Technische Universität München
Fakultät für Elektrotechnik und Informationstechnik
Lehrstuhl für Schaltungsentwurf

Concepts for Multi-Level Cell Operation of Embedded Flash in Automotive Applications

Sebastian Kiesel

Vollständiger Abdruck der von der Fakultät für Elektrotechnik und Informationstechnik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs

genehmigten Dissertation.

Vorsitzender: Prof. Dr.-Ing. Ralph Brederlow

Prüfer der Dissertation:

1. apl. Prof. Dr.-Ing. habil. Helmut Gräb
2. Prof. Dr.-Ing. Bernhard Wicht

Die Dissertation wurde am 25.09.2019 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 10.02.20 angenommen.

Abstract

The advance in automotive technology results a large scope of applications for microcontroller units (MCUs) with growing complexity. Trends like autonomous driving and sensor fusion are pushing the demand for high performance controllers with increased embedded memory capacity, for code and data storage. The storage of multiple bits in one memory cell (multi-level cell (MLC) operation) is a proven method to increase memory density and hence reduce cost for a required amount of memory. Due to high requirements regarding temperature range, long operating lifetimes and absolute code integrity, MLC operation has not been applied for embedded Flash (eFlash) memories in automotive applications, so far.

This work evaluates different Flash cell biasing schemes for the usage in a sensing scheme for MLC eFlash targeting automotive requirements. Besides the implementation complexity and area consumption, sensing robustness and speed are analyzed. For this purpose, a time-domain voltage sensing principle is chosen, and the analytically derived transfer characteristics of the investigated biasing approaches are taken for assessment.

Based on the findings of the investigation, a novel time-domain sensing scheme using a ramped gate cell biasing is introduced. It comprises an improved voltage mode sense amplifier design with offset compensation capability for better distinction of multiple programmed cell states at high speed operation. The components of the sensing scheme's implementation, including sense amplifier, ramp generation and strobing circuitry are presented and analyzed. Measurements conducted on a test chip verify the functionality and reveal the overall read performance of the sensing scheme, which is compared to state-of-the-art memory designs.

Further, the biasing scheme's implications on the programming performance is investigated and disturbing effects like cell-to-cell interference and temperature shifts are highlighted. According to the outcome of the conducted measurements, guidelines for development of an efficient programming algorithm are given.

The present work demonstrates that modern eFlash technology is capable for robust MLC operation that meets stringent automotive requirements. The advance towards multiple bits per cell is a viable option for memory density increase and cost reduction in eFlash for automotive applications.

Contents

Chapter 1

Introduction

1.1 Evolution of Multi-Level Flash Memory

In the mid-1980s, a new type of non-volatile memory evolved from the predominant erasable programmable read only memory (EPROM). In search of a relief from the lengthy erase procedure, Masuoka et al. from Toshiba introduced the first modern electrically erasable programmable read only memory (EEPROM) in 1984 [1]. The memory array was erased within one operation rather than byte per byte, as it was common for conventional EEPROMs. It was given the name "Flash", as it could be erased within less than a second; fully electrical and without need for ultraviolet (UV) light. With this evident advantage, Flash memory quickly displaced the EPROM in embedded applications, requiring economical reprogramming. New features like independently erasable blocks and an embedded high voltage supply enabled Flash to become more than an EPROM replacement and it quickly entered other markets. In the 1990s, digital cameras, portable MP3 players and cell phones were the drivers for the rapidly growing Flash market. Today, Flash has become a multi-billion dollar business.

With different requirements, two major Flash memory technologies evolved. The NOR-type Flash was the first one in the market. It provides fast random read access and good reliability. Consequently, it is well suited as code memory for direct code execution (Execute-in-place). The NAND-type Flash, which was introduced later, has its benefits in fast write operation and high-density storage. Though it has a slow random access time, its sequential read bandwidth and programming throughput is very good, which makes it the ideal candidate for data storage.

To reduce cost per bit, the memory transistor was scaled down aggressively. In the early 1990s, engineers put a lot of effort in the development of technologies for storage of multiple bits per memory cell, to accelerate the cost reduction even

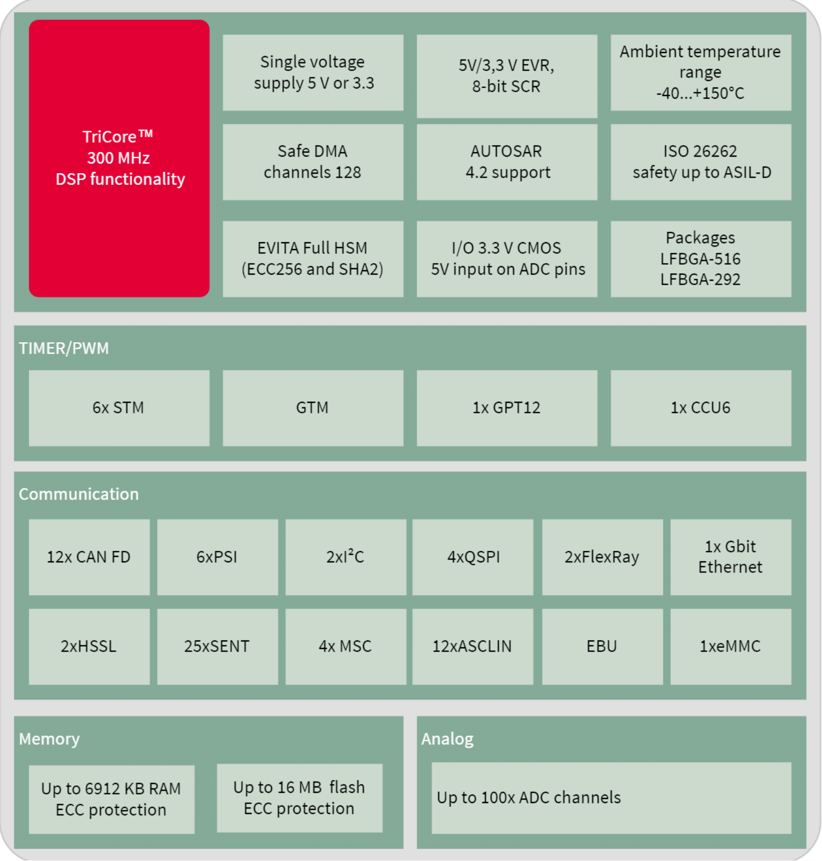


Figure 1.1: Feature set of the AURIX$^{\text{TM}}$ TC39x automotive microcontroller. Source: Infineon Technologies AG

more. In 1995 the breakthrough was accomplished, and Intel presented their first MLC product with 32 Mbit stored as 2-bit/cell [2]. From then on, MLC technology evolved rapidly and many features and improvements were added. For NOR-type Flash, sensing speed was increased by introduction of new sensing schemes, that used parallel comparison against multiple references [3] or a serial sensing, selecting the reference in binary search fashion [4, 5]. Further, biasing schemes with variable voltage were developed [6, 7], to improve the sensing robustness at continuously shrinking Flash cell size. Advances in NAND-type Flash were achieved in the increase of memory density and programming throughput. The All Bit Line architecture, introduced by Cernea et al. [8] boosted the programming speed significantly by raising the parallelism of the verify read. Together with several improvements including novel programming algorithms [9, 10], this enabled the development of NAND Flash memories with triple-level cell (TLC)- (three bits per cell) and even quad-level cell (QLC)-operation (four bits per cell).

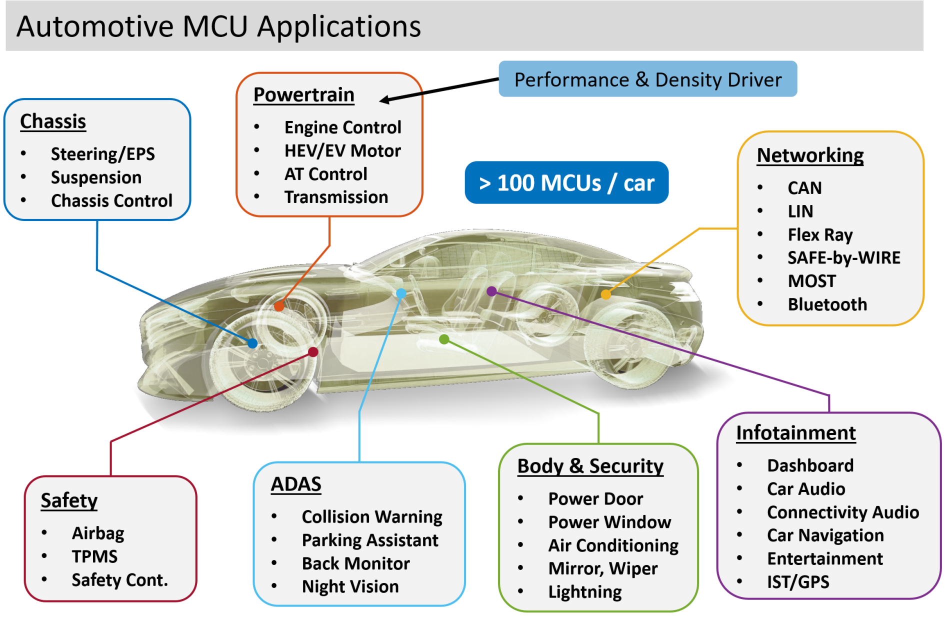


Figure 1.2: Application scope of automotive microcontrollers in a modern car. Image source: Pixabay.com

Microcontrollers have gained in importance, as their range of applications is continuously growing. Industrial, Internet of Things (IoT) and automotive are important market segments, that gather pace as society is moving forward into a digital future with autonomously driving vehicles. Besides SRAM, eFlash is the predominant memory used in MCUs. As relevant performance factors for eFlash are fast random access and absolute code integrity, mainly NOR-type Flash is used [11]. Automotive microcontrollers have the highest requirements regarding operation conditions. Very low failure rates must be guaranteed over lifetimes of 15-20 years and operation at junction temperatures up to 175 °C must be supported. An example for a state-of-the-art microcontroller for automotive application is the Infineon AURIX$^{\text{TM}}$ TC39x. Figure 1.1 shows the feature set of the MCU. It is fabricated in 40 nm technology and includes six cores, operated with 300 MHz. The embedded memory comprises up to 6912 kB of SRAM and 16 MB of Flash. It is equipped with high performance, numerous analog to digital converter channels and various communication protocols to address in particular automotive requirements.

One of the factors driving the market for eFlash of automotive MCUs is their growing scope of applications (see Fig. 1.2). In a modern vehicle today, over 100 MCUs are operated within its control units for powertrain, chassis, body, airbag etc. [12]. The rising requirement for more personalization, comfort and safety, but also the trend towards autonomous driving, including Advanced Driver Assistance Systems (ADAS) and sensor fusion will increase this amount and it will push the application complexity [13]. The main drivers for performance and memory density are powertrain and transmission. Current high-end microcontrollers for this application scope comprise flash memories of 8 MB to 16 MB. To meet the needs, upcoming generations of automotive microcontrollers have to be equipped with larger memories. Embedded Flash memories with a density up to 32 MB are already in development, to be integrated into future MCUs [14] and the requirements will continue to rise.

As in standalone NOR and NAND Flash memories, the cost position of high density eFlash for automotive application could also be drastically improved with multi-level cell operation. However, the high requirements for read performance and data integrity have prevented the usage of MLC technology for automotive MCUs so far. Consequently, research for reliable memory technologies and sensing schemes with increased robustness are required to bring MLC operation into high-end MCUs for automotive applications.

1.2 Contribution of This Work

The scope of this work is the development and analysis of concepts and circuits for the MLC operation of eFlash in automotive applications. Thereby, the focus is put on a robust sensing scheme that is able to meet the speed and reliability requirements of automotive applications. The main contributions of this work can be summarized as follows:

- A new time-domain voltage sensing method for MLC operation is developed, that uses dynamic voltage ramps at the Flash cells' control gates.
- The time-domain transfer characteristics of ramped gate biasing and fixed gate biasing is derived analytically. Both characteristics are compared regarding their suitability for MLC operation.
- An enhanced voltage mode sense amplifier design is presented, which extends a state-of-the-art design with an offset compensation and precharge acceleration mechanisms. Its improved sensing accuracy is shown in simulation.
- A ramp generator design with fast recovery from discharge is presented. It implements the ramped gate biasing.
- A time-domain strobing concept is presented, which allows to analyze the sensing scheme with a time scan method.

- The analytically derived transfer characteristic of the ramped gate biasing is verified by measurement.
- Experimental results demonstrate the functionality and performance of the sensing scheme, including the sense amplifier enhancements.
- The impact of the cell's control gate biasing on the programming is analyzed. The benefits of the ramped gate biasing are shown with measurement.
- Guidelines for the design of an efficient programming algorithm are developed.

1.3 Outline of the Thesis

The thesis is organized as follows: Chapter 2 introduces the basic principles for storage and readout of multiple bits per Flash memory cell. Different MLC cell types are presented and techniques for sensing and programming are explained.

Chapter 3 gives an overview of the state-of-the art cell biasing schemes and compares them with respect to complexity, area and their suitability for high speed MLC operation. To assess the sensing robustness of the biasing approaches, their time-domain transfer characteristic is derived analytically.

Chapter 4 describes the MLC sensing scheme developed in this work. The sense amplifier design using a current-biased slope detection is introduced and its improvements over the state-of-the-art voltage-biased design are analyzed. Mechanisms for acceleration of the sense amplifier's precharge process are presented and its high speed capability is shown. Further, the design of the deployed ramp generator for implementation of the ramped gate cell biasing is explained. The strobing concept of the sensing scheme is introduced and the analysis features implemented for investigation of the sensing performance are explained. Finally, the sensing architecture of the implemented test chip is discussed and measurement results are shown.

Chapter 5 discusses important aspects for the development of a precise programming algorithm for placement of multiple cell states. The impact of the used memory cell biasing on the programming performance is analyzed by means of experiments conducted on the fabricated test chip.

Chapter 6 summarizes the thesis' results and gives a short outlook on embedded memory technologies for future applications.

Chapter 2

Fundamentals of Multi-Level Cell Flash Memories

2.1 Cell Technologies for Multibit Storage

The floating gate field effect transistor was proposed by Kahng and Sze in 1967 [15]. It established the basis for all modern flash memories. The underlying principle is the storage of charge on a fully insulated electrode, located in between the channel and the normal gate electrode of a metal-oxide-semiconductor field effect transistor (MOSFET). The intermediate electrode, called floating gate (FG), usually is made of polysilicon. Since it is completely encapsulated by isolating layers, it can retain the stored charge for a very long period, which allows to store data, without need of constant power (non-volatility). The upper gate electrode is called control gate (CG). The voltage applied at this node shifts the potential of the FG by capacitive coupling and therefore controls the forming of a conducting channel in between the transistors drain and source contacts. Figure 2.1 shows the cross section of a floating gate transistor. It has the same characteristics as a

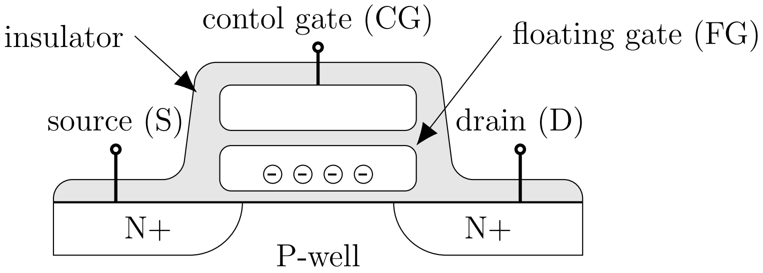


Figure 2.1: Cross section of a floating gate transistor. Electrons are stored on the fully isolated FG.

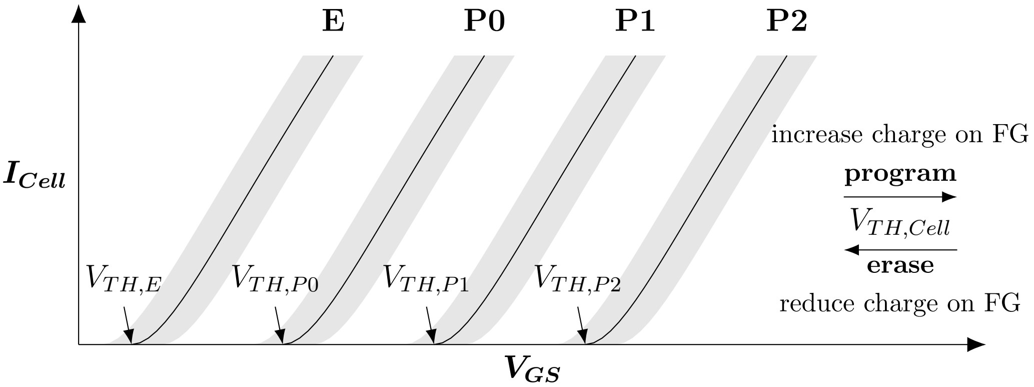


Figure 2.2: Current-voltage characteristics of four distinct cell states (2 bit per cell). Threshold variations due to inaccurate programming are highlighted in gray.

common MOS transistor, except for its variable threshold voltage ($V_{TH,Cell}$), which is modulated by the amount of charge that is stored on the floating gate. The threshold voltage change effectively results in a parallel shift of the transistor's transfer curve, as it can be seen in Fig. 2.2. It calculates as

$$\Delta V_{TH} = -\frac{\Delta Q_{FG}}{C_{tot}} \tag{2.1}$$

with the change of floating gate charge ΔQ_{FG}, and the total floating gate capacitance C_{tot}, which calculates from the sum of all coupling capacitances between floating gate to the other transistor nodes. To program the cell, electrons have to be injected into the floating gate, resulting in a positive threshold shift. Therefore, high voltages have to be applied to allow the electrons to overcome or tunnel through the energy barrier in between channel and floating gate. Channel hot-electron (CHE) injection and Fowler-Nordheim (FN) tunneling are the two common programming mechanisms used. For erasure of the cell, voltage polarities are reversed, and electrons can exit the floating gate again; the threshold voltage reduces. The most common mechanism used for erase is FN tunneling [16, 17]. Typical voltages and durations for the programming/erase mechanisms are listed in Table 2.1. FN tunneling is slow, but very power efficient. CHE injection is the faster mechanism, but it is power inefficient, as just a few of the electrons traveling through the conducting channel are injected into the FG.

The floating gate transistors can be arranged to a memory array in various ways. The predominant two types used today are NOR Flash and NAND Flash. The NOR architecture connects the cells of each column with their drain terminal to the same bitline (BL). The source terminal of all cells is connected to one common source line. In this way each cell can be addressed individually, resulting in a byte-oriented memory. The NAND architecture contains strings of cells (typical

	CG	D	S	P-well	Duration
Programming with CHE injection	10 V	5 V	0 V	0 V	~1 µs
Programming with FN tunneling	20 V	0 V	0 V	0 V	~1 ms
Erasing with FN tunneling	0 V	floating	floating	20 V	~1 ms

Table 2.1: Typical bias conditions and durations for the most common programming and erasing mechanisms.

string lengths are 8,16 or 32), that are connected in between BL and source line. Both string ends have an additional select transistor (string select and ground select). This array organization yields a higher packing density than achieved with the NOR organization. However, cells cannot be addressed individually, but in blocks, resulting in a block-oriented memory.

Since the amount of charge on the floating gate can be varied continuously, the storage principle of floating gate flash memories is perfectly suitable for MLC operation. That means every memory cell can store more than one bit of data, by modulating its threshold such that multiple distinct V_{TH} ranges can be distinguished. The storage of two bit per cell for example requires four distinct memory states. In Fig. 2.2 the shifted I-V characteristics needed for a 2 bit per cell operation is shown. With precise placement of charge Q_{FG}, the cells' thresholds are shaped to four groups representing the four cell states E, P0, P1, P2, which can be mapped to the 2-bit binary representation: 00, 01, 10, 11. The areas highlighted in grey indicate the uncertainty of the cell states arising from inaccurate programming. For reliable multilevel operation it is important to achieve very narrow V_{TH} distributions, to improve read windows. The distribution widths can be reduced by finer voltage stepping used for programming or sophisticated algorithms [9]. This aspect is discussed in Chapter 5. However, factors like random telegraph signal (RTS) [18] or inaccuracy of the sense amplifiers limit the precise placement of the cell threshold voltage.

Besides the floating gate technology, NROM [19] is a further Flash memory technology which is capable of MLC operation. The memory cell differs from the conventional floating gate transistor in the used charge storage medium. Instead of a polysilicon floating gate electrode, an oxide-nitride-oxide (ONO) stack is placed above the channel. It stores charge localized inside nitride traps. Figure 2.3 shows a cross section of an NROM cell. By programming with CHE injection, the electrons are trapped in the nitride above the channel close to the junctions. This spatially confined injection of charges allows to store two bits of data inside one cell, physically separated at both ends of the N-channel transistor. The second bit

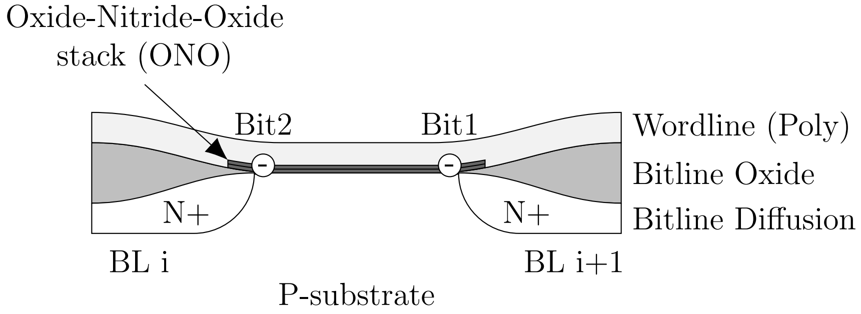


Figure 2.3: Cross section of an NROM cell along a wordline. Electrons are stored in the ONO stack.

is stored by reversal of the BL bias voltages for programming. The concept uses a reverse read method. That means the current direction through the channel during read operation is inverse to the direction during programming. In this way, the addressed bit is located at the electrical source of the cell during read operation. The large depletion zone induced by the high drain voltage at the opposite junction suppresses the second bit's channel impact and hence screens it out. This "local drain-induced barrier lowering effect" (DIBL) allows a reliable readout without significant cross talk between the bits [20]. It has been shown that the NROM cell is capable to store even 4 bit of data [21]. The high density is achieved by combining the concept of two spatially separated storage locations inside one cell with the concept of multiple charge levels programmed to the storage medium. In this way only four different charge states have to be distinguished at each end of the channel region. A conventional floating gate approach would require 16 distinct charge levels. However, the close proximity of both storage regions requires a special programming algorithm to overcome the cross talk issue (2nd bit effect) [22].

2.2 Sensing Principles

The readout of data from a memory cell, called sensing, is performed by a sense amplifier (SA). Two different sensing approaches used with flash memories are state-of-the-art, current and voltage sensing. Both methods are based on a measurement of current flowing through the memory cell. The only difference basically is their way to convert the cell current into a voltage, that then is amplified to full logic level and latched as digital data.

2.2.1 Voltage Sensing

Figure 2.4 shows the typical voltage sensing principle. The read operation is split into two phases, a precharge phase and a sensing phase. At the beginning of the precharge phase the new address is decoded and the cells in the selected row (wordline) and column (bitline) are connected to the sense amplifier such that the cell current I_{Cell} can flow. Furthermore, a reference bitline is connected to the other input of the differential sense amplifier. Usually a normal bitline from another part of the memory array serves as reference here. In this way, good matching with respect of capacitance and resistance is achieved and the sense amplifier has symmetrical conditions at its input. The sensing works with reference cells or with a reference current. In the latter case, all cells along the reference bitline have to be deselected to prevent any extra current charging this node. At the same time, a precharge circuit preconditions both selected bitlines to a read bias voltage. When all voltages in the sensing path have settled, the precharge phase ends (stop of signal PRE) and the sensing phase starts. Both currents I_{Cell} and I_{ref} are converted into a voltage by integration on the bitline capacitance. Hence the current difference translates into a differential voltage ΔV_{BL}. The sense amplifier is activated, detects the voltage difference and amplifies it to a digital output

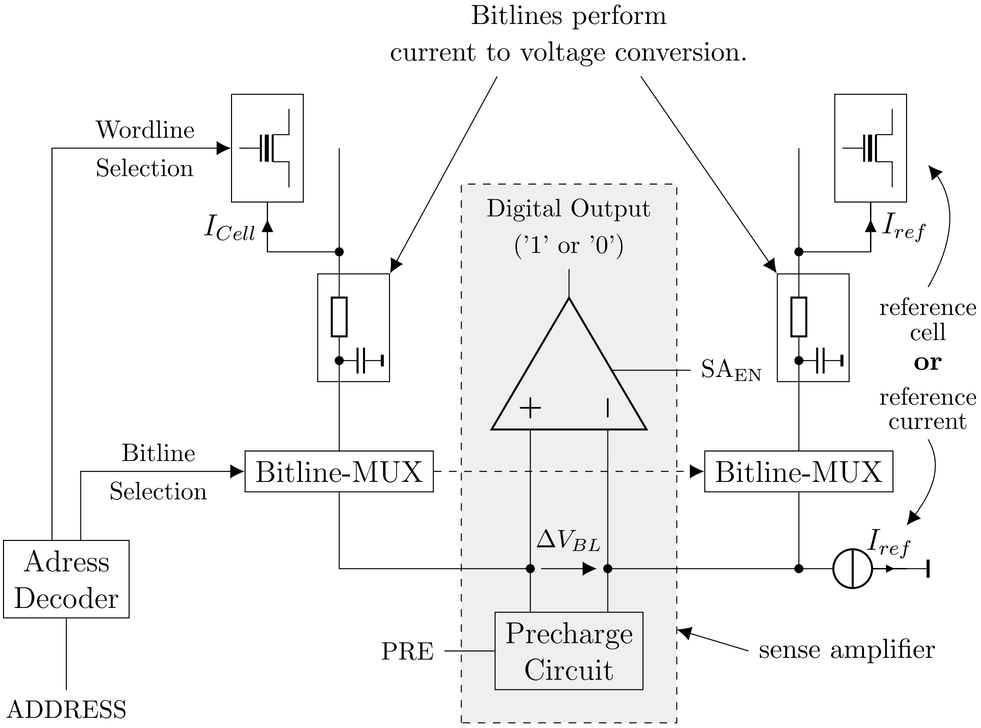


Figure 2.4: Block diagram of the voltage sensing principle.

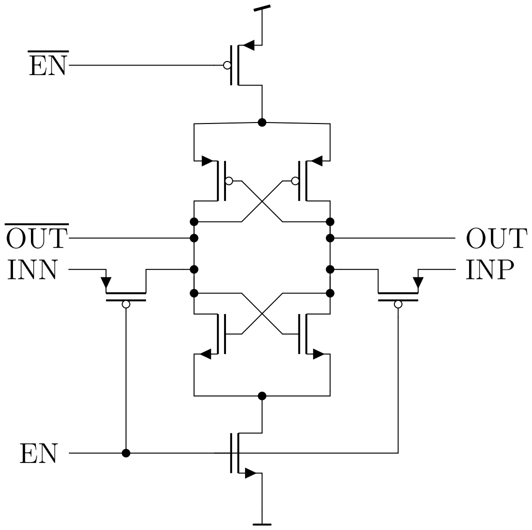


Figure 2.5: Latch type voltage comparator (dyamic comparator).

signal. It basically acts as simple voltage comparator, outputting a logic '1' or '0', depending on which of both bitline voltages is higher.

Voltage sensing is widely used in single-level cell (SLC) flash memory designs [23, 24, 25]. It works with a low complexity sense amplifier and can achieve high-speed read operation. A latch-type voltage comparator as shown in Fig. 2.5 is commonly used for sensing. However, the achievable sensing speed is limited by long bitlines with high capacitance. A long integration time is required to develop sufficient voltage difference ΔV_{BL} at the comparator's input before sensing can be finished. Many designs additionally deploy an offset compensation to overcome missmatch between the transistors inside the latch and hence to reduce the voltage difference required for proper amplification. For MLC operation the sensing procedure described above has to be executed several times either with different reference currents or different cell bias. In Chapter 3 this aspect is discussed in more detail. Different to MLC NAND flash designs [26], voltage sensing is rarely used in MLC NOR flash designs. For this type of memory, the approach of choice is current sensing.

2.2.2 Current Sensing

The current sensing approach performs the current to voltage conversion inside the sense amplifier. Figure 2.6 shows the basic current sensing principle. A current conveyor circuit first precharges the bitline up to its bias level right after the address change. For the rest of the read operation, it keeps the bitline voltage constant and passes both input currents on to a current-to-voltage converter circuit. Here the difference between both currents (I_{Cell} and I_{ref}) is translated to

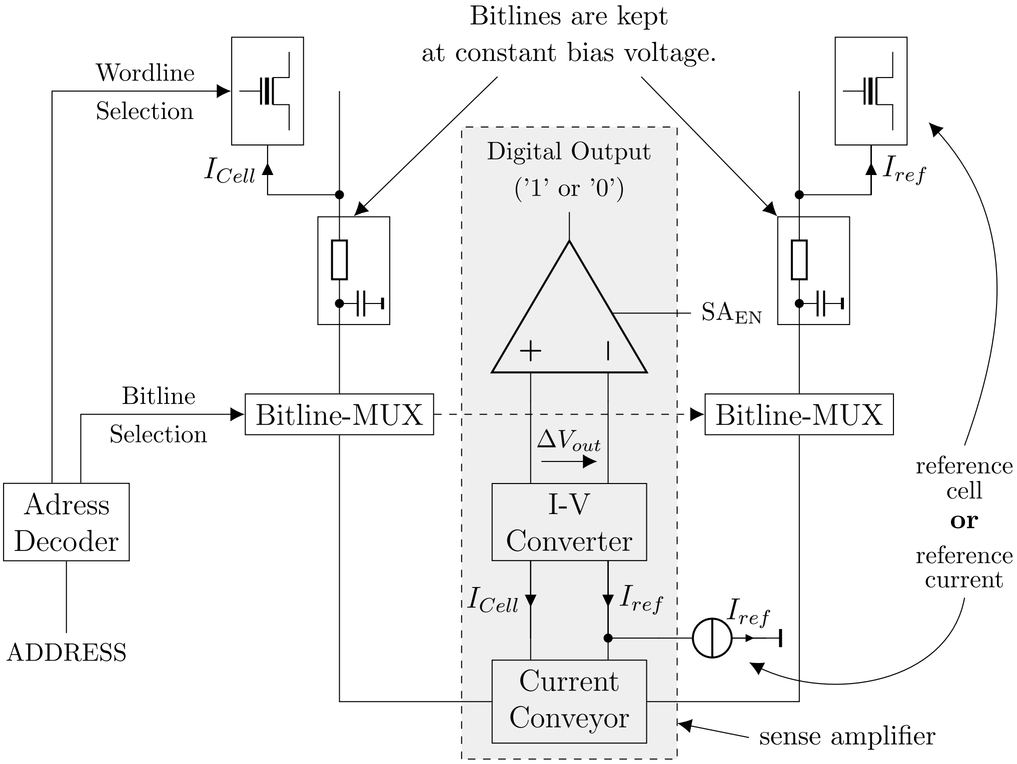


Figure 2.6: Block diagram of the current sensing principle.

a voltage difference ΔV_{out}. Finally, this difference is detected and amplified by a voltage comparator (as in voltage sensing) and afterwards latched as digital data. The sensing method also works with a current generated by a reference cell or with a constant reference current source. If no reference cells are used, the whole reference bitline path including conveyor can be omitted.

To keep the bitline voltage constant, the current conveyor needs an output impedance as small as possible. Common current conveyor implementations used for current sense amplifiers can be found in references [27, 28]. The most commonly used current conveyor for sensing in flash memories is the bitline voltage clamp, implemented as simple cascode (see Fig. 2.7). The scheme, often referred to as drain biasing, can be operated either with a fixed gate bias (static BL clamp) or regulated cascode with negative feedback inverter (dynamic bitline clamping). The latter implementation speeds up the bitline precharge at the cost of additional area and power overhead.

The voltage-to-current conversion is performed by a load attached to the current conveyor output. Typical loads found in sense amplifiers for non-volatile

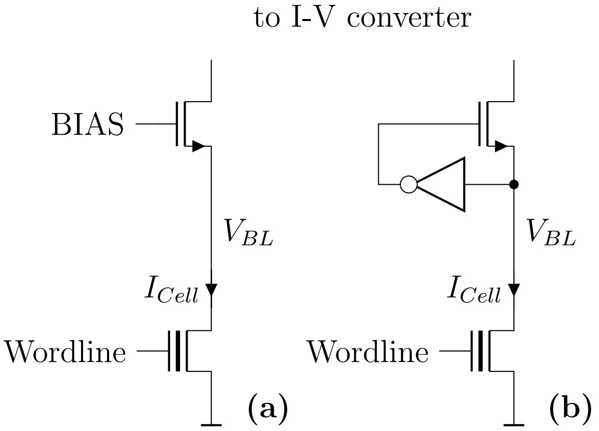


Figure 2.7: Current conveyor implementations commonly used in flash memories: (a) Static BL clamping; (b) Dynamic BL clamping.

memory designs are simple resistor loads [29, 4], diode-connected transistor loads [30] or current mirror loads [31]. The most flexible of them is the current mirror load. It additionally allows to amplify the cell current and to distribute it to multiple voltage comparators. This is necessary if a multi-level cell is to be read with a parallel sensing scheme [3], where the cell current is compared simultaneously with multiple reference currents. Current sensing has a speed advantage over voltage sensing when the memory array has long bitlines with high capacitance [32]. Since the bitlines are kept at constant voltage and the voltage difference only develops between low capacitive internal nodes, the sensing operation is very fast. However, in comparison to that, precharging the bitline before the actual sensing takes long and is associated with significant power consumption. It makes up most of the timing budget of the whole read operation and hence has to be optimized in order not to lose the speed advantage. Furthermore, the constant operation condition of the cell is of advantage for small current windows between the cell states. Cell current stays constant during the whole read operation. With voltage sensing, it drops with the falling bitline voltage and hence results an effectively smaller current window. This is one of the reasons why current sensing is mostly used with MLC flash memories.

2.2.3 Time-Domain Sensing

Current and voltage sensing differ in the way of biasing the memory cells and the location of the performed current to voltage conversion. However, they have in common that the voltage comparator inside the sense amplifier is activated at a specific point in time during the read operation. This activation time depends on the precharge duration of the bitline and the time needed for development of a sufficiently high differential voltage ΔV_{BL} or ΔV_{out}. Once activated, the latch-type sense amplifier very quickly falls into one of its stable states and the read operation is finished. Therefore, both sensing approaches have a well-defined sensing duration.

Time-domain sensing is a method that has a variable sensing duration. It can be combined with voltage or current sensing. In both cases a sense amplifier is used that is active during the whole sensing phase, continuously monitoring the bitline voltage (when combined with voltage sensing) or the cell current (when combined with current sensing). The more obvious approach is the combination with voltage sensing. Figure 2.8 shows the basic differences between conventional voltage sensing and time-domain voltage sensing. First, all bitlines are precharged to a common value V_{PRE}. After settling, the bitlines are released and hence discharged by the cell current or the reference current. Programmed cells (P) conduct almost no current and discharge the bitline very slowly. Erased cells (E) conduct more than the reference current and hence discharge the bitline faster. In the case of conventional voltage sensing, the sense amplifier is enabled at a certain time t_{ref}, when a sufficiently high voltage difference ΔV with regard to the reference has developed. The voltage comparator then detects whether the input voltage is higher (programmed cell) or lower (erased cell) compared to the reference. Time-domain voltage sensing deploys a voltage comparator, that instantly reacts on the bitline voltage crossing a reference voltage V_{ref}. This leads to sense amplifier output signals carrying the information in their timing. Erased cells produce a signal

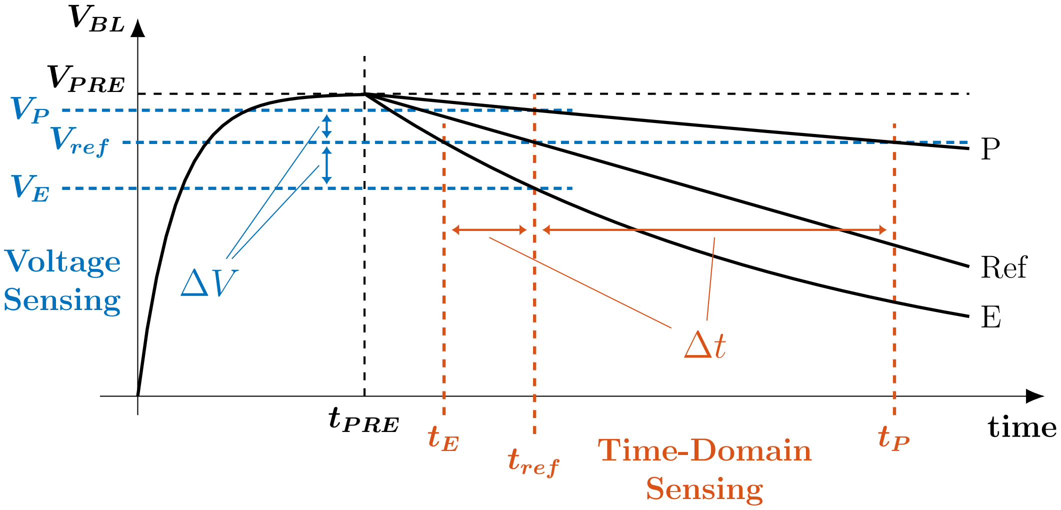


Figure 2.8: Comparison of conventional voltage sensing and time-domain voltage sensing.

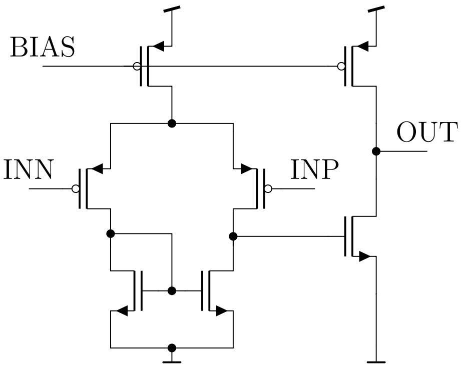


Figure 2.9: Continuous time voltage comparator (static comparator).

with short delay t_E, whereas programmed cells generate signals with long delay t_P. The data can be reconstructed from the time difference Δt towards the reference using a time to digital converter. Common latch type voltage comparators as used in conventional voltage and current sensing are not suitable for this operation, since they are clocked and hence cannot directly react on a voltage crossing at the input. Instead, static voltage comparators have to be used for this sensing approach. A suitable comparator implementation widely discussed in literature is an uncompensated two-stage operational amplifier [33, 34] (see Fig. 2.9).

Conventional current sensing cannot directly be combined with the time-domain principle. Since cell bias conditions are kept constant (including the cell current), there is no time varying trigger condition to be monitored by the sense amplifier. To enable time-domain current sensing, the cell bias can be varied dynamically. Figure 2.10 shows an approach that uses a linear voltage ramp at the control gate to combine advantages from current and time-domain sensing [35]. With the ramp, the cell current rises dynamically during the sense operation and triggers a current mode sense amplifier once it reaches the reference level I_{ref}. The first stage implements a current conveyor that precharges the bitline and converts the current difference $\Delta I = I_{Cell} - I_{ref}$ into a voltage difference ΔV_{out}. This voltage is detected by a voltage comparator, whose digital output then is evaluated in the time-domain. As in time-domain voltage sensing, the deployed sense amplifier has to have a static output stage and cannot use a clocked latch type voltage comparator.

In general, time-domain sensing benefits from the local translation of the cell state (either by cell current or bitline voltage) to a digital signal carrying the information in its timing. The concept is robust since the sensing decision is taken in the digital domain, and it allows an efficient reference distribution throughout the memory module [36]. As described in Section 4.3.2, the reference time can be generated very flexibly and not necessarily is linked to reference cells or a reference

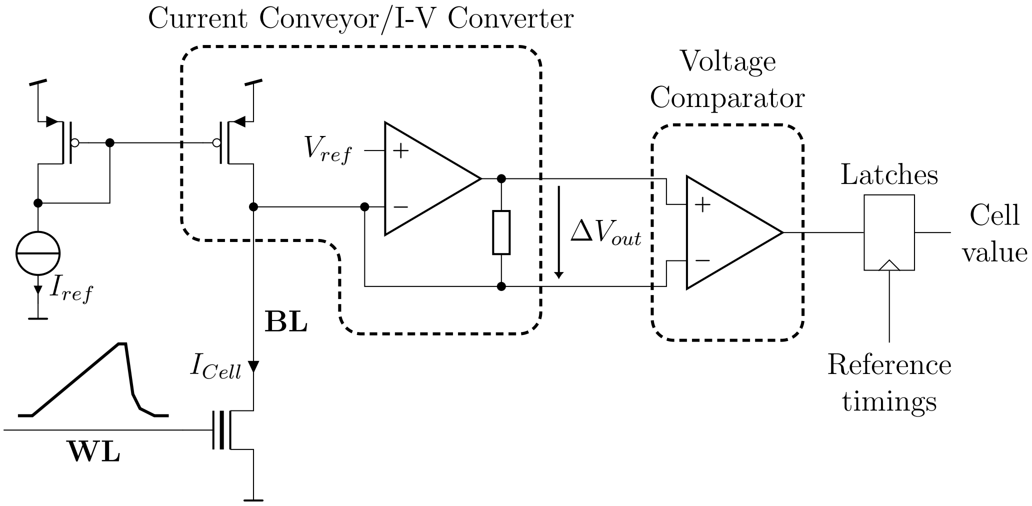


Figure 2.10: Time-domain current sensing scheme using a linear voltage ramp at the word line [35].

current like used in conventional current or voltage sensing.

2.3 Incremental Step Pulse Programming

The key for a reliable MLC operation is a well-controlled threshold voltage (V_{TH}) distribution of the memory cells. That means, the threshold voltage of each cell representing one memory state (E, P0, P1, P2) has to stay within a certain range, in order to prevent the states to overlap each other. Early SLC NAND EEPROM designs faced this issue already in the late 1980s. To program the V_{TH} of the cells to a range in between 0 V and the supply voltage $V_{CC} = 5\,\mathrm{V}$ or 3 V, programming techniques including verify mechanisms were developed [37, 38]. Their underlying principle is a series of short programming pulses (steps) each followed by a verify read, which is a normal read operation executed at the target CG bias. As this verify read is intended to check whether a cell's V_{TH} is programmed sufficiently, it is referred to as program verify. With the program verify, the algorithm excludes cells which already reached the target V_{TH} from further programming pulses. If the program verify yields zero fails, the programming operation is finished and the sequence can proceed to the next memory address to be programmed. With this technique much tighter V_{TH} distributions could be achieved since program duration was controlled individually, which could compensate for varying cell parameters and varying program voltage levels. A tight threshold control however required many programming steps leading to a low programming throughput. To optimize programming speed, a scheme with quasi-logarithmically increasing pro-

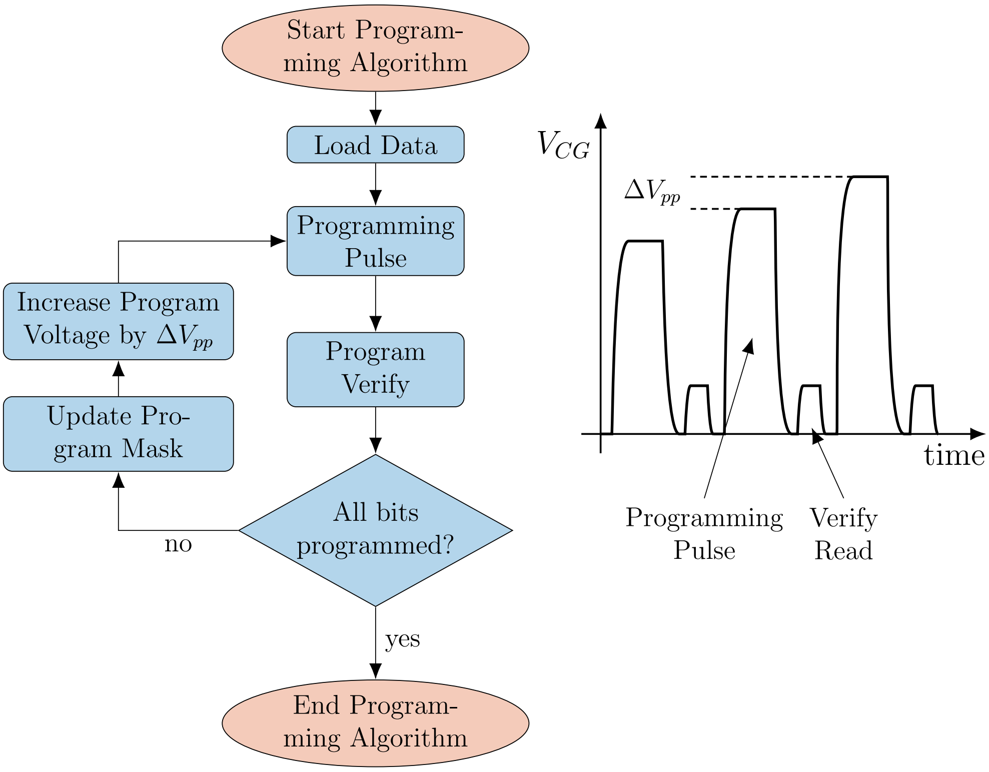


Figure 2.11: Flowchart and control gate waveform of an incremental step pulse programming algorithm.

gramming pulse width was introduced, which reduced the verification overhead by two orders of magnitude [39]. The CG voltage applied during the programming pulses (program voltage or V_{pp}) remained constant. In the mid of the 1990s, a scheme combining the program verify methodology with a step by step increasing program voltage was introduced [40]. It formed the basis for all modern, sophisticated incremental step programming pulse (ISPP) schemes. The program verify prevents the fast cells to be overprogrammed, whereas the increased voltage of the subsequent pulses accelerates the programming of the slow cells. Figure 2.11 shows the basic ISPP algorithm. After each program verify the cells which already reached the program verify level are excluded from the subsequent programming pulse (masked out), the program voltage is increased by ΔV_{pp} and the next programming step is executed started. In theory, this approach yields threshold distributions with a width of ΔV_{pp}. Real designs however do not achieve this theoretical minimum. Their distribution width is wider due to inaccuracies in the sensing path, deviations of the V_{pp} level and RTS [41]. The process of forming a narrow program distribution (distribution of the cell threshold voltages of all cells in the programmed state) step by step, starting from a broad erase distribu-

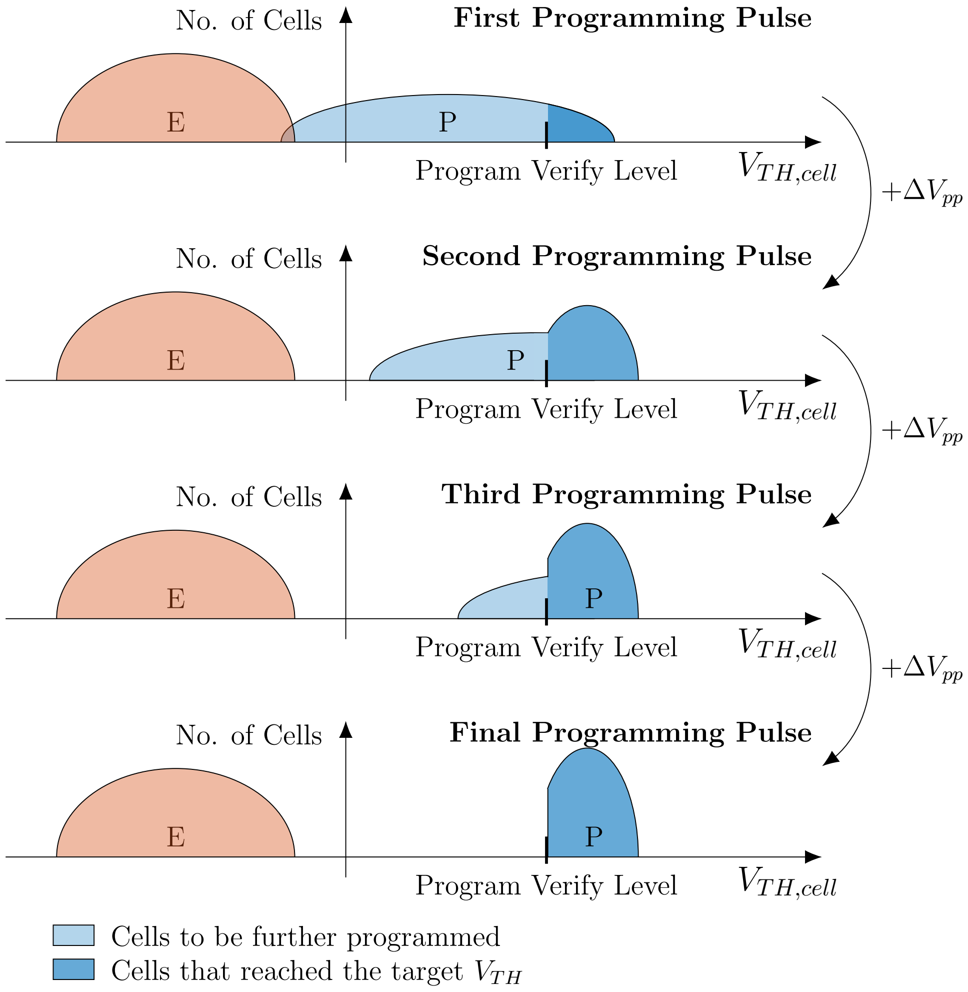


Figure 2.12: Distribution compaction during the incremental step pulse programming algorithm. Starting from the erased state (E), the cells to be programmed (P) get shifted step by step to their target threshold voltage, above the program verify level.

tion (distribution of the cell threshold voltages of all cells in the erased state) is illustrated in Fig. 2.12. Staring from the erased state (E), the threshold voltages of cells to be programmed (P) are shifted step by step to the right. Once they reach the program verify level, they are masked out and their V_{TH} remains at this position. After the final step, all cells to be programmed have a V_{TH} above the program verify level.

The storage of multiple bits per memory transistor in very dense cell arrays requires to consider interferences between adjacent cells, like floating gate to floating gate coupling. Due to the diminishing space for placement of multiple threshold

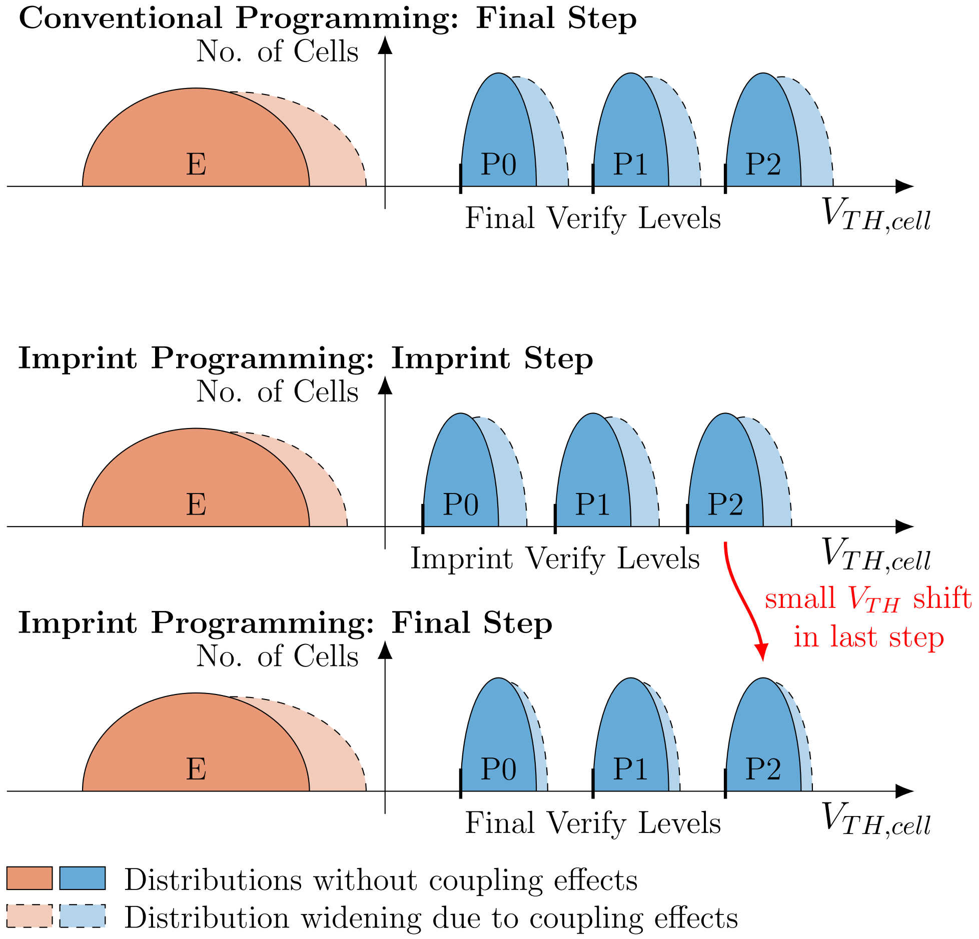


Figure 2.13: Methodology of imprint programming: Programming overshoot is reduced by minimizing the V_{TH} shift caused by the final programming step.

distributions, V_{TH}-shifts caused by programming of adjacent cells have to be minimized, to keep up the already narrow read windows. At the top of Fig. 2.13 the impact of coupling effects on the read window can be seen. Basically, there are two algorithm related measures to reduce the shifts. The first one is a width reduction of the initial erase distribution which reduces the V_{TH} difference between deeply erased cells and strongest programmed cells. A control of the erase distribution can be realized by a erase-verify methodology in the same manner as the ISPP. The other measure is an imprint programming method [42]. It is depicted in the lower two diagrams of Fig. 2.13.

The idea is to program the cells of multiple, physically adjacent pages to a lower target level (the imprint level) first. After this, the cells are shifted to their final V_{TH} level within one last programming operation. In this way, a margin for coupling effects is left in the first step (Imprint Step). The V_{TH}-shift of the last programming step is minimized and hence an overshoot due to FG coupling is reduced. Over the years, algorithms for ISPP have been consequently improved to increase programming throughput, reduce bit error rate to allow for three or even four bit per cell operation [43, 9, 10].

2.4 Summary

The floating gate field effect transistor and NROM are two non-volatile memory technologies that are suitable for storage of multiple bits of data per cell. Both are based on shifting the threshold voltage of a field effect transistor by introduction of charge carriers into a layer above the channel. Either the amount of charge or its specific location can modulate multiple cell states and hence enable multi-level operation. The NROM transistor allows to combine both parameters to achieve a density of 4 bit per cell with 2 bits at each end of its channel.

There are two sensing methods commonly used to read out flash memory cells, voltage sensing and current sensing. With voltage sensing, the cell current and the reference current both are integrated on a separate bitline. The resulting voltage difference is sensed by a voltage comparator. The approach can be realized with low complexity. It reaches fast sensing speeds when bitline capacitances are low. In NOR flash designs, voltage sensing is only rarely used in combination with multi-level cell operation, since it can only be realized in a serial sensing manner. In current sensing, the cell bias conditions are kept constant and the cell current is compared directly inside the current sense amplifier. This enables accurate sensing even with small current windows and fast sensing at long bitlines with high capacitance. It is the common sensing approach for MLC NOR flash designs. Time-domain sensing can be combined with both voltage and current sensing. It uses a time continuous sense amplifier, which triggers when the bitline breaks though a reference voltage or when the cell current breaks through a reference level. The sensing decision is done by a time to digital converter in fully digital fashion. It allows for a flexible and robust distribution of the reference signal.

A good control of the programmed threshold is the key for a robust multi-level operation. ISPP is the method for shaping of narrow threshold distributions. It is based on a programming algorithm, which increases the program voltage from step to step and excludes cells from subsequent programming pulses, once their threshold is shifted to the target value. To handle multiple threshold distributions in dense cell arrays, effects of floating gate to floating gate coupling are reduced by introduction of methods like erase threshold control or imprint programming. Over time, programming throughput and bit error rates have been optimized by various sophisticated programming algorithms.

Chapter 3

Memory Cell Biasing

Flash memory cells can provide long data retention time, provided that they are biased correctly to ensure a reliable programming, erasing and read operation. Depending on cell and memory technology type, biasing schemes differ in complexity. Cell operations in NAND-type flash memories generally have more complex schemes than in NOR-type memories. The NAND arrangement in strings requires all unselected cells in the selected string to be biased in on-state to allow readout of the selected cell. NOR-type memories have to select only one wordline at a time, but on the other hand have to cope with an overerase issue, if 1-transistor cells (1T cells) are used. Hence, measures like preprogramming or soft programming [44] have to be deployed and the control gate voltage has to be chosen carefully. Use of 2-transistor cells (2T cells) eliminate the need for a controlled erase threshold and their suitability for low voltage read operation [45] makes them a good choice for embedded Flash memory application. However, no matter which cell type is used, the voltages conditions applied to the cell terminals during read operation are decisive for the usable read window, which is of big importance for MLC memories. This chapter discusses the impact of different cell biasing schemes on the read conditions of NOR-type Flash cells and investigates their suitability for MLC operation. For assessment of the resulting sensing robustness, the time-domain transfer characteristic obtained with the different cell biasing schemes is derived analytically.

3.1 Concepts

An important aspect of a biasing scheme is the voltage applied to the CG terminal of a memory cell. According to this aspect, the schemes from literature shall be classified in three major groups.

3.1.1 Fixed Gate Biasing

The simplest CG biasing approach is to use one constant voltage level $V_{CG,read}$ during read operation. This method is widely used in SLC as well as MLC applications. The voltage level is placed between the threshold voltages of erased and programmed cells. In this way, cells with a threshold voltage $V_{TH,Cell}$ lower than $V_{CG,read}$ - erased cells (E) - will conduct a high current and programmed cells (P2) with a threshold higher than $V_{CG,read}$ will conduct virtually no current (see Fig. 3.1). The resulting relative cell current window calculates as

$$W_{I,rel} = \frac{I_{Cell,LVT} - I_{Cell,HVT}}{I_{Cell,LVT}}, \tag{3.1}$$

where $I_{Cell,LVT}$ corresponds to the current of the cell state with lower threshold and $I_{Cell,HVT}$ to the current of the cell state with higher threshold. For SLC flash memories, this leads to a comfortable window close to 100%, since the current of the programmed cell is zero with a correct placement of $V_{CG,read}$. To distinguish multiple cell states with only one bias voltage, it has to be placed in between the threshold distributions of the two stronger programmed states (P1, P2), as it is shown in Fig. 3.1. The result is several cell current levels that have to be distinguished. Either, this is accomplished by comparing those currents to multiple reference currents sequentially (serial sensing scheme) [4, 46, 31] or in parallel (parallel sensing scheme) [3]. The latter implementation has to provide parallel sensing and reference structures and hence has significant area overhead. However, it can cut down access time by up to a factor of three within a 2 bit per cell operation. For distinction of the high current states (E, P0), the constant biasing provides a significantly reduced relative cell current window, which is the major issue with this approach, when used for MLC implementations. 2T cell configurations suffer most from this penalty, as their available cell current range is

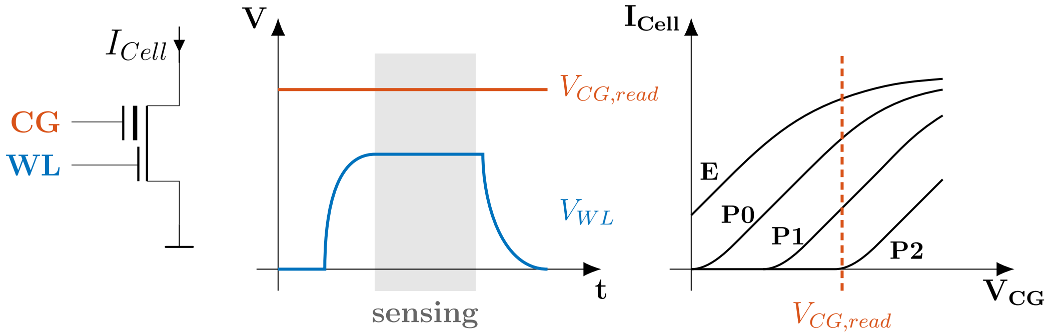


Figure 3.1: Gate voltages and operation point during a read operation with fixed gate biasing concept (2T cell).

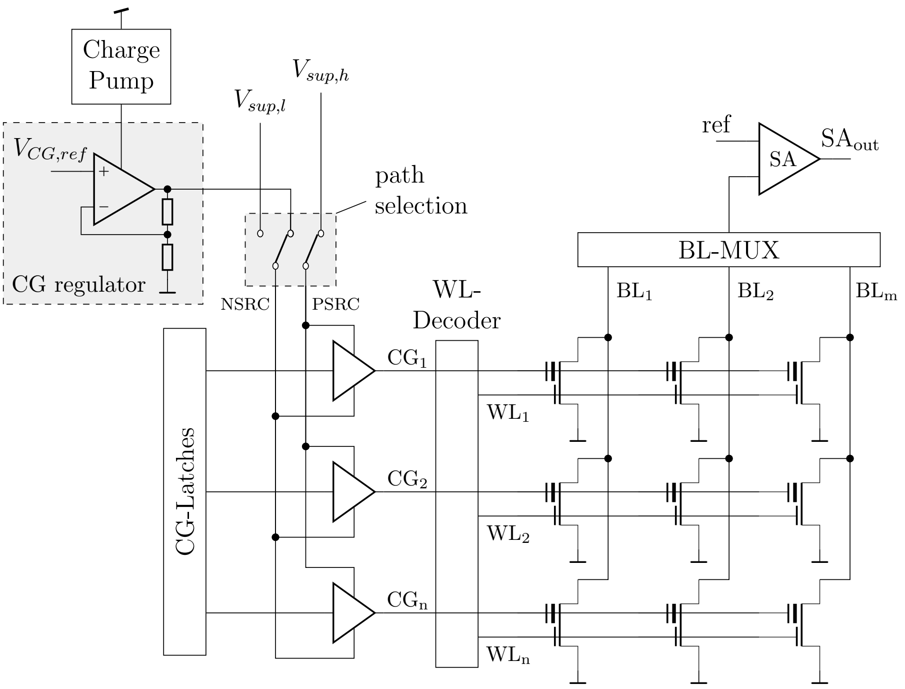


Figure 3.2: Simplified structure of a fixed gate biasing scheme using split-gate 2-transistor flash cells.

limited by the select device of the cell (select transistor) connected to the wordline (WL) (which can be seen in Fig. 3.1 from the saturating transfer curve of the states E and P0). Hence, a precise SA with larger dynamic range is needed to ensure robust read operation. Another drawback are the voltage drops in the array as well as at selection transistors, caused by high cell currents. Also cell gain variations have a big impact on the current in this operating point. Both effects broaden the distributions and are temperature dependent. So, the read windows severely degrade, which further complicates distinction.

Figure 3.2 shows the simplified architecture with a constant CG biasing using split gate, 2T cells. A single voltage regulator is capable to supply the bias voltages for normal read mode as well as for verify modes as used in a ISPP algorithm. The voltage level is distributed via the decoding path to each of the array's CG lines. Depending on the range of voltage level, either the NMOS path (NSRC, used for lower range) or the PMOS path (PSRC, used for upper range) of the driver is chosen. Read or verify levels that exceed the supply voltage of the chip have to be provided by a charge pump, which is also required for programming the flash cells in case of common supply levels. When 2T cells are used, the row selection is done with the select transistor at the WL [47] and the driving strength of the CG path, which has to be designed for high voltage operation, can be reduced to a minimum, as all CG terminals globally stay at $V_{CG,read}$ during read mode. Programming and erasing are slow speed operations (in the range of micro seconds) and hence do not require strong drivers. To avoid sudden erroneous switching within the high voltage modes, the CG selection signals usually are latched. Latching is required for parallel erasing of multiple blocks, too. The WL-decoder needed for row selection can be designed in a lower voltage domain and hence does not contribute much to the area consumption. Therefore, this biasing scheme has a big area advantage over other schemes; the slow high voltage path is fully decoupled from the fast row selection path. This advantage in area disappears with the use of 1T cells. Here, the row selection for read and programming operations are done via the same drivers, which require them to be big, due to large high voltage (HV) devices and high W/L for low ohmic read.

3.1.2 Stepped Gate Biasing

Biasing schemes with variable CG voltage aim for a better operating point of the memory cell during read operation. Stepped gate biasing schemes therefore apply multiple read voltages sequentially to the control gate [6, 48]. For each sense operation, a specific level is chosen, which is optimal for distinction of two adjacent cell states (see Fig. 3.3). With a CG voltage just below the threshold of the stronger programmed cell, the relative cell current window gets close to 100%. Thus, the dynamic range requirement of the sense amplifiers is heavily reduced compared to an approach using fixed gate biasing. Each comparison can be carried out with one very low reference current, which considerably reduces the impact of resitive voltage drops in the read path and cell gain variations. The use of multiple cell operation points requires a serial sensing scheme and hence suffers from large time overhead compared to fixed gate biasing. Therefore, it is not fully

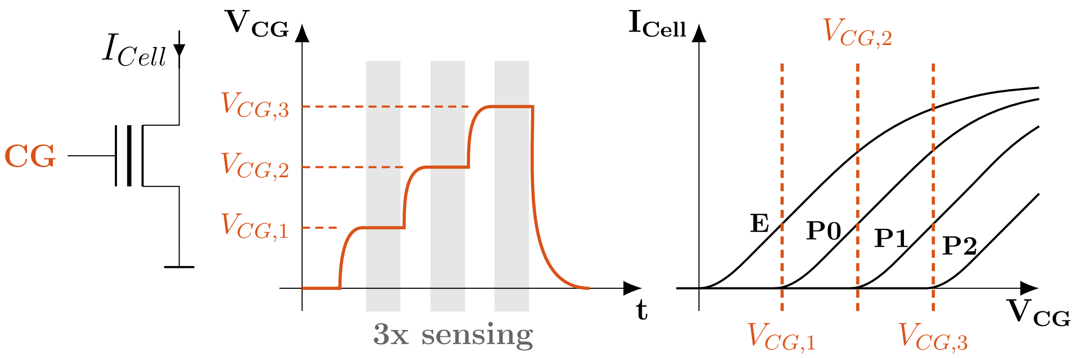


Figure 3.3: Gate voltages and operation point of a read operation with stepped-gate biasing concept (1T cell).

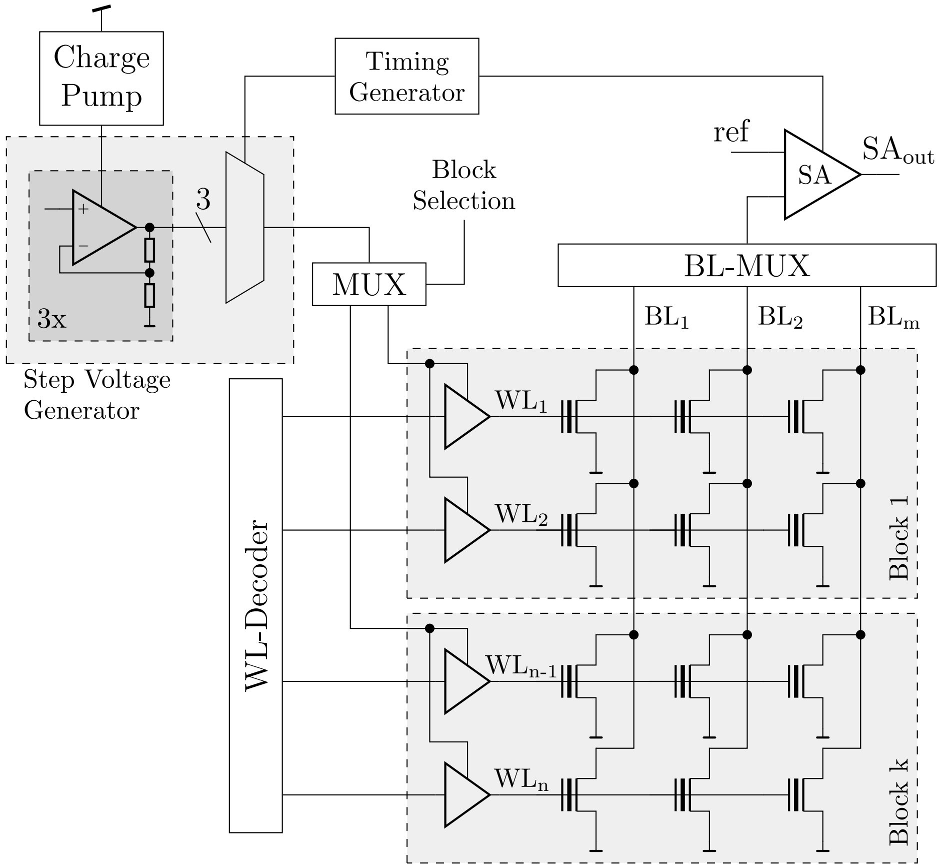


Figure 3.4: Simplified structure of a stepped gate biasing scheme using 1-transistor flash cells.

suitable for high speed operation. Tightly spaced threshold distributions and the requirement of very low read currents make this the state of the art approach for high density NAND flash memories achieving up to four bits per cell [43].

The implementation of a serial sensing with stepped CG bias voltage can be done in a similar way as the fixed gate biasing. Figure 3.4 illustrates a typical stepped gate biasing setup [6, 48]. The most critical aspect for fast memory access is the change of V_{CG} within every read operation. Each individual voltage level has to be provided by its own regulator. A source follower output ensures a low output impedance and the capacitance along the wordline selection path is reduced by a partitioning into small memory blocks, that are all, except the selected one, separated from the voltage supplies. A separation between slow HV path for programming and fast selection path during read operation is not possible here. Therefore, independent of the used flash cell type (1T or 2T), the WL-driver has to be designed with high driving strength. Otherwise, its RC time constant becomes too big for quick voltage stepping. Consequently, this biasing approach requires more area than the fixed gate schemes. Furthermore, a logic is needed, which consecutively selects the different bias levels and triggers the individual sense operations. As long as there is sufficient wait time for all changing nodes to settle, an exact synchronization is not needed. The complexity added by this sequencing logic therefore is low.

3.1.3 Ramped Gate Biasing

The third biasing scheme found in literature dynamically applies a linear voltage ramp to the cells' CG terminals [7, 35]. After a precharge phase, the ramp starts at the beginning of the sensing phase and varies the cell biasing during the whole operation (see Fig. 3.5). In this way, erased cells start with an initial current which increases with time. Programmed cells start with zero current and reach conductivity once the bias ramp hits their threshold voltage. Hence the comparison between different cell states cannot be done in terms of the cell current level itself. The decision has to be transferred into the time domain. This can be done by integrating the cell current on the BL capacitance (or on another dedicated integration capacitance) up to a certain voltage level (time-domain voltage sensing) or by directly comparing the instantaneous current against a certain level (time-domain current sensing). With its inherent time-domain nature, the ramped gate biasing is capable of parallel sensing; multiple sense operations, including precharge phase, are not needed. Therefore, low access times can be achieved. In addition, the advantages of an increased CG bias range apply for this biasing scheme. It provides a compromise between sensing speed and accuracy and is well suited for the use in MLC flash memories.

The structural difference between a ramped gate and a stepped gate biasing

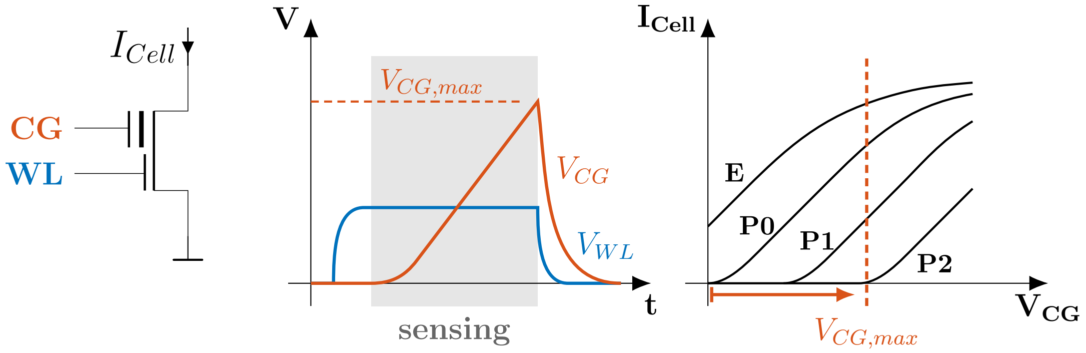


Figure 3.5: Gate voltages and operation point of a read operation with ramped gate biasing concept (2T cell).

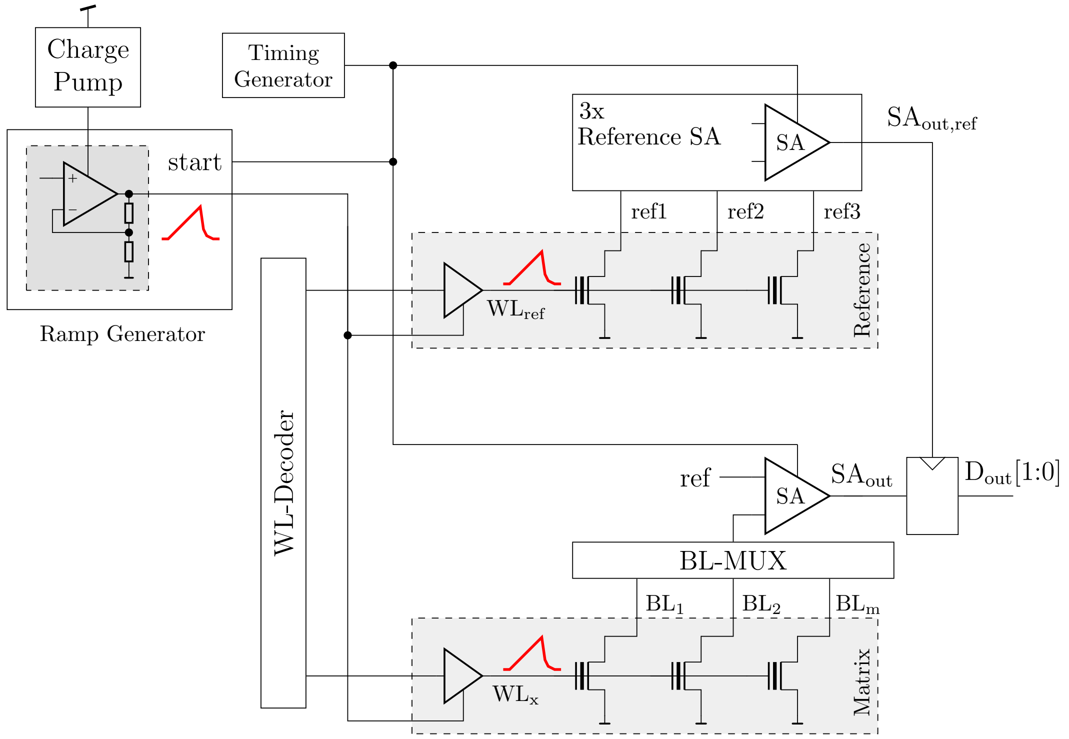


Figure 3.6: Simplified structure of a ramped gate biasing scheme including reference path.

implementation is small. However, the linear voltage ramp requires more complex row selection and sensing circuitry. Figure 3.6 shows a corresponding implementation. To propagate the voltage ramp quickly, strong WL-drivers are as important for this approach as for stepped gate biasing schemes. As the threshold voltage of the erased state (lowest V_{TH} state) usually is negative for embedded NOR Flash cells, it is a good choice to start the ramp at 0 V. Therefore it has to be transferred to the CG via the NMOS path of the driver or via a special transmission gate. In the first case, a sufficient gate overdrive has to be provided, which might require a pumped voltage. The latter solution considerably increases the area of the HV driver structure.

For best match to cell properties, reference cells should be used and sensing paths (matrix and reference) have to be matched in terms of resistance and capacitance, when a ramped gate biasing is deployed. With the dynamic CG biasing, this task requires special attention, as there are transient events both on WL and BL that have to be properly synchronized. Since both circuit parts (WL-driver and SA) are not necessarily located nearby, their signal paths should be well aligned. Timing signals, like the precharge enable, the SA strobes and the voltage ramp itself, hence should have the same traveling directions and a matched RC delay along the memory array. The reference cell array and the highly dynamic ramp generator lead to a considerable circuit and testing overhead. Reference cells have to be preconditioned and afterwards protected against any disturbance, since they must not change their properties during the whole lifetime. These requirements make this approach the most expensive in terms of area and complexity.

3.2 Speed Considerations

Depending on its application, a flash memory faces different speed requirements. One key performance parameter is the read throughput, which is significant for storage applications with large and preferably linearly accessed data. It is mainly increased by parallelization of read accesses, requiring more sense amplifiers and a faster data bus to transfer the data. The memory's read latency, described by its random access time t_{acc}, is another important performance parameter. It has large relevance in applications with many random memory accesses. Execution of code directly from the non-volatile memory (Execute-in-place), as commonly done inside MCUs, also requires a fast random memory access with low latency. Both performance parameters, read throughput and read latency, are affected by precharge duration of the memory cells in the array, but the latter one cannot be improved by placing more sense amplifiers in parallel. Latency is only dependent on the access time of one single read. Therefore, the impact of the three different cell biasing schemes on the achievable access time shall be analyzed.

The speed of the variable CG voltage bias schemes is mainly limited by the dynamics of the CG lines. Within this section they are referred to as wordlines, although in 2T cell configurations the select gate is associated with this term. As the dynamics of the select gate usually is higher compared to the CG, they are not included in the speed considerations.

3.2.1 Wordline Model

The wordline of a flash array consists of a high resistive poly silicon line, with equally spaced contacts (stitches), connecting to a low resistive metal line, that is routed over the whole width of the array. Figure 3.7 shows a part of a wordline model using two distributed RC lines for metal and poly routing, regularly interconnected with metal stitches. $R_{WL,md}$ represents the metal resistance, $C_{WL,md}$ the metal capacitance per unit length. The poly resistance in between the cells is modeled by $R_{WL,pd}$ and the cells' gate capacitance by C_{Cell}. Coupling capacitances to adjacent BLs are included in this value. With a resistance $R_{WL,pd}$ in the range of kilo ohms, the poly line RC time constant quickly becomes dominant and hence

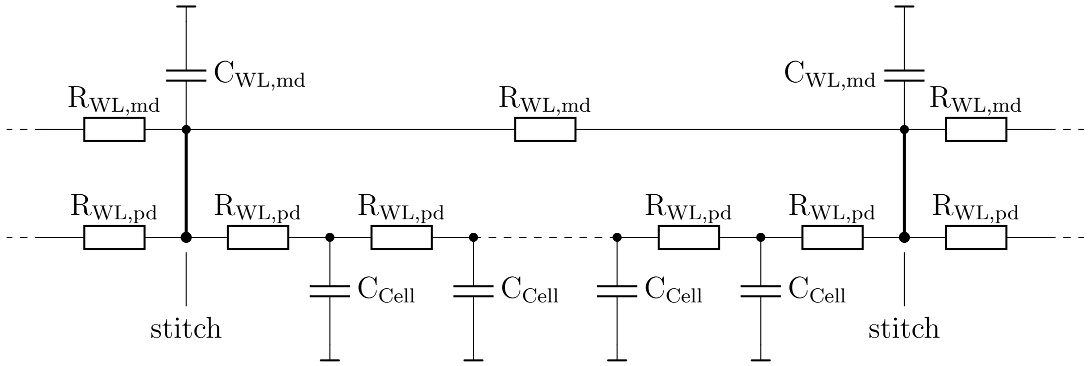


Figure 3.7: Delay line WL model including metal and polysilicon path regularly connected by stitches.

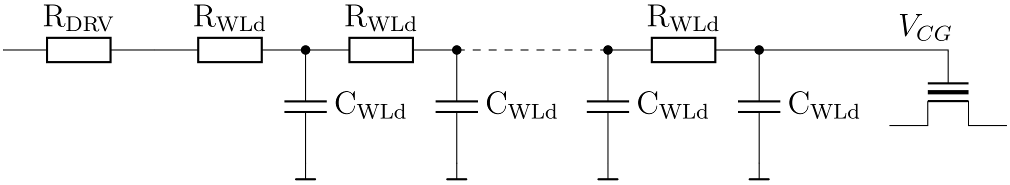


Figure 3.8: Simplified delay line WL model including driver resistance.

limits the dynamics of the cells' control gate voltage. To allow quick changes of the CG potential, the distance of the stitches has to be low, in order to achieve a poly line delay, which is small compared to the total RC time constant of the WL. Thus, the model can be simplified significantly, by merging all capacitances into one lumped capacitor $C_{WL,d}$ per stitch. This simplification results in the model shown in Fig. 3.8. It includes the wordline driver resistance R_{DRV}. The floating gate transistor symbolizes the cell at the far end of the wordline. It experiences voltage changes with the worst-case delay. To get an appropriate estimation, the delay line can be modeled using a first order π-approximation [49]. Neglecting the source resistance (internal resistance of the driving source), the time constant of the distributed RC line then is described by its lumped resistance R_{WL} and its lumped capacitance C_{WL} as

$$\tau_{WL} = \frac{R_{WL}C_{WL}}{2}. \tag{3.2}$$

Though the settling time of the step response is overestimated, using this approximation, it gets more accurate, as the source resistance rises. Since the WL-driver uses high voltage transistors, R_{DRV} accounts for the major portion of the path resistance, making it a good approximation for the source resistance. A reasonable estimation of the settling time hence can be obtained replacing the RC line

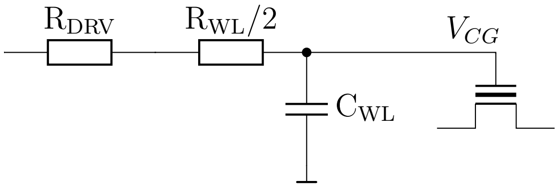


Figure 3.9: Strongly simplified, lumped WL model including driver resistance.

by the resistance $R_{WL}/2$ and the capacitance C_{WL}. With this strongly simplified WL-path model shown in Fig. 3.9, the time needed for proper control gate voltage changes can easily be compared among the different biasing schemes. Its overall RC time constant calculates as

$$\tau_{CG} = \left(R_{DRV} + \frac{R_{WL}}{2}\right) C_{WL}. \tag{3.3}$$

The error introduced by the simplifications made above, does mainly affect the cells that are located far from the WL stitches. They experience the changing bias voltage with a slightly higher delay than the cells directly at the stitches. For a general comparison of the different biasing schemes, this is of lower relevance, but anyhow has to be considered for the design of the memory array.

3.2.2 Access Time

Fixed Gate Biasing

A constant CG voltage scheme using a 1T cell requires the WL to be raised to its bias level for every read operation. Consequently, a precharge time of at least three times τ_{CG} has to be spent prior to the sense phase, to ensure sufficient settling. For typical values ($C_{WL} = 1\,\mathrm{pF}$, $R_{DRV} = 2\,\mathrm{k\Omega}$, $R_{WL} = 1\,\mathrm{k\Omega}$) the precharge time needed yields $t_{pre,WL,fix} = 3\tau_{CG} = 7.5\,\mathrm{ns}$. The BL is precharged simultaneously. As the BL does not consist of polysilicon but of metal, it is precharged faster. Thus, the read access can be approximated to

$$t_{acc} = t_{pre,WL} + t_{sense}. \tag{3.4}$$

The sense delay t_{sense} represents the time needed by the sense amplifier to compare the cell current to the reference current. A typical sense delay of $t_{sense} = 10\,\mathrm{ns}$ [50] would then result in an access time of $t_{acc,fix,1T} = 17.5\,\mathrm{ns}$. The sense delay of an MLC memory will not be the same as in an SLC implementation like referenced from [50], since the current read windows are much tighter and the respective signals (e.g. cell current differences) are smaller. However, it affects the absolute duration of the read access for all biasing schemes the same and hence this assumption still is suited for a comparison among the three different approaches.

Using a 2T cell yields the advantage that the whole array can be biased at the target control gate voltage during read mode and the row selection is done with the select devices. Since these are typically lower voltage class devices, they can be selected much faster. In this case, the BL precharge usually defines the precharge duration and the result is an access time of about

$$t_{acc,fix,2T} = t_{pre,BL} + t_{sense} < t_{pre,WL} + t_{sense} = 17.5\,\text{ns}, \tag{3.5}$$

depending on the time needed for BL precharge $t_{pre,BL}$.

Stepped Gate Biasing

The stepped gate approaches change the CG voltage multiple times for one read. Given there is sufficient gate overdrive for the WL-driver, the same values as above can be taken for calculation of the precharge time needed. A two bit per cell operation requires three consecutive sense operations each preceded by a precharge phase. So every read requires in total a precharge time of

$$t_{pre,step} = 3t_{pre,WL,fix} = 22.5\,\text{ns}. \tag{3.6}$$

This time only considers the stepping of the WL. Charging of the path from step voltage generator to the selected row driver (compare Fig. 3.4) might prolong the needed wait time. To estimate the possible read access time, three times the sense delay has to be added. Considering the changed bias conditions for each consecutive sensing, the relative current read window is better than in the fixed gate scheme. Absolute cell currents however are in the same range as the current of the intermediately programmed cells (P1) when using only one gate bias (compare Figs. 3.1 and 3.3). So, the sensing gets more robust, but the sensing delay will stay in the same ballpark. Therefore, the same sense delay is assumed, which results in a total access time of

$$t_{acc,step} = t_{pre,step} + 3t_{sense} = 52.5\,\text{ns}. \tag{3.7}$$

Ramped Gate Biasing

Ramping the CG voltage yields a compromise between sensing speed and optimum read bias. In contrast to the stepped gate scheme, it is not necessary to wait for the settling of the bias voltage. Instead, sensing phase can start simultaneously with the rise of the applied voltage ramp. However, the selected CG terminals end at a high voltage level after each read operation and they have to be precharged at least once, in order to ensure same starting conditions for each ramp. Using the same resistance and capacitance values as above results in $t_{pre,ramp} = 3\tau_{CG} = 7.5\,\text{ns}$. After this wait time the ramp generator starts to raise the voltage. Along the path from the generator to the cells (compare Fig. 3.6), the ramp gets delayed. In general, the major part of the delay is caused by the WL-driver and the WL itself. WL-drivers are placed at every WL and thus contribute significantly to the total area, the remaining blocks however are placed less often and hence can be scaled up sufficiently. So the delay is determined by the RC time constant τ_{CG} from (3.3). To get to the WL's response to a linear voltage ramp at its input, we start with the step response of the RC lowpass with $V_{CG}(t=0)=0$ [51]:

$$V_{CG,step}(t) = V_{CG,read}\left(1-\exp\left(-\frac{t}{\tau_{CG}}\right)\right). \tag{3.8}$$

Since a linear ramp input signal $v(t) = m \cdot t$ is obtained by integration of a step signal with height m, the ramp response of the RC lowpass can be calculated using the integration property of linear time-invariant systems [52]. So the ramp response results from the integration of (3.8) to

$$V_{CG,ramp}(t) = m \cdot t - m \cdot \tau_{CG}\left(1-\exp\left(-\frac{t}{\tau_{CG}}\right)\right) \tag{3.9}$$

with $m = V_{CG,read}$. Under the condition that the input voltage ramp lasts long enough compared to the time constant of the wordline ($t > 3\tau_{CG}$), the exponential term approaches zero and simply leaves a delayed version of the input. This means that as long as the ramp slope is chosen moderately, as

$$m = \frac{V_{CG,read,max}}{6\tau_{CG}} \tag{3.10}$$

for example, the signal reaches the control gates of the memory cells with a delay of τ_{CG} approaching the same slope (see Fig. 3.10). Since the sensing phase starts simultaneously with the ramp start, the read access is finished latest one sense delay after the control gate voltage reached the maximum read bias target $V_{CG,read}$ (corresponding to the third read bias level in the stepped gate approach; compare Figs. 3.1 and 3.5). According to (3.9) and the required precharge time $t_{pre,ramp}$, this accumulates to an access time of

$$t_{acc,ramp} = t_{pre,ramp} + 6\tau_{CG} + \tau_{CG} + t_{sense} = 10\tau_{CG} = 35\,\text{ns}. \tag{3.11}$$

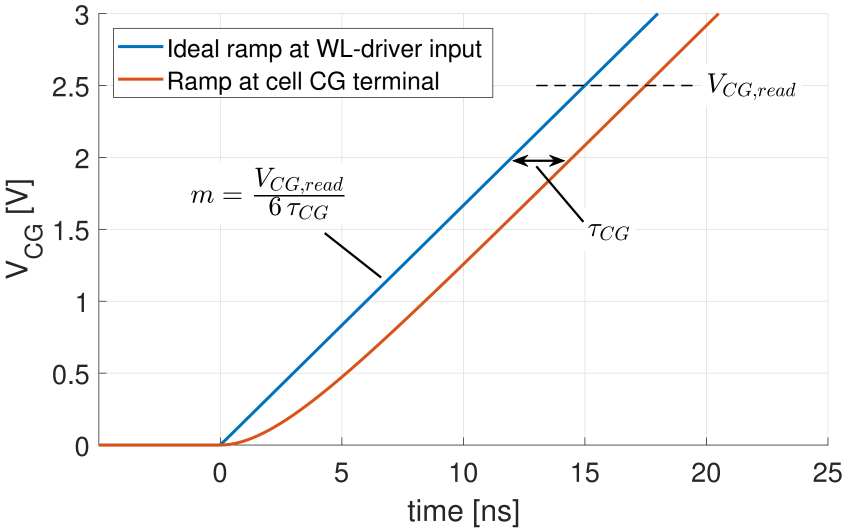


Figure 3.10: Ramp propagation on WL according to lumped RC model.

3.2.3 Comparison

Comparison of the calculated read access times shows a clear speed advantage of the constant gate bias scheme. Especially when using a 2T flash cell type with low voltage select transistor, this biasing reduces the speed limitation of the WL-path to a minimum and read speed becomes determined by the BL precharge and sensing speed. Stepped gate approaches suffer from significant time penalty due to their need of multiple precharge phases. They benefit from better cell biasing, but their total sensing time remains high. A good compromise between both is given with the ramped gate system, which merges all sensing phases into a long one, eliminating the required wait time of all intermediate precharge phases. The sensing operations are delayed by the slowly rising gate bias. This preserves the better cell biasing conditions and yet allows simultaneous sensing of all cell states. However, both schemes with variable CG voltage have to ensure a proper connection of step voltage generator and ramp generator to the WL-driver. Otherwise they lose additional time during change of the cell bias, which is the major contributor of time overhead here.

3.3 Time-Domain Transfer Characteristic

To compare all three biasing schemes with respect to their sensing robustness with MLC operation, a common sensing method has to be selected. The ramped gate biasing scheme requires to use a timing-based concept. Therefore a time-domain

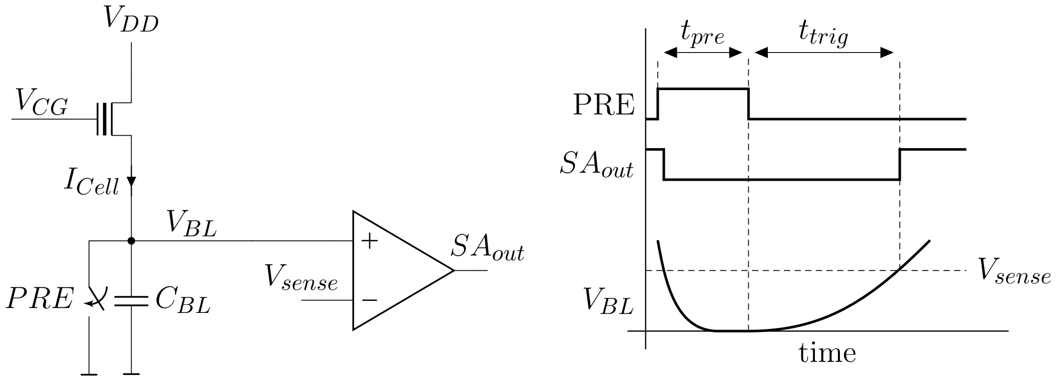


Figure 3.11: Basic principle of the analyzed time-domain voltage sensing.

voltage sensing approach as used in [53, 21] is taken for assessment (see Fig. 2.8).

Figure 3.11 depicts the principle of the cell current to time conversion, which establishes the basis of the analyzed sensing scheme. At the beginning of each read operation the bitline is discharged to V_{SS}. The actual sensing is started by disabling the discharge switch, and the cell current I_{Cell} accumulates charge on the bitline capacitance C_{BL}. As soon as the bitline voltage V_{BL} reaches the reference level V_{sense}, the sense amplifier output SA_{out} flips from low to high and the sense operation is completed. The time elapsed between the start of integration and the trigger event shall be defined as trigger time t_{trig}.

As evaluation criterion for the sensing robustness of all three biasing approaches, the transfer characteristic from cell threshold $V_{TH,Cell}$ to trigger time t_{trig} (time-domain transfer characteristic) is derived in the following. The calculations are done for two different CG stimuli: a constant voltage bias (fixed gate biasing) and a linear voltage ramp (ramped gate biasing). For assessment of the robustness obtained by stepped gate biasing, the characteristic of fixed gate biasing can be taken three times, each with a different bias level $V_{CG,read}$. This section includes only the initial assumptions of the derivation, the important steps for solving of the problem and the final result. A detailed description of the derivation can be found in Appendix A.

With the assumption of no current flowing into the sense amplifier's input terminals, the transient behavior of the bitline voltage during sensing phase is described by the differential equation

$$\frac{dV_{BL}}{dt} = \frac{I_{Cell}(t)}{C_{BL}}. \tag{3.12}$$

In order to relate $V_{TH,Cell}$ with V_{BL}, the I-V-characteristic of the flash cell is needed. Advanced eFlash cell types usually are operated with low drain-source

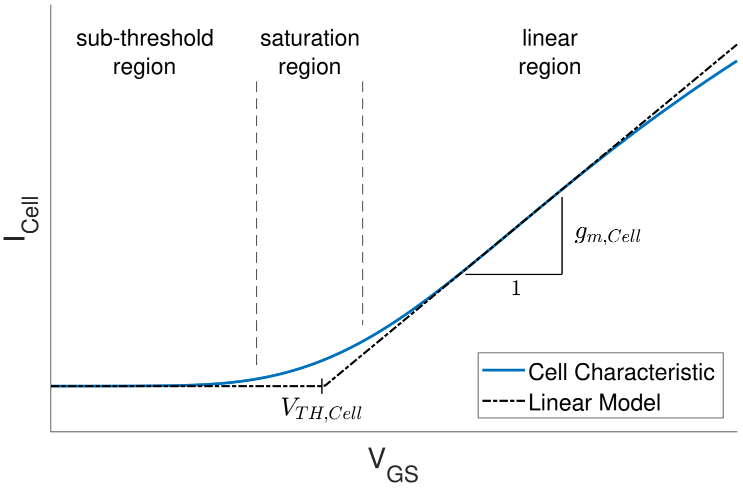


Figure 3.12: Simple model for the I-V-characteristic of a flash cell with low drain-source bias.

bias voltage in read operation [45]. Hence the CG voltage range of their saturation region is very small and a simple piecewise linear model can sufficiently describe the I-V-characteristic of the flash cell. As shown in Fig. 3.12, it is described by the following equations:

$$I_{Cell}(V_{CG}) = \begin{cases} (V_{CG} - V_{BL} - V_{TH,Cell}) \cdot g_{m,Cell} & \text{for} \quad V_{CG} - V_{BL} \geq V_{TH,Cell} \\ 0 & \text{for} \quad V_{CG} - V_{BL} \leq V_{TH,Cell} \end{cases} \tag{3.13}$$

Combining (3.12) with the cell current in the conducting case yields the first-order linear differential equation

$$\frac{dV_{BL}}{dt} = \alpha \left(V_{CG}(t) - V_{BL}(t) - V_{TH,Cell}\right) \quad \text{with} \quad \alpha = \frac{g_{m,Cell}}{C_{BL}}. \tag{3.14}$$

Without knowledge of the particular waveform of $V_{CG}(t)$, (3.14) can be solved for $V_{BL}(t)$ resulting in

$$V_{BL}(t) = e^{-\alpha t} \int \alpha e^{\alpha t} \left(V_{CG}(t) - V_{TH,Cell}\right) dt + c_1 e^{-\alpha t}. \tag{3.15}$$

Further analysis requires the integral to be solved for both input stimuli.

3.3.1 Fixed Gate Biasing

For a constant control gate biasing, $V_{CG}(t)$ simply is replaced by $V_{CG,read}$. With the boundary condition $V_{BL}(t=0)=0$, (3.15) can then be solved as

$$V_{BL}(t) = (V_{CG,read} - V_{TH,Cell}) \cdot \left(1 - e^{-\alpha t}\right). \tag{3.16}$$

Evaluating (3.16) for $V_{BL}(t_{trig}) = V_{sense}$ finally results the transfer function that connects the cell threshold voltage $V_{TH,Cell}$ and the trigger time t_{trig}

$$t_{trig}(V_{TH,Cell}) = -\frac{1}{\alpha} \cdot \ln\left(1 - \frac{V_{sense}}{V_{CG,read} - V_{TH,Cell}}\right). \tag{3.17}$$

It only yields a valid result if the logarithm's argument becomes positive. This is reasonable, since a threshold voltage, which is too close to the read bias $V_{CG,read}$, does not deliver a cell current and thus would theoretically never cause a trigger event. Cell thresholds above the read bias result in negative trigger times having no causal relation. (3.17) describes the transfer characteristic for constant CG bias within the range of $-\infty < V_{TH,Cell} < V_{CG,read} - V_{sense}$.

3.3.2 Ramped Gate Biasing

For the second scenario, a dynamic linear voltage ramp is applied to the control gate. Defining the start of the integration as $t = 0$, a voltage ramp with slope m, starting at $V_{CG}(0) = 0$ can be described by $V_{CG}(t) = m \cdot t$. Inserted in (3.15), the BL voltage is then be solved as

$$V_{BL}(t) = c_1 e^{-\alpha t} + m \cdot t - \left(V_{TH} + \frac{m}{\alpha}\right). \tag{3.18}$$

In order to eliminate coefficient c_1, again, the boundary conditions have to be considered. Since the control gate voltage now changes dynamically, the conditions are more complex. Two cases leading to different boundary conditions can be distinguished:

Case I

The cell's threshold voltage is below the initial control gate voltage and the cell current flows immediately:

$$V_{TH,Cell} < V_{CG}(0) \quad \Rightarrow \quad V_{BL}(t=0) = 0. \tag{3.19}$$

Then, the bitline voltage starts at zero and rises continuously in the integration phase. Using this boundary condition, (3.18) yields the constant $c_1 = \frac{m}{\alpha} + V_{TH,Cell}$ and thus results in

$$V_{BL}(t) = \left(\frac{m}{\alpha} + V_{TH,Cell}\right) \cdot \left(e^{-\alpha t} - 1\right) + m \cdot t \tag{3.20}$$

for the bitline voltage curve.

Case II

The cell's threshold voltage is above the initial control gate voltage and the cell current starts flowing only after the voltage ramp crosses the threshold voltage

$$\begin{aligned} V_{TH,Cell} \geq V_{CG}(0) \quad &\Rightarrow \quad V_{BL}(t \leq t_0) = 0 \\ V_{CG}(t_0) = V_{TH,Cell} \quad &\Rightarrow \quad t_0 = \frac{V_{TH,Cell}}{m}. \end{aligned} \tag{3.21}$$

Here, t_0 represents the point in time when V_{CG} exactly reaches $V_{TH,Cell}$. Until then, the bitline voltage stays at zero. Considering these boundaries, the constant solves for $c_1 = \frac{m}{\alpha} \cdot e^{\alpha t_0}$ and the bitline voltage curve can be found as:

$$V_{BL}(t) = \begin{cases} \frac{m}{\alpha} \cdot \left(e^{-\alpha(t-t_0)} - 1\right) + m \cdot (t - t_0) & \text{for} \quad t \geq t_0 \\ 0 & \text{for} \quad t < t_0 \end{cases} \tag{3.22}$$

In order to retrieve the full transfer function from cell threshold voltage $V_{TH,Cell}$ to trigger time t_{trig}, the solutions from both boundary cases ((3.20) and (3.22)) are evaluated for $V_{BL}(t_{trig}) = V_{sense}$. Solving both equations for the time yields the following piecewise defined function:

$$t_{trig}(V_{TH,Cell}) =$$

$$\begin{cases} \frac{1}{\alpha} \cdot \left(W\left(-\left(1 + \alpha \frac{V_{TH,Cell}}{m}\right) e^{-\left(1 + \alpha \frac{V_{sense}+V_{TH,Cell}}{m}\right)} \right) + 1 \right) \\ \quad + \frac{V_{sense} + V_{TH,Cell}}{m} & (I) \\ \\ \frac{1}{\alpha} \cdot \left(W\left(-e^{-\left(\alpha \frac{V_{sense}}{m} + 1\right)} \right) + 1 \right) \\ \quad + \frac{V_{sense} + V_{TH,Cell}}{m} & (II) \end{cases}$$

$$(I) \quad \text{for} \quad V_{TH,Cell} < V_{CG}(0) \qquad (II) \quad \text{for} \quad V_{TH,Cell} \geq V_{CG}(0) \tag{3.23}$$

where $W(x)$ represents the Lambert W function, which is the inverse relation of $f(x) = x \cdot e^x$. The two branches complement each other to a continuous waveform. The second branch (II) is perfectly linear with respect to $V_{TH,Cell}$, since the Lambert function term reduces to a constant. The linear increase is determined by the ramp slope m only. The first branch (I), by contrast, shows a non-linear portion, since $V_{TH,Cell}$ appears in the Lambert function's argument.

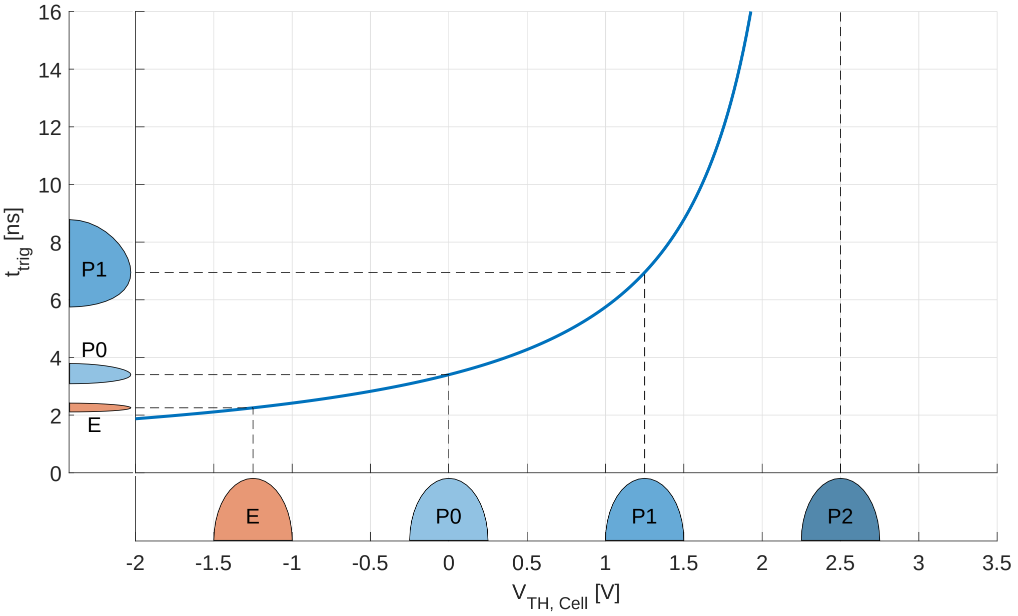
3.3.3 Comparison

The analytically derived time-domain transfer characteristic, relating the cell current to a trigger time ((3.17) and (3.23)), show considerable differences for the two CG bias conditions. For better illustration, the characteristics are plotted using a common set of typical parameters $g_{m,Cell} = 3\,\mu\text{S}$, $C_{BL} = 250\,\text{fF}$, $V_{sense} = 100\,\text{mV}$, a constant bias voltage $V_{CG,read} = 2.5\,\text{V}$ and a ramp slope $m = V_{CG,read}/6\tau_{CG}$ (compare Section 3.2.2). Figure 3.13 shows the curves of both characteristics. Each plot includes an exemplary cell threshold distribution for 2-bit per cell operation below the x-axis (colored lobes, compare Fig. 2.13). The mapped time-domain representative is depicted left of the y-axis. The mapped distributions are not normalized uniformly, they shall just show the resulting spread in time-domain.

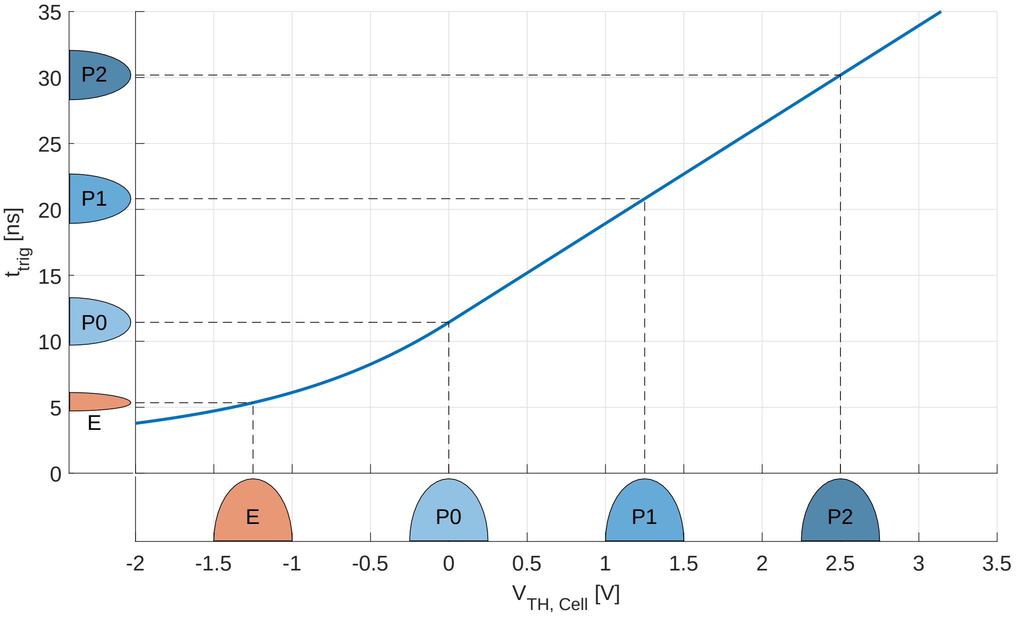
Due to the non-linear characteristic, the fixed gate biasing perfectly separates the erased state (E) from the strongly programmed state (P2). Since $V_{CG,read}$ is placed just below $V_{TH,Cell}$ of state P2, its time-domain representative theoretically is shifted to infinity, whereas the sensing of erased cells (E) is accelerated by a high overdrive. This shows its suitability for SLC operation. Intermediately programmed cells (P0 and P1), however, map to narrow spaced trigger times, which makes them prone to even small timing errors in the strobing path. Cell states close to the read level (P1, P2) on the other hand are very sensitive to threshold voltage shifts. Consequently the constant CG bias allows a fast sensing, but its robustness suffers from the non-linear transfer characteristic. To obtain a better spaced time-domain representation, the cell states could be programmed non-equidistant in the voltage domain. This however increases programming effort significantly, resulting into lower programming throughput.

A stepped gate approach benefits from the good time-domain separation of the cells with $V_{TH,Cell} < V_{CG,read}$ from the ones with $V_{TH,Cell} > V_{CG,read}$ in each of the consecutive sense operations. Yet, it cannot be as fast as an SLC implementation, since the windows in between the cell states are much closer. According to Fig. 3.13(a) the distinction between P1 and P2 would take about 10 ns and the decision between E and P2 in comparison could be finished in about 3 ns. If the CG read bias for all consecutive sense operations is chosen equivalently (0 V, 1.25 V and 2.5 V for the given example), an equal separation within an equal sense delay is achieved. This provides good sense conditions and makes the biasing the preferred one regarding sensing robustness. Programming conditions are the same as for constant biasing, since program verify is done with one bias level as well.

The ramped gate scheme maps the equidistant cell threshold distribution relatively linear to the time-domain (see Fig. 3.13(b)). In this way, it avoids the disadvantages of a constant biasing scheme. The timing distance of the low threshold cell states is stretched, which relaxes the timing requirements of the sensing path. The high threshold cell states on the other hand move closer together, consequently



(a) Time-domain transfer characteristic for fixed gate biasing.



(b) Time-domain transfer characteristic for ramped gate biasing.

Figure 3.13: Time-domain transfer characteristics for different CG bias scenarios. The colored lobes below the x-axis represent an exemplary cell threshold voltage distribution (4 cell states). The mapped time-domain representative of this distribution is depicted left of the y-axis.

decreasing sensing robustness here. However, the sensitivity to threshold shifts is much lower, since it translates linearly to the time-domain. The advantage of this biasing scheme is the parallel sensing capability, reducing the number of necessary sense operations to a single one. Even if the access time is much longer compared to a constant CG biasing scheme, it improves sensing conditions significantly and is still capable to outperform the access time of a stepped gate scheme. In addition to that, using the variable ramp slope, the mapping characteristic can be tuned to a certain extent, in order to fit particular timing needs. The flexibility in the time-domain translation also eases a precise placement of the cell states with an ISPP scheme, which allows for higher programming throughput. In Chapter 5, this aspect is covered in more detail. Altogether, the ramped gate scheme offers a good trade-off between sensing robustness and speed, making it a suitable approach for the use with MLC flash.

3.4 Summary

Three different CG biasing schemes for MLC usage have been compared regarding circuit complexity, area consumption, speed, sensing robustness and programming effort. The results are summarized in Fig. 3.14.

Constant voltage biasing can be implemented with least circuit complexity. Used with a time-domain voltage sensing, it provides robust biasing conditions for the distinction of two well-spaced cell states but has difficulties with the separation of a higher count of programmed cell states. The combination with 1T flash cells has drawbacks in the area consumption of the WL-driver, when high speed operation is required. Used with 2T flash cells it benefits from the constant bias level regarding area and speed. Due to the non-linear transfer characteristic, an exact placement of multiple cell states is hard to achieve with reasonable programming effort. It is the preferred choice only for high speed SLC flash memories.

Stepped gate control gate biasing requires more circuit effort and area, since bias changes have to reach the cells quickly. In this way, however, it provides the most favorable sensing conditions, with robust separation of adjacent cell states. The required serial sensing reduces its speed and programming requires high effort, when an exact placement is needed. For slower operation, it is best suited regarding sensing robustness.

The ramped gate approach offers a trade-off between speed and accuracy. It has highest circuit complexity and area requirement compared to the other schemes. With the dynamic voltage ramp, it requires only one sense operation and hence reduces the precharge time overhead. It is faster and it translates the cells threshold voltages linearly to the time-domain, which increases sensing robustness. Programming effort also benefits from the linear transfer characteristic.

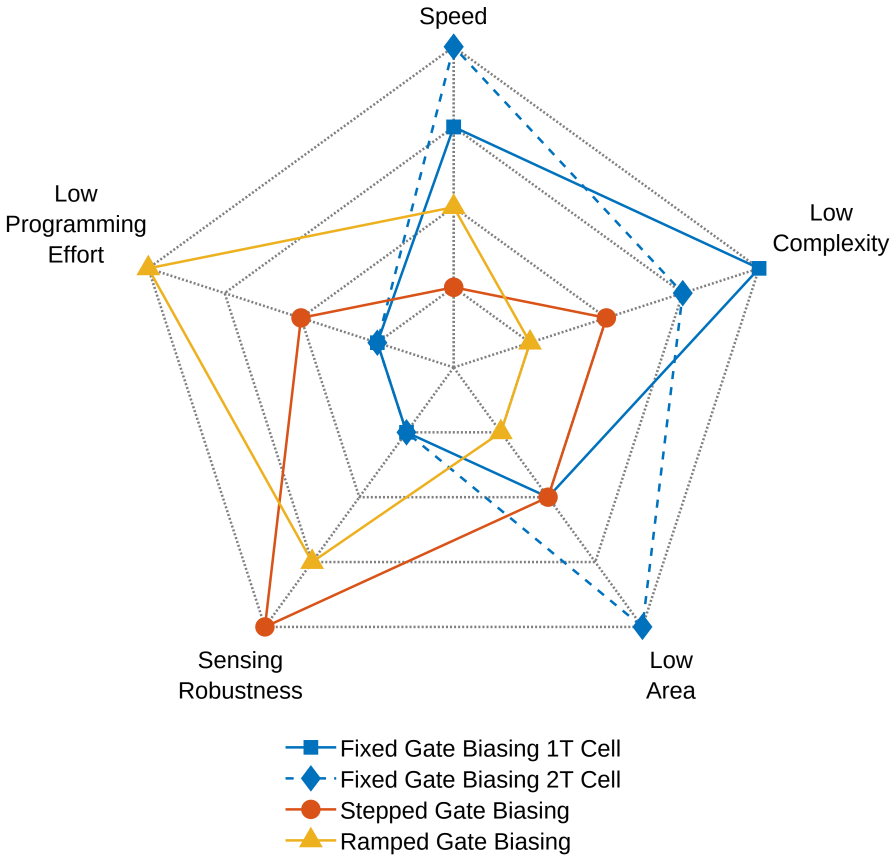


Figure 3.14: Comparison of the different cell biasing schemes.

According to the findings of the comparison, the ramped gate biasing is chosen for this work. In the next chapter the development of a new time-domain voltage sensing scheme for MLC operation is described. It is based of the ramped gate cell biasing.

Chapter 4

Time-Domain Ramped Gate Sensing for MLC Operation

This chapter describes the ramped gate sensing scheme developed in the scope of this work. It comprises a new voltage mode sense amplifier with an improved slope detection. The design allows for offset compensation and it includes precharge acceleration mechanisms, which enable the sense amplifier for high speed operation and precise sensing for MLC operation. Further, the ramp generation circuitry with fast recovery from discharge is introduced. It implements the ramped gate biasing scheme. The strobing concept is explained and an overview over the whole sensing architecture is given. Measurement results from the fabricated test chip are shown at the end of this chapter.

4.1 Time-Domain Sense Amplifier with Offset Compensated Slope Detection

The ramped gate control gate biasing as described in Chapter 3 requires a static comparator to generate the trigger events from the cell information (cell current or bitline voltage). In [50] a time-domain voltage sensing scheme is introduced, which uses a corresponding sense amplifier. Compared to a conventional voltage comparator realized as single ended OTA, it has a very compact, power saving design and at the same time offers an input voltage range down to V_{SS} ($= 0\,\mathrm{V}$). Figure 4.1 shows the structure of the sense amplifier realized as a common-gate amplifier biased with an adaptive current source P2 that is controlled by a slope detection circuit. The sensing voltage V_{sense} is defined as

$$V_{sense} = V_{Ngate} - V_{TH,N1} - V_{OD,N1} \tag{4.1}$$

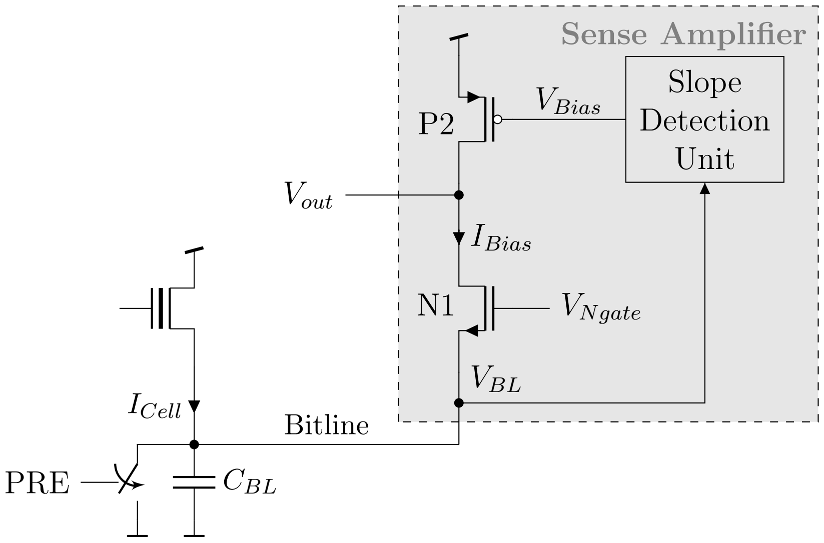


Figure 4.1: Common gate sense amplifier with slope detection unit [50].

with the overdrive of transistor N1 $V_{OD,N1}$. As soon as the bitline voltage V_{BL} reaches V_{sense}, N1 changes from linear to saturation region and thus decouples the amplifier's output (V_{out}) from its input node (V_{BL}), which is connected to the large bitline capacitance (C_{BL}). A further rising bitline voltage puts N1 in the off-state and the output can be charged quickly by the bias current provided by P2, since the capacitive load at the output is low.

The drawback of the common-gate amplifier structure is its bias current I_{bias} charging up the BL in parallel to the actual cell current I_{Cell}. As a result, the relative cell current window is diminished to (compare (3.1))

$$W^*_{I,rel} = \frac{I_{Cell,LVT} - I_{Cell,HVT}}{I_{Cell,LVT} + I_{bias}}. \tag{4.2}$$

The difference between the cell currents in the numerator stays the same, as I_{bias} adds to both currents. The absolute value of the high current value however rises, resulting in a smaller value of the fraction. The slope detection is capable to adapt I_{bias} depending on the voltage slope at the input of the amplifier. This allows to keep the constant portion of the bias current low unless a voltage slope is detected at the BL. That means, in case of a programmed cell (low voltage slope at the BL), the amplifier's bias current is reduced to its minimum, to keep transistor N1 at its operation point. In case of an erased cell (high voltage slope at the BL), the amplifier is accelerated by a boosted bias current that additionally speeds up the charging of the bitline. The sense amplifier is also suitable for MLC operation with a ramped control gate voltage. However, the I_{bias} has to be minimized in order to maintain the reading window.

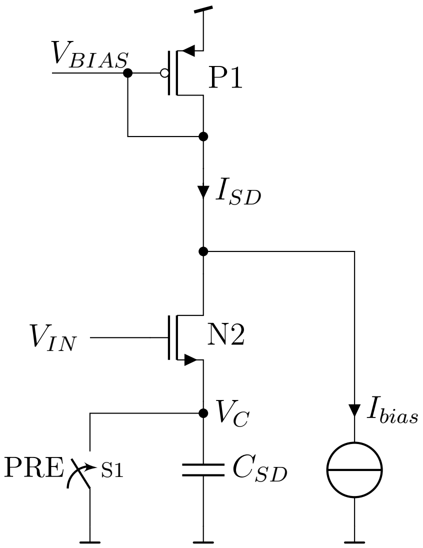


Figure 4.2: Voltage-biased slope detection (VBSD) circuit[50].

4.1.1 Voltage-Biased Slope Detection vs. Current-Biased slope detection

The main part of the slope detection described in [50] can be seen in Fig. 4.2. It is a current source dependent on the input slope $\partial V_{IN}/\partial t$, consisting of an NMOS source follower N2 connected to a capacitor C_{SD} at the source node V_C and a diode-connected PMOS transistor P1 between the supply voltage and the drain contact of N2. P1 acts as load for N2 and as input transistor of the current mirror formed together with P2 of the common-gate amplifier (see Fig. 4.1). In addition to that, a constant bias current source I_{bias} is connected to P1. It provides the slope-independent portion of the bias current that is needed to keep the amplifier biased, even if no voltage slope is detected at the bitline. The switch S1 is needed to discharge the capacitor in order to reset the circuit to a well-defined starting condition ($= V_{SS}$) before each read operation. It is activated by the signal PRE that also discharges the selected bitline (see Fig. 4.1).

The slope-dependent current is generated by the source follower N2 keeping its source voltage V_C at a value of about $V_{IN} - V_{TH}$. In this way, the differential component of the input voltage is copied to the capacitor which then translates the voltage's rate of change proportionally into a current. It is the opposite procedure of the cell current translating into the bitline voltage by the bitline capacitance (see (3.12)) and can in first order approximation be described by

$$I_{SD} = \frac{\partial V_C}{\partial t} \cdot C_{SD} + I_{bias} \approx \frac{\partial V_{IN}}{\partial t} \cdot C_{SD} + I_{bias}. \tag{4.3}$$

To ensure proper functionality, the input voltage level V_{IN} has to stay above the threshold voltage of N2. The bitline starts at V_{SS} and hence cannot directly be connected to the slope detection's input. It first has to be raised by a PMOS source follower acting as a level shifter, which is not depicted in Fig. 4.2. Using its small signal equivalent circuit, the transfer function of the slope detection can be determined as

$$I_{SD}(s) = g_{m,N2}\left(\frac{s}{s+1/\tau_{SD}}\right)\cdot v_{in}(s). \tag{4.4}$$

The time constant τ_{SD} determines the settling speed of the slope detection. It calculates as

$$\tau_{SD} = \frac{C_{SD}}{g_{m,N2}+g_{mb,N2}}. \tag{4.5}$$

An application of a voltage ramp $V_{IN}(t) = V_{IN,0} + t \cdot S$, with the ramp slope S, would then lead to the following transient response [50]:

$$\begin{aligned} i_{SD}(t) \quad &= \quad SC_{SD}\frac{g_{m,N2}}{g_{m,N2}+g_{mb,N2}}\left(1-\exp\left(-\frac{t}{\tau_{SD}}\right)\right)+I_{bias} \\ &+ \quad i_{SD,0}\cdot\exp\left(-\frac{t}{\tau_{SD}}\right). \end{aligned} \tag{4.6}$$

The first part of the equation yields the slope-dependent current portion, which matches the time-varying part of the approximation in (4.3), except for the attenuation caused by the body-effect of N2. The second part accounts for the initial bias condition of N2, which is determined by the starting input voltage level:

$$i_{SD,0} = I_{D,N2}\left(V_{IN,0}\right). \tag{4.7}$$

Consequently the circuit is referred to as voltage-biased slope detection (VBSD). The current $i_{SD,0}$ offsets the amplifier's bias current at the start of the sensing phase and thus accumulates additional charge on the bitline. This effect is unwanted, since it further diminishes the reading window. A simulation of this initial behavior at two different input voltage slopes can be seen in Fig. 4.3. The ideal transient response of the slope detection would be an instantaneous step right to the final value at

$$\underset{t\to\infty}{i_{SD}}(S) = SC_{CD}\frac{g_{m,N2}}{g_{m,N2}+g_{mb,N2}}+I_{bias}. \tag{4.8}$$

The area above this line makes up the unwanted additional charge that is copied to the bitline via the current mirror P1/P2. The smaller the input voltage slope S, the larger is the area. Therefore, it affects (speeds up) mostly the sensing of strongly programmed cells. The diagram also reveals that the small signal model

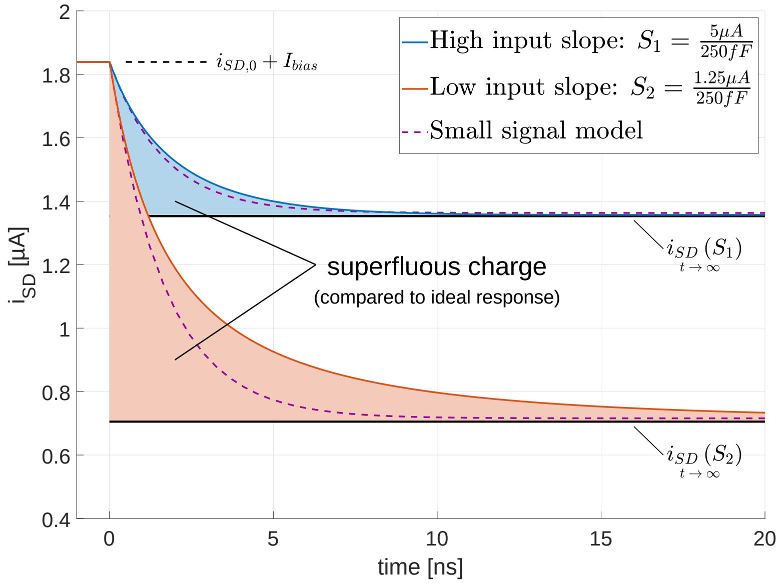


Figure 4.3: Transient response of the voltage-biased slope detection circuit with $V_{IN,0} = 400\,\text{mV}$, $I_{bias} = 500\,\text{nA}$.

matches the simulation only at higher input slopes. With low input slopes the current through N2 (and thus its transconductance $g_{m,N2}$) drops quickly leading to a dynamically rising time constant τ_{SD}. As a consequence, the settling time is underestimated.

By reducing $V_{IN,0}$ the slope-independent initial current could be lowered. However, this would also lower the initial g_m of transistor N2 and hence slow down the whole operation considerably, since both current portions settle with the same time constant. Process variations do not allow to place $V_{IN,0}$ very precise, since it is determined by the threshold value of the preceding PMOS source follower performing the level shift. Actually, a PMOS threshold would have to be matched to an NMOS threshold which requires an offset compensation technique. A compromise has to be found fulfilling both contradictory objectives, a fast response of the slope detection and a low initial offset current disturbing the sense operation.

To maximize the read window for an MLC operation, the focus is put on the reduction of the offset current. For this reason, a new scheme to bias the slope detection was developed as part of this work. The new biasing scheme is current-based and allows for offset compensation (Fig. 4.4). By omitting the switch and moving the connection of the constant bias current I_{bias}, the discussed voltage offset gets sampled at C_{SD} and only the AC part of the input voltage is translated into an increase of I_{SD}. Hence the undesired current spike at the startup of the slope detection mechanism vanishes completely. As long as the initial input voltage

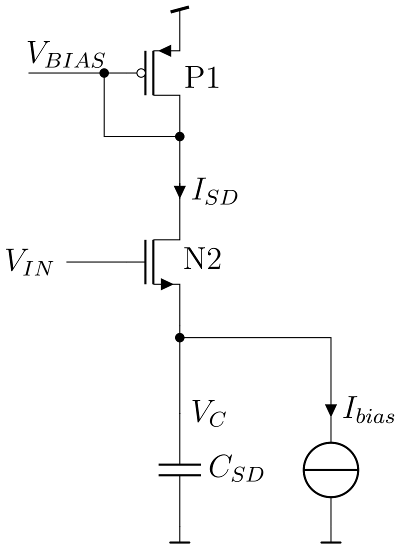


Figure 4.4: Current-biased slope detection (CBSD) circuit.

$V_{IN,0}$ is chosen high enough to leave some voltage headroom for the current source, the operation point of N2 now is controlled by its constant bias current. Hence, the circuit it is referred to as current-biased slope detection (CBSD). After each read operation, when the bitline is reset to the starting point (V_{SS}), the slope detection capacitor C_{SD} is discharged by the current source and the circuit returns to its initial operation point automatically.

The small signal analysis of the new biasing scheme yields the same transfer function (see (4.4)), the transient response though lacks the decaying offset current from the voltage-biased version of the slope detection circuit. So only the differential slope-dependent portion is remaining which is described by

$$i_{SD}(t) = SC_{SD}\frac{g_{m,N2}}{g_{m,N2}+g_{mb,N2}}\left(1-\exp\left(-\frac{t}{\tau_{SD}}\right)\right)+I_{bias}. \tag{4.9}$$

Hence, the small signal parameters are defined more conveniently by the constant bias current I_{bias}, which is not dependent on the matching of threshold voltages of two complementary transistors. However, also with this scheme, the small signal equivalent circuit cannot reproduce the actual behavior of the circuit at all input slopes. In this case, the settling duration is overestimated at high input slopes. This can be identified by the transient simulation result of the current-biased slope detection, which is depicted in Fig. 4.5. It has the same reason as explained for the voltage-based biasing; a dynamically changing time constant τ_{SD} due to the quickly rising current through the source follower.

The current finally reaches the same value $i_{SD,t\rightarrow\infty}$, but it approaches always

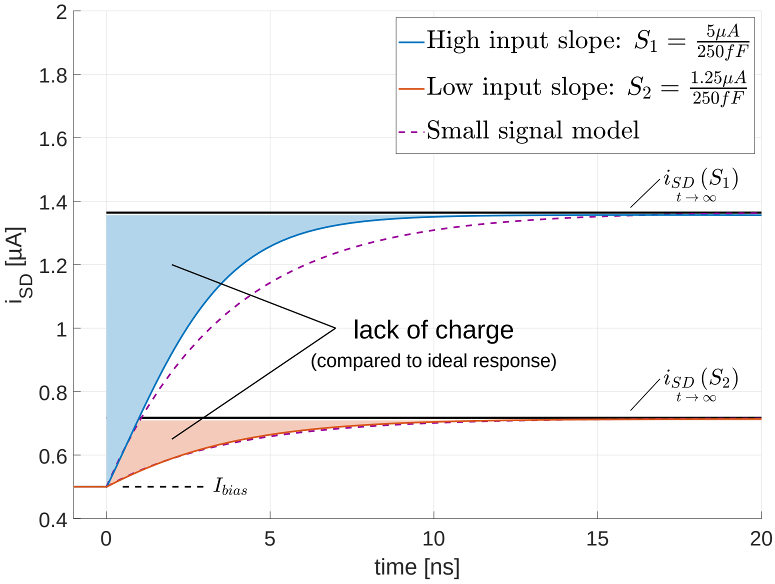


Figure 4.5: Transient response of the current-biased slope detection circuit with $V_{IN,0} = 400\,\mathrm{mV}$, $I_{bias} = 500\,\mathrm{nA}$.

from below. This results in a theoretical lack of charge transferred to the bitline when compared to an ideal slope-dependent current source (instantaneous step from I_{bias} to $i_{SD,t\to\infty}$). The lack has its minimum at zero input slope $S = 0$ and rises together with an increasing input slope. So, it affects (slows down) the sensing of highly conductive cells (with high slope), but not the sensing of programmed cells (with low slope).

The comparison of both transient responses shows that the voltage-based biasing can be more suitable for detection of higher bitline slopes than the current-based biasing scheme. The latter one puts the source follower in a better controlled operating point and allows an offset compensation. Since the slope-dependent current can only approach the final value from below, it will be slower for steep input slopes, but much more accurate for very low slopes. In addition to that, its compensation cancels out the threshold variation of the source follower N2 and the preceding PMOS level shifter.

This improvement can be shown with a simulation of slope detection unit attached to the sense amplifier as seen in Fig. 4.6. The level shift of the bitline voltage for simplicity reasons is done with an ideal dc voltage source ($V_{IN,0}$). In a transient simulation, the bitline capacitance C_{BL} first is discharged before the sensing phase is started by opening the precharge switch. Then a constant cell current I_{cell} charges the bitline until the sense amplifier output flips high and stops the time measurement. In order to assess the read window improvement,

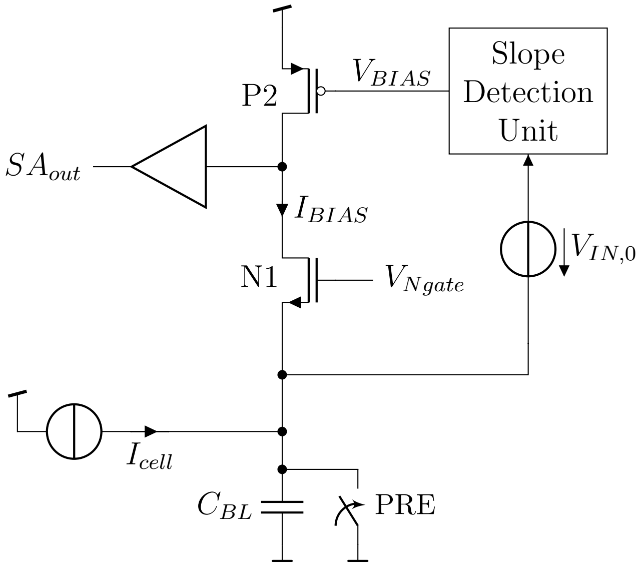


Figure 4.6: Test bench for the slope detection circuit.

I_{cell} was set to 10 µA representing a weakly erased cell, to 1 µA accounting for a considerably high leakage current on the bitline and to 0 µA for the ideal case of a fully programmed cell.

Figure 4.7 shows the simulation results of the comparison of both biasing schemes. The timing window between programmed and weakly erased cells depends strongly on $V_{IN,0}$ using the VBSD scheme. Within the selected sweep range, it drops by about 30%. The newly proposed CBSD scheme eliminates this dependency by offset compensation. It keeps the timing window widely open and hence is more suitable for precise distinction of multiple cell states as it is needed for MLC operation.

Another simulation shows the effect of the slope detection's biasing on the whole sensing scheme, including a ramped gate cell biasing. For this purpose the sense amplifier test bench of Fig. 4.6 is extended by the PMOS source follower replacing $V_{IN,0}$, the constant cell current is replaced by a linear cell model (according to Fig. 3.12) and an ideal voltage ramp is applied to its CG terminal. A transient Monte Carlo simulation of the sensing operation with a sweep over the cell threshold then yields the transfer characteristics shown in Fig. 4.8. Both curves show a linear translation of $V_{TH,Cell}$ to the time domain in an intermediate range before they saturate at the high cell thresholds. This happens, since the sense amplifier's bias current I_{bias} charges the BL up to the trigger point V_{sense} before the voltage ramp reaches the cell's threshold. So, the cell current stays zero and t_{trig} only depends on the sense amplifier's characteristic itself. Saturation starts earlier using the VBSD, since the initial settling of the slope detection

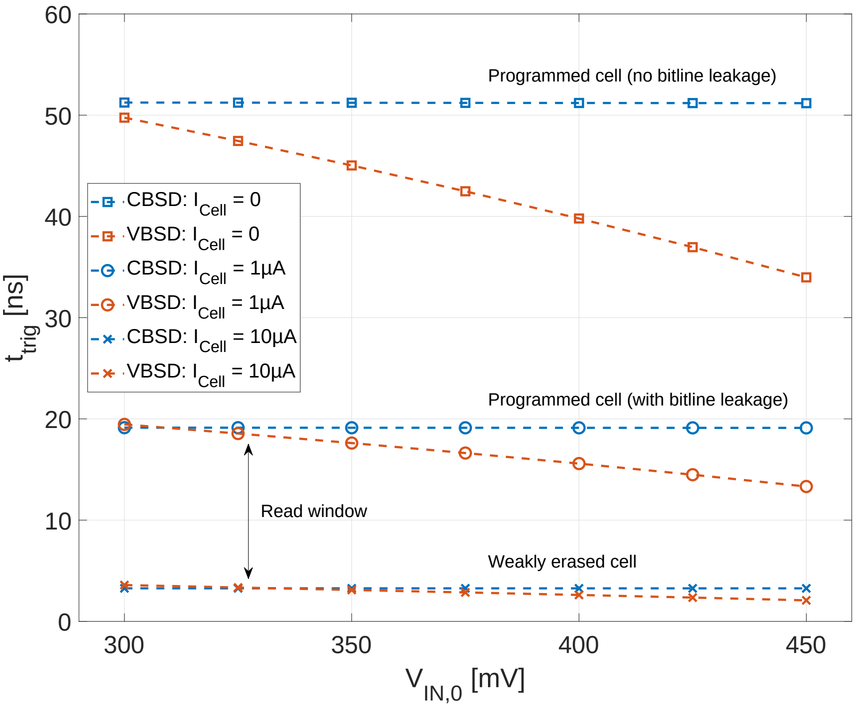


Figure 4.7: Simulation of the common-gate sense amplifier with both slope detection schemes, with $C_{BL} = 200\,\text{fF}$, $V_{sense} = 100\,\text{mV}$, $C_{SD} = 50\,\text{fF}$, $I_{bias} = 500\,\text{nA}$.

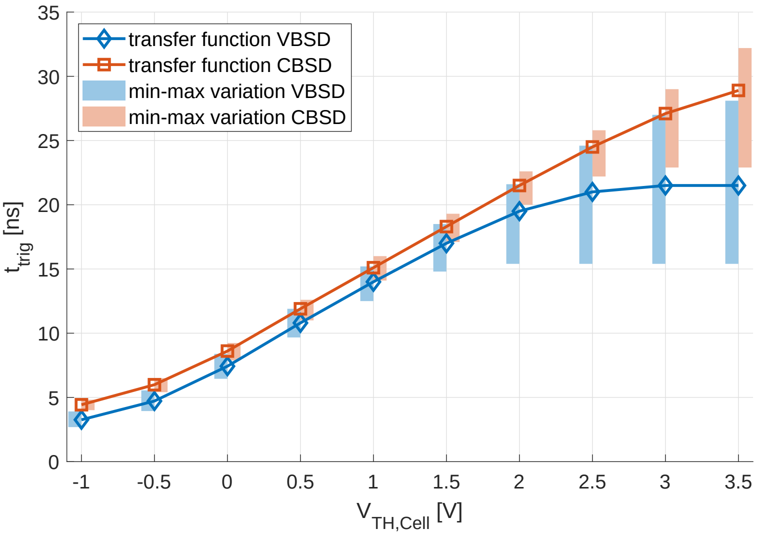


Figure 4.8: Simulated ramped gate time-domain transfer function of the common-gate sense amplifier with both slope detection schemes.

circuit accumulates more charge on the BL and hence speeds up the self-charging process. CBSD cannot prevent this effect, but it minimizes the self-charging and therefore prolongs the linear characteristic to higher cell thresholds by about 1 V. In this way, the new scheme enlarges the timing range, usable for placement of the intermediate cell states (P0, P1), by over 30 % [54]. The impact of the offset cancellation capability can be observed looking at the trigger time variation. CBSD reduces the time-domain spread by about one third compared to VBSD. Both schemes show a dramatic increase of variation in the upper range. It can be explained by the amplifying transistor N1 being operated with very low current in the moment of transition from low to high. In this operation point, its threshold variation almost directly translates into the sense amplifiers input voltage offset and impact on t_{trig} is high. The worst case is reached when there is no cell current at all, and the slope detection fully settled supplying the lowest I_{bias} for N1.

4.1.2 Precharge Acceleration Mechanisms

For accurate sensing it is crucial to start every sensing operation with identical operating conditions. Consequently, the precharge time has to be chosen long enough for all nodes to settle in their starting point. This time shall be minimized to achieve a fast read access. According to Section 4.1.1, the CBSD scheme returns to its initial operation point automatically once the bitline is precharged to V_{SS}. This however happens slowly, since the discharging of C_{SD} is limited by two factors. In the first place, the negative slew rate of the NMOS source follower's output limits the process. Calculated as

$$SR_{SD\downarrow} = \frac{I_{bias}}{C_{SD}} \, , \tag{4.10}$$

it is determined by the constant bias current which is required to be as small as possible. Secondly, the settling behavior adds a considerable time overhead prolonging the needed precharge pulse length. The time constant τ_{SD} (4.5) suffers from a low transconductance of N2 and hence is also not favored by a low bias current.

Common offset cancellation schemes, sampling SA variations on capacitors, either put the capacitors into the feedback path of the amplifier during an explicit offset nulling phase [24] or connect them with the diode-connected input transistors of the input current mirror [55] to keep the timing overhead low. Others fully eliminate the overhead by a digital current offset cancellation setting up at the system's initialization [23]. The structure of the slope detection circuit however does not allow a similar approach. Hence alternative circuit solutions were investigated and a mechanism accelerating the precharge process was developed within this work.

Capacitor Swapping

The VBSD scheme does not suffer from a long precharge duration since it discharges the slope detection capacitance very quickly by simply shorting it to V_{SS}. Though this is very effective, it eliminates the offset cancellation capability. With a very brief activation of the shorting switch just at the beginning of the precharge phase, the settling process could be turned around (charge up the capacitor, instead of discharging it) accelerating it considerably. Figure 4.9 shows an approach using such short discharge pulses in combination with a copy of the slope detection circuit precharging in an interleaved manner.

During each read operation one of the capacitors is actively used for slope detection. The other one simultaneously gets precharged by a replica branch, reproducing the same precharge conditions as in the active branch. In between two consecutive read operations the active capacitor is discharged by a very short pulse (DIS) and swapped with the other capacitor. With this procedure, the time available for returning to the slope detection's initial operation point effectively gets prolonged by a full read cycle. Further it now approaches the operating point from below, which results in a faster settling due to the big initial overdrive of the NMOS source follower (N2 and N2'). The solution however needs two additional control signals increasing the sense amplifier's complexity. To avoid charge losses, swapping the capacitors would even require a non-overlapped clock-

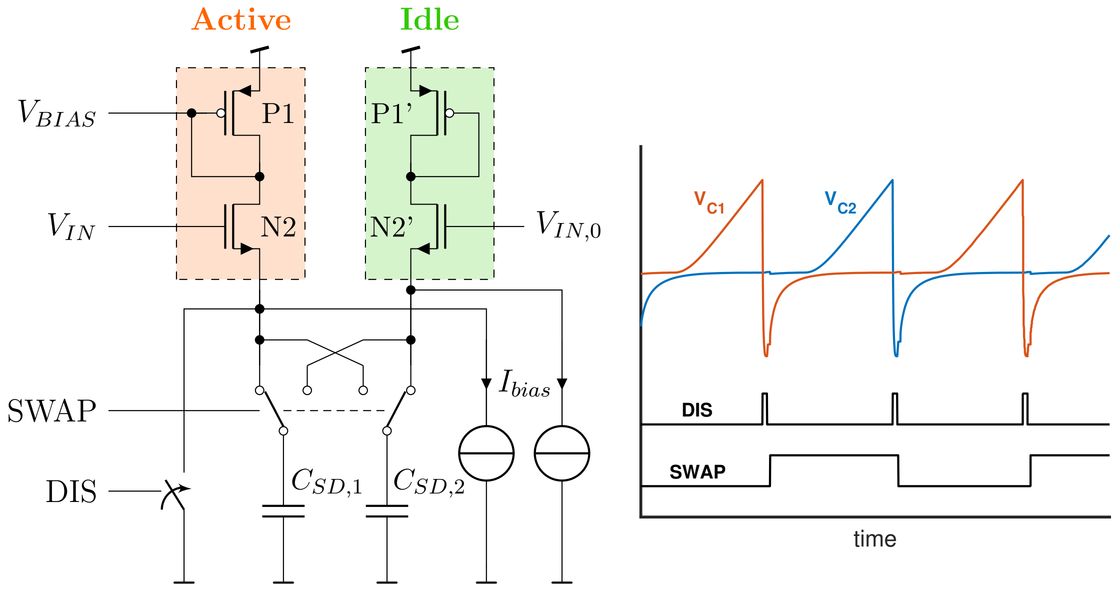


Figure 4.9: Schematic and signal waveforms of a capacitor swapping approach with interleaved precharging.

ing scheme which has to be generated locally due to the distributed nature of the sense amplifiers. Another issue is supply noise generated by the discharge pulses locally injecting considerable amounts of charge into the V_{SS} net followed by quick charging of the capacitors.

Gm-Boost

Another approach, which avoids the switching noise is depicted in Fig. 4.10. It uses n-1 replicas of the slope detection branch, which are connected in parallel during the precharge phase. In this way, it does not change the operation point voltage at C_{SD}, and at the same time, it accelerates the precharge process in two respects. On one hand it increases the discharge current for C_{SD} by connecting n bias current sources in parallel. As a consequence, the slew rate at the capacitor node $S_{SD\downarrow}$ increases by a factor of n. On the other hand, it diminishes the time constant that limits the settling behavior. Both transistors of the replica branches are sized equally as P1 and N2. In case of perfect matching, the new time constant results in $\tau'_{SD} = \tau_{SD}/n$. This effectively corresponds to a temporary boost of the g_m of the NMOS source follower. With appropriate sizing, the parasitic capacitance added to the integration node is negligible compared to C_{SD} and hence does not affect

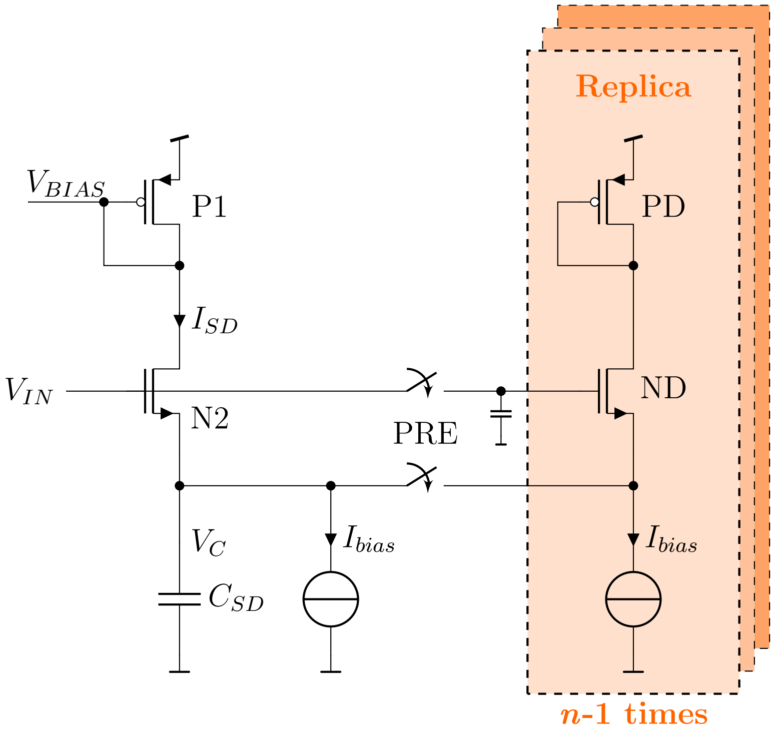


Figure 4.10: Schematic of the gm-boost approach.

the settling behavior. At the input node V_{IN}, the situation is different. Here, an additional capacitance would slow down the slope detection's dynamics. Hence, the input gates of the replica branches are disconnected during sensing phase. To keep the voltage at this node constant in between the precharge phases, a small capacitance is added here. It reduces the effect of charge injection and leakage currents on the voltage of this floating node.

This approach can accelerate the precharge process substantially. However, it leads to a considerable overhead in current consumption and area, which limits the number n of replica branches.

Local Precharge Scheme

In case an erased cell is sensed, the bitline voltage rises fast and the voltage accross the slope detection capacitance C_{SD} saturates quickly at its maximum level ($\sim V_{DD} - V_{TH}$). This is the worst starting condition for the subsequent precharge process, since the voltage difference to its precharged state is at the maximum. A temporary bias current raise as used in the approach described above can mitigate this problem only with considerable circuit overhead. Keeping the voltage swing at the capacitor's node small in the first place can reduce the precharge time much better. It reduces the amount of charge needed to be discharged and therefore minimizes the slewing time.

In normal operation, all sense amplifiers pass through precharge and sensing phase simultaneously. Since the time-domain sensing concept does not work with a fixed timing, but rather translates the cell information into a continuous time information, each sense amplifier finishes its sensing (i.e. its output flips from low to high) at its individual point time. As long as the amplifiers keep their output value, they can locally be put into precharge state directly after they finished their sensing. Figure 4.11 shows the block diagram of an approach using this local precharge concept. Within a feedback loop the sense amplifier resets the input of its slope detection to V_{SS} once its output goes high. A small delay stage ensures enough time for internal nodes to stabilize before the local precharge state is entered. If the bitline voltage does not reach V_{sense} within the sensing period (i.e. a fully programmed cell is selected) the next global precharge pulse ($\mathrm{PRE_{BL}}$) brings the sense amplifier to its precharge state.

Since the voltage at the slope detection capacitor V_C follows the bitline voltage V_{BL} differentially, this local precharge mechanism effectively limits the internal voltage swing to

$$\Delta V_C = V_{sense} + Delay \cdot \frac{dV_{BL}}{dt}. \tag{4.11}$$

Further it prolongs the time available for precharge of the slope detection depending on the state of the connected cell. Sense amplifiers connected to erased cells

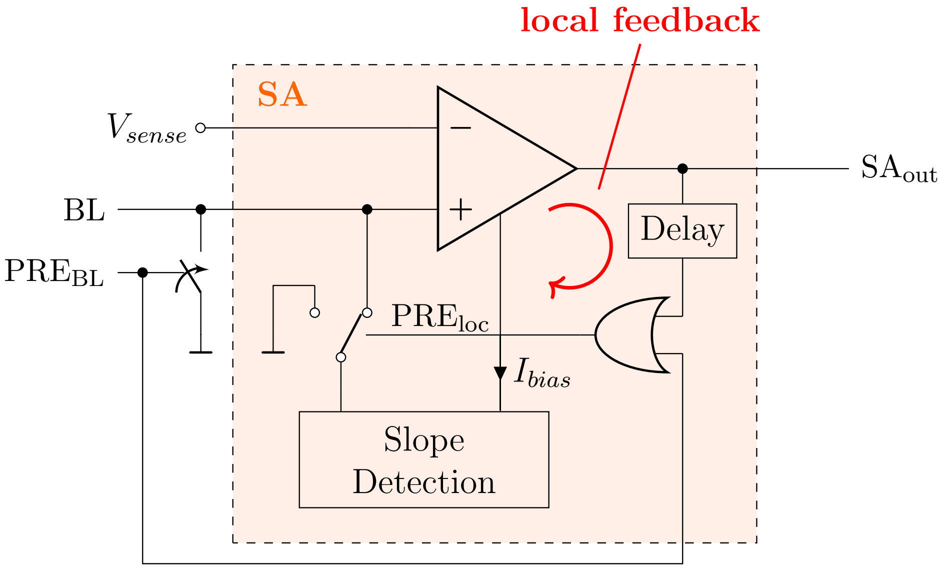


Figure 4.11: Block diagram of the local precharge scheme. The delayed sense amplifier output triggers the precharge of the slope detection circuit before the global precharge signal ($\mathrm{PRE_{BL}}$) starts the subsequent read operation.

enter the precharge state very early. Programmed cells let them enter very late. This however is not a drawback, since internal nodes do not charge quickly, if a programmed cell is sensed. Consequently, their difference to the initial operating point is small and the precharge process does not take a long time.

4.1.3 Sense Amplifier Circuit

As discussed in Section 4.1.1, the proposed new slope detection biasing scheme is capable to improve the selectivity of the common-gate sense amplifier introduced in [50]. However, it needs modifications to speed up its precharge process. Approaches addressing this issue were introduced in Section 4.1.2. In consideration of the findings, both parts, the sense amplifier with adapted slope detection and the precharge acceleration measures, are combined into the new time-domain sense amplifier with offset-compensated slope detection. Figure 4.12 shows the schematic of the resulting design. It deploys CBSD by connecting transistor N3 at the source terminal of the source follower N2 instead of its drain node. In addition to that, two of the three precharge acceleration schemes presented in the previous subsection are incorporated. Firstly, the gm-boost scheme is added, which consists of the replica structure P1', N2' and N3'. It is connected to the slope detection path only during precharge phase and speeds up both the discharging of C_{SD} through N3 and the settling of the source follower. Secondly, the

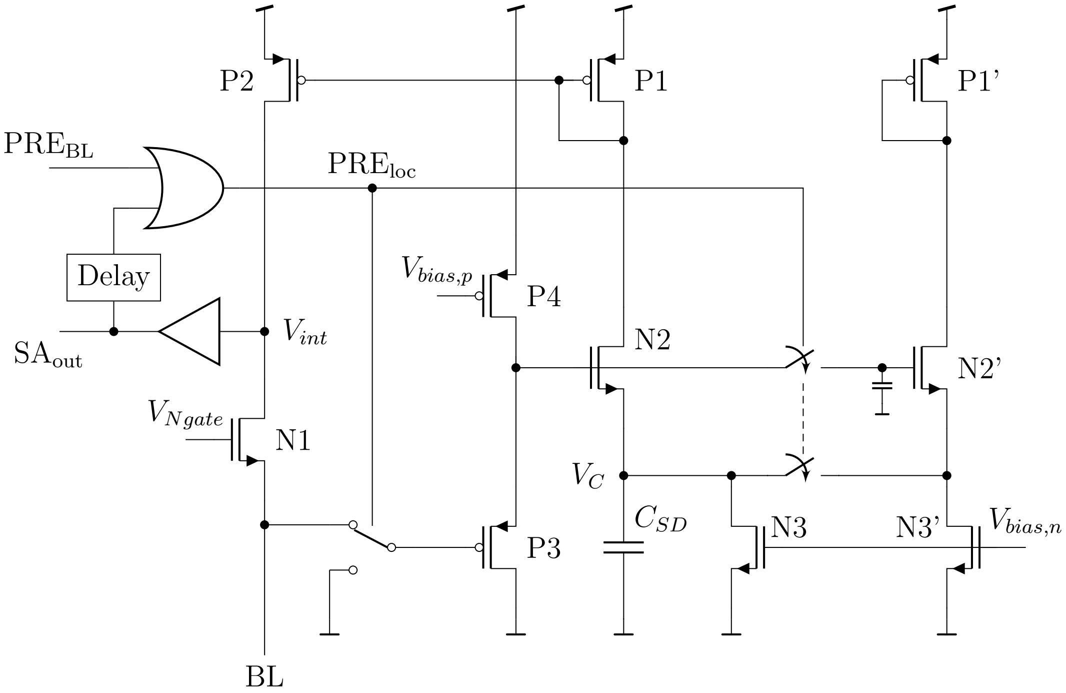


Figure 4.12: Sense amplifier with current-biased slope detection and precharge acceleration mechanisms.

local precharge scheme is integrated. It uses a delayed version of the digital sense amplifier output to start its precharge process right after it finished sensing. The capacitor swapping was not incorporated, as increased complexity and generated noise on the supply outweighed its advantage.

The sense amplifier is simulated for a burst read, to illustrate the interaction between sense amplifier, slope detection and precharge acceleration mechanisms. A burst read is a sequence of consecutively triggered read operations with a fixed frequency. In each cycle, data from the previous read is fetched and a new read operation is issued directly. This poses the most challenging requirements to sense amplifier and precharge circuits, since there is no time to settle in between the reads. The general timing of a burst read sequence is shown figure Fig. 4.13. A read cycle, started with a short read trigger pulse, is divided into precharge phase, sensing phase and a short time slot for fetch of read data. The sense amplifier has to finish before the end of sensing phase, or wrong data will be fetched. Within the duration of the next precharge phase it has to return to its initial operation point and be ready for the next sensing.

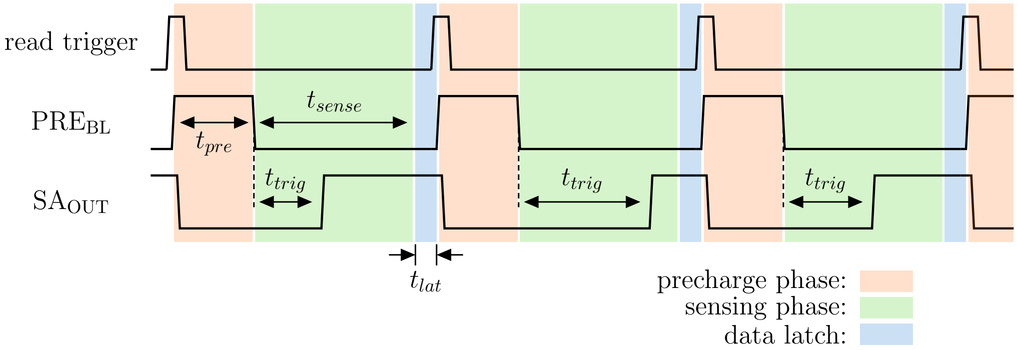


Figure 4.13: Timing of a burst read sequence. The trigger time t_{trig} depends on the read cell state.

The improvements by the newly introduced circuit parts are well visible within a sequence alternating between read out of an erased cell (E) and an intermediately programmed cell (P1). Simulated curves of relevant internal sense amplifier nodes during a corresponding burst read are shown in Fig. 4.14. The subplots show waveforms of three node voltages: The bitline voltage V_{BL}, which gets discharged to V_{SS} during the precharge phase (signaled by PRE$_{\mathrm{BL}}$) and which then is released for integration of the cell current. The voltage at the amplifier's unbuffered output V_{int}, which follows the bitline voltage for low values and charges up to V_{DD} once the sensing threshold V_{sense} is reached. And the voltage V_C across the slope detection's capacitor C_{SD}. It rises together with the bitline voltage during the sensing phase and gets reset to its idle value $V_{C,\infty}$ (highlighted by the dashed line) during precharge. Each subplot shows the simulated curves of a burst read, executed for any combination of both accelerating mechanism either activated or disabled.

The first scenario, shown in the uppermost subplot, has both acceleration mechanisms switched off. The precharge time is too short for the slope detection capacitor to reset to its proper starting voltage. Its voltage V_C consequently accumulates from cycle to cycle depending on the state of previously read cells. As a result, the following sense operations in the burst are slowed down, since the slope detection is taking effect considerably delayed. The timing penalty affects the sensing of erased cells only marginally. Intermediately programmed cells however considerably suffer from the delay and can be read as strongly programmed. This can be seen in the third read operation as the (unbuffered) SA output V_{int} barely reaches the trigger point before the fourth read is started.

For the second scenario, the feedback path of the local precharge mechanism is activated. It extends the time available for discharge of C_{SD} and limits the ripple of V_C. The sense amplifier's precharge time now is not constant anymore.

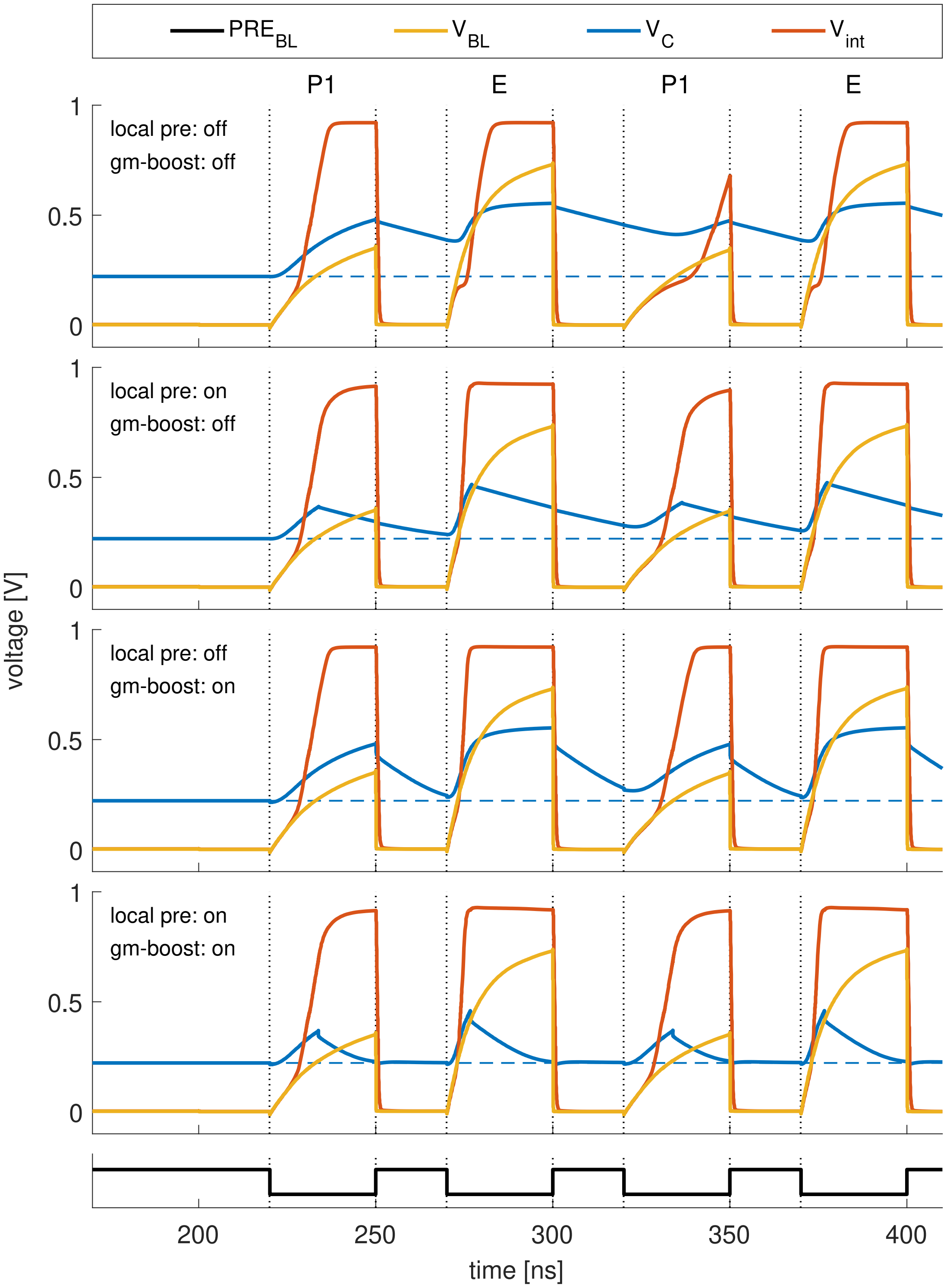


Figure 4.14: Simulation of relevant sense amplifier nodes during a read burst alternating between an erased (E) and an intermediately programmed (P1) cell. Both precharge acceleration mechanisms are toggled on or off.

It depends on the content of the previously read cell. Sensing of an erased cell is followed by a long precharge period; a programmed cell leaves less time bonus for the subsequent precharge phase. The settling of the slope detection works better, and the delaying effect is reduced. Due to its low bias current, the sense amplifier still reaches its final operating point only very slowly and the achievable burst frequency is limited.

The third subplot shows the curves for activated gm-boost without local precharge scheme. The temporary bias current increase during precharge clearly speeds up discharging and settling behavior. So, the necessary precharge duration is reduced. As in the first scenario, the worst-case starting point of the discharge of C_{SD} is reached when reading an erased cell. V_C then approaches its maximum at about $V_{DD} - V_{TH}$, which has to be ramped down the following precharge phase. The minimum precharge pulse duration with activated gm-boost then approximately is determined by

$$t_{pre,min} = \frac{V_{DD} - V_{TH} - V_{C,\infty}}{nI_{bias}} C_{SD}, \tag{4.12}$$

where $V_{C,\infty}$ is the value that V_C finally settles at and n is the factor of the gm-boost (see Section 4.1.2). The sudden voltage jumps apparent in the plot are caused by charge sharing when the dummy branches of the gm-boost circuit are connected to C_{SD}.

The last subplot shows the simulated read burst with both precharge acceleration mechanisms activated. Precharge time is extended and capacitor discharge is boosted with increased bias current. V_C reaches its idle state long before the sensing phase starts, and no delaying effect can be observed at the common-gate amplifier's output V_{int}. Both mechanisms combined hence give room for improvement in burst frequency.

To analyze the achieved improvement more in detail, the burst read simulation was conducted for various burst frequencies f_{burst} and different precharge times t_{pre}. The settling error e_s at the positive node of the slope detection capacitor is taken for assessment of a properly precharged slope detection circuit. It is determined as

$$e_s = \frac{V_C - V_{C,\infty}}{V_{C,\infty}} \tag{4.13}$$

evaluated at the end of the precharge phase that follows the sensing of an erased cell within the read burst. Figure 4.15 shows the simulated settling error for a burst read with both acceleration mechanisms switched off. At precharge times of 75 ns and less, the settling error increases significantly and hence sensing is distorted as

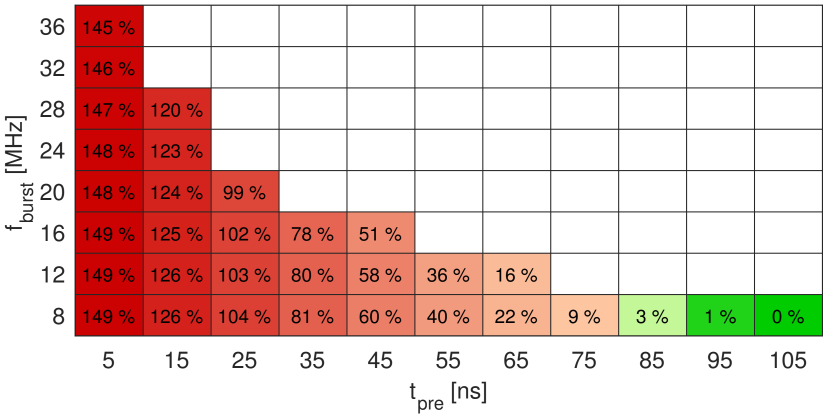


Figure 4.15: Settling error, measured in a burst read simulation with both precharge acceleration mechanisms switched off.

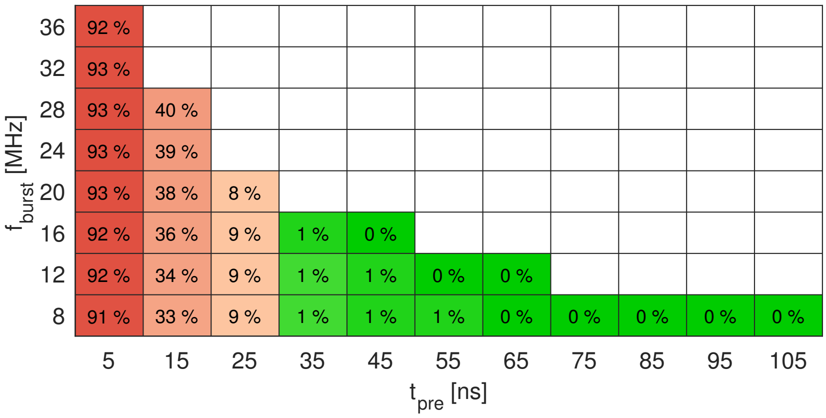
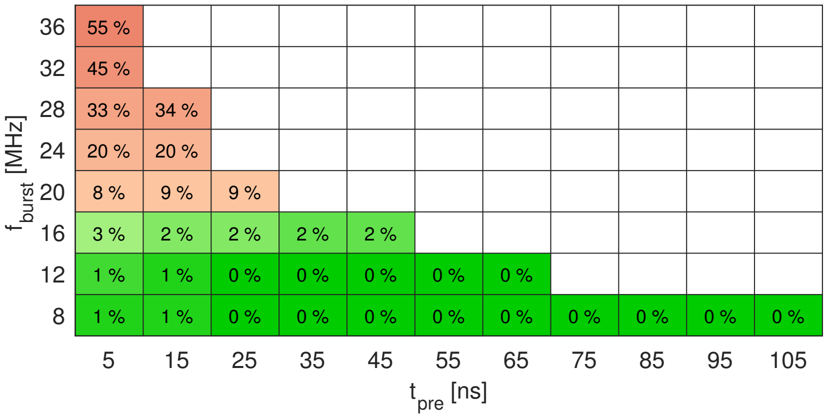


Figure 4.16: Settling error, measured in a burst read simulation with activted gm-boost.

shown in Fig. 4.14. To leave enough time for the sensing phase, the burst frequency cannot be increased above 8 MHz. The simulation results for the same burst read with activated gm-boost are shown in Fig. 4.16. Here the settling error stays low for precharge times above 25 ns. This reduction by a factor three reflects a boost factor $n = 3$ of the implemented gm-boost mechanism. Consequently, the burst frequency can be increased up to 16 MHz without noticeable distortion. Both results reveal that only the precharge duration affects the settling progress of the slope detection. Different burst frequencies do not change the resulting settling error a lot. This is because in both cases the time available for discharge of C_{SD} is only determined by t_{pre}. Also the worst case starting point $V_C \approx V_{DD} - V_{TH}$ is independent of the burst frequency. This is different when the local precharge



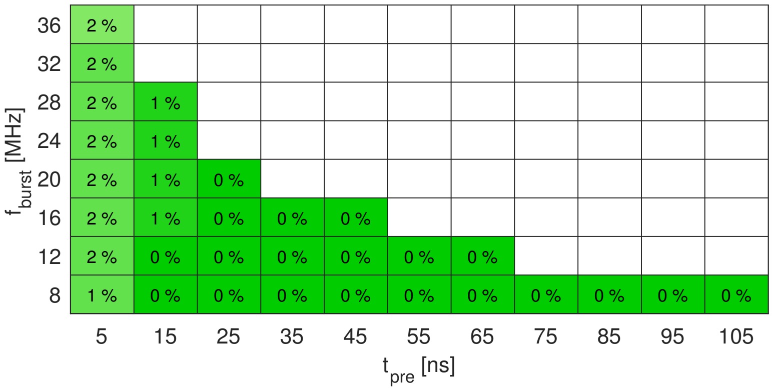
	5	15	25	35	45	55	65	75	85	95	105
36	55 %										
32	45 %										
28	33 %	34 %									
24	20 %	20 %									
20	8 %	9 %	9 %								
16	3 %	2 %	2 %	2 %	2 %						
12	1 %	1 %	0 %	0 %	0 %	0 %	0 %				
8	1 %	1 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %

Figure 4.17: Settling error, measured in a burst read simulation with activated local precharge scheme.

scheme is activated. The time available for precharge in each individual sense amplifier then starts when its (delayed) output goes high and stops with the end of the following precharge phase. It calculates as

$$t_{pre,loc} = t_{burst} - t_{trig,k-1} - t_{delay} \tag{4.14}$$

with burst period t_{burst}, the trigger time of the previous read operation $t_{trig,k-1}$ and the delay t_{delay} used by the local precharge scheme. According to (4.14), the settling of the slope detection is fully independent of the global precharge time t_{pre}. The simulation results for activated local precharge scheme, shown in Fig. 4.17, confirm this statement. The settling error does not change with t_{pre}. It only increases with rising burst frequency. Although gm-boost is not activated for this simulation, 16 MHz can be achieved. This corresponds a precharge time smaller than 62 ns, which is an acceleration of more than 25 % compared to the global precharge scheme (85 ns). The combination of both precharge acceleration mechanisms speeds up the settling significantly. Corresponding simulation results are shown in Fig. 4.18. At a burst frequency of 36 MHz with global precharge time of 5 ns the settling error still makes up only 2 %, which shows the sense amplifier's capability of high speed sensing. It is able to reach read access times below 30 ns according to simulation.



f_{burst} [MHz]	5	15	25	35	45	55	65	75	85	95	105
36	2 %										
32	2 %										
28	2 %	1 %									
24	2 %	1 %									
20	2 %	1 %	0 %								
16	2 %	1 %	0 %	0 %	0 %						
12	2 %	0 %	0 %	0 %	0 %	0 %	0 %				
8	1 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %

Figure 4.18: Settling error, measured in a burst read simulation with both precharge acceleration mechanisms activated.

4.2 Ramp Generation Circuitry

Driving a dynamic voltage ramp to the cells' control gates requires special attention for the memory macro design. Dynamic events at bitline and wordline path have to be well synchronized. Consequently, an embedded ramp generator has to be placed close to the array. This section discusses the ramp generation and its distribution via the wordline driver.

4.2.1 Load Requirements

At first, an analysis of the load helps to take important decisions regarding design and placement of the ramp generator. As highlighted in Section 3.2, the control gate lines inside the array constitute a complex RC network which, can be simplified to a lumped RC lowpass with time constant τ_{CG}. This time constant is the limiting factor for the maximum ramp slope that can be driven to the control gate lines. To keep the ramp propagation along the line fast compared to the sensing duration, the wordline length has to be chosen at a reasonable length. Further, the CG lines connected to many flash cells pose a considerable capacitive load to the ramp generators output, especially because two wordlines have to be charged up simultaneously at each read operation; The active and the reference wordline. This load can cause stability issues for the ramp generator's regulator. The on-resistance of the multiplexer devices for hierarchical decoding partly decouple the capacitive load from the regulator's output and ease the situation slightly. However, a frequency compensation has to take care of the capacitive load characteristic.

As discussed in Section 3.2.2, a reasonable ramp slope of $m = V_{CG,read,max}/6\tau_{CG}$ should be provided from the ramp generator. A further requirement is a fast dis-

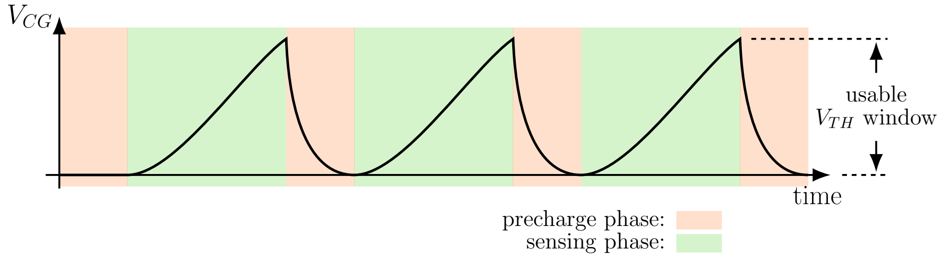


Figure 4.19: Control gate curve of a ramped gate burst read sequence.

charge of the CG lines after sensing, for the subsequent read operation. Figure 4.19 shows the output curve of the ramp generator in the worst case, regarding available precharge time. It occurs in a burst read with maximum read frequency on one wordline. After each read operation, the selected CG has to be discharged to its idle voltage within the short precharge phase and the ramp generator has to return to its target operation point, to start the next sensing with exactly the same biasing conditions. The idle point not necessarily lies at V_{SS}, but this choice allows to increase the usable CG bias window and hence the usable V_{TH} window (see Fig. 4.19), the cell states can be placed in. Consequently, a topology allowing for an output voltage range down to the negative rail (V_{SS}) has to be used.

4.2.2 Generator Design

The structure of the embedded ramp generator implemented for this work is shown in Fig. 4.20. It consists of a tunable current source i_{ramp}, an integration capacitor C_{int}, a buffer amplifier and two switches for discharge. During precharge phase (PRE), the switches $SW1$ and $SW2$ discharge C_{int} as well as the generator's output to V_{SS}. To start the ramp, both switches are opened, and the current source linearly charges up the node n_{int}. The buffer amplifier operates in unity gain configuration and drives the voltage ramp to the load, consisting of the CG decoding path, including its parasitic capacitance C_{dec} and two selected CG lines within the array. A special feedback output is added to the buffer to improve stability. The generator's ramp slope is determined by

$$m = \frac{i_{ramp}}{C_{int}}. \tag{4.15}$$

According to the estimation from (3.10), a time constant of $\tau_{CG} = 2.5\,\text{ns}$ and a maximum CG read voltage of $V_{CG,read,max} = 4\,\text{V}$, a selectable range between $100\,\text{mV/ns}$ and $\leq 350\,\text{mV/ns}$ is implemented. The series resistance of the decoding switches SW_{dec} can be neglected. Different to the CG driver transistors M_{CGdrv},

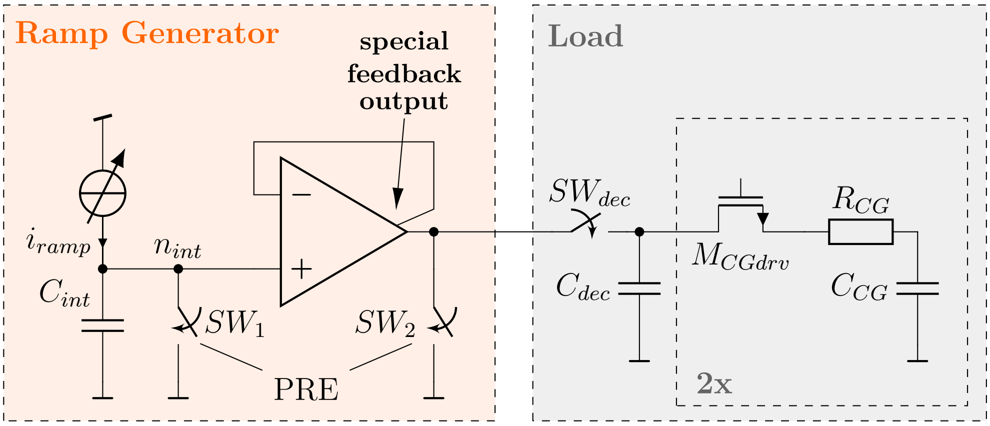


Figure 4.20: Block diagram of the ramp generator including load. Always two control gate lines are selected simultaneously, reference and regular line.

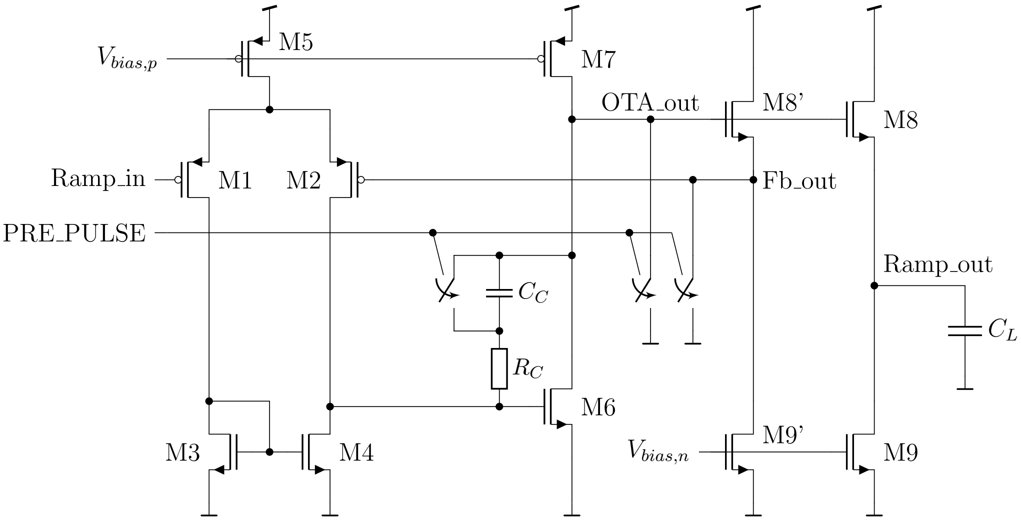


Figure 4.21: Miller opamp with explicit feedback output (Fb_out) in unity gain configuration. Switches allow to discharge the output quickly after each ramp.

they are placed much less often and hence can be implemented as transmission gates, respectively can be sized for a low on resistance.

The design of the buffer amplifier can be depicted from Fig. 4.21. It is implemented as two-stage Miller operational transconductance amplifier (OTA) followed by a source follower buffer stage. The PMOS differential pair at the input supports a common-mode input range down to the lower rail, which is necessary for ramps starting at V_{SS}. The output stage (M8, M9) is added to reduce the power consumption. Without, the Miller OTA would require a considerably high bias current to drive the fast ramp to the load, which is in the range of several picofarads. To achieve decent phase margin, usually a compensation capacitance C_C in the range of 0.3 to 0.5 times the load capacitance C_L is chosen and the transconductance of the output transistor should be in the range of $g_{m8} = 8...12 \cdot g_{m1}$ [56]. According to

$$GBW = \frac{g_{m1}}{2\pi C_C}, \tag{4.16}$$

this would result into $g_{m1} \geq 1.3\,\mathrm{mS}$ and $g_{m5} \geq 10.4\,\mathrm{mS}$, if $C_L = 3\,\mathrm{pF}$ and a gain-bandwidth product (GBW) of 200 MHz is required for example. To achieve these values, a high bias current has to flow permanently through both stages, even if there is no ramp triggered. With the added source follower stage, the output of M6 (OTA_out) is loaded much less. The buffer stage has an input impedance at low frequencies of

$$Z_{in,8,low} \approx \frac{1}{sC_{GS,8}}\left(1 + \frac{g_{m8}}{g_{mb8}}\right) + \frac{1}{g_{mb8}}, \tag{4.17}$$

which indicates, that it loads the output with a fraction of $C_{GS,8}$. At high frequencies, the input impedance approximates with

$$Z_{in,8,high} \approx \frac{1}{sC_{GS,8}} + \frac{1}{sC_L} + \frac{g_{m8}}{s^2 C_{GS,8} C_L}, \tag{4.18}$$

a series connection of the capacitors $C_{GS,8}$ and C_L, that also combines to an effective value smaller than $C_{GS,8}$ [57]. Consequently the buffer M8 decouples the big load capacitance from the output of the two stage Miller amplifier and hence reduces the current needed to achieve stable operation at the same GBW. The last term in (4.18) actually forms a negative resistance with the value $-g_{m8}/(C_{GS,8}C_L\omega^2)$. Its combination with the inductive component in the source follower's output impedance can cause ringing. Especially if the input is driven by a high impedance source and the output is loaded by a large capacitance. Some oscillator designs actually utilize this behavior [58]. This property is unwanted in the ramp generator buffer, since it degrades the overall stability of the regulation loop.

To mitigate the stability issue, a smaller copy of the output buffer is added (M8' and M9') to the circuit. It drives the second output, which provides the feedback for the regulation loop. Since the feedback output (Fb_out) is decoupled from the big load capacity, the additional pole added at the output stage is shifted to higher frequencies and stability is improved. This effect can be seen in Fig. 4.22.

Without using the feedback output (both outputs shorted and directly connected the negative input), a pole is located close to the gain crossover point and diminishes stability by introducing additional phase shift. When the feedback loop is disconnected from the loaded output, the pole moves to higher frequencies, phase shift is reduced and a good phase margin (PM) of about 85° is achieved.

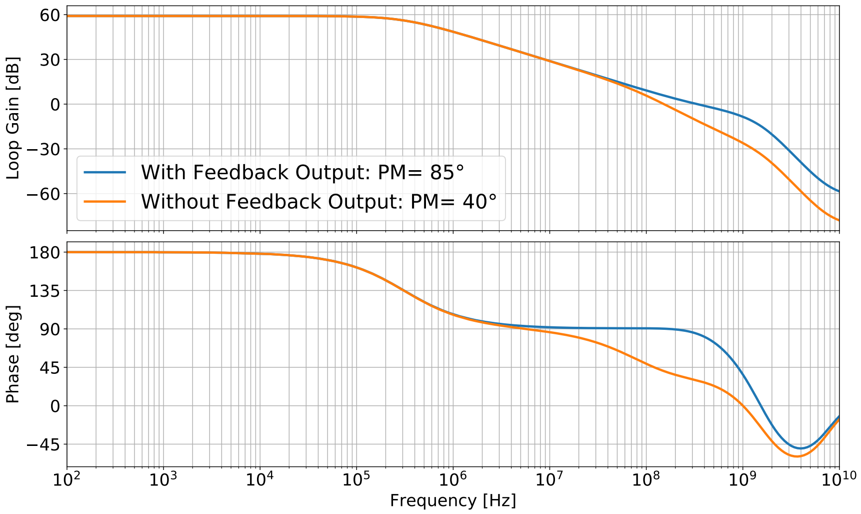


Figure 4.22: Loopgain analysis result from the ramp generator's buffer amplifier with and without explicit feedback output.

A disadvantage of this technique is that only the voltage of the feedback output is regulated, not the voltage at the load. Hence the CGs will not exactly see the voltage ramp as it is generated inside the ramp generator. Mismatch between transistors M8/M8' and M9/M9' causes a DC settling error and the difference in capacitive load causes dynamic errors. As all cells see the same ramp, the different ramp shape causes a common-mode shift, which can only affect the absolute sensing duration, but does not affect sensing accuracy. Thus, stability of the buffer was prioritized. Figure 4.23 shows the resulting output ramp.

To allow for a short precharge phase, the voltage ramp has to be reset quickly. As depicted in Fig. 4.20, there are two switches at the buffer amplifier's input and output to drain the charge from C_{int} and from the load capacitance after completion of the sensing phase. During this discharge process, the buffer amplifier changes its operating point significantly in short time. However, its dynamics are limited by the frequency compensation and therefore, the buffer needs some time to settle at the new operation point. Further, the low ohmic output stage draws high current when simply shorted by SW2. The strong supply current pulse is clearly visible in the transient simulation results, shown in Fig. 4.23. The spike decays only slowly together with the internal nodes settling to the discharged operating point. To accelerate the discharge process, some switches were added to internal nodes of the amplifier (see Fig. 4.21). They are activated with a short pulse (PRE_PULSE), just at the start of the precharge phase. One switch shorts

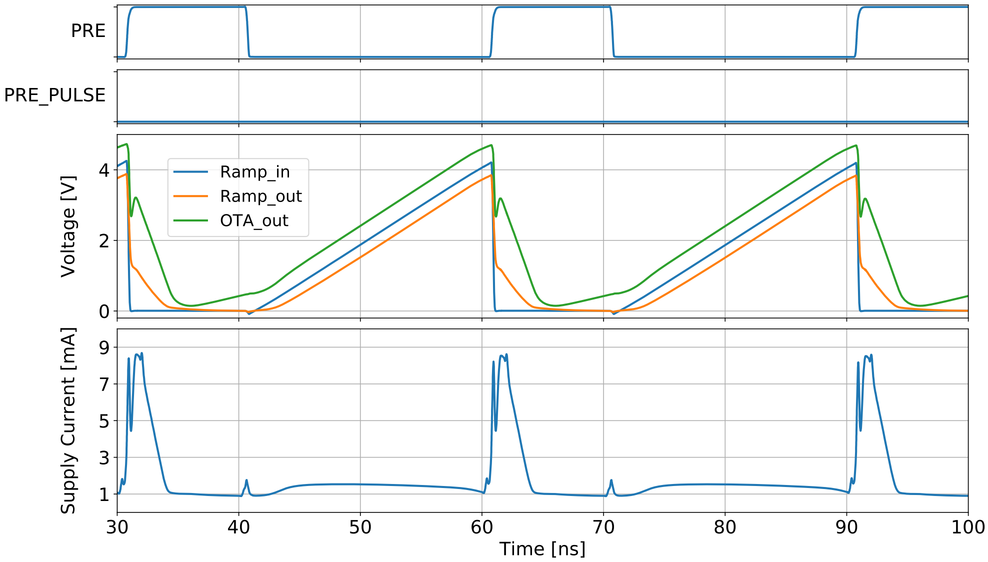


Figure 4.23: Transient simulation result of the ramp generator during a ramped gate burst read sequence. The discharge switches inside the buffer amplifier are not activated.

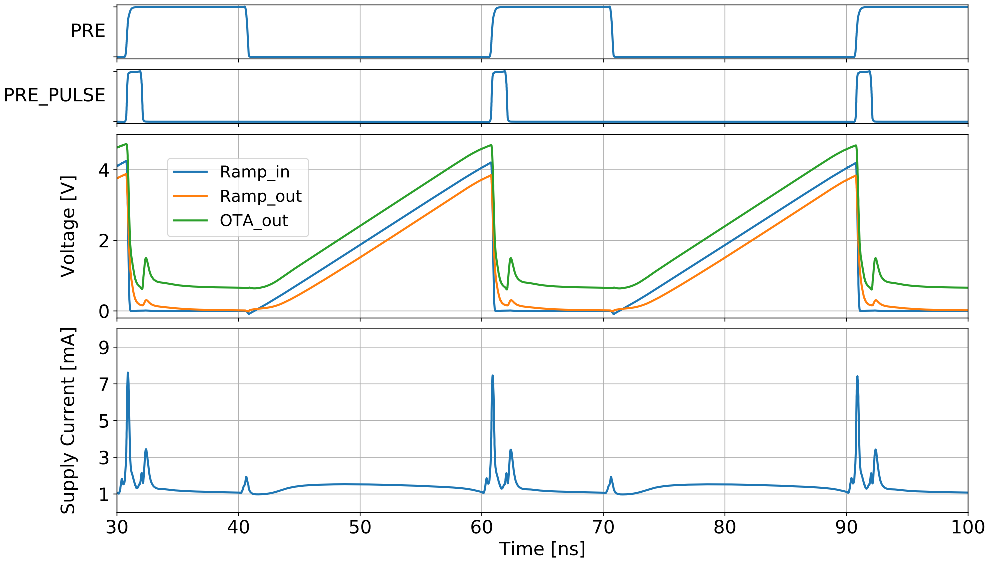


Figure 4.24: Transient simulation result of the ramp generator during a ramped gate burst read sequence. The discharge switches inside the buffer amplifier are activated shortly with the signal PRE_PULSE.

both ends of the compensation capacitor, the other two pull down the OTA- and the feedback output. In this way, the circuit is quickly reset to its idle state and the high output current spike is avoided. After the discharge pulse, the nodes are released again and amplifier returns to its steady idle state, before the next ramp is issued. Figure 4.24 shows a ramped gate burst sequence with the accelerated discharge process. The current spike at the start of the precharge phase is clearly reduced resulting in a lower current consumption.

4.3 Strobing Concept

The previous sections concentrated on the translation from memory cell parameters ($V_{TH,Cell}$, I_{Cell}) to a digital signal carrying the information in its timing (t_{trig}). A time-domain sensing scheme needs a time reference to regain the stored data from that signal. Usually, a reference strobe pulse latches all sense amplifier outputs within flip-flops (see Section 2.2). This section describes the methods used for generation of the reference strobe and their application within the presented sensing scheme.

4.3.1 Reference Sensing Path

An effective way to generate the timing reference is the use of a reference sensing path. It ideally comprises the same circuit parts as the regular sensing path, aiming for perfect matching regarding resistance and capacitance. The output of the reference sense amplifier then directly can be taken as reference strobe. It operates exactly as the regular sense amplifiers except from its input signal, which can have different sources.

Reference Current

One option is the use of a reference current source, supplying the constant current I_{ref} directly to the reference bitline. It is integrated on the reference bitline capacitance $C_{BL,ref}$ and generates the reference timing $t_{trig,ref}$ just as the regular cell currents do (see Section 3.3). Since the current is constant, the resulting trigger time can simply be calculated as

$$t_{trig,ref} = \frac{1}{I_{ref}} V_{sense} C_{BL,ref}. \tag{4.19}$$

The current source can easily be implemented widely adjustable, stable over temperature and over lifetime. So, this reference strobe generation is very flexible, reliable and it is fully functional, right after powerup. Its symmetry to the regular sensing paths allows a direct comparison of the reference current with the

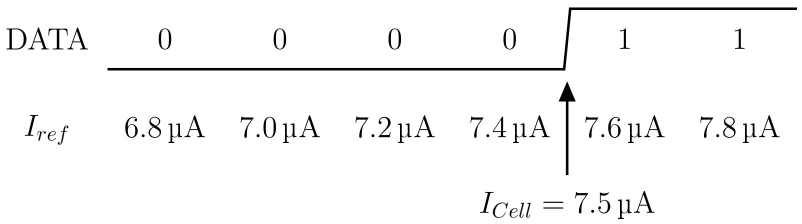
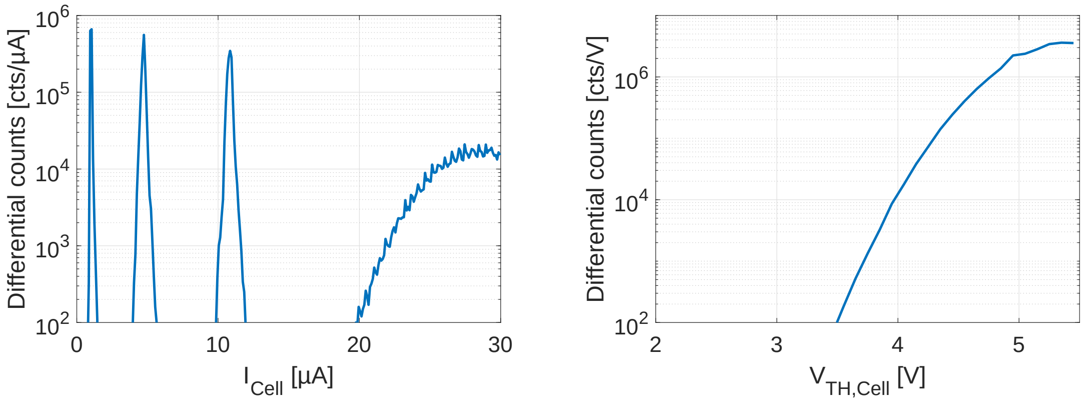


Figure 4.25: I_{ref}-scan method: Data is read multiple times with constant control gate bias, but consecutively increasing i_{ref}. The cell current lies in between both reference currents, corresponding to the data points next to the edge.



(a) I_{ref}-scan of a cell distribution containing four distict states

(b) V_{TH}-scan of programmed cells in SLC operation

Figure 4.26: Cell distributions obtained by scan methods I_{ref}- and V_{TH}-scan.

cell current and hence opens up some useful analysis opportunities. Although the whole sensing principle is based on time-domain, this reference scheme enables a measurement of the cell currents, without a direct memory access. Therefore, the cells are read several times using a step-wise increasing reference current for reference strobe generation while keeping the cells' control gate bias constant. The cell current then is in between both values of I_{ref} corresponding to the read data points at which the SA output (DATA) flips. This is illustrated in Fig. 4.25. With this method, in the following referred to as I_{ref}-scan, cell current distributions can be recorded. To extend the visibility towards programmed cell states, this sweep method can also be executed with a fixed reference current and variable control gate bias. The result is a cell threshold distribution, also called V_{TH}-scan. Figure 4.26 shows two exemplary distributions obtained by these two methods.

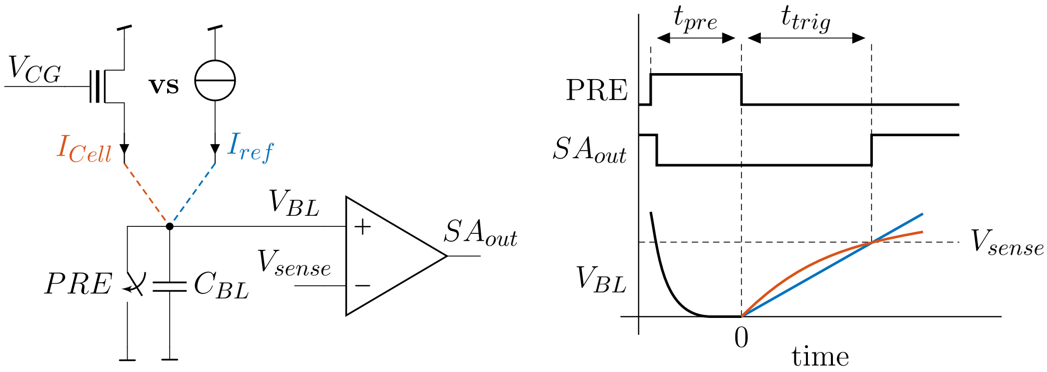


Figure 4.27: Time-domain sensing with different input sources. A constant reference current charges the bitline linearly whereas the cell current decreases with rising bitline voltage.

One weak spot of this reference approach is the different characteristics of the currents that are compared. The reference current I_{ref} is constant, but the cell current is not. It decreases with rising bitline voltage, since voltage sensing is employed. The reference current therefore leads to a linear increasing bitline voltage, the voltage on the regular bitline resembles a RC charging curve. This situation is illustrated in Fig. 4.27. Both charging processes lead to the same t_{trig}, since their average current, evaluated in the time span $t = [0, t_{trig}]$, is equal. The initial cell current, though, is higher than the reference current, apparent from the steeper slope at the start. So the actual cell current is underestimated and the result has an offset compared to a direct memory access measurement, which commonly is done at constant cell bias conditions (at $V_{BL} = 0\,\mathrm{V}$ in this case). The error can be minimized by lowering of the trigger point V_{sense}. This however reduces the amount of charge integrated on C_{BL} and thereby the noise robustness of the sensing. As described in Section 4.1.1, the employed common-gate sense amplifier particularly suffers from a low V_{sense}, since it determines the available timing window. A V_{TH}-scan is affected by the decreasing cell current as well. The underestimated cell current in this case translates into an overestimated cell threshold.

Besides the analysis features of I_{ref}-scan and V_{TH}-scan, the reference strobe generation with constant current can also be utilized for normal read mode. In case of a 2-bit multi-level cell operation, three consecutive strobes can be generated with three reference sensing paths and three different reference currents. This scheme however is not very suitable for the use with a ramped gate cell biasing. According to (4.19), a linearly adjustable I_{ref} results in a hyperbolic characteristic for $t_{trig,ref}$. Consequently, it has the same issues as a fixed gate biasing regarding the exact placement of the cell timings and reference timings respectively (see Section 3.3).

Low values for I_{ref} have to be very precise whereas high values produce narrow spaced timings. A temperature compensation by temperature dependent reference current in general is possible, but very complex, since multiple parameters like cell transconductance, $V_{TH,Cell}$ and ramp slope m have an influence on the time-domain shift caused by the temperature variation.

Reference Cell

Memory cells for generating a reference current are widely used in flash memory modules. The key argument for their usage is the good match to the regular cell characteristics. Properly placed and biased equally, they can serve as very good reference.

The use of reference cells in combination with a ramped gate biasing scheme eliminates the mismatch in time-domain characteristic, as it occurs with a constant reference current. To achieve a good correlation, the bitline path and the wordline path of the reference branch and the regular branches have to be well matched. Another important point is the placement of the reference cells within the module. As in [5], reference cells can be organized in a small dedicated reference array placed next to the sense amplifiers. In this way, there are enough cells to place several references needed for distinction of multiple states in MLC operation and special references for program or erase verify operations. Moreover, this mini-array does not add much area overhead. However, its structural difference to the regular array makes it impossible to match the sensing paths in terms of resistance and capacitance. Another argument against special reference arrays are possible process induced effects, as reported in [59]. As the cells are placed at different spots on the chip, the chance for such effects altering the cells characteristics is much more likely.

Voltage sensing schemes, using a constant reference current, need a perfectly matched reference bitline. It is normally implemented as regular array bitline with cells, that are deactivated by left out drain contacts [50]. In this way, the bitline capacitance is matched to the others as good as possible and there is no cell current obstructing the reference sensing path. By simply keeping the drain contacts of these cells, they can be used as proper reference cells. For the ramped gate biasing approach, a good matching along bitline and wordline is crucial. So the best approach regarding sensing robustness would be to place one reference cell (or multiple ones for MLC operation) per wordline (see Fig. 4.28). This ensures, that reference cells and regular cells see the same voltage ramp at their control gates and at the same time have a well matched bitline path. Common module sizes however would then require thousands of reference cells that all have to be preconditioned. In addition to this maintenance effort, the reference cells would get disturbed significantly during programming of the regular cells on the same

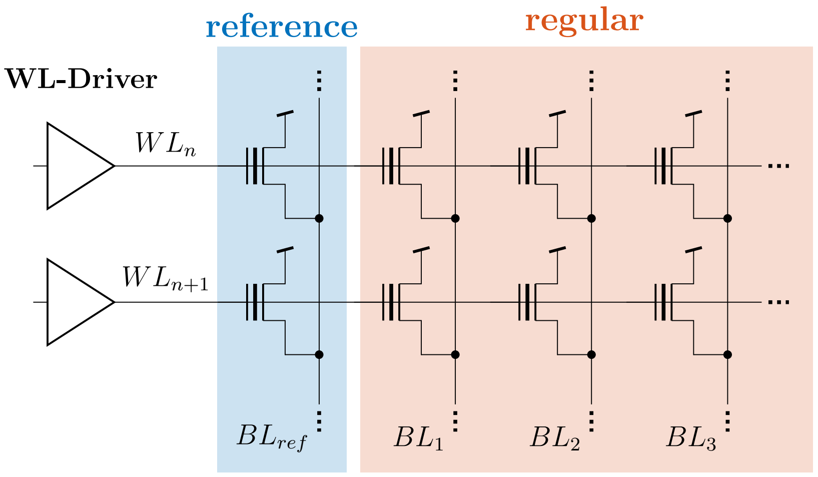


Figure 4.28: Reference cell placement with one reference cell per wordline.

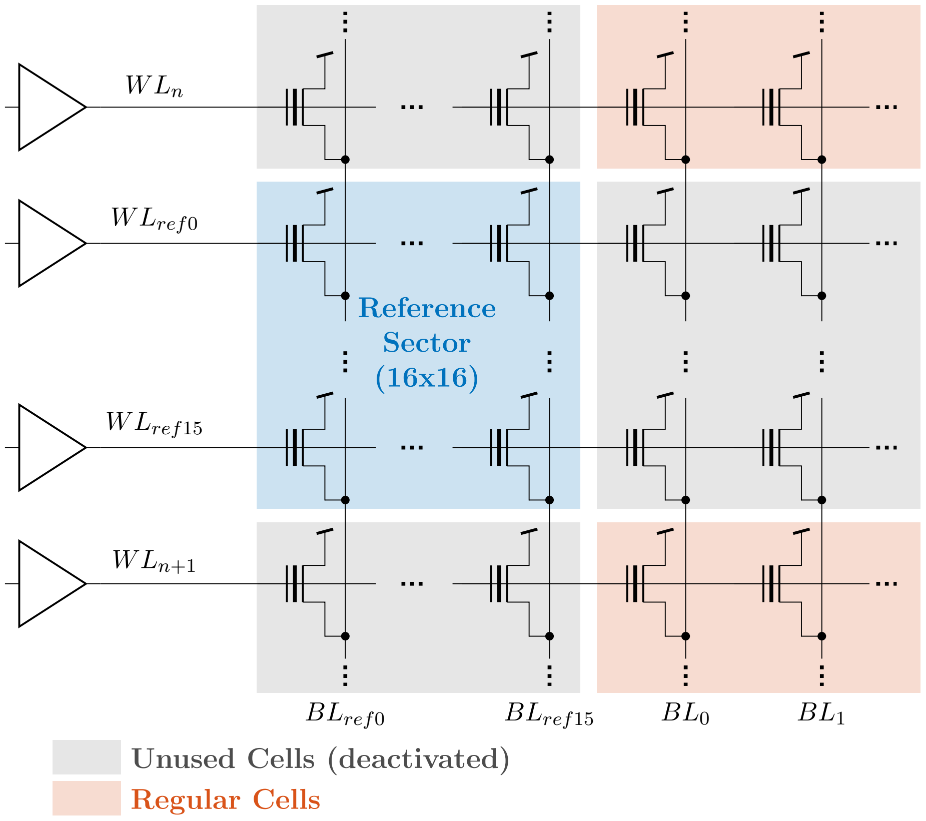


Figure 4.29: Reference cell placement with dedicated reference sector.

wordline. Since the normal embedded Flash array structure only allows to erase all cells inside certain blocks at once, the reference cells would have to be rewritten every time the data from inside their block gets erased.

To eliminate this enormous effort, the reference cells can be placed in a small block inside the regular array that has its wordlines and bitlines reserved for reference only. Figure 4.29 shows the placement of the reference block as implemented for the sensing scheme presented in this work. The remaining Cells along the reference BLs and WLs (indicated with grey boxes) are unused. The electrical separation between reference and regular cells makes sure that the reference cells do not see any stress induced by programming of adjacent cells and hence do not have to be rewritten once they have been preconditioned at the beginning. Since the reference wordline has to be active on every read operation, measures have to be taken to avoid selection of two cells along one bitline at once. The unused cells either can be deactivated by omitting their drain contact or a special selection scheme among different sectors (see Section 4.4.1) can avoid the double selection.

4.3.2 Reference Time

All schemes for reference time generation discussed in previous sections rely on a sensing process that shall match to the sensing of the regular cells as good as possible. Their parameters for tuning are reference current or reference cell state (represented by the reference cell's threshold V_{TH}); both quantities that are only indirectly measured by the time-domain sense amplifier. There are many factors influencing their translation into time, which is the actual quantity deciding whether a cell is read as programmed or erased. Especially the use of a ramped gate biasing in combination with reference cells requires special effort to match regular and reference sensing path. Complexity can be reduced significantly using an absolute time reference, generated in a digital fashion, by synchronous logic. Furthermore, such an approach offers additional analysis capabilities.

A common way to generate timing events used in microcontrollers is the capture/compare unit of internal timers [60]. It can either measure the timings by starting and stopping a running counter on certain events (capture mode) or generate timing events by matching the counter value against certain set values (compare mode). The general working principle of a timer's compare mode is sketched in Fig. 4.30. The core of the timer unit is a counter that increments by one on every rising edge of a clock signal at its input. Once the counter value reaches the set trigger value, an event is triggered. The event could be a set, or a reset of an output signal. On a counter overflow, or a counter reset event its value starts from zero again and timing generation starts all over. As the scheme works with synchronous logic, the minimum timing resolution achievable depends on the

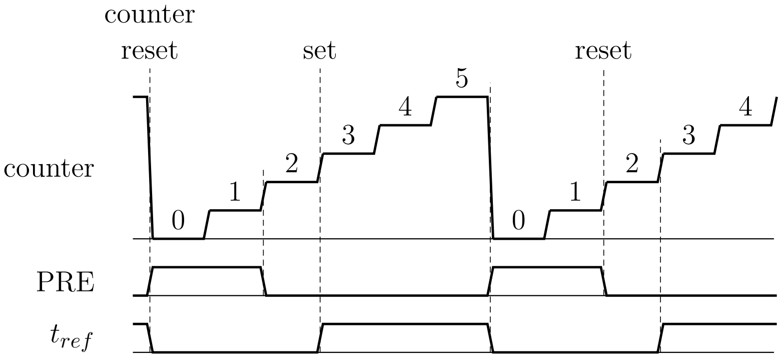


Figure 4.30: Compare mode of a timer.

input clock frequency, which provides the time base. As long as there is a stable clock provided, this approach can produce robust timing signals for sense amplifier control.

A capture/compare unit is used within this work to generate the precharge pulse PRE and an arbitrary sense amplifier strobe, refered to as ”external trigger”. The former one defines the duration of the precharge phase. The latter one is used to strobe the sense amplifier outputs at a certain time within the sensing phase. The counter of the unit is clocked by an internal 1 GHz oscillator. Both signals can be synchronized to either the rising or the falling clock edge and are set with an 8-bit register value. Hence they can be tuned in the range of 0 and 127.5 ns with 0.5 ns resolution. Since both signals are generated synchronously inside the same logic block, their time relation is well controllable, which is important for the time-domain sense operation. Figure 4.31 shows the time relation of the relevant signals during sensing with external trigger. After the precharge pulse, the sense amplifier outputs are strobed with a delay of t_{ext}. Since this delay is fully tunable, it allows to perform a scan in time-domain, which works in the same fashion as the I_{ref}- and V_{TH}-scan.

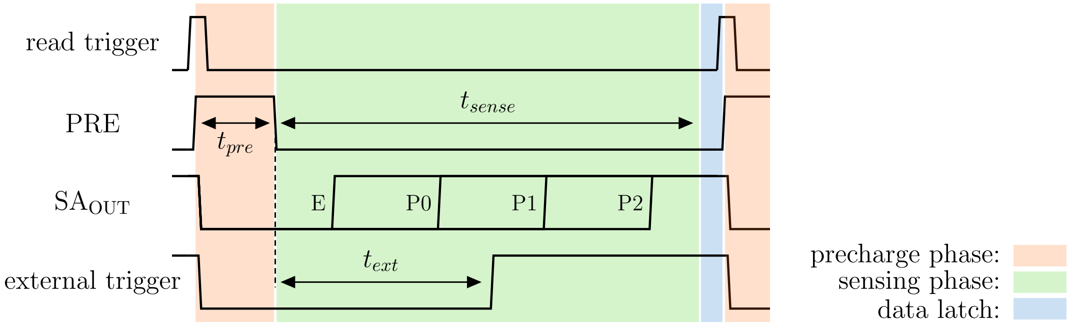


Figure 4.31: Timing of a read operation with external trigger as reference source.

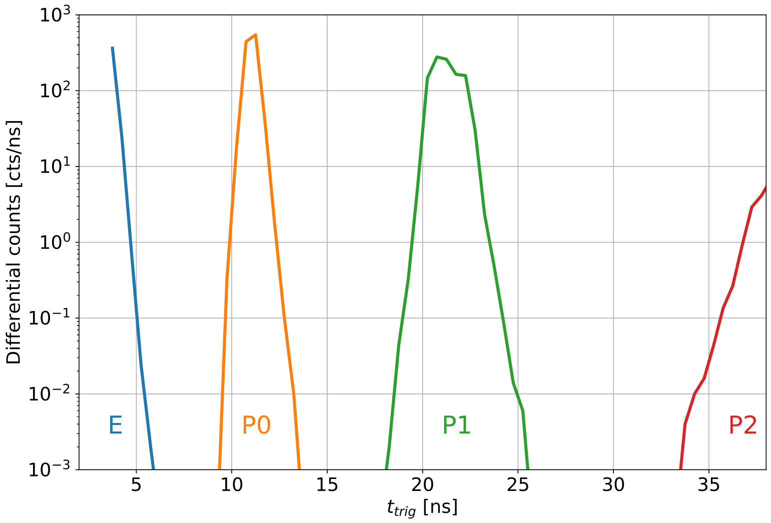


Figure 4.32: Time scan of a cell distribution containing four distinct states.

Figure 4.32 shows an exemplary time scan, obtained by successively stepping through the tuning range of the external trigger signal. It was measured on the test chip, which was designed within this work (Section 4.4.1).

For regular read operation with a single external trigger signal, multi-level operation is only possible in combination with a serial sensing scheme. To support a fast parallel sensing, the system has to be expanded for multiple external triggers (2-bit per cell requires three external triggers). The capture/compare unit approach allows this adaption easily. In the test chip, the reference time strobe is only used for time scan analyses. However, a fully digital time strobing of the sense amplifiers allows to replace reference cells and hence reduces complexity, maintenance and testing effort significantly. Its tunability also enables a digital temperature compensation of the cell threshold.

4.4 Sensing Architecture and Measurement Results

4.4.1 Test Chip Architecture

A memory macro implementing the sensing scheme presented in the previous sections has been fabricated in 28nm CMOS flash technology. Figure 4.33 shows a micrograph of the test chip. It comprises 8M×2-bit MLC flash memory, arranged in two sectors with two 512 kB wings each. Each sector contains 1024 wordlines (WL) plus 16 reference WL and 4096 bitlines (BL) plus 128 reference BL. The sense amplifiers are placed between both sectors and are multiplexed to 16 × 2 local

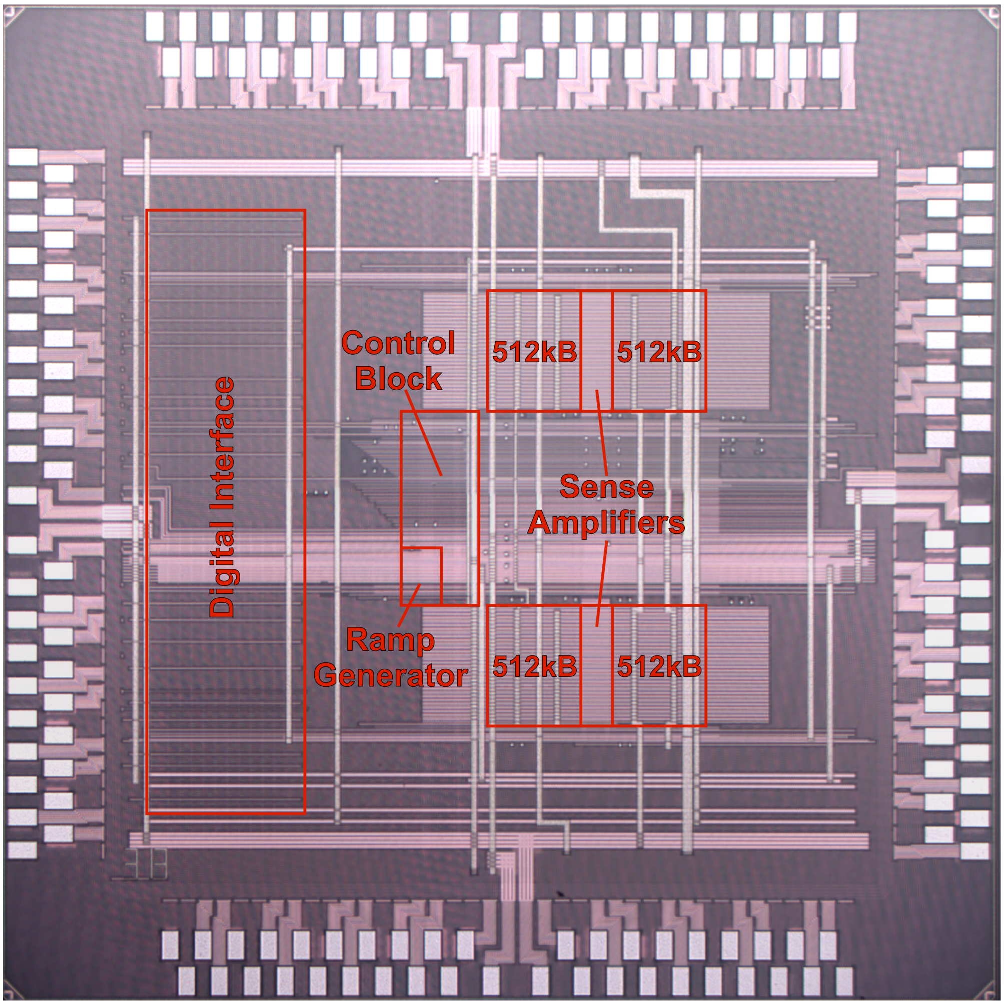


Figure 4.33: Micrograph of the 2 MB test chip.

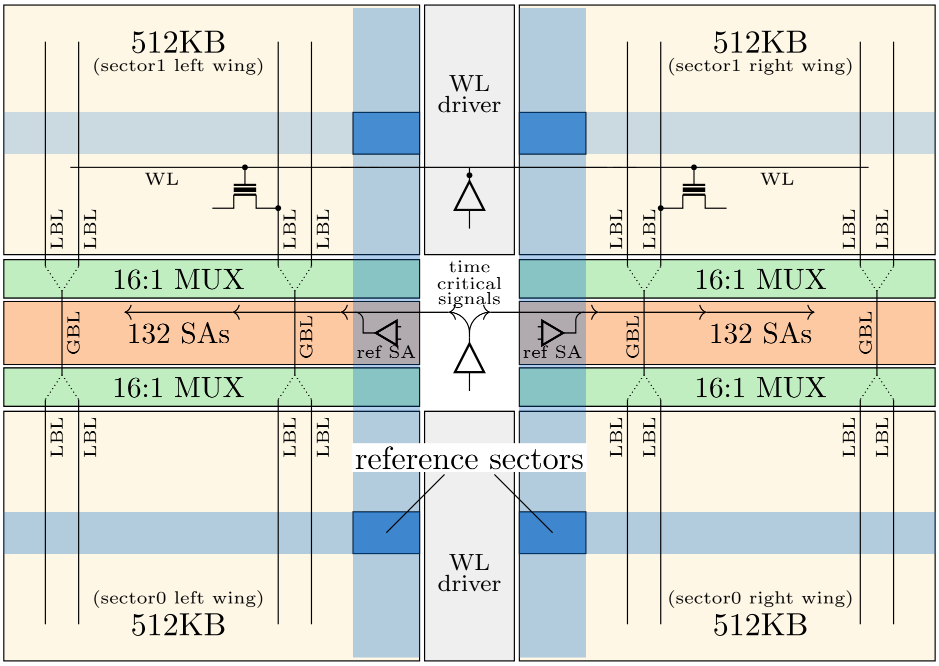


Figure 4.34: Array structure of the implemented memory including indicated reference sectors. Time critical signals are routed from the inner to the outer array edges.

bitlines (LBL) (1×16 in for each sector). In this way, the global bitlines (GBL) are kept very short and hence do not contribute much to the total BL capacitance. The WL-drivers are located between both wings. An additional control block, including switches, analog reference and the ramp generator, is placed at the edge of the memory array. A digital interface for communication with the test chip is placed at the left edge of the test chip.

Figure 4.34 depicts the array structure in more detail. There are four small sectors dedicated for the reference cells; one in each of the four memory tiles. By their integration into the regular cell arrays, the reference cells have very good matching to the other cells. However, they have to be electrically separated from them, to prevent disturbs during programming and erasing. This is why the reference cells have their dedicated WL- and BL-slices including drivers and multiplexers. A double selection on one local bitline (LBL) is avoided by WL selection from opposite sectors (reference from the upper, active from the lower; or vice versa). All dynamic signals like the bias voltage ramp, bitline address, precharge pulse, and the read strobes generated by the reference sense amplifiers are propagated in the same direction from inner to outer array edge. By this means, timing skew among all signals related to the read operation is reduced.

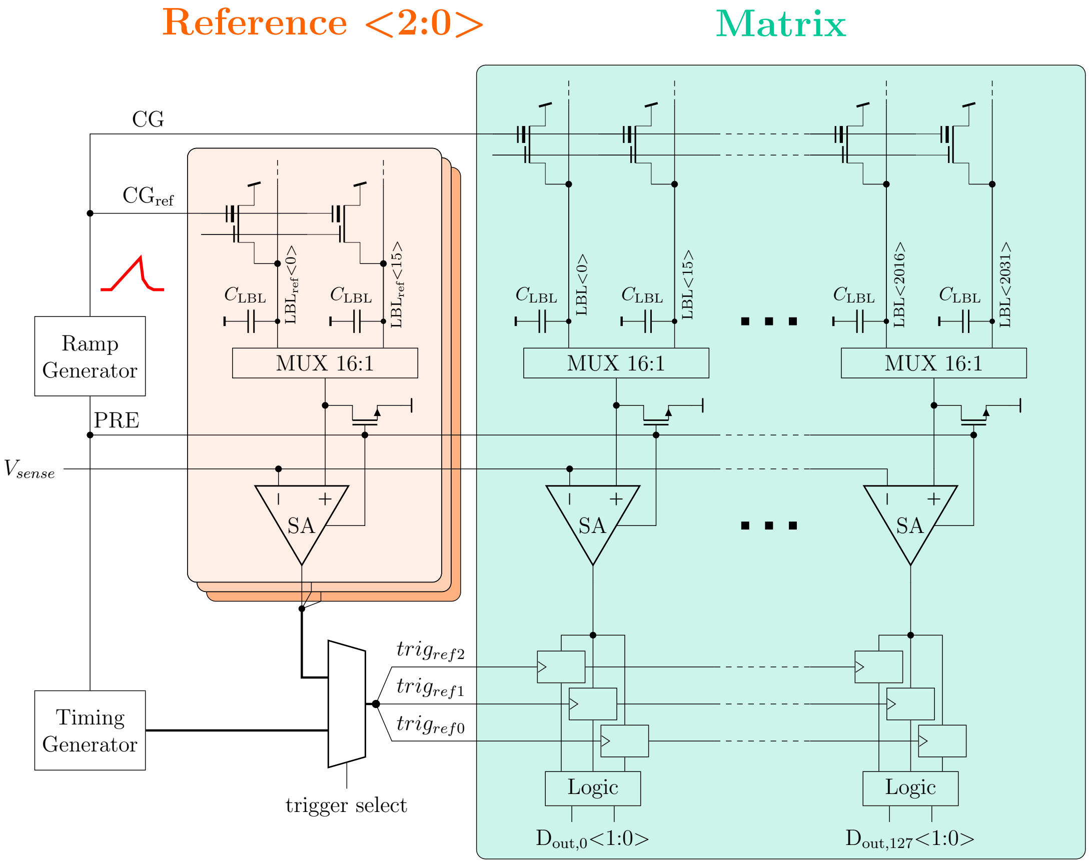


Figure 4.35: Sensing path of one memory wing including ramp generation.

To eliminate the delay from one wing to the other, each wing has its own set of three reference sense amplifiers and hence generates the reference strobes locally. A schematic of the sensing path from one of the two symmetrical memory wings including the ramp generation is shown in Fig. 4.35.

Each read operation is started by a precharge pulse (PRE) generated by the timing generator. After the precharge, the ramp generator drives the dynamic voltage ramp to the selected CG line and to the corresponding reference CG line. All selected BLs are charged by the related cell current and the sense amplifiers translate the BL voltages to time-related digital signals, by comparison against V_{sense}. The outputs of the three reference sense amplifiers consecutively strobe all regular sense amplifier outputs and a digital logic decodes the two data bits from the resulting thermometer code according to Table 4.1. A gray coding scheme is used in order to produce only single-bit-errors for each cell whose state is misinterpreted as one of its two adjacent cell states. For analysis purpose, the reference strobe can be selected from the timing generator. It generates a reference strobe, which is time-correlated to the precharge pulse. This enables the generation of time scans as they are described in Section 4.3.2.

Cell state	$\mathrm{SA}_{\mathrm{out}}$ @ $t =$			$\mathrm{D}_{\mathrm{out}}$
	$t_{trig,ref0}$	$t_{trig,ref1}$	$t_{trig,ref2}$	
E	1	1	1	0 0
P0	0	1	1	0 1
P1	0	0	1	1 1
P2	0	0	0	1 0

Table 4.1: Mapping table for conversion from thermometer code to binary output. The stronger the cell is programmed, the later the corresponding sense amplifier output flips from 0 to 1.

4.4.2 Time-Domain Transfer Function

For verification of the analytical results from Section 3.3, several circuit and cell parameters had to be measured. The time-domain transfer functions ((3.17) and (3.23)) depend on the following parameters:

- Reference voltage level V_{sense}
- Control gate read bias level $V_{CG,read}$ (fixed gate biasing)
- Ramp slope m (ramped gate biasing)
- Bitline capacitance C_{BL}
- Cell transconductance g_m
- Cell threshold $V_{TH,Cell}$

The voltages V_{sense} and $V_{CG,read}$ were provided externally. The ramp slope m is set by a digital register setting. It was verified by wafer-level probing. The bitline capacitance C_{BL} was indirectly determined using a time scan on the reference sensing branch. According to (4.19), the capacitance could be extracted from multiple measurements with different reference currents. For the actual measurement of the time-domain transfer function, memory cells were preconditioned to different threshold levels using an ISPP scheme (see Chapter 5). The resulting cell parameters g_m and $V_{TH,Cell}$ were obtained with a direct memory access measurement of the cells transfer characteristic. Figure 4.36 shows measured I-V curves and corresponding linear fits for extraction of both cell parameters. The

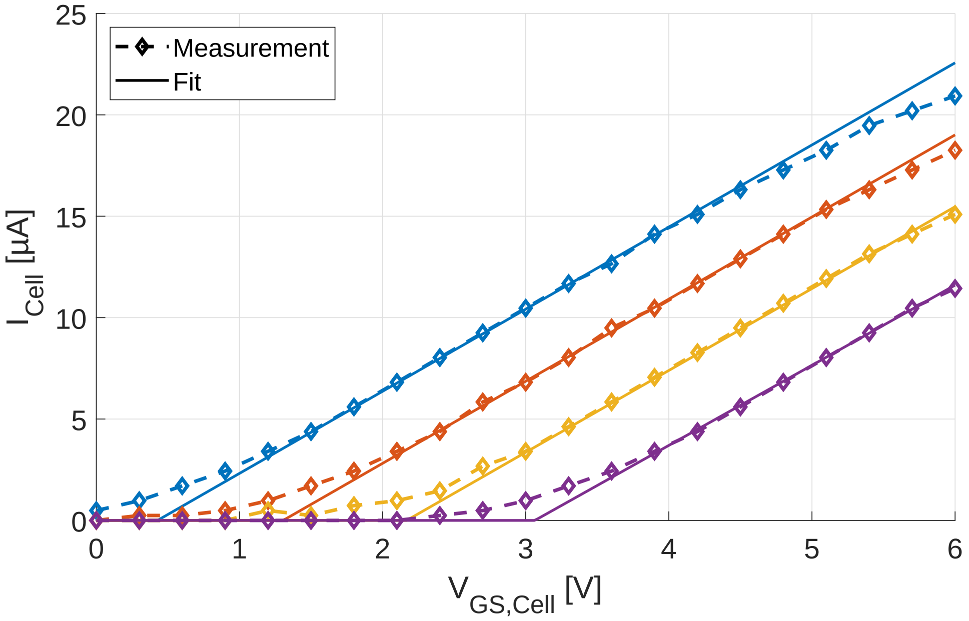


Figure 4.36: DMA measurement for extraction of cell parameters $V_{TH,Cell}$ and g_m.

linear fits match the measured characteristics in a large range. At the low end, the measured current approaches zero much smoother than the model. This is due to operation in saturation region, which is neglected by the linear approximation. At high currents, the measured characteristic flattens slowly, which is explained by current limitation of the select device of the cell (select transistor).

Figure 4.37 shows the measured time-domain transfer functions for both biasing schemes and the corresponding analytical solutions from (3.17) and (3.23). As expected from the analytical analysis in Section 3.3, the fixed gate biasing scheme results a nonlinear curve whereas the ramped gate biasing yields a linear increase. Measurement data has a longer delay as predicted by the model, looking at the low end of $V_{TH,Cell}$. This is due to the model's assumption of an ideal comparator (see Fig. 3.11). Simulations of the implemented sense amplifier (Fig. 4.12) showed a delay of about $t_{delay,SA} = 2\,\text{ns}$ with respect to the actual crossing point of V_{BL} and V_{sense}. Another significant mismatch between model and simulation can be seen for fixed gate biasing at thresholds close to the read bias level $V_{CG,read} = 3\,\text{V}$. It is explained by insufficient modeling of the cell current in the vicinity to the threshold voltage (see Fig. 4.36). Consequently, the analytical model for fixed gate biasing is not applicable in this range. The ramped gate transfer characteristic is flatter than expected from the model. Although its increase should only be determined by the ramp slope m according to the equation, it does not match

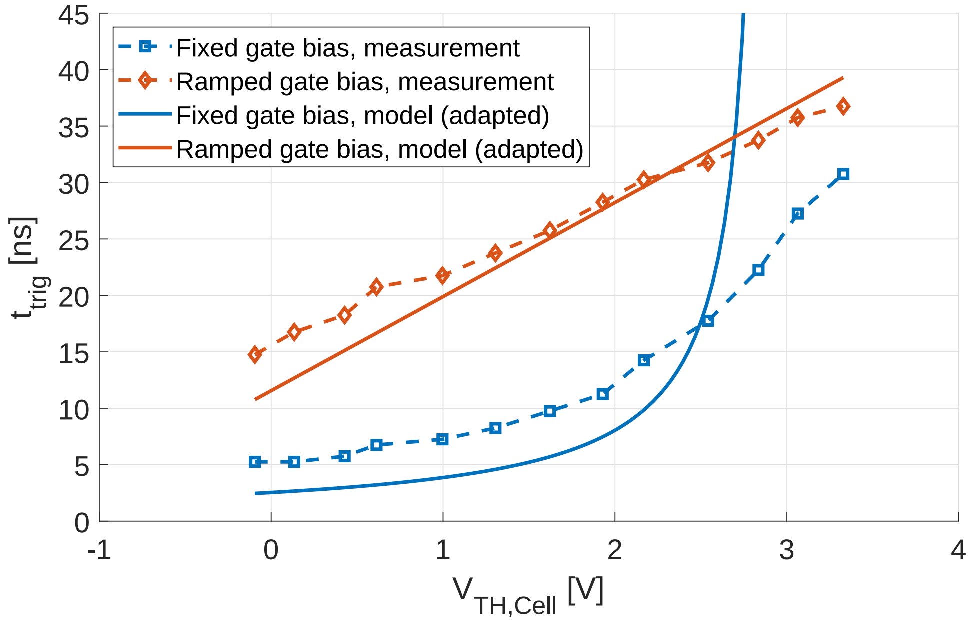


Figure 4.37: Comparison between measured and modeled time-domain transfer function.

the measured bias ramp. The reason for that mismatch is the common-gate sense amplifier's bias current, adding charge to the bitline capacitance in parallel to the cell current. Since the model is based on an ideal comparator (without bias/offset current), it does not cover this effect. To adapt the model to the implemented sense amplifier, the analytical derivation from Section 3.3 has to be redone with a modification of the originating differential equation (3.12). It is expanded by the bias current and yields

$$\frac{dV_{BL}}{dt} = \frac{I_{Cell}(t) + \boldsymbol{I_{bias}}}{C_{BL}}. \tag{4.20}$$

This seemingly small adaption leads to a lot of changes in the solution of the equation. A detailed analytical derivation can be found in Appendix A. Generally, the resulting transfer function has a slightly derated slope, since the bitline voltage rises linearly already before the cell starts to conduct. Consequently, the voltage difference between gate (CG terminal) and source (BL terminal) effectively rises slower, which is the main reason of the changed characteristic. With the adapted transfer functions ((A.28), (A.32) and (A.35)) and a constant offset of $t_{delay,SA} = 2\,\text{ns}$, accounting for internal sense amplifier delay, a more accurate model is obtained. The new comparison between measurement and adapted model is shown in Fig. 4.38. The inaccuracy of the model for fixed gate biasing close to

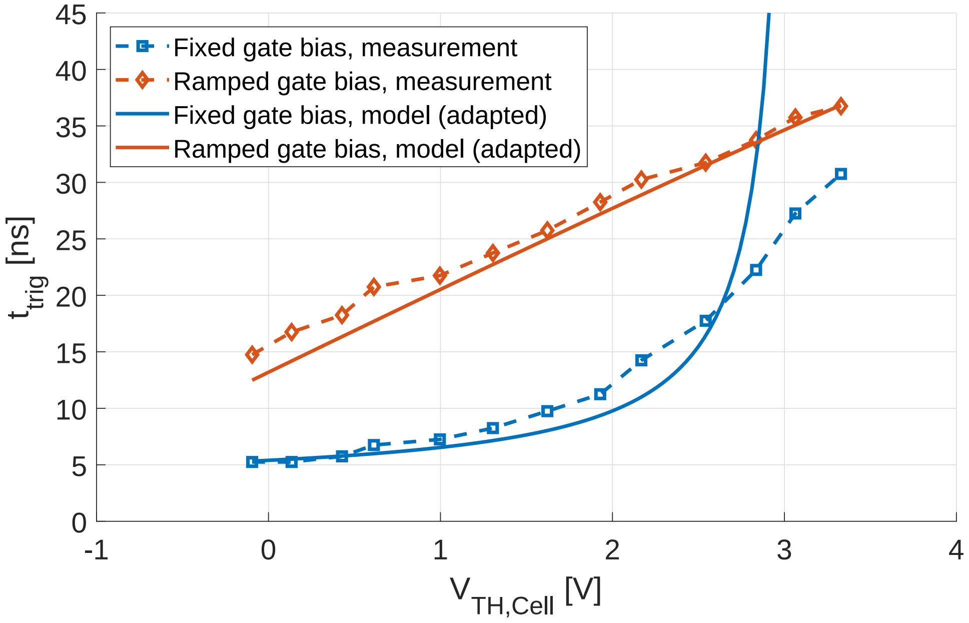


Figure 4.38: Comparison between measured and modeled time-domain transfer function with adapted model, accounting for sense amplifier delay and bias current.

the read bias level $V_{CG,read} = 3\,\mathrm{V}$ stays, as the new model is still based on the linear cell current model. Trigger times at the lower $V_{TH,Cell}$ range are modeled more precise and the slope of the ramped gate transfer characteristic matches better to the measurement.

4.4.3 Precharge Acceleration

Two of the three discussed mechanisms for accelerating the precharge process were integrated into the sense amplifier deployed on the test chip (see Section 4.1.2). To measure their impact on the achievable burst read frequency, the cells were preconditioned with an alternating pattern of erased (E) and intermediately programmed (P1) cell state. As shown in Section 4.1.3, this pattern is well suited to detect too low precharge time available for the return to the proper operation point of the slope detection circuit. Trigger times of the individual reads within the executed read bursts were determined using a time scan (Section 4.3.2). Since the cells were precisely preconditioned to result in specific trigger times, the values resulting from the burst read should be very reproducible for low burst frequencies. The difference of the trigger times from the first and the third read within

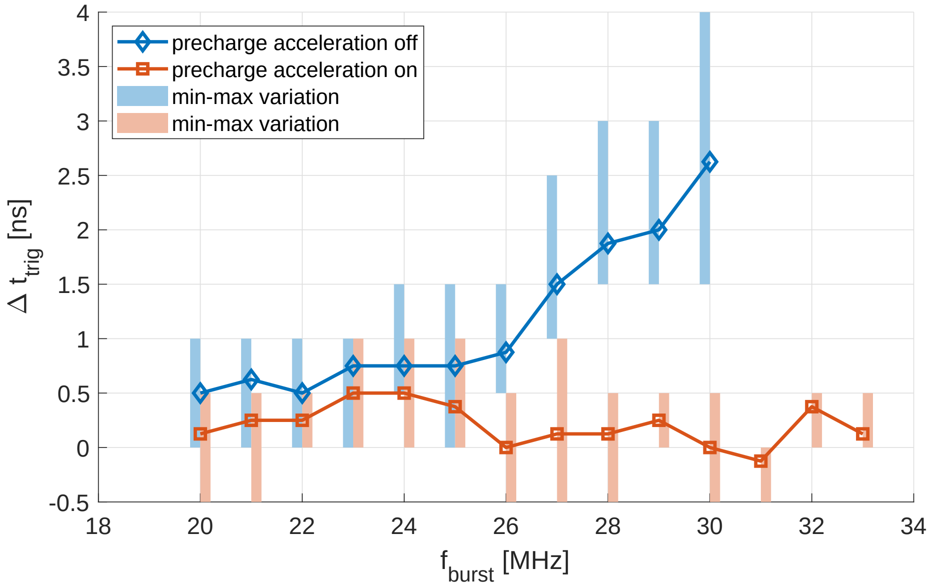


Figure 4.39: Measured timing difference between first and third read operation within a burst read. Activation of precharge acceleration mechanisms enables higher burst frequencies without noticeable distortion caused by insufficient settling.

the burst (the first two reads on P1 cells), calculated as

$$\Delta t_{trig} = t_{trig,3} - t_{trig,1} \tag{4.21}$$

is taken as evaluation criterion for proper settling of the operation point. In the ideal case, the difference should vanish, given the cells could be preconditioned to exactly the same state and the sense amplifiers in both reads start from an identical initial point. As shown by simulation (Fig. 4.14), insufficient settling time would result in a significant difference between read results of the first and the third read. The reason for that is the second read operation in between, which is on an erased cell. It drives the sense amplifiers' internal nodes far from their operation point in idle. If there is not enough time to return to the idle point before the next read operation, the following sensing is slowed down considerably.

The time scans were performed with burst frequencies in the range from 20 MHz to 33 MHz (corresponding to a read period ranging from 50 ns to 30 ns) and a constant precharge pulse length of $t_{PRE} = 11\,\text{ns}$. For the first measurement, both precharge acceleration mechanisms were switched off. For the second one, both were activated. The results are shown in Fig. 4.39. At burst frequencies lower than 26 MHz the difference between both read results are small, no matter if precharge acceleration is activated or not. That means consecutive reads yield reproducible results and the sense amplifier settles properly before the subsequent sense operation is started. At frequencies above, the timing difference Δt_{trig} starts rising if precharge acceleration is switched off. With activated precharge acceleration, the difference between first and third read stays within the range of $+1/-0.5$ns for burst frequencies up to 33 MHz. This shows that the implemented acceleration mechanisms enables the sense amplifier to work not less than 30 % faster and allows random access times down to 30 ns. A comparison of the measurement results without precharge acceleration against the simulation from Section 4.1.3 shows a clear deviation. Measured results indicate proper operation up to 25 MHz burst frequency, whereas simulations predicted a degradation starting already at 8 MHz with about 80 ns precharge time. This can be explained by the fact that the implementation on the test chip does not allow to set a constant precharge time when a time scan is executed. The arbitrary timing strobe, fetching data from all sense amplifiers, puts them into the precharge state already before the end of the read period. Therefore, the sense amplifiers have extra time to return to their operation point, similar to the situation with local precharge activated. Hence the speed increase of the sense amplifiers gained by the added circuitry is even higher than measurement can show.

4.4.4 Read Access Time

For measurement of the achievable read access time, multiple parameters of the sensing scheme have to be optimized. The read access is limited by the following aspects:

- Minimum precharge time t_{PRE}, needed to bring control gate and internal sense amplifier nodes back to idle level
- Slope of the voltage ramp at the array's control gates
- Sense amplifier delay
- Compactness of the cells' trigger time distribution

All four points are mutually dependent within the sensing scheme and hence cannot be distinguished easily. Measurements discussed in the previous section showed that precharge times of about 10 ns result in reproducible sensing results if the employed acceleration mechanisms are activated. The read access time can be calculated as

$$t_{acc} = t_{PRE} + t_{sense}. \tag{4.22}$$

The sense delay t_{sense} depends on the delay of the sense amplifier itself and the achievable compactness of the cells' trigger time distribution at a certain voltage ramp slope. The distribution's compactness shall be defined as

$$C_d(m) = \frac{1}{t_{max,P1} - t_{min,E}} \tag{4.23}$$

with the upper edge of the P1 distribution $t_{max,P1}$ and the lower edge of the erased distribution $t_{min,E}$. A reasonable compactness can only be specified if all four cell states are fully separated. Otherwise correct read out is not ensured. Since the third reference timing, located in between states P1 and P2, strobes the sense amplifier outputs last, the timings of the strongest programmed cells (P2) are not decisive for the achievable access time. Consequently, they do not appear in the equation. According to (3.23), the ramp slope directly affects the translation from cell threshold to trigger time. Hence, the compactness also depends strongly on the applied ramp slope m. The time scans shown in Fig. 4.40 illustrate the compactness and how it can be modulated by the ramp.

All three scans are applied on the same set of cells after programming a pattern containing an equal amount of all four cell states (E,P0,P1,P2). The first subplot shows a time scan recorded with fixed gate biasing. It results in a very compact distribution, which is good for a fast sensing operation. However, the overlap of the erased state with state P0 yields read errors. Consequently, the compactness achieved by this bias setting is not acceptable for robust operation. The second scan is recorded using a ramped gate biasing, starting at $V_{CG} = 0\,\mathrm{V}$ and rising with a slope of $m = 350\,\mathrm{mV/ns}$. The lower initial bias voltage does not affect the erased state since cell current of the 2T cell is limited by the select transistor rather than the control gate bias of the floating gate transistor. Another reason is the internal sense amplifier delay. At very high bitline voltage slopes, its sensitivity saturates, and the delay stays constant. The programmed states P0 and P1 are shifted right to higher trigger times and a lower compactness is obtained, but there is no overlap. All three reference timings can be placed in between the distributions and a sense delay $t_{sense} < 25\,\mathrm{ns}$ can be reached. With a reduced ramp slope $m = 250\,\mathrm{mV/ns}$ in the last measurement, the compactness is further relaxed. This yields a longer sense delay $t_{sense} = 25...30\,\mathrm{ns}$, but widens up the read windows in between the distributions and hence gives more margin for a robust read.

Not only the read bias conditions are relevant to the achievable distribution compactness. The programming algorithm also plays an important role to the shape of the distributions. However it is closely connected to the read operation since every programming pulse is followed by a program verify operation within

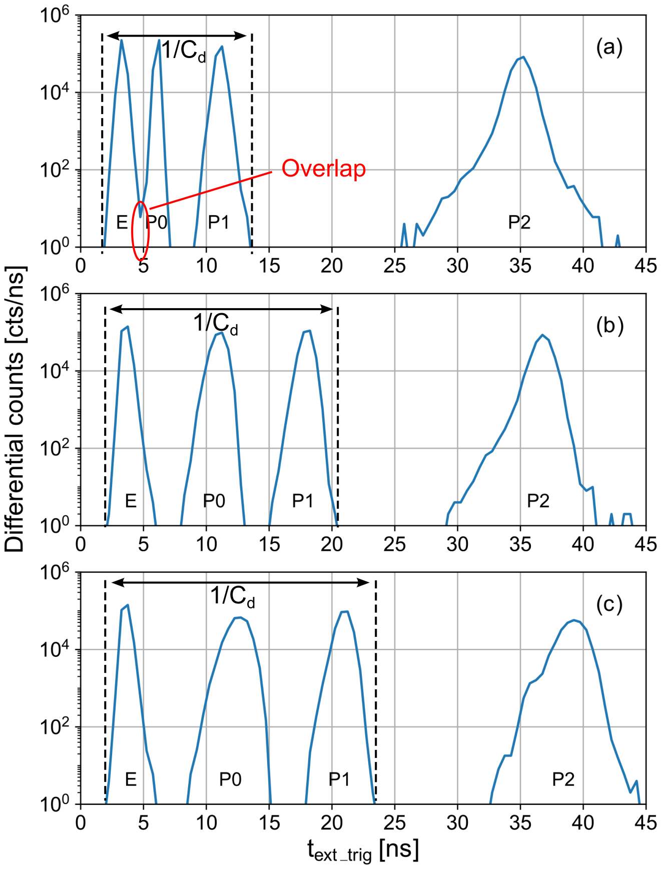


Figure 4.40: Cell distribution measurement result in time domain: (a) with fixed gate biasing; (b) with ramped gate biasing $m = 350\,\mathrm{mV/ns}$; (c) with ramped gate biasing $m = 250\,\mathrm{mV/ns}$.

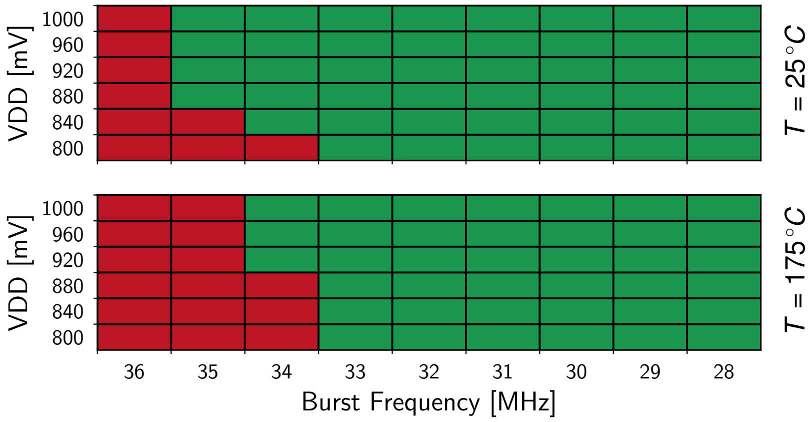


Figure 4.41: Shmoo plot of the measured read access time.

	This work	ISSCC [48] 2005	JSSC [35] 2008
Memory type	NOR Flash	NOR Flash	NOR Flash
Process	28 nm	90 nm	65 nm
Capacity	16 Mbit (2b/cell)	512 Mbit (2b/cell)	1 Gbit (2b/cell)
CG biasing method	ramped gate	stepped gate	ramped gate
Sensing principle	time-domain voltage sensing	current sensing	time-domain current sensing
Random access time	30 ns	65 ns	70 ns
Read throughput	1.98GB/s	470MB/s	400MB/s
Operating temp	-40/175 °C	-	-25/85 °C

Table 4.2: Design comparison with other MLC flash designs.

the ISPP algorithm. For the read access time measurements, a very fine stepped algorithm was used to obtain distributions as narrow as possible. A detailed discussion of the ISPP algorithm can be found in Chapter 5.

With an optimized set of read parameters, the read access time can be brought down to 30 ns [61]. The shmoo plots of the according measurement at room temperature and at the worst case temperature 175 °C are shown in Fig. 4.41. Data could be read out correctly with burst frequencies up to at least 33 MHz, even at reduced supply voltage. The sensitivity towards V_{DD} and temperature variation is low, which confirms the sensing scheme's robustness and its high temperature capability. A read throughput of 1.98 GB/s is achieved by parallel read out of 256 sense amplifiers. Table 4.2 compares this work with two former MLC NOR Flash designs. The time-domain voltage sensing applied in this work achieves a random access time, which is less than half compared to the sensing principles of the other designs. With the higher parallelization, it outperforms the other memories by a factor of four in read throughput and can be operated in a broader temperature range.

4.5 Summary

The time-domain ramped gate sensing scheme introduced in this work offers a robust sensing operation at moderately high read speed. It uses a common-gate sense amplifier with improved slope detection circuitry. By changing the biasing to a current-defined scheme, an offset compensation is added to the sense amplifier. In this way, matching is enhanced and a more selective slope detection characteristic is achieved [54]. The amplifier structure requires a low biasing current to keep up the relative current read window. Since this is in contrast to a fast settling of the offset compensation, acceleration mechanisms are introduced allowing for a shorter precharge phase.

A ramp generator, driving the voltage ramps to the control gate lines of the array, is developed. The deployed amplifier uses a source follower buffer stage to reduce power consumption and a special feedback output to improve stability by decoupling the feedback loop from the big capacitive load. To ensure a quick discharge after each ramp, the buffer amplifier has internal switches to short-circuit the Miller compensation network. By this means, the reset of the ramp generator is accelerated considerably, and current consumption is further reduced.

The flexible strobing concept implemented on the test chip allows for a detailed analysis of the cells and the sensing scheme. A reference sensing path enables an indirect comparison against dedicated reference cells or a tunable reference current. Cell current and threshold distributions can be obtained by I_{ref}- and V_{TH}-scans without direct memory access. Furthermore, a tunable reference time strobe generated on chip can be used for absolute time measurements of the sense amplifier outputs.

Measurements were done on a 16 Mbit MLC memory macro implementing the time-domain ramped gate sensing scheme. The test chip was fabricated in 28 nm CMOS. Results confirm the linear time-domain transfer characteristic of the ramped gate biasing and an achievable random access time of 30 ns up to a temperature of 175 °C.

Chapter 5

Distribution Shaping

The MLC operation of Flash requires a precise placement of the threshold voltage, which is dependent on the floating gate charge. In contrast to the NAND architecture, SLC NOR Flash has no issue with overprogramming. The situation is different in MLC Flash. If a cell gets programmed too much (i.e. too long or with a too high program voltage V_{PP}), its threshold voltage can overshoot to the adjacent cell state, immediately resulting into incorrect data. Since there is no erasing operation on individual bits, the excess electrons cannot be removed from that particular floating gate to undo the error. Therefore, cells have to be programmed carefully and well controlled. In this chapter, the impact of the ramped gate biasing scheme (and its linear transfer characteristic) on the programming efficiency is investigated. Further, coupling effects and the temperature dependency of the memory cell are highlighted and guidelines for the design of a precise programming algorithm are developed.

In general, the amount of electrons injected into the FG can be controlled by the voltage levels applied to the flash cells for a specific amount of time. The parameters depend on the charge injection mechanism and the cell type (1T cell or 2T cell, floating gate or charge trapping layer). Since the implemented memory macro in this work deploys a 2T split-gate memory cell, source-side channel hot-electron (SS-CHE) injection is used for programming and FN tunneling for erasing. In this chapter, the focus shall be on the programming algorithm rather than the physics of programming and erasing mechanisms. Charge transfer to the FG can be controlled most reliably by variation of the applied control gate voltage. Hence the ISPP algorithm developed in this work uses consecutive programming pulses with increasing control gate bias $V_{CG,n+1} = V_{CG,n} + \Delta V_{pp}$. All other biasing parameters remain unchanged. A detailed description of the injection mechanisms and their dependencies can be found in literature [62, 63, 16].

5.1 Cell Programming Algorithm

5.1.1 Cell Biasing During Program Verify

The verify read carried out after each programming step of an ISPP algorithm ensures that the threshold voltages of all cells to be programmed reach their specific target range. In case of an MLC memory with two bits per cell, the target ranges of both outer cell states (E and P2) only have one limit; the inner states (P0 and P1) have a lower and an upper limit which have to be met to make sure there is a sufficient read window in between the states (see Fig. 5.1). Writing data in Flash memories happens in a unidirectional fashion. That means, the programming operation can be executed on individual bits. The erasing operation, however, can only be executed for a whole block (usually comprising thousands of bits). Consequently, if a cell was not programmed strong enough to meet the lower threshold target, it will be detected during verify read and marked for further programming pulses. If it was programmed too much to meet the upper threshold limit, this could also be detected during verify read, but there are no reasonable measures against the issue, except erasing and reprogram. Therefore, the verify read usually only screens for cells that are below their targeted threshold voltage. Compliance with the upper threshold voltage limit has to be ensured by sufficiently fine programming steps ΔV_{pp}.

SLC Flash memories with voltage or current sensing schemes usually set the programming targets in the threshold voltage domain. Accordingly, the cells are biased with the program verify level $V_{CG} = \mathrm{PV}$ and the sense amplifier compares the cell currents against the reference current $i_{ref,verify}$ to determine if the cell is programmed sufficiently. This reference current level commonly is set below the normal read level, to add some programming margin. Hence both parameters, the control gate bias voltage V_{CG} and the reference current i_{ref}, have different levels during verify read compared to normal read. For MLC operation, it is helpful to keep at least one of the two parameter values equal for both read operations. In

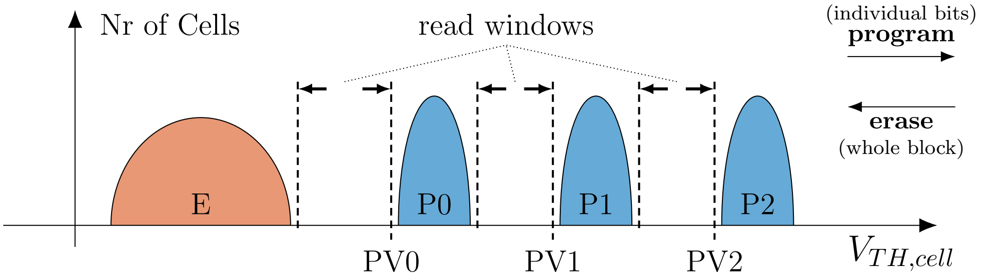


Figure 5.1: Threshold distribution placement for two bit per cell operation.

this way, the number of parameters affecting the sensing result is reduced and the distinction of the cell states gets more robust, as programming can concentrate on the precise adjustment of the quantity, which is the decisive one for the sensing process. It depends on the used cell biasing scheme (Chapter 3) which biasing parameter is to be kept constant. In a stepped gate biasing scheme, the reference current should remain the same and the cell states are separated by different control gate bias levels $V_{CG} = PV0|PV1|PV2$ (distinction in threshold voltage domain). A fixed gate biasing scheme should stick to its normal control gate bias and separate the cell states by different reference current levels $i_{ref,verify\,0|1|2}$ (distinction in cell current domain).

For a ramped gate scheme, as used in this work, the situation is different. The cell bias during normal read is a dynamic voltage ramp and the sensing decision is taken in the time domain (according to the trigger time t_{trig}). To get reproducible read conditions, the same voltage ramp has to be applied for read verify and the programming targets can be directly specified in the time-domain (as verify times $t_{PV0|PV1|PV2}$). The corresponding verify read can then be executed as read operation with external trigger (as described in Section 4.3.2).

Due to its linear transfer characteristic (see Section 3.3), the ramped gate biasing is better suited for program verify than a fixed gate biasing. It allows a more precise cell state placement in time domain. The following example highlights the differences between both biasing approaches with respect to programming: A series of 40 weak programming pulses is executed on a block with 8 kB. The control gate programming bias is increased by $\Delta V_{pp} = 50\,\text{mV}$ from step to step. After each programming pulse, the trigger time t_{trig} corresponding to each cell is recorded using a time scan. The whole series is carried out once with a constant CG bias and once with a ramped CG bias during the time scan.

The resulting trigger times are plotted in Fig. 5.2. In plot (a), the non-linear transfer characteristic of the fixed gate biasing can be recognized in the shape of the distribution range. As the mean value remains low throughout the series, the majority of cells have their threshold voltage in the low sensitive region. The slow cell, indicated with a dash-dotted line, barely changes its trigger time, whereas fast cells at the upper end of the distribution reach the bending part of the transfer curve and quickly drift apart to high trigger times. The results from the series with ramped gate biasing in the lower subplot show a stable linear progression. Cells located at the upper end of the distribution (referred to a fast cells) are relevant for the amount of overprogramming and hence for the achievable distribution with. Since the maximum value in both biasing schemes has a similar gradient in the range of $10\,\text{ns} < t_{trig} < 15\,\text{ns}$, they perform comparable in this target region. For trigger times of $t_{trig} > 20\,\text{ns}$ overprogramming gets worse when fixed gate biasing used, as the slope of the upper end of the distribution range increases significantly

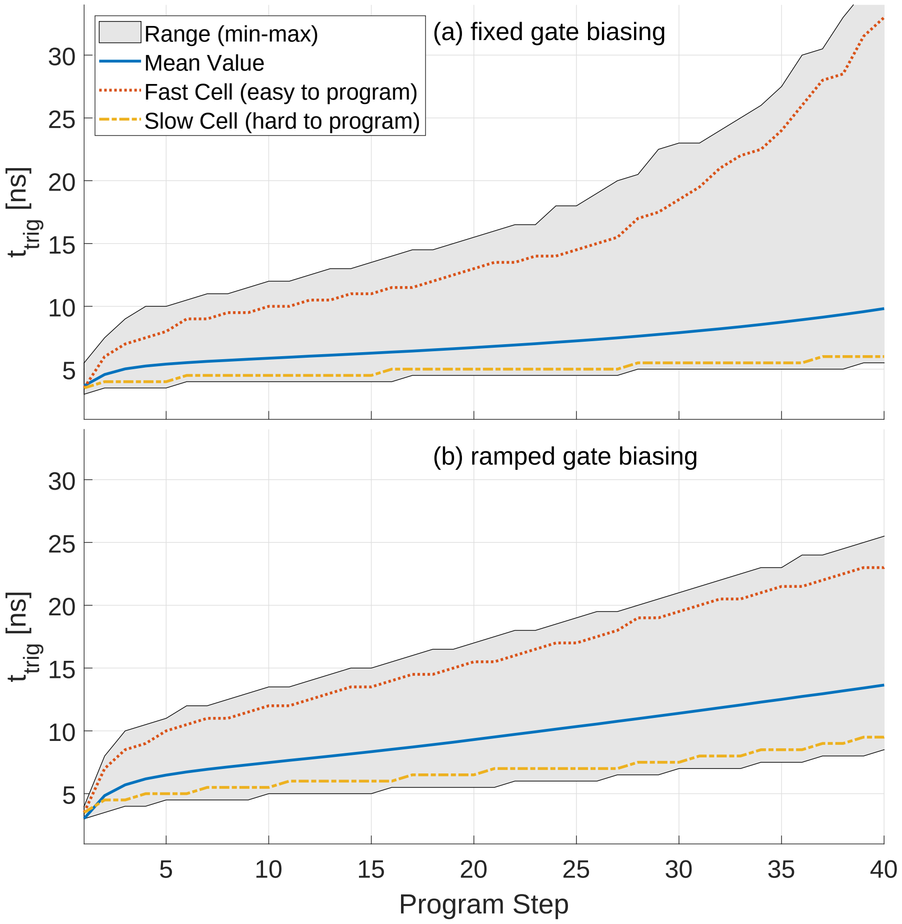


Figure 5.2: Trigger time t_{trig} measurement results, recorded after each step of the programming pulse series: (a) Verify read with fixed gate cell biasing; (b) Verify read with ramped gate cell biasing. The dotted and dash-dotted lines highlight the t_{trig} progression of exemplary cells.

to a slope of about Δt_{trig} = 1...1.5ns per step. The ramped gate biasing in Fig. 5.2(b) keeps the slope below 0.5 ns per step and thus can either achieve two to three times narrower distributions with the same step width ΔV_{pp} or in turn achieve the same distribution width with two to three times less programming steps. Cells located at the lower end of the distribution range are hard to program. They need significantly more steps to reach the same trigger time target and hence are referred to as slow cells. A steeper slope of the lower edge in the diagram indicates, that the slower cells can reach the target faster and the ISPP algorithm needs less programming steps to finish. In the shown range, the slope of the bottom edge in the lower subplot is two to three times larger than in the upper subplot. This means the algorithm using a ramped gate biasing during verify has a further advantage over the fixed gate algorithm than calculated from the progression of the fast cells.

To sum up, the experiment shows that the ramped gate biasing is more suitable for precise placement than the fixed gate biasing. The linear transfer characteristic allows to produce narrower time-domain distributions with less programming steps. Consequently, a higher programming throughput can be achieved.

5.1.2 Programming Order

With decreasing feature size, parasitic effects of the memory cell become more relevant for performance and reliability. Cell-to-cell interference (in particular FG-FG coupling) has been the most dominant parasitic effect diminishing the read window of NAND Flash memories for many technology nodes. Due to a lower density in the array structure, NOR Flash has been affected less. Especially modern eFlash cells like ESF2 or ESF3 [64] do not suffer much from coupling between floating gates along the bitline, as they are shielded by the other cell contacts such as wordline or source line. As SLC Flash has much more read window margin, cell-to-cell interference has not been a big issue in eFlash for automotive applications. For reliable MLC operation however, it has to be considered.

A measurement is conducted, to determine the interference effect for neighboring cells of adjacent bitlines of the memory macro, designed within this work. In the experiment (see Figure Fig. 5.3), a set of victim cells is programmed to an intermediate threshold level. After recording a V_{TH}-scan, the neighboring cells (aggressor cells), located on the same WL but on adjacent BL, are programmed to a broad range of threshold voltages. At the end, a final V_{TH}-scan is recorded, and the shift of threshold voltages of each cell is calculated.

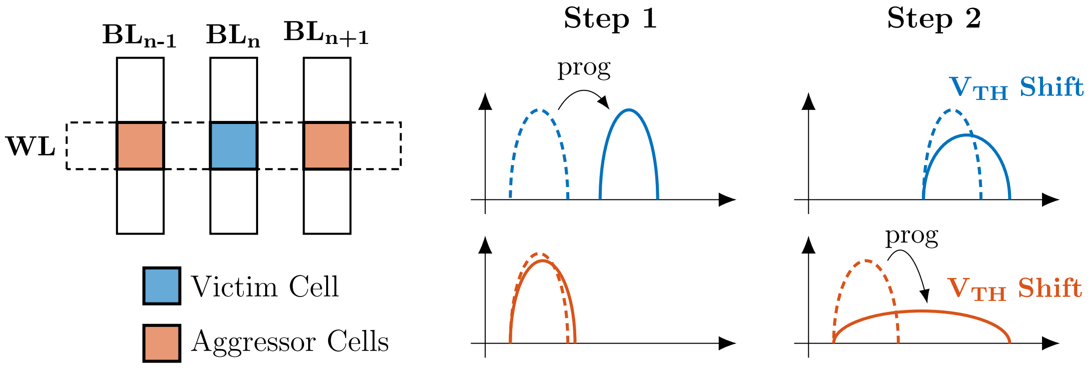


Figure 5.3: Programming procedure of the experiment, conducted to determine the interference effect of neighboring cell along one WL.

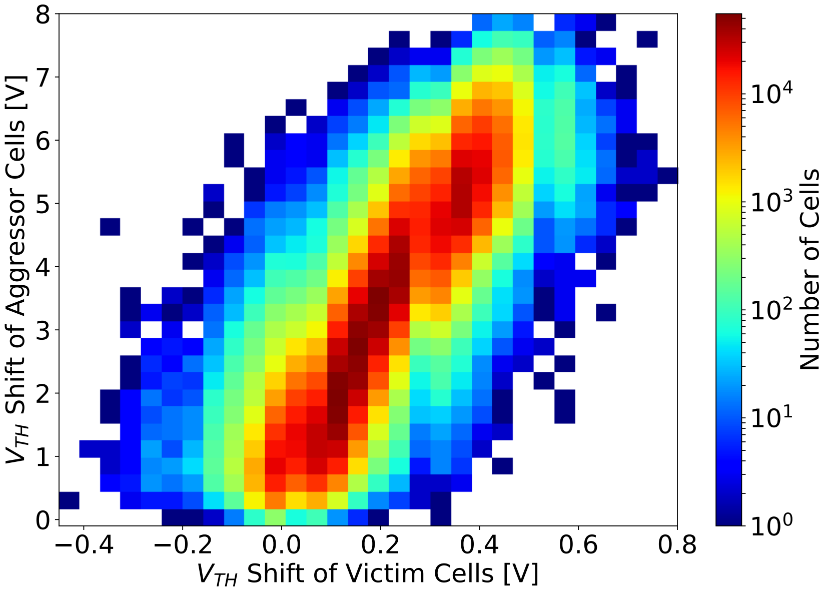


Figure 5.4: Measured cell-to-cell interference between adjacent cells along a WL.

Figure 5.4 shows both V_{TH} shifts plotted into a two-dimensional histogram with color map. A clear correlation can be concluded from the plot. The coupling ratio between adjacent cells on the same WL can be calculated as

$$k = \frac{\Delta V_{TH,victim}}{\Delta V_{TH,aggressor}} \approx \frac{6\,\text{V}}{0.4\,\text{V}} = 6.7\,\%. \tag{5.1}$$

This result gives evidence and shows the significance of inter-cell interference for MLC operation of eFlash, as a considerable part of the available cell threshold range can already be consumed by widening of distributions during programming.

There are numerous approaches tackling the interference issue in NAND Flash technology. They either improve the device geometries to reduce coupling ratios [65], or they alter the programming sequence to reduce the V_{TH} shift in the final step. Among the latter approaches, program-after-erase schemes [26], improved programming orders [66] and imprint programming schemes [10, 9] can be found. The findings from the NAND schemes altering the programming schemes are considered for the development of a programming algorithm, suitable for the time-domain sensing scheme of this work. Further the impact of the different order on the time-domain behavior with ramped gate biasing is analyzed.

Different to NAND Flash memories, overprogramming of the strongest programmed cell state (P2) is not an issue in NOR Flash, as each cell is directly accessible. Consequently, a positive threshold shift of this state, caused by cell-to-cell interference, is tolerable. It does not diminish the read window. Therefore, it is of advantage to first place state P2 as the first order aggressor. Placing the intermediate states (P0, P1) afterwards does not suffer from the strong coupling of cells in state P2. As shown in Section 5.1.1, using ramped gate cell biasing for the program verify operation yields a more precise placement. Combining this verify method with the imprint programming methodology, allows to merge the placement of states P0 and P1 into one step of the ISPP algorithm.

Imprint programming methods for NAND Flash set several intermediate threshold voltage targets for each step, leaving some margin to the final target V_{TH}. Each C_{TH} level is aimed for by changing the control gate bias to the specific program verify (PV) level. After each programming target is reached, the algorithm proceeds with to the V_{TH} level. The requested V_{TH} shifts for the last step are to be minimized in order to reduce the inter-cell coupling effects to a minimum.

The ramped gate time-domain sensing scheme in this work allows, to set the programming targets in the time-domain without changing control gate bias conditions from step to step. Therefore, the verify times t_{PV0} and t_{PV1} are set once at the beginning and cells are shifted to the target within one series of programming pulses. The verify operation consecutively marks the cells that reached their designated target and the distributions are continuously shaped to their target. The step is finished once both states (P0 and P1) are placed. With this continuous shaping, the V_{TH} shifts from pulse to pulse are small, reducing the inter-cell interference and the algorithm can fully benefit from the linear time-domain characteristic of the ramped gate biasing scheme.

To verify the chosen programming order and the shaping methodology, the following measurement is conducted. A block of 64 kB is programmed to a target pattern comprising all four cell states. To cover the effect of cell-to-cell interference, the pattern alters through all states along each WL. The used pattern can be seen in Fig. 5.5. For comparison, the programming operation is executed with two different programming sequences:

- **ISPP sequence 1**
  1. Program states P0 and P1 with the continuous shaping method.
  2. Program state P2.
- **ISPP sequence 2** (chosen for this work)
  1. Program state P2.
  2. Program states P0 and P1 with the continuous shaping method.

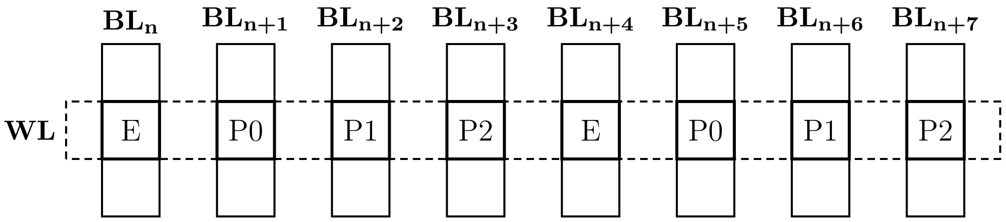


Figure 5.5: Programming pattern used for the experiment, conducted to determine the time-domain performance of the chosen programming sequence.

The resulting time-domain distributions can be seen in Fig. 5.6. The cell distribution obtained with ISPP sequence 1 has state P1 shifted by 1.5 ns to the right from the original target. This leads to a diminished read window between P1 and P2. The result from ISPP sequence 2 shows no shift of P1, but P2 shifted by 1 ns to the right, which widens this window. Both shifts are as expected and can be explained by the cell-to-cell interference. Having a closer look on the

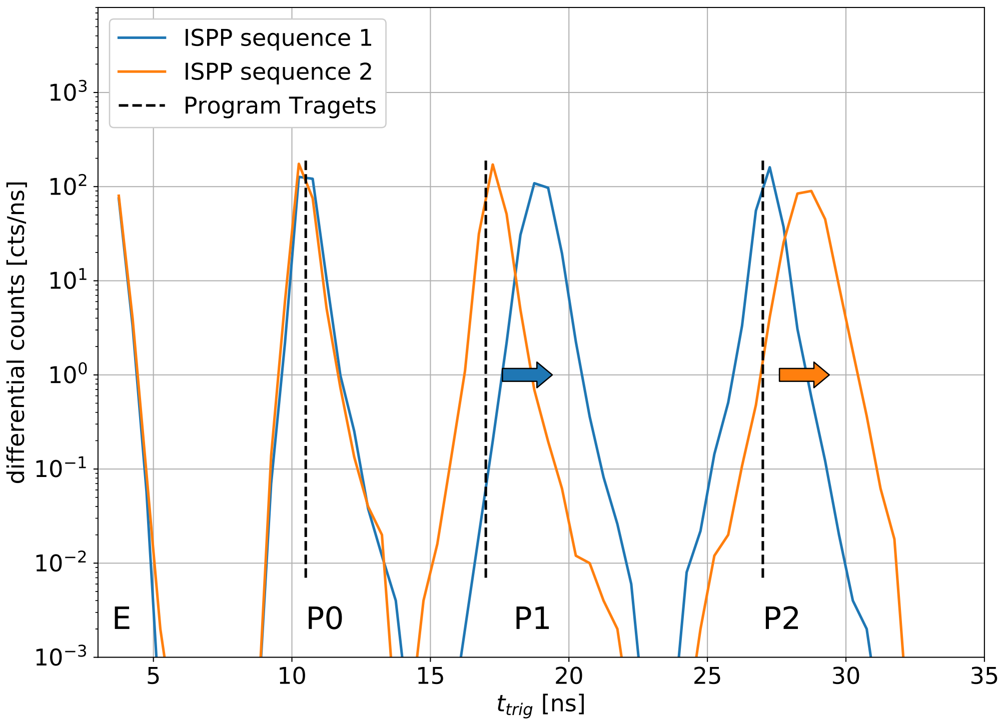


Figure 5.6: Time-domain distribution, measured in the experiment, conducted to verify the performance of the chosen programming sequence. The colored arrows indicate V_{TH} shifts, caused by cell-to-cell interference.

plot, the opposite effect can be observed for the read window in between P0 and P1; sequence 1 results a wider distance than sequence 2. This effect, however, is caused by the regularity of the used programming pattern. Every cell has the same coupling effects as the others programmed to the corresponding state. Under realistic conditions, the programmed data is random, and the distribution would rather be widened than shifted. The result would hence be a combination of the curves from both ISPP sequences.

The experiment shows that ISPP sequence 2 is the better choice, as it has the wider read window and the shift of P2 to the right does not affect the performance of the NOR Flash. Looking at state P0, no significant shift can be observed, which means the continuous programming method of P0 and P1 also performs well in reducing the coupling effect.

5.2 Temperature Shift

Device characteristics as threshold voltage V_{TH} and transconductance g_m of MOSFETs have a rather high temperature dependency. As automotive applications require a broad operating temperature range from $-40\,°\mathrm{C}$ to $175\,°\mathrm{C}$, it is important to analyze and to understand the impact of temperature on the cell and the sensing process, to develop a programming algorithm for a robust Flash memory design. In the following, the temperature effects of the Flash memory cell are discussed and their impact on the sensing scheme is analyzed.

There are two major parameters affecting a MOSFET's operation, that vary with temperature. The threshold voltage V_{TH}, having a negative temperature coefficient and the carrier mobility μ_s, which also degrades with increasing temperature. Both parameters cause the transfer curve to flatten out at higher temperatures and there is a bias point with zero temperature coefficient (ZTC) where both effects cancel each other out [67, 68]. This operation point is commonly used for high temperature analog circuits.

A Flash cell has the same conduction principle as a MOSFET and hence its characteristics change in the same way, when temperature changes. Figure 5.7 shows the measured temperature dependency of a programmed cell's transfer curve. At high operating temperatures, the read window generally decreases. The currents of programmed cells has a positive temperature coefficient (PTC) characteristic as V_{TH} drops with increasing temperature. The currents of erased cells decrease. They are usually operated with a bias above the ZTC point. Due to the degrading carrier mobility their current has a negative temperature coefficient (NTC) characteristic.

The biasing scheme used in this work does not operate the cells in one specific control gate voltage during read. The applied ramp sweeps through a voltage

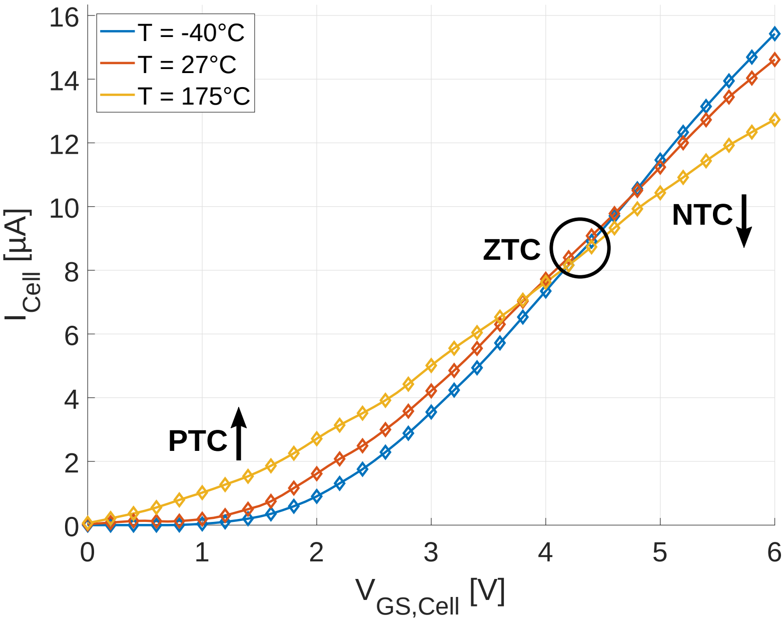


Figure 5.7: DMA Measurement a of programmed cell's transfer characteristic. At low gate bias, the current has PTC characteristic. At high gate bias it has NTC characteristic. The ZTC bias point lies in between.

range starting from 0 V up to 4 V. Consequently, the effects of temperature on the resulting time-domain windows cannot directly be seen from that observation. The decision criterion of the applied sensing scheme is the time t_{trig}, needed to accumulate the charge

$$Q_{trig} = V_{sense} \cdot C_{BL} \tag{5.2}$$

on the BL capacitance (given the bitline is discharged to 0 V at the start of the integration). The time depends on the cell current I_{Cell} and can be calculated from the integral

$$Q_{trig} = \int_0^{t_{trig}} I_{Cell}\, dt. \tag{5.3}$$

As the ramped gate biasing scheme linearly sweeps through the gate bias voltage range, Q_{trig} can be visualized as the area under the transfer curve, which is plotted versus time (see Fig. 5.8). On the left side, the current of an erased cell is shown. Both surfaces marked under the curves symbolize the charge Q_{trig}. They have the same area. At high temperature, the area extends to the right, which means the sensing results in a larger time t_{trig} compared to the cold case. On the right, the situation for programmed cells is plotted. In this case, high temperature results in a lower integration time. The plot in the middle shows that there is a ZTC

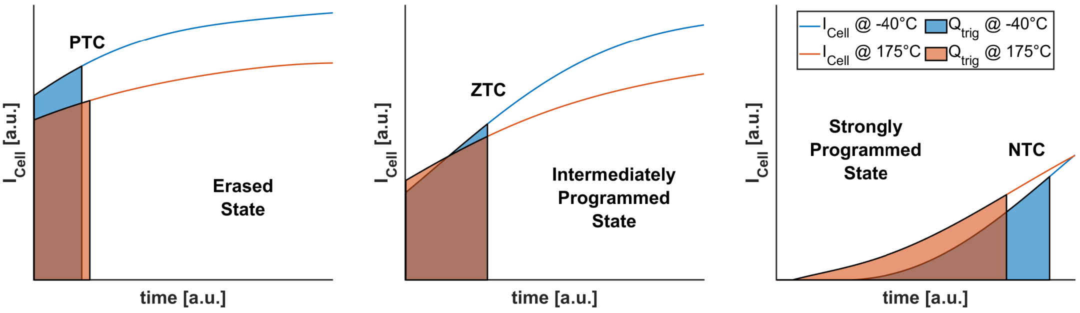


Figure 5.8: Visualization of the accumulated bitline charge during the sensing process of different cell states at hot and cold temperature. The highlighted areas below the curves represent the charge Q_{trig}, needed to trigger the sense amplifier. In each case the marked surfaces extend differently in time.

point in time-domain like in the cell current domain. To verify this temperature behavior and to see its impact on the time-domain read windows, the cells of a test chip were programmed with the same pattern as used in the experiment described in Section 5.1.2. Time scans were recorded at different temperatures using the ramped gate cell biasing scheme. The resulting distributions are shown in Fig. 5.9. The cells states show the expected temperature behavior. Erased cells have a PTC time-domain characteristic, strong programmed cells (P1) have a NTC characteristic and intermediate cells (P0) could be placed near the ZTC point of the sensing mechanism. Cells from state P2 were out of the time scan range. With rising temperature, the time-domain read windows get smaller, as the outer states (E and P1) move towards the inner state (P0). This has to be considered for placement of cell states. Placing one of the states exactly at the ZTC point of the time-domain sensing, has the advantage that read windows do only shrink, but do not move their absolute position with varying temperature. Such a temperature shift of the whole window can be compensated by reference cells like it is done in this work. As they are matched to the regular cells, they experience the same shifts and hence stay in the middle of the read windows over the whole temperature range. However, since initialization and maintenance of the reference cells has high effort, it would be very beneficial to avoid them by use of the ZTC point. As mentioned in Section 4.3.2, the tunable time reference can be used as flexible sense amplifier strobe. When the intermediately programmed state P0 is placed at the ZTC point and the artificially generated reference timings are placed in appropriate distance to the left and right, a sensing scheme without reference cells would be feasible. As the reference timings can be tuned fully digital, an additional temperature compensation of the time strobes could be achieved with

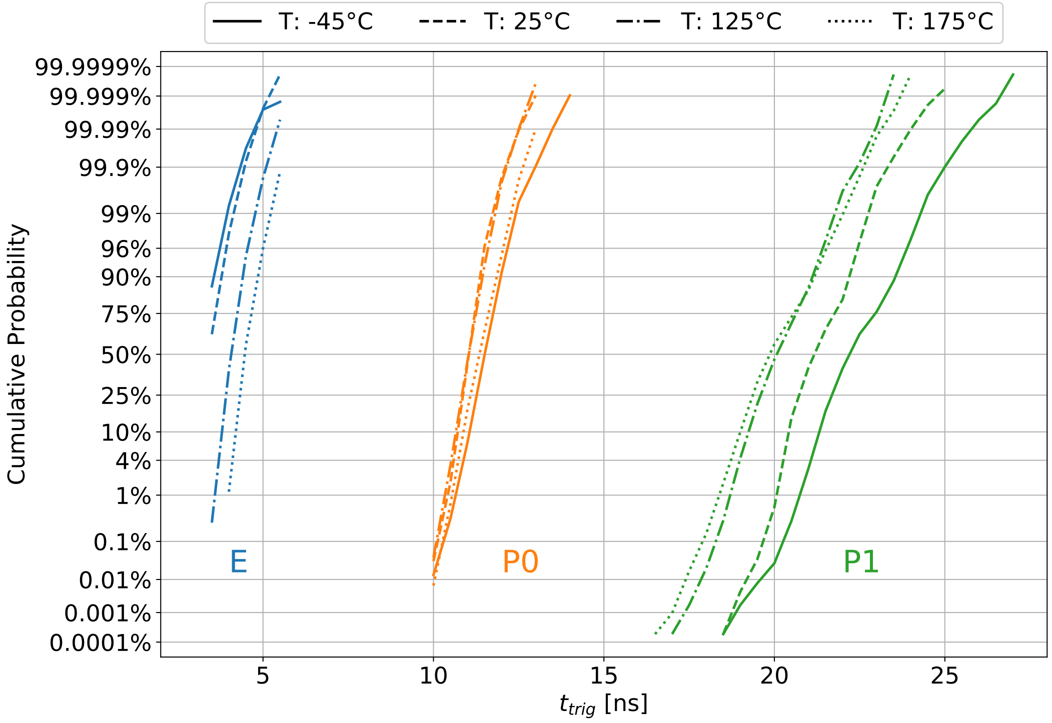


Figure 5.9: Measured time-domain shift due to temperature variation.

an on-chip thermometer.

5.3 Summary

Besides accurate sensing, a precise placement of memory states is key for a robust MLC Flash memory. The cell's programming characteristic and its impact on the sensing mechanism have to be considered thoroughly, when designing a programming algorithm. The ramped gate biasing scheme implemented in this work reduces programming effort. With its linear time-domain transfer characteristic it supports an effective placement of the cell states. Measurements show the advantages of the proposed scheme over conventional cell biasing methods during program verify. An ISPP algorithm using the ramped gate scheme could finish programming with two to three times less programming steps compared to a conventional fixed gate scheme.

Cell-to-cell interference is a limiting factor for MLC operation of Flash memories. Although its impact on embedded NOR Flash cells is less than in a NAND Flash array, it has to be considered for the development of a robust programming algorithm. Measurement reveal a significant V_{TH} shift due to coupling from adjacent cells on the same WL. Measures as imprint programming and the right programming order can mitigate the issue of widening cell distributions caused by inter-cell interference.

The temperature dependence of the cell characteristics has a significant impact on read windows. In time-domain, the shifts caused by temperature variation behave similar as in current-domain. Read windows are closing at high temperatures, but there is a point with ZTC characteristic. It can be used to shape a cell distribution, which is more robust against temperature shifts.

Chapter 6

Conclusion and Outlook

In the present work, different cell biasing schemes for Flash memories were investigated and compared with regard to their suitability for MLC operation. The fixed gate scheme, commonly used for SLC Flash memories, potentially yields the fastest read speed with low complexity sensing schemes. For MLC operation however, it provides less sensing robustness compared to biasing schemes with variable control gate voltage, like stepped gate and ramped gate biasing. Their variable cell operating point provides a much more reliable sense operation at the cost of larger area, higher circuit complexity and lower speed. The ramped gate approach offers a lower programming effort and has less speed penalty, but still results in a robust sensing. Therefore, it proves to be a good candidate for a fast and reliable biasing scheme in an embedded MLC Flash.

The time-domain sensing for MLC operation, introduced in this work, greatly benefits from the ramped gate cell biasing. It improves a state-of-the-art voltage sensing approach [69] by adding an offset cancellation and precharge acceleration mechanisms. The proposed sense amplifier design provides lower mismatch and a better resolution in the programmed cell threshold range. It is suitable for distinction of multiple cell states at high-speed read access.

An integrated ramp generator circuit was developed for the implementation of the ramped gate biasing scheme. It is optimized for highly dynamic ramps and quick recovery from discharge after each sense operation. The strobing concept of the developed memory design is implemented with high flexibility. The reference path allows the usage of a constant reference current, reference cells and a tunable reference time. In this way, the cells can be analyzed with various features in current, voltage, and time domain.

The sensing scheme developed in this work was implemented on a test chip and the claimed improvements were validated. The linear transfer characteristic of the ramped gate cell biasing scheme was shown by measurement. A random read access time of 30 ns within the temperature range from −40 °C to 175 °C was achieved [61].

The placement of the cell states is key for a robust MLC Flash memory. The linear transfer characteristic of the ramped gate biasing scheme improves the programming of multiple cell states in the time-domain. Measurements showed the advantages of the ramped gate scheme over conventional constant gate bias schemes. Either the programming results distributions which are two to three times narrower, or same distribution width is achieved with two to three times less programming pulses. Further, cell-to-cell interference and temperature shifts have to be considered in the design of an efficient programming algorithm. Measures like imprint programming and a placement at the ZTC point of the cells' transfer characteristic should be used to obtain a cell state distribution for robust MLC operation.

It was shown that MLC operation is a viable option to increase memory density and to bring down cost of embedded Flash for automotive usage. However, the higher circuit complexity and the variable voltage biasing scheme increase the development effort and limit the achievable read speed, which is a key performance factor in many automotive applications.

Another promising option for cost reduction is to move to emerging memory technologies such as resistive RAM (RRAM). RRAM is placed in the backend and hence can be easily integrated into existing CMOS logic processes, including FinFET technologies [70]. Its low manufacturing cost and the fairly low operation voltages makes RRAM a good candidate for future embedded non-volatile memories.

Appendix A

Time-Domain Transfer Function for Different Control Gate Biasing

A.1 Derivation for Scenario Without Sense Amplifier Bias Current

(A.1) describes the integration process on the bitline according to Fig. 3.11 and the simplified cell current model (3.13)

$$\frac{dV_{BL}}{dt} = \alpha \left(V_{CG}(t) - V_{BL}(t) - V_{TH,Cell} \right) \quad \text{with} \quad \alpha = \frac{g_{m,Cell}}{C_{BL}}. \tag{A.1}$$

The equation has the form of an ordinary first-order differential equation:

$$y'(t) + f(t)y(t) = g(t) \tag{A.2}$$

$$\begin{aligned} \text{with} \quad y(t) &= V_{BL}(t), \\ f(t) &= \alpha, \\ g(t) &= \alpha \left(V_{CG}(t) - V_{TH,Cell} \right). \end{aligned}$$

It can be solved to [71]

$$y(t) = e^{-F(t)} \left(\int e^{F(t)} g(t) \, dt + C \right) \tag{A.3}$$

$$\text{with} \quad F(t) = \int f(t) \, dt.$$

Backward substitution then yields the general solution of the problem.

$$V_{BL}(t) = e^{-\alpha t} \int \alpha e^{\alpha t} \left(V_{CG}(t) - V_{TH,Cell}\right) dt + Ce^{-\alpha t} \tag{A.4}$$

A.1.1 Fixed Gate Biasing

For fixed gate biasing ($V_{CG}(t) = V_{CG,read}$), (A.4) simplifies significantly and can be solved the following:

$$\begin{aligned} V_{BL}(t) &= ke^{-\alpha t} \int \alpha e^{\alpha t}\, dt + Ce^{-\alpha t} \\ &= ke^{-\alpha t} e^{\alpha t} + Ce^{-\alpha t} \\ &= k + Ce^{-\alpha t} \end{aligned} \tag{A.5}$$

$$\text{with } k = \left(V_{CG,read} - V_{TH,Cell}\right).$$

Using the boundary condition $V_{BL}(t = 0) = 0$, the constant C is eliminated

$$\begin{aligned} 0 &= k + Ce^{-\alpha 0} \\ C &= -k \end{aligned} \tag{A.6}$$

and the BL voltage yields

$$\begin{aligned} V_{BL}(t) &= k\left(1 - e^{-\alpha t}\right) \\ &= \left(V_{CG,read} - V_{TH,Cell}\right) \cdot \left(1 - e^{-\alpha t}\right). \end{aligned} \tag{A.7}$$

To obtain the time-domain transfer function, the crossing point $V_{BL}(t_{trig}) = V_{sense}$ is evaluated:

$$\begin{aligned} V_{sense} &= \left(V_{CG,read} - V_{TH,Cell}\right)\left(1 - e^{-\alpha t_{trig}}\right) \\ e^{-\alpha t_{trig}} &= \left(1 - \frac{V_{sense}}{V_{CG,read} - V_{TH,Cell}}\right) \\ t_{trig} &= -\frac{1}{\alpha} \ln\left(1 - \frac{V_{sense}}{V_{CG,read} - V_{TH,Cell}}\right). \end{aligned} \tag{A.8}$$

A.1.2 Ramped Gate Biasing

For ramped gate biasing ($V_{CG}(t) = mt$), (A.4) can be rearranged to the linear combination

$$\begin{aligned} V_{BL}(t) &= e^{-\alpha t}\left(k_1 \int \alpha e^{\alpha t}\, dt + k_2 \int t e^{\alpha t}\, dt\right) + Ce^{-\alpha t} \\ &= e^{-\alpha t}\left(k_1 e^{\alpha t} + k_2 \frac{1}{\alpha^2} e^{\alpha t}\left(\alpha t - 1\right)\right) + Ce^{-\alpha t} \\ &= k_1 + \frac{k_2}{\alpha^2}\left(\alpha t - 1\right) + Ce^{-\alpha t} \qquad \text{(A.9)} \end{aligned}$$

$$\text{with} \quad k_1 = -V_{TH,Cell}, \quad k_2 = m\alpha.$$

Case I

With the boundary conditions $V_{TH,Cell} < V_{CG}(t = 0) \;\Rightarrow\; V_{BL}(t = 0) = 0$, the constant C is eliminated as follows:

$$\begin{aligned} 0 &= k_1 + \frac{k_2}{\alpha^2}\left(\alpha 0 - 1\right) + Ce^{-\alpha 0} \\ C &= \frac{k_2}{\alpha^2} - k_1. \qquad \text{(A.10)} \end{aligned}$$

According to (A.9) the BL voltage then results:

$$\begin{aligned} V_{BL}(t) &= k_1 + \frac{k_2}{\alpha^2}\left(\alpha t - 1\right) + \left(\frac{k_2}{\alpha^2} - k_1\right) e^{-\alpha t} \\ &= \left(\frac{k_2}{\alpha^2} - k_1\right) \cdot \left(e^{-\alpha t} - 1\right) + \frac{k_2 t}{\alpha} \\ &= \left(\frac{m}{\alpha} + V_{TH,Cell}\right) \cdot \left(e^{-\alpha t} - 1\right) + mt \qquad \text{(A.11)} \end{aligned}$$

For evaluation of this term at $V_{BL}(t_{trig}) = V_{sense}$, the inverse relation of $f(x) = xe^x$ (the Lambert W function or product logarithm) is needed [72]. The term is rearranged and two substitutions are made to solve the problem:

$$\begin{aligned} V_{sense} &= \left(\frac{m}{\alpha} + V_{TH,Cell}\right) \cdot \left(e^{-\alpha t_{trig}} - 1\right) + mt_{trig} \qquad \text{(A.12)} \\ e^{-\alpha t_{trig}} &= 1 + \frac{\alpha V_{sense}}{m + \alpha V_{TH,Cell}} - \frac{m\alpha}{m + \alpha V_{TH,Cell}} t_{trig} \\ &= h + g\, t_{trig} \qquad \text{(A.13)} \end{aligned}$$

$$\text{with}\quad h = 1 + \frac{V_{sense}\,\alpha}{m + \alpha V_{TH,Cell}},$$
$$g = -\frac{m\alpha}{m + \alpha V_{TH,Cell}}.$$

For further steps t_{trig} is replaced by x. The term (A.13) can be rearranged to

$$\begin{aligned} e^{-\alpha x} &= gx + h \\ 1 &= (gx + h)\,e^{\alpha x} \\ \frac{\alpha}{g} &= \left(\alpha x + \alpha\frac{h}{g}\right) e^{\alpha x} \\ \frac{\alpha}{g} e^{\alpha\frac{h}{g}} &= \left(\alpha x + \alpha\frac{h}{g}\right) e^{\alpha x + \alpha\frac{h}{g}}. \end{aligned} \tag{A.14}$$

For typical numbers of α, m and $V_{TH,Cell}$ both sides stay within the domain of Lambert W function, which is $[-e^{-1}, \infty[$. Therefore, it can be applied and the term can be solved for x:

$$W\left(\frac{\alpha}{g} e^{\alpha\frac{h}{g}}\right) = \alpha x + \alpha\frac{h}{g} \tag{A.15}$$

$$x = \frac{1}{\alpha} W\left(\frac{\alpha}{g} e^{\alpha\frac{h}{g}}\right) - \frac{h}{g}. \tag{A.16}$$

Backwards substitution yields

$$\begin{aligned} t_{trig} = \frac{1}{\alpha}&\left(W\left(-\left(1 + \alpha\frac{V_{TH,Cell}}{m}\right) e^{-\left(1 + \alpha\frac{V_{sense} + V_{TH,Cell}}{m}\right)}\right) + 1\right), \\ &+ \frac{V_{sense} + V_{TH,Cell}}{m} \end{aligned} \tag{A.17}$$

which is the first branch of time-domain transfer function with ramped control gate bias.

Case II

The boundary conditions for the second case are:

$$\begin{aligned} V_{TH,Cell} \geq V_{CG}(t = 0) \quad &\Rightarrow \quad V_{BL}(t \leq t_0) = 0, \\ V_{CG}(t_0) = V_{TH,Cell} \quad &\Rightarrow \quad t_0 = \frac{V_{TH,Cell}}{m}. \end{aligned} \tag{A.18}$$

The constant C is eliminated

$$\begin{aligned} 0 &= k_1 + \frac{k_2}{\alpha^2}\left(\alpha t_0 - 1\right) + Ce^{-\alpha t_0} \\ C &= -e^{\alpha t_0}\left(k_1 + \frac{k_2}{\alpha^2}\left(\alpha t_0 - 1\right)\right) \\ &= -e^{\alpha t_0}\left(-V_{TH,Cell} + \frac{m\alpha}{\alpha^2}\left(\alpha \frac{V_{TH,Cell}}{m} - 1\right)\right) \\ &= \frac{m}{\alpha}e^{\alpha t_0} \end{aligned} \tag{A.19}$$

and the term for the bitline voltage results from (A.9) as

$$\begin{aligned} V_{BL}(t) &= -V_{TH,Cell} + \frac{m\alpha}{\alpha^2}\left(\alpha t - 1\right) + \frac{m}{\alpha}e^{\alpha t_0}e^{-\alpha t} \\ &= \frac{m}{\alpha}\left(e^{-\alpha(t-t_0)} - 1\right) + m\left(t - t_0\right). \end{aligned} \tag{A.20}$$

The evaluation of $V_{BL}(t_{trig}) = V_{sense}$ is simplified by substitution of $\hat{t} = t_{trig} - t_0$. As in (A.12) and (A.13), the equation is brought to the form

$$e^{-\alpha\hat{t}} = h + g\hat{t} \tag{A.21}$$

$$\begin{aligned} \text{with} \quad h &= \frac{V_{sense}\alpha}{m} + 1, \\ g &= -\alpha. \end{aligned}$$

As in case I, $\hat{t}$ is substituted by x and according to (A.14) to (A.16) the solution is found as

$$x = \frac{1}{\alpha}W\left(\frac{\alpha}{g}e^{\alpha\frac{h}{g}}\right) - \frac{h}{g} \tag{A.22}$$

The second branch of the time-domain transfer function is obtained by backward substitution as

$$\begin{aligned} \hat{t} &= \frac{1}{\alpha}W\left(\frac{\alpha}{g}e^{\alpha\frac{h}{g}}\right) - \frac{h}{g} \\ &= \frac{1}{\alpha}\left(W\left(-e^{-\left(1+\alpha\frac{V_{sense}}{m}\right)}\right) + 1\right) + \frac{V_{sense}}{m}, \end{aligned} \tag{A.23}$$

$$\begin{aligned} t_{trig} &= \frac{1}{\alpha}\left(W\left(-e^{-\left(1+\alpha\frac{V_{sense}}{m}\right)}\right) + 1\right) + \frac{V_{sense}}{m} + t_0 \\ &= \frac{1}{\alpha}\left(W\left(-e^{-\left(1+\alpha\frac{V_{sense}}{m}\right)}\right) + 1\right) + \frac{V_{sense} + V_{TH,Cell}}{m}. \end{aligned} \tag{A.24}$$

A.2 Derivation for Scenario With Sense Amplifier Bias Current

With the sense amplifier bias current I_{bias} added to the differential equation (4.20), the integration process on the bitline is described by

$$\frac{dV_{BL}}{dt} = \alpha \left(V_{CG} - V_{BL} - V_{TH,Cell}\right) + \beta \quad \text{with } \alpha = \frac{g_{m,Cell}}{C_{BL}}; \beta = \frac{I_{bias}}{C_{BL}}. \tag{A.25}$$

Analogous to (A.1) to (A.4), (A.25) can be solved to

$$V_{BL}(t) = e^{-\alpha t} \int \alpha e^{\alpha t} \left(V_{CG}(t) - V_{TH,Cell} + \frac{\beta}{\alpha}\right) dt + Ce^{-\alpha t}. \tag{A.26}$$

A.2.1 Fixed Gate Biasing

For constant CG bias ($V_{CG}(t) = V_{CG,read}$), (A.26) is solved in the same way as (A.4), according to (A.5) to (A.7) with the only difference in the substitution

$$k^* = \left(V_{CG,read} - V_{TH,Cell} + \frac{\beta}{\alpha}\right).$$

The bitline voltage V_{BL} is now described by

$$V_{BL}(t) = \left(V_{CG,read} - V_{TH,Cell} + \frac{\beta}{\alpha}\right) \cdot \left(1 - e^{-\alpha t}\right), \tag{A.27}$$

which can be solved for $V_{BL}\left(t_{trig}\right) = V_{sense}$ as

$$t_{trig}(V_{TH,Cell}) = -\frac{1}{\alpha} \cdot \ln\left(1 - \frac{V_{sense}}{V_{CG,read} - V_{TH,Cell} + \frac{\beta}{\alpha}}\right). \tag{A.28}$$

The model (A.28) is only valid for threshold voltages $V_{TH,Cell}$ that are smaller than $V_{CG,read} - V_{sense} + \beta/\alpha$. The pole at this position yields infinite trigger times t_{trig} which is not reasonable, since the sense amplifier bias current I_{bias} will eventually charge the bitline up to the sensing level V_{sense}. The inconsistency of the model results from the cell current model I_{Cell}, which was only considered for the conducting branch $V_{GS} > V_{TH,Cell}$.

A.2.2 Ramped Gate Biasing

For ramped gate biasing ($V_{CG}(t) = mt$) (A.26) can be rearranged analogous to (A.9). The solution differs only in the substitution. The result is

$$V_{BL}(t) = k_1^* + \frac{k_2}{\alpha^2}(\alpha t - 1) + Ce^{-\alpha t} \tag{A.29}$$

$$\begin{aligned} \text{with} \quad k_1^* &= \frac{\beta}{\alpha} - V_{TH,Cell}, \\ k_2 &= m\alpha. \end{aligned}$$

As in Appendix A.1.2, the boundary conditions of two cases have to be considered.

Case I

The boundary conditions are $V_{TH,Cell} < V_{CG}(t = 0) \;\Rightarrow\; V_{BL}(t = 0) = 0$. According to (A.10) and (A.11), (A.29) is solved for the bitline voltage as

$$V_{BL}(t) = \left(\frac{m}{\alpha} + V_{TH,Cell} - \frac{\beta}{\alpha}\right) \cdot \left(e^{-\alpha t} - 1\right) + mt. \tag{A.30}$$

Analogous to (A.12) to (A.17), (A.30) can be solved for the trigger time t_{trig}. The made substitutions are

$$\begin{aligned} h &= 1 + \frac{V_{sense}\,\alpha}{m + \alpha V_{TH,Cell} - \beta}, \\ g &= -\frac{m\alpha}{m + \alpha V_{TH,Cell} - \beta}, \\ \eta &= \frac{m - \beta}{m}. \end{aligned} \tag{A.31}$$

The solution is found as

$$\begin{aligned} t_{trig} = {} & \frac{1}{\alpha}\left(W\left(-\left(\eta + \alpha\frac{V_{TH,Cell}}{m}\right)e^{-\left(\eta + \alpha\frac{V_{sense}+V_{TH,Cell}}{m}\right)}\right) + \eta\right) \\ & + \frac{V_{sense} + V_{TH,Cell}}{m}. \end{aligned} \tag{A.32}$$

Case II

Considering the sense amplifier bias current I_{bias}, the boundary conditions of the second case look different. As long as the control gate voltage V_{CG} stays below the cell's threshold voltage $V_{TH,Cell}$, the cell current stays zero, but I_{bias} charges up the bitline. The moment when the cell current starts to flow (t_0^*) is slightly delayed compared to the situation without I_{bias}, as the source node of the cell (V_{BL}) rises slowly. The new boundary conditions for the second case are:

$$\begin{aligned}
&V_{TH,Cell} \geq V_{CG}(t=0),\\
&V_{BL}(t_0^*) = \beta \cdot t_0^*,\\
&V_{CG}(t_0^*) = V_{BL}(t_0^*) + V_{TH,Cell}\\
&\quad m \cdot t_0^* = \frac{I_{bias}}{C_{BL}} t_0^* + V_{TH,Cell} = \beta \cdot t_0^* + V_{TH,Cell},\\
&\Rightarrow \quad t_0^* = \frac{V_{TH,Cell}}{m-\beta}.
\end{aligned}$$

According to the boundary condition $V_{BL}(t_0^*) = \beta \cdot t_0^*$, (A.29) is solved for the bitline voltage as

$$V_{BL}(t) = \left(\frac{m}{\alpha} - \frac{\beta}{\alpha}\right)\left(e^{-\alpha\left(t-t_0^*\right)} - 1\right) + m\left(t - t_0^*\right) + t_0^*\beta. \tag{A.33}$$

Analogous to (A.12) to (A.17) and (A.21) to (A.22), (A.33) can be solved for the trigger time t_{trig}. The made substitutions are

$$\begin{aligned}
h &= 1 + \frac{\alpha}{m-\beta}\left(V_{sense} - t_0^*\beta\right)\\
g &= -\frac{m}{m-\beta}\alpha\\
\eta &= \frac{m-\beta}{m}.
\end{aligned} \tag{A.34}$$

The solution is found as

$$\begin{aligned}
t_{trig} = {} & \frac{1}{\alpha} \cdot \left(W\left(-\eta e^{-\left(\eta + \frac{\alpha}{m}\left(V_{sense} + V_{TH,Cell}\left(1 - \frac{1}{\eta}\right)\right)\right)}\right) + \eta\right)\\
& + \frac{V_{sense} + V_{TH,Cell}}{m}.
\end{aligned} \tag{A.35}$$

Interpretation

The solutions derived from the differential equation which considers the sense amplifier bias current I_{bias} ((A.32) and (A.35)) look similar to the former solutions ((A.17) and (A.24)). The linear characteristic of the second branch (Case II) is still apparent, but the slope is decreased.

Acronyms

1T cell 1-transistor cell

2T cell 2-transistor cell

ADAS Advanced Driver Assistance Systems

BL bitline

CBSD current-biased slope detection

CG control gate

CHE channel hot-electron

DIBL drain-induced barrier lowering

EEPROM electrically erasable programmable read only memory

eFlash embedded Flash

EPROM erasable programmable read only memory

FG floating gate

FN Fowler-Nordheim

GBL global bitline

GBW gain-bandwidth product

HV high voltage

IoT Internet of Things

ISPP incremental step programming pulse

LBL local bitline

MCU microcontroller unit

MLC multi-level cell

NTC negative temperature coefficient

ONO oxide-nitride-oxide

OTA operational transconductance amplifier

PM phase margin

PTC positive temperature coefficient

QLC quad-level cell

RRAM resistive RAM

RTS random telegraph signal

SA sense amplifier

SLC single-level cell

SS-CHE source-side channel hot-electron

TLC triple-level cell

UV ultraviolet

VBSD voltage-biased slope detection

WL wordline

ZTC zero temperature coefficient

List of Figures

List of Tables

Publications of the Author

S.Kiesel, T.Kern, and B.Wicht. ”Time-domain Ramped Gate Sensing for Embedded Multi-level Flash in Automotive Applications”. *2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*. IEEE, 2017, pp. 691-694

S.Kiesel, T.Kern, B.Wicht, and H. Graeb.”A 30 ns 16 Mb 2 b/cell Embedded Flash with Ramped Gate Time-Domain Sensing Scheme for Automotive Application”. *2019 International Symposium on VLSI Design, Automation and Test (VLSI-DAT)*. IEEE, 2019, pp. 1-4

Bibliography

[1] F. Masuoka, M. Asano, H. Iwahashi, T. Komuro, and S. Tanaka. “A new flash E2PROM cell using triple polysilicon technology”. *1984 International Electron Devices Meeting*. Vol. 30. IRE, 1984, pp. 464–467.

[2] M. Bauer, R. Alexis, G. Atwood, B. Baltar, A. Fazio, K. Frary, M. Hensel, M. Ishac, J. Javanifard, M. Landgraf, D. Leak, K. Loe, D. Mills, P. Ruby, R. Rozman, S. Sweha, S. Talreja, and K. Wojciechowski. “A multilevel-cell 32 Mb flash memory”. *Solid-State Circuits Conference, 1995. Digest of Technical Papers. 41st ISSCC, 1995 IEEE International* (1995), pp. 132–133.

[3] G. Campardo, R. Micheloni, S. Commodaro, E. Yero, M. Zammattio, S. Mognoni, A. Sacco, M. Picca, A. Manstretta, M. Scotti, I. Motta, C. Golla, A. Pierin, A. Ohba, T. Futatsuya, R. Makabe, S. Kawai, Y. Kai, S. Shimizu, T. Ohnakado, T. Sugihara, R. Bez, A. Grossi, A. Modelli, O. Khouri, and G. Torelli. “A 40 mm/sup 2/ 3 V 50 MHz 64 Mb 4-level cell NOR-type flash memory”. *Solid-State Circuits Conference, 2000. Digest of Technical Papers. ISSCC. 2000 IEEE International* (2000), pp. 274–275.

[4] H. Castro, K. Augustine, S. Balasubrahmanyam, T. Bressie, S. Chandramouli, G. Christensen, M. Dayley, D. Elmhurst, K. Fan, M. Goldman, C. Haid, R. Haque, M. Ishac, M. Khandaker, J. Kreifels, B. Li, K. Loe, T. Ly, F. Marvin, R. Melcher, S. Monasa, R. Nambiar, Q. Ngo, R. Padilla, B. Pathak, A. Rahman, R. Rajagopal, K. Ramamurthi, S. Saini, A. Sayed, I. Sharif, B. Srinivasan, M. Szwarc, G. Vadlamudi, V. Viajedor, and R. Zeng. “A 125MHz burst mode 0.18μm 128Mbit 2 bits per cell flash memory”. *2002 Symposium on VLSI Circuits. Digest of Technical Papers (Cat. No.02CH37302).* IEEE, 2002, pp. 304–307.

[5] D. Elmhurst and M. Goldman. “A 1.8-V 128-Mb 125-MHz multilevel cell memory with flexible read while write”. *IEEE Journal of Solid-State Circuits* 38.11 (2003), pp. 1929–1933.

[6] M. Bauer and K. Tedrow. “A scalable stepped gate sensing scheme for sub-100nm multilevel flash memory”. *2005 International Conference on Integrated Circuit Design and Technology, 2005. ICICDT 2005.* IEEE, 2005, pp. 23–26.

[7] C. Villa, D. Vimercati, S. Schippers, E. Confalonieri, M. Sforzin, S. Polizzi, M. La Placa, C. Lisi, A. Magnavacca, E. Bolandrina, A. Martinelli, V. Dima, A. Scavuzzo, B. Calandrino, N. Gatto, M. Scardaci, F. Mastroianni, M. Pisasale, A. Geraci, M. Gaibotti, and M. Sali. “A 125MHz burst-mode flexible readwhile-write 256Mbit 2b/c 1.8V NOR flash memory”. *ISSCC. 2005 IEEE International Digest of Technical Papers. Solid-State Circuits Conference, 2005.* IEEE, 2005, pp. 52–54.

[8] R.-A. Cernea, L. Pham, F. Moogat, S. Chan, B. Le, Y. Li, S. Tsao, T.-Y. Tseng, K. Nguyen, J. Li, J. Hu, J. H. Yuh, C. Hsu, F. Zhang, T. Kamei, H. Nasu, P. Kliza, K. Htoo, J. Lutze, Y. Dong, M. Higashitani, J. Yang, H.-s. Lin, V. Sakhamuri, A. Li, F. Pan, S. Yadala, S. Taigor, K. Pradhan, J. Lan, J. Chan, T. Abe, Y. Fukuda, H. Mukai, K. Kawakami, C. Liang, T. Ip, S.-F. Chang, J. Lakshmipathi, S. Huynh, D. Pantelakis, S. Member, M. Mofidi, and K. Quader. “A 34 MB/s MLC Write Throughput 16 Gb NAND With All Bit Line Architecture on 56 nm Technology”. *IEEE Journal of Solid-State Circuits* 44.1 (2009), pp. 186–194.

[9] S.-H. Shin, D.-K. Shim, J.-Y. Jeong, O.-S. Kwon, S.-Y. Yoon, T.-Y. Kim, H.-W. Park, H.-J. Yoon, Y.-S. Song, Y.-H. Choi, Y.-L. Ahn, K.-T. Park, J.-M. Han, K.-H. Kyung, and Y.-H. Jun. “A New 3-bit Programming Algorithm using SLC-to-TLC Migration for 8MB / s High Performance TLC NAND Flash Memory”. *2012 Symposium on VLSI Circuits (VLSIC)* 6 (2012), pp. 132–133.

[10] Y. S. Cho, I. H. Park, S. Y. Yoon, N. H. Lee, S. H. Joo, K.-W. Song, K. Choi, J.-M. Han, K. H. Kyung, and Y.-H. Jun. “Adaptive Multi-Pulse Program Scheme Based on Tunneling Speed Classification for Next Generation Multi-Bit/Cell NAND FLASH”. *IEEE Journal of Solid-State Circuits* 48.4 (2013), pp. 948–959.

[11] R. Strenz. “Embedded Flash technologies and their applications: Status & outlook”. *2011 International Electron Devices Meeting.* IEEE, 2011, pp. 9.4.1–9.4.4.

[12] T. Yamauchi, Y. Yamaguchi, T. Kono, and H. Hidaka. “Embedded flash technology for automotive applications”. *2016 IEEE International Electron Devices Meeting (IEDM).* IEEE, 2016, pp. 28.6.1–28.6.4.

[13] T. Jew. "Embedded Microcontroller Memories: Application Memory Usage". *2015 IEEE International Memory Workshop (IMW).* Vol. 129. IEEE, 2015, pp. 1–4.

[14] Y. Taito, T. Kono, M. Nakano, T. Saito, T. Ito, and K. Noguchi. "A 28 nm Embedded Split-Gate MONOS (SG-MONOS) Flash Macro for Automotive Achieving 6.4 GB/s Read Throughput by 200 MHz No-Wait Read Operation and 2.0 MB/s Write Throughput at Tj of 170C". *IEEE Journal of Solid-State Circuits* 51.1 (2016), pp. 213–221.

[15] D. Kahng and S. M. Sze. "A Floating Gate and Its Application to Memory Devices". *Bell System Technical Journal* 46.6 (1967), pp. 1288–1295.

[16] J. E. Brewer and M. Gill. *Nonvolatile Memory Technologies with Emphasis on Flash.* Wiley-IEEE Press, 2008, p. 792.

[17] H. Hidaka. "Evolution of embedded flash memory technology for MCU". *2011 IEEE International Conference on IC Design & Technology.* IEEE, 2011, pp. 1–4.

[18] H. Kurata, K. Otsuga, A. Kotabe, S. Kajiyama, T. Osabe, Y. Sasago, S. Narumi, K. Tokami, S. Kamohara, and O. Tsuchiya. "Random telegraph signal in flash memory: Its impact on scaling of multilevel flash memory beyond the 90-nm node". *IEEE Journal of Solid-State Circuits* 42.6 (2007), pp. 1362–1369.

[19] B. Eitan, P. Pavan, I. Bloom, E. Aloni, A. Frommer, and D. Finzi. "NROM: A novel localized trapping, 2-bit nonvolatile memory cell". *IEEE Electron Device Letters* 21.11 (2000), pp. 543–545.

[20] Hang-Ting Lue, Tzu-Hsuan Hsu, Min-Ta Wu, Kuang-Yeu Hsieh, R. Liu, and Chih-Yuan Lu. "Studies of the reverse read method and second-bit effect of 2-bit/cell nitride-trapping device by quasi-two-dimensional model". *IEEE Transactions on Electron Devices* 53.1 (2006), pp. 119–125.

[21] Y. Polansky, A. Lavan, R. Sahar, O. Dadashev, Y. Betser, G. Cohen, E. Maayan, B. Eitan, N. Ful-Long, K. Yen-Hui Joseph, L. Chih-Yuan, C. Tim Chang-Ting, L. Chun-Yu, C. Chin-Hung, C. Chung-Kuang, H. Wen-Chiao, S. Yite, T. WenChi, and L. Wenpin. "A 4b/cell NROM 1Gb Data-Storage Memory". *Solid-State Circuits Conference, 2006. ISSCC 2006. Digest of Technical Papers. IEEE International* (2006), pp. 448–458.

[22] B. Eitan, G. Cohen, A. Shappir, E. Lusky, A. Givant, M. Janai, I. Bloom, Yan Polansky, O. Dadashev, A. Lavan, Ran Sahar, and E. Maayan. "4-bit per cell NROM reliability". *IEEE InternationalElectron Devices Meeting, 2005. IEDM Technical Digest.* Vol. 00. c. IEEE, 2005, pp. 539–542.

[23] T. Kono, T. Ito, T. Tsuruda, T. Nishiyama, T. Nagasawa, T. Ogawa, Y. Kawashima, H. Hidaka, and T. Yamauchi. "40-nm Embedded Split-Gate MONOS (SG-MONOS) Flash Macros for Automotive With 160-MHz Random Access for Code and Endurance Over 10 M Cycles for Data at the Junction Temperature of 170 °C". *IEEE Journal of Solid-State Circuits* 49.1 (2014), pp. 154–166.

[24] J. Javanifard, T. Tanadi, H. Giduturi, K. Loe, R. L. Melcher, S. Khabiri, N. T. Hendrickson, A. D. Proescholdt, D. A. Ward, and M. A. Taylor. "A 45nm Self-Aligned-Contact Process 1Gb NOR Flash with 5MB/s Program Speed". *2008 IEEE International Solid-State Circuits Conference - Digest of Technical Papers.* Vol. 51. IEEE, 2008, pp. 424–624.

[25] Y. Sofer, M. Edan, Y. Betser, M. Grossgold, E. Maayan, and B. Eitan. "A 55 mm/sup 2/ 256 Mb NROM flash memory with embedded microcontroller using an NROM-based program file ROM". *2004 IEEE International Solid-State Circuits Conference (IEEE Cat. No.04CH37519).* IEEE, 2004, pp. 48–512.

[26] Seungjae Lee, Young-Taek Lee, Wook-Kee Han, Dong-Hwan Kim, Moo-Sung Kim, Seung-Hyun Moon, Hyun Chul Cho, Jung-Woo Lee, Dae-Seok Byeon, Young-Ho Lim, Hyung-Suk Kim, Sung-Hoi Hur, and Kang-Deog Suh. "A 3.3 V 4 Gb four-level NAND flash memory with 90 nm CMOS technology". *2004 IEEE International Solid-State Circuits Conference (IEEE Cat. No.04CH37519).* Vol. 36. 11. IEEE, 2004, pp. 52–513.

[27] E. Seevinck, P. van Beers, and H. Ontrop. "Current-mode techniques for high-speed VLSI circuits with application to current sense amplifier for CMOS SRAM's". *IEEE Journal of Solid-State Circuits* 26.4 (1991), pp. 525–536.

[28] R. Micheloni, L. Crippa, M. Sangalli, and G. Campardo. "The flash memory read path: building blocks and critical aspects". *Proceedings of the IEEE* 91.4 (2003), pp. 537–553.

[29] B. Pathak, A. Cabrera, G. Christensen, A. Darwish, M. Goldman, R. Haque, J. Jorgensen, R. Kajley, T. Ly, F. Marvin, S. Monasa, Q. Nguyen, D. Pierce, A. Sendrowski, I. Sharif, H. Shimoyoshi, A. Smidt, R. Sundaram, M. Taub, W. Tran, R. Trivedi, P. Walimbe, and E. Yu. "A 1.8 V 64 Mb 100 MHz flexible read while write flash memory [in CMOS]". *2001 IEEE International Solid-State Circuits Conference. Digest of Technical Papers. ISSCC (Cat. No.01CH37177).* IEEE, 2001, pp. 32–33,

[30] S. Dietrich, M. Angerbauer, M. Ivanov, D. Gogl, H. Hoenigschmid, M. Kund, C. Liaw, M. Markert, R. Symanczyk, L. Altimime, S. Bournat, and G. Mueller. "A Nonvolatile 2-Mbit CBRAM Memory Core Featuring Advanced Read and Program Control". *IEEE Journal of Solid-State Circuits* 42.4 (2007), pp. 839–845.

[31] T. Ogura, M. Hosoda, T. Ogawa, T. Kato, A. Kanda, T. Fujisawa, S. Shimizu, and M. Katsumata. "A 1.8-V 256-Mb Multilevel Cell NOR Flash Memory With BGO Function". *IEEE Journal of Solid-State Circuits* 41.11 (2006), pp. 2589–2600.

[32] B. Wicht. *Current Sense Amplifiers: For Embedded SRAM in High-Performance System-on-a-Chip Designs.* 12th ed. Berlin: Springer, 2003, p. 164.

[33] B. Goll and Z. Horst. *Comparators in Nanometer CMOS Technology.* 50th ed. Berlin: Springer Berlin Heidelberg, 2015.

[34] P. E. Allen and D. R. Holberg. *CMOS Analog Circuit Design.* New York: Oxford University Press.

[35] 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 (2008), pp. 132–140.

[36] D. Miyashita, R. Yamaki, K. Hashiyoshi, H. Kobayashi, S. Kousai, Y. Oowaki, and Y. Unekawa. "An LDPC decoder with time-domain analog and digital mixed-signal processing". *IEEE Journal of Solid-State Circuits* 49.1 (2014), pp. 73–83.

[37] M. Momodomi, T. Tanaka, Y. Iwata, Y. Tanaka, H. Oodaira, Y. Itoh, R. Shirota, K. Ohuchi, and F. Masuoka. "A 4 Mb NAND EEPROM with tight programmed V/sub t/ distribution". *IEEE Journal of Solid-State Circuits* 26.4 (1991), pp. 492–496.

[38] T. Tanaka, Y. Tanaka, H. Nakamura, H. Oodaira, S. Aritome, R. Shirota, and F. Masuoka. "A quick intelligent program architecture for 3 V-only NAND-EEPROMs". *1992 Symposium on VLSI Circuits Digest of Technical Papers.* Vol. 26. 4. IEEE, 1992, pp. 20–21.

[39] T. Tanaka, M. Kato, T. Adachi, K. Ogura, K. Kimura, and H. Kume. "High-Speed Programming and Program-Verify Methods Suitable for Low-Voltage Flash Memories". *Proceedings of 1994 IEEE Symposium on VLSI Circuits.* Vol. 5. IEEE, 1994, pp. 61–62.

[40] K. D. Suh, B. H. Suh, Y. H. Lim, J. K. Kim, Y. J. Choi, Y. N. Koh, S. S. Lee, S. C. Kwon, B. S. Choi, J. S. Yum, J. H. Choi, J. R. Kim, and H. K. Lim. "A 3.3 V 32 Mb NAND Flash Memory with Incremental Step Pulse Programming Scheme". *IEEE Journal of Solid-State Circuits* 30.11 (1995), pp. 1149–1156.

[41] H. Kurata, K. Otsuga, a. Kotabe, S. Kajiyama, T. Osabe, Y. Sasago, S. Narumi, K. Tokami, S. Kamohara, and O. Tsuchiya. "The Impact of Random Telegraph Signals on the Scaling of Multilevel Flash Memories". *2006 Symposium on VLSI Circuits, 2006. Digest of Technical Papers.* 00.c (2006), pp. 105–106.

[42] D. Richter. *Flash Memories: Economic Principles of Performance, Cost and Reliability Optimization.* 40th ed. Dodrecht: Springer, 2014, p. 268.

[43] C. Trinh, N. Shibata, T. Nakano, M. Ogawa, J. Sato, Y. Takeyama, K. Isobe, B. Le, F. Moogat, N. Mokhlesi, K. Kozakai, P. Hong, T. Kamei, K. Iwasa, J. Nakai, T. Shimizu, M. Honma, S. Sakai, T. Kawaai, S. Hoshi, J. Yuh, C. Hsu, T. Tseng, J. Li, J. Hu, M. Liu, S. Khalid, J. Chen, M. Watanabe, H. Lin, J. Yang, K. McKay, K. Nguyen, T. Pham, Y. Matsuda, K. Nakamura, K. Kanebako, S. Yoshikawa, W. Igarashi, A. Inoue, T. Takahashi, Y. Komatsu, C. Suzuki, K. Kanazawa, M. Higashitani, S. Lee, T. Murai, K. Nguyen, J. Lan, S. Huynh, M. Murin, M. Shlick, M. Lasser, R. Cernea, M. Mofidi, K. Schuegraf, and K. Quader. "A 5.6MB/s 64Gb 4b/Cell NAND flash memory in 43nm CMOS". *Digest of Technical Papers - IEEE International Solid-State Circuits Conference* (2009), pp. 246–248.

[44] K. Seki, H. Kume, Y. Ohji, T. Tanaka, T. Adachi, M. Ushiyama, K. Shimohigashi, T. Wada, K. Komori, T. Nishimoto, K. Izawa, T. Hagiwara, Y. Kubota, K. Shohji, N. Miyamoto, S. Saeki, and N. Ogawa. "An 80 ns 1 Mb flash memory with on-chip erase/erase-verify controller". *1990 37th IEEE International Conference on Solid-State Circuits.* Vol. 1. IEEE, 1990, pp. 60–61.

[45] N. Do. "eNVM Technologies Scaling Outlook and Emerging NVM Technologies for Embedded Applications". *2016 IEEE 8th International Memory Workshop (IMW).* IEEE, 2016, pp. 1–4.

[46] D. Elmhurst, R. Bains, T. Bressie, C. Bueb, E. Carrieri, B. Chauhan, N. Chrisman, M. Dayley, R. De Luna, K. Fan, M. Goldman, P. Govindu, A. Huq, M. Khandaker, J. Kreifels, S. Krishnamachari, P. Lavapie, K. Loe, T. Ly, F. Marvin, R. Melcher, S. Monasa, Q. Nguyen, B. Pathak, A. Proescholdt, T. Rahman, B. Srinivasan, R. Sundaram, P. Walimbe, D. Ward, D. Zeng, and H. Zhang. "A 1.8 V 128 Mb 125 MHz multi-level cell flash memory with

flexible read while write". *2003 IEEE International Solid-State Circuits Conference, 2003. Digest of Technical Papers. ISSCC.* Vol. 1. July 2002. IEEE, 2003, pp. 286–287.

[47] Y. S. Chu, Y. H. Wang, C. Y. Wang, Y. H. Lee, A. C. Kang, R. Ranjan, W. T. Chu, T. C. Ong, H. W. Chin, and K. Wu. "Split-Gate Flash Memory for Automotive Embedded Applications". *International Reliability Physics Symposium.* 2011, pp. 636–640.

[48] M. Taub, R. Bains, G. Barkley, H. Castro, G. Christensen, S. Eilert, R. Fackenthal, H. Giduturi, M. Goldman, C. Haid, R. Haque, K. Parat, S. Peterson, A. Proescholdt, K. Ramamurthi, P. Ruby, B. Sivakumar, A. Smidt, B. Srinivasan, M. Szwarc, K. Tedrow, and D. Young. "A 90nm 512mb 166MHz multilevel cell flash memory with 1.5MByte/s programming". *ISSCC. 2005 IEEE International Digest of Technical Papers. Solid-State Circuits Conference, 2005.* IEEE, 2005, pp. 54–56.

[49] Vasant B. Rao. "Delay Analysis of the Distributed RC Line". *32nd Design Automation Conference.* 7. ACM, 1995, pp. 370–375.

[50] M. Jefremow. "Power Efficient and Robust Sense Amplifiers for Embedded Non-Volatile Memories in High-Speed Microcontrollers for Automotive Applications". PhD thesis. Technische Universität München, 2014, p. 123.

[51] A. Agarwal and J. Lang. *Foundation of Analog and Digital Electronic Circuits.* Elsevier Science and Technology, 2005, p. 1009.

[52] G. Li. *Signals and Systems: Fundamentals.* Berlin: De Gruyter, 2015.

[53] M. Jefremow, T. Kern, U. Backhausen, J. Elbs, B. Rousseau, C. Roll, L. Castro, T. Roehr, E. Paparisto, K. Herfurth, R. Bartenschlager, S. Thierold, R. Renardy, S. Kassenetter, N. Lawal, M. Strasser, W. Trottmann, and D. Schmitt-Landsiedel. "A 65nm 4MB embedded flash macro for automotive achieving a read throughput of 5.7GB/s and a write throughput of 1.4MB/s". *2013 Proceedings of the ESSCIRC (ESSCIRC).* IEEE, 2013, pp. 193–196.

[54] S. Kiesel, T. Kern, and B. Wicht. "Time-Domain Ramped Gate Sensing for Embedded Multi-level Flash in Automotive Applications". *2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS).* IEEE, 2017, pp. 691–694.

[55] M.-F. Chang, S.-J. Shen, C.-C. Liu, C.-W. Wu, Y.-F. Lin, Y.-C. King, C.-J. Lin, H.-J. Liao, Y.-D. Chih, and H. Yamauchi. "An Offset-Tolerant Fast-Random-Read Current-Sampling-Based Sense Amplifier for Small-Cell-Current Nonvolatile Memory". *IEEE Journal of Solid-State Circuits* 48.3 (2013), pp. 864–877.

[56] W. M. C. Sansen. *Analog Design Essentials.* 857th ed. Vol. 859. Dodrecht: Springer, 2006, p. 777.

[57] B. Razavi. *Design of Analog CMOS Integrated Circuits.* Boston: McGraw-Hill, 2001, p. 684.

[58] B. Razavi. *RF Microelectronics.* Upper Saddle River, NJ: Prentice Hall, 1998.

[59] Sang-Pil Sim, Wook Hyun Kwon, Heon Kyu Lee, Jee Hoon Han, Seung Boo Jeon, Bong Yong Lee, Jae Hoon Kim, Jung In Han, Byoung Moon Yoon, Wook H. Lee, Chan-Kwang Park, and Kinam Kim. "Anomalous charge loss of reference cell in MLC flash memory due to process-induced mobile ion". *Proceedings of 35th European Solid-State Device Research Conference, 2005. ESSDERC 2005.* Vol. 2005. IEEE, 2005, pp. 321–324.

[60] J. H. Davies. *MSP430 microcontroller basics.* Burlington: Newnes, 2008, p. 688.

[61] S. Kiesel, T. Kern, B. Wicht, and H. Graeb. "A 30 ns 16 Mb 2 b/cell Embedded Flash with Ramped Gate Time-Domain Sensing Scheme for Automotive Application". *2019 International Symposium on VLSI Design, Automation and Test (VLSI-DAT).* IEEE, 2019, pp. 1–4.

[62] P. Pavan, L. Larcher, and A. Marmiroli. *Floating Gate Devices: Operation and Compact Modeling.* 1st ed. Boston: Kluwer Academic Publishers, 2004.

[63] P. Dimitrakis. *Charge-Trapping Non-Volatile Memories.* Ed. by P. Dimitrakis. 1st ed. Cham: Springer International Publishing, 2015.

[64] H. Hidaka. *Embedded Flash Memory for Embedded Systems: Technology, Design for Sub-systems, and Innovations.* Ed. by H. Hidaka. Integrated Circuits and Systems. Cham: Springer International Publishing, 2018.

[65] Dae-Seok Byeon, Sung-Soo Lee, Young-Ho Lim, Jin-Sung Park, Wook-Kee Han, Pan-Suk Kwak, Dong-Hwan Kim, Dong-Hyuk Chae, Seung-Hyun Moon, Seung-Jae Lee, Hyun-Chul Cho, Jung-Woo Lee, Moo-Sung Kim, Joon-Sung Yang, Young-Woo Park, Duk-Won Bae, Jung-Dal Choi, Sung-Hoi Hur, and Kang-Deog Suh. "An 8gb multi-level NAND flash memory with 63nm STI CMOS process technology". *ISSCC. 2005 IEEE International Digest of Technical Papers. Solid-State Circuits Conference, 2005.* IEEE, 2005, pp. 46–47.

[66] K.-T. Park, M. Kang, D. Kim, S.-W. Hwang, B. Y. Choi, Y.-T. Lee, C. Kim, and K. Kim. "A Zeroing Cell-to-Cell Interference Page Architecture With Temporary LSB Storing and Parallel MSB Program Scheme for MLC NAND Flash Memories". *IEEE Journal of Solid-State Circuits* 43.4 (2008), pp. 919–928.

[67] J. Power, R. Clancy, W. Wall, A. Mathewson, and W. Lane. “An investigation of MOSFET statistical and temperature effects”. *ICMTS 92 Proceedings of the 1992 International Conference on Microelectronic Test Structures.* IEEE, pp. 202–207.

[68] A. Osman and M. Osman. “Investigation of High Temperature Effects on MOSFET Transconductance (gm)”. *IEEE International High Temperature Electronics Conference (HITEC)* 4 (1998), pp. 301–304.

[69] M. Jefremow, T. Kern, U. Backhausen, C. Peters, C. Parzinger, C. Roll, S. Kassenetter, S. Thierold, and D. Schmitt-Landsiedel. “Bitline-Capacitance-Cancelation Sensing Scheme with 11ns Read Latency and Maximum Read Throughput of 2 . 9GB/s in 65nm Embedded Flash for”. *IEEE International Solid-State Circuits Conference (ISSCC)* (2012), pp. 428–430.

[70] O. Golonzka, U. Arslan, P. Bai, M. Bohr, O. Baykan, Y. Chang, A. Chaudhari, A. Chen, N. Das, C. English, P. Jain, H. Kothari, B. Lin, J. Clarke, C. Connor, T. Ghani, F. Hamzaoglu, P. Hentges, C. Jezewski, I. Karpov, R. Kotlyar, M. Metz, J. Odonnell, D. Ouellette, J. Park, A. Pirkle, P. Quintero, D. Seghete, M. Sekhar, A. S. Gupta, M. Seth, N. Strutt, C. Wiegand, H. J. Yoo, and K. Fischer. “Non-Volatile RRAM Embedded into 22FFL FinFET Technology”. *2019 Symposium on VLSI Technology.* IEEE, 2019, T230–T231.

[71] L. Råde and B. Westergren. *Mathematics Handbook for Science and Engineering.* Berlin, Heidelberg: Springer Berlin Heidelberg, 2004.

[72] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth. “On the LambertW function”. *Advances in Computational Mathematics* 5.1 (1996), pp. 329–359.

Danksagung

Die vorliegende Arbeit entstand während meiner Tätigkeit als wissenschaftlicher Mitarbeiter am Lehrstuhl für Technische Elektronik der Technischen Universität München. Der Lehrstuhl stand damals unter der Leitung von Prof. Doris Schmitt-Landsiedel, der ich hiermit ganz herzlich danken möchte. Sie hat mir diese Arbeit ermöglicht und mir ihr Vertrauen für all meine Entscheidungen während dem Projekt geschenkt.

Ebenso danken möchte ich Prof. Helmut Gräb, der die Betreuung meiner Arbeit übernommen hat, nachdem Prof. Schmitt-Landsiedel krankheitsbedingt leider frühzeitig in den Ruhestand getreten ist. Danke für die vielen Ratschläge zu Papern und Konferenzbesuchen.

Mein besonderer Dank gilt meinem Mentor Dr. Thomas Kern aus der Flash Abteilung von Infineon, in deren Zusammenarbeit das Forschungsprojekt durchgeführt wurde. Er unterstützte mich stets mit vollem Einsatz, stand mir mit technischem als auch menschlichem Rat beiseite und motivierte mich in den schwierigen Momenten des Projekts. Sein Vertrauen in meine Entscheidungen schätze ich sehr und danke ihm für die erfolgreiche Zusammenarbeit.

Ein weiterer Dank geht an Prof. Bernhard Wicht, der mir mit seiner fachlichen Erfahrung hilfreiche Tipps gab und mich sehr motivierte die Arbeit fertig zu stellen.

Weiterhin möchte ich meinem Masterstudenten und späteren Lehrstuhlkollegen Eugen Egel danken. Mit seiner Zielstrebigkeit war und ist er mir ein gutes Vorbild. Danke für die hilfreichen Diskussionen in der Lehrstuhlküche und die schönen Lehrstuhlexkursionen.

Ein herzlicher Dank geht an meine weiteren Lehrstuhlkollegen. Vielen Dank an Cenk Yilmaz, der mich während meinem Studium für die Mikroelektronik begeisterte und mich an den Lehrstuhl holte. Weiterer Dank an Stephan Breitkreutz-von Gamm, Leonhard Heiß, Irina Eichwald, Max Stelzer für die kurzweilige Zeit am Lehrstuhl (und die schönen Lehrstuhlausflüge), Grazvydas Ziemys und Rainer Emling, für den kompetenten IT Support. Danke an Bettina Cutrupia, Marta Giunta und Iris Artinger, für die Hilfe bei Organisatorischem und bei den Abrechnungen.

Ein großer Dank geht an meine Kollegen bei Infineon, Martin Stiftinger, für die

stete Hilfe im Labor, Thomas Röhr, für die Hilfe und die vielen Ratschläge zum Toplevel-Design des Test Chips. Danke an Christoph Roll für die unnachlässige Unterstützung bei den Layouts. Danke an Volker Pissors, für den herzlichen Empfang in Dresden und die Hilfe bei Sequenzdefinitionen.

Weiterhin möchte ich meinem besten Freund Marcus Ossiander danken, der mich erst auf die Idee gebracht hat, die Promotionsstelle anzunehmen. Vielen Dank für die vielen Diskussionen, Tipps und dein immer offenes Ohr.

Ein besonderer Dank gilt meinen Eltern und meiner Familie, die stets am mich glaubt und mich in all meinen Entscheidungen unterstützt.

Abschließend möchte ich den größten Dank meiner Frau Esther aussprechen, mit der ich zusammen die Höhen und Tiefen unserer beider Promotion durchschreiten durfte. Zusammen schaffen wir alles! Danke, dass es dich gibt!