

NUMERICAL SIMULATION OF EMBEDDED NON-VOLATILE MEMORY

Ann Concannon (aconcan@nmrc.ucc.ie), Russell Duane, Ray Duffy, Mike O'Shea, and Diarmuid McCarthy and Alan Mathewson

Technology Characterisation and Modelling Group, National Microelectronics Research Centre, Ireland.

Abstract

Non volatile memories have in recent years become an established component of all electronic systems, especially portable systems. The general trend towards a higher level of integration and speed are pushing towards their integration with complex logic – *embedded* non-volatile memory. Non Volatile memory research in the NMRC, which was stimulated by a Marie Curie Fellowship, has focused on the model development which has facilitated the use simulation in the exploration and analysis of advanced embedded flash memory solutions.

1. INTRODUCTION

Embedded memory is the first step on the path to “system on a chip”, the ultimate destination of semiconductor integration. The enhanced flexibility that is offered by non-volatile memory, particularly for on-chip data and code storage, and for user-programmable micro-controllers, is now demanded by IC and system designers. Logic and memory do not naturally co-exist on the same silicon, requiring developers to make trade-offs between cost and performance. In the past, non-volatile memory technology has lagged a couple of generations behind the leading edge CMOS technology development; but in today's market, foundries now expect to offer NVM solutions very shortly after CMOS in order to gain a competitive edge.

The role of technology computer aided design (TCAD) in device engineering and technology development has matured significantly in the past few years. Numerical process and device simulation of CMOS silicon technology is the main driver of this development, as this is the mainstream choice for the semiconductor industry. The work described in this abstract further enhances Europe's reputation in the TCAD field by developing TCAD methods for the simulation of NVM devices. The application of the enhanced tools to investigate NVM device engineering, ultimate scaling limitations and novel

devices is also described. This has enabled better understanding of the device physics, device operation and is continuing to make a significant contribution to the development cycle of new embedded memory processes.

2. NON VOLATILE MEMORY

Non volatile memory describes a system where the contents of the memory are retained after the power is switched off. An example of a non-volatile memory is a floppy disk or a compact disk. In silicon integrated circuit technology non-volatile memory is realised through the use of an electrically isolated “floating gate”. A cross-section schematic shown in Figure 1 illustrates a conventional stacked gate flash memory. The memory state depends on the charge on the floating gate. The insulator around the floating gate must be thick enough to prevent the floating gate discharging when the power is removed; but it also must be thin enough to allow the transfer of charge on and off the floating gate; under appropriate bias configuration. In addition, a change in the charge on the floating gate should have a significant change in the electrical properties of the device so that the circuitry can sense the memory state.

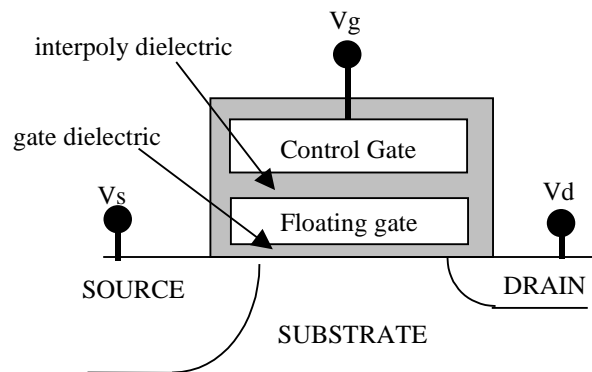


Figure 1. Cross-section schematic of a NVM transistor.

MOSFET: A MOS field effect transistor is the basic building block of all digital integrated circuits. A MOS transistor is a miniature electronic switch that is on and conducts electricity when the gate voltage is high and is off when the gate voltage is low. The binary operation of a transistor gives microprocessors the ability to perform many complex tasks, utilising millions of transistors in one circuit and executing millions of instructions per second.

NVM: An ideal memory would be fast, cheap, reliable, dense, reprogrammable and non volatile. Non-volatility refers to the ability of the memory to retain its contents when power is removed. In solid state non-volatile memory, charge is stored on a floating polysilicon gate, which is electrically isolated from the gate and substrate by an insulator. The charge on the floating gate transforms the MOS transistor into a *programmable switch*.

The basic operations of the memory can be described as (a) programming b) erasing and c) reading. Programming usually refers to putting charge on the floating gate; erasing usually refers to removing the charge from the gate and reading involves sensing the electrical signal of the cell to determine if the device is programmed or erased.

Numerical process and device simulation are extensively used to determine the best design of memory for a particular application, how to integrate the memory in the IC process and how best to bias the memory to achieve the programming/erasing and reading electrical targets. As with all simulation tools, the models must first be developed to account for all the relevant phenomena before the tool can be used to accurately represent the system. The key requirements for NVM device simulation (in addition to the MOSFET requirements) are the ability to simulate the transient programming and erasing operations, where charge is transferred to a floating gate, causing a change in the electrical characteristics of the memory. In most flash memories, this requires models for a) charge storage on a floating gate; b) Fowler-Nordheim tunnelling; c) Band-to-band tunnelling; and d) hot carrier injection. Since 1989 researchers at the NMRC have been working on these topics, in close co-operation with ST microelectronics, who are Europe's largest NVM manufacturers. In 1994 a Human Capital and Mobility award enabled the author to pursue this work with Professors Baccarani and Rudan at the University of Bologna in Italy towards the development of numerical models to describe some of the complex phenomena that occur in state-of-the-art NVM devices [1]. These models were then implemented in the robust hydrodynamic device simulator of the Italian group. Subsequently these models and models developed previously in the NMRC [2,3,4] have been taken up by the commercial TCAD vendors and implemented in their software in response to a world-wide customer demand for this facility.

3. THE MODEL DEVELOPMENT

The models that were developed for the NVM device simulation are also very useful in advanced CMOS analysis. "Hot carriers" describe electrons or holes that have significantly more energy than the average energy. It usually refers to a small percentage of the carrier population, but due to the high energy, these carriers can have a large effect on the devices. The carriers are given energy by applying high electric fields between the source and drain which accelerates the carriers and increases the chances of a carrier reaching the drain without sustaining energy loss collisions on route. Hot carrier effects play an increasingly important role in many semiconductor devices, because they provide either a useful characteristic of the device or an unwanted parasitic effect. The programming mechanism of many flash EEPROMs is based on hot electron injection due to generation of hot carriers near the drain.

This device is therefore designed to enhance hot electron injection to a floating gate by having a very abrupt drain junction. However, in the case of MOSFETs, hot carrier generation plays a role in device degradation and therefore MOS devices are designed to avoid this unwanted phenomenon. Hot electron and hot hole trapping in the gate oxide can cause a shift in the threshold voltage over time and can cause parametric shifts resulting in failures in circuits and substrate currents can cause bipolar in the MOS transistor system.

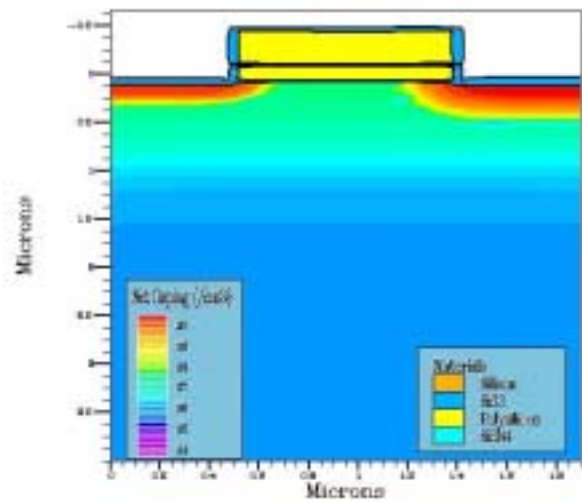


Figure 2: Two dimensional doping profile of 0.9um gate length n-channel flash EEPROM computed using process simulation.

Numerical process simulation is used to build up a numerical model of the device structure in two dimensions, as shown in Figure 2, providing information on geometry and doping profiles that reflect the processing the real device will receive. This is used as input to the device simulator to provide valuable insight into the device action, enabling analysis of hot carrier phenomena in MOS and EEPROM devices. However, accurate simulation of hot carrier effects in flash EEPROMs and MOSFETs are complicated, because the calculation of the hot carrier energy distribution requires detailed knowledge of electron transport in a complex silicon band structure. Ideally this should be simulated using a Monte Carlo approach, incorporating a full band description of the silicon (e.g.: DAMOCLES of IBM). However, this approach is computationally too expensive to be considered as a realistic TCAD tool. A more general approach has been to develop models for hot electron and hot hole gate currents based on analytical expressions which determine the probability of an energetic electron reaching the gate. In these models, numerical device simulators are used to calculate the electric fields and carrier distributions in the silicon.

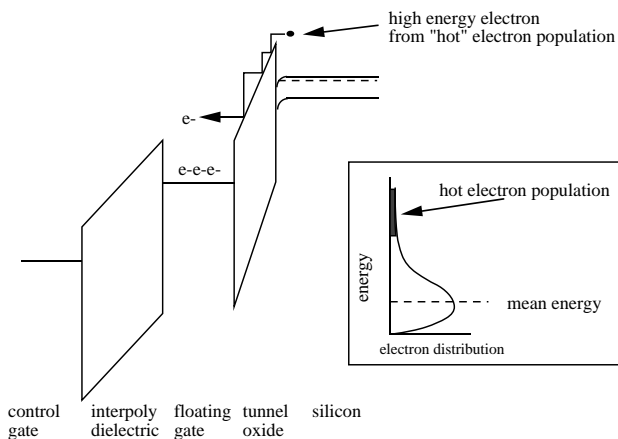


Figure 3: Band Energy diagram of a double poly flash EEPROM system showing the programming due to hot electron injection at the drain edge.

Hydrodynamic numerical device simulators solve for the potential, carrier concentration and carrier temperature at each point in the simulation grid [5]. However, the value of the carrier temperature calculated is an average value and because hot carrier processes involve carriers with energies above the average value, an analytical description of the high tail of the energy distribution function is required. The model developed in this work relies on the consistent description of the energy distribution function for both electrons and holes in the substrate current model and the gate current model. The model parameters were calibrated for a particular technology by extensive comparison with experimental data [3]. This model relies on the consistent description of the energy distribution function for both electrons and holes in the substrate current model and the gate current model. The model that was developed and implemented in the device simulator in this work allows, for the first time, analysis of the gate current of the MOS system by computing the electron and hole components individually. The match that is achieved between measured and simulated currents demonstrate the accuracy of the model over a large voltage range as illustrated in Figures 4 and 5. This gives the user confidence that the model can be used in engineering applications.

4. DEVICE ENGINEERING

The main author completed the Marie-Curie Fellowship in the University of Bologna in May 1995 and as awarded a PhD from the National University of Ireland in 1996 [3]. Building on the strong foundation of NVM TCAD model development, she now has assembled a significant group working on different related applications in the NMRC. The following sections outline the contribution to the current state-of-the-art by 4 PhD and 1 M.Eng.Sc. students who are working in flash memory technology development, model development and exploratory device concepts.

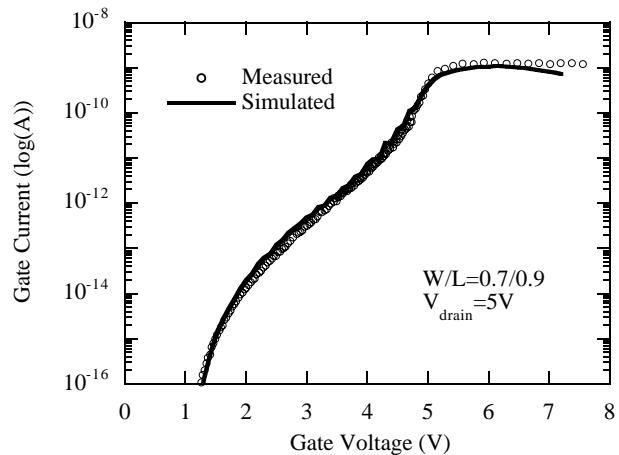


Figure 4. Floating gate current vs floating gate current at drain voltage of 5V on 0.9um gate length flash EEPROM.

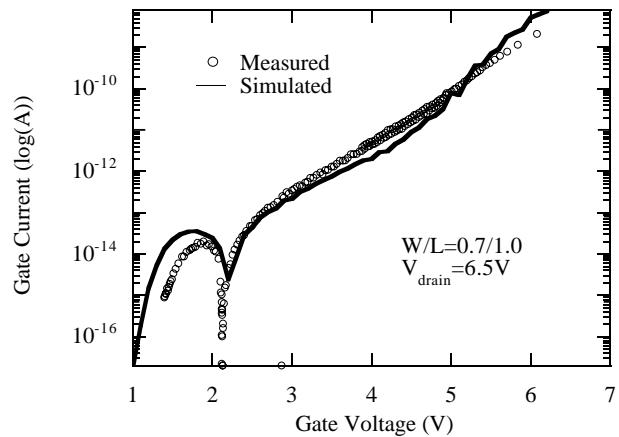


Figure 5. Floating gate current vs floating gate current at drain voltage of 6.5V, measured on a 1.0um gate length flash EEPROM cell.

As mentioned previously, embedding flash memory in logic CMOS processes involve trade-off between cost and performance at the development stage. The NMRC, having already developed the appropriate models to enable NVM simulation, are now in a position to apply the tools to engineering applications. This section describes a PhD work, the aim of which was to simulate and evaluate a low cost embedded flash EEPROM which relies on Fowler-Nordheim tunnelling for both programming and erasing [6,7]. The baseline technology is a state-of-the-art 0.18um CMOS technology. This PhD project is sponsored by Philips Research. During the course of the project, different cell options were examined using simulation and comparisons made between alternative cell designs. Furthermore process variations were investigated and process splits simulated. Based on the most conclusive trends in the simulation results, splits were defined in process runs for a 0.18um process, to efficiently optimise the embedded flash EEPROMs. One of the key parameters optimised was programming speed. This was of critical importance in the target applications,

but the baseline CMOS process could not be affected. After much investigation, a thermal oxidation step, near the end of the process was altered so as to fractionally reduce the encroaching oxide under the gate. The very small change in the oxide thickness did not affect the DC characteristics, but during programming the electric field increased a little which improved the programming speed by an order of magnitude, as illustrated in Figure 6. Many simulations were performed using statistical design of experiment techniques to determine the optimised processing conditions, cell layout, gate lengths and operating voltages, in order to meet the specifications for power consumption, drive current and speed. Some results are summarised in Figure 7. The use of TCAD in this project has resulted in a very efficient design cycle, in terms of both fabrication costs, development time, and optimised device parameters.

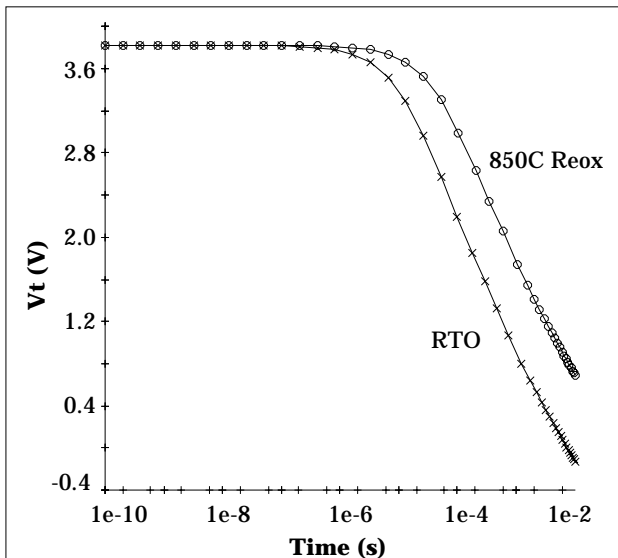


Figure 6. Simulated results showing flash EEPROM threshold voltage vs time during programming for two different poly reoxidation strategies resulting in an order of magnitude improvement in the speed.

5. CHARACTERISATION TECHNIQUES

In conventional NVM the charged stored on an electrically isolated floating gate changes the electrical characteristics of the cell and thus defines the “memory action”. Because the floating gate is not accessible (by definition) many techniques have been developed to indirectly calculate the floating gate voltage in order to determine the efficiency of the memory. The common technique is to determine the coupling ratio: which is a measure of the capacitive coupling of the voltage on each terminal (ie gate, source, substrate and drain) to the floating gate (normalised to the total capacitance). However, using TCAD, the floating gate voltage of a non-volatile memory cell directly accessed during a device simulation. This allows numerical simulation to provide a benchmark for coupling ratio measurement

methodologies. Simulation has been used to determine the relative accuracy for different coupling ratio extraction methodologies, and the results analysed to provide guidelines for valid measurements. This has had a significant contribution to the understanding of the voltage dependency of the coupling ratios [8]. The ability of the simulator to directly access the floating gate voltage has also enabled evaluation of the “polysilicon depletion” effect in NVM.

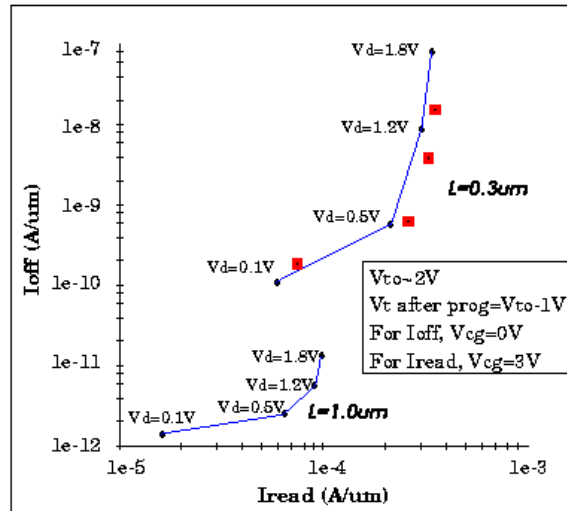


Figure 7. Measured (symbols) and simulated (lines) results showing the relationship between I_{off} and I_{read} for different gate lengths and operating voltages.

6. COMPACT MODEL DEVELOPMENT

The work described in the previous sections has concentrated on model development and engineering applications at device level. The next logical step is to develop models that can be used at circuit level, to simulate memory array architectures. A PhD work that has recently started in the NMRC aims to link the previously described topics together. The objective is to develop a general compact model for flash NVM and a parameter extraction scheme based on numerical device simulation results and accurate voltage dependant coupling ratio extraction. The circuit schematic of the macro model is shown in Figure 8. The transistor model BSIM3 is used as the core of the EEPROM macro model. Using calibrated device simulation the surface potential at three points along the channel is extracted and linearly fitted to the floating gate and drain voltages. A least squares routine is used to calculate the effective area of the Fowler-Nordheim diodes such that the total current through the three diodes is equal to the Fowler-Nordheim current simulated in the device simulation. This exciting work is at an early stage, but the results are already very promising, as shown in Figure 9.

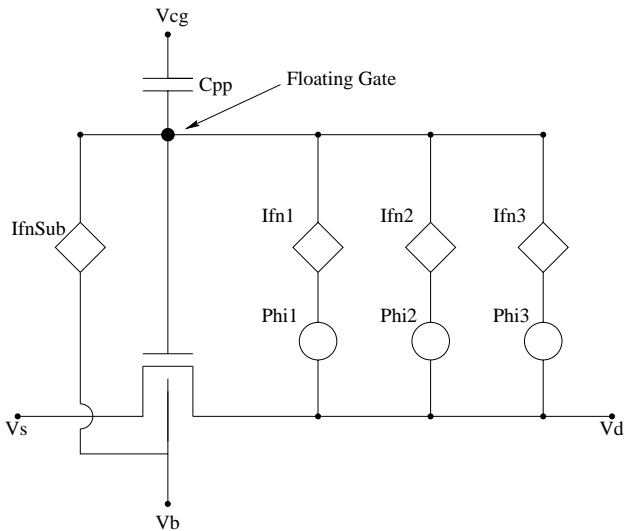


Figure 8. Simple circuit schematic showing the macro-model components of the flash model

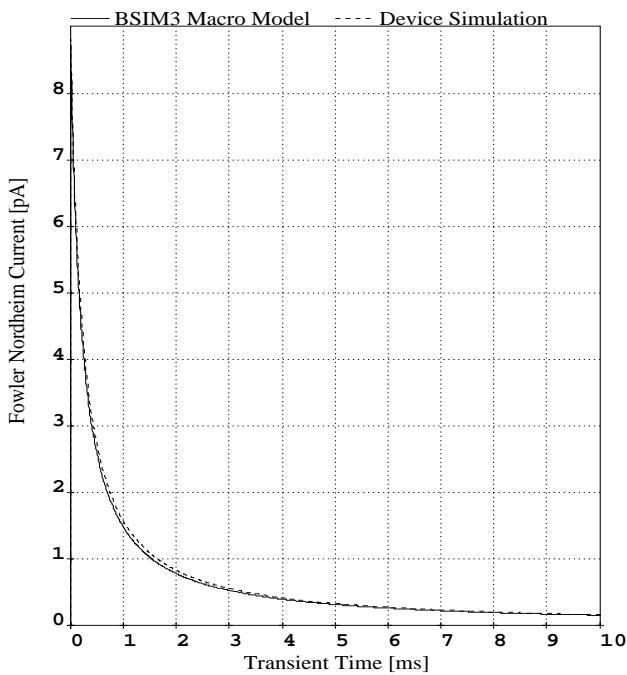


Figure 9. Initial macro-model vs device simulation results showing floating gate current vs time during programming.

7. NOVEL DEVICES

The development of embedded non-volatile memory (NVM) is becoming increasingly important for low power mobile systems, and in particular for mobile communications. An example of this is in advanced mobile phones where there is a dual requirement for a high performance core logic running DSP capability as well as CMOS embedded storage elements for protocols and executable codes on-chip. The aim of the work described in this section, also the subject of a PhD thesis, is to develop a novel flash memory with low additional mask count, with very low power consumption, suitable

for embedded applications using an original approach. The pseudo-floating gate flash EEPROM (PSI) cell looks similar to a MOSFET with polysilicon spacers flanking the gate instead of oxide spacers. This is consistent with the polysilicon emitter technology developed for a biCMOS process. The “memory action” in a conventional stacked gate flash EEPROM is realised by using the stored charge on the floating gate to change the threshold voltage of the cell. This is illustrated schematically in Figure 10. The memory action in the PSI-cell is achieved by a modification of the drain series resistance by the charge on the floating polysilicon as illustrated in Figure 11. As the maximum threshold and transconductance are determined by the intrinsic properties of the MOSFET the cell will not suffer from over-programming. The power consumption decreases and the speed increases as the cell resembles closer the intrinsic MOSFET.

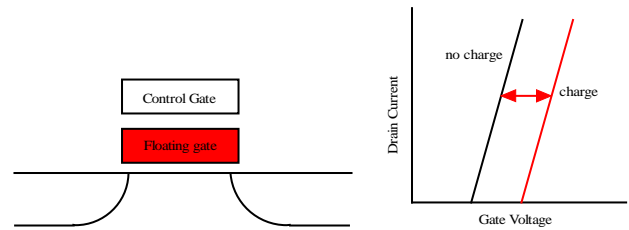


Figure 10. Cross section schematic of standard stacked gate cell and the "memory action"

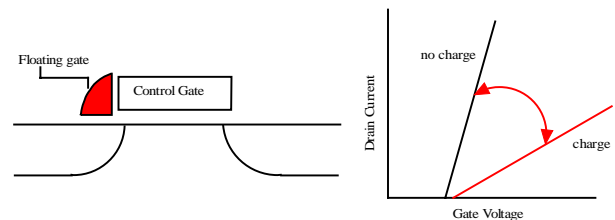


Figure 11. Cross section schematic of PSI-cell and the "memory action".

The PSI cell flash concept has been effectively demonstrated using process, device and mixed-mode circuit simulation in 0.35um and 0.18um CMOS technology and analysis of simulation results shows a possibility of scaling with CMOS in future generations, promising results for low power mobile applications of the future [9,10]. This work is now the subject of an ESPRIT research project entitled PANORAMA, and an Irish government funded project. The scanning electron microscope (SEM) cross-section shown in Figure 13 is the first silicon of this cell concept, fabricated by ST Microelectronics in Grenoble.

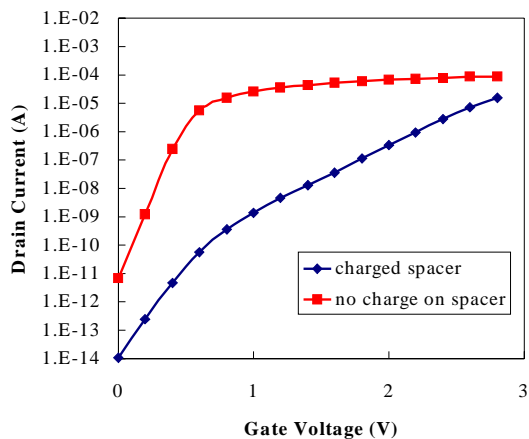


Figure 12. Simulated drain current vs gate voltage for PSI-cell with and without charge on the floating polysilicon spacer

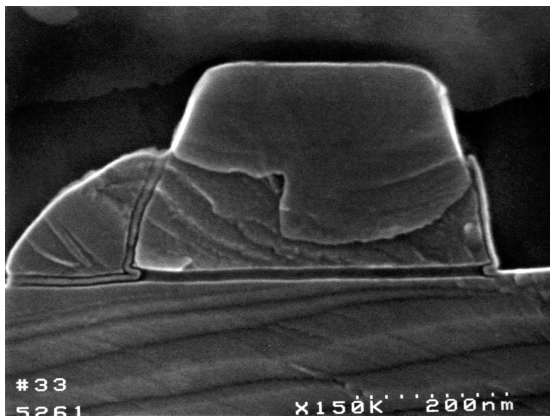


Figure 13. Cross-section SEM of first silicon fabricated of PSI-cell.

8. HIGH TEMPERATURE SOI-EEPROM

Silicon-on-insulator (SOI) technology has been in development for over 30 years, but due to technological problems in making the substrates reliable and cost effective, SOI has remained a technology with many advantages over silicon, but not exploited. IBM have recently changed that, by innovative substrate manufacture and have pioneered a new state of the art microprocessor production using SOI. SOI has a faster switching speed, lower power consumption and resistance to cosmic rays and background radioactive material. In the NMRC, SOI was identified as a solution for NVM memory applications in harsh environment applications, and in particularly high temperature. A 1kbit test memory was designed using a combination of process, device and circuit simulation to evaluate a novel SOI FLOTOX-type memory cell for high temperature applications. One major problem with high temperature is the charge leakage from the floating gate and this reduces the retention of the memory. To counteract this a refresh mechanism has been designed and implemented. This is an unusual feature because in normal operation EEPROMs are expected to

retain charge for at least ten years after programming. The refresh mechanism and the design using ZTC (zero temperature coefficient) bias points have made possible a memory which has the minimum sensitivity to the temperature of operation. To reduce electromigration effects at high temperatures tungsten metalisation has been used.

9. SUMMARY

The work described in this article represents the state of the art in non-volatile memory modelling. Following the successful Marie Curie Fellowship in 1995, the research activity in this topic has grown. Successful EU Framework 4 and Irish government research projects have funded this research and close collaboration with European industry has ensured that the results of the research activities are contributing to the European economy. As silicon technology is constantly evolving, incorporating new materials and processes, and ever decreasing geometries, so also are the models to describe the device and circuit operation evolving. This should ensure interesting challenges and research topics for our group well into the 3rd millennium.

10. REFERENCES

- [1] "Three dimensional simulation of non-volatile memory", by Ann Concannon, Technical Report, Human Capital and Mobility Contract ERBCHICT941284, May 1995.
- [2] "The Numerical Simulation of Floating Gate Non-Volatile Memory Devices", S. Keeney, PhD thesis, National University of Ireland, 1991.
- [3] "Technology Computer Aided Design of Non Volatile Memory" A. Concannon, PhD thesis, National University of Ireland, 1996
- [4] "The Numerical Simulation of Substrate and Gate Currents in MOS and EPROMs", by Ann Concannon, Francesco Piccinini, Alan Mathewson and Claudio Lombardi, in the IEDM Technical Digest, p289-292, Dec. 1995.
- [5] A. Gnudi et al, The European School of Device Modelling, Bologna, 1991
- [6] "Simulation based development of EEPROM devices within a 0.35um process", R. Duffy, A. Concannon, A. Mathewson, C. de Graaf, M. Slotboom and R. Verhaar, Proceedings of the Simulation of Semiconductor Processes and Devices Conference September 1998 p376-379.
- [7] "Scaling low power embedded flash EEPROM to 0.18um" R. Duffy, A. Concannon, A. Mathewson, M. Slotboom D. Dormans, N. Wils and R. Verhaar, in the Proceedings of Proceedings of 29th European Solid-State Device Research Conference (ESSDERC), Leuven, Belgium, p620-623 Sept 1999
- [8] "Advanced Numerical Modelling of Non-Volatile Memory Cells", R. Duane, A. Concannon, P. O'Sullivan, A Mathewson in Proceedings of 28th European Solid-State Device Research Conference (ESSDERC), Bordeaux, France, pp304-307, September 1998.
- [9] "A novel CMOS Compatible Multi -Level FLASH EEPROM For Embedded Applications", by A.Concannon, D.McCarthy, A.Mathewson, B.Guillaumot, C.Papadas and C.Kelaidis Presented at the 56th Annual Device Research Conference (DRC) in University of Virginia, Charlottesville, VA. on June 23rd 1998
- [10] "Theoretical Analysis of a Pseudo-Floating Gate Flash EEPROM Device", A. Concannon, A. Mathewson, C. Papadas, B. Guillaumot, C. Kelaidis, Proceedings of the 27th European Solid-State Device Research Conference (ESSDERC), Stuttgart, Germany, p320-323, Sept. 1997