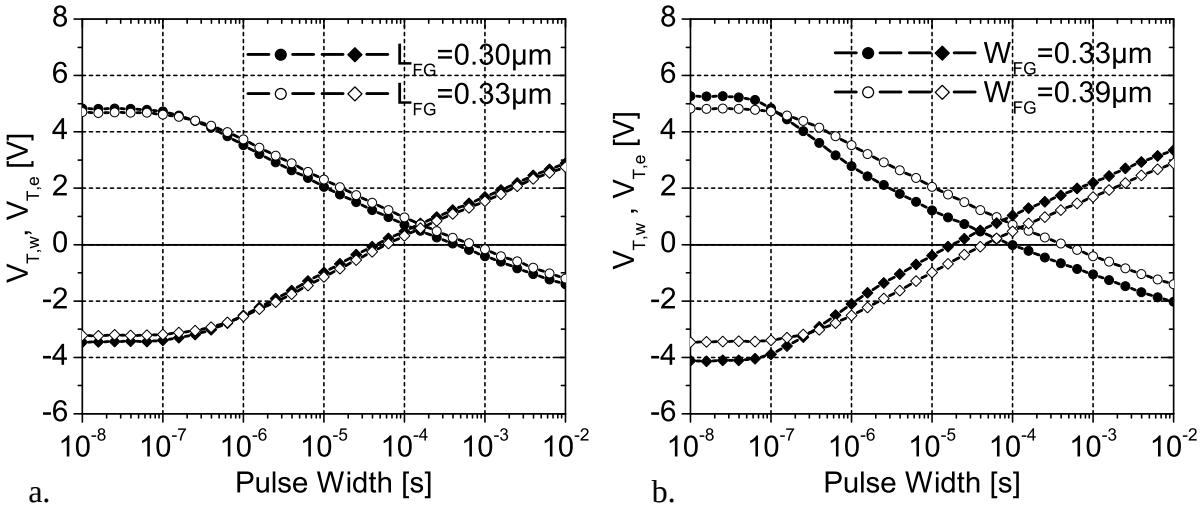


Figure 66: Influence of the layout parameters L_{FG} (a.) and W_{FG} (b.) on the transient cell behaviour; programming voltage V_{CG}=+/-12V; t_{oxg}=8.3nm; 2-transistor cell

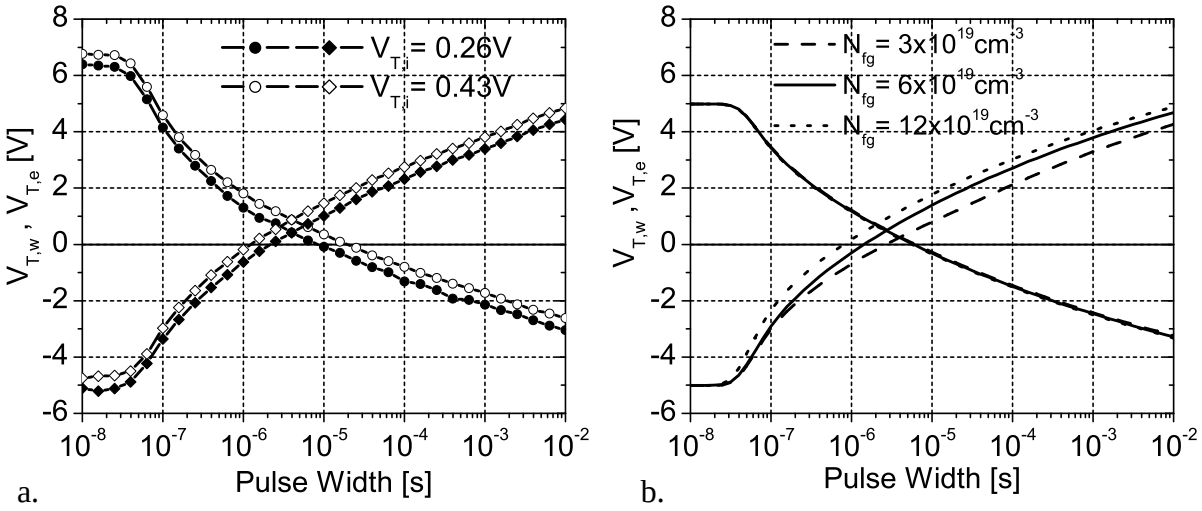


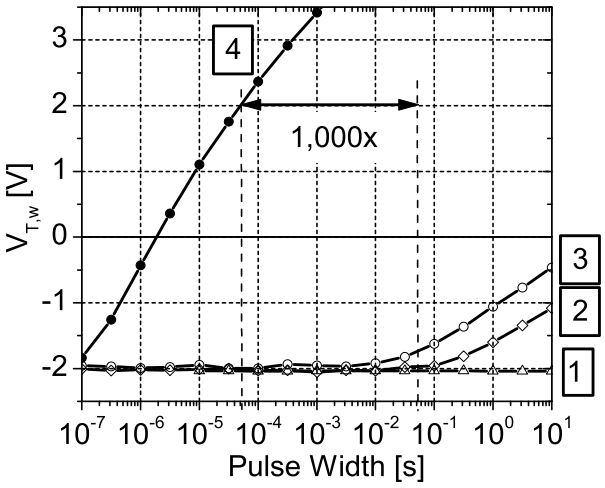
Figure 67: Influence of the p-well doping (a.) and the FG doping (b.) on the transient behaviour; programming voltage: V_{CG}=+/-14V ; (a) measured at a 1-transistor cell; (b) calculated after appendix A; t_{oxg}=8.3nm

given by reliability constraints.
It should be noted that the different initial values of $V_{T,w}$ and $V_{T,e}$ differ for the different oxide layer thicknesses, because the measurements of both cells have been done with the same re-set pulses.
Fig. 66 shows the influence of the layout parameters L_{FG} and W_{FG}. As these parameters are usually chosen as small as possible with respect to the design rules in order to achieve a small cell size, only small variations have been done here. Nevertheless the general trend of the behaviour is visible. For both parameters it can be seen that a higher value leads to slower programming. This effect is due to the reduced floating gate potential, similar to the effect of a raised interpoly oxide thickness. The control gate coupling ratio k_{CG} is reduced here, because the floating gate has a stronger capacitive coupling to the channel as a result of a a larger tunnel oxide area.

The influence of the intrinsic threshold voltage V_{Ti}, which is changed with the p-well doping concentration, and the influence of the FG doping concentration on the transient behaviour are shown in Fig. 67. The FG doping concentration has not been varied experimentally. The influence is shown by calculated curves (after appendix A), to complete the picture of the technological influence on the cell behaviour.
By varying V_{Ti} the V_T-window is shifted towards higher V_T-values. For a higher V_{Ti} erasing becomes slower and writing becomes faster. This is the case, because the electric field in the tunnel oxide for the same applied programming voltage is different for the same V_T state of the cells, as a different amount of FG charge is present at this V_T state.
A higher doping concentration in the FG poly-Si leads to a lower voltage drop in the poly-Si in depletion, leading to a higher electric field in the tunnel oxide at the same programming voltage. Since the poly-Si is in depletion only during cell writing, while the surface towards the tunnel oxide is in accumulation during cell erasing, an influence is only seen for the writing curves (Fig. 67b).

The sensitivity of the programming behaviour on the different cell parameters leads to a certain distribution of $V_{T,w}$ and $V_{T,e}$ within the cells of one memory, within cells of different memories on the same wafer and within cells on different wafers, which are programmed with the same programming pulse. Thus the most critical parameters, first of all the tunnel oxide

thickness, must be carefully controlled during the process. Since another very sensitive parameter is the programming voltage, an additional variation of $V_{T,w}$ and $V_{T,e}$ in different circuits results from differences in the generated high voltage in a memory chip. This has to be kept in mind during the circuit design of the memory chip, and measures have to be applied to minimize the variation (e.g. by using bandgap references). The final overall distribution of $V_{T,w}$ and $V_{T,e}$ has to be taken into account when defining the operating conditions of the memory. Especially the V_T-window and the reading conditions have to be chosen with respect to this variation, so that the memory operates correctly also for cells at the edges of the distribution. A further broadening of the distribution by so-called tail-cells due to oxide defects is an additional issue that is a major concern regarding the reliability of a flash memory.



1. non-selected WL non-selected BL
2. selected WL non-selected BL
3. non-selected WL selected BL
4. selected WL selected BL

Figure 68: Write disturb effect; programming voltages $V_{CG,w}$=7V, $V_{SDW,w}$=-7V; measured at a 1-transistor cell; t_{oxg}=7.5nm

5.1.5. Flash cell reliability

It must again be mentioned here that only basic test on the reliability of single flash cells are done within this work. A full characterisation of the reliability of complete memory circuits with arrays of millions of cells is a major research project by itself, requiring extended long-term investigations and statistical methods to obtain information about the behaviour of so-called tail-cells.
The results presented here cover 1. disturb effects, 2. endurance behaviour and 3. retention measurements. These investigations give a picture of the behaviour of a typical memory cell. As the process is not optimized with respect to the split-gate cell, the retention tests have only been performed at 1-transistor and 2-transistor cells.

1. Disturb effects

Disturb effects can be separated into two groups, write-disturbs and read-disturbs. A write-disturb means the unintended influence of cell writing on the un-selected memory cells of an array, while read disturb usually refers to the unintended influence of cell reading on selected cells of an array. Disturb effects are thus always related to the array organization and chosen way of memory operation. Here, the disturb effects that appear for cell and array operation as described in chapter 2 have been investigated.

When looking at write-disturb effects, the differently biased un-selected cells that exist within an array must be distinguished. These are un-selected cells in the same bitline as the selected cell, un-selected cells that are in the same wordline as the selected cell, and all other un-selected cells (Fig. 13, chapter 2). The transient behaviour of a 1-transistor cell under the bias conditions of these three different types of un-selected cells is presented in Fig. 68. It is compared to the programming curve of a selected cell. The worst disturb is seen at the un-selected cell belonging to the same bitline as the selected cell. Un-selected cells in the same wordline as the selected cell show less disturb, while no write-disturb has been measured up to 10s disturb time for cells that are neither in the same bitline nor in the same wordline as the selected cell. The reason for the disturb effect is the voltage between the channel and the control gate. For cells belonging to the same bitline as the selected cell this voltage is $V_{SDW,w}$. Source, drain and p-well are on the same potential in this case. For cells of the same wordline as the selected cell, this voltage is $V_{CG,w}$. Although $V_{CG,w} = -V_{SDW,w}$ in this case, the disturb is less for cells in the same wordline as the selected cell. The reason is that the p-well has a

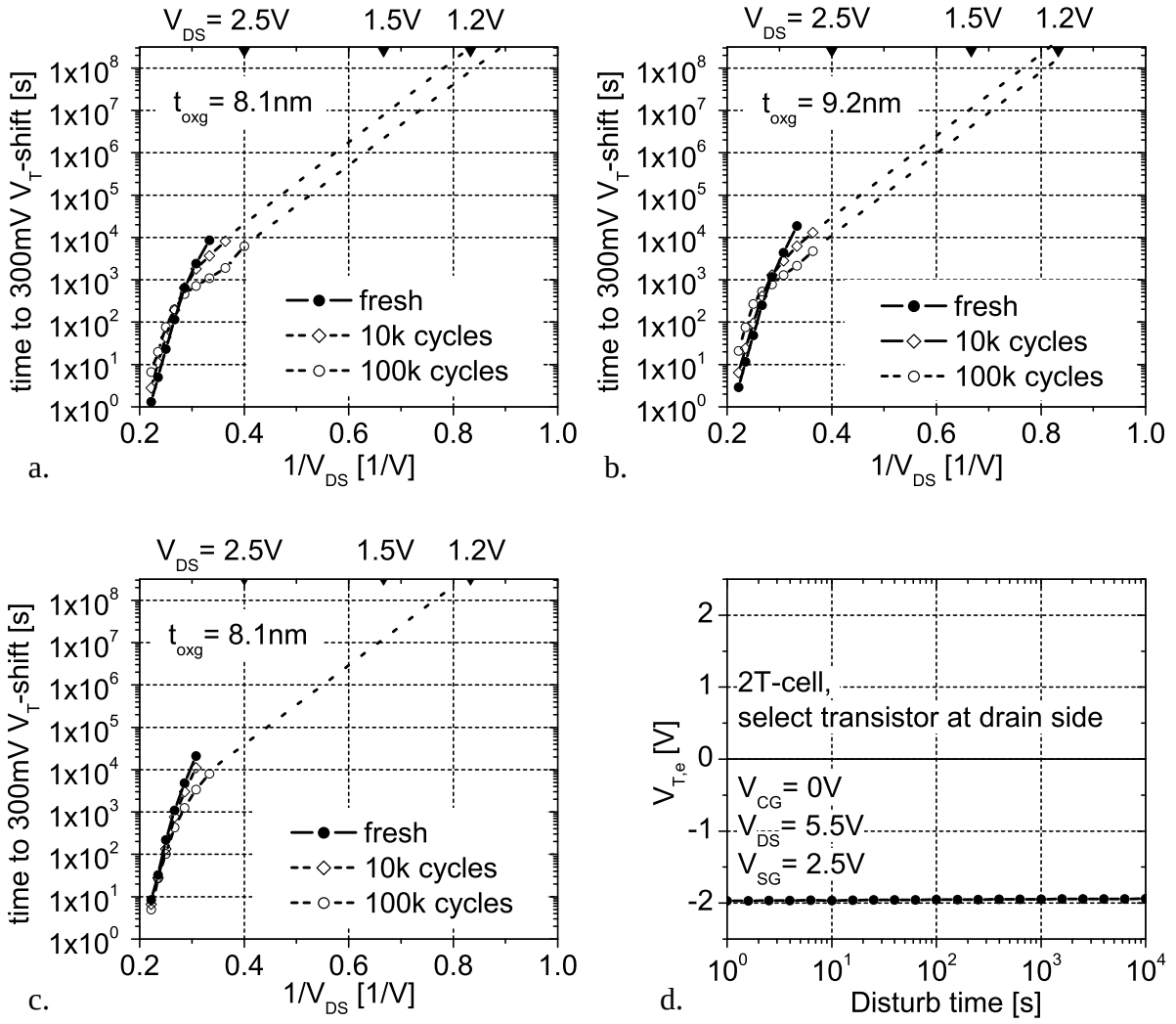


Figure 69: Results of read disturb measurements at fresh and cycled cells; a. 1-transistor cell, t_{oxg}=8.1nm; b. 1-transistor cell t_{oxg}=9.2nm; c. 2-transistor cell, t_{oxg}=8.1nm (select transistor at source side); d. 2-transistor cell, t_{oxg}=8.1nm (select transistor at drain side); write-erase cycling with constant pulses, V_{CG}=+/-14V, initial V_T-window: $V_{T,w}$=2V, $V_{T,e}$=-2V

negative voltage with respect to the S/D and channel potential. Outside the actual transistor area, where the floating gate resides on the field oxide, the floating gate is capacitive coupled to the p-well. This leads to a reduced floating gate potential and thus a reduced FN-current through the tunnel oxide. At all other unselected cells the channel and the control gate are on the same potential. In cells that neither belong to the same wordline, nor to the same bitline as the selected cell, the electric field in the tunnel oxide, which is present due to the charging state of the floating gate, is even slightly reduced by the capacitive coupling to the p-well via the field oxide.
The measurement result leads to the definition of a maximum allowed number of cells that are connected in one bitline. Each cell must withstand the writing of all other cells in its bitline, so the disturb time is the number of cells minus one times the writing time. In Fig. 68 it can be seen that a shift of around 300mV of an unselected cell is observed after a disturb time of around 1000 times the writing time of the selected cell. This means that the maximum number of cells allowed in one bitline is according to this measurement around 1000 cells, if a shift of 300mV is regarded as being acceptable.

The read disturb effect is a change of the cells V_T due to the bias applied to the selected cell during memory reading. It must be guaranteed that this V_T-shift is low enough for a specified

maximum reading time, so that the information can be read correctly. Typically memories are specified for 10 years of continuous reading.
The V_T of a written cell can theoretically be reduced by the two effects: 1) a positive drain voltage that is applied to the drain of a selected cell or 2) by the negative voltage applied to the control gates of the unselected cells in a 1-transistor memory. These two disturb effects have been found to be so small that they can be neglected.
An important effect is the soft-writing of selected erased cells. The drain current leads also at low drain voltages to a charging of the floating gate. This disturb effect is measured at elevated drain voltages for acceleration. A typical procedure is to determine the time for a defined V_T-shift (typically 300mV) for a range of drain voltages. The values are plotted versus $1/V_{DS}$ and in this scale linearly extrapolated to the specified maximum reading time [76]. The result gives the maximum allowed drain voltage.
Fig. 69 shows results of read disturb measurements. The strong influence of stressing the cell by repeated writing and erasing (cycling) can be seen. As done for endurance testing described below, the cells have been written and erased for a number of programming cycles with constant pulses. The pulses were adjusted for an initial V_T-window of $V_{T,w}$=2V at V_{CG}=14V for writing and $V_{T,e}$= -2V at V_{CG}=-14V for erasing. For the 1-transistor cell with t_{oxg}=8.1nm the maximum allowed drain voltage after 10k cycles is around V_{DS}=1.2V, after 100k cycles around V_{DS}=1.1V. A thicker tunnel oxide reduces the disturb-effect and allows a maximum drain voltage of V_{DS}=1.18V after 100k cycles. The same measurements, done at the 2-transistor cell, show a significant dependence of the disturb-effect on the position of the select transistor. In the case of placing the select transistor at the source side of the cell, a maximum drain voltage of V_{DS}=1.2V is allowed after 100k cycles, even for t_{oxg}=8.1nm. In the case of placing the select transistor at the drain side of the cell, almost no change in the cell's V_T could be measured after 10,000s continuous reading up to V_{DS}=5.5V.
The results show that for this technology the read disturb-effect limits the minimum allowed oxide thickness of the 1-transistor cell for a given V_{DS} needed for the reading operation and for a required number of allowed programming cycles. The 2-transistor cell gives slightly better results when the select transistor is placed at the source side of the cell, while the read-disturb is effectively suppressed by placing the select transistor at the drain side of the cell.

2. Endurance

The endurance has been measured by subsequently applying fixed write and erase pulses and monitoring the evolution of the threshold voltage of the written and the erased state. Different programming conditions as well as tunnel oxide thicknesses have been compared. The degradation of the cell's transfer characteristics with repeated cycling has been investigated. Since the endurance first of all reflects the degradation of the tunnel oxide, which is qualitatively the same for 1-transistor and 2-transistor cells, the investigations presented here concentrate on measurements done at the 2-transistor cell only.
Fig. 70 shows examples of typical endurance curves. First of all it can be seen that the V_T of the erased state is more affected than the V_T of the written state. This behaviour is typical for flash cells with uniform channel programming. The V_T-window closure, which is ascribed to a built up of trapped negative charge within the tunnel oxide that reduces the electric field at the electron injecting surface, is accompanied by an upward-shift of the whole V_T-window due to the same charge. The latter can be explained by the fact that the oxide charge is present in the channel area, and thus it raises the intrinsic V_T of the cell as seen from the floating gate (V_{Ti}). The combination of these two effects leads to an almost constant $V_{T,w}$ value of the written cell. Furthermore the influence of interface-trapped charge is present, which does not lead to a window-closure but merely to a shift of the V_T-window. This is discussed more in detail in appendix B, where the behaviour is investigated also by simulations of the transient

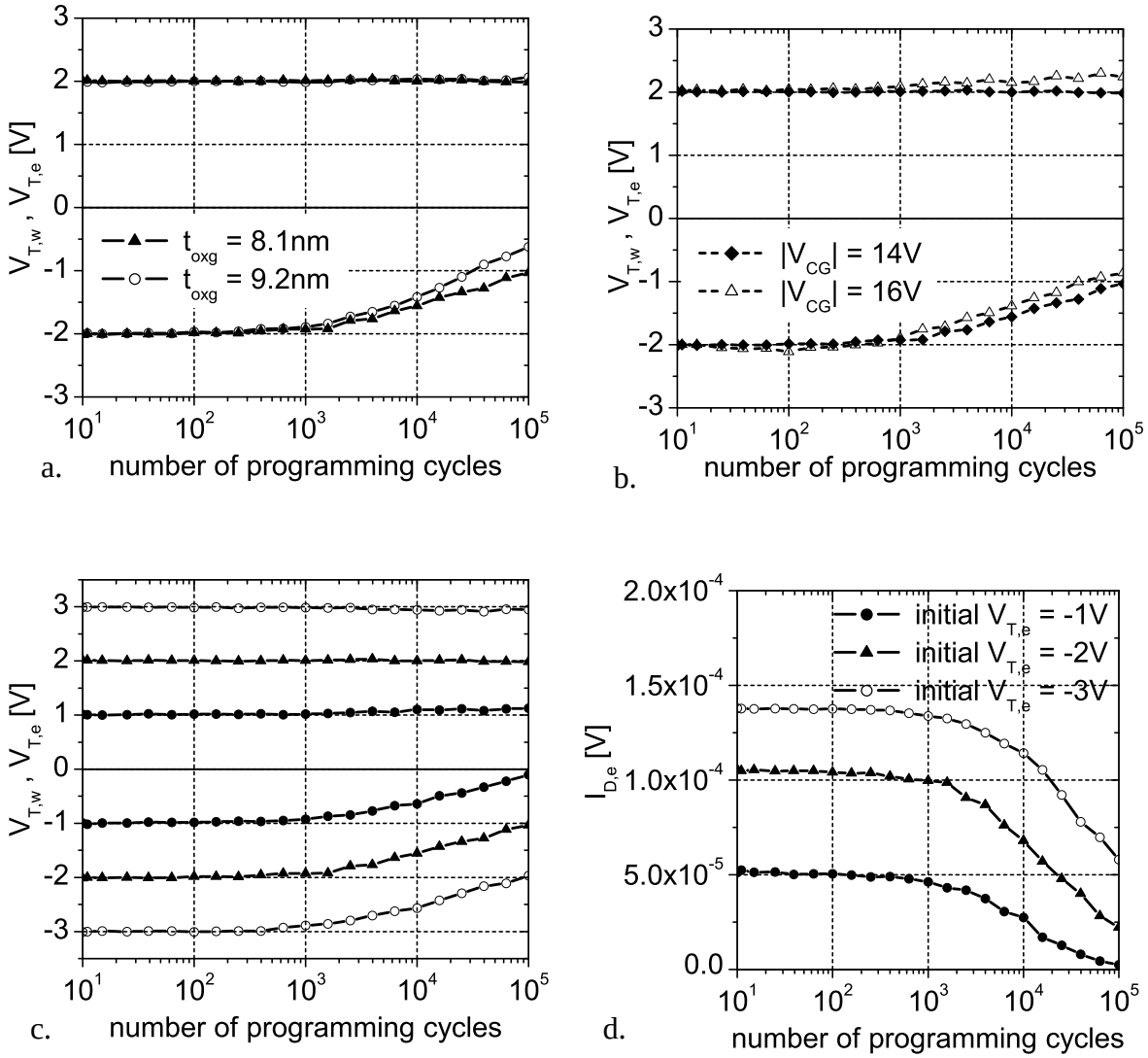


Figure 70: Results of endurance measurements at 2-transistor cells: a. V_{CG}=+14V for writing, V_{CG}=-14V for erasing, cells with different t_{oxg}; b. t_{oxg}=8.1nm, cells cycled with different programming voltages; c. t_{oxg}=8.1nm, V_{CG}=+/-14V, length of write and erase pulses adjusted for different initial programming windows (+/- 1V, +/- 2V, +/- 3V); d. evolution of the drain current of erased cells at V_{CG}=0V and V_{DS}=1.5V with repeated programming under the conditions of c.

cell behaviour at the presence of oxide and interface charges, and a new method is proposed to extract these charges from measured endurance curves.

When comparing cells from two different wafers with different tunnel oxide thicknesses (Fig. 70a), it can be seen that the cell with the thicker oxide shows a slightly worse degradation with respect to the $V_{T,e}$ of the erased cell. The V_T-shift of the erased cell after 100k cycles is 1V for the cell with t_{oxg}=8.1nm, while it is 1.2V for the cell with t_{oxg}=9.2nm. Stressing has been done with V_{CG}=+14V for writing, V_{CG}=-14V for erasing and an initial V_T-window of $V_{T,w}$=2V and $V_{T,e}$=-2V.

For a higher programming voltage of V_{CG}=16V for writing and V_{CG}=-16V for erasing (Fig. 70 b), it can be seen that for the same initial V_T-window both, the V_T of the written cell and the erased cell is shifted to higher values after repeated cycling compared to programming with +/-14V. This indicates that more interface-trapped charge is present after cycling with the higher voltage (appendix B).

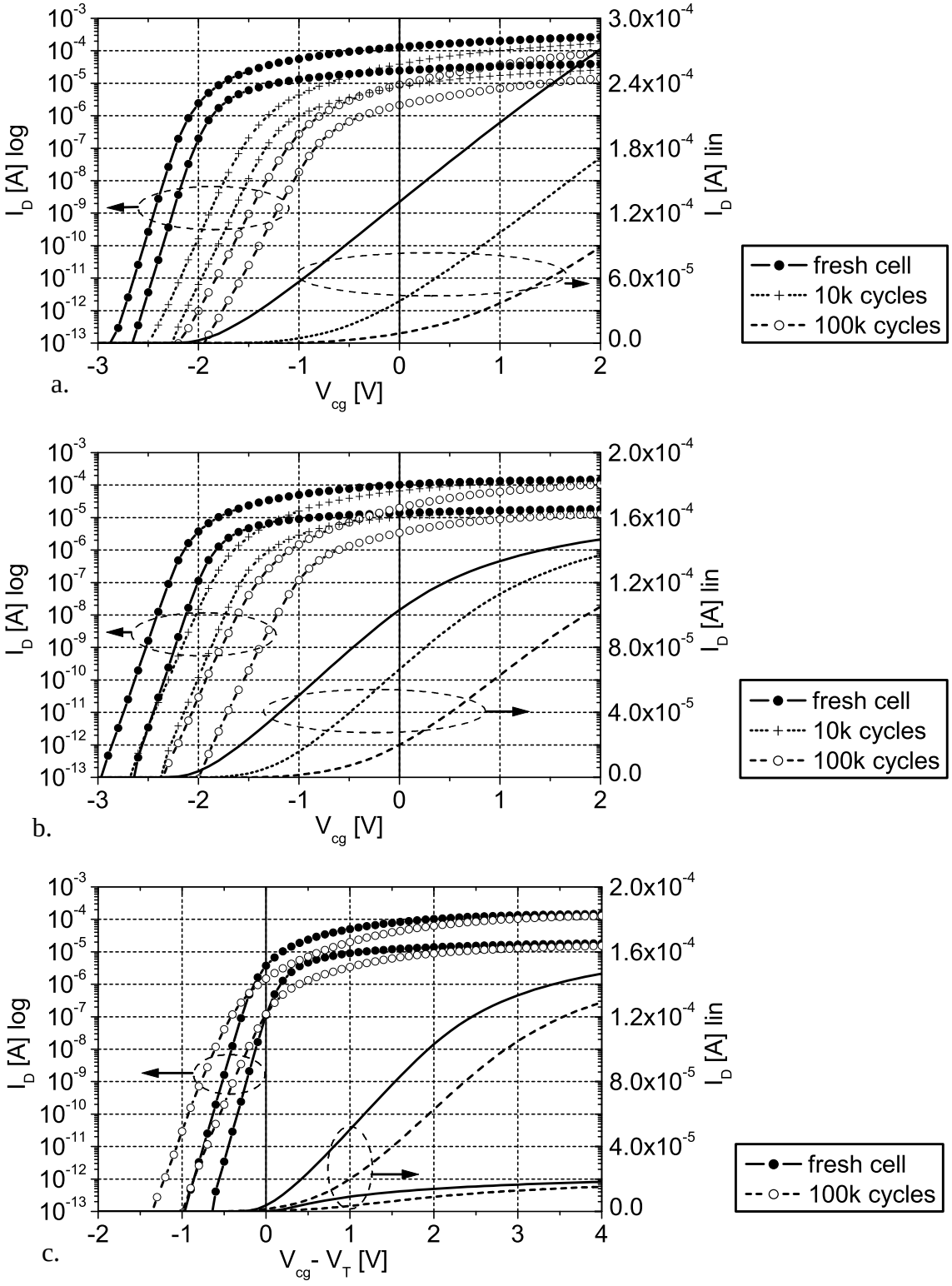


Figure 71: Degradation of the transfer characteristics with repeated cell cycling: a.1-transistor cell, t_{oxg}=8.1nm; b. 2-transistor cell, t_{oxg}=8.1nm; c. 2-transistor cell, drain current plotted versus ($V_{CG} - V_T$)

Comparing the degradation after cycling with different programming windows (Fig. 70c) it can be seen that a wider programming window leads to a slightly worse degradation with respect to the V_T-window-closure, which is understandable as more charge is transferred through the tunnel oxide at each cycle.

Fig. 70d shows the evolution of the drain current of erased cells with repeated cycling, measured at the bias conditions of cell reading. It can be seen that the reading current drops drastically. This can not only be explained by the raised $V_{T,e}$. The $V_{T,e}$ of the cell with an initial $V_{T,e}$=-3V reaches after 100k cycles about the same $V_{T,e}$ as a fresh cell with an initial $V_{T,e}$ =-2V, but the stressed cell shows a significantly lower drain current. An explanation is a reduced carrier mobility due to a degraded silicon/oxide interface.
The degradation of the drain current is also visible in transfer-characteristics measured at cells before and after cycling, as shown in Fig. 71. In Fig.71a and Fig. 71b the shift of $V_{T,e}$ the degraded characteristics can be seen for a 1-transistor cell and a 2-transistor cell, respectively. In Fig. 71c The current is drawn versus (V_{CG}- V_T), which makes the reduced subthreshold slope in logarithmic scale and the reduced saturation current in linear scale visible without the effect of a raised $V_{T,e}$.
The strong reduction of the reading current after cycling can be partly compensated by a verified erasing, that adjusts the erasing time so that the cell is always erased to the specified $V_{T,e}$. In this case only the degradation as indicated in Fig. 71c has to be taken into account. The reduction of the reading current limits the maximum allowed programming cycles.

3. Retention

The retention needs to be measured under accelerated conditions to make a prediction of the long-term behaviour possible within a reasonable time. A widely used way is to measure the charge loss of the floating gate under high temperature conditions and extrapolate the results to the operating temperature by assuming a constant activation energy. This approach is in question due to two reasons: firstly the activation energy has been found to change with temperature [86], and secondly some degradation, especially the behaviour of tail-cells, is already annealed at temperatures usually used during such tests [33]. In this work the cells have been investigated under a raised electric field by programming the cells to V_T states well above the usual operating conditions.
The cells have been programmed to $V_{T,w}$=+7V and $V_{T,e}$=-7V to investigate the retention of written and erased cells, respectively. Then, during the first 1000s after programming, the V_T has been measured with the cells being constantly connected (all terminals grounded during the time between the V_T-measurements). Then the wafers have been stored at room temperature and the V_T has been measured again after different times. It must be mentioned that during the storage of the wafers the control gate was floating. The CG potential can be nevertheless assumed to be the same as the wafer substrate, as the CG is connected to a metal contact pad, which has a capacitance versus the substrate that is significantly higher than the internal capacitances of the flash memory cell.
Fig. 72 shows the results of the retention measurements. The patterned box shows the usual V_T-window of +/-2V and a usually specified retention time of 10 years. The extrapolated curves should stay outside of this box to guarantee the specified retention. For a tunnel oxide thickness of t_{oxg}=8.1nm this is only valid for fresh cells. Already after 10^3 programming cycles the extrapolated curve of an erased cell reaches $V_{T,e}$=-2V after around 10^8s. After 10^4 programming cycles this is already reached after a time of less than 10^7s. For a tunnel oxide thickness of 9.2nm even after 10^4 programming cycles the extrapolated curves of the written and the erased cells stay outside the programming window up to 10 years.
It can be seen that the V_T of the erased cells changes faster than that of the written cells. This is due to the intrinsic V_T of the flash cells, which is above 0V. As discussed in chapter 4, the intrinsic V_T should be in the middle of the V_T-window to achieve the best retention results.

In summary it has been found that the reliability strongly depends on technological and memory operation parameters. According to the measurement results for a tunnel oxide

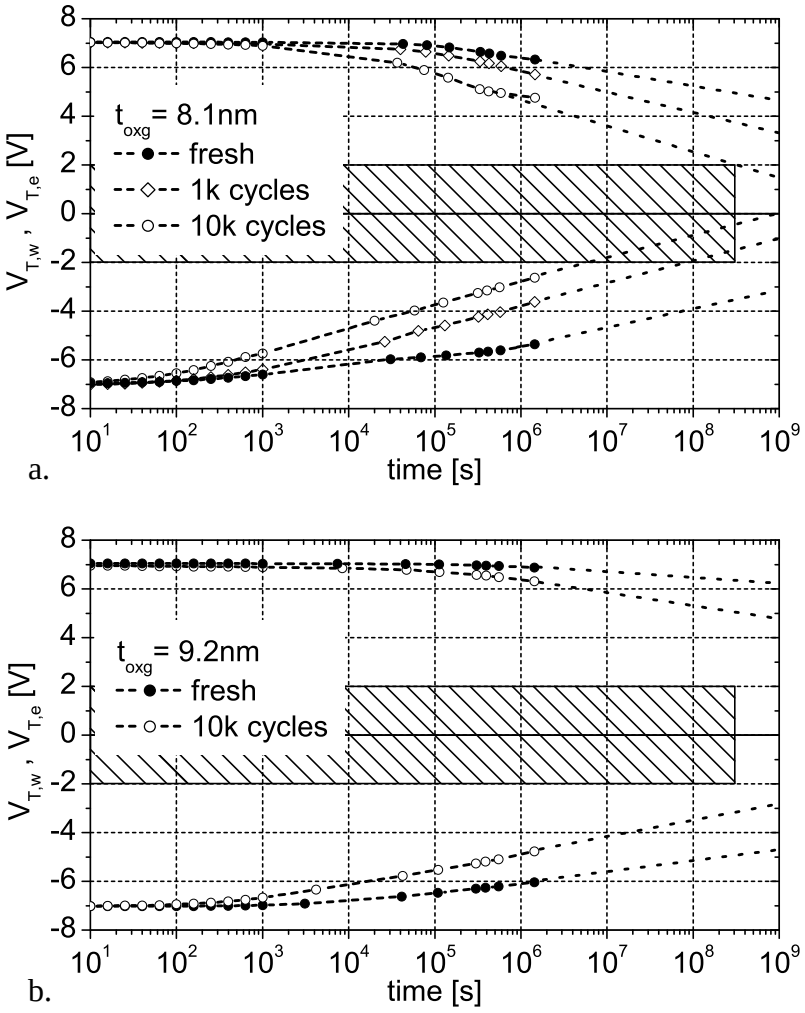


Figure 72: Results of the retention measurements of 1-transistor cells, comparing fresh cells and cells after repeated programming with V_{CG}=+14V for writing, V_{CG}=-14V for erasing, initial V_T-window: $V_{T,e}$=-2V, $V_{T,w}$=+2V; a. t_{oxg}=8.1nm; b. t_{oxg}=9.2nm

thickness of 9.2nm 10 years retention time and 10 years read disturb immunity can be predicted for a 1-transistor cell, also after 10^4 programming cycles.

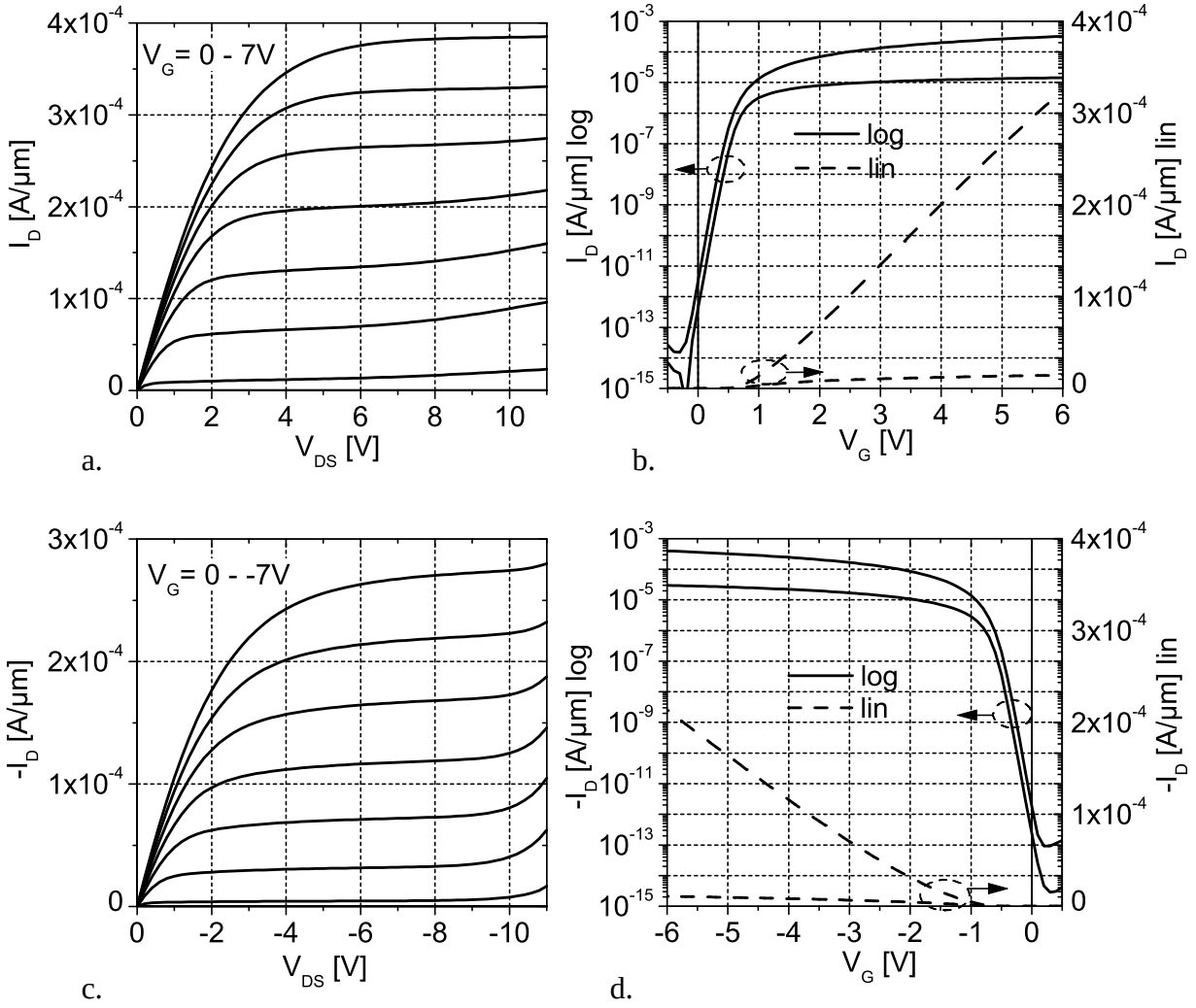


Figure 73: DC-characteristics of the HVMOS devices; output and transfer-characteristics of the HVNMOS (a., b.) and of the HVPMOS (c., d.); device dimensions see appendix D; output characteristics with 1V-steps in V_G; transfer characteristics with V_{DS}=0.1V and V_{DS}=6V (HVNMOS), V_{DS}=-0.1V and V_{DS}=-6V (HVPMOS)

5.2. High voltage MOSFETs

In following, the DC-characteristics of the high voltage MOS transistors will be presented and discussed. The HVMOS transistors (HVPMOS and HVNMOS) are required for memory operation, i.e. for generating and switching the high voltage needed for programming the memory. A very important parameter is thus the drain-to-source breakdown voltage, which has to be safely above the applied voltages. Here, a breakdown voltage of more than 10V has been targeted. Furthermore the drain-leakage current at high drain voltages should be low to achieve low current consumption during the write operation.

The devices under investigation are a HVNMOS with LDD extension areas at the source and the drain side of the gate, as described in chapter 3, with a gate-length of 0.6µm and a length of the LDD areas of 0.6µm on each side of the gate; a HVNMOS without LDD extension area, with a gate-length of 0.8µm; and a HVPMOS with a gate-length of 0.8µm. The implantation doses of the wells, which is closely related to the well of the flash cells as described in chapter 3, are adjusted for a process that has been optimized for a 1-transistor cell memory and for getting about the same absolute V_T-value of the HVNMOS and HVPMOS.

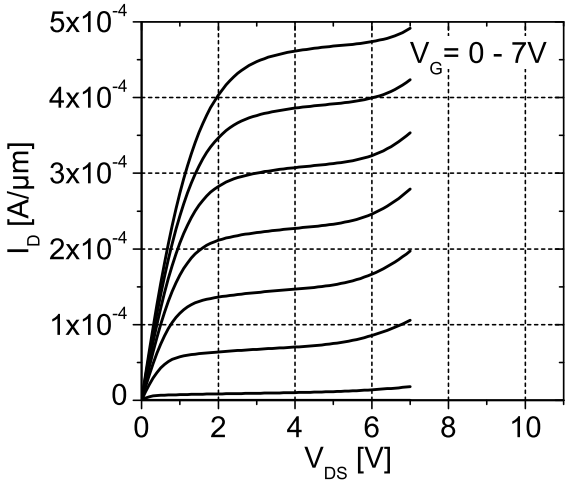


Figure 74: Output-characteristic of the HVNMOS without LDD area; 1V-steps in V_G

Fig. 73 shows the DC characteristics of the HVNMOS and of the HVPMOS. In the output characteristics it can be seen for both devices that the drain-source breakdown voltage is above V_{DS}=10V for the investigated range of gate voltages (V_{GS}=0V – V_{GS}=7V), as it has been targeted. Regarding the breakdown voltage, a memory operation with voltages up to 6V or 7V is possible, leading to a total programming voltage of 12V – 14V. These values have yet to be verified also by HVMOS reliability investigations like hot-carrier stressing. The transfer-characteristics have been measured at two values for V_{DS}. At V_{DS}=0.1V the threshold voltage of the devices has been determined to be V_T=0.5V for the HVNMOS and V_T=-0.6V for the HVPMOS. At V_{DS}=6V, which is a typical memory operating voltage the leakage current at V_{GS}=0V is important. For the HVNMOS the leakage is I_L/W_G=3pA/µm and for the HVPMOS it is I_L/W_G=1pA/µm.
The significant influence of the LDD area that has been introduced between the poly-Si gate edge and the highly doped S/D areas can be seen in Fig. 74. A HVNMOS without LDD areas, where the source and drain doping is implanted self-aligned to the gate poly-Si, has been built and measured. In the output characteristic it can be seen that the drain current significantly rises above V_{DS}=6V, which is the case for the HVNMOS with LDD area only above V_{DS}=10V. This can be explained by a reduced electric field at the drain edge due to smoother doping profiles. The main drawback of the construction with LDD areas, besides the larger

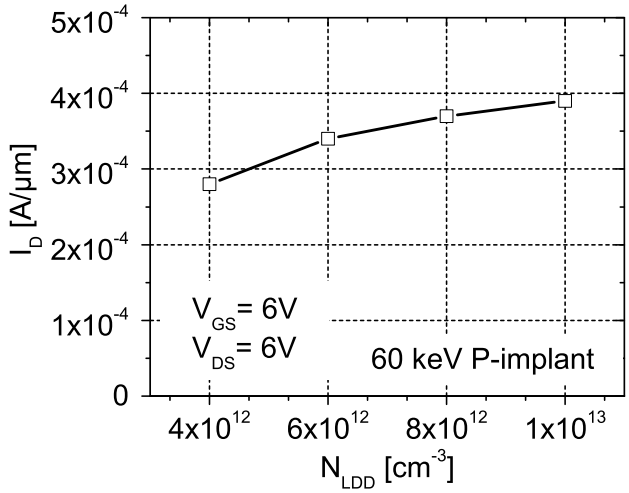


Figure 75: Drain current of the HVNMOS transistor with LDD areas at the source and at the drain side of the gate for different values of the dose of the implanted P-ions (N_{LDD})

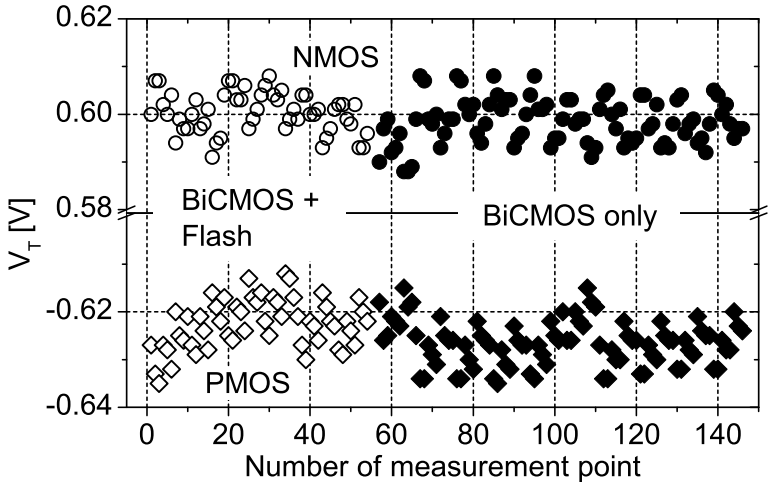


Figure 76: Threshold voltages of long-channel NMOS and PMOS transistors (L_G=25µm), measured on wafers prepared with a BiCMOS process only and on wafers prepared with a BiCMOS process including additional the embedded flash process modules

area consumption, is the reduced drain current due to the additional series resistance at the source and the drain. The dose of the implanted donors, which controls the doping level of the LDD areas, has to be adjusted to achieve a compromise between a high breakdown voltage and a high drain current. Fig. 75 shows the dependence of the drain current at V_{DS}=V_{GS}=6V for different P-implant doses (N_{LDD}). It can be seen that the drain current tends to saturate towards higher doses. A dose of N_{LDD}=8x$10^{12}$$cm^{-3}$ has finally been chosen, leading to the DC-characteristics of Fig. 73. The different values for I_D comparing Fig.75 to Fig. 73 results from slightly different p-well implant doses, which were not yet optimized in the experiment of Fig. 75.

5.3. BiCMOS devices

The influence of the flash memory integration on the electrical characteristics of the CMOS transistors and the HBTs will now be discussed. Important parameters of devices prepared with the original BiCMOS process, without the additional process steps for embedded flash memory fabrication, will be compared to the parameters of devices prepared with the presented process flow.

5.3.1. CMOS transistors

It is important that the electric characteristics of the CMOS devices are still well within the given specifications after the flash memory integration. One important benefit from such a modular integration is that the digital cell library can still be used for circuit design and circuit simulation, as the development of digital libraries is costly and time consuming.
Since all geometrical structuring of the flash memory process modules is done while the areas of CMOS devices on the wafer are completely covered by poly-Si, and process steps for the gate formation of the CMOS devices have not been changed, the major influence on device parameters is only by the added thermal budget of the new process steps. To investigate possible changes in the well doping profiles due to possible thermal diffusion of dopands, the threshold voltage of long-channel NMOS and PMOS transistors has been measured on wafers processed with the original BiCMOS process flow only and on wafers processed with the developed BiCMOS process flow with embedded flash modules. Fig. 76 shows the result of measurements of transistors with W_G=25µm and L_G=25µm. V_T-values measured on different wafers with 9 data points belonging to one wafer are presented. It can be seen that both, the

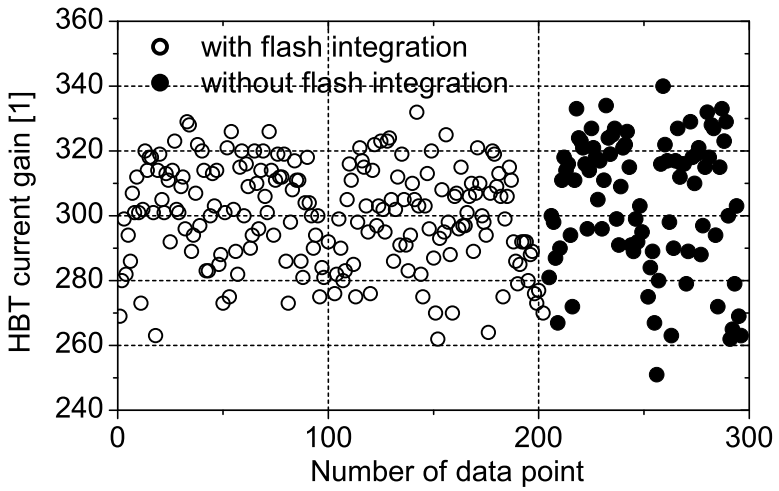


Figure 77: Current gain of SiGe:C HBTs, measured on wafers prepared with a BiCMOS process only and on wafers prepared with a BiCMOS process including the additional embedded flash process modules; current gain measured at V_{BE}=0.7V and V_{CB}=0V

mean value and the scattering of the threshold voltage are not influenced by the flash memory integration.
It can be concluded that the presented flash memory integration scheme has no negative impact on the characteristics of the CMOS devices. Due to the integration before CMOS gate structuring, the possible impact is reduced to a minimum, and is according to the measurements in an acceptable range.
This is also confirmed by the fact that the circuit simulation of the CMOS parts of the memory chip presented in chapter 6 could be done with the original device models of the BiCMOS flow.
An open question is the yield of high-density CMOS circuits, which could be affected by the changed cleaning regime due to the tunnel-oxide formation. But as dual gate-oxide processes are industry-standard today, the tunnel-oxide formation cannot be a principle problem of this integration scheme. Another open question is the impact of the prolonged nitride wet etching after the gate structuring. As it does not lead to geometrical changes in the CMOS transistor, no impact on device parameters is expected. A possible influence on the yield of high-density CMOS circuits has nevertheless strictly speaking still to be investigated.

5.3.2. SiGe:C HBT

The possible impact of the flash memory integration on the SiGe:C HBTs of the BiCMOS process can be regarded as more critical than the impact on the CMOS transistors, as the almost fully processed HBT has to withstand the geometrical structuring of the flash memory devices as well as the added thermal budget (see chapter 4).
As a sensitive parameter for monitoring changes in the device behaviour due to changed doping profiles, the current gain of single HBTs has been compared for wafers processed with and without the additional process modules of the embedded flash integration. Fig 77 shows the result of comparing four wafers, with 100 HBTs measured on each wafer with BiCMOS and embedded flash process, and 45 HBTs measured on each wafer with BICMOS processing only. The results of two wafers with additional flash integration and two wafers with the original BiCMOS processing are shown. It can be seen that no impact, neither on the mean value nor on the scattering of the data points is visible.
To investigate a possible impact of the flash integration on the yield of HBTs, especially as a consequence of the steps that lead to geometrical changes, e.g. the partly removing of the

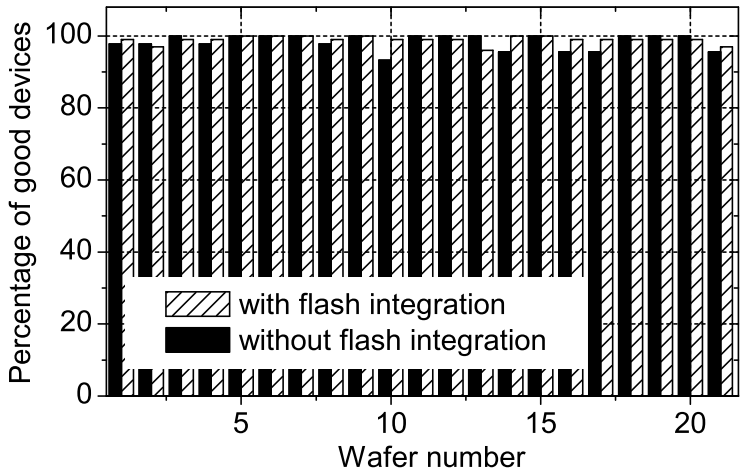


Figure 78: Yield of 4k HBT arrays, measured on wafers prepared with a BiCMOS process only and on wafers prepared with a BiCMOS process including the embedded flash process modules; a "good device" is defined as an HBT with I_{ECs}<1nA at V_{EB}=0.4V

oxide spacers (see chapter 4), the wafer-yield of HBT arrays has been measured. HBT arrays consisting of 4k HBTs connected in parallel have been measured with respect to their leakage behaviour. Fig. 78 shows the percentage of good devices found on wafers that have been processed with and without integrated flash memory. On each wafer with embedded flash processing 100 arrays have been measured; on each wafer without embedded flash processing 45 arrays have been measured. As a result, no impact of the additional flash memory process modules can be seen on the yield of the HBT arrays.

5.3.4. Modularity of the technology

In summary the results of comparing CMOS transistors and HBTs of a BiMCOS process with and without the additional embedded flash memory modules indicate that the process integration can be regarded as being modular. This is especially important for a flexible technology for SOC design. Different systems, with and without embedded flash memory can be developed based on the same baseline technology.

Chapter 6

Full Circuit Demonstration

A 1-Mbit memory chip has been developed, based on the 1-transistor cell, to demonstrate the feasibility of the technology for such memory densities. The circuit design has been done in cooperation with the CAD-department of the Technical University of Kiev.
As the circuit design itself is not part of this thesis, only a brief general description of the memory chip will be given, followed by a presentation of the most important measurement results that show the principle functionality of the device.
It should be mentioned that, as an interface between technology and circuit-design, firstly a set of design-rules for the newly introduced mask-layers had to be defined, and secondly the electrical modelling of the new devices had to be done, which are the high-voltage transistors and the flash cell itself. The high-voltage transistors have been modelled with a usual SPICE transistor model (BSIM3v3). The flash-cell has only been modelled for simulating the reading operation, as a compact model that includes the transient behaviour has not been available at the IHP at this time. To simulate the reading operation of erased and written cells, two sets of SPICE MOSFET model parameters have been extracted, one representing an erased cell and the other representing a written cell.

6.1. Building blocks and memory organization

Fig. 79 shows a micrograph of the memory chip after full processing. The memory matrix is visible as black area. It is separated into two equal parts, to reduce the length of the wordlines, which are oriented horizontally with respect to the figure. The row-decoder is placed between the two cell arrays. It consists of a low-voltage decoder using the standard CMOS transistors, which is connected to the wordlines via level-shifters built with the high-voltage MOS transistors [90]. The column-decoder is placed underneath the cell matrix. It consists of two 3-address-decoders, each having 8 output signals, a multiplexer circuit, high voltage shifter circuits, sense-amplifiers, output latches and output buffers. The shifter circuits do again the connection of a low-voltage decoder with the memory matrix, in this case with the source and drain connections. The sense amplifiers are differential 2-stage amplifiers that detect the drain current of erased cells. The information that is read from the cells is kept in output-latches. The output-latches are connected to the output data-pads via the output-buffers, which produce the required fan-out and protect the inner circuit (e.g. from ESD). The last important building blocks are the charge-pumps for the on-chip generation of the positive and negative high voltage. Two concurrent Dickson-charge-pumps are used to generate the positive high voltage, while the negative high voltage is generated using the high-voltage PMOS transistors [89]. Both charge-pumps are designed to have the same timing-parameters. The voltages generated for writing and erasing are targeted to be +6V and –6V, which are added to achieve in total +12V and –12V for writing and erasing, respectively. A negative voltage of V_{off}=–3V is generated and applied to the non-selected wordlines during reading to suppress the current in erased cells.
The memory architecture is NOR-type and it is operated as described in chapter 2. The memory matrix consists of 1024 rows and 1024 columns. One word consists of 16 bits, which

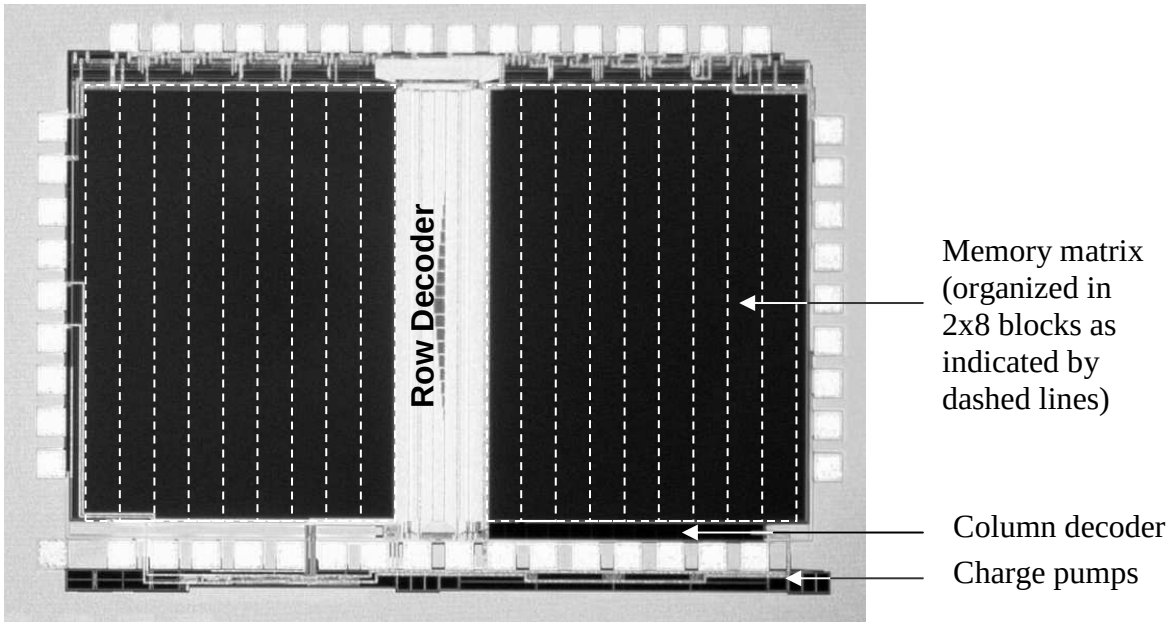


Figure 79: Micrograph of the 1-Mbit embedded flash memory chip after full processing

are addressed simultaneously during programming and reading. Due to this, the whole matrix can be imaginarily divided into 16 blocks, with 64 columns belong to one block. Each bit of one word is stored in a different block, each at the same place within its block. Each block has it's own sense-amplifier. The dashed lines in Fig.79 indicate this memory-organization.
Besides the 16 I/O data pads and the 16 address pads, the memory is controlled by several control pads, which define the different operation modes of the memory (read, write, erase, common-write). In the read-mode the information of the 16 cells belonging to the word that is defined by the 16 address pads can be read at the 16 data pads. In the write-mode the information at the 16 data pads is written into the 16 cells belonging to the word that is defined by the 16 address pads. In the erase-mode all cells of the memory are erased simultaneously. In the common-write-mode all cells of the memory are written simultaneously. The latter is done before memory erasing, in order to get a tighter distribution of the threshold-voltages within the cells of the memory matrix. Table 1 shows how the different modes are switched by applying the respective signals to the control pads. These operation modes are the needed for the basic memory operation. Additional modes, control pads and pads connected to internal circuit nodes exist that allow to get more detailed information about the circuit, which is not discussed here. Finally, two pads are reserved for the power supply (V_{DD} = 2.5V) and the ground connection.

	Operation Mode			
Control Pad	*Read*	*Write*	*Erase*	*Common-Write*
Read	1	0	0	0
Write	0	1	0	1
Erase	0	0	1	0
Common	0	0	1	1

Table 1: Switching of the memory operation modes by the control pad signals

6.2. Functional testing

The functional testing has been done using an Agilent SOC93K Test system, combined with a TSK UF200 wafer-prober for on-wafer measurements. All tests were done at a supply voltage of V_{DD}=2.5V. The high voltage required for cell writing and erasing was in all cases generated on-chip.
A number of basic tests have been done. The results will be presented and discussed. Some general description and definition of important terms will be given first.

Memory erasing
The erasing is always done in a 2-step procedure. First, the memory is set to the common-write mode for a defined time, and then the memory is set to the erase-mode for a defined time, which is the erasing time. The data and address pads are not needed during this operation.

Memory writing
Before the memory can be written, all cells have to be erased first. When this is done, the memory is set to the write-mode, an address is applied to the address pads and the data that is to be saved is applied to the data pads. This configuration is kept for a certain time, which is the writing time. During this time the state of the addressed cells is changed according to the applied data. After this, both, the address and the data are changed to write the information of the next word into the memory. After switching to the write-mode and before writing the first cell, a delay time has to be introduced to make sure that the charge-pumps have reached the final level of the programming voltage.

Memory reading
The memory is set to the read-mode. The address of the word that is to be read is applied to the address pads, and after some time, which is the minimum reading time, the information that is stored at this address is present at the data pads. After getting this information the address can be changed to read the contents of the next word. The time difference between two times switching the address is later on called the reading time.

6.2.1. Test sequence and results

The following sequence of basic tests has been applied to the memory chip:

1. Memory erasing
2. Full memory writing (all cells “1”)
3. Writing a single row / single columns
4. Writing a “checker-board” pattern (every second cell “1”)
5. Determination of the read access time after writing a “checker-board” pattern
6. Writing one word with long writing times

1. Memory erasing

The memory has been erased with a common-write time of 1ms and a different values of the erasing time. After erasing, all addresses have been read with a reading time of 500ns.
As a result it was observed that after an erasing time of >10ms all cells of a memory were read as “0”. Comparing this value to the transient characteristics of a single cell of the same wafer at V_{CG} = -12V, it could be seen that a cell is read as “0” already at a threshold voltage of about $V_{T,e}$ = -1V.

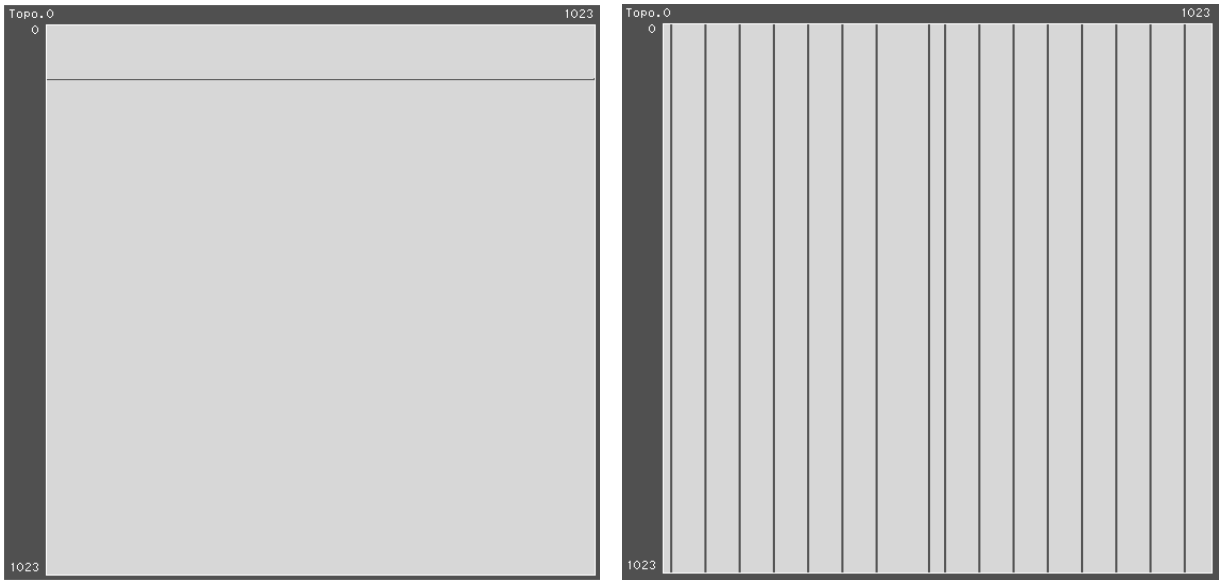


Figure 80: Screenshots of bitmaps generated by the evaluation-software for functional testing, taken after writing single rows (left) and columns (right) and reading the full memory

2. Full memory writing

The memory has been erased with a common-write time of 1ms and an erasing time of 10ms. After this, all cells were written to "1" (which means that a value of 65535 is written to all addresses) with different values of the writing time. Finally, all cells were read with a reading time of 500ns. As a result it was observed, that after a writing time of 20us all cells of the matrix were read as "1". Again, comparing this result to the transient characteristics measured at a single cell of the same wafer, it could be seen that a cell is read as "1" already at a threshold voltage slightly above $V_{T,w} = 0V$.

3. Writing single rows / single columns

The memory has been erased with a common-write time of 1ms and an erasing time of 10ms. Then, a single row or a single column (in each block) has been written to "1", respectively, with 20µs writing time. Fig. 80 shows screenshots of the evaluation-software after reading the full memory with 500 ns reading time. Bitmaps are shown, where the logical addresses have been translated to the physical position of the cells within the matrix. A black spot on the bitmap is a written cell, while the erased cells are white. By writing a single row, possible problems concerning the leakage of the unselected cells and overerase problems can be detected. All cells in one bitline, except one, are erased to a low V_T-state, which is the worst-case regarding leakage current. The total leakage of all unselected cells must still be clearly distinguishable from the reading current of a selected cell. Furthermore, if only one cell in the bitline would be overerased, this means that its V_T is so low that also an unselected cell has a significant current contribution, then all cells of the bitline would be read as "0", also the one that is actually written. None of this has been observed here. In the picture showing the example of writing single columns it can be seen that one column is shifted to the right with respect to the other columns. This is due to a mistake in the circuit design that caused a different connection of the columns belonging to the 8^{th} bit of one word. The columns are ordered in this block in the opposite direction.

4. Writing a "checker-board" pattern

The memory has been erased with a common-write time of 1ms and an erasing time of 10ms. After this, a "checker-board" pattern has been written into the memory with 20 µs writing time. The "checker-board" pattern means that every second cell is written, so that finally a

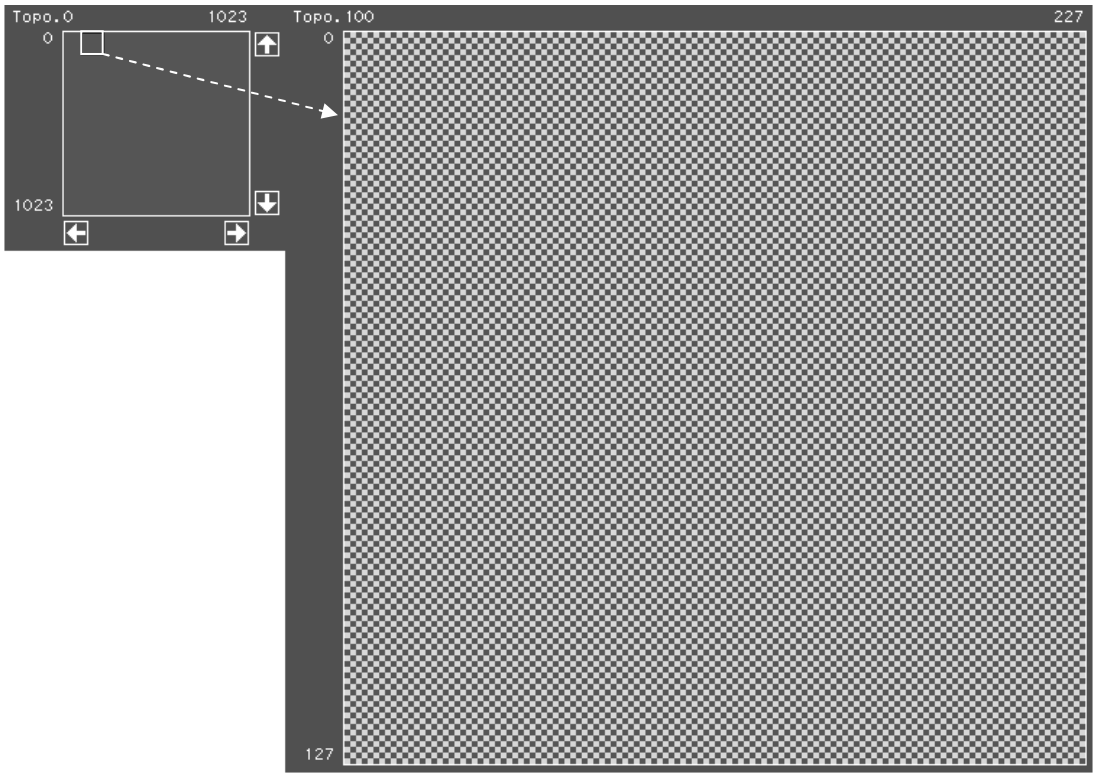


Figure 81: Screenshot of bitmaps generated by the evaluation-software for functional testing, taken after writing a “checker-board” pattern and reading the full memory

chess-board-like pattern is visible. This pattern is widely used for memory testing, as it is in some respect a worst-case scenario. For example every cell is surrounded by cells of the opposite state, so that a possible influence between neighbouring cells becomes visible. Fig. 81 shows a screenshot of the evaluation-software after reading the full memory with 500 ns reading time. The “checker-board” could be written without errors.

5. Determination of the read access time after writing a “checker-board” pattern
After writing a “checker-board” pattern into the memory, the read access time has been directly measured at selected bits. To do this, a timing diagram has been generated. This

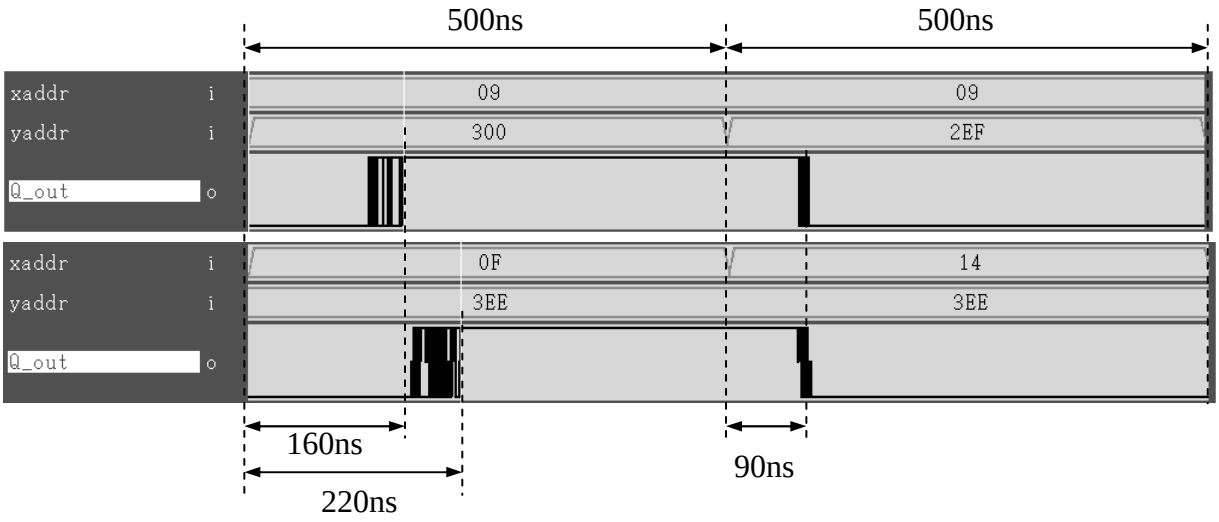


Figure 82: Screenshots of timing diagrams generated by the evaluation-software for functional testing, taken after writing a “checker-board” pattern and reading cells in y-direction row by row (upper diagram) and in x-direction column by column (lower diagram)

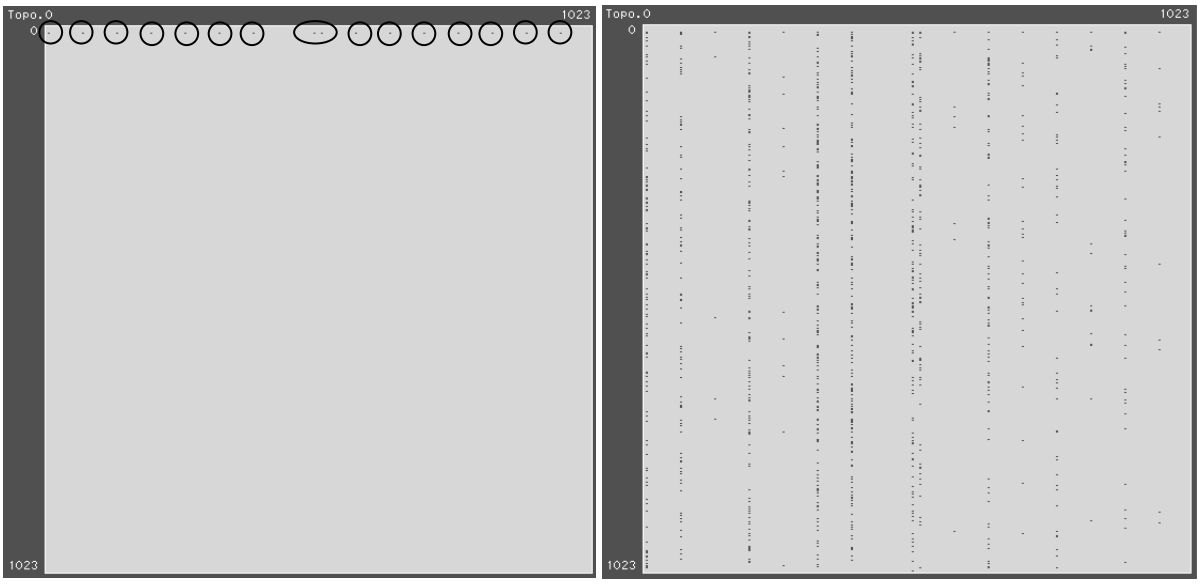


Figure 83: Screenshot of bitmaps generated by the evaluation-software for functional testing, taken after writing "1" into one word with 1ms (left) and 10ms (right) writing time, and reading the full memory

diagram shows the status ("0" or "1") of all data pads versus time during reading. As in the "checker-board" pattern every cell has the opposite state compared to the cell read before, the time until the right state is read at the data pads after a change in the address can be measured. Fig. 82 shows screenshots taken after measuring different timing diagrams. Q_out is the status of all 16 data pads drawn together. Xaddr and yaddr are the addresses in hexadecimal numbers, separated in column number and row number, respectively. The upper timing diagram shows the states of the data pads when switching from one row to another by changing the y-address; the lower diagram shows the same for switching from one column to another by changing the x-address. It can be seen that the "0" is already read correctly after 90ns, independent on the direction of address changing. The correct reading of a "1" is slower for reading cells in different columns after each other, requiring a minimum reading time of 220ns. When changing the to another cell in y-direction, the slowest data pad changes its information 160ns after changing the address. The slowest of these times defines the minimum read access time, which is 220ns here. This has to be verified on all cells of the memory by reading the whole memory without errors at this reading time with a "checker-board" and the respective inverse "checker-board" pattern written into the memory.

6. Writing one word with long writing times

This test is used to get an information about write disturb effects. The memory has been erased with a common-write time of 1ms and an erasing time of 10ms. After this the 16 cells of one word have been written to "1" with writing times of 1ms and 10ms. Fig. 83 shows the bitmaps after reading the whole memory with 500ns reading time. After writing with 1ms writing time the pattern is read correctly. After writing with 10ms other cells in the same column as the cells that have actually been written are also read as "1 ". This can be explained by the onset of write disturb. As already observed in chapter 5, the cells in the same bitline as the selected cell are stronger affected by this disturb effect compared to all other cells in the memory array. This picture is also observed here. Cells in the same wordline as the selected cells are not yet affected.

Write disturb is a severe problem here. The value of 10 ms is only 500 times the usual writing time of 20µs. This means that the number of cells in one bitline has to be less than 500 to

Memory size	2.28 x 1.55 mm^2
Cell size	2.14μm^2
Memory architecture	NOR
Flash cell type	1-transistor cell
Capacity	1 Mbit
No. of data pads	16
No. of address pads	16
Internal high voltage	+/- 6V (target values)
Writing time	20 µs
Erasing time	10 ms
Read access time	>220 ns

Table 2: Summary of important parameters of the embedded flash memory chip

guarantee full functionality at all possible patterns written to the memory. This needs to be either improved or taken into account in future realisations of the memory.
The write disturb effect is much stronger than expected. The reason for this is an unwanted unsymmetrical distribution of the generated high voltage in this realization of the memory. The generated negative high voltage is in absolute values higher than the positive high voltage. This leads to a stronger disturb at the unselected cells in the bitline of a selected cell compared to the symmetrical case, or compared to the opposite case of an absolutely higher negative voltage. As explained before (chapter 2 and chapter 5), the cells in a bitline of a selected cell have to withstand $V_{SDW,w}$, which is the full generated negative high voltage. Such a constellation needs to be avoided to improve the memory in this respect, and the circuit design of the charge-pumps needs to be reviewed.

6.3. Summary of the memory chip functional testing

In summary the results of the functional testing of the memory chip that has been built with the developed technological process flow has given positive results in the basic functions. Different patterns could be written and read without any errors. The writing time was in the expected range. It has been observed that the right state can be determined with a small V_T-window only, but to compensate for parameter variations in different memories a wider window needs to be specified for reliably using the memory. The reading time of >220ns is significantly longer than expected from circuit simulations, which needs to be reviewed in next design iterations. Write disturb has been identified to be a major issue, which requires to be dealt with in new versions of the memory chip. One possible reason could be identified.
Table 2 summarizes the parameters of the memory chip.
The next steps that have to be done are a statistical analysis of the memory's yield and investigations of the reliability of the complete circuit, which is not covered by this dissertation.
The first successful realization of a 1-Mbit embedded flash memory in a SiGe:C BICMOS process has been done. The results of the basic functional testing show the feasibility of the process and its individual devices for memories of such densities.

Chapter 7

Summary and Conclusions

This dissertation describes for the first time in detail the integration of an embedded flash memory module into a 0.25µm SiGe:C RF-BiCMOS process. The combination of a high-performance BiCMOS process with a non-volatile memory is important for SOC solutions, resulting in both, reduced fabrication cost and enhanced system functionality. Similar solutions are commonly used for RF-CMOS processes, while no publications are known to me about a flash memory integration into a SiGe:C BiCMOS technology platform, which is needed for applications requiring highest RF-performance.

In this work a process integration scheme has been proposed and the fabrication process has been implemented in the pilot line of the IHP. The process has been characterized by means of SEM and TEM at different stages. Overall integration issues as well as the main process steps have been discussed. The individual devices have been characterised electrically, including basic reliability investigations of single flash memory cells. The impact of the process integration on the original BiCMOS process and its devices has been discussed and evaluated by electrical measurements. Finally the results of successful functional testing of a 1-Mbit memory chip, which has been developed in cooperation with the NTU Kiev and fabricated with the presented process flow, have been shown, demonstrating the feasibility of the process for such memory densities.

It has been found in a comparison of CMOS embedded flash memory solutions reported in literature that a wide variety of concepts exist, with different mechanisms for writing and erasing the memory and with different ways of process integration. Dedicated processes have been developed during the last 10-15 years to meet the specific requirements of embedded flash memories, which are first of all reduced processing-costs and a possible low-power memory operation. Features like cell size and programming speed play only minor roles here. The cost is reflected in the number of required additional mask levels, and the lowest numbers reported for embedded stacked gate flash memories are 3-4 additional mask levels. The lowest power consumption is achieved by flash memories using FN-tunnelling for cell programming, with the drawbacks of a slower programming speed and high programming voltages. Especially embedded NOR-type 2-transistor flash memories that are written and erased by FN-tunnelling have been reported as a preferred solution during the recent years. More advanced NVM concepts are not yet mature enough or have not the CMOS compatibility of a standard floating gate approach.

The chosen memory concept is an FN-programmed, double poly-silicon, floating-gate, NOR embedded flash memory. The advantages and drawbacks of this approach have been discussed. The developed process integration scheme is a low-cost approach leading to an embedded flash memory process requiring only 4 additional mask levels on top of the baseline BiCMOS process flow. This has been achieved by sharing mask layers for flash cell and peripheral HVMOS transistor fabrication and by using the interpoly oxide of the flash

cells simultaneously as gate dielectric for the HVMOS transistors. The process allows the fabrication of different types of flash cells, which are 1-transistor cells, 2-transistor cells and split-gate cells. It has been shown that the process has to be adjusted with respect to the chosen cell type, especially regarding implantation parameters.

SEM and TEM images demonstrate how the different devices evolve during the fabrication process. Considerations of adjusting different parts of the process have been discussed in detail, these parts are the tunnel oxide formation, the interpoly oxide formation, the well formation, the RIE of the floating gate, the RIE of the control gate and the introduction of an ARC for the CG-lithography. This gives a deeper insight into the process and into the interaction of the different steps with each other and with the baseline process.

The tunnel oxide, which is most important for the transient and reliability characteristics of the flash cells, has been investigated with respect to its current-voltage behaviour, also after FN-stress. The measurements have been compared with simulations according to models published in the literature, which shows a good agreement.
The DC-behaviour of the flash cells, which is important for the cell reading, has been presented for the different cell types. From this, the operating conditions of a memory using the different cells have been be determined. As a result it can be seen that especially the 2-transistor cell offers a robust solution for achieving high reading currents, which is important for a fast read access in a memory chip.
The transient behaviour of the flash cells has been presented for the different cell types. In addition, the influence of various technological parameters has been demonstrated and discussed. The results show both, the possible ways for technological improvements, and the need for carefully controlling a number of sensitive parameters during processing with respect to homogeneity and reproducibility, in order to get a tight enough distribution of the programmed and erased V_T of the cells.
The presented reliability investigations of single memory cells show that a careful choice of the tunnel oxide thickness is necessary to fulfil the required specifications. A too thin oxide leads to severe problems regarding read-disturb effects and memory cell retention, especially after repeated cell programming. A compromise needs to be found with respect to the acceptable programming time for the target application.
The investigations done within this work only give a basic picture of the reliability behaviour of the developed flash memory cells. A full reliability characterisation taking statistical effects into account is another major project by itself.

The HVMOS transistors have been found to have a sufficiently high drain-source breakdown voltage of >10V. It has been demonstrated that in the presented technology this can be achieved for both types of transistors (n-channel and p-channel), if an LDD-area is introduced at the source and drain side of the HVNMOS gate.

The electrical measurements done at the CMOS transistors and HBTs demonstrate the modular character of the presented integration scheme. The presented results indicate, that the integration of the additional process modules can be done without a significant influence on both, the device parameters and the yield. The modularity is important to allow using the same CMOS and HBT cell library for circuit design with and without an embedded flash memory, thus offering more flexibility. Open issues in this respect that have to be investigated in future projects have been identified.

The successful fabrication of a 1-Mbit memory-chip with the proposed process technology has been demonstrated by a set of basic functional tests. Memories without any failing bit

during the presented functional test flow have been measured. The minimum writing time has been determined to be 20µs, and the read access time to be >220ns.
The writing time is in the expected range, while the read access time is slower than in circuit simulations, which has to be further investigated.
Write disturb measurements have been found to be a severe issue for the presented memory chip. A reason could be identified and has to be carefully taken into account for future memory designs.

Altogether the results of this dissertation indicate that the developed process technology is able to produce embedded FN-programmed flash memories for medium density applications within a SiGe:C BiCMOS platform. This opens new possibilities for future SOC developments.

Appendix A: Calculating the transient cell behaviour

Calculation of the oxide electric field

To calculate the transient cell behaviour, it is first of all necessary to calculate the electric fields in the tunnel oxide and in the interpoly oxide, as these determine the charging and leakage currents. These fields are dependent (1) on the applied voltages at the control gate and source/drain (source and drain have always the same potential here) versus the isolated p-well, (2) on the charge residing on the floating gate, (3) on charges in the oxides due to oxide degradation, (4) on the technological parameters (doping concentrations in the isolated p-well, s/d, floating gate and control gate) and (5) on the cell's layout dimensions. An exact solution requires a numerical 2-D device simulation, so for efficient calculation of the cell behaviour some approximations are needed.

The simplified structure used here is presented in Fig. 84. The MOSFET part of the cell up to the floating gate is modelled as three capacitors in parallel, one representing the area where the tunnel oxide covers the isolated p-well area (= channel area, A_{pw}); the second the area where the tunnel oxide covers the under-diffused source/drain doping (S/D-overlap, A_{sd}); and the third the area where the floating gate is outside the active area residing on the STI oxide (A_{fox}). These three capacitors are connected in series with the capacitor formed by the interpoly dielectric (area A_{cg}) to complete the flash cell. The capacitors each consist of the respective silicon oxide layers (thicknesses t_{oxg}, t_{oxcg}, t_{oxfox}), having constant capacitances, and the silicon surfaces, having voltage-dependent capacitances (as indicated by the bent lines in Fig. 84). The doping concentrations in the different silicon or poly-silicon areas (N_{pw}, N_{sd}, N_{fox}, N_{fg}, N_{cg}) are regarded as being homogenous (not location dependent within the respective areas). Any oxide charge (Q_{oxg}, Q_{oxcg}) is regarded to be a 2-D-charge layer located in a certain depth in the oxide. The location is given by the parameters x_{oxg} and x_{oxcg}, respectively; $x_{oxg} = 0$ means that the charge is at the oxide/bulk-silicon interface, $x_{oxg} = 1$ means that the charge is at the oxide/FG interface; $x_{oxcg} = 0$ refers to the FG/interpoly oxide interface and $x_{oxcg} = 1$ to the interpoly-oxide/CG interface (Fig. 86). Interface trapped charge (Q_{itpw}, Q_{itsd}) is regarded as a 2-D charge layer at the bulk-silicon/oxide interface, and it is not

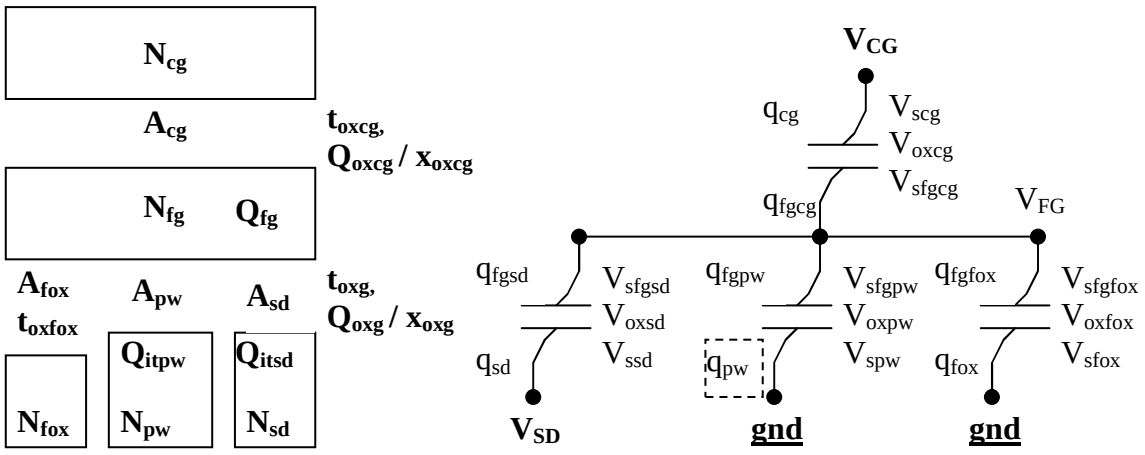


Figure 84: Simplified structure of a flash cell used for the calculations, the **bold** printed are the input parameters, and the regular printed are the calculated values

counted in p-well-accumulation (as it is regarded as trapped charge, positioned energetically in the silicon's energy gap).
The solution of this configuration, which means calculating the electric fields in the different oxide regions out of the input parameters (printed bold in Fig. 84), is still not easily done analytically, and it is done here in a numerical way by a least-squares routine. This routine uses the possibility of a straightforward calculation of V_{CG} out of a given charge residing on the silicon side of the floating-gate/p-well capacitor, q_{pw}. Thus, q_{pw} is initially guessed, and then systematically varied until the right V_{cg} comes out (+/- an allowed error, V_{error}).
The main formulas needed are for the calculation of the conduction band bending in the silicon (surface potential V_s) for a given charge at the respective silicon surface (q_s):

$$V_s = -\frac{kT}{q_0}\left(\frac{q_s}{A}\frac{L_D q_0}{\sqrt{2}\varepsilon_0\varepsilon_{si}kT}\right)^2 \qquad \text{p-type semiconductor in depletion} \qquad (1)$$

$$V_s = -2\frac{kT}{q_0}LN\left(\frac{q_s}{A}\frac{N_D}{n_i}\frac{L_D q_0}{\sqrt{2}\varepsilon_0\varepsilon_{si}kT}\right) - V_{sd} \qquad \text{p-type semiconductor in inversion,} \qquad (2)$$

⇨ V_{sd} must only be added if S/D-junctions are present (not in the FG or the CG)!

$$V_s = \frac{kT}{q_0}LN\left[\left(\frac{q_s}{A}\frac{L_D q_0}{\sqrt{2}\varepsilon_0\varepsilon_{si}kT}\right)^2 + 1\right] \qquad \text{p-type semiconductor in accumulation} \qquad (3)$$

with A being the capacitor's area, N_D the silicon doping concentration, T the temperature, L_D the Debye-length, and n_i the silicon intrinsic carrier concentration. The latter are defined as

$$L_D = \sqrt{\frac{kT\varepsilon_0\varepsilon_{si}}{q_0^2 N_D}} \qquad (4) \qquad n_i = \sqrt{N_L N_V}\, e^{\frac{-E_{gsi}q_0}{2kT}} \qquad (5)$$

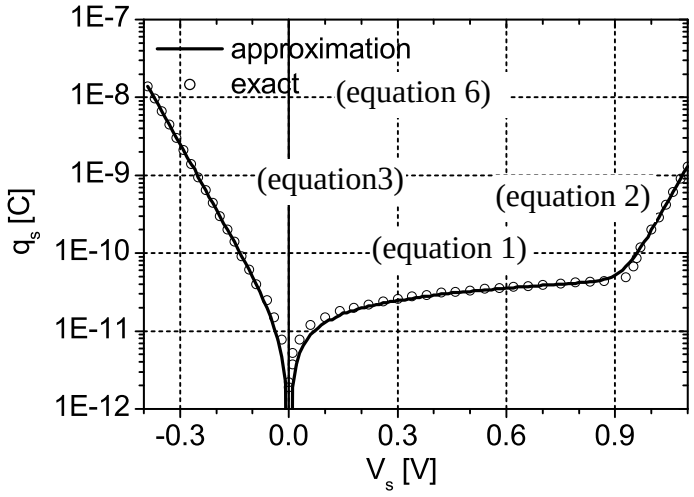
These calculations are an approximation to exact formulas, e.g. [83], p.368, which gives the opposite dependence, the change of the surface charge with the surface voltage (here for an n-type semiconductor) for the case of constant quasi-Fermi levels in all areas:

$$q_s = A\frac{\sqrt{2}\varepsilon_0\varepsilon_{si}kT}{L_D q_0}\left[\left(e^{\frac{V_s q_0}{kT}} - \frac{V_s q_0}{kT} - 1\right) + \frac{n_i^2}{N_D^2}\left(e^{\frac{-V_s q_0}{kT}} + \frac{V_s q_0}{kT} - 1\right)\right]^{1/2} \qquad (6)$$

A comparison between the approximation and the exact formula is shown in Fig. 85. Around $V_s = 0$V and in the regime around the onset of inversion some deviation can be seen.
Other formulas needed are for the calculation of the flatband voltages in the different areas, V_{fbpw}, V_{fbsd}, V_{fbfox} and V_{fbcg}. These are:

$$V_{fbpw} = \frac{kT}{q_0}LN\left(\frac{N_{pw}N_{fg}}{n_i^2}\right) \qquad (7) \qquad V_{fbfox} = \frac{kT}{q_0}LN\left(\frac{N_{fox}N_{fg}}{n_i^2}\right) \qquad (8)$$

$$V_{fbsd} = \frac{kT}{q_0}LN\left(\frac{N_{fg}}{N_{sd}}\right) \qquad (9) \qquad V_{fbcg} = \frac{kT}{q_0}LN\left(\frac{N_{cg}}{N_{fg}}\right) \qquad (10)$$



t_{oxg} = 7.9nm
N_{fg} = 2x10^{20} cm^{-3}
no oxide or interface charges
other simulation parameters
see app.C

Figure 85: Comparison of exact calculation (equation 6) and approximated calculation (equations 1, 2, 3) of the silicon surface charge which is present at a given surface potential

The complete flow of the numerical solution can be seen in figures 89 and 90. The input parameters q_{min} and q_{max} define an interval in which the q_{pw} is searched until V_{cg} is in the allowed limits. The variables q_{up} and q_{low} represent the search interval during the calculation, which is narrowed after every iteration. The parameters Q_{oxgpw} and Q_{oxgsd} are the fraction of the total oxide charge that resides in the p-well and source/drain areas, respectively, and they are calculated by the total oxide charge and the ratio of the areas.
At the end of the routine the electric fields can be calculated from the present q_s values in the different areas:

$$E_{ox} = -\frac{q_s + Q_{it}}{A_{ox}\varepsilon_0\varepsilon_{ox}} \qquad (11)$$

This is the electric field in the oxide at the interface belonging to the respective q_s. At the opposite interface the electric field is different at the presence of oxide charges. This is illustrated in Fig. 86. Only the electric field at the electron-injecting interface is relevant for the tunnelling current (at high electric fields).
The voltage drop across the oxide can be calculated from the different electric fields in the

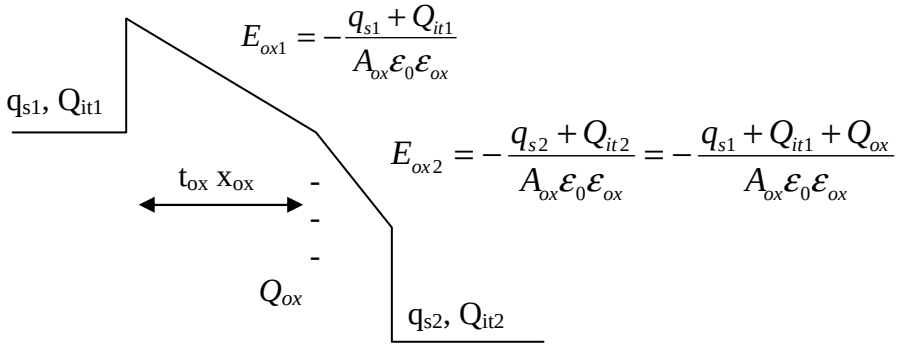
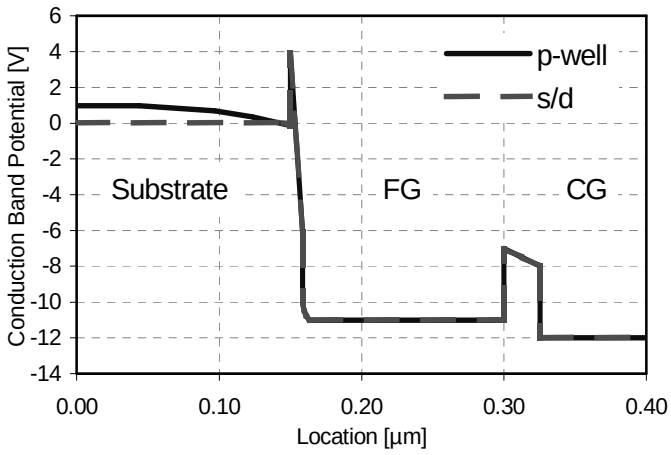


Figure 86: Illustration of the electric field in a silicon oxide layer between two silicon layers at the presence of oxide and interface charge



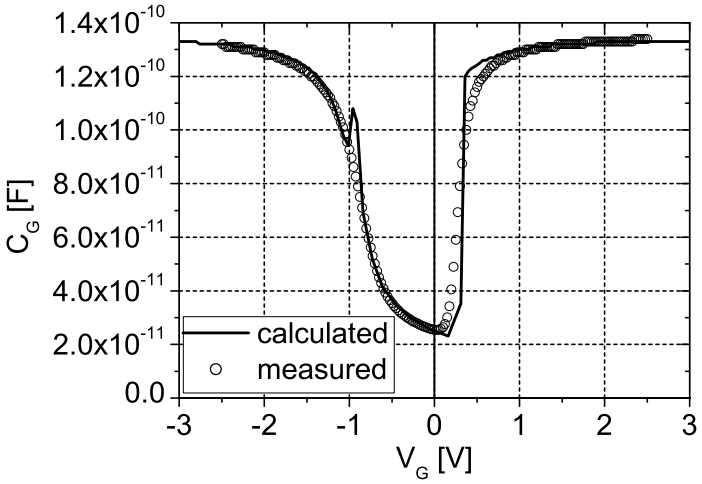
V_{CG} = 12V
V_T = -2V
no oxide or interface charges
other simulation parameters
see app. C

Figure 87: Example of a calculated conduction band potential in a Flash structure

oxide (Fig. 86):

$$V_{ox} = \frac{q_{s1} + Q_{it1}}{A_{ox}\varepsilon_0\varepsilon_{ox}} x_{ox} t_{ox} + \frac{q_{s1} + Q_{it1} + Q_{ox}}{A_{ox}\varepsilon_0\varepsilon_{ox}} (1 - x_{ox}) t_{ox} \qquad (12)$$

The conduction band potential along a vertical cut through the flash cell can easily be drawn after the routine, as the complete conduction band bending is then known. An example is shown in Fig. 87, showing the conduction band of an erased cell (V_T = -2V), when a programming voltage of V_{CG} = 12V is applied.
If the presented calculation is stopped after the floating gate potential has been calculated, also a CV-curve of the MOS capacitor formed by the floating gate can be printed. To do this, V_{FG} is calculated for different q_{pw}. The capacitance is the change of the resulting total q_{sfg} with V_{FG}. The calculated CV curve can be compared with a measured one to calibrate the doping of the isolated p-well and oxide thickness. Fig. 88 shows a measured and a calculated CV curve. Oxide and interface charges are not counted there, as the oxide is not stressed. The oxide thickness and doping concentrations are adjusted to match the curve. The regions the $V_s(q_s)$ approximation is less good can identified. This is between the accumulation and the depletion regions (where V_s becomes 0V) and at the onset of inversion.



t_{oxg} = 7.9nm
N_{fg} = 2x10^{20} cm^{-3}
no oxide or interface charges
other simulation parameters
see app. C

Figure 88: Calculated versus measured CV-curve: t_{oxg}, N_{pw} and N_{fg} have been fitted

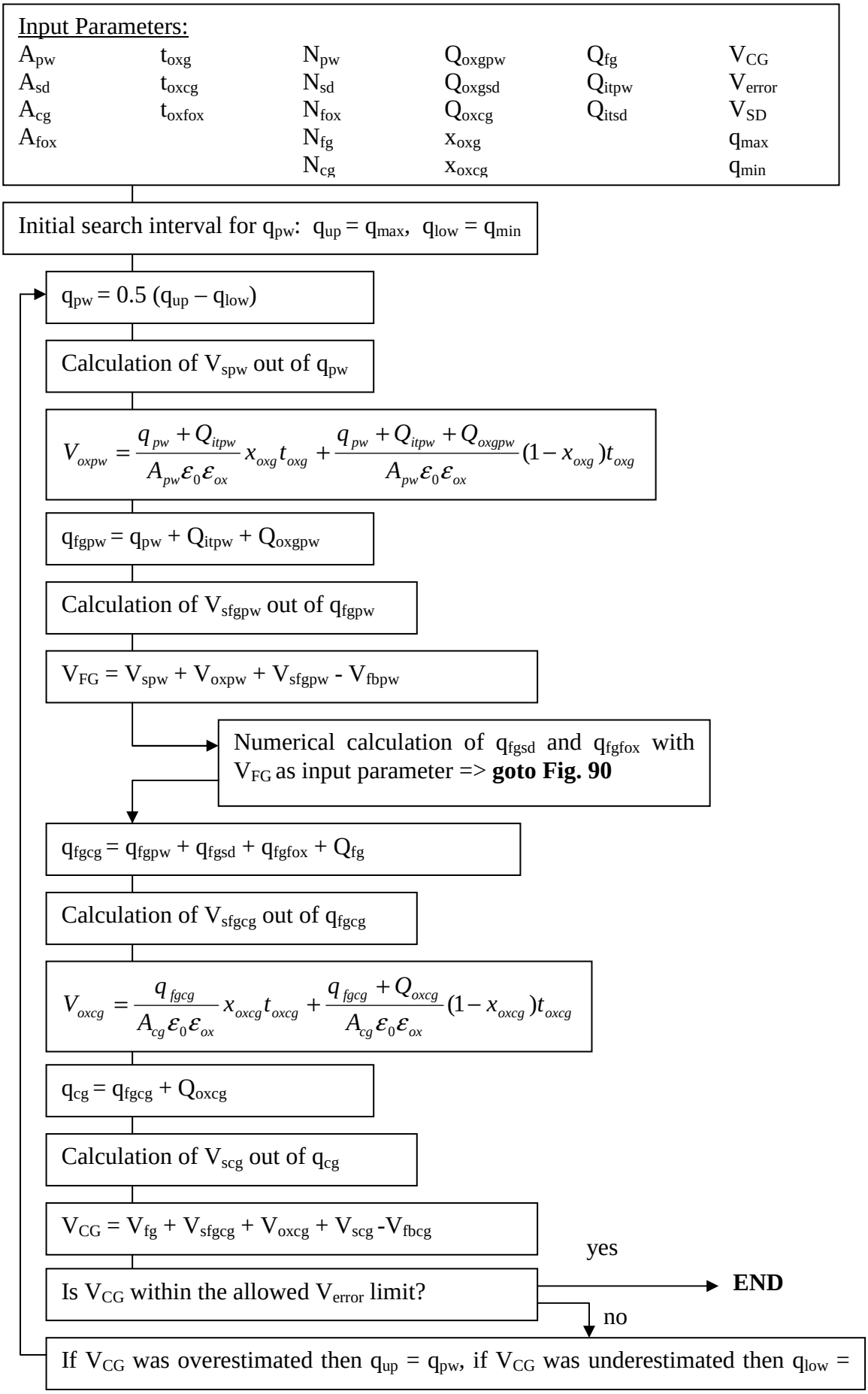


Figure 89: Routine used for calculating the electric field in the oxide – main part

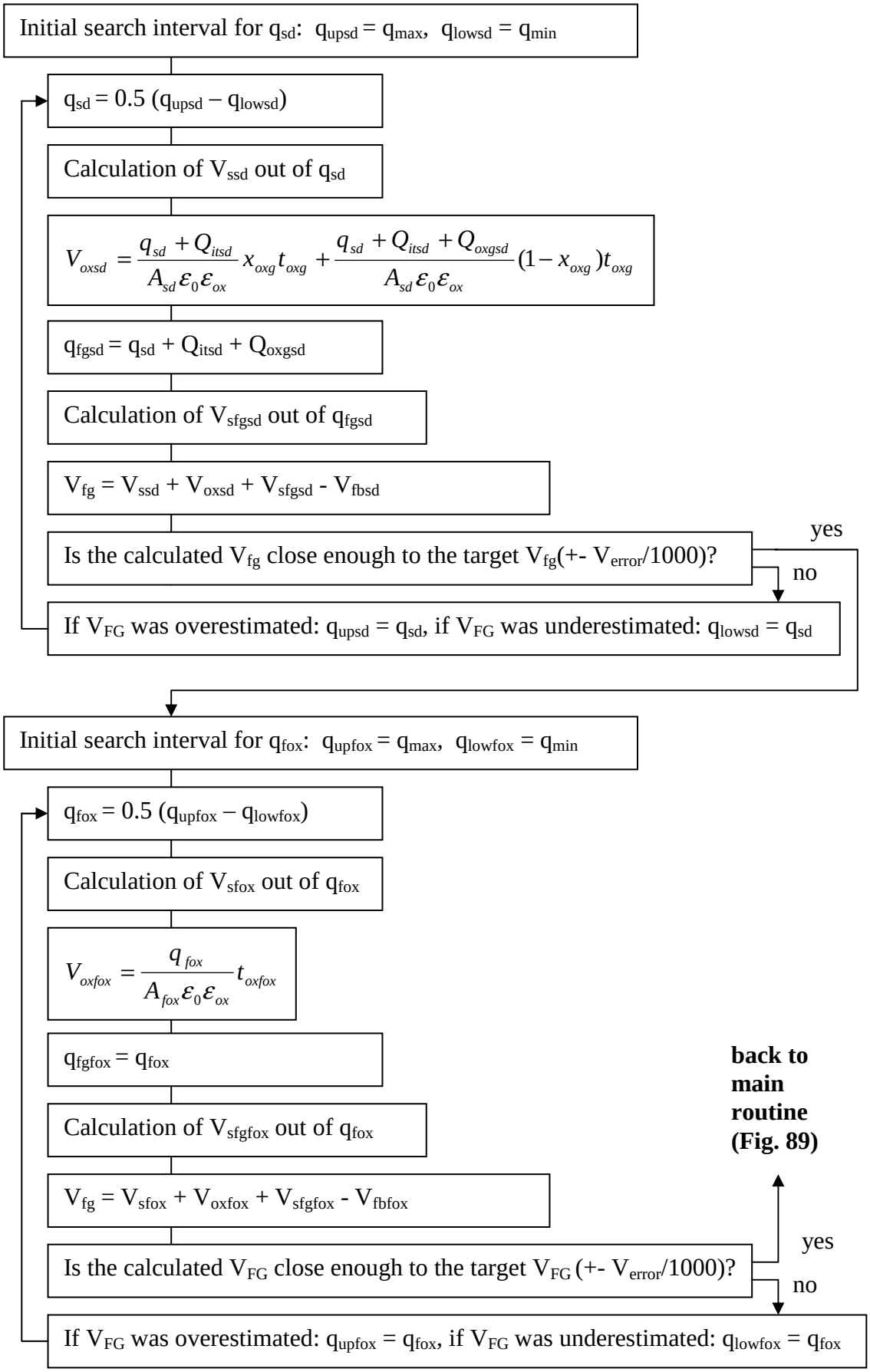


Figure 90: Subroutine used for calculating the electric field in the S/D and STI area

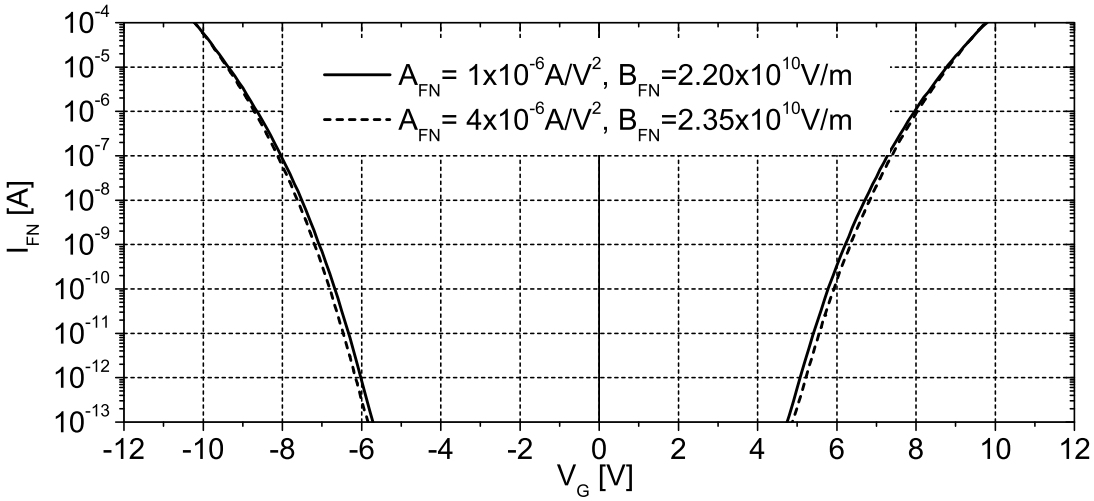


Figure 91: Calculated FN current in a MOS capacitor for different FN-parameters A_{FN} and B_{FN} (other simulation parameters see app. C; no oxide or interface charges)

The current through the tunnel oxide: FN and SILC

The tunnelling current in the tunnel oxide follows in quite good agreement the well-known Fowler-Nordheim equation:

$$I_{FN} = A_{ox} A_{FN} E_{ox}^2 e^{-\frac{B_{FN}}{E_{ox}}} \qquad (13)$$

with A_{ox} being the oxide area, A_{FN} and B_{FN} being FN-parameters that are used as fitting parameters here, but can in principal also be calculated for the silicon/silicon oxide case. A more exact analysis can be found in [5], where also effects like e.g. barrier lowering at high electric fields are explained, which is not taken into account here.
A comparison between the measured and calculated oxide current is shown in chapter 5, Fig. 50. The FN-parameters were fitted to match the curves. The values

$A_{FN} = 1.0\text{x}10^{-6}\text{A/V}^2$
$B_{FN} = 2.2\text{x}10^{10}\text{V/m}$

lead to a good fit for the observed oxide thicknesses and voltage range at room temperature. The influence of the FN-parameters on the curve is shown in Fig. 91, where both have been varied.
The technological parameters that dominate the tunnelling current are the oxide thickness and the poly silicon doping concentration. The doping level of the p-well has a minor influence on the current. This is because the for positive gate voltages the well is shielded by the inversion layer, while for negative gate voltages the surface is accumulated, and the voltage drop in the silicon is small in this case. The doping level of the poly silicon gate is used to fit the current for both, positive and negative voltages simultaneously. The influence can be seen in Fig. 92. The difficulty is that the doping concentration is not homogeneous in the poly silicon of the floating gate. So the current is slightly overestimated for lower voltages and slightly underestimated for higher voltages compared to measurements. This is compensated partly by choosing values for A_{FN} and B_{FN} that result in a not so steep curve.

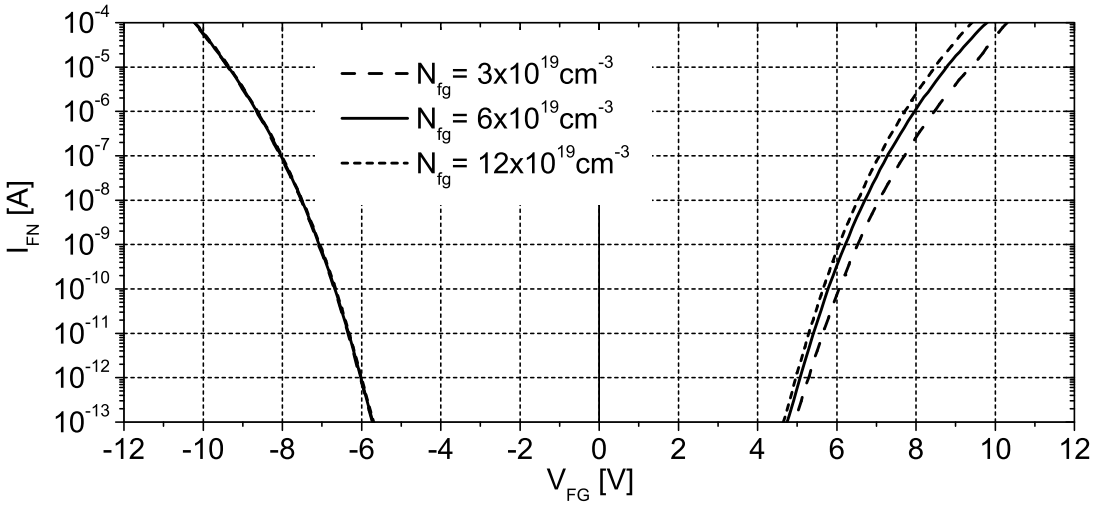


Figure 92: Calculated FN current in a MOS capacitor for different doping levels of the gate poly silicon N_{fg} (other simulation parameters see app. C; no oxide or interface charges)

The steady state SILC at low oxide electric fields can be empirically modelled as a current that is proportional to the density of oxide charge and has FN-like electric field dependence with reduced oxide barrier height [40]. This translates to an FN equation with adjusted FN-parameters, A_{SILC} and B_{SILC}, and which is multiplied with the of oxide charge per oxide area Q_{ox}/A_{ox} and a proportionality factor c_{SILC}:

$$I_{SILC} = c_{SILC} \frac{Q_{ox}}{A_{ox}} A_{ox} A_{SILC} E_{ox}^2 e^{-\frac{B_{SILC}}{E_{ox}}} \tag{14}$$

For a barrier height of 0.9eV, as proposed in [40], and an effective electron mass in the oxide of $m_{ox} = 0.5m_0$ the SILC – parameters calculate to [84]

$A_{SILC} = 3.4\text{x}10^{-6}$ A/V^2
$B_{SILC} = 4.1\text{x}10^{9}$ V/m.

The proportionality factor c_{SILC} has been used as fitting factor. A good agreement with the measurements has been found for

$c_{SILC} = 8.0\text{x}10^{-11}$ m^2/C.

It should be noted that the oxide charge Q_{oxg} also has an impact on the electric field E_{ox}, which is taken into account in the calculation described in the first part of this appendix.
The oxide charge accumulation during an FN-stress, which leads to the SILC, can also be calculated from the stress conditions by an empirical equation [40]:

$$Q_{ox} = A_{ox} k_{ox} \left[\frac{Q_{inj}}{A_{ox}} \right]^{ai} J_{inj}^{bi} \tag{15}$$

where Q_{inj} is the injected charge, J_{inj} the current density of a constant-current stress. The parameters ai and bi are fitting factors. Good agreement with measurements has been found for

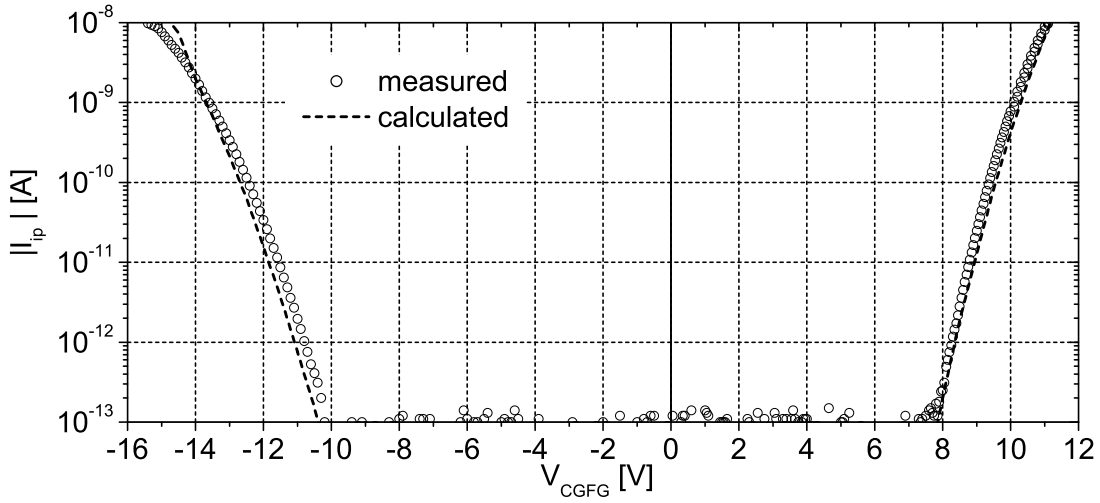


Figure 93: Calculated FN current in an FG/CG capacitor (simulation parameters see app. C)

ai = 0.5
bi = 0.3

which is close to values reported in literature [40]. The proportionality factor k_{ox} is also used for fitting, with

$k_{ox} = 2.8\text{x}10^{-6}\ C^{0.5}m^{-1}(A/m^2)^{-0.3}$

giving good results for ai = 0.5 and bi = 0.3.
A comparison between calculated and measured SILC can be seen in chapter 5, Fig. 53.
The SILC component of the current is not needed in the calculation of the programming behaviour of the cell, as there the electric field is in a range where the current is dominated by the FN component. But by including the SILC in the transient cell calculations, the cells retention can be calculated (at least for cells with regular SILC, of course not for the tail-cells showing anomalous SILC).

The current through the interpoly oxide: modified FN tunnelling

The tunnelling current through the interpoly oxide has a strongly unsymmetrical behaviour for positive and negative voltages with respect to V_{CGFG} = 0V. The reason for this is the geometrically enhanced electric field at the corner of the floating gate. When this corner is the electron-injecting surface (for positive V_{CGFG}) the current density is significantly raised at this point by several orders of magnitude compared to the other areas of the oxide. Also for negative gate voltages a higher current as calculated by a regular FN-equation for this oxide thickness has been found. The measured curve nevertheless has FN-like I-V behaviour. Thus it was tried to empirically model the measured curve by using two fitting parameters. One is f_{ip} by which the oxide electric field is multiplied. The second is a_{ip}, which is an effective width that is multiplied with the length of the FG corner, L_{fge}, in order to get an effective area, by which the total FN current can be calculated.

$$I_{ip} = a_{ip} L_{fge} A_{SILC} \left(f_{ip} E_{ox}\right)^2 e^{-\frac{B_{SILC}}{\left(f_{ip} E_{ox}\right)}} \qquad (16)$$

The parameters f_{ip} and a_{ip} have different values for positive and negative gate voltages, thus resulting in two pairs of parameters, f_{ipp} / a_{ipp} and f_{ipn} / a_{ipn}. Fig. 93 shows simulation results in comparison with measured curves for

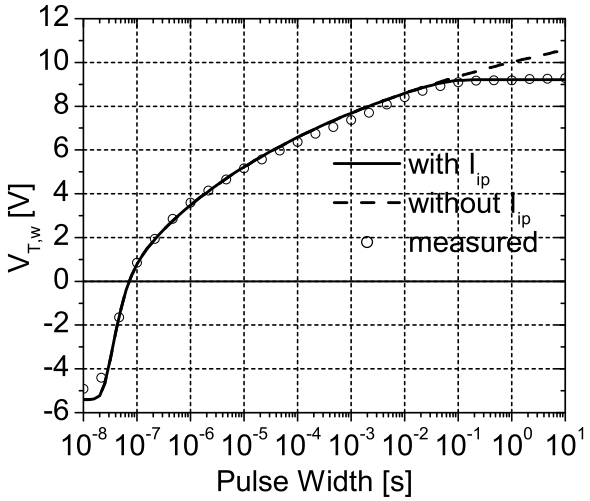


Figure 94: Comparison of a transient flash cell simulation with and without taking the interpoly oxide current into account; V_{CG}=18V, no oxide or interface charges

f_{ipp} = 1.85 / a_{ipp} = 100nm
f_{ipn} = 1.40 / a_{ipn} = 100nm

The current through the interpoly oxide has to be taken into account in the transient simulation of the flash cells when trying to simulate the V_T - saturation observed at high programming voltages when writing the cell to high V_T-states. As the amount of negative charge rises on the floating gate during the writing process, the absolute FG potential rises and the current through the interpoly oxide rises as well, while the current through the tunnel oxide becomes lower. At some point both currents have equal values and the floating gate is not further charged and the V_T saturates. This effect could be simulated with the presented model for the current in the interpoly oxide. For fitting the flash cell transient measurements the field enhancement could be modelled with the same parameters f_{ipp} / a_{ipp}. The erase curves did not show saturation in the investigated ranges of programming voltage and programming times, as the interpoly oxide current is much lower in this case. Therefore only f_{ipp} and a_{ipp} could be verified here, and interpoly leakage is not taken into account in the calculation of the flash cell erasing behaviour.
A comparison between a transient simulation with and without taking the current through the interpoly oxide into account is presented in Fig. 94.

Transient simulation

When the different oxide fields and the tunnelling currents are known, a transient simulation of the flash cell can be done. This is done by a numerical step-by-step integration of the floating gate charge. First, the electric fields and the resulting tunnelling currents are calculated as described above. The currents are approximated to be constant for a small time step of the duration dt. The change of the floating gate charge in that time step is calculated by multiplying the net charging current of the floating gate with dt. With this new floating gate charge as input parameter, the electric fields are calculated again, leading to new tunnelling currents, which then change the floating gate charge in the next time step. This is done until the final programming time is reached. A non-constant control-gate voltage (a ramped or otherwise shaped programming pulse) can be taken into account during the integration over time.

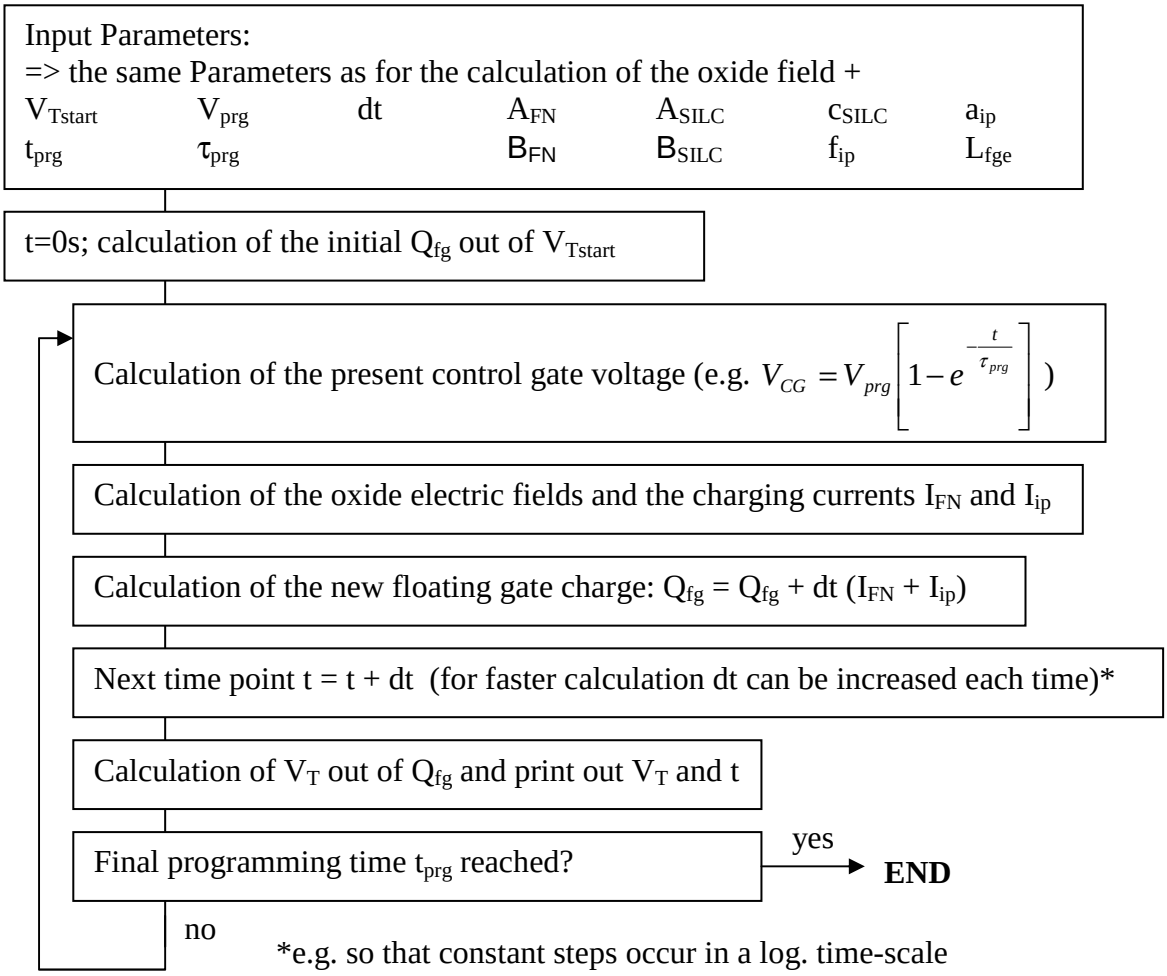


Figure 95: Routine for the calculation of the transient cell behaviour

The result of the calculation is the change of floating gate charge Q_{fg} over time. The floating gate charge can be translated into the cell's threshold voltage V_T and vice versa. The threshold voltage is characterised by a specific surface charge q_{pwVT} in the channel area of the flash cell. This is described by the following equation [83]:

$$q_{pwVT} = -\sqrt{4\varepsilon_0\varepsilon_{si}A_{pw}^2 q_0 N_{pw}\left(\frac{kT}{q_0}LN\left(\frac{N_{pw}}{n_i}\right)\right)} \qquad (17)$$

Knowing the depletion charge in the p-well, V_{cg} can be calculated by the routine described in the first part of this appendix. The V_T of the flash cell is equal to the resulting V_{cg}. The other direction (calculating Q_{fg} out of V_T) is not so straightforward. It can be done numerically by calculating V_{cg} out of q_{pwVT} and varying Q_{fg} until the resulting V_{cg} is equal to the given V_T (within allowed error limits).
Fig. 95 shows the complete flow of the calculation of the transient flash cell behaviour. Figure 63 in chapter 5 presents calculation results compared with measurements, showing an excellent agreement over a wide range of programming voltages and flash cell threshold voltages.

Appendix B: Oxide Charge Extraction

The presented routine for calculating the transient characteristics of the flash cells presented in appendix A is used here for a closer investigation of the endurance curves, especially how the closure and shift of the V_T-window depends on the charging state of the oxide and the interface to the silicon substrate. The oxide charge density and the interface charge density can be identified as the origins for the V_T-window closure and shift, and can separately be determined from the measured endurance curves, thus giving a better idea of the state of a degraded cell.

Fig. 96a and Fig. 97a show the results of calculations of $V_{T,w}$ and $V_{T,e}$ for different oxide and interface charge densities. The parameters used for simulation are given in appendix C. The programming conditions have been chosen to get a 4V-V_T-window with $V_{T,w}$ = 2V and $V_{T,e}$ = -2V in the initial state (without charges). Then, $V_{T,w}$ and $V_{T,e}$ have been calculated with the same programming conditions, with changed oxide and interface charge densities. The result indicates that, independent from interface charge densities, the oxide charge leads to a closure of the V_T-window, $dV_{T,close}$ (= difference between the V_T-window of the degraded cell and the initial V_T-window), while interface charge, independent from oxide charge density, leads to a shift of the whole V_T-window. Independent means that the closure is the same for different interface charge densities, and the shift is the same for different oxide charge densities. Furthermore, the (negative) oxide charge leads to a lowering of $V_{T,w}$. These dependencies are shown in Fig. 96b and Fig. 97b. They have been fitted by second order polynomial equations:
Dependence of oxide charge density on V_T-window closure:

$$Q'_{oxg}\left[\frac{C}{cm^2}\right] = -2.45x10^{-12}\left(dV_{T,close}[V]\right)^2 - 2.63x10^{-7}\left(V_{T,close}[V]\right) \qquad (18)$$

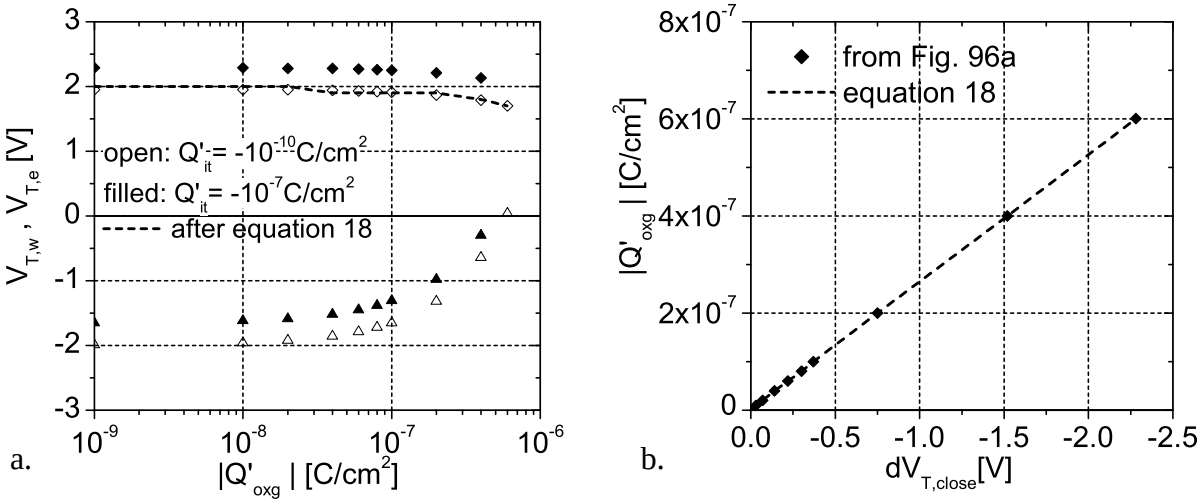


Figure 96: a. Calculated dependence of $V_{T,w}$ and $V_{T,e}$ on Q'_{oxg} for 2 different values of Q'_{it}
b. Relationship between V_T-window closure and Q'_{oxg} (equal for both Q'_{it})

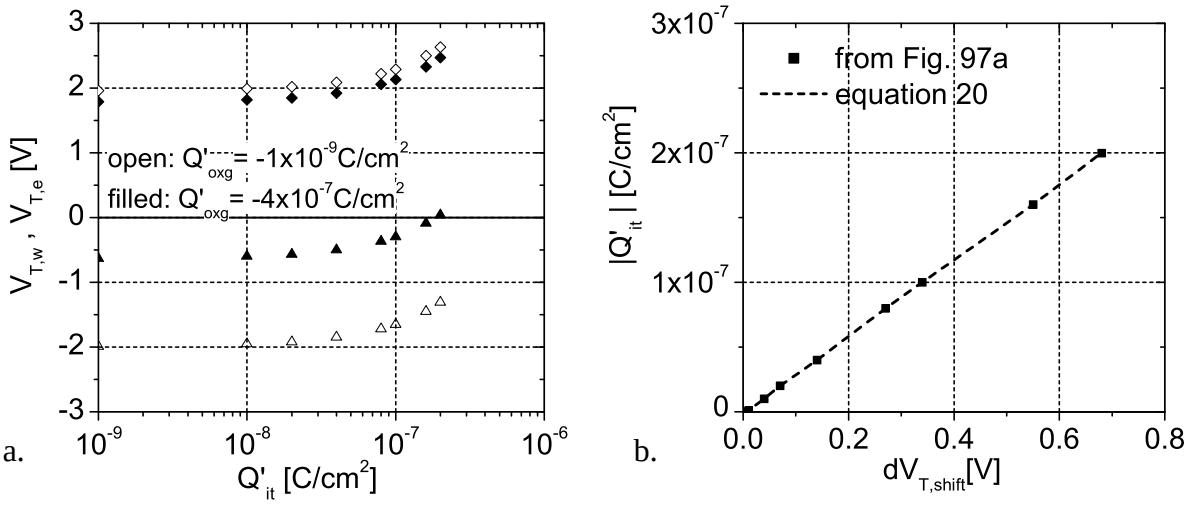


Figure 97: a. Calculated dependence of $V_{T,w}$ and $V_{T,e}$ on Q'_{it} for 2 different values of Q'_{oxg}
b. Relationship between V_T-window shift and Q'_{it} (equal for both Q'_{oxg})

Dependence of $V_{T,w}$-lowering on oxide charge:

$$dV_{T,w,c}[V] = 1.44x10^{10} Q'^{2}_{oxg}\left[\frac{C}{cm^2}\right] - 4.28x10^{5} Q'_{oxg}\left[\frac{C}{cm^2}\right] \qquad (19)$$

Dependence of interface charge on V_T-window shift:

$$Q'_{it}\left[\frac{C}{cm^2}\right] = -2.44x10^{-9}\left(dV_{T,shift}[V]\right)^2 - 2.97x10^{-7}\left(dV_{T,shift}[V]\right) \qquad (20)$$

Note that these equations are only valid for this unique memory cell, modelled with the given set of parameters.
With help of these equations, a measured endurance curve can now be analysed. From the V_T-

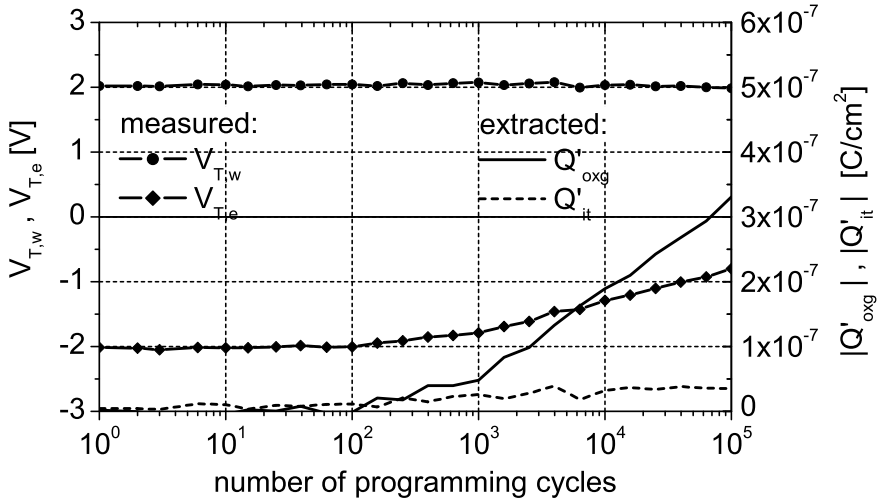


Figure 98: Analysis of a measured endurance curve: extraction of Q'_{oxg} and Q'_{it} from the measured V_T-window closure and V_T-window shift by the described method

window closure the oxide charge can be calculated using equation 17. This oxide charge leads to a theoretical $V_{T,w}$ shift $dV_{T,w,c}$, given by equation 18, if no interface charge are taken into account. The difference between this result and the actually measured shift $dV_{T,w,m}$ of $V_{T,w}$ after endurance is the V_T-window shift: $dV_{T,shift} = dV_{T,w,m} - dV_{T,w,c}$. From this shift the interface charge can be calculated using equation 19.
Fig. 98 shows the result of this chain of calculations. Starting from a measured endurance curve, first the evolution of the oxide charge is calculated out of the V_T-window closure. Using this result and the $V_{T,w}$ behaviour, the V_T-window shift is calculated, which finally gives the interface charge.
It can be seen that the major degradation is due to oxide charges. After 100k cycles under the given programming conditions, a value of $Q'_{oxg} = 3.3x10^{-7}$ C/cm^2 is reached. The interface charge is found to be about $Q'_{it} = 3.5x10^{-8}$ C/cm^2. It must be noted that the resulting curves are not very smooth. This routine can only give a qualitative picture of what is going on, rather than quantitative values.

Appendix C: Summary of Simulation Parameters

If not stated differently in the respective figures, the following parameters have been used for the different simulations:

Table 3: Parameters for simulating the behaviour of the tunnel oxide MOS capacitor (calculation of $q_s(V_s)$; CV; FN-current; SILC)

A_{pw}	31,506	μm^2		N_{fg}	$6.6x10^{19}$	cm^{-3}		ai	0.5	1
A_{sd}	29.2	μm^2		A_{FN}	$1.0x10^{-6}$	A/V^2		bi	0.3	1
t_{oxg}	8.1	nm		B_{FN}	$2.2x10^{10}$	V/m		k_{ox}	$2.8x10^{-6}$	$C^{0.5}m^{-1}(A/m^2)^{-0.3}$
x_{oxg}	0.5	1		A_{SILC}	$3.4x10^{-6}$	A/V^2		V_{error}	0.01	V
N_{pw}	$6.6x10^{16}$	cm^{-3}		B_{SILC}	$4.1x10^{9}$	V/m		q_{max}	$3x10^{-8}$	C
N_{sd}	$2.0x10^{19}$	cm^{-3}		c_{SILC}	$8.0x10^{-11}$	m^2/C		q_{min}	$-3x10^{-8}$	C

No oxide and interface charges unless stated in the respective figures.

Table 4: Parameters for simulating the behaviour of the CG/FG capacitor (calculation of the current through the interpoly oxide)

A_{cg}	16,800	μm^2		a_{ipn}	100	nm		B_{FN}	$2.2x10^{10}$	V/m
t_{oxcg}	24	nm		L_{fge}	24,000	µm		q_{max}	$2x10^{-8}$	C
f_{ipp}	1.85	1		N_{cg}	$1.0x10^{20}$	cm^{-3}		q_{min}	$-2x10^{-8}$	C
a_{ipp}	100	nm		N_{fg}	$5.0x10^{19}$	cm^{-3}		V_{error}	0.01	V
f_{ipn}	1.40	1		A_{FN}	$1.0x10^{-6}$	A/V^2				

Table 5: Parameters for simulating the behaviour of the 1-transistor flash memory cell (calculation of the transient behaviour; oxide charge extraction from endurance curves)

A_{pw}	0.108	μm^2		N_{sd}	$2.0x10^{19}$	cm^{-3}		f_{ipp}	1.85	1
A_{sd}	0.007	μm^2		N_{fg}	$5.0x10^{19}$	cm^{-3}		a_{ipp}	100	nm
A_{cg}	1.2	μm^2		N_{cg}	$1.0x10^{20}$	cm^{-3}		L_{fge}	0.28	µm
A_{fox}	1.2	μm^2		N_{fox}	$6.6x10^{16}$	cm^{-3}		τ_{prg}	$1.5x10^{-8}$	s
t_{oxg}	8.6	nm		x_{oxg}	0.5	1		dt	$10^{-11*)}$	s
t_{oxcg}	25.5	nm		x_{oxcg}	0.5	1		V_{error}	0.01	V
t_{oxfox}	500	nm		A_{FN}	$1.0x10^{-6}$	A/V^2		q_{max}	10^{-12}	C
N_{pw}	$1.0x10^{17}$	cm^{-3}		B_{FN}	$2.2x10^{10}$	V/m		q_{min}	-10^{-12}	C

No oxide and interface charges, unless stated in the respective figures.
*) this is the initial value; dt is raised after each iteration so that in a logarithmic time-scale 100 calculations are done per time-decade

Appendix D: Dimensions of the Test Devices

In the following tables important dimensions of the different devices that have been used for electrical characterisation are listed.
The oxide thicknesses are given in the respective figures.

Table 6: Dimensions of the flash cells

	1-T cell		2-T cell		Split-gate cell	
L_{fg}	0.32	µm	0.28	µm	0.26	µm
W_{fg}	0.36	µm	0.33	µm	0.33	µm
A_{cg}	1.25	$µm^2$	0.75	$µm^2$	1.27	$µm^2$
L_{fge}	0.56	µm	0.93	µm	3.39	µm
A_{fox}	1.19	$µm^2$	0.50	$µm^2$	0.54	$µm^2$
t_{oxfox}	500	nm	500	nm	500	nm
L_{sg}	-		0.24	µm	0.40	µm

Table 8: Dimensions of the MOS capacitors

	Large area		Large S/D junction	
A_{fg}	31,500	$µm^2$	92.4	µm2
L_{sd}	1460	µm	660	µm

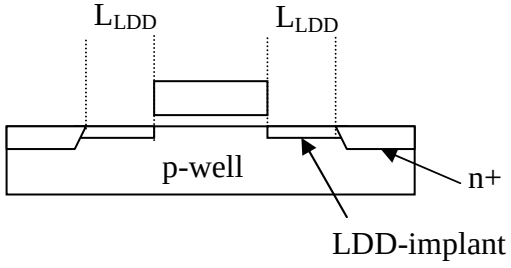
Table 7: Dimensions of the interpoly capacitor

	Interpoly capacitor	
A_{cg}	19,800	$µm^2$
L_{fge}	24,000	µm

Table 9: Dimensions of the HVMOS transistors

	HVNMOS without LDD		HVNMOS		HVPMOS	
L_G	0.8	µm	0.6	µm	0.8	µm
W_G	25	µm	25	µm	25	µm
L_{LDD}	-		0.6	µm	-	

HVNMOS:



Appendix E: Determination of V_T and t_{OX}

V_T measurement

The threshold voltage V_T of MOS transistors and flash cells has been determined by finding the gate voltage V_G or V_{CG}, respectively, that leads to a defined value of the drain current at a given drain bias:

$I_D = (W_{FG}/L_{FG})$ x 0.1µA (21)

for n-channel devices at V_{DS} = 0.1V and

$I_D = -(W_{FG}/L_{FG})$ x 0.1µA (22)

for p-channel devices at V_{DS} = -0.1V.
This procedure is e.g. in agreement with [87] and [88].

Electrical C_{OX} measurement

The electrical measurement of the tunnel oxide thickness has been done by an analysis of CV-measurements of MOS-capacitors in accumulation. At V_G=-2.5V for n-channel devices and V_G=+2.5V for p-channel devices the capacitance C_{MOS} and its deviation dC_{MOS}/dV_G has been determined. From this the oxide capacitance has been approximated as [91]:

$$C_{ox} = C_{MOS}\left(1-\frac{2\frac{kT}{q_0}\left|\frac{dC_{MOS}}{dV_G}\right|}{C_{MOS}}\right)^{-1}\left(1+\left(\frac{C_{MOS}}{2\frac{kT}{q_0}\left|\frac{dC_{MOS}}{dV_G}\right|}\right)^{-0.5}\right). \quad (23)$$

The oxide thickness has been calculated from C_{ox}, assuming a dielectric constant of ε_{ox}=3.9:

$$t_{OX} = \varepsilon_0\varepsilon_{ox}\frac{A_{ox}}{C_{ox}} \quad (24)$$

This method does not take quantum mechanical effects into account. This is the reason for a difference in the order of 0.5nm when comparing this result to other values obtained by TEM or ellipsometry.

List of Abbreviations

III-V	refers to compound semiconductors consisting of group III and V elements
AG-AND	Assist Gate AND (memory array arrangement)
Al	Aluminum
ARC	Anti-Reflective Coating
As	Arsenic
B	Boron
BiCMOS	Bipolar-CMOS (combination of bipolar and CMOS devices on one chip)
BL	Bit Line
BSIM3v3	MOS transistor model by Berkley university
BTBT	Band To Band Tunneling
C	Carbon
CAD	Computer Aided Design
CHISEL	Channel Initiated Secondary Electron
CMOS	Complementary MOS
CMP	Chemical Mechanical Polishing
CG	Control Gate
Co	Cobalt
CPU	Central Processing Unit
CV	Capacitance Voltage
CVD	Chemical Vapour Deposition
D	Drain
DC	Direct Current
DE	Drain Erase
DGT	Dual Gate oxide
DiNOR	Divided bitline NOR (memory array arrangement)
DRAM	Dynamic Random Access Memory
DUV	Deep Ultra Violet
E	refers to the erase operation
EEPROM	Electrically Erasable Programmable Read Only Memory
EPROM	Electrically Programmable Read Only Memory
EOT	Equivalent Oxide Thickness
ESD	ElectroStatic Discharge
ETOX	EPROM Tunnel OXide (a type of stacked gate flash cell)
FeRAM	Ferroelectric RAM
FET	Field-Effect-Transistor
FG	Floating Gate
FImp	Flash Implant (mask step of the flash process)
FLOTOX	FLOating gate Tunnel OXide (a type of EEPROM cell)
FN	Fowler-Nordheim
G	Gate
Ge	Germanium
gnd	Ground
HBT	Hetero Bipolar Transistor
HDP	High Density Plasma
HF	Hydro-Fluoric acid

HIMOS	High Injection MOS (a type of split-gate Flash cell invented by IMEC)
HV	High Voltage
HVMOS	High-Voltage MOSFET
HVNMOS	High-Voltage NMOS transistor
HVPMOS	High-Voltage PMOS transistor
IC	Integrated Cicuit
ILD	Inter-Layer-Dielectric
ipw	isolated p-well
LDMOS	Lateral Diffused MOS
LDD	Lightly Doped Drain (here: a separately implanted area in the HVMOS)
LPCVD	Low Pressure Chemical Vapour Deposition
MEMS	Micro Electro-Mechanical System
MIM	Metal-Isolator-Metal layer stack
MOS	Metal-Oxide-Semiconductor
MOSFET	Metal-Oxide-Semiconductor Field-Effect-Transistor
MRAM	Magnetic RAM
NAND	memory array arrangement
NMOS	MOSFET with n-type Channel
NOR	memory array arrangement
NVM	Non-Volatile Memory
n-well	substrate area with n-type doping
ONO	Oxide Nitride Oxide
P	Phosphorus / or: refers to the program operation
PCM	Phase Change Memory
PE	Poly Erase
PECVD	Plasma Enhanced Chemical Vapour Deposition
PG	Program Gate
PMOS	MOSFET with p-type Channel
PN	refers to a junction between p-type and n-type silicon
Poly-Si	Poly silicon
p-well	substrate area with p-type doping
R	refers to the read operation
RAM	Random Access Memory
RF	Radio Frequency
RIE	Reactive Ion Etching
RTA	Rapid Thermal Annealing
S	Source
Salicide	Self-Aligned silicide
S/D	Source and Drain
SEM	Scanning Electron Microscopy
SG	Select Gate
Si	Silicon
Si_3N_4	Silicon Nitride
SIC	Selectively Implanted Collector
SiGe	Silicon-Germanium
SiGe:C	Carbon doped Silicon-Germanium
SILC	Stress Induced Leakage Current
SOC	System On Chip
SONOS	Silicon-Oxide-Nitride-Oxide-Silicon layer stack
SPICE	Simulation Program for Integrated Circuits Emphasis
SRAM	Static Random Access Memory

STI	Shallow Trench Isolation
TEM	Transmission Electron Microscopy
TEOS	Tetraethylorthosilicate = $SiO_4C_8H_20$
UV	Ultra Violett
VLSI	Very Large Scale Integration
W	refers to the write operation
WL	Word Line
WLAN	Wireless Local Area Network

List of Symbols

A	area
ai	fitting parameter used in the calculation of the oxide degradation
a_{ip}	fitting parameter used in the calculation of I_{ip}
a_{ipp}	a_{ip} for positive V_{CGFG}
a_{ipn}	a_{ip} for negative V_{CGFG}
A_{cg}	area of the interpoly capacitor formed between the CG and the FG
A_{fox}	area of the capacitor formed by the part of the FG that resides on STI oxide
A_{ox}	area of an oxide layer
A_{pw}	area of the tunnel oxide capacitor formed between the FG and the p-well
A_{sd}	area of the tunnel oxide capacitor formed between the FG and the S/D junction
A_{FN}	linear FN parameter
A_{SILC}	modified linear FN parameter for SILC modelling
B_{FN}	exponential FN parameter
bi	fitting parameter used in the calculation of the oxide degradation
B_{SILC}	modified exponential FN parameter for SILC modelling
BV_{CEO}	breakdown voltage between Collector and Emitter with open base
C_G	gate capacitance
c_{SILC}	proportionality factor for SILC modelling
dt	initial time interval used for transient flash cell simulation
$dV_{T,close}$	V_T-window closure during endurance testing
$dV_{T,w,c}$	calculated decrease of $V_{T,w}$ during endurance testing due to oxide charges
$dV_{T,w,m}$	measured shift of $V_{T,w}$ during endurance testing
$dV_{T,shift}$	V_T-window shift during endurance testing due to interface trapped charges
ε_{si}	permittivity of silicon
ε_{ox}	permittivity of silicon oxide
E_{ox}	electric field in an oxide layer
f_{ip}	fitting parameter used during the calculation of I_{ip}
f_{ipp}	f_{ip} for positive V_{CGFG}
f_{ipn}	f_{ip} for negative V_{CGFG}
f_T	transit frequency
I_{BL}	current measured at the contact to the buried layer
I_D	drain current
$I_{D,e}$	drain current of an erased flash memory cell
I_{ECs}	HBT emitter current with collector and base having the same potential
I_{FN}	FN current component of the tunnel oxide current
I_{ip}	current through the interpoly oxide
I_{ipw}	current measured at the contact to the isolated p-well
I_L	leakage current
I_{SD}	total current at the S/D junctions in the case of connected S/D
I_{SILC}	SILC component of the tunnel oxide current
J_{inj}	current density during an FN-stress of an oxide layer
k_{CG}	control Gate coupling ratio
k_{ox}	proportionality factor for calculating the oxide degradation during FN-stress
L_D	Debye length
L_{FG}	gate length of the floating gate

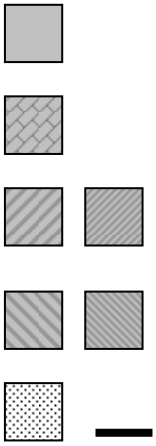
L_G	gate length
L_{fge}	length of the FG/CG corner in a flash cell or a FG/CG capacitor structure
m_{ox}	effective electron mass in an oxide layer
m_0	electron mass in vacuum
N_{cg}	poly-Si doping concentration within the control gate
N_D	doping concentration
N_{fg}	poly-Si doping concentration within the floating gate
N_{fox}	doping concentration in p-well regions under the STI oxide
N_{LDD}	dose of implanted ions in the LDD regions of the HVNMOS
N_{pw}	doping concentration in p-well regions
n_i	intrinsic carrier concentration in silicon
N_{sd}	doping concentration in the S/D – FG overlap region of a flash cell
Q_{fg}	charge residing on the floating gate
Q_{inj}	injected charge during an FN-stress of an oxide layer
Q_{it}	interface trapped charge
Q'_{it}	interface trapped charge per area
Q_{itpw}	interface trapped charge in a MOS structure at the oxide / p-well interface
Q_{itsd}	interface trapped charge in a MOS structure at the oxide / S/D interface
Q_{ox}	oxide charge
Q_{oxcg}	oxide charge in the interpoly oxide
Q_{oxg}	oxide charge in the tunnel oxide
Q'_{oxg}	Q_{oxg} per area
Q_{oxgpw}	fraction of the oxide charge residing over the p-well region
Q_{oxgsd}	fraction of the oxide charge residing over the S/D region
q_{cg}	surface charge at the CG side of the CG/FG capacitor
q_{fox}	surface charge at the substrate side of the FG/substrate capacitor (STI region)
q_{fgcg}	surface charge at the FG side of the CG/FG capacitor
q_{fgfox}	surface charge at the FG side of the FG/substrate capacitor (STI region)
q_{fgpw}	surface charge at the FG side of the FG/substrate capacitor (p-well region)
q_{fgsd}	surface charge at the FG side of the FG/substrate capacitor (S/D region)
q_{pw}	surface charge at the substrate side of the FG/substr. capacitor (p-well region)
q_{pwVT}	q_{pw} if $V_{CG} = V_T$ is applied to the control gate
q_s	surface charge
q_{sd}	surface charge at the substrate side of the FG/substrate capacitor (S/D region)
q_{max}, q_{min}	initial search interval for q_{pw} in the numerical calculation of the oxide field
q_{up}, q_{low}	search interval for q_{pw} during the numerical calculation of the oxide field
q_{upfox}, q_{lowfox}	search interval for q_{fox} during the numerical calculation of the oxide field
q_{upsd}, q_{lowsd}	search interval for q_{sd} during the numerical calculation of the oxide field
T	absolute temperature
t_{ox}	thickness of an oxide layer
t_{oxcg}	thickness of the interpoly oxide
t_{oxfox}	thickness of the STI oxide (= field oxide)
t_{oxg}	thickness of the tunnel oxide
τ_{prg}	parameter for modelling the pulse shape during transient flash cell simulation
t_{prg}	programming pulse width
t	time
V_{BE}	voltage between base and emitter
V_{BL}	voltage (versus substrate) applied to the buried layer
V_{CB}	voltage between collector and base
V_{CG}	control gate voltage (versus substrate)
$V_{CG,e}$	control gate voltage (versus substrate) applied for cell erasing

V_{CGFG}	voltage between the CG and the FG
$V_{CG,r}$	control gate voltage (versus substrate) applied for cell reading
$V_{CG,w}$	control gate voltage (versus substrate) applied for cell writing
V_D	drain voltage (versus substrate)
V_{DD}	positive supply voltage in an electric circuit
V_{DS}	drain voltage (versus source)
$V_{DS,r}$	drain voltage (versus source) applied for cell reading
V_{error}	allowed error at the numerical calculation of voltages
V_{fbcg}	flatband voltage of the FG / CG capacitor
V_{fbpw}	flatband voltage of a MOS capacitor in the p-well region
V_{fbfox}	flatband voltage of a MOS capacitor in the STI oxide region
V_{fbsd}	flatband voltage of a MOS capacitor in the S/D region
V_{FG}	floating gate voltage (versus substrate)
V_G	gate voltage (versus substrate)
V_{ipw}	voltage (versus substrate) applied to an isolated p-well
V_{off}	V_{CG} applied to the non-selected cells in a 1-transistor NOR memory
V_{ox}	voltage drop across an oxide layer
V_{oxcg}	voltage drop across the interpoly oxide
V_{oxfox}	voltage drop across the STI oxide
V_{oxpw}	voltage drop across the tunnel oxide (p-well region)
V_{oxsd}	voltage drop across the tunnel oxide (S/D region)
V_{prg}	parameter for calculating time-dependent V_{CG} during transient simulations
V_s	surface potential
V_{scg}	surface potential at the CG side of the CG/FG capacitor
V_{SD}	voltage of the S/D junction (versus substrate) in the case of connected S/D
V_{SDW}	voltage applied to source, drain and isolated p-well (versus substrate)
$V_{SDW,e}$	voltage applied to source, drain and isolated p-well for cell erasing
$V_{SDW,w}$	voltage applied to source, drain and isolated p-well for cell writing
V_{sfgcg}	surface potential at the FG side of the CG/FG capacitor
V_{sfgfox}	surface potential at the FG side of the FG/substrate capacitor (STI region)
V_{sfgpw}	surface potential at the FG side of the FG/substrate capacitor (p-well region)
V_{sfgsd}	surface potential at the FG side of the FG/substrate capacitor (S/D region)
V_{sfox}	surface potential at the substrate side of the FG/substrate capacitor (STI region)
V_{SG}	select gate voltage (versus substrate)
V_{spw}	surface potential at the substr. side of the FG/substr. capacitor (p-well region)
V_{ssd}	surface potential at the substrate side of the FG/substrate capacitor (S/D region)
V_T	threshold voltage
V_{T0}	V_T of a flash memory cell with no net charge on the floating gate
$V_{T,e}$	threshold voltage of an erased flash cell
V_{Ti}	intrinsic threshold voltage of a flash cell (measured at cells with contacted FG)
V_{Tstart}	initial V_T for transient flash cell simulation
$V_{T,w}$	threshold voltage of a written memory cell
W_{FG}	gate width of the floating gate
W_G	gate width
W_{ref}	gate width of the reference transistor in a 2-transistor memory cell
W_{SG}	gate width of the select transistor in a 2-transistor memory cell
x_{ox}	parameter defining the location of oxide charge
x_{oxg}	parameter defining the location of oxide charge in the tunnel oxide
x_{oxcg}	parameter defining the location of oxide charge in the interpoly oxide

Physical Constants and Material Parameters

$k = 1.38066 \times 10\text{-}^{23}$	$J\ K^{-1}$	Boltzmann constant
$q_0 = 1.60218 \times 10^{-19}$	C	Elementary charge
$\varepsilon_0 = 8.85418 \times 10^{-12}$	$F\ m^{-1}$	Permittivity in vacuum
$h = 6.62617 \times 10^{-34}$	J s	Planck constant
$N_L = 2.80 \times 10^{+25}$	m^{-3}	Effective density of states conduction band (at 300K)
$N_V = 1.04 \times 10^{+25}$	m^{-3}	Effective density of states in valence band (at 300K)
$E_{gsi} = 1.12$	eV	Silicon bandgap at room temperature (at 300K)

Legend: Cross Section Views



Si, undoped

Poly-Si, undoped

Si or poly-Si, with n-type doping

Si or poly-Si, with p-type doping

Dielectric (silicon oxide)

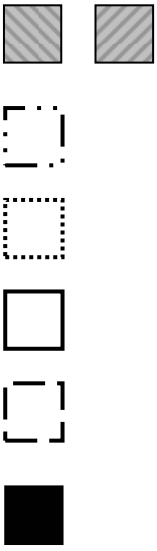
Dielectric (silicon nitride or oxide/nitride layer stack)

Metal (W - contact plug)

Metal (Al – interconnect layer)

Resist

Legend: Layout Views



Active areas, with underlying p-well or n-well, respectively
(remark: inverse = STI)

FImp

FG-etch

CG-etch

CMOS gate etch

Contact

Index of Figures

Index of Tables

References

[1] U. König, “SiGe and GaAs as competitive technologies for RF-applications”, Proceedings of the BCTM, p. 87-92, 1998

[2] D. Harame, L. Larson, M. Case, S. Kovacic, S. Voinigescu, T. Tewksbury, D. Nguyen-Ngoc, K. Stein, J. Cressler, S.-J. Jeng, J. Malinowski, R. Groves, E. Eld, D. Sunderland, D. Rensch, M. Gilbert, K. Schonenberg, D. Ahlgren, S. Rosenbaum, J. Glenn, and B. Meyerson, “SiGe HBT technology: Device and application issues”, IEDM Tech. Dig., pp. 731-734, 1995

[3] R. A. Johnson, M. J. Zierak, K. B. Outama, T. C. Bahn, A. J. Joseph, C. N. Cordero, J. Malinowski, K. A. Bardt, T. W. Weeks, R. A. Milliken, T. J. Medve, G. A. May, W. Chongt, K. M. Waltert, S. L. Tempestt, B. B. Chaut, M. Boenket, M. W. Nelson, D. L. Harame, “1.8 Million Transistor CMOS ASIC Fabricated in a SiGe BiCMOS Technology”, IEDM Tech. Dig., pp. 217 – 220, 1998

[4] D. Chin, “Executing system on a chip: Requirements for a successful SOC implementation “, IEDM Tech. Dig., pp. 3 – 8, 1998

[5] P.Cappelletti, C. Golla, P. Olivo, E. Zanoni, “Flash Memories”, Kluwer Academic Publishers, 1999

[6] C. Contiero, P. Galbiati, M. Palmieri, L. Vecchi, “Characeristics and Applications of a 0.6µm Bipolar-CMOS-DMOS Technology combining VLSI Non-Volatile Memories”, IEDM Technical Digest., pp. 465 – 468, 1996

[7] C. Contiero, P. Galbiati, A. Merlini, A. Moscatelli, F. Tampellini, L. Vecchi, “Trends and Issues in BCD Smart Power Technologies”, Proc. ESSDERC, pp. 111 – 118, 1999

[8] http://www.st.com/stonline/prodpres/dedicate/soc/process/feature.htm

[9] D. Knoll, K. E. Ehwald, B. Heinemann, A. Fox, K. Blum, H. Rücker, F. Fürnhammer, B. Senapati, R. Barth, U. Haak, W. Höppner, J. Drews, R. Kurps, S. Marschmeyer, H. H. Richter, T. Grabolla, B. Kuck, O. Fursenko, P. Schley, R. Scholz, B. Tillack, Y. Yamamoto, K. Köpke, H. E. Wulf, D. Wolansky, and W. Winkler, “A Flexible, Low Cost, High Performance SiGe:C BICMOS Process with a One-Mask HBT Module”, IEDM Tech. Dig., pp. 783-786, 2002

[10] A. Fox, K. E. Ehwald, P. Schley, R. Barth, S. Marschmeyer, C. Wolf, V. E. Stikanov, A. Gromovyy, A. Hudyryev, „Cost-effective Integration of an FN-programmed Embedded Flash Memory into a 0.25µm RF-BiCMOS Technology”, Proc. ICM, pp. 463 - 466, 2004

[11] D. Knoll, A. Fox, K.E. Ehwald, B. Heinemann, R. Barth, A. Fischer, H. Rücker, P. Schley, R. Scholz, F. Korndörfer, B. Senapati, V.E. Stikanov, B. Tillack, W. Winkler, Ch. Wolf, P. Zaumseil, “A Low-Cost SiGe:C BiCMOS Technology with Embedded Flash Memory and Complementary LDMOS Module”, accepted for presentation at the BCTM, 2005

[12] W. D. Brown, J. E. Brewer, “Nonvolatile Semiconductor Memory Technology: A comprehensive Guide to Understanding and Using NVSM Devices”, IEEE Press, 1998

[13] C. Hu, “Nonvolatile Semiconductor Memories: Technologies, Design, and Applications”, IEEE Press, 1991

[14] R. Zambrano, G. Casagrande, R. Bez, “Non-volatile Memory Technology”, in IEDM Short Course Notes, 2001

[15] C. Kou, “Embedded Flash memory: Applications, technology, and design,” in IEDM Short Course Notes, 1995.

[16] P. Pavan, E. Zanoni, “Flash Memory Cells – An overview”, Proceedings of the IEEE, vol. 85, No. 8, pp. 1248 – 1271, August 1997
[17] S. Lai, “Flash memories: where we are and where we are going”, IEDM Tech. Dig., p.971 - 974, 1998
[18] J. G. Ganssle, “Flash Memory: Past, Present, and Future”, Embedded Systems Programming, pp. 59-67, July 1999
[19] R. Bez, E. Camerlenghi, A. Modelli, A. Visconti, “Introduction to Flash Memory”, Proc. Of the IEEE, Vol. 91, No. 4, pp. 489 – 502, April 2003
[20] K. Yoshikawa, “Embedded Flash memories – Technology assessment and future”, Int. Symp. on VLSI Technology, Systems and Applications, pp. 183-186, 1999
[21] L. Baldi, A. Maurelli, “ Embedded non-volatile memories in deep submicron CMOS”, Proc. ESSDERC, pp. 127 – 134, 1999
[22] D. Frohmann-Bentchkowski, “A Fully Decoded 2048-Bit Electrically Progammable FAMOS Read-Only memory”, J. Solid-State Circuits, vol. SC-6, no.5, pp. 301-306, 1971
[23] E. Harari, L. Schmitz, B. Troutman, S. Wang, “A 256 bit nonvolatile static RAM”, IEEE ISSCC Tech. Dig., p.108, 1978
[24] D. C. Gutermannn, I. H. Rimawi, T. L. Chiu, R. D. Halvorson, and D. J. McElroy, „An electrically alterable nonvolatile memory cell using a floating gate structure”, IEEE Trans. Electron Devices, vol. ED-26, no. 4, pp. 576-586, 1979
[25] G. Verma, N. Mielke, “Reliability performance of ETOX based Flash memories”, Proc. IRPS, p. 158, 1988
[26] S. Mahapatra, S. Shukuri, J. Bude, “CHISEL Flash EEPROM – Part I: Performance and Scaling”, IEEE Trans. Electron Devices, vol. 49, No. 7, pp. 1296 – 1301, July 2002
[27] T. Hara, K. Fukuda, K. Kanazawa, N. Shibata, K. Hosono, H. Maejima, M. Nakagawa, T. Abe, M. Kojima, M. Fujiu, Y. Takeuchi, K. Amemiya, M. Morooka, T. Kamei, H. Nasu, K. Kawano, C.-M. Wang, K. Sakurai, N. Tokiwa, H. Waki, T. Maruyama, S. Yoshikawa, M. Higashitani, T. D. Pham, T. Watanabe “A 146mm2 8Gb NAND Flash Memory with 70nm CMOS Technology”, Proceedings of the ISSCC, pp. 44 – 45, 2005
[28] H. Kume, M. Kato, T. Adachi, T. Tanaka, T. Sasaki, T. Okazaki, N. Miyamoto, S. Saeki, Y. Ohji, M. Ushiyama, J. Yugami, T. Morimoto, Takashi Nishida, „A 1.28µm2 contactless memory cell technology for a 3V only 64 Mb EEPROM“, IEDM Tech. Dig., pp. 991-993, 1992
[29] T. Kobayashi, Y. Sasago, H. Kurata, S. Saeki, Y. Goto, T. Arigane, Y. Okuyama, H. Kume, K. Kimura, “A giga-scale assist-gate (AG)-AND-type flash memory cell with 20-MB/s programming throughput for content-downloading applications”, IEDM Tech. Dig., pp. 29 - 32, 2001
[30] H. Onoda, Y. Kunori, S. Kobayashi, M. Ohi, A. Fukumoto, N. Ajika, H. Miyoshi, “A novel cell structure suitable for a 3V operation, sector erase Flash memory”, IEDM Tech. Dig., pp. 599-602, 1992
[31] K. Yoshikawa, S. Mori, E. Sakagami, N. Arai, Y. Kaneko, Y. Oshima, “Flash EEPROM cell scaling based on tunnel oxide thinning limitations”, Symp. On VLSI Technology, pp. 79 – 80, 1991
[32] P. Cappelletti, R. Bez, D. Cantarelli, L. Fratin, “Failure Mechanisms of Flash Cell in Program/Erase Cycling”, IEDM Tech. Dig., pp. 291 – 294, 1994
[33] P. Cappelletti, R. Bez, A. Modelli, A. Visconti, “What we have learned on Flash Memory Reliability in the Last Ten Years”, IEDM Tech. Dig., pp. 489 - 492, 2004
[34] A. Spinelli, “Flash: NOR Reliability”, Tutorial Notes at the International Reliability Physics Symposium (IRPS), 2005

[35] R. Shirota, “Flash: NAND Reliability”, Tutorial Notes at the International Reliability Physics Symposium (IRPS), 2005

[36] B. De Salvo, G. Ghibaudo, G. Pananakakis, G. Reimbold, F. Mondond, B. Guillaumot, P. Candelier, “Experimental and Theoretical Investigation of Nonvolatile Memory Data-Retention”, IEEE Trans. on Electron Devices, vol. 46, no. 7, pp. 1518 – 1524, July 1999

[37] K. Naruke, S. Taguchi, M. Wada, Stress induced leakage current limiting to scale down EEPROM tunnel oxide thickness“, IEDM Tech. Dig., pp. 424 – 427, 1988

[38] B. Ricco, G. Gozzi, M. Lanzoni, “Modeling and Simulation of Stress-Induced Leackage Current in Ultrathin SiO_2 Films”, IEEE Trans. On Electron Devices, vol. 45, no. 7, pp. 1554 – 1560, July 1998

[39] D. Ielmini, A. Spinelli, M. A. Rigamonti, A. L. Lacaita, “Modeling of SILC Based on Electron and Hole Tunneling – Part I: Transient Effects” and , “Modeling of SILC Based on Electron and Hole Tunneling – Part II: Steady State”, IEEE Trans. On Electron Devices, vol. 47, no. 6, pp. 1258 – 1272, June 2000

[40] J. De Blauwe, J. Van Houdt, D. Wellekens, G. Groeseneken, H. E. Maes, “SILC-Related Effects in Flash E^2PROM’s – Part I: A Quantitative Model for Steady-State SILC“ and “SILC-Related Effects in Flash E^2PROM’s – Part II: Prediction of Steady-State SILC-Related Disturb Characteristics“, IEEE Trans. On Electron Devices, vol. 45, no. 8, pp. 1745 – 1760, August 1998

[41] S. Mori, N. Arai, Y. Kaneko, K. Yoshikawa, “Polyoxide Thinning Limitation and Superior ONO Interpoly Dielectric for Nonvolatile Memory Devices“, IEEE Trans. On Electron Devices, vol. 38, No. 2, pp. 270 – 276, February 1991

[42] S. Mori, et.al., “Thickness scaling limitation factors of ONO interpoly dielectric for nonvolatile memory devices”. IEEE Transactions on Electron Devices, vol 43, no. 1, pp 47 –53, 1996

[43] P. Candelier, B. De Salvo, F. Martin, B. Guillaumont, F. Mondon, G. Reimbold, “Thinning Oxide-Nitride-Oxide Interpoly Dielectric (11-13nm) for 0.25µm Flash Cell Memories”, Proc. ESSDERC, pp. 264 – 267, 1997

[44] J.-H. Kim, J.-B. Choi, “Long-Term Electron Leakage Mechanisms Through ONO Interpoly Dielectric in Stacked-Gate EEPROM Cells”, IEEE Trans. Electron Devices, vol. 51, No. 12, pp. 2048 – 2053, 2004

[45] Y.-H. Shih, H.-T. Lue, K.-Y. Hsieh, R. Liu, C.-Y. Lu, “A novel 2-bit/cell nitride storage flash memory with greater than 1M P/E-cycle endurance”, IEDM Tech. Dig., pp. 881-884, 2004

[46] Y. N. Tan, W. K. Chim, W. K. Choi, M. S. Joo, T. H. Ng, B. J. Cho, ”High-K HfAlO charge trapping layer in SONOS-type nonvolatile memory device for high speed operation”, IEDM Tech. Dig., pp. 889-892, 2004

[47] Y.-H. Lin, C.-H. Chien, C.-T. Lin, C.-W. Chen, C.-Y. Chang, T.-F. Lei, “High performance multi-bit nonvolatile HfO2 nanocrystal memory using spinodal phase separation of hafnium silicate”, IEDM Tech. Dig., pp. 1080-1082, 2004

[48] G. Müller, T. Happ, M. Kund, G. Y. Lee, N. Nagel, R. Sezi, ”Status and outlook of emerging nonvolatile memory technologies”, IEDM Tech. Dig, pp. 567-570, 2004

[49] J. Van Houdt, D. Wellekens, Luc Haspeslagh, “The HIMOS Flash Technology: The Alternative Solution for Low-Cost Embedded Memory”, Proceedings of the IEEE, Vol.91, No 4, pp. 627 - 635, April 2003

[50] J. W. Zahlmann-Nowitzki, “Ein Beitrag zur Zuverlässigkeit von EEPROM Zellen für den erweiterten Temperaturbereich”, Dissertation, Christian Albrechts Universität zu Kiel, 2001

[51] L. Nebrich, “Entwicklung eines Makromodells für die Schaltungs- und Zuverlässigkeitssimulation von EEPROM-Zellen im erhöhten Temperaturbereich”, Dissertation, Christian Albrechts Universität zu Kiel, 2001

[52] L. Baldi, G. Dallalibera, M. Patelmo, B. Vajana, P. Zuliani, " High Density EEPROM Cell for 0.35µm Embedded Applications", Proc. ESSDERC, pp. 624 – 627, 1999

[53] V. N. Kynett, A. Baker, M. Fandrich, G. Hoekstra, O. Jungroth, J. Kreifels, S. Wells, "An in-system reprogrammable 256K CMOS Flash memory", Proc. ISSCC, p.123, 1988

[54] S. N. Keeney, "A 130nm generation high density ETOX™ flash memory technology", IEDM Tech. Dig., pp. 41 - 44, 2001

[55] Intel press release "Intel Introduces World's First NOR Flash Memory On 90-Nanometer Manufacturing Technology", February 19th, 2004: http://www.intel.com/pressroom/archive/releases/20040219comp.htm; See also: http://www.intel.com/design/flcomp/prodbref/M18.htm

[56] H. Watanabe, S. Yamada, M. Tanimoto, M. Matsui, S. Kitamura, K.Amemiya, T. Tanzawa, E. Sakagami, M. Kurata, K. Isobe, M. Takebuchi, M. Kanda, S. Mori, T. Watanabe, "Novel 0.44µm^2 Ti-Salicided STI Cell Technology for High-Density NOR Flash Memories and High Performance Embedded Application", IEDM Tech. Dig., pp. 975 – 978, 1998

[57] F. Piazza, P. Colombo, P. Ghezzi, V. Lista, A. Maurelli, E. Palumbo, D. Peschiaroli, S. Soleri, G. Di Biase, A. Silvagni, C. Torti, M. Olivo, L. Baldi, "1.8µm^2 High Density Flash Memory for 0.35µm Embedded Applications", Proc. ESSDERC, pp. 616 – 619, 1999

[58] J. Van Houdt, L. Haspeslagh, D. Wellekens, L. Deferm, G. Groeseneken, H.E. Maes, "HIMOS-A High Efficiency Flash E^{2}PROM Cell for Embedded Memory Applications", IEEE Trans. Electron Devices, vol. 40, pp. 2255 – 2263, 1993

[59] D. Wellekens, J. Van Houdt, L. Haspeslagh, J. Tsouhlarakis, P. Hendrickx, L. Deferm, H.E. Maes, "Embedded HIMOS flash memory in 0.35-µm and 0.25-µm CMOS technologies", Trans. Electron Devices, vol. 47, p. 2153, November 2000

[60] J. De Vos, L. Haspeslagh, M. Demand, A. Redolfi, C. Baerts, S. Beckx, F. Vleugels, J. Van Houdt, "Integration of HIMOS™ Flash Memory in a 90nm CMOS Technology", Electrochemical Society Proceedings, vol. 2005-06, pp. 306 – 314, 2005

[61] J. Tsouhlarakis, G. Vanhorebeek, G. Verhoeven, J. De Blauwe, S. Kim, D. Wellekens, P. Hendrickx, L. Haspeslagh, J. Van Houdt, H. Maes, "A Flash Memory Technology with Quasi-Virtual Ground Array for Low-cost Embedded Applications", IEEE Journal of Solid-State Circuits, vol. 36, no. 6, June 2001

[62] S. Kianian, A. Levi, D. Lee, Y.-W. Hu, "A Novel 3 Volts-Only, Small Sector Erase, High Density Flash E2PROM", Symp. on VLSI Technology, pp. 71 – 72, 1994

[63] A. Kotov, A. Levi, Y. Tkachev, V. Markov, "Tunneling Phenomenon in SuperFlash Cell", at Nonvolatile Memory Technology Symposium (NVMTS), 2002

[64] H. Guan, D. Lee, G. P. Li, "An Analytical Model for Optimization of Programming Efficiency and Uniformity of Split Gate Source-Side Injection Superflash Memory", IEEE Trans. On Electron Devices, vol 50, no. 3, pp. 809 – 815, March 2003

[65] K. Ohsaki, N. Asamoto, S. Takagaki, "A Single Poly EEPROM Cell Structure for Use in Standard CMOS Processes", IEEE J. Solid State Circuits, vol. 29, no. 3, pp. 311 – 316, March 1994

[66] R. J. McPartland, R. Singh, "1.25 Volt, Low Cost, Embedded Flash Memory for Low Density Applications", Tech. Dig. Of Symposium on VLSI Circuits, pp. 158 – 161, 2000

[67] L. Baldi, A. Cascella, B. Vajana, "A scalable Single Poly EEPROM Cell for Embedded Memory Applications", ESSDERC 1995

[68] R. Heinrich, W. Heinrigs, G. Tempel, J. Winnerl, T. Zettler, "A 0.5µm CMOS Technology for Multifunctional Applications with Embedded FN-Flash Memory and Linear R and C Modules

[69] J. K. Yeh, H. D. Su, Y. F. Lin, C. D. Shieh, D. S. Kuo, M. S. Liang, G. Q. Tao, F. J. List, L. Shi, R. Colclaser, N. Tandan, K. Chen, M. Chen, A. van Gorkum, "A 0.5µm Flash Technology suitable for Low Voltage Embedded Applications", Proc. ESSDERC, pp. 260 – 263, 1997

[70] K. K. Takahashi, H. Doi, N. Tamura, K. Mimuro, T. Hashizume, Y. Moriyama, Y. Okuda, "A 0.9V Operation 2-Transistor Flash Memory for Embedded Logic LSIs", Symp. VLSI Technology, pp. 21 – 22, 1999

[71] C. J. Huang, Y. C. Liu, P. C. Lin, A. Wu, H. H. Chen, W. C. Ting, G. Hong, „ A 0.25µm Embedded Flash Technology with Shallow Trench Isolation Using Channel FN Operation", at the Nonvolatile Memory Technology Symposium (NVMTS), 2000

[72] C.-N. B. Li, D. Farenc, R. Singh, J. Yater, S. Liu, C.-L. Chang, S. Bagchi, K. Chen, P. Ingersoll, K.-T. Chang, "A Novel Uniform-Channel-Program-Erase (UCPE) Flash EEPROM Using An Isolated P-well Structure", IEDM Tech. Dig, pp. 779 – 782, 2000

[73] A. R. Khan, N. Tandan, "Description and Reliability of a Robust 0.18 micron Low-Voltage and Low-Power Embedded Flash Technology", at the Nonvolatile Memory Technology Symposium (NVMTS), 2002

[74] G. Tao et al., "Reliability aspects of advanced embedded floating-gate non-volatile memories with uniform channel FN tunneling for both program and erase", Proc. *NVSMW*, pp 130-131, 2001

[75] A. Scarpa, G. Tao, J. Dijkstra, F.G. Kuper, "Reliability implications in advanced embedded two-transistor-Fowler-Nordheim-NOR flash memory devices", Solid-State Electronics, vol. 46, pp. 1765-1773, 2002

[76] J, Caywood, et al., "IEEE Standard Definitions and Characterization of Floating Gate Semiconductor Arrays", Standards Committee of the IEEE Electron Devices society, IEEE 1005-1998, June 1998

[77] R. E. Shiner, J. M. Caywood, B. L. Euzent, "Data Retention in EPROMS", Proc. 18th annual Reliability Physics Symposium, pp. 238 – 243, 1980

[78] P. K. Roy, R. H. Doklan, E. P. Martin, S. F. Shive, A. K. Sinha, "Synthesis and Characterization of high Quality Ultrathin Gate Oxides for VLSI/ULSI Circuits", IEDM Tech. Dig., pp. 714 – 717, 1988

[79] H.-H. Tseng, P. J. Tobin, J. D. Hayden, K.-M. Chang, "Advantages of CVD Stacked Gate Oxide for Robust 0.5µm Transistors", IEDM Tech. Dig., pp. 75 – 78, 1991

[80] R. Moazzami, C. Hu, "A High-Quality Stacked Thermal/LPCVD Gate Oxide Technology for ULSI", IEEE Electron Device Letters, vol. 14, no. 2, pp. 72 – 73, February 1993

[81] H.-H. Tseng, P. J. Tobin, J. D. Hayden, K.-M. Chang, J. W. Miller, "A Comparison of CVD Stacked Gate Oxide and Thermal Gate Oxide for 0.5µm Transistors Subjected to Process-Induced Damage", IEEE Trans. On Electron Devices, vol. 40, no. 3, pp. 613 – 618, March 1993

[82] C. Lam, T. Sunaga, Y. Igarashi, M. Ichinose, K. Kitamura, C. Willets, J. Johnson, S. Mittl, F. White,H. Tang, T.-C. Chen, "Anomalous Low Temperature Charge Leackage Mechanism in ULSI Flash Memories", IEDM Tech. Dig., pp. 335 – 338, 2000

[83] S.M. Sze, "Physics of Semiconductor Devices", 2nd Edition, Wiley-Interscience, 1981

[84] M. Lenzlinger, E.H. Snow, "Fowler-Nordheim tunneling into thermally grown SiO_2", Journal of Applied Physics, vol. 40, pp. 278 – 283, 1969

[85] P. Palestri, A.D. Serra, L. Selmi, M. Pavesi, P.L. Rigolli, A. Abramo, F. Widdershoven, E. Sangiorgi, "A Comparative Analysis of Substrate Current Generation Mechanisms in Tunneling MOS Capacitors", IEEE Trans. On Electron Dev. Vol. 49, no.8, pp. 1427 - 1435, 2002

[86] B. De Salvo, G. Ghibaudo, G. Pananakakis, G. Reimbold, F. Mondond, B. Guillaumot, P. andelier, "Experimental and Theoretical Investigation of Nonvolatile Memory Data-Retention", IEEE Trans. On Electron Dev. Vol. 46, no.7, pp. 1581 – 1524, 1999

[87] JEDEC standard, EIA/JESD60, „A procedure for measuring p-channel MOSFET hot carrier induced degradation at maximum gate current DC stress“, April 1997
[88] JEDEC standard, JESD28-A, „Procedure for measuring n-channel MOSFET hot carrier induced degradation under DC stress“, December 2001
[89] T.Kawahara, N. Kovayashi, Y. Jyouno, S. Saeki, N. Miyamoto, T. Adachi, M. Kato, A. Sato, J. Yugami , H. Kukme , K. Kimura , Bit-Line clamped Sensing Multiplex and Accurate High Voltage Generator for Quarter-Micron Flash Memories, IEEE J. Solid-State Circuits, no. 31 (11), pp. 1590-1600, 1996
[90] N. Otsuka and M. A. Horowitz, Circuit Techniques for 1.5 – V Power Supply Flash Memory, IEEE J. Solid-State Circuits, no. 32 (8), pp. 1217-1230, 1997
[91] R. Sorge, "Implant dose monitoring by MOS C-V measurement", Microelectronics Reliability, no. 43, pp.167-171, 2003

Publications

[p1]* A.Fox, K.E. Ehwald, P. Schley, R. Barth, S. Marschmeyer, V.E. Stikanov, A. Gromovyy, A Hudyryev, “Cost-Effective Integration of an FN-programmed Embedded Flash Memory into a 0.25 µm RF-BiCMOS Technology”, Proc. International Conference on Microelectronics, p. 463, 2004

[p2] D. Knoll, A. Fox, K.-E. Ehwald, B. Heinemann, R. Barth, A. Fischer, H. Rücker, P. Schley, R. Scholz, F. Korndörfer, B. Senapati, V.E. Stikanov, B. Tillack, W. Winkler, Ch. Wolf, P. Zaumseil, ”A Low-Cost SiGe:C BiCMOS Technology with Embedded Flash Memory and Complementary LDMOS Module”, Proc. of BCTM 2005, Santa Barbara, October 10-11, 2005 USA, p. 132, 2005

[p3] D. Knoll, K. E. Ehwald, B. Heinemann, A. Fox, K. Blum, H. Rücker, F. Fürnhammer, B. Senapati, R. Barth, U. Haak, W. Höppner, J. Drews, R. Kurps, S. Marschmeyer, H. H. Richter, T. Grabolla, B. Kuck, O. Fursenko, P. Schley, R. Scholz, B. Tillack, Y. Yamamoto, K. Köpke, H. E. Wulf, D. Wolansky, and W. Winkler, “A Flexible, Low Cost, High Performance SiGe:C BICMOS Process with a One-Mask HBT Module”, IEDM Tech. Dig., pp. 783-786, 2002

* Extended version of [p1] is in the selection of best papers of the ICM 2004 and accepted for publication in a special issue of the Microelectronics Journal

Index