Universita' degli Studi di Pavia

PhD. Thesis in Microelectronics

Integrated system for fast parametric tests and characterization of Phase Change Memory cells

Author:

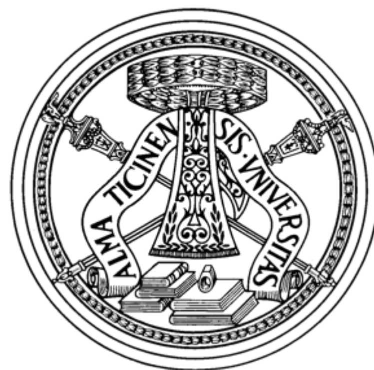
Erika Covi

Supervisor:

Prof. Guido Torelli

Co-supervisor:

Ing. Alessandro Cabrini



Academic Year 2012/2013

Contents

Introduction iv

1 Non-volatile memories 1
1.1 Non-volatile memories: an overview 1
1.2 Operating principle of Phase Change Memories 4
1.3 Programming algorithm 8
1.3.1 Partial-SET programming 9
1.3.2 Partial-RESET programming 10
1.4 Usual Automated Test Equipment 12

2 On-wafer pulse generator for Phase Change Memory cells 18
2.1 Target specifications and high-level description 18
2.2 First solution 22
2.2.1 Basic operating principle 22
2.2.2 System description 23
2.2.3 Accuracy of main parameters 26
2.3 Second solution 26
2.3.1 Basic operating principle 26
2.3.2 System description 27
2.3.3 Accuracy of main parameters 30
2.4 Third solution 31
2.4.1 Basic operating principle 31
2.4.2 System description 31
2.4.3 Accuracy of main parameters 34
2.5 Comparison of the conceived solutions 35

3 Analysis and test chip of the main blocks 37
3.1 Aim of the prototype 37
3.2 Circuit design 38
3.2.1 Output buffer 38
3.2.1.1 Preliminary version of the output buffer 39
3.2.1.2 Final version of the output buffer 43
3.2.2 Delay time processing 47
3.2.3 Pulse time duration generation 50

3.2.4 Pulse fall time . . . 53
3.3 Manual calibration procedure . . . 55
3.3.1 Calibration equations . . . 57
3.3.1.1 Calibration equation for delay time . . . 57
3.3.1.2 Calibration equation for time duration . . . 59
3.3.1.3 Calibration equation for fall time . . . 60
3.3.2 Calibration procedure . . . 62

4 Final implementation **64**
4.1 Introduction . . . 64
4.2 High-level description . . . 64
4.3 Basic operating principle . . . 66
4.3.1 Pulse generation . . . 66
4.3.2 Interface with the test equipment . . . 66
4.3.3 Automatic calibration procedure . . . 67
4.4 Description of the system . . . 67
4.4.1 Interface with the ATE . . . 68
4.4.2 Pulse generation . . . 69
4.5 Circuit design . . . 72
4.5.1 Features and aims . . . 72
4.5.2 Interface with the test equipment . . . 75
4.5.2.1 Enable signal generation . . . 75
Clock generation . . . 75
Synchronizer . . . 77
Monostable . . . 78
4.5.2.2 Current Track-and-Hold circuit . . . 78
4.5.3 Pulse generation . . . 81
4.5.3.1 Output buffer . . . 81
4.5.3.2 Delay time . . . 82
4.5.3.3 Time duration . . . 85
4.5.3.4 Fall time . . . 88
4.6 Calibration procedure . . . 90
4.6.1 Calibration equations for time duration and fall time . . . 91
4.6.2 Accuracy considerations . . . 93
4.7 Simplified version of the system for different test equipment . . . 95

5 Experimental results **98**
5.1 Preliminary buffer . . . 98
5.1.1 Simulations . . . 98
5.1.2 Measurements . . . 100
5.2 First fabricated test chip . . . 102
5.2.1 Simulation of the output buffer . . . 102
5.2.2 Measurements . . . 104
5.2.2.1 Output buffer . . . 104

5.2.2.2 Pulse Generator 107
5.2.2.3 Calibration 112
5.3 Final implementation 115
5.3.1 Simulations of the final implementation 115
5.3.1.1 Output buffer 115
5.3.2 Whole system 118
5.3.3 Experimental results of the final implementation 120
5.3.3.1 Measurements of the current Track-and-Hold circuit 122
5.3.4 Calibration procedure 123

6 Model for Phase Change Memories **125**
6.1 Introduction 125
6.2 Crystallization and amorphization kinetics 129
6.2.1 Crystallization kinetics 129
6.2.2 Amorphization kinetics 131
6.3 Model implementation 132
6.3.1 Calculation of the temperature at the heater-GST interface . 132
6.3.2 RESET modelling 133
6.3.2.1 Growth of the amorphous cap during a RESET operation 133
6.3.3 Analysis of the crystallization phenomenon 135
6.3.4 SET modelling 143
6.4 Model validation 145
6.4.1 RESET operation 147
6.4.2 SET operation 150

Conclusions **153**

Bibliography **155**

Introduction

In semiconductor market, memories are becoming more and more important thanks to the increasing number of portable devices that need a memory device (digital cameras, music players, smartphones, laptops, tablets,. . .) and to the introduction of concepts as Cloud computing, which is defined by the National Institute of Standards and Technology (NIST) as "a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction" [1].

The main challenge in this scenario is developing new memory technologies which keep on enhancing their performance while decreasing their energy consumption. In this respect, new kinds of non-volatile memories are being studied, such as Conductive Bridge RAMs (CBRAMs), Resistive RAMs (ReRAMs), Spin-Torque-Transfer RAMs (STT-RAMs), and Phase Change Memories (PCMs).

Among them, PCMs turned out to be the most promising technology, which may be able to replace current Flash technology [2]. Some application fields for PCMs are wireless systems, solid state storage system, and computing platforms, where the power consumption has to be reduced. However, Flash memory market is well established: to be competitive, PMCs should match the cost of the existing technologies (in terms of both cell size and process complexity), find applications in which they can optimize the overall memory system, and be extremely scalable.

Up to now, PCMs are being able to fulfil all the above requirements. However, continuous efforts must still be devoted to keep on the line and optimize materials, cell architecture, and programming algorithms, also aiming at production yield improvement. To this end, the investigation should be carried on following two directions: the development of a PCM cell model which can reliably simulate the cell behaviour under different programming conditions and the characterization of the cell at wafer level, so as to perform an accurate statistical evaluation of the

cell performance. In fact, the final cell state depends on parameters of the applied programming pulse(s) such as amplitude, duration, and fall time.

On the one hand, the conventional instrumentation available for on-wafer automatic parametric testing features high accuracy, but is mainly conceived for DC measurements of elementary devices. The speed of this instrumentation is therefore not sufficient to allow the fast parametric tests required to perform statistical studies of the I-V characteristic of PCM cells, with the aim of investigating the cell behaviour and improving both the cell geometry and the materials used. On the other hand, long cables are necessary to connect the selected cell(s) on the wafer to commercial pulse generators. Non negligible noise can therefore be added to the generated waveforms, thus degrading the signal and, hence, limiting both the accuracy and the controllability of program pulses. Moreover, the usual hybrid configuration (an Automatic Parametric Wafer Testing System, APWTS, and a Pulse Generator, PG, that cooperate to perform tests) uses a switch matrix to switch the connections of the pads of the memory chip under characterization to the different signals provided by the test equipment, thus decreasing testing speed due to switching and settling times. In fact, with the standard configuration, a write-and-read cycle can last from few hundreds of ms up to 1 s.

Moreover reading the programming current at the pulse plateau is fundamental to getting useful information about the cell programming performance. However, conventional ATE is not able to read the PCM cell programming current at pulse plateau, since reading time may vary between 1 ms and 10 ms, whereas the programming pulse duration is on the order of tens to hundreds of nanoseconds. An interface be able to sample and hold the programming current at pulse plateau and feed the test equipment with a replica of the sampled current should be integrated.

The performance and the flexibility of conventional Automated Test Equipment (ATE) should therefore be enhanced to carry out reliable tests on different cell implementations. Aiming at this goal, it is thus highly desirable to have a custom on-wafer system able to generate the required programming pulses with high accuracy and flexibility, controlled by commercial ATE, which is able to provide

voltage pulses with programmable amplitude, duration, and fall time. A key design requirement was to allocate the generator in the wafer scribe lanes, in order to reduce the impact of the proposed approach on testing cost. However, this requirement implies that the system must feature reduced silicon area occupation and have a very disadvantageous aspect ratio as well as limited pad count. Since unavoidable process spreads and non-idealities can affect the accuracy of the on-wafer pulse generator, a calibration procedure for pulse parameters should be conceived.

This Thesis is organized as follows. Chapter 1 provides an introduction on emerging memory technologies, focusing on PCMs, and on commercial ATE. Chapter 2 evaluates and compares different possible implementation of an on-chip pulse generator, controlled by usual commercial ATE, to perform fast parametric tests on PCM cells. The best solution was developed, fabricated and characterized. Two test-chip were fabricated. The first one, presented in Chapter 3, aims at the characterization and debug of the main blocks of the system. The second one focuses on the interfacing with the ATE, and it is discussed in Chapter 4, as well as a simplified version of the test chip conceived to be used with different test equipment. Experimental results of the test-chips are presented in Chapter 5. Finally, Chapter 6 deals with a model for PCM cells, together with an analysis of the crystallization process.

Chapter 1

Non-volatile memories

1.1 Non-volatile memories: an overview

In the last decades, memory products are gaining a significant portion of the semiconductor market. The two predominant technologies are the DRAM and the NAND Flash. The latter has become the leading technology since 2005, when the increasing amount of advanced mobile applications (e.g. smart-phones, tablets, digital cameras, mp3 players,...) started demanding high density, low power and cheap storage hardware. The scaling path of NAND technology is remarkable: a new technology node is introduced almost every year. Nowadays, the 2x nm technology has already been entered into and next nodes are under development [3].

Despite many efforts are made to face the challenges of NAND Flash scaling down [4], new memory technologies based on innovative materials are being investigated. Among them, there are Conductive Bridge RAMs (CBRAMs), Resistive RAMs (ReRAMs), Spin-Torque-Transfer RAMs (STT-RAMs), and Phase Change Memories (PCMs).

The information storage of CBRAMs (Fig. 1.1) takes place in a solid electrolyte. The memory effect relies on a polarity-dependent resistive switching induced by appropriate voltages across the two electrodes of the device. The low-resistance

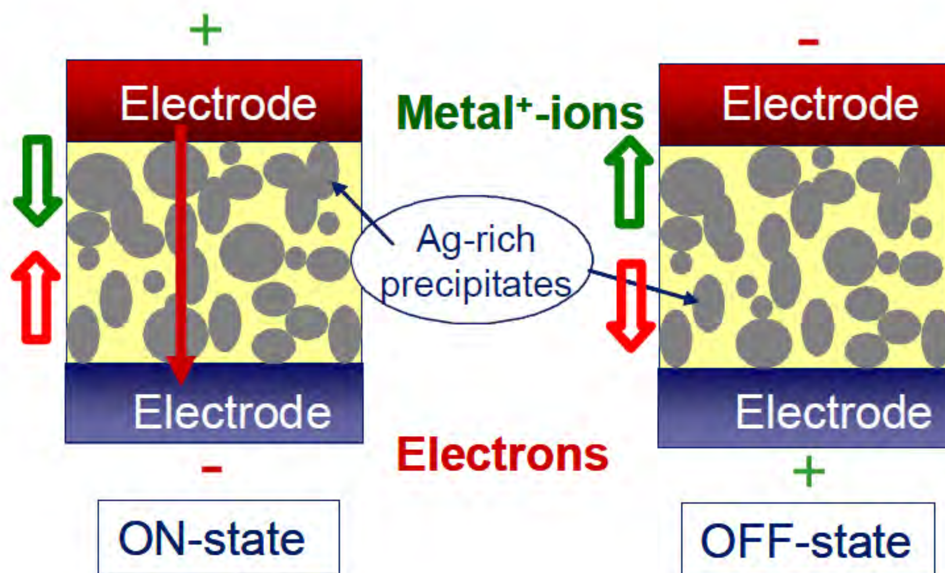


FIGURE 1.1: CBRAM operating principle [7].

state (ON-state) is obtained by applying a low positive voltage (typically about 250 mV) across the electrolyte, thus inducing the migration of metallic ions which leads to the formation of a conductive bridge between the electrodes. The high-resistance state (OFF-state) is obtained by applying a reverse bias voltage (typically about -80 mV) which removes the metal ions, thus reducing the conductive filament cross section [5], [6].

The operating principle of ReRAMs (Fig. 1.2) is based on a dielectric, typically insulated, which may be made conductive by generating a conductive filament. The formation of this filament can be induced by applying a sufficiently high voltage across the dielectric, thus generating a controlled current flow between the two terminals of the device. Several physical mechanisms cause the conductive path formation, including material defects and metal migration. The conductive filament can be broken (RESET state, higher resistance) or re-formed (SET state, lower resistance) by applying adequate voltage pulses [8], [9].

The storage element of STT-MRAM (Fig. 1.3) is a thin insulating tunnelling barrier placed in the middle of two ferromagnetic layers (Magnetic Tunnel Junction, MTJ) [11], [12]. One layer has a fixed magnetization orientation whereas the

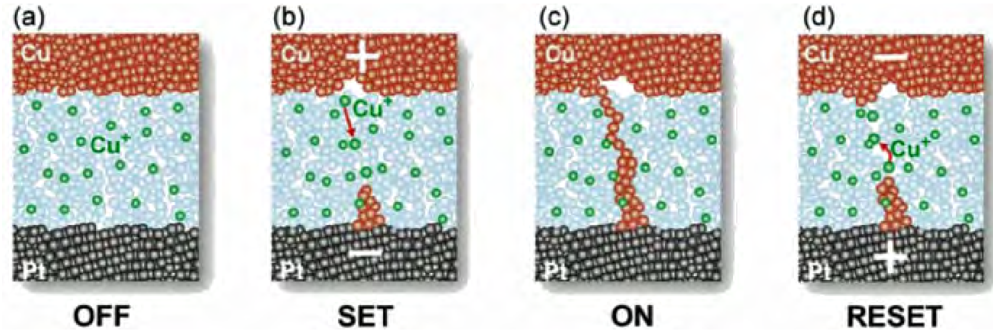


FIGURE 1.2: ReRAM operating principle [10].

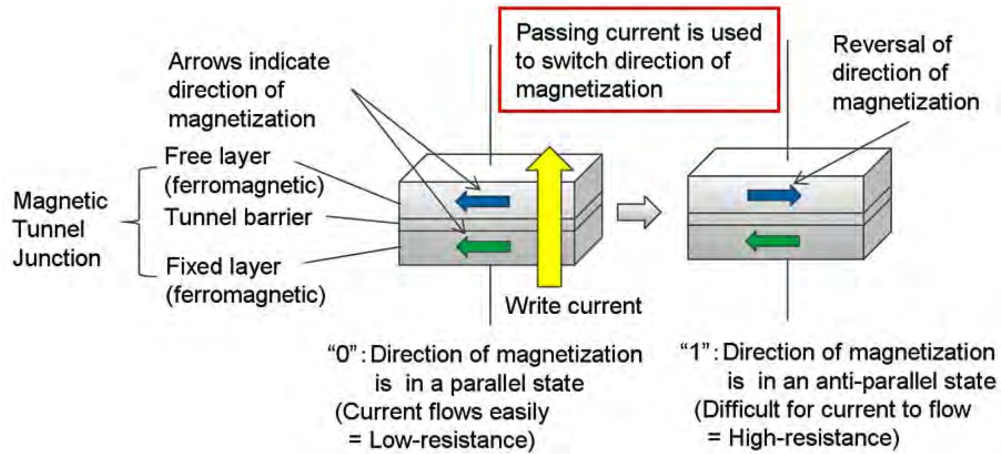


FIGURE 1.3: STT-RAM operating principle [14].

other is free to switch it. A current with a spin polarization transfers the angular momentum to the free layer, switching its magnetization orientation and thus changing the cell resistance. In fact, the resistance of a STT-MRAM cell depends on the relative magnetization orientation of the two ferromagnetic layers [13].

The information in a PCM cell is stored by exploiting two different structural solid-state phases (namely, the amorphous and the (poly-)crystalline phase) of a small portion of chalcogenide alloy, typically $Ge_2Sb_2Te_5$ (GST), which have different electrical resistivity. More specifically, the resistivity is higher (RESET state) for the amorphous phase and lower (SET state) for the crystalline phase. The typical PCM cell resistance in the SET state is a few kΩ, whereas in the RESET state it increases up to a few MΩ [5], [15]. The phase transition, which is very fast, is thermally induced by Joule effect and is controlled by the current flowing through the GST [16], [17].

1.2 Operating principle of Phase Change Memories

As explained above, the basic operating principle principle of a PCM cell relies on the physical properties of chalcogenide materials, which can switch between two phases (amorphous and crystalline) when stimulated by suitable electrical pulses. A PCM cell is composed of a thin chalcogenide layer, a resistive element named heater (typically made of TiN or W), and two metal electrodes, i.e., the top electrode contact (TEC) and the bottom electrode contact (BEC).

In PCM devices, Joule heating is fundamental to perform the phase transition, which occurs when the temperature inside the chalcogenide material reaches the proper value.

The crystalline-to-amorphous phase transition, also known as RESET operation, is obtained by bringing the temperature inside the GST above the melting point T_{melt} (about 600 $°C$ for the GST alloy) [18] by applying an electrical pulse with high amplitude and short duration (few tens of ns are typically sufficient [19]) to the cell, and then rapidly cooling it down (fast fall time). This way, the GST is frozen into a disordered (i.e. amorphous) structure.

The amorphous-to-crystalline phase transition, also known as SET operation, can be obtained with two different kinds of pulses. The first one has a lower amplitude and a longer time duration than in the RESET operation. The GST is so heated to a temperature below the melting point but above the crystallization temperature T_{cryst} (depending on the alloy, T_{cryst} can range from 150 $°C$ [20] to 200 $°C$ [21]), that is the minimum temperature required to activate the crystallization process. The thermal energy provided enables to restore the crystalline lattice, which is an ordered structure and therefore it is a minimum-energy structural configuration. Another way to perform a SET operation is to bring the chalcogenide alloy above its melting point with a high-amplitude pulse, then to cool it down slowly (slow fall time) so that the cell remains at a temperature between the melting and the

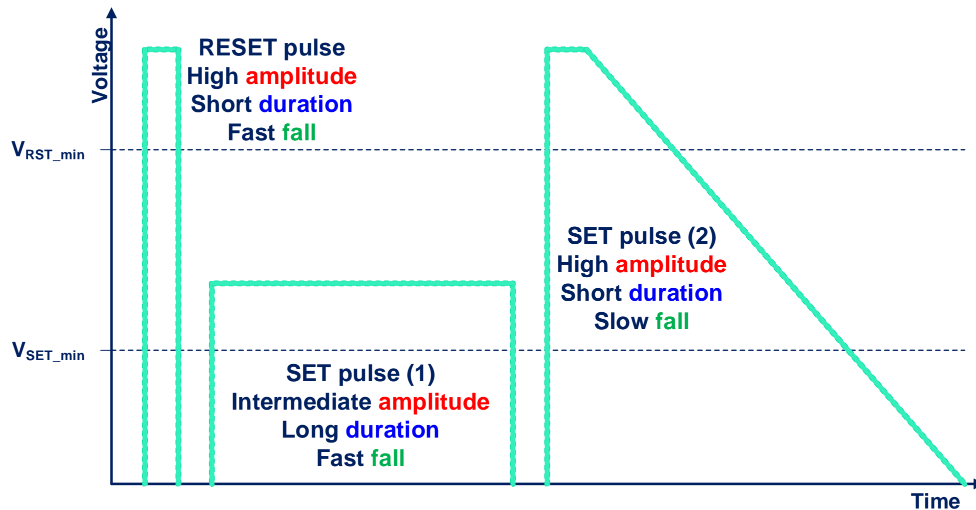


FIGURE 1.4: Typical voltage pulses for SET and RESET operations.

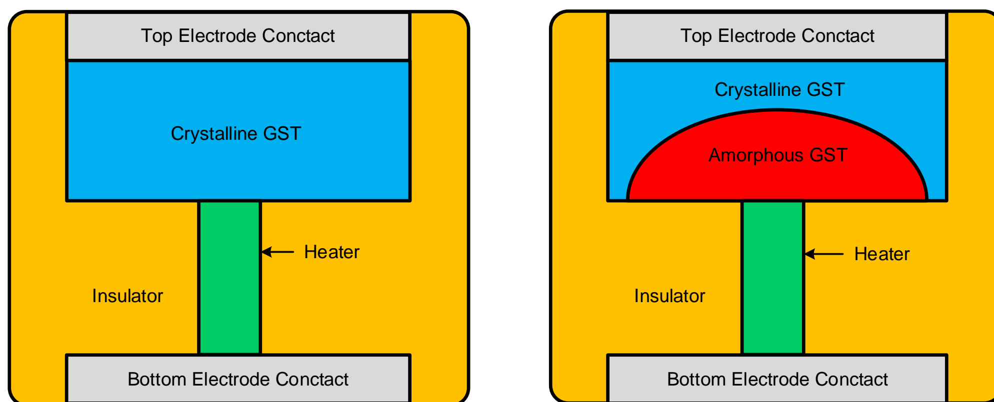


FIGURE 1.5: Conceptual scheme of a PCM cell.

crystallization ($T_{cryst} < T < T_{melt}$) for a time sufficient for the GST to re-dispose its atoms in a lattice.

Typical voltage pulses for SET and RESET operations are shown in Fig. 1.4.

The conceptual scheme of a PCM cell is shown in Fig. 1.5, from which it is apparent that only a portion of the chalcogenide layer, located at the interface between GST and heater, undergoes phase transition. The phase state of this portion, referred to as active chalcogenide, determines the value of the memory cell resistance.

The typical I-V characteristics of the PCM cell in the SET and RESET states are shown in Fig. 1.6. When the cell is in its full-SET state, the resistance of the cell decreases with the increase of the applied voltage. An S-shaped behaviour can

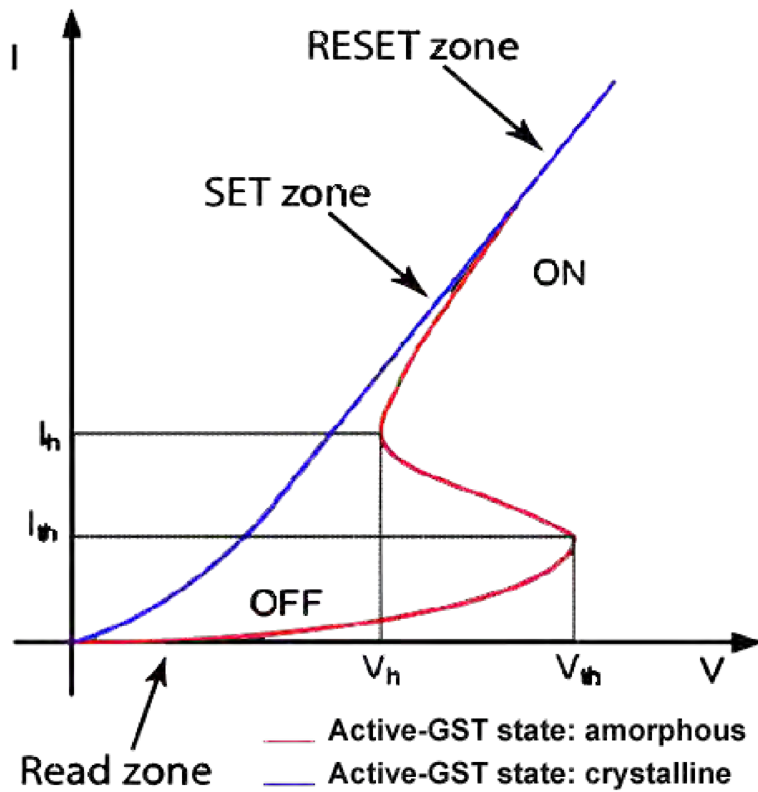


FIGURE 1.6: I-V characteristics of the PCM cell in the SET and RESET states.

be observed in the I-V curve of the cell in its RESET state. This effect is due to a phenomenon named threshold switching [22], [23], [24], [25]. This phenomenon consists in a sudden drop of the amorphous chalcogenide resistivity as the voltage (current) across the PCM cell exceeds a critical value, V_{th} (corresponding to a current I_{th}). This resistance drop allows providing the electrical power required to achieve the desired state transition, which would be otherwise impossible.

On the one hand, when low-amplitude electric pulses are applied to the cell in its high-resistance state (OFF region in Fig. 1.6), a low current flows through the device. On the other hand, when a pulse with higher voltage amplitude ($V_{pulse} > V_{th}$) is applied to the cell, threshold switching takes place and the device shows a much lower resistance (ON region in Fig. 1.6). For the feasibility of PCM technology, threshold switching phenomenon is fundamental. In fact, without threshold switching, the power to be delivered to the device in order to activate phase transition would require very high voltage pulses. This way, despite the very high

resistance of the cell in the OFF region, the threshold switching makes the programming operation possible by requiring only few Volts for programming the cell.

In the ON region, the I-V curves of the cell in SET and RESET state are superimposed, whereas in the OFF region they are very different. Therefore, the programming operation takes place when operating in the ON region, so as to provide the device with an energy sufficient to induce phase change. Indeed, reading operation takes place while the cell is in the OFF region, so as not to change the phase state while reading the cell state: a predetermined read voltage, low enough to prevent any unwanted change in the alloy phase, is usually applied to the cell and the current flowing through the device, named read current, is sensed.

The range from the minimum (RESET) and the maximum (SET) read current is called read window and it is considerably wide, which allows both safe storage of an information bit in the cell and the possibility to explore the multilevel (ML) approach, which is intended to achieve low-cost high-density storage [26].

However, some critical issues must be taken into account when implementing ML programming. Among them, there is the capability to program and read the memory cells with an accuracy sufficiently high, the reproducibility of the programming operation, and the stability of programmed levels. Moreover, ML programming algorithm must be robust with respect to variations of parameters of the memory cell and the surrounding circuitry in order to be effective.

Each memory cell consists of a PCM storage element connected to a selection transistor, which can be either a bipolar device as depicted in Fig. 1.7, or an MOS transistor. The base or the gate of all select transistors of the same row are connected to the same Word Line, whereas the TECs of the PCM cells belonging to the same column are connected to the same Bit Line. The memory cells can be easily addressed by means of a decoding network.

The decoding approach changes depending on the kind of memory cell to be characterized [27], [28]. Basically, cell selection is performed by applying proper voltage

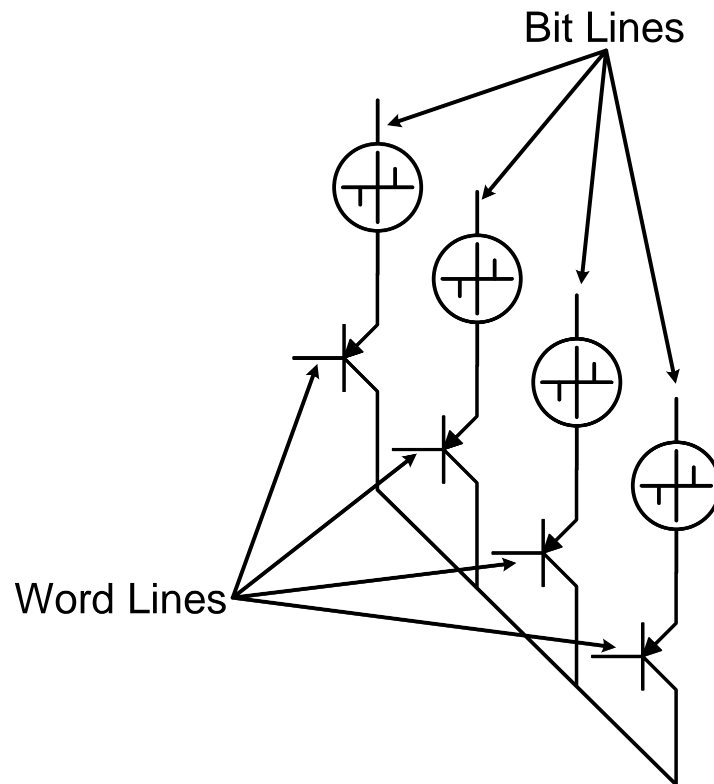


Figure 1.7: Word Lines and Bit Lines in a portion of array of PCM cells with a bipolar selector.

levels to the Word Line and the Bit Line of the desired memory cell and different voltage levels to the remaining Word Lines and Bit Lines; in some cases unselected Bit Lines may also be floating.

1.3 Programming algorithm

In the literature, several programming algorithms are proposed, especially for multilevel-cell programming. They can be divided into two macro-categories: partial-SET and partial-RESET programming.

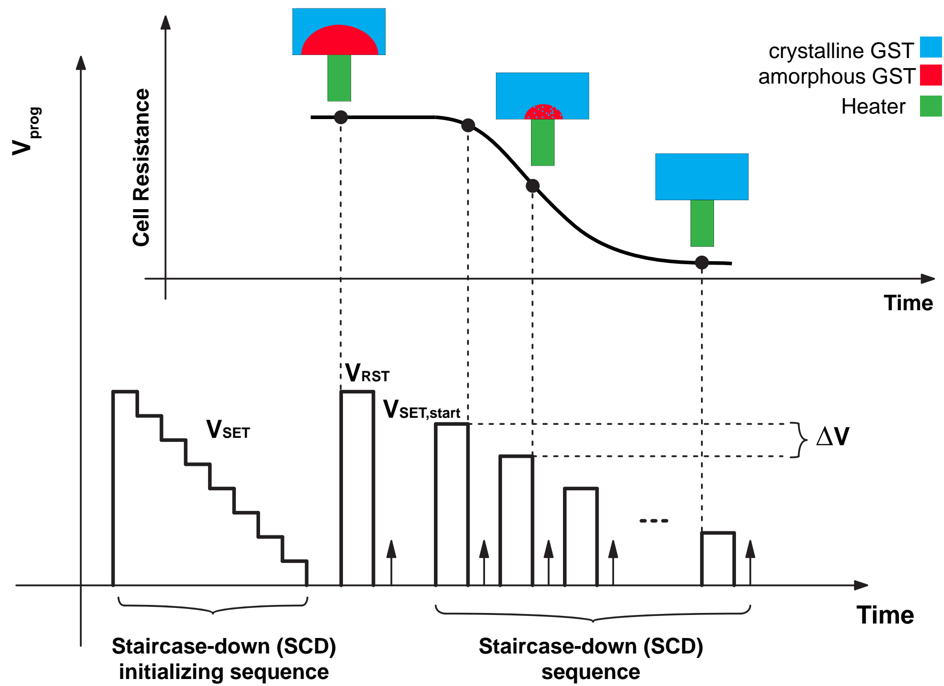


FIGURE 1.8: Sequence of program and read pulses for the partial-SET staircase down programming algorithm.

1.3.1 Partial-SET programming

In partial-SET the two main algorithms are staircase down and single pulse programming.

The staircase-down algorithm is made of a RESET initializing pulse followed by a sequence of pulses having the same duration and decreasing amplitudes (Fig. 1.8).

This algorithm shows the cumulative nature of the SET process: even though a single pulse of the sequence applied to a cell does not make it switch from the RESET to the SET state, the whole sequence of the SET pulses can make it change state and/or crystallize it better, bringing it to a resistance state lower than the state reachable by the cell by applying a single programming pulse.

In order to extensively study the crystallization kinetics of PCM cells, the single pulse programming algorithm is preferable. In fact, the basic idea is similar to the partial-RESET single pulse programming algorithm. At the beginning of

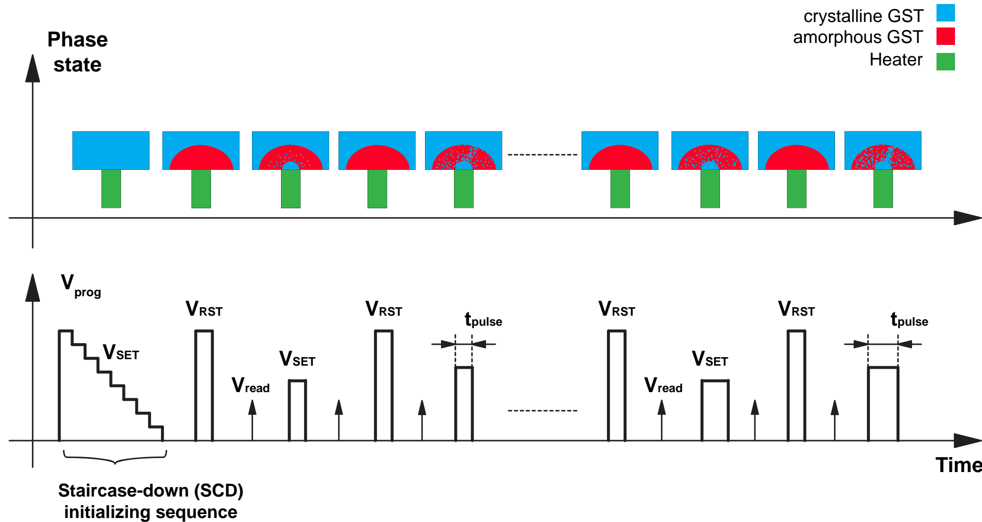


FIGURE 1.9: Sequence of program and read pulses for the partial-SET single pulse programming algorithm.

the sequence, the cell is initialized with a staircase-down sequence, then a RESET pulse is applied, followed by a SET pulse (Fig. 1.9). The RESET pulse - SET pulse sequence should be repeated with SET pulses having different duration and/or amplitude, so as to characterize the cell under a wide range of different programming conditions.

1.3.2 Partial-RESET programming

As for partial-SET programming, also in partial-RESET programming the two main algorithms are the single pulse and the stair case up.

In single pulse algorithm, only one rectangular programming pulse is given to the cell. Consequently, to perform a PCM cell characterization, the algorithm to be used is the one described in Fig. 1.10.

After having initialized the memory cell to its full-SET state by means of a staircase-down initializing sequence, a RESET pulse is fed to the cell, then the state of the cell is read.

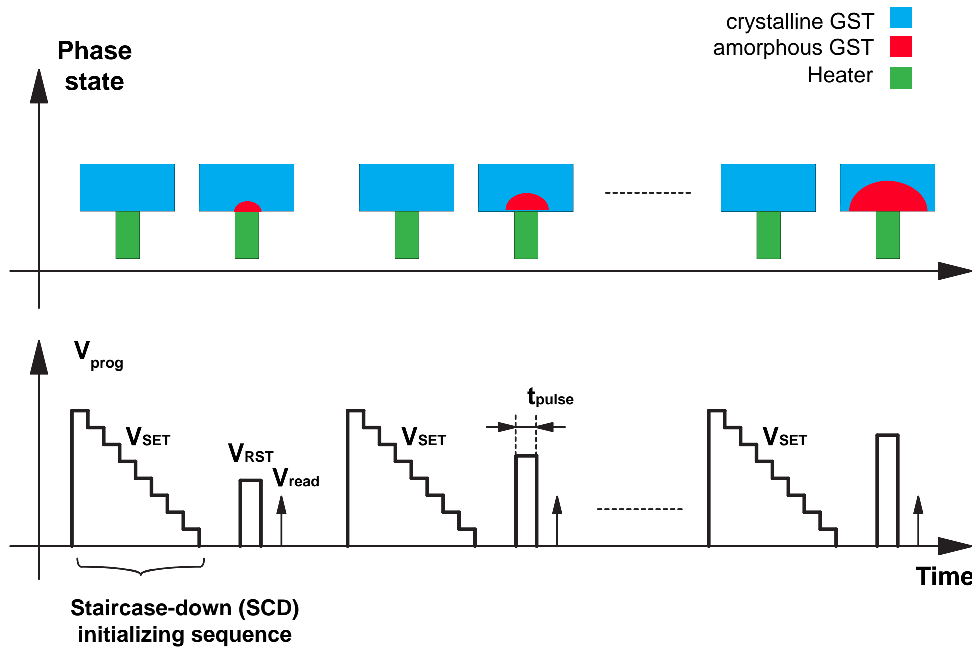


FIGURE 1.10: Sequence of program and read pulses of the single pulse programming algorithm.

The algorithm can be performed either with RESET pulses having the same duration but increasing amplitude of a certain voltage ΔV or with RESET pulses having the same amplitude but different duration, in the order of few tens of ns.

In the first case, the higher is the pulse amplitude, the thicker is the resulting amorphous cap in the GST and, consequently, the higher is the cell resistance.

In the second case, the impact of the amorphization dynamics over the cell resistance can be highlighted. It can be observed that the final cell resistance increases with increasing programming pulse duration, until it approaches a saturation level for sufficiently long pulses, which increases with increasing the amplitude of the program pulse.

In staircase-up programming algorithm, the memory cell is initialized to its full-SET state by means of a staircase-down initializing sequence, then a series of RESET pulses is applied to the cell under test. The pulses have the same duration, but the pulse voltage increases each time by ΔV (Fig. 1.11) and the cell state is read after every pulse.

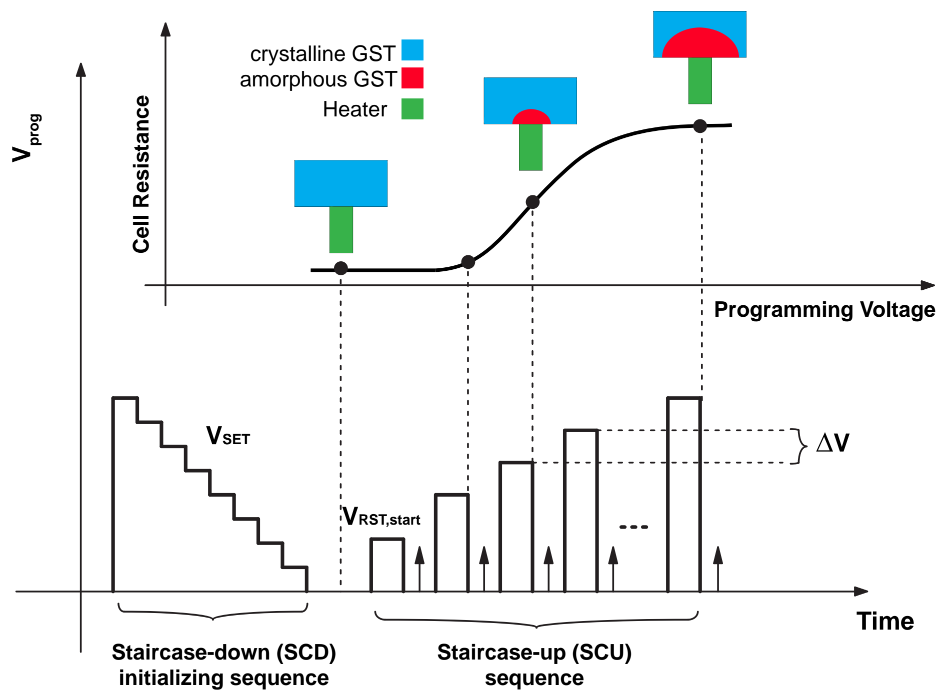


FIGURE 1.11: Sequence of program and read pulses for the staircase-up programming algorithm.

When reducing the time duration of the programming pulses, the minimum voltage to be applied to the cell in order to start the amorphization process slightly increases. On the contrary, the amorphization kinetics is faster if the time duration is longer [29].

1.4 Usual Automated Test Equipment

In emerging non-volatile memories, as already explained, intensive efforts must be devoted to optimize the composition of the active material (for instance, to improve data retention capability), the cell architecture (which, typically, has an important impact on the programming current and, hence, on power consumption), and/or programming algorithms (that affect programming throughput and accuracy). Clearly, optimization efforts also aim at improving production yield and process reliability. In particular, accurate and controllable pulses are required to experimentally investigate the programming performance of such memories [16]

during the research phase, whereas extensive characterization at wafer level is highly desirable so as to collect the large amount of data required to monitor the cell characteristics during the production phase. The focus of this statistical characterization is generally intended to carefully evaluate the "analog" performance of selected cells rather than the "digital" bitmap of large arrays.

It has been already pointed out that the shape of the pulse applied to the memory cell determines the physical state of the active material after programming [30], [31], [32]. During development and research phases, high accuracy and flexibility in controlling the amplitude, the time duration, and the rise and the fall time of programming pulses is therefore essential to investigate the cell behaviour under the widest possible range of programming conditions. Undesired and unpredictable disturbances (such as ringing and overshoots or undershoots that inevitably arise when fast signals are applied through standard testing equipments for on-wafer testing) limit the controllability of programming pulse parameters and the reliability of the testing procedure and the experimental analysis. Indeed, all the disturbances that may occur during programming alter the final cell state and thus adversely affect the quality of the program operation, leading to unexpected results which can limit the efficiency of the testing methodology. The applied pulses must therefore be carefully controlled in order to perform successful programming operations. This aspect is even more important when multilevel-cell programming is addressed [33], [27], [26]. In fact, in this case, the value of the programmed resistance must be controlled much more strictly than in the case of conventional bi-level programming.

Conventional automated test equipments (ATEs) for on-wafer parametric testing typically includes an Automatic Parametric Wafer Testing System (APWTS) and a Pulse Generator (PG) that cooperate to perform tests.

On the one hand, the APWTSs feature high accuracy in both generating and reading voltages and currents. They are usually equipped with a ground reference (GND) and four Source Measurement Units (SMUs), which are able to provide either a DC voltage level or a DC current level. The voltage (current) level can be

changed during a test sequence, but the settling time of a SMU was experimentally evaluated to be around 200 μs.

In addition, APWTSs are able to read a current, when forcing a voltage, and to read a voltage, when forcing a current. However, when reading an electrical variable, the voltage (current) under measure needs to be at a steady level for the instrumentation acquisition time, which may vary between 1 ms to 10 ms.

The high-accuracy of the APWTSs is therefore mainly conceived for DC characterization of elementary device parameters and is not adequate to implement parametric tests on new PCM devices, which require fast and accurate signals when characterizing their I-V response.

On the other hand, the PGs have two Pulse Generation Units (PGUs) which can provide voltage pulses with programmable amplitude, rise time, fall time, and time duration. However, long cables are necessary to connect the on-wafer Device Under Test (DUT) to the PGs. This required interconnection chain is typically the limiting factor of the effective bandwidth achievable by using conventional PGs. In fact, non negligible (also including ringing transients) noise is typically coupled to the generated waveforms, limiting both the accuracy and the controllability of programming pulses. It was experimentally observed that the PGUs settling time is about 100 ns when the pulse rise and fall time is about 80 ns, which is still not adequate for PCM characterization.

Moreover, the usual hybrid configuration (APWTS and PG) makes use of a switch matrix to connect the pads of the DUT under characterization to the different signals provided by the test equipment, thus decreasing testing speed due to switching and settling times. This heavily affects overall testing costs when a huge amount of measurements is required to carry out exhaustive analysis. In fact, with the standard configuration, a write-and-read cycle can last from few hundreds of ms up to 1 s.

Although instrumentation with enhanced performance is available on the market, it is intended more for use in laboratory than for statistical analysis, and is therefore not suited to the investigation at hand.

There is another important feature to be taken into account when characterizing PCM devices, that is reading the programming current at the pulse plateau. This is fundamental to get useful information about the cell programming performance. As already pointed out, APWTSs are able to read currents with high accuracy, provided that the current value remains stable for an adequate amount of time, which is much longer than the programming pulse duration (few ms for reading compared to tens to hundreds of nanoseconds of pulse duration).

It is thus highly desirable to have a custom on-wafer system able to generate the required programming pulses with high accuracy and flexibility. The system must be easy to use, but it also has to be non-invasive. In order to ease the testing process and save the silicon area required to implement the above ad-hoc circuitry, the best solution is to drive the on-wafer system by using the capabilities of existing commercial instrumentation as much as possible. Furthermore, in order to limit the impact of this approach on testing costs on dedicated test structures or limiting the testing directly to the final product, the proposed on-wafer pulse generator should be allocated within the wafer scribe lanes (Fig. 1.12). This implies the need for providing the system with low silicon area occupation and a particularly disadvantageous aspect ratio.

In addition to having the capability to generate the desired programming pulses, the system should be provided with the possibility of selecting cells located in different positions within an array (a mini-array is sufficient) so as to allow exploring side effects related to the cell location. Moreover, an interface that allows the programming current to be read by an ATE should be integrated. Essentially, this interface must be able to sample and hold the programming current at pulse plateau and feed the test equipment with a replica of the sampled current.

All the above requirements should be met while still minimizing both the number of pads used to configure the different parameters of the pulse to be generated (i.e.,

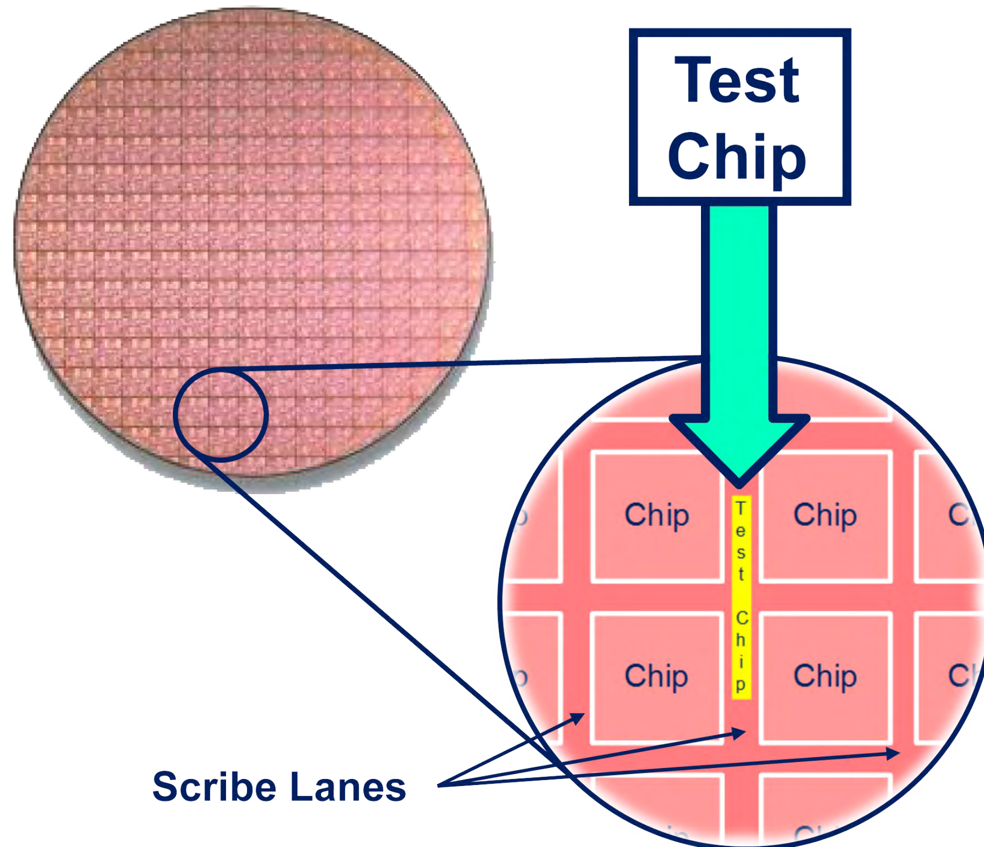


FIGURE 1.12: Placing of the test-chip.

amplitude, pulse duration, and fall time) and the pads used to address the cells in the mini-array. A possible solution should be using digital configuration bits, which means giving the pulse information to the test-chip by means of a digital word, and convert it internally into an analog information. Unfortunately, when using a pad limited digital system, the required digital control data must necessarily be provided through a serial communication, which compromises testing speed. Moreover, interfacing the integrated system with the external instrumentation when using a digital equipment, which is not included in standard ATEs, is a very challenging task, due the need for a time consuming communication protocol. In addition, it is very difficult to read the cell current with a standard digital tester due to the intrinsic limitation of this equipment in the low current range. The best solution is therefore to implement the system as a fully analog circuit, which allows a write-and-read cycle to be performed in tens of ms by switching among the two operating phases using internal pass-gates. This way, the use of the switch matrix

is limited to set cell selection at the beginning of a test sequence by connecting the address pads to the desired voltage levels, which makes the test sequence faster.

Finally, component non-idealities and process spreads will limit system accuracy, even though the test chip is designed to reduce these effects. An automatic procedure to calibrate the system, compatible with the available instrumentation characteristics (such as analog variable ranges and resolution), has therefore to be conceived in order to enhance the overall accuracy of the entire measurement chain. This calibration procedure should be easily implemented in a test sequence and adequately fast so as not to affect testing speed. This means that the number of measurements needed to calibrate the system must be negligible with respect to the total number of measurements required to perform a test sequence.

Chapter 2

On-wafer pulse generator for Phase Change Memory cells

2.1 Target specifications and high-level description

As explained in Chapter 1, having a custom on-wafer system to perform massive test on Resistive Switching Memories is highly desirable.

This on-chip pulse generator must be able to generate the required programming pulses with high accuracy and flexibility exploiting the external ATE.

The program pulse to be applied to the selected memory cell must have a trapezoidal shape (Fig. 2.1) with the following features:

- pulse amplitude programmable from 0.5 V to 4.5 V;
- pulse fall time programmable from 10 ns to several μs;
- pulse time duration programmable from 50 ns to 350 ns;
- pulse rise time set to about 15 ns;

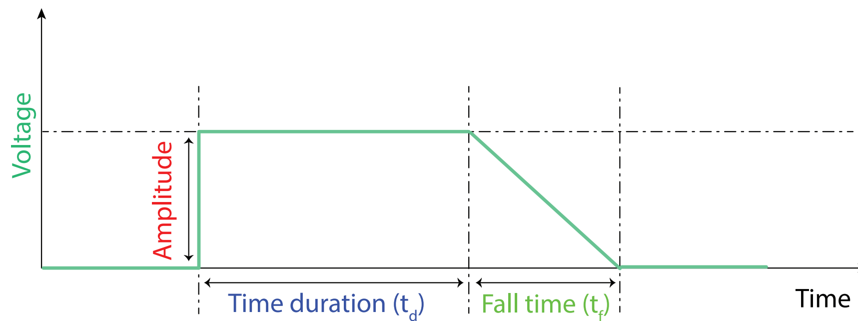


Figure 2.1: Programmable parameters of the program pulse.

- accuracy in generating the pulse parameters within ±10%.

The programmable pulse parameters are externally set by the amplitude of three analog signals.

An accuracy within ±10% in generating the pulse timing parameters is considered adequate for the investigation of the I-V characteristics of emerging memory cells and their performance even with very tightly controlled programming conditions as, according to experimental observations, a variation of these parameters on this order gives rise to no repeatable differences in the obtained cell resistance. Moreover, conventional ATEs are usually affected by uncertainties on the order of a few percent, and fabrication process variations contribute to further spreads: due to the sensitivity of the transfer function of the designed circuit from the input control voltages to the output pulse shape, these uncertainties result in a final inaccuracy in pulse timing parameters in the 10% range.

Moreover, the programming pulse should be fed to the cell after an externally programmable delay, Δt, which takes the transients of external control signals at the input pads of the test chip into account: more specifically, Δt must be greater than or equal to the minimum time necessary for control signals from the PG to reach their steady-state voltage level.

A high-level scheme of the core of the on-chip pulse generator is illustrated in Fig. 2.2. The analog signals from the external instrumentation (which can be either voltage or current signals) are recombined in the on-chip pulse generator

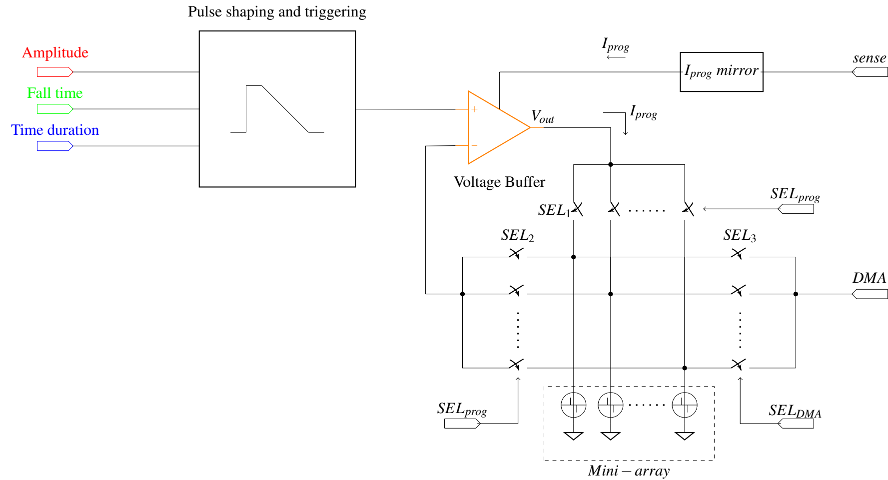


FIGURE 2.2: High-level scheme of the proposed system for memory cell characterization.

in order to generate the desired pulse, which is then applied to the selected cell through a voltage buffer able to deliver the amount of current needed for the programming operation. The voltage buffer is an operational amplifier in non-inverting unity gain configuration, whose aim is to deliver the desired power to the cell with the adequate voltage accuracy. This aspect will be investigated better in Chapter 3.

In Fig. 2.2, the load of the voltage buffer is represented as a number of addressable memory cells but, actually, these cells are placed in a mini-array. The cells can be easily addressed by means of a decoding network: the number of addressable cells depends on the number of pads available for this purpose. In our test-chip implementation, two cells, placed within a mini-array, can be selected (one address pad is present).

The decoding approach changes depending on the kind of memory cell to be characterized [27], [28]. Basically, cell selection is performed by applying proper voltage levels to the Word Line and the Bit Line of the desired memory cell and different voltage levels to the remaining Word Lines and Bit Lines; in some cases unselected Bit Lines may also be floating.

Cell selection is performed at the beginning of a test sequence by suitable digital circuits that drive few pass-gates, indicated as switches in Fig. 2.2, under the

control of dedicated external signals. Once cell selection is performed, the test sequence starts and evaluates the characteristics of the selected cell with negligible time overhead between the programming and the read sequence. A current mirror replicates the programming current, which is sensed by the APWTS. The APWTS can also read the current of the selected cell in Direct Memory Access (DMA) mode when required. The flexibility of this solution is that even the (analog) current of the cell, not only its programmed state, is measured. This way, a careful analysis of the I-V cell characteristic can be performed.

The conceived system allows performing extensive and accurate analyses on a limited number of cells. In particular, only one cell can be characterized during each test sequence (the address pad is kept to a constant voltage level during the whole test sequence, so that the switch matrix is used only at the beginning of the sequence, thus significantly decreasing testing times).

However, further developments can be easily implemented. As an example, by replicating the output buffer, more cells can be analysed simultaneously. This solution can be adopted when testing speed has to be increased significantly. Basically, the number of cells which can be programmed simultaneously during a test sequence is limited by the number of pads available and by the number of available reading channels in the ATE.

Another application that can be easily implemented is studying whether programming one cell affects the neighbouring cells, which is of utmost importance in the case of aggressive technology scaling down. To this end, the test chip should be designed so as to include the possibility to choose the cell to be read in DMA mode, which can be either the cell under investigation or one of the neighbouring cells.

Three solutions were conceived to implement the above principle scheme, and their advantages and disadvantages were analysed in order to identify the most suitable. In the following Sections, an overview of the three solutions is first provided. Then, these solutions are compared and the best one is chosen. The

chosen solution, which was designed and integrated, is finally discussed in detail in Chapter 3.

2.2 First solution

2.2.1 Basic operating principle

In this solution, information about pulse amplitude and pulse fall time are fed to the test chip by two different PG channels (V_{ampl} and V_{tf}, respectively), whereas information about pulse time duration is provided by the APTWS as a DC current level, I_{td}^{dch}.

- The voltage amplitude of signal V_{ampl} exactly represents the desired program pulse amplitude, and is stored in a storage element (a capacitor named C_{ampl}) after twice the externally programmable delay Δt.
- The voltage amplitude of signal V_{tf} is converted (after twice the delay Δt) into a current I_{tf} by applying V_{tf} at the gate of a saturated NMOS transistor connected in common source configuration ($I_{tf} = \frac{\beta}{2}(V_{tf} - V_{th})^2$, with the usual meaning of symbols). The time required to discharge C_{ampl} represents the pulse fall time.
- Signal I_{td}^{dch} is a constant current which charges a capacitor named C_{td}. The time spent to charge C_{td} until its voltage reaches a predetermined reference level $V_{ref,td}$ (internally generated from supply voltage V_{CC} by means of a resistive divider) represents the pulse duration.

Time delay Δt is programmed as a function of the settling time of pulses V_{ampl} and V_{tf} (more specifically, Δt is set larger than both the above settling times). From above, it is clear that Δt must be processed twice, namely once per each pulsed control signal (V_{ampl} and V_{tf}), before the programming pulse can be fed to the cell.

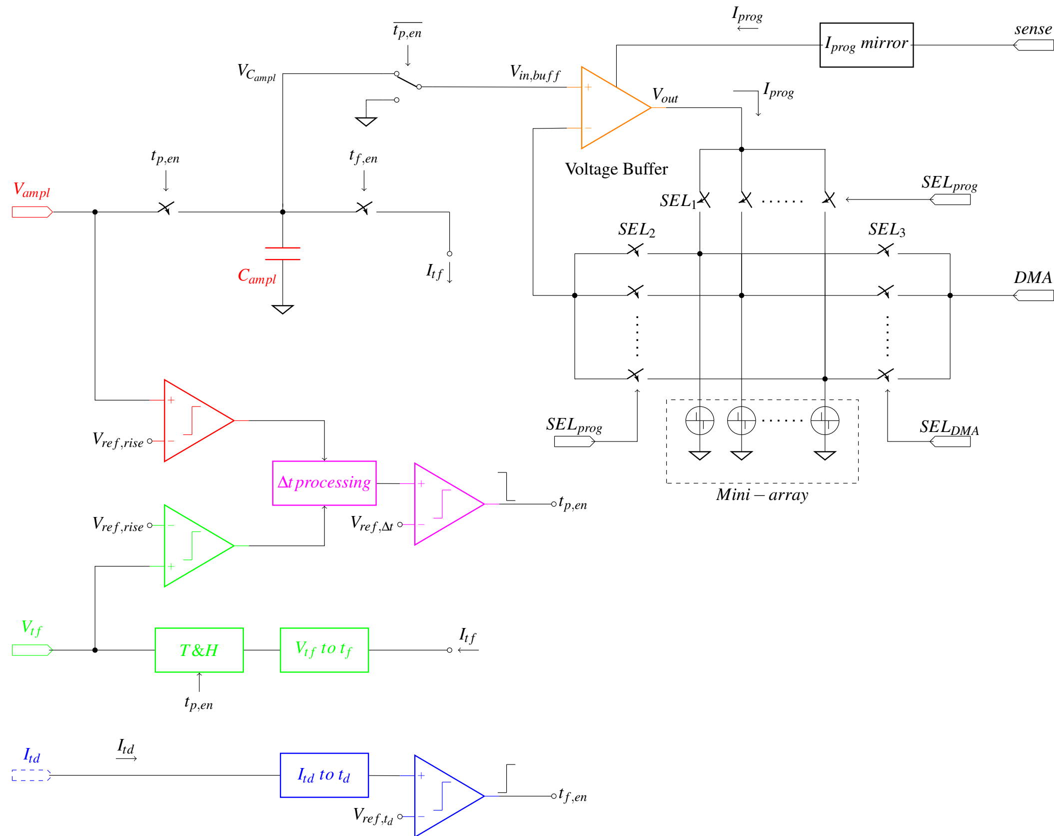


FIGURE 2.3: Block diagram of the 1st conceived solution

In all the three solutions, all external voltage signals are first tracked and then held across three respective capacitors after $2\Delta t$. The value of Δt is established just once, based on experimental evaluation, depending on the equipment used, so as to ensure negligible distortion of signals from the PG due to interconnection cables.

Using Track-and-Hold circuits assures that a stable voltage value is fed to the cascaded blocks for pulse generation. The use of Track-and-Hold rather than Sample-and-Hold circuits increases speed in the voltage level acquisition.

2.2.2 System description

The system will be described referring to the block diagram in Fig. 2.3 and the waveforms in Fig. 2.4.

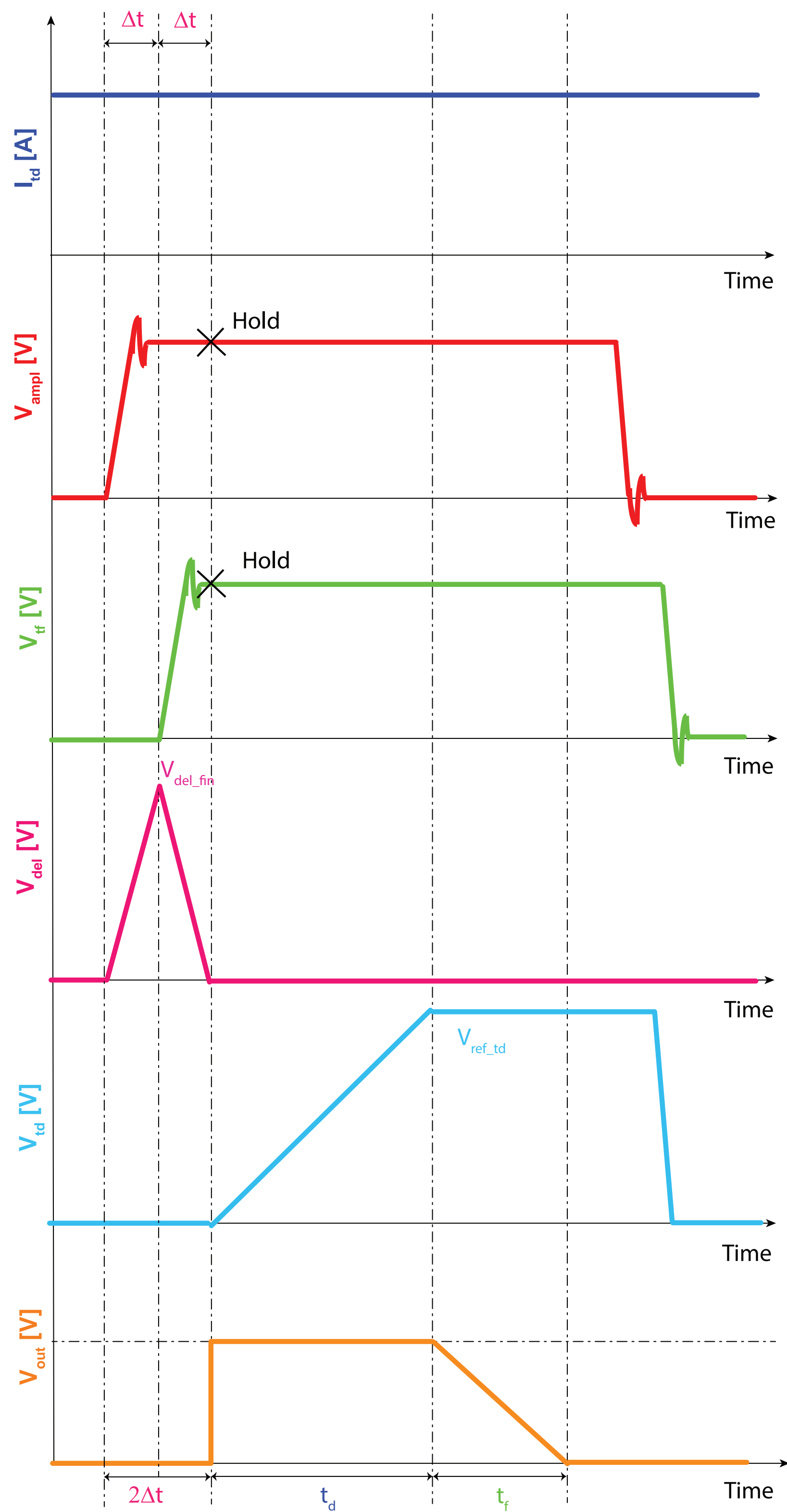


FIGURE 2.4: External (I_{td}^{dch}, V_{ampl}, V_{tf}) and internal ($V_{\Delta t}$, V_{Ctd}, V_{out}) waveforms of the 1st conceived solution

Storage capacitor C_{ampl} is initially discharged and, hence, the output voltage V_{out} is initially forced to 0 V through the output buffer. After the rising edge of V_{ampl} is detected, a capacitor, $C_{\Delta t}$, which is initially discharged ($V_{\Delta t} = 0\ V$), is charged with an internally generated constant current, $I_{\Delta t}$, until the rising edge of V_{tf} is detected, after a time interval Δt. The final value of $V_{\Delta t}$ is

$$V_{\Delta t_fin} = \frac{I_{\Delta t}\ \Delta t}{C_{\Delta t}} \tag{2.1}$$

At this instant ($t = \Delta t$), the system begins to discharge $C_{\Delta t}$ with the same current $I_{\Delta t}$. Voltage $V_{\Delta t}$ then approaches 0 V after an additional time delay

$$t_{add} = V_{\Delta t_fin} \frac{C_{\Delta t}}{I_{\Delta t}} = \Delta t \tag{2.2}$$

i.e., at $t = 2\Delta t$.

The value of $I_{\Delta t}$ does not need to be very accurate: the key factor for this conversion is an adequate matching between the above charging and discharging currents, which is easily achieved on-chip by means of current mirrors.

Voltage V_{tf} is held at $t = 2\Delta t$. This voltage is converted into a current

$$I_{tf} = \frac{\beta}{2}(V_{tf} - V_{th})^2 \tag{2.3}$$

by means of an NMOS transistor, M_{tf} in its common source configuration and of current mirrors. Current I_{tf} is used to control the falling slope of V_{out}, as explained below.

At $t = 2\Delta t$, V_{ampl} is also held across capacitor C_{ampl}. The output buffer therefore makes V_{out} rise from 0 V to V_{ampl}. The current I_{td}^{dch}, provided by the APWTS, begins now to charge C_{td}, thus producing an increasing voltage ramp V_{td}. When V_{td} becomes higher than reference voltage $V_{ref,td}$, capacitor C_{ampl} is connected to the discharging current I_{tf}, and, hence, V_{out} begins to decrease, thus determining

the pulse duration. As mentioned above, the value of I_{tf} controls the falling slope of the generated pulse V_{out}.

When the falling edge of V_{tf} is detected, V_{td} is reinitialized to 0 V.

2.2.3 Accuracy of main parameters

As already pointed out, the program pulse accuracy is affected by component non-idealities. In detail:

- pulse amplitude is programmed by means of V_{ampl}; its accuracy is affected by charge injection effects (sampling on capacitor C_{ampl}) and output buffer performance;
- pulse duration is programmed by means of I_{td}^{dch}; its accuracy is affected by capacitor spreads (C_{td}), comparator performance, and resistor matching ($V_{ref,td}$);
- pulse fall time is programmed by means of V_{tf}; its accuracy is affected capacitor spreads (C_{ampl}), charge injection effects (sampling on capacitor C_{tfs} and connection of capacitor C_{ampl}), transistor spreads (M_{tf}), and transistor matching (current mirrors).

2.3 Second solution

2.3.1 Basic operating principle

In this solution, information about pulse amplitude and pulse time duration are encoded by the amplitude of signals V_{ampl} and V_{td}, respectively, which are provided by the PG, whereas pulse fall time is fed to the test chip by the APWTS as a DC current, I_{tf}.

- As in the first solution, the voltage amplitude of signal V_{ampl} represents the program pulse amplitude and is stored across capacitor C_{ampl} after twice the externally programmable delay Δt.
- Pulse time duration is programmed by means of signal V_{td}, whose voltage amplitude is converted (after twice the delay Δt) into a current I_{td}^{dch} by applying V_{td} over a resistor ($I_{td}^{dch} = \frac{V_{td}}{R_{td}^{dch}}$). I_{td}^{dch} is used to discharge a capacitor, C_{td}, which was previously charged for a time interval Δt with an internally generated constant current I_{td}^{ch}. The time required to discharge voltage V_{Ctd} across C_{td} represents the pulse time duration.
- DC current I_{tf} provides a constant current which discharges C_{ampl} at a constant rate, thus determining the desired fall time.

As in the first solution, the delay Δt is programmed as a function of the settling times of pulses from the PG, V_{ampl} and V_{td}.

The main difference between this solution and the first one consists in the generation of time duration and fall time.

2.3.2 System description

The system will be described referring to the block diagram in Fig. 2.5 and the waveforms in Fig. 2.6.

As in the first solution, C_{ampl} is initially discharged and, hence, the output voltage V_{out} is initially forced to 0 V through the output buffer. After the rising edge of V_{ampl} is detected, capacitor $C_{\Delta t}$, which is initially discharged ($V_{\Delta t} = 0\ V$), is charged with an internally generated constant current, $I_{\Delta t}$, until the rising edge of V_{td} is detected, after a time interval Δt. At this instant ($t = \Delta t$), the system begins to discharge $C_{\Delta t}$ with the same current $I_{\Delta t}$. Voltage $V_{\Delta t}$ approaches 0 V after an additional time delay Δt, i.e., at $t = 2\Delta t$. During the time interval from Δt to 2Δt, C_{td} is also charged at a constant rate by an internally generated current

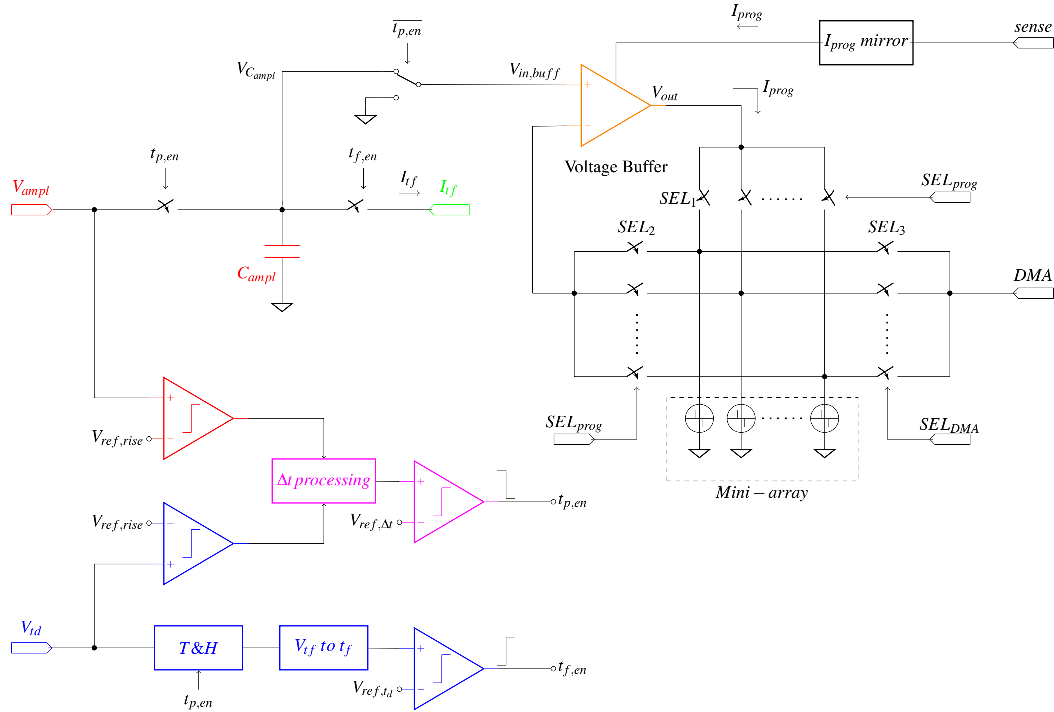


FIGURE 2.5: Block diagram of the 2nd conceived solution

$$I_{td}^{ch} = \frac{V_{CC} - V_{b,td}}{R_{td}^{ch}} \tag{2.4}$$

where $V_{b,td}$ is the gate to source voltage of an MOS transistor and R_{td}^{ch} is an n-well resistor, thereby reaching a final voltage level

$$V_{Ctd_fin} = \frac{I_{td}^{ch}}{C_{Ctd}}\Delta t = \frac{V_{CC} - V_{b,td}}{R_{td}^{ch}} \frac{\Delta t}{C_{Ctd}} \tag{2.5}$$

As in the first solution, at $t = 2\Delta t$, V_{ampl} is held across storage capacitor C_{ampl}. The output buffer therefore makes V_{out} rise from 0 V to V_{ampl}.

In the same instant, voltage V_{td} is also held across a dedicated capacitor, C_{tds}. This held voltage is converted into a current

$$I_{td}^{dch} = \frac{V_{td}}{R_{td}^{dch}} \tag{2.6}$$

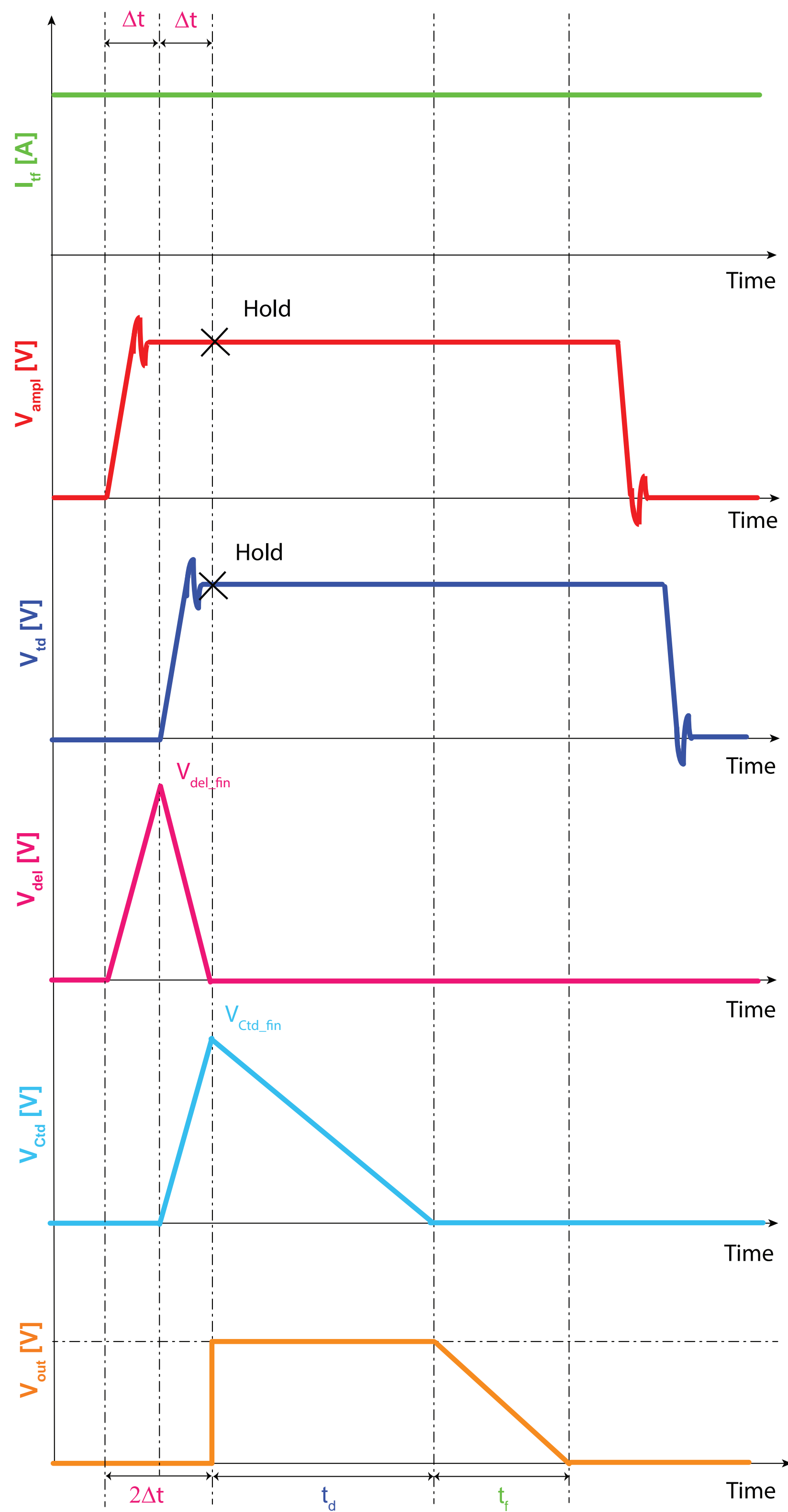


FIGURE 2.6: External (I_{tf}, V_{ampl}, V_{td}) and internal ($V_{\Delta t}$, V_{Ctd}, V_{out}) waveforms of the 2nd conceived solution

by means of an operational amplifier and an n-well resistor, R_{td}^{dch}, matched to R_{td}^{ch}. Current I_{td}^{dch} discharges capacitor C_{Ctd}, so that V_{Ctd} approaches 0 V after a time interval t_d equal to

$$t_d = V_{Ctd_fin} \frac{C_{Ctd}}{I_{td}^{dch}} = \frac{V_{CC} - V_{b,td}}{V_{td}} \frac{R_{td}^{dch}}{R_{td}^{ch}} \Delta t \tag{2.7}$$

As equation (2.7) clearly shows, pulse duration t_d is dependent on ratio $\frac{R_{td}^{dch}}{R_{td}^{ch}}$, which allows minimizing any problem due to fabrication process spreads.

At this time instant, capacitor C_{ampl} is connected to discharging current I_{tf} provided by the APWTS, thus giving rise to the beginning of the falling slope of the pulse and, hence, determining the pulse duration.

The fall time is obviously controlled by the programmed current I_{tf}.

It is worth to underline that the time duration of the pulse is controlled not only by the pulse amplitude of signal V_{td}, but also by the value of the external delay Δt, which adds a degree of freedom with respect to the case of the first solution.

2.3.3 Accuracy of main parameters

In this second solution, the program pulse accuracy is affected by the following non-idealities:

- pulse amplitude is programmed by means of V_{ampl}; its accuracy is affected by charge injection effects (sampling on C_{ampl}) and output buffer performance;
- pulse duration is programmed by means of V_{td}; its accuracy is affected by resistor matching (R_{td}^{ch}, R_{td}^{dch}), transistor matching (current mirrors), charge injection effects (sampling on capacitor C_{tds}), comparator performance, operational amplifier offset and DC gain ($\frac{V_{td}}{R_{td}^{dch}}$) and transistor spreads ($\frac{V_{CC}-V_{b,td}}{R_{td}^{ch}}$);
- pulse fall time is programmed by means of I_{tf}; its accuracy is affected by capacitor spreads (C_{ampl}) and charge injection effects (connection of C_{ampl}).

2.4 Third solution

2.4.1 Basic operating principle

In this solution, the required value of pulse amplitude is directly fed to the test chip by the APWTS as a DC level, V_{ampl}, whereas the pulse fall time and the pulse time duration are encoded by means of the amplitude of pulses V_{tf} and V_{td}, respectively, which are provided by the PG.

- The voltage amplitude of signal V_{ampl} exactly represents the program pulse amplitude, and is therefore stored across capacitor C_{ampl} after twice the delay time Δt and directly fed to the cell through the output buffer.
- The voltage amplitude of signal V_{tf} is converted into a current I_{tf} by applying V_{tf} (after twice the delay Δt) at the gate of a saturated NMOS transistor connected in common source configuration. The time I_{tf} takes to discharge C_{ampl} represents the pulse fall time.
- The voltage amplitude of signal V_{td} is converted into a current I_{td}^{dch} (after a time equal to 2Δt) by applying V_{td} across a resistor R_{td}^{dch} ($I_{td}^{dch} = \frac{V_{td}}{R_{td}^{dch}}$). I_{td}^{dch} discharges a capacitor, C_{Ctd}, which was previously charged for a time interval Δt with an internally generated constant current I_{td}^{ch}. The time I_{td}^{dch} takes to discharge C_{td} represents the pulse time duration.

2.4.2 System description

The third solution will be described referring to the block diagram in Fig. 2.7 and the waveforms in Fig. 2.8.

As in the previous solution, voltage V_{ampl} needs no conversion to set the amplitude of the generated pulse, so it is held across C_{ampl}, which is initially discharged C_{ampl}. This way, the output voltage V_{out} is initially forced to 0 V through the output

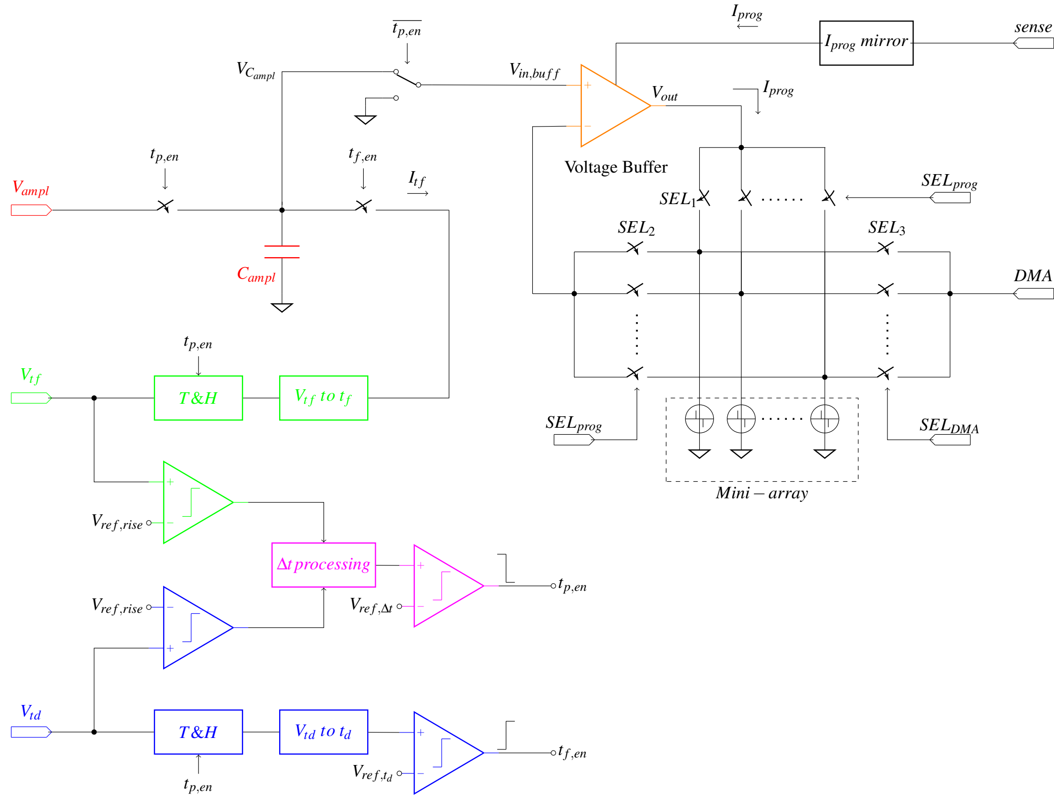


FIGURE 2.7: Block diagram of the 3rd conceived solution.

buffer. As in previous solutions, time delay $2\Delta t$ is obtained by first charging and then discharging a capacitor ($C_{\Delta t}$): after the rising edge of V_{tf} is detected, capacitor $C_{\Delta t}$, which is initially discharged ($V_{\Delta t} = 0$ V), is charged with a constant current ($I_{\Delta t}$) until the rising edge of V_{td} is detected, after a time interval Δt. At this instant ($t = \Delta t$), the system begins to discharge $C_{\Delta t}$ with the same current $I_{\Delta t}$. Voltage $V_{\Delta t}$ will then approach 0 V after an additional time delay Δt, i.e., at $t = 2\Delta t$.

During the time interval from Δt to 2Δt, C_{td} is also charged at a constant rate by an internally generated current

$$I_{td}^{ch} = \frac{V_{CC} - V_{b,td}}{R_{td}^{ch}} \tag{2.8}$$

thereby reaching a final voltage level V_{Ctd_fin}. At $t = 2\Delta t$, the output buffer is enabled and makes V_{out} rise from 0 V to V_{ampl}.

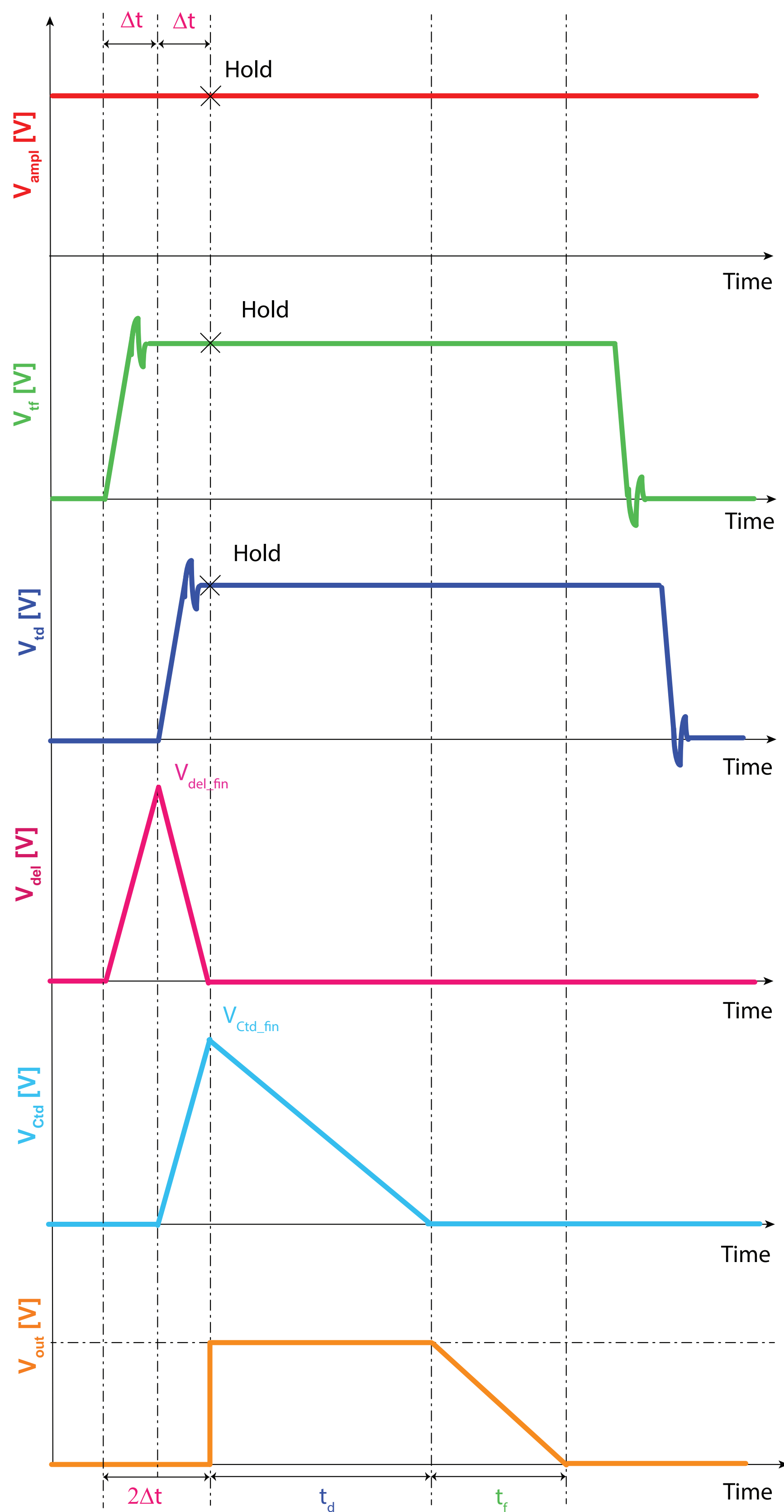


FIGURE 2.8: External (V_{ampl}, V_{tf}, V_{td}) and internal ($V_{\Delta t}$, V_{Ctd}, V_{out}) waveforms of the 3rd conceived solution.

At $t = 2\Delta t$, the output buffer makes V_{out} rise from 0 V to V_{ampl}.

In the same instant, voltages V_{td} and V_{tf} are also held across two dedicated capacitors, C_{tds} and C_{tfs}, respectively, and C_{Ctd} begins to be discharged with an internally generated current

$$I_{td}^{dch} = \frac{V_{td}}{R_{td}^{dch}} \tag{2.9}$$

The voltage, V_{Ctd}, across this capacitor will approach 0 V after a time interval

$$t_d = \frac{I_{td}^{ch}}{I_{td}^{dch}}\Delta t = \frac{R_{td}^{dch}}{R_{td}^{ch}}\frac{V_{CC}}{V_{td}}\Delta t \tag{2.10}$$

At this time instant, capacitor C_{ampl} and discharging current I_{tf} are connected, thus starting the discharge of capacitor C_{ampl} and, hence, setting the length of the generated pulse.

The desired program pulse fall time is obtained by means of a voltage-to-time conversion, which is performed by discharging C_{ampl} with a current, I_{tf}, generated by an MOS transistor, M_{tf}, operated in the saturation region under the control of voltage V_{tf}:

$$I_{tf} = \frac{\beta}{2}(V_{tf} - V_{th})^2 \tag{2.11}$$

with the usual meaning of symbols.

2.4.3 Accuracy of main parameters

In this solution, the non-idealities affecting the program pulse accuracy are the following:

- pulse amplitude is programmed by means of V_{ampl}; its accuracy is affected by charge injection effects (sampling on capacitor C_{ampl}) and output buffer performance;
- pulse duration is programmed by means of V_{td}; its accuracy is affected by by resistor matching (R_{td}^{ch}, R_{td}^{dch}), transistor matching (current mirrors), charge injection effects (sampling on capacitor C_{tds}), comparator performance, operational amplifier offset and DC gain ($\frac{V_{td}}{R_{td}^{dch}}$) and transistor spreads ($\frac{V_{CC}-V_{b,td}}{R_{td}^{ch}}$);
- pulse fall time is programmed by means of V_{tf}; its accuracy is affected by transistor spreads and matching (M_{tf}, current mirrors), capacitor spreads (C_{ampl}) and charge injection effects (sampling across capacitor C_{tfs} and connection of capacitor C_{ampl}).

2.5 Comparison of the conceived solutions

The following table (Table 2.1) provides a comparative summary of the characteristics of the three solutions described above.

In the first section of the table, symbol "X" indicates whether the parameter information is given by a signal from the APWTS or by a signal from the PG. In the second section, symbol "X" indicates whether a voltage or a current control signal is fed to the pulse generation circuit. In the third section, symbol "X" indicates the factors affecting the accuracy of pulse parameters in each solution.

The second and third solutions are substantially equivalent as far as accuracy issues are concerned. However, the third solution has been preferred because it better exploits the flexibility of the used equipment. In fact, signals from the PG cannot be programmed in order to vary the pulse amplitude while running, whereas signals from the APWTS can. The third solution is therefore more flexible. For instance, it also allows the implementation of algorithms such as stair-case up and stair-case down programming. This solution was therefore chosen to be implemented.

Table 2.1: Comparison of the conceived solutions

Solution	First solution			Second solution			Third solution		
Parameter	Ampl.	Dur.	Fall	Ampl.	Dur.	Fall	Ampl.	Dur.	Fall
SMU		X			X		X		
PGU	X		X	X		X		X	X
Voltage input	X		X	X	X		X	X	X
Current input		X				X			
Output buffer performance	X			X			X		
Charge injection	X		X	X	X	X	X	X	X
Capacitor spreads		X	X			X			X
Resistor matching		X			X			X	
Comparator performance		X			X			X	
Op-amp offset					X			X	X
Transistor spreads			X		X			X	X
Transistor matching			X		X			X	X

From Table 2.1, it is clear that, although the described solutions are designed so as to reduce the effects due to component non-idealities and process spreads, still some impact of these effects limits system accuracy. A procedure to calibrate the system implementing the chosen solution, compatible with the instrumentation characteristics (such as analog variable ranges and resolution), was then conceived in order to enhance the overall accuracy of the entire measurement chain.

The implementation of the third solution, as well as the corresponding calibration procedure, will be shown in Chapter 3.

Chapter 3

Analysis and test chip of the main blocks

3.1 Aim of the prototype

A first prototype of the on-chip pulse generator was designed and fabricated for debugging purposes. The main goal of this prototype is the experimental evaluation of the pulse generator (third solution) conceived and shown in Chapter 2. In this prototype, the interface with the available ATE was therefore reduced to a minimum. Besides, after a preliminary on-wafer experimental analysis carried out by using the above ATE, the integrated prototype was assembled in a Dual-In-Line (DIL) ceramic package so as to allow the evaluation of the generated waveforms at the output of the output buffer by means of an active microprobe, which is not possible when using commercial ATE.

In addition, the output buffer was provided with the possibility to read the current fed to the load with an external equipment so as to allow monitoring the behaviour of the memory cell during experimental investigation. In particular, it is important to monitor the amplitude as well as the falling edge of SET current pulses. In fact, as pointed out in Chapter 1, the final state of the cell after the SET operation strongly depends on the pulse fall time (from $\approx$ 100 ns up to several μs), since the

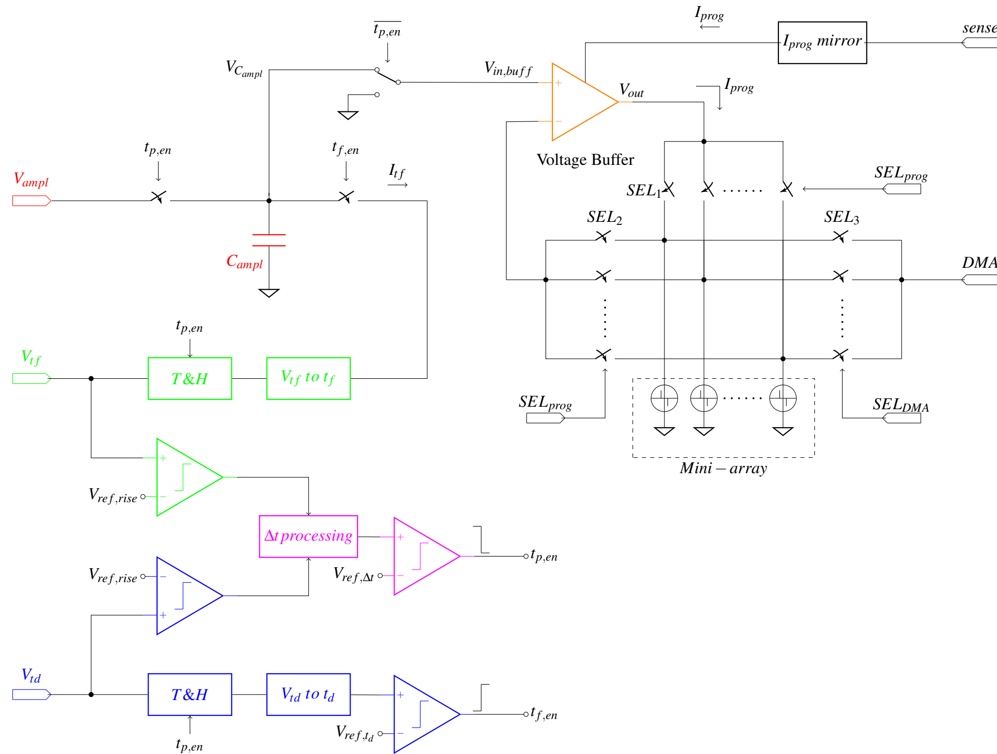


FIGURE 3.1: Block diagram of the proposed integrated system.

crystallization of the active chalcogenide portion takes place in this part of the programming pulse [32].

Moreover, in this first prototype, a manual calibration procedure was conceived and experimentally evaluated.

The overall structure is repeated here in Fig. 3.1 for convenience.

3.2 Circuit design

In the following of this Chapter, circuital details of the main blocks of the implemented system (third solution) are provided and analysed.

3.2.1 Output buffer

As explained in Chapter 1, during a high-to-low resistance (SET) programming operation, the PCM cell resistance abruptly drops under certain bias conditions.

The buffer which drives the memory cell therefore plays a key role during programming operations, as it must be able to feed the optimized pulse to the memory cell under test with adequate accuracy while providing the required amount of current. The large variability of the cell resistance must be taken into account during the design phase.

A preliminary buffer targeted at this application was therefore designed, fabricated and experimentally characterized (for experimental results, see Chapter 5). A final version was then developed and included in the whole pulse generator system.

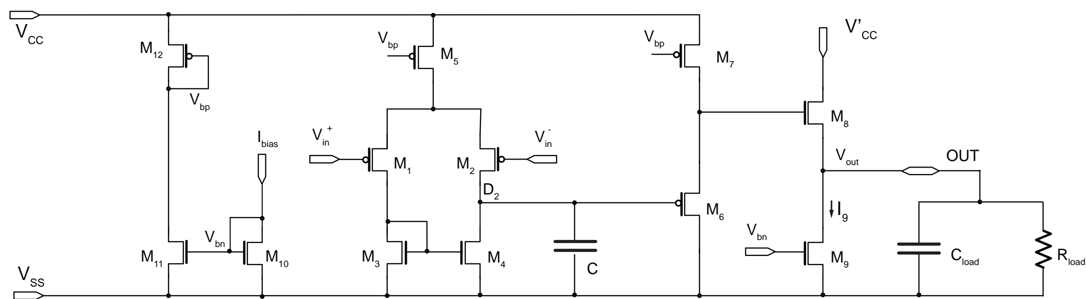


FIGURE 3.2: Circuit schematic of the proposed amplifier.

3.2.1.1 Preliminary version of the output buffer

As the typical SET resistance of a PCM memory cell in the used fabrication technology is about 10 kΩ, the buffer must have a current drive capability up to 0.45 mA. Further requirements are the ability of reproducing pulses with rise time and fall times down to 15 ns and an input/output voltage range from $\approx$ 0.4 V to 4.5 V. The lower bound was set so as to be able to control the memory cell at least down to a safe voltage (lower than the hold voltage of the cell), which is not able to (unintentionally) degrade the programmed cell state.

A moderate DC gain of the amplifier (about 30 dB) is sufficient, as the key electrical output variable under test is the programming current. The buffer offset is thus a negligible feature and was not considered during the design.

The high current drive capability requirement leads to the need for an output stage implemented with an n-type source follower.

To cope with the required minimum input voltage level ($V_{i-cm,min} \approx 400$ mV), a p-type input differential stage was chosen.

Rise and fall times essentially determine the required bandwidth of the amplifier. However, in order to ensure stability even in the presence of a light load, the amplifier bandwidth needs to be limited. In fact, in the case of a light load, the secondary pole (which is due to the output node) is pushed towards low frequencies. Both the above opposed requirements were taken into account and will be dealt with later in this Subsection.

An adequate phase margin must be provided to avoid overshoots. This target was achieved by using a single gain stage amplifier. More specifically, a simple differential stage was used. In the further development of the buffer, where high DC values are required, a cascode differential stage is used (Section 3.2.1.2).

The circuit schematic of the amplifier designed for the first version of the buffer is depicted in Fig. 3.2. This amplifier is basically made up by three cascaded sections, namely a differential input stage and two cascaded source followers. The input stage consists of a PMOS differential pair (devices M_1 and M_2 biased by current source M_5) and an NMOS active load (devices M_3 and M_4), which also performs differential-to-single-ended conversion. As pointed out above, a p-type differential stage was chosen to cope with the requirements for the input common-mode range.

The section on the left side of the scheme in Fig. 3.2 generates the bias levels for the amplifier. The bias circuitry is controlled by current I_{bias} (nominal value 100 μA).

As pointed out above, since the buffer must be able to feed up to 0.45 mA to the load, an n-type follower is required as a final stage (transistor M_8 biased by M_9). However, an n-type follower cannot be directly cascaded to the output node of the differential stage. In fact, when connecting the amplifier in buffer configuration, we have $V_{in}^- = V_{out}$. The drain voltage of M_2, V_{d2}, can be expressed as

$$V_{d2} = V_{out} + \Delta V = V_{in}^{-} + \Delta V \tag{3.1}$$

When D_2 is directly connected to the gate terminal of M_8, ΔV is equal to

$$\Delta V = V_{th8} + V_{ov8} \tag{3.2}$$

which gives

$$V_{d2} = V_{in}^{-} + V_{th8} + V_{ov8} \tag{3.3}$$

In equation (3.2) and (3.3) and in the following, V_{thi} and V_{ovi} are the threshold voltage and the overdrive voltage of M_i (i = integer). The same way, g_{mi} will represent the transconductance of M_i.

From (3.3), when V_{in} is high, M_2 may be forced to work in its triode region. To ensure saturated operation of this transistor under any operating and process conditions, an additional p-type follower (transistor M_6 biased by M_7), which acts as a level shifter, was included in front of the n-type follower. This way, ΔV becomes equal to

$$\Delta V = V_{th8} + V_{ov8} - |V_{th6}| - |V_{ov6}| \tag{3.4}$$

and V_{d2}, turns out to be equal to

$$V_{d2} = V_{out} + V_{th8} + V_{ov8} - |V_{th6}| - |V_{ov6}| = V_{in} + V_{th8} + V_{ov8} - |V_{th6}| - |V_{ov6}| \tag{3.5}$$

which ensures saturated operation of M_2 for reasonable values of V_{th8}, $|V_{th6}|$, V_{ov8}, and $|V_{ov6}|$.

The output swing bounds are set by M_9 and M_5, which are brought out of their saturation region for low and high input voltage levels, respectively.

The supply line, V'_{CC}, of the output branch was kept separate from the supply line, V_{CC}, of the rest of the amplifier in order to allow the current flowing through the n-type follower to be measured, as required for PCM cells characterization purposes.

The slew rate, SR, of the amplifier must ensure the above requirements for t_r and t_f to be met. To this end, we set

$$SR = \frac{I_5}{C_2} > \frac{0.8\ V_{ampl,max}}{t_r\ (t_f)} = \frac{3.6}{15}\ \frac{V}{ns} = 0.24\ \frac{V}{ns} \tag{3.6}$$

where I_5 is the biasing current for the input stage, C_2 is the overall capacitance at node D_2, and $V_{ampl,max}$ is the maximum pulse amplitude.

An adequate unity-gain angular frequency, UGF, as well as a sufficiently large phase margin of the amplifier must also be provided, so as to ensure fast settling of the output pulse while still avoiding any risk of large overshoots and ringing transients. The first and the second pole of the amplifier are associated to the output nodes of the differential stage (node D_2) and the second follower (node OUT), respectively. Assuming dominant-pole approximation, we set

$$\frac{5}{t_r\ (t_f)} < UGF = \frac{g_{m2}}{C_2} \leq \frac{1}{10}\omega_{p2} = \frac{1}{10}\ \frac{g_{m8} + \frac{1}{R_{load}}}{C_{out}} \tag{3.7}$$

where C_{out}, which includes the bit-line capacitance C_{load}, is the overall capacitance at node OUT and ω_{p2} is the angular frequency of the second pole (a settling time constant of 5 ns, which is equal to 10% of the minimum time duration, was considered adequate for the purpose, whereas a UGF one decade below the second pole was assumed to ensure adequate phase margin). To prevent ω_{p2} from being shifted to excessively low values in the presence of the highest load resistances, a sufficiently high biasing current I_9 is required. From (3.7), in order to limit

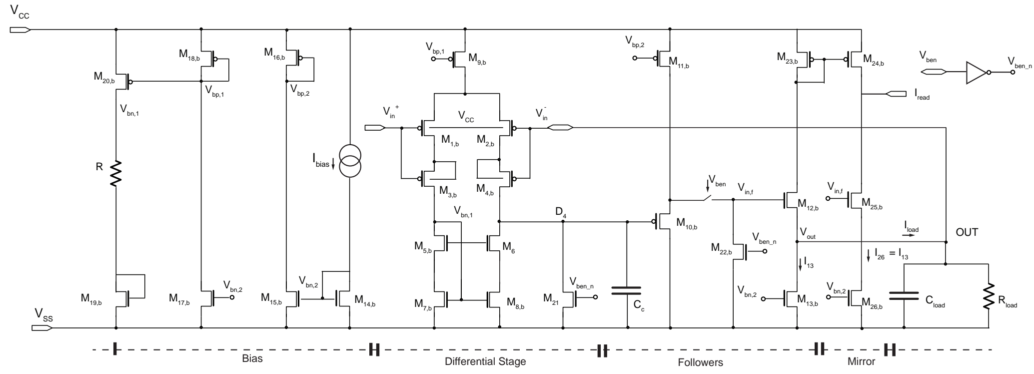


FIGURE 3.3: Circuit schematic of the final version of the buffer.

the required values of g_{m8} (and, hence, of the sizes of M_8 and I_9), we included a capacitor C between node D_2 and ground. The value of I_5 was then determined from (3.6). The width of devices M_2 and M_8 and the value of I_9 where finally chosen so as to meet (3.7).

3.2.1.2 Final version of the output buffer

After the first buffer was experimentally characterized, a second buffer, with a higher open-loop DC gain, was designed and included in the on-chip pulse generator system.

In this implementation, the variables under test are both the programming current and the voltage programming pulse. An error within ±0.5% between the input and the output voltage was considered adequate for the target application. Moreover, also in this design the buffer should be provided with the possibility to read the current fed to the load with an external equipment so as to allow monitoring the behaviour of the memory cell during experimental investigation.

The proposed buffer schematic is depicted in Fig. 3.3. The same way as for the buffer shown in Section 3.2.1.1, also this buffer is basically made up by three cascaded sections, namely a differential input stage and two cascaded followers. However, the input stage was improved by using a PMOS cascode differential pair (devices $M_{1,b}$ to $M_{4,b}$ biased by current source $M_{9,b}$). The NMOS active load

(devices $M_{5,b}$ to $M_{8,b}$) of this stage, which also performs differential-to-single-ended conversion, uses a conventional low-drop cascode configuration.

As in the preliminary buffer, the allowed upper bound of the input common-mode voltage range is set by the need for maintaining $M_{9,b}$ in its pinch-off region, whereas the lower bound is due to $M_{13,b}$.

We will now analyse the biasing requirements of the input pair referring to Fig. 3.4. The gate bias voltage, V_B, of $M_{4,b}$ must be set to a value that ensures saturated operation of $M_{2,b}$ and $M_{4,b}$ over the whole specified input voltage range when the amplifier is connected in unity-gain buffer configuration. To this end, we must have

$$\begin{cases} V_B + |V_{ov4}| + |V_{th4}| & \leq & V_{in}^- + |V_{th2}| \\ V_{d4} & \leq & V_B + |V_{th4}| \end{cases} \tag{3.8}$$

From (3.8), we obtain

$$V_{d4} - |V_{th4}| \leq V_B \leq V_{in}^- - |V_{ov4}| + |V_{th2}| - |V_{th4}| \tag{3.9}$$

When the amplifier is connected in buffer configuration ($V_{out} = V_{in}^-$), V_{d4} tracks V_{out} and, hence, V_{in}^-. More specifically, as in the case of the preliminary version of the buffer (Section 3.2.1.1), we have $V_{d4} = V_{in}^- + \Delta V$, where

$$\Delta V = V_{th12} + V_{ov12} - |V_{th10}| - |V_{ov10}| \tag{3.10}$$

and, hence

$$V_{in}^- + \Delta V - |V_{th4}| \leq V_B \leq V_{in}^- - |V_{ov4}| + |V_{th2}| - |V_{th4}| \tag{3.11}$$

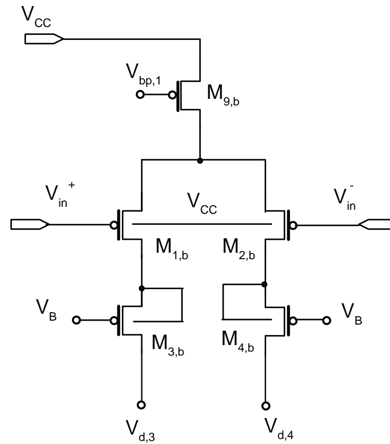


FIGURE 3.4: Analysing the biasing conditions of the differential pair.

It is then apparent that, in the presence of the large required voltage swing of V_{in}^{-}, bias voltage V_B cannot be kept constant. An adaptive bias voltage, which tracks V_{in}^{-}, is needed.

From (3.10), the upper bound of ΔV is lower than $|V_{th4}|$ for reasonable values of V_{th12}, V_{ov12}, $|V_{th10}|$, and $|V_{ov10}|$ and, hence, the left-side inequality in (3.11) is easily met by setting $V_B = V_{in}^{-}$. When $V_B = V_{in}^{-}$, the right-side inequality in (3.11) can be met by setting $|V_{th2}| - |V_{th4}| > |V_{ov4}|$, which is achieved by connecting the body terminal of $M_{2,b}$ to V_{CC} and short circuiting the body and source terminals of M_4, b (the size of the latter device ensures a sufficiently low value of $|V_{ov4}|$). It is worth to point out that, in practice, the difference between V_{th2} and V_{th4} (due to the body effect) is exploited to ensure saturated operations of the two transistors biased with the same gate voltage [34]. Although the voltage swing at the drain terminal of $M_{3,b}$, V_{d3}, is very small, the same biasing scheme is used for $M_{3,b}$ for matching purposes.

Let us now turn our attention again to the whole buffer schematic in Fig. 3.3. Since a current up to 0.45 mA must be fed to the load, an n-type follower ($M_{12,b}$ biased by $M_{13,b}$) was used as a final stage. However, as already explained for the preliminary version of the buffer in Section 3.2.1.1, an n-type follower cannot be directly cascaded to the output node of the differential stage because, when the buffer operates in unity-gain negative feedback configuration ($V_{out} = V_{in}^{-}$), $M_{2,b}$

and $M_{4,b}$ may enter their triode region. Also in this version, an additional p-type follower ($M_{10,b}$ biased by $M_{11,b}$), which acts as a level shifter, was included in front of the n-type follower. When the amplifier is connected in unity-gain non inverting buffer configuration, the voltage at the drain node of $M_{4,b}$, V_{d4}, is therefore equal to:

$$\begin{aligned} V_{d4} &= V_{out} + V_{th12} + V_{ov12} - |V_{th10}| - |V_{ov10}| = \\ &= V_{in}^{+} + V_{th12} + V_{ov12} - |V_{th10}| - |V_{ov10}| \end{aligned} \tag{3.12}$$

which, as pointed out above, ensures saturated operation of $M_{4,b}$ and $M_{2,b}$ under any process and operating conditions for reasonable values of V_{th12}, $|V_{th10}|$, V_{ov12}, and $|V_{ov10}|$.

The SR and the UGF of the amplifier were set following the criteria explained in Section 3.2.1.1 for the preliminary version of the buffer and using equations (3.6) and (3.7), respectively, as follows

$$SR = \frac{I_9}{C_4} > \frac{0.8\ V_{ampl,max}}{t_r\ \ (t_f)} = \frac{3.6}{15}\ \frac{V}{ns} = 0.24\ \frac{V}{ns} \tag{3.13}$$

$$\frac{5}{t_r\ \ (t_f)} < UGF = \frac{g_{m2}}{C_4} \leq \frac{1}{10}\omega_{p2} = \frac{1}{10}\ \frac{g_{m12} + \frac{1}{R_{load}}}{C_{out}} \tag{3.14}$$

In (3.13), I_9 is the biasing current for the input stage and C_4 is the overall capacitance at node D_4.

The current flowing through the n-type follower is mirrored through $M_{24,b}$, which delivers its drain current to a pad, I_{read}. This current can therefore be read by an external equipment without adversely affecting the pulse applied to the memory cell. Transistor $M_{26,b}$ in the rightmost branch subtracts a current equal to I_{13} from the mirrored current ($M_{26,b}$ was included for matching purposes). This way, a current equal to the programming current through the memory cell is delivered to the output, provided that all devices are adequately matched.

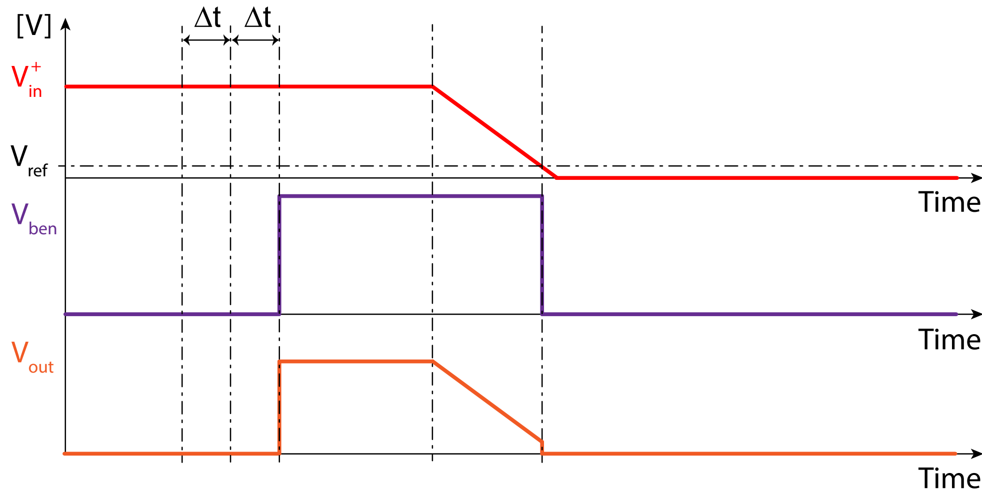


FIGURE 3.5: Timing diagram of the buffer (final version).

An enable signal (V_{ben}, active high) enables the buffer when the input voltage, V_{in}^{+}, has to be applied to the load. When V_{ben} is low, the gate of $M_{12,b}$ is shorted to ground through $M_{22,b}$ and, hence, the output node of the buffer is grounded through $M_{13,b}$ (Fig. 3.5). In the on-chip pulse generator, a circuit (not shown in Fig. 3.3) automatically drives V_{ben} low when the voltage applied to the buffer input gets lower than 400 mV, thus ensuring the control of the output voltage over the whole input voltage range 0 V to 4.5 V. This solution prevents any risk of applying unintentional voltages to the memory cell.

The section on the left side of the scheme in Fig. 3.3 generates the bias voltages for the buffer. This bias circuitry is controlled by current I_{bias} (nominal value 100 μA).

3.2.2 Delay time processing

The transistor-level schematic of the circuit designed to obtain delay time $2\Delta t$ is illustrated in Fig. 3.6.

Time delay $2\Delta t$ is obtained starting from the (intentional) skew Δt between V_{tf} and V_{td}. To this end, a capacitor, $C_{\Delta t}$, is first charged and then discharged with two nominally equal currents, $I_{\Delta t}^{ch}$ and $I_{\Delta t}^{dch}$, respectively, through the following procedure. At the beginning, switches $S_{1,\Delta t}$ and $S_{2,\Delta t}$ are off and a capacitor $C_{\Delta t}$

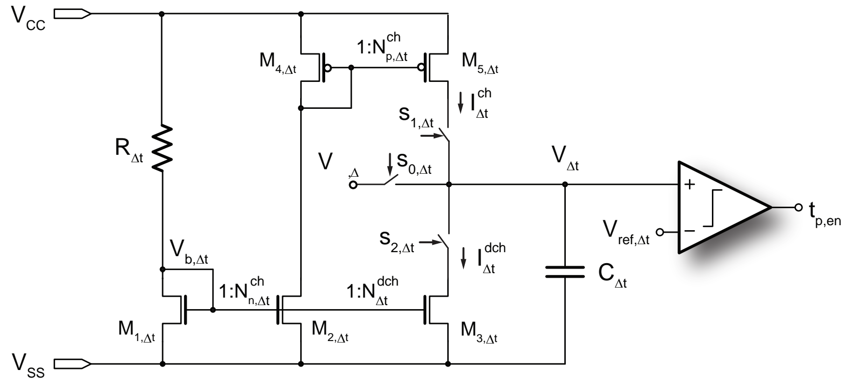


FIGURE 3.6: Circuit scheme for delay time processing.

is initially precharged at a predetermined low voltage $V_{ref,\Delta t}$ (switch $S_{1,\Delta t}$ on). This precharge level allows a better operation of the comparator which generates $t_{p,en}$. After the rising edge of V_{tf} is detected, switch $S_{0,\Delta t}$ is turned off and switch $S_{1,\Delta t}$ is turned on (Fig. 3.7). $M_{5,\Delta t}$ starts charging $C_{\Delta t}$ wit a current $I^{ch}_{\Delta t}$ equal to

$$I^{ch}_{\Delta t} = \frac{V_{CC} - V_{b,\Delta t}}{R_{\Delta t}} N^{ch}_{\Delta t} \tag{3.15}$$

where $N^{ch}_{\Delta t} = N^{ch}_{n,\Delta t} N^{ch}_{p,\Delta t}$ is the mirror factor of the charging circuitry ($N^{ch}_{n,\Delta t}$ and $N^{ch}_{p,\Delta t}$ being the mirror factors of current mirrors $M_{1,\Delta t}$, $M_{2,\Delta t}$ and $M_{4,\Delta t}$, $M_{5,\Delta t}$, respectively), $V_{b,\Delta t}$ is the gate voltage of $M_{1,\Delta t}$, and $R_{\Delta t}$ is an n-well resistor.

The voltage, $V_{\Delta t}$, across capacitor $C_{\Delta t}$ therefore increases at a constant rate. When, after a time interval Δt, the rising edge of V_{td} takes place, $V_{\Delta t}$ has reached a value $V_{\Delta t_fin}$ equal to

$$V_{\Delta t_fin} = V_{ref,\Delta t} + \frac{I^{ch}_{\Delta t}}{C_{\Delta t}} \Delta t \tag{3.16}$$

At this instant $t = \Delta t$, $S_{1,\Delta t}$ is turned off and $S_{2,\Delta t}$ is turned on, so that $M_{3,\Delta t}$ starts discharging $C_{\Delta t}$ with a current

$$I^{dch}_{\Delta t} = \frac{V_{CC} - V_{b,\Delta t}}{R_{\Delta t}} N^{dch}_{\Delta t} \tag{3.17}$$

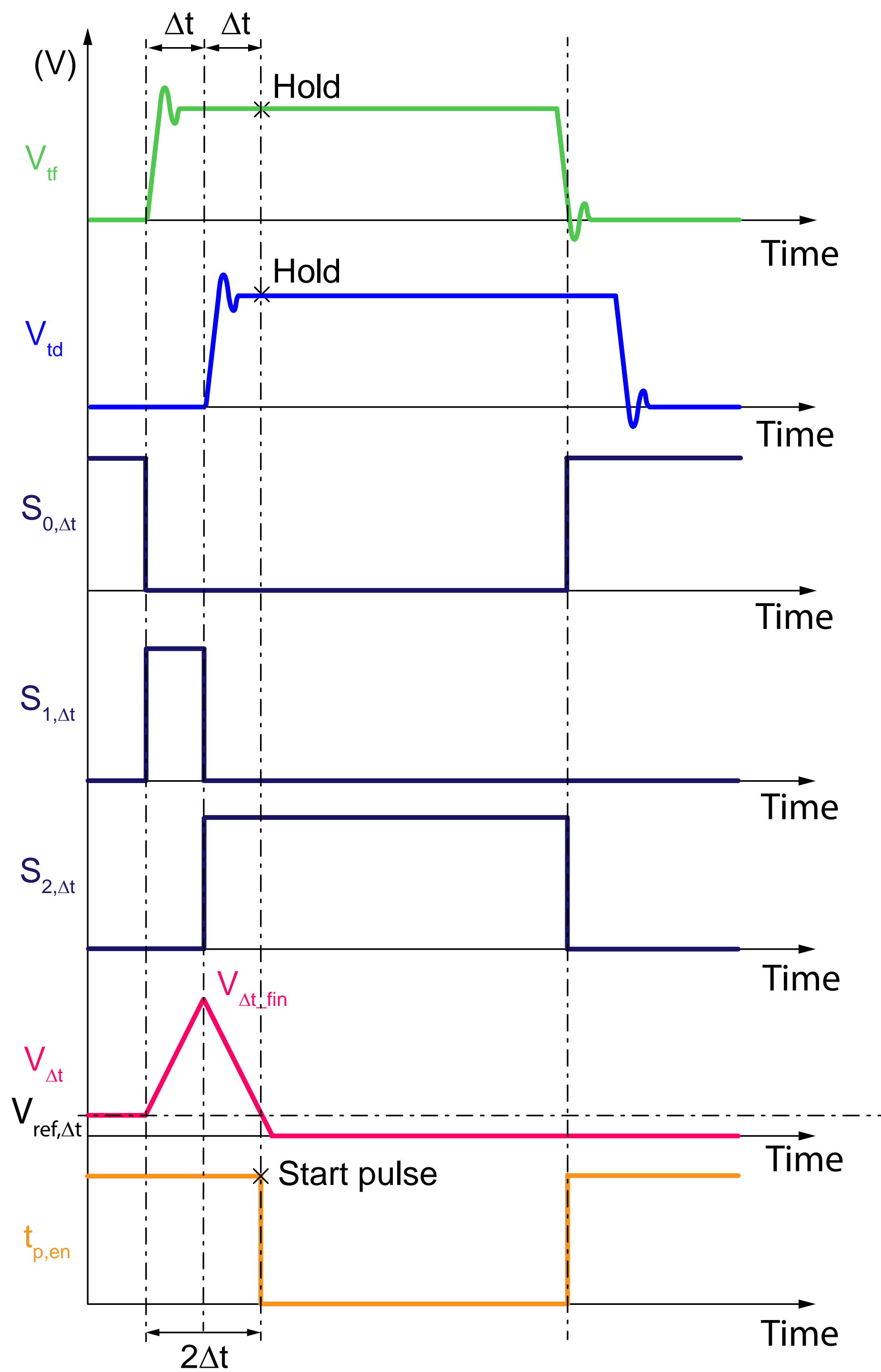


FIGURE 3.7: Timing diagram of the processing of time delay Δt.

where $N_{\Delta t}^{dch}$ is the mirror factor of the discharging circuitry (current mirror $M_{1,\Delta t}$, $M_{3,\Delta t}$).

Voltage $V_{\Delta t}$ then decreases at an ideally constant rate and reaches $V_{ref,\Delta t}$ after an additional time delay Δt_{dch}

$$\Delta t_{dch} = \frac{I_{\Delta t}^{ch}}{I_{\Delta t}^{dch}} \Delta t \tag{3.18}$$

From (3.18), $\Delta t_{dch} \cong \Delta t$ to a first order. When $V_{\Delta t}$ decreases below $V_{ref,\Delta t}$, the output voltage of the comparator, $t_{p,en}$, is driven low, thus enabling the output buffer, which makes V_{out} rise from 0 V to V_{ampl}, and starting the holding phase of voltage Track-and-Hold circuits. A simple logic forces the comparator output to V_{CC} when precharging $C_{\Delta t}$ to $V_{ref,\Delta t}$ and during the first part of the charging phase of this capacitor, so as to prevent signal $t_{p,en}$ from being unintentionally driven low.

In the proposed circuit, $V_{\Delta t}$ must never reach too high levels, so as to ensure that $M_{5,\Delta t}$ operates in its saturation region under any operating and fabrication process conditions. The values of $I_{\Delta t}^{ch}$ and $C_{\Delta t}$ were chosen so as to meet the above constraint.

3.2.3 Pulse time duration generation

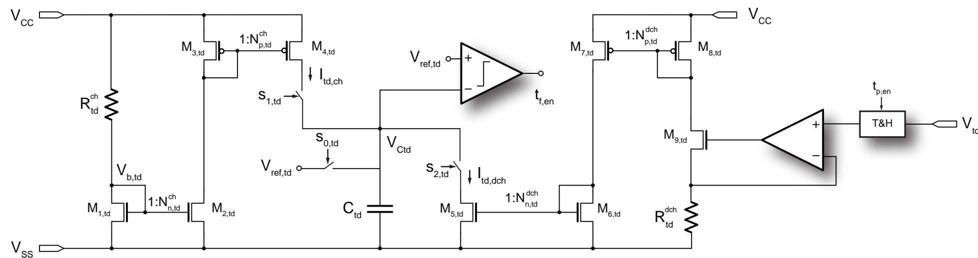


FIGURE 3.8: Circuit schematic for time duration t_d: currents for charging (I_{Ctd}) and discharging (I_{td}) capacitor C_{Ctd}.

Fig. 3.8 shows the circuit scheme developed to generate the pulse time duration, whereas Fig. 3.9 illustrates the corresponding timing diagram.

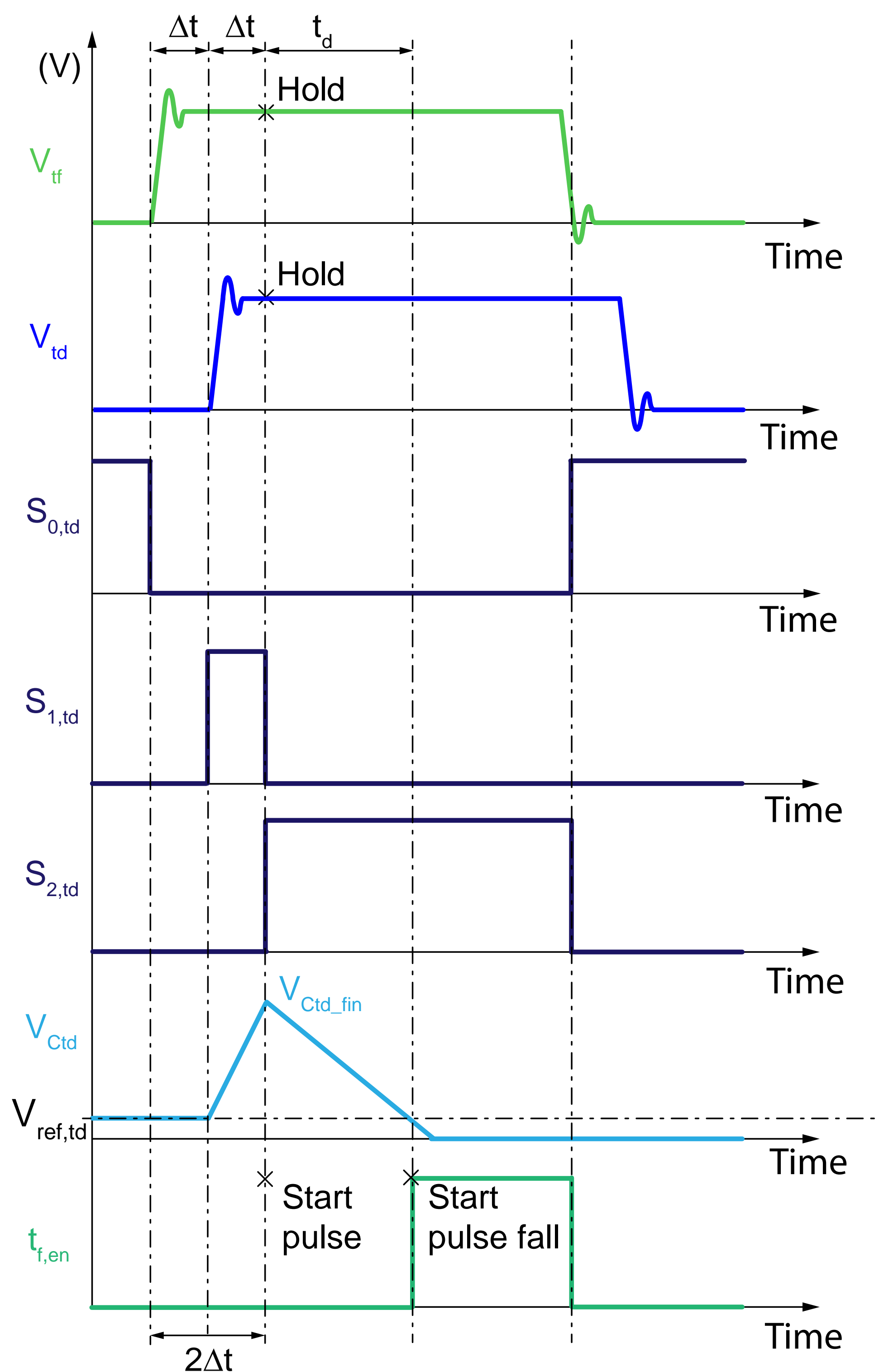


FIGURE 3.9: Timing diagram for pulse time duration (t_d) generation.

The programming pulse duration, t_d, is obtained by means of voltage-to-time conversion which, as in the circuit for determining time delay $2\Delta t$, is obtained by first charging and then discharging a capacitor by means of predetermined constant currents. Initially, a capacitor, C_{td}, is precharged to a voltage level $V_{ref,td}$ (switches $S_{1,td}$ and $S_{2,td}$ are off, whereas switch $S_{0,td}$ is on). At $t = \Delta t$, switch $S_{0,td}$ is turned off and switch $S_{1,td}$ is turned on, thus allowing $M_{4,td}$ to charge C_{td} at a constant rate, during the time interval from Δt to $2\Delta t$, with current I_{td}^{ch}, which is equal to

$$I_{td}^{ch} = \frac{\Delta V}{R_{td}^{ch}} N_{td}^{ch} = \frac{V_{CC} - V_{b,td}}{R_{td}^{ch}} N_{td}^{ch} \tag{3.19}$$

where $V_{b,td}$ is the gate voltage of transistor $M_{1,td}$, R_{td}^{ch} is an n-well resistor, and $N_{td}^{ch} = N_{n,td}^{ch} N_{p,td}^{ch}$ (with obvious meaning of symbols $N_{n,td}^{ch}$ and $N_{p,td}^{ch}$) is the mirror ratio of the charging circuitry.

When $t = 2\Delta t$, the voltage, V_{Ctd}, across C_{td} reaches a final voltage level

$$V_{Ctd_fin} = V_{ref,td} + \frac{I_{td}^{ch}}{C_{td}} \Delta t \tag{3.20}$$

At this instant, switch $S_{1,td}$ is turned off and switch $S_{2,td}$ is turned on. $M_{5,td}$ begins discharging C_{td} with current I_{td}^{dch}, which is proportional to V_{td} according to the following relation:

$$I_{td}^{dch} = \frac{V_{td}}{R_{td}^{dch}} N_{td}^{dch} \tag{3.21}$$

where R_{td}^{dch} is an n-well resistor matched to R_{td}^{ch} and $N_{td}^{dch} = N_{p,td}^{dch} N_{n,td}^{dch}$ (with obvious meaning of symbols $N_{n,td}^{dch}$ and $N_{p,td}^{dch}$) is the mirror ratio of the discharging circuitry.

Voltage V_{Ctd} will decrease at a constant rate and it will reach $V_{ref,td}$ after a time interval

$$t_d = \frac{I_{td}^{ch}}{I_{td}^{dch}} \Delta t = \frac{R_{td}^{dch}}{R_{td}^{ch}} \frac{V_{CC} - V_{b,td}}{V_{td}} \frac{N_{td}^{ch}}{N_{td}^{dch}} \Delta t \tag{3.22}$$

A simple comparator is able to detect the time instant t_d, thus starting the discharge of capacitor C_{ampl} (by setting signal $t_{f,en}$ high) and, hence, determining the length of the generated pulse. As in the case of the circuit in Fig. 3.6, a simple logic is provided to prevent the comparator output from being driven high when precharging C_{td} and during the charging phase of this capacitor.

From (3.22), pulse duration t_d is dependent on ratio $\frac{R_{td}^{dch}}{R_{td}^{ch}}$, which in principle allows minimizing any problem due to fabrication process spreads.

Passive integrators were chosen in this implementation, although they are less accurate than active integrators, so as to minimize silicon area. It will be shown in Section 3.3 that the developed calibration procedure is able to overcome the ensuing inaccuracy issue.

3.2.4 Pulse fall time

FIGURE 3.10: Circuit schematic for fall time generation.

The desired programming pulse fall time is obtained by converting the amplitude of signal V_{tf} into a current, I_{tf}, which discharges C_{ampl} at a constant rate after the programming pulse time duration is over, thus controlling the fall slope

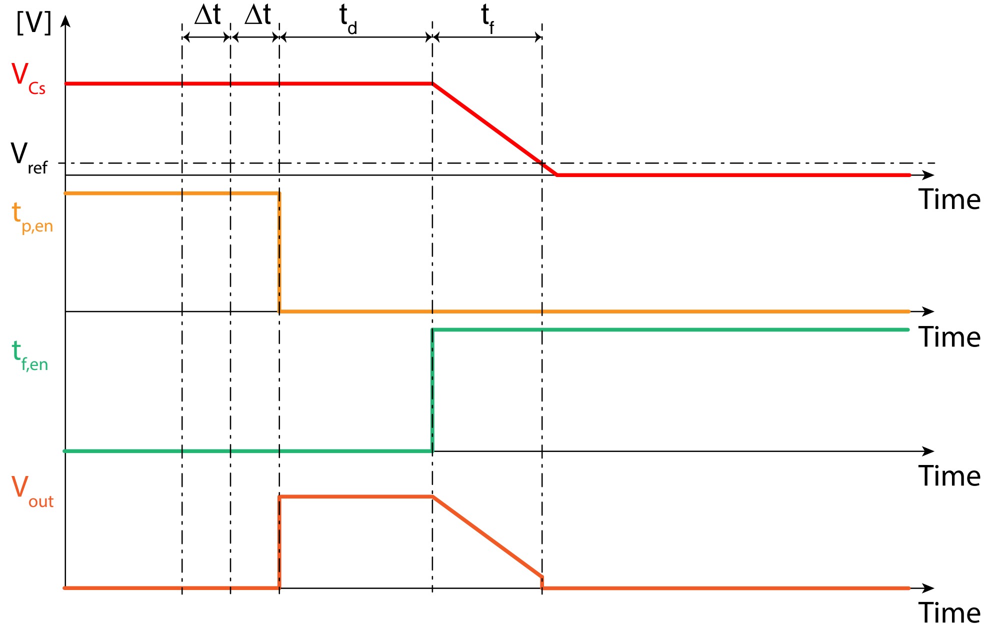


FIGURE 3.11: Timing diagram for pulse fall time (t_f) generation.

(Figures 3.10 and 3.11). I_{tf} is generated by an MOS transistor, $M_{1,tf}$, operated in the saturation region under the control of voltage V_{tf}

$$I_{tf} = \frac{\beta_{1,tf}}{2}(V_{tf} - V_{th1,tf})^2 N_{p,tf} N_{n,tf} \tag{3.23}$$

where $V_{th1,tf}$ is the threshold voltage of transistor $M_{1,tf}$, $N_{p,tf}$ and $N_{n,tf}$ are the mirror ratios of the PMOS and the NMOS current mirror, respectively, and $\beta_{1,tf}$ has the usual meaning.

To a first order, the pulse fall time is given by

$$t_f = 0.8\frac{C_{ampl}}{I_{tf}}V_{ampl} = 0.8\frac{2\ C_{ampl}}{\beta_{1,tf}}\frac{V_{ampl}}{(V_{tf} - V_{th1,tf})^2} \tag{3.24}$$

where the factor 0.8 accounts for the usual definition of fall time, which is measured from 90% to 10% of the output voltage swing.

When the size of the MOS transistor $M_{1,tf}$ was chosen, the key target was obtaining a moderate value of its transconductance, which implies moderate sensitivity of

current I_{tf} to input voltage variations (this feature is important so as to reduce effects due to disturbances).

When the size of device $M1_{tf}$ was chosen, the key target was obtaining a moderate value of its transconductance, which implies moderate sensitivity of current I_{tf} to input voltage variations (this feature is important so as to reduce effects due to disturbances). As a consequence, considering the required range of the converted current, the allowed range of input signal V_{tf} has to be kept as wide as possible.

Equations (3.22) and (3.24) provide the relationships between control voltages V_{td} and V_{tf} and target pulse parameters t_d and t_f. Once design and process parameters are known, the values of the control voltages to be fed to the on-chip pulse generator in order to obtain the desired values of t_d and t_f can be easily determined by inverting the two above equations.

3.3 Manual calibration procedure

As already explained in Chapters 1 and 2, despite the care taken during design to limit the sensitivity of the system to process spreads and component non-idealities, still the processing of the control variables (Δt, t_d, and t_f) and the uncertainties of the ATE introduce some inaccuracies over the timing parameters (t_d and t_f) of the generated pulse. Indeed, the values of timing parameters as given by equations (3.18), (3.22), and (3.24) are affected by process spreads and mismatches. In addition, the above equations are based on first-order models for the used active and passive components, whose real behaviour will therefore result in deviations in the values of the generated pulse parameters. Further inaccuracy is contributed by non-idealities such as comparator offset and operational amplifier offset and finite DC gain. Finally, unavoidable ATE uncertainties also result in inaccuracies in the generated control voltages.

As multiple inaccuracy sources affect the parameters of the generated pulse, characterizing single components for the specific chip under test to obtain their actual

I-V characteristics and, then, adjust the control variables so as to achieve the required pulse accuracy is not practically feasible. A dedicated calibration procedure should therefore be conceived to obtain a global compensation of all non-ideal effects, including inaccuracy contributions due the used instrumentation.

In the following of this Section, the above aspects aiming at developing the required calibration procedure are analysed.

Thanks to the calibration procedure, for any chip under test, the control voltages to be applied to the pulse generator are adjusted so as to compensate for its specific non-idealities, thus obtaining the desired pulse timing parameters.

In fact, the actual relationships between the control variables (V_{td}, V_{tf}, Δt) and the timing parameters (t_d, t_f) of the generated pulse are experimentally estimated for the specific chip under test: for any desired pulse to be generated during the parametric test of the cells in the considered chip, the external instrumentation will then be programmed to feed the corresponding values of the control variables V_{td} and V_{tf} (for the chosen Δt).

To develop the proposed calibration procedure, the three equations (3.18), (3.22), and (3.24) were analysed taking non-idealities into account. These equations were manipulated so as to obtain simple equations where non-ideal contributions are grouped in a number n of unknowns. For any chip under test, the values of these unknowns must be determined, so that the values of control inputs V_{td} and V_{tf} required to obtain the desired values of timing parameters t_d and t_f can be found by inverting the simplified equations corresponding to (3.22) and (3.24). For this purpose, first an adequate number of pulses are generated with the chip under the control of predetermined values of input signals and the values of their real timing parameters t_d and t_f are measured, then the measured values are substituted in the simplified equations and, finally, the ensuing equation system is solved for the unknowns. Simplifying assumptions were made so as to reduce n to a minimum, in order to reach a good trade-off between ease and speed of implementation in a test sequence on the one hand and accuracy on the other hand, bearing in mind

that a pulse accuracy within ±10% in the timing parameters of the generated programming pulse is required.

The proposed calibration procedure would require the ATE to measure times, which is not possible. In order to overcome this issue, timing parameters should be converted on-chip to voltages, so as to avoid the need for time measurements. This feature is implemented in the final version of the system and it will be therefore dealt with in Chapter 4. In Chapter 5, the effectiveness of the calibration procedure is demonstrated by using an active microprobe and an oscilloscope to measure timing parameters.

3.3.1 Calibration equations

In the following, the dependence of each parameter in equations (3.22) and (3.24) upon process spreads, mismatches, and operating conditions are considered (different operating conditions refer to different circuit operation as a consequence of different values of Δt and/or analog control voltages). Parameters depending only on spreads and/or mismatches are considered constant within any test sequence carried out on a given die. Parameter dependence on operating conditions are taken into account in deriving calibration equations when these conditions may change during a test sequence.

Supply voltage V_{CC} and temperature are considered constant, as the proposed system is intended for use in factory under controlled supply voltage and environment conditions. Time reference Δt is also considered constant during a test sequence (or even during a whole test session).

3.3.1.1 Calibration equation for delay time

Since t_d depends on Δt [eq. (3.22)], non-idealities in processing the latter variable (Fig. 3.6) were first considered, thus obtaining Table 3.1.

TABLE 3.1: Non-idealities in Δt processing (Legend: S. = process spreads, M. = mismatches, O. C. = operating conditions)

Parameter	S.	M.	O. C.	Comments
$N^{ch}_{n,\Delta t}$	X	X		
$N^{ch}_{p,\Delta t}$	X	X	X	Depends on $V_{\Delta t_fin}$ and, hence, on Δt
$N^{dch}_{\Delta t}$	X	X	X	Depends on $V_{\Delta t_fin}$ 0 and, hence, on Δt
$R_{\Delta t}$	X			
$C_{\Delta t}$	X			
$V_{b,\Delta t}$	X			
$V_{os,comp,\Delta t}$		X		

In addition to spreads and mismatches that affect current mirrors (the spread in the channel modulation effect over mirror ratios $N^{ch}_{n,\Delta t}$, $N^{ch}_{p,\Delta t}$, and $N^{dch}_{\Delta t}$ was also taken into account), $C_{\Delta t}$, $R_{\Delta t}$, and $V_{b,\Delta t}$, thus impacting over the ratio $\frac{I^{ch}_{\Delta t}}{I^{dch}_{\Delta t}}$, the offset, $V_{os,comp,\Delta t}$, of the comparator in Fig. 3.6 must be considered. Equation (3.18) then becomes

$$\Delta t_{dch} = \frac{I^{ch}_{\Delta t}}{I^{dch}_{\Delta t}}\Delta t + \frac{C_{\Delta t}}{I^{dch}_{\Delta t}}V_{os,comp,\Delta t} \tag{3.25}$$

However, in our system, even when considering a value as high as 10 mV for $V_{os,comp,\Delta t}$, the second term on the right side of (3.25) has a maximum impact of 0.7% on Δt_{dch} and can therefore be neglected. In addition, it should pointed out that mirror ratios $N^{ch}_{p,\Delta t}$ and $N^{dch}_{\Delta t}$ can be considered constant within a test sequence as Δt and, hence, $V_{\Delta t_fin}$ are constant.

Δt_{dch} can be thus expressed as

$$\Delta t_{dch} = \alpha_{\Delta t}\Delta t \tag{3.26}$$

TABLE 3.2: Non-idealities in t_d generation (Legend: S. = process spreads, M. = mismatches, O. C. = operating conditions)

Parameter	S.	M.	O. C.	Comments
$N^{ch}_{n,td}$	X	X		
$N^{ch}_{p,td}$	X	X	X	Depends on V_{Ctd_fin}
$N^{dch}_{p,td}$	X	X	X	Depends on V_{td}
$N^{dch}_{n,td}$	X	X	X	Depends on V_{Ctd_fin}
$\frac{R^{dch}_{td}}{R^{ch}_{td}}$	X	X	X	R^{dch}_{td} depends on V_{td}
C_{td}	X			
$V_{b,td}$	X			
$V_{er,opamp,td}$	X	X		
$V_{os,comp,td}$		X		

where $\alpha_{\Delta t}$ accounts for mismatches between $I^{dch}_{\Delta t}$ and $I^{ch}_{\Delta t}$ and non-idealities in charging and discharging operations, and is substantially constant during a test sequence.

3.3.1.2 Calibration equation for time duration

Then, equation (3.22) was considered, referring to Fig. 3.8: non-idealities affecting this equation are summarized in Table 3.2, where the errors of the operational amplifier ($V_{er,opamp,td}$, due to its offset and finite DC gain) and the comparator (offset voltage $V_{os,comp,td}$) are also included.

Equation (3.22) can therefore be rewritten as

$$t_d = \begin{array}{l} \frac{R^{dch}_{td}}{R^{ch}_{td}} \frac{V_{CC}-V_{b,td}}{V_{td}-V_{er,opamp,td}} \frac{N^{ch}_{td}}{N^{dch}_{td}} \alpha_{\Delta t} \Delta t + \\ + C_{td} \frac{R^{dch}_{td}}{V_{td}-V_{er,opamp,td}} V_{os,comp,td} \frac{N^{ch}_{td}}{N^{dch}_{td}} \end{array} \tag{3.27}$$

If we set $V_{os,comp,td} = 10$ mV and consider a 2% error of the operational amplifier, which are very relaxed specifications, the worst-case impact of these two non-idealities on t_d in our test-chip is equal to 2.3%, which is well within the target specification and can therefore be neglected.

Ratios $N_{p,td}^{ch}$ and $N_{n,td}^{dch}$, which are affected by channel length modulation effects, depend on V_{Ctd_fin}, which is determined by Δt and is therefore constant during a test sequence. Ratio $N_{p,td}^{dch}$ depends on V_{td} due to channel length modulation: indeed, V_{td} controls the current through R_{td}^{dch} and, hence, affects the drain voltage of $M_{8,td}$. However, the variation in this voltage is very small, which in turn causes negligible variation in $N_{p,td}^{dch}$.

The ratio $\frac{R_{td}^{dch}}{R_{td}^{ch}}$, instead, depends on V_{td}: indeed, as a consequence of the dependence of the resistance of n-well resistors upon the applied voltage, R_{td}^{dch} depends on V_{td} whereas R_{td}^{ch} does not. As the above dependence is linear to a first approximation, we can write

$$\frac{R_{td}^{dch}}{R_{td}^{ch}} = m'V_{td} + q' \tag{3.28}$$

where m' and q' are two coefficients affected by process spreads.

Grouping all constant (and quasi-constant) parameters in an unknown, α_{td}, equation (3.27) becomes

$$t_d = \frac{\alpha_{td}(m'V_{td} + q')\Delta t}{V_{td}} \tag{3.29}$$

The above equation can be rewritten as

$$t_d = (m + \frac{q}{V_{td}})\Delta t \tag{3.30}$$

where $m = \alpha_{td}m'$ and $q = \alpha_{td}q'$. Two unknowns, namely m and q, must therefore be determined.

3.3.1.3 Calibration equation for fall time

Non-idealities affecting the generation of t_f (Fig. 3.10) will be now discussed.

TABLE 3.3: Non-idealities in t_f generation (Legend: S. = process spreads, M. = mismatches, O. C. = operating conditions)

Parameter	S.	M.	O. C.	Comments
$N_{n,tf}$	X	X	X	Depends on V_{tf} and V_{ampl}
$N_{p,tf}$	X	X	X	Depends on V_{tf}
μ	X		X	Depends on V_{tf}
C_{ox}	X			
$(\frac{W}{L})_{1,tf}$	X			
$V_{th1,tf}$	X			
C_{ampl}	X			

As shown in Table 3.3, parameters that depend on operating conditions are mirror factors $N_{p,tf}$ and $N_{n,tf}$, which are affected by channel length modulation effects, and mobility μ. Their variation over the whole operating range of V_{tf} was therefore analysed.

Channel length modulation effects in current mirror $M_{2,tf}$, $M_{3,tf}$ are very small as a consequence of the very small variation in the drain voltage of these two transistors and hence $N_{p,tf}$ can be considered substantially constant. Channel length modulation effects of $M_{5,tf}$ depend on voltage V_{Campl} and on the channel length modulation parameter, λ_n. Assuming $\lambda_n = 0.02\ V^{-1}$, the impact of the channel length modulation on t_f turns out to be within ±1.9%. $N_{n,tf}$ can then be considered substantially constant and set equal to its mid-range value. Finally, the variation of μ in the range of interest has an impact within ±3% on t_f, assuming that μ follows its conventional equation

$$\mu = \frac{\mu_0}{1 + \vartheta(V_{tf} - V_{th1,tf})} \tag{3.31}$$

where a typical value for ϑ is $0.075\ V^{-1}$. Also parameter μ can therefore be considered constant and set equal to its mid-range value.

Including all constant (and quasi constant) parameters except $V_{th1,tf}$ in an unknown, α_{tf}, we find the calibration equation for t_f, which is

$$t_f = \frac{0.8 \ V_{ampl} \ \alpha_{tf}}{(V_{tf} - V_{th,n})^2} \tag{3.32}$$

Equation (3.32) shows that there are two unknowns to be found, namely α_{tf} and $V_{th,n}$.

Based on equations (3.30) and (3.32), the calibration procedure illustrated in next Subsection has been developed.

3.3.2 Calibration procedure

The proposed calibration procedure is illustrated in Fig. 3.12.

Once the number n of unknowns in the calibration equation for t_d (t_f) is determined (in our case, $n = 2$ for both equations), n pulses are generated. The voltage levels of the input signals for these pulses are chosen so as to adequately evaluate the real system response over the whole specified range of the parameter under consideration (t_d, t_f).

The values of the timing parameter t_d (t_f) of the n generated pulses are measured and then substituted for the corresponding variable in (3.30) [(3.32)], thus obtaining a system of n equations. The system is solved so as to find the values of the unknowns. The obtained values are then substituted for the unknowns in (3.30) [(3.32)]. These equations are finally inverted in order to express V_{td} (V_{tf}) as a function of t_d (t_f), thus allowing us to obtain the control parameters to be used during normal test mode (parametric test and characterization of memory cells).

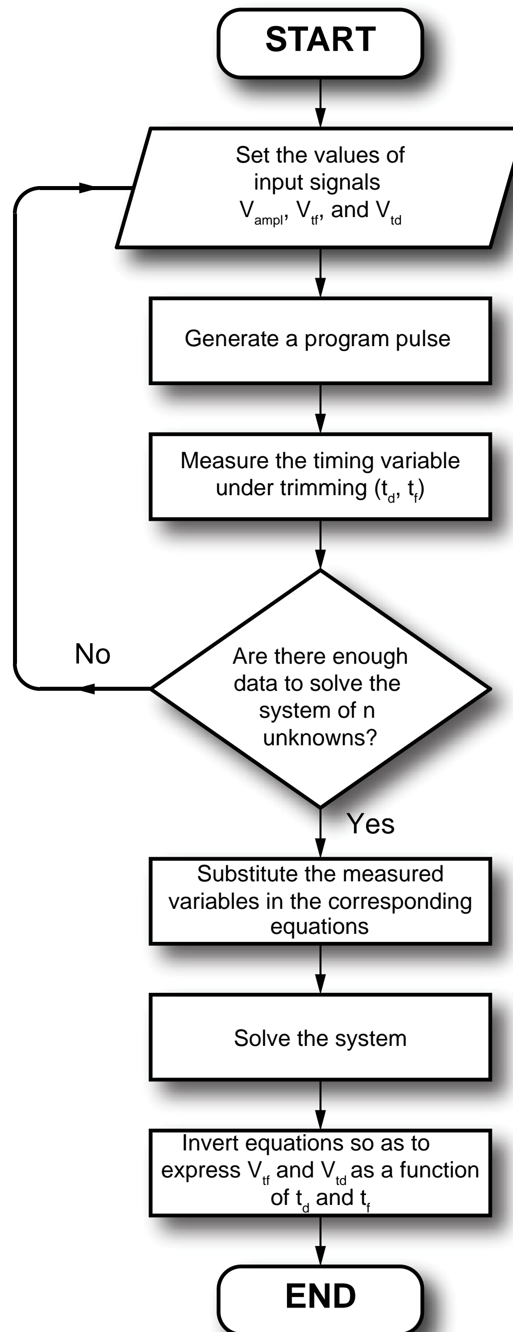


FIGURE 3.12: Flowchart describing the calibration procedure algorithm.

Chapter 4

Final implementation

4.1 Introduction

The system described in Chapter 3 was designed focusing on the core of the system for debugging purposes rather than accounting for the efficiency of the interface between the ATE and the system. This leads to the development of a new system which both optimizes the previous one and implements an efficient interface with the ATE. In the following of this Chapter, an overview of the new system is provided together with a comparison between the two versions of the system. The final implementation is then described in detail, as well as the developed automatic calibration procedure. Last Section, finally, present a simplified version of the system, conceived to be used with commercial ATE which does not suffer from disturbances due to interconnection cables.

4.2 High-level description

A high-level scheme of the system is illustrated in Fig. 4.1.

As pointed out in Chapter 2, the external signals provided by the ATE are used in the system in order to generate the desired programming pulse, which is applied

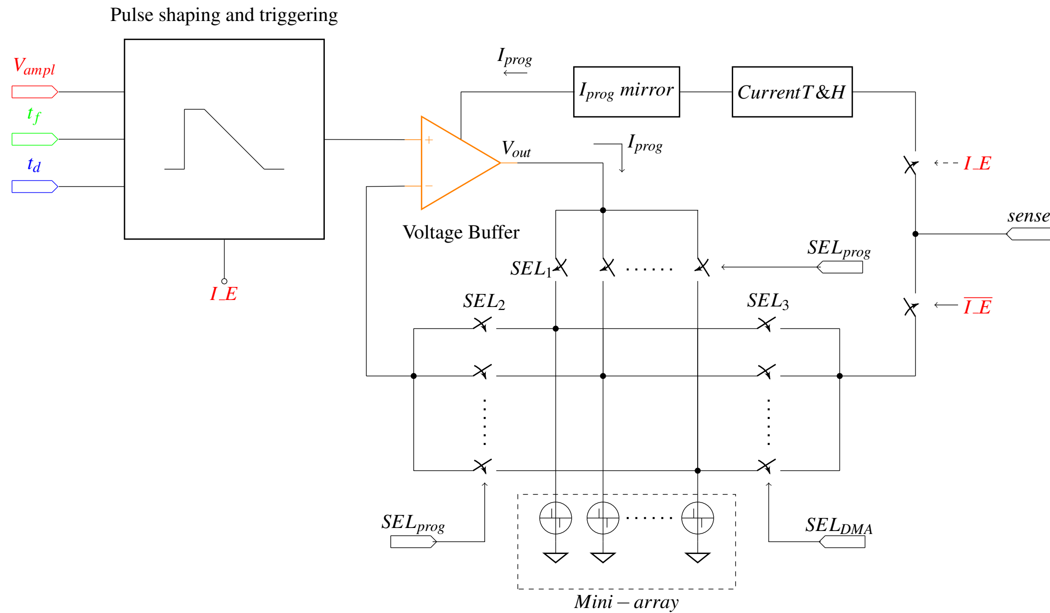


FIGURE 4.1: High-level scheme of the system for PCM cell characterization.

to the cell through an output buffer, which can feed the cell with the amount of current needed for program operation. A current mirror replicates the programming current, which is first tracked and then held by a dedicated circuit for the time necessary to allow external measurement by the APWTS. The APTWS can also read the state of the selected cell in DMA mode when required. In order to get rid of the transients of external signals, the program pulse is available after a time interval equal to twice an externally programmable delay Δt.

The main difference between this version and the one described in Chapter 3 is the interface with the test equipment. In fact, an internal switch allows both reading the programming current and reading the cell resistance from the same channel of the commercial instrument. This feature allows long sequences of repeated program-and-read cycles to be performed without using the external switch matrix, thus significantly decreasing the execution time of test operations.

Moreover, an interface that allows the programming current to be read by an ATE was integrated. Essentially, a current mirror replicates the programming current, which is first tracked and then held by a dedicated circuit for the time necessary to allow external measurement by the APWTS.

4.3 Basic operating principle

4.3.1 Pulse generation

According to the first implementation, the information about pulse fall time and pulse time duration is provided by the PG, and pulse amplitude is provided by the APWTS. The basic operating principle of the pulse generation has already been widely described in Chapters 2 and 3, so it will not be described in detail in this Section.

- Pulse amplitude is encoded by the amplitude of signal V_{ampl}. Its voltage amplitude exactly represents the program pulse amplitude and is stored in a capacitor, C_{ampl}.
- Pulse fall time is programmed by signal V_{tf}. Its voltage amplitude is converted into a current, I_{tf}, which discharges C_{ampl}. The time that I_{tf} requires to discharge capacitor C_{ampl} represents the pulse fall time.
- Pulse time duration is programmed by signal V_{td}. Its voltage amplitude is converted into a current, I_{td}^{dch}, which is used to discharge a capacitor, C_{td} (previously charged to a voltage dependent on Δt). The time required to discharge the voltage V_{Ctd} across C_{td} represents the pulse time duration.

The time delay between the rising edges of signals V_{tf} and V_{td} represent the externally programmable delay Δt.

4.3.2 Interface with the test equipment

It has been experimentally observed that the settling time of the signals from the APWTS is much longer, about 200 μs (Chapter 1). Pulse generation must start only after all signals from APWTS reach their programmed voltage. In order to meet this requirement, a counter, which provides an enabling signal for pulse generation, is used.

As pointed out in Chapter 1, reading the programming current at the pulse plateau is fundamental to getting useful information about the cell programming performance. To this end, a current Track-and-Hold circuit was designed so as to ensure a stable output voltage for the time required by the ATE for read-out.

As already explained (Chapter 1), the external switch matrix heavily reduces testing speed. This switch has therefore to be bypassed during a test sequence such as repeated read-and-write cycles in order to save time. An internal switch controlled by the digitalized value of V_{ampl}, I_E, has been implemented to enable either a read or a program operation, thus improving testing speed.

4.3.3 Automatic calibration procedure

As pointed out in Chapters 2 and 3, even when a careful design is carried out, the achievable accuracy of the generated pulses is limited by unavoidable fabrication process spreads and non-idealities. The calibration procedure described in Chapter 3 allowed overcoming accuracy issues due to process spreads and mismatches. However, the ATE can not measure times, but it can easily read currents and voltages. A dedicated hardware, which performs an internal time-to-voltage conversion, was therefore designed and included in the system so as to be able to automatically perform the calibration procedure.

4.4 Description of the system

The system will be now described referring to the block diagram in Fig. 4.2 and the waveforms in Figs. 4.3 and 4.4. Section 4.4.1 focuses on the interfacing with the ATE, whereas Section 4.4.2 deals with the programming pulse generation.

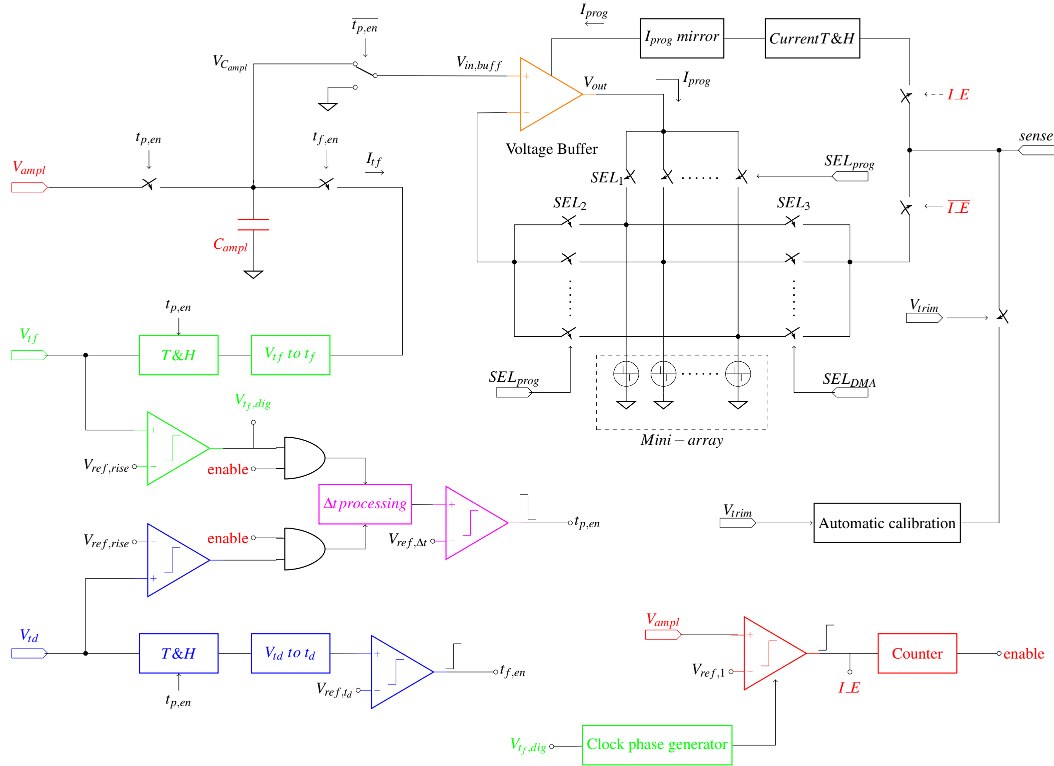


FIGURE 4.2: Detailed block diagram of the implemented system for PCM cell characterization.

4.4.1 Interface with the ATE

Channels V_{tf} and V_{td} are in free running mode during the whole test sequence. V_{td} is a train of pulses with a delay of Δt with respect to the V_{tf} train. V_{tf} acts as the counter clock for system enable signal generation.

As soon as the V_{ampl} rising edge is detected, logic signal I_E (which is initially low) becomes high and the counter counts twenty rising edges of V_{tf}. When the twenty-first rising edge of V_{tf} is detected, a signal referred to as *enable*, which enables program pulse generation, set to its high level. The program pulse is thus generated and fed to the cell. Just before the program pulse starts to fall, the current Track-and-Hold circuit is set in the hold mode and, hence, provides the programming current to the test equipment for a fine interval adequate to allow its read-out.

After the program current is read, V_{ampl} is forced low and, consequently, I_E becomes low, thus enabling the DMA mode. Since the level of V_{ampl} indicates

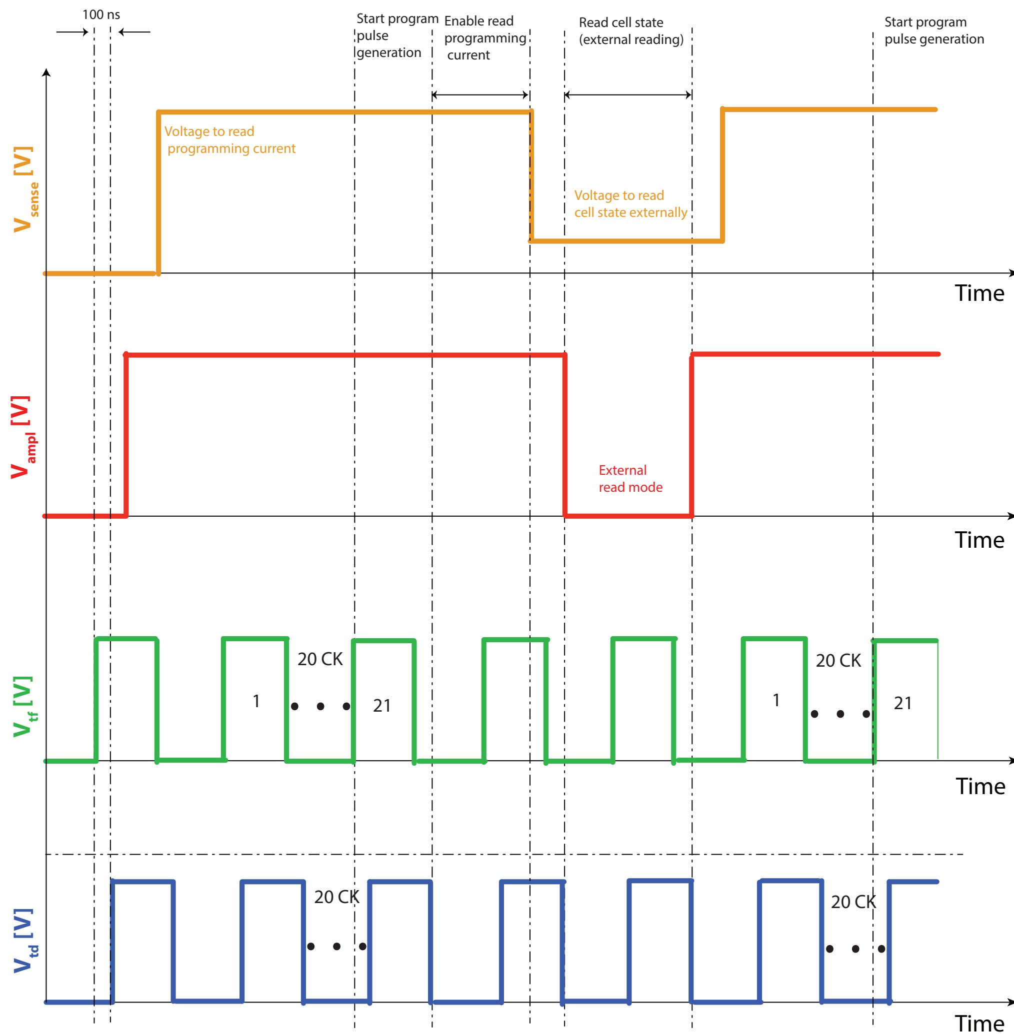


FIGURE 4.3: Example of the waveforms applied to the on-chip pulse generator during a read-and-write cycle.

which operation is being performed, it must remain at a high logic level (which implies $I_E = 1$) until the program current is read. V_{sense} performs both the program current reading and the DMA reading of the cell state. Since it has to force a voltage in the order of 2 V to read the programming current and a voltage in the order of 1 V when in DMA mode, it must be raised by the test equipment to the program current reading voltage after the V_{ampl} rising edge has been detected and must fall to the cell state reading voltage before V_{ampl} is forced to a low level (which implies $I_E = 0$).

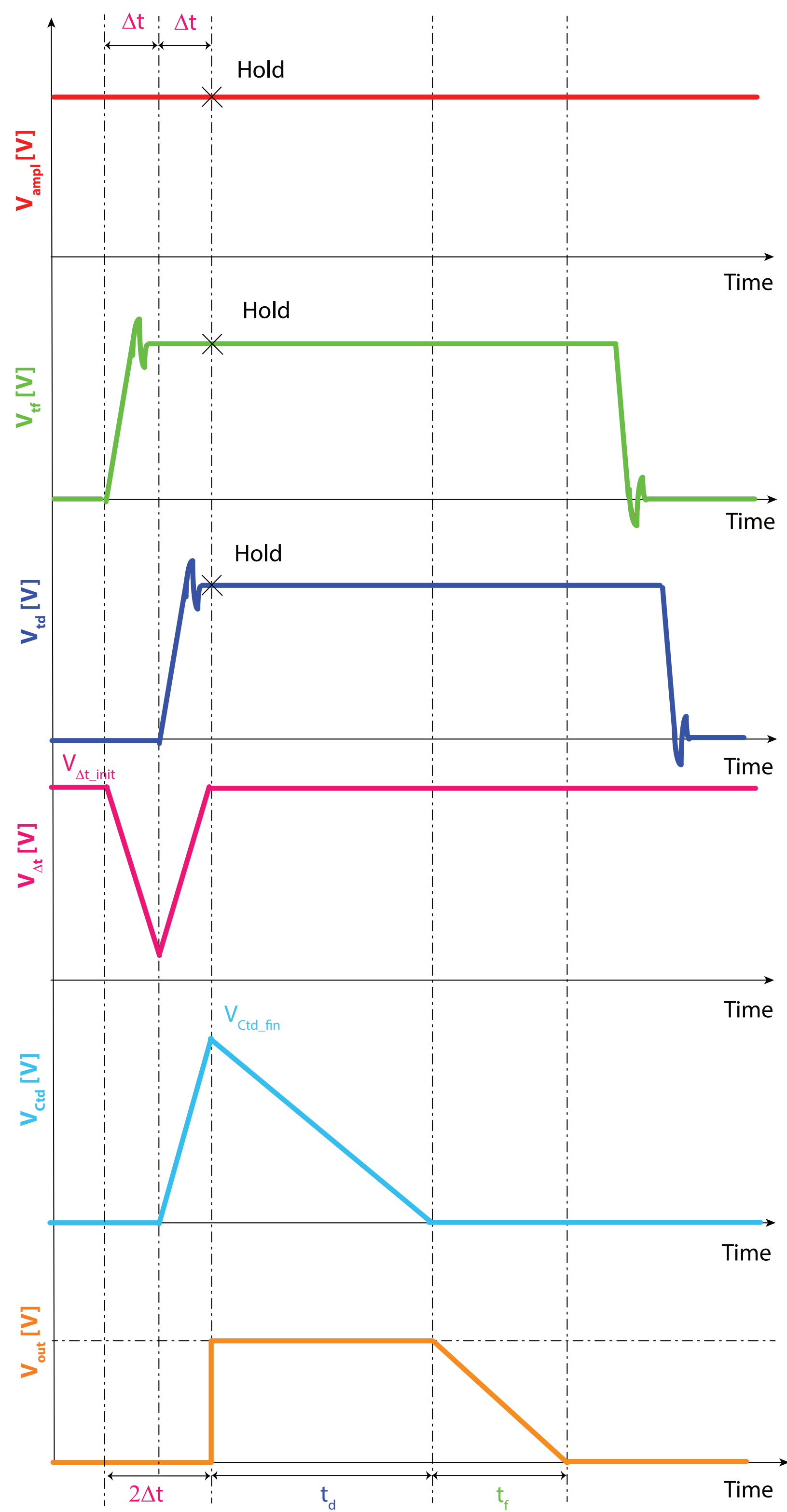


FIGURE 4.4: Example of the internal waveforms of the on-chip pulse generator during program pulse generation.

4.4.2 Pulse generation

The pulse generation scheme in this system is almost the same to the one described in Chapter 3. The main difference is that pulse generation is now conditioned by the enable signal. In the following description, we assume that the enable signal is at a high logic level.

Capacitor C_{ampl} is initially discharged. The output voltage V_{out} is initially forced to 0 V. Voltage V_{ampl} is tracked as soon as the rising edge of V_{tf} is detected. In order to hold V_{out} at 0 V when the buffer input rises from 0 V to V_{ampl}, the buffer is disabled by an internally generated signal which enables the buffer only when the generated pulse is ready to be applied to the load, i.e., after $2\Delta t$. After the rising edge of V_{tf} is detected, capacitor $C_{\Delta t}$, which is initially precharged at $V_{\Delta t_init}$ ($V_{\Delta t_init}$ = 3.5 V), is discharged with an internally generated constant current, $I_{\Delta t}^{dch}$, until the rising edge of V_{td} is detected after a time interval Δt. At this instant (t = Δt), the system begins to charge $C_{\Delta t}$ with a current $I_{\Delta t}^{ch}$ equal to $I_{\Delta t}^{dch}$. Voltage $V_{\Delta t}$ will then approach again $V_{\Delta t_init}$ after an additional time delay Δt, i.e., at t = $2\Delta t$. During the time interval from Δt to $2\Delta t$, capacitor C_{td} is also charged at a constant rate by an internally generated current I_{td}^{ch}

$$I_{td}^{ch} = \frac{V_{CC} - V_{ref2}}{R_{td}^{ch}} \tag{4.1}$$

thereby reaching a final voltage level V_{td_fin}. V_{ref2} is a constant voltage generated internally from V_{CC} by means of a resistive divider.

At t = $2\Delta t$, V_{ampl} is held across storage capacitor C_{ampl}. The output buffer is enabled and makes V_{out} rise from 0 V to V_{ampl}.

Voltages V_{td} and V_{tf} are also stored across respective capacitors C_{tds} and C_{tfs} at t = $2\Delta t$ and are then converted into respective currents

$$I_{td}^{dch} = \frac{V_{CC} - V_{td}}{R_{td}^{dch}} \tag{4.2}$$

and

$$I_{tf} = \frac{\beta}{2}(V_{tf} - V_{th})^2 \tag{4.3}$$

At t = $2\Delta t$, I_{td}^{dch} begins to discharge capacitor C_{td}. When V_{Ctd} approaches zero, capacitor C_{ampl} is connected to discharging current I_{tf}, thus giving rise to the falling slope of the pulse and, hence, determining the pulse duration

$$t_d = \frac{V_{CC} - V_{ref2}}{V_{CC} - V_{td}} \frac{R_{td}^{dch}}{R_{td}^{ch}} \Delta t \tag{4.4}$$

The value of current I_{tf} controls the falling slope of the pulse and, hence, its fall time, which is equal to

$$t_f = \frac{0.8\ C_{ampl}}{I_{tf}} V_{ampl} \tag{4.5}$$

4.5 Circuit design

4.5.1 Features and aims

The system implemented focuses on the characterization of PCM cells and the synchronization and interfacing with the external test equipment.

The main features of version, compared to the one described previously (Chapter 3), are listed in Table 4.1.

The operating principle of the system has been described in Section 4.4. Referring to Fig. 4.2 further circuital details will be given.

The voltage to deselect unaddressed Word Lines (V_{high}) can be tuned between 4 V and 5 V without affecting circuit operation, whereas the supply voltage V_{CC} is kept constant at 6 V. Biasing current I_{bias} is generated by means of an n-well resistor, R_{bias}. One of its terminals is connected either to V_{high} or to V_{CC}. During programming and current reading operations ($I_E = 1$), biasing current I_{bias} is

Table 4.1: Enhanced and added features of the new system

Feature	**Enhanced feature**	**New feature**
Delay time (Δt) generated under external control	X	
External signal V_{tf} internally converted to pulse fall time	X	
External signal V_{td} internally converted to pulse time duration	X	
Output buffer applying the programming pulse to the selected load	X	
External DMA reading of the load performed without using the external switch matrix	X	
Counter for synchronization with the external equipment		X
Track-and-hold circuit for programming current read-out		X
Hardware for automatic calibration procedure		X

derived by tuning V_{high}, as high accuracy is required; this way, I_{bias} can be tuned to the desired value by adjusting V_{high} within the allowed range. In contrast, I_{bias} is derived from supply voltage V_{CC} when the system operates in DMA mode ($I_E = 0$), because in this phase the bias current does not need to be very accurate.

All reference voltages used for the comparators (V_{ref}, V_{ref1}, V_{ref2}) are generated internally from V_{CC} by means of a resistive divider.

The external switch matrix is only used to select the cell to be characterized at the beginning of a test sequence and to select the calibration procedure, which is performed at the beginning of a test session.

Finally, in this implementation, only two cells can be selected, therefore only one addressing pad (sel_{prog} is needed.

In the following of this Section, signal functions and the correspondence between test chip pads are summarized (Table 4.2). Then, circuit details of the main blocks of the designed test chip, which allow program pulse generation, are discussed.

Table 4.2: Functions of pads

Pad	**Function**
V_{ampl}	• Pulse amplitude ($V_{ampl} \geq 2$ V) • Program mode ($V_{ampl} > 1.5$ V) • DMA mode ($V_{ampl} < 1.5$ V)
V_{tf}	• Pulse fall time (voltage-to-time conversion) • System start (rising edge $\rightarrow$ start Δt processing) • Clock (V_{tf} digitalized)
V_{td}	• Pulse time duration (voltage-to-time conversion) • Rising edge $\rightarrow$ end Δt processing
V_{CC}	• Supply voltage ($V_{CC} = 6$ V) • Biasing current generation when in DMA mode
V_{SS}	• Ground ($V_{SS} = 0$ V)
V_{high}	• Deselect unaddressed Word Lines • Biasing current generation when in program mode
sel_{prog}	• Cell selection
V_{trim}	• Calibration mode
$sense$	• Read programming current ($V_{sense} = 2.5$ V) • Read cell state ($V_{sense} = 1.2$ V) • Read voltage during calibration ($I_{sense} = 0$ A)

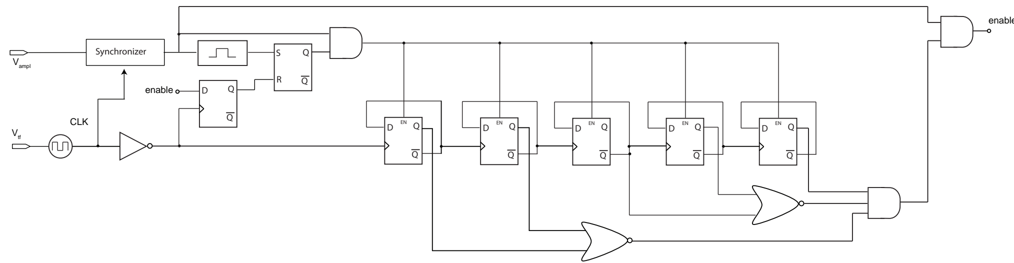


FIGURE 4.5: Counter schematic.

4.5.2 Interface with the test equipment

4.5.2.1 Enable signal generation

The generation of the enable signal for the whole system (*enable*) is proved by a digital block, which is composed by a synchronizer, a monostable, a Set-Reset latch (SR latch), and a digital counter (Fig. 4.5). The delay flip-flops employed in the counter and the SR latch are not described in detail because a standard design was used.

First, the general operation of this circuit block is described and, then, circuital details of the clock generation, the synchronizer, and the monostable are provided.

The clock is derived from signal V_{tf}. It drives the synchronizer, which detects when V_{ampl} rises above a reference voltage $V_{ref1} = 1.5$ V. Signal V_{ampl} is thus digitized. This digital signal is detected by a monostable circuit, which sets an SR latch. The SR latch enables a digital asynchronous counter, whose least significant bit is driven by the inverted clock. When the 20th clock cycle is detected, the enable signal rises from a logical "0" to a logical "1". The *enable* is used both to start the program pulse generation and to reset the SR latch after a clock cycle delay. The reset of the SR latch disables and reinitializes the counter, which is thus ready to start counting again when the next rising edge of V_{ampl} is detected.

Clock generation Non-overlapping clock phases are used to drive the synchronizer in order to avoid any risk of race. The generator of these clock phases ($CLK1$, $CLK2$, and the corresponding inverted phases) is depicted in Fig. 4.6.

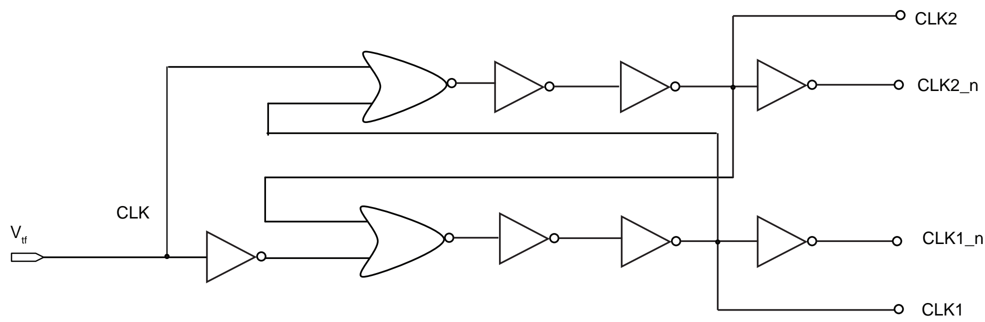


FIGURE 4.6: Generator of non-overlapping phases.

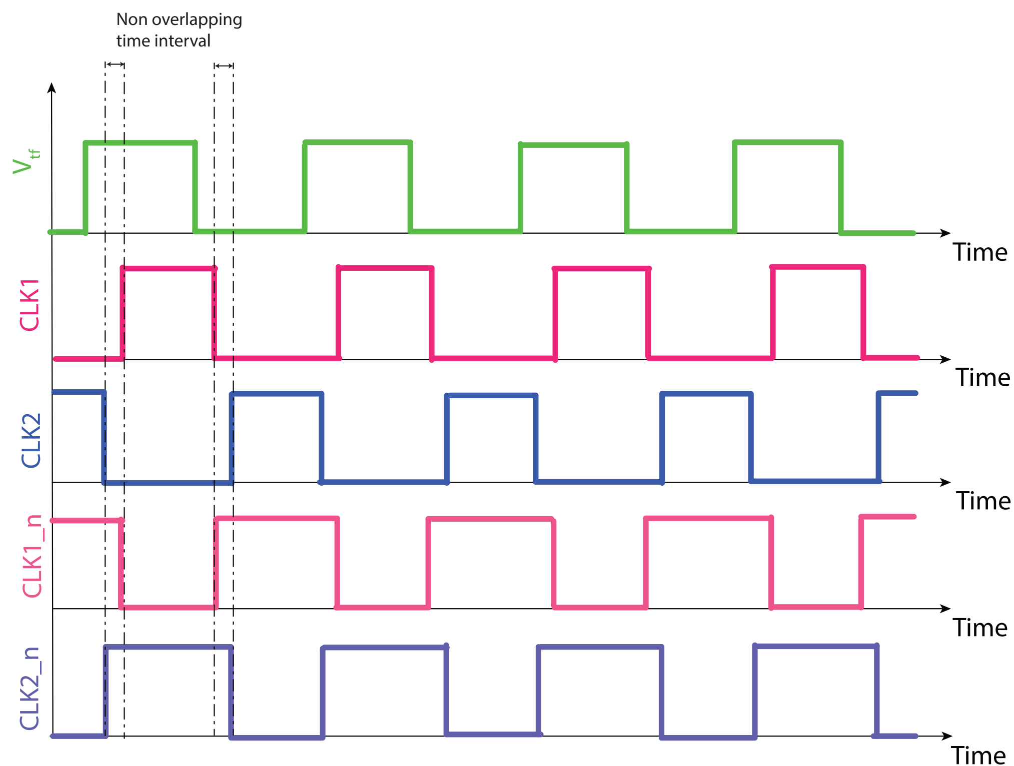


FIGURE 4.7: Clock phase waveforms.

The generated waveforms are illustrated in Fig. 4.7, where the non-overlapping time intervals are highlighted.

Clock phases $CLK1$ and $CLK2$ (non-overlapping high level) are used to drive NMOS transistors, whereas $CLK1_n$ and $CLK2_n$ (non-overlapping low level) are used to drive PMOS transistors.

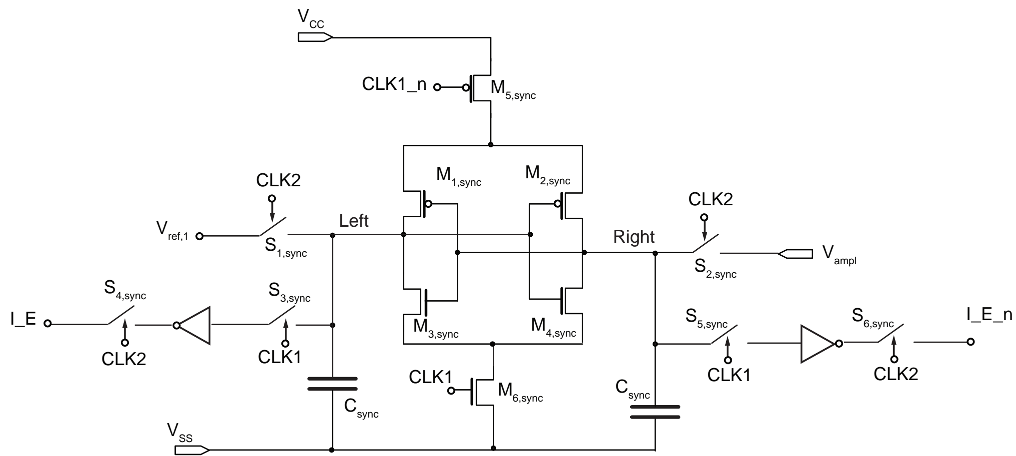


FIGURE 4.8: Circuit schematic of the synchronizer.

Synchronizer As pointed out in Chapter 1, the settling time of V_{ampl} is very long (about 200 μs), which can cause problems in detecting the rising edge of this signal with a simple comparator. A synchronizer has been included in order to overcome the above drawback (Fig. 4.8).

Initially, we have $CKL1 = 0$ and $CLK2 = 1$. Thus, nodes $Left$ and $Right$ are connected to V_{ref1} (through switch $S_{1,sync}$), and V_{ampl} (through switch $S_{2,sync}$) respectively. When the next transition of clock signals occurs ($CKL1 = 1$ and $CLK2 = 0$), $S_{1,sync}$ and $S_{2,sync}$ are turned off and the latch (transistors $M_{1,sync}$ to $M_{4,sync}$), which is connected to the output inverters by switches $S_{3,sync}$ and $S_{5,sync}$, is activated by $M_{5,sync}$ and $M_{6,sync}$. The positive feedback makes nodes $Left$ and $Right$ evolve from their initial state. If $V_{ampl} < V_{ref1}$, node $Left$ is brought to V_{CC} and node $Right$ is brought to V_{SS}. On the contrary, if $V_{ampl} > V_{ref1}$, node $Left$ is brought to V_{SS} and node $Right$ is brought to V_{CC}. During the first part of the next clock cycle ($CLK1 = 0$, $CLK2 = 1$), $S_{3,sync}$ and $S_{5,sync}$ turn off whereas switches $S_{4,sync}$ and $S_{6,sync}$ turn on, thus connecting signals I_E and I_E_n to the outputs of the inverters. Signals I_E and I_E_n assume the opposite values of nodes $Left$ and node $Right$ respectively: thus, I_E is at a high level for values of V_{ampl} higher than V_{ref1}.

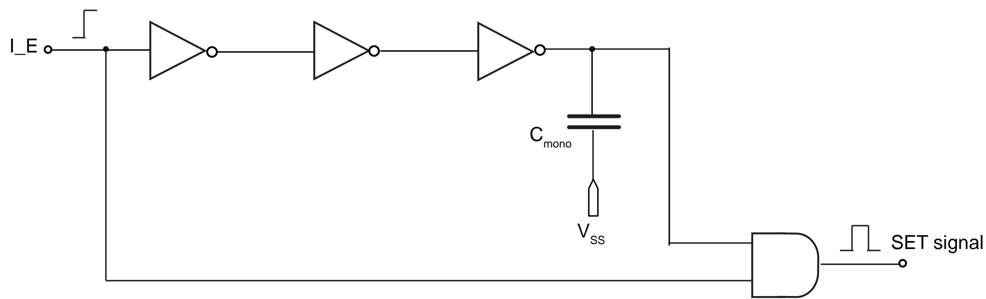


FIGURE 4.9: Circuit schematic of the monostable in Fig. 4.5.

Monostable The monostable circuit shown in Fig. 4.9 detects the rising edge of signal I_E and generates a square pulse of 1 ns, which sets the SR latch in Fig. 4.5 and starts the counter.

Capacitor C_{mono} makes the upper AND gate input to fall slowly, thus ensuring that the pulse width of the Set signal is never less than 1 ns.

4.5.2.2 Current Track-and-Hold circuit

Measuring the programming current is fundamental to adequately characterize the PCM cell. However, as underlined in Chapter 1, available instrumentation can only read a current that does not change significantly for about 1 ms to 10 ms. This time is huge with respect to the duration of a programming pulse, which is expected to vary from 50 ns to 350 ns. A current Track-and-Hold circuit is thus necessary in order to generate a replica of the programming current and to maintain it at a stable value for the amount of time required by the external instrumentation.

A conventional solution [35] was chosen for the current Track-and-Hold circuit (Fig. 4.10). At the beginning of a program operation, all switches are open and a DC voltage, V_{sense}, is applied to pad *sense* by the external equipment. When the pulse generator system is enabled, switches $S_{1,T\&H}$ and $S_{2,T\&H}$ are turned on (Fig. 4.11), so that the tracking phase for I_{prog} starts. The voltage across capacitor $C_{T\&H}$ increases until it reaches the level that allows the diode-connected

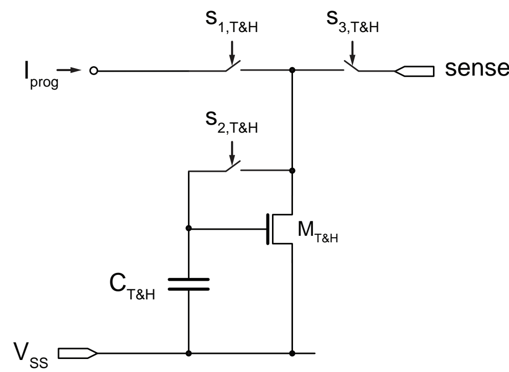


FIGURE 4.10: Circuit schematic of the current Track-and-Hold circuit.

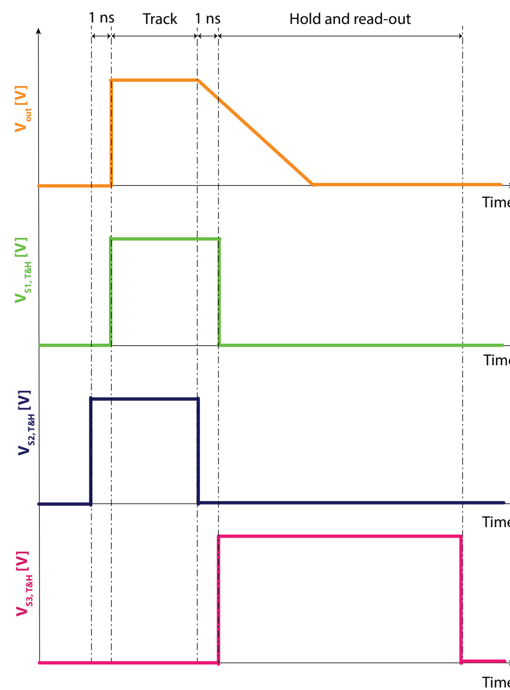


FIGURE 4.11: Timing diagram of the current Track-and-Hold circuit.

transistor $M_{T\&H}$ to draw current I_{prog}. When the programming pulse duration is over, $S_{2,T\&H}$ is turned off. $S_{1,T\&H}$ is turned off after about one nanosecond to avoid charge injection effects from this device into $C_{T\&H}$, which could change the voltage at the gate of $M_{T\&H}$, thus affecting the accuracy of the Track-and-Hold circuit. For ease of implementation, the (intentional) skew between the falling edges of the control signals of switches $S_{1,T\&H}$ and $S_{2,T\&H}$ is also present between the rising edges of these two signals.

After $S_{1,T\&H}$ and $S_{2,T\&H}$ are turned off, $S_{3,T\&H}$ is turned on, thus connecting pad

sense to the drain of $M_{T\&H}$. This transistor then sinks a current I_{sense} equal to the stored current I_{prog}. The external instrumentation connected to pad I_{sense} is thus able to read the current I_{sense} flowing through $M_{T\&H}$.

As mentioned above, when reading the programming current, a DC voltage is applied to pad *sense* by the external equipment. To improve readout accuracy, the applied DC voltage should be equal to the voltage at the drain node of $M_{T\&H}$ (hence, to the voltage across $C_{T\&H}$) at the end of tracking mode operation. As this voltage depends on the value of I_{prog} and, therefore, is not known in advance, its estimated average value will be applied to pad *sense*.

Capacitor $C_{T\&H}$ holds the information about the programming current. On the one hand, increasing the size of $C_{T\&H}$ ensures that this information remains substantially unchanged for a longer time in spite of unavoidable leakage currents. On the other hand, increasing $C_{T\&H}$ also implies that a longer time is needed to charge this capacitance to the correct value during tracking mode operation. It should be pointed out that the shortest programming pulse in our implementation has a time duration of 50 ns, which therefore represents the upper bound of the allowed charging time of $C_{T\&H}$. The used value of $C_{T\&H}$ results from a tradeoff choice between the two above opposite requirements.

Switch $S_{2,T\&H}$ must feature sufficiently low on-resistance to ensure fast charging of capacitor $C_{T\&H}$ in tracking mode ($S_{2,T\&H}$ on) on the one hand, and minimum junction leakage and subthreshold currents when operating in holding mode, so as to provide adequately constant output current for correct read-out, on the other hand. Increasing the length and reducing the width of this switch minimize subthreshold and leakage currents, but also increase the switch on-resistance (and, hence, the charging time of $C_{T\&H}$) during the tracking phase. The above opposite requirements were taken into account in the design phase when choosing the sizes of $S_{2,T\&H}$.

FIGURE 4.12: Circuit schematic of the output buffer.

4.5.3 Pulse generation

4.5.3.1 Output buffer

As explained in Chapter 3, the output buffer (Fig. 4.12) applies the programming pulse to the load under test (a PCM cell). The buffer must allow fast rise time and, at the same time, manage the highly variable load due to the memory cell, and must therefore be provided with adequate current drive capability. In addition, the output voltage must be well controlled down to about 400 mV, so as to avoid the risk of unintentional changes in the cell state.

The basic operating principle has already been described in Chapter 3. However, some enhancements were designed. Here only these improvement are reported in detail.

The first enhancement is that, in this output buffer, a cascade scheme ($M_{27,b}$, $M_{13,b}$) has been used to bias the n-type follower ($M_{12,b}$). This way, the drain voltage of $M_{13,b}$ remains substantially constant even in the presence of change in the output voltage V_{out}. The bias current of the follower, I_{13}, is thus better matched with I_{bias_buffer} and $I_{26,b}$, thus enhancing mirror accuracy.

The second improvement regards the mirrored current I_{26}. Even in this implementation, the current flowing through the n-type follower is mirrored through transistor $M_{24,b}$. However, this transistor delivers its current I_{prog} to a current

Track-and-Hold circuit (Section 4.5.2.2), whereas in Chapter 3 the mirrored programming current was directly fed to the ATE. The mirrored current can be read externally by connecting the APWTS to this current Track-and-Hold circuit through pad *sense*.

The logic threshold voltage of the inverter connecting the source and the gate of $M_{25,b}$ is equal to 4.5 V, which is the expected average voltage at the drain of $M_{23,b}$. This way, the drain of $M_{24,b}$ is set to the same voltage as the drain of $M_{23,b}$, thus minimizing errors in the mirror factor due to channel length modulation effects. The transistors of mirror branch $M_{24,b}$, $M_{25,b}$, $M_{26,b}$, $M_{28,b}$, have a $\frac{W}{L}$ ratio which is one half the corresponding transistors in the n-type follower branch ($M_{23,b}$, $M_{12,b}$, $M_{13,b}$, $M_{27,b}$, respectively), which results in a current mirror factor ($M_{24,b}$, $M_{13,b}$) of 0.5. It should be pointed out that the current through $M_{12,b}$ and $M_{23,b}$ is actually the sum of the programming current injected into the cell under testing and bias current I_{13}. Transistor $M_{26,b}$ was included to subtract a current equal to 0.5 I_{13} from the mirrored current so that only a scaled copy of the programming current is delivered to the Track-and-Hold circuit. Transistor $M_{28,b}$ was included for matching purposes. From the above discussion, the Track-and-Hold circuit stores a current equal to one half the programming current, provided that all devices are adequately matched.

4.5.3.2 Delay time

Delay time Δt is processed by charging and discharging capacitor $C_{\Delta t}$ with two predetermined constant currents, namely equal, $I_{\Delta t}^{ch}$ and $I_{\Delta t}^{dch}$ (Fig. 4.13).

A Miller integrator implements charge and discharge operation of the approximate integrator used in the previous test chip (see Chapter 3). The operational amplifier of the integrator is similar to the output buffer input stage. Capacitor $C_{\Delta t}$ is initially discharged by turning switch $S_{0,\Delta t}$ on (Fig. 4.14). When the rising edge of V_{tf} is detected, $S_{0,\Delta t}$ is turned off and switch $S_{1,\Delta t}$ is turned on. $M_{12,\Delta t}$ starts charging (negatively) the capacitor: voltage $V_{\Delta t}$ therefore decreases with a constant slope. When the rising edge of V_{td} is detected (after a delay time Δt), $S_{1,\Delta t}$

FIGURE 4.13: Circuit schematic of the delay time counter.

is turned off and $S_{2,\Delta t}$ is turned on. $M_{13,\Delta t}$ then starts drawing current from $C_{\Delta t}$ and $V_{\Delta t}$ increases with a constant slope. When $V_{\Delta t}$ increases above the reference voltage V_{ref2}, twice the delay time Δt has passed. A digital signal, $t_{p,en}$, generated from $V_{\Delta t}$ enables the output buffer and starts the holding phase of the voltage Track-and-Hold circuits (Fig. 4.2).

After the falling edge of V_{tf} is detected, $S_{2,\Delta t}$ is turned off and $S_{0,\Delta t}$ is turned on, thus re-initializing $C_{\Delta t}$.

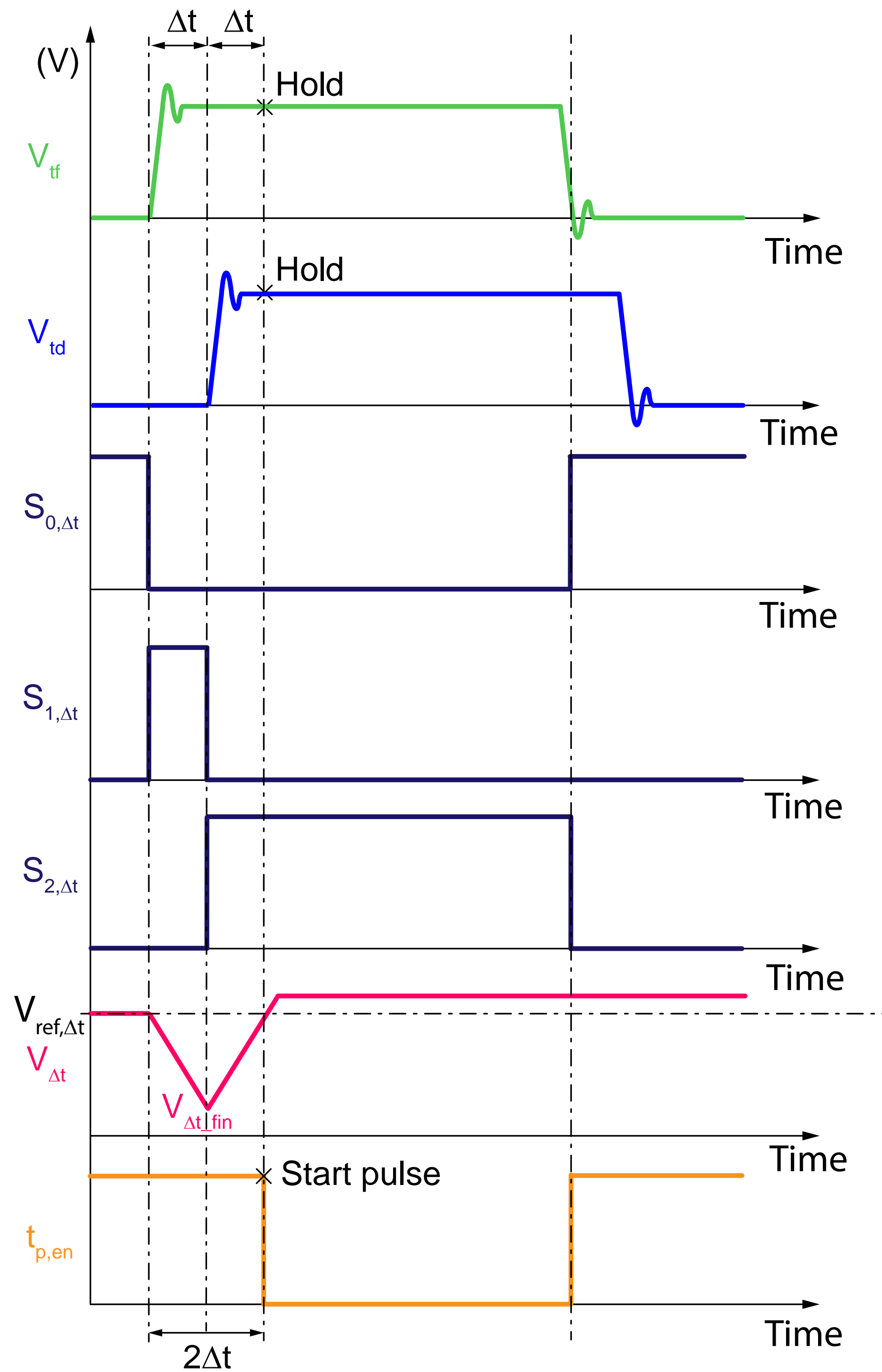


FIGURE 4.14: Timing diagram of the processing of time delay Δt.

Currents $I_{\Delta t}^{dch}$ and $I_{\Delta t}^{ch}$ are generated internally by the branch consisting of $R_{\Delta t}$, $M_{10,\Delta t}$, and $M_{11,\Delta t}$.

They are respectively equal to

$$I_{\Delta t}^{dch} = \frac{V_{CC} - V_{ref2}}{R_{td}^{ch}} N_{\Delta t}^{dch} \tag{4.6}$$

and

$$I_{\Delta t}^{cch} = \frac{V_{CC} - V_{td}}{R_{td}^{dch}} N_{\Delta t}^{dch} \tag{4.7}$$

where $N_{\Delta t}^{dch}$ and $N_{\Delta t}^{ch}$ account for the mirror factors in the discharging ($M_{11,\Delta t}$, $M_{12,\Delta t}$) and charging circuitry ($M_{10,\Delta t}$, $M_{13,\Delta t}$).

Even in this implementation, the value of this current is not critical, as the same current is used to charge and discharge capacitor $C_{\Delta t}$. The only constraint is that $V_{\Delta t}$ must never reach either too high or too low levels, so that the integrator never saturates under any operating and fabrication process conditions.

4.5.3.3 Time duration

Pulse time duration is obtained by converting the amplitude of pulse V_{td} to a time information by first charging capacitor C_{td} with a given constant current I_{td}^{ch} and then discharging it with a current I_{td}^{dch} proportional to the amplitude of V_{td} (Fig. 4.15). Capacitor C_{td} is initially discharged ($S_{0,td}$ is turned on, Fig. 4.16).

While $C_{\Delta t}$ is being charged by current $I_{\Delta t}^{ch}$ (second time interval Δt), C_{td} is charged positively with current I_{td}^{ch} ($S_{0,td}$ is turned off and $S_{1,td}$ is turned on). C_{td} is then discharged by means of a current I_{td}^{dch} whose value is proportional to the amplitude of V_{td} ($S_{1,td}$ is turned off and $S_{2,td}$ is turned on). A component detects when V_{Ctd} falls below V_{ref} and generates a signal, $t_{f,en}$, which determines the pulse duration end.

As in the Δt processing circuit, when the falling edge of V_{tf} is detected, C_{td} is re-initialized by turning $S_{2,td}$ off and $S_{0,td}$ on.

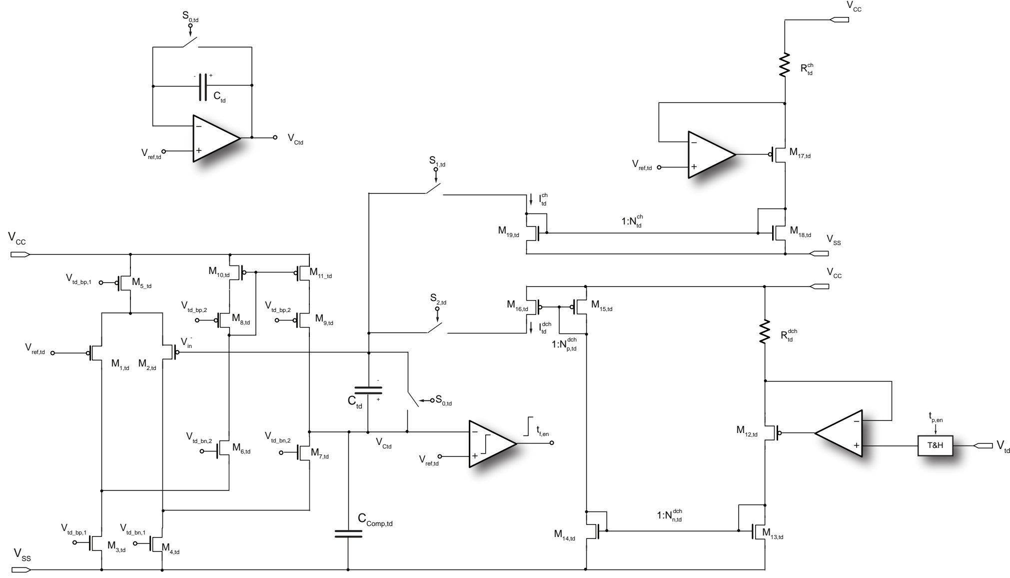


Figure 4.15: Circuit schematic for time duration generation: charging (I_{td}^{ch}) and discharging (I_{td}^{dch}) capacitor C_{td}.

Also in this case, a Miller integrator is used. Since V_{Ctd} and, consequently, the pulse time duration, must vary over a large range, wide input and output swings of the integrator are required. A folded-cascode topology has been chosen for the operational amplifier in this integration to meet this requirement.

I_{td}^{ch} and I_{td}^{dch} are obtained by Ohm law and can be rewritten, according to (4.1) and (4.2), as

$$I_{td}^{ch} = \frac{V_{CC} - V_{ref2}}{R_{td}^{ch}} N_{td}^{ch} \tag{4.8}$$

and

$$I_{td}^{dch} = \frac{V_{CC} - V_{td}}{R_{td}^{dch}} N_{td}^{dch} \tag{4.9}$$

where N_{td}^{ch} is the NMOS mirror factor in the charging circuitry ($M_{18,td}$, $M_{19,td}$) and $N_{td}^{dch} = N_{td}^{p,dch}\ N_{n,td}^{dch}$ is the product of the NMOS and PMOS mirror factors in the discharging circuitry ($M_{13,td}$, $M_{14,td}$ and $M_{15,td}$, $M_{16,td}$).

The above equations (4.8) and (4.9) do not take into account non-idealities such as inaccuracies due to the operational amplifiers.

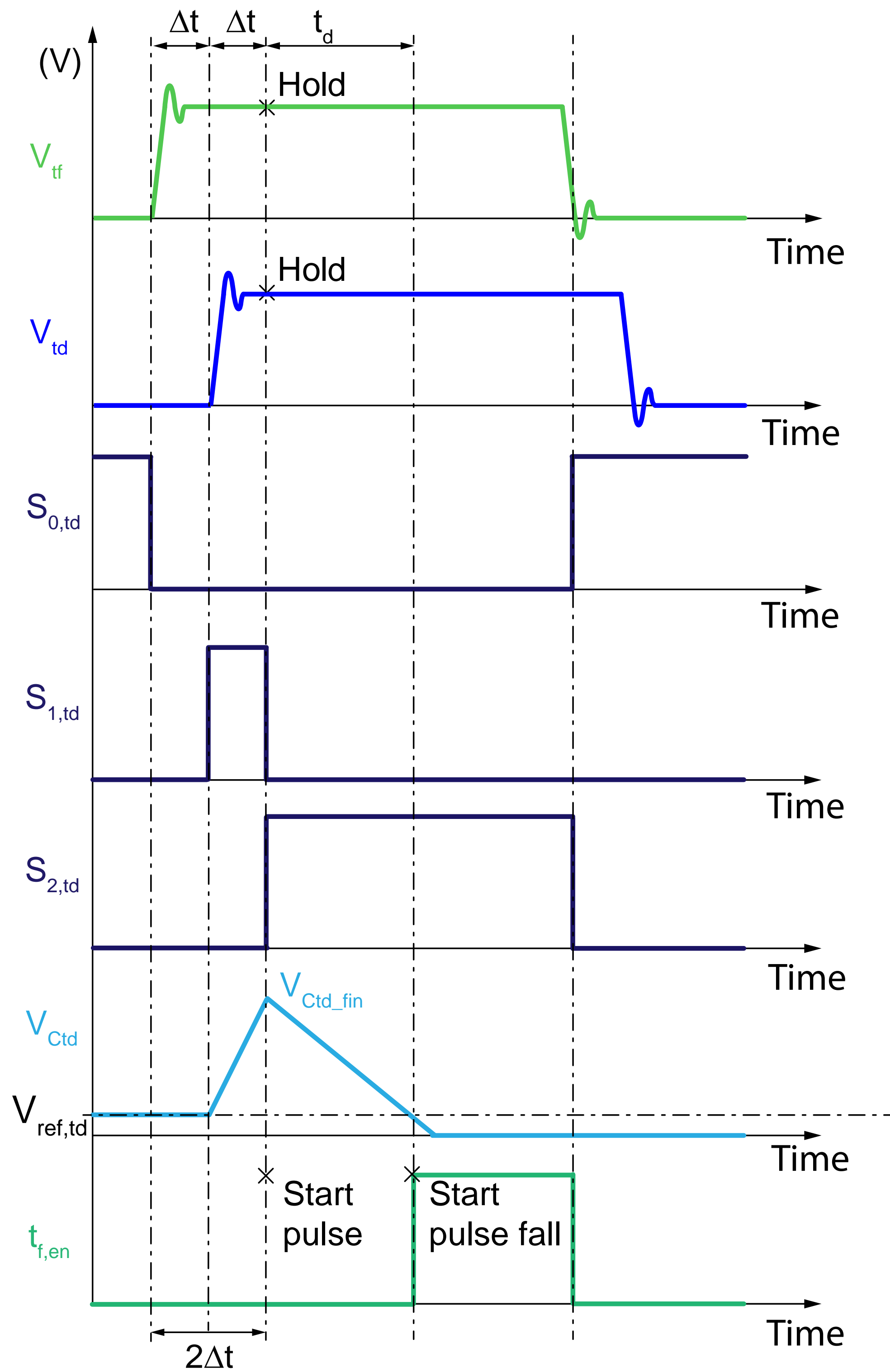


FIGURE 4.16: Timing diagram for pulse time duration (t_d) generation.

To a first order, time duration t_d is given by rewriting (4.4) as the following expression:

$$t_d = \frac{I_{td}^{ch}}{I_{td}^{dch}} \Delta t = \frac{V_{CC} - V_{ref2}}{V_{CC} - V_{td}} \frac{R_{td}^{dch}}{R_{td}^{ch}} \Delta t \frac{N_{td}^{ch}}{N_{td}^{dch}} \tag{4.10}$$

which shows that time duration depends on both V_{td} and Δt. As the absolute value of n-well resistors depend on the applied voltage, it is very important to ensure that the operating conditions of R_{td}^{ch} and R_{td}^{dch} be as similar as possible. In fact, the better the matching of these two resistors, the better the accuracy in obtaining t_d.

In order to minimize the resistance value mismatch due to operating conditions, in this final implementation both resistors have a terminal connected to V_{CC}; as the voltages of the other terminal of R_{td}^{ch} and R_{td}^{dch} are V_{ref2} and V_{td}, respectively, V_{ref2} was set to 3.5 V, which corresponds to about half the swing of V_{td} (1 V to 5.5 V), thus minimizing the mismatch between the two resistors due to operating conditions.

4.5.3.4 Fall time

According to the target specification, the fall time of the program pulse must vary between 10 ns and several μs.

The shortest fall time is obtained by setting V_{tf} to 6 V. The presence of this setting generates a digital signal which forces V_{ben} to 0 V, thus disabling the output buffer after the programmed pulse time duration is reached. The other fall time values are obtained by converting the amplitude of signal V_{tf} into current I_{tf}. This current discharges the storage capacitor, C_{ampl} (Fig. 4.2), at a constant rate after the programming pulse time duration is reached, thus controlling the fall slope. To a first order, the fall time t_f is given by (4.5).

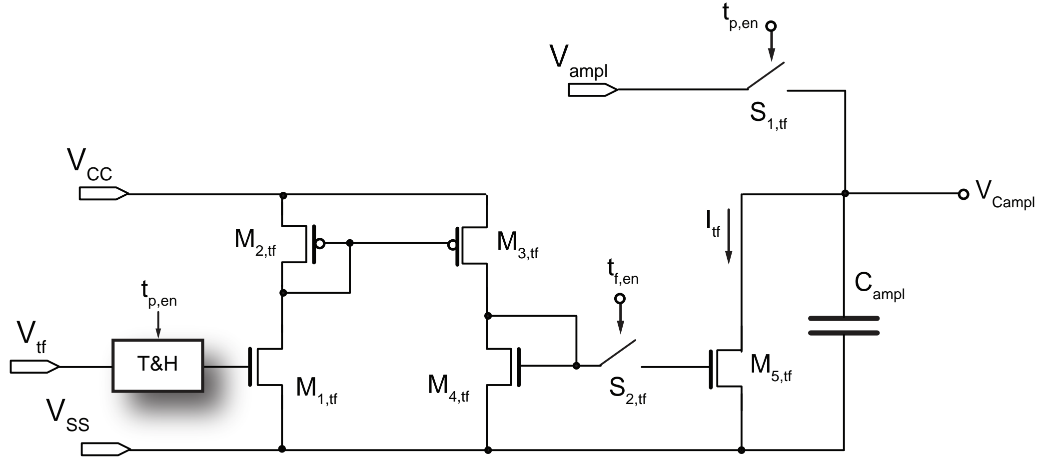


FIGURE 4.17: Circuit schematic of the voltage-to-current conversion for fall time generation.

The current interval required to achieve the specified range of t_f is very wide. As in the case of the fall time generation described in Chapter 3, the voltage-to-current conversion is carried out by using an NMOS transistor, $M_{1,tf}$, working in its pinch-off region (Fig. 4.17).

Discharging current I_{tf} is thus given by rewriting (4.3) as

$$I_{tf} = \frac{\beta}{2}(V_{tf} - V_{th})^2 N_{p,tf} N_{n,tf} \tag{4.11}$$

where $N_{p,tf}$ and $N_{n,tf}$ are the mirror ratios of the PMOS ($M_{2,tf}$, $M_{2,tf}$) and the NMOS ($M_{4,tf}$, $M_{5,tf}$) current mirror, respectively.

Fall time is thus expressed as

$$t_f = \frac{0.8\ C_{ampl}}{\frac{\beta}{2}(V_{tf} - V_{th})^2 N_{p,tf} N_{n,tf}} V_{ampl} \tag{4.12}$$

Differently than in the scheme described in Chapter 3, switch $S_{2,tf}$ connects the gates of transistors $M_{4,tf}$ and $M_{5,tf}$ so as to begin the discharge of C_{ampl}. This solution significantly reduces charge injection effects with respect to the solution where the switch connects storage capacitor C_{ampl} to the drain of $M_{5,tf}$.

4.6 Calibration procedure

The guidelines of the calibration procedure explained in Chapter 3 were followed even for this implementation.

As already explained, the calibration algorithm described in Chapter 3 is based on the possibility of sensing and storing information about time, which is not possible when using conventional ATE. However, as mentioned in Chapter 1, conventional ATE can very accurately read both currents and voltages. A dedicated interface hardware was therefore designed and integrated to allow timing parameters to be measured. The basic idea is to charge a capacitor C_{cal} with an internally generated constant current (I_{cal}) for a time interval equal to the timing parameter under consideration (either t_d or t_f) and then deliver the final voltage V_{cal_fin} developed across C_{cal} to the ATE, so as to exploit its accuracy in voltage measurements. The voltage read by the instrumentation can be easily re-converted to the corresponding timing information by using the conventional capacitor equation

$$t_{cal} = \frac{C_{cal}}{I_{cal}} V_{cal_fin} \tag{4.13}$$

where t_{cal} is the timing parameter which is being measured. However, the ratio $\frac{C_{cal}}{I_{cal}}$ is affected by process spreads. A pre-calibration procedure is therefore necessary to determine the exact value of this ratio. The pre-calibration procedure is performed by charging C_{cal} for a known time interval t^*_{cal} and then sensing the final voltage across C_{cal}, $V^*_{cal_fin}$. The value of $\frac{C_{cal}}{I_{cal}}$ is found as

$$\frac{C_{cal}}{I_{cal}} = \frac{t^*_{cal}}{V^*_{cal_fin}} \tag{4.14}$$

The internal digital signals used to generate the programming pulse during normal test mode operation of the chip are exploited to obtain the charging time of C_{cal}, t_{cal}, in calibration mode. This approach allowed us to avoid the generation of

Table 4.3: Non-idealities in Δt processing (Legend: S. = process spreads, M. = mismatches, O. C. = operating conditions)

Parameter	S.	M.	O. C.	Comments
$N_{\Delta t}^{ch}$	X	X		
$N_{\Delta t}^{dch}$	X	X		
$R_{\Delta t}$	X			
$C_{\Delta t}$	X			
$V_{bn,\Delta t}$	X			
$V_{bp,\Delta t}$	X			
$V_{os,comp,\Delta t}$		X		

any further digital signal devoted to determine t_{cal}, thus optimizing overall hardware. In order to use the same calibration hardware for measuring both timing parameters (t_d and t_f), only one of these two parameters per generated pulse is measured. Two predetermined analog values of V_{td} and V_{tf} out of their respective normal operating ranges indicate which parameter is being sensed (t_f and t_d, respectively) during each calibration operation. The above choice obviously leads to longer calibration time, but has the advantage of silicon area saving (calibration hardware must not be duplicated).

In the following of this Chapter, the equations for t_d and t_f are found, using the approach described in Chapter 3.

4.6.1 Calibration equations for time duration and fall time

First, non-idealities in processing Δt were considered. They are summarized in Table 4.3.

Compared to the test chip described in Chapter 3, and consequently Table 4.3 and Table 3.1, it can be observed that the dependence of the parameters on the operating condition is no longer present, thanks to the active integrator. All the parameters in Table 4.3 can thus be considered constants, and they can be grouped in an unknown, $\alpha_{\Delta t}$.

Table 4.4: Non-idealities in t_d generation (Legend: S. = process spreads, M. = mismatches, O. C. = operating conditions)

Parameter	S.	M.	O. C.	Comments
N_{td}^{ch}	X	X		
$N_{n,td}^{dch}$	X	X		
$N_{p,td}^{dch}$	X	X		
$\frac{R_{td}^{dch}}{R_{td}^{ch}}$	X	X	X	R_{td}^{dch} depends on V_{td}
C_{td}	X			
$V_{er,opamp,td}^{dch}$	X	X		
$V_{er,opamp,td}^{dch}$	X	X		
$V_{os,comp,td}$		X		

The remarks made in Chapter 3.3.1.1 for $V_os, comp, \Delta t$ are valid even in this case, so the effect of $V_os, comp, \Delta t$ has been neglected.

Δt_{dch} can be thus expressed as

$$\Delta t_{dch} = \alpha_{\Delta t} \Delta t \tag{4.15}$$

Then, non-idealities affecting t_d are considered and summarized in Table 3.2.

Equation (4.10) can be rewritten as

$$\begin{aligned} t_d = & \frac{R_{td}^{dch}}{R_{td}^{ch}} \frac{V_{CC}-(V_{ref,td}-V_{er,opamp,td}^{dch})}{V_{CC}-(V_{td}-V_{er,opamp,td})} \frac{N_{td}^{ch}}{N_{td}^{dch}} \alpha_{\Delta t} \Delta t + \\ & + C_{td} \frac{R_{td}^{dch}}{V_{CC}-(V_{td}-V_{er,opamp,td})} V_{os,comp,td} \end{aligned} \tag{4.16}$$

If we set $V_{os,comp,td} = 10$ mV and consider a 2% error of the operational amplifiers, as it has been done in Chapter 3.3.1.2, the worst-case impact of these non-idealities on t_d in our test-chip is negligible also in this case.

Moreover, the dependence on operating conditions of N_{td}^{ch} and N_{td}^{dch} are overcome thanks to the active integrator. Still, the dependence of $\frac{R_{td}^{dch}}{R_{td}^{ch}}$ on V_{td} persists.

If all constant (and quasi-constant) parameters are grouped in an unknown, α_{td}, and if the linear dependence of R_{td}^{dch} on V_{td} is taken into accounte, equation (4.16) can be written as

$$t_d = \frac{\alpha_{td}(m'V_{td} + q')\Delta t}{V_{CC} - V_{td}} \tag{4.17}$$

Equation (4.17) can be then rewritten as

$$t_d = \frac{m\ V_{td} + q}{V_{CC} - V_{td}}\Delta t \tag{4.18}$$

where $m = \alpha_{td}m'$ and $q = \alpha_{td}q'$. As in Chapter 3.3.1, the number of unknowns to be determined is two, namely m and q.

Finally, non-idealities regarding the generation of t_f will not be discussed, since the circuits of Fig. 3.10 and 4.17 are substantially equivalent.

Here only the resulting equation is given.

$$t_f = \frac{0.8\ V_{ampl}\ \alpha_{tf}}{(V_{tf} - V_{th,n})^2} \tag{4.19}$$

Equation (4.19) shows that the unknowns to be found are two, namely α_{tf} and $V_{th,n}$.

4.6.2 Accuracy considerations

First, the error due to external equipment inaccuracy was investigated. From datasheets, the typical voltage error from the pulse generator is $\pm 1\%$. First-order equations (4.10) and (4.12) were taken into account to determine the sensitivity of pulse parameters t_d and t_f to voltage variations of control signals V_{td} and V_{tf}, respectively. It was calculated that the maximum impact on t_d of a $\pm 1\%$ inaccuracy in V_{td} is around $\mp 1\%$. Similarly, the maximum variation of t_f for a $\pm 1\%$

variation in V_{tf} was found to be +6.9% and -6.5%, respectively. Then, through circuit simulations, it was found that the variation of ±1% on the input voltage level results in a maximum variation between +1.5% and -2% for t_d and between +6.2% and -7.5% for t_f, which are both within target specifications. The difference between calculated and simulated results is due to second-order effects which are not taken into account in (4.10) and (4.12). In this respect, it should be pointed out that, assuming that most of the equipment inaccuracy has a very slow drift over time, any error in analog control voltages is substantially compensated for by the proposed calibration procedure, provided that calibration is carried out just before characterization with the same equipment used in test mode.

Similar observations can be made for the uncertainty of the external pulse generator: the system was designed so as to operate with input voltage ranges which minimize the above uncertainty by using the different scales of the instrumentation. In addition, in our design, the lowest input voltage levels, which correspond to the highest sensitivity of the system to input voltage variations, give rise to longest time durations and fall times.

Second, through simulations, we performed the proposed calibration procedure for a typical-case process, thus extracting the unknown parameters (m, q, α_{tf}, and $V_{th,n}$), which were then used to generate pulses in worst-case process corners ($\pm 3\sigma$). The deviations in pulse parameter values obtained for different process corners can be remarkable, especially as far as pulse fall time is concerned: indeed, deviations of up to 40% were observed in the case of long fall times, which correspond to low voltage levels of V_{tf}. Depending on the maturity of the process under test, the calibration procedure can thus be performed once per site, once per wafer, or even once per batch. The adopted strategy will result as a trade-off choice between testing speed and target accuracy.

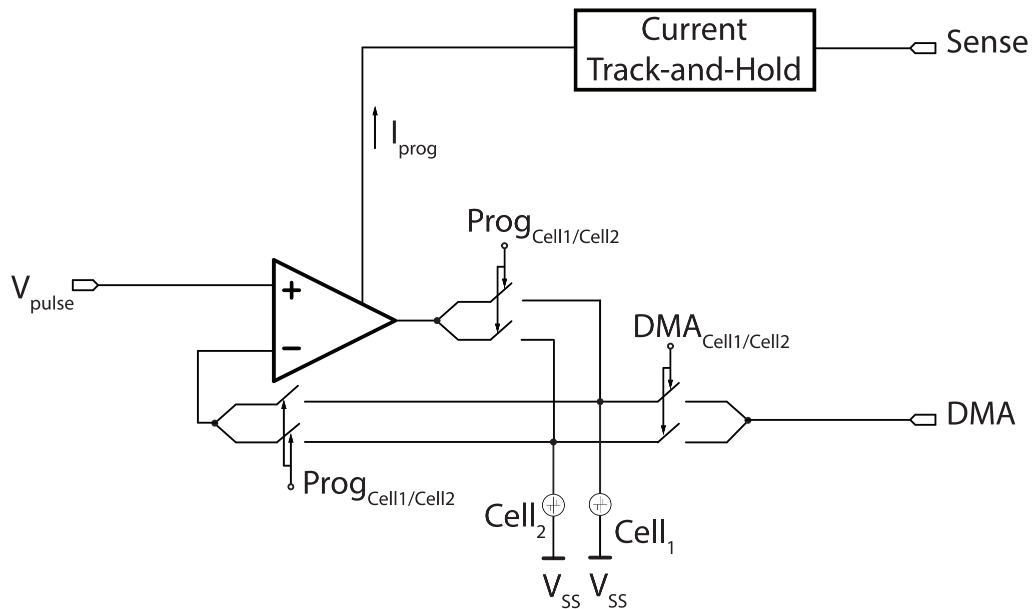


FIGURE 4.18: High-level description of the simplified version.

4.7 Simplified version of the system for different test equipment

Instrumentation able to perform very accurate analysis is on the market. However, some analysis may have to be performed on test-benches with limited equipment, but where the disturbances introduced by interconnections are within acceptable levels. In this case, the main challenge is to be able to read the programming current of the cell at the pulse plateau, rather than generating an accurate and flexible program pulse. To this end, a simplified version of the system was conceived.

The basic idea of this version is to feed the cell with a programming pulse V_{pulse} from the external PG through the buffer and sense the programming current with the APWTS by means of the current Track-and-Hold (Fig. 4.18). The programming current is then read by the ATE through pad *sense*.

In addition to V_{pulse}, another signal from the PG, $V_{p,ampl}$ is used so as to generate the internal signals needed by the Track-and-Hold. In addition, as it will be explained, the amplitude of $V_{p,ampl}$ is set to a voltage equal to V_{pulse}- 200 mV, so as to detect the falling edge of V_{pulse}.

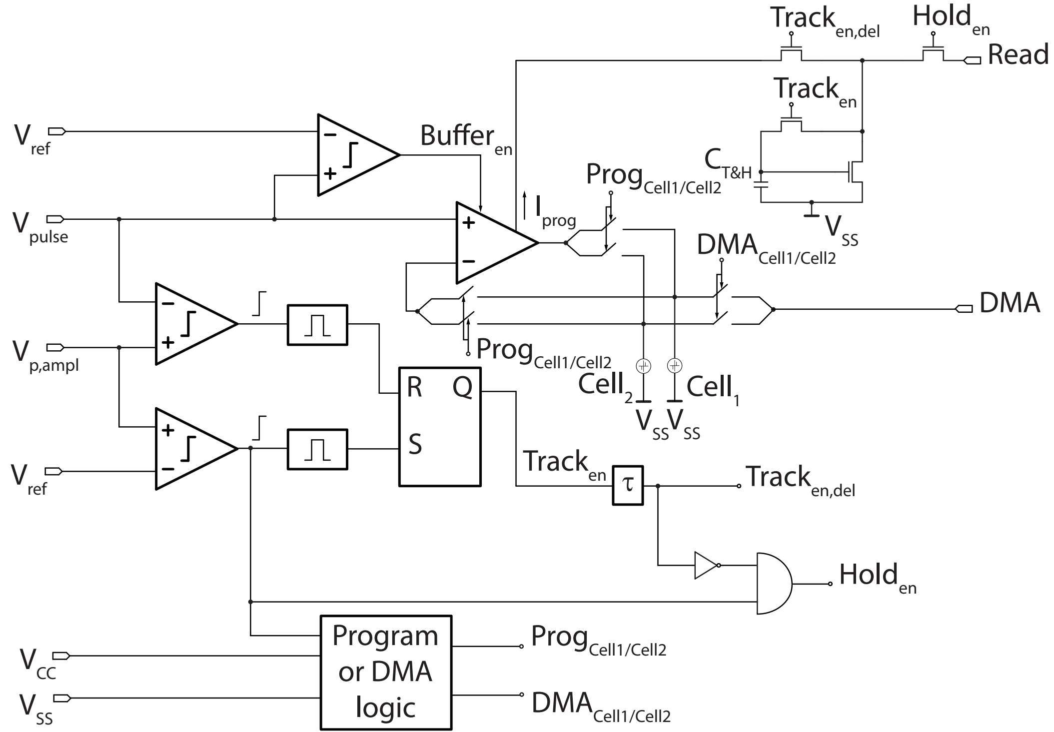


FIGURE 4.19: Scheme of the simplified version.

The system basic operating principle will be now described referring to the scheme in Fig. 4.19 and to the waveforms in Fig. 4.20.

When the rising edge of $V_{p,ampl}$ is detected ($V_{p,ampl} > V_{ref}$, where $V_{ref} = 0.5$ V is a reference voltage from the APWTS), a monostable circuit generates a pulse which sets the output of a SR latch to a logic '1' ($Q = 1$), corresponding to the voltage supply V_{CC}. The tracking phase of the Track-and-Hold is thus enabled (signal $Track_{en}$ high, signal $Track_{en,del}$ high about 3 ns after $Track_{en}$).

When the rising edge of V_{pulse} is detected ($V_{pulse} > V_{ref}$), the buffer is enabled and the programming pulse is fed to the cell, while the programming current is being traced by the Track-and-Hold. As soon as the amplitude of V_{pulse} decreases below $V_{p,ampl}$, a pulse generated by a monostable circuit resets the latch, thus setting $Q = 0$. Signal $Track_{en}$ is set low; after a delay of about 3 ns, signal $Track_{en,del}$ becomes low too, whereas signal $Hold_{en}$ is set high, thus determining the beginning of the Track-and-Hold holding phase: a replica of the programming current is fed to the APTWS, and it can thus be sensed.

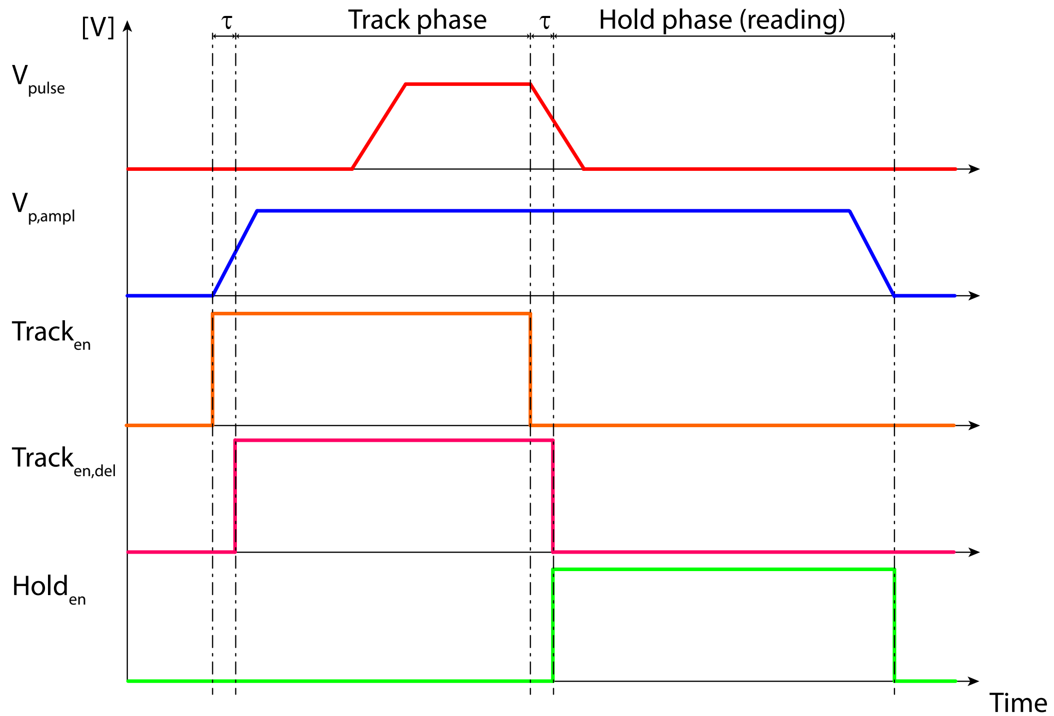


FIGURE 4.20: Internal and external waveforms of the simplified version.

The falling edge of $V_{p,ampl}$ triggers the reinitialization of the system: once the programming current is read, $V_{p,ampl}$ is driven low, thus ending the holding phase of the current Track-and-Hold circuit ($Hold_{en}$ is driven low) and allowing the final state of the cell to be read in DMA mode.

In this implementation, the system was provided of two sensing pads (one to read the programming current and one for DAM mode) in order to simplify the logic.

Chapter 5

Experimental results

All the test chip described were designed for a 180 nm CMOS fabrication process featuring high-voltage (HV) devices withstanding up to 6 V, which are generally available in emerging memory technologies.

5.1 Preliminary buffer

5.1.1 Simulations

In order to evaluate the performance of the preliminary buffer described in Chapter 3.2.1.1, simulations with a load capacitance of 250 fF and load resistances of 10 kΩ, 100 kΩ, 1 MΩ, and 10 MΩ were carried out.

Fig. 5.1 shows the simulated Bode diagram of the gain magnitude with different values of the load resistance. The input DC voltage was set at mid-supply (3 V). The DC gain is about 27 dB. A variation less than 1.5 dB was observed in the whole considered range of load resistance. The error between the input and the output voltage of the operational amplifier connected in buffer configuration was found to be below 6% over the whole common mode voltage of interest, under any process and operating conditions.

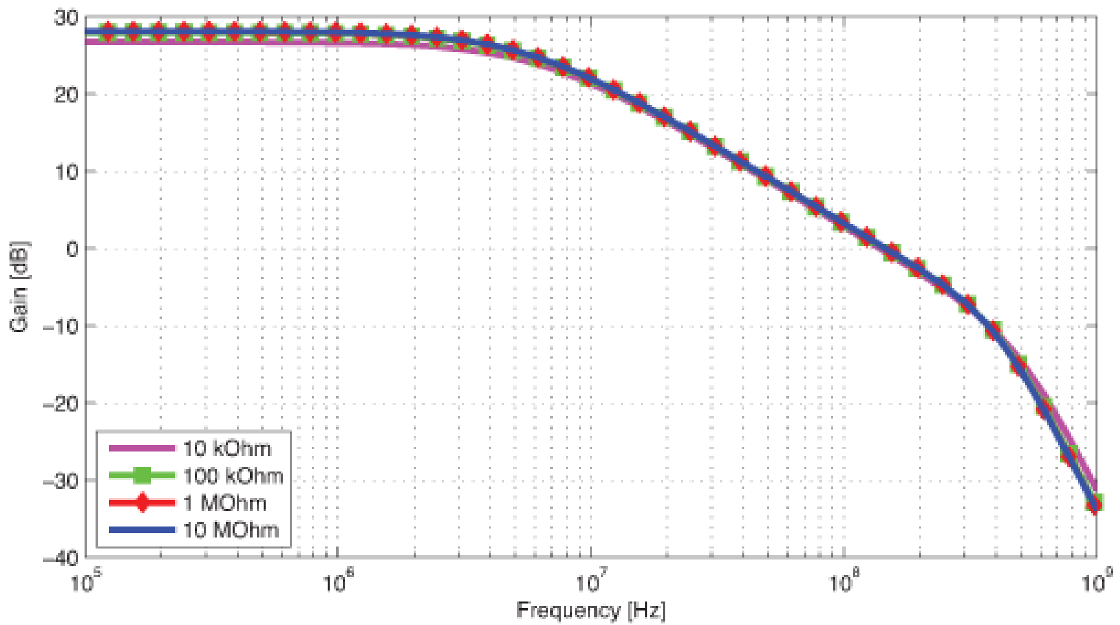


FIGURE 5.1: Frequency response of the designed HV amplifier in open-loop configuration simulated for different values of the load resistor (10 kΩ, 100 kΩ, 1 MΩ, and 10 MΩ).

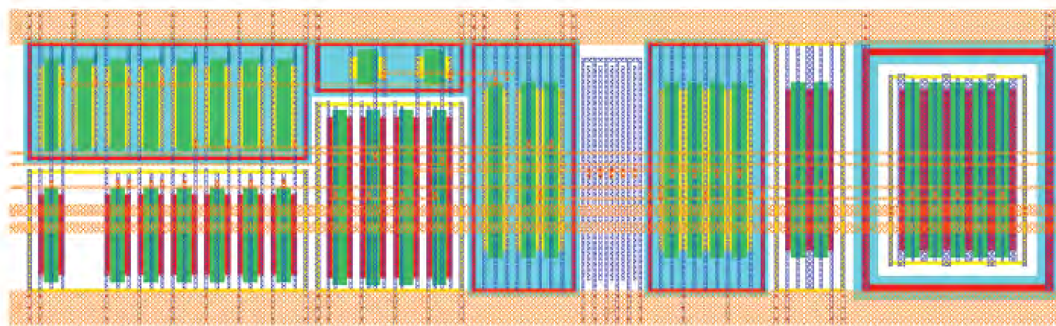


FIGURE 5.2: Buffer layout. The total area (not including bias circuits) is 880 μm^2.

From Fig. 5.1, the unity gain frequency is 140 MHz and is substantially independent of the load resistance value. The minimum phase margin in worst-case fabrication process and operating conditions was found to be 60°.

Further simulations showed a common mode gain ranging from -30 dB to -10 dB, thus leading to a CMRR ranging from 37 dB to 57 dB.

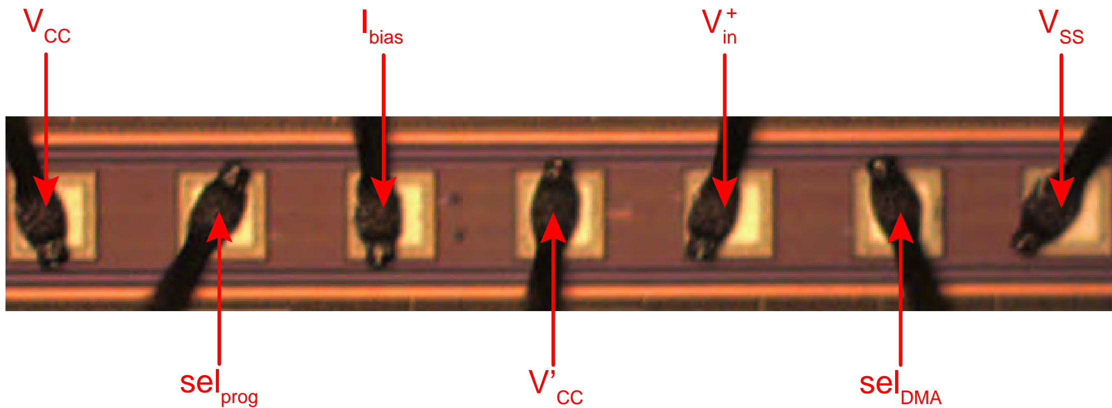


FIGURE 5.3: Buffer microphotograph.

5.1.2 Measurements

The resulting total area of the fabricated buffer (not including bias circuits) is 880 μm^2. Total power consumption is 2.55 mW, not including bias circuitry. The amplifier layout is shown in Fig. 5.2, whereas its microphotograph is shown in Fig. 5.3.

The buffer was provided with the possibility to vary the integrated resistance. In the following of this Section, results obtained with an equivalent resistance of about 13 kΩ, which represents the heaviest resistive load condition, will be shown.

Fig. 5.4 shows the measured (I_{M8-m}) and simulated (I_{M8-s}) currents at pulse plateau for different input pulse amplitudes. A comparison between the simulated (V_{out-s}) and the measured (V_{out-m}) output voltage at pulse plateau is provided in Fig. 5.5. Very good agreement between measured data and simulated results is apparent. As highlighted in Fig. 5.6, the error (ΔV_{out}) between the output and the input voltage is below 6% in the whole voltage range of interest for any load condition.

Transient analysis of the amplifier connected in unity-gain configuration was carried out. First, the buffer swing capability was measured by applying three pulse waveforms (blue curve in Fig. 5.7, plots A, B, C) having different amplitudes and the same values of time duration, rise time, and fall time (Table 5.1, pulses A, B, C). In these measurements, rise time, fall time, and time duration were relaxed in

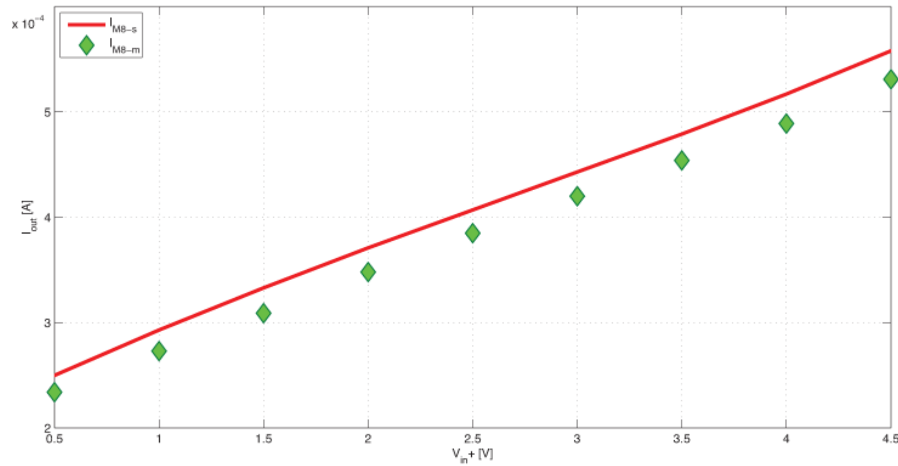


FIGURE 5.4: Comparisons between simulated (I_{M8-s}) and measured (I_{M8-m}) output current at the pulse plateau.

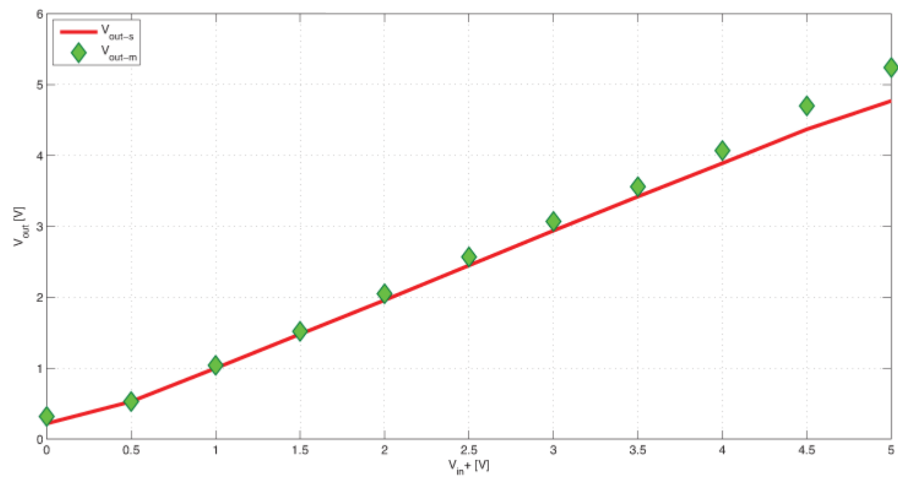


FIGURE 5.5: Comparison between simulated (V_{out-s}, continuous line) and experimental (V_{out-m}, diamonds) output voltage.

order to better evaluate the current at plateau. The buffer response to a fast pulse (Table 5.1, pulse D) was also measured (Fig. 5.7, plot D). It should be pointed out that, due to our experimental setup, ringing transients are present in the input waveform in this case, which gives rise to corresponding transients in the measured current.

Measurements were carried out in AC coupled mode. In Fig. 5.7, the output signal (red curve) represents the voltage pulse $\Delta V'_{CC}$ observed at pad V'_{CC}, which is connected to 6.3 V through a resistor, R_M, of 2 kΩ. A higher value than 6 V was used for V'_{CC} to account for the voltage drop $\Delta V'_{CC} = I_{M8}\ R_M$ across the resistor,

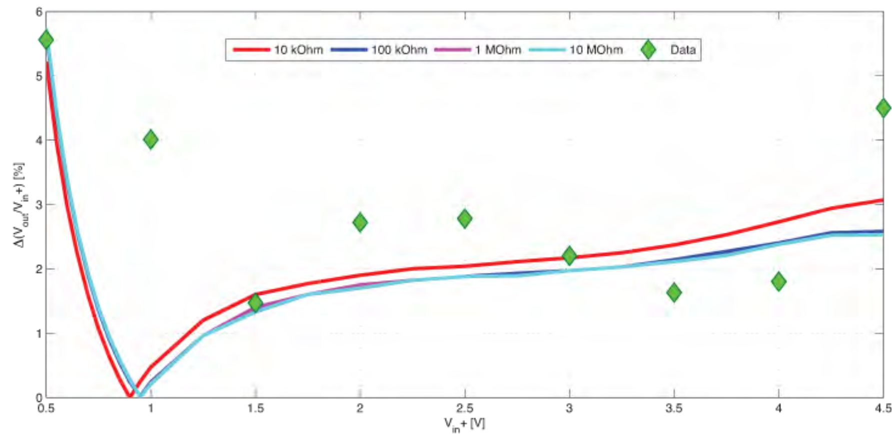


FIGURE 5.6: Output voltage error ($\frac{V_{out}-V_{in}^{+}}{V_{in}^{+}}$): simulated (lines) and experimental data (diamonds).

TABLE 5.1: Parameters of the applied pulse waveforms.

Pulse Name	Ampl. [V]	Pulse Length [ns]	Rise [ns]	Fall [ns]
A	0.4	400	100	100
B	3	400	100	100
C	4.5	400	100	100
D	4.5	500	15	15

where I_{M8} is the current through $M8$ (this current is the sum of the current fed to the load and the biasing current of the n-type follower).

Transient measurements demonstrate that the pulse waveforms are adequately replicated and meet target specifications both in speed and in amplitude.

5.2 First fabricated test chip

5.2.1 Simulation of the output buffer

First, the amplifier described in Chapter 3.2.1 was analysed through simulations of its open-loop gain magnitude and phase. They were carried out with a load capacitance of about 260 fF and load resistances of 10 kΩ and 10 MΩ, which correspond to typical values of the minimum and the maximum resistance of a memory cell in our application. The input DC voltage was set at mid-supply (3 V).

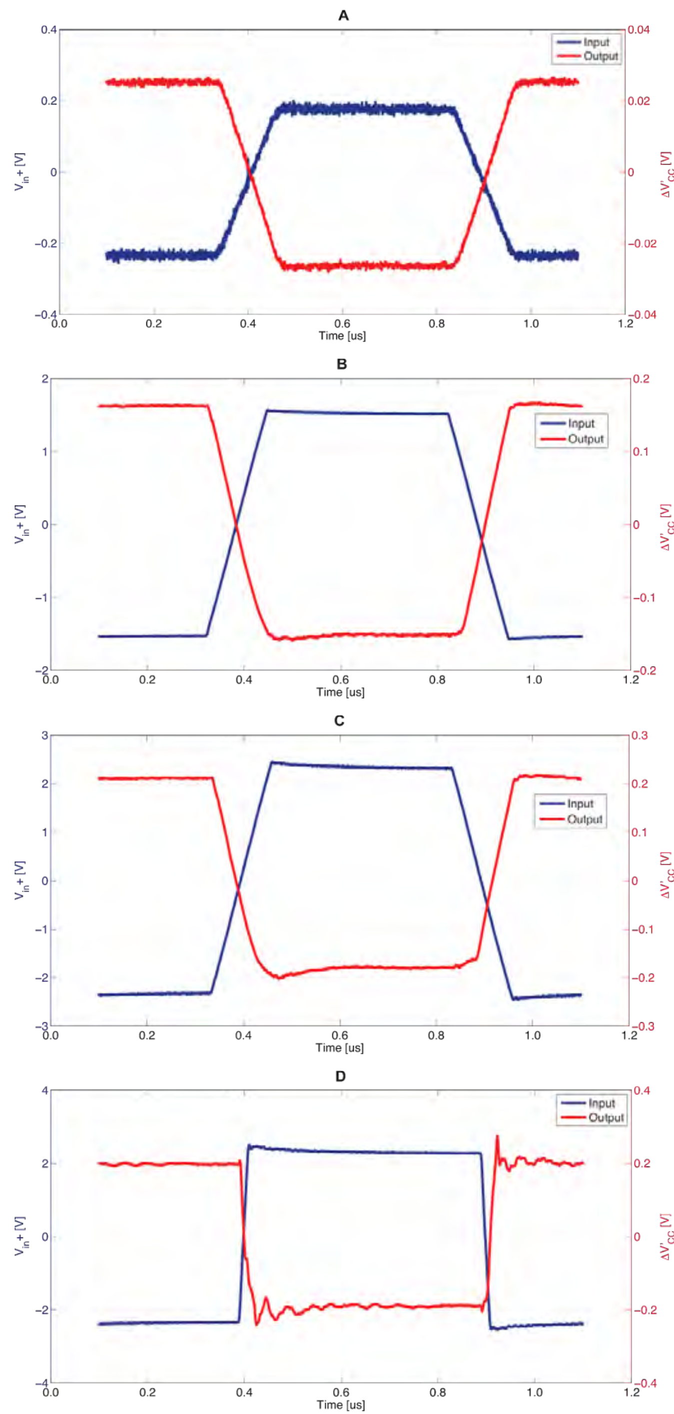


FIGURE 5.7: Waveform A: pulse amplitude = 400 mV, Waveform B: pulse amplitude = 3 V, Waveform C: pulse amplitude = 4.5 V, Waveform D: fast pulse (rise and fall time = 15 ns). Details of the applied waveforms are provided in Table 5.1.

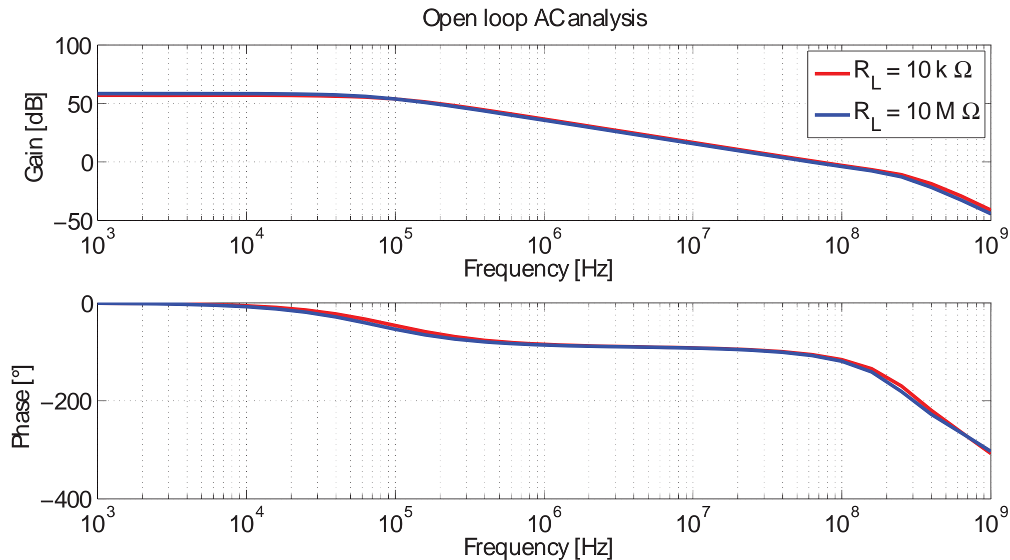


FIGURE 5.8: Simulated open-loop gain magnitude and phase of the amplifier (nominal conditions)

From the Bode plots in Fig. 5.8 (nominal process and operating conditions), the DC gain is about 57 dB for $R_L = 10$ kΩ and 58.5 dB for $R_L = 10$ MΩ. The unity-gain frequency ranges from 62 MHz ($R_L = 10$ kΩ) to 69 MHz ($R_L = 10$ MΩ). The phase margin is 73° in both cases and was found to range from 70° to 87° when considering all fabrication process and operating conditions. Finally, simulations showed a CMRR of 76 dB.

5.2.2 Measurements

5.2.2.1 Output buffer

The silicon area of the fabricated amplifier, not including pads, is 2710 μm^2. Power consumption, not including bias circuitry, is 3.7 mW. In the following of this Section, results obtained with a load resistance of 14.5 kΩ (which emulates a memory cell in its SET state), are shown.

All voltage measurements were carried out by means of an active microprobe. Fig. 5.9 shows the measured (I_{out-m}) and the simulated (I_{out-s}) current at pulse plateau for different values of V_{ampl}. A comparison between the experimental (V_{out-m}) and the simulated (V_{out-s}) output voltage at pulse plateau is provided in

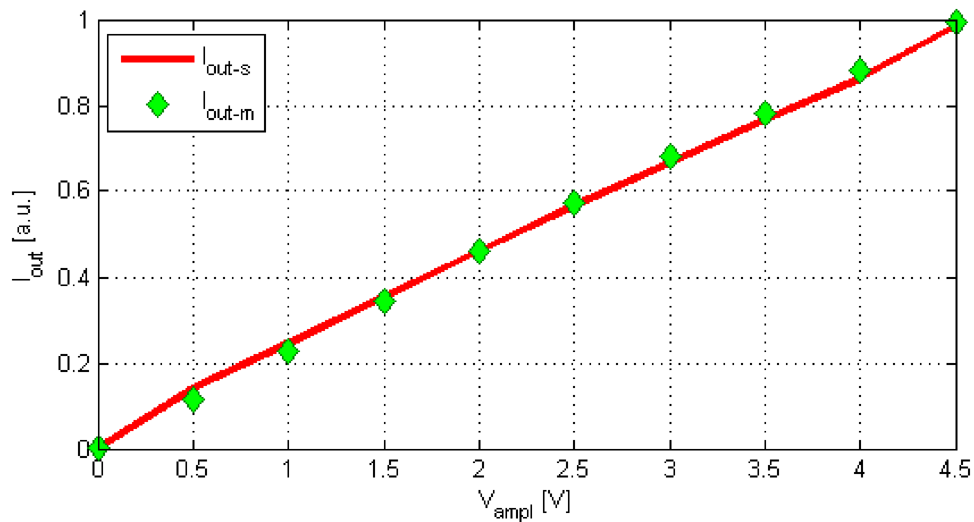


FIGURE 5.9: Comparison between the measured (I_{out-m}) and the simulated (I_{out-s}) output current at pulse plateau.

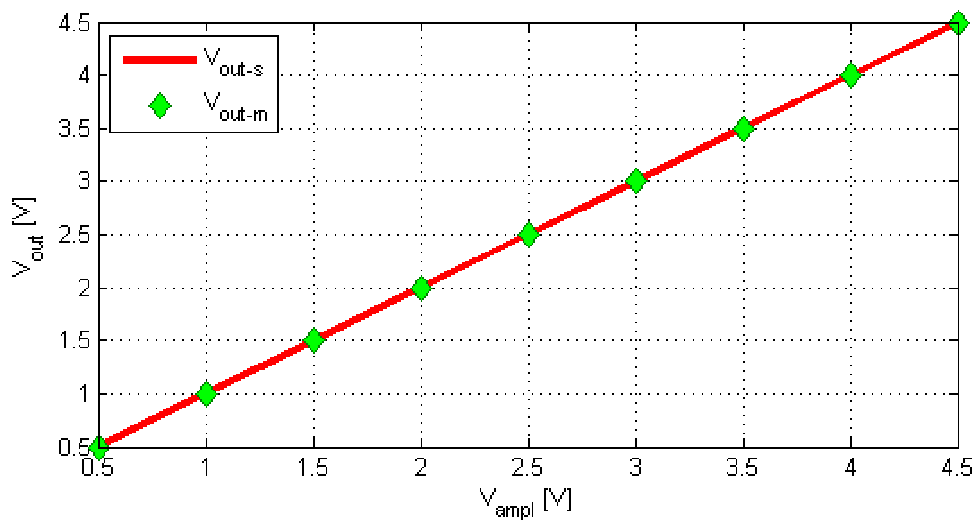


FIGURE 5.10: Comparison between the measured (V_{out-m}) and the simulated (V_{out-s}) output current at pulse plateau.

Fig. 5.10. Very good agreement between experimental data and simulated results is apparent in both figures. Fig. 5.10 also shows the pulse amplitude capability of the buffer: pulses with different amplitudes were provided to the buffer input and the buffer is shown to be able to replicate pulses with an amplitude from 0.5 V up to 4.5 V, thus meeting target specifications. The error between the voltage level of V_{ampl} and the amplitude at program pulse plateau was investigated to prove that no calibration procedure for pulse amplitude is needed. According to simulations, the open-loop DC gain of the output buffer is about 56 dB, which ideally leads

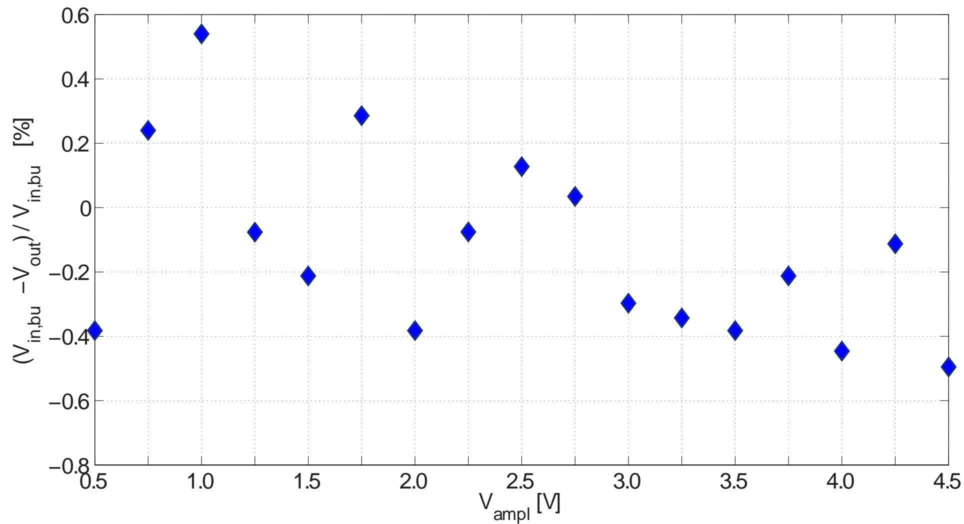


FIGURE 5.11: Measured relative error $\frac{V_{in}^{+}-V_{out}}{V_{in}^{+}}$ between the input and the output voltage of the buffer at pulse plateau

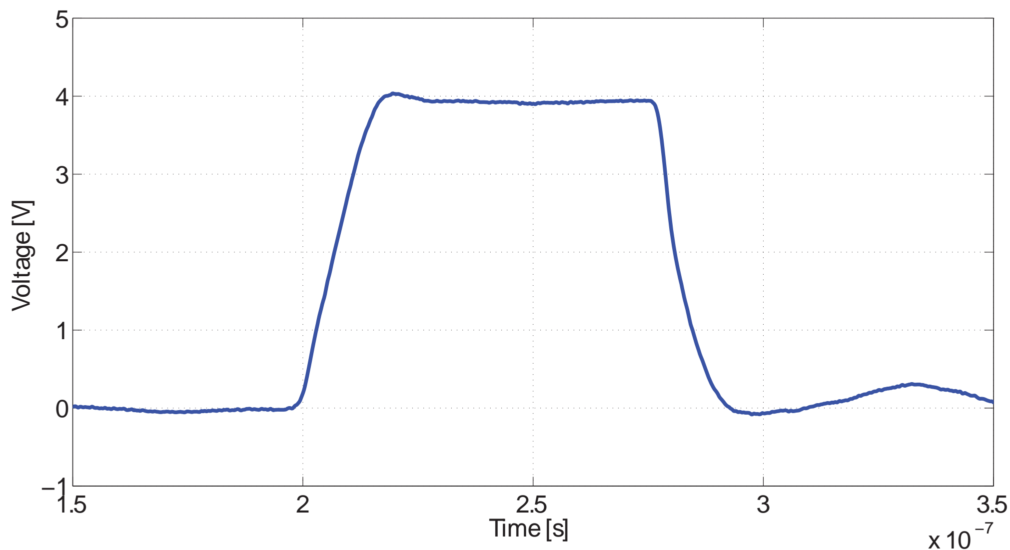


FIGURE 5.12: Measured output pulse with an amplitude of 4 V, a time duration of 53 ns, and rise and fall times of 12 ns

to an input-to-output voltage error of about ±0.2%. The experimental results in Fig. 5.11 show that the error in the range of interest is confined between -0.5% and +0.53%, which is more than adequate for our purposes. Then, the buffer ability to replicate pulses having short duration and fast rising and falling edges was experimentally evaluated (Fig. 5.12): the time duration is 53 ns and t_r and t_f are both about 9 ns, which fully meets target requirements.

Finally, Fig. 5.13 shows the output voltage and the voltage at pad I_{read}, V_{rd},

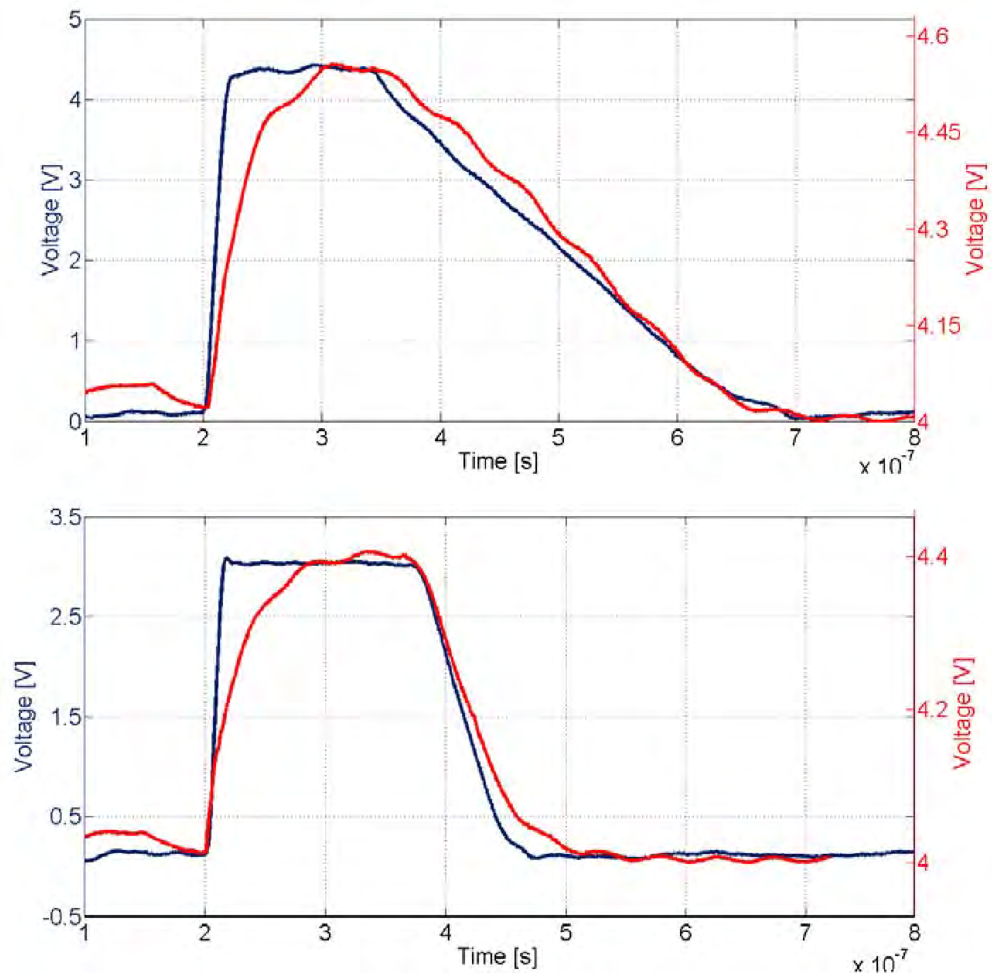


FIGURE 5.13: Measured output voltage (blue) and voltage at pad I_{read} (red) of two typical SET pulses. Upper plot: $V_{ampl} = 4.5$ V, $t_d = 130$ ns, $t_r = 14$ ns, $t_f = 280$ ns. Lower plot: $V_{ampl} = 3$ V, $t_d = 220$ ns, $t_r = 9$ ns, $t_f = 100$ ns.

(measured by means of a passive probe) of typical SET pulses. Pad *I_read* was connected to a bias voltage of 4.0 V through a resistor (R_{rd}) of 1.8 kΩ in order to set the drain of $M_{24,b}$ to the average voltage expected at the drain of M23. The measured current ($I_{rd} = \frac{V_{rd}}{R_{rd}}$) reaches its plateau after ≈ 80 ns, which is more than adequate for our application. Moreover, the pulse falling edge is replicated with adequate accuracy.

5.2.2.2 Pulse Generator

The resulting total area of the pulse generator is 15,165 μm^2. A microphotograph of the fabricated test chip is shown in Fig. 5.14 together with the corresponding

layout. The chip was bonded for debug and characterization purposes. The biasing current for analog circuits is provided externally through pad I_{bias}.

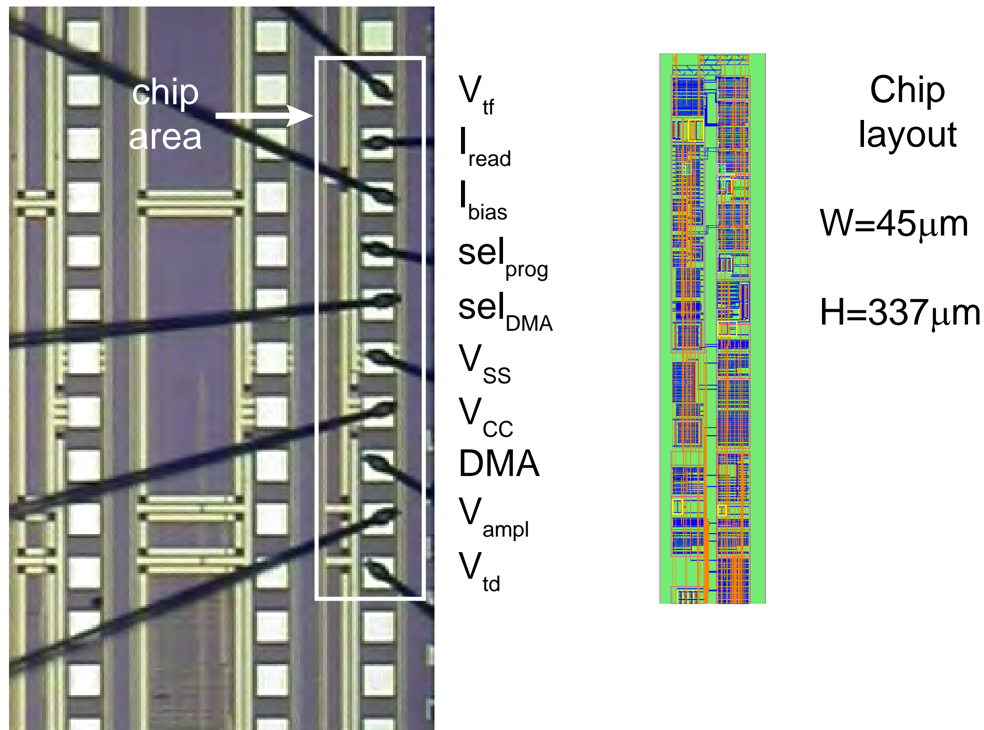


FIGURE 5.14: Chip microphotograph (left) and layout (right).

The test chip was provided with the possibility to test two integrated load resistors (which emulate the memory cells), so as to allow its performance to be evaluated under different load conditions. These resistors can be accessed in DMA mode to allow high-accuracy measurements. Since only two passive resistors are integrated, only one pad is provided to address the load. In the following, experimental results obtained with a supply voltage of 6 V, an equivalent load capacitance of 260 fF, which emulates the bit line capacitance, and an equivalent load resistance of 14.5 kΩ, which corresponds to the heaviest current sinking in our application, will be illustrated.

All the following experimental analysis, except measurements related to the current Track-and-Hold circuit, was carried out at the buffer output, after the cascaded selector SEL_1 (Fig. 3.1), by means of an active microprobe.

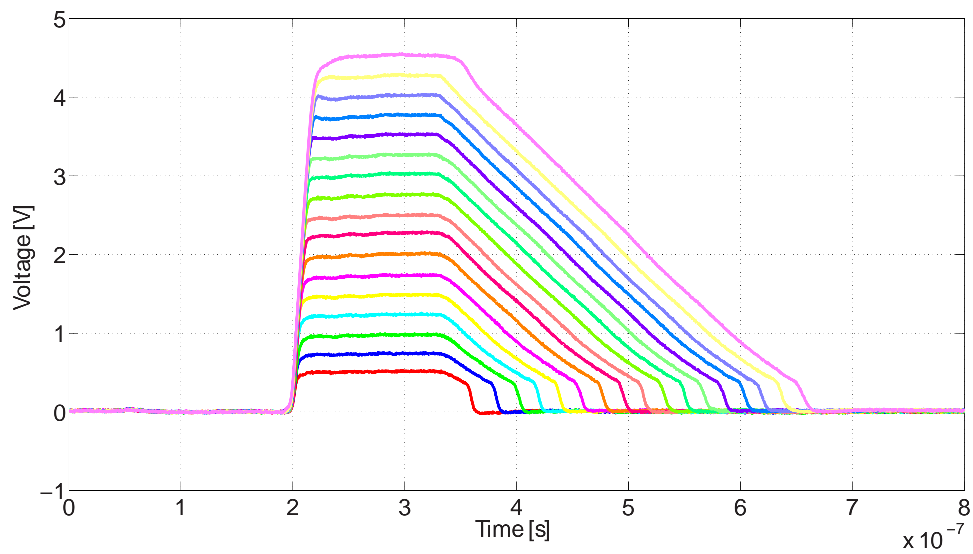


FIGURE 5.15: Measured voltage program pulse generated by varying the amplitude of V_{ampl} (500 mV step) while keeping V_{tf} and V_{td} constant ($\Delta t = 100$ ns).

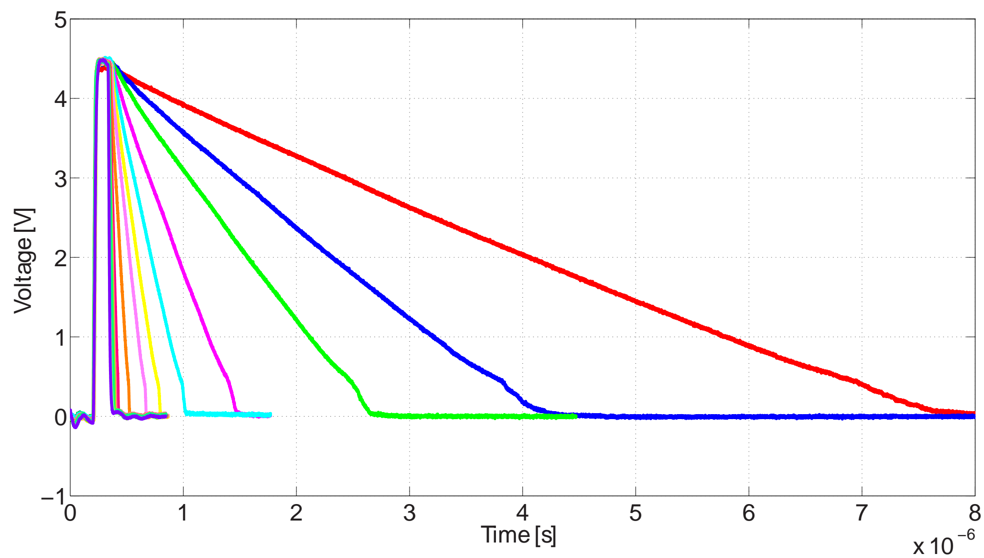


FIGURE 5.16: Measured voltage program pulse generated with different values of voltage V_{tf} (variable step) while keeping V_{ampl} and V_{td} constant ($\Delta t = 100$ ns).

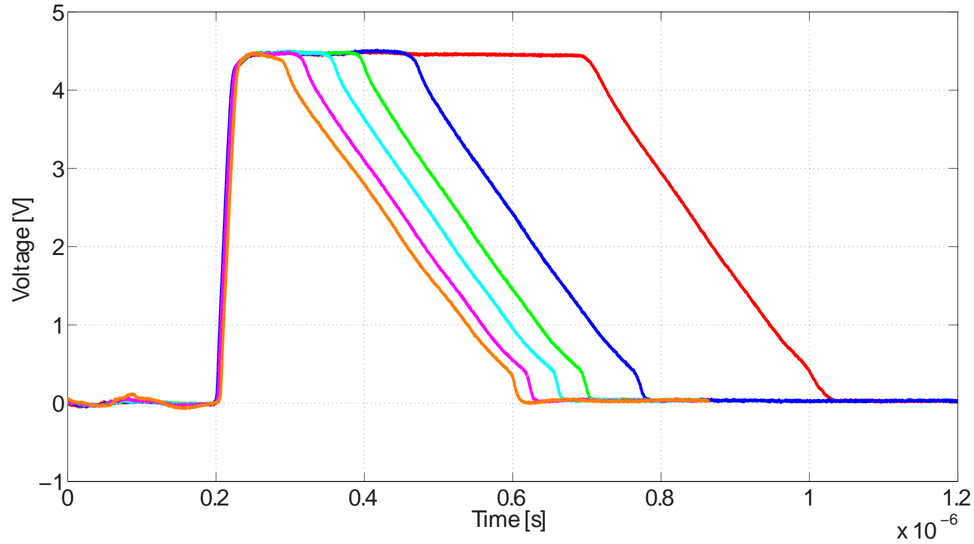


FIGURE 5.17: Measured voltage program pulse generated by varying V_{td} (500 mV step) while keeping V_{ampl} and V_{tf} constant (Δt = 100 ns).

Each of the program pulse control parameters (V_{ampl}, V_{tf}, V_{td}, and Δt) was separately varied in order to explore program pulse flexibility, whereas the other control parameters were kept constant.

The pulse voltage amplitude capability was investigated by keeping V_{tf} and V_{td} at constant values and varying V_{ampl} with steps of 250 mV ($\Delta t = 100$ ns). It is apparent from Fig. 5.15 that design specifications are fully met, since the generated program pulse amplitude ranges from 0.5 V up to 4.5 V. In addition, the system (internal pulse generation and pulse buffering) is proven not to give rise to overshoots.

The pulse fall time variability was investigated by setting $V_{ampl} = 4.5$ V, keeping V_{td} at a constant value, and varying V_{tf} ($\Delta t = 100$ ns). Since the performed voltage-to-time conversion is non-linear, a variable step of V_{tf} was chosen. Measurements showed that, with the used range of V_{tf}, a minimum and a maximum fall time of about 9 ns and 6 μs, respectively, are achieved (Fig. 5.16), which fully meets design targets.

The pulse time duration performance was investigated by setting $V_{ampl} = 4.5$ V, keeping V_{tf} at a constant value, and varying V_{td} ($\Delta t = 100$ ns). Measurements (Fig. 5.17) showed that the specified minimum (50 ns) and maximum (350 ns)

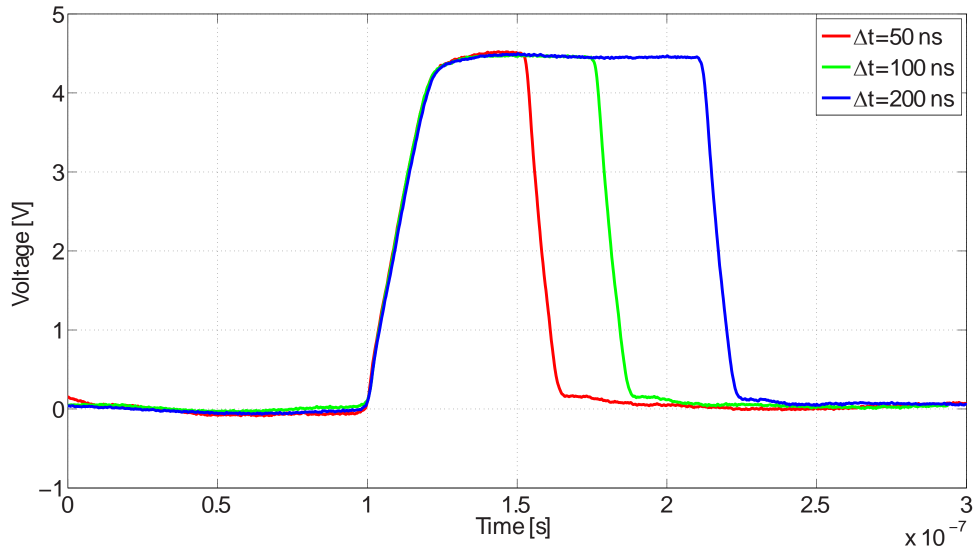


FIGURE 5.18: Measured voltage program pulse generated by varying Δt (50 ns step) while keeping V_{ampl}, V_{tf} and V_{td} constant (the trigger occurs in different time instants for the three pulses).

TABLE 5.2: Measured fall times and time durations without and with calibration.

	Pulse A $V_{ampl} = 2$ V		Pulse B $V_{ampl} = 4.5$ V	
	t_d [ns]	t_f [μs]	t_d [ns]	t_f [ns]
Target	350	2	100	80
Non-calibrated	463	3.07	140	125
Error [%]	+32.3	+53.5	+40.0	+56.3
Calibrated	342	2.05	103	85
Error [%]	-2.3	+2.5	+3.0	+6.3

time duration values are fully achieved (the longest pulse in Fig. 5.17 is about 440 ns). Further characterization was carried out by varying Δt from 50 ns to 200 ns, while keeping V_{td} constant. This set of pulse time duration measurements confirmed two effects: *i*) the variation of t_d is proportional to Δt (Fig. 5.18), as shown by (2.10), thus enhancing the programmable range of t_d, and *ii*) the time instant in which the program pulse is generated varies with Δt (Fig. 5.19), in agreement with Fig. 2.8.

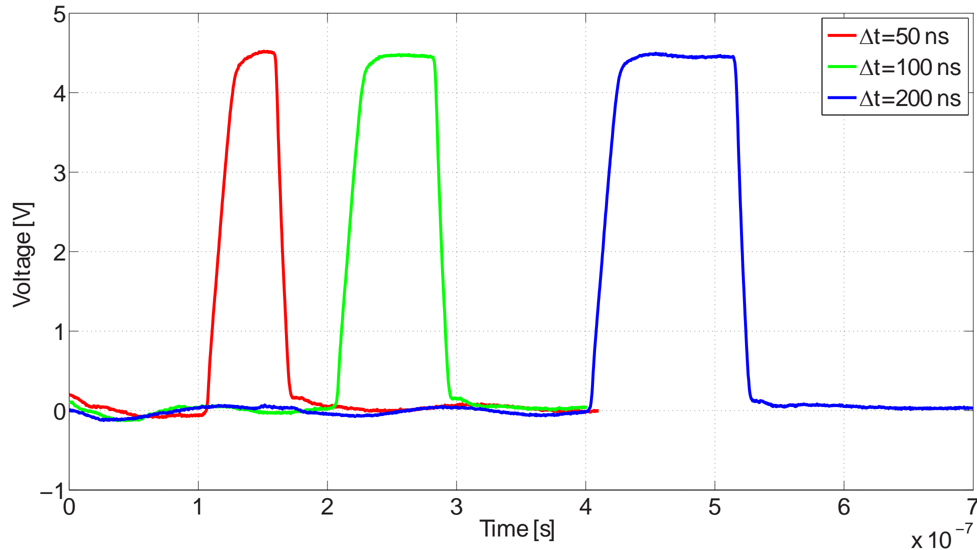


FIGURE 5.19: Measured voltage program pulse generated by varying Δt (50 ns step) while keeping V_{ampl}, V_{tf} and V_{td} constant (the trigger occurs in the same time instant for all pulses).

5.2.2.3 Calibration

After demonstrating the flexibility of the system in generating the pulses, the calibration procedure was evaluated.

Since four measurements are required to carry out the calibration procedure, the four pulses in Fig. 5.20 were generated and their t_d and t_f were measured so as to determine the values of the unknowns in (3.30) and (3.32). These four pulses were chosen so as to take the specified minimum and maximum values of time duration and fall time into account.

In order to test the effectiveness of the calibration procedure, two pulses, namely pulse A and pulse B, with the target timing parameters shown in the first row of Table 5.2, were generated.

Pulse A aims at testing the accuracy in generating a long pulse ($t_d = 350$ ns) with a small amplitude ($V_{ampl} = 2$ V) and a long fall time ($t_f = 2$ μs), which corresponds to a low value of I_{tf}. Pulse B, instead, aims at testing system accuracy in generating a short pulse ($t_d = 100$ ns) with a short fall time ($t_f = 80$ ns) and a large amplitude ($V_{ampl} = 4.5$ V), which corresponds to a high value of I_{tf}.

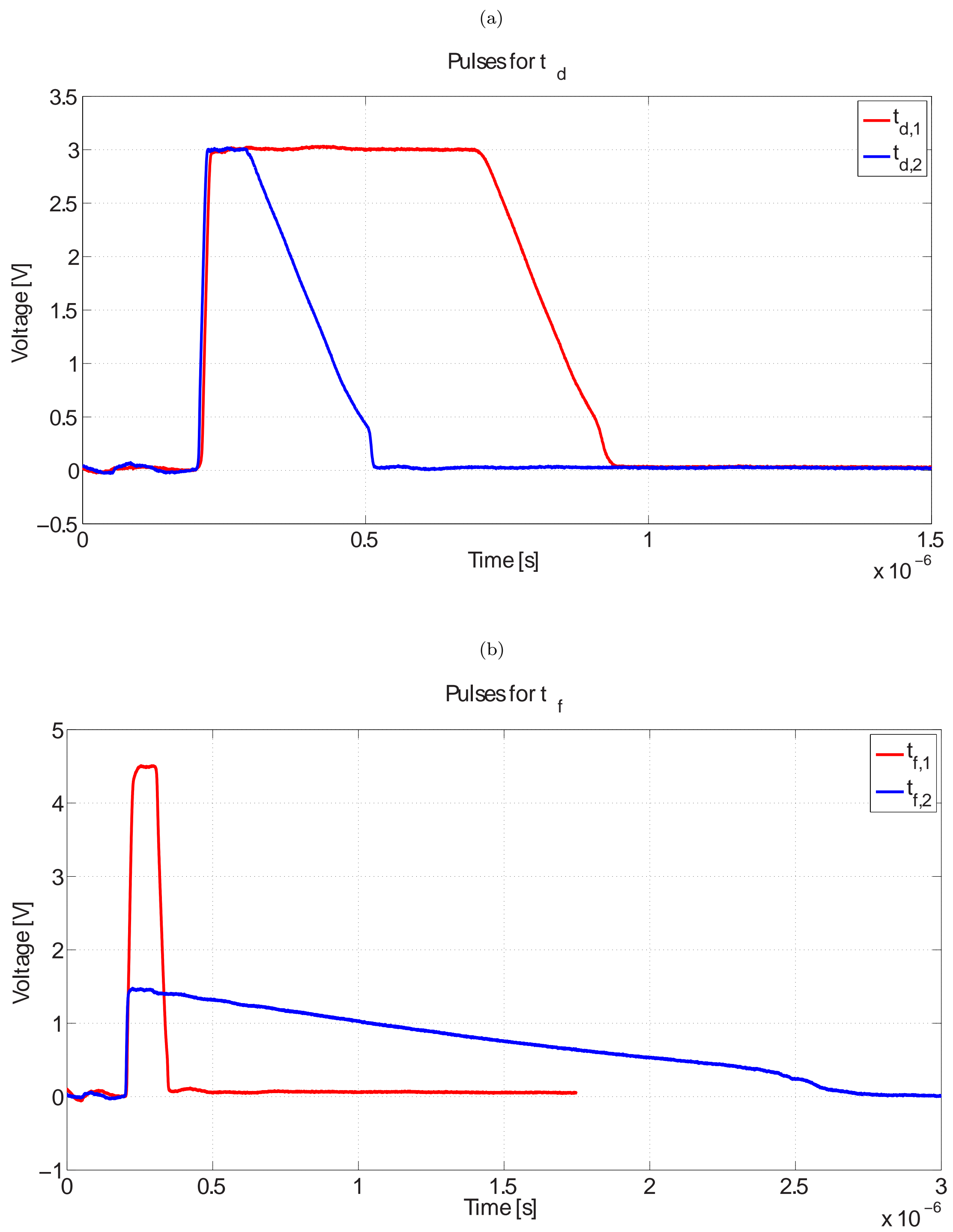


FIGURE 5.20: Pulses generated to perform the trimming procedure: pulses for t_d evaluation (a); pulses for t_f evaluation (b).

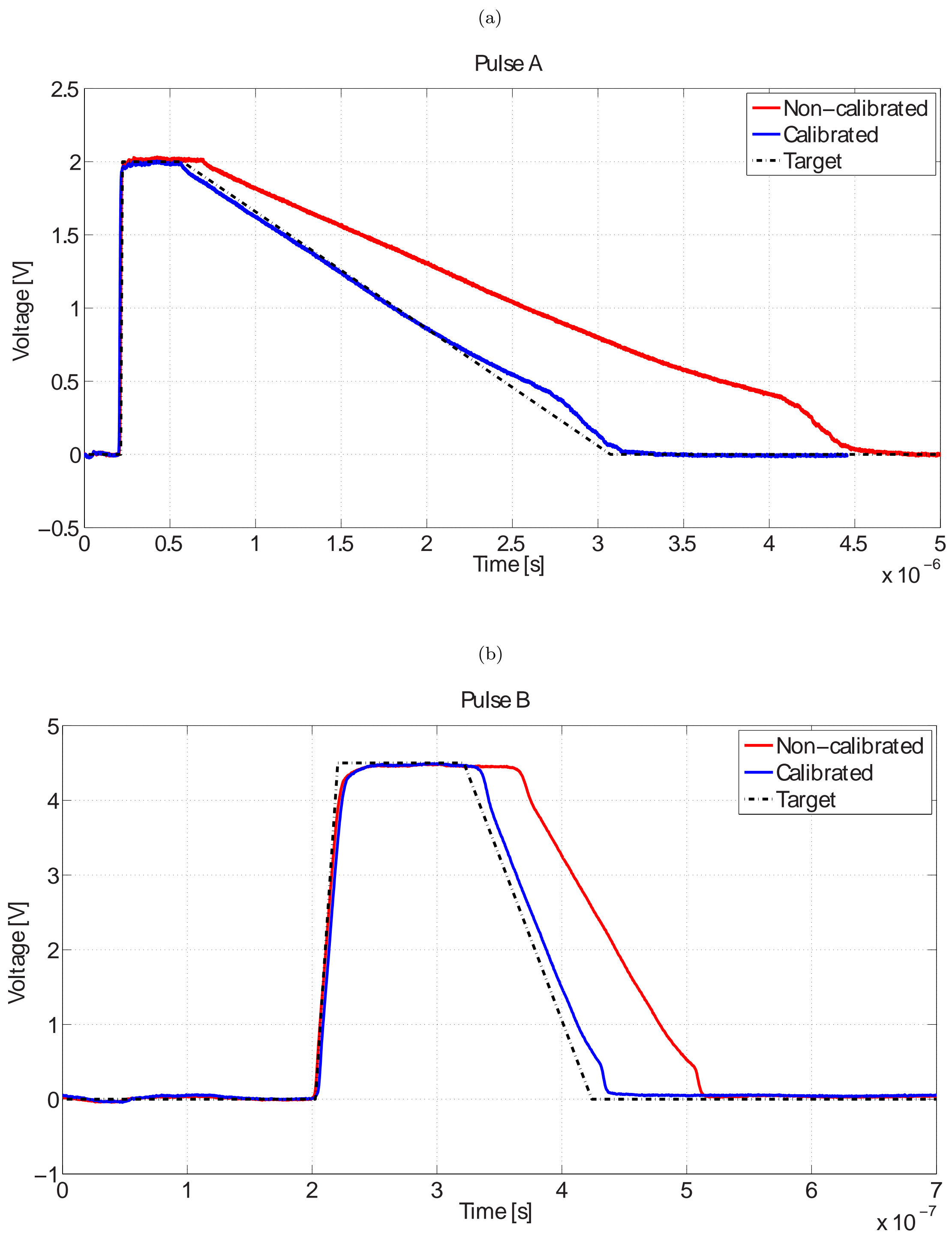


FIGURE 5.21: Measured calibrated and non-calibrated pulses compared to target pulse: pulse A (a) and pulse B (b).

Figures 5.21(a) and 5.21(b) compare the measured calibrated and non-calibrated pulses to the target pulses. The obtained results are summarized in Table 5.2. Non-calibrated pulses were obtained by substituting the values corresponding to the nominal process parameters for the unknowns in (3.30) and (3.32).

From both Fig. 5.21 and Table 5.2, the accuracy improvement due to the calibration procedure is apparent: the error is always kept within target specifications.

5.3 Final implementation

The final test-chip was fabricated and characterized using the commercial ATE, so no waveform are available. In this Section, therefore, the most significant simulation results of the test-chip are shown first. All simulations refer to the typical case of wafer fabrication process. Then, measurement are discussed.

5.3.1 Simulations of the final implementation

5.3.1.1 Output buffer

The simulated output buffer performance is shown first.

AC analysis of the buffer in open-loop configuration was performed. Simulation conditions for AC analysis are summarized in Table 5.3.

Fig. 5.22 illustrates the magnitude of the output buffer open-loop gain.

The load resistance R_{load} was set to 13 kΩs.

The DC gain of the buffer is 60 dB, the -3 dB frequency is 50 kHz and the unity gain frequency is 50 MHz.

Fig. 5.23 shows the open-loop phase response of the output buffer. The phase margin is 75°, which is reduced to 73° in worst-case fabrication process conditions. Further simulations with a load resistance of 100 MΩ were carried in order to verify

TABLE 5.3: Simulation conditions for output buffer AC analysis

Name	**Value**	**Notes**
V_{CC}	6 V	Power supply
V_{SS}	0 V	Ground (GND)
V_{high}	4.5 V	Voltage applied to unaddressed Word Lines; I_{bias}
V_{ampl}	3 V	Buffer DC input (all transistors in saturation)
V_{ben}	6 V	Buffer always enabled
R_{load}	13 kΩ	PCM cell resistance during programming
C_{load}	250 fF	Capacitive load of a Bit Line

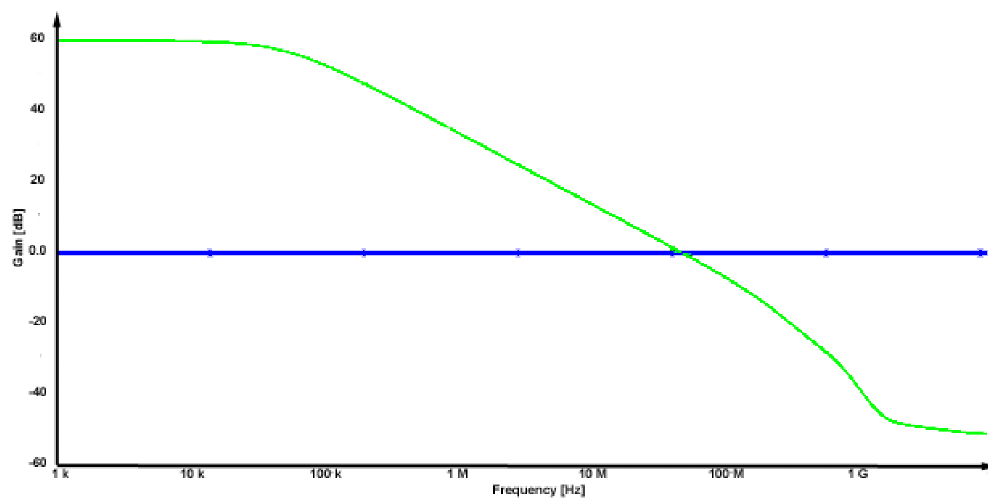


FIGURE 5.22: Simulated magnitude of the output buffer open-loop gain.

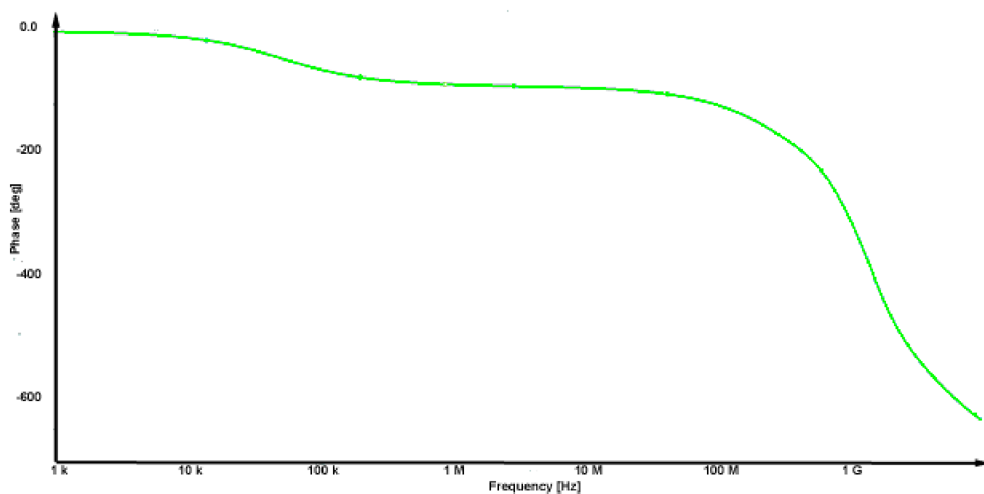


FIGURE 5.23: Simulated open-loop phase response of the output buffer.

Table 5.4: Simulation conditions for transient analysis of the output buffer

Name	**Value**	**Notes**
V_{CC}	6 V	Power supply
V_{SS}	0 V	Ground (GND)
V_{high}	4.5 V	Voltage applied to unaddressed Word Lines; $I_{bias} = 100\ \mu$A
R_{load}	13 kΩ	PCM cell resistance during programming
C_{load}	250 fF	Capacitive load of a Bit Line
L_R	0 V	Choice of the left-side load
sense	2.25 V	Voltage provided by the external equipment to read the program current
Δt	100 ns	Delay time (typical value)

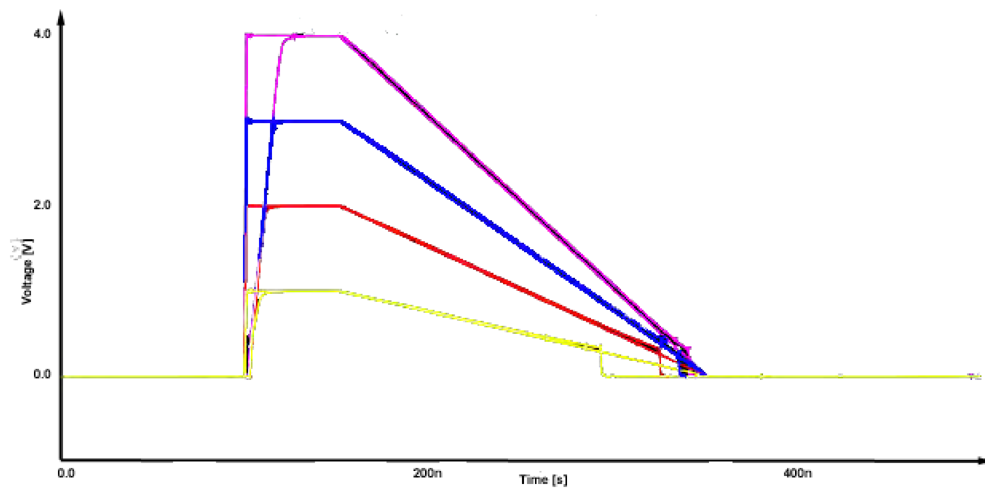


Figure 5.24: Simulated program pulses.

the buffer stability even when reading a memory cell is in its reset state, which represents the worst case from the frequency stability point of view because a light load pushes the secondary pole (due to the output node) towards low frequencies. The minimum phase margin obtained in worst-case process conditions is 71° (the open-load DC gain magnitude was again 60 dB).

Transient analysis of the test chip output buffer has then been performed. Simulation conditions are summarized in Table 5.4.

Fig. 5.24 shows simulated program pulses with different values of amplitude and falling time. The waveforms with the faster rising slope are the simulated inputs of the output buffer, whereas the others (slower rising slope) are the simulated

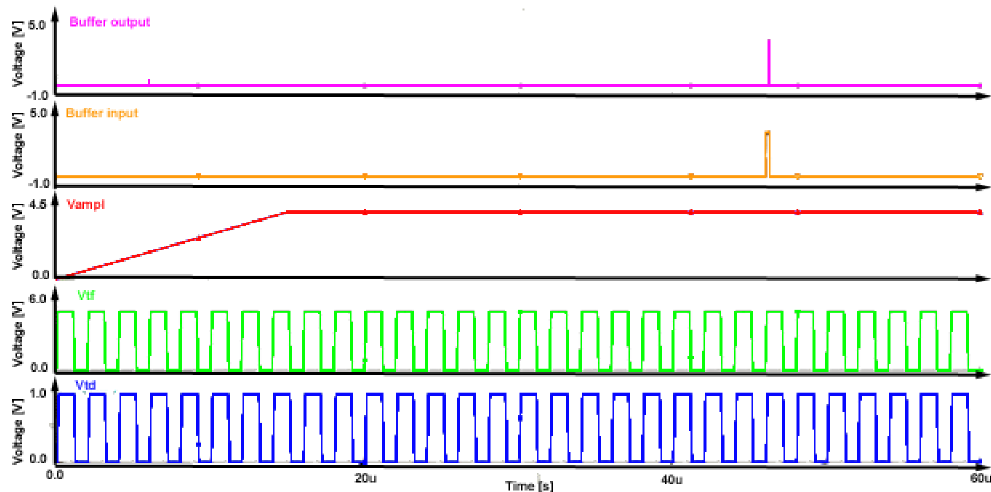


FIGURE 5.25: Input and output waveforms.

outputs. The maximum delay time between the corresponding input and output waveforms is 3 ns. The maximum estimated rise time of the output waveform in the typical case is 8 ns when $V_{ampl} = 4$ V, which demonstrates that the rising time specifications are fully satisfied. The maximum simulated rise time was found to be 12 ns in worst-case process conditions.

5.3.2 Whole system

Fig. 5.25 shows an example of pulse generation. The pink waveform represents the generated pulse; the orange waveform represents the waveform at the buffer input; the red waveform represents signal V_{ampl}. A 100 ns delay is present between the green and the blue waveform (signal V_{tf} and V_{td}, respectively).

V_{ampl} has the highest program pulse amplitude (4.5 V). Its rise and fall times were set to 200 μs in order to simulate the rise and fall time of the correspondent external signal from the external equipment. The value of its period was chosen to allow a complete write-read cycle.

When V_{ampl} rises above 1.5 V, the counter is enabled and the pulse is applied to the cell after 20 clock cycles, according to specifications.

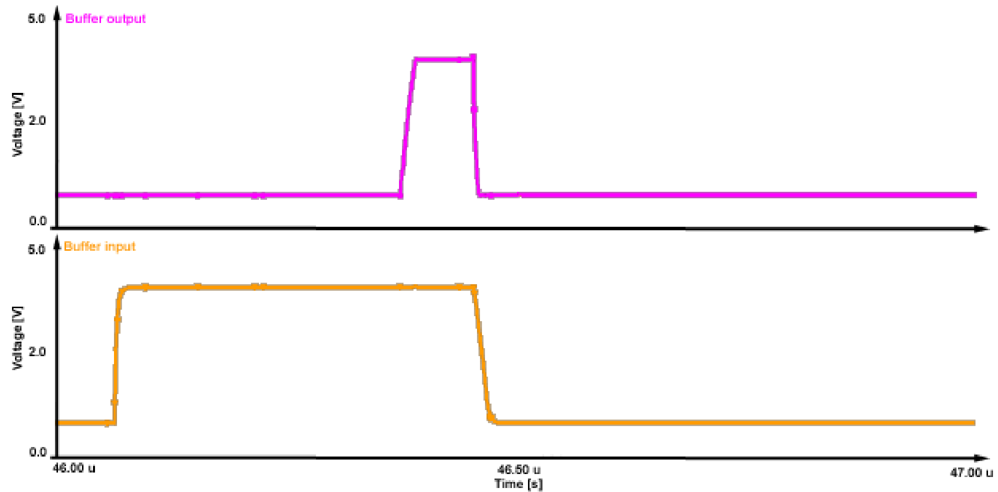


FIGURE 5.26: Detail of a pulse with fast fall.

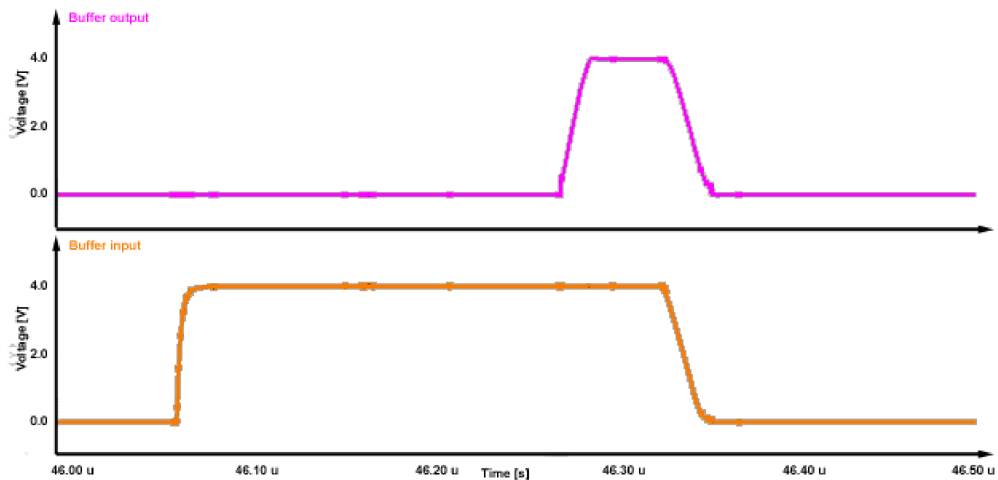


FIGURE 5.27: Detail of a pulse with $t_f = 20$ ns.

Two details of the buffer input and the pulse applied to the cell are shown in Fig. 5.26 and Fig. 5.27. V_{tf} was set to 6 V (fast fall, Fig. 5.26) and 5 V (Fig. 5.27). In Fig. 5.26, the buffer output voltage falls faster than the buffer input voltage, whereas in Fig. 5.27, the buffer input and the pulse applied to the cell fall together.

Some significant cases were also explored in order to evaluate the accuracy of the system without any calibration operation.

The used input signals lead to programming pulses with the same t_d, different V_{ampl} (4 V for the blue line, 3 V for the green line, and 2 V for the red line), and different fall times.

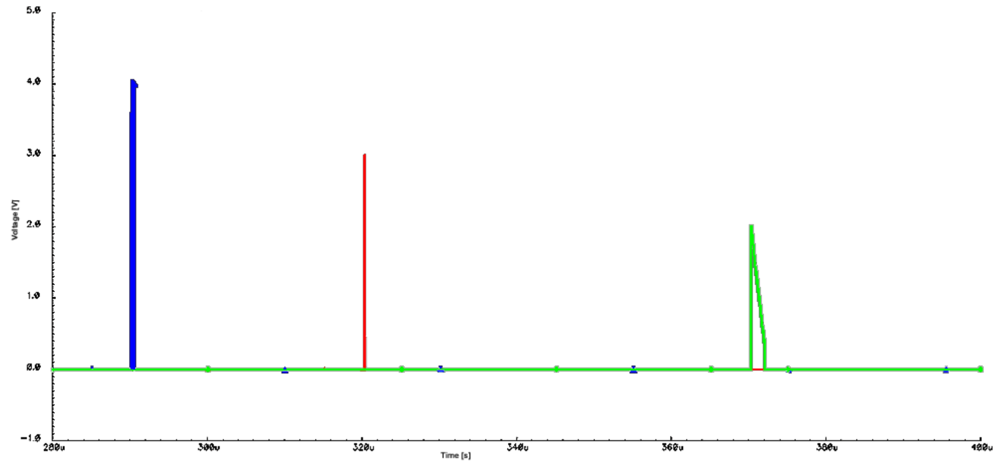


FIGURE 5.28: Simulated program pulses (blue line: V_{ampl} = 4 V; green line: V_{ampl} = 3 V; red line: V_{ampl} = 2 V).

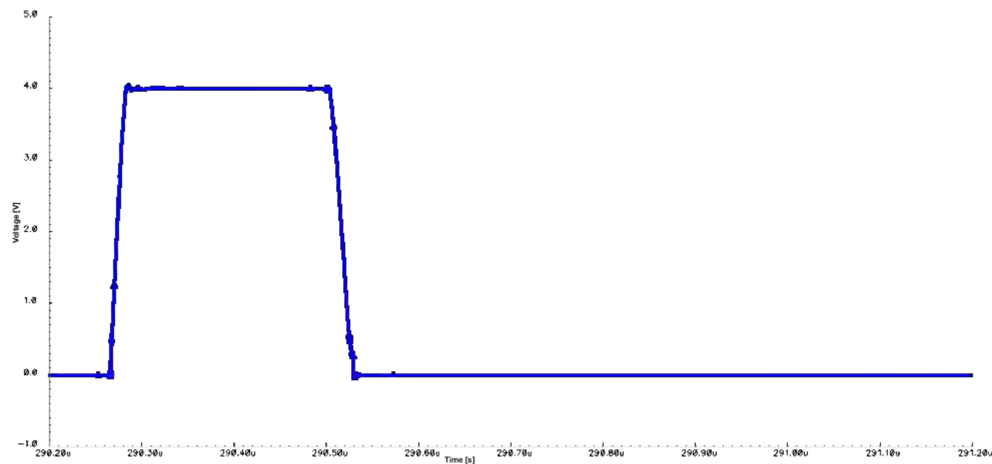


FIGURE 5.29: Zoom of Pulse I.

According to experimental observations of the used ATE, the rising time of V_{ampl} was set to 200 μs in every simulation (Fig. 5.28). The three pulses occur in different instants because V_{ampl} has different rising slopes depending on its amplitude.

The generated pulses are shown in detail in Fig. 5.29 to 5.31.

5.3.3 Experimental results of the final implementation

The fabricated test-chip microphotograph is shown in Fig. 5.32. First, the overall functioning of the system was evaluated by generating RESET and SET programming pulses, that were applied to the selected memory cell. After every pulse, the

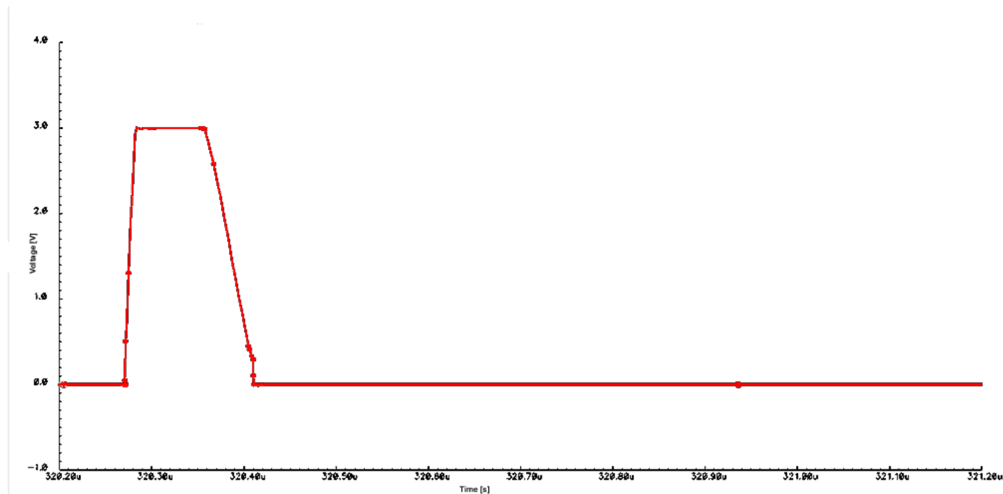


FIGURE 5.30: Zoom of Pulse II.

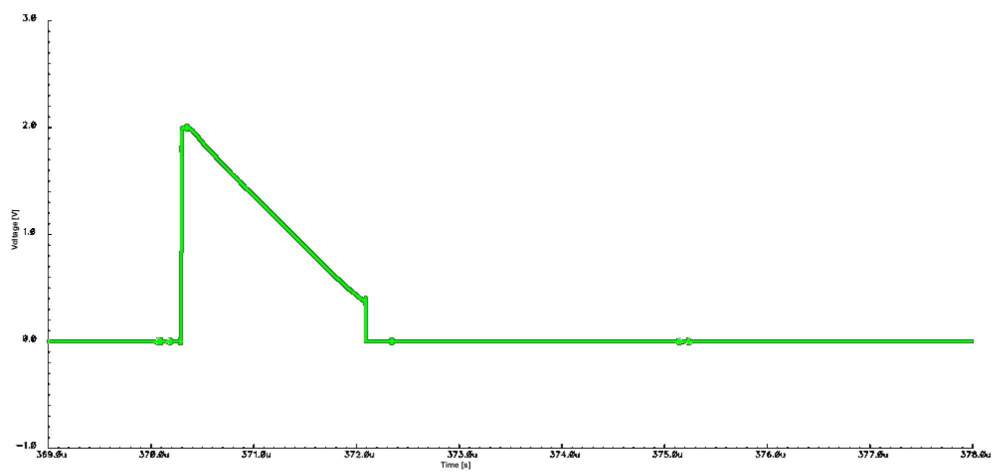


FIGURE 5.31: Zoom of Pulse III.

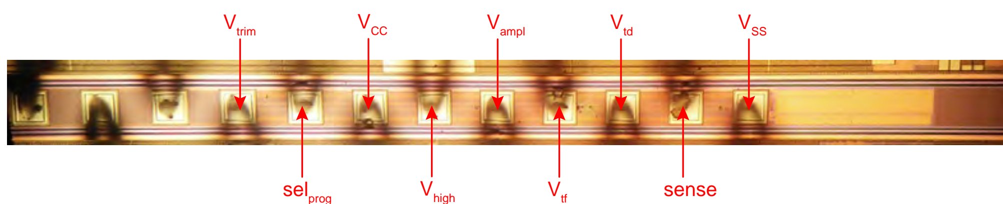


FIGURE 5.32: Test-chip microphotograph.

state of the cell was read in DMA mode. The reading current after a full-RESET operation was 130 nA, whereas the reading current after a full-SET operation turned out to be 23.5 μA, thus confirming the programming operations were both successful.

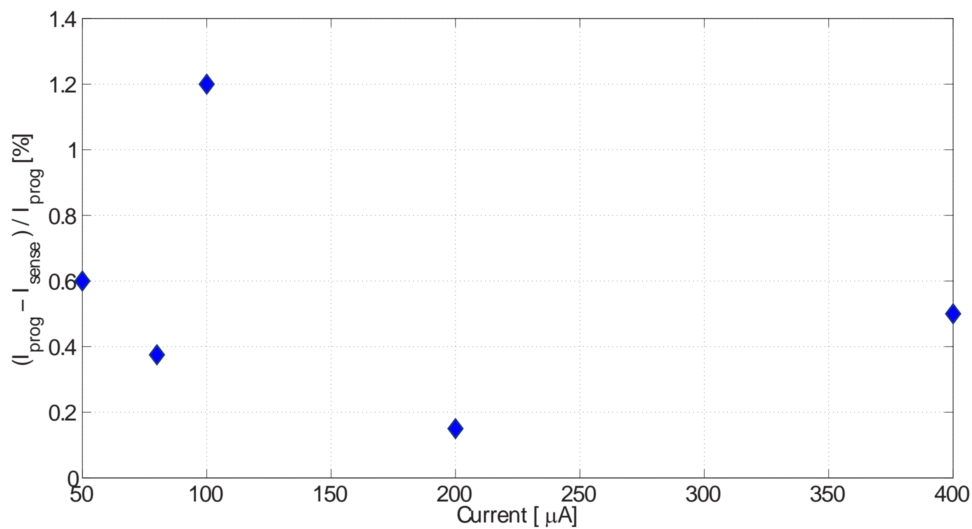


FIGURE 5.33: Error between the mirrored current I_{prog} and the current read by test equipment, I_{sense}, when forcing a voltage equal to the one expected on pad I_{sense}.

5.3.3.1 Measurements of the current Track-and-Hold circuit

To evaluate the current Track-and-Hold circuit performance, several tests were performed on a stand alone Track-and-Hold cell integrated for characterization purposes. This way, the performance of the Track-and-Hold circuit were explored by applying ranges of currents I_{prog} wider than expected in normal operation. In fact, the currents used for these measurements range from 50 μA to 400 μA, whereas expected program currents range from 70 μA to 225 μA.

First, readout accuracy was tested for different values of current I_{prog}. For any current value, the voltage expected on the drain/gate node of transistor $M_{T\&H}$ at the end of the tracking phase was forced on pad I_{sense}. As shown in Fig. 5.33, the accuracy is within ±1.2% for currents from 50 μA to 400 μA, which is considered adequate for our application.

Then, the voltage applied to pad I_{sense} was changed while keeping test current I_{prog} constant, so as to estimate the error introduced by channel length modulation effects (four different values of I_{prog} in the range from 50 μA to 400 μA were tested). The error is within the range from -2.1% to +5% (Fig. 5.34), and is minimized when forcing a voltage of about 3 V on pad I_{sense}.

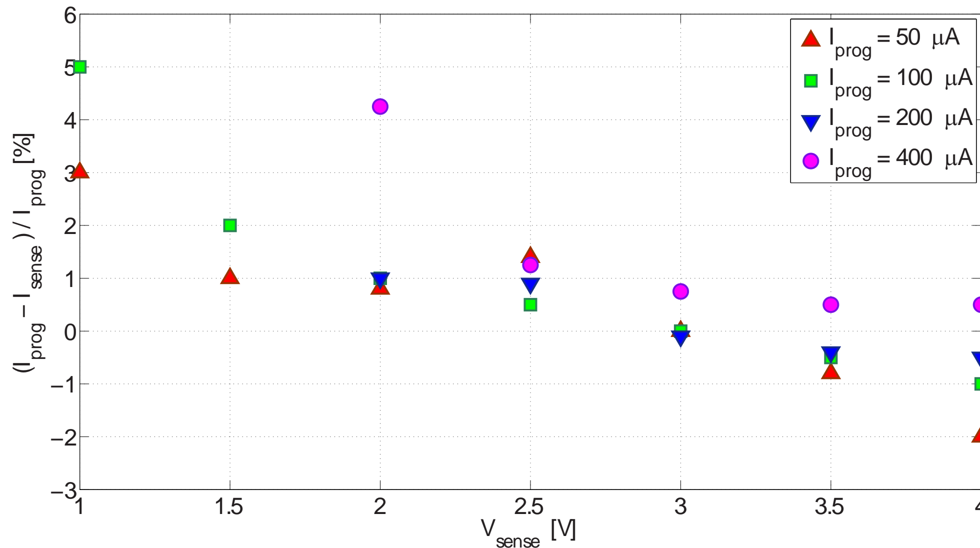


FIGURE 5.34: Error between the mirrored current I_{prog} and the current read by test equipment, I_{sense}, when varying the voltage on pad I_{sense}.

5.3.4 Calibration procedure

TABLE 5.5: Measured fall times after calibration.

V_{ampl}	Target t_f [μs]	Measured t_f [μs]	Error [%]
2.0	2.00	2.16	+8
2.5	1.00	1.02	+2
3.0	0.08	0.076	-5

TABLE 5.6: Measured time durations after calibration.

Target t_d [ns]	Measured t_d [ns]	Error [%]
200	194	-3.0
150	196	-2.7
90	91	+1.1

Finally, the automatic calibration procedure was evaluated. The pre-calibration was performed in order to find the ratio $\frac{C_{cal}}{I_{cal}}$, according to (4.14), then the four measurements required to carry out the calibration procedure were generated and their t_d and t_f were derived from (4.13), so as to determine the values of the unknowns in (4.18) and (4.19).

In order to test the effectiveness of the calibration procedure, six pulses were generated, three for the fall time and three for the time duration. Those pulses could not be observed directly because the programming pulse can only be accessed

by means of a microprobe, which can not be used with commercial ATE. In order to be able to measure these pulses, they were generated while the system was still in calibration mode. This way, the calibration hardware was exploited to read the timing parameter under test. The results for the fall time calibration are shown in Table 5.5, whereas the results for the time duration calibration are shown in Table 5.6.

From both Tables, the automatic calibration procedure is validated: the error is always kept within target specifications.

Chapter 6

Model for Phase Change Memories

6.1 Introduction

In this Chapter, a compact model for PCM cells, which evaluates the state of the cell (given as the current flowing across the cell when a reading voltage V_{read} of 0.3 V is applied between the TEC and the BEC) during and after a programming operation, is described. First, the amorphization and crystallization kinetics were taken into account. A continuous-time MatLab model, which makes use of first-order differential equations, was then implemented so as to describe the programming operation as a dynamic system.

The basic idea of the proposed model is shown in the flow-chart in Fig. 6.1. Once a pulse is applied to the cell, the voltage determines the ensuing temperature increase at the heater-GST interface. This temperature determines whether a RESET operation, a SET operation or a reading operation is being performed: the reading current, which gives information about the state of the cell, is then derived. The reading current, and hence the state of the cell, is continuously updated during the whole simulation.

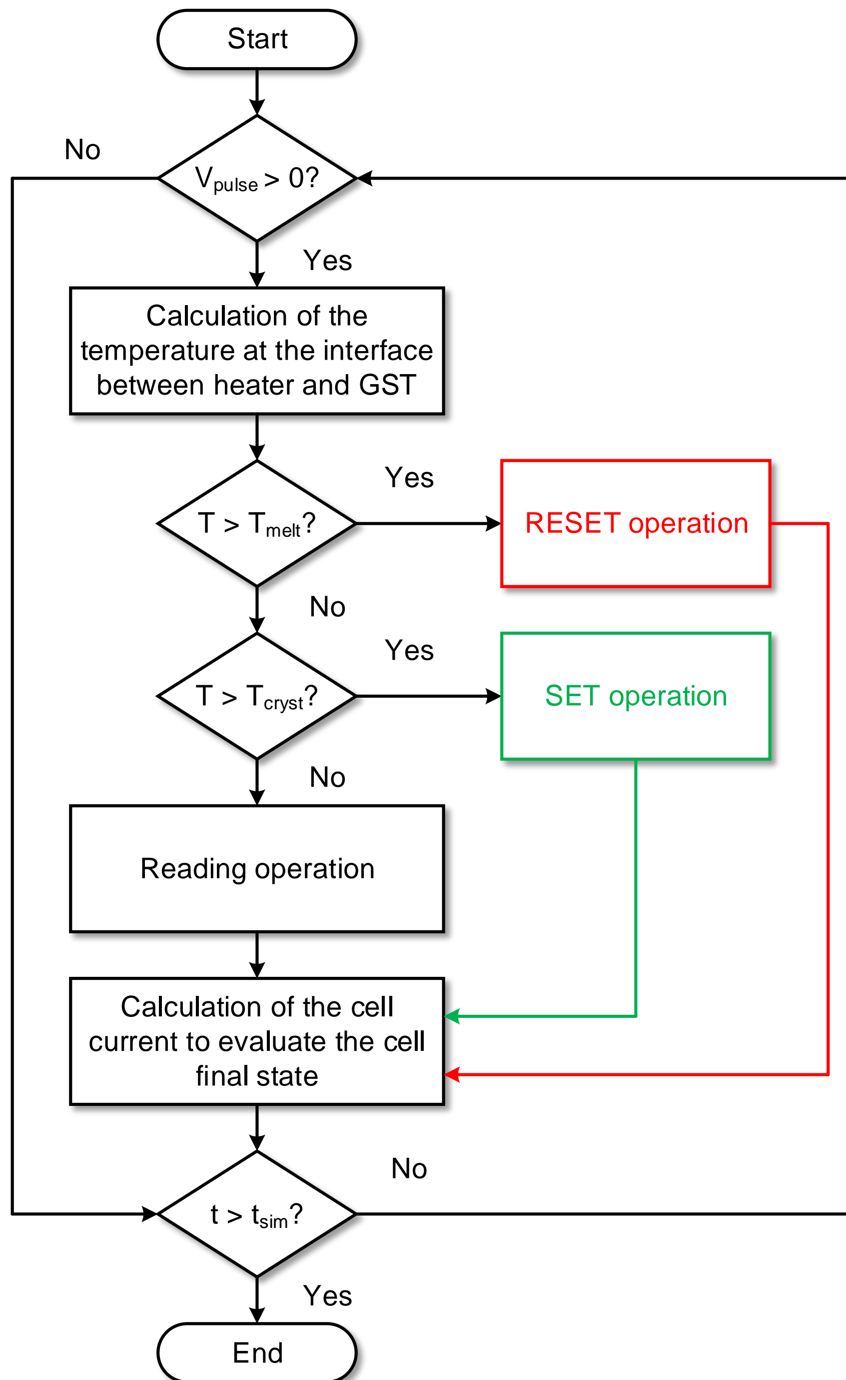


FIGURE 6.1: Flow-chart describing the basic idea of the model.

The increase of the temperature, ΔT_J, at the heater-GST interface is due to Joule heating and can be derived as

$$\Delta T_J = R_{th} \frac{V_{pulse}^2}{R_h} \tag{6.1}$$

where R_{th} is the equivalent thermal resistance of the memory cell (whose value depends on the GST phase), V_{pulse} is the programming voltage and R_h is the resistance of the heater. The resistance of the GST is neglected, since during a programming operation the GST is in its ON region, as explained in Chapter 1. The absolute temperature at the interface is easily found by adding the operating temperature to ΔT_J. Once the temperature at the interface is known, it is possible to evaluate which portion of the GST layer crystallizes or amorphizes, depending on the GST thermal profile.

In both partial-SET and partial-RESET programming algorithms described in Chapter 1, the final state of the cell depends on the distribution of the amorphous and crystalline phases inside the GST layer. In this respect, a key role is played by the thickness of the amorphous cap obtained after the RESET operation. In fact, the amorphous cap thickness is fundamental in both partial-RESET and in partial-SET programming. In the first case, since the resistivity of the amorphous phase is much greater than the resistivity of the crystalline phase, the contribution of the resistivity of the amorphous phase, which is proportional to the amorphous cap thickness, dominates the overall GST resistance to a first order. In the second case, the amorphous cap thickness, resulting from a previous RESET operation on the cell, affects the maximum value of the cell resistance and the resistance range where programmed states can be placed (programming window).

During the RESET operation, the GST starts to amorphize. At the end of the programming operation, the GST resistance will be equal to the series of the amorphous cap resistance and the resistance of the residual crystalline portion (Fig. 6.2(a)).

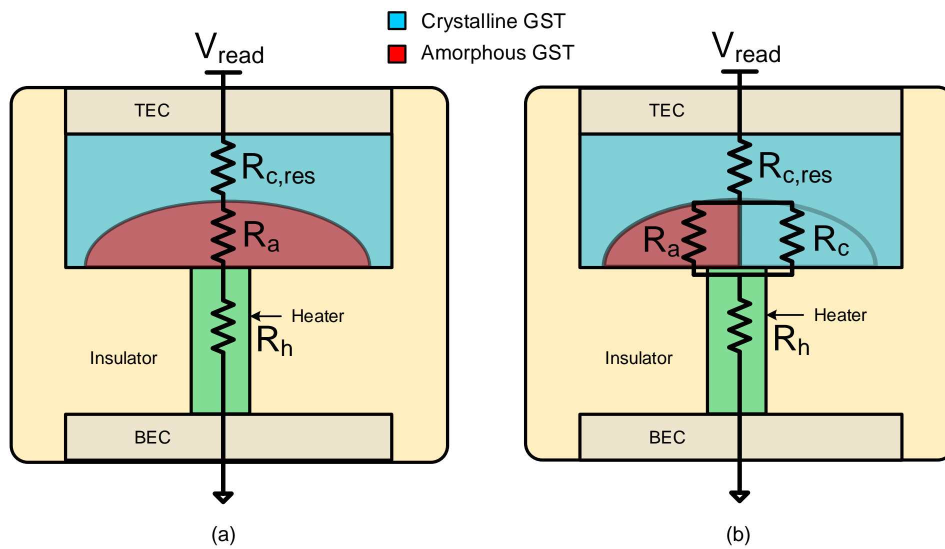


FIGURE 6.2: Equivalent resistance of GST after a RESET (a) or a partial-SET (b) operation.

In contrast, during the SET operation, the amorphous cap will start crystallizing. At the end of the programming operation, the GST resistance will be evaluated considering that a portion of the amorphous cap can still be present (this is the case of partial-SET programming). The resistance of the GST can be modelled as the series of the resistance of the residual crystalline portion and the resistance resulting from the parallel connection of the resistance of the residual amorphous cap and the resistance of the re-crystallized portion (Fig. 6.2(b)).

The cell behaviour is described in the model mainly by using the Johnson-Mehl-Avrami (JMA) theory [36], [37], [38] for the crystallization mechanism (SET operation) and the amorphization equation (RESET operation), as will be explained later in this Chapter.

Since in 1937, Kolmogorov, a Russian mathematician, conceived the same theory by using a probabilistic approach, the JMA theory is therefore usually addressed as Johnson-Mehl-Avrami-Kolmogorov (JMAK) theory.

The model was validated by comparison with experimental data from μtrench PCM cells [39].

This Chapter is organized as follows. In Section 6.2, crystallization and amorphization kinetics are described. In Section 6.3, the modelling of the RESET operation is described. An analysis on PCM cells crystallization based on experimental data is then carried out and the modelling of the SET operation is discussed. Finally, the whole model is validated in Section 6.4.

6.2 Crystallization and amorphization kinetics

6.2.1 Crystallization kinetics

The crystallization process is characterized by a re-arrangement of the atoms of the material in an ordered structure. In the case of GST, the atoms previously arranged in a disordered structure (amorphous phase), rearrange themselves in an ordered lattice (crystalline phase). This process, which takes few tens of nanoseconds, is driven by two phenomena, namely, nucleation and growth.

The nucleation begins with the formation of new small crystal grains (embryo clusters) inside the amorphous material, when the temperature inside the GST rises above the crystallization temperature T_{cryst} ($\approx 200\ ^{\circ}C$). The nucleation rate depends both on the temperature of the material and the time the material remains at this temperature.

The crystalline grains are unstable and, once they have reached a critical size, they either become stable nuclei or they vanish. Once the nuclei are stable, the growth process begins and the grains become bigger and bigger.

The JMAK theory is typically used to quantitatively describe the crystallization process because of its flexibility and simplicity.

JMAK model describes the temporal evolution of the Crystal Fraction (CF), which is the volumetric percentage of the crystallized material with respect to the total amorphous volume at the beginning of the crystallization process.

$$CF = \frac{final\ crystallized\ volume}{initial\ amorphous\ volume}\ [\%] \tag{6.2}$$

The JMAK model is based on three hypotheses:

1. the nucleation probability of the new crystalline phase is uniform in the volume and it depends on temperature;
2. nucleation and growth laws are known a priori;
3. the nuclei critical size, i.e. the size beyond which only diffusion processes come into play, is zero.

From these hypotheses, Avrami theory derives the CF [36] as

$$CF = 1 - e^{-V_e(t)} \tag{6.3}$$

where $V_e(t)$ is the extended volume, i.e. the volume of the crystalline phase which would have been formed if the GST had been still entirely amorphous. This way, the model does not account for the overlapping of two grains, whereas actually two grains stop growing when they impinge upon each other.

At a later stage, Avrami expresses $V_e(t)$ under isothermal conditions [37] as

$$V_e(t) = K_c\ t^n \tag{6.4}$$

where n is the Avrami exponent, which depends on nucleation and growth laws specific for the considered material, and K_c is called reaction constant and it depends on temperature (T), crystallization activation energy (E_a), and crystallization rate (K_0) following the Arrhenius-like equation [40]

$$K_c = K_0\ e^{-\frac{Ea}{k_B\ T}} \tag{6.5}$$

where k_B is Boltzmann constant.

Although the JMAK model is very simple and flexible, it can be critical when used to estimate the electrical resistance of the GST layer. In fact, the state of the material is described only macroscopically and the model does not account for the microstructure, which is of utmost importance in a SET operation, since the distribution of the crystal grains and their size determine the resistance of the material.

6.2.2 Amorphization kinetics

The amorphization kinetics was studied by applying the Foubert model [41], [42] to chalcogenide materials [43].

During amorphization, the formation of an amorphous cap in the GST layer leads to a decrease of the CF. This decrease can be expressed as

$$\frac{dCF}{dt} = -K_a \; CF \tag{6.6}$$

where K_a is the amorphization constant, which determines the speed of the amorphization process. From (6.6), CF has an exponential behaviour over time:

$$CF = e^{-K_a \; t} \tag{6.7}$$

It is thus possible to include both the amorphization and the crystallization phenomena in the same differential equation as follows

$$\frac{dCF}{dt} = K_c \; (1 - CF) - K_a \; CF \tag{6.8}$$

where K_c and K_a account for the crystallization and the amorphization process, respectively.

6.3 Model implementation

Since the implemented model is a dynamic system, the GST state can be simulated during all the whole programming operation.

The model can be conceptually broken into four main parts:

1. calculation of the temperature at the heater-GST interface;
2. calculation of the thickness of the amorphous cap;
3. calculation of the CF (during both crystallization and amorphization processes);
4. calculation of the GST resistance and, consequently, of the reading current.

While a simulation is running, all those four parts cooperate to derive the state of the cell during a programming operation.

6.3.1 Calculation of the temperature at the heater-GST interface

Since the modelled system is a dynamic one, the first item was implemented according to (6.1) and taking the temperature dynamics into account. The temperature dynamics can be modelled as a first-order Ordinary Differential Equation (ODE), so as to reproduce the transient of the cell temperature:

$$\frac{dT_{if}}{dt} = \frac{\Delta T_J - T_{if}}{\tau_T} \tag{6.9}$$

where τ_T is a time constant which models the temperature dynamics and T_{if} is the instantaneous temperature at the GST-heater interface. The thermal profile over time of the temperature at the GST-heater interface is obtained solving (6.9), thus obtaining

$$\Delta T_{if}(t) = \Delta T_J (1 - e^{-\frac{t}{\tau_T}}) \tag{6.10}$$

The instantaneous temperature at the GST-heater interface, T_{inst}, is found by adding ΔT_{if} to the operating temperature, $T_o p$

$$\Delta T_{inst} = \Delta T_{if} + T_o p \tag{6.11}$$

6.3.2 RESET modelling

The principle of the proposed RESET model is shown in Fig. 6.3. During a RESET operation, the GST is heated above its melting point. The chalcogenide starts melting, and its atoms begin to move into a chaotic configuration, typical of the amorphous state. The temperature T_{inst}, as it will be explained later in this Section, determines the maximum thickness of the amorphous cap. Since this thickness directly depends on temperature, it is also dependent on the programming pulse voltage.

The dynamics of the amorphization process is therefore affected by two contributions: the dynamics of the temperature at the GST-heater interface and the dynamics of the amorphous cap growth. In the proposed model, both these contributions are taken into account.

6.3.2.1 Growth of the amorphous cap during a RESET operation

As already explained, it is fundamental to know how thick the amorphous cap in the GST layer is during a RESET operation in order to be able to derive the state of the cell. Moreover, since the PCM cell has a thermal capacitance which delays the amorphization process, a dynamic model must account for both the maximum thickness of the amorphous cap ($z_{A,max}$) and the time the cap takes to reach this asymptotic value. The maximum thickness of the amorphous cap is expressed as a function of the temperature at the GST-heater interface [44] as

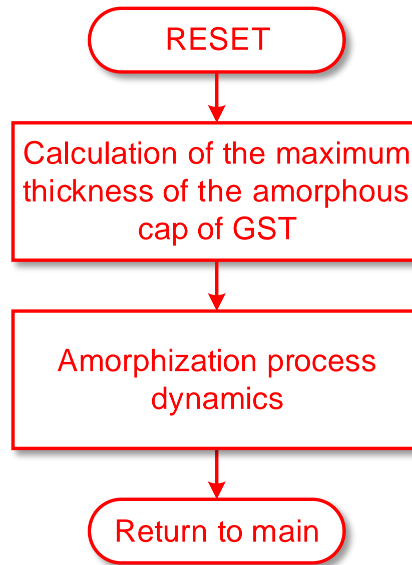


FIGURE 6.3: Flow-chart describing the basic idea of the proposed modelling for the RESET operation.

$$z_{A,max} = h\frac{T_{inst} - T_{melt}}{T_{inst} - T_{op}} \tag{6.12}$$

where h is the thickness of the GST layer and T_{melt} is the GST melting temperature ($\approx 600\ °C$).

The instantaneous thickness, z_A, of the amorphous cap depends on T_{inst}, $z_{A,max}$, and K_z, which is an amorphization constant determining the speed of the process. Even in this case, a first-order ODE can describe this dynamics:

$$\frac{dz_A(t)}{dt} = K_z\ (z_A - z_{A,max}) \tag{6.13}$$

From (6.13), z_A turns out to be

$$z_A = z_{A,max}\ (1 - e^{K_z\ t}) \tag{6.14}$$

Finally, the resistance of the GST after a RESET operation ($R_{GST,rst}$) and, hence, the reading current ($I_{GST,rst}$), can be derived:

$$R_{GST,rst} = \frac{\rho_A \ z_A}{S} + \frac{\rho_C \ (h - z_A)}{S} = \frac{\rho_C \ h}{S} + \frac{(\rho_A - \rho_C) \ z_A}{S} \tag{6.15}$$

$$I_{GST,rst} = \frac{V_{read}}{R_{GST,rst} + R_h} \tag{6.16}$$

where S is the area of the GST layer, ρ_C and ρ_A are the resistivities of the crystalline and the amorphous phases, respectively, and $V_{read} = 300$ mV is the reading voltage.

6.3.3 Analysis of the crystallization phenomenon

During a RESET operation, the grown amorphous cap is usually considered homogeneous. Actually, some crystal grains may remain buried in the cap. However, these grains do not appreciably affect the overall resistance of the GST layer after a RESET operation, since their resistivity is much less than that of the amorphous phase. From a macroscopic point of view, the residual crystal grains can therefore be neglected. Although this is true when a crystalline-to-amorphous transition occurs, during a SET operation, the residual crystalline grains in the amorphous cap play a decisive role. In fact, their number and their position relative to each other can lead to an early or late formation of a crystalline path in the amorphous cap during a SET operation. This leads to the formation of a parallel low-resistive path in the amorphous GST, thus determining an abrupt drop of the overall GST resistance and, consequently, a relevant increase of the reading current.

Before developing a model for the SET operation, it is thus important to study the repeatability of the crystallization process. To this end, SET pulses with different amplitude, from 1.4 V to 3.6 V ($\Delta V = 100$ mV step), and different time duration, from 50 ns to 500 ns ($\Delta t = 50$ ns step), were applied to a PCM cell in its full-RESET state. The devices used to perform the analysis is described in

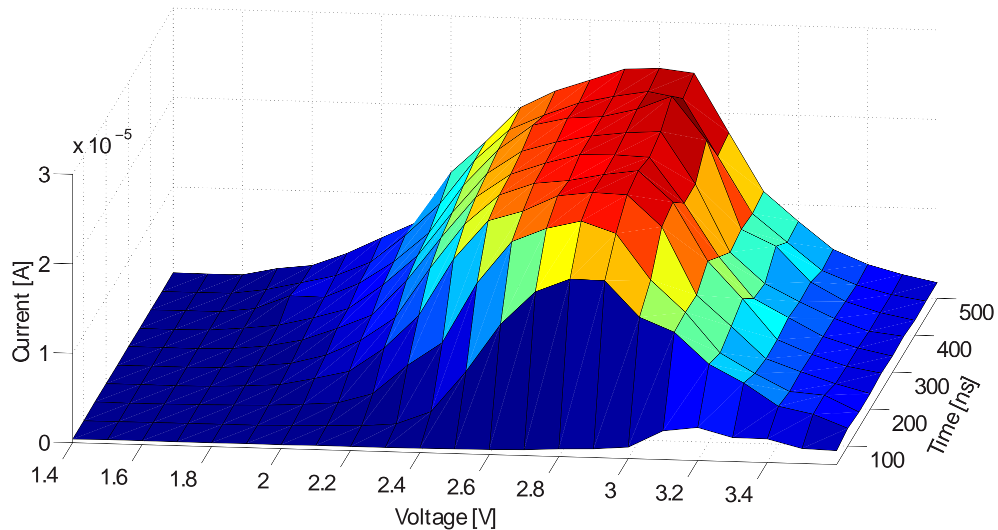


FIGURE 6.4: 3-D representation of the average reading current as a function of programming voltage and time duration.

[45]. The used programming algorithm is a SP partial-SET programming sequence with rectangular pulses (which means fast rise and fall times). At the beginning, a full-RESET pulses initialize the cell, then a partial-SET pulse is applied. Each partial-SET pulse was therefore applied after returning the cell to its full-RESET state, and the cell current was read after every applied pulse (readout after each RESET pulse allowed us to verify that the cell was returned to its initial condition). The reading operation is performed by applying a reading voltage V_{read} of about 300 mV to the cell and then sensing the corresponding current. Reading currents in the order of few hundreds of nA are associated to a full-RESET state, whereas reading currents greater than few μA are associated to a partial-SET state. Each measurement was repeated five times to investigate the repeatability of programming operation.

The five measured currents corresponding to the same programming pulse were averaged, and the resulting average current was plotted as a function of pulse amplitude and pulse duration, as shown in the 3-D plot of Fig. 6.4.

From Fig. 6.4, it can be observed that the starting point for the crystallization process depends on both applied voltage and time duration. In fact, first the amorphous GST has to reach a temperature greater than T_{cryst}, then it needs a certain amount of energy to be able to begin the crystallization process. For programming

pulses lower than 1.7 V, the inner temperature of the GST is lower than T_{cryst}, and hence the crystallization does not occur. Moreover, the energy delivered by short programming pulses, in the order of 50 ns, is demonstrated to be insufficient to start the process. The crystallization process, therefore, starts either in presence of pulses with moderate amplitude (V_{pulse} = 1.7 V) and long time duration (t_{pulse} = 500 ns) or in presence of pulses with higher amplitude (V_{pulse} = 2.3 V) and shorter time duration (t_{pulse} = 100 ns). It should be noted that the reading current increases with the increase of when increasing time duration and/or applied voltage, the latter effect being observed until an applied voltage of 2.9 V. When $V_{pulse} > 2.9$ V, a decrease of the reading current is observed, thus demonstrating the bi-directionality of the RESET operation. In fact, when the applied voltage exceeds the above value, the temperature at the heater-GST interface rises above T_{melt}, thus starting a partial-RESET operation, which follows equations (6.12) and (6.14), as it will be demonstrated in Section 6.4.1.

Fig. 6.5(a) shows that the dependence of the GST resistance on programming pulse amplitude is stronger when shorter pulses are applied ($t_{pulse} \leq 150$ ns), whereas for longer pulses the curves tend to get close to each other, thus demonstrating a strong dependence of the crystallization process on the temperature, according to (6.5).

The variation among the five measurements corresponding to the same programming conditions is shown in Fig. 6.5(b), where only four curves were selected for easier reading of the plot. In Fig. 6.5(b), the minimum and the maximum currents for the same programming pulse conditions are plotted together with the average current. It is apparent that the relative difference between measurements is larger when the pulse time duration is shorter. In fact, in the latter case, the programmed state of the cell is more sensitive to the initial microscopic structure. This is also demonstrated by Fig. 6.6, where the standard deviation σ_I (Fig. 6.6(a)) and the normalized deviation with respect to the average current, I_{avg}, (Fig. 6.6(b)) are shown. The standard deviation σ_I is higher for shorter pulses and for higher voltages. However, these data must be compared to the normalized deviation. In

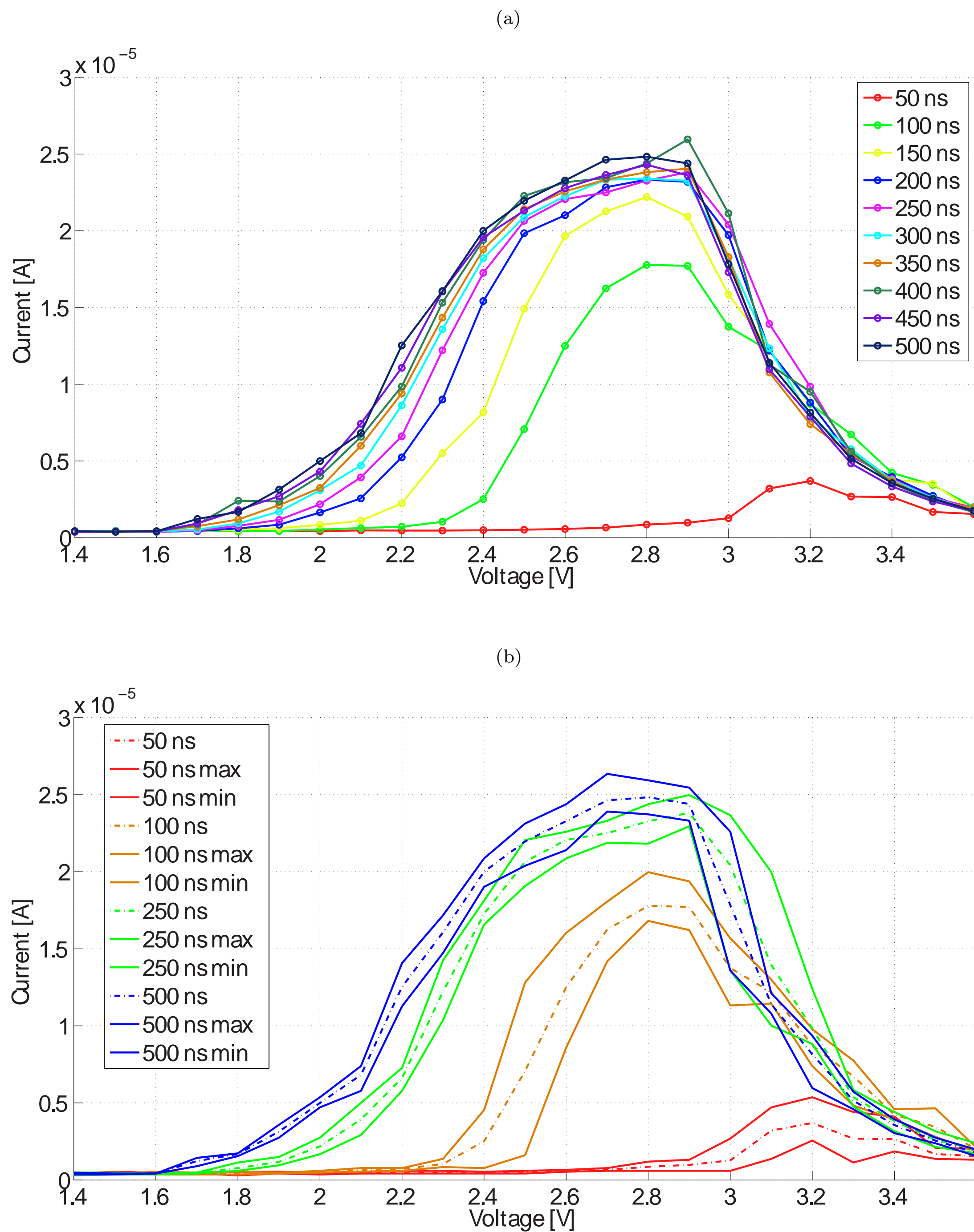


FIGURE 6.5: SET pulses with different amplitude and the same time duration: average of five reading current measurements (a) and average reading current compared to the highest and the lowest current measured for the same programming conditions (b).

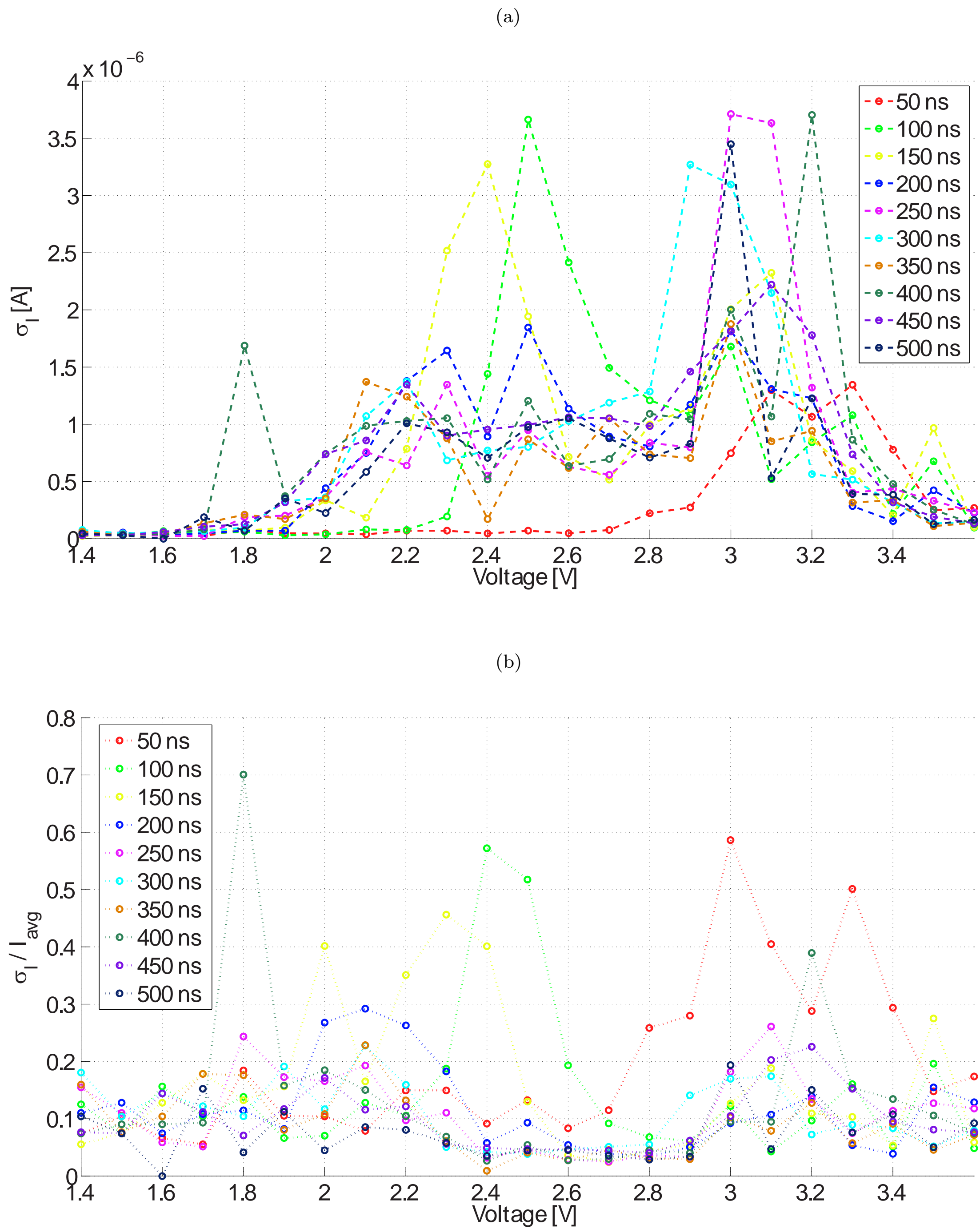


FIGURE 6.6: SET pulses with different amplitude and the same time duration: standard deviation (a) and maximum normalized error with respect to the average reading current, I_{avg} (b) for five measurements corresponding to the same programming conditions.

fact, although the absolute deviation σ_I increases for increasing pulse voltage amplitude, the relative deviation is mainly within 0.10 - 0.25 (which correspond to a percentage of 10% - 25%) for pulses time durations longer than 200 ns, thus demonstrating that, when time duration is long, the crystallization process is less affected by the microscopic initial condition.

The above observations find a further confirmation in Fig. 6.7(a), where where the measured reading curve is plotted as a function of the pulse time duration for different values of pulse amplitude. In this Figure, a strong dependence of the programmed cell state on the applied pulse amplitude is apparent. The major increase of reading current mainly occurs in the first 200 ns after the SET operation begins, then the reading current reaches an asymptotic level which strongly depends on the applied programming voltage. This confirms that a crystalline path in the amorphous cap is likely to have been formed in the first part of the crystallization process, thus heavily affecting the overall cell resistance and greatly increasing the reading current.

Fig. 6.7(b) shows the variation among the five measurements carried out with the same programming conditions. Only five curves are presented for easier reading of the plot. Again, it can be noted that the maximum spread among measurements occurs when short pulses with moderate amplitude are applied (e.g. the green curve at 150 ns), thus confirming the higher sensitivity of the cell to its microscopic structure in this range.

Finally, Fig. 6.8 also confirms that the repeatability of the SET operation tends to be better when programming with longer pulses whereas, for shorter pulses, variations in the final state of the cell up to 60% can occur. In fact, for longer pulses, σ_I is around 1 μA for any voltage amplitude, which corresponds to a percentage error below 15%. The 1.8 V curve is the only one which presents an abnormal trend at 400 ns but this can reasonably considered an outlier since one measurement among the five repeated turned out to be extremely far from the others, thus greatly affecting the results.

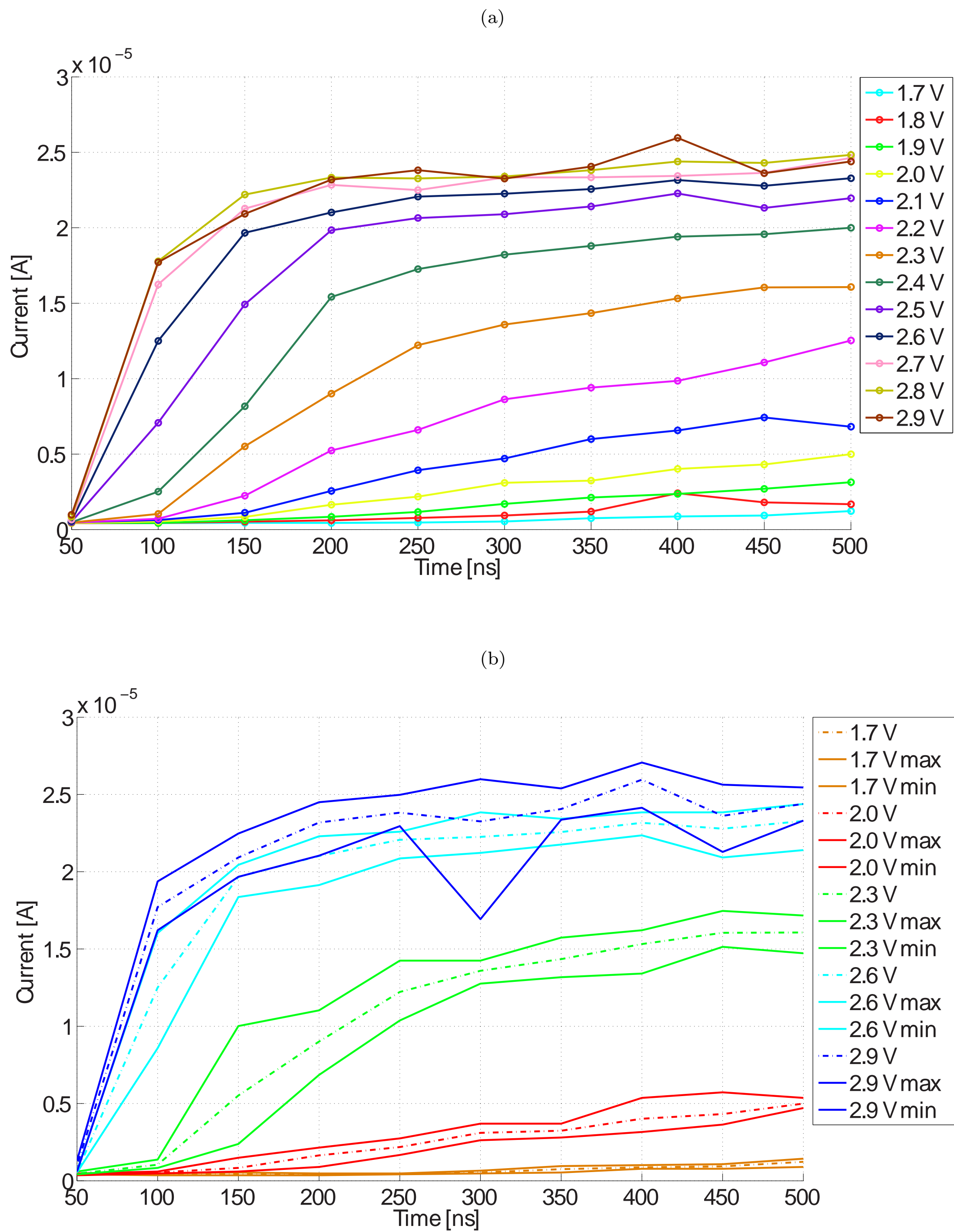


FIGURE 6.7: SET pulses with different time duration and the same amplitude: average of five reading measurements (a) and average reading current compared to the highest and the lowest current measured for the same programming conditions (b).

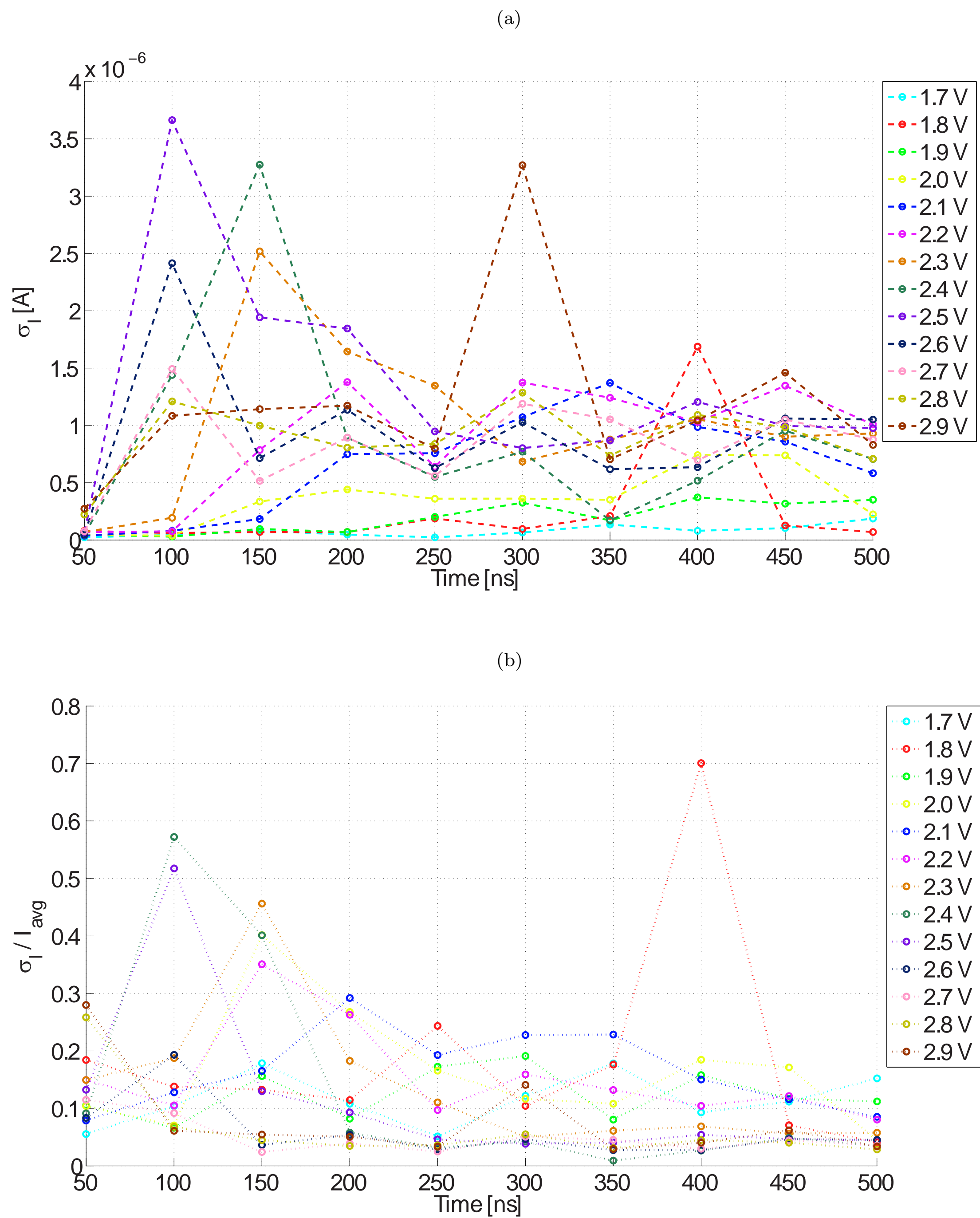


FIGURE 6.8: SET pulses with different time duration and the same amplitude: standard deviation (a) and maximum normalized deviation with respect to the average reading current, I_{avg} (b) for five measurements corresponding to the same programming conditions.

This study on the crystallization process led to the implementation of a model for the SET operation which aims at fitting the average behaviour of a PCM cell within the measured standard deviation $\pm\sigma_I$. In fact, unlike in the case of the RESET operation, which is repeatable [29] and does not need multiple measurements to be taken into account when validating the model, the observed dependence of the crystallization process on the GST microstructure led to a different approach when addressing the model of the SET operation.

6.3.4 SET modelling

The principle of the proposed SET model is shown in Fig. 6.9. Equation (6.8) was implemented to describe the evolution of the CF over time. It is worth noticing that K_a can be treated as a constant, since it has no dependence on time, whereas K_c is variable both over time and temperature (see equation (6.5)). More specifically, during a SET operation, K_a is set equal to 0, since it only has effect in the RESET operation to decrease the crystal fraction. CF can be thus expressed as

$$CF = 1 - e^{-K_0 \; e^{-(\frac{E_a}{k_B \; T}) \; t}} \tag{6.17}$$

Moreover, it was experimentally observed that also K_0 shows an exponential dependence on temperature

$$K_0 = 10^{\; a \; T_{inst}^2 + b \; T_{inst} + c} \tag{6.18}$$

where a, b, and c are three coefficients that can be obtained through a fitting operation on experimental data.

The crystallization of the amorphous GST in any part of its volume depends on the phase of the neighbouring material. The probability of phase change is greater in regions near crystalline grains rather than in a fully amorphous portion.

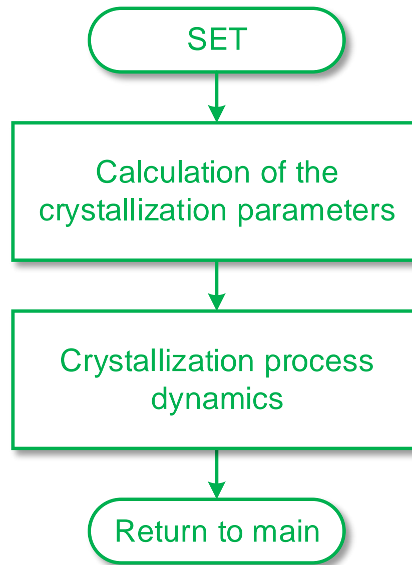


FIGURE 6.9: Flow-chart describing the basic idea of the modelling of the SET operation.

As discussed in Section 6.3.3, it can be assumed that a crystalline filament can be formed in the amorphous cap during the SET operation, which explains the abrupt resistance drop during a SET operation. In fact, the formation of a crystalline path in the amorphous region is electrically equivalent to connecting two resistors in parallel, one of which has a resistance much lower that the other. The final resistance of the cell thus decreases very abruptly.

In this case, it is convenient to calculate the GST conductance (and, hence, the reading current), rather than its resistance

$$G_{GST,set} = G_{SET}\ CF + G_{RST}\ (1 - CF) \tag{6.19}$$

where G_{SET} and G_{RST} are the conductances of the full-SET and the full-RESET states, respectively. These two conductances are equal to

$$G_{SET} = \sigma_C \frac{S}{h} \tag{6.20}$$

$$G_{RST} = \frac{\sigma_C \ \sigma_A \ S}{\sigma_A \ (h - z_A) + \sigma_C \ z_A} \tag{6.21}$$

where σ_C and σ_A are the conductances of the crystalline and the amorphous phases, respectively. As a consequence, the reading current $I_{GST,set}$ can be derived as

$$I_{GST,set} = \frac{V_{read}}{\frac{1}{G_{GST,set}} + R_h} \tag{6.22}$$

6.4 Model validation

In this Subsection, the model is validated through comparison with experimental data. Table 6.1 shows the parameters used and their values, which were obtained by means of a fitting operation with measurements on real cells, used also to validate the model. The memory device used, which includes an array of PCM μtrench cells, was fabricated in a 180 nm CMOS technology [45].

Before presenting the validation, an overall view of the model is given in Fig. 6.10

Once a pulse V_{oulse} is given to the cell, the temperature at the heater-GST interface is calculated, thus determining which operation is being performed (RESET, SET, or read). If a RESET operation is taking place, the CF decrease according to (6.8) and the maximum thickness of the amorphous cap and the dynamics of its growth are calculated through (6.12) and (6.13). Then, the value of the GST resistance, R_{GST}, is updated by using (6.15). If a SET operation is being performed, the CF grows with a dynamics given by (6.8), then the value of the GST conductance, G_{GST}, and consequently of R_{GST}, is updated by using (6.19).

After calculating R_{GST}, and when a reading operation is taking place, the cell reading current and, hence, the cell state, is easily calculated by

$$I_{read} = \frac{V_{read}}{R_{GST} + R_h} \tag{6.23}$$

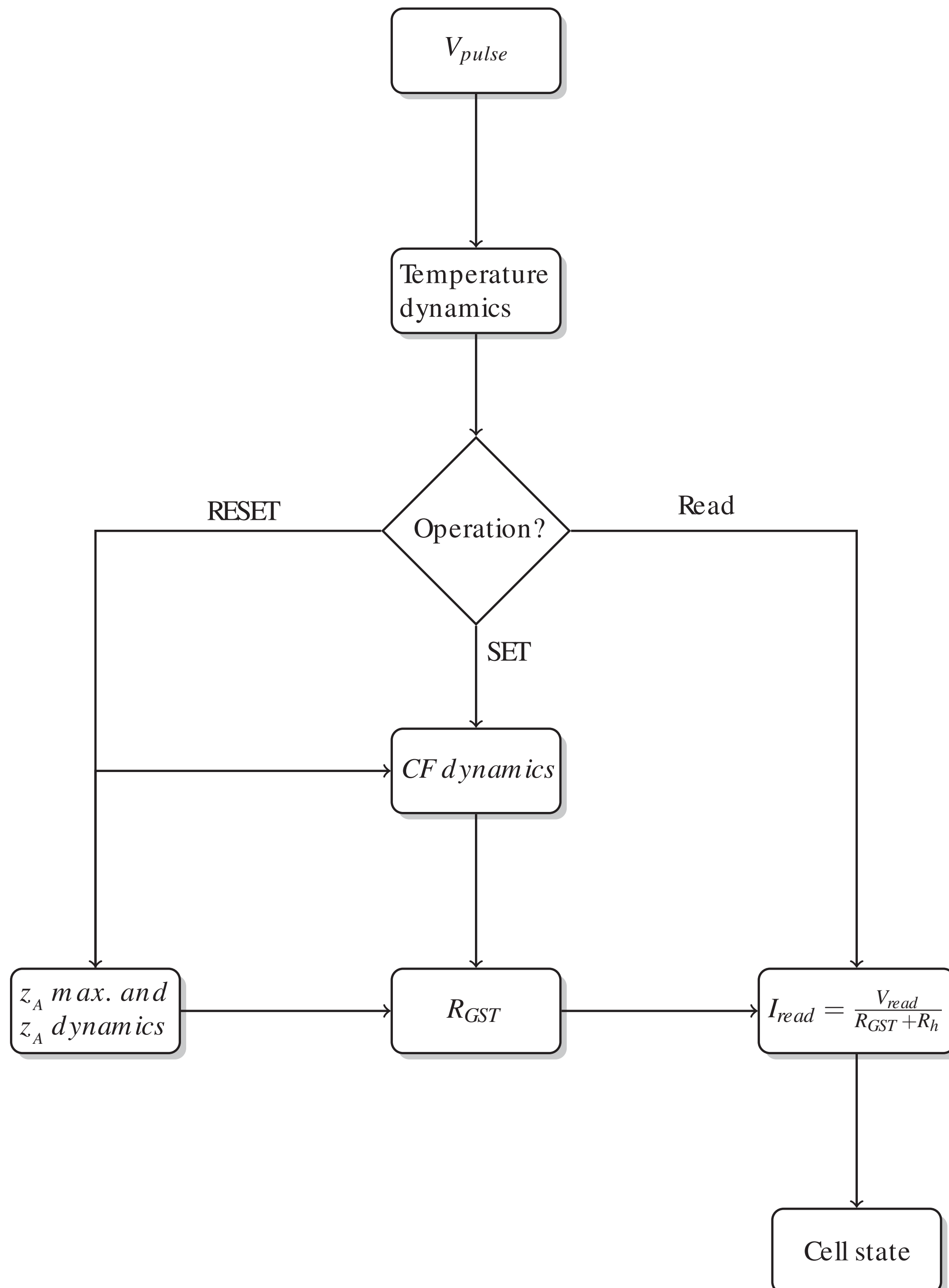


FIGURE 6.10: Overall model scheme.

TABLE 6.1: Values of the used parameters (obtained by means of a fitting operation on experimental data)

Parameter name	**Symbol**	**Value**
Activation energy	E_a	2.132 eV
Thickness of GST layer	h	80 nm
Area of GST layer	S	1600 nm^2
Operating temperature	T_op	20 ° C
Melting temperature	T_{melt}	600 ° C
Crystallization temperature	T_{cryst}	200 ° C
Constant for the temperature dynamics	τ_T	15 ns
Amorphization parameter	K_a	1.5 10^7 s^{-1}
Amorphization process speed	K_z	14.6 10^6 s^{-1}
Crystallization process speed	K_0	[s^{-1}]
	a	27 10^{-6} K^{-2}
	b	-54 10^{-3} K^{-1}
	c	47
Heater resistance	R_h	5 kΩ
Thermal resistance (crystalline GST)	$R_{th,c}$	320 $\frac{K}{mW}$
Thermal resistance (amorphous GST)	$R_{th,c}$	600 $\frac{K}{mW}$
Resistivity of crystalline GST	ρ_C	0.1 Ω mm
Resistivity of amorphous GST	ρ_A	2 Ω mm
Conductivity of crystalline GST	σ_C	10 $\frac{S}{mm}$
Conductivity of amorphous GST	σ_A	0.5 $\frac{S}{mm}$

6.4.1 RESET operation

The RESET operation was validated with a SP algorithm, which has been explained in Chapter 1. The used SP programming algorithm is repeated for convenience in Fig. 6.11.

It is expected that the RESET operation is very well repeatable, since the major contribution to the GST resistance is given by the amorphous cap. The model can

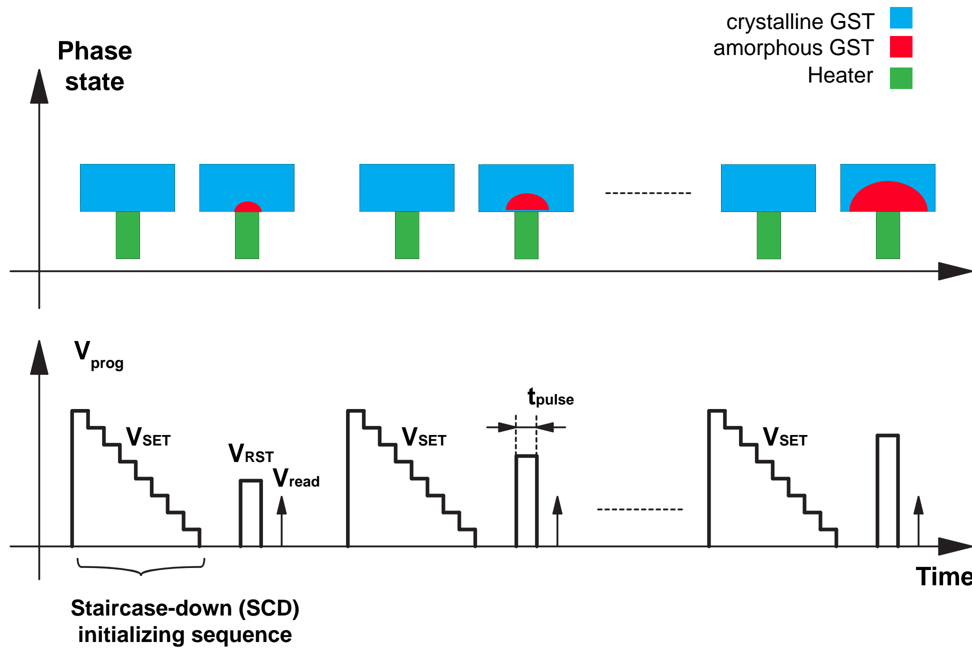


FIGURE 6.11: Sequence of program and read pulses of the single pulse RESET programming algorithm.

be thus validated by using a single set of measurements on a cell.

Pulses with different amplitudes, from 4 V to 5 V (ΔV = 250 mV step), and durations, from 50 ns to 400 ns (variable step), were applied to the memory cell, then the cell was simulated by means of the proposed model applying the same sequence of programming pulses. The comparison between model and measurements is shown in Figs. 6.12 and 6.13.

From Fig. 6.12, it is apparent that, for any value of pulse time duration, the higher the programming voltage, the lower the reading current (and, hence, the higher the GST resistance), which is in agreement with (6.12). From Fig. 6.13, it is observed that, for any value of pulse amplitude, the reading current reaches a stable value for long pulses, which demonstrates that the amorphization process dynamics follows equations (6.13) and (6.14).

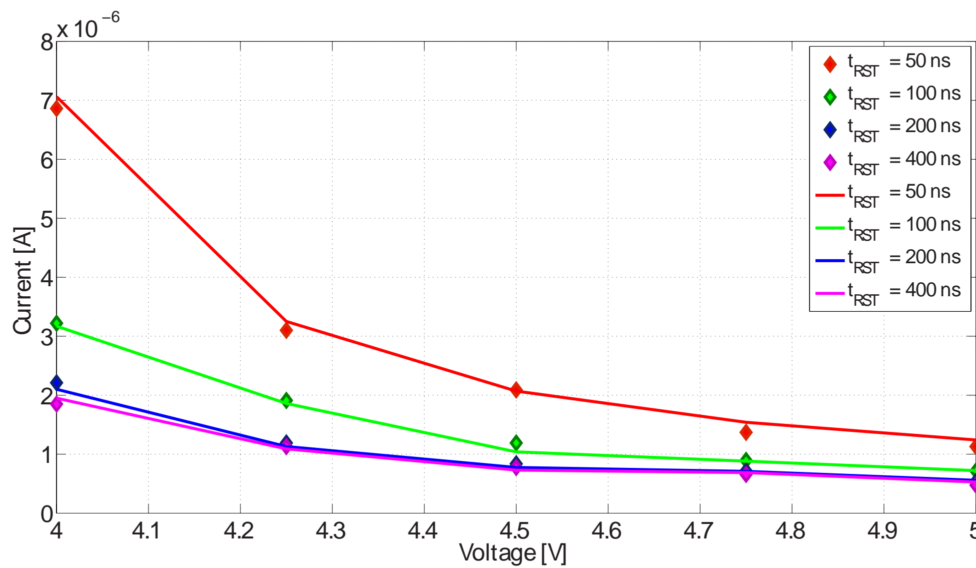


FIGURE 6.12: RESET dynamics: comparison between simulations (solid line) and measurements (dots) for RESET pulses with different time duration.

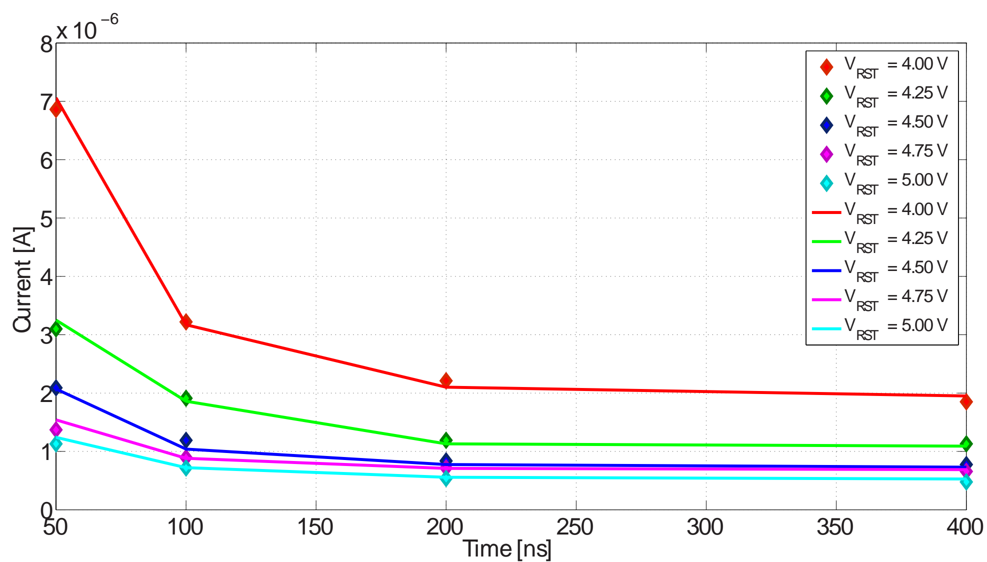


FIGURE 6.13: RESET dynamics: comparison between simulations (solid line) and measurements (dots) for RESET pulses with different amplitude.

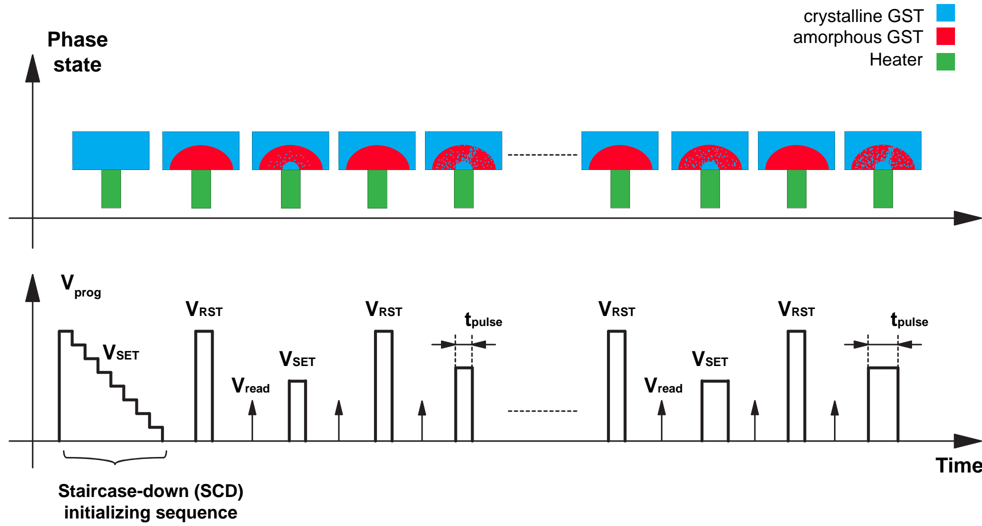


FIGURE 6.14: Sequence of program and read pulses for the partial-SET single pulse programming algorithm.

6.4.2 SET operation

As the RESET operation, also the SET operation was validated with an SP algorithm, which was explained in Chapter 1. Also in this case, the used programming algorithm is repeated here for convenience (Fig. 6.14).

Unlike the RESET operation, the SET operation is more sensitive to the microstructure of GST, as explained in Section 6.3.3. Therefore, in Figs. 6.15 and 6.16, the simulations with the proposed model are compared with five sets of measurements carried out on the same cell. In order to validate the model, error bars (length $\pm\sigma_I$) were added in both Figures.

Figs. 6.15 and 6.16 were obtained by applying pulses with different amplitude, from 1.7 V to 2.9 V (ΔV = 300 mV step), and different time duration, from 50 ns to 500 ns (Δt = 50 ns step).

From Fig. 6.15, it is clear that the crystallization process speed increases with increasing amplitude of the programming voltage and, hence, of the temperature at the heater-GST interface, as expressed in (6.5) and (6.18). The effect of the decrease of the GST resistance, which follows equation (6.19), is shown in Fig. 6.16.

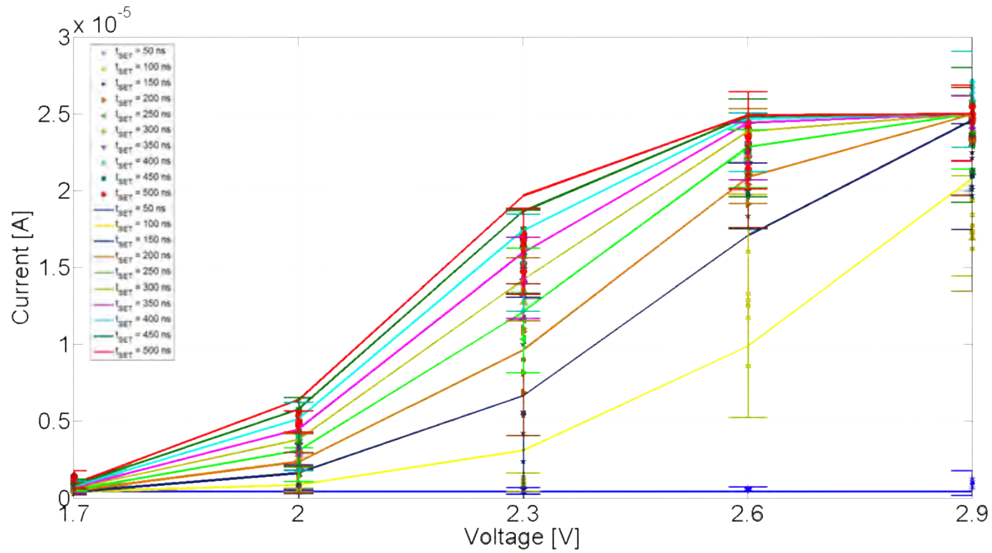


FIGURE 6.15: SET dynamics: comparison between simulations (solid line) and measurements (dots) for SET pulses with different time duration.

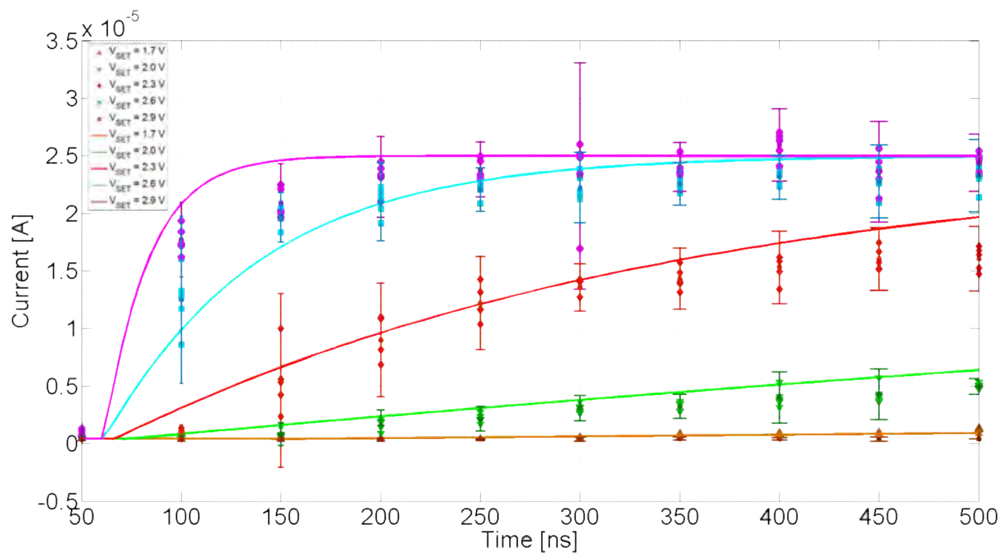


FIGURE 6.16: SET dynamics: comparison between simulations (solid line) and measurements (dots) for SET pulses with different amplitude.

Moreover, when the crystallization process starts (t = about 60 ns), if the temperature is sufficiently high (e.g. V_{SET} = 2.6 V or V_{SET} = 2.9 V), a more significant increase of the reading current in the first 200 ns is observed. This effect is ascribed to the dependence of the probability of formation of a crystalline path upon temperature.

If the spread in the SET operation, already discussed, is considered, the model results mostly to lie within the error bars in both Fig. 6.15 and 6.16. Therefore,

the model turns out to be successfully validated.

Conclusions

This Thesis presents an on-wafer analog pulse generator for fast parametric tests and characterization of Phase Change Memory (PCM) cells and a model to simulate the behaviour of PCM cells.

In Chapter 1 emerging non-volatile memories were introduced, focusing on PCMs. The usual commercial Automated Test Equipment (ATE) is also presented, as well as the need to design an accurate on-chip pulse generator able to exploit the conventional ATE and enhance its performance.

Chapter 2 shows three possible implementations for the on-chip pulse generator. All the three approaches are controlled by the ATE, so as to exploit and enhance its performance. The implementations are described, analysed and compared. The best one was chosen to be fabricated.

Two test-chips were fabricated. The first one is described in Chapter 3. Its aim is to debug and experimentally evaluate the main blocks of the system. This system allows generating pulses with different values of amplitude, duration, and fall time in order to meet the flexibility, controllability, and accuracy specifications required to massively characterize new-generation Non Volatile Memories, thus overcoming the limits of commercial ATEs. A manual calibration procedure able to limit the inaccuracies of the on-chip pulse generator due to fabrication process spreads and non-idealities as well as to the uncertainties of test equipment was also conceived and evaluated.

In Chapter 4 a second version of the pulse generator is presented. This version focuses on the interfacing of the on-chip pulse generator with the ATE, which can be very challenging especially when short current pulses have to be read. Thanks to this system, a complete read-and-write cycle can be executed in few tens of milliseconds by limiting the use of the switch matrix to set cell selection at the beginning of a test sequence. The previously presented calibration procedure was automatised, included in this test chip, and evaluated. A simplified version of the test-chip developed to be used with different test equipment was also presented.

Chapter 5 showed the experimental results of all test chips. The designed systems are able to generate pulses with an amplitude from 0.5 V up to 4.5 V, a pulse duration from 50 ns to 350 ns, and a fall time from 10 ns to several μs (the variability of the last parameter is essential in order to analyse the response of PCM cells to different quenching times). By using the proposed calibration procedure, it is possible to obtain an accuracy better than $\pm 10\%$ in all the parameters that define the shape of the generated programming pulse.

Finally, a model to simulate the behaviour of PCM cells under different programming conditions was presented in Chapter 6. The amorphization and crystallization kinetics were described and the derived model was validated through comparison with experimental data. To this end, a memory device fabricated in a 180 nm CMOS technology featuring an array of μtrench PCM cells was used.

Bibliography

[1] Peter Mell and Timothy Grance. The NIST definition of cloud computing. National Institute of Standards and Technology (NIST), Sep. 2011. Special Publication 800-145.

[2] R. Bez, P. Cappelletti, G. Servalli, and A. Pirovano. Phase change memories have taken the field. In *Proceedings of IEEE International Memory Workshop (IMW)*, pages 13–16, 2013.

[3] Roberto Bez and Paolo Cappelletti. Emerging memory technology perspective. In *IEEE International Symposium on VLSI Design, Automation, and Test (VLSI-DAT)*, pages 1–2, 2012.

[4] Stefan K. Lai. Floating gate memories: Moore's law continues. In *IEEE International Symposium on VLSI Technology (VLSI-TSA)*, pages 74–77, 2005.

[5] F. Wang and X. Wu. Non-volatile memory devices based on chalcogenide materials. In *Proceedings of International Conference on Information Technology: New Generations*, pages 5–9, 2009.

[6] U. Russo, D. Kamalanathan, D. Ielmini, A. L. Lacaita, and M. N. Kozicki. Study of multilevel programming in programmable metallization cell (PMC) memory. *IEEE Transactions on Electron Devices*, 56(5):1040–1047, May 2009.

[7] Michael Kund, Gerhard Beitel, Cay-Uwe Pinnow, Thomas Rhr, Jrg Schumann, Ralf Symanczyk, Klaus-Dieter Ufert, and Gerhard Mller. Conductive bridging RAM (CBRAM): An emerging non-volatile memory technology scalable to sub 20nm. In *IEEE Technical Digest of International Electron Devices Meeting (IEDM)*, pages 754–757, 2005.

[8] D. Lee, D.-J. Seong, I. Jo, F. Xiang, R. Dong, S. Oh, and H. Hwang. Resistance switching of copper doped moox films for nonvolatile memory applications. *Applied Physics Letters*, 90(12):2237–2251, Mar. 2007.

[9] H. Akinaga and H. Shima. Resistive random access memory (ReRAM) based on metal oxides. *Proceedings of the IEEE*, 98(12):2237–2251, Dec. 2010.

[10] Ilia Valov and Michael N. Kozicki. Cation-based resistance change memory. *Journal of Physics D: Applied Physics*, 46(7):doi:10.1088/0022–3727/46/7/074005, Feb. 2013.

[11] Yiming Huai, Yuchen Zhou, I. Tudosa, R. Malmhall, R. Ranjan, and Jing Zhang. Progress and outlook for STT-MRAM. In *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, page 235, 2011.

[12] R. Buhrman. Spin torque MRAM challenges and prospects. In *Device Research Conference (DRC)*, page 33, 2009.

[13] SangBum Kim and Chung H. Lam. Transition of memory technologies. In *IEEE International Symposium on VLSI Technology (VLSI-TSA)*, pages 1–3, 2012.

[14] Fujistu Laboratories and University of Toronto. Fujitsu and university of toronto develop high-reliability read-method for spin-torque-transfer MRAM, next-generation non-volatile memory. http://www.fujitsu.com/global/news/pr/archives/month/2010/20100210-03.html, Feb. 2010.

[15] K. Byeungchul, S. Yoonjong, A. Sujin, K. Younseon, J. Hoon, A. Dongho, N. Seokwoo, J. Gitae, and C. Chilhee. Current status and future prospect of phase change memory. In *Proceedings of IEEE International Conference on ASIC (ASICON)*, pages 279–282, 2011.

[16] A. L. Lacaita, A. Redaelli, D. Ielmini, F. Pellizzer, A. Pirovano, A. Benvenuti, and R. Bez. Electrothermal and phase-change dynamics in chalcogenide-based

memories. In *IEEE International Electron Device Meeting (IEDM) Technical Digest*, pages 911–914, 2004.

[17] A. Redaelli, A. Pirovano, F. Pellizzer, A. L. Lacaita, D. Ielmini, and R. Bez. Electronic switching effect and phase-change transition in chalcogenide materials. *IEEE Electron Device Letters*, 2004.

[18] C. Peng, L. Cheng, and M. Mansuripur. Experimental and theoretical investigations of laser-induced crystallization and amorphization in phase-change optical recording media. *Journal of Applied Physics*, 82(9):4183–4191, 1997.

[19] V. Weidenhof, N. Pirch, I. Friedrich, S. Ziegler, and M. Wuttig. Minimum time for laser induced amorphization of $ge_2sb_2te_5$ films. *Journal of Applied Physics*, 88(2):657–664, 2000.

[20] Matthias Wuttig et al. The role of vacancies and local distortions in the design of new phase-change materials. *Nature Materials*, 6:122–128, Dec. 2006.

[21] X.Q. Wei, L.P. Shi, W. Rajan, R. Zhao, B.S. Quek, X.S. Miao, and T.C. Chong. Hspice macromodel of pcram for binary and multilevel storage. *IEEE Transactions on Electronic Devices*, 53(1):56–62, Jan. 2006.

[22] S.R. Ovshinsky. Reversible electrical switching phenomena in disordered structures. *Physical Review Letters*, 21(20):1450–1453, 1968.

[23] C. B. Thomas, B. D. Rogers, and A. H. Lettington. Monostable switching in amorphous chalcogenide semiconductors. *Journal of Applied Physics*, 9(18): 2571–2586, 1976.

[24] D. Adler, M. S. Shur, M. Silver, and S. R. Ovshinsky. Threshold switchingin chalcogenide-glass thin films. *Journal of Applied Physics*, 59(6):3289–3309, 1980.

[25] A. Pirovano, A. L. Lacaita, A. Benvenuti, F. Pellizzer, and R. Bez. Electronic switching in phase-change memories. *IEEE Transactions on Electronic Devices*, 51(3):452–459, Mar. 2004.

[26] Stefania Braga, Alessandro Cabrini, and Guido Torelli. Voltage-driven multilevel programming in phase change memories. In *Proceedings of IEEE International Workshop on Memory Technology, Design, and Testing (MTDT)*, pages 3–6, 2009.

[27] F. Bedeschi, R. Fackenthal, C. Resta, E. M. Donze', M. Jagasivamani, E. C. Buda, F. Pellizzer, D. W. Chow, A. Cabrini, G. M. A. Calvi, R. Faravelli, A. Fantini, G. Torelli, D. Mills, R. Gastaldi, and G. Casagrande. A bipolar-selected phase change memory featuring multi-level cell storage. *IEEE Journal of Solid-State Circuits*, 44(1):217–227, Jan. 2009.

[28] C. Villa, D. Vimercati, S. Schippers, S. Polizzi, A. Scavuzzo, M. Perroni, M. Gaibotti, and M.L. Sali. A 65 nm 1 Gb 2b/cell NOR flash with 2.25 MB/s program throughput and 400 MB/s DDR interface. *IEEE Journal of Solid-State Circuits*, 43(1):132–140, Jan. 2008.

[29] Stefania Braga. *Characterization and Modeling of Phase Change Memories*. PhD thesis, Ph.D. Thesis in Microelectronics, University of Pavia, 2010.

[30] G. De Sandre, L. Bettini, E. Calvetti, G. Giacomi, M. Pasotti, M. Borghi, P. Zuliani, R. Annunziata, I. Tortorelli, F. Pellizzer, and R. Bez. Program circuit for a phase change memory array with 2 MB/s write throughput for embedded applications. In *Proceedings of European Solid-State Circuits Conference (ESSCIRC)*, pages 198–201, 2008.

[31] J.-T. Lin, Y.-B. Liao, M.-H. Chiang, and W.-C. Hsu. Operation of multi-level phase change memory using various programming techniques. In *Proceedings of IEEE International Conference on IC Design and Technology (ICICDT)*, pages 199–202, 2009.

[32] C. Ahn, B. Lee, R. G. D. Jeyasingh, M. Asheghi, G.A.M. Hurkx, K. E. Goodson, and H.-S. Philip Wong. Crystallization properties and their drift dependence in phase-change memory studied with a micro-thermal stage. *Journal of Applied Physics*, 110(11):114520 – 114520–6, 2011.

[33] T. Nirschl, J. B. Philipp, T D. Happ, G. W Burr, B. Rajendran, M.-H. Lee, A. Schrottt, M. Yang, M. Breitwisch, C.-F. Chen, E. Joseph, M. Lamorey, R. Cheek, S.-H. Chen, S. Zaidi, S. Raoux, Y. C. Chen, YZhu, R. Bergmann, H.-L. Lunge, and C. Lam. Write strategies for 2- and 4-bit multi-level phase change memory. In *Proceedings of IEEE International Electron Devices Meeting (IEDM)*, pages 461–464, 2007.

[34] Daniele Devecchi and Guido Torelli. MOS stage with high output resistance particularly for integrated circuits. SG-Thomson Microelectronics srl, Milan, Italy, Aug. 1990. 4952885.

[35] John B. Hughes, Neil C. Bird, and Ian C. Macbeth. Switched currents - a new technique for analog sampled-data signal processing. In *Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1584–1587, 1989.

[36] Melvin Avrami. Kinetics of phase change. I general theory. *Journal of Chemical Physics*, 7(12):doi:10.1063/1.1750380, Dec. 1939.

[37] Melvin Avrami. Kinetics of phase change. II transformationtime relations for random distribution of nuclei. *Journal of Chemical Physics*, 8(2): doi:10.1063/1.1750631, Feb. 1940.

[38] Melvin Avrami. Granulation, phase change, and microstructure kinetics of phase change. III. *Journal of Chemical Physics*, 9(2):doi:10.1063/1.1750872, Feb. 1941.

[39] F. Bedeschi, R. Bez, C. Boffino, E. Bonizzoni, E. C. Buda, G. Casagrande, L. Costa, M. Ferraro, R. Gastaldi, O. Khouri, F. Ottogalli, F. Pellizzer, A. Pirovano, C. Resta, G. Torelli, and M. Tosi. 4-Mb MOSFET-Selected μtrench phase-change memory experimental chip. *IEEE Journal of Solid-State Circuits*, 40(7):1557–1565, Jul. 2005.

[40] N. Mehta and A. Kumar. Observation of phase separation in some seteag chalcogenide glasses. *Materials Chemistry and Physics*, 96(1):73–78, Mar. 2006.

[41] Imogen Foubert, Peter A. Vanrolleghem, Bert Vanhoutte, and Koen Dewettinck. Dynamic mathematical model of the crystallization kinetics of fats. *Food Research International*, 35(10):945–956, 2002.

[42] Imogen Foubert, Peter A. Vanrolleghem, Bert Vanhoutte, and Koen Dewettinck. Modelling of the crystallization kinetics of fats. *Trends in Food Science and Technology*, 14(3):79–92, Mar. 2003.

[43] Stefania Braga, Alessandro Cabrini, and Guido Torelli. An integrated multiphysics approach to the modeling of a phase-change memory device. In *Proceedings of IEEE European Solid-State Device Research Conference (ESSDERC)*, pages 154–157, 2008.

[44] Stefania Braga, Alessandro Cabrini, and Guido Torelli. Theoretical analysis of the reset operation in phase-change memories. *Semiconductor Science and Technology*, 24(11):doi:10.1088/0268–1242/24/11/115008, Nov. 2009.

[45] F. Pellizzer, A. Pirovano, F. Ottogalli, M. Magistretti, M. Scaravaggi, P. Zuliani, M. Tosi, A. Benvenuti, P. Besana, S. Cadeo, T. Marangon, R. Morandi, R. Piva, A. Spandre, R. Zonca, A. Modelli, E. Varesi, T. Lowrey, A. Lacaita, G. Casagrande, P. Cappelletti, and R. Bez. Novel μtrench phase-change memory cell for embedded and stand-alone non-volatile memory applications. In *Digest of Technical Papers of IEEE Symposium on VLSI Technology*, pages 18–19, 2004.