



UNIVERSITÀ DEGLI STUDI DI PAVIA

FACULTY OF ENGINEERING

Department of Electrical, Computer and Biomedical Engineering

Ph.D school in Microelectronics
XXX Cycle

**Analog Circuit Design for Non-Volatile
Memories**

Advisor:

Prof. Guido Torelli

Prof. Alessandro Cabrini

Ing. Marco Pasotti

Ph.D Coordinator:

Prof. Guido Torelli

Ph.D thesis of
Riccardo Zurla

Academic year 2016/2017

Contents

| | | |
|----------|-------------------------------------------------------|-----------|
| 1 | Solid-State Memories | 3 |
| 1.1 | Static RAM | 6 |
| 1.2 | Dynamic RAM | 9 |
| 1.3 | Flash memory | 12 |
| 1.3.1 | Flash-memory scalability | 17 |
| 1.4 | Phase Change Memory | 20 |
| 2 | The Spider-Mem Test Chip | 27 |
| 2.1 | Write Circuitry and Procedure | 31 |
| 2.2 | Read Circuitry and Procedure | 37 |
| 3 | Design of Analog Circuits | 41 |
| 3.1 | V_Y Voltage Regulator Design | 42 |
| 3.1.1 | Voltage Regulator Topology | 44 |
| 3.1.2 | The V_Y Regulator Operational Amplifier | 49 |
| 3.1.2.1 | Compensation | 51 |
| 3.1.2.2 | Component Sizing | 57 |
| 3.1.3 | Simulation Results | 59 |
| 3.2 | Improved Current Mirror for PCM Programming | 62 |
| 3.2.1 | Simulation Results | 67 |
| 3.3 | Charge Pump Design | 69 |
| 3.3.1 | Proposed Charge Pump Architecture | 76 |
| 3.4 | Enhanced Voltage Buffer Compensation | 82 |
| 3.4.1 | Simulation Results | 91 |
| 3.5 | Bandwidth Optimization | 92 |
| 3.5.1 | Large Capacitive Loads | 101 |
| 3.5.2 | Comparison with Circuit Simulation | 102 |

| | | |
|----------|--------------------------------------|------------|
| 4 | Experimental Characterization | 105 |
| 4.1 | V_Y Voltage regulator | 107 |
| 4.2 | Improved current mirror | 112 |

Introduction

Nowadays, the pervasiveness of integrated circuits is so massive that it is common for a person to have in her/his own pockets at least one electronic device, which includes dozens of integrated circuits (e.g., microprocessors, radio-frequency modules, memories, image sensors, inertial sensors) each of them containing up to billions of transistors. This huge number of devices is expected to increase since the market is pushing towards the Internet of Things that would include electronics circuits in objects that historically did not contain them. Another range of applications that are showing a strong growth in the demand for integrated circuits belong to the automotive market, which is driven by the increasing number of electronic devices included in conventional cars and next-generation vehicles: electric and self-driving cars. In particular, the latter require a large variety of integrated sensors and the former need the aid of power electronic devices to control the electric engines. These factors contributed to the significant growth in the request of solid-state memories, which typically are included (embedded) in these more and more complex systems to store information and implement smart devices. In this scenario, STMicroelectronics has chosen Phase Change Memory (PCM) as the non-volatile memory to be implemented in its next-generation automotive and smart-power products. Phase Change Memory is a very attractive emerging non-volatile memory due to its desirable properties such as high density, long endurance, good read/write performance, and high scalability potential.

During my Ph.D. activity, I was involved in a cooperation between the Smart-Power Technology Group of STMicroelectronics (Agrate Brianza, Italy) and the University of Pavia (Pavia, Italy) that led to the design of a 32 KB embedded-PCM macrocell, which was integrated in a test-chip (named Spider-Mem) using a 180 nm BCD (Bipolar, CMOS, and Drift-MOS) technology.

This thesis presents the design of analog blocks that were included in the Spider-Mem test-chip, in particular in the programming circuitry. To be more specific, an improved current mirror featuring enhanced recovery time and a voltage regulator, which supplies the write circuits, are described from the

design phase to the experimental characterization. Beside above circuits, a novel charge pump architecture that guarantees enhanced power efficiency is also described, analyzed, and simulated. In addition, two theoretical analyses, whose conclusions provide guidelines to optimize the design of two-stage CMOS operational amplifiers, as required for the next version of the chip, were carried out: an enhanced frequency compensation scheme that achieves larger bandwidth and reduced silicon area with respect to conventional compensation schemes and a design strategy that optimizes the operational amplifier bandwidth under silicon area and power consumption constraints.

Chapter 1 gives an overview of the solid-state semiconductor memory scenario and describes the characteristics of the most utilized memories, aiming at providing an explanation of the driving forces that determine the suitability of PCM as the embedded non-volatile memory to be implemented in next-generation products. Chapter 2 describes the test-chip architecture giving an overview of the macrocell building blocks, focusing in particular on programming and sensing circuits. Chapter 3 contains the theoretical analysis, the design choices, and the simulation results of both the designed analog circuits and the theoretical studies and, finally, Chapter 4 shows the experimental characterization of the designed analog blocks integrated in the Spider-Mem test chip.

Chapter 1

Solid-State Memories

Since it was established, the memory market includes a large variety of storage elements (and, thus, several types of memories), each one having different characteristics and performance. It is, therefore, useful to classify them in categories, and to do so the most important parameters are the following.

- **Cost per bit:** it is one of the most important characteristic of a memory because it allows to be competitive on the market; it is strictly correlated with single cell area occupation and with the number of additional masks required to implement memory cells with respect to standard CMOS process. It is important to notice that, while the former is independent from wafer size, the impact of the latter on the single die can be reduced by increasing the wafer size. The lower, the cost per bit, the better: indeed, a low relative cost means a lower final price with the same storage capacity, or higher storage capacity at the same price.
- **Storage capacity:** it indicates how many data can be saved inside the memory. It is usually measured in Byte (B) or in its multiple.
- **Latency** (also known as **access time**, or **read time**): it is defined as the amount of time between data request and return.
- **Program throughput:** it corresponds to the amount of data that is possible to store in a memory in a given time period. It is usually measured in Byte per second (B/s) or in its multiple.
- **Power consumption:** it is the power needed to ensure the correct behavior of the memory in both read and write phases; the latter is usually the most demanding one. Power consumption is becoming more and more critical due to the growth of battery-powered systems.

- **Endurance:** it is the characteristic that indicates how many times it is possible to rewrite data without compromising the functionality of a memory cell. It is measured in number of write cycles.
- **Reliability:** it expresses how long data can be stored and then successfully read without errors. The accepted standard for consumer electronics is typically 10 years at the maximum operating temperature allowed by the application.
- **Non-volatility:** it is the ability to retain data even in the absence of power supply. It only depends on the physical mechanism on which the storing capability is based. A memory can only be either volatile or non-volatile, the latter feature being obviously highly desirable.

Two of these parameters stand out from the others: cost-per-bit and latency. The former characteristic is crucial on the market: the history of integrated circuits (ICs) has shown that a low cost is an essential factor (sometimes the only one) to be successful. The latter is important since it states in which applications the memory can be used. The other parameters could be critical in particular applications, however there is not a general demanding rule.

Analyzing the listed parameters, it is straightforward to understand that an ideal memory should be non-volatile and should have minimum cost per bit, read time, and power consumption as well as maximum storage capacity, program throughput, reliability, and endurance. Unfortunately, such utopian device, also called universal memory, has not been produced yet. It is possible, however, to find on the market examples of memories that have some of these desired characteristics while, simultaneously, they are also missing at least one of the remaining features. To overcome this limitation, a typical solution is to use more than one type of memory to benefit from their best characteristics.

As a result of this approach, a memory hierarchy took shape and it is, nowadays, well defined: a conventional way to represent it is through a pyramidal structure (see Fig. 1.1). Each floor of the pyramid depicts a hierarchical level, the upper being occupied by memories with the smallest latency, and the inferior stages comprehending memories with the largest storage capacity available. This structure descends from the fact that the faster memories on the market have also the higher cost per bit, implying that it is not affordable nor convenient to build large arrays of such type of memories. To better understand the hierarchical structure, it is fundamental to comprehend the idea behind it: in place of an ideal universal memory, a system composed by various memories is designed so as to have small memories that are as fast as the lowest-latency memories, and other memory devices that provide large

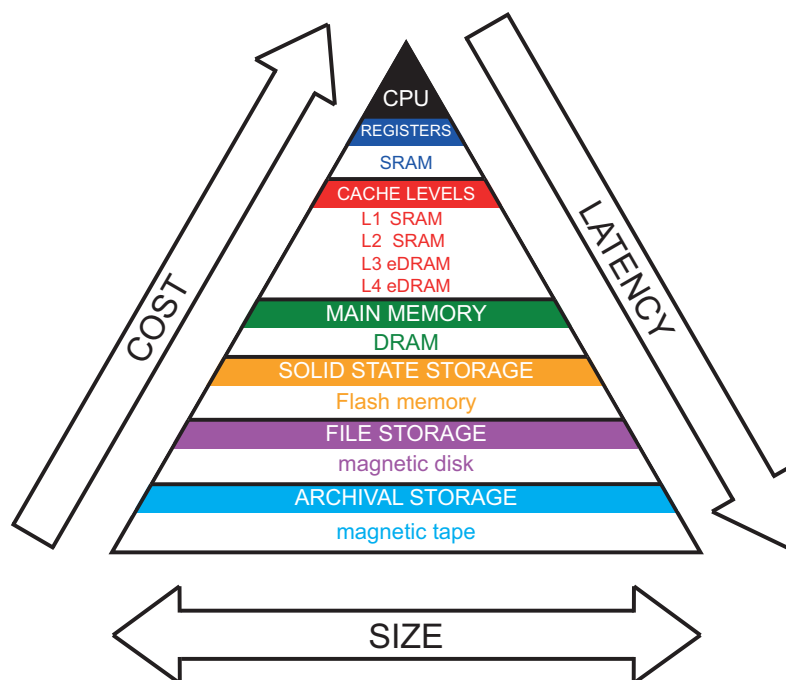


Figure 1.1: Memory hierarchy.

capacity at the same cost-per-bit of the cheapest memories. This principle is particularly effective thanks to the different probability with which data are required by the Central Processing Unit (CPU): the majority of data, indeed, is rarely requested, whereas a minor part is repeatedly sought. Let us now have a better look at the memory pyramid as depicted in Fig. 1.1: the Central Processing Unit is placed on the top of the pyramid, whereas the fastest (and most expensive) memories are placed closer to the CPU, and are used to store a limited amount of the whole data (i.e., the more frequently requested portion). The remaining data are stored in the following levels, starting from the top to the bottom, based on their probability of being accessed. In this way, the greater part of operations can be run by the CPU interacting only with the upper (and faster) layers of the memory hierarchy. Simultaneously, a large data capacity is available in the lowest memory levels. Moreover, storage operations are carried out in parallel, since these memories have very poor program throughput, in order to allow the CPU to run other operations, thus avoiding huge idle time and, hence, inefficiency.

Memories are also divided in two main groups: memory-type and storage-

type. The first class includes fast, small-capacity, and volatile memories, such as Static Random Access Memory (SRAM) and Dynamic Random Access Memory (DRAM), that correspond to the upper layers of the pyramid, whereas the second class includes slow, large, and non-volatile memories (NVMs). During the past years, the gap between these two groups has significantly grown due, on the one hand, to the strong increase of memory-type speed performance and, on the other hand, to the enormous storage capability reached by storage-type memories driven by the strong reduction of their cost-per-bit. This cost decrement is mainly caused, as will be explained in the following sections, by the increase in density that Flash memories were able to provide in the past years.

As shown in Fig. 1.2, the gap between these two groups is so pronounced that there is room in the memory market for a third category, defined storage-class memory, that has to have particular characteristics in order to fit in the unoccupied space: a smaller latency with respect to Flash memories combined with a lower cost-per-bit when compared with DRAMs and SRAMs. In fact, in today's market, these three types of memory are dominant and only with the mentioned specifications it would be possible to obtain a significant market share.

In the following part of the chapter these types of memory are analyzed and compared. Finally, the last section of the chapter is dedicated to the emerging Phase Change Memory (PCM), which is an excellent candidate to stand up in the storage-type class memory and possibly to conquer a significant market share.

1.1 Static RAM

Static RAM is the fastest memory on the market (i.e., the memory with the smallest read and write time) and, therefore, its main application is to be used to implement CPU instruction registers, which are required to be as fast as possible since they are continuously accessed by the CPU when executing even the simplest operations. The SRAM cell, depicted in a symbolic representation in Fig. 1.3, is substantially made up of one bistable latch and two access transistors that act as a selector. In order to keep the density small, the latch is implemented with the smallest possible transistor scheme, which is basically composed by two inverters connected in crosscoupled configuration (Fig. 1.4). Therefore, the total number of transistor needed to implement a typical SRAM cell is six, which is usually indicated as a 6T cell. Due to this reason, SRAMs turn out to have the highest cost per bit on the market.

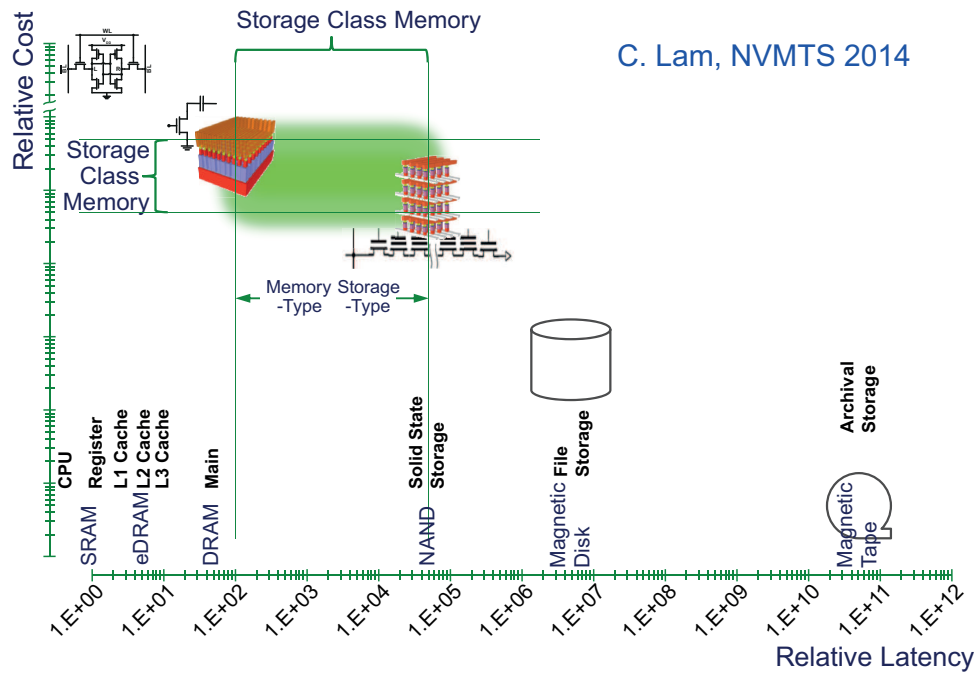


Figure 1.2: Representation of each kind of memory based on its normalized latency (x axis) and cost (y axis).

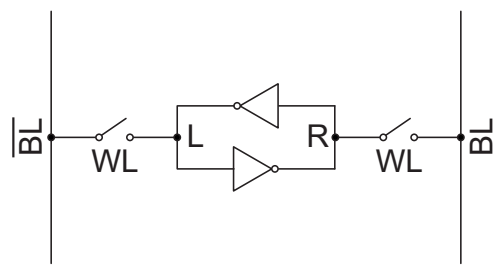


Figure 1.3: Symbolic representation of a SRAM cell.

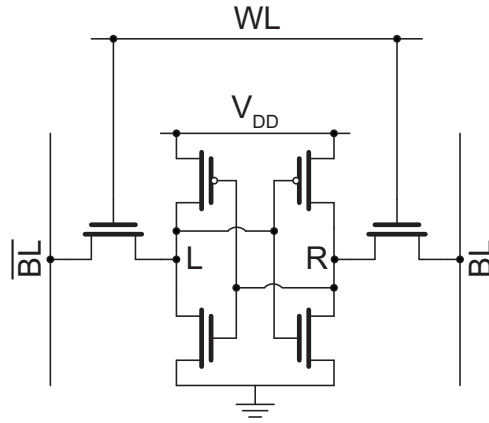


Figure 1.4: Circuit representation of a 6T SRAM cell.

The latch acts as the core memory element: in fact, it is possible to write the data by, firstly, setting the desired voltage on the selected Bit Line (BL) and its complementary value on \overline{BL} (i.e., the complementary Bit Line); then, the two selectors are turned on by raising the selected Word Line (WL); finally, the two inverters actively flip (if needed) the voltages on nodes L and R. After a short transient, the WL is driven low and the data is stored. Since the storage ability is due to the positive feedback produced by the two inverters, it is straightforward to see that, if the power supply is removed, the data is immediately lost, hence the volatility of this type of memory.

The reading operation in SRAM consists in raising the selected WL to allow the voltages on nodes R and L to be transferred to the selected BL and \overline{BL} , respectively: a sense amplifier can then detect the data and deliver it to the output. Static RAMs are excellent from a power consumption point of view, since not only they require little current to be programmed, but also they have almost-zero static power consumption, since, as in standard logic gates, leakage current is the main contributor in static conditions.

Analyzing the write and the read procedure, it is clear that they both are, intrinsically, very fast operations, since they simply require to wait a transient time in which a positive-gain bistable circuit has to flip (i.e., program phase) or to charge the BL parasitic capacitance (i.e., read phase). From the performance point of view, SRAMs are comparable with a standard logic gate implemented in the same technology node, thus their usage allows avoiding bottlenecks while running CPU instructions. Moreover, having active components that handle data transfer is the key characteristic that makes SRAM to stand out,

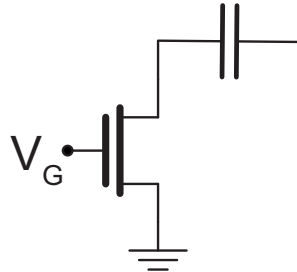


Figure 1.5: Circuit representation of a 1C 1T DRAM cell.

from access-time point of view, when compared to memories that rely only on passive data transfer (i.e., DRAM).

1.2 Dynamic RAM

The idea behind Dynamic RAMs is very simple, and is based on charge storage: the presence or the absence of charges correspond to a logical ‘1’ or ‘0’, respectively. As depicted in Fig. 1.5, the element used to store the charge is a standard capacitor, whereas the selector is an NMOS transistor and, therefore, the cell results a 1 Transistor 1 Capacitor (1T 1C) type.

In Fig. 1.6 a typical DRAM array is depicted. The top plates of the capacitors of several DRAM cells are connected to the same Bit Line, while the NMOS selectors of these cells are connected to different Word Lines. This architecture allows accessing in parallel all the cells in the same row (i.e., all the cells sharing the same WL) to either read or write the data on the corresponding BLs.

To store a logical ‘1’ in a DRAM cell, it is necessary to force the selected Bit Line to V_{dd} , then the selected cell is connected to the selected BL by raising the corresponding WL. In this way, the NMOS transistor acts as a resistor (since it works in its triode region) and transfers the charge from the selected Bit Line to the selected capacitor. Vice versa, in the dual case, the selected BL is driven to ground (GND) and, once the WL is activated the charge is sunk from the capacitor, thus, discharging it.

The read operation is performed by pre-charging the selected BL to a voltage midway between GND and V_{DD} (usually $V_{DD}/2$) and then leaving it floating. Immediately after, the chosen cell is selected by raising the selected WL. The parasitic capacitance of the selected BL and the selected storage

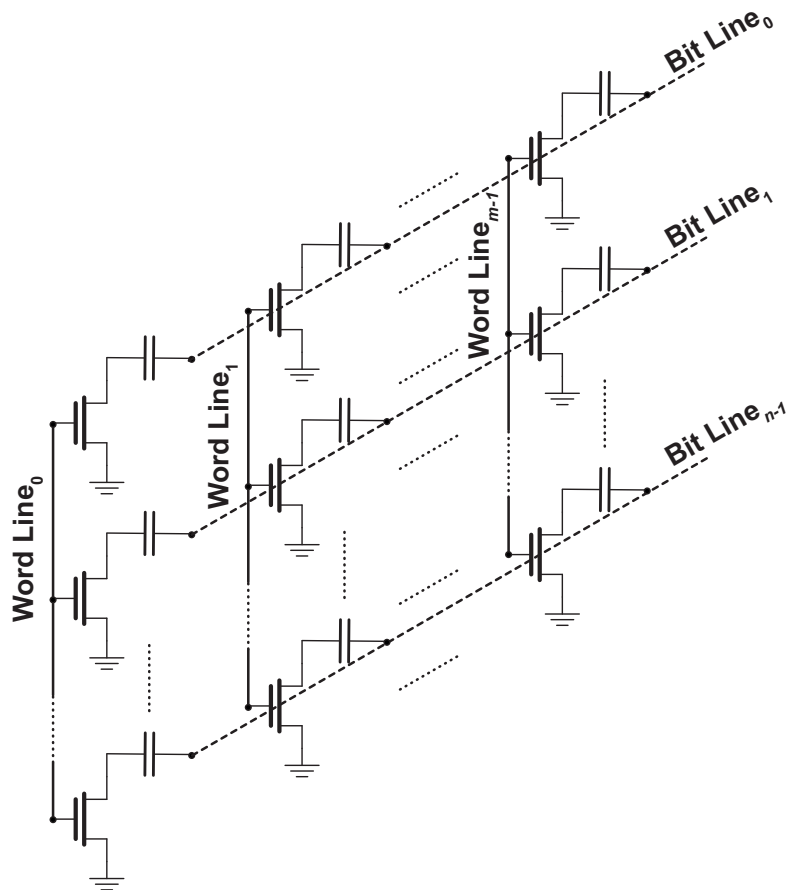


Figure 1.6: Typical $n \times m$ DRAM array.

capacitor are now connected in parallel and, therefore, after a short transient, the voltage across the Bit Line rises, if the cell content is a logical '1', or falls if the cell content is a logical '0'. A sense amplifier is then used to compare the selected BL with an unselected Bit Line charged to $V_{DD}/2$, measure the voltage difference, and finally provide the correct data to the output.

The bits stored in an ideal DRAM cell are available for an indefinitely long time, provided that the power supply is kept active: this condition is mandatory due to the volatile nature of the memory cell. However, in a real integrated circuit, a leakage current is always present inside a DRAM cell and discharges the storage capacitor, thus causing data degradation. To overcome

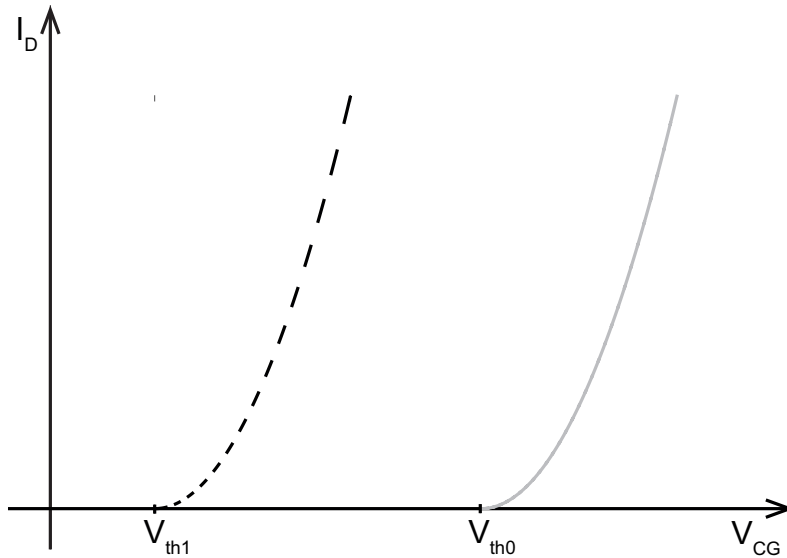


Figure 1.8: Erased-cell current (black dashed curve) and programmed-cell current (grey solid curve) in function of control-gate voltage.

1.3 Flash memory

Starting from its invention in early 80's [1], Flash memory has flourished into a dominant position in the non-volatile memory market. The first company to produce a commercial chip was Intel Corporation with a 256 kb memory. Flash-memory cell is the natural evolution of EPROM and EEPROM cell, since both are based on the floating-gate technology.

The floating-gate MOS structure, shown in Fig. 1.7, consists in an MOS transistor provided with a first polycrystalline silicon (usually called polysilicon or poly) layer (Floating Gate) which is deposited on top of a thin silicon-oxide layer (tunnel oxide), as in standard MOS transistor, and a second polysilicon layer (Control Gate) which is placed on top on the first, with an interposed silicon-oxide layer (interpoly oxide).

Therefore, there is a galvanic isolation of the Floating Gate (FG hence its name). However, it is still possible to drive it through the Control Gate (CG) not with a direct connection but thanks to capacitance coupling. This particular structure allows, in the proper conditions, storing electrons inside the FG. These charged particles shield the electric field imposed by the CG and thus alter the effective threshold voltage of the transistor, which results

higher when there are charges trapped into the FG.

The I-V characteristic of a typical cell is shown in Fig. 1.8: the black dashed curve depicts the current flowing into a flash cell when it stores a logical ‘1’, whereas the grey solid line corresponds to a logical ‘0’. It is worth to point out that, in this example, the alteration of the threshold voltage is equal to $V_{th1} - V_{th0}$. In Flash memory terminology, setting a cell to a logical value of ‘0’ is called program because correspond to trap electrons in the Floating Gate, whereas the reciprocal operation (i.e., write a logical ‘1’) is defined erase since it removes the trapped charges from the FG.

In Flash memory, the program operation is achieved in two possible manners. The first technique consists in accelerate channel electrons, through the generation of a sufficiently high electric field (>100 kV/cm [2]), in order to reach an unbalanced condition between the energy lost by the negative charges, due to their interaction with the lattice, and the energy gained from the electric field. Under this circumstance, some electrons reach particularly high energy levels (i.e., become ‘hot’ electrons) and, therefore, have enough energy to surpass the oxide barrier; this technique is, thus, known as channel hot electrons injection [3]. The second approach exploits a quantum-mechanical effect, called Fowler-Nordheim tunneling, that grants to electrons a non-zero probability to pass through the oxide barrier even if their energy is not sufficiently high to jump it. In order to let hot electron injection to take place, it is needed to connect the Flash-cell drain terminal to a high voltage, the source to ground and the control gate to a sufficiently high potential (typically higher than the nominal V_{DD}). In the case of Fowler-Nordheim tunneling, high voltages are also required, however source terminal is not forced to a specific potential but it is left unconnected, while the Control Gate is risen to a high voltage and the drain terminal is connected to ground.

The erase operation also use quantum-tunneling effects to remove the stored charges from the Floating Gate, however, in this case, the gate-to-drain biasing is reversed (i.e., CG is connected to ground and the drain terminal is connected to a high potential). Both techniques require high voltages. As a consequence, in the last years it has become very common to integrated a charge pump or, in general, a DC-DC converter in flash memory ICs, since they are typically powered with a single low-voltage supply especially in the case of battery-powered systems.

Flash memories can be divided in two main categories depending on the architecture used to implement them: NOR Flash and NAND Flash memories. The names of these architecture descend from their circuit similarity with logical gates (NOR and NAND, respectively). Indeed, in NOR Flash memory,

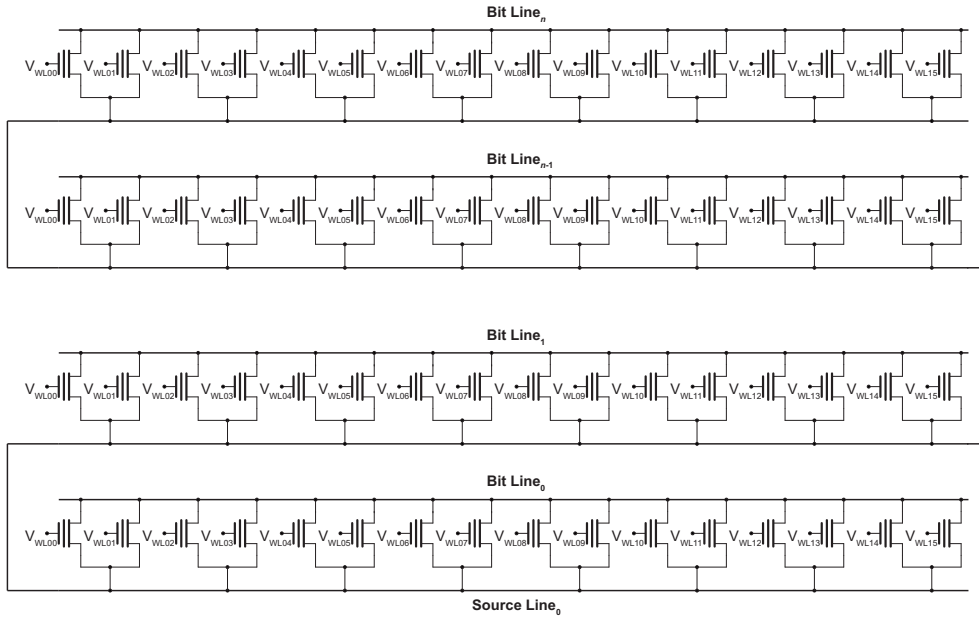


Figure 1.9: Example of a NOR Flash memory block (i.e., cells that share the same source line) containing n Bit Lines each with 16 cells.

cells belonging to a given Bit Line are connected in parallel between a source line (common to all the cells in a memory sector) and Bit Line (see Fig. 1.9), whereas NAND Flash memories have several cells (e.g., 16 or 32) connected in series forming a string; at the two ends of this group, there are two NMOS transistors that act as selectors, since they are able to connect the string between the associated BL and the source line, thus, addressing it. As shown in Fig. 1.10, several of these groups are then connected in parallel, in a NOR-like configuration, to the same Bit Line. NOR and NAND Flash memories have several differences regarding their characteristics and performance, and these dissimilarities are mainly due to the electrical connections.

On the one hand, the NOR architecture allows a random access of data, since all cells are independently addressable. The cells belonging to the same WL can, indeed, be read at the same time by forcing the Word Line to a voltage between V_{th0} and V_{th1} . The corresponding Bit Lines (i.e., the BLs connected to the drain terminal of the selected cells) are then driven to a low voltage only if a logical ‘1’ is stored in the associated cell, otherwise the voltage of the Bit Lines stays stable at the pre-charged value. In the NOR Flash memory, the

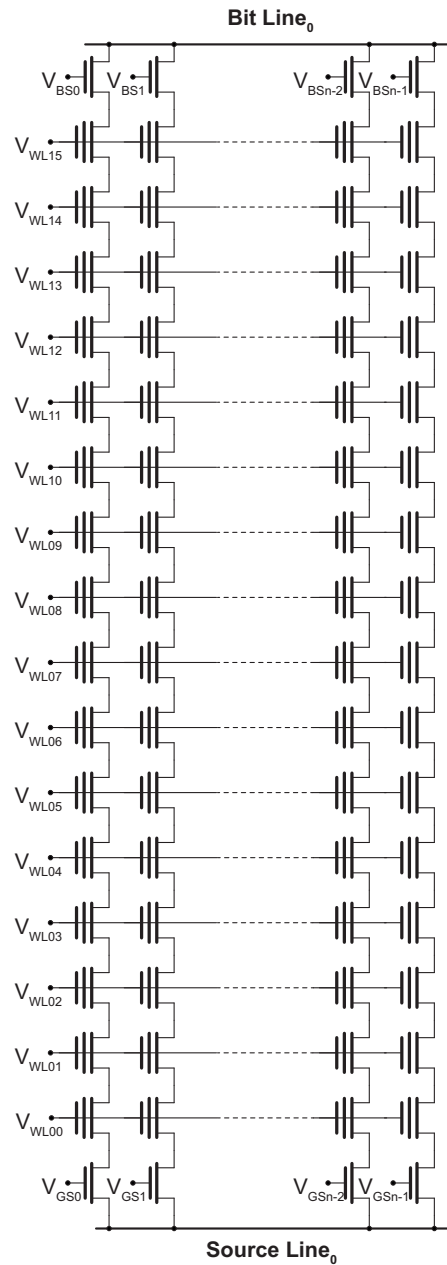


Figure 1.10: Example of a NAND Flash memory block containing 16 pages with n cells each.

program operation can be carried out with a single cell granularity, whereas the erase operation affects indiscriminately all the cells belonging to the same block. Moreover, IC manufacturers are usually able to guarantee the intrinsic data correctness and retention (without the aid of Error Correction Codes, ECCs, or other techniques).

On the other hand, the NAND architecture, when performing a read operation, requires to enable a block (i.e. to connect several string of cells, which belong to different BLs, by turning on the NMOS selector transistor of each string). Then, the selected Word Line is raised, while the unselected WLs of the same block are driven to a voltage that is much higher than V_{th1} , thus forcing all the unselected cells in the same string in the ON state (ie., the conductive state, regardless of the stored data). As in the case of NOR architecture, the voltage value chosen for the selected Word Line is in between V_{th0} and V_{th1} . By monitoring the corresponding Bit Lines, it is possible to read the cells and then provide the data to the output. This operation takes a significant amount of time that directly impacts on the final access time. As a consequence, to reduce the weight of this overhead, the read operation is typically performed on the whole block of strings, and the read data is then sequentially transferred to the output. NAND Flash memory program operations can be performed at the page level, whereas the erase procedure addresses an entire block. Furthermore, it is very common to include an ECC algorithm to improve die yield and data retention, because, due to the complexity of read and write operation and the need to shrink the array cells as much as possible, IC manufacturers usually do not guarantee the perfect behavior of every single cell. These characteristics make NAND Flash Memory highly suitable for large storage applications (i.e., for the applications shown at the very bottom of memory hierarchy in Fig. 1.1), since it provides high data throughput (when large data are requested) and poor latency, which can be mitigated by sequential accesses, which is indeed very common in the mentioned utilizations.

The final, and most relevant, layout difference of the two architectures is related to silicon implementation. NOR Flash memory requires one contact every two cells to connect their drain to the corresponding Bit Line, whereas the source line has one contact shared between, typically, 16 or 32 BLs. Therefore, in first order approximation, one contacts every two cells is needed to ensure the correct electrical behavior. NAND Flash memory, on the contrary, exploits its series organization to minimize the number of contacts needed, because the electric inter-cell connection takes place directly in the silicon surface by sharing source/drain diffusion, as depicted in Fig. 1.11. In this case, one contact is needed every two strings to connect them to the correspondent BL. Thus, this

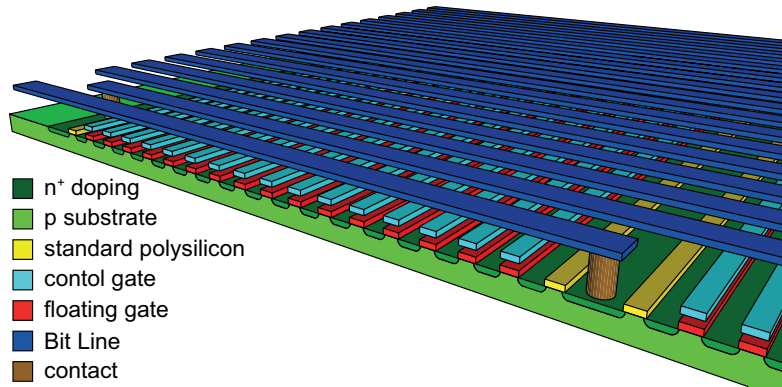


Figure 1.11: Example of a planar NAND Flash memory silicon implementation with 16 cells per string. For the sake of clarity, both interpoly- and tunnel-oxide layers are not shown.

architecture requires a contact every $2m$ cells, where m is the number of cells contained in a string (typical values are $m = 16, 32, 64$). Due to design rules imposed by the planar technology, to form a contact between the substrate and the first level of metal it is mandatory not only to leave additional space for the contact itself, but also additional extra area to counteract intrinsic inaccuracies (e.g., masks misalignment, process random variations, etc.). Thus, the strong reduction of contacts provided by the NAND architecture allows a significant decrease of area occupation with respect to the NOR architecture, which directly translates in higher density and, hence, in lower cost-per-bit.

Even though, at its origin, NAND Flash memory was considered less promising, thanks to its high density, it was able to reach an impressive diffusion in a lot of different application, to gain a larger and larger share in the non-volatile memory, and finally to conquer an uncontested dominant position.

1.3.1 Flash-memory scalability

According to the Flash-memory cell working principle, the logic state of the cell is determined by the presence or absence of electrons in the Floating Gate. As a consequence, data retention is strictly related with the probability that these charges leave the FG. In particular, the main parameter that controls this effect is the thickness of the tunnel-oxide layer (t_{ox}), so that the thicker the layer, the lesser the probability that electrons escape and, thus, the longer the data retention.

The correlation between gate oxide thickness and data reliability started to become critical when the technology scaling down pushed process nodes below 65 nm ($t_{ox} \approx 1.2$ nm), where data retention was barely acceptable. The solution implemented by technologists to mitigate this problem in more advanced technology nodes was to replace silicon oxide with high- κ material (usually hafnium-based compounds), where κ is their dielectric constant. A higher dielectric constant allows using a thicker layer having the same (or sometimes even larger) coupling capacitance between the Control and the Floating Gate. Later on, the whole polysilicon layer that used to implement the floating gate was replaced with a charge-trapping layer that, being dielectric, provides a higher yield, endurance and retention thanks to its better behavior against oxide defects that can give rise to short circuits between the floating gate and the transistor channel. Applying these techniques, the data retention was therefore ensured for more advanced technologies, however, it was also foreseen that no Flash memory could be implemented in CMOS technology nodes below 20 nm.

To further increase the memory density, designers were able to develop multi-level cells: firstly, four-level cells, which store two bits per cell, were developed and, then, cells that can accommodate up to 8 discernible states and, therefore, allow storing 3 bits per cell. The drawback of these techniques are an increasing programming time, latency, and power consumption as well as a lower endurance. The reason is mainly due to the higher number of write pulses required to program the cell and to the slower reading procedure (more levels have to be detected). Nevertheless, the cost-per-bit reduction led these solutions to a great success on the memory market, which is not surprising since history has proven that in storage-type memories this is, by far, the most important characteristic.

In 2012, when the 20 nm technological barrier was approaching, Samsung developed a new architecture based on a groundbreaking idea that was around from some years, but nobody was yet able to successfully implement: develop in the third dimension the historically planar IC technology. In fact, from that moment on, the flash cell went through a substantial revolution, moving transistor channel along the vertical direction, as shown in Fig. 1.12. The charge-trapping layer, as well as the tunnel-oxide layer, became hollow cylinders wrapped around the vertical cylindrical-shaped channel. Finally, the old polysilicon stripes that used to form control gates and WLS were turned into horizontal planes and stacked on top of each other. This new approach was so convenient that it is nowadays a standard and all the memory manufacturers produce devices with 3D approach. To better understand the new structure,

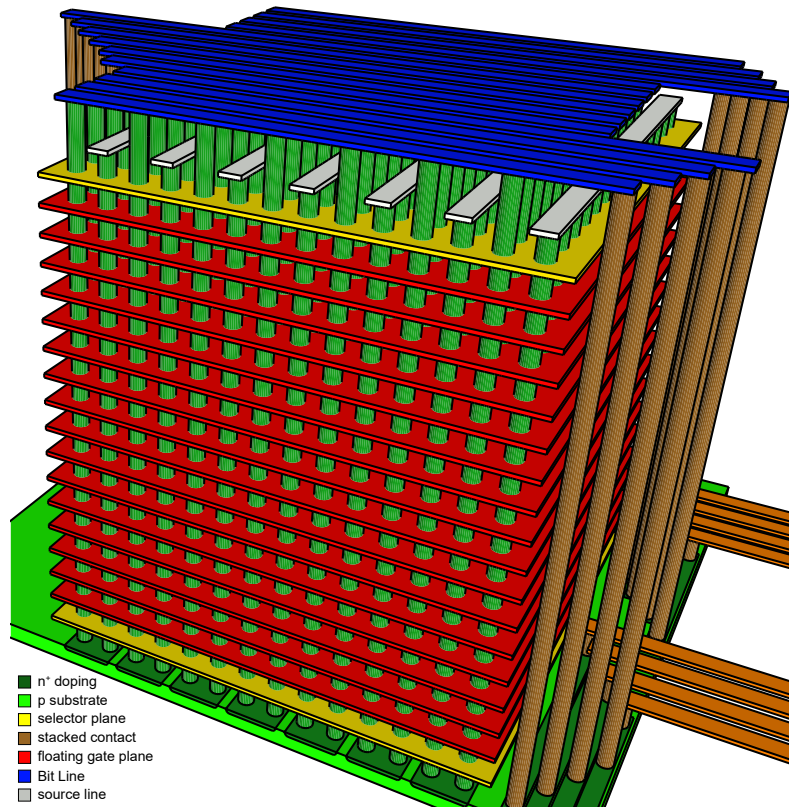


Figure 1.12: 3D NAND Flash architecture with 32 cells per string.

the reader can imagine a standard planar NAND Flash memory cell implementation, such as the one represented in Fig. 1.11, fold it in the middle point of the string, and rotate it 90° to obtain the structure depicted in Fig. 1.12.

From an intuitive point of view, the transition from 2D NAND to 3D NAND memory is comparable with the introduction of skyscrapers in the building market: the possibility to stack floors on top of each other enabled the opportunity to increase the density of a given area almost indefinitely.

The memory manufacturers able to bear the huge development cost of these complex structures are nowadays only two companies (i.e., Samsung and SK Hynix) and two joint ventures (i.e., Intel-Micron and Toshiba-SanDisk), therefore the open information available is very modest.

Thanks to its very high density and very low cost-per-bit (especially for

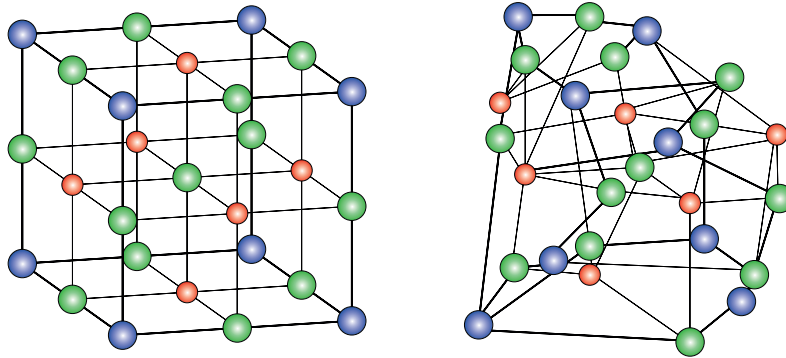


Figure 1.13: Example of two atomic structures of a Germanium (red spheres) Antimony (blue spheres) Tellurium (green sphere) alloy in two different phases: crystalline (on the left side) and amorphous (on the right side).

3D NAND architecture) Flash memory was able to take over magnetic-based storage memories and it is expected to be the dominant technology in the storage-type memory market for next years, since the main technological limitations seem to be overcome with the presented 3D architectures.

1.4 Phase Change Memory

The Phase Change Memory (PCM) cell is typically made of a phase change layer, a heater, and a selector.

The phase change material is a chalcogenide alloy that (Fig. 1.13) has two distinct, reversible, and solid states: crystalline and amorphous. The former is characterized by an ordered structure: the atoms are, indeed, organized in a crystalline lattice, which corresponds to the minimum-energy molecular structure. The latter has, instead, a random atomic arrangement without any notable symmetry. This two phases have very different properties, in particular the crystalline state shows a low resistivity as well as a high refractive index and, vice versa, the amorphous phase exhibits a much higher resistivity and a lower refractive index.

The phase change material is usually a Germanium (Ge), Antimony (Sb), and Tellurium (Te) alloy. This alloy is very well known in electronics industry, especially with stoichiometry $\text{Ge}_2\text{Sb}_2\text{Te}_5$, often referred to as GST, since it was largely used in Compact Disc-Recordable (CD-RW) and Digital Video Disc-Recordable (DVD-RW), due to its aforementioned optical properties. The phase

change layer is, hence, the core part of a PCM cell, because it is where the logical information is stored: the crystalline state conventionally corresponds to a logical ‘1’, whereas the amorphous phase is associated with a logical ‘0’.

The heater is needed to carry out write operation. It is a resistive element that increases its temperature when a current is forced to flow into it, due to Joule effect. By controlling the current it is, thus, possible to control the heater temperature and, since it is placed right below the phase change layer, the GST temperature can be set accordingly. In order to write a logical ‘0’ (i.e., to drive GST into the amorphous phase) in a PCM cell, it is necessary to increase the GST temperature above its melting value (T_{melt}) by setting a proper current, higher than I_{melt} (where I_{melt} is the current that imposes the heater temperature to be equal to T_{melt}). Under these conditions, the chalcogenide alloy reaches a liquid form, in which its atoms have sufficiently energy to freely move inside the volume. Then, the current is abruptly turned off and, thus, the temperature in both the heater and the GST layer is sharply decreased. The atoms are now frozen in the random position assumed during the melt state, which causes an amorphous arrangement of the physical structure. This operation is called RESET. The dual procedure (i.e., writing a logical ‘1’) is defined as SET operation and can be carried out in two different manners. The first is based on the fact that GST naturally tends to reach the crystalline state, because, as any physical system, it spontaneously converges towards the minimum energy state. Therefore, if the temperature is risen sufficiently high, the natural crystallization tendency is exponentially accelerated according to Arrhenius law. In this way, a process that would take decades at room temperature can be concluded in few μs or even less. However, it is fundamental to keep the GST temperature below T_{melt} , during this type of SET operation, to ensure that no portion of the phase change layer reaches the melting point, thus going back to an undesired amorphous part once cooled. To accomplish this operation, a rectangular-shaped pulse is needed in order to achieve the right temperature profile in the GST alloy. The alternative SET procedure is performed by increasing the GST temperature to a sufficiently high value (sometimes even above T_{melt}), and then smoothly reducing it down to the room value. If the temperature gradient is small enough, the atoms of the chalcogenide alloy have an adequate time to settle and reach the minimum-energy state and, thus, to assume a crystalline arrangement. Hence, the pulse needed to obtain the desired temperature profile has a triangular shape. Figure 1.14 shows three illustrative programming pulses that correspond to the three described techniques. In particular, the RESET pulse is depicted in blue, the first SET pulse (i.e., the rectangular shaped one) described is illustrated

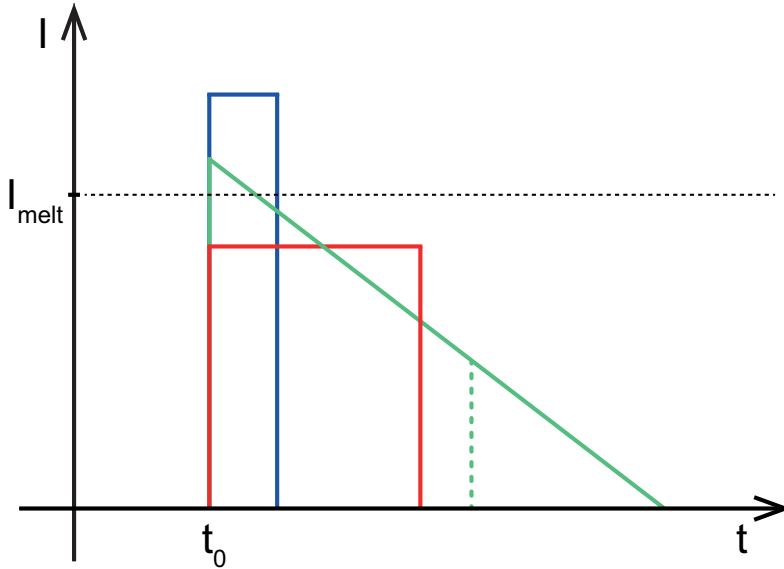


Figure 1.14: RESET pulse (blue curve), rectangular SET pulse (red curve), triangular/trapezoidal SET pulse (green solid/dashed curve).

in red and, finally, the green pulse represents the alternative SET procedure (i.e., the triangular shaped waveform). It is worth to point out that the triangular pulse can be truncated below a certain current value (green dashed line), which corresponds to a temperature too low to deliver significant energy to the atomic structure, thereby resulting in a trapezoidal-shaped pulse. Since the GST phase (at least in first order approximation) is stable during the guaranteed lifetime (usually 10 years), PCM is a non-volatile memory.

The selector is the only active element in a PCM cell, and determines whether the current can flow into the heater and the GST layer. When implementing such device, the main constraint is to allow the current to flow without introducing a significant voltage drop across terminals of the selector, while still keeping its dimensions limited in order to fit the memory array pitch. This condition it is not trivial to achieve since RESET programming current can be easily in the order of several hundreds μA . Initially, MOS transistors were not able to satisfy these specifications and, hence, mainly bipolar transistors (or even diodes) were used to implement the PCM selector. However, in recent years, several IC manufacturers implemented PCM selectors with MOS transistors, which have lower leakage current and also enable full

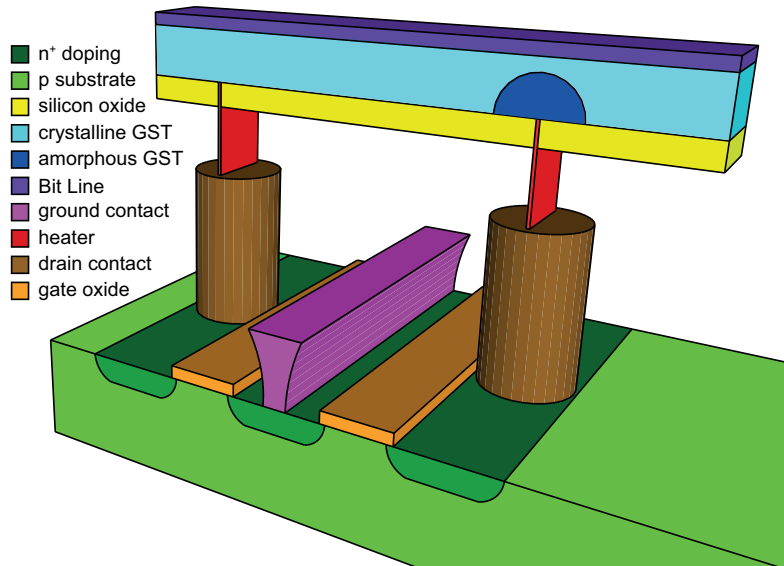


Figure 1.15: 3D view of two PCM cells which share the source diffusion of the selector and the ground contact. The cell on the left side is in the SET state, whereas the cell on the right side is in the RESET state

compatibility with standard (and cheapest) CMOS technologies.

Given the PCM-cell structure, it is straightforward to observe that it can be seen mainly as the series connection of a variable resistor, a fixed resistor and a switch, as illustrated in Fig. 1.15. The GST can be, indeed, modeled as a resistor that can assume a very wide variety of resistance values (typically from tens of $k\Omega$ to few $M\Omega$), depending on its phase and on reading temperature. From an electrical point of view, the heater is a simple ohmic resistor, whereas the selector can be seen as an ideal switch that can either be closed to create a conductive path between the heater and ground or be open to isolate the corresponding memory cell.

As depicted in Fig. 1.16, in a typical PCM array, the top electrodes of several cells are connected to the same Bit Line, thus forming a column of the array matrix. The Word Lines run in the orthogonal direction and are connected to the gate terminals of the selectors belonging to cells placed on the same row. To read a PCM cell, it is therefore necessary to raise the selected WL and impose an adequate voltage on the corresponding Bit Line, which causes a (small) read current to flow through the cell. Simultaneously, a fixed (reference) current is generated in a separated path and is then fed to the first

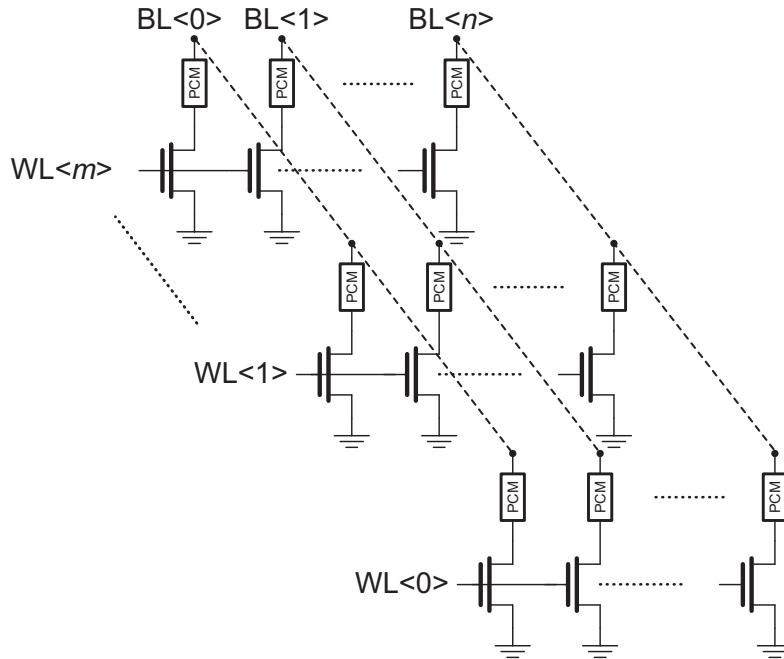


Figure 1.16: Example of a PCM MOS-selected array with m Word Lines (rows) and n Bit Lines (columns). For the sake of simplicity, the heater is embedded in the PCM block.

branch of a current comparator. The read current is applied to the second branch and, therefore, the comparator is unbalanced by the current difference. In this manner it is possible to detect whether the reading current is higher or lower than the reference current and, hence, sense the conductance of the cell (i.e. high conductance, SET state, or low conductance, RESET state). The reference current is chosen to have the optimal value: it has to be in between the minimum expected SET current and the maximum expected RESET current. The read operation can be executed, simultaneously, on different cells placed on the same row (i.e. on cells sharing the same Word Line). Usually the number of cells involved in the read procedure is equal to the word length (typically 16, 32 or 64). It is worth to point out that PCM can reach very low read latencies (even few ns), which is a performance that places it close to DRAM, with the additional advantage of being non-volatile.

The program operation is performed by applying either voltage or current pulses with the shape discussed above. The former technique is defined voltage

mode programming, whereas the latter current mode programming. There are not significant differences between the two approaches, apart from the electronic circuitry that needs to be implemented to manage write operations. Moreover, PCM can be written with a single cell granularity without the necessity of erasing an entire sector and then reprogramming it, as in the case of Flash memory. Obviously, this is a substantial advantage from the program throughput point of view, especially when only few bits have to be modified in a sector.

In the early phases of its development, PCM was considered the best candidate to overtake Flash memory, once those would have faced their foreseen scalability problems. In fact, PCM does not suffer from problems related to the shrink dictated by newer technology nodes. Even more, PCM can take advantage from technology scale down: the programming current is reduced, the selector is a standard MOS transistor (and therefore has the same performance of the other transistor implemented in the same technology), and the GST volume can be easily scaled. However, as explained in the previous section, the huge amount of investment that memory manufacturers spent in NAND Flash memory research and development allowed solving the majority of scalability issues and, thus, avoided the need for replacing them with PCM at least in the case of storage applications. Nowadays, PCM still has the possibility to fit in the aforementioned storage-class scenario. Indeed, as shown in Fig. 1.15, the phase change layer can be implemented in the Back End Of Line (BEOL) stage of IC fabrication process, i.e. it is deposited when the most critical layers (i.e., the ones in direct contact with the substrate) have already been fabricated. This characteristic is fundamental because the cost of BEOL masks is lower, since they do not have to cope with the smallest lithographic length. Moreover, PCM does not require complex structures, as 3D-NAND Flash memory does: it can therefore be used as an embedded memory (i.e., a memory incorporated in the same die with the rest of the system) for instance, embedded PCM has been implemented by STMicroelectronics only by using 3 additional masks on top of the standard CMOS fabrication process flow. Finally, PCM fits also very well as non-volatile memory in space electronics applications, since it shows much higher resistance to radiation (i.e., it is much more radiation hard) with respect to NAND Flash memories, which are based on electrons trapping and, therefore, can be easily corrupted in this harsh environment.

Chapter 2

The Spider-Mem Test Chip

Spider-Mem is a test chip designed and fabricated by STMicroelectronics with the purpose of experimentally investigating the performance of phase-change-memory arrays (featuring NMOS selectors) as embedded non-volatile memory.

Nowadays, automotive and smart-power systems are becoming more and more complex since they have to offer several functionalities, in order to be competitive on the market, while still providing extremely high reliability and flexibility. Therefore, it is more and more common to integrate a micro-controller to fully manage all the internal components and be able to adapt the system to most disparate conditions and environments. Moreover, the choice to integrate the micro-controller together with the rest of the system, instead of having two separated ICs, allows producing smaller and, thus, cheaper devices with higher power efficiency and smaller parasitic effects due to shorter internal connections.

Along with the integration of a micro-controller, it comes, simultaneously, the necessity of a non-volatile memory able to store its firmware. The required NVM has to comply with the strict specifications dictated by automotive standards and minimize the impact on the other integrated components, which actually are the main contributors to the added value of the product. Therefore, memories that are implemented in the BEOL are preferred with respect to memories that have components located in the front end of the line (FEOL), e.g. EPROMs, EEPROMs, and Flash memories. The simplest solution would be to use a masked Read Only Memory (ROM), however, this choice inhibits the possibility to change the micro-controller firmware, which is a highly desirable feature for both IC manufacturers and users, in order to facilitate the firmware development and provide customers with the opportunity to write and, when required, modify their own personalized code.

For the above reasons, PCM has been chosen by STMicroelectronics to be the embedded memory implemented in its next-generation automotive and smart-power next-generation devices. Even though PCM has the attractive feature of being implemented in the BEOL, this choice may not seem cost-effective when compared to the use of high-density and low-cost-per-bit memory such as Flash memory. However, it is important to notice that Flash memory has a very low cost per bit only when implementing a large-capacity stand-alone memory, otherwise its cost can increase substantially, especially in the case of small-size (≤ 128 KB) embedded array. The increased cost per bit is mainly due to two reasons: firstly, some of the area-reduction techniques (e.g., 3D architectures) can not be implemented in embedded solutions; secondly, the available area dedicated to the complex overhead circuits has a larger impact on total silicon area occupation when the array area is smaller. Since 128 KB is a storage array size that is more than sufficient for containing the firmware of an embedded micro-controller, PCM represents the most cost-effective non-volatile memory for this application.

The Spider-Mem chip has been developed as a vehicle to show the feasibility of PCM integration as well as to study and experimentally characterize its performance in the chosen STMicroelectronics BCD9s technology. BCD9s, like all BCD technologies, offers the possibility to integrate, together with standard CMOS transistors, BJTs (Bipolar Junction Transistors) and DMOS (Double Diffused Metal Oxide Semiconductor) transistors, which are widely used in power electronics thanks to their ability to manage high voltages and drive large currents. Therefore, this technology has been chosen by STMicroelectronics as a standard when developing ICs for both automotive and smart-power applications. BCD technologies are also designed to comply with automotive specifications: for instance, ICs fabricated in this technology are able to operate in a temperature range from -40 °C to $+150$ °C, inside harsh environment (high moisture and salinity), and in the presence of mechanical stresses (vibrations, pressures, and shocks). BCD9s technology has two power supplies available: a low-voltage $V_{dd} = 1.8$ V, and a high-voltage $V_{pp} = 5$ V. Both power supplies are granted with an accuracy of about $\pm 10\%$ (i.e., $V_{dd} = 1.6 \div 2.0$ V and $V_{pp} = 4.5 \div 5.5$ V).

To do so, a macrocell [4] (sometimes referred to as IP, Intellectual Property) has been developed. This macrocell integrates a PCM array and all the circuitry required to write and read data as well as to test all the memory and system parameters. The macrocell was then included in the chip in 16 independent instances in order to increase the total number of cells in a single Spider-Mem chip, to reduce testing cost and improve the consistency of the ex-

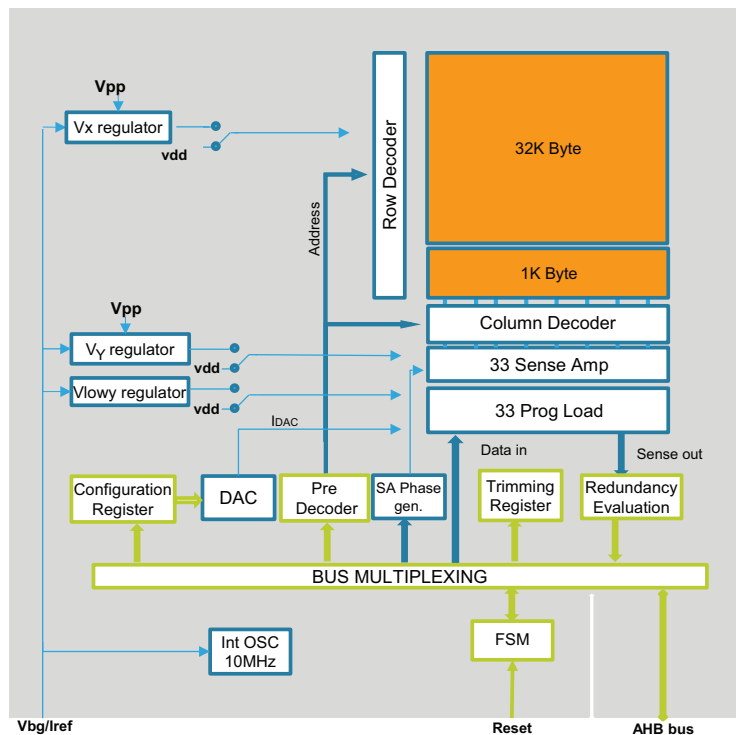


Figure 2.1: Macrocell block diagram.

tracted statistical data. In addition to the 16 PCM macrocells, the developed IC also includes a Built-In Self-Test (BIST) block, several test-chip memory registers, a reference generator block, and the circuitry that manages the input/output interface. The reference generator block provides the macrocells with a bandgap voltage, $V_{BG} = 1.2$ V, as well as with a reference current, $I_{ref} = 10$ μ A. This block is not included in each single macrocell, since the designed IP shares the reference generator block with the target system in order to save silicon area.

The schematic block of the macrocell is depicted in Fig. 2.1, and includes: the memory array, a finite state machine, 2 decoders (namely, the row and the column decoder), 3 voltage regulators, an internal oscillator, a digital-to-analog converter, and several internal registers.

The majority of the macrocell area is occupied by the PCM array, which contains 557,568 cells divided in 528 row (WLs) and 1056 column (BLs). The array is divided in 32 groups of 33 columns. One column for each of these

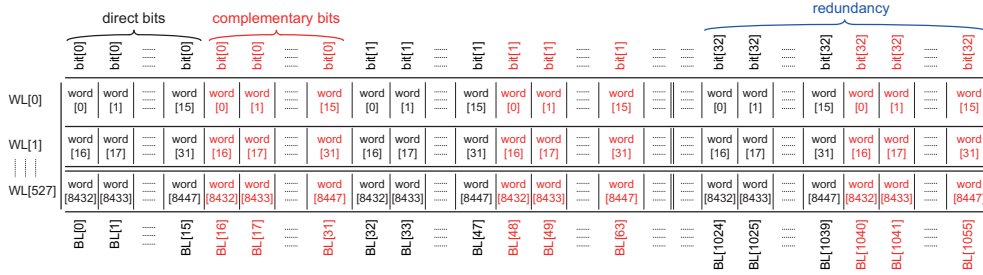


Figure 2.2: Schematic representation of the PCM array that shows the physical bit scrambling.

groups is not used as standard storage space but is actually utilized to implement redundancy: if a standard column (i.e., a data storage column) does not operate correctly after fabrication, it is possible, during Electrical Wafer Sort (EWS), to redirect its address to the address of the redundancy column present in the same group, thus increasing the yield. The readdressing process is fully transparent to the user. The standard columns are therefore 1024, and the redundancy columns are 32. Data are stored in the memory using a differential approach and, therefore, cells are split into two categories: Direct Cells (DCs) and Complementary Cells (CCs). When a program operation is issued in order to store a given data, the information is written as it was received into DCs, but is also complemented and, then, written into the corresponding CCs. In this way, each data present in the memory has a complementary counterpart, which has been programmed shortly after it. Therefore, the effective storage capacity of the PCM array is 33 KB, namely 32 KB (WL[0÷511]) dedicated to user's data and 1 KB (WL[512÷527]) used as a reserved sector to store system data (e.g., trimming configurations, voltage values, malfunctioning columns to be readdressed, etc.). The minimum data portion that can be addressed correspond to a word of 32 (differentially stored) bits and, thus, consists of 64 cells (32 DCs and 32 CCs that store 32 direct bits and 32 complementary bits, respectively). A Word Line contains 16 words and the total number of words and, hence, addresses in the array are 8448 (8192 dedicated to users and 256 reserved to the system or the application).

The memory bits are physically scrambled throughout the array, as shown in Fig. 2.2. This is the outcome of layout optimization aimed at minimizing the parasitic effects (due to interconnections) and allowing the BLs routing to the sense amplifiers (as will be explained in Section 2.2) and the column decoder. The direct bit[0] of word[0] occupies the leftmost cell of WL[0] (i.e.,

of the top row). Moving toward the right, all the direct bit[0] of the successive 15 words (word[1÷15]) are stored in the next memory cells in an ascending word order. The following 16 cells are occupied by the complementary bit[0] in the same order as in the case of direct bits. The subsequent 32 cells are assigned to all bit[1] (first the direct and, then, the complementary bit) of word[0÷15]. The same pattern is repeated along the whole WL[0] with the remaining bits (bit[2÷31]). It is worth to point out that the rightmost 32 cells are occupied by bit[32] of word[0÷15], which correspond to redundancy bits. The succeeding Word Lines (WL[1÷527]) exhibit the exact same pattern organization explained for the case of WL[0].

A finite state machine (FSM) controls the internal operations of the macrocell (e.g., boot, write, read, etc.), and manages data receiving and sending through the 32 bit Advanced High-performance Bus (AHB), which exploits the AMBA (Advanced Microcontroller Bus Architecture) protocol. This protocol is open source and is commonly used to carry out the communications inside a SoC: it therefore guarantees an easier interaction of the macrocell with the rest of the blocks, once employed in the target integrated system. A clock is provided to the FSM by an internal oscillator that is designed to work at 10 MHz and is trimmed during the EWS phase to adjust its frequency in order to attenuate the effect of process spreads. Furthermore, during boot operation, the FSM loads the trimming configurations and the pulse settings (explained in details in the following section) from the memory reserved sector to the macrocell registers. To guarantee that the data contained in the reserved sector are valid, a verification code has to be written in the first reserved memory location. The FSM proceeds in loading the configurations only if the verification code is correctly recognized, otherwise it uses the default value for each setting. All the data stored in the reserved sector are written with a majority voting approach (and, hence, each information is copied in 3 different memory locations) to ensure an enhanced robustness of the system.

The following sections will focus on write and read procedures and circuits, which are the most significant from the analog design point of view.

2.1 Write Circuitry and Procedure

The write operation on PCM cells is carried out exploiting the current mode programming approach mentioned in Chapter 1. The implemented program and verify algorithm demands a group of short rectangular-shaped current pulses (i.e., RESET pulses) and a class of long trapezoidal-shaped pulses (i.e., SET pulses). Therefore, one the most critical aspects of the program circuitry

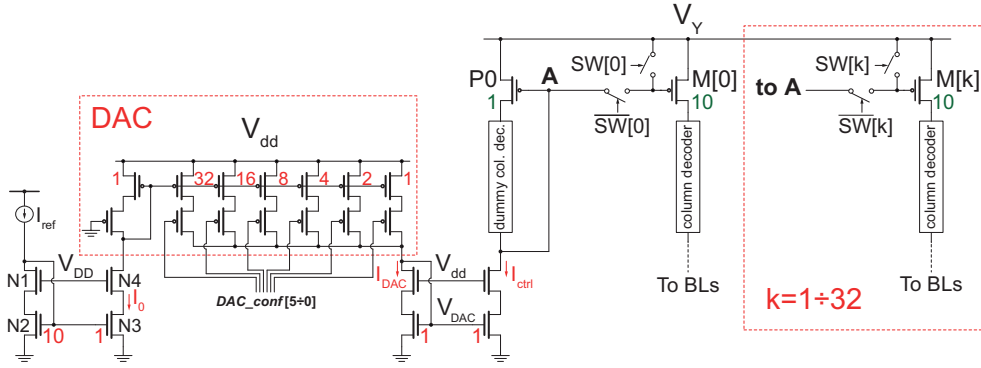


Figure 2.3: Schematic of the programming circuitry.

is the generation of a precise and variable current to be employed by the finite state machine to accurately reproduce the desired current pulses, which have to be fed to the PCM cells for successful programming.

The circuit that generates and shapes suitable programming current pulses is shown in Fig. 2.3. The circuit receives, as an input, an external current, $I_{ref} = 10 \mu\text{A}$, that is provided to the macrocell by the above mentioned reference generator located in the periphery of the chip. The reference current is divided by 10 thanks to the different aspect ratios of NMOS transistors $N2$ and $N3$, which implement a current mirror. Transistors $N1$ and $N4$ act as cascode devices, thus, improving the current mirror accuracy. The output current of the mirror ($I_0 = I_{ref}/10 = 1 \mu\text{A}$) is fed to the control branch of a 6-bit Digital to Analog Converter (DAC), implemented with PMOS transistors, to be used as a reference current to be multiplied [5]. The DAC has six branches that implement six independent current mirrors. Each of these branches can generate a copy of I_0 with a different mirroring ratio (i.e. 1, 2, 4, 8, 16, and 32) and is controlled by a dedicated digital signal ($\text{DAC_conf}[i]$ with $i = 0, 1, 2, 3, 4, 5$). In this way, when the i -th bit is active, a current equal to $2^i \mu\text{A}$ is delivered to the DAC output node. It is worth to point out that the bits of signal DAC_conf are considered active when they are set to a ‘0’ logical value, since they are connected to the terminal gate of PMOS transistors, and inactive when they are in the complementary state. By superimposing the effect of multiple active bits, it is thus possible to obtain any output current (I_{DAC}) between 0 (i.e., $\text{DAC_conf} = 111111$) and $63 \mu\text{A}$ (i.e., $\text{DAC_conf} = 000000$) with a nominal resolution equal to $I_0 = 1 \mu\text{A}$. I_{DAC} is fed to another NMOS cascoded current mirror with unity gain. The output

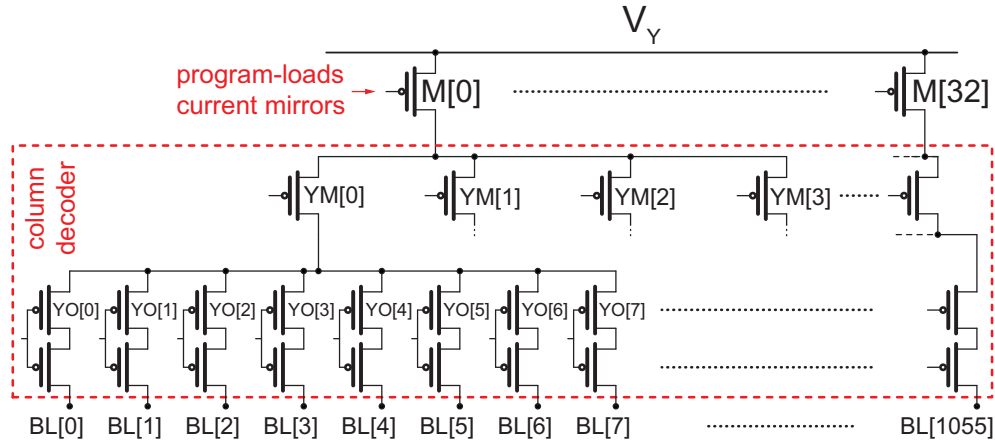


Figure 2.4: Circuit schematic of the macrocell column decoder. For the sake of simplicity, only the leftmost branch (i.e., the branch that drives the first 8 Bit Lines) is completely shown.

current, I_{ctrl} , of this current mirror is used to feed a branch that will be referred to as “program-control branch” hereinafter since it controls the current fed to the cells under programming by means of dedicated current mirrors. These current mirrors (including the program-control branch) are powered by V_Y , which is a programmable biasing voltage, higher than V_{dd} , generated by a dedicated voltage regulator. This choice is made to allow the FSM to set the voltage V_Y depending on the programming pulse that has to be delivered to the cells: indeed, the voltage across the selected Bit Lines is a function of the programming current and can easily exceed 2 V. In this way, the programming circuits can operate correctly and generate the required current pulses. The characteristics, the circuits, and the design choices of the V_Y voltage regulator will be described in Chapter 3. The program-control branch is connected, by means of 33 independent switches, to the output section of 33 program-load current mirrors (transistors $M[0 \div 32]$), which feed the programming current to the memory cells selected by the column decoder. The mirroring ratio between the program control and each program load is 1 to 10, thus allowing a substantial reduction of the macrocell power consumption during write operations. A dummy column decoder is placed on the program-control branch in order to improve the matching with the output sections of the current mirrors, thus providing adequate mirror accuracy.

The circuit used to implement the column decoder is depicted in Fig. 2.4:

even though is powered by a voltage higher than V_{dd} , it is composed by low-voltage transistors. This design choice is dictated by layout requirements: the column decoder has to fit in a given area in order to not increase the memory array pitch. This condition can hardly be met when using high-voltage MOS transistors, which are intrinsically larger than low-voltage transistors. Being forced to use low-voltage MOS transistors imposes, therefore, a careful design and management of the power supply, which has to be sufficiently high to guarantee the correct circuit behavior and, at the same time, adequately low to avoid any damage to transistors.

Two addressing levels have been introduced to obtain the decoded path from the 33 program-load current mirrors to the 1056 BLs. Any Bit Line is selected only when the corresponding path to V_Y is enabled. The first addressing level is composed by a total of $4 \times 32 = 132$ $YM[j]$ PMOS transistors ($j = 0$ to 3): 4 YM transistors $\{YM[0]$ to $YM[3]\}$ are connected to the drain terminal of each of the 33 $M[i]$ transistors. In this way, 2 address bits allow the current generated by one program-load current mirror to be fed to in one of the four paths toward 8 YO transistors. Moreover, $8 \times 4 \times 32 = 1056$ $YO[k]$ ($k = 0$ to 7) PMOS transistors (i.e., one transistor for each Bit Line) implement the second addressing level. An additional PMOS transistor is connected in series with each $YO[k]$, so as to reduce the voltage drop across each activated low-voltage transistor in the column decoder, thereby achieving a better reliability of the circuit. In order to guarantee that the programming current generated by transistors $M[i]$ is fed correctly into the selected PCM cell, only one of the four $YM[j]$ transistors, connected to its drain terminal, as well as only one of the eight $YO[k]$ transistors that are connected to the active YM transistor can be turned on at a given time. Furthermore, thanks to this architecture, the $M[i]$ transistor can program one cell belonging to any of the 32 associated Bit Lines, i.e. of Bit Lines from $BL[32i]$ to $BL[32i + 31]$ (e.g. $BL[0 \div 31]$ for $i = 0$, $BL[32 \div 63]$ for $i = 1$, $BL[64 \div 95]$ for $i = 2$, etc). The digital signals that drive the two addressing levels can swing between V_Y and V_{lowY} , where V_{lowY} is a biasing reference generated by a dedicated programmable voltage regulator. During write operation, V_{lowY} can be set by the FSM to a suitable value that is a function of the BL voltage and, thus, of the programming current.

A row decoder selects the addressed cell by raising the corresponding Word Line from ground to V_X , which is a bias voltage generated by the row voltage regulator. V_X , which is also programmable, is applied directly to the gate terminal of the NMOS transistors that act as cell selectors.

The described programming circuits are able to write from 0 to 33 cells in parallel with a programmable current between 0 and 630 μA and a cur-

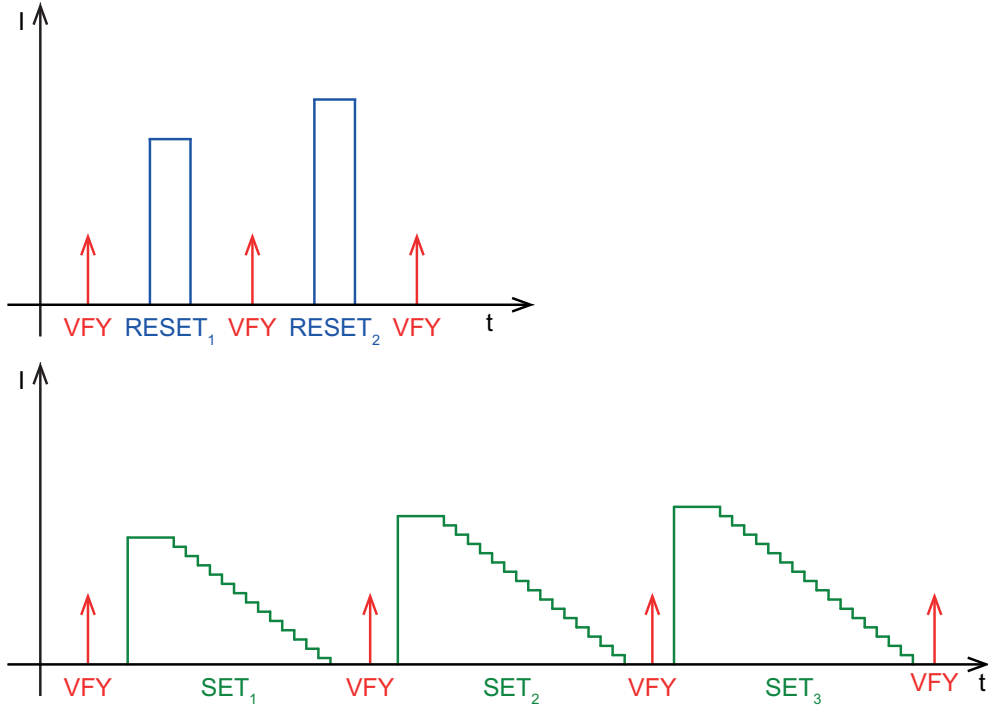


Figure 2.5: Macrocell RESET (top) and SET (bottom) program algorithms.

rent resolution of $10 \mu\text{A}$. Another important feature of the write circuitry is the possibility to dynamically change the DAC configuration during a current pulse, to modify the programming current that flows through the cells. In this way, it is possible to generate triangular- and trapezoidal-shaped current pulses, which are needed to crystallize the phase change material: this can be obtained starting from the desired current plateau level (i.e., from the desired `DAC_conf` plateau value) and, then, decreasing it (i.e., incrementing, in modulus, the `DAC_conf` value). Due to the discrete nature of the programming current generation, it is not possible to achieve a constant and smooth slope in the falling edge of SET pulses. The linear ideal slope is approximated with a sequence of small and short current steps (i.e., the step amplitude is much smaller than the SET pulse amplitude and the step time length is much shorter than the SET pulse duration). The programming results obtained with these discrete pulses and with standard SET pulses were compared by technologists: the two obtained cell-resistance distributions did not show significant differences.

The complete program algorithm is shown in Fig. 2.5. When a cell has to be programmed, its content is firstly read (i.e. verified, VFY): if the obtained data match the bit that has to be written, the procedure is ended, otherwise a first pulse (i.e., SET₁ or RESET₁, depending on the data) is delivered to the cell. The cell current, I_{cell} , is now verified by means of a comparison with a reference current, I_{vfy} : if the verify operation succeeds (i.e., $I_{cell} > I_{vfy}$ or $I_{cell} < I_{vfy}$ in the case of SET or RESET operation, respectively) the write procedure is concluded. Otherwise, another pulse (i.e., SET₂ or RESET₂, depending on the data) is applied to the cell and, then, the cell content is verified again. This procedure is iterated until verify succeeds or the number of applied pulses reaches a predetermined upper bound (in this case, the memory cell is considered to fail). The SET operation can require up to three different pulses, whereas in the case of RESET procedure the maximum number pulses is two. The full pulses specification are summarized in Tab. 2.1. It is worth to notice that each pulse applied after the first has a higher current amplitude when compared to the previous pulse and, thus, can affect a larger portion of the PCM layer, resulting, hence, in a more effective crystallization or amorphization of the phase change material in the case of SET or RESET operation, respectively.

The same write circuitry is also utilized, during EWS, to carry out additional operations, namely the forming and the pre-soldering procedure. On the one hand, the forming operation is required to correctly initialize the phase change material, prior to any other programming pulse, so as to drive the alloy atomic structure of the cells into an optimal crystalline state. From an operative point of view, the forming pulse is equivalent to a SET_{*i*} pulse (where $i = 1$ to 3) with a higher plateau current values. On the other hand, the pre-soldering procedure is a programming operation that guarantees an optimal data retention. This feature is achieved by using stronger SET and RESET pulses (i.e., pre-soldering SET and pre-soldering RESET, respectively) than in standard programming and, thus, acting on a larger volume of the phase change material. The purpose of the latter high-energy pulses is to write, in PCM cells, data that can be successfully read after the industrial soldering procedure (e.g., wave soldering) of the IC on the application board. The standard user mode pulses can not guarantee data retention in these conditions, since the soldering process can typically reach temperatures well above the nominal operative range (i.e., temperature larger than 220°) for about 2 hours. It is worth to point out that both pre-soldering pulses are designed to be applied very few times during the cell lifetime, because they can severely damage the cell heater and, thus, cause memory failures when utilized frequently.

Table 2.1: Pulse specifications.

| | Pulse | I plateau (μA) | T pulse (μs) |
|-----------|---------------------|-----------------------------|---------------------------|
| User mode | SET ₁ | 300 | 0.5–5 |
| | SET ₂ | 350 | 0.5–5 |
| | SET ₃ | 400 | 0.5–5 |
| | RESET ₁ | 450 | 0.1–0.5 |
| | RESET ₂ | 500 | 0.1–0.5 |
| EWS | Pre-Soldering SET | 550 | 0.5–5 |
| | Pre-Soldering RESET | 600 | 0.1–5 |
| | Forming | 600 | 0.5–5 |

2.2 Read Circuitry and Procedure

The read circuits, shown in Fig. 2.6, are powered with the low voltage supply V_{dd} and, thus, do not require any voltage regulator.

The purpose of the reading circuit is to compare two currents, detect which is larger, and provide the information to the output as a digital data. 33 sense amplifiers are implemented in the macrocell in order to guarantee the parallel readout of an entire word. Even though a word is composed by 32 bits, one additional sense amplifier is needed to manage the redundancy: when a column is readdressed in the redundant memory space, the sense amplifier of the malfunctioning Bit Line is deactivated and the sense amplifier associated to the corresponding redundant Bit Line is turned on.

The 1056 BLs have to be connected to the 33 sense amplifiers and, therefore, two level of addressing are implemented, similarly to the case of programming circuit. However, the reading addressing circuits are composed by NMOS transistors and, since they are driven by low-voltage signals, do not require an additional transistor connected in series to decrease the voltage drop across their source-drain terminals.

The first level of read addressing is made of 1056 transistor YO_N , which have their source terminals connected to a single BL, whereas their drain terminal are gathered in groups of 8 and connected to a sense-amplifier Main Bit Line (MBL_{SA}). Similarly, 2 YM_N transistors have their drain terminal connected to the direct input node of a sense-amplifier (SA_{INd}), while their source terminal is connected to 2 distinct MBL_{SA} . The complementary input node (SA_{INc}) of the same sense amplifier is connected, in an equivalent manner, to

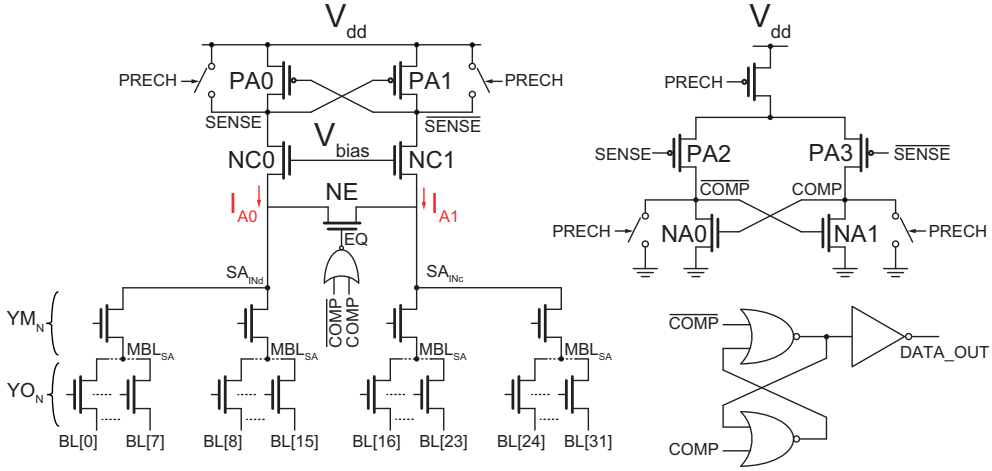


Figure 2.6: Circuit schematic of a macrocell sense amplifier. This schematic depicts the connection of sense amplifier[0] to the corresponding Bit Lines.

2 other MBL_{SA} . Therefore, the i -th sense amplifier has the direct input node that can be connected, by properly driving transistors YO_N and YM_N , to one of the Bit Lines between $BL[32i]$ and $BL[32i + 15]$, whereas its complementary input node can be connected to any Bit Line between $BL[32i + 16]$ and $BL[32i + 31]$. Since a sense amplifier has to compare a direct cell with its complementary copy, the aforementioned physical scrambling (Fig. 2.2) has been implemented in order to correctly pair a DC with the corresponding CC (i.e. the complementary cell that contains the complementary bit with respect to the corresponding direct-cell data).

When a sense amplifier is in idle state, the pre-charge signal ($PRECH$) is active and inhibits the transistors perform the current comparison, namely $PA0$, $PA1$, $NA0$, $NA1$, $PA2$, and $PA3$. In this way, when the sense amplifier is activated and the comparison phase begins, the internal nodes of the circuit are biased with supply voltages (i.e., $SENSE = \overline{SENSE} = V_{dd}$ and $COMP = \overline{COMP} = GND$) and the two branches are balanced. To address the selected cells, the corresponding WL is raised to the low-voltage power supply value (V_{dd}) while, simultaneously, one YM_N and one YO_N transistors are turned on, in order to select the two desired BLs (one for a direct, the other for the complementary cell) and connect them to the sense-amplifier inputs. The bias voltage V_{bias} is set to turn on transistors $NC0$ and $NC1$ and impose the same voltage value (i.e., $V_{bias} - V_{th}$) on the two sense-amplifier inputs. Assuming

that the active transistors YM_N and YO_N have a negligible voltage drop, since they both are driven to reach their triode state, the selected Bit Lines are also raised by $NC0$ and $NC1$ to the same voltage value as the sense-amplifier inputs (i.e., $V_{bias} - V_{th}$). This architecture therefore allows biasing the selected BLs with an equal voltage, which is fundamental to precisely generate two I_{cell} with the purpose of correctly determining which is larger.

Once all node voltages are settled, the pre-charge signal is deactivated (i.e., $PRECH = GND$) and the comparison phase starts. Node **SENSE** is discharged by I_{A0} (i.e., the current sunk from the direct cell) whereas, in the opposite branch, $\overline{\text{SENSE}}$ is simultaneously discharged by I_{A1} (i.e., the current that flows through the complementary cell). Since the two branches are perfectly symmetrical (from both circuit and layout points of view), the different amplitudes of the two discharging currents is the only parameter that influences how fast the voltages on nodes **SENSE** and $\overline{\text{SENSE}}$ drop. The node that first reaches a voltage equal to $V_{dd} - V_{thP}$, where V_{thP} is the threshold voltage of a low-voltage PMOS transistor, immediately turns on the PMOS transistor of the opposite branch (i.e., either $PA0$ or $PA1$) and, hence, triggers the two cross-coupled transistors (i.e., $NA0$ - $NA1$). Thanks to this approach, the small unbalancing of the two cell currents is amplified and, then, fed to a set-reset NOR latch, which provides the digital signal to the output (i.e., **DATA_OUT**).

It is worth to point out that, when a read operation is complete, signal **PRECH** is reactivated and, thus, transistor NE is activated, thus equalizing the two sense-amplifier inputs. This operation is carried out to cancel the effect of the previous read operation on the following one.

The described circuitry not only manages the read operation, but also carries out the verify step during the write algorithm. In the latter case, however, the current generated in the selected cell is compared with a programmable current. The programmable current is generated with the DAC circuit used during program operations (see Fig. 2.3), and is then fed to the one input of the sense-amplifier. The FSM can set **DAC_conf** to obtain the desired value I_{vy} , which can be different depending on the programming pulse just applied or changed in order to obtain a larger margin as may be required in particular applications.

Chapter 3

Design of Analog Circuits

In the following Sections, the design of two analog blocks (i.e., a voltage regulator and an enhanced current mirror) that were implemented in the Spider-Mem chip will be described and analyzed. In particular, the V_Y voltage regulator, which supplies the Spider-Mem programming circuits, will be presented in Section 3.1, whereas the enhanced current mirror that reduces the delay time in programming current generation will be described in Section 3.2. Section 3.3 will present a novel architecture for charge pumps that exploits an optimized charge-transfer technique to achieve a higher driving capability and an improved power efficiency. The two following sections (i.e., Section 3.4 and Section 3.5) will show two theoretical studies: an enhanced voltage buffer compensation technique for two-stage CMOS amplifiers and an analysis aimed at maximizing the bandwidth of a two-stage operational amplifier under power consumption and area constraints. All the theoretical results presented in the last three sections of this chapter will be used to provide guidelines to designers during the development of the next version of Spider-Mem.

As mentioned in Chapter 2, the design and the implementation of the Spider-Mem chip have been carried out using the BCD9s technology. This technology offers the possibility to integrate both high- and low-voltage MOS transistors. The main difference among these kinds of transistors is the maximum voltage that they can manage without going in breakdown and, thus, being permanently damaged. On the one hand, high-voltage transistors have a thicker gate oxide, that is sized to safely operate at the high-voltage power supply, $V_{pp} = 4.5 \div 5.5$ V. As a drawback, these transistors have an increased minimum channel length $L_{min,HV} = 600$ nm and a larger threshold voltage $V_{th,HV} \approx 800$ mV (these values hold for both NMOS and PMOS transistors). On the other hand, low-voltage transistors are designed to safely work

at nominal low-voltage supply $V_{dd} = 1.6 \div 2.0$ V. These transistors have a thinner gate-oxide layer with respect to their high-voltage counterpart, and their minimum channel length is $L_{min,LV} = 180$ nm. Low-voltage transistors are also able to achieve a better transconductance than high-voltage transistors with the same dimensions. Finally, they exhibit a lower threshold voltage $V_{th,LV} \approx 500$ mV. It is worth to notice that BJTs and DMOS transistors, which are available in this technology, are not used in the design of the macrocell, mainly due to area constraint.

3.1 V_Y Voltage Regulator Design

To guarantee the correct behavior of all the components as well as to ensure successful programming operations, the V_Y voltage regulators have to satisfy the following specifications and requirements.

- **Voltage programmability** of the regulated output is required since V_Y has to be set to different values depending on the programming current amplitude, which causes different voltage drops across the PCM cell. Therefore, a higher V_Y is required for higher current pulses to provide the write circuitry with a sufficiently large voltage difference to correctly operate. V_Y has to be digitally programmable by the finite state machine to provide voltages from 4.0 V to 5.2 V.
- The **minimum accuracy** required in static conditions by the application is $\pm 2\%$ independently from the V_Y value set by the FSM.
- The regulator has to be able to guarantee all the functionalities in the **temperature range** from -40 °C to $+150$ °C and under all specified **supply conditions** (i.e., $V_{dd} = 1.55 \div 2.00$ V and $V_{pp} = 4.5 \div 5.5$ V). It is worth to notice that the minimum value of V_{dd} is smaller and, thus, more conservative than the minimum allowed voltage applied to the circuit (i.e., $V_{dd} = 1.6$ V) to take into account the voltage drop due to interconnections.
- A maximum **static power consumption** of $750 \mu\text{W}$ as well as an almost-zero power consumption when the FSM is not running write operations are required.

It is important to notice that it could be requested by the FSM to provide a regulated voltage higher than the high supply voltage (e.g., $V_Y = 5$ V when $V_{pp} = 4.5$ V). Obviously, it is not possible, under these conditions, to achieve

Table 3.1: Pulse specifications and total current requirements.

| Pulse | | I plateau (μA) | V_Y (V) | Regulator current (mA) | | | |
|-----------|---------------------|-----------------------------|-----------|------------------------|----------|----------|----------|
| | | | | PAR = 2 | PAR = 16 | PAR = 32 | PAR = 33 |
| User mode | SET ₁ | 300 | 3.7 | 0.6 | – | 9.3 | – |
| | SET ₂ | 350 | 3.9 | 0.7 | – | 11.2 | – |
| | SET ₃ | 400 | 4.1 | 0.8 | – | 12.8 | – |
| | RESET ₁ | 450 | 4.3 | 0.9 | 7.2 | – | – |
| | RESET ₂ | 500 | 4.5 | 1 | 8.0 | – | – |
| EWS | Pre-Soldering SET | 550 | 4.7 | 1.1 | – | – | 18.2 |
| | Pre-Soldering RESET | 600 | 5.2 | 1.2 | – | – | 19.8 |
| | <i>Forming</i> | 600 | 5.2 | 1.2 | – | – | 19.8 |

the set values of V_Y , therefore, it is expected that the regulator provides a voltage output close to V_{pp} . It is worth to point out that the only operations that require a value of V_Y higher than the minimum allowed value of V_{pp} (i.e., 4.5 V) are carried out during electrical wafer sort and can take advantage of programmable and well-defined power supplies (i.e., both V_{dd} and V_{pp} can be set at the desired value with $\pm 1\%$ accuracy) and controlled temperature (i.e., $T \approx 25^\circ$). Therefore, the case in which $V_Y > V_{pp}$ can be easily avoided.

The V_Y voltage regulator has to comply with additional requirements since it carries out critical tasks, described in the following.

- It has not only to power all the programming circuitry (described in Chapter 2), but also to bias the n -well diffusions where all the PMOS transistors that carry out programming operations are implemented. Therefore, the V_Y voltage regulator has to drive a large and variable capacitive load: $C_L = 66 \div 86$ pF (depending on process variations and biasing conditions).
- The V_Y voltage regulator has to provide the programming current to the selected PCM cells and, thus, has to guarantee a sufficiently high driving capability to ensure the possibility to program up to 33 cells in parallel (i.e., ≈ 20 mA during EWS, and ≈ 13 mA in user mode, as shown in Tab. 3.1).
- Finally, a limited output voltage drop must be ensured when a large current is delivered to the load so as to allow the programming circuits to operate correctly even at the beginning of the output current pulse. This is a particularly demanding task, especially when associated with low static-power consumption.

Table 3.1 shows the specifications requested to successfully generate the current pulses: the plateau current, the V_Y voltage, and the total current that

the voltage regulator has to deliver. The voltage V_Y varies, for each pulse, depending on the amount of programming current that has to be provided to a single PCM cell. The voltage of a selected Bit Line, V_{BL} , is a function of both the cell resistance and the programming current: therefore, a pulse with a higher plateau current generates a higher V_{BL} and, thus, requires a larger V_Y to guarantee the correct behavior of the write circuit.

Moreover, the programming operations can be carried out with different parallelism, which can be set to several values (i.e. $\text{PAR} = 1, 2, 4, 8, 16, 32, 33$). However, Tab. 3.1 only shows the voltage-regulators total currents that correspond to the two configurations available to the final users: low parallelism (i.e., $\text{PAR} = 2$ for both SET and RESET operations), which is an option designed for systems powered with a supply voltage unable to provide large currents to the macrocell; and high parallelism, which is enabled when the program throughput is a critical aspect (i.e., $\text{PAR} = 32$ and $\text{PAR} = 16$ for SET and RESET operations, respectively). The last portion of Tab. 3.1 is dedicated to programming pulses that are used during EWS. In this phase, the chosen parallelism is 33 since both the standard and the redundant cells have to be programmed and read in order to detect the faulty cells and, thus, activate the redundancy.

3.1.1 Voltage Regulator Topology

The choice of the voltage regulator topology plays a key role in achieving the programmability and the accuracy dictated by specifications. First of all, since a bandgap voltage generator is embedded in the test chip along with the 16 macrocell and is designed to intrinsically attenuate the effect of temperature variations, it has been exploited as reference to generate the regulated output voltage. This approach does not compromise the compatibility of the macrocell when it is included in a generic SoC, since it is very common to share a reference voltage generator among several circuit blocks.

A simple idea to obtain a programmable, linear, and accurate output voltage [6, 7] is to firstly generate a programmable current I_{R1} and, then, force it to flow into a resistor that closes a negative feedback loop around an operational amplifier, as shown in Fig. 3.1. The generated voltage V_{out} can be set to different values by modifying the amplitude of I_{R1} . In particular, assuming that the open loop gain of the operational amplifier is sufficiently high, voltage v^- turns out to be equal to V_{BG} and the obtained output voltage is

$$V_{out} = V_{BG} + I_{R1}R_1. \quad (3.1)$$

A voltage generated with this approach is, therefore, a function of the resistor

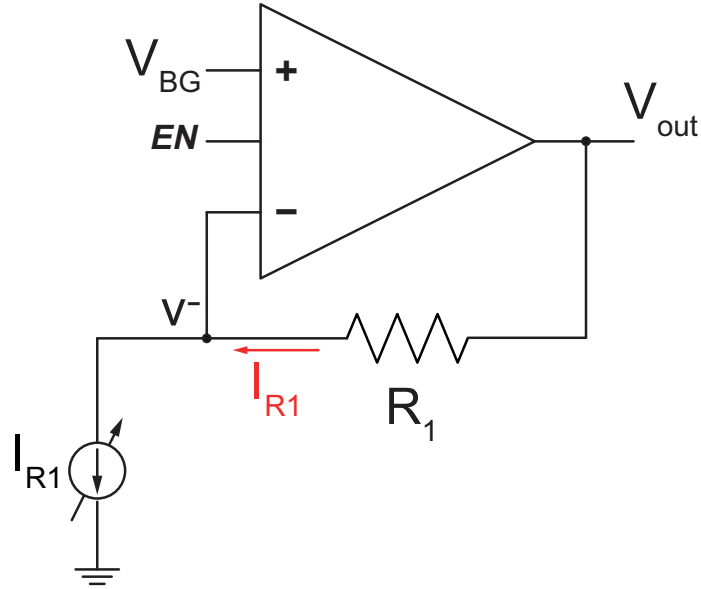


Figure 3.1: Ideal schematic of a voltage regulator.

value and, thus, is prone to any resistance fluctuation (e.g., due to thermal variations). Moreover, the distribution of the effective resistance value of an integrated resistor has a large statistical standard deviation due to IC fabrication inaccuracies (e.g., irregular dopant concentrations, fabrication process random variations, masks misalignment, etc.). Hence, the output voltage of this simple schematic is also affected by poor accuracy and significant temperature dependence.

The schematic shown in Fig. 3.2 has been designed to overcome the above limitations. In this topology, voltage V_{BG} is exploited to generate a current, I_{R0} , proportional to the sum of two series-connected additional resistors, R_0 and R_{trim} . Operational amplifier A0 is used to impose a negative feedback, which generates I_{R0} by maintaining a constant voltage drop equal to $V_{BG} = 1.2$ V across R_0 and R_{trim} . Both R_0 and R_{trim} are implemented using a polysilicon resistor: R_0 was sized to achieve a nominal 110 k Ω resistance, whereas R_{trim} can vary from 0 to 17.5 k Ω with a nominal resolution of 2.5 k Ω . Indeed, R_{trim} consists of three series-connected resistors, sized to achieve a resistance of 10 k Ω , 5 k Ω , and 2.5 k Ω , respectively, that can be individually short circuited by turning on three corresponding switches. These switches are activated by a 3-bit digital signal controlled by the finite state machine. The

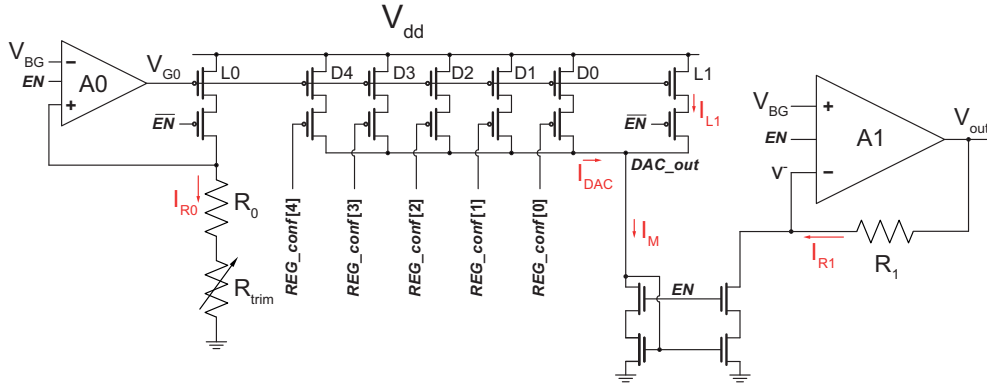


Figure 3.2: Circuit schematic of the V_Y voltage regulator.

optimal configuration of these three bits (i.e., the configuration that provides the value of $R_0 + R_{trim}$ closest to the desired resistance of 120 k Ω) is calculated, for each macrocell, during EWS and is then stored in a dedicated portion of the reserved sector to make it available to the FSM. The trimming procedure allows any variation between the actual and the desired resistance value to be further attenuated.

Therefore, the nominal value of the generated current is equal to $I_{R0} = V_{BG}/(R_0 + R_{trim}) = 10 \mu\text{A}$. A low-voltage PMOS transistor, $L0$, is included in the feedback loop around $A0$ to provide a means to mirror current I_{R0} , as detailed below. The purpose of the described network is to ensure that the main fluctuations in the current amplitude are only due to the variations of resistances R_0 and R_{trim} , since the voltage variations across the two resistors, in first order approximation, are negligible.

The voltage on the gate terminal of $L0$ (V_{G0}), imposed by $A0$, is used as a reference to generate several copies of I_{R0} . V_{G0} is set by the negative feedback loop to $V_{dd} - V_{ov0}$, where V_{ov0} is the overdrive voltage (in absolute value) that forces a current equal to I_{R0} to flow through a PMOS transistor that has the same aspect ratio as $L0$. The length and the width of transistor $L0$ were set to $L_{L0} = 1 \mu\text{m}$ and $W_{L0} = 10 \mu\text{m}$, respectively. V_{G0} is fed to a circuit that acts as a digital-to-analog converter.

This circuit has five branches: each of them generates a scaled copy of I_{R0} and can be, individually, either turned on or disabled by a dedicated 5-bit digital signal (i.e., $\text{REG_conf}[4 \div 0]$). The i -th branch is composed by a series connection of two PMOS transistors: the top transistor, Di , has its gate terminal biased by V_{G0} , whereas the bottom transistor acts as a switch driven

by the i -th bit of signal `REG_conf[4÷0]` (i.e., `REG_conf[i]`). Hence, a branch can be either disabled or activated by driving high (i.e., to V_{dd}) or low (i.e., to ground) the gate terminal of the correspondent switch transistor. Each top transistor has a different aspect ratio in order to set the desired current in the corresponding branch. In particular, the i -th top transistor (i.e., D_i) was designed with length $L_i = L_{L0}$ and width $W_i = (2^i/10)W_{L0}$. The current delivered by this circuit (I_{DAC}) to the output node, `DAC_OUT`, can therefore vary from 0 to $31 \mu\text{A}$ depending on the value of `REG_conf`. A PMOS transistor, $L1$, generates an additional scaled copy of I_{R0} , which is independent from the value of signal `REG_conf`. Being $L1$ aspect ratio sized as $11 \mu\text{m}/1 \mu\text{m}$, its current turns out to be $I_{L1} = 11 \mu\text{A}$.

The total current that flows through node `DAC_OUT`, I_M , is mirrored with unity gain in order to generate I_{R1} , which, thus, can be expressed as

$$I_{R1} = I_M = I_{DAC} + I_{L1} = 11 \mu\text{A} + (0 \div 31 \mu\text{A}) = 11 \div 42 \mu\text{A}. \quad (3.2)$$

This current is an amplified copy of I_{R0} and can be rewritten as $I_{R1} = \alpha I_{R0}$, where α is a rational parameter (i.e., $\alpha \in \mathbb{Q}$) that represents the variable mirroring ratio and is equal to

$$\alpha = \frac{11}{10} \div \frac{42}{10}. \quad (3.3)$$

As already mentioned, the mirroring ratio and, thus, α are set by signal `REG_conf`, which is controlled by the finite state machine.

So, finally, I_{R1} is sunk from node v^- and, therefore, is forced to flow from V_{out} through R_1 to set the output voltage:

$$\begin{aligned} V_{out} &= V_{BG} + I_{R1}R_1 = V_{BG} + \alpha V_{BG} \frac{R_1}{R_0 + R_{trim}} \\ &= V_{BG} \left(1 + \alpha \frac{R_1}{R_0 + R_{trim}} \right). \end{aligned} \quad (3.4)$$

Table 3.2 shows all the V_{out} values that is possible to achieve through different configurations of `REG_conf` and, thus, through different values of I_{R1} .

Analyzing equation (3.4), it is worth to point out that V_{out} depends not only by V_{BG} but also on α and the ratio of R_1 to $R_0 + R_{trim}$. In order to make this circuit effective, it is therefore extremely important to carefully design the silicon layout implementation of the transistors involved in the current mirroring as well as the polysilicon resistors. The former were, thus, designed with a channel length suitably large (i.e., $L \geq 1 \mu\text{m}$) and with their channel width split in several fingers and then arranged in symmetrical structures

Table 3.2: V_{out} and I_{R1} as a function of REG_conf.

| REG_conf | I_{R1} | V_{out} | REG_conf | I_{R1} | V_{out} |
|----------|------------------|-----------|----------|------------------|-----------|
| 11111 | 11 μA | 2.3 V | 01111 | 27 μA | 3.9 V |
| 11110 | 12 μA | 2.4 V | 01110 | 28 μA | 4.0 V |
| 11101 | 13 μA | 2.5 V | 01101 | 29 μA | 4.1 V |
| 11100 | 14 μA | 2.6 V | 01100 | 30 μA | 4.2 V |
| 11011 | 15 μA | 2.7 V | 01011 | 31 μA | 4.3 V |
| 11010 | 16 μA | 2.8 V | 01010 | 32 μA | 4.4 V |
| 11001 | 17 μA | 2.9 V | 01001 | 33 μA | 4.5 V |
| 11000 | 18 μA | 3.0 V | 01000 | 34 μA | 4.6 V |
| 10111 | 19 μA | 3.1 V | 00111 | 35 μA | 4.7 V |
| 10110 | 20 μA | 3.2 V | 00110 | 36 μA | 4.8 V |
| 10101 | 21 μA | 3.3 V | 00101 | 37 μA | 4.9 V |
| 10100 | 22 μA | 3.4 V | 00100 | 38 μA | 5.0 V |
| 10011 | 23 μA | 3.5 V | 00011 | 39 μA | 5.1 V |
| 10010 | 24 μA | 3.6 V | 00010 | 40 μA | 5.2 V |
| 10001 | 25 μA | 3.7 V | 00001 | 41 μA | 5.3 V |
| 10000 | 26 μA | 3.8 V | 00000 | 42 μA | 5.4 V |

(i.e., common centroid arrangement) in order to provide a better transistor matching. The polysilicon resistors were also implemented as close as possible and in a shared symmetrical structure. In this way, matching is improved and any resistance variation (e.g., due to thermal fluctuation) affects R_0 , R_1 and R_{trim} in a similar manner.

Providing that the mentioned layout precautions have been taken, this circuit offers, ideally, a cancellation (or, in real ICs, a strong reduction) of the impact on the regulated voltage of any resistance variation, since V_{out} is a function of the resistance ratio. To better understand the behavior of the circuit, let us suppose that a decrease of R_0 due to a temperature increment occurs. Since R_0 and R_{trim} are implemented on silicon in a structure shared with R_1 , it is reasonable to assume that the temperature of the resistors is the same and, thus, their resistance is decreased by the same factor. Therefore, the current flowing through R_0 and R_{trim} increases, since the bandgap voltage is temperature compensated and $I_{R0} = V_{BG}/(R_0 + R_{trim})$, and so does I_{R1} , which is a (scaled) copy of I_{R0} . The voltage drop across R_1 is not affected by the temperature increment, since the resistance reduction and the increase of I_{R1} cancel each other. Thus, there is no significant dependence of the output voltage with respect to temperature.

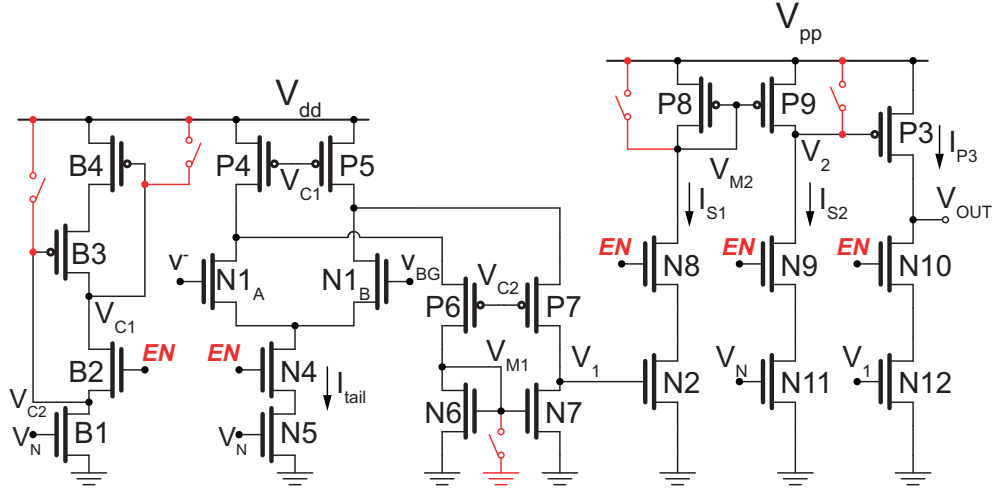


Figure 3.3: Circuit schematic of the three-stage V_Y operational amplifier.

3.1.2 The V_Y Regulator Operational Amplifier

A key block of the designed voltage regulator is operational amplifier $A1$, since its characteristics are the major contributors to the voltage regulator overall performance. The operational amplifier receives an enable signal, EN , and exploits V_{BG} and V_N as reference voltages, where V_N is a voltage generated by a diode-connected NMOS transistor biased with the aforementioned external current reference I_{ref} . A three-stage topology, shown in Fig. 3.3, was chosen to implement the V_Y operational amplifier, since the load, during the programming phase, assumes a low impedance (i.e., a few $k\Omega$), which is mainly due to the heater resistance. This aspect prevents the use of a two-stage topology, since it severely reduces the output stage gain and, in this condition, the total gain of a two-stage topology turns out to be insufficient. On the contrary, with a three-stage solution, even in the case of a low-impedance load, the gain is kept adequately high. As a drawback, this topology requires more power and has worst speed performance with respect to a two-stage operational amplifier.

The first stage of the operational amplifier is powered by the low-voltage supply in order to reduce power consumption, whereas the second and third stages are supplied with V_{pp} to be able to drive V_{out} in the desired voltage range. The first stage consists of an NMOS differential pair (transistors $N1_A$ and $N1_B$) connected to an active load made of two PMOS transistors ($P4$ and $P5$) and two PMOS transistors ($P6$ and $P7$) that implement a folded cascode

scheme. Moreover, a current mirror, made of two NMOS transistors ($N6$ and $N7$) converts the differential signal into a single-ended output, V_1 , that is fed to the second stage. All the transistors (i.e. both the NMOS and the PMOS transistors) that belong to the first stage are implemented with low voltage devices. The differential-pair tail current, I_{tail} , is set by transistor $N5$, which has its gate terminal biased at V_N , and is equal to $10 \mu\text{A}$. A bias branch, consisting in transistors $B1$, $B2$, $B3$, and $B4$, was inserted in the operational amplifier in order to generate both V_{C1} and V_{C2} . Transistor $B1$, thanks to V_N , sets a current equal to $10 \mu\text{A}$ to feed the biasing branch. Therefore, V_{C1} imposes both transistors $P4$ and $P5$ to carry the same amount of current (i.e., $10 \mu\text{A}$), which is then split into two halves: a current of $5 \mu\text{A}$ flows through each of transistors $N1_A$ and $N1_B$, and a current of $5 \mu\text{A}$ flows through each of the two folded-cascode transistors (i.e., $P6$ and $P7$). Transistor $N4$ has two main purposes: firstly, it helps to achieve a more accurate tail current, I_{tail} , and, secondly, acts as a switch, driven by EN , that is able to turn off I_{tail} when the regulator is disabled. Three additional switches, which force V_{C1} and V_{C2} to V_{dd} and V_{M1} to GND , are placed to ensure an almost-zero static power consumption when the programming phase is over.

The input of the second stage corresponds to the gate terminal of an NMOS low-voltage transistor, $N2$, which is connected to V_1 in order to implement a common-source gain stage. The voltage signal on node V_1 is, therefore, converted to a current signal by $N2$ and, then, fed to a PMOS unity-gain current mirror (composed by high-voltage transistors $P8$ and $P9$). The mirrored current signal is injected into a high impedance node, V_2 , thus achieving the second-stage voltage gain, A_2 . Thanks to the double inversion of the signal (i.e., one inversion through $N2$ and another through $P9$), voltage gain A_2 turns out to be positive. This is a fundamental characteristic since it is required to achieve stability when using a classic compensation network, as will be explained later in detail. The quiescent current through the output branch of the second stage, I_{s2} , is set by the aspect ratio of transistor $N11$ to be $I_{s2} = (3/4) I_{ref} = 7.5 \mu\text{A}$. Current I_{s1} is forced to be equal to I_{s2} by the outer negative-feedback loop, which imposes to node V_1 to assume an adequate value in order to set $N2$ to carry $I_{s2} = I_{s1}$. It is worth to point out that both $N2$ and $N11$ are low-voltage transistors, even though their branches are powered by the high voltage supply V_{pp} . Therefore, two additional high-voltage NMOS transistors, $N8$ and $N9$, biased by EN , are placed to form a cascode configuration. When the regulator is active (i.e., $\text{EN} = V_{dd}$), the drain terminals of both $N2$ and $N11$ are kept in the safe operating region by the voltage drop across the gate-source terminals of $N8$ and $N9$, respectively. When EN

is driven to GND, $N8$ and $N9$ are turned off and a switch drives node V_{M2} to V_{pp} , thus allowing the power consumption of the second stage to be reduced to almost zero.

A class-AB output was chosen as the third stage to be able to use a small quiescent current and, thus, save power, while concurrently guaranteeing a very large dynamic output current. The gate terminal of the high-voltage PMOS transistor $P3$, indeed, is connected to node V_2 , whereas transistor $N12$ is driven by V_1 . Since V_1 and V_2 are in phase, each output transistor (i.e., $P3$ and $N12$) manages the amount current flowing through V_{out} in only one specific direction: in detail, $P3$ can only increase the current flowing into the output node (i.e., the current fed to the load), whereas $N12$ can raise the current sunk from V_{out} (i.e., the current sunk from the load). The high-voltage NMOS transistor $N10$ accomplishes the same task as second-stage transistors $N8$ and $N9$. When the macrocell is not executing a program operation, i.e. $EN=0$, an additional switch forces V_2 to V_{pp} to shut off transistor $P3$.

3.1.2.1 Compensation

In the majority of multiple-stage CMOS operational amplifiers, a compensation network is needed to relocate their internal poles, which are intrinsically close to each other. The role of a compensation network is to move the operational amplifier poles to frequency positions that guarantee closed-loop stability, and, thus, ensure a sufficiently large phase margin, φ_m .

The standard and simplest choice to compensate a three-stage operational amplifier is a nested Miller technique [8–10], depicted in Fig. 3.4 referring to the operational amplifier small-signal circuit. In the chosen notation, r_i and g_{mi} are the output resistance and transconductance of the i -th stage, respectively, while C_i is the parasitic capacitance associated to the output node of the i -th stage. To better understand nested Miller compensation, let us consider an operational amplifier composed only by two stages (i.e., the second and the third stages of the operational amplifier of Fig. 3.3) and compensated only with capacitor C_{C1} . It is straightforward to see that, without the presence of C_{C1} , the two main poles of the circuits are associated to nodes V_2 and V_{out} , respectively, and their frequency positions are:

$$\omega_{V_2} = \frac{1}{r_2 C_2} \quad (3.5)$$

and

$$\omega_{V_{out}} = \frac{1}{r_3(C_L + C_3)} \approx \frac{1}{r_3 C_L}, \quad (3.6)$$

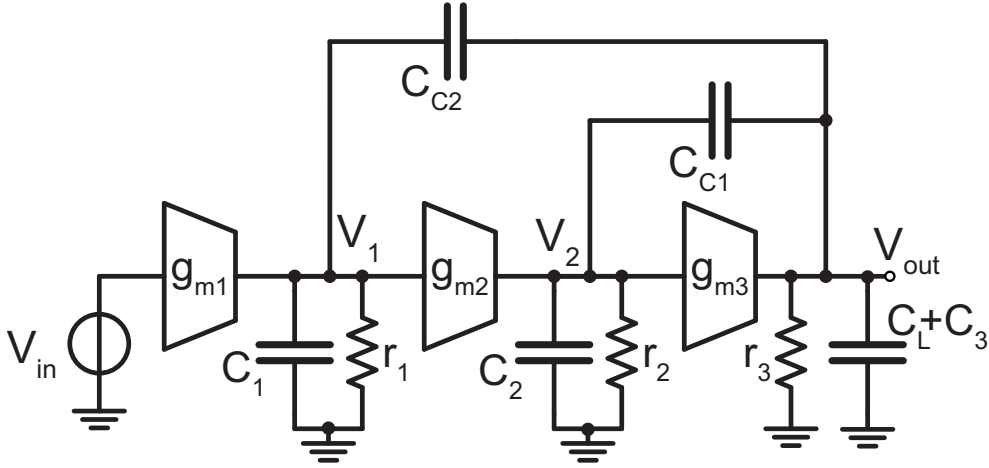


Figure 3.4: Small-signal circuit of the operational amplifier shown in Fig. 3.3 compensated with a nested Miller technique.

where the approximation holds under the hypothesis $C_L \gg C_3$, which is, typically, ensured. To guarantee circuit stability, the frequency distance of the two poles has to be sufficiently large to ensure an adequately large φ_m . This requirement is almost impossible to achieve without a compensation network in integrated CMOS operational amplifier with a reasonable DC gain. The consequences that C_{C1} produces on the circuit poles are mainly two. The first effect is to increase the equivalent capacitance on node V_2 by exploiting Miller effect. Capacitor C_{C1} , thus, results as an equivalent larger capacitance and, hence, the frequency of pole on V_2 is reduced and can be expressed as

$$\omega_{V_2} = \frac{1}{r_2(C_2 + C_{C1}(1 + A_3))} \approx \frac{1}{r_2 C_{C1} A_3} = \frac{1}{r_2 C_{C1} g_{m3} r_3}, \quad (3.7)$$

where $A_3 = g_{m3} r_3$ is the gain of the third stage. The unity gain frequency, ω_U , of the two-stage amplifier can now be calculated as the product between the DC gain of this circuit and the first pole frequency:

$$\omega_U = A_2 A_3 \omega_{V_2} = (g_{m2} r_2)(g_{m3} r_3) \frac{1}{r_2 C_{C1} g_{m3} r_3} = \frac{g_{m2}}{C_{C1}} \quad (3.8)$$

The second effect takes place at high frequency. Under this condition, C_{C1} reaches a so low impedance that it can be approximated as a short circuit and, thus, connects node V_{out} and V_2 . The output impedance, hence, becomes equal

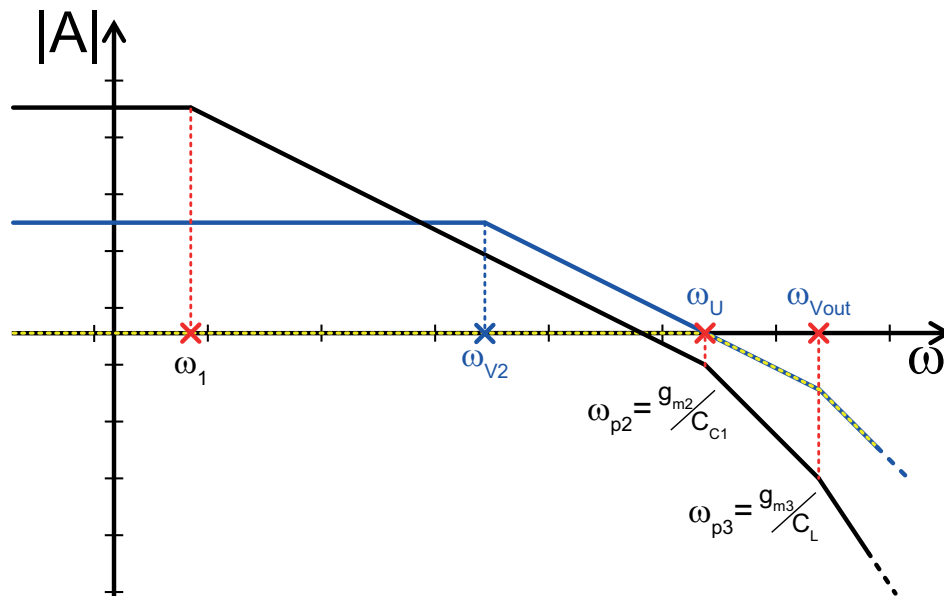


Figure 3.5: Transfer function of the operational amplifier shown in Fig. 3.4 (black line). Transfer function of an operational amplifier composed by the second stage, the third stage, and C_{C1} of the topology in Fig. 3.4 in open-loop configuration (blue line) and closed in unity-gain buffer configuration (yellow dashed line).

to $1/g_{m3}$ due to the presence of C_{C1} , and the expression of the pole associated to V_{out} becomes

$$\omega_{V_{out}} = \frac{g_{m3}}{C_L + C_3} \approx \frac{g_{m3}}{C_L}. \quad (3.9)$$

The transfer function of this partial circuit is depicted with a blue line in Fig. 3.5.

Taking now in consideration the whole circuit of Fig. 3.4, it is easier to analyze nested Miller compensation. Indeed, the outer compensation capacitor, C_{C2} , behaves similarly to C_{C1} : the dominant pole (ω_{p1}), which is associated to node V_1 , is moved to low frequencies due to the Miller effect generated by the second- and third-stage gains ($A_2 \cdot A_3$). Therefore, we have

$$\omega_{p1} = \frac{1}{r_1 [C_1 + C_{C2} (1 + A_2 A_3)]} \approx \frac{1}{r_1 C_{C2} A_2 A_3} = \frac{1}{r_1 r_2 r_3 g_{m2} g_{m3} C_{C2}}, \quad (3.10)$$

where this approximation holds under the additional hypotheses that $g_{m2} r_2 \gg 1$ and $g_{m3} r_3 \gg 1$. The unity-gain frequency of the three stage amplifier (ω_0) can be easily calculated as

$$\omega_0 = A_1 A_2 A_3 \omega_{p1} = (g_{m1} r_1)(g_{m2} r_2)(g_{m3} r_3) \frac{1}{r_1 r_2 r_3 g_{m2} g_{m3} C_{C2}} = \frac{g_{m1}}{C_{C2}}. \quad (3.11)$$

At sufficiently high frequency, C_{C2} can also be approximated as a short circuit, which connects the second-stage input node (i.e., V_1) with the third-stage output node (i.e., V_{out}). In this conditions, the obtained circuit corresponds to the first stage followed by the partial amplifier that was previously analyzed (i.e., the operational amplifier composed by C_{C1} and the second and the third stage), which is now closed in unity buffer configuration. It is possible, thus, to use the results obtained before, i.e. equations (3.8) (3.9), to understand the high-frequency behavior of the circuit. The closed-loop transfer function of the partial amplifier (second and third stage in Fig. 3.4) connected in unity-gain configuration is depicted with a yellow dashed curve in Fig. 3.5. Therefore, the second-pole frequency of the main three-stage operational amplifier, ω_{p2} , corresponds to the unity gain frequency of the two-stage amplifier:

$$\omega_{p2} = \omega_U = \frac{g_{m2}}{C_{C1}}. \quad (3.12)$$

Similarly, the third-pole frequency corresponds to the frequency associated to node V_{out} [equation (3.9)]:

$$\omega_{p3} = \omega_{V_{out}} \approx \frac{g_{m3}}{C_L}. \quad (3.13)$$

The second and the third pole can be considered independent from each other, and can therefore be represented with equations (3.12), and (3.13), only when $g_{m2} \ll g_{m3}$, otherwise the two poles interact with each other and become complex conjugate. The transfer function of the three-stage operational amplifier compensated with the nested Miller technique is depicted in Fig. 3.5 with a black line.

Analyzing equations (3.11), (3.12), and (3.13), it is straightforward to find the conditions that ensures circuit stability. To guarantee stability, indeed, the second pole has to be placed at a frequency higher than ω_0 :

$$\omega_{p2} = m \omega_0, \quad (3.14)$$

where m is a factor, greater than unity, that is chosen by the designer to achieve the desired phase margin, i.e. $\varphi_m = \arctan(m)$. A typically chosen value is $m = 2$, since it sets the phase margin close to 60° . In addition to this constraint, the second and the third poles have to be set sufficiently apart from each other:

$$\omega_{p3} = p \omega_{p2}, \quad (3.15)$$

where p is a parameter, greater than unity, that can be established by designers to set the amplifier damping ratio, ξ . When both m and n are set equal to 2 ($\xi = 0.5$), the circuit poles assume, in a closed loop configuration, a third-order Butterworth constellation.

The nested Miller compensation has the great advantage to be simple and widely used by analog designers, however it has several drawbacks that make it inadequate for critical applications. The bandwidth achieved through this compensation is, in fact, only a quarter (at best) with respect to the bandwidth obtained with a single stage amplifier implemented with a stage equivalent to the third stage. Moreover, the feed-forward path, which can be exploited by the input signal by passing through the compensation capacitors, generates two zeros: one located in the right and the other in the left half-plane. The right half-plane zero is typically close to the unity-gain frequency and, thus, can severely degrade φ_m , jeopardizing the operational amplifier stability. Therefore, a zero nulling resistor has to be inserted in series connection with the compensation capacitors to move the zero frequency far from ω_0 . Finally, the additional two major drawbacks of the nested Miller compensation are related to the placement of C_{C1} in parallel with the third stage (i.e. between the gate and the drain terminal of transistor $P3$ in Fig. 3.3). Indeed, when a large positive current is suddenly requested by the load, the gate terminal of transistor $P3$ is moved toward ground. As already mentioned, C_{C2} behaves as a low-impedance path to high-frequency signals and, thus, transfers the same

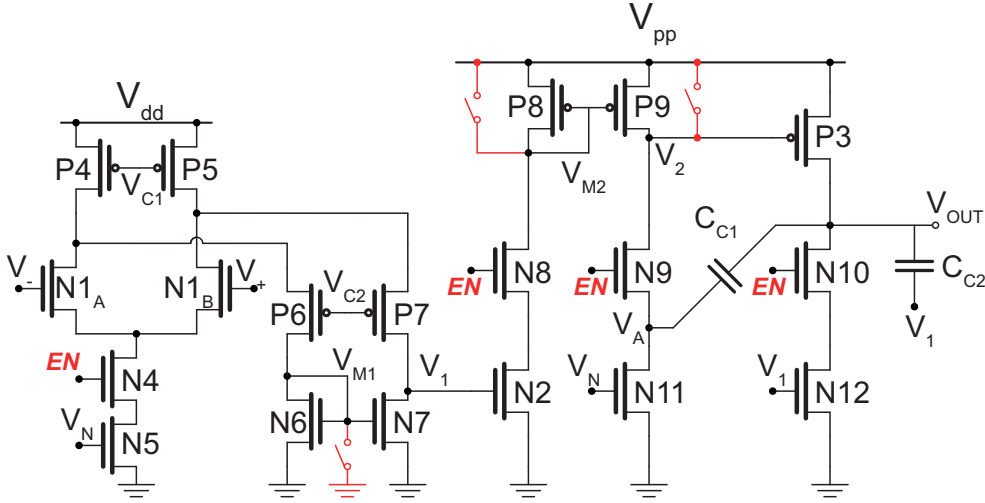


Figure 3.6: Circuit schematic of the three-stage V_T operational amplifier with the implemented compensation network.

voltage drop suffered by V_2 to the output node. The regulated voltage, hence, shows an undesired voltage drop in this operating condition. The second drawback associated to the presence of C_{C2} is related to the positive power supply rejection ratio (PSRR^+): in the presence of a noise disturbance superimposed on V_{pp} , the voltage on the gate terminal of transistor $P3$ is modulated with the same noise pattern in order to keep its V_{gs} constant and, thus, its current unchanged. As in the previous case, C_{C2} transfers the noise disturbance from V_2 to the output node. Therefore, the regulator with an operational amplifier with nested Miller compensation shows poor PSRR^+ .

To solve the limitations introduced by the nested Miller compensation, a cascode compensation technique [11], shown in Fig. 3.6, was implemented in the V_T operational amplifier. The compensation capacitor C_{C1} , indeed, is not directly connected to node V_2 , as in standard Nested Miller compensation, but to the source terminal of common-gate transistor $N9$. This transistor has its drain terminal connected to node V_2 and offers a low-impedance node (V_A) as an input for the compensation current, $I_{C_{C1}}$, generated by C_{C1} . If this input impedance of common-gate transistor $N9$ is sufficiently low, the current through C_{C1} can be expressed as

$$I_{C_{C1}} = V_{out} sC_{C1} \quad (3.16)$$

$I_{C_{C1}}$ is, then, injected into node V_2 similarly to the case of conventional nested

Miller compensation. The main difference can be appreciated by analyzing the voltage generated on V_2 at sufficiently high frequency (i.e., at frequency at which the impedance on node V_2 is dominated by the parasitic capacitance C_2):

$$V_2 = \frac{I_{C_{C1}}}{sC_2} = V_{out} \frac{C_{C1}}{C_2}, \quad (3.17)$$

which is higher than in the case of conventional nested Miller compensation (i.e., $V_2 = V_{out}$). Therefore, the circuit modification can be seen as a voltage gain (equal to C_{C1}/C_2) introduced from node V_{out} to node V_2 . From the compensation point of view, this high-frequency gain causes a lower impedance on the output node: transistor $P3$ shows a diode connection with a voltage gain (that can be referred to as a super-diode connection). Thus, the high-frequency impedance is equal to $(1/g_{m3})(C_{C1}/C_2)$ and, hence, the third pole frequency is moved to

$$\omega_{p3} = \frac{g_{m3} C_{C1}}{C_L C_2}, \quad (3.18)$$

whereas ω_{p1} and ω_{p2} have the same expressions as in conventional nested Miller compensation case.

It is possible to exploit the higher ω_{p3} to obtain a higher bandwidth with respect to the previous case with the same power consumption. Moreover, the cascode compensation technique allows solving the most critical drawbacks of conventional nested Miller compensation. Firstly, it cuts the feed-forward path of the signal and, therefore, the creation of right half-plane zeros is prevented without the need for any zero nulling resistors. Secondly, both the effects due to the connection of C_{C1} are intrinsically solved by the proposed topology. In particular, when a large current has to be fed to the output, any variation of V_2 is not immediately transferred on V_{out} . Therefore, the voltage drop on the regulated current is highly reduced, which is a key characteristic for our application. Furthermore, PSSR^+ is enhanced due to the removal of the high frequency path from V_2 to V_{out} , which is fundamental when the macrocell is employed in a system that generates V_{pp} with a charge pump, since it produces a significant amount of noise on the power supply.

3.1.2.2 Component Sizing

The most critical transistor to size is $P3$, since it drives the programming current. The length of its channel was set to the minimum allowed value, $L_{P3} = L_{min,HV} = 600$ nm. This choice is forced by area constraint: indeed, any increase above $L_{min,HV}$ has to be applied, with the same multiplying factor, to W_{P3} so as to maintain the W/L ratio unchanged. The channel width

of $P3$ was sized to be $W_{P3} = 375 \mu\text{m}$ so that the transistor is able to drive up to 20 mA assuming to apply the maximum V_{GS} achievable (i.e., gate terminal connected to ground) in the worst-case condition (i.e., slowest transistor corner, maximum operating temperature $T = +150 \text{ }^\circ\text{C}$, and minimum power supply $V_{pp} = 4.5 \text{ V}$). The quiescent current I_{P3} is the sum of the current set by transistor $N12$ and the current requested by the external bias network I_{R1} (see Subsection 3.1.1). Transistor $N12$ was sized to carry $50 \mu\text{A}$ and, thus, in static conditions $I_{P3} = 61 \div 92 \mu\text{A}$, which results in a adequately large value of the transconductance of transistor $P3$ (i.e., $g_{m3} = 1.1 \div 1.45 \text{ mA/V}$). The total current requested by the operational amplifier in static conditions is, therefore, $I_{op-amp} = 106 \div 137 \mu\text{A}$. The bias current sunk from V_{dd} is $I_{dd} = 30 \mu\text{A}$, whereas the current drawn from V_{pp} is $I_{pp} = 76 \div 107 \mu\text{A}$. Hence, the static power consumption of the operational amplifier, P_{op-amp} , at nominal values of the voltage supplies, is

$$P_{op-amp} = V_{dd}I_{dd} + V_{pp}I_{pp} = 434 \div 589 \mu\text{W}. \quad (3.19)$$

The total power consumption of the V_Y voltage regulator can be easily calculated as

$$P_{tot} = P_{op-amp} + P_{net} = 508 \div 719 \mu\text{W}, \quad (3.20)$$

where $P_{net} = 74 \div 130 \mu\text{W}$ is the static power consumption of the voltage regulator bias network.

The aspect ratio of transistor $N2$ was set to $3 \mu\text{m}/1 \mu\text{m}$, so that the transconductance of the second stage is $g_{m2} = 98 \mu\text{A/V}$, which is both sufficiently high and compliant with the hypothesis $g_{m2} \ll g_{m3}$ used during the pole analysis. The large size of transistor $P3$ makes this device the major contributor from the capacitance point of view at node V_2 . C_2 can be approximated as the gate capacitance of $P3$, thus, $C_2 \approx 350 \text{ fF}$. With these parameters, it was possible to size the compensation capacitance C_{C1} in order to separate ω_{p2} and ω_{p3} . Indeed, combining equations (3.12),(3.15), and (3.18) it is possible to obtain the expression of C_{C1} that guarantees stability:

$$C_{C1} = \sqrt{p \frac{g_{m2}}{g_{m3}} C_2 C_L} \quad (3.21)$$

Thus, the inner compensation capacitance was sized $C_{C1} = 2 \text{ pF}$, which sets the frequency of the second and the third pole to

$$\omega_{p2} = \frac{g_{m2}}{C_{C1}} \approx 49 \text{ Mrad/s} = 7.8 \text{ MHz} \quad (3.22)$$

and

$$\omega_{p3} = \frac{g_{m3}}{C_L} \frac{C_{C1}}{C_2} \approx 102.5 \text{ Mrad/s} = 16.3 \text{ MHz}, \quad (3.23)$$

Table 3.3: Summary of the V_Y operational amplifier transistors sizes.

| Stage | Transistor | W (μm) | L (μm) | Transistor | W (μm) | L (μm) |
|-----------------|------------|--------------------------|--------------------------|------------|--------------------------|--------------------------|
| 1 st | $N1_A$ | 6 | 1 | $N1_B$ | 6 | 1 |
| | $N4$ | 6 | 1 | $N5$ | 6 | 1 |
| | $N6$ | 3 | 1 | $N7$ | 3 | 1 |
| | $P4$ | 4 | 1 | $P5$ | 4 | 1 |
| | $P6$ | 4 | 0.2 | $P7$ | 4 | 0.2 |
| 2 nd | $N2$ | 3 | 1 | $N11$ | 3 | 1 |
| | $N8$ | 6 | 0.6 | $N9$ | 18 | 0.6 |
| | $P8$ | 6 | 1 | $P9$ | 6 | 1 |
| 3 rd | $N12$ | 20 | 1 | $N10$ | 20 | 0.6 |
| | $P3$ | 375 | 0.6 | | | |

respectively. The transistors that form the differential pair were equally sized to have an aspect ratio of $6 \mu\text{m}/1 \mu\text{m}$, which ensures the possibility to achieve good matching and a first stage transconductance $g_{m1} = 84 \mu\text{A}/\text{V}$. Finally, the value of the outer compensation capacitor was sized to set the unity-gain frequency m -times lower than ω_{p2} :

$$C_{C2} = m \frac{g_{m1}}{g_{m2}} C_{C1} = 4.3 \text{ pF}, \quad (3.24)$$

where m was set equal to ≈ 2.5 to be more conservative and have a larger phase margin with respect to the canonical choice $m = 2$. The unity gain frequency, therefore, turns out to be

$$\omega_0 = \frac{g_{m1}}{C_{C2}} \approx 19.6 \text{ Mrad/s} = 3.1 \text{ MHz} \quad (3.25)$$

Table 3.3 shows the aspect ratio of all the transistor in the V_Y operational amplifier.

3.1.3 Simulation Results

To validate the design and the effectiveness of the theoretical analysis, the voltage regulator was simulated in the *Cadence Virtuoso*[®] environment, which allows studying the circuitry behavior in all the requested conditions. The simulations, indeed, cover all variations due to process, voltages, and temperature (PVT), allowing designers to look for and optimize critical behaviors that can, otherwise, lead to circuit malfunctions.

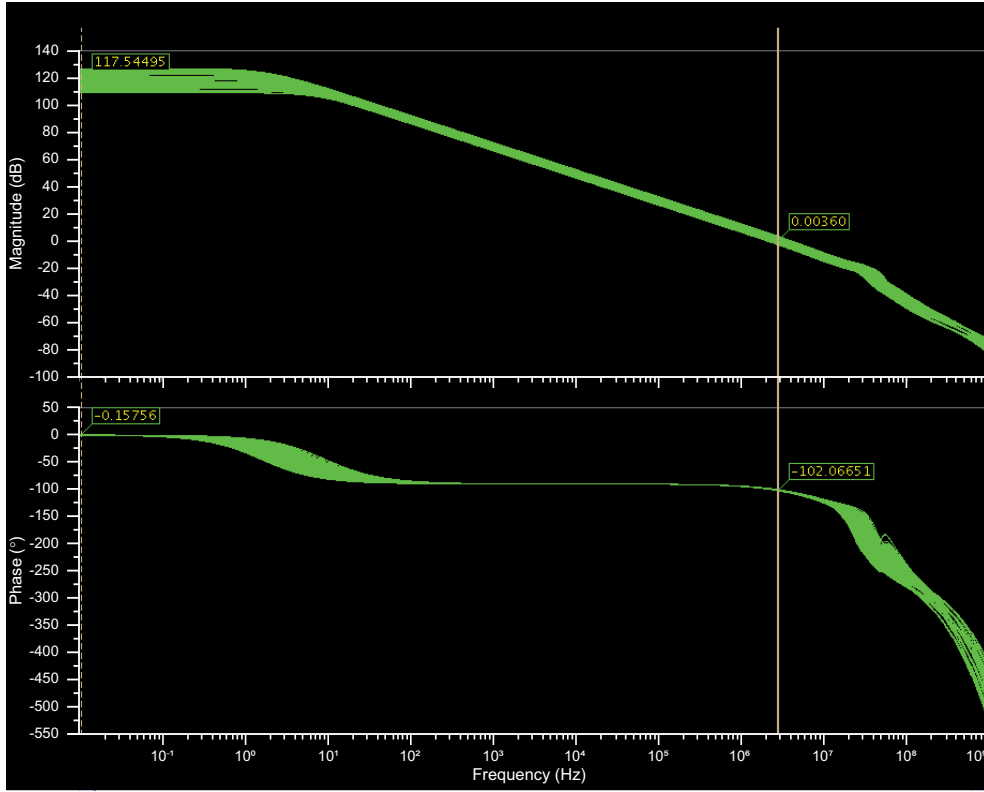


Figure 3.7: Simulated open-loop transfer functions of the implemented V_Y operational amplifier obtained under different conditions (i.e. with all the possible permutations of the parameters listed in in Tab. 3.4).

Particular attention was paid to stability analysis since it is one of the most critical points. An unstable regulator is indeed catastrophic because it inevitably leads to the failure of any program operation and, thus, the memory becomes unusable.

To ensure a closed-loop stability, AC simulations were run in several conditions: not only all the PVT specification range has been span (i.e. MOS transistor simulated with fast, typical, and slow models; $1.55 < V_{dd} < 1.95$ V and $4.5 < V_{pp} < 5.5$ V; $-40 < T < +150$ °C), but also all the V_Y values and I_{reg} currents were considered since they vary substantially with the pulse type and the applied parallelism.

Fig. 3.7 shows the Bode plots of a set of open-loop transfer functions of V_Y operational amplifier. Each curve represents the transfer function calculated

Table 3.4: Simulation parameters of Fig. 3.7.

| T | $-40\text{ }^\circ\text{C}$ | $27\text{ }^\circ\text{C}$ | $+150\text{ }^\circ\text{C}$ |
|-------------|-----------------------------|----------------------------|------------------------------|
| MOS models | fast-fast | typical-typical | slow-slow |
| V_{dd} | 1.55 V | – | 1.95 V |
| V_{pp} | 4.5 V | – | 5.5 V |
| I_{prog} | 450 μA | – | – |
| Parallelism | 1 | 2 | 4 |

with a specific set of parameters. In particular, the 108 curves depicted in Fig. 3.7 are obtained through all the possible permutation of the parameters shown in Tab. 3.4.

Fig. 3.8 depicts the open-loop transfer functions of V_Y operational amplifier, obtained with the same parameters used to obtain Fig. 3.7 with the only exception of the applied parallelism. Indeed, in this case, the simulation were run with parallelism equal to 8, 16, and 32.

It is straightforward to notice that the phase margin is sufficiently high in all cases, the unity gain frequency is well above 1 MHz, and the DC gain is sufficiently high, since is always higher than 90 dB. It is also interesting to notice that for higher parallelism and, thus, for higher current load, the DC gain of the operational amplifier decreases, whereas the first pole frequency increases, as expected due to the smaller load resistance.

Furthermore, several transient simulation were run in different conditions, similarly to the case of AC simulations. In particular, Fig. 3.9 shows the evolution during time of V_Y and the programming current during a RESET operation carried out on 33 PCM cells written in parallel. The blue curves correspond to the proposed voltage regulator, whereas the yellow curves were obtained with the same voltage regulator with the exception of its operational amplifier which, in this case, is compensated with conventional nested Miller technique. It is important to notice that the voltage drop (ΔV), when the current load is requested, is significantly smaller in the proposed solution ($\Delta V \approx 100\text{ mV}$) with respect to the case of conventional nested-Miller compensated operational amplifier ($\Delta V \approx 600\text{ mV}$). Moreover, the designed solution shows a better rectangular-shaped pulses (the trailing edge is sharper and the overshoot is smaller).

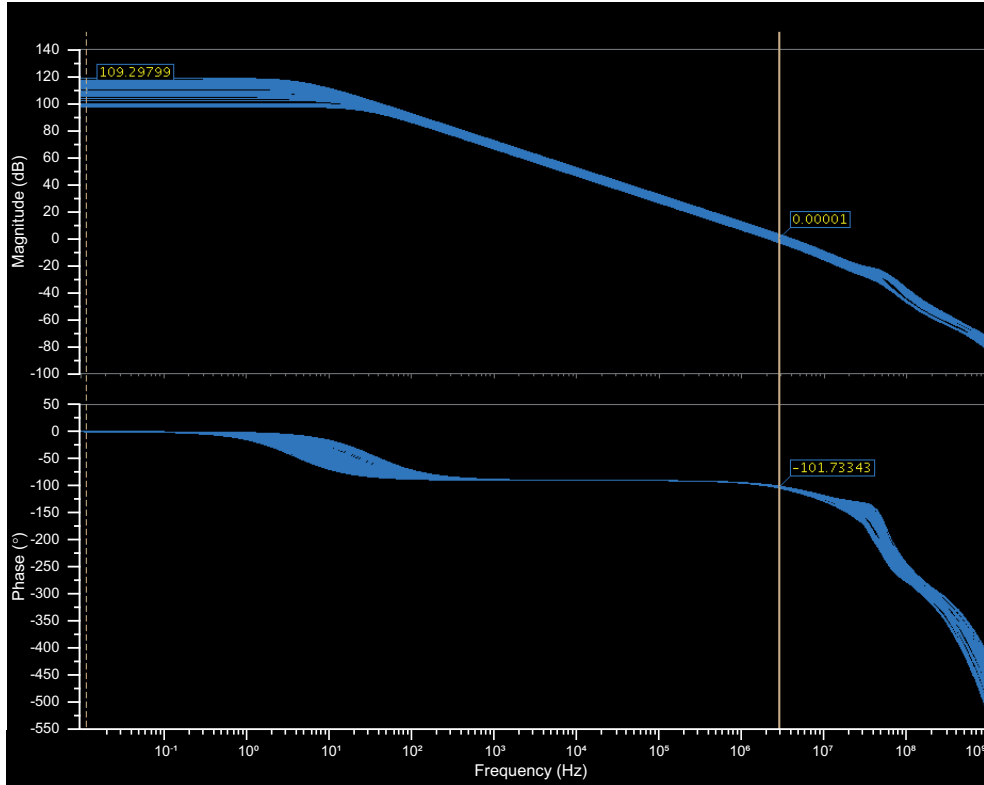


Figure 3.8: Simulated open-loop transfer functions of the implemented V_Y operational amplifier obtained with all the possible permutations of the parameters listed in in Tab. 3.4 with the exception of the parallelism, which in this case was set equal to 8, 16, and 32.

3.2 Improved Current Mirror for PCM Programming

The V_Y regulated voltage is then utilized to power all the programming circuitry, as mentioned in Chapter 2. From the design point of view, the most critical block of the programming circuitry is the current mirror that has to generate up to 33 programming current pulses in parallel. In particular, in the case of RESET pulses, the current has to be shaped with steep edges and brief duration (as short as 100 ns). These specifications are very hard to meet with the standard circuit shown in the previous Chapter.

To better understand the performance limitations, a standard current mir-

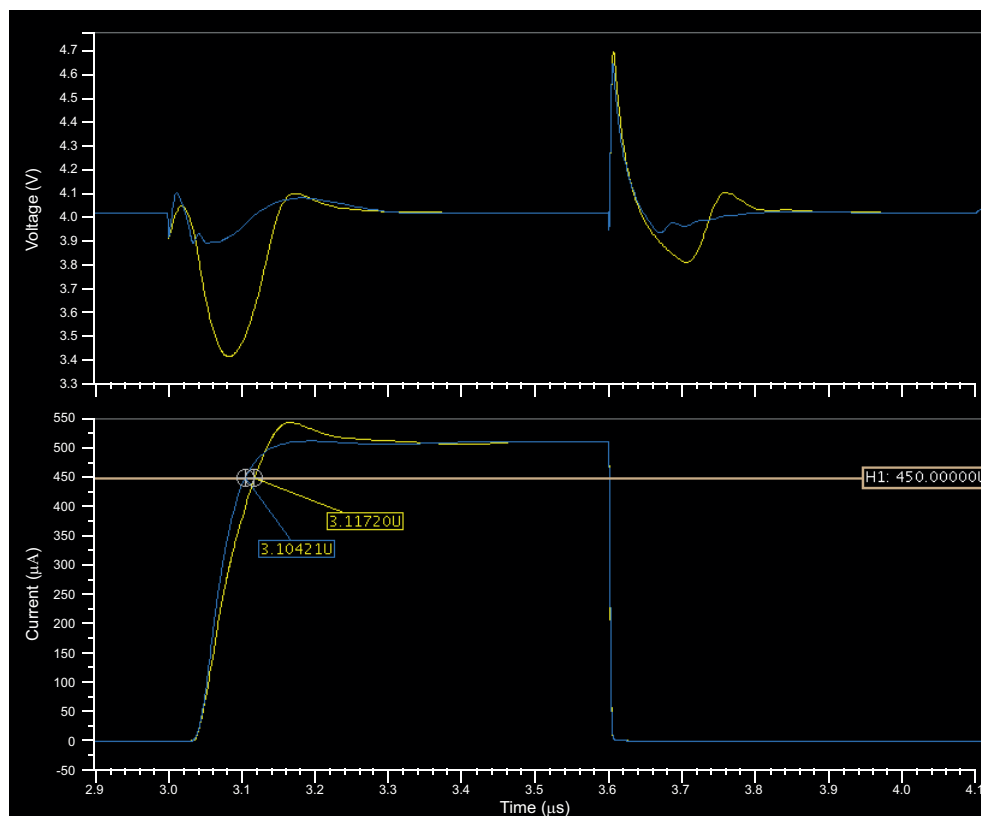


Figure 3.9: Simulated V_Y (top) and I_{prog} (bottom) during a RESET pulse in the case of the proposed voltage regulator (blue curves), which includes the cascode-compensated operational amplifier, and in the case of the same regulator with a conventional nested-Miller compensated operational amplifier (yellow curves). In both cases, the simulation parameters were: $V_{dd} = 1.8$ V, $V_{pp} = 5$ V, $T = 27$ °C, MOS transistors model = typical-typical, parallelism = 33, $T_{pulse} = 600$ ns, DAC_conf = 001101 (i.e., $I_{plateau} = 500$ μ A), and REG_conf = 01110 (i.e., $V_Y = 4$ V).

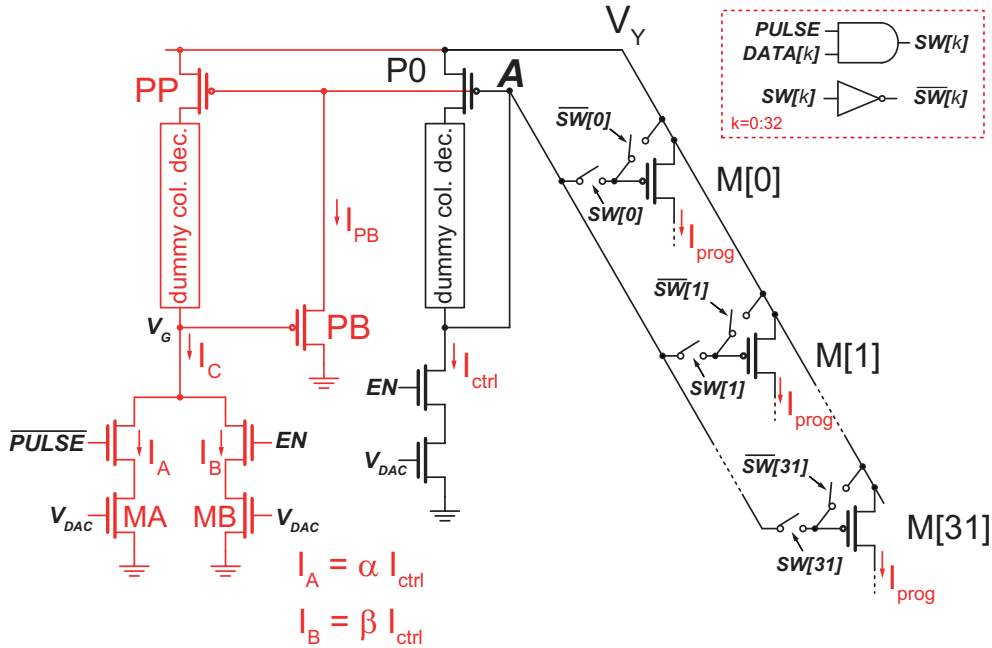


Figure 3.10: Circuit schematic of the improved recovery circuit (red) applied to a standard current mirror (black).

ror is depicted in black in Fig. 3.10. It is important to point out that this circuit, as well as the rest of the programming circuitry, can be turned off by driving the enable signal, EN, low to guarantee an almost-zero power consumption of these blocks when the memory is not performing a write operation. At the beginning of the programming phase, indeed, the finite state machine raises signal EN before any other programming signal, in order to turn the write circuitry on. Then, the FSM runs other internal operations and, concurrently, gives sufficient time to the circuit nodes and currents in the circuit in Fig. 3.10 to settle. During this idle phase, as explained in Chapter 2, I_{ctrl} is generated (as a scaled copy of I_{DAC}) and biases the program-control branch of the current mirror. In particular, node A settles to the voltage that allows transistor $P0$ to carry I_{ctrl} . It has to be noticed that the gate terminal of each transistor $M[k]$ (with $k = 0$ to 32) is controlled by a pair of switches, which are driven by complementary phases ($SW[k]$ and $\overline{SW}[k]$), and can be connected to either V_Y or node A. During this idle phase, all gate terminals are connected to V_Y , because $\overline{SW}[k]$ is set high by signal PULSE (see the inset in Fig. 3.10),

which is directly driven by the FSM to a ‘0’ logic state.

When a programming pulse has to be delivered, the finite state machine sets $PULSE = '1'$ and, concurrently, turns the 33 bit of signal $DATA$ into the desired configuration. $DATA$ is a digital signal that contains a number of ‘1’ equal to the number of cells that have to receive the following programming current pulse. Each bit of $DATA$ is, indeed, used to discriminate how many and which cells have to take the following programming pulse. The configuration of signal $DATA$ is a function of three independent aspects:

- the **chosen maximum parallelism**: indeed the number of ones in signal $DATA$ has to be less or equal to the maximum allowed parallelism;
- the **information that has to be written**, i.e. the k -th bit of $DATA$ is equal to ‘1’ only if the pulse type (SET or RESET) and the k -th bit of the information that has to be written match (e.g., SET and ‘1’ is considered a match for a direct cell); and
- the **current state of the cell** that is selected for the program: indeed, if the cell content already matches the data that has to be programmed (i.e. the verify operation has succeeded) the write operation is concluded for that specific cell.

Both signals SW and \overline{SW} consist of 33 bits that are generated by 33 AND gates and 33 inverters, as shown in Fig. 3.10. The k -th AND gate receives $DATA[k]$ and $PULSE$ as an input, and provides $SW[k]$ to the output, which is, then, inverted to obtain $\overline{SW}[k]$.

It is worth to point out that, due to the mentioned aspects, it is not possible to know, a priori, the number of cells that are going to be programmed in each cycle. Therefore, when signal $PULSE$ is driven high by the finite state machine, the gate terminal of a certain number of transistors $M[k]$ is disconnected from V_Y and, immediately, connected to node **A**. This switch of connections causes an increment in the voltage at node **A**: indeed, a capacitive load is connected to node **A** and injects its charges into this node. The added capacitive load is large with respect to the capacitance at node **A**: indeed, the program-load transistors $M[k]$ are ten times bigger than transistor $P0$. Therefore, the added capacitance can easily raise the voltage of node **A** toward V_Y . This voltage increase reduces the current through $P0$, which can be brought easily into cut-off. The only means to discharge node **A** is current I_{ctrl} , and only when this current has sufficiently discharged the capacitive load, the programming current starts to flow into the selected cells. This operation causes a delay in the generation of the programming pulse that increases with increasing number

of program-load transistors activated in the write procedure and, thus, it is worst in the high parallelism case. This behavior of the circuit can make the programming pulses ineffective, since the delay introduced can severely reduce the pulse time length and excessively smooth the rising edge especially in the case of narrow RESET pulses. Moreover, it is important to remark that the delay can not be compensated for by simply increasing the pulse length, since this effect is data dependent.

The circuit shown in red in Fig. 3.10 has been designed to be applied to the standard current mirror to overcome this limitation. The idea is to rapidly discharge the added capacitance on node A. The problem is that, in the standard case, the maximum current available to discharge the node is set by the value of I_{ctrl} , which is dependent from the type of applied pulse (i.e. SET₁, RESET₁, SET₂, etc.). Moreover, the effective portion of I_{ctrl} that is involved in the discharge of node A is actually reduced when transistors $P0$ starts to conduct. Therefore, a PMOS transistor, PB , with its source terminal connected to node A was added to the circuit, in order to be activated by the large overdrive voltage determined by the voltage increase at node A. In this way, transistor PB guarantees an additional current, which is generated by the voltage variation of node A, to be sunk to discharge the added capacitive load. To bias this high-voltage PMOS transistor, an additional branch was placed in parallel to the program-control branch and used to drive the gate terminal of PB . The added branch is similar to the program-control branch, however, it is biased with a current, I_C , which is generated by two NMOS transistors, MA and MB . To properly size I_C , let us assume to choose $I_C = I_{ctrl}$: the voltage drop across the source and gate terminal of PB results ideally zero and, thus, PB is not able to react until the voltage increment on node A is larger than $V_{th,HV}$. However, if I_C is set larger than I_{ctrl} [i.e. $I_C = (\alpha + \beta)I_{ctrl}$, where α and β are two positive arbitrary parameters that satisfy conditions $\alpha + \beta > 1$ and $\beta < 1$], then node A is forced to settle on a lower voltage in order to accommodate the higher current and, thus, $P0$ has also to carry a current equal to I_C . The exceeding part of the current, i.e. $I_{PB} = I_C - I_{ctrl} = (\alpha + \beta - 1)I_{ctrl}$, flows through $P0$ to bias PB . In this case, PB is on when PULSE is activated (i.e., its V_{SG} is already larger than zero, since it has to accommodate the bias current) and, therefore, can immediately react to a voltage increment at node A.

Using this approach (i.e., $I_C > I_{ctrl}$), the programming current generated by transistors $M[k]$ turns out to be a multiplied copy of I_C , which is not the desired value. To overcome this unwanted effect, a portion of I_C (i.e., $I_A = \alpha I_{ctrl}$) is shut off, concurrently with the rising edge of signal PULSE, by

turning off transistor MA . Therefore, after the transient is settled, I_C results equal to $I_B = \beta I_{ctrl}$, the current through $P0$ is I_{ctrl} , and, thus, transistor PB is automatically turned off. To achieve this purpose, transistor MA generates a current equal to $I_A = \alpha I_{ctrl}$, whereas transistor MB sets a current equal to $I_B = \beta I_{ctrl}$.

On top of each of two NMOS transistors there is a cascode NMOS transistor. The cascode transistor in series with MA is driven by \overline{PULSE} in order to turn off the entire bias branch when the programming pulse has to be generated (i.e., when $PULSE='1'$). By choosing the values of α and β (i.e., by sizing transistors MA and MB) the designer can finely tune the transient behavior of node A: the larger the sum of the two parameters, the faster the reaction of PB , and the higher the value of β , the longer PB is active. In the implemented chip, α and β were chosen equal to $2/3$ and $1/2$, respectively, since this was found to be the best trade-off in our application between the delay in the generation of I_{prog} and the current overshoot.

3.2.1 Simulation Results

The presented circuit was simulated in *Cadence Virtuoso*[®] to validate its correct behavior and estimate the improvement with respect to a standard solution.

Figure 3.11 shows the most significant signals, and their evolution over time, involved in a program operation that has to deliver a $RESET_1$ pulse to 33 cells in parallel. In particular, the figure shows the time window in which the signal $PULSE$ is activated and the current pulse is driven into the cells. To better appreciate the obtained improvement, our solution is compared to a standard current mirror: in all the plots of Fig. 3.11, the black curves represent signals that correspond to the proposed circuit, whereas the blue dashed lines shows the signals obtained with the standard solution..

As can be seen in Fig. 3.11(a), t_1 indicates the instant when the rising edge of signal $PULSE$ takes place. In this simulation, it is ensured that signal EN is active since t_0 , where t_0 is chosen in order to guarantee that the interval $t_0 - t_1$ is sufficiently long to let the circuit nodes to settle to their quiescent values. Moreover, I_{ctrl} is set by the FSM through the DAC circuit to be equal to $45 \mu A$, since the $RESET_1$ specifications requires a programming current $I_{prog} = 450 \mu A$, and V_Y is set to 4.6 V. The current flowing into transistor PB is depicted in Fig. 3.11(b) and starts from its designed initial value $I_{PB} = I_{ctrl}/6 \approx 7.5 \mu A$.

When $t = t_1$, all the 33 $\overline{SW}[k]$ switches are opened and all the 33 $SW[k]$ switches are connected to node A. It is clearly visible, in Fig. 3.11(d), that

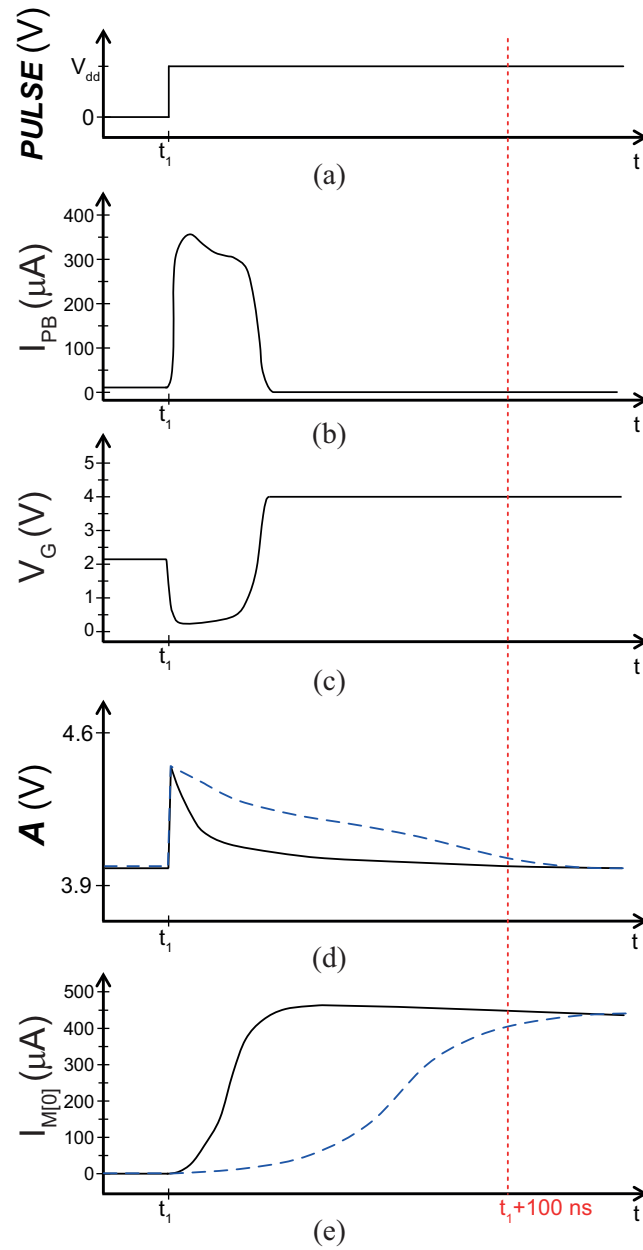


Figure 3.11: Evolution over time of voltages and currents during the initial part of a RESET₁ pulse (parallelism = 33) in the case of a standard current mirror (blue dashed curves) and the proposed circuit (black solid curves).

the voltage at node A, when $t = t_1$, is immediately raised in both cases, thus preventing transistors $M[k]$ ($k = 0 \div 32$) to turn on and generate I_{prog} . After t_1 the difference between the behavior of the two circuits is substantial. On the one hand, in the standard current mirror case, node A is slowly discharged by a limited maximum current equal to $I_{ctrl} = 45 \mu\text{A}$. Therefore, the programming current reaches 80% of its final value (i.e., $360 \mu\text{A}$) in approximately 114 ns, as shown in Fig. 3.11(e), which shows the program current through program-load transistor $M[0]$. On the other hand, in the proposed solution, node A is discharged much faster due to the large current (I_{PB}), shown in Fig. 3.11(b), provided by transistor PB . Indeed, the voltage at the gate terminal of transistor PB , shown in Fig. 3.11(c), is driven toward ground by I_C , which can not flow through transistor PP , since the voltage on node A forces it into cut-off. The source terminal of transistor PB is connected to node A, which, as mentioned, is increasing. Therefore, the source-to-gate voltage of transistor PB is largely increased in a very short time and causes I_{PB} to reach a peak value of $\approx 365 \mu\text{A}$. It is important to notice that I_{PB} is turned off by the circuit before I_{prog} reaches its plateau value, thanks to the chosen values of $\alpha = 2/3$ and $\beta = 1/2$. The improved current mirror is able to reach 80% of the plateau current in ≈ 23 ns, and is therefore almost 5 times faster than the standard current mirror, as it is possible to appreciate by looking at Fig. 3.11(e).

It is also fundamental to point out that the minimum duration of a RESET pulse is 100 ns (indicated in the Fig. 3.11 with a red dotted line), which makes the standard solution unable to program the cells with high parallelism.

The simulation results show a substantial and finely tunable improvement in the recovery time of the current mirrors, thus a US patent application [12] was filed to protect this solutions.

3.3 Charge Pump Design

As already mentioned, the macrocell implemented in Spider-Mem requires two power supplies: a low-voltage V_{dd} and a high-voltage V_{pp} . To make the PCM macrocell usable by a larger variety of applications, it has been decided to implement, in the future version of the chip, the internal generation of supply voltage V_{pp} . In this way, a single external low-power supply is necessary for the correct behavior of the circuit and, thus, all the applications that do not provide a double supply are enabled. To do so, a charge pump block will be included in the next version of Spider-Mem. The objective of this section is to present a novel charge-pump structure aimed at improving the charge transfer efficiency and provide a theoretical analysis as well as an exemplifying design

to appreciate the offered advantages.

A charge pump (CP) is an switched-capacitor circuit that provides at its output a voltage, V_{out} , that is, ideally, multiple of an input voltage, V_{in} . Typically, V_{out} is larger, as an absolute value, than V_{in} , while its sign can be either positive or negative, depending on the CP topology. Charge pumps are, therefore, a particular type of the DC-DC converter, and are very useful to generate, inside a chip, a different voltage starting from the given power supply, V_{dd} . CPs are highly suitable to be implemented in integrated circuits thanks to the availability of integrated capacitors with adequate characteristics. On the contrary, other types of DC-DC converters exploit magnetic field, instead of charges, to store and release energy and use, to perform these tasks, conductive coils that, even though they are widely used as discrete components, are very difficult to implement in ICs. Indeed, inductive DC-DC converters require long metallic turns stacked on top of each other, which can not be realized effectively in planar technologies. Integrated inductors and transformers have to have poor performance (i.e., low-quality factor, high series resistance, and high silicon area occupation).

A charge pump provides, in general, a voltage elevation (or gain, A), by delivering charges from the input terminal (typically connected to V_{dd}) to the output terminal through a cascade of N stages. Each stage is composed by a pump capacitor, C_p , and one (or more) transfer switches that are driven by adequate clock phases: by properly timing the control signals, it is possible to store a given amount of charge in C_p in one phase and, in the next phase, deliver this charge amount to the following stage or, in the case of the last stage, to the output.

To better understand the charge-pump working principle, it is useful to refer to the simplified schematic shown in Fig. 3.12(a). In the first phase, Φ_1 , the top plate of the pump capacitor is connected to the input-stage voltage, while the bottom plate is connected to ground, as depicted in Fig. 3.12(c). At the end of Φ_1 , the voltage across pump capacitor is equal to V_{in} , meaning that a charge packet $Q = C_p V_{in}$ was stored in this capacitor. Since, typically, the input voltage is connected to the power supply, Q is equal to $C_p V_{dd}$. During the second phase, Φ_2 , the bottom plate of C_p is connected to V_{dd} , while its top plate is connected to the output voltage, V_{out} . This operation, illustrated in Fig. 3.12(d), is usually defined capacitor boosting. The output capacitor C_{out} (assumed initially discharged) receives a charge packet that can be calculated

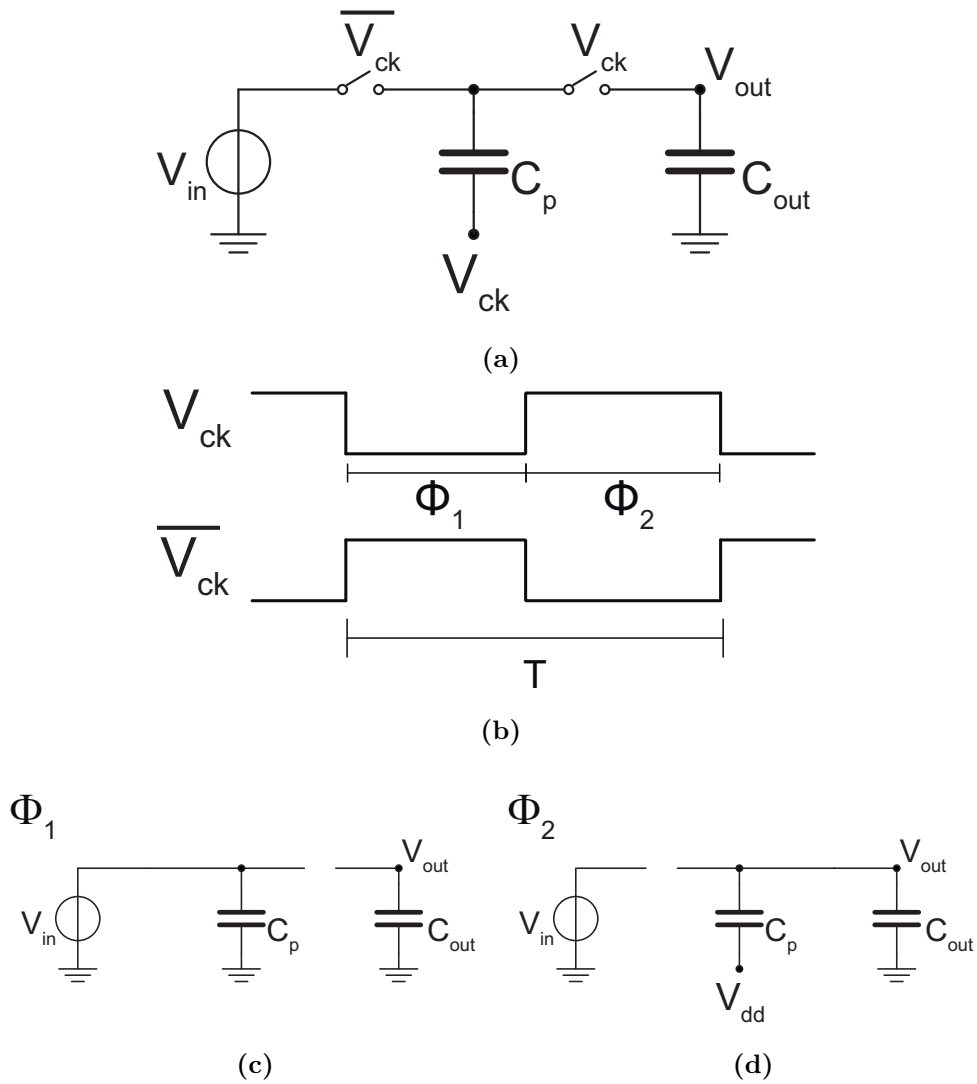


Figure 3.12: Ideal charge pump schematic (a) and corresponding phase signals (b). Equivalent charge pump circuit during Φ_1 (c) and Φ_2 (d).

by applying the charge conservation principle:

$$\begin{aligned} C_p V_{dd} &= C_{out} V_x + C_p (V_x - V_{dd}) \\ Q_{out} = C_{out} V_x &= 2V_{dd} \frac{C_p C_{out}}{C_p + C_{out}} \end{aligned} \quad (3.26)$$

where V_x is the voltage across C_{out} at the end of Φ_2 . As shown in equation (3.26), the amount of charges transferred in a cycle depends on the size of both the pump and the output capacitor. In particular, the larger the pump capacitor, the more charges are transferred in the same phase. The switching frequency, f , is also an important parameter since it determines the charge transfer rate. Obviously, repeating phase one and two with a higher frequency allows to proportionally increase the number of charges delivered to the output in a given time. Finally, capacitor boosting is a key aspect in the charge transferring procedure because it enables to store a voltage higher than the power supply in C_{out} . By repeating Φ_1 and Φ_2 multiple times, the voltage across the output capacitor rises until it asymptotically reaches $2V_{dd}$. The reason for this value can be also understood using an intuitive point of view: since, during Φ_1 , C_p is always recharged to V_{dd} , then, once it is boosted (i.e. when the bottom plate of C_p is connected to the power supply), the upper plate instantly reaches $2V_{dd}$, which is, thus, the asymptotic value calculated as the sum of the stored and boost voltages. On the contrary, without the aid of the capacitor boosting technique, the maximum voltage available at the output is equal to V_{dd} , resulting in $A = 1$ and, hence, jeopardizing the usefulness of the circuit.

The above considerations hold under the hypothesis that no load current is drawn from the output node. When a constant load current, I_{load} , is considered, the expression of the output voltage becomes [13]

$$V_{out} = 2V_{in} - \frac{I_{load}}{fC_p}. \quad (3.27)$$

N stages can be cascaded in order to obtain a charge pump that provides a output voltage gain equal to

$$A = \left. \left| \frac{V_{out}}{V_{in}} \right| \right|_{I_{load}=0} = N + 1, \quad (3.28)$$

and an output voltage gain is given by

$$V_{out} = (N + 1)V_{in} - \frac{NI_{load}}{fC_p}. \quad (3.29)$$

The derivative of the output voltage with respect to the load current is defined as the charge-pump output resistance, R_{out} , which can be easily calculated starting from equation (3.29):

$$R_{out} = \frac{N}{fC_p}. \quad (3.30)$$

It is worth to point out that R_{out} can be reduced by increasing the switching frequency: indeed, the number of charge transfers per time unit is increased and so does the number of charge packets delivered to the output. Equation (3.30) is obtained assuming ideal switches (i.e., switches with zero ON-resistance, r_{on}), whereas to carry out the analysis on a broader and more general case is useful to include the contribution of a non-zero r_{on} (as occurs in the case of switches implemented with MOS transistors). The main effect caused by the presence of this resistance is the change in the time constant, τ , associated with the charge transfer in one half-period, $T_{hp} = 1/2f$, that now is

$$\tau = r_{on}C_p. \quad (3.31)$$

The charge transfer is not accomplished instantaneously, but follows an exponential behavior $\propto (1 - e^{-t/\tau})$. To deliver all the charges in one half-period, it is necessary an infinite time, however, from an engineering point of view, the charge transfer is considered complete when $T_{hp} > 4 \div 5\tau$, which corresponds to $\sim 98\%$ and $\sim 99\%$ of the total charge transfer, respectively. When the switching frequency is excessively high, this condition is not verified and, thus, the charge transfer can not be completed in a half-period due to the insufficient time interval dedicated to charge transfer. Therefore, the ON-resistance of MOS switches, r_{on} , has to be taken in account in order to calculate the CP output resistance. r_{on} is in practice equal to the drain-source resistance of a MOS transistor that operates in the triode region and can therefore be represented as

$$r_{on} = \frac{1}{\mu C_{ox}(W/L)V_{ov}}, \quad (3.32)$$

where μ is the electrical mobility of the charge carriers in the channel, C_{ox} is the gate-oxide capacitance per unit area, W and L are the width and length, respectively, of the switch transistor, and V_{ov} is the overdrive voltage. Finally, the impact of the non-zero r_{on} on R_{out} can be expressed as [14]

$$R_{out} = \frac{N}{fC_p} \coth\left(\frac{T_{hp}}{r_{out}C_p}\right). \quad (3.33)$$

It is worth to point out that a complete charge transfer during a half-period implies that $r_{out}C_p$ is much smaller than T_{hp} and, thus, $\coth\left(\frac{T_{hp}}{r_{out}C_p}\right) \rightarrow 1$. Therefore, in this case, equation (3.33) coincides with equation (3.30).

Designers, typically, choose a number of stages, N , according to the desired output voltage. Then, they size C_p and f to sustain a given maximum load current (or, equivalently, to obtain the desired output resistance), design MOS switches to achieve a sufficiently small τ , and, finally, determinate C_{out} depending on the maximum output ripple allowed by specifications.

Another key charge-pump parameter is power efficiency, η , defined as the ratio of output power, P_{out} to input power, P_{in} . η is a fundamental element, especially in battery-supplied systems, since it expresses the amount of power delivered to the output with respect to the power absorbed by the input and, indirectly, the power wasted inside the circuit. To emphasize this fact, it is useful to indicate P_{in} as the sum of P_{out} and P_l , where P_l represents the system power losses. The two main contributors to system power losses are the parasitic capacitances, C_{par} , associated to the CP internal nodes, and the ohmic losses, P_r , due to the aforementioned non-zero CP output resistance. The former are usually defined dynamic losses, P_d , and can be measured by analyzing the input power of a charge pump when the load is disconnected ($I_{load} = 0$), which corresponds to the power needed to charge and discharge the parasitic capacitances. Therefore, P_d can be expressed as

$$P_d \propto fC_{par}V_{dd}^2, \quad (3.34)$$

where the proportionality factor is determined by several parameters such as the charge pump architecture and the number of stages.

In a design that aims at achieving a high transfer efficiency, it is fundamental to minimize both P_r and P_d . A good strategy to reduce dynamic losses is to have the least parasitic capacitance on the internal nodes. The charge pump parasitic capacitance is, typically, dependent on the capacitance contribution of the switches, $C_{par,sw}$, and the pump capacitor plates $C_{par,C}$. $C_{par,C}$ is mainly due to the capacitance coupling of the plates to the silicon substrate and, thus, this parameter is not under the control of designers since it is technology related. Usually, $C_{par,C}$ is expressed as the sum of the top- and bottom-plate parasitic capacitances, C_{top} and C_{bot} , respectively. When designing a charge pump for optimal efficiency, it is important to size the switches sufficiently small to ensure that their capacitance contribution is negligible with respect to the parasitic plate capacitance (i.e., $C_{par,sw} \ll C_{par,C}$). Once the parasitic capacitance are minimized, the other parameter that can reduce P_d is the switching frequency as is obvious from equation (3.34). However, a

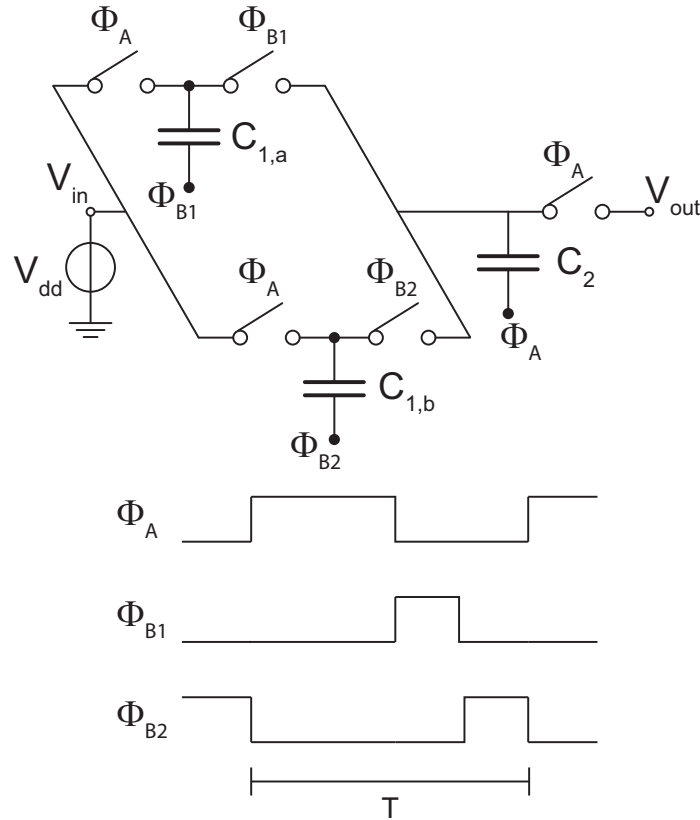


Figure 3.13: Optimized charge transfer scheme applied to a two-stage charge pump with ideal switches.

reduction of f adversely affects R_{out} and, thus, increases ohmic losses.

The main idea behind the proposed CP architecture is to exploit an optimized charge-transfer technique to decrease the charge-pump output resistance without incrementing dynamic power losses. The improved charge transfer technique consists in splitting the pump capacitor of each stage in multiple capacitors in order to allow several charge transfers to take place in the same half-period. The charge pump total capacitance is kept equal with respect to a conventional CP and, thus, the silicon occupation of the two solutions is the substantially same.

The operation principle of the proposed scheme will be now explained referring to Fig. 3.13, where $C_{1,a} = C/2$, $C_{1,b} = C/2$, and $C_2 = C$. When

Φ_A is high, $C_{1,a}$ and $C_{1,b}$ are connected in parallel between the power supply voltage V_{dd} and ground, so that the total amount of charges stored in these capacitors is

$$Q_I = (C_{1,a} + C_{1,b})V_{dd} = CV_{dd}. \quad (3.35)$$

During the next phase (i.e., when Φ_{B1} is high), $C_{1,a}$ is boosted and connected to C_2 to transfer charges. Under the assumption that, initially, C_2 is completely discharged, the voltage across C_2 becomes

$$V_1 = V_{dd} \frac{C_{1,a}}{C_{1,a} + C_2} = \frac{V_{dd}}{3}. \quad (3.36)$$

In an equivalent manner, when Φ_{B2} is high, $C_{1,b}$ is boosted to transfer charges to C_2 . The resulting voltage across capacitor C_2 is

$$V_2 = \frac{C_{1,b}V_{dd} + C_2V_1}{C_{1,b} + C_2} = V_{dd} \frac{C_{1,b} + C_2/3}{C_{1,b} + C_2} = \frac{5}{9}V_{dd}. \quad (3.37)$$

Therefore, the amount of charges stored in capacitor C_2 is $\frac{5}{9}CV_{dd}$, which is equal to $\sim 55.6\%$ of Q_I , whereas, in a conventional charge transfer that employs two equally-sized capacitor, the total amount of charges in the second capacitor is $\frac{1}{2}CV_{dd}$ which corresponds to 50% of the initial charge Q_I .

It is possible to generalize the effect of this technique by splitting the pump capacitor in k parts, and then complete k sequential charge transfers. The charge transfer improvement obtained using this transfer scheme with respect to a conventional transfer scheme is [15]

$$2 \sum_{m=1}^k \frac{k^{m-1}}{(1+k)^m}, \quad (3.38)$$

which it is an increasing function of k . Thus, the higher the number of capacitors in which C_p is split, the better the charge-transfer efficiency improvement with respect to the conventional case of equally-sized C_p . It has been demonstrated [15] that the same benefit is obtained when the charge transfer takes place in the opposite direction, i.e. when sequentially transferring charges from a regularly sized pump capacitor to k fractionated pumping capacitors.

3.3.1 Proposed Charge Pump Architecture

The proposed CP architecture, shown in Fig. 3.14, takes advantage of the improved charge transfer technique introduced in previous Section, which is

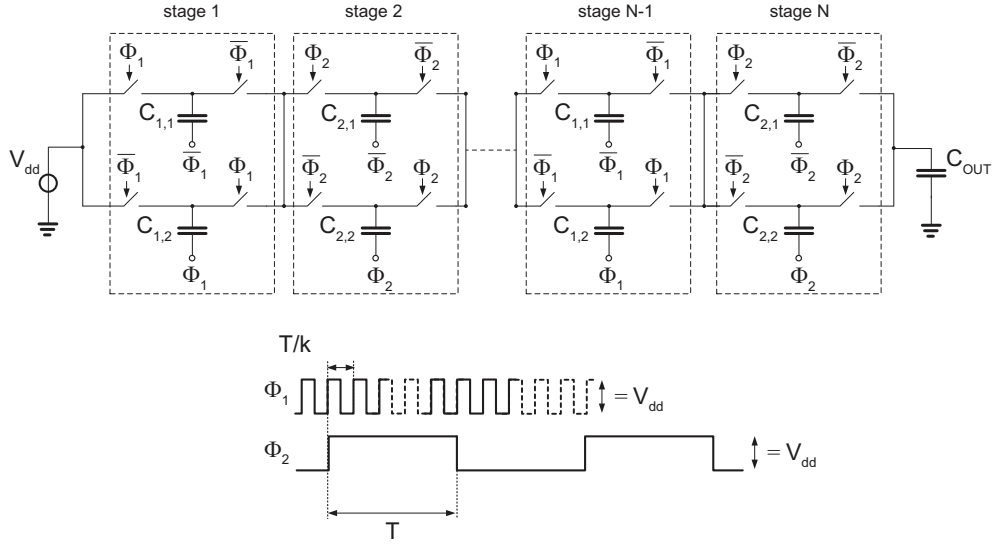


Figure 3.14: Ideal schematic of the proposed charge architecture implemented with N cross-coupled cascaded stages.

exploited to obtain a lower output resistance and, thus, a higher driving capability when compared with a conventional charge pump architecture. The fundamental idea is to achieve k charge-transfers in one period T , by driving a single small pump capacitor ($C_p = C/k$) with a switching frequency equal to kf to deliver charge packets to a large pump capacitor ($C_p = C$). In this way, instead of using k small capacitors, which are involved in the charge transfer only for a small portion of the period (i.e., T_{hp}/k) and are inactive for the remaining time [i.e. for $T_{hp}(k-1)/k$], as shown when describing the operation principle of the technique, a single small capacitor is exploited multiple times in a single period T . This approach not only allows reducing silicon area occupation, but also greatly simplifies the network that generates the control phases. The proposed architecture consists of a cascade of two types of cross-coupled CP stages, which have been chosen as building blocks: the odd stages are composed of two small pump capacitors (i.e., $C_{1,1} = C_{1,2} = C/k$) switched at frequency kf , whereas the even stages are made of two standard-sized capacitors (i.e., $C_{2,1} = C_{2,2} = C$) controlled with a switching frequency equal to f , as in a conventional CP. In this way, the two fractional capacitors $C_{1,1}$ and $C_{1,1}$ transfer, alternatively, the charge packets to $C_{2,1}$ (when Φ_2 is high) and to $C_{2,2}$ (when Φ_2 is low). Moreover, when one fractional capacitor is de-

livering charges, the other capacitor belonging to the same stage is charged by either $C_{2,1}$ or $C_{2,2}$ of the previous stage (depending on the level of Φ_2) with the exception of the first-stage capacitors, which are charged directly by V_{dd} . Therefore, all the charge transfers are achieved according to the optimized technique explained above and, thus, benefit from the same advantages. It is important to notice that the frequency-capacitance product is equal to $2fC$ in both the even and the odd stages, independently from the chosen k and, thus, according to equation (3.33), the charge pump output resistance should be equivalent to the R_{out} of a conventional CP composed by equally-sized pump capacitors. Hence, any improvement in the output resistance shown by the proposed charge pump is directly ascribed to the improved charge transfer technique.

To prove the effectiveness of the proposed architecture, a three-stage charge pump with ideal switches (i.e., $r_{on} = 0$) has been simulated with different values of the splitting factor k from 1 to 6 and with an input voltage equal to $V_{dd} = 3.3$ V. It is worth to point out that the charge pump with $k = 1$ corresponds to a conventional CP with equally sized pump capacitors. The two pump capacitors in the even stage were sized to be $C_{2,1} = C_{2,2} = 50$ pF and the switching frequency was set equal to $f = 1$ MHz. Thus, the odd stages have $C_{1,1} = C_{1,2} = (50/k)$ pF and $f = k$ MHz. The top- and bottom-plate parasitic capacitances, not shown in Fig. 3.14, were set to be equivalent to 5% and 3%, respectively, of the corresponding pump capacitance. The obtained I-V characteristics, represented in Fig. 3.15, show that the charge pump with the lowest output resistance corresponds to the highest splitting factor (i.e., $k = 6$) and, in general, a higher k leads to a lower R_{out} . Analogously, the power efficiency of the simulated charge pumps, shown in Fig. 3.16, is higher for larger values of k for the whole output current range. Therefore, the simulation results demonstrate that the proposed architecture is able to achieve a lower output resistance, which directly translates into lower ohmic losses and, thus, higher power efficiency.

A charge pump exploiting the proposed architecture was designed in a CMOS $0.35\mu\text{m}$ high-voltage technology. The designed CP, shown in Fig. 3.17, is a three-stage structure based on a cross-coupled CMOS topology [16]. The splitting factor was chosen to be $k = 4$ as a good trade-off between maximum switching frequency and efficiency improvement. This stage topology is an enhanced version of the standard voltage doubler structure [17], where the latched connection of the four switch transistor (i.e. N_i , P_i , N'_i , and P'_i with $i = 1$ to 3) is replaced with a small boosting circuit (i.e., N_{bi} , N'_{bi} , C_{bi} , and C'_{bi} with $i = 1$ to 3) in order to avoid the undesired dependence of the switch-

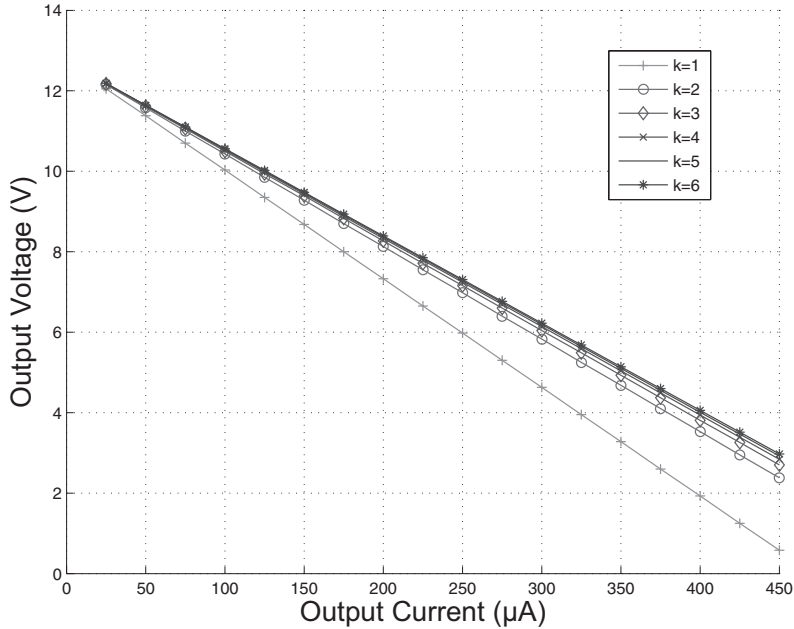


Figure 3.15: Simulated voltage output as a function of the output current of several cross-coupled charge pumps simulated with different values of parameter k .

transistors overdrive voltage with respect to the CP output current, which leads to an increase in the output resistance in the presence of a high output current.

The pump capacitors and the switching frequency were designed with the same values of the analysis presented above (case $k = 4$), therefore having: $C_{p,1} = C'_{p,1} = C_{p,3} = C'_{p,3} = C/4 = 12.5$ pF, $C_{p,2} = C'_{p,2} = C = 50$ pF, $f = 4$ MHz in the first and in the third stage, and $f = 1$ MHz in the second stage. Both the NMOS and PMOS switch transistors are equally sized in the three stages: indeed, the stages that are driven at $4f$ (i.e., the odd stages) also have 4-times smaller capacitors and, thus, τ results also reduced by a factor of 4. The NMOS switches were designed with a channel width $W_N = W_{N'} = 25$ μm , whereas the channel width of PMOS switches was set to $W_P = W_{P'} = dW_N = dW_{N'} = 85$ μm , where $d = 3.4$ is the ratio between the electrical mobilities of electrons and holes in the channel. This choice was made in order to obtain an equal ON-resistance of PMOS and NMOS switches, and, thus, have a symmetrical behavior when charging and discharging pump capacitors.

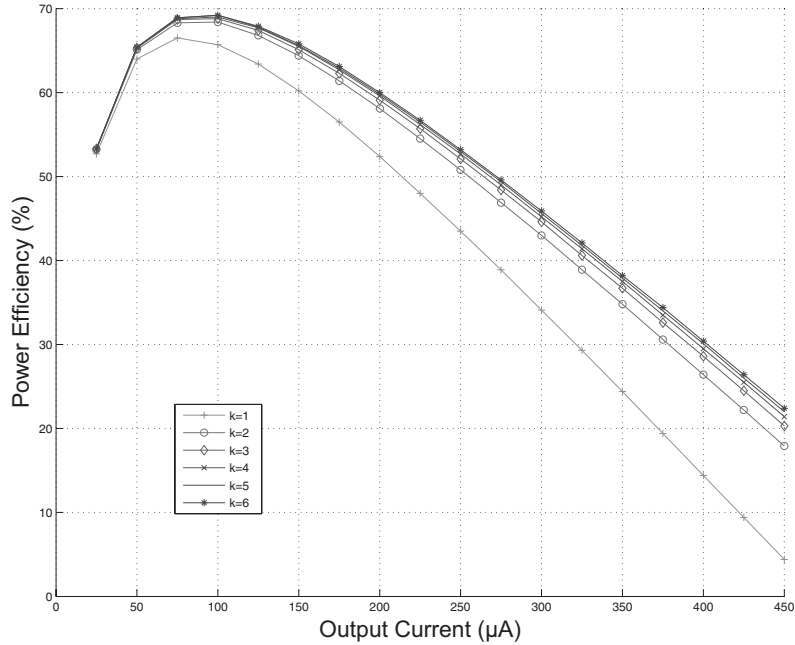


Figure 3.16: Simulated power efficiency as a function of the output current of several cross-coupled charge pumps simulated with different values of the splitting parameter k .

Table 3.5: Summary of the different CPs topologies simulated.

| Topology | Stage 1 | | Stage 2 | | Stage 3 | | Total Area |
|---------------------------|---------|-------|---------|-------|---------|-------|------------|
| Proposed | $4f$ | $C/4$ | f | C | $4f$ | $C/4$ | $3C$ |
| Conventional ₁ | f | C | f | C | f | C | $6C$ |
| Conventional ₂ | $2f$ | $C/2$ | $2f$ | $C/2$ | $2f$ | $C/2$ | $3C$ |
| Conventional ₃ | $4f$ | $C/4$ | $4f$ | $C/4$ | $4f$ | $C/4$ | $3C/2$ |

The capacitance contribution of MOS switches is negligible with respect to the top- and bottom-plate parasitic capacitance of the pump capacitors.

To compare the performance of the proposed architecture charge pump to the performance of conventional schemes, three additional CPs were designed in the same technology. The additional charge pumps were obtained by cascading the same building block (i.e. the cross-coupled stage) with the same values of switching frequency and pump capacitance throughout the stages. For fair comparison, the frequency-capacitance-product was set to the same

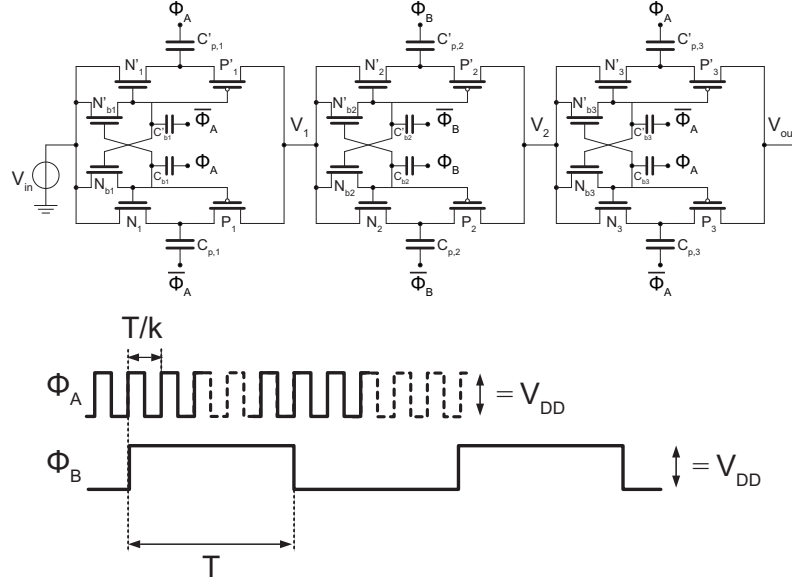


Figure 3.17: Circuit schematic of the CMOS implementation (three-stage charge pump) of the conceptual scheme in Fig. 3.14.

value in all stages of the four considered charge pumps. In particular, the three conventional CPs (namely, Conventional₁, Conventional₂, and Conventional₃) were designed to have an output resistance $R_{out} = 28.6 \text{ k}\Omega$. However, each conventional charge pump has different values of C and f , as shown in Tab. 3.5. An estimation of the silicon area occupation of the four simulated CPs is also provided in the Table, under the assumption that the pump capacitors are much larger than the rest of the components.

Figure 3.18 shows the I-V characteristics of the four simulated charge pumps. We can observe that the curves representing the three conventional CPs have the same slope and, thus, the same output resistance $R_{out} = 28 \text{ k}\Omega$, which is in good agreement with the theoretical calculation. The proposed charge pump, depicted with empty-circles markers, exhibits the lowest output resistance (about $22 \text{ k}\Omega$), which corresponds to an improvement by more than 20% with respect to the other three CPs.

The lower R_{out} directly translates into an overall better η , thanks to the reduction of the ohmic losses. Indeed, as depicted in Fig. 3.19, the proposed charge pump shows a better η over the whole output current range, thus proving the effectiveness of the presented architecture.

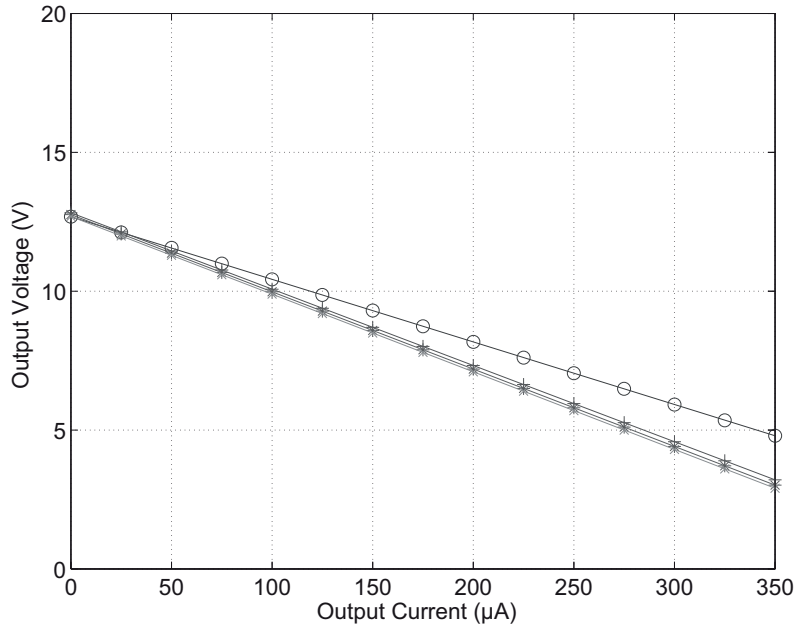


Figure 3.18: Voltage output as a function of the output current of the four simulated charge pumps: Conventional₁ (+ markers), Conventional₂ (* markers), Conventional₃ (× markers), and Proposed (○ markers).

3.4 Enhanced Voltage Buffer Compensation

An improved compensation technique for two-stage CMOS operational amplifiers, which can be extended to three-stage operational amplifiers, was developed for a second version of Spider-Mem chip.

The most common compensation technique for a two-stage operational amplifier is Miller compensation, whose operation principle has been introduced above (see Subsection 3.1.2.1), even though, in that case, a three-stage amplifier was considered.

Figure 3.20 depicts a two-stage CMOS operational amplifier in which the compensation network is not shown, but is represented as a generic block “comp. network”. If a compensation capacitance, C_C , is included in place of the generic block, standard Miller compensation is obtained. One of the main drawbacks of Miller compensation is the presence of a right half-plane zero [18], ω_z , due to the decreased impedance of C_C at sufficiently high frequency. In fact, the output signal of the first stage can take two different paths to reach

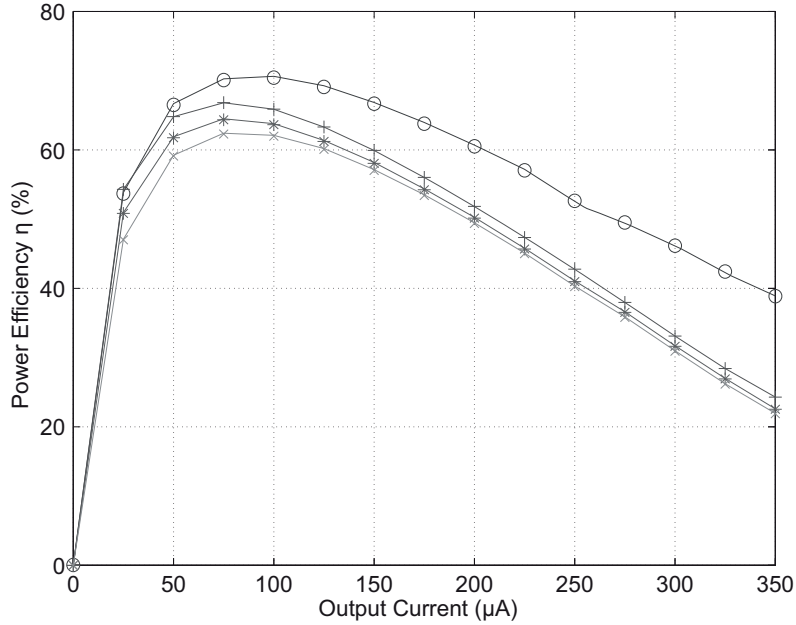


Figure 3.19: Power efficiency of the four simulated CPs as a function of the output current: Conventional₁ (+ markers), Conventional₂ (* markers), Conventional₃ (× markers), and Proposed (○ markers).

the output node. The first path takes place through transistor P_2 , which adds a 180° phase shift to the signal, whereas the second path is directly through the compensation capacitor. The current able to flow into the second path is a function of the signal frequency: indeed, C_C becomes more conductive at higher frequencies and imposing no phase shift to the signal. Therefore, the zero is located at the frequency at which the current generated by P_2 (i.e., $I_2 = -g_{m2}V_1$ where g_{m2} is the transconductance of the second stage) is equal in module to the current passing through C_C (i.e., $I_C = j\omega C_C V_1$), and is located in the right half-plane due to the different phases of the two currents. The right half-plane frequency is therefore expressed as

$$\omega_z = g_{m2}/C_C, \quad (3.39)$$

which typically results close to the unity-gain (angular) frequency $\omega_0 = g_{m1}/C_C$ and, thus, degrades the phase margin.

Several compensation techniques, depicted in Fig. 3.21, have been proposed during the past years to overcome this limitation. The simplest technique [19],

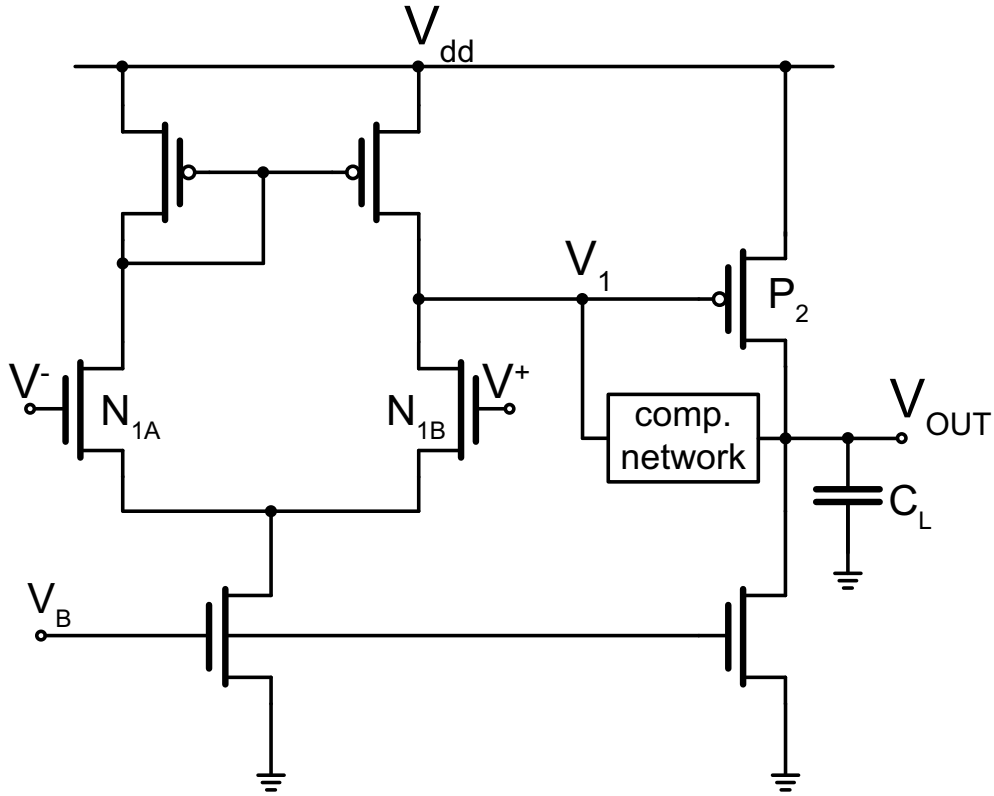


Figure 3.20: Circuit schematic of a two-stage CMOS operational amplifier with a generic compensation network.

shown in Fig. 3.21(a), exploits a resistor R_z to modify the impedance of the compensation network in order to control the frequency at which the zero occurs

$$\omega_z = \frac{1}{(1/g_{m2} - R_z)C_C}, \quad (3.40)$$

and allows moving it at very high frequency by sizing $R_z \approx 1/g_{m2}$.

Two additional techniques solve the problem from the root cause and avoid the zero formation by removing the signal feed-forward path through C_C . Both techniques, indeed, implement an active circuit that is unidirectional (i.e., that injects signals coming from the output node into node V_1), thus preventing any signal from node V_1 from going toward the load through C_C . The compensation network shown in Fig. 3.21(b) [20] exploits a current buffer, placed between C_C and V_1 , to inject the current generated by the compensation capacitor into

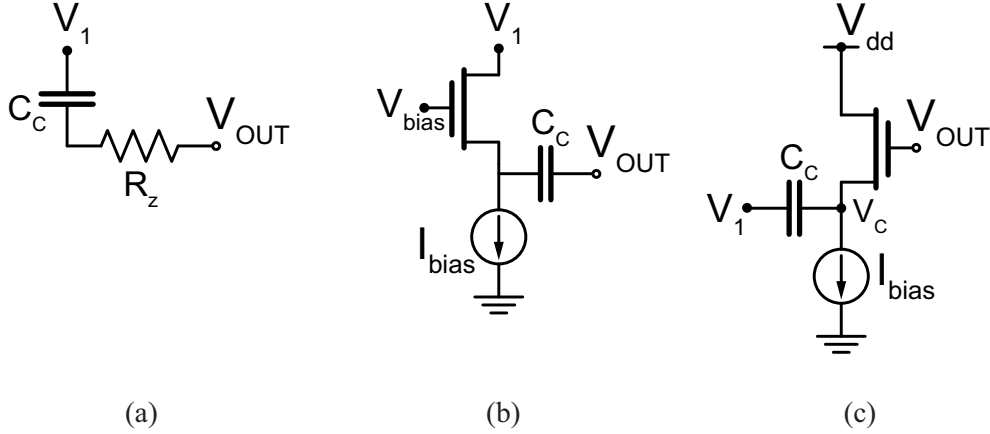


Figure 3.21: Compensation networks to be inserted in place of the generic ‘comp. network’ block in Fig. 3.20 in order to obtain: a zero-nulling compensation (a), a current-buffer compensation (b), and a voltage-buffer compensation (c).

node V_1 . This technique not only cuts the feed-forward path of the signal, but also achieves a voltage gain that, at high frequency, is equal to the ratio of the compensation capacitor to the parasitic capacitance on node V_1 , C_1 :

$$\frac{V_1}{V_{out}} = \frac{C_C}{C_1} \quad (3.41)$$

This voltage gain can be exploited to increase the gain-bandwidth product, GBW. The third method shown in Fig. 3.21(c) [21] uses a unity-gain voltage buffer, typically implemented with a source-follower transistor, placed in the compensation branch between the output node and the compensation capacitor. The unity-gain buffer transfers the output voltage to node V_C and, thus, an operational amplifier compensated with this approach has the same performance of a Miller-compensated one without the right half plane zero:

$$\omega_{p1} = \frac{1}{r_1 C_C g_{m2} r_2}, \quad (3.42)$$

$$\omega_0 = \frac{g_{m1}}{C_C}, \quad (3.43)$$

and

$$\omega_{p2} = \frac{g_{m2}}{C_{out}}, \quad (3.44)$$

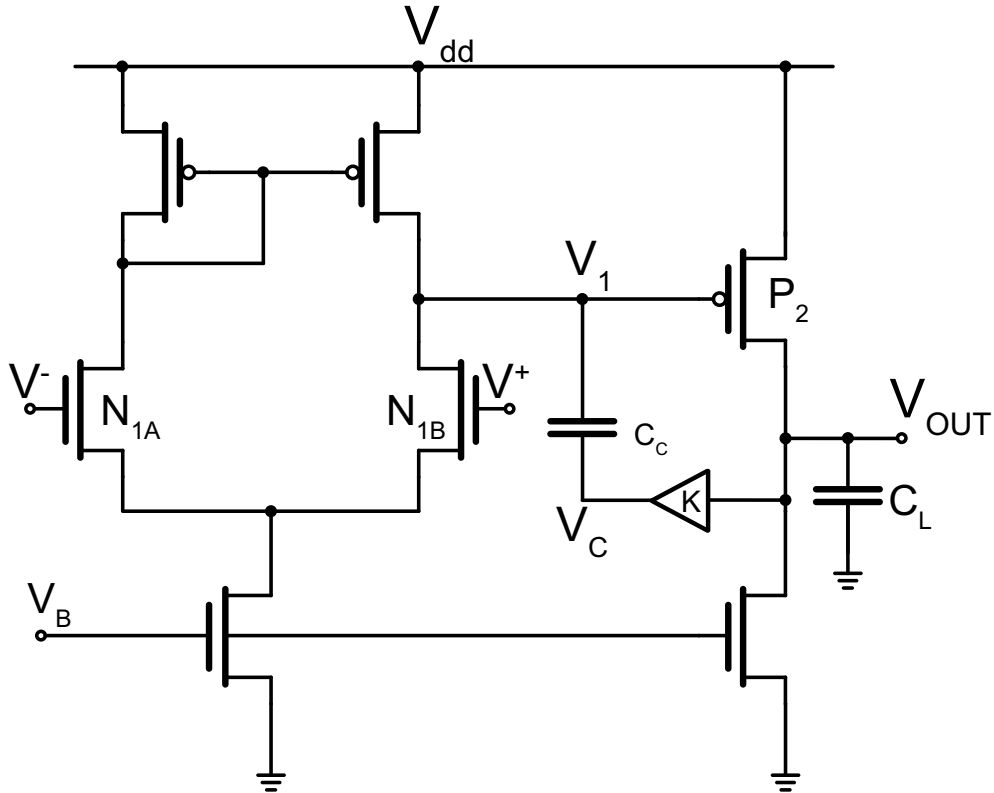


Figure 3.22: Circuit schematic of a two-stage operational amplifier with an enhanced voltage-buffer compensation.

where ω_{p1} and ω_{p2} are the first and second pole frequencies, respectively, r_1 and r_2 are the output resistances of the first and second stage, respectively, and g_{m1} is the transconductance of the first stage.

The technique proposed in this Thesis work [22] aims at improving the voltage-buffer compensation by replacing the unity-gain voltage buffer with a non-inverting voltage amplifier having a gain $K > 1$, as shown in Fig. 3.22. The idea is to exploit an adequate voltage gain in the compensation network to relocate the poles. Indeed, the addition of this voltage amplifier has a twofold impact on the operational-amplifier poles. The first takes place at low frequency and corresponds to an improvement of Miller effect: the voltage gain across the compensation capacitor is not equal to the gain of the second stage, $A_2 = g_{m2}r_2$, as in the case of standard compensation, but is increased by a

factor K . Therefore, the effective capacitance at node V_1 is increased by the same factor, so that the first pole is moved accordingly:

$$\omega_{p1} = \frac{1}{Kr_1C_Cg_{m2}r_2}. \quad (3.45)$$

For better understanding of the following part, it is useful to point out that if this were the only implication on the frequency response of the circuit, then it would be possible to reduce the size of the compensation capacitance by a factor K to achieve the same pole placement of Miller or voltage-buffer compensation techniques.

There is a second effect that takes place at high frequency: the low impedance shown by C_C at sufficiently high frequency, causes a direct anti-parallel connection of the voltage amplifier with respect to the second stage (i.e., V_C is short-circuited to V_1). Similarly to the effect obtained in a cascode compensation, this condition allows achieving an even lower high-frequency impedance at the output node and, thus, pushing the first non-dominant pole ω_{p2} to even higher frequency. To better understand this effect, let us suppose to apply a small high-frequency voltage signal, v_x , on V_{out} . The signal is immediately amplified by a factor K and provided to node V_C and, thanks to the above hypothesis, to node V_1 . The second-stage input is, therefore, driven with a voltage equal to Kv_x and, thus, injects a signal current $i_x = Kv_xg_{m2}$ into node V_{out} . The resulting high-frequency output impedance is

$$Z_{out,HF} = \frac{v_x}{i_x} = \frac{v_x}{Kv_xg_{m2}} = \frac{1}{Kg_{m2}}, \quad (3.46)$$

which allows the first non-dominant pole frequency to be calculated as

$$\omega_{p2} = \frac{1}{Z_{out,HV}C_{out}} = \frac{K}{g_{m2}C_{out}}. \quad (3.47)$$

The enhanced voltage-buffer compensation allows the second pole to be placed at a frequency K times higher with respect to the case of voltage-buffer compensation. This characteristic is particularly interesting since it is ω_{p2} that limits the gain-bandwidth product (i.e., ω_0 is then moved by the designer to ω_{p2}/m , where m represents the ratio ω_{p2}/ω_0 , with the aid of a suitable value of C_C). Therefore, having a higher ω_{p2} can be easily exploited to obtain a K -times higher gain-bandwidth product by simply setting C_C smaller by a factor K . However, as mentioned above, thanks to the enhanced Miller effect provided by the added positive gain amplifier, the size of compensation

capacitance can be further reduced by an additional factor K and, thus, we have

$$C_C = \frac{m}{K^2} \frac{g_{m1}}{g_{m2}} C_{out}. \quad (3.48)$$

The proposed compensation technique is therefore able to achieve a K times higher gain-bandwidth product by implementing a K^2 smaller compensation capacitance. It has to be noticed that equation (3.47) is obtained under the hypothesis $C_C \gg C_1$. Therefore, in the case of an excessive reduction of the compensation capacitance, the second pole may be pushed to a lower frequency than the value indicated by this equation. This effect is mainly due to the series connection of capacitances C_C and C_1 , which acts as a voltage divider: if C_C is much larger than C_1 , it is possible to approximate $V_1 \approx V_C$. However, if the hypothesis is not verified, we have $V_1 = V_C C_C / (C_C + C_1)$ and, thus, the second pole frequency can be approximated as

$$\omega_{p2} = \frac{K}{g_{m2} C_{out}} \frac{C_C}{C_C + C_1}. \quad (3.49)$$

Alternatively, in applications in which the GBW improvement is not desired, the designer can exploit the enhanced voltage-buffer compensation to set ω_{p2} equal to the second-pole frequency obtained through a voltage-buffer compensation and, concurrently, decrease g_{m2} by a factor K [equation (3.47)]. In this manner, the capacitance size reduction is limited to K , but the current needed to bias the second stage, I_2 , is decreased by a factor K^2 , since $g_{m2} \propto \sqrt{I_2}$ in pinch-off region. To make an accurate calculation about the power saved thanks to this approach, the power consumption of the block that achieves the gain K has to be taken into account. However, for applications that require a large value of I_2 should be straightforward to design a block biased with a current lower than $I_2 - (I_2/K^2)$, thus allowing a significant power reduction.

It is important to point out that the voltage gain in current-buffer compensation is a function of C_C , whereas in the proposed solution the two parameters are completely independent. For this reason, in the current-buffer compensation, there is a trade-off between the voltage gain (thus, the gain-bandwidth product improvement) and the size of C_C . Hence, it is not possible to fully exploit the voltage gain to achieve a corresponding area reduction.

In Tab. 3.6, the main parameters of the proposed compensation technique are summarized and compared to the parameters obtained with voltage-buffer compensation and current-buffer compensations.

The non-inverting voltage amplifier is a very simple block. The specifications that this block has to meet are mainly the following:

Table 3.6: Summary of the design parameters obtained with different compensations: enhanced voltage buffer (EVBC), voltage buffer (VBC), and current buffer (CBC) compensation

| | EVBC | VBC | CBC |
|---------------|---------------------------------------------|-------------------------------------|-------------------------------------------|
| ω_{p1} | $\frac{1}{Kr_1C_Cg_{m2}r_2}$ | $\frac{1}{r_1C_Cg_{m2}r_2}$ | $\frac{1}{r_1C_Cg_{m2}r_2}$ |
| ω_{p2} | $K\frac{g_{m2}}{C_{out}}$ | $\frac{g_{m2}}{C_{out}}$ | $\frac{C_C}{C_1}\frac{g_{m2}}{C_{out}}$ |
| C_C | $\frac{m}{K^2}\frac{g_{m1}}{g_{m2}}C_{out}$ | $m\frac{g_{m1}}{g_{m2}}C_{out}$ | $\sqrt{m\frac{g_{m1}}{g_{m2}}C_{out}C_1}$ |
| GBW | $\frac{K}{m}\frac{g_{m2}}{C_{out}}$ | $\frac{1}{m}\frac{g_{m2}}{C_{out}}$ | $\sqrt{\frac{g_{m1}g_{m2}}{mC_{out}C_1}}$ |

- it needs an **accurate gain**, which has to be robust against fabrication process variations, otherwise the sizing of the compensation capacitor is not able to provide the desired poles placement over the whole range of process parameters;
- a sufficiently **high bandwidth** of the voltage amplifier is required to effectively relocate the poles; in particular, its bandwidth has to be larger than the higher pole that has to be moved (i.e., the voltage amplifier bandwidth has to be larger than ω_{p2}).

While the latter constraint is mandatory, the former can be more relaxed if a tunable compensation capacitor is used. In this case, indeed, it is possible to obtain a stable pole constellation by trimming the value of C_C against a variation of K . In particular, to counteract a variation of the voltage gain ΔK from the expected value, from equation (3.48), it is necessary to have a compensation-capacitor tunable range of

$$\Delta C_C = (\Delta K)^2. \quad (3.50)$$

For instance, with $\Delta K = \pm 5\%$ than C_C must be able to vary by $\pm 25\%$. This constraint is easy to be met, since the expected value of C_C is largely reduced by the proposed compensation technique.

Figure 3.23 shows three possible circuits to implement the non-inverting voltage amplifier. The circuit shown in Fig. 3.23(a) exploits an operational amplifier with a resistive feedback to set the desired gain (i.e., $K = 1 +$

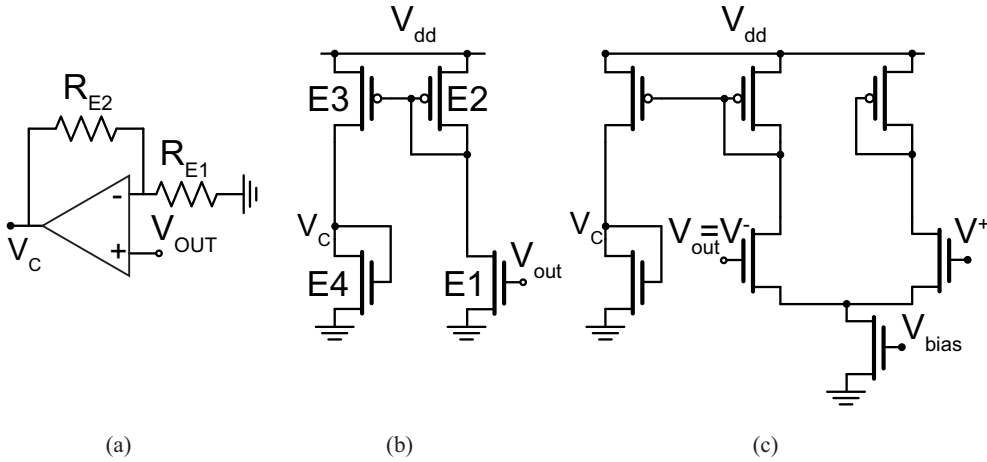


Figure 3.23: Three examples of CMOS implementation of the voltage gain stage K .

R_{E2}/R_{E1}). Note that this compensation is effective even with low values of K , so a single stage amplifier can be sufficient. The circuit represented in Fig. 3.23(b) exploits an NMOS transistor to convert the voltage signal to the output node, V_{out} , to a current $V_{out} g_{m,E1}$, which is fed to a current mirror that has a mirroring factor F . Therefore, a current $F V_{out} g_{m,E1}$ is injected into a diode-connected NMOS transistor, $E4$, which shows an impedance $1/g_{m,E4}$. Therefore, the gain K can be calculated as

$$K = F \frac{g_{m,E1}}{g_{m,E4}} \quad (3.51)$$

which can be set with high accuracy if a careful layout is done, with particular attention to the transistor matching. Moreover, this circuit has the advantage to work in an open-loop configuration, which helps the designer to satisfy the bandwidth requirements for this block. Finally, the circuit shown in Fig. 3.23(c) shares a lot of similarities with the solution Fig. 3.23(b). However, it has a crucial advantage when the power consumption is critical for the application, or a variable input common mode is expected: in this case, this circuit offers a controlled (and limited) bias current in the input stage, which is equal to the tail current of the differential pair. The drawbacks are the necessity to use the main two-stage amplifier in a unity-gain configuration (i.e., $V^- = V_{out}$), since it is required that V^- and V^+ are sufficiently close to each other, and a reduction of a factor 2 of the voltage gain obtained for the

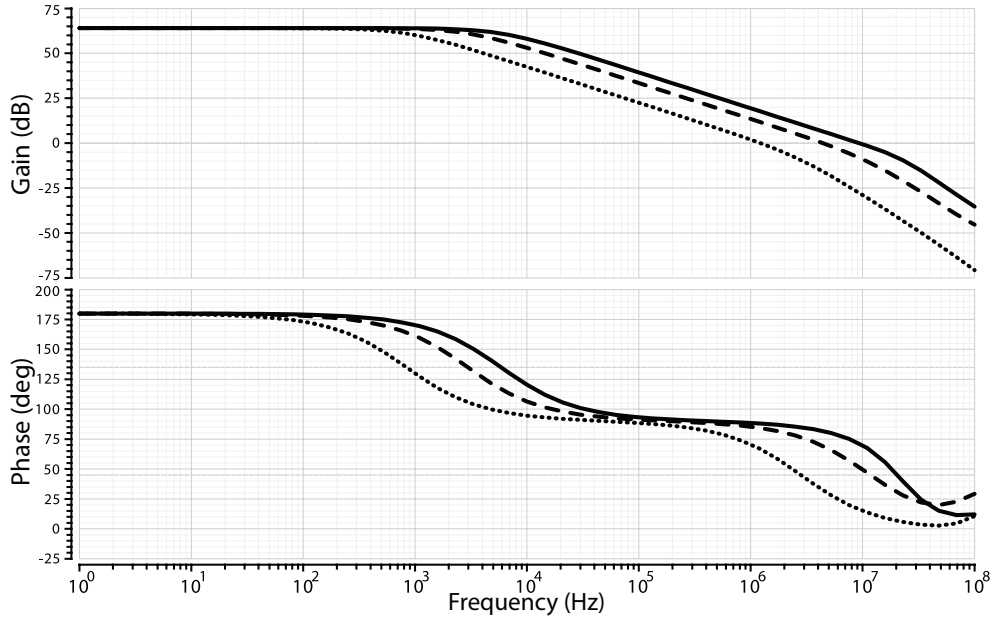


Figure 3.24: Simulated bode plots of the two-stage operational amplifier compensated with different solutions ($C_{out} = 50$ pF; dc gain $A_0 = 65$ dB). Voltage-buffer compensation: $K = 1$, $C_C = 15.5$ pF (dotted line). Enhanced voltage-buffer compensation: $K = 3.47$, $C_C = 1.29$ pF (dashed line); $K = 7.2$, $C_C = 0.3$ pF (solid line).

same power consumption.

3.4.1 Simulation Results

To prove the effectiveness of the proposed compensation technique, a two-stage operational amplifier was designed and then compensated with three different solutions: a voltage-buffer compensation, an enhanced voltage-buffer compensation with $K \approx 3.5$, and an enhanced-buffer compensation with $K \approx 7$. In both enhanced voltage-buffer compensations, the circuit in Fig. 3.23(b) was chosen and designed to implement the voltage amplifier. The capacitive load used for the comparison was set to $C_L = 50$ pF in all the three cases, and the transconductances of the two stages of the operational amplifier were designed as $g_{m1} \approx 118 \mu\text{A}/\text{V}$ and $g_{m2} \approx 900 \mu\text{A}/\text{V}$. The three designed operational amplifier are therefore identical with the exception of the compensation network, which allows a fair performance comparison of the compensation techniques.

Table 3.7: Summary of parameters extracted from Bode plots in Fig. 3.24 ($C_{out} = 50$ pF; op-amp dc gain $A_0 = 65$ dB)

| Compensation | K | C_C (pF) | φ_m | f_0 (MHz) | f_{p1} (kHz) | f_{p2} (MHz) |
|--------------|------|---------------|-------------|----------------|-------------------|-------------------|
| VBC | 1 | 15.5 | 66° | 1.2 | 0.85 | 2.9 |
| EVBC | 3.47 | 1.29 | 69° | 4.5 | 3.0 | 11.9 |
| EVBC | 7.2 | 0.3 | 70° | 9.5 | 6.0 | 20.6 |

The Bode plots shown in Fig. 3.24 represent the open-loop transfer functions of the three operational amplifiers. In the upper plot, the gain amplitude is depicted, whereas the gain phase is shown in the lower plot. It is easily observed that the solution that achieves the smallest bandwidth and, since the DC gain is equal in the three cases, the smallest gain-bandwidth product is the voltage-buffer compensation (dotted line) with $C_C = 15.5$ pF and $K = 1$. The dashed line represents the case of an enhanced voltage-buffer compensation exploiting a $K = 3.47$ and $C_C = 1.29$ pF, whereas the solid line indicates the case of $K = 7.2$ $C_C = 0.3$ pF.

The parameters extracted from the Bode plots of Fig. 3.24 are summarized in Tab. 3.7 in order to better appreciate the difference in performance and the obtained improvements. In particular, it is easy to appreciate the effective relocation of the poles and, at the same time, the reduction of compensation capacitance. The latter advantage is particularly attractive when the operational amplifier has to drive a large capacitive load, since in this circumstance the silicon area needed to implement C_C occupies a significant portion of the total circuit area and, in some case, can even be larger than the area of the rest of the operational amplifier.

Due to the good obtained results, not only this technique will be utilized during the design of the next version of chip Spider-Mem, but also a US Patent Application [23] was filed.

3.5 Bandwidth Optimization

The second study, carried out to achieve better design strategies for the future Spider-Mem version, was driven by the lack in literature, to the best of authors' knowledge, of a design procedure aimed at addressing a common problem faced by ICs designers: how to achieve the maximum gain-bandwidth product in a two-stage operational amplifier under given area and power consumption constraints.

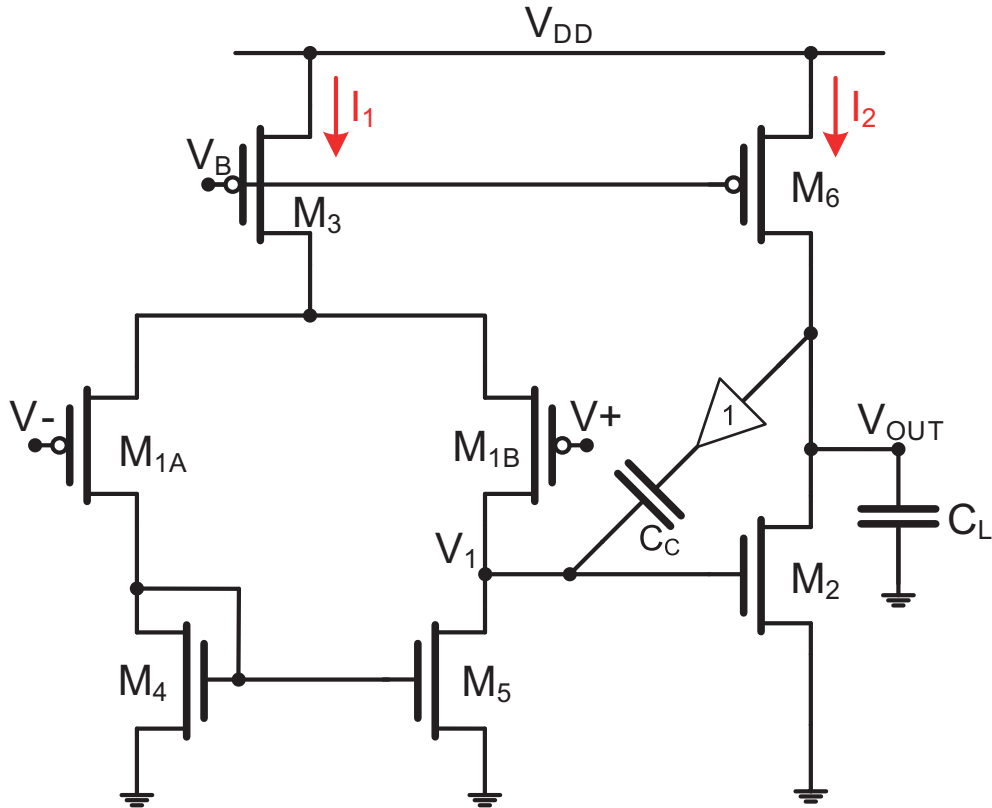


Figure 3.25: Standard CMOS two-stage operational amplifier with voltage-buffer Miller compensation.

In this respect, our idea is to provide a theoretical analysis that can guide the designer to achieve the above result starting from the total silicon area available to implement all the operational amplifier components (A_{TOT}), the power budget that can be consumed by the circuit (P_{TOT}), and the capacitive load that has to be driven (C_L). Moreover, in the following it will be assumed that all technology parameters are known, including the electrical mobility of the surface charge carriers (μ_n and μ_p for electrons and holes, respectively), the gate-oxide capacitance per unit area (C_{ox}), the minimum channel length (L_{min}), the nominal power supply (V_{dd}).

For the sake of simplicity, the analysis is carried out by assuming that both the first and the second stages of the operational amplifier are powered by V_{dd} and are implemented with the simplest possible topology: a differential pair

with an active load as an input stage and a common-source transistor topology as an output stage. Both stages obviously have an additional transistor that sets the bias current. It is worth to point out that the following optimization study does not make any assumption regarding the implementation of the two stages (i.e., NMOS or PMOS transistors). The only hypothesis consists in the duality of the two stages: an NMOS differential pair cascaded by a PMOS common source, or vice versa. Thus, the choice in the circuit representation of Fig. 3.25 is arbitrary. The compensation network is also included in the analysis and is assumed to be implemented with a Miller compensation technique with the right half-plane zero issue solved with a unity-gain buffer, so that it can be ignored from the stability point of view (as explained in previous Section).

Thanks to the hypothesis that both stages are supplied with V_{dd} , it is possible to translate the power consumption constraint into a current consumption constraint. Indeed, we can define a total current available to power the operational amplifier:

$$I_{TOT} = \frac{P_{TOT}}{V_{dd}} = I_1 + I_2, \quad (3.52)$$

where I_1 and I_2 are the bias currents of the first and the second stage, respectively.

The maximization of the gain-bandwidth product, when considering a two-stage operational amplifier, is strictly related with the frequency position of the first non-dominant pole, ω_{p2} . This constraint is dictated by the stability requirements of the amplifier in a closed-loop configuration. Therefore, in order to optimize GBW, we need to maximize ω_{p2} . Under the assumption $\omega_{p1} \ll \omega_{p2}$, the second pole frequency can be expressed as

$$\omega_{p2} = \frac{C_C g_{m2}}{C_1 C_C + C_1 C_L + C_C C_L}, \quad (3.53)$$

where C_1 is the parasitic capacitance on node V_1 and g_{m2} is the transconductance of the output stage. C_1 can be calculated as the sum of the capacitance contributions of transistors M_{1B} , M_5 and M_2 . However, M_2 is typically much bigger than the other transistors and, then, C_1 can be approximated, without compromising accuracy, as the gate capacitance of M_2 :

$$C_1 = C_{ox} W_2 L_2, \quad (3.54)$$

where W_2 and L_2 are the channel width and the channel length of transistor M_2 .

Intuitively, by looking at equation (3.53), it may seem possible that an arbitrary increment of ω_{p2} can be obtained by increasing g_{m2} . However, to achieve this goal either I_2 or W_2 have to be increased. The former solution is limited by the power consumption constraint, expressed by equation (3.52), whereas the latter approach leads to a larger C_1 , which adversely affects the value of ω_{p2} . These brief considerations suggests that not only the GBW enhancement is a not a straightforward task, but also that there is an optimum design point.

The transconductances of the two stages are fundamental parameters for both the operational-amplifier stability and the gain-bandwidth product maximization. As mentioned above, g_{m2} depends on the bias current of the output stage and on the aspect ratio of M_2 . To be more specific, the transconductance of the second stage can be expressed as

$$g_{m2} = \sqrt{\frac{2\mu_2 C_{ox} I_2 W_2}{L_2}}, \quad (3.55)$$

where μ_2 is the electrical mobility of the surface charge carriers of transistors M_2 . By applying similar considerations to the first stage, and observing that a current $I_1/2$ flows through each transistor of the differential pair, it is straightforward to calculate the first-stage transconductance:

$$g_{m1} = \sqrt{\frac{\mu_1 C_{ox} I_1 W_1}{L_1}}, \quad (3.56)$$

where μ_1 is the electrical mobility of the surface charge carriers, W_1 and L_1 are, respectively, the channel width and the channel length, respectively, of both transistors M_{A1} and M_{B1} . Note that a generic notation has been assigned to μ_1 and μ_2 for the aforementioned purpose of modeling both the complementary operational-amplifier implementation topologies.

The closed-loop stability of the operational amplifier is a function of the frequency separation of the unity-gain frequency, ω_0 , from the first non-dominant pole frequency. In particular, the phase margin of the operational amplifier can be expressed as

$$\varphi_m = \arctan\left(\frac{\omega_{p2}}{\omega_0}\right). \quad (3.57)$$

For this reason, a supplementary parameter, $\alpha = \omega_{p2}/\omega_0$, is added to our analysis in order to allow the designer to set φ_m as the best value for his/her needs. Therefore, the stability condition is imposed by forcing ω_{p2} to be placed at a frequency α times higher than $\omega_0 = g_{m1}/C_C$:

$$\frac{C_C g_{m2}}{C_1 C_C + C_1 C_L + C_C C_L} = \alpha \frac{g_{m1}}{C_C}. \quad (3.58)$$

From the silicon area occupation point of view, it is required to force the operational-amplifier components to fit in A_{TOT} . To do so, it is assumed that the biasing transistors (i.e., M_3 and M_6), as well as the transistors that implement the first-stage active load (i.e., M_4 and M_5) occupy a silicon surface portion that is negligible with respect to the other transistors. Therefore, the significant contributors to area occupation are the differential-pair transistors (i.e., M_{1A} and M_{1B}), the common source transistor (i.e. M_2), and the compensation capacitance. Thus, the silicon area constraint can be expressed as

$$A_{TOT} = A_1 + A_2 + A_C = 2W_1L_1 + W_2L_2 + \frac{C_C}{C_{ox}}, \quad (3.59)$$

where A_1 , A_2 , and A_C are the silicon areas occupied, respectively, by the first stage, the second stage, and C_C .

By set a system of equations that includes (3.52), (3.54), (3.55), (3.56), (3.58), and (3.59), we obtain the expression of the unity-gain frequency, which is compliant with all the above mentioned constraints:

$$\omega_0 = \sqrt{\frac{2\mu_2 \frac{W_2}{L_2} C_{ox} I_{TOT}}{2\frac{\mu_2}{\mu_1} \frac{W_2}{W_1} \frac{L_1}{L_2} C_{ox}^2 A_C^2 + \alpha^2 [C_L(1 + \frac{A_2}{A_C}) + C_{ox} A_2]^2}}. \quad (3.60)$$

It has to be pointed out that the analysis has been carried out without considering the values of the output resistance of both stages. This choice was made in order to obtain simpler equations, which are easier to be exploited by the designer. However, this approach does not limit the analysis usefulness: indeed, for a number of applications the DC gain is not an important parameter, provided that it is sufficiently high. Designers can, therefore, size L_1 and L_2 so as to achieve adequate values of DC gain and flicker noise reduction. In any case, it is required that the output resistance of the first stage (r_1) and the second stage (r_2) satisfy the following conditions: $g_{m1}r_1 \gg 1$, $g_{m2} \gg 1$.

A numerical analysis was carried out to study equation (3.60), which is not straightforward to solve with a standard algebraic computation and without losing generality by adding further hypotheses. The only variables that are unknown in equation (3.60) are the channel widths of transistors M_{1A} , M_{1B} , and M_2 (i.e., W_1 and W_2), indeed:

- μ_1 , μ_2 , C_{ox} are provided by technologists;
- C_L , I_{TOT} , and A_{TOT} are constraints and specifications given to the designer; and

- α , L_1 and L_2 are chosen, as mentioned above, by the designer to meet the specific application requirements (i.e., phase margin and DC gain).

Thus, the numerical analysis was carried out by running several iterations of the same cycle:

- firstly, a pair of W_1 and W_2 values is chosen;
- W_1 and W_2 are used to immediately calculate C_1 , C_C , and ω_0 using equations (3.54), (3.59), and (3.60), respectively;
- having obtained ω_0 allows calculating g_{m2} through $\omega_0 = g_{m1}/C_C$ and equation (3.58);
- I_2 is now determined through equation (3.55), since g_{m2} has been delivered, and I_1 is calculated through equation (3.52);
- finally, g_{m1} can be calculated using the value of I_1 in equation (3.56).

For each cycle the pair of W_1 and W_2 is changed, and the operation is carried on until all the possible pairs of values are scanned. The values of both W_1 and W_2 are chosen between the minimum value allowed by the technology and the maximum value allowed by the area constraint (i.e., $W_{1,max} = A_{TOT}/2L_1$ and $W_{2,max} = A_{TOT}/L_2$, respectively).

The chart in Fig. 3.26 depicts the top view of a 3D plot that represents, using a color scale, the obtained value of ω_0 as a function of W_1 and W_2 with the following set of parameters: $C_{ox} = 3.6 \text{ fF}/\mu\text{m}^2$, $\mu_1 = \mu_p = \mu_2/3 = \mu_n/3 = 9.26 \times 10^{-3} \text{ m}^2/\text{Vs}$, $L_1 = L_2 = 1 \mu\text{m}$. These parameters have been extracted from the models of a CMOS high-voltage 350 nm technology. Additionally, I_{TOT} , C_L , and I_{TOT} have been arbitrarily chosen assuming to have to drive a 36 pF capacitive load with 200 μA with a silicon area 10 times smaller than A_C . The whole graph area is not actually available for the design: indeed, the highlighted region above the segment AB does not satisfy the silicon area constraint (i.e., it corresponds to $A_1 + A_2 > A_{TOT}$). The three segments that enclose the valid design space (i.e., segments AB, BO, and OA) corresponds to circuits in which at least one of the three main components of the operational amplifier (from the silicon area occupation point of view) is set to have zero area occupation (i.e., $A_1 = 0$, $A_2 = 0$, or $A_C = 0$). Therefore, none of the points belonging to these segments can be considered valid. Thanks to the color scale, it is possible to appreciate how ω_0 varies using different pairs of W_1 and W_2 . In particular, the red area corresponds to the pairs of width values that provide the larger unity-gain frequency. It is fundamental to remark that,

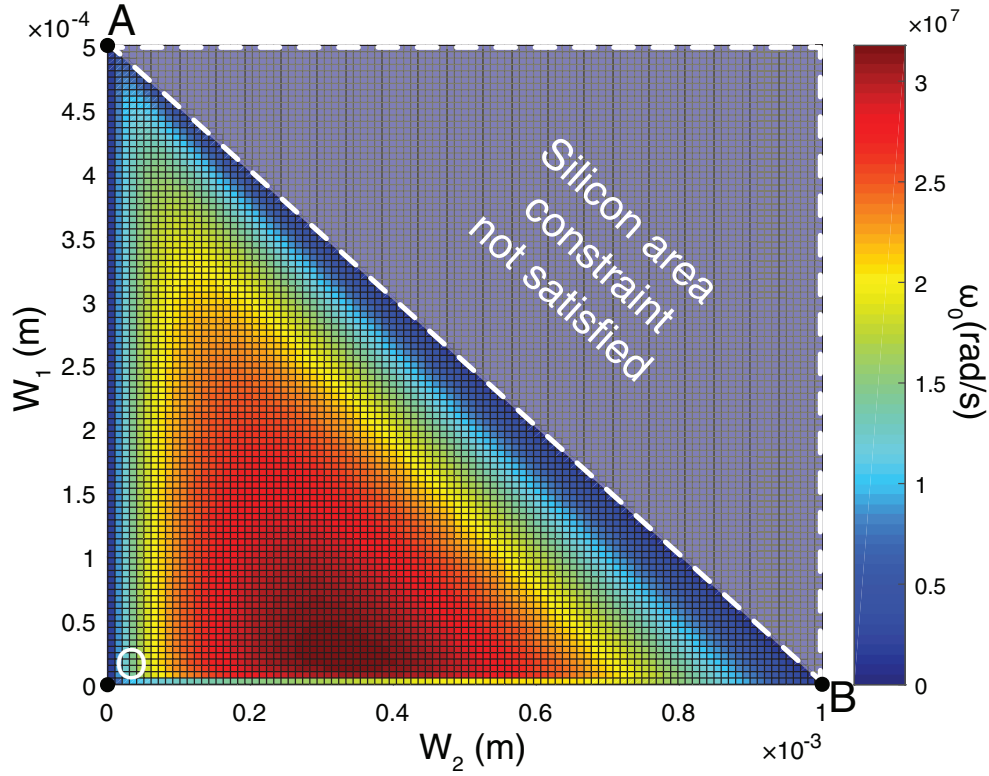


Figure 3.26: Unity-gain bandwidth as a function of W_1 and W_2 with the following set of specifications: $\alpha = 2$, $C_L = 36$ pF, $I_{TOT} = 200$ μ A, $L_1 = L_2 = 1$ μ m, $A_{TOT} = 1000$ μ m², $C_{ox} = 3.6$ fF/ μ m², $\mu_1 = \mu_p = \mu_2/3 = \mu_n/3 = 9.26 \times 10^{-3}$ m²/Vs.

in each point, the values of W_1 and W_2 are not the only variables to take into considerations: indeed, the full set of parameters to carry out a complete design (e.g. I_1 , I_2 , C_C , etc.) are also associated to each point. Furthermore, the surface region of the plot that corresponds to the optimum unity-gain frequency values is sufficiently large and its slope smooth enough to be easily targeted by the designer.

The successive step in our analysis was to carry out a simulation with a given set of parameters, find the point that represents the design with the highest unity-gain frequency (i.e., $\omega_0 = \omega_{0max}$), and extract the values of W_1 , W_2 , I_1 , I_2 , and C_C that are needed to design an operational amplifier with unity-gain frequency equal to ω_{0max} , namely W_{1opt} , W_{2opt} , I_{1opt} , I_{2opt} , and C_{Copt} . Then, the same simulation as above has been run with different values

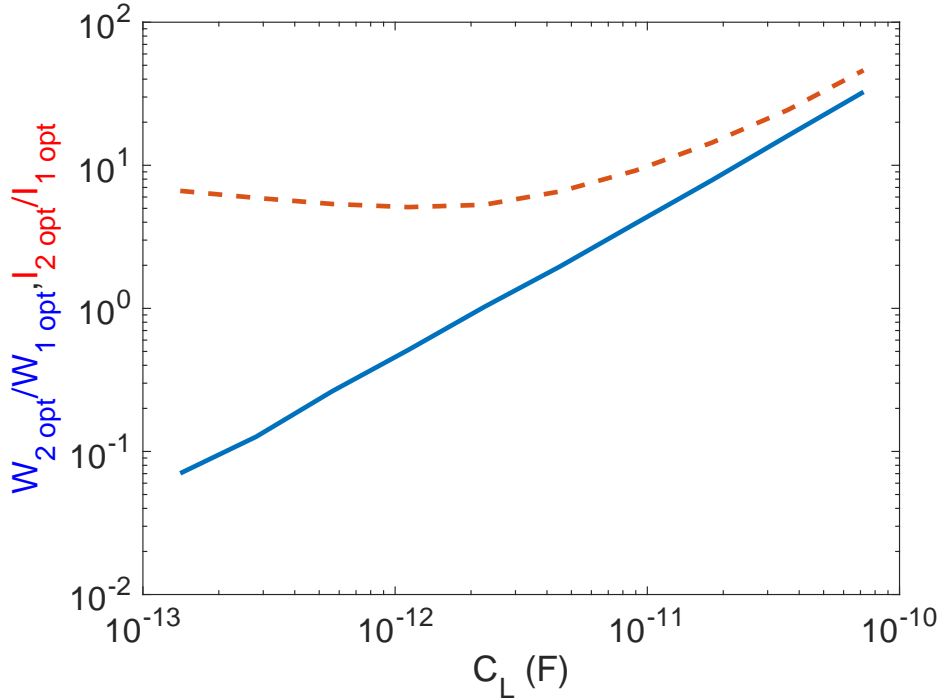


Figure 3.27: Ratios W_{2opt}/W_{1opt} (blue solid line) and I_{2opt}/I_{1opt} (red dashed line) as a function of load capacitance and the same set of specifications used for Fig. 3.26.

of C_L with the rest of initial parameters unchanged. Several simulations were carried out to span a very large range of capacitive load ($C_L = 100 \text{ fF} \div 100 \text{ pF}$), and the optimum design-parameters values were collected for each value of C_L . In the following, the results of these simulations (carried out with the same set of parameters of Fig. 3.26) are presented.

The ratio W_{2opt}/W_{1opt} , represented in Fig. 3.27 and depicted with a blue solid line, shows an interesting proportional dependence upon the load capacitance for more than three decades variation of C_L . In the same figure, the ratio I_{2opt}/I_{1opt} is depicted with a red dashed line and shows an almost constant behavior of this ratio is observed for small values of C_L , whereas a proportional dependency is apparent for large values of the load capacitance. The region in which the slope of I_{2opt}/I_{1opt} changes is in proximity of the load capacitance value that corresponds to $C_L = C_{ox}A_{TOT} = C_{TOT}$, where C_{TOT} represent the maximum capacitance that is possible to implement, as a MOS

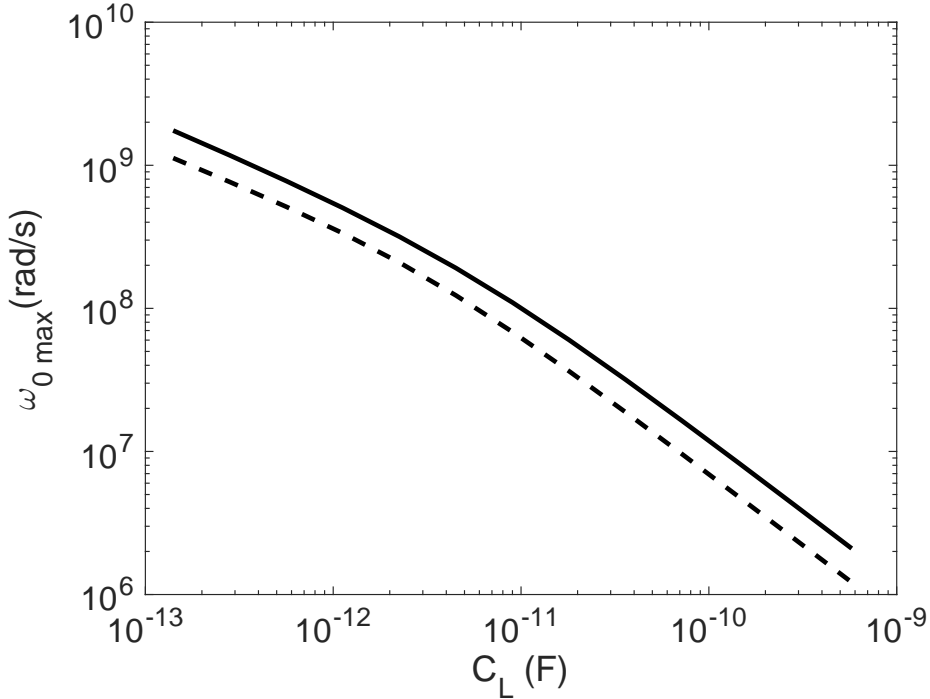


Figure 3.28: Dependency of the maximized unity-gain bandwidth upon load capacitance with the same set of specifications used for Fig. 3.26 (solid line). The dashed curve corresponds to the same set of specifications with the exception $\mu_2 = \mu_p = \mu_1/3 = \mu_n/3 = 9.26 \times 10^{-3}$.

capacitor, in A_{TOT} .

Exploiting the plots shown in Fig. 3.27, the designer can extract the values of ratios W_{2opt}/W_{1opt} and I_{2opt}/I_{1opt} that correspond to the considered C_L . Then it is possible to calculate I_{1opt} , I_{2opt} , W_{1opt} , W_{2opt} , and C_{Copt} using equations (3.52), (3.54), (3.55), (3.56), (3.58), and (3.59). With this approach, the designer can avoid the burden to face complicated and not intuitive equations.

The results of the above simulations that represents ω_{0max} are depicted with a solid curve in Fig. 3.28. Obviously, ω_{0max} displays a monotonic decreasing behavior for increasing values of C_L . For large capacitive loads (i.e., in the rightmost part of the plot), ω_{0max} becomes inversely proportional to C_L . It is worth to notice that ω_{0max} achieves lower values if simulations are carried out with the same parameters excepts for the μ_1 and μ_2 values, which are reversed (i.e. $\mu_1 = 3\mu_2 > \mu_2$). This case is represented by a black dashed

curve in Fig. 3.28 and corresponds to the case of an input differential pair implemented with NMOS transistors, since electrons have a higher mobility than holes. It is apparent that a higher gain-bandwidth product can be obtained choosing a PMOS implementation of the differential pair transistors, which, at the same time, help to achieve a lower flicker noise.

3.5.1 Large Capacitive Loads

An interesting aspect of the presented results is the particular dependence of the design parameters in the case of large capacitive loads. In this analysis, a capacitive load is considered *large* when it is much bigger than the equivalent total capacitance that can be implemented when using an MOS capacitor that occupies an area A_{TOT} . This case is particularly interesting for designers, since it is very common to have a limited silicon area to drive a large capacitive load.

By applying the hypothesis $C_L \gg C_{ox}A_{TOT}$ to equation (3.60), it is possible to neglect the term $C_{ox}A_2$ since it is much smaller than C_L . Therefore, equation (3.60) can be simplified and, then, rewritten (applying some mathematical manipulation) as

$$\omega_0 = \sqrt{\frac{2\mu_2 W_2 C_{ox} I_{TOT} A_C^2}{\alpha^2 L_2 C_L^2 (A_C + A_2)^2}} \frac{1}{\sqrt{1 + \frac{2}{\alpha^2} \frac{\mu_2}{\mu_1} \frac{W_2}{W_1} \frac{L_1}{L_2} \frac{C_{ox}^2}{C_L^2} \frac{A_C^4}{(A_C + A_2)^2}}} \quad (3.61)$$

A Taylor series expansion can be applied to the second factor of (3.61) to simplify the equation. This approximation can be applied since the second addend under the square root is much smaller than unity, due to the fact that $\frac{C_{ox}^2 A_C^4}{C_L^2} \ll 1$. The expression of ω_{0opt} obtained with a first-order expansion is

$$\omega_0 = \sqrt{\frac{2\mu_2 W_2 C_{ox} I_{TOT} A_C^2}{\alpha^2 L_2 C_L^2 (A_C + A_2)^2}} \left(1 - \frac{1}{\alpha^2} \frac{\mu_2}{\mu_1} \frac{W_2}{W_1} \frac{L_1}{L_2} \frac{C_{ox}^2}{C_L^2} \frac{A_C^4}{(A_C + A_2)^2} \right) \quad (3.62)$$

Finally, in order to find the values of W_1 and W_2 that maximize the unity-gain frequency (W_{1opt} and W_{2opt}), the two components of the gradient ($\nabla\omega_0$) were first calculated and then set equal to zero. The pair of optimum values is:

$$W_{1opt} \approx \frac{4}{9\alpha} \sqrt{\frac{\mu_2}{3\mu_1}} \frac{C_{ox} A_{TOT}^2}{C_L L_2} \quad (3.63)$$

$$W_{2opt} \approx \frac{A_{TOT} - 2W_1 L_1}{3L_2} \approx \frac{A_{TOT}}{3L_2}. \quad (3.64)$$

Having W_{1opt} and W_{2opt} allows calculating the optimum compensation capacitance

$$C_{Copt} = C_{ox}(A_{TOT} - 2W_1L_1 - W_2L_2) \approx \frac{2}{3}C_{ox}A_{TOT}, \quad (3.65)$$

as well as I_{1opt} and I_{2opt}

$$I_{1opt} \approx \frac{8}{27\alpha} \sqrt{\frac{3\mu_2}{\mu_1} \frac{L_1}{L_2} \frac{C_{ox}A_{TOT}}{C_L}} I_{TOT}, \quad (3.66)$$

$$I_{2opt} = I_{TOT} - I_{1opt} \approx I_{TOT}. \quad (3.67)$$

Equations (3.63), (3.64), (3.65), (3.66), and (3.67) represent the final set of expressions required to fully design the operational amplifier. Therefore, designing an operational amplifier with a large capacitive load and using the provided optimum parameters, allows obtaining a maximum gain-bandwidth product that can be expressed as

$$\omega_{0max} \approx \frac{1}{\alpha C_L L_2} \sqrt{\frac{8}{27} \mu_2 C_{ox} A_{TOT} I_{TOT}}. \quad (3.68)$$

Equation (3.68) allows the op-amp unity-gain frequency to be easily calculated when silicon area and power consumption are given, however, the designer could also use it to estimate the $A_{TOT}I_{TOT}$ product needed to achieve the desired ω_0 .

It is interesting to notice that, in the optimum case, the majority of the area available is occupied by the compensation and the second stage of the operational amplifier [as can be clearly seen in (3.65) and (3.64)], whereas the area occupation of the first stage is negligible, to a first-order approximation. In a similar manner, in the optimum case, the majority of I_{TOT} is used to bias the second stage. It is also interesting to notice that the ratios g_{m2opt}/g_{m1opt} , I_{2opt}/I_{1opt} , and W_{2opt}/W_{1opt} are proportional to the factor $C_L/(A_{TOT}C_{ox})$.

3.5.2 Comparison with Circuit Simulation

To validate the optimization analysis, a small-signal circuit, shown in Fig. 3.29, has been simulated in *Cadence Virtuoso*[®] environment. The parameters extracted from the optimum point of Fig. 3.26 (i.e., $I_{1opt} \approx 7.89 \mu\text{A}$, $I_{2opt} \approx 192.11 \mu\text{A}$, $W_{1opt} \approx 20.1 \mu\text{m}$, $W_{2opt} \approx 323.5 \mu\text{m}$, $C_{Copt} \approx 2.29 \text{ pF}$) have been used to calculate the small-signal parameters of the simulated circuit: $g_{m1opt} \approx 7.28 \mu\text{A/V}$, $g_{m2opt} \approx 3.525 \text{ mA/V}$, $C_{1opt} \approx 1.16 \text{ pF}$. The output resistances of the first and second stage (r_1 , r_2) were sized in order to achieve a sufficiently large DC gain (i.e., about 70 dB).

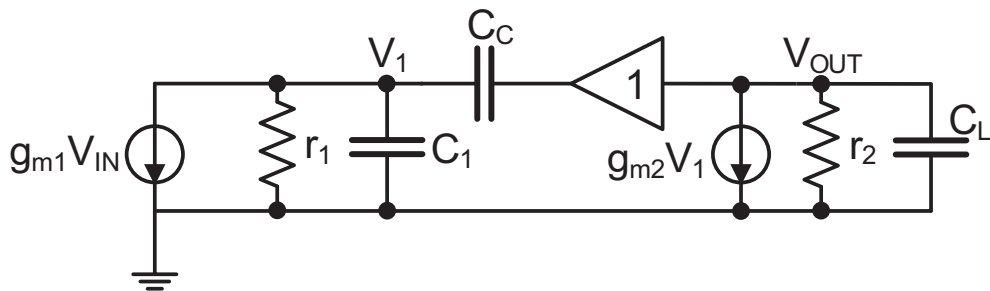


Figure 3.29: Equivalent small-signal circuit that models a standard CMOS two-stage operational amplifier with voltage-buffer compensation.

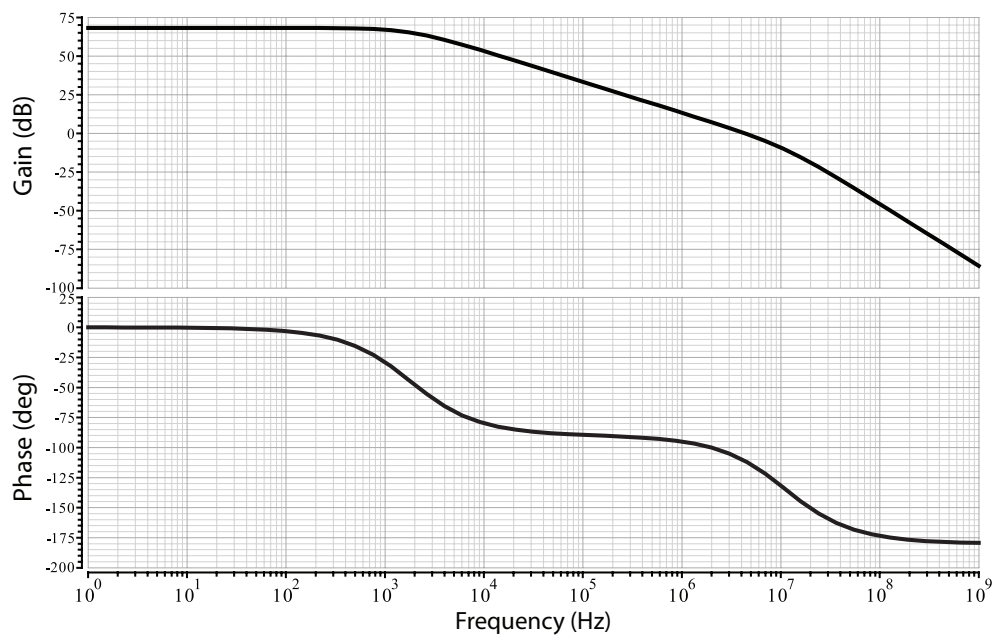


Figure 3.30: Bode plot of the circuit in Fig. 3.29 sized to target the optimum point of the plot in Fig. 3.26.

The Bode plots, shown in Fig. 3.30, depict the magnitude and the phase of the transfer function of the small-signal circuit. The unity-gain frequency and the first non-dominant pole frequency calculated with the aid of the presented numerical analysis are:

$$f_{0max} = \frac{\omega_{0max}}{2\pi} = 5.06 \text{ MHz} \quad (3.69)$$

and, since $\alpha = 2$,

$$f_{p2max} = 10.12 \text{ MHz}, \quad (3.70)$$

whereas the values extracted from the Bode plots obtained with CAD simulation are:

$$f_{0CAD} = 4.43 \text{ MHz} \quad (3.71)$$

and

$$f_{p2CAD} = 11.46 \text{ MHz}. \quad (3.72)$$

The results are therefore in good agreement. The difference between the two values is ascribed to the approximation which is typically used to calculate the dominant pole position and, thus, the unity-gain frequency:

$$g_{m2}r_1r_2C_C + r_2(C_L + C_C) + r_1(C_1 + C_C) \approx g_{m2}r_1r_2C_C. \quad (3.73)$$

The only way to increase the accuracy of our analysis is to take the output resistance of the two stages in consideration. However, as already mentioned, this approach leads to equations that are not intuitive and, thus, are difficult to exploit by the designer.

Chapter 4

Experimental Characterization

In this Chapter, the results of the experimental characterization of the Spider-Mem chip are presented. The micrograph of the Spider-Mem test chip is depicted in Fig. 4.1, where the most significant blocks are highlighted and the

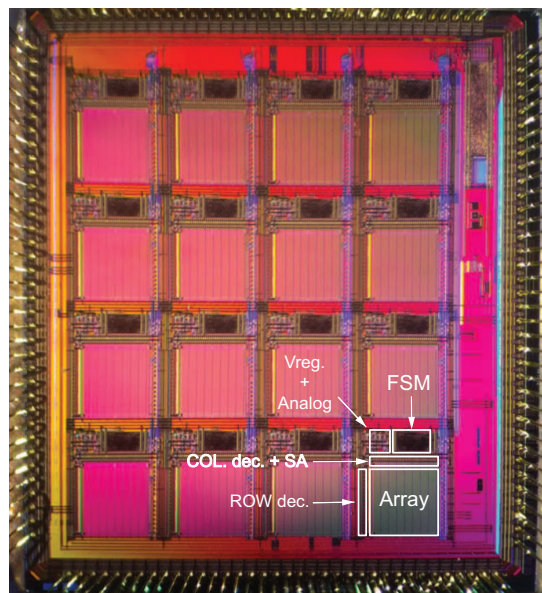


Figure 4.1: Test-chip micrograph (main blocks are identified). Sixteen 32-KB ePCMs are integrated in a 0.7 mm^2 silicon area.

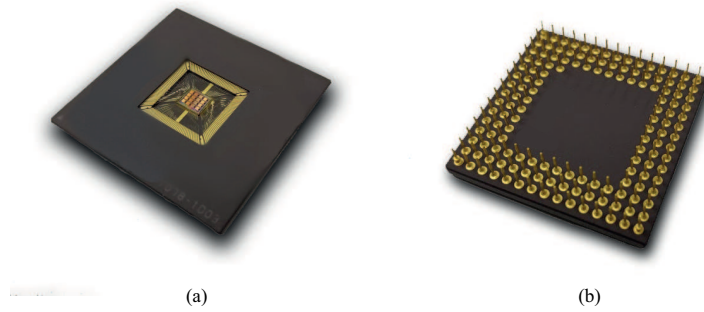


Figure 4.2: Top- (a) and bottom- (b) sides of the test-chip package.

sixteen 32 KB macrocells are clearly recognizable. The test-chips were assembled on Ceramic Pin Grid Array (CPGA) packages, shown in Fig. 4.2, that feature a removable ceramic lid, which is very convenient to protect the assembled silicon die, while allowing the possibility to access the top layer of the chip when required. Indeed, to measure the dynamic evolution over time of internal signal, several micro-pads were included in the design. Therefore, with the aid of a microscope, it is possible to place thin active micro-probes on the micro-pads in order to measure the desired signal without adding a significant capacitive load (as happens when passive micro-probes are used) and, thus, without altering the shape of the signals.

The experimental characterization of the test-chip has been carried out with the aid of a custom evaluation board designed by STMicroelectronics. The board can be controlled, through a Universal Serial Bus (USB) connection, by a custom software, also developed by STMicroelectronics, installed on a computer. The custom software includes a graphic user interface, which simplifies the measurements and provides a high flexibility, since the code can be easily modified. The board includes an STM32 microcontroller, to provide the test chip with the required signals, and several DAC generators, which are digitally programmable to feed the chip with the desired voltage supplies and references.

A Temptronic TP04390A ThermoStream® has been used to generate a temperature-controlled air flow that can be oriented to hit, directly, the silicon-die surface in order to reach the desired temperature. In this way, it was possible to carry out several measurements to validate the design of the test-chip as well as to measure its performance at different temperatures.

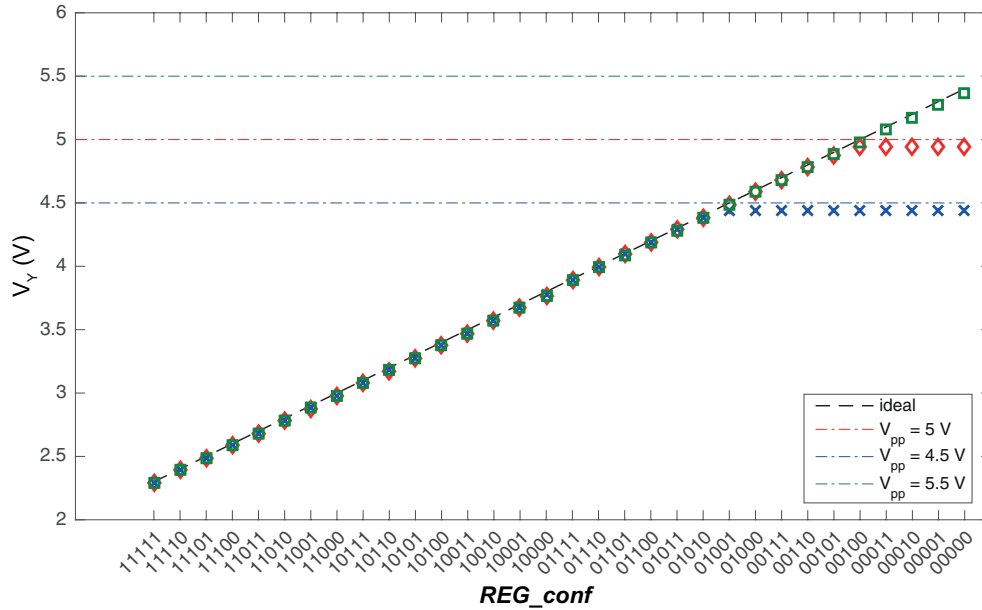


Figure 4.3: Measured V_Y regulator output voltage as a function of signal `REG_conf` obtained at room temperature and with three different high-power supplies values: $V_{pp} = 4.5$ V (blue \times markers), $V_{pp} = 5.0$ V (red \diamond markers), and $V_{pp} = 5.5$ V (green \square markers). The black dashed line represents the desired value of V_Y as a function of signal `REG_conf`.

4.1 V_Y Voltage regulator

The output voltage of the V_Y regulator was measured: the obtained results are shown in Fig. 4.3. In particular, three sets of measurement, obtained with three different values of the high-voltage power supply [i.e. the minimum (blue \times markers), the nominal (red \diamond markers), and the maximum (green \square markers) V_{pp} values] are displayed. Each set of measurement was collected with all the possible configurations of signal `REG_conf`, which are shown on the x-axis. It is straightforward to notice that each set of measurements saturates when the desired V_Y value is sufficiently close (i.e., different by less than 100 mV) to the high-voltage power supply value, which is represented with a dash-dotted line depicted with the same color of the corresponding markers. The black dashed line represents the expected values of V_Y as a function of signal `REG_conf`.

Eight sets of measurements, shown in Fig. 4.4, represent the output voltage of the V_Y regulator obtained by varying the 3-bit signal that controls the value

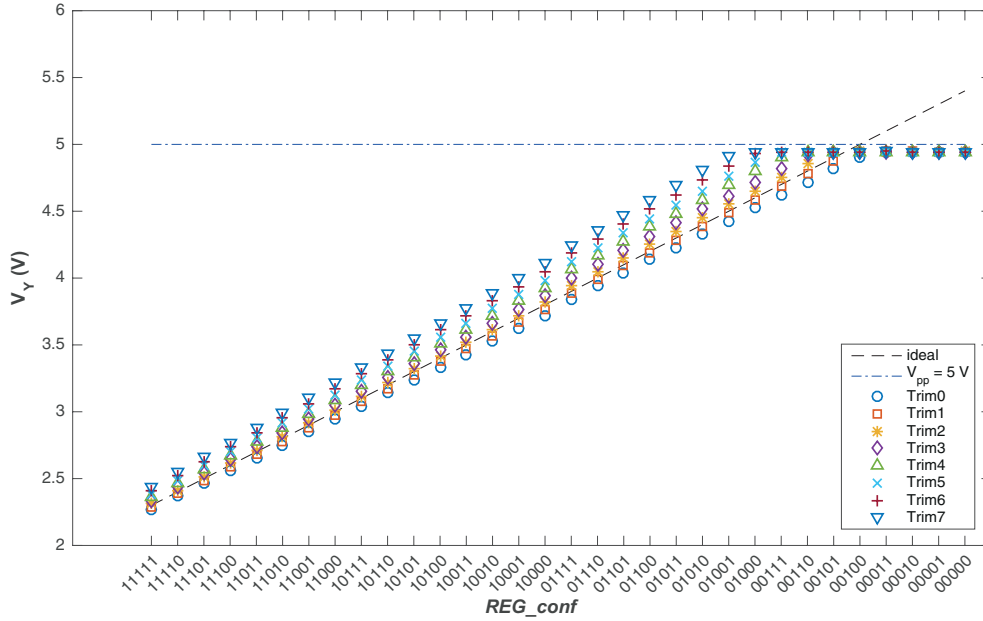


Figure 4.4: Measured V_Y regulator output voltage as a function of signal `REG_conf` obtained at room temperature, with the eight available trimming configurations (i.e., Trim_i with $i = 0$ to 7), and $V_{pp} = 5$ V. The black dashed line represents the desired value of V_Y as a function of signal `REG_conf`.

of resistor R_{trim} (see Subsection 3.1.1), i.e. Trim_i with $i = 0$ to 7 . Trim_0 represents the configuration that achieves the maximum value of R_{trim} (i.e., ≈ 17.5 k Ω) and, thus, the minimum V_Y in good agreement with equation (3.4). On the contrary, Trim_7 corresponds to the minimum value of R_{trim} (i.e., ≈ 0) and, hence, the maximum V_Y .

The measured absolute error (i.e. the absolute error of the measured V_Y regulator output voltage with respect to the desired value), shown in Fig. 4.5, helps to verify which trimming configuration is the optimum for the voltage regulator. Each of the 16 voltage regulators integrated in the test chip (one regulator is included in each macro-cell) can be trimmed independently from the others. The above measurements are carried out during the EWS phase and the result is stored in the reserved sector of each PCM memory, so that the finite state machine can load the optimum configuration every time that the voltage regulator is activated. It is straightforward to notice that the optimum trimming configuration, for the experimentally-characterized voltage

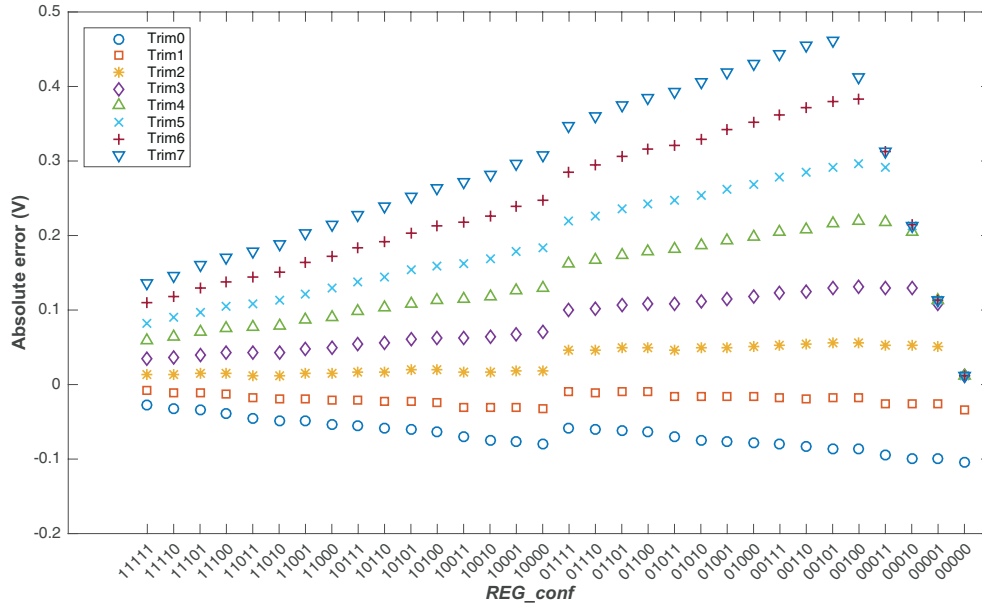


Figure 4.5: Measured absolute error of the V_Y regulator output voltage with respect to the desired value as a function of signal `REG_conf` obtained with the eight available trimming configurations (i.e., `Trim i` with $i = 0$ to 7), $V_{pp} = 5.5$ V, and $T = 24$ °C.

regulator, is `Trim1`.

The relative error of voltage V_Y (generated with the optimal trimming configuration, i.e. with configuration `Trim1`) with respect to the desired value, shown in Fig. 4.6, was measured at the minimum operating temperature (i.e., $T = -40$ °C), at room temperature (i.e., $T = +24$ °C), and at the maximum operating temperature (i.e., $T = +150$ °C). It is important to point out that the values of signal `REG_conf` used for these particular measurements are between the minimum configuration (i.e. 11111) and the configuration that correspond to $V_Y = 4.9$ V (i.e., 00101), since the high-voltage power supply was set equal to $V_{pp} = 5$ V. The specifications have been satisfied since the relative error of the regulated voltage is between 0 and -1% in all the measured cases.

In order to compare the performance of the chosen cascode-compensation technique (see Subsection 3.1.2.1) with respect to a standard technique, a test-chip has been provided with one macrocell that includes a voltage regulator with a nested-Miller compensated operational amplifier. The waveforms

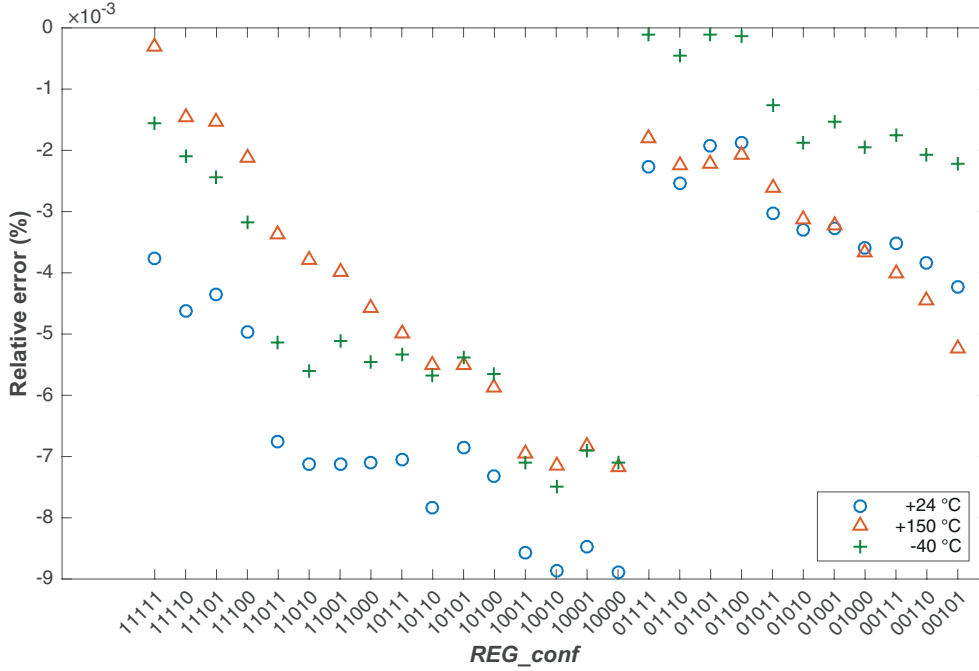
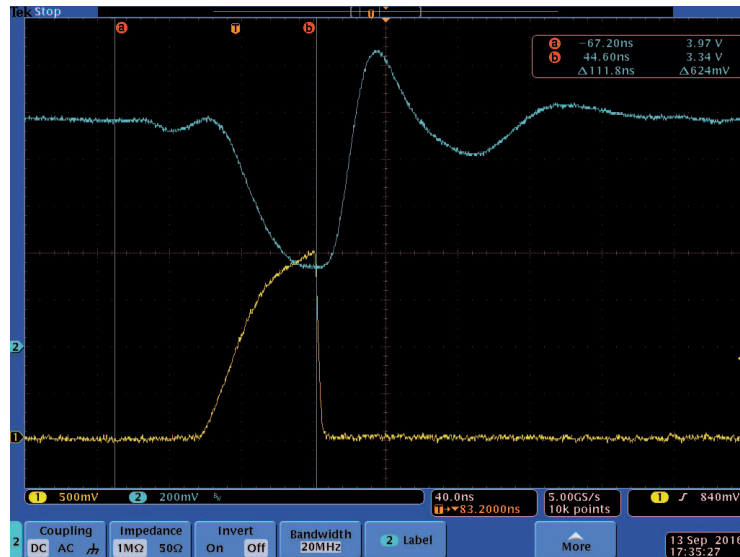
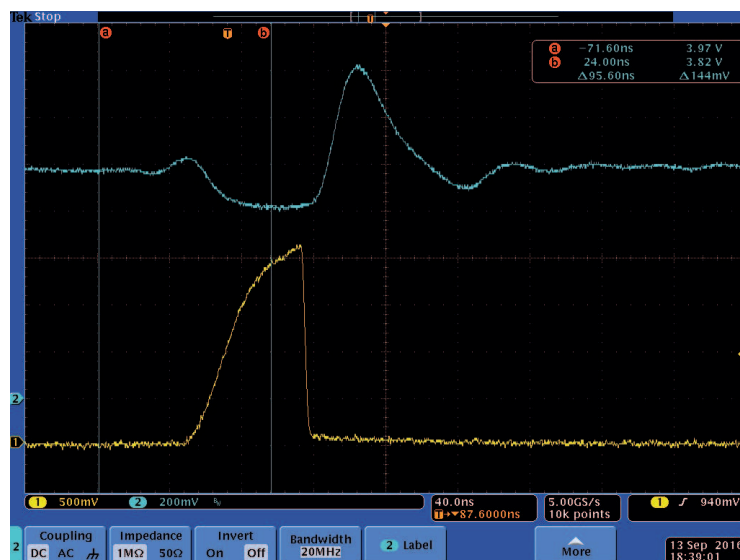


Figure 4.6: Measured relative error of the V_Y regulator output voltage with respect to the desired value as a function of signal `REG_conf` obtained with the optimum trimming configuration (i.e., Trim1) and $V_{pp} = 5$ V at three different operative temperatures: $T = -40$ °C, $T = +24$ °C, and $T = +150$ °C.

shown in Figure 4.7 were collected by means of active micro-probes and represent the output voltage of the V_Y regulator (cyan curves) and voltage BL[0] (yellow curves) measured during a write operation, i.e. during a RESET_1 programming pulse, carried out with parallelism = 33 (it is worth to remind that the voltage BL[0]) corresponds to the write pulse applied to the addressed bit-line). Voltage. The voltage V_Y corresponding to the solution that uses the cascode-compensation technique, depicted in Fig. 4.7(b), shows a 77% smaller voltage drop (≈ 144 mV) with respect to the case of the standard nested-Miller compensated operational amplifier (≈ 624 mV), represented in Fig. 4.7(a). It is important to point out that, in both cases, the shape of the RESET pulse is not compliant with the specifications. The improved current mirror, as will be shown in the following Section, overcomes this limitation.



(a)



(b)

Figure 4.7: V_Y (cyan curves) and $BL[0]$ (yellow curves) voltages measured during a $RESET_1$ programming pulse in the case of two voltage regulators that include a nested-Miller compensated (a) and a cascode-compensated (b) operational amplifier.

4.2 Improved current mirror

As mentioned above, the improved current mirror allows reducing the delay time in the generation of the required programming current with respect to the case of a standard current mirror. Figures 4.8, 4.9, 4.10, 4.11, 4.12, and 4.13 show the V_Y (cyan curves) and $BL[0]$ (yellow curves) voltages measured with active micro-probes during a $RESET_1$ programming pulse with parallelism = 1, 2, 4, 8, 16, 33, respectively. All these figures represent $RESET$ pulses with $T_{pulse} = 100$ ns, since this is the most challenging pulse to achieve due to the fact that it features the highest I_{pulse} and the shortest T_{pulse} . It is worth to point out that the pulse shapes achieve the required sharp rising and falling edges and present similar duration in all the cases independently from the chosen parallelism.

It is straightforward to notice, by comparing Fig. 4.7(b) and Fig. 4.13, that the voltage drop experienced by V_Y is increased (144 mV to be compared to 400 mV) when the improved current mirror is active due the more rapid load-current request. This effect highlights the importance to have a voltage regulator with a fast response to an abrupt change of the current load: a standard voltage regulator would show a much larger voltage drop and, hence, would not allow the correct behavior of the improved current mirror. All the pulses were collected at room temperature since, unfortunately, our experimental set-up does not allow to use, at the same time, both the ThermoStream® and the micro-probes.

To better appreciate the improvement introduced by the proposed programming circuitry, the BitLine voltages collected during a $RESET$ pulse as well as during a SET pulse (both carried out with parallelism 33) are superimposed in Fig. 4.14: the red curves correspond to programming pulses obtained with a standard current mirror and a voltage regulator that includes a nested-Miller compensated operational amplifier, whereas the blue curves correspond to write pulses acquired from a macrocell that implements both the proposed voltage regulator and the improved current mirror. The proposed programming circuitry is able to reach 80% of the required programming current in about 26 ns, whereas the standard solution takes more than T_{pulse} (i.e., ≈ 100 ns), since the measured peak value of the current corresponds to only 72% of the required $I_{plateau}$.

Finally, to prove that high-parallelism programming, enabled by the proposed circuitry, does not degrade the obtained cell-resistance distributions, 8 Mcells were measured after being written using both high- and low- parallelism programming algorithms. The measured normalized current distributions obtained by applying the user-mode programming algorithm and the

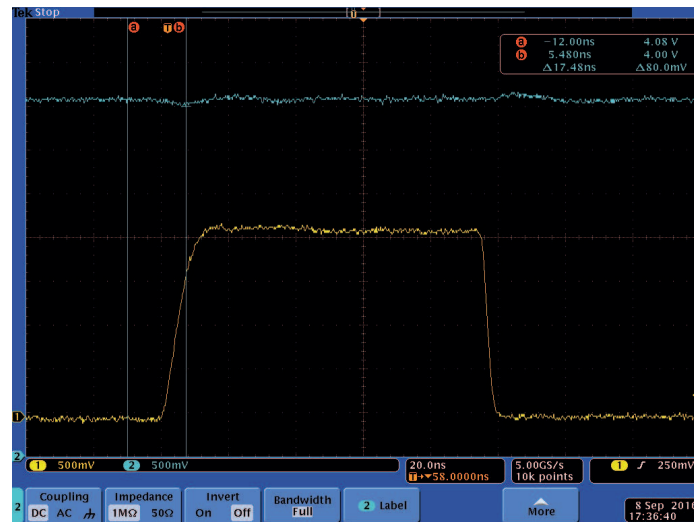


Figure 4.8: V_Y (cyan curve) and $BL[0]$ (yellow curve) voltages measured during a $RESET_1$ pulse with parallelism = 1.

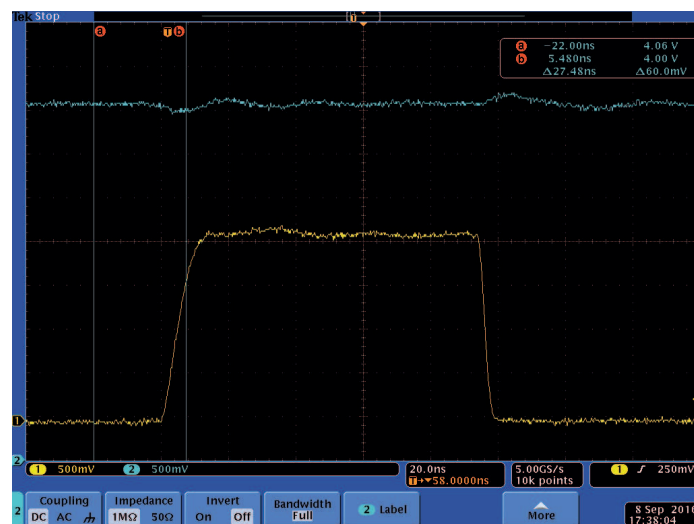


Figure 4.9: V_Y (cyan curve) and $BL[0]$ (yellow curve) voltages measured during a $RESET_1$ pulse with parallelism = 2.

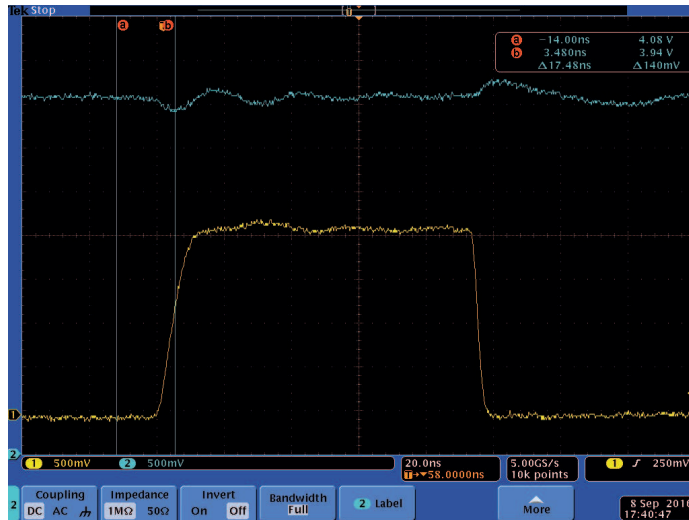


Figure 4.10: V_Y (cyan curve) and $BL[0]$ (yellow curve) voltages measured during a $RESET_1$ pulse with parallelism = 4.

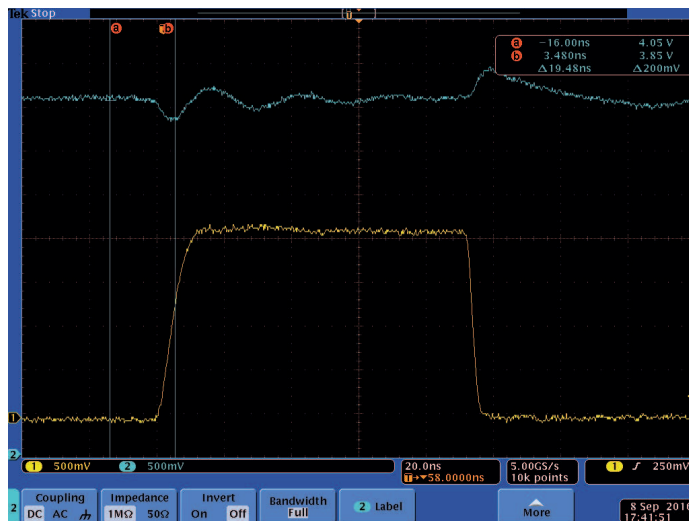


Figure 4.11: V_Y (cyan curve) and $BL[0]$ (yellow curve) voltages measured during a $RESET_1$ pulse with parallelism = 8.

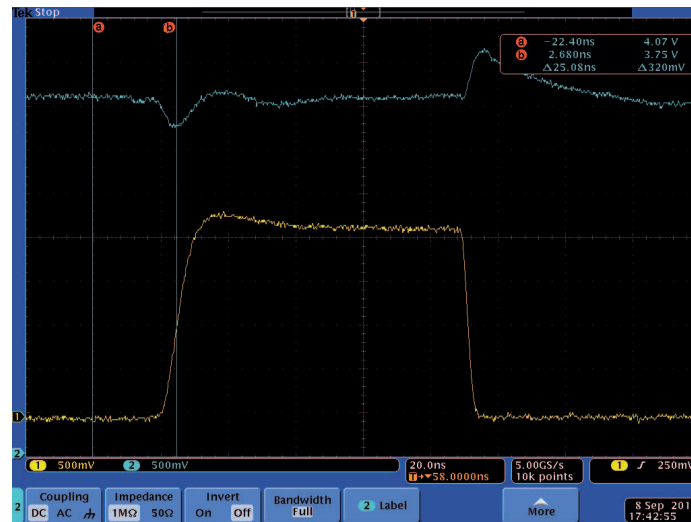


Figure 4.12: V_Y (cyan curve) and $BL[0]$ (yellow curve) voltages measured during a $RESET_1$ pulse with parallelism = 16.

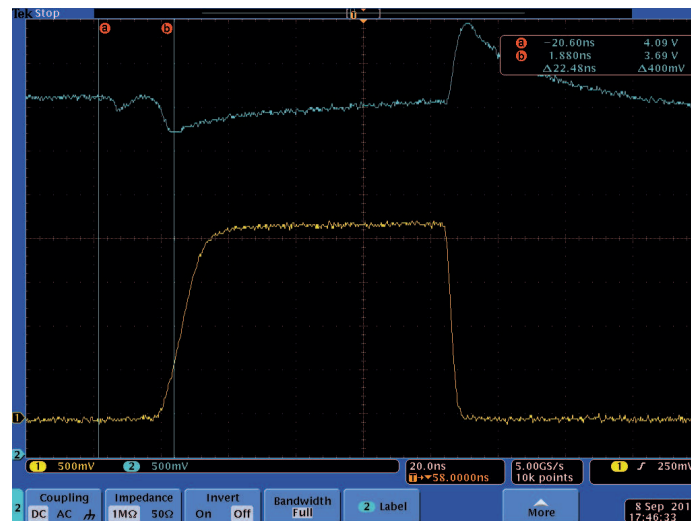


Figure 4.13: V_Y (cyan curve) and $BL[0]$ (yellow curve) voltages measured during a $RESET_1$ pulse with parallelism = 33.

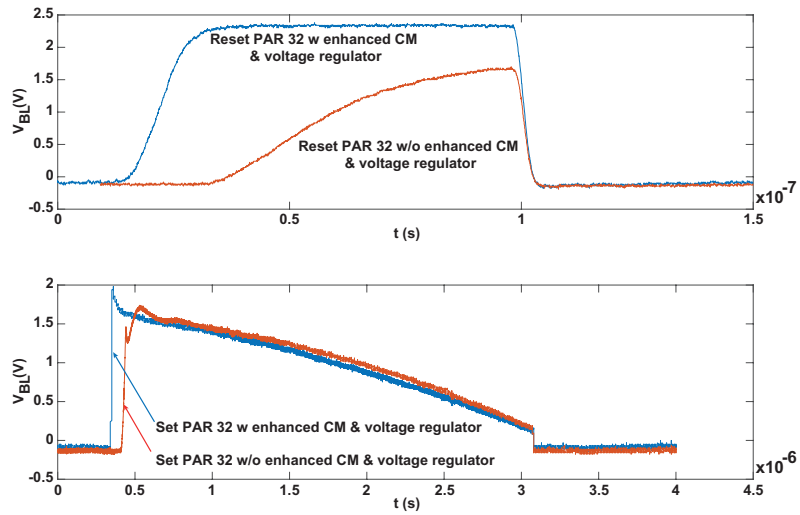


Figure 4.14: Measured $RESET_1$ (top) and SET_1 (bottom) pulses at high parallelism (32 PCM cells) with (blue curves) and without (red curves) the enhanced programming circuits (i.e., the proposed voltage regulator and the improved current mirror).

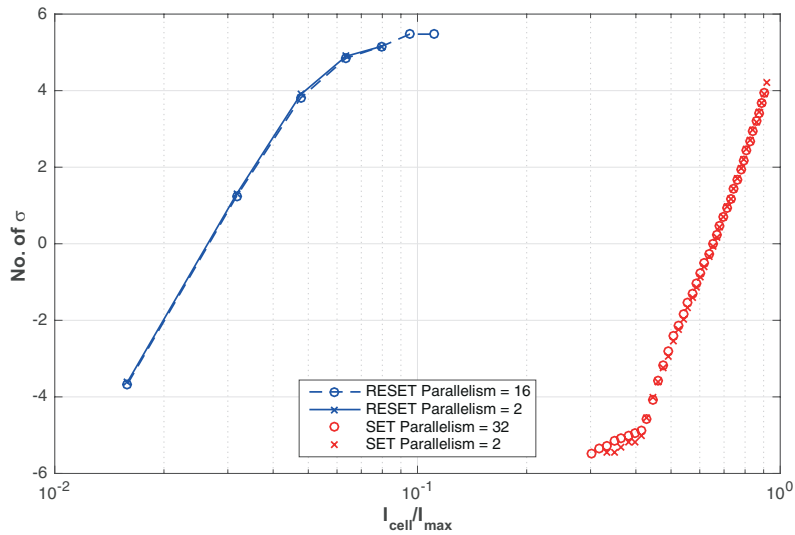


Figure 4.15: Measured normal inverse distributions of cell current after different user-mode programming algorithms carried out with high (\circ markers) and low (\times markers) parallelism. 8 Mcells per distribution.

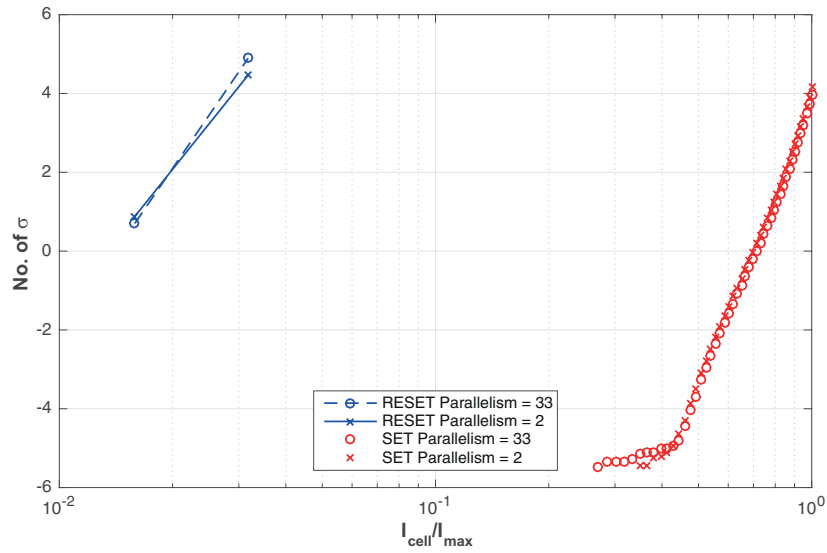


Figure 4.16: Measured normal inverse distributions of cell current after different pre-soldering programming pulses carried out with high (\circ markers) and low (\times markers) parallelism. 8 Mcells per distribution.

pre-soldering pulses are shown in Fig. 4.15 and Fig. 4.16, respectively. It is fundamental to notice that there is no significant difference in the obtained results between the high- and the low-parallelism case.

Conclusions

In this thesis, the design and the experimental characterization of the programming circuitry of a test-chip including an embedded PCM were presented. It is worth noting that the designed circuit solutions are suitable to solve limitations commonly encountered during analog integrated circuit design and, therefore, can be also included in a large variety of applications achieving the same benefit as in the case of the presented PCM test-chip.

The designed circuits include a voltage regulator implemented by means of operational amplifier that features a cascode-compensation technique and shows a faster recovery and a reduced output voltage drop when compared with a regulator that includes a conventional nested-Miller compensated operational amplifier. This was essential in order to obtain a sufficiently high bandwidth and the desired dynamic response in the case of a sudden change of the load current. The experimental results showed a 77% reduction of the output voltage drop with respect to a conventional voltage regulator in the case of a current load of about 15 mA, corresponding to 33 cells programmed in parallel.

In addition, an improved current mirror that minimizes the delay time in the generation of the programming current with respect to conventional solutions was presented. The experimental characterization of the proposed circuit showed an almost 5 times shorter recovery time with respect to a conventional current mirror. Thus, the proposed improved current mirror enabled the possibility to use a high-parallelism programming algorithm that allows improving the memory program throughput, which is a key feature in a number of applications.

Both the proposed voltage regulator and the improved current mirror were included in a 32 KB embedded-PCM macrocell and implemented in a test-chip, fabricated by STMicroelectronics to carry out experimental investigations on embedded phase change memories.

Besides the above circuits, a novel charge pump architecture, featuring an enhanced power efficiency, was analyzed and a CMOS realization was de-

signed and simulated. The simulation results of the proposed charge pump exhibit an output resistance reduced by more than 20% when compared to a conventional charge pump. The reduced output resistance directly translates into higher power efficiency, since the proposed solution does not impact on dynamic power losses.

Furthermore, an enhanced voltage-buffer compensation that can be applied to two-stage operational amplifiers was described and analyzed. The proposed solution exploits an additional voltage amplifier (with $K > 1$), placed in the compensation path, to obtain a gain-bandwidth product K times higher with respect to a conventional voltage-buffer compensation while using a K^2 smaller compensation capacitance. The simulation results of an operational amplifier based on the proposed technique (with $K = 7$) show a 7 times higher gain-bandwidth product with a compensation capacitor reduced by a factor of about 50 with respect to the case of the same operational amplifier compensated with a conventional voltage-buffer technique, in excellent agreement with theoretical results.

Finally, a design strategy aimed at optimizing the gain-bandwidth product of two-stage CMOS operational amplifiers for given specifications in terms of maximum power consumption and total silicon area was developed. This theoretical analysis provides the designer with simple equations (especially in the case of large capacitive loads) that represent the optimum sizing of transistors and compensation capacitor, as well as the optimum split ratio of the available bias current between the two operational-amplifier stages in order to achieve the maximum gain-bandwidth product.

Bibliography

- [1] F. Masuoka and H. Iizuka, “Semiconductor memory device and method for manufacturing the same,” Jul. 23 1985, uS Patent 4,531,203. [Online]. Available: <https://www.google.it/patents/US4531203>
- [2] P. E. Cottrell, R. R. Troutman, and T. H. Ning, “Hot-electron emission in n-channel igfet’s,” *IEEE Transactions on Electron Devices*, vol. 26, no. 4, pp. 520–533, Apr 1979.
- [3] P. Pavan, R. Bez, P. Olivo, and E. Zanoni, “Flash memory cells-an overview,” *Proceedings of the IEEE*, vol. 85, no. 8, pp. 1248–1271, Aug 1997.
- [4] M. Pasotti, M. Carissimi, C. Auricchio, D. Brambilla, E. Calvetti, L. Capecci, L. Croce, D. Gallinari, C. Mazzaglia, V. Rana, R. Zurla, A. Cabrini, and G. Torelli, “A 32kb 18ns random access time embedded pcm with enhanced program throughput for automotive and smart power applications,” in *ESSCIRC 2017 - 43rd IEEE European Solid State Circuits Conference*, Sept 2017, pp. 320–323.
- [5] F. Bedeschi, R. Fackenthal, C. Resta, E. M. Donze, M. Jagasivamani, E. C. Buda, F. Pellizzer, D. W. Chow, A. Cabrini, G. M. A. Calvi, R. Faravelli, A. Fantini, G. Torelli, D. Mills, R. Gastaldi, and G. Casagrande, “A bipolar-selected phase change memory featuring multi-level cell storage,” *IEEE Journal of Solid-State Circuits*, vol. 44, no. 1, pp. 217–227, Jan 2009.
- [6] F. Goodenough, “Low dropout linear regulators,” *Electronic Design*, pp. 65–67, May 1996.
- [7] R. G. H. Eschauzier, R. Hogervorst, and J. H. Huijsing, “A programmable 1.5 V CMOS class-AB operational amplifier with hybrid nested miller

- compensation for 120 dB gain and 6 MHz UGF,” *IEEE Journal of Solid-State Circuits*, vol. 29, no. 12, pp. 1497–1504, Dec 1994.
- [8] S. O. Cannizzaro, A. D. Grasso, R. Mita, G. Palumbo, and S. Pennisi, “Design procedures for three-stage CMOS OTAs with nested-miller compensation,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 54, no. 5, pp. 933–940, May 2007.
- [9] J. Huijsing, “Multi-stage amplifier with capacitive nesting for frequency compensation,” Dec. 17 1985, uS Patent 4,559,502. [Online]. Available: <https://www.google.com/patents/US4559502>
- [10] R. G. H. Eschauzier, L. P. T. Kerklaan, and J. H. Huijsing, “A 100-MHz 100-dB operational amplifier with multipath nested miller compensation structure,” *IEEE Journal of Solid-State Circuits*, vol. 27, no. 12, pp. 1709–1717, Dec 1992.
- [11] X. Peng and W. Sansen, “Transconductance with capacitances feedback compensation for multistage amplifiers,” *IEEE Journal of Solid-State Circuits*, vol. 40, no. 7, pp. 1514–1520, July 2005.
- [12] M. Pasotti, L. Capecchi, and R. Zurla, “Analog boost circuit for fast recovery of mirrored current,” Jan. 3 2017, US Patent Application No. 15/397137.
- [13] J. F. Dickson, “On-chip high-voltage generation in mnos integrated circuits using an improved voltage multiplier technique,” *IEEE Journal of Solid-State Circuits*, vol. 11, no. 3, pp. 374–378, Jun 1976.
- [14] G. van Steenwijk, K. Hoen, and H. Wallinga, “Analysis and design of a charge pump circuit for high output current applications,” in *Solid-State Circuits Conference, 1993. ESSCIRC '93. Nineteenth European*, vol. 1, Sept 1993, pp. 118–121.
- [15] A. Cabrini, L. Gobbi, and G. Torelli, “A theoretical charge transfer scheme for efficiency optimization of integrated charge pumps,” in *2014 21st IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, Dec 2014, pp. 303–306.
- [16] —, “Enhanced charge pump for ultra-low-voltage applications,” *Electronics Letters*, vol. 42, no. 9, pp. 512–514, April 2006.

- [17] R. Gariboldi and F. Pulvirenti, "A 70 m Ω intelligent high side switch with full diagnostics," in *Solid-State Circuits Conference, 1995. ESSCIRC '95. Twenty-first European*, Sept 1995, pp. 262–265.
- [18] G. Palmisano and G. Palumbo, "An optimized compensation strategy for two-stage CMOS op amps," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 42, no. 3, pp. 178–182, Mar 1995.
- [19] D. Senderowicz, D. A. Hodges, and P. R. Gray, "High-performance NMOS operational amplifier," *IEEE Journal of Solid-State Circuits*, vol. 13, no. 6, pp. 760–766, Dec 1978.
- [20] B. K. Ahuja, "An improved frequency compensation technique for CMOS operational amplifiers," *IEEE Journal of Solid-State Circuits*, vol. 18, no. 6, pp. 629–633, Dec 1983.
- [21] Y. P. Tsividis and P. R. Gray, "An integrated NMOS operational amplifier with internal compensation," *IEEE Journal of Solid-State Circuits*, vol. 11, no. 6, pp. 748–753, Dec 1976.
- [22] R. Zurla, A. Cabrini, M. Pasotti, and G. Torelli, "Enhanced voltage buffer compensation technique for two-stage CMOS operational amplifiers," *2016 IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, pp. 121–124, Dec 2016.
- [23] M. Pasotti, R. Zurla, A. Cabrini, and G. Torelli, "System and method for a multistage operational amplifier," Jun. 24 2016, US Patent Application No. 15/192863.