Design of Low-cost One-time
Programmable Memory

by

Man Chiu, LEE

A Thesis Submitted to
The Hong Kong University of Science and Technology
in Partial Fulfillment of the Requirements for
the Degree of Master of Philosophy
in the Department of Electronic and Computer Engineering

August, 2009, Hong Kong
Authorization

I hereby declare that I am the sole author of the thesis.

I authorize the Hong Kong University of Science and Technology to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize the Hong Kong University of Science and Technology to reproduce the thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

[Signature]

Man Chiu, LEE

19 August 2009
Design of Low-cost One-time Programmable Memory

by

Man Chiu, LEE

This is to certify that I have examined the above MPhil thesis and have found that it is complete and satisfactory in all respects, and that any and all revisions required by the thesis examination committee have been made.

Thesis Supervisor: Professor Mansun Chan

Committee Chairman: Professor Amine Bermak

Committee Member: Professor Howard Luong

Head of Department: Professor Ross D. Murch

The Department of Electronic and Computer Engineering

19 August 2009
Acknowledgements

First of all, I wish to express my appreciation to my thesis supervisor, Prof. Mansun Chan, who has helped me understand the methodologies of learning, researching, and thinking. He also encouraged me to explore more to broaden my horizon so that I now have a clearer picture of what I should do in the future. Mansun has facilitated a very free environment for students in the group to enjoy the process of research.

I would also like to thank Prof. Howard Luong for his valuable support in the RFID project. He has provided great opportunities for people in the project team to learn and gain experience. Another professor I want to bring my gratitude to is Prof. Philip Mok, who was my final year project supervisor in my undergraduate. He has inspired so many of us in his projects and classes. Those were certainly the greatest courses we have ever had. Next, I would also like to thank Prof. Amine Bermak for being my thesis examination committee member.

I would like to thank my labmate Dr. Scottie Man, who has given huge amount of technical inputs as well as career advices to the members in the group and even people in other groups. He’s a great mentor without any doubt. I wish him a very happy life and a successful career. Meanwhile, I wish to thank Tim Chan, Ricky Ng, Ben Ng, Henry Kwong, Dong Wei and Liu Feng for being my dinner buddies and having those nice talks with me. Special thanks to Ben Ng and Dong Wei for their help in measurements.

As my partner in the RFID project, Eric Yi has given me lots of help and advices. His kindness to other people has impacted me tremendously. I wish he can find his eternity very soon. Another partner I want to thank is Matthew Law, who has helped me in technical aspects and given me useful advices in my career planning.

I also want to express my gratitude to Mr. Allen Ng and Mr. Siu Fai Luk for their technical support, help and care during my MPhil study.
Last but not least, I want to thank my parents and all my other friends for their support in these few years. Without them, my study life here wouldn’t be as easy as it has been.
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title Page</td>
<td>i</td>
</tr>
<tr>
<td>Authorization Page</td>
<td>ii</td>
</tr>
<tr>
<td>Signature Page</td>
<td>iii</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>iv</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>vi</td>
</tr>
<tr>
<td>List of Figures</td>
<td>ix</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xii</td>
</tr>
<tr>
<td>Abstract</td>
<td>xiii</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Flash Memory and EEPROM</td>
<td>3</td>
</tr>
<tr>
<td>1.2 Amorphous Silicon Anti-fuse</td>
<td>6</td>
</tr>
<tr>
<td>1.3 3D OTP Memory</td>
<td>7</td>
</tr>
<tr>
<td>1.4 Thesis Structure and Organization</td>
<td>9</td>
</tr>
<tr>
<td>2 Selection of Memory Elements</td>
<td>12</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>------------------------------------------------------------</td>
</tr>
<tr>
<td>2.1</td>
<td>Diode Anti-fuse Memories</td>
</tr>
<tr>
<td>2.1.1</td>
<td>Diode Elements in a Standard CMOS Process</td>
</tr>
<tr>
<td>2.1.2</td>
<td>Diode Breakdown Process</td>
</tr>
<tr>
<td>2.1.3</td>
<td>Diode as Anti-fuse Memory</td>
</tr>
<tr>
<td>2.2</td>
<td>Gate-oxide Anti-fuse Memory</td>
</tr>
<tr>
<td>2.2.1</td>
<td>Gate-oxides in a Standard CMOS Process</td>
</tr>
<tr>
<td>2.2.2</td>
<td>Gate-oxide Breakdown Process</td>
</tr>
<tr>
<td>2.2.3</td>
<td>Gate-oxide as Anti-fuse Memory</td>
</tr>
<tr>
<td>2.3</td>
<td>Summary</td>
</tr>
<tr>
<td>3</td>
<td>Design of OTP Memory Cells</td>
</tr>
<tr>
<td>3.1</td>
<td>Overview of Read and Write Operations</td>
</tr>
<tr>
<td>3.2</td>
<td>Cascode GOAF Memory Cell</td>
</tr>
<tr>
<td>3.3</td>
<td>Pulse Write Operation</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Determination of Pulse Height</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Determination of Pulse Width</td>
</tr>
<tr>
<td>3.4</td>
<td>Sensing of Memory Contents</td>
</tr>
<tr>
<td>3.4.1</td>
<td>On and Off Currents</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Post-breakdown Resistance</td>
</tr>
<tr>
<td>3.4.3</td>
<td>Sensing with Driving Devices</td>
</tr>
<tr>
<td>3.5</td>
<td>Summary</td>
</tr>
<tr>
<td>4</td>
<td>Embedded OTP Memory for Low-power Applications</td>
</tr>
<tr>
<td>4.1</td>
<td>Overview of the OTP Memory</td>
</tr>
<tr>
<td>4.2</td>
<td>Memory Array, Decoding and Level-shifting</td>
</tr>
<tr>
<td>4.3</td>
<td>High Voltage Tolerant Circuit</td>
</tr>
<tr>
<td>4.4</td>
<td>Sensing Path</td>
</tr>
<tr>
<td>4.5</td>
<td>Write Operation</td>
</tr>
</tbody>
</table>

vii
4.6 Read Operation ................................................. 54
4.7 Interface of OTP Memory inside RFID Tag ................. 55
4.8 Summary ................................................... 55

5 Conclusions .............................................. 58
  5.1 Summary of Works Done .................................... 59
  5.2 Future Works ............................................. 60
  5.3 Appendix ................................................ 60
    5.3.1 Pre-charge Sensing Scheme ......................... 60
    5.3.2 Cell Layout ........................................ 61
# List of Figures

1.1 The Floating Gate Structure ........................................... 3  
1.2 (a) Program Operation; (b) Erase Operation ........................ 4  
1.3 Band Diagram for FN Tunneling ...................................... 5  
1.4 Id-Vg Characteristics Before and After Programming .............. 6  
1.5 The ViaLink Technology ............................................ 7  
1.6 Illustration of a 3D OTP Memory Cell .............................. 8  
1.7 Cross-section of the 3D OTP Memory ................................ 9  
1.8 IV Characteristics of 3D OTP Memory Cells ......................... 10  
1.9 Design Flow of an OTP Memory .................................... 11  
2.1 Diode Anti-fuses in a Standard CMOS Process ..................... 13  
2.2 Poly-silicon Diode Anti-fuse in a Standard CMOS Process ....... 14  
2.3 IV Characteristics of a Diode ...................................... 15  
2.4 (a) Energy Band Diagram for Zener Breakdown; (b) Energy Band Diagram for Avalanche Breakdown ................................. 16  
2.5 Diode Currents Before and After Poly-diode Permanent Breakdown .... 17  
2.6 Cross Section of a CMOS Process Showing the NMOS, PMOS, and DNW NMOS Gate-oxide Anti-fuses .......................... 19  
2.7 The Oxide Bond Breaking ........................................... 20
3.13 Cell Currents in Log and Linear Scales under Various Post-breakdown Resistances for Minimum-size Driving Devices ............................................. 40

4.1 System Block Diagram of the GOAF Memory Array .......................... 43
4.2 A 16x8 NMOS GOAF Memory Array ................................................. 44
4.3 Decoding and 3-stage Level-shifting .................................................. 45
4.4 A Simple Level-shifter used in the OTP Memory .................................. 46
4.5 The Half $V_{PP}$ Shifter with Input and Output Signals ......................... 47
4.6 The Zero Static Current High Voltage Switch ...................................... 48
4.7 Simulation Result of the Zero Static Current High Voltage Switch .......... 48
4.8 The High Voltage Multiplexer ......................................................... 49
4.9 The Complete Sensing Path for Memory Read Operation with Constant Current Discharging ................................................................. 51
4.10 Push-pull Sensing of Memory Cell Output Current .............................. 52
4.11 Simulation Result for Write Operation of the Whole Memory Circuit 3 Pulses 53
4.12 Measurement Result for Write Operation of the Whole Memory Circuit with On-chip Charge Pump, Voltage Supplies, and References - 3 Cells .............. 53
4.13 Measured Time-to-breakdown Distribution with 100 Bits of NMOS GOAF OTP 54
4.14 Simulation Result for Read Operation of the Whole Memory Circuit with 1M ohm Resistors Replacing Half of the GOAF Elements ................................. 55
4.15 Measurement Result for Read Operation of the Whole Memory Circuit with On-chip Charge Pump, Voltage Supplies, and References ................................. 56
4.16 Interfacing of the OTP Memory with On-chip Power Management Unit and Digital Baseband .............................................................................. 56
4.17 Chip Micrograph of the 128-bit OTP Memory in a RFID Tag .................. 57

5.1 Chip Micrograph of the 128-bit OTP Memory in a RFID Tag .................. 62
5.2 (a)Cell Layout with Sharing of Contact; (b)Elimination of Source or Drain .................................................. 63
List of Tables

2.1  Breakdown Currents and Voltages of 3 Diode Anti-fuses .................. 18

5.1  A Table Summarizing Various Parameters for the Embedded NMOS GOAF
     OTP Memory ................................................................. 59
Design of Low-cost One-time Programmable Memory

by Man Chiu, LEE

The Department of Electronic and Computer Engineering

The Hong Kong University of Science and Technology

Abstract

Non-volatile memories (NVMs) are essential for a variety of embedded applications such as encryption keys, code storage, and analog trimming. One of the lowest-cost non-volatile memories is the one-time programmable (OTP) memory. Most of the existing OTP memories are fabricated with special processes where extra masks are added to a standard CMOS process. It is, however, still necessary to develop an OTP memory in a pure CMOS process to further reduce the cost for many applications such as radio frequency identification (RFID) tags.

Different from conventional OTP memory designs, the targeted embedded applications in this work require ultra-low power consumption in both the read and the write operations. Because of that, power reduction strategies are provided during memory cell and circuit designs. Owing to the need for high voltages during write operations, gate-oxide reliability is also a major concern. To solve all the high voltage problems in the memory circuit, the Cascode technique is used extensively to maintain programming reliability.

Various anti-fuse memory elements manufactured with a standard 0.18µm CMOS process from a major foundry are compared and the gate-oxide anti-fuse (GOAF) is chosen as the
most suitable one when programming current is seriously constrained. Different combinations of GOAF memory cells have been fabricated, measured, and studied. A complete GOAF OTP memory circuit has been designed and fabricated with the above process. The functions of each part are described and measurement results are shown and analyzed in this thesis.
Chapter 1

Introduction

The demand for low cost electronic equipments continues to drive CMOS technology to its physical limits. Among all technologies, non-volatile memory (NVM) becomes more and more essential for various applications, such as encryption keys, code storage, and analog trimming. All these applications require permanent data storage as well as field programmable capability. One important issue in today’s memory module design for integrated circuits is the manufacturing cost. For example, in radio frequency identification (RFID) transponder design, the cost for each additional RFID tag produced is extremely important as it affects the feasibility of large scale implementation. If barcode labels on commercial products are to be replaced, the unit cost must be reduced to a few US cents or even lower. One way of reducing this cost is to use low cost standard CMOS process, where the number of masks is minimal because only those essential layers for high performance RF and logic circuits are used. The other issue is the minimization of the manufacturing cost requires the RFID tags to be passive, which means there is no battery inside and the whole chip is powered by the energy stored in the storage capacitor, so power consumption becomes critical.

The cost and power consumption issues mentioned above are not unique to RFID tag design, they are also essential for other applications. This is why there must be conscious effort to look for solutions using standard CMOS without any additional mask and low power design techniques. These usages of NVMs are collectively called embedded applications
throughout this thesis. In this chapter, the working principles of some conventional non-volatile memories will be introduced. This gives an overview of the mainstream technologies and also emphasizes their inadequacy, and thus results in the need for the selection of the most suitable non-volatile memory in later chapters.
1.1 Flash Memory and EEPROM

Flash memory and EEPROM are now the most popular non-volatile memories in the global market. While Flash memory is better for large-scale data storage such as those in USB drives due to its higher density, EEPROM is suitable for embedded applications like the security code storage in smart cards, where a lower bit-count is needed. Both of these make use of the phenomenon of charge injection into the floating gate of a double-poly MOS transistor [1]. As shown in Figure 1.1, a Flash memory cell is made up of a floating gate transistor with two layers of poly-silicon, of which one acts as the control gate (CG) and the other acts as the floating gate (FG).

![Figure 1.1: The Floating Gate Structure](image)

Figure 1.2 exhibits the programming and erasing mechanisms of a commercial Flash memory. During the program operation, a high voltage is applied to the CG, a median voltage is applied to the drain and the source is grounded. During the program operation, electrons are accelerated by the horizontal electric field and move towards the drain side. While gaining energy from the lateral electric field, electrons will also lose energy by collisions with the lattice ions. With the help of the vertical electric field from the high CG voltage, a small fraction of the electrons have enough energy to overcome the barrier between oxide and silicon conduction band edges [2]. This process is called hot electron injection (HEI).
Figure 1.2: (a) Program Operation; (b) Erase Operation

To erase the memory cell, the effect called Fowler-Nordheim (FN) tunneling is used. Figure 1.3 shows an energy-band diagram of a MOS capacitor structure with negative bias applied to the poly electrode with respect to the p-substrate. In the quantum mechanical framework, there is certain amount of probability that an electron in the poly-gate region can penetrate through the oxide barrier and move to the p-substrate. The current density passing through the oxide can be expressed as

$$J = \frac{q^3 F^2}{16\pi^2 h^2 \Phi_B} \exp\left[-4\left(2m^*_o\right)^{1/2}\Phi_B^{3/2}/3hqF\right]$$

(1.1)

where $\Phi_B$ is the barrier height, $m^*_o$ is the effective mass of the electron in the forbidden gap of the oxide, $h$ is the Planck’s constant, $q$ is the electronic charge, and $F$ is the electric field through the oxide.

One way to achieve the erase operation by FN tunneling in a Flash memory cell is to apply a negative high voltage to the CG and a positive median voltage to the source while the drain is floating.

Since the floating poly layer is enclosed by dielectric material such as silicon dioxide, electrons injected into the FG will stay inside and can only leak out very slowly. Electrons collected will then cause an increase in the threshold voltage $V_{th}$ of the transistor. Therefore,
1.1. Flash Memory and EEPROM

the $I_d-V_g$ characteristic will shift to the right side, as shown in Figure 1.4, and this change can be easily detected by comparing the two current levels at the same gate voltage. This creates two different states in the memory cells, in digital terms, 0 and 1. The amount of threshold voltage shift is dependent on the how much charge is inside the FG and it’s given by $V_{th} = Q/C_{FC}$, where $Q$ is the magnitude of charge stored inside the floating gate and $C_{FC}$ is the capacitance between the floating gate and the control gate.

The operations of EEPROM is similar to the Flash memory except that EEPROM can be erased bit-by-bit thanks to the existence of the access transistor in series with the FG transistor. In spite of their re-writability, both of these memories require extra masks compared with the standard CMOS process and thus a higher manufacturing cost. In addition to that, these memories suffer from serious problems due to technology scaling. Data retention in a FG cell is getting worse when the gate-oxide thickness keeps on down-scaling. To avoid gate leakage current, the equivalent oxide thickness (EOT) of Flash memory devices has lagged behind that of logic MOS devices for a few generations down the road of scaling[3]. This is definitely undesirable if higher memory density is needed in applications.
1.2 Amorphous Silicon Anti-fuse

In order to lower the manufacturing cost, for many applications where rewritability is not a concern, one-time programmable (OTP) memories are definitely the most suitable choices. Conventionally, anti-fuses (AFs) are the most widely used elements for configuring an OTP memory IC. An anti-fuse is an electrical device that performs the opposite function to a fuse. Whereas a fuse starts with a low resistance and is designed to permanently break an electrically conductive path, an anti-fuse starts with a high resistance and is designed to permanently create an electrically conductive path.

Figure 1.5 illustrates a type of AF memory, the ViaLink [4], invented by a company called QuickLogic. The ViaLink is a metal electrode amorphous silicon anti-fuse. As shown on the figure, the ViaLink and the tungsten via plug are sandwiched between two metal layers. The ViaLink is programmed by applying a voltage to the top electrode while the bottom electrode is grounded. The conducting programmed filament is formed by moving electrode material in the electron flow direction into the amorphous silicon. Analysis indicates that the filament consists of polycrystalline tungsten silicide. Near the bottom electrode the filament contains higher amounts of Ti in the form of titanium silicide or Ti-W silicide. The programming
mechanism generates enough heat to move the electrode material into the amorphous silicon, react the metal to form a silicide, and recrystallize the adjacent silicon.

Like many other anti-fuses, this type of memory has been used as field programmable gate array (FPGAs) elements. It is attractive because it meets the low capacitance and low resistance requirements for high speed applications. Although ViaLink can be embedded into a CMOS process, it requires a few more masks than a standard CMOS process. Clearly, it is not the inexpensive NVM required in this work.

1.3 3D OTP Memory

Figure 1.6 shows another type of OTP memory which makes use of the oxide breakdown phenomenon to form anti-fuses [5], [6]. It employs a multiple-layer stacked structure to increase the memory density. Figure 1.7 illustrates the cross-section of the 3D OTP memory [7]. Each layer of the 3D OTP memory consists of a group of word-lines and a group of bit-lines perpendicular to each other. Both the word-lines and bit-lines are laid out as parallel lines with minimum pitch to maximize the memory density. The doped poly-silicon memory cells are located at every intersection between a word-line and a bit-line. A $SiO_2$ anti-fuse is embedded in the memory cell. The entire memory cell can be modeled as a capacitor in
series with a diode.

Out of the factory, all memory cells are in an un-programmed virgin state, or a "1". Programming of the memory cell is accomplished by applying a high voltage programming pulse to the selected memory cell and an inhibit voltage to the un-selected bit-lines. The $SiO_2$ anti-fuse breaks down under the programming pulse and creates a conduction path between the anode and cathode. After programming, the doped poly-silicon forms a junction diode and the memory cell becomes a "0".

![Figure 1.6: Illustration of a 3D OTP Memory Cell](image)

Figure 1.6 compares the currents before and after programming. At the same read voltage, the programmed memory cell gives a current level a few orders of magnitude larger than the current from an un-programmed memory cell.

Despite the high density and sufficiently large on/off current difference ratio, the 3D OTP memory has to be manufactured with a special process. This non-standard CMOS requirement inevitably raises the cost, and it is obvious that this memory is not suitable for

---

**Man Chiu, LEE**  The Department of Electronic and Computer Engineering, HKUST
low density applications such as smart cards.

1.4 Thesis Structure and Organization

In this thesis, a detailed study of standard CMOS OTP memories is given. Design guidelines on the selection of memory elements, the memory cell formation, determination of design parameters, and the construction of complete memory arrays are presented.

In Chapter 2, several low cost memories in a standard CMOS process are discussed. These can be categorized into 2 groups: the diode anti-fuse memories and the gate-oxide anti-fuse (GOAF) memory. The working principles behind these memory elements are described in this part. Samples of these memories are fabricated, measured and evaluated. Comparisons are made to select the most suitable memory for low power embedded applications.

Chapter 3 focuses on the development of the memory cell. Based on a Cascode GOAF OTP memory cell structure, various 3-transistor cells are fabricated and discussed. Design

Device Characterization Laboratory

Man Chiu, LEE
parameters are determined by the analysis of experimental results. Different from stand-alone memory designs, the targeted memory is constrained by currents and voltages generated from the on-chip power management unit (PMU) in the system.

An embedded GOAF OTP memory for low cost low power RFID applications is presented together with the measurement results in Chapter 4. This part covers the arrangement of the memory array, the decoding circuit, the level-shifting circuits, the design of high voltage tolerant circuits, and the sensing path. The Cascode technique is also used in the high voltage section of the memory circuit. With proper array configuration and the use of serial operations, the power and area can be significantly reduce while maintaining high voltage reliability of the whole memory.

Finally, in Chapter 5, the results in this thesis are summarized and some future research directions are presented.

Figure 1.9 illustrates the design flow of the OTP memory. The design process starts from the basic memory element which can permanently store information, to the cell design with determination of read/write parameters, and finally to the formation of a complete memory.
with the peripheral circuits. Constraints from the system outside the memory need to be taken care of throughout the whole process.

![Design Flow of an OTP Memory](image)

**Figure 1.9:** Design Flow of an OTP Memory
Chapter 2

Selection of Memory Elements

To select the most suitable memory element for low cost low power embedded applications, two categories of OTP memories are presented and compared in this chapter. They are the diode anti-fuse memories and the gate-oxide anti-fuse (GOAF) memory. These are basically originated from the parasitic elements and MOS-capacitors that can be found in a standard CMOS process. For low power embedded applications, the design criteria are low programming voltage and current. Only if the programming voltage is within the breakdown limits of source/drain to substrate/n-well junctions in a CMOS process, the anti-fuse element is considered a valid candidate. As for programming current, the value is limited by the maximum loading current of the on-chip charge pump to maintain its output voltage. In low power embedded applications, this value is on the order of 10µA due to the limited area of on-chip capacitor allowed.
2.1 Diode Anti-fuse Memories

2.1.1 Diode Elements in a Standard CMOS Process

Reverse breakdown has been used in the designs of anti-fuses, both standard CMOS and non-standard CMOS designs[8][9]. Figure 2.1 illustrates a cross-section of a standard CMOS process. Instead of showing the regular NMOS and PMOS devices, the figure shows all the diode elements which can be found in such a process. These diodes are named D1 to D5 here.

![Diode Anti-fuses in a Standard CMOS Process](image)

Figure 2.1: Diode Anti-fuses in a Standard CMOS Process

D1 to D4 are PN-junctions on the silicon substrate or inside the N-well. D5 is built with the poly-silicon layer and it’s isolated from the substrate by the field oxide. The diode is formed by opposite implantation types on the same poly-silicon bar, as shown in Figure 2.2. In a modern CMOS process, to reduce parasitic resistance, silicide is used on both the poly-silicon layer and the active area of a MOSFET[10]. However, resistance reduction is undesirable in this work since it is equivalent to a small resistor in parallel with the poly-diode and this effectively eliminates the diode as no current will flow through it. To reserve the
diode, a silicide block, which is originally employed to create poly resistor with high resistivity (high-resistive poly), is used to avoid the formation of silicide. Now all the possible diode anti-fuses in a standard CMOS process are located, the next step is to study their suitability as memory elements for the embedded applications.

2.1.2 Diode Breakdown Process

Figure 2.3 shows the typical IV characteristics of a diode. Reverse breakdown occurs at $V_{br}$ where the potential at the n-doped region is higher than the p-doped region. This reverse biasing can be a few volts to a few thousand volts before breakdown happens. For designing standard CMOS anti-fuse memories, it is necessary to acquire permanent breakdown so that in the reverse-biasing region the diode is changed from high impedance state to low impedance state indefinitely.

There are two types of breakdown mechanisms in a PN-junction, namely, the Zener breakdown and the Avalanche breakdown. These phenomena exist in all diodes and one
of them is more prominent than the other depending on doping concentrations and biasing voltages\cite{11}.

In Zener breakdown, electrons tunnel directly from the valence band of the p-side to the conduction band of the n-side. For tunneling current to occur, there should be a large number of electrons separated from a large number of empty states by a narrow barrier of finite height(Figure 2.4(a)). That means the diode should be heavily doped to keep the depletion region narrow enough. Zener breakdown normally occurs at a lower reverse-biasing voltage than Avalanche breakdown and the latter requires a lower doping concentration.

Avalanche breakdown is caused by impact ionization of electron-hole pairs(Figure 2.4(b)). The electric field in the depletion region of a diode can be very high. Electron/holes that enter the depletion region will be accelerated. As they collide with the atoms, electrons can be knocked out from their bonds, creating additional electron/hole pairs and thus additional current. When these secondary carriers are swept into the depletion region, they too are accelerated and the process repeats itself. This phenomenon is just like an avalanche. The efficiency of the avalanche effect is characterized by a multiplication factor $M$, which depends

\begin{figure}
\centering
\includegraphics[width=0.5\textwidth]{diode_characteristics.png}
\caption{IV Characteristics of a Diode}
\end{figure}
Figure 2.4: (a) Energy Band Diagram for Zener Breakdown; (b) Energy Band Diagram for Avalanche Breakdown

on the reverse voltage (Equation 2.1).

\[
M = \frac{1}{1 - \left(\frac{V}{V_{br}}\right)^n}
\]  

(2.1)

where \( n \) ranges from 3 to 6, \( V \) is the applied (reverse) voltage, and \( V_{br} \) is the breakdown voltage.

Neither Zener nor avalanche breakdown are inherently destructive in that the crystal lattice is damaged. However, when the reverse breakdown current increases to a certain critical value, second breakdown will happen. It is a phenomenon called filamentation [12] where the diode is melted permanently. This process consists of three stages: nucleation, growth and melt transition. Only when all the steps have finished does the junction change its state. The voltage drop across the diode will become smaller than that in a non-permanent breakdown. Effectively, there is a resistor in parallel with the diode and the current at a certain reverse-biasing voltage will be much larger than the original reverse saturation current. Taking poly-diode as an example, Figure 2.5 shows the difference between the diode currents before and after permanent breakdown.
2.1.3 Diode as Anti-fuse Memory

Breakdown voltages and currents are critical for evaluating the performance of a diode as an anti-fuse.

Table 2.1 summarizes the breakdown characteristics of all the diode anti-fuses [13]. D2 and D4 have been eliminated because of their lack of isolation from the substrate. The breakdown voltages of all 3 diodes are lower than 6V. This is an advantage because the number of stages needed for charge pump design is small.

However, all 3 diodes require breakdown currents in the order of 1mA or even higher. In fact, during some measurements, the contacts connecting the diodes to the outside of the chip became opened before the diodes broke down. For the targeted passive RFID tag application, this large write current requirement is clearly not suitable. The reason is charge pump load current of that order implies large capacitors, which have total silicon area dominating the whole RFID tag.
### Table 2.1: Breakdown Currents and Voltages of 3 Diode Anti-fuses

<table>
<thead>
<tr>
<th>Doping Combination</th>
<th>Breakdown Voltage</th>
<th>Breakdown Current</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1 n+/p+</td>
<td>4-5V</td>
<td>&gt;10mA</td>
</tr>
<tr>
<td>D3 n/p+</td>
<td>5-6V</td>
<td>&gt;10mA</td>
</tr>
<tr>
<td>D5 n+/p+</td>
<td>4-5V</td>
<td>~1mA</td>
</tr>
</tbody>
</table>

#### 2.2 Gate-oxide Anti-fuse Memory

##### 2.2.1 Gate-oxides in a Standard CMOS Process

In addition to the diode anti-fuses, one other important anti-fuse element that can be found in a standard CMOS process is the gate-oxide anti-fuse (GOAF). In modern CMOS processes, such as those at the 0.18\(\mu\)m technology node, the gate-oxides become thin enough to be broken down by a supply voltage a few times higher than the nominal \(V_{DD}\). Figure 2.6 illustrates two cross-sections of a CMOS process where the MOS structures act as the anti-fuse elements. The deep n-well (DNW) NMOS can be a candidate because the DNW is usually used in RF circuits in the system, so it’s free of extra masks.

There are usually two types of transistors with different oxide thicknesses: the thin-oxide core devices and the thick-oxide input/output (I/O) devices. For this particular 0.18\(\mu\)m process, the thin-oxide has a thickness of 4nm and the thick-oxide has a thickness of 7nm. Thus, only the thin-oxide devices will be used as anti-fuse elements because their breakdown voltages are much lower.

##### 2.2.2 Gate-oxide Breakdown Process

Although the exact physical mechanism responsible for oxide breakdown is still an open question, it is generally believed that an applied voltage and thus the resulting tunneling electrons create defects in the oxide film. The number of defects accumulate over time and finally reach a critical density to trigger an abrupt loss of the capacitive property. A sudden increase of gate current induces local hot spots, which lead to permanent damage in the gate
2.2. Gate-oxide Anti-fuse Memory

Figure 2.6: Cross Section of a CMOS Process Showing the NMOS, PMOS, and DNW NMOS Gate-oxide Anti-fuses
oxide.

This subsection will go over some information that is essential for the understanding of the CMOS gate-oxide breakdown process. Oxide traps are defects within the gate oxide. They are named this way because the degraded oxide can trap charges. Gate-oxide breakdown starts with the forming of oxide traps. It can been thought that the Si-O-Si bond is broken and replaced by the Si-Si bond during oxide breakdown [14], as illustrated in Figure 2.7. The Si-Si bond is very weak and it can be either an electron trap or a hole trap.

![Figure 2.7: The Oxide Bond Breaking](image)

Figure 2.8 shows the detailed process of oxide wearout, which explains how oxide breakdown is induced by oxide traps. At the first step, only a few traps exist and they are non-overlapping, so they do not conduct. At the second step, as more and more traps are formed due to the stress on the oxide, these traps become overlapping. At the third step, a conducting path is created so that charges can move between the anode and the cathode. Gate-oxide breakdown occurs [15] and this type of breakdown is called soft breakdown (SBD).

When there is conduction, new traps are created by thermal damage, which in turn increases the conductance as shown in Figure 2.9. The increased conduction leads to thermal runaway [16] and finally to a lateral propagation of the breakdown spot. The silicon within the breakdown spot starts to melt, and oxygen is released, and a silicon filament is formed.

Man Chiu, LEE  
The Department of Electronic and Computer Engineering, HKUST
2.2. Gate-oxide Anti-fuse Memory

**Figure 2.8:** The Oxide Wearout Process

**Figure 2.9:** The Process of Hard Breakdown in Gate-oxide
in the breakdown spot. This is called hard breakdown (HBD).

There are many different models on the generation of defects in gate-oxides. Two important ones are the anode hole injection (AHI) model and the thermochemical model. It is still controversial as to which model is correct since both of them are supported by experimental data [17] under different conditions.

The AHI model is also named the 1/E-model because of its relationship with the Fowler-Nordheim (FN) electron tunneling current, as shown by Equation 1.1 in Chapter 1.

The AHI model is based on the process of electron injection into the oxide, which generates holes at the anode. These hole will then get trapped in the oxide. At high fields, this model shows better agreement to experimental data because, at such fields the electron tunneling is significant, and hole generation dominates over the thermochemical model.

At high electric fields, the electrons arriving at the gate have a high kinetic energy (> 8MV/cm). When these hot electrons reach the gate electrode, they transfer their entire energy to a deep-valence band electron, and then this electron is promoted to the lowest available electron energy state, which is the conduction band edge of the anode. Once the electron reaches the conduction band, it creates a hot hole, which tunnels into the oxide. The holes which enter the oxide, allow for increased current density due to hole-induced trap generation. Once the trap has been created, and there is increased current density, there are more high energy electrons entering the gate which can create more hot holes, and thus there is positive feedback until breakdown occurs.

The thermochemical model is also named the E-model because it predicts the time-to-breakdown, \( t_{BD} \), with the following relationship

\[
\log(t_{BD}) \propto \frac{\Delta H_0}{k_BT} - \beta E
\]  

(2.2)

where \( \Delta H_0 \) is the activation energy, \( k_B \) is the Boltzmann’s constant, \( T \) is the absolute temperature, \( \beta \) is the field acceleration factor, and \( E \) is the electric field strength applied on the gate-oxide.

Man Chiu, LEE  
The Department of Electronic and Computer Engineering, HKUST
2.2.3 Gate-oxide as Anti-fuse Memory

The breakdown voltages of both thin and thick gate-oxide are shown on Figure 2.10. A semiconductor parameter analyzer was used in this test and the setup is also shown on the figure. Oxide breakdown happens at around 6.5V and 11V for thin-oxide and thick-oxide NMOS devices. Thus, the thin-oxide device is suitable for anti-fuse application because only a few stages are required in the charge pump design.

Note that the nominal supply voltages for thin-oxide and thick-oxide devices are 1.8V and 3.3V, respectively. These voltages are too low for oxide breakdown to occur. In fact, even if these numbers are increased by 20%, to about 2.2V and 4V, oxide reliability can still be maintained after long-term operations. There are not many defects in the oxide under such gate voltages. This is verified by the small gate currents on the plot. This property of the thick-oxide device was used for building high voltage tolerant circuits, as will be shown in later chapters.

![Figure 2.10: IV Characteristics during Oxide Breakdown of Both Thin-oxide and Thick-oxide NMOS Transistors](image)

Although the breakdown voltage for thin gate-oxide is a little higher than some diode anti-fuses, the current limit for permanent breakdown is much lower. As shown in Figure 2.11, constant current stress in sampling mode is used to find out the current for hard breakdown. At 10nA constant current stress, several SBD events occur, during which the
anode voltage drops by a certain amount. However, this is not large enough to create HBD. When the current is increased to 100nA, HBD occurs and the voltage drops to 0V, which means the post-breakdown resistance is very small compared with the internal resistance of the current source. Figure 2.12 plots the read currents after both cases. It’s clear that the post-breakdown resistance for the 100nA case is much smaller than that for the 10nA case. This further confirms 100nA is the current level for GOAF element breakdown.

\[ \text{Figure 2.12: IV Characteristics of NMOS GOAF after 10nA and 100nA Constant Current Stress} \]

\[ \text{Figure 2.11: Time-to-breakdown Characteristics of NMOS GOAF with Constant Current Stress at 10nA and 100nA} \]

From the previous discussion, GOAF memories are obviously better than diode anti-
2.2. Gate-oxide Anti-fuse Memory

fuses for low power applications like passive RFID tags. However, that’s not the end of the selection process because there are 3 types of GOAF elements as shown before. Memory design normally involves the minimization of cell area. Among the 3 types of GOAF elements, PMOS and DNW NMOS require the use of n-well and deep n-well. That results in much larger area compared with NMOS GOAF. For this particular process, the area ratio for the 3 is approximately 1:2:60. The deep n-well design rule clearly makes DNW NMOS an undesirable choice. NMOS GOAF, however, has a potential problem compared with the other 2 choices. Breakdown from the gate to the substrate might occur to create leakage current. In the following part, it will be shown that this is not a serious issue and simple NMOS GOAF is still the best choice.

After oxide breakdown, the NMOS GOAF element can be modeled by the 3 structures in Figure 2.13. Models (a) and (b) are the same structure representing gate-to-source/drain breakdown and the resulted current path is almost purely resistive; model (c) represents gate-to-body breakdown and the resulted current path contains a reverse-biased diode.

The reverse diode in Model (c) is formed by the n+ poly-silicon gate and the p substrate, as shown in the cross-section of the NMOS GOAF. Under high voltage stress, only leakage current on the order of 1pA can pass through the reverse diode. As shown in Figure 2.14, the gate current is less than 10pA under 9V stress. This effectively stops the breakdown process due to the lack of charge to create defects. As a result, breakdown is inefficient from the gate to the substrate. Even if the oxide in this middle region breaks down, there is still a reverse diode to stop current from flowing through.

It can also be shown that breakdown will occur much more easily near the source and drain regions. For example, if only the source or the drain region is connected to the ground and the substrate is floating, when high voltage stress is applied at the gate, the resulting breakdown curve will be the same as the one in Figure 2.10. From the gate to the body, there is still a reverse diode. One important reason for easier source/drain breakdown is in these regions, the doping concentration is higher and doping atoms worsen the oxide quality.
Chapter 2. Selection of Memory Elements

Figure 2.13: Post-breakdown Model for NMOS GOAF: (a) Gate-to-Source Breakdown, (b) Gate-to-Drain Breakdown, and (c) Gate-to-Body Breakdown

Figure 2.14: Measured Gate-to-Body Currents during Write Operations
grown on top of it. Therefore, it can be concluded that Models (a) and (b) are sufficient for a NMOS GOAF after breakdown.

![Image of Oxide Interface at Gate-to-Drain Overlap]

**Figure 2.15:** Oxide Interface at Gate-to-Drain Overlap

### 2.3 Summary

In this chapter, the selection process of the most suitable standard CMOS anti-fuse was presented under the context of passive RFID tag. The detailed breakdown processes in both diode anti-fuse and gate-oxide anti-fuse were described. It has been shown that the GOAF is a better choice than the diode anti-fuse due to its much lower write current requirement. Among 3 types of GOAF elements, the simple NMOS is the best because of its smallest area and most importantly breakdown can be assumed to happened only near the source and drain regions.
Design of OTP Memory Cells

With the understanding of the gate-oxide anti-fuse (GOAF) in the previous chapter, the scope of study will be extended and more insight will be presented in this chapter. A reliable memory cell using standard CMOS processes will be presented and design parameters will be determined by experiments. The Cascode GOAF memory cell structure forms the basis for the study in this chapter [19].

The need for Cascode memory cells stems from the fact that supply voltages higher than the nominal $V_{DD}$ for normal operations are required during write operation. This leads to one important issue in the design of the memory - gate-oxide reliability for the driving devices. The second major concern here is the limited current available for the write operation. For low power embedded applications, the design concern here is significantly different from those mass storage usages where sufficient power always exists. More specifically, the transient current available for programming a GOAF element is seriously limited. This current limit $I_{LIMIT}$ is defined to be 20µA here for the passive RFID application.
3.1 Overview of Read and Write Operations

To begin with, it’s useful to show the fundamental ideas of read and write operations here. The conceptual write process is shown in Figure 3.1. The high voltage switching circuit necessary for write operation is represented by a single MOSFET and the GOAF memory cell is represented by a MOS capacitor (MOSCap). During write operation, a single bit is selected first, by the use of word-line and bit-line select transistors. The bottom plate of the MOSCap is pulled down to ground level while a high voltage pulse is applied to the top plate. The switch input controls the pulse width for programming the GOAF cell and write pulse is of smaller amplitude than the high voltage pulse so that the switch accomplishes an upward level-shifting action. The important parameters here are the programming pulse height \( V_{PP} \) and width \( t_{PW} \).

![Figure 3.1: Conceptual Write Operation of a GOAF Memory Cell](image)

In read operation, as shown in Figure 3.2, a constant voltage \( V_{DD} \) is applied at the top plate of the anti-fuse. The bottom plate of the anti-fuse is connected to the sensing module. Since the memory cell can be capacitive or resistive, the cell current \( I_{CELL} \) will be at different levels. This current information determines the output voltage level and thus the memory content is known.
3.2 Cascode GOAF Memory Cell

Showing in Figure 3.3 is a Cascode 3-transistor GOAF OTP memory cell with a thin-oxide NMOS transistor acting as the anti-fuse element. M1 is the anti-fuse and both M2 and M3 are thick-oxide driving devices. M3 functions as the word-line select transistor while M2 is responsible for high voltage blocking.

Figure 3.4 illustrates a 2x2 memory array formed by connecting all the anti-fuse elements to the same node $V_{ARRAY}$, which represents the high voltage pulses applied to that node during write operation or the nominal supply voltage for thin-oxide devices during read operation. $V_M$ is a constant voltage which biases all the gates of the upper transistors at the Cascode structure and it is equal to $V_{PP}/2$.

Suppose now Cell A is being written. For the selected row, $WL_0$ and $BL_0$ are equal to $V_{PP}/2$ and 0V respectively; for the unselected row, $WL_1$ is equal to 0V and $BL_1$ is floating. Cell B has already been written, so its GOAF element is in the resistive state. A high voltage pulse with amplitude $V_{PP}$ appears at the node $V_{ARRAY}$. Since the bottom plate of the MOSCap at Cell A has been pulled to ground, the thin gate-oxide is broken down by the pulse. For Cell B, since the resistance after breakdown is small, the high voltage pulse

Man Chiu, LEE
The Department of Electronic and Computer Engineering, HKUST
will appear at the drain of the blocking transistor. However, the thick gate-oxide will not suffer from high voltage stress because the gate voltage is at $V_{PP}/2$ and the voltage across the oxide is also $V_{PP}/2$. The drain voltage of the word-line select transistor in Cell A is limited by $V_M - V_T$ because this node is the source of the blocking transistor. Thus, the select transistor is protected. As for Cell C and D, the word-line transistors are off, so there is no stress problem on the gate-oxides.

### 3.3 Pulse Write Operation

#### 3.3.1 Determination of Pulse Height

As mentioned above, write operation requires a pulse of height $V_{PP}$ and width $t_{PW}$. These are important parameters for the memory circuit design, so they should be determined carefully by with the help of measurement results in the cell level.

The pulse height $V_{PP}$ for programming has a lower bound $V_{PP,MIN}$ and an upper bound $V_{PP,MAX}$. The lower bound is due to the largest voltage required to breakdown the gate-oxide. The upper bound is a limit related to the drain to substrate reverse diode. Figure 3.5
Figure 3.4: A 2x2 GOAF OTP Memory Array with NMOS Transistors and a Resistor Representing a Written Anti-fuse
3.3. Pulse Write Operation

shows the result of the DC voltage sweep tests. For all 6 samples, the breakdown voltages are lower than 6.5V. That implies \( V_{PP,MIN} \) have to be equal to 6.5V if 100% programming yield is needed. In order to confirm this is the case, tests on 30 samples were performed and the distribution of breakdown voltages are shown in Figure 3.6. Again, the result verifies that at 6.5V, all the GOAF elements will be broken down.

![Figure 3.5: Measured Voltage Sweep Programming of NMOS Gate-oxide Anti-fuses](image)

Figure 3.7 shows the schematic of a programmed Cascode memory cell. Notice that there is a diode formed by the n+ drain and the p substrate. If the post-breakdown resistance \( R_{OX} \) is small, during programming of other cells, \( V_{PP} \) will appear at the drain node \( V_D \) and the reverse diode current will increase. According to the measured result in Figure 3.8, the reverse current becomes larger than 1nA at around 9V. That will increase the loading current of the charge pump circuit for high voltage generation. If that happens at many programmed cells, the \( V_{PP} \) value will drop due to the aggregate of these leakage currents. Therefore, as a rule of thumb, the programming pulse height \( V_{PP} \) has to be smaller than 9V.

Now, \( V_{PP,MIN} \) and \( V_{PP,MAX} \) are determined, so the next step is to determine the exact value of \( V_{PP} \) that will be used in the memory circuit design. As shown in the circuit model
Chapter 3. Design of OTP Memory Cells

Figure 3.6: Distribution of Breakdown Voltages of 30 NMOS GOAF Elements

Figure 3.7: A Programmed Cascode GOAF Memory with Drain to Substrate Diode Shown
3.3. Pulse Write Operation

3.3.1 Determination of Pulse Height

In Figure 3.9, not the whole $V_{PP}$ is dropped across the GOAF element. The reason is there is some leakage current flowing through the memory array and this can be modeled by a resistor $R_{ARRAY}$. Even though this problem can be alleviated by increasing the size of the switching and pull-down transistors, the voltage across the GOAF element can never reach $V_{PP}$. Since breakdown is more efficient under higher voltages, a $V_{PP}$ value of 7.8V is chosen as the pulse height. This will give a margin of 20% over $V_{PP,MIN}$ and at this voltage level, the reverse diode current mentioned above will not be significant.

3.3.2 Determination of Pulse Width

To determine a suitable pulse width for write operation in the circuit level, it is necessary to carry out tests in the cell level to obtain data in advance. The programming conditions are $V_{PP}=7.8V$ and $I_{LIMIT}=20\mu A$.

At first, the approximate order of magnitude for the pulse width was found by trial pulse write operations on a few anti-fuse elements. As shown in Figure 3.10 is a group of
output currents after applying various pulse widths: 10µs, 100µs, and 1ms. Only when the pulse width lies between 100µs and 1ms can the GOAF element be broken. After that, the searching of pulse width was performed under 200µs, 300µs, 400µs, and 500µs, each with 16 samples from 4x4 arrays constructed with the same NMOS GOAF cells. It was found that for the first 3 cases, successful breakdowns happened more and more frequently and until 500µs, all the 16 samples were broken.

3.4 Sensing of Memory Contents

3.4.1 On and Off Currents

Figure 3.11 plots the On/Off currents for a GOAF memory element. The ratio between the on and off currents at a nominal supply voltage of 1.8V is more than 8 orders of magnitude. This is another advantage for this type of memory except for the low breakdown current.

To distinguish between a “0” and a “1”, it is necessary to compare the cell current $I_{\text{CELL}}$
3.4 Sensing of Memory Contents

with a reference current \( I_{REF} \). \( I_{REF} \) should be lower than the smallest read current from a “1” cell and higher than the largest read current from a “0” cell. Therefore, the reference current is bounded from above by the cell current when the post-breakdown resistance is the largest. Similarly, the reference current is bounded from below by the cell current when the pre-breakdown resistance is the smallest. In the next part, the upper limit will be found by measuring the worst-case post-breakdown resistance. However, the lower limit is not that important since the Off-current is always much smaller than 1nA and it is not easy to set such a small reference current.

3.4.2 Post-breakdown Resistance

Figure 3.12 plots the cumulative distribution of post-breakdown resistance for 20 GOAF elements when \( V_{DD} \) equals 1.8V. All samples have resistance lower than 15KΩ, so it’s safe to assume that 100KΩ is the worst-case post-breakdown. This value of \( R_{OX} \) will be used in the next part to find out the upper limit for \( I_{REF} \) using simulation.

*Device Characterization Laboratory*

*Man Chiu, LEE*
Figure 3.11: Comparison of On/Off Currents for a GOAF Memory

Figure 3.12: Cumulative Distribution of Post-breakdown Resistance for 20 GOAF Elements at $V_{DD} = 1.8V$
3.4.3 Sensing with Driving Devices

With the post-breakdown resistance of a GOAF element, it is time to include the effect when the driving devices are added. Figure 3.13 shows the simulated drain currents through different post-breakdown resistances $R_{OX}$ when $V_{ARRAY}$ is at sweeping mode. This is equivalent to the $I_D-V_D$ characteristic of a NMOS transistor with resistive load. The transistor sizes of the driving devices are minimum in this case. The gate voltages $V_{WORD}$ and $V_M$ of driving devices are both 1.8V and $V_{ARRAY}$ sweeps from 0V to 1.8V. The arrows on the plots indicate the range of feasible reference currents. A cell current of 15$\mu$A at 1.8V corresponds to the worst-case post-breakdown resistance 100K$\Omega$. Thus, $I_{REF}$ is required to be smaller than 15$\mu$A if a “1” is not to be misinterpreted as a “0”.

Another concern when determining the reference current is the power consumption for read operation. A very small $I_{REF}$ can be used since a stable band-gap reference circuit exists in the system only if it’s smaller than the upper limit mentioned above. Specifically, $I_{REF}$ equals 26nA in the GOAF OTP memory for the example application - the passive RFID tag.

3.5 Summary

The operation principles of read and write were explained in this chapter. The oxide reliability problem in write operation has been solved by the use of Cascode technique in the memory cell level. Design parameters including write pulse height, write pulse width and reference current were determined with the help of experiments.
Figure 3.13: Cell Currents in Log and Linear Scales under Various Post-breakdown Resistances for Minimum-size Driving Devices
After selecting and designing the suitable memory cell, the next problem to solve is the formation of the whole memory array with peripheral circuitry, which includes the decoding and level-shifting module, the high voltage writing module, and the sensing module. Writing of memory cells requires high voltage pulses to be transferred from the on-chip charge pump circuit to the memory core. In this chapter, the configuration of the memory core with the high voltage tolerant circuits and the sensing path will be shown. Reliable read and write operations will also be described in detail. Throughout this chapter, a 128-bit OTP memory used in passive RFID tags will be presented as an example to verify the functionalities and performance of the design. Design guidelines on how to maintain reliability and low power consumption will be provided as important results of this study.
4.1 Overview of the OTP Memory

The memory designed consists of a 128-bit memory core arranged in a 16x8 array. Figure 4.1 depicts the simplified system block diagram for the GOAF memory array. It mainly consists of a memory core, decoders and level-shifters, a high voltage multiplexer and sensing path circuits. There are row and column decoders to select individual row and column, so the memory can achieve bit-by-bit read and write operations.

It is important to note that the peripheral circuits are divided into two groups: the high voltage (HV) section and the low voltage (LV) sections. To maintain oxide reliability of all MOSFETs other than the GOAF element, thick-oxide devices are used in the whole HV section. The voltage across every oxide is limited to 4V in this part. The LV sections include decoders, level-shifters, and sensing path. To interface with the low power digital control circuits outside the memory, the input signals are designed to have amplitude of 0.6V. However, signals arriving at the memory core and the high voltage module have to be higher than 3.3V, which is the nominal supply voltage for thick-oxide devices. Therefore, the last stage of the level-shifting module is also required to be implemented with thick-oxide devices.

To lower the power consumption and at the same time reduce the silicon area, memory cells are written and read in a serial manner. The high voltage pulse for write operation and the constant read voltage for read operation are transferred from the external voltage supplies through the high voltage multiplexer (HVM). In the write mode, a memory cell is selected by the decoding circuits, then high voltage pulse stresses and breaks the anti-fuse element due to the high electric field across the gate-oxide. Similarly, in the read mode, cells are chosen one-by-one because of the continuously changing address. Each time when significant amount of cell current passes through a cell, it will flow all the way down the sensing path to the sense amplifier. If the anti-fuse is correctly written, it will output a high pulse instead of the zero voltage level for the case of reading a virgin cell.
4.1. Overview of the OTP Memory

Figure 4.1: System Block Diagram of the GOAF Memory Array
4.2 Memory Array, Decoding and Level-shifting

As mentioned in the previous chapter, the thick-oxide blocking transistor is used to improve reliability of the select transistor. The high voltage reliability problem is not unique to the cell level design. Therefore, when designing memory core and high voltage switching circuit, all nodes with possibility of high voltage stress must be taken care of.

Figure 4.2 illustrates a 16x8 GOAF memory array constructed by NMOS GOAF cells. All the gate terminals of the NMOS transistors are connected to the same node, where the high voltage pulse and the constant read voltage will reside during the write operation and the read operation, respectively.

Figure 4.2: A 16x8 NMOS GOAF Memory Array
Decoding for a low bit-count memory can be implemented by an AND structure without sacrificing much speed. Since the memory has to interface with low voltage digital base-band control circuits, the input signals are shifted from 0.6V to 3.9V in 3 stages, as shown in Figure 4.3. The other two levels in the middle are 1V and 1.8V. Each level-shifting stage consists of a simple level-shifter shown in Figure 4.4. For level-shifter design, the NMOS transistors MN1 and MN2 have to be large enough to over-drive the positive feedback loop formed by the PMOS transistors MP1 and MP2. However, when the ratio between the 2 voltages attains a certain level, it requires huge NMOS transistors, which are not suitable for memory design because of the pitch matching consideration and the over-sized NMOS transistors will slow down switching due to increased gate capacitance. These are the reasons for the need of multi-stage level-shifting.

![Figure 4.3: Decoding and 3-stage Level-shifting](image)

### 4.3 High Voltage Tolerant Circuit

High voltage switching modules, which transfer the voltages for writing memory cells, are essential parts inside a non-volatile memory using high voltage supplies generated from on-chip charge pump circuits. As mentioned previously, high voltage stress across a thick-oxide I/O device in a standard CMOS process is the major concern in the design of a reliable OTP memory. In conventional non-volatile memories, since the processes used are not standard CMOS, there can be high voltage options to implement these switches [20] and the structure of these switches is actually the level-shifter shown in Figure 4.4.

In this section, a fully standard CMOS high voltage multiplexer will be presented. Not
only does this topology requires no extra mask other than CMOS, it also consumes zero DC current.

In lieu of using high voltage devices and resistors, a fully CMOS high voltage switch (HVS) is used [13], which is shown in Figure 4.6. The cross-coupled design forms a feedback loop which prevents direct connection between $V_{PP}$ and ground, thus minimizes power consumption. The top PMOS MP3 switches the output to $V_{PP}$, while the bottom NMOS MN3 switches the output to 0V. The devices in between are used to protect the drains of both MP3 and MN3 from absorbing 0V and 7.8V directly. When the input signal WRITE is 0V, the voltage at the gate of MP3 should not be lower than 3.9V to avoid over-stressing the gate-oxide. In a conventional cross-coupled topology, a 0V at the output $VOUT−$ will appear at the gate of MP3 and the condition above is violated. To solve this problem, two pairs of PMOS transistors are adopted to form the level shifting structure. Each of them is called a half $V_{PP}$ shifter as shown in Figure 4.5. Since the gate of MF2 is at $V_M$, it passes
the 7.8V from the $V_{OUT}^{-}$ to the gate of MP3. When the $V_{OUT}^{-}$ is at 0V, MF2 is turned off and MF1 is turned on, switching its source voltage of 3.9V to the gate of MP3.

The simulation result in Cadence environment is shown in Figure 4.7 which shows the correct function of the circuit. Essentially, the circuit shifts a 0V - 3.9V pulse ($WRITE$) to a 0V - 7.8V pulse ($V_{OUT}^{+}$) and at the same time provides a 7.8V - 3.9V pulse output ($HV_{OUT}$). The latter is used to implement the high voltage multiplexer, which has been shown in the system block diagram already (Figure 4.1).

![Figure 4.5: The Half $V_{PP}$ Shifter with Input and Output Signals](image)

The high voltage multiplexer is needed for this design because the high voltage switch only transfers 0V - 7.8V pulses to the memory cells during write operations. It is crucial that the read voltage is developed on the gate oxides of the cells after a read enable signal has been sent from the digital control circuits. Therefore, essentially, a high voltage multiplexer which is able to transfer 0V, $V_{DD}$ and $V_{PP}$ has to be implemented. Figure 4.8 illustrate one way to achieve this requirement. Using the same high voltage protection strategy, the stacking of transistors avoids the voltage across the oxide from exceeding $V_{PP}/2$ [21].

When $WRITE$ is 0V, the internal signal $HV_{OUT}$ is equal to 7.8V, if at this moment the signal $READ$ is also 0V, the output node $V_{ARRAY}$ will be at ground level because of the signal $WRITE\&READ$; if the $READ$ is $V_{DD}$, then $V_{ARRAY}$ will be equal to $V_{DD}$, which is
Figure 4.6: The Zero Static Current High Voltage Switch

Figure 4.7: Simulation Result of the Zero Static Current High Voltage Switch
4.4 Sensing Path

Sensing of memory cell content is mainly performed by current comparison. Figure 4.9 shows the schematic of the complete sensing path, starting from the read voltage $V_{DD1}$ to the output of the sense amplifier or comparator.

![The High Voltage Multiplexer](image)

**Figure 4.8: The High Voltage Multiplexer**
Here, a written cell is defined to be a “1” and otherwise it’s a “0”. They are represented by $R_{CELL}$ or $C_{CELL}$ in the circuit. Due to the existence of 127 un-selected cells, a resistor $R_{ARRAY}$ is used to model the leakage current in those cells and a capacitor $C_{ARRAY}$ is used to model the total capacitance of all other cells.

Since a written cell contains a resistive anti-fuse, when the cell is selected, a significant amount of current will flow through the memory cell to the input of the comparator. Referring to Figure 4.10, if $I_{UP}$ is significantly larger than $I_{DOWN}$, the $V_{COMP}$ node will be charged up in a certain length of time, which depends on the resistance of the anti-fuse. When the voltage of the positive input to the comparator exceed the reference voltage $V_{REF}$, the comparator output will change from 0V to $V_{DD2}$. Note here that $I_{UP}$ and $I_{DOWN}$ are actually the cell current $I_{CELL}$ and the reference current $I_{REF}$, respectively. The size ratio of M6 to M7 is 1:1, so $I_{BIAS1}$ is equal to the reference current, assuming the current mirror has perfect matching.

$I_{BIAS1}$ is scaled up by 18 times with another current mirror. This branch will form the discharging path whenever the clock signal is high. Constant current discharging is used here to keep the discharging current stable because $I_{BIAS1}$ is generated from a band-gap reference, which rejects variations resulting from change of temperature and supply noise.

Note also that the biasing current $I_{BIAS2}$ for the comparator will be cut off if the memory is not in read mode, so the memory consumes less power during non-read mode.

### 4.5 Write Operation

Simulation result of continuously applying 3 write pulses to the $V_{ARRAY}$ node is shown in Figure 4.11. The write pulses are 7.8V in height and 500µs in width. Since the true breakdown process won’t happen during simulation, the high voltage pulses at $V_{ARRAY}$ keep constant during the write period.

The measured write operation on 3 memory cells is shown in Figure 4.12. The most important phenomenon in this test is the sudden drop of voltage at the $V_{ARRAY}$ node. During write operation on a memory cell, a 7.8V pulse of 500µs is applied on this node. The designed
4.5. Write Operation

Figure 4.9: The Complete Sensing Path for Memory Read Operation with Constant Current Discharging
pulse width $t_{PW}$ is controlled by the WRITE signal. Note that for convenience, signal amplitude of 1V were used instead of 0.6V.

If the memory cell is broken down, its resistance will drop from more than 1GΩ to less than 100KΩ. This results in a sudden drop of $V_{ARRAY}$ from 7.8V to much lower voltage, depending on the actual resistance on the Cascode memory cell. The time from the rising edge of the pulse to the sudden drop is defined as the time-to-breakdown $t_{BD}$.

It is not easy to guarantee 100% success rate with only one high voltage pulse applied on a memory cell during write operation. In order to increase the yield of the write operation, it is necessary to create a feedback loop when writing a cell. After each write pulse, when $V_{ARRAY}$ drops to the ground, a read operation on this particular cell can be used to test whether the cell has been successfully written. The signal $Data_{out}$ is sent back to the control circuit and this procedure determines if another write pulse is required to be sent to the array. A counter will be used to keep track of the total number of write pulses sent and when it reach
4.5. Write Operation

Figure 4.11: Simulation Result for Write Operation of the Whole Memory Circuit 3 Pulses

Figure 4.12: Measurement Result for Write Operation of the Whole Memory Circuit with On-chip Charge Pump, Voltage Supplies, and References - 3 Cells
a preset maximum, the write operation on the memory cell will be terminated, and the cell is recorded as a malfunctioning cell.

Figure 4.13 illustrates a distribution of $t_{BD}$ for 100 memory cells. 97 of the cells were written successfully after the first round of programming. The remaining 3 were written with second write pulses.

![Figure 4.13: Measured Time-to-breakdown Distribution with 100 Bits of NMOS GOAF OTP](image)

**4.6 Read Operation**

Figure 4.14 shows the simulation result of reading all the 128 bits consecutively. It is assumed that half of the memory cells are written and thus these GOAF elements are replaced resistors of 1MΩ, which is larger than the measured post-breakdown resistance in the worst-case. The delay from the change of address to the start of the output pulse is around 500ns.

Figure 4.15 is the measurement result corresponding to the above simulation. The read delay is around 600ns and the difference from the simulation result is mainly due to probe
4.7 Interface of OTP Memory inside RFID Tag

As shown in Figure 4.16, A0-A6 are the address signals from the Base-band; READ, WRITE and \( V_{CLK} \) are control signals from the Base-band; Data_Out is the output to the Base-band; \( V_{PP}, V_M \) and \( V_{WORD} \) are generated by the charge pumps in the Power Management Unit (PMU); \( V_{REF}, I_{BIAS1} \) and \( I_{BIAS2} \) are provided by the reference circuits in the PMU; \( V_{DD1}, I_{DD2} \) and \( I_{DD3} \) are voltage supplied by the low-dropout regulators (LDR) in the PMU.

4.8 Summary

In this chapter, the proposed memory architecture has been described in detail. The separation of high voltage and low voltage sections ensures gate-oxide reliability of all devices.
Figure 4.15: Measurement Result for Read Operation of the Whole Memory Circuit with On-chip Charge Pump, Voltage Supplies, and References

Figure 4.16: Interfacing of the OTP Memory with On-chip Power Management Unit and Digital Baseband
except those GOAF elements during write operations. The cascode technique was used in the high voltage section to protect the gate-oxide wherever $V_{PP}$ exists. The sensing path of the memory was also demonstrated. The write and read operation were verified by measurements with internal supplies and references generated by on-chip power management circuits.

The rectangle in Figure 4.17 highlights the chip micrograph of the 128-bit DNW NMOS GOAF memory in a complete passive RFID tag. The memory occupies an area of $0.2\,\text{mm} \times 0.5\,\text{mm} = 0.1\,\text{mm}^2$ or around 20% of the total RFID tag area.

**Figure 4.17:** Chip Micrograph of the 128-bit OTP Memory in a RFID Tag
Conclusions

In this chapter, a summary of the works done in this thesis is given. After that, some future research directions are described for the improvement of gate-oxide anti-fuse memories. Specifically, the yield issue and design for other applications will be discussed. Some additional information is provided at the end.
5.1 Summary of Works Done

In this work, detailed study of anti-fuse memories in standard CMOS processes has been presented. The use of standard CMOS provides a low-cost solution for many applications. Design guidelines on the selection of memory elements, the memory cell formation, and the construction of complete memory arrays with all the peripheral circuits were demonstrated.

A few low-cost anti-fuse memories found in a standard CMOS process were discussed. These include the diode anti-fuse memories and the gate-oxide anti-fuse (GOAF) memory. The performance of each type of memory was analyzed with the measurement results. By making comparisons, the NMOS GOAF was proved to be the most suitable memory for low-power embedded applications.

After that, the development of the memory cell was presented. Based on a Cascode GOAF OTP memory cell structure, various 3-transistor cells were fabricated and discussed in detail. Design parameters including the write pulse height, write pulse width and reference current magnitude were determined by analyzing measurement results.

A GOAF OTP memory array for RFID application was also discussed together with the measurement results. This part has covered the arrangement of the memory array, the level shifting circuit, the high voltage tolerant circuits built with the Cascode technique, and the sensing path with power consumption awareness. With the use of serial operation, the silicon area was significantly reduce while high voltage reliability of the whole array was ensured.

Table 5.1: A Table Summarizing Various Parameters for the Embedded NMOS GOAF OTP Memory

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell Size</td>
<td>7 $\mu^2$</td>
<td>Write Current Limit</td>
<td>20 $\mu$A</td>
</tr>
<tr>
<td>Bit-count</td>
<td>128</td>
<td>Write Period</td>
<td>500 $\mu$s</td>
</tr>
<tr>
<td>Array Size</td>
<td>0.1 mm$^2$</td>
<td>Read Power</td>
<td>4 $\mu$W</td>
</tr>
<tr>
<td>Write Pulse Height</td>
<td>7.8 V</td>
<td>Read Period</td>
<td>1.6 $\mu$s</td>
</tr>
</tbody>
</table>
5.2 Future Works

In Chapter 5, it has been shown that read operation is successful with half of the memory cells written. However, this was performed when the read voltage from the top of the sensing path was fixed. To obtain more information on the yield of the memory circuit, it will be useful if read operation is measured under various read voltages.

In this work, only passive RFID tag was used as the example application. The OTP memory presented is, however, suitable for many different applications. As a result, it will be interesting if this memory is redesigned under other constraints. One other application is the addition of a small number of bits of data to an analog amplifier or other precision analog chip. Less than 100 bits of data are added to the analog amplifier or precision analog chip to allow it to be digitally tuned. Traditionally, analog trimming is done by laser cutting of metal or poly-silicon links, opening poly fuses with a current, and Zener diode zapping to adjust resistance, etc. Just like the existing non-volatile memories mentioned in Chapter 1, these methods incur extra processing cost. It will be helpful if the standard CMOS OTP memory developed here is used to replace those techniques.

5.3 Appendix

5.3.1 Pre-charge Sensing Scheme

Sensing of the memory can also be done in a pre-charge topology as shown in Figure 5.1. The comparator is placed at the top of the sensing path. The read voltage $V_{DD1}$ is used to pre-charge the $V_{ARRAY}$ node before evaluation of the memory content. If a cell is written, the cell current will discharge this node so that the voltage drops below $V_{DD1}$. When the node voltage drops below the reference voltage, output will become “1”. The reference voltage for the comparator input will be close to $V_{DD1}$ instead of the ground.

However, this topology requires isolation between the output of the high voltage multi-
plexer to the input of the comparator. In another word, the high voltage section and low voltage section are not separated.

5.3.2 Cell Layout

The aim of this work is not on the optimization of memory cell size, but two techniques on reducing the size is worth further discussion here. Firstly, sharing of contact every 2 row has been used in this work, as shown in Figure 5.2. Secondly, since only breakdown at source or drain of the GOAF is required, one of them can be eliminated in the layout and that reduces the area of the GOAF by 30%.
Figure 5.1: Chip Micrograph of the 128-bit OTP Memory in a RFID Tag
Figure 5.2: (a) Cell Layout with Sharing of Contact; (b) Elimination of Source or Drain
Bibliography


