

École polytechnique de Louvain

Modeling of integrated Single Photon Avalanche Diodes and design of an ultra-low-power analog readout circuit

Author: **Tom KLAUNER**
Supervisor: **Denis FLANDRE**
Readers: **David BOL, Nicolas ROISIN**
Academic year 2022–2023
Master [120] in Electrical Engineering

Abstract

Single-photon avalanche diodes (SPADs) are highly sensitive photodetectors capable of detecting light at the photon level. SPADs have a wide range of applications, including biomedical imaging techniques like fluorescence lifetime imaging microscopy (FLIM) or Positron Emission Tomography (PET) scanners. In this context, the ICTEAM (UCLouvain) has developed a thin ultraviolet-enhanced backside-illuminated single-photon avalanche diode. The device requires the integration of a dedicated CMOS interface circuit to read and process the signal. The design of the front-end circuitry heavily depends on the characteristics of the photodetector and an equivalent model of the SPAD is often required to facilitate the procedure. Hence, the objective of the thesis is twofold; modeling the SPAD behavior and designing a readout circuit. In the first step, this work proposes a novel compact SPICE model based on TCAD simulations (Silvaco) to accurately reproduce the transient response of the studied SPAD to a light pulse. In the second step, this thesis presents the design of the first stage of the readout electronics. The proposed first amplification stage is achieved by a transimpedance amplifier using a folded-cascode configuration. The design is conducted at the transistor level using the g_m/I_D methodology. The circuit is simulated and validated at the schematic level using Cadence Virtuoso. It shows a transimpedance gain of 66dB Ω , a bandwidth of 48MHz, and a power consumption of 760 μ W. It offers similar performance as the state-of-the-art sub-mW transimpedance amplifiers, where the gain is traded for bandwidth.

Acknowledgements

I would like to sincerely thank my supervisor, Prof. Denis Flandre, for his unwavering support and guidance during the completion of my thesis. Despite his busy schedule, he always found the time to provide me with valuable insights.

I am especially grateful to Nicolas Roisin for his continuous guidance and assistance. His help greatly enhanced the quality of this work.

I want to thank Marco Gonzalez for his help and patience in answering my numerous questions about designing analog circuits. I also want to thank Nicolas André for his assistance using *Silvaco* and Roselien Vercauteren for conducting and providing useful measurements.

I would like to further thank Eléonore, Jérémy, Loïc, Nicolas, and Roselien for offering me a pleasant working environment in their office during the accomplishment of my thesis.

Finally, I am grateful to my family for their unfailing support, and in particular to my brother, for all the help we have given each other throughout our years of study.

ChatGPT [1] was used to assist in the writing process.

Contents

1 State of the art	2
1.1 SPAD	2
1.1.1 Principle	2
1.1.2 Figure of merits	4
1.1.3 Equivalent model	6
1.2 SPAD interface	9
1.2.1 Front-end approach	9
1.2.2 Quenching circuit	9
1.3 Silicon Photon Multiplier (SiPM)	10
1.3.1 Principle	10
1.3.2 Equivalent model	11
1.4 SiPM front-end approaches	12
1.4.1 Preamplifier	13
1.4.2 Pulse shaper	16
1.4.3 Amplifier topology	17
1.5 Conclusion	19
2 SPAD Modeling	20
2.1 Device structure and TCAD modeling	20
2.1.1 Structure	20
2.1.2 TCAD layout	21
2.2 TCAD simulations	24
2.2.1 Mixed mode	24
2.2.2 DC simulations	24
2.2.3 AC simulations	30
2.2.4 Transient simulation	32
2.3 SPICE compact modeling	36
2.3.1 Proposed model	36
2.3.2 Response to a light pulse	38
2.3.3 Adaption to a SiPM model	39
2.3.4 Comparison with the literature	40
2.4 Conclusion	40

3 Transimpedance amplifier design	42
3.1 Structure and specs.	42
3.1.1 Front-end Architecture	42
3.1.2 Amplifier topology	43
3.1.3 Specifications	43
3.1.4 Design methodology: gm over ID	44
3.2 Folded-cascode	45
3.2.1 Equations	46
3.2.2 Design flow	48
3.2.3 Design choices	50
3.3 Bias circuit	51
3.3.1 Architecture	51
3.3.2 Design flow	52
3.3.3 DC point	53
3.4 Output buffer	55
3.4.1 Architecture	55
3.4.2 Design flow	56
3.5 Feedback structure	57
3.5.1 Feedback resistance	57
3.5.2 Feedback capacitance	58
3.6 Conclusion	59
4 Readout Implementation	60
4.1 Amplifier characterization	60
4.1.1 Final design	60
4.1.2 Frequency response	62
4.1.3 Transient response	64
4.1.4 PVT corners	65
4.1.5 Noise	67
4.1.6 Comparison with the literature	69
4.2 Modeled SiPM readout	70
4.2.1 Setup	70
4.2.2 Pulse response	71
4.3 Conclusion	74
A Appendix	77
A.1 SPAD equivalent width	77
A.2 Implantation level	78
A.3 Time constant vs. junction capacitance	78
A.4 Quenching circuit	80
A.5 Output referred noise	81
A.6 PSRR and CMRR	82

Introduction

Single-photon avalanche diodes (SPADs) have gained significant interest in various fields due to their capability to detect extremely low light levels. These photodiodes find applications in Light Detection and Ranging (LiDAR), Time-of-Flight (TOF) spectroscopy, and particularly in the biomedical field for technologies like fluorescence lifetime imaging microscopy (FLIM) or Positron Emission Tomography (PET) scanners [2]. In the latter case, scintillator crystals are used to emit light that typically ranges in the UV/near-UV peak wavelength [3]. It motivates the development of SPADs that are designed and optimized to detect photons in this wavelength range. Addressing this requirement, the ICTEAM (UCLouvain) has developed a thin ultraviolet-enhanced backside-illuminated SPAD using a complementary metal-oxide-semiconductor (CMOS) compatible technology [4].

The targeted application behind the photodetector development includes the quality control of radioactive tracers used in the biomedical field. A dedicated front-end circuit is essential for reading and processing signals from SPAD arrays. Hence, this thesis focuses on designing an integrated low-power front-end circuit for the SPAD developed by the ICTEAM. A complete knowledge of the SPAD characteristics is required to design a proper circuit. To do so, this thesis is built on the continuity of the work conducted in [5], which focused on the simulation and characterization of the photodiode.

Chapter 1 introduces the fundamental concepts behind the SPAD mechanisms and typical front-end circuits with their associated challenges.

Chapter 2 presents the characterization of the proposed SPAD using TCAD simulations. This TCAD modeling approach allows extracting the required parameters to develop an equivalent model. A novel compact SPICE model is proposed to embed the intrinsic response of the SPAD, facilitating the design of a compatible readout circuit.

Chapter 3 focuses on the design of the first amplification stage of the readout electronics. It covers the amplifier specifications and design methodology at the transistor level. It provides a comprehensive overview of the design process and the optimization of the circuit.

Chapter 4 characterizes the simulated amplification stage. It presents the performance of the circuit and emphasizes the response of the readout signal when coupled with the SPICE model of the SPAD array.

Chapter 1

State of the art

1.1 SPAD

1.1.1 Principle

A single-photon avalanche diode (SPAD) is a type of PN junction (Figure 1.1). The diode operates in reverse bias at a voltage exceeding the breakdown voltage V_{bd} , known as the Geiger mode [6]. In this mode, the depletion region within the diode is fully depleted, producing a strong electric field. Due to the high field, a single photon generating an electron-hole (e-h) pair can cause a sustainable current through the avalanche process. Hence, it enables light detection at the level of a few photons. For instance, the SPAD under investigation exhibits a probability of detecting a single incident photon of 22.31 % at 423 nm wavelength. In contrast, commercially available SPAD arrays can reach a probability of 65 % at similar wavelengths [7]. The enrichment p-layer (Figure 1.1) is heavily doped to increase the electric field further, facilitating the avalanche process. In the absence of an avalanche event, the current flowing through the system is referred to as the leakage current. This current arises due to spontaneous generation-recombination mechanisms, but its magnitude is negligible compared to the avalanche current [8]. However, these generated carriers can initiate an avalanche process in the dark, leading to false detection [9]. The SPAD operates in a bistable condition, wherein the device is either quiescent or in a state of avalanche breakdown [6]. The SPAD operation can be separated into three stages: trigger, quenching, and recharge, as depicted in Figure 1.2.

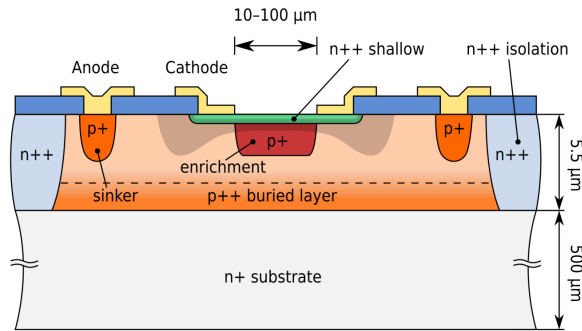


Figure 1.1: Cross-section of a typical planar SPAD structure. Reproduced from [10].

- Trigger: The SPAD is reverse-biased above V_{bd} , at a bias voltage of $V_A = V_{bd} + V_{ex}$, with V_{ex} the excess voltage. When the SPAD is triggered, the avalanche process starts, and the current rises promptly to a dynamic current level. The signal amplitude should be high enough to be read by a readout circuit.
- Quenching: The self-sustaining avalanche current has to be restrained to avoid the destruction of the device due to the high current pulse. It is also necessary to stop the avalanche process to be able to detect the next incoming photons. This operation is done using a quenching circuit. A simple solution is to use a so-called quenching resistor, in series with the SPAD. The high current flowing through the resistor causes the voltage across the SPAD to drop down to the breakdown voltage, stopping the avalanche.
- Recharge: The voltage across the SPAD has to be restored to its original bias to detect new photons.

During quenching and recharge, the SPAD is nearly insensitive to photon excitation. The time at which the SPAD cannot detect new photons is called the dead time. The total process typically ranges from hundreds of nanoseconds to a few microseconds in case of passive recharge [11].

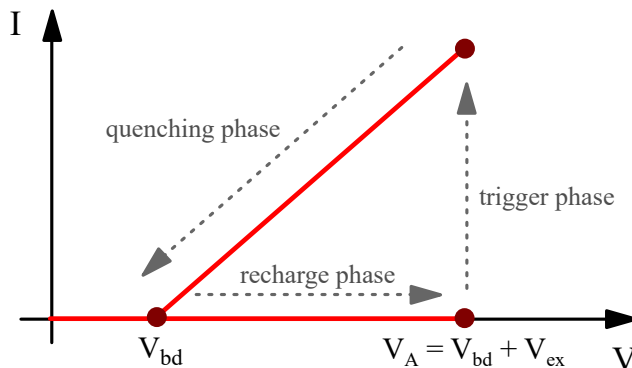


Figure 1.2: IV SPAD characteristic. Adapted from [6].

SPADs are similar to Avalanche photodiodes (APDs) as they rely on the avalanche mechanism. The primary difference between the two is their operating point. The APD operates before the breakdown voltage, while SPAD operates in Geiger mode after the breakdown (Figure 1.3). At this bias, a trigger event induces a self-sustainable avalanche that needs to be stopped externally. The output of the APD is linear to the emitting light, with a gain factor M ranging from 10 to 100. The SPAD response is rather binary, with typical gain factors ranging from $1e5$ to $1e7$ [12]. With this high gain, SPADs are more light-sensitive, allowing photon-level detection.

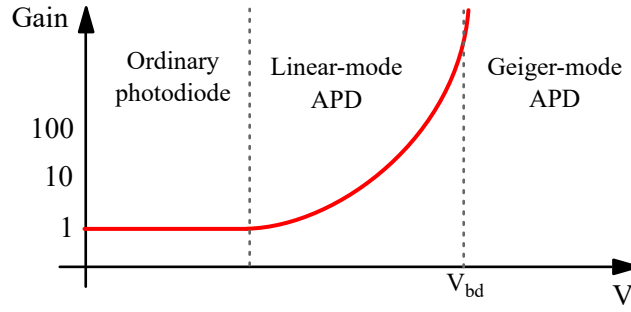


Figure 1.3: APD gain as a function of its reverse bias voltage. Adapted from [13].

1.1.2 Figure of merits

Gain factor

Due to the high electric field in the reverse-biased PN junction, a generated e-h pair can acquire enough energy to create a second e-h pair through impact ionization. This process continues as each newly generated pair gains acceleration and generates another pair, exponentially increasing the number of free carriers. This phenomenon is known as the avalanche mechanism, whereby a single generated carrier can result in a large number of free carriers. The gain factor M is the ratio of total charges flowing to the terminal for each e-h pair generated. Its rigorous expression is given for a 2D PN junction by [14]:

$$M_n = \frac{1}{1 - \int_{z_n}^{z_p} \alpha_n \exp\left(-\int_z^{z_p} (\alpha_n - \alpha_p) dz'\right) dz}, \quad (1.1)$$

$$M_p = \frac{1}{1 - \int_{z_n}^{z_p} \alpha_p \exp\left(-\int_{z_n}^z (\alpha_p - \alpha_n) dz'\right) dz},$$

with M_n representing the gain factor for electron entering at the p-side of the junction, whereas M_p refers to the gain for the holes coming at the n-side of the junction. z_n and z_p are the boundaries of the depletion layer at the n-type and p-type sides of the junction, respectively. $\alpha_{n/p}$ are the ionization coefficients for electrons and holes and represent the number of e-h pairs generated by a carrier per unit distance traveled [15]. The standard model used to describe these parameters rely on the Chynoweth model [16], which gives its dependency associated with the electric field:

$$\alpha_{n/p}(E) = \alpha_{\infty, n/p} e^{-\frac{b_{n/p}}{E}}, \quad (1.2)$$

where $\alpha_{\infty, n/p}$ and $b_{n/p}$ are found empirically. The previously defined gain factor theoretically tends to infinity when reaching avalanche breakdown because of the self-sustaining current. In the case of a SPAD, this is prevented by the quenching mechanism, which results in a finite number of charges being generated for a given event. In this situation, a less intrinsic gain factor M_{SPAD} can be computed using the previously stated definition, that is, the total charges flowing to the terminal (Q_{tot}) divided by the charge of an electron $q = 1.6 \times 10^{-19}$ C [17]:

$$M_{SPAD} = \frac{Q_{tot}}{q} = \frac{\int I_{pulse} dt}{q}, \quad (1.3)$$

with I_{pulse} the current pulse following the trigger event.

Breakdown voltage

As mentioned, the gain factor tends to infinity at the breakdown voltage. With this assumption, a condition for avalanche breakdown can be set when the denominator of Equation [1.1](#) becomes zero:

$$\begin{aligned} \int_{z_n}^{z_p} \alpha_n \exp \left(- \int_z^{z_p} (\alpha_n - \alpha_p) dz' \right) dz &= 1, \\ \int_{z_n}^{z_p} \alpha_p \exp \left(- \int_{z_n}^z (\alpha_p - \alpha_n) dz' \right) dz &= 1. \end{aligned} \quad (1.4)$$

These are the so-called ionization integrals [\[14\]](#). They state that the breakdown voltage depends on the electric field generated within the depletion region through $\alpha_{n/p}(E)$. This field is influenced by several factors, such as the SPAD geometry, the doping concentration, or the presence of traps.

Photon Detection Efficiency (PDE)

Several figures of merits built on top of each other exist to describe the efficiency of SPADs in detecting photons:

- The quantum efficiency QE is the probability that an incident photon generates an e-h pair in a region where the carriers can cause an avalanche [\[18\]](#). It depends on the light wavelength λ , as shorter wavelengths are absorbed less deeply into semiconducting materials than longer wavelengths. Hence, SPADs are designed for specific wavelength ranges to optimize detection capabilities over the required light spectra.
- The avalanche triggering probability P_T is the probability that a generated e-h pair in the depletion region initiate an avalanche breakdown. It depends on the excess voltage V_{ex} through the impact ionization coefficients $\alpha_{n/p}(E)$ [\[19\]](#). A higher electrical field leads to more accelerated carriers, thereby raising the probability of an avalanche event.
- The Photon detection probability (PDP) is the number of detected photons compared to the number of incident photons for a given active area [\[18\]](#). The Photon Detection Efficiency (PDE) is a normalized version of PDP that considers the number of incident photons on the total pixel area. It is obtained by multiplying the PDP by the fill factor (FF), which refers to the ratio of the active area compared to the total pixel size. The PDE can be expressed as:

$$PDE(V_{ex}, \lambda) = QE(\lambda) P_T(V_{ex}) FF. \quad (1.5)$$

Dark Count Rate (DCR)

The dark count refers to a false detection that occurs when an avalanche breakdown happens without a photon causing it. The dark count rate (DCR) is used to measure the frequency of this event in counts per second (cps) and can be normalized by the active area size (cps/ μm^2).

Two principal mechanisms cause dark counts; thermal generation and band-to-band tunneling. The former is mainly associated with Shockley-Read-Hall (SRH) trap-assisted thermal generation and trap-assisted tunneling (TAT) [20]. Dark counts also result from band-to-band tunneling (BTBT). Because of the high electric field, charges can travel from the valence band to the conduction band through the tunneling effect [20]. These carrier generations are depicted in Figure 1.4. The dark count rate also depends on the excess voltage, as it increases the avalanche-triggering probability P_T .

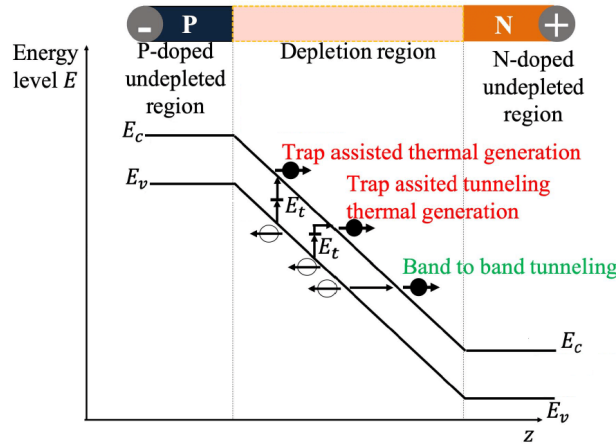


Figure 1.4: Main DCR contributions for a reverse-biased PN junction. Adapted from [20].

Afterpulsing

Afterpulsing is a case of dark count that arises just after a triggering event. During the avalanche process, some carriers become trapped at the junction and can be released afterward. As the SPAD has recharged, the released carriers can trigger a subsequent avalanche. It results in afterpulses correlated with the prior pulse [21].

1.1.3 Equivalent model

Various equivalent models have already been proposed to describe the SPAD dynamic behavior [22] [23] [24]. A classical model of the SPAD pixel is presented in Figure 1.5, which includes the diode in series with its quenching resistor. The latter is expressed by a parallel resistance R_q and a parasitic capacitance C_q . The photodiode resistance R_d models both the space charge resistance and the resistance of neutral regions crossed by the current [22]. C_d represents the junction capacitance of the diode. C_m is the parasitic capacitance between the cathode and anode of the pixel due to metal lines and bonding pads. A pulse of typically a few hundred picoseconds [22] controls the switch to simulate the photon trigger. The leakage current of the quiescent diode is not considered in the model as it is usually insignificant compared to the current generated by the avalanche [8].

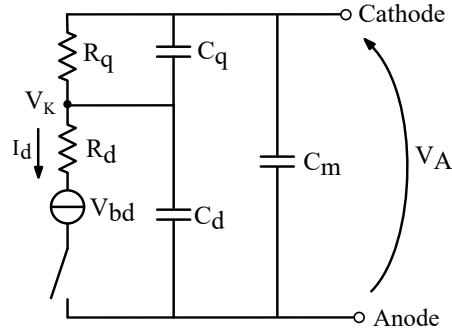


Figure 1.5: Typical pixel SPAD equivalent model.

When the switch is open, no current flows through the diode, and the voltage V_K equals the bias voltage applied between the anode and cathode (V_A). Then, the switch is closed to initiate a trigger event. The resulting high current rise induces a voltage drop across the quenching resistance, causing the diode voltage to drop down to the breakdown voltage (Figure 1.6). Because of the high current dynamic and speed, the resistance is shorted by C_q , and most of the current induced by the pulse flows through the parasitic capacitance. Hence the latter impact directly the current pulse peak amplitude I_{peak} , which is approximated by [25]:

$$I_{peak} \approx \frac{V_{ex}}{R_d(1 + C_d/C_q)}. \quad (1.6)$$

That is why in some devices, C_q is considered as a parasitic component, whereas in others, it is enhanced to increase the peaking [26]. Both the voltage drop and the induced current pulses are characterized by a time constant τ_q , given by:

$$\tau_q = R_d(C_d + C_q). \quad (1.7)$$

In practice, as the voltage V_K approaches V_{bd} , the internal current I_d decreases, resulting in fewer carriers crossing the high-field region. Since the avalanche multiplication mechanism is a statistical process, there is a probability that no new free carriers are generated by impact ionization, stopping the self-sustaining process. From [21], it is found experimentally that this probability becomes significant when the internal current I_d drops below $100 \mu A$. The switch is then reopened to model the extinction of the avalanche mechanism. The voltage V_K slowly recharges to its initial bias point, with an associated time constant τ_r , given by:

$$\tau_r = R_q(C_d + C_q). \quad (1.8)$$

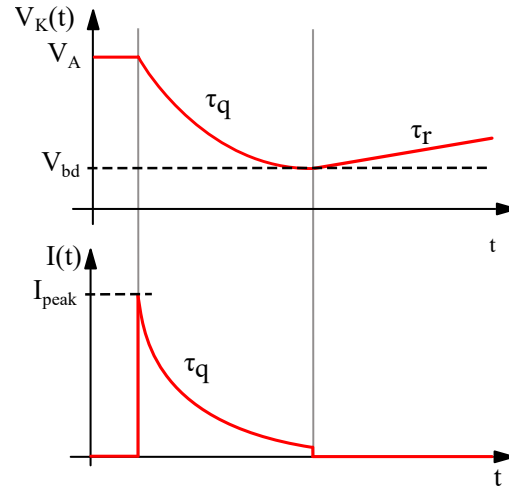


Figure 1.6: Transient response of the classical SPAD model after a trigger event. Adapted from [21].

One issue with this model is that the width of the pulse controlling the switch has to be arbitrarily set to initiate the recharge process. More advanced models [22] address this issue by incorporating a current-controlled switch (Figure 1.7) that opens the current-generating branch when the internal current I_d falls below a specified threshold (e.g., 100 μA).

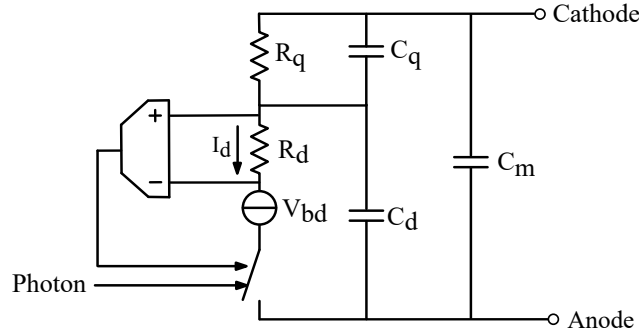


Figure 1.7: SPAD model considering the extinction of the self-sustaining current using a voltage-controlled switch. Adapted from [26].

The evaluation of the model parameters generally relies on measurements. The internal resistance R_d and breakdown voltage V_{bd} are estimated from the dark IV curve of the reverse-biased diode [22]. The slope of the breakdown curve is used to retrieve R_d , while the intercept of the extrapolated linear fit with the x-axis gives V_{bd} . The extraction of the equivalent capacitance ($C_q + C_d$) is done through the time constants τ_q and τ_r from the transient response of the SPAD pixel to a photon excitation (Figure 1.6). However, it is not possible to distinguish C_d from C_q based on the time constants. Their values are typically assessed by fitting the complex impedance of the SPAD model to the measured impedance of the reverse-biased device [27].

1.2 SPAD interface

1.2.1 Front-end approach

The typical structure of the SPAD front-end circuit is presented in Figure 1.8. The quenching and recharge are done by a resistor, which is placed either at the cathode or the anode. The signal is read out by probing the voltage drop across the resistor. Because the output of the SPAD is binary, a comparator (e.g., an inverter) is typically used [21]. Active quenching and recharge circuits can also be implemented to reduce dead time.

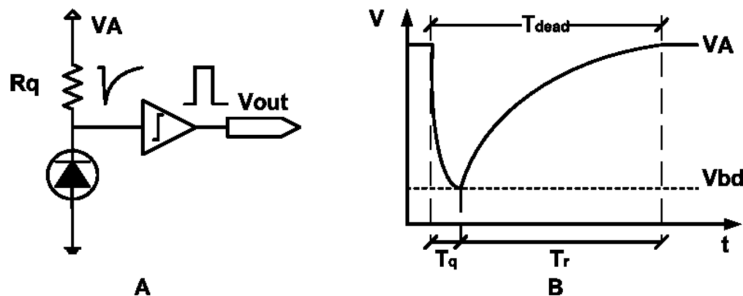


Figure 1.8: (A) Representation of a typical SPAD front-end. The quenching is passively done with a resistor. The readout is performed with an inverter. (B) Timing behavior of the diode voltage bias. Adapted from [28].

1.2.2 Quenching circuit

As mentioned, quenching and recharging are required to stop the avalanche mechanism and reset the SPAD. A simple approach to perform both operations consists of a high-value quenching resistance R_q in series with the SPAD (Figure 1.8). When the current begins to flow, the resistance causes the diode voltage to drop to the breakdown voltage V_{bd} , stopping the avalanche process. Then, the residual current crossing the quenching resistor slowly recharges the diode bias from V_{bd} to V_A . The recharge time is significantly higher than the quenching time due to the low recharge current compared to the avalanche one [28].

An active recharge circuit can reduce the dead time by short-circuiting the quenching resistance (Figure 1.9a). The trigger event is detected through the readout electronics, and R_q is shortened by closing the switch AR after a given delay through control logic. It yields a large recharge current to quickly restore the SPAD initial bias voltage. The switch is opened once the recharge is completed, and the diode can detect new photons [28].

Passive quenching circuits also have drawbacks, such as a lack of well-defined quenching time due to the stochastic process [29] or limited control over the afterpulsing. Active quenching can address both issues by eliminating the dependence of the avalanche process on the quenching time [30] and adding a hold-off step before the recharge. Its implementation is similar to passive quenching (Figure 1.9b). The switch AQ shorts the diode node to a voltage V_q ($\leq V_{bd}$) after the pulse detection to stop the avalanche process. The switch is kept closed for an extra delay T_{hold} to prevent afterpulsing (Figure 1.9b).

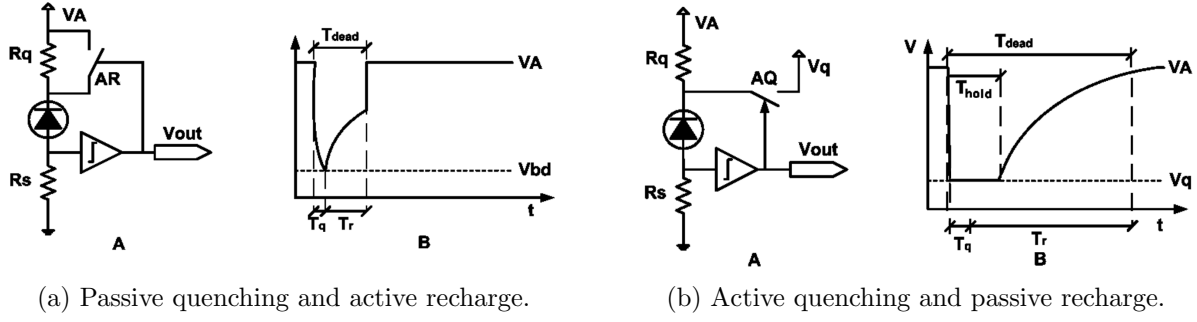


Figure 1.9: Implementation of (a) active recharge and (b) active quenching, and the associated timing behavior of the diode voltage bias. Adapted from [28].

Both circuits can be implemented simultaneously or using mixed passive and active implementations. In both cases, the readout comparator is used to control the switches. Timing control is usually achieved using monostable circuits, whereas delays, like hold-off time, are induced through RC circuits [30]. Although relying on an active circuit provides better control over the operation, it increases the complexity and area occupation of the circuit compared to the passive approach.

1.3 Silicon Photon Multiplier (SiPM)

1.3.1 Principle

A Silicon Photon Multiplier (SiPM) consists of the parallelization of several SPADs pixels, where each SPAD is associated with a quenching resistor (Figure 1.10). It can be made up of thousands of individual cells. Within the SiPM, several individual SPADs can be activated by different photons simultaneously. The SiPM can detect photons even if certain pixels experience a dead time, as the quenching and recharge processes are performed independently for each SPAD. The current generated from each fired cell is summed up. When several photons strike different pixels simultaneously, the resulting pulse becomes proportional to the number of incoming photons. Therefore, the SiPM produces an analog output current as a function of the number of incident photons (Figure 1.11). Because of the numerous SPADs in parallel, SiPMs suffer from high output capacitance, which can reach the hundreds of picofarad [31].

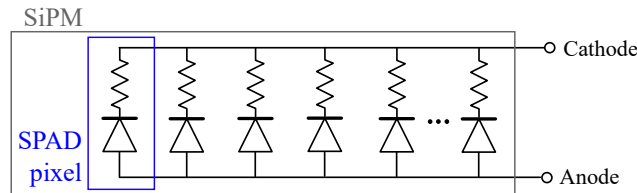


Figure 1.10: Schematic of a SiPM device, made of parallelizing multiple SPAD pixels.

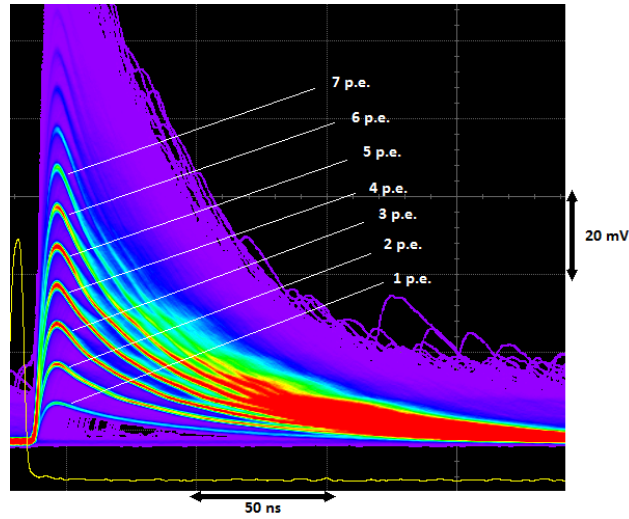


Figure 1.11: Transient response for Hamamatsu S13360-4929 SiPM with 1 to 7 simultaneously fired SPADs in response to low incoming light. Reproduced from [32].

1.3.2 Equivalent model

As depicted in Figure 1.12, two kinds of representations exist for SiPM. Either a two macrocell approach or an array of individually triggered microcells [26]. The first method consists of having in parallel one macrocell representing n simultaneously fired SPAD pixels and another cell representing $(N - n)$ quiescent ones, with a total of N microcells. This model is more suitable when it is required to simulate a large number of SPADs. However, this approach cannot distinguish which cell is fired, neither in time nor in the space domain. To deal with this, the second model simulates the whole array, where each microcell can be triggered individually. Although the computation time is longer, it can consider other effects, like mismatches among microcells.

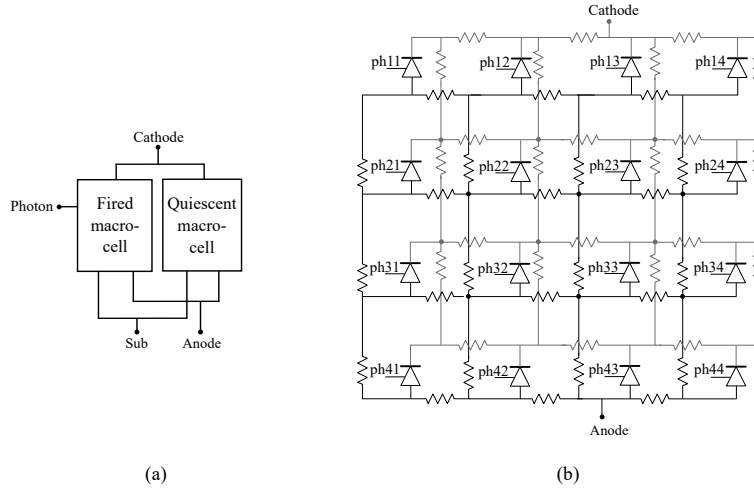


Figure 1.12: SiPM is modeled as (a) two macrocells, one for the simultaneously-fired microcells and another for those remaining quiescent, or as (b) an array of individually triggered microcells. Adapted from [26].

Because of working with typically hundreds of SPADs, the macrocell approach is often preferred. The typical model is represented in Figure 1.13. The fired macrocell relies on the same equivalent circuit as the SPAD of Section 1.1.3. To consider the effects of the numerous fired SPADs, the resistance of the circuit is divided by the number of fired SPADs n , while the capacitance is multiplied by n . The non-fired macrocell is based on the same principle: the circuit is equivalent to a quiescent photodiode with $(N - n)$ non-fired SPADs.

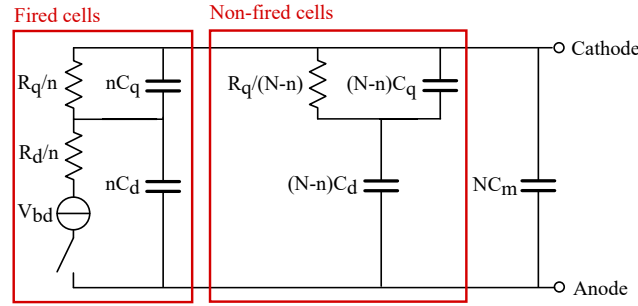


Figure 1.13: Typical SiPM model, composed of n fired cells and $(N - n)$ non-fired cells. Adapted from [26].

1.4 SiPM front-end approaches

The readout structure of a SiPM is distinct from that of a SPAD. Because of the SiPM analog output, its readout cannot be done using a comparator but must rely on an amplification chain. Three key quantities can be obtained from SiPMs: energy, arrival time, and position. Each extracted information requires an adapted readout circuit, as presented in Figure 1.14.

It is difficult to extract energy and timing/position simultaneously [33]. The energy is usually retrieved thanks to an integrator. This device is not suitable for precise timing because of its slow time response [34]. Thus, the information is often read out by two paths, a slow one for energy integration and a fast one for timing and position purposes. Timing extraction requires high accuracy. After being preamplified, the signal is fed to a fast pulse shaper. It maintains a short rising time and reduces the pulse tail. The timing information is then acquired using a comparator, as the amplitude is not required. Position information becomes useful when dealing with multiple SiPMs. The process follows a similar chain but is less strict regarding pulse width requirements. The signal is also shaped to minimize the influence of one pulse on the subsequent [33]. Once the signal is amplified and shaped, it is digitized for further processing.

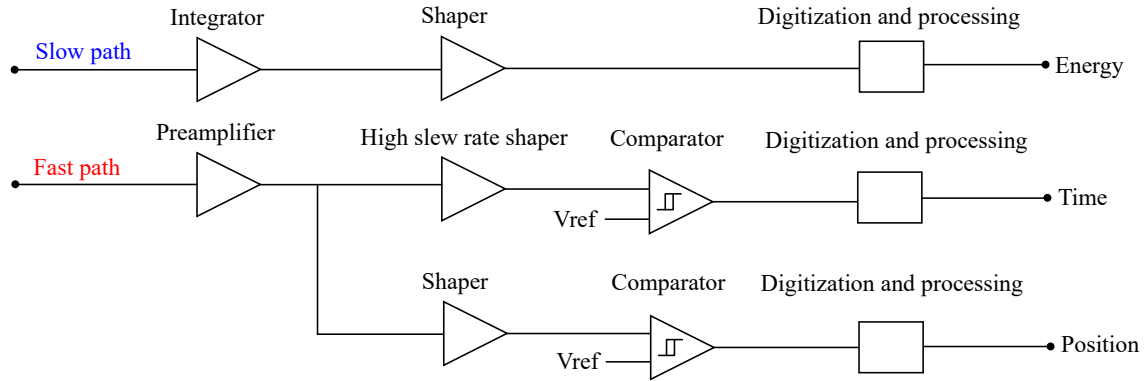


Figure 1.14: Front-end structures for SiPMs to extract either the energy, time, or position information. Adapted from [33].

1.4.1 Preamplifier

Regardless of the extracted information, a first amplification stage is required before processing the signal. The slow path is typically made of a charge-sensitive amplifier (Figure 1.15a) as it can extract the energy by integrating the signal. While the latter choice is relatively straightforward, it is not the case for the fast path. The three common architectures are the transimpedance amplifier, voltage amplifier, and current buffer. [34]. These are summarized in Figure 1.15. The characteristics of the SiPM make the choice of the most suitable preamplifier difficult [34]. For instance, it should have sufficient gain whereas exhibiting high bandwidth to detect the fast rise of current. It also requires a low input impedance to prevent time response degradation induced by the large SiPM capacitance. Finally, the amplifier should have good noise characteristics compared to the input signal.

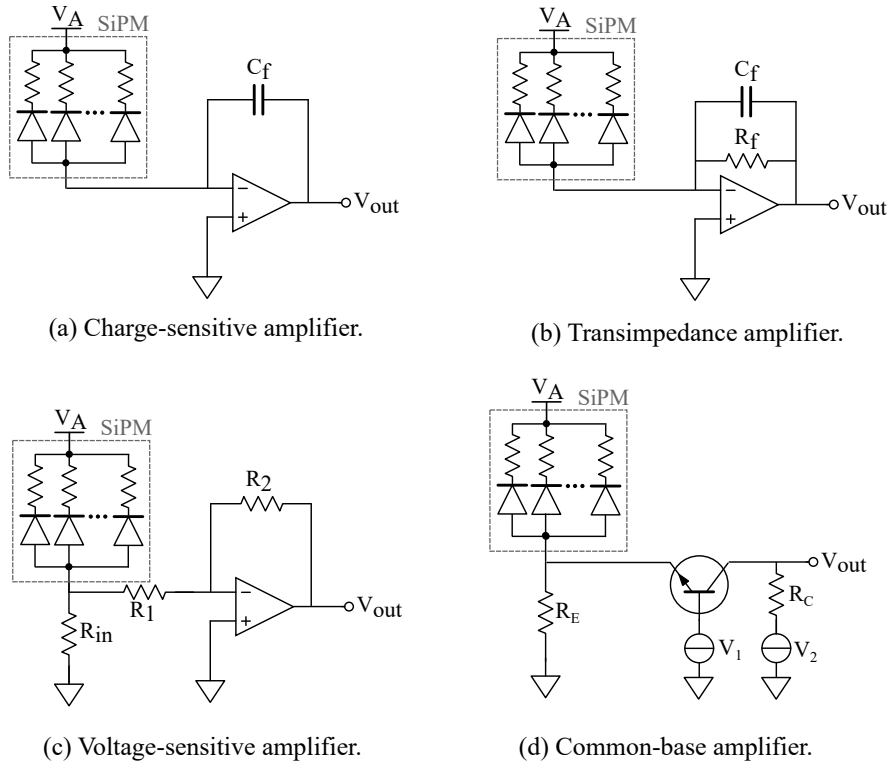


Figure 1.15: Schema of typical preamplifiers used to readout SiPM. Adapted from [33] and [35].

Charge sensitive preamplifier (CSA)

The charge-sensitive amplifier (Figure 1.15a) delivers an output voltage proportional to the integrated current through a feedback capacitor C_f . A reset circuit, consisting of a switch in parallel with C_f , is typically included to prevent saturation. An advantage of this device is the limited impact of the SiPM capacitance C_{in} using the Miller effect. Thanks to the open-loop amplifier gain A_v , the feedback capacitance appears large compared to the input one ($C_{eff} = (1 + A_v)C_f$). This device is also the best for noise performance because it relies on a capacitor [33]. However, when reading a wide array of SiPMs, the dynamic range can represent a limitation and would require a large area for the capacitor [36]. As mentioned, it also suffers from a slow response time, making it unsuitable for timing recovery. Speaking in terms of charges, the associated transfer function and input impedance are given by:

$$A_{CSA} = \frac{v_{out}}{Q_{in}} = \frac{1}{j\omega C_f + \frac{j\omega}{A_v}(C_f + C_{in})}, \quad (1.9)$$

$$Z_{in}^{CSA} = \frac{1}{j\omega C_f} \frac{1}{1 + A_v}. \quad (1.10)$$

Transimpedance preamplifier (TIA)

The transimpedance amplifier (Figure 1.15b) converts the input current into an output voltage. Thanks to the feedback structure, the input impedance is expressed as $R_{in} = R_f / (A_v + 1)$. Hence,

this device tackles the issue of the large RC delay due to the input capacitance. However, the pole associated with R_f and the SiPM capacitance could cause stability issues. To mitigate the effect, a compensation capacitance C_f can be added in parallel with R_f , at the cost of a reduction of the closed-loop bandwidth and on the timing accuracy [33]. Considering the feedback capacitance, the transfer function is given by:

$$A_{TIA} = \frac{v_{out}}{i_{in}} = \frac{-R_f}{1 + \frac{1}{A_v\beta}} \quad \text{with} \quad \beta = \frac{1 + j\omega R_f C_{in}}{1 + j\omega R_f (C_{in} + C_f)}. \quad (1.11)$$

Another advantage of the TIA compared to the other structure is the enhancement of its bandwidth, thanks to the feedback structure. Its bandwidth is approximated by:

$$BW_{TIA} = \sqrt{\frac{GBW}{2\pi R_f C_{in}}}, \quad (1.12)$$

with GBW , the gain bandwidth product of the amplifier. From the above equation, the bandwidth depends on the feedback resistance and the SiPM capacitance.

Voltage sensitive preamplifier (VA)

The voltage-sensitive amplifier (Figure 1.15c) delivers a voltage proportional to the input one. Since the SiPM delivers a current-based signal, the conversion is done by a sensing resistor R_s . Compared to the CSA, it has a better dynamic range and smaller area occupation [36]. Its main drawback is its large input impedance compared to the other designs. To limit the RC delay, the input resistance R_S should be very low (e.g., 10 Ω), which results in a low input voltage. If the SiPM gain is not large enough, it could be required to cascade multiple gain stages, reducing the timing resolution [35]. Its associated transfer function is given as:

$$A_{VA} = \frac{v_{out}}{v_{in}} = \frac{-A_v R_2}{(1 + A_v)R_1 + R_2} \quad \text{with} \quad v_{in} = \frac{i_{in} R_{in}}{1 + j\omega C_{in} R_{in} + 1}. \quad (1.13)$$

Current buffer (e.g., CBA)

The Current buffer is commonly used to avoid multiple conversions between the current and the voltage domain [36]. For instance, it is employed to copy the SiPM output current when multiple readout electronics are required [37]. It can also function as a transimpedance amplifier when combined with a sensing resistor. An example is given in Figure 1.15d with the common-based amplifier (CBA). CBA has a small input impedance (a few ohms) thanks to the absence of a feedback structure [35]. Thus, low impedance is also achieved at high frequencies. Compared to the VA, the sensing resistance at the collector can be higher (e.g., $R_s=100\Omega$) since the collector capacitance is small compared to the SiPM one. Ultimately, the signal can be further amplified with a classical voltage amplifier. The current gain and input impedance of the CBA are given by:

$$A_{CBA} = \frac{i_{out}}{i_{in}} = 1 \quad \text{if} \quad \beta \gg 1, \quad (1.14)$$

$$Z_{in}^{CBA} = r_e \left(\approx \frac{1}{g_m} \right) \quad \text{if} \quad \beta \gg 1. \quad (1.15)$$

CMOS-based alternatives to the BJT current buffer exist, as presented in [36].

The architectures and performance of SiPM readout circuits are compared in Table 1.1.

	Refs.	Structure	Input Resistance/capacitance	Gain	Bandwidth	Power
CSA	[38]	JFET op-amp	$(1 + A_v)10 \text{ pF}$	10.2 V/pC	38 MHz	-
	[39]	Folded cascode	$(1 + A_v)100 \text{ fF}$	10 V/pC	500 kHz	2 mW
	[40]	Regulated cascode + CG	$(1 + A_v)20 \text{ fF}$	37.5 V/pC	-	$37.5 \text{ }\mu\text{W}$
Current buffer	[41]	CMOS buffer + VA	$17 \text{ }\Omega \text{ (DC)}$	$40.8 \text{ dB}\Omega$	250 MHz	3.64 mW (buffer only)
	[34]	CBA + Rs	$3 \text{ }\Omega \text{ (DC)}$	- dB Ω	794 MHz	-
TIA	[31]	Regulated cascode	$200 \text{ k}\Omega / (A_v + 1)$	$106 \text{ dB}\Omega$	15.9 MHz	$660 \text{ }\mu\text{W}$
	[42]	Regulated CG	$20 \text{ k}\Omega / (A_v + 1)$	$86 \text{ dB}\Omega$	12 MHz	$336 \text{ }\mu\text{W}$
	[43]	Three stage CS	$14 \text{ k}\Omega / (A_v + 1)$	$83 \text{ dB}\Omega$	1.75 GHz	48.6 mW
VA	[44]	Schmitt Trigger	-	37 dB	30 MHz	$130 \text{ }\mu\text{W}$
	[34]	VA + Rs	$10 \text{ }\Omega$	$33 \text{ dB}\Omega$	2.4 GHz	-
	[27]	VA + Rs	$50 \text{ }\Omega$	$74 \text{ dB}\Omega$	50 MHz	-

Table 1.1: Summary of preamplifier implementations for SiPM readout.

1.4.2 Pulse shaper

A pulse shaper is often added after the preamplifier to improve the signal quality. It provides a second amplification stage, improves the signal-to-noise ratio (SNR), and shortens the signal response to minimize the pile-up effect. The latter is depicted in Figure 1.16a, where the decaying tail of one pulse impacts the amplitude of the subsequent one. A popular method for pulse shaping involves using a semi-Gaussian filter with a CR-RC_m design (Figure 1.17) [45]. This architecture involves a first-stage differentiator, acting as a high pass filter to reduce the pulse tail and noise. It is followed by m integrators, which perform a m -order low-pass filtering to improve the SNR further. A pole-zero cancellation (PZC) can be applied on the differentiator input capacitor to prevent an undershoot (Figure 1.16b) resulting from the zero induced by the capacitor. [46]. Digital alternatives of the pulse shaper are also used in practice [47].

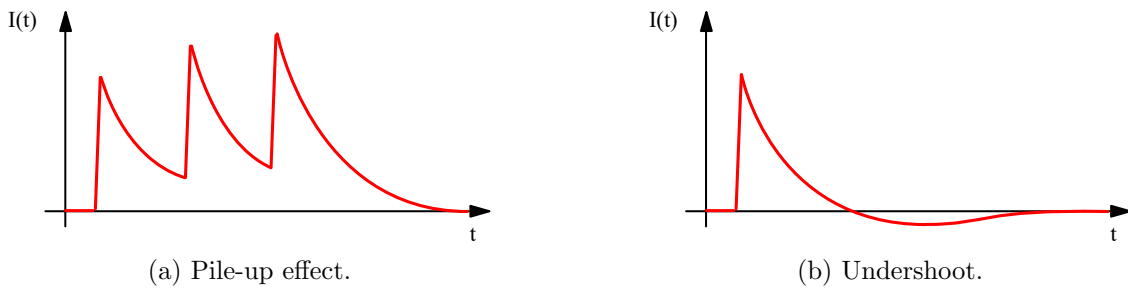


Figure 1.16: Distortion of the SiPM output signal.

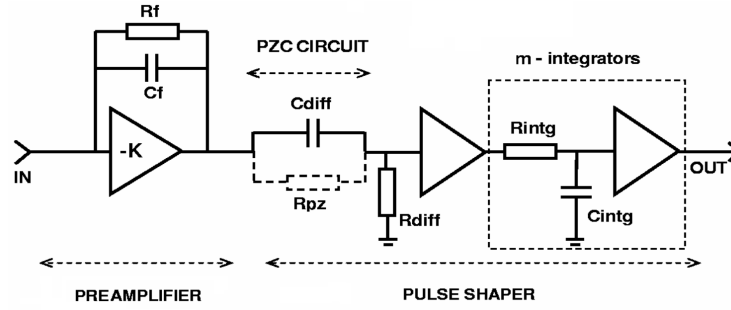


Figure 1.17: Implementation of a CR-RC_m shaping filter with an eventual pole-zero cancellation. Adapted from [46].

1.4.3 Amplifier topology

Various topologies of operational amplifiers exist to implement the amplification stage. A common implementation is based on the two-stage amplifier (Figure 1.19) or one of its associated variants [48]. These include the m -stage, folded-cascode, telescopic, and gain-boosted amplifiers, mainly done to increase the gain. Table 1.18 summarizes a comparison of these structures, showing the trade-off between their characteristics.

	Gain	Speed	Output swing	Noise	Power
Telescopic					
Folded-cascode					
m -stage					
Gain-boosted					

Figure 1.18: Comparison of performance of various operational amplifier typologies. Adapted from [49].

m -stage amplifier

The two-stage amplifier, and more generally the m -stage amplifier, consists of a first differential input stage followed by $(m - 1)$ additional amplification stages, typically common sources [50]. By cascading the amplifier stages, it is possible to achieve higher gain at the cost of more power. Because the output stage is a simple common source, the amplifier exhibits a high output swing [49]. The issue is that each additional stage introduces a zero, which can cause stability concerns that must be addressed through techniques such as zero-pole compensation. Compensation techniques reduce the bandwidth, whereas each added stage increases circuitry complexity. As a result, multi-stage amplifiers are typically limited to 2-stage or 3-stage gain [50].

Telescopic amplifier

A solution to achieve high gain without the need for cascading amplification stages is to increase the output impedance. It can be done using a telescopic amplifier (Figure 1.20). It relies on a single-stage amplifier where the output node is cascoded. The drawback of this design is the low output swing due to the number of transistors seen at the output branch [48]. This design typically does not require frequency compensation because the amplification is done with a single stage. It leads to a higher bandwidth compared to the two-stage amplifier [51].

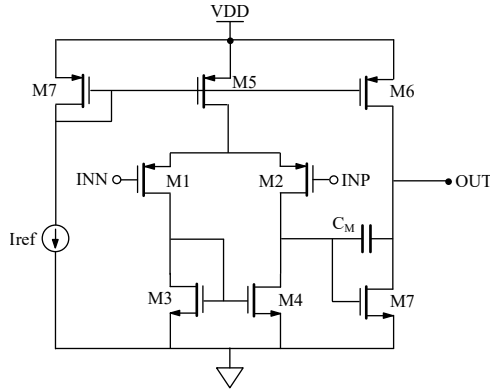


Figure 1.19: 2-stage amplifier. Adapted from [48].

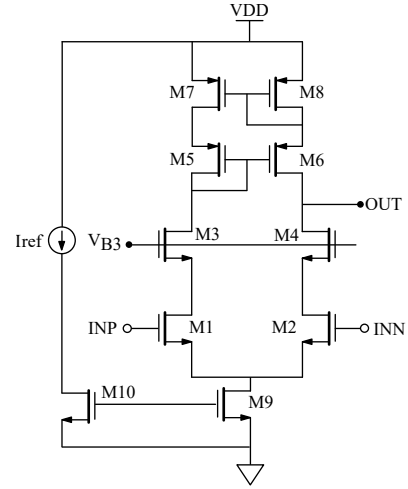


Figure 1.20: Telescopic amplifier. Adapted from [51].

Folded-cascode amplifier

The folded-cascode amplifier (Figure 1.21) operates on the same principle as the telescopic amplifier by cascoding the output. However, the design is adapted to enhance the output swing by reducing the number of transistors seen at the output [48]. It has the same advantages as the telescopic amplifier but offers a better output swing at the expense of higher power consumption [49].

Gain-boosted amplifier

The gain-boosted topology (Figure 1.22) uses the same structure as the folded cascode. However, it relies on active cascode transistors to further increase the output impedance without degrading the output swing. The active cascode configuration increases the effective transconductance of the cascode transistors, resulting in a higher output impedance [48]. Due to its active implementation, the gain increase is done at the cost of large power consumption.

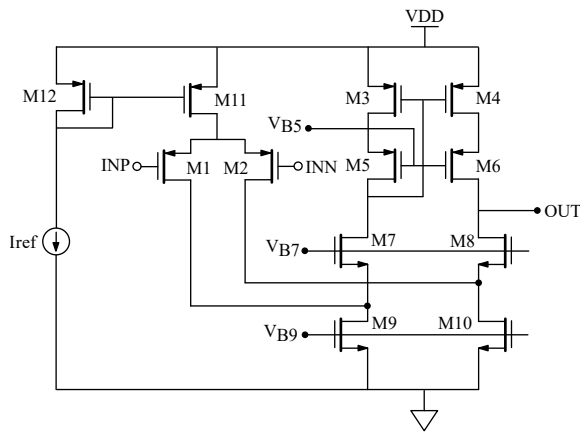


Figure 1.21: Folded-cascode amplifier.
Adapted from [48].

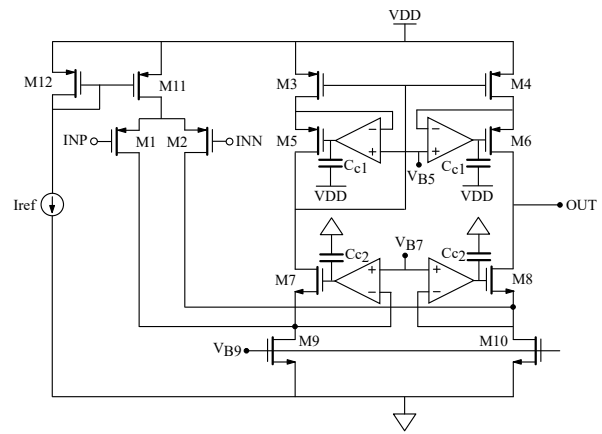


Figure 1.22: Gain-boosted amplifier. Adapted from [48].

1.5 Conclusion of the chapter

This chapter presents the key idea to understanding the mechanisms behind SPAD and SiPM with their main characteristics. It highlights the process employed to detect light at a few photon levels, relying on the avalanche breakdown. It presents the required front-end circuit to operate properly, consisting of a quenching and recharge circuit. Additionally, it introduces the standard modeling approaches for SPAD and SiPM. The models simulate the photodiode behavior to ease the design of a tailored readout circuit. They will serve as a background for a novel compact model presented in this work. It also explores typical readout circuits and their potential implementations. It emphasizes the challenges associated with the SiPM readout structure. Furthermore, it provides an overview of the methods for extracting the usual quantities of interest, such as energy, time, or position, from the incident light. It focuses on the design possibilities of the first amplification stage and suggests candidate structures for the studied SPAD array.

Chapter 2

SPAD Modeling

As discussed in Chapter 1, the classical models rely on measurements to extract their parameters, which are affected by parasitic elements or the pulse shaping of the readout circuitry. Here, the parameters are rather extracted through mixed-mode TCAD simulations of the SPAD structure. Then, an equivalent two-step SPICE model is built to tackle the intrinsic response.

The detector studied here is an ultra-thin ultraviolet backside-illuminated SPAD with 650 nm-thin silicon body fabricated in complementary metal-oxide-semiconductor (CMOS) using silicon-on-insulator (SOI) technology, based on [4]. Measurements from [5] have shown a low breakdown voltage of 11.2 V, and photon-detection probability (PDP) of 60 % under a wavelength of 423 nm at 2 V excess voltage. A SiPM based on these SPADs has also been developed and is composed of 234 SPAD pixels. The fabricated layout contains the actual SiPM, a SPAD pixel, a separate photodiode and quenching resistance, allowing individual measurements.

2.1 Device structure and TCAD modeling

2.1.1 Structure

The pixel layout is described in Figure 2.1a. It is composed of a SPAD in series with an integrated quenching resistor. The actual pixel size is $55\text{ }\mu\text{m} \times 75\text{ }\mu\text{m}$. The quenching resistor is connected to the cathode of the SPAD to control the avalanche. The resistance should be in the hundred $\text{k}\Omega$ range to quench the avalanche process [41]. Regarding the nomenclature, the electrodes known as *cathode* and *anode* refer to the terminal electrodes of either the SPAD or the pixel. For the pixel, the node between the diode and the integrated resistor is called the *diode cathode*.

The different layers of the photodiode are shown in Figure 2.1b. The silicon body is 650 nm thick, the buried oxide (BOX) is $1\text{ }\mu\text{m}$ thick, and the upper oxide layer is 200 nm thick. The anode-cathode pitch distance is $20\text{ }\mu\text{m}$. The active region is based on an n^{++}/p -well junction, where a gradient profile in the p-well generates a strong electric field at the junction, initiating an impact ionization mechanism. The guard ring is done with n-well regions to avoid a premature breakdown at the edges of the depletion region [52]. The integrated quenching resistor is done by a 350 nm thick layer of boron p-doped polysilicon (Figure 2.1b). Polysilicon is chosen due to its high resistivity to be able to quench the diode with a limited footprint.

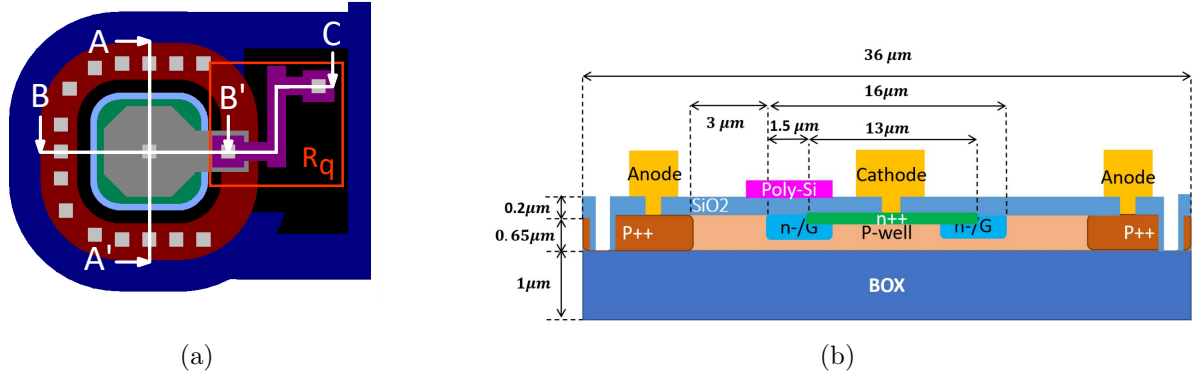


Figure 2.1: (a) Pixel layout inspired from [4]. The red square locates the integrated quenching resistor. The different cross-section lines show the section used for the 2D TCAD simulations. (b) Schematic of the physical 2D cross-section AA' of the SPAD, adapted from [5].

2.1.2 TCAD layout

The device is modeled, and its parameters are extracted using the TCAD software *Silvaco*, following the steps outlined in Figure 2.2. The structural modeling of the SPAD is performed using *Athena*, while the device simulation is carried out using either *Atlas* or mixed-mode simulation. Mixed-mode simulation allows the combination of device simulation and compact circuit models [15]. In particular, *Atlas* is used to simulate the intrinsic behavior of the SPAD, while the mixed-mode simulation is used to acquire the interaction between the SPAD and its quenching resistor. Table 2.1 summarizes the models considered in *Atlas* to simulate the device.

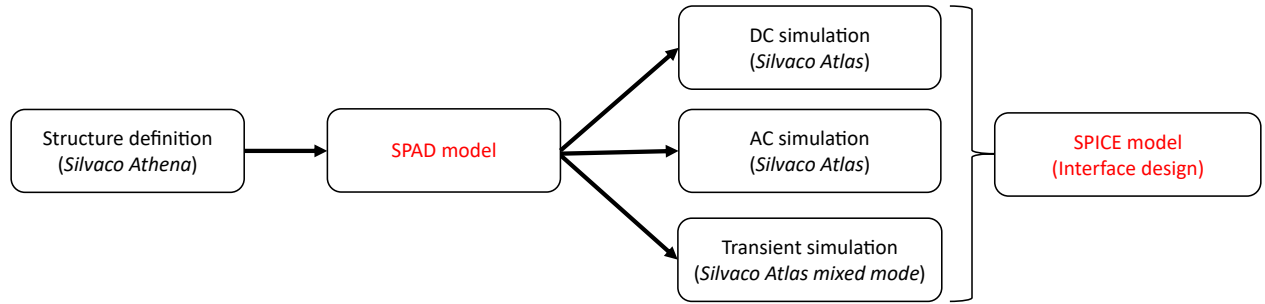


Figure 2.2: Simulation procedure used to obtain the SPAD parameters from the TCAD model.

Transport	Recombination	Carrier statistics	Mobility	Impact ionization
default (Drift-diffusion)	consrh (concentration dependent SRH)	boltzman (Boltzmann approximation)	conmob (Low-field mobility)	selb (Selberherr)
	auger (Auger)	bgn (Bandgap narrowing)	fldmob (Saturation velocity)	

Table 2.1: Main models used in *Atlas* for simulating the device. Adapted from [5].

Compared to the 3D physical SPAD pixel, the simulated diode is modeled in 2D for time efficiency. The TCAD script is based on what was done by the research team of the ICTEAM (UCLouvain) [4] and is designed with the dimensions of the actual pixel (Figure 2.1). The fabricated circular diode is not symmetrical regarding its revolution axis due to the required contact for the quenching resistor (Figure 2.1a). Hence, several cross-sections are considered to represent the different geometries of the pixel structure.

The cross-section AA' (Figure 2.3) is a two-sided anode structure of the diode and represents most of the actual SPAD geometry. According to the equivalent perimeter of the 3D actual diode (see Section A.1), its width is 35 μm . This model can retrieve the intrinsic characteristics of the diode, such as its DC and AC behaviors. However, Since the quenching resistor is not considered in this layout, it cannot account for the parasitic effects caused by its connections, nor can obtain the transient response without a quenching mechanism.

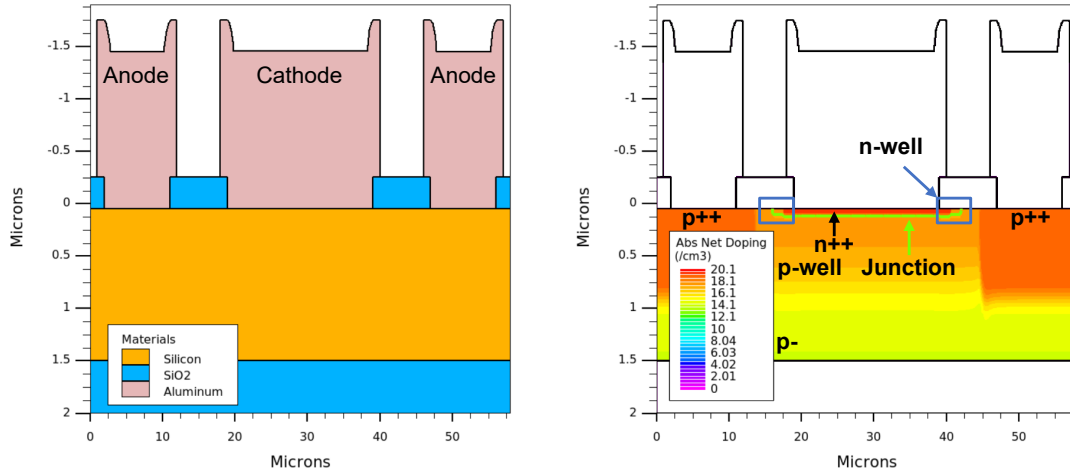


Figure 2.3: Cross section AA': (Right) 2D TCAD-simulated layout of the SPAD. (Left) related net doping $\log(/cm^3)$.

The cross-section B'C (Figure 2.4) represents the integrated polysilicon quenching resistor. Compared to the exact geometry (Figure 2.1a), the resistor is unfolded to fit in one plane. Its equivalent drain-source pitch distance is 35 μm , its width is 4 μm , and it is set atop a 200 nm-thick oxide layer. The substrate is connected to the ground from above through the gate electrode to limit the noise in the substrate that could impact the resistor characteristics. A thick layer of air is added to consider the eventual parasitic capacitance between the two electrodes from the air.

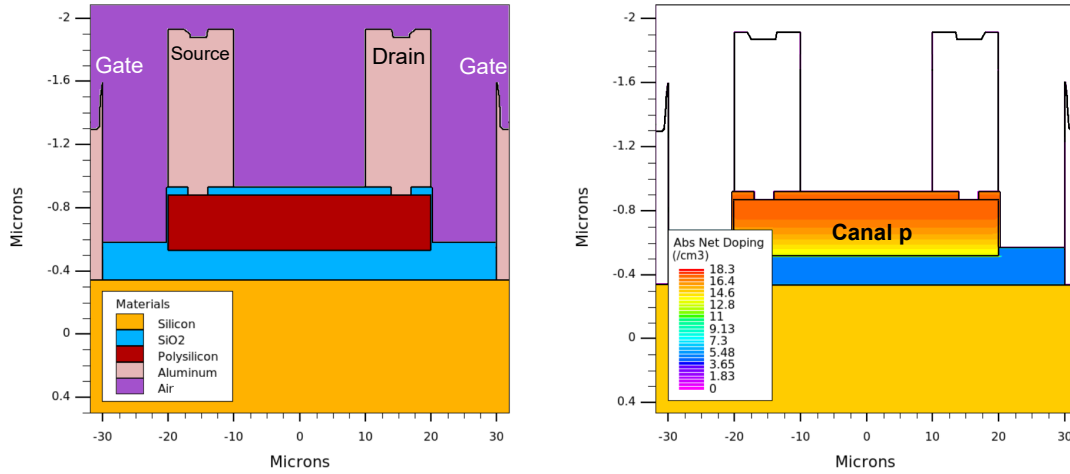


Figure 2.4: Cross section B'C: (Right) 2D TCAD-simulated layout of the integrated PSi resistor. (Left) related net doping ($/\log(\text{cm}^3)$).

The cross-section BC (Figure 2.5) combines the two precedent layouts to account for the parasitic components linked to the connection of the diode and the quenching resistor. Its width is considered the same as the resistor, that is, $4\mu\text{m}$. The diode structure is single anode-sided to include the connection with the quenching resistor in the 2D plane. This layout can retrieve the parasitic capacitance between the diode and the resistor. However, this cross-section cannot model the complete pixel due to the difference in width between the diode and resistor compared to the actual 3D geometry.

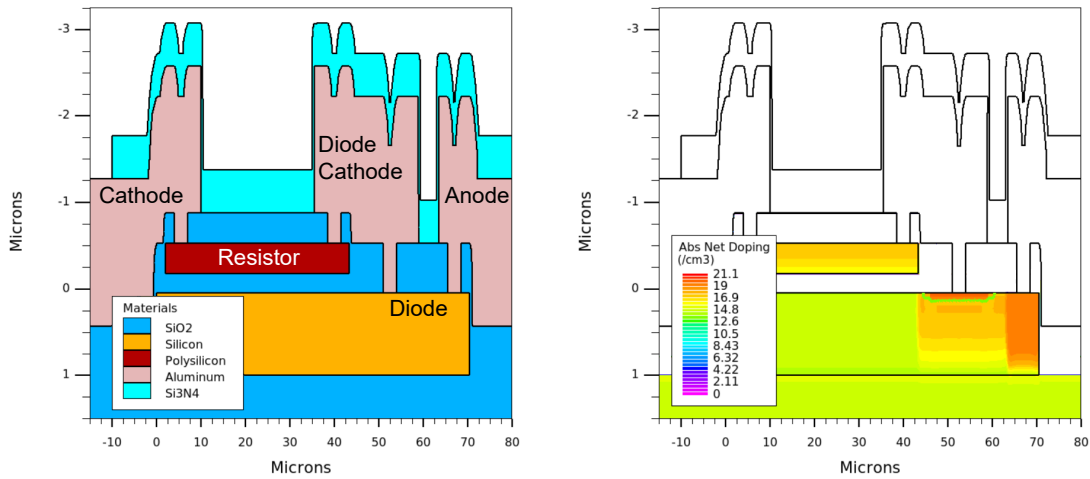


Figure 2.5: Cross section BC: (Right) 2D TCAD-simulated layout of the SPAD pixel (SPAD & quenching resistor). (Left) related net doping ($/\text{cm}^3$).

In practice, it is possible to obtain an equivalent layout that simulates the pixel structure based on cross-section BC (Figure 2.5). Its width should be twice the one of cross-section AA' (Figure 2.3) to

obtain the same equivalent width as the double-sided anode diode. It gives an equivalent perimeter of $70\text{ }\mu\text{m}$. The resistivity of the polysilicon should be enhanced by a factor of 17.5 to account for the difference in width compared to the actual resistor ($4\text{ }\mu\text{m}$). Although it enables the simulation of the integrated pixel at the layout level, this model would be less accurate in describing the actual behavior. This is due to differences in the configuration, including electrode positions and sizes, compared to the existing device. Hence a mixed-mode simulation is preferred to simulate the SPAD pixel.

2.2 TCAD simulations

2.2.1 Mixed mode

A mixed-mode model is employed to simulate the pixel structure (Figure 2.6a). The diode is integrated at the layout level (cross-section AA', Figure 2.3), while the quenching resistor is modeled either by a layout (cross-section B'C, Figure 2.5) or lumped elements. This model is used to simulate the interaction of the SPAD with its quenching resistor, enabling the retrieval of both the DC and transient characteristics of the complete pixel. One limitation of this model is that the metal capacitance between the pixel electrodes cannot be considered at the layout level. In the case of a pixel SPAD alone, this capacitance has a negligible impact on the pixel characteristics, and it will not be considered in the pixel model [41].

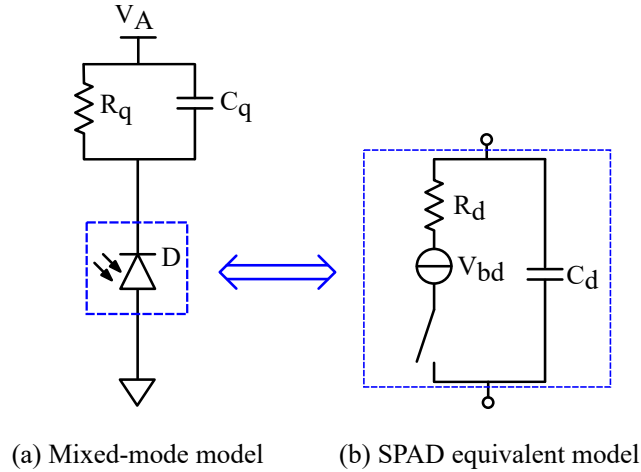


Figure 2.6: (a) Schematic of the mixed-mode circuit with the SPAD and its quenching circuit, formed by a resistor R_q and its associated capacitance C_q . (b) Equivalent circuit of the avalanche diode, as a switch and voltage generator with its internal resistance R_d and capacitance C_d

2.2.2 DC simulations

The quenching resistor is done with a thin layer of polysilicon p-doped with boron. As depicted in Figure 2.7a, the IV curve of the measured resistor shows a linear characteristic up to at least 10 V. The measured resistance is about $440\text{ k}\Omega$ by performing a least-square linear approximation of the IV curve.

The doping of the polysilicon rod is simulated through ion implantation in *Silvaco*. The statement involves two parameters: the dose and the energy of the implantation. The dose is directly related to the impurity concentration, while the energy determines the depth of the impurity diffusion in the structure. Higher doping concentration induces more free carriers, leading to a decrease in the resistivity (Figure 2.7a). Higher implantation energies result in deeper impurity diffusion, lowering the resistance (Figure 2.7b). The measured resistance is achieved with a dose of boron of $5 \times 10^{12} \text{ cm}^{-2}$ and an implantation energy of 10 keV. The simulated IV curve associated with the same parameters exhibits a resistance of 46 k Ω .

The doping concentration of the TCAD-simulated resistor can be adjusted to fit the measured resistance, but it is done at the expense of linearity (Figure 2.7a). With a dose of $4.5 \times 10^{11} \text{ cm}^{-2}$, the characteristic of the simulated resistor presents a higher resistance than the targeted value at low voltage but suffers from a large non-linear characteristic. The same conclusion can be drawn with the reduction of the implantation energy level, as depicted in Figure 2.7b. The measured resistance has a linear response up to 10 V, which is larger than its utilization range (since the voltage drop is limited to the excess voltage, which is a few volts). It is appropriate to consider the quenching resistor as an ideal one. As depicted in Figure 2.7, a lumped resistance of 440 k Ω fits closely with the measured one.

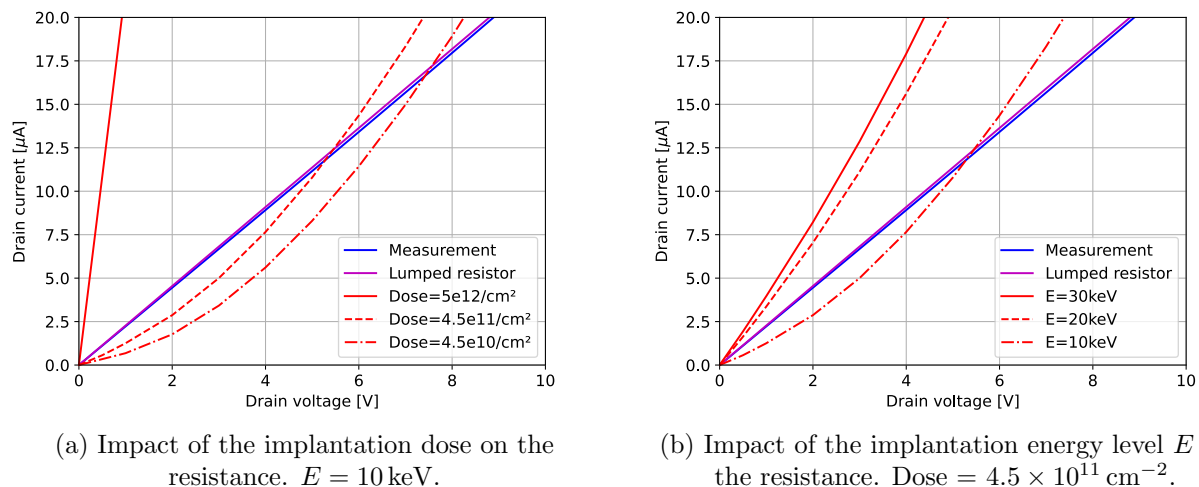


Figure 2.7: Comparison of the IV curves of the measured PSi resistor (blue) with TCAD simulations (red). In magenta, an ideal resistance of 440 k Ω .

The difference in the resistivity between the simulation and the measurement can be attributed to the use of default silicon parameters for the polysilicon in *Silvaco*. In practice, the polysilicon has grain boundaries that affect the properties of the material, including its mobility [53]. The resistivity ρ can be expressed for a p-doped material as:

$$\rho = \frac{1}{q\mu N_A}, \quad (2.1)$$

where q is the electron charge, μ is the overall mobility and N_A the number of acceptors. Using Drude model [54] and under the assumption that the different mechanisms that impact the mobility are independent, the overall mobility μ_i is given by:

$$\frac{1}{\mu_i} = \sum_{k=1}^K \frac{1}{\mu_{i,k}} = \frac{1}{\mu_{i,lattice}} + \frac{1}{\mu_{i,impurity}} + \dots \quad (2.2)$$

with i either refers to electron (n) or holes (p) and where K independent mechanisms are considered. The mobility model used in *Silvaco* depends on the doping concentration and accounts for the different mobility contributions. Its dependency is provided by empirical data only for silicon at $T = 300$ K [15]. The impurity mobility decreases with doping as an increase in the impurity concentration in the material result in a higher collision probability between carriers and defects. Compared to its monocrystalline counterpart, polysilicon is composed of multiple silicon crystals. It further reduces $\mu_{i,impurity}$ due to grain boundaries. For instance, the electron mobility in *Silvaco* is set to $1000 \text{ cm}^2/\text{Vs}$ at a doping concentration of $1 \times 10^{15} \text{ cm}^{-3}$. As presented in Figure 2.8, polysilicon mobility is much slower than pure silicon. In the case of bulk polysilicon, its value is around $300 \text{ cm}^2/\text{Vs}$, considering the same doping concentration.

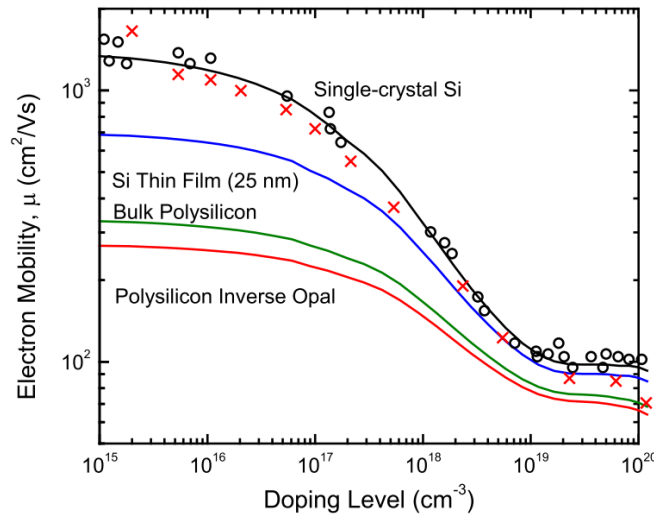
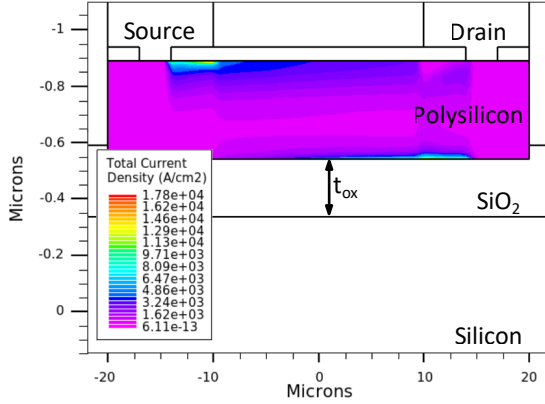
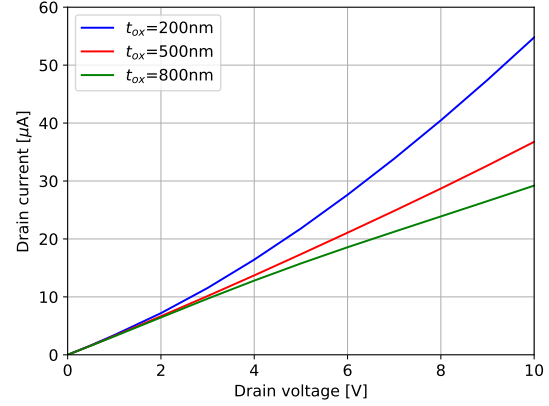


Figure 2.8: Electron mobility in bulk and thin-film single-crystal silicon, bulk polysilicon, and polysilicon inverse opals at 300 K. Reproduced from [53].

The model employed in *Silvaco* states a linear dependence of the mobility regarding the voltage because of the low-field assumption. The non-linearity of the response is associated with the parasitic capacitance of the oxide layer. It can be explained by a phenomenon similar to what is observed in a MOS transistor. Increasing the bias voltage induces an accumulation of charges near the bottom oxide. As seen in Figure 2.9a, it reduces the resistance locally, creating a secondary channel in parallel with the first one. A simple way to limit the non-linearity is to increase the oxide thickness t_{ox} (Figure 2.9b), as it reduces the parasitic capacitance. With $t_{ox} = 500 \text{ nm}$, a dose of $4.5 \times 10^{11} \text{ cm}^{-2}$, and an implantation energy of 20 keV, the resistor shows a linear IV curve up to 10 V. The associated resistance is about 370 kΩ, which is close to the target resistance of 440 kΩ.



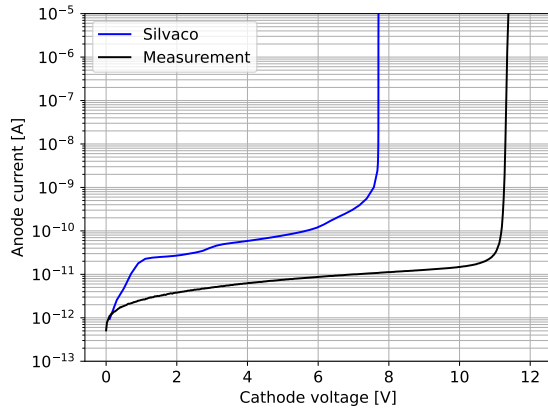
(a) Current density in the PSi resistor, with a drain voltage of 5 V. $t_{ox} = 200$ nm.



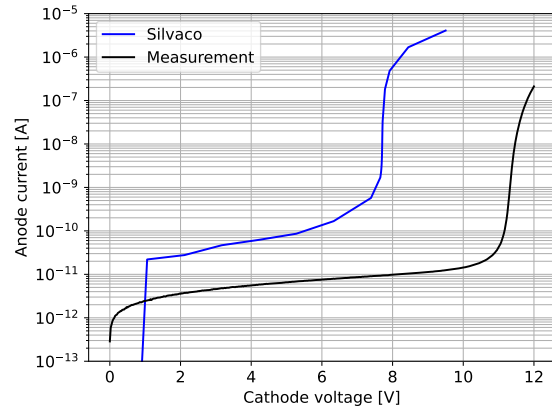
(b) PSi resistor IV curve for several oxide thicknesses.

Figure 2.9: Impact of the oxide thickness t_{ox} on the PSi resistor characteristics.
Dose = $4.5 \times 10^{11} \text{ cm}^{-2}$, $E = 20 \text{ keV}$.

The TCAD-simulated IV curve of the SPAD and pixel are presented in Figure 2.10 and compared to the measured curves. The latter are obtained with a probe PM8PS and a b1500 semiconductor analyzer. The estimations of the internal resistance R_d and breakdown voltage V_{bd} are performed through least-square linear approximation of the diode breakdown dark IV curve (Figure 2.10a) [22]. The slope of the curve after breakdown is used to retrieve R_d , while the intercept of the extrapolated linear fit with the x-axis gives V_{bd} . Similarly, the quenching resistance can be estimated from the slope of the dark IV curve of the SPAD pixel (Figure 2.10b).



(a) SPAD IV curve.



(b) SPAD pixel (SPAD and quenching resistor) IV curve.

Figure 2.10: Comparison of the log-scale dark DC IV curve of the TCAD-simulated SPAD and measurements.

The simulated breakdown voltage and internal resistance are about 7.7 V and 142 Ω , respectively. The measurements lead to values around 11.3 V for the breakdown voltage and about 1 k Ω for the internal resistance. However, the data available for the IV curve are restricted by the current compliance applied during measurements (50 μ A). Reaching a higher current level could provide a more accurate estimation of the actual internal resistance. The disparity between simulations and measurements, such as in terms of the breakdown voltage or dark current, can be attributed to various factors, including differences in geometry, doping profile, trap concentration, and default parameters set by *Silvaco*. However, the transient response of the device in Geiger mode rather depends on the excess voltage than on the breakdown voltage [26]. Hence, the subsequent comparison will be conducted considering similar excess voltages between the model and the actual diode. From [4], the excess voltage considered for the studied device is around 3 V.

The extracted quenching resistance from the SPAD pixel (Figure 2.10b) is about 1.7 M Ω , which differs from the separately measured value of about 440 k Ω . The measured IV curve is here limited to a bias voltage of 12 V, and reaching a higher voltage could lead to better estimations. Because the SPAD operates up to $V_{bd} + V_{ex}$, the IV characteristics should be taken up to this value to obtain a representative first-order estimation of the SPAD behavior in operation. The extracted values are summarized in Table 2.2.

	V_{bd} [V]	R_d [Ω]	R_q [Ω]
Measurement	11.36	1k	1.7M
Simulation	7.7	142	440k

Table 2.2: Static parameter extraction from the SPAD dark IV curve.

As Figure 2.10a shows, the simulated dark current is one order of magnitude higher than the measured one and presents a steeper slope. The dark current I_{dark} is given by [55]:

$$I_{dark} = I_{ds} + MI_{dg}, \quad (2.3)$$

with I_{ds} the surface leakage current and I_{dg} the current generated inside the substrate. The latter results in random carrier generation within the depletion region, amplified by the gain factor M . Because of this multiplication, it is usually the primary component of the dark current [55]. Multiple mechanisms contribute to I_{dg} , including thermally driven SRH generation and band-to-band tunneling through band gap narrowing [56]. The indirect thermal generation rate of non-equilibrium carriers due to traps G_{SRH} can be expressed as [57]:

$$G_{SRH} = \frac{pn - n_i^2}{\tau_p \left[n + n_i e^{\frac{E_{trap}}{kT_L}} \right] + \tau_n \left[n + n_i e^{\frac{-E_{trap}}{kT_L}} \right]}, \quad (2.4)$$

with E_{trap} the difference between the trap energy and intrinsic Fermi levels. Its default value in *Silvaco* is 0 eV. T_L is the lattice temperature and k the Boltzmann's constant. τ_n and τ_p are expressed as [15]:

$$\tau_n = \frac{\tau_{n0}}{1 + \frac{N}{N_{SRHn}}} \quad \text{and} \quad \tau_p = \frac{\tau_{p0}}{1 + \frac{N}{N_{SRHp}}}, \quad (2.5)$$

where N is the impurity concentration and $N_{SRH_{n/p}}$ a concentration parameter set by default to $1 \times 10^{15} \text{ cm}^{-3}$. $\tau_{n0/p0}$ are the intrinsic electron and hole lifetimes in a steady state. Typical values range from $1 \times 10^{-7} \text{ s}$ to $1 \times 10^{-5} \text{ s}$ in the case of TCAD SPAD models [58][9]. As done in [9], the lifetime can be adjusted in the model to fit the measured dark current levels. Figure 2.11 shows the impact of the carrier lifetime on the dark current. Higher values reduce the SRH generation and, by extension, the dark current. It also depends on the temperature, as higher values increase the carrier energy and thus facilitate their generation through SRH. It is worth noting that setting E_{trap} to 0 eV corresponds to the worst-case scenario, and thus the dark current level can also be reduced by setting a more representative value.

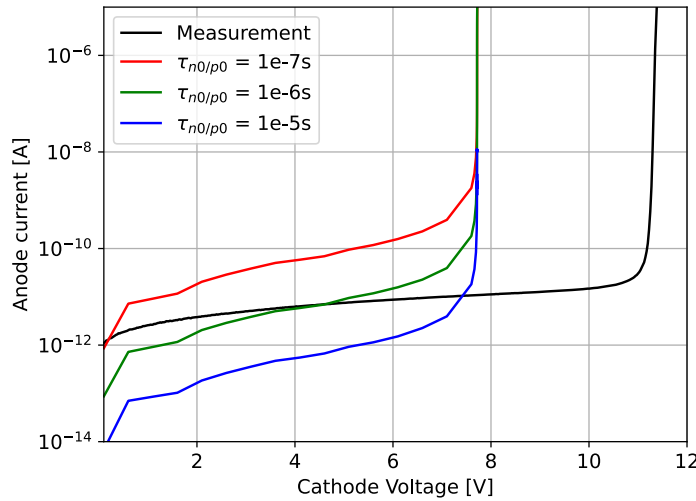


Figure 2.11: Impact of the SRH lifetime on the dark current compared to the measurement.

The steeper slope of the simulated dark current can be associated with two mechanisms: the band-to-band tunneling and the ionization coefficient [59]. The generation rate associated to the band-to-band tunneling G_{BTBT} is nearly insensitive to temperature [9] and is implemented by the Hurkx tunneling model as [60]:

$$G_{BTBT} = DA_{BB}E^{\Gamma_{BB}}e^{-B_{BB}/E}, \quad (2.6)$$

with E the electric field, D a statistical factor and A_{BB} , B_{BB} , Γ_{BB} model parameters with default values set in *Silvaco* [15]. Increasing the bias voltage narrows the bandgap, promoting generation through tunneling. The slope of the dark IV curve is also associated with the ionization coefficient as it sets the gain factor and breakdown voltage. Because the ionization coefficient depends on the electric field within the depletion region, the p-well doping concentration and the n+/p-well grading coefficient can be tuned to match the measurements. In particular, they can be adjusted to obtain a similar breakdown voltage, as done in [61].

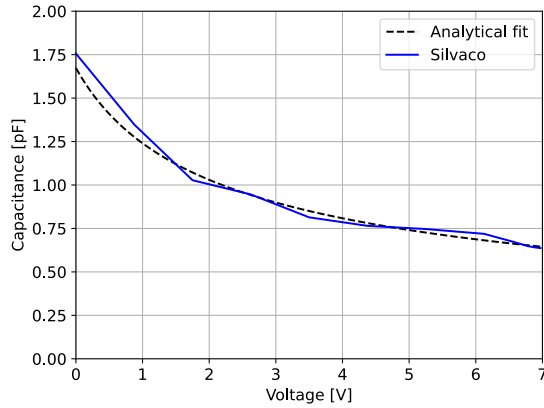
2.2.3 AC simulations

It is possible to obtain the different capacitances of the SPAD pixel by performing small-signal AC simulations. As outlined in Section 1.1.3, the SPAD pixel includes three capacitances: the diode junction capacitance C_d , the quenching resistor capacitance C_q , and the metal capacitance C_m . The main contribution of the junction capacitance is the depletion capacitance of the PN junction, which depends on the bias voltage. For a reverse-biased step-graded PN junction, its value can be approximated analytically by [62]:

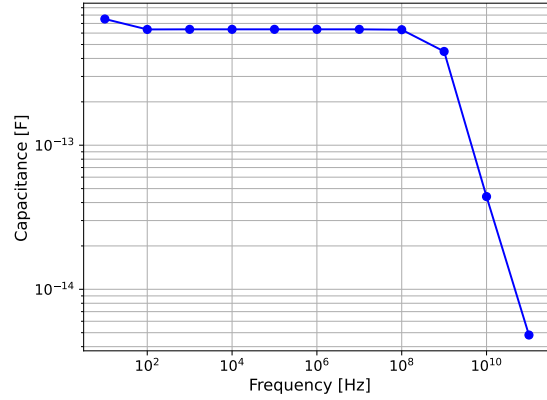
$$C_d = \frac{A\epsilon_s}{W} = A\sqrt{\frac{q\epsilon_s N_d}{2(V_{bi} - V)}} \quad \text{where} \quad V_{bi} = \frac{kT}{q} \ln\left(\frac{N_D N_A}{n_i^2}\right), \quad (2.7)$$

with A the effective area of the junction, ϵ_s the permittivity, W the depletion width and V_{bi} the built-in potential. Figure 2.12a shows the TCAD-simulated junction capacitance before breakdown. It is not possible to obtain the capacitance value through AC simulation at biases beyond the breakdown voltage due to the avalanche phenomenon that occurs. Compared to the assumption done in the analytical expression, the simulated PN junction is more complex than a 1D step-graded PN junction because of its 2D structure and gradual doping. However, the simulated capacitance decreases in the square root of the applied reversed voltage, as shown by comparison with a least-squared linear approximation of $C_d^{-2}(V)$ (Figure 2.12a). Because the SPAD operates in Geiger mode, the capacitance value should be obtained close to its operating point. The capacitance is extracted at 7 V, as its impact of bias on the capacitance is limited around the breakdown voltage (Figure 2.12a, left). At a frequency of 100 kHz, C_d is about 0.63 pF.

In addition to a voltage dependency, small-signal TCAD AC simulation shows in Figure 2.12b a frequency dependency of the junction capacitance, taken just before the breakdown voltage. At low and moderate frequencies, the capacitance is independent of the frequency. At higher frequencies, the minority carriers cannot respond to the rapidly varying voltage across the device due to their finite mobility and recombination lifetime. As a result, the device capacitance drops to a low value. The frequency at which the capacitance drops is associated with the cut-off frequency of the diode in the GHz range [9].



(a) Junction capacitance C_d at 100 kHz and its related least-square linear fit vs. the reverse bias voltage, before the breakdown.



(b) Junction capacitance C_d vs the frequency. The value is taken at a biased voltage just before the breakdown.

Figure 2.12: Extraction of the junction capacitance C_d using TCAD simulation.

An AC simulation of the TCAD resistor layout can be conducted to determine the quenching capacitance C_q . However, the simulation yields a negative capacitance. It can be attributed to the effect of the charges present in the substrate on the dynamic within the polysilicon, as already observed in the IV curves. Hence, the quenching capacitance is rather retrieved from the literature. Considering quenching resistors in the hundreds of $\text{k}\Omega$, their associated capacitance is typically in the femtofarad range [63][27][64]. Figure 2.13 presents the parametrization of the transient response to a light pulse for several C_q . It is done using the simple SPICE SPAD model presented in Chapter 1 (Figure 1.5). The associated parameters corresponds to the up-to-now values extracted from the TCAD simulations, that is, $R_d = 142 \Omega$, $R_q = 440 \text{ k}\Omega$, $V_{bd} = 7.7 \text{ V}$, $V_{ex} = 3 \text{ V}$, $C_d = 0.63 \text{ pF}$. As seen, larger capacitance induces a higher peak and slower decay. The quenching capacitance of this model is set to a typical value of 6 fF.

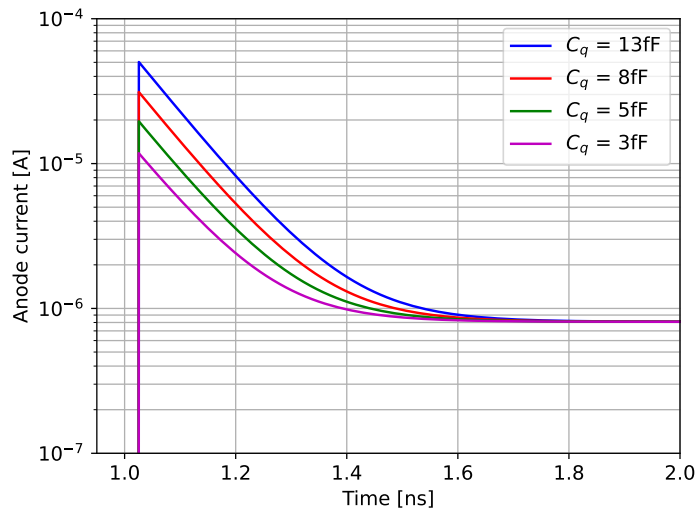


Figure 2.13: Parametrization of quenching capacitance on the transient response to a light pulse of a SPICE SPAD model. The latter relies on the TCAD parameters and literature-based typical values.

The cross-section BC' (Figure 2.5) gives the metal capacitance C_m between the cathode and the anode of the pixel. From AC simulation, its value is independent of the bias voltage and is about 13.4 fF. Although this value is neglected in the SPAD model, it will be useful when dealing with the SiPM.

2.2.4 Transient simulation

The transient simulation is done to simulate the response of the SPAD to a light pulse. The diode is reverse-biased above breakdown voltage, at a value of $V_A = V_{bd} + V_{ex}$, where the excess voltage V_{ex} is set around 3 V [4]. A first quasi-static simulation is performed to reach this bias point without triggering the diode [9]. The thermal-driven SRH generation is neglected only to consider the avalanche triggered by photons. Next, Figure 2.14 describes the transient response of the SPAD after a 1 ns light pulse at 480 nm wavelength with a power of $1 \times 10^{-4} \text{ W/cm}^2$. The light triggers the avalanche mechanism, which generates a high current peak up to 1.95 mA. It induces a voltage drop across the quenching resistor and lowers the cathode voltage down to the breakdown voltage, stopping the process. The breakdown voltage retrieved from the transient simulation is about 6.3 V, which is lower than that of the dark DC IV curve. This difference can be attributed to the non-equilibrium increase in carrier concentration due to the avalanche triggered by light. At this bias, the self-sustaining current is about 7 μA (Figure 2.14, left). Then, the residual current slowly recharges the cathode node through the quenching resistor up to its initial bias voltage (Figure 2.14, right). The recharge takes about 1 μs . The gain factor M_{SPAD} of the simulated SPAD pixel is retrieved from the transient simulation by integrating the current arising after the light pulse. M_{SPAD} is about 6.5×10^6 , which falls within the common range of values found in practice (1×10^5 to 1×10^7) [12].

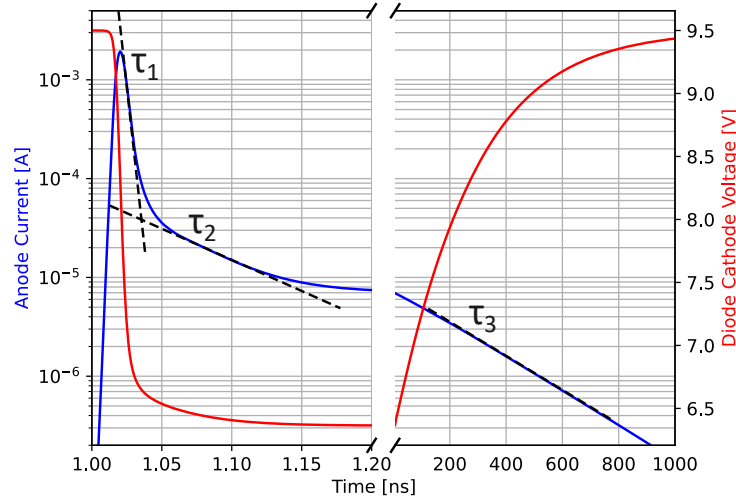


Figure 2.14: Transient response of the SPAD to a light pulse sent after 1 ns. The anode current is represented in blue (left axis), and the cathode voltage in red (right axis). The left part of the graph is a 200 ps-long zoom on the fast response induced by the light excitation. The black lines describe the different slopes linked to the time constants of the drift (τ_1) and diffusion (τ_2) currents. The right part of the graph shows the slow recharge (τ_3) of the avalanche diode by the quenching resistor.

The two slopes shown in Figure 2.14 (left) can be explained by drift and diffusion mechanisms [4]. As outlined in Figure 2.15 (left), the carriers located in the depletion region are rapidly extracted, while a second contribution based on the diffusion of the carriers outside the depletion region causes a delayed current pulse. A similar TCAD-simulated response has already been reported in [9] in the case of quenching at the anode, with a current peak in the mA range and double slopes in the quenching phase.

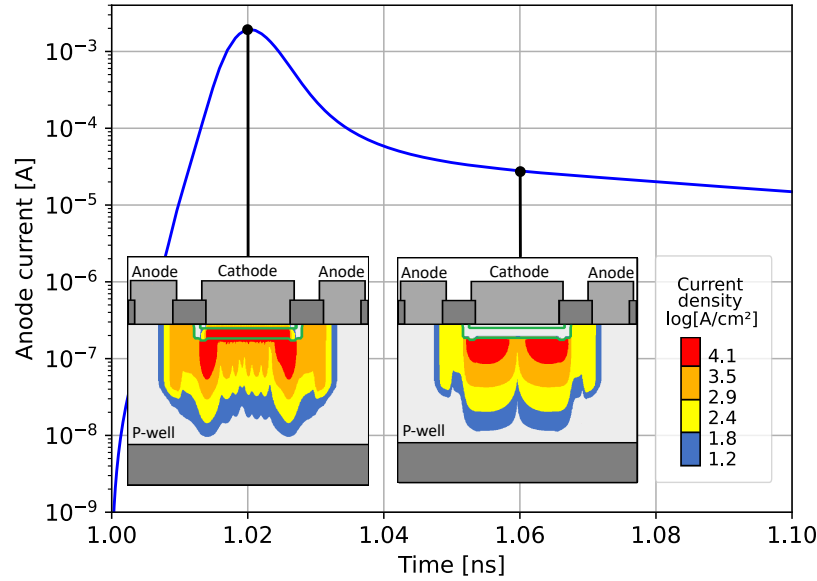
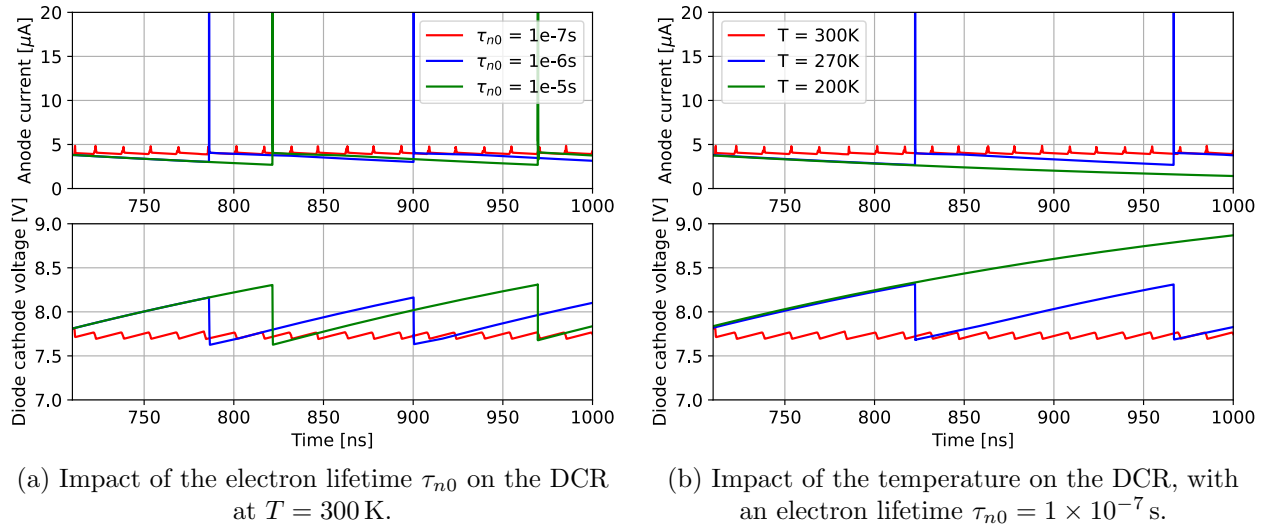


Figure 2.15: Temporal evolution of the current density in the TCAD-simulated layout after a light pulse excitation. The depletion region is delimited by a green line. At 1.02 ns, the carriers in the depletion region are rapidly swept out due to the high electric field (left inset). In contrast, diffusion carriers from the quasi-neutral region take longer to extract (right inset).

Considering the generation associated with SRH in the TCAD simulation, the transient simulation experiences dark counts, as shown in Figure 2.16. In particular, they arise around the DC breakdown voltage (7.7 V). Dark counts occur when spontaneously generated carriers trigger an avalanche mechanism [9]. Hence it is the same origin as the dark current and is primarily due to thermal SRH generation and band-to-band tunneling [57]. The ionization coefficient directly impacts the DCR, as a higher value increases the avalanche-triggering probability. Focusing on the SRH contribution, temperature, and carrier lifetime affects the DCR since they set the generation rate (Equation 2.4). As presented in Figure 2.16, increasing the electron lifetime or decreasing the temperature decreases the DCR by reducing trap-assisted thermal generation. On the other hand, BTBT generation is the main contribution of DCR at low temperature [9]. As seen in Figure 2.16b, the simulation displays no dark count at 200 K. It suggests that BTBT has a negligible impact on the simulated DCR compared to SRH generation.



(a) Impact of the electron lifetime τ_{n0} on the DCR at $T = 300$ K.

(b) Impact of the temperature on the DCR, with an electron lifetime $\tau_{n0} = 1 \times 10^{-7}$ s.

Figure 2.16: TCAD transient simulation of SPAD recharge in the dark with an excess voltage of 1.8 V compared to the static breakdown voltage ($V_{bd} = 7.7$ V), when enabling SRH generation.

Table 2.3 summarizes the characterization of the DCR related to SRH generation and compares it with the measured DCR. It is expressed in terms of cps/ μm^2 by considering the dark counts per active area. The active area of the TCAD-simulated structure is about $910 \mu\text{m}^2$. By default, the simulated DCR is several orders of magnitude higher than the measured one. This contrast can be attributed to the difference observed in the dark IV characteristics of the diode (Figure 2.11). The TCAD model exhibits a higher dark current and gain factor than the measured SPAD. It results in more generated carriers, enhancing the DCR.

	T [K]	τ_{n0} [s]	V_{ex}	DCR [cps/ μm^2]
<i>Silvaco</i>	300	1e-7	1.8	71868
	300	1e-6	1.8	9615
	300	1e-5	1.8	8121
	270	1e-7	1.8	7604
	200	1e-7	1.8	NA*
Measurement [5]	300	-	2.5	136

Table 2.3: Comparison of the DCR between the simulation and the measurement. *No DCR is observed in the simulation within a time frame of up to 1 s.

It is worth noting that the simulated dark counts are deterministic. The simulation appears to trigger the avalanche when the dark current reaches a specific level, resulting in consistent peak amplitudes and inter-arrival times (e.g., Figure 2.16a). By reducing the SRH generation, the dark current is decreased. Consequently, the TCAD model can recharge to a higher bias before the number of generated carriers exceeds the threshold and triggers an avalanche. In practice, both dark current and DCR are random. Dark current is driven by spontaneous generation, while DCR additionally involves the triggering probability. Therefore, the comparison between the measured

and simulated DCR is limited. Further investigations employing a statistical approach would offer a better representation of this phenomenon, as done in [20]. However, this work focuses on the transient response to a light pulse to design a dedicated readout circuit. Hence, the behavior of the DCR is not required, and SRH is neglected in the TCAD model to avoid dark count.

2.3 SPICE compact modeling

2.3.1 Proposed model

As mentioned, classical SPICE models rely on measured data, which are affected by parasitic elements and the pulse shaping of the instrumentation circuitry. These non-idealities tend to hide the different current mechanisms, such as drift and diffusion, highlighted in Figure 2.15. To address the issue, a novel compact model is proposed to represent the transient intrinsic response of a SPAD to a light pulse based on the TCAD simulations. The proposed SPICE macro-model is presented in Figure 2.17 and comprises three equivalent diode sub-circuits. Each one is modeled by the equivalent circuit shown in Figure 2.6b. The switch is closed by a signal in the ps range to simulate a photon trigger. Compared to the conventional approach, the switch is here modeled in SPICE by a voltage-controlled resistance to obtain a physical rise of the current. When off, the resistance is considered an open circuit. When on, it decreases exponentially to simulate the fast rise of the avalanche process.

Each equivalent diode sub-circuit describes a different physical mechanism that occurs after the light excitation. Because these mechanisms arise at different timescales and speeds, the junction capacitance will take a different value, as can be inferred from Figure 2.12b. These values are obtained by extracting the time constants τ from the transient TCAD simulation (Figure 2.14) through the slope of the log-scale current taken at half of the current dynamic range of each region. The capacitances are retrieved through the following relation:

$$\tau = C_{eq}R_{eq}, \quad (2.8)$$

with the equivalent capacitance $C_{eq} = (C_q + C_d)$ and the equivalent resistance $R_{eq} = (R_q || R_d)$.

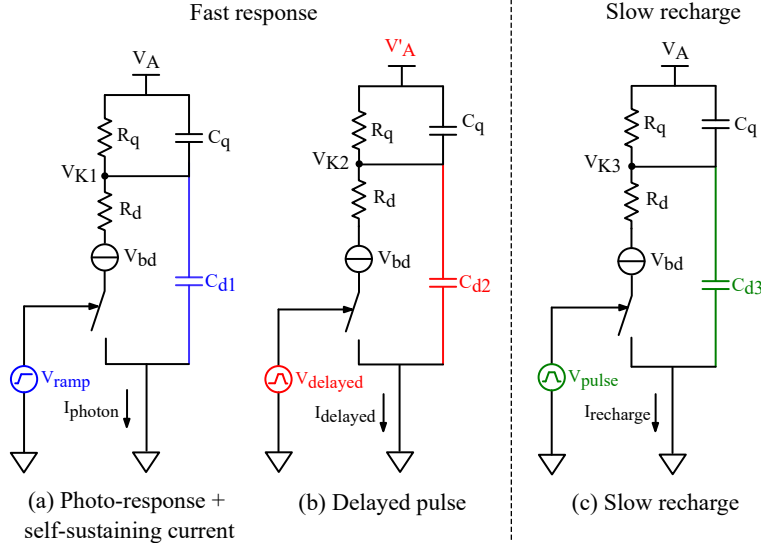


Figure 2.17: Proposed SPICE two-step model of the SPAD, describing both fast pulse and slow recharge. V_{ramp} and V_{pulse} rise at $t = 0$ while the $V_{delayed}$ rises at $t = T_{delayed}$.

The first diode sub-circuit (Figure 2.17a) models the primary photo-response current. The switch is closed by a ramp signal V_{ramp} to start the avalanche process induced by the light. When closed, the equivalent circuit generates a high current pulse, and the cathode voltage V_{K1} drops abruptly. After the pulse, the command signal V_{ramp} is kept high to induce the self-sustaining current of 7 μA , generated through the voltage drop across R_q . Given a time constant of about 3.7 ps for the fast pulse related to drift, the equivalent capacitance is about 18.7 fF. This corresponds to a depletion capacitance C_{d1} of about 12.7 fF, as observed at very high frequency in Figure 2.12b.

The second diode sub-circuit (Figure 2.17b) simulates the diffusion process of the carriers generated outside the depletion region. Since slower, its contribution is delayed compared to the first pulse. Hence, the second switch is closed after a delay $T_{delayed}$ of 25 ps estimated from Figure 2.14, through the signal $V_{delayed}$. The diffusion process has a time constant τ_2 of about 42 ps after removing the contribution of the current flowing through R_q (detailed computation done in Annex A.3). It yields an estimated junction capacitance C_{d2} of about 289 fF. Since this current contribution is delayed, the voltage at the cathode has already dropped. This is considered by applying a bias voltage V'_A set as: $V'_A = V_{K1}(t = T_{delayed}) = 6.7 \text{ V}$. The sum of the currents set by the two equivalent sub-circuits expresses the total generated pulse.

The third diode sub-circuit (Figure 2.17c) focuses only on recharge. The switch is closed to produce the voltage drop from the initial bias to the breakdown voltage. After 1 ns, the switch is opened to represent the end of the avalanche process by the quenching resistor. Since the diode is switched off, the mechanism responsible for the recharge operates at a lower frequency, which results in a high capacitance C_{d3} (Figure 2.12b). The associated time constant τ_3 is about 277 ns in the simulated transient recharge. When off, the diode equivalent resistance is very large compared to R_q , which gives an equivalent capacitance of about 0.63 pF. It corresponds to the low-frequency value of C_d in Figure 2.12b.

In the last step, the above estimated sub-model parameters have been adjusted to fit the full SPICE simulation on the transient response of the TCAD simulation, yielding the final SPICE parameters given Tab. 2.4.

V_A [V]	V'_A [V]	V_{bd} [V]	R_q [k Ω]	C_q [fF]
9.5	6.7	6.3	440	6
R_d [Ω]	C_{d1} [fF]	C_{d2} [fF]	C_{d3} [pF]	
142	22	250	0.6	

Table 2.4: Summary of SPICE model parameters.

2.3.2 Response to a light pulse

Figure 2.18 presents the response of the SPICE model compared to the TCAD simulation. The first phase (Figure 2.18, left) describes the physical current rise induced by the light-triggered avalanche mechanism and correctly reproduces the different current contributions. A first pulse of 1.97 mA arises 20 ps after the light excitation. The pulse is followed by a delayed response after 25 ps and a self-sustaining current of 8.5 μ A. It accurately represents the TCAD-simulated current peak (1.95 mA) but slightly overestimates the self-sustaining current (7 μ A). The cathode voltage drops from its initial bias voltage of 9.5 V to the non-equilibrium breakdown voltage of 6.3 V. The first part of the drop (down to 6.7 V) is dominated by the fast pulse, while the delayed response gives the second part. In the second phase (Figure 2.18, right), the model shows the slow 1 μ s-recharge of the cathode done through the quenching resistor.

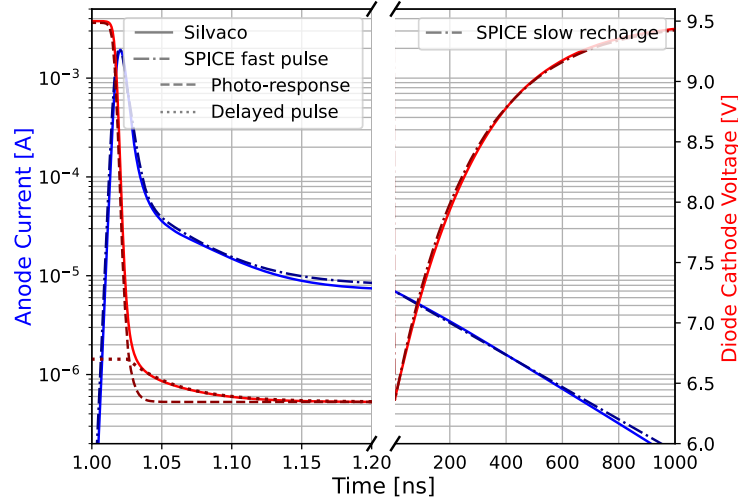


Figure 2.18: Transient response of the proposed model vs. the mixed-mode simulation. On the left part of the graph, the red dashed line is the voltage drop induced by the first current pulse, while the dotted line is the voltage drop due to the delayed pulse. The blue dot-dashed line is the sum of the currents induced by the two first equivalent diodes.

2.3.3 Adaption to a SiPM model

Once the model of one SPAD is done, the circuit can be easily adapted to an equivalent SiPM model, as presented in Figure 2.19. The fired SPADs are only characterized by the first diode sub-circuit, with a junction capacitance C_{d1} . The second mechanism related to diffusion can be neglected as the delays are increased due to the large capacitance of the SiPM. The non-fired cells are described by the capacitance of the quiescent SPAD C_{d3} .

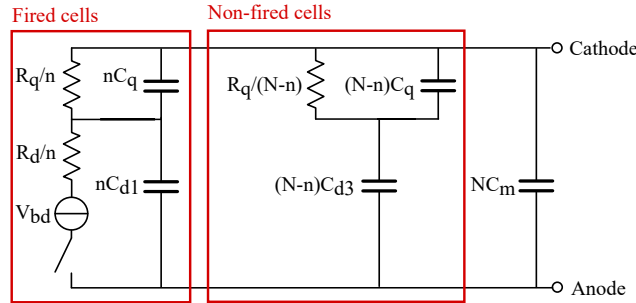


Figure 2.19: Equivalent SiPM model adapted from the proposed SPAD model, with n firing microcells and $(N - n)$ quiescent microcells.

The equivalent impedance of the quiescent macrocell is expressed as:

$$Z_{eq} = \frac{1}{(N - 1)} \frac{j\omega R_q(C_{d3} + C_q) + 1}{(j\omega)^2 R_q(C_{d3}C_q + C_mC_{d3} + C_mC_q) + j\omega(C_{d3} + C_d + C_m)}. \quad (2.9)$$

Given that the pulses occur in the nanosecond range, the quenching resistors of the quiescent cells have a limited impact compared to the capacitances. In the first step, they can be neglected [26]. With this assumption, Equation 2.9 is simplified to an equivalent capacitance C_{eq} as:

$$C_{eq} = (N - 1) \frac{C_q C_{d3}}{C_q + C_{d3}} + (N - 1)C_m. \quad (2.10)$$

With $C_{d3} = 0.6$ pF, $N = 234$, $C_q = 6$ fF and $C_m = 13.4$ fF, the equivalent capacitance is about 4.3 pF. As seen in Figure 2.20a, above 100 MHz, the equivalent impedance Z_{eq} is close to the equivalent capacitance C_{eq} . Figure 2.20b presents the transient response of the SiPM model to several simultaneous photons. The peak amplitudes are directly proportional to the number of fired cells.

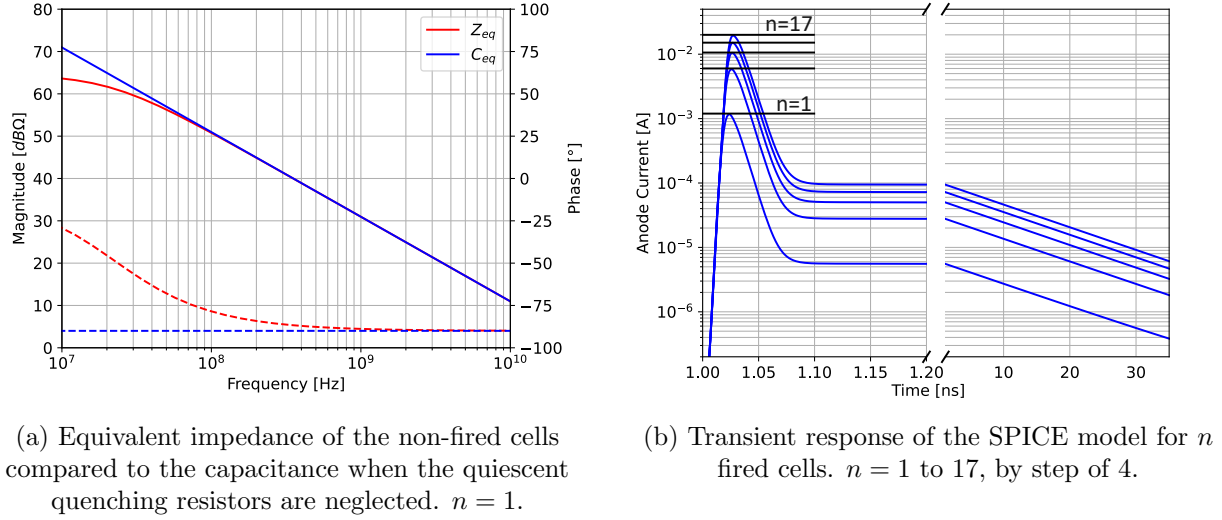


Figure 2.20: Response of the derived SPICE SiPM model for $N = 234$ cells, with the values extracted from the TCAD SPAD simulations.

2.3.4 Comparison with the literature

Table 2.5 displays extracted parameters from various SPAD models. The full width at half maximum (FWHM) of the signal pulse significantly varies across the different applications, ranging from 6 ns to 1 ps. SiPMs typically exhibit a wider pulse due to their large number of capacitances in parallel. The pulse width obtained from the TCAD simulation is narrower than the other models as it represents a more intrinsic response. Literature-based models usually derive parameter values from measurements where the pulse width increases due to parasitic capacitances or limited readout bandwidth.

		V_{bd} [V]	V_{ex} [V]	R_d [Ω]	C_d [F]	R_q [Ω]	C_m [F]	C_q [F]	# Cells	Pixel [μm]	FHMW [s]
SiPM	Hamamatsu S10362-33-050c [63]	69.5	1.5	1k	65.5f	30M	16f	50f	3600	50	110-140n
	Hamamatsu S11828-3344MX [65]	70.5	2	1k	78f	300k	1f	8f	3600	50	30n
	Hamamatsu [27]	65.2	2.8	-	86.1f	256k	0.66f	12.8f	3600	50	31n
	IRST [64] (research)	31.2	3.8	-	34.13f	394k	38.94f	4.95f	625	-	-
	Photonique [64] (research)	61	2.8	-	40.3f	774k	47.17f	5.40f	516	-	-
	UCLouvain BSI SOI [4] (measured)	11.4	3	1k	-	440k	-	-	234	55	-
SPAD	28 nm FD-SOI CMOS SPAD [9] (simulated)	9.5	0.5	-	-	220k*	/	3f*	1	-	-
	CMOS UTBB FDSOI [66] (research)	13.5	0.2-4	-	1p	200k	/	/	1	-	1p
	[22] (research)	21.54	4-6	390	4.33p	220k	/	/	1	-	6n
	This work (simulated)	6.3-7.7	3.1-1.8	142	22f-250f-0.6p	440k*	/	6f*	1	-	10p

Table 2.5: Comparison of extracted parameters from different SPAD and SiPM models. *The parameters are chosen rather than extracted.

2.4 Conclusion of the chapter

Despite not fully matching the measured data, the TCAD model provides a general understanding of the SPAD pixel behavior. It enables its complete characterization by retrieving its DC, AC, and

transient characteristics. In particular, the mixed-mode model simulates the transient response of the SPAD pixel to a light pulse. The parameters of the SPAD pixel are extracted from the simulated characteristics to build an equivalent model. However, the TCAD simulation depends on the default parameters established in *Silvaco*, which sets discrepancies between the simulation and the measured device. These values can be tuned to reproduce the measured characteristics. For instance, the polysilicon mobility can be adjusted to match the IV curve of the quenching resistor. The lifetime of the silicon can be tuned to achieve a closer dark current level compared to the measured SPAD. The doping profile and gradient can also be adjusted to obtain a similar multiplication gain and breakdown voltage. Further investigations on the DCR could also be done, including extracting the different generation contributions and the triggering probability. However, because this work focuses on the design of a dedicated readout circuit, only the photo-response of the SPAD is necessary, and the impact of dark counts on SPAD behavior can be discarded.

The presented compact SPICE model accurately represents the intrinsic transient response of a TCAD-simulated SPAD to a light pulse. The developed two-phase macro-model describes the fast double-response to a photon excitation and the slow recharge induced by the quenching resistor. This is achieved using three parallel compact equivalent sub-circuits of the diode, each characterizing a different physical mechanism. Their associated junction capacitance is linked to the physical time constant of each phenomenon. The first diode sub-circuit, describing a fast pulse reaching about 2 mA, has a capacitance in the tens of femtofarad and a time constant of about 3 ps, corresponding to a drift process with a very high-frequency depletion capacitance. The second diode sub-circuit, representing a slower diffusion current induced after the fast pulse, has a capacitance in the hundreds of femtofarad and a time constant of 35 ps leading to a self-sustaining current of about 8.5 μ A. The third diode sub-circuit, modeling the slow recharge, has a capacitance of about tenths of a picofarad linked to the low-frequency diode depletion capacitance and a time constant of 264 ns. This SPICE model can be used to design a dedicated integrated (at pixel level) readout circuit with active recharge by accurately reproducing the current and voltage induced by the SPAD and the quenching resistor after a photon excitation.

Chapter 3

Transimpedance amplifier design

This chapter focuses on the design methodology of the integrated preamplification stage for a SiPM, specifically targeting fast readout applications. The transistor-level design is carried out using the g_m/I_D methodology. All the design flow is achieved using a *Python* script based on analytical expressions, where the required transistor parameters are obtained from simulated curves generated by *Cadence Virtuoso*. The transistor technology is based on the 1.8 V, 0.18 μm on SOI CMOS technology from XFAB (XT018) [67]. This choice is motivated by its compatibility with the fabrication of the studied SiPM [4].

3.1 Structure and specifications

3.1.1 Front-end Architecture

The transimpedance amplifier (TIA) is selected as the preamplifier structure based on the state-of-the-art analysis (Chapter [1]). The choice is mainly motivated by its high bandwidth and short RC time associated with its low input impedance. These characteristics make it suitable for photon-counting applications. Although a common-base amplifier would have been a good option, the use of CMOS technology in this application makes its implementation less practical. From Chapter [1] the gain of the transimpedance amplifier is recalled as:

$$A_{TIA} = \frac{v_{out}}{i_{in}} = \frac{-R_f}{1 + \frac{1}{A_v\beta}} \quad \text{with} \quad \beta = \frac{1 + j\omega R_f C_{in}}{1 + j\omega R_f (C_{in} + C_f)}, \quad (3.1)$$

where β is the feedback factor and A_v designates the open-loop gain at the output of the operational amplifier. $A_v\beta$ is the loop gain, which also considers the feedback structure.

The front-end architecture is shown in Figure [3.1]. In this chapter, the SiPM is modeled by an ideal current source, in parallel with a capacitor $C_{in} = 4.3 \text{ pF}$. The equivalent capacitance of the studied SiPM has been determined in Chapter [2] (Equation [2.10]). The input impedance of the next block is also important to consider, as it affects the transfer function of the amplifier. As discussed in Chapter [1], a pulse shaper is typically added after the preamplifier, where a common implementation relies on a RC-CRm filter. Hence, a capacitance C_{hp} in series with a resistance R_{hp} is added at the output of the preamplifier. However, this work will not cover the implementation of the pulse shaper.

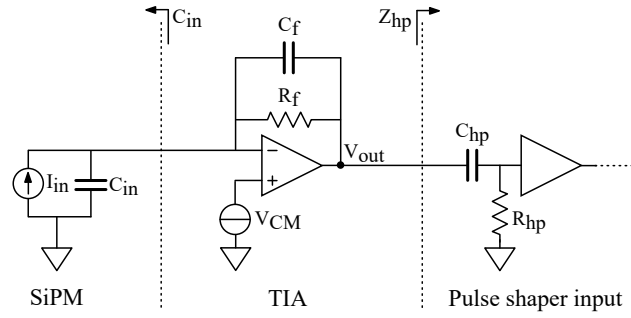


Figure 3.1: Front-end setup. The TIA sees at its input the equivalent capacitance of the SiPM and at its output the input impedance of the pulse shaper.

3.1.2 Amplifier topology

The operational amplifier is implemented using a folded-cascode architecture. It offers sufficient gain and speed for a limited power while exhibiting a higher output swing than the telescopic configuration [49]. The amplifier structure is presented in Figure 3.2. The folded cascode is connected to an output buffer and a low-voltage bias circuit. The buffer is implemented by a source follower. It is added to drive the required current in R_f and to decrease the output impedance. The output impedance should be low compared to the load impedance Z_L to limit the impact of the latter on the amplifier characteristics. The low-voltage bias circuit enhances the output swing to counteract the decrease caused by the cascode and the connection with the buffer. For the same reason, the design is implemented using low-threshold voltage transistors.

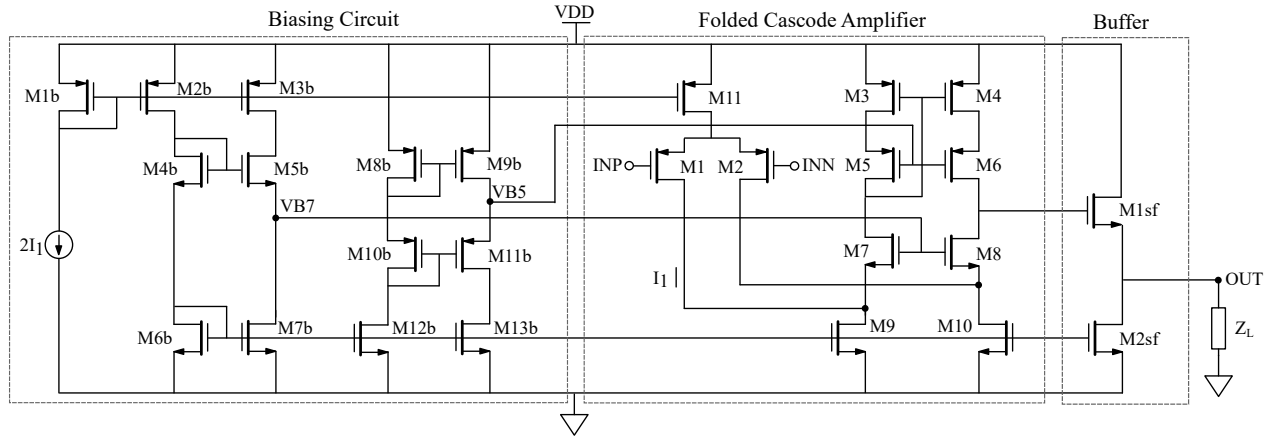


Figure 3.2: Considered operational amplifier, using a folded-cascode amplifier with a low-voltage bias circuit and an output buffer implemented with a source follower.

3.1.3 Specifications

The SiPM current peak amplitude sets the specifications in terms of driving current and maximum transimpedance gain. The current peak depends on both the current of one pixel and the number of simultaneously fired cells. As shown in Chapter 2, the SPICE SiPM model generates a pulse peak in the mA range for a single-photon event. However, this range differs from the literature, which

rather considers peaks in the μA range in the case of the single-photon detection [31]. For more realistic specifications, the maximum current is based on the amplitude of the slower decay observed in the TCAD-simulated pulse, which is in the tens of μA . Regarding the number of simultaneously fired cells, it follows a Poisson distribution [26]. Its expectation depends on the application and the number of pixels in the SiPM. Typical numbers of simultaneous photons range in the tens of events [68]. Taking both factors into account, the maximum current I_{max} is set at $100\text{ }\mu\text{A}$.

The transimpedance gain must be limited to prevent the output voltage from falling outside the dynamic range and distorting the signal. A gain of between $60\text{ dB}\Omega$ and $80\text{ dB}\Omega$ is considered sufficient while ensuring an acceptable input dynamic range. Additionally, the operational amplifier gain A_v should be high to reach low input impedance. Assuming a transimpedance gain of $80\text{ dB}\Omega$ ($R_f = 10\text{ k}\Omega$), A_v should be at least 60 dB to obtain a DC input impedance ($= R_f/(A_v + 1)$) below the tens of Ω [35].

The amplifier should have high bandwidth and, by extension, a large gain bandwidth product (GBW) to achieve photon counting. High bandwidth is required to detect short consecutive pulses and to sense the fast rise of the signal. According to the literature, the bandwidth typically ranges above 10 MHz [31] and can go beyond 1 GHz [43]. Large bandwidths are achieved at the expense of higher power or more advanced configurations. Since the application aims for low power, the consumption should fall within the sub-mW range. With this constraint, typical bandwidths range in the tens of MHz [31] [42].

A summary of the general specifications is reported in Table 3.1. However, these specifications remain broad. In practice, they heavily rely on the specific application and its related characteristics, such as the maximum number of simultaneous photons or the frequency of consecutive distinct pulses. A better knowledge of the application would lead to more stringent specifications in terms of gain, maximum current, or bandwidth.

I_{max}	R_f	A_v	Bandwidth	Power
$100\text{ }\mu\text{A}$	$1\text{-}10\text{ k}\Omega$	$> 60\text{ dB}$	$> 10\text{ MHz}$	$< 1\text{ mW}$

Table 3.1: Summary of the preamplifier specifications.

3.1.4 Design methodology: gm over ID

The g_m/I_D methodology is a design tool to size transistors. It uses the fact that the relationship between g_m/I_D and the normalized drain current, $I_{Dn} = I_D/(W/L)$, can be considered independent of the transistor dimensions (W , L) for a given technology [69]. The relationship between these two parameters can be obtained by simulations, as shown in Figure 3.3. It allows working in all the levels of inversion of a MOSFET, where a low g_m/I_D indicates strong inversion (SI), and a high g_m/I_D implies weak inversion (WI). Strong inversion provides advantages in terms of size and speed, while weak inversion has benefits in terms of gain and power. For a given g_m/I_D and L , parameters such as the saturation voltage V_{DSsat} , the Early voltage V_{EA} , and parasitic capacitances per unit area can also be estimated from similar curves. It allows a complete estimation of the transistor characteristics.

Because of the cascoded structure, the body effect should be considered. When the body of a cascode transistor is connected to the small-signal ground rather than to its source, the body-source voltage is not constant. It induces a second transconductance contribution $g_{mb} = \chi g_m$, with χ the body factor coefficient [48]. The total transconductance is enhanced and is expressed as $g_{ms} = g_m + g_{mb} = n g_m$, with $n = 1 + \chi$ the linearized body factor. Hence, when working with a cascode transistor, g_{ms}/I_D is used instead (Figure 3.3, red).

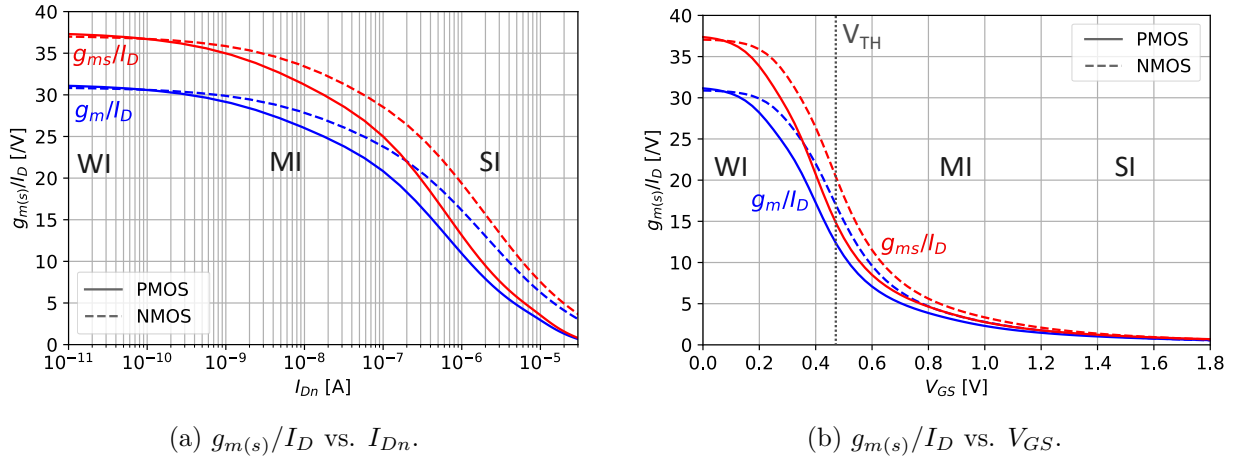


Figure 3.3: Simulated $g_{m(s)}/I_D$ characteristics of the XT018 technology low-threshold voltage transistors, with $V_{DS} = 0.45$ V, $L = 0.5$ μ m. (SI: strong inversion, MI: moderate inversion, WI: weak inversion).

3.2 Folded-cascode amplifier

This section focuses on the implementation of the folded-cascode topology. The amplifier architecture is shown in Figure 3.4 and is set as follows: M1-M2 form the differential pair. M11 functions as a current source and divides the reference current into the two branches of the input pair. M9-M10 sets the reference current in their respective branch, while M3-M4 form a current mirror. The NMOS side of the output is cascoded by M7-M8, and the PMOS side is cascoded by M5-M6, with M7 and M5 added for matching purposes. The relationship between the static drain current of the differential pair I_{D1} and the cascode transistors I_{D3} is set by the factor B , such as $I_{D3} = B I_{D1}$. The positive input is found by small-signal analysis. It is provided by M1, as increasing the gate voltage reduces I_{D1} . It increases I_{D3} as I_{D9} is fixed by M9. I_{D3} is mirrored to I_{D4} , leading to a charge of the output node. A similar analysis can be done for the negative input set by M2. The output load is set by the input capacitance of the source follower, represented by C_{sf} , which corresponds to the parasitic capacitance of its input gate.

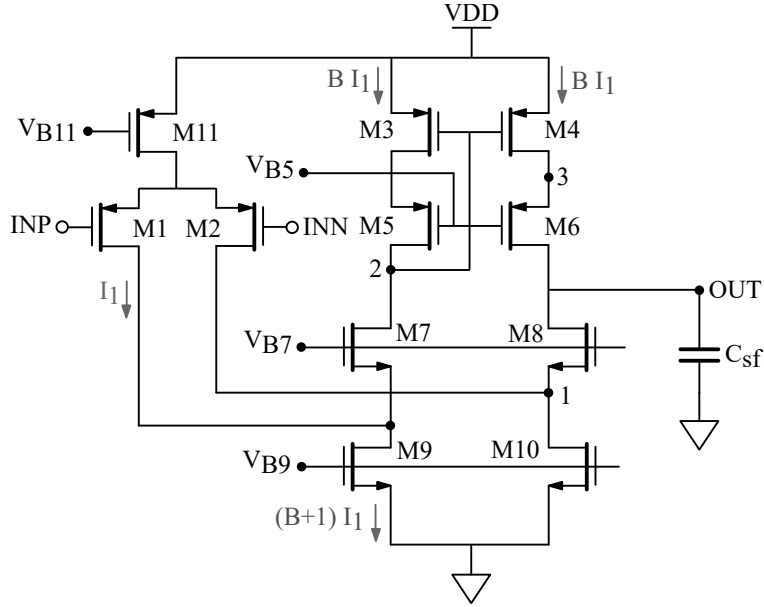


Figure 3.4: Folded-cascode amplifier. The load capacitance corresponds to the input gate capacitance of output buffer C_{sf} .

Because of the symmetry of the structure, the pairs M1-M2, M3-M4, M5-M6, M7-M8, and M9-M10 are equally sized. For the sake of simplicity, their parameters are identified as the value of the odd transistor, such as $g_{m1,2} = g_{m1}$.

3.2.1 Equations

Transfer function

The small-signal analysis is employed to approximate the transfer function of the folded cascode. As detailed in [48], the DC gain A_{v0}^{DC} can be expressed as:

$$A_{v0}^{DC} = g_{m1} R_o, \quad (3.2)$$

where R_o is the output impedance and is approximated by:

$$\frac{1}{R_o} = g_o \approx \frac{g_{d8}(g_{d2} + g_{d10})}{g_{ms8}} + \frac{g_{d6}g_{d4}}{g_{ms6}}. \quad (3.3)$$

The dominant pole of the transfer function is associated with the output node of the folded cascode. It is determined by its output impedance R_o and the load capacitance C_{sf} :

$$\omega_{p0} = \frac{1}{R_o C_{sf}}. \quad (3.4)$$

The three nodes highlighted in Figure 3.4 give the non-dominant poles. They are approximated as follows:

$$\begin{cases} \omega_{p1} = \frac{g_{ms8}}{C_1} \approx \frac{g_{ms8}}{C_{gs8}}, \\ \omega_{p2} = \frac{g_{ms5}}{C_2} \approx \frac{g_{ms5}}{C_{gs3} + C_{sg4}}, \\ \omega_{p3} = \frac{g_{ms6}}{C_3} \approx \frac{g_{ms6}}{C_{sg6}}, \end{cases} \quad (3.5)$$

by only considering the parasitic capacitances associated with C_{gs} , the gate-to-source intrinsic capacitance. As discussed in [70], the current mirror M3-M4 introduces a doublet zero-pole. The associated zero is given by:

$$\omega_{z1} = 2\omega_{p2}. \quad (3.6)$$

Knowing the DC gain and the main poles and zeros, the transfer function of the folded cascode, denoted A_{v0} , is set by:

$$A_{v0} = \frac{g_{m1}R_o}{1 + j\omega R_o C_{sf}} \overbrace{\frac{\sum_{i=1}^{n_z} \left(1 + \frac{j\omega}{\omega_{zi}}\right)}{\sum_{i=1}^{n_p} \left(1 + \frac{j\omega}{\omega_{pi}}\right)}}^{\text{Secondary poles and zeros}}. \quad (3.7)$$

Phase margin

The stability of a closed-loop system is measured by the phase margin ϕ_M . It is defined as the phase difference between the open-loop transfer function at unity gain and -180° . The unity gain arises at a frequency called the transition frequency f_T . As a practical rule, a margin of 45° is acceptable [71]. Considering the main poles and zeros induced by the folded cascode and supposing that the dominant pole results in a phase loss of 90° at f_T , the phase margin is given by:

$$\phi_M = 90^\circ - \arctan\left(\frac{2\pi f_T}{\omega_{p1}}\right) - \arctan\left(\frac{2\pi f_T}{\omega_{p2}}\right) + \arctan\left(\frac{2\pi f_T}{\omega_{z2}}\right) - \arctan\left(\frac{2\pi f_T}{\omega_{p3}}\right). \quad (3.8)$$

As Equation 3.8 indicates, setting the parasitic poles away from f_T increases the phase margin. However, it only gives a condition on the stability of the folded-cascode alone. One should ensure the loop gain $A_v\beta$ is still stable after the buffer and the feedback structure.

Slew rate

The slew rate is defined as the highest change rate of the output voltage and is determined by the maximum current in the output branch. It depends on the ratio B , which sets the current in the cascode transistors ($I_{D4} = BI_{D1}$). By construction, the slew rate is given by:

$$\begin{aligned}
SR &\approx \frac{2I_{D1}}{C_{sf}} && \text{for } B \geq 1, \\
SR &\approx \frac{(1+B)I_{D1}}{C_{sf}} && \text{for } B < 1.
\end{aligned} \tag{3.9}$$

Selecting a value of $B < 1$ results in a deterioration of the slew rate, while opting for $B > 1$ is less optimal in terms of power consumption without improving the slew rate. Therefore, in this work, the value of B is set to 1.

Noise

Two main noise sources can be considered in a transistor: thermal noise and flicker noise. The former is associated with the resistive channel between the source and drain, while the flicker noise is associated with defects and shows a $1/f$ dependency of the noise. The equivalent input noise power spectral density (PSD) for the i^{th} transistor is set as [72]:

$$S_{n,i} = \underbrace{\frac{4kT\Gamma_i n}{g_{m,i}}}_{\text{Thermal noise}} + \underbrace{\frac{K_{f,i}}{C_{ox}W_iL_i f}}_{\text{Flicker noise}} \left[\frac{V^2}{Hz} \right], \tag{3.10}$$

where Γ is the thermal noise coefficient which depends on the channel inversion level, and K_f is the empirical flicker noise coefficient. As detailed in [48], the noise contribution of the cascode transistors does not contribute significantly to the equivalent noise of the circuit. The total PSD at the input of the folded-cascode amplifier S_a can be written as:

$$S_a = 2 \left[S_{n,1} + \left(\frac{g_{m3}}{g_{m1}} \right)^2 S_{n,3} + \left(\frac{g_{m9}}{g_{m1}} \right)^2 S_{n,9} \right]. \tag{3.11}$$

By replacing the noise contribution of each transistor from Equation 3.10 into Equation 3.11 the thermal noise contribution of the amplifier ($S_{a,th}$) can be separated from the total flicker noise ($S_{a,1/f}$). They are respectively given by:

$$S_{a,th} = \frac{8kTn}{g_{m1}} \left[\Gamma_1 + \Gamma_3 \frac{g_{m3}}{g_{m1}} + \Gamma_9 \frac{g_{m9}}{g_{m1}} \right], \tag{3.12}$$

$$S_{a,1/f} = \frac{2}{C_{ox}f} \left[\frac{K_{f1}}{W_1L_1} + \left(\frac{g_{m3}}{g_{m1}} \right)^2 \frac{K_{f3}}{W_3L_3} + \left(\frac{g_{m9}}{g_{m1}} \right)^2 \frac{K_{f9}}{W_9L_9} \right]. \tag{3.13}$$

3.2.2 Design flow

Once all the equations are set up, the design can be assessed. Figure 3.5 summarizes the design flow of the folded cascode. It starts by fixing the transition frequency f_T , the capacitive load C_{sf} , as well as the length L and $[g_m/I_D]$ of each transistor pair. Assuming that the non-dominant poles are located away from the transition frequency, the latter equals the GBW . Under the same assumption, the gain bandwidth product is given by:

3.2.3 Design choices

The choice of g_m/I_D sets the characteristics of the amplifier. It is crucial to choose its value adequately for each transistor. It is done by considering its impact on the different aspects of the amplifier and the related trade-offs. To help the selection, the characteristics of the amplifier are expressed as a function of g_m/I_D when feasible.

M1-M2 differential pair design

The value of g_{m1} is fixed and set by Equation 3.14. Hence $[g_m/I_D]_1$ affects directly the value of I_{D1} . By substituting the expression of R_0 from Equation 3.3 into Equation 3.2, the impact of $[g_m/I_D]_1$ on the total gain is given by:

$$A_{v0}^{DC} = g_{m1}R_o = \frac{\left[\frac{g_m}{I_D}\right]_1}{\left[\frac{g_m}{I_D}\right]_7^{-1} (g_{d8}(g_{d2} + g_{d10})) + \left[\frac{g_m}{I_D}\right]_5^{-1} g_{d6}g_{d4}}. \quad (3.18)$$

Regarding the power, all the current contributions are set by I_{D1} . Therefore, it is inversely proportional to $[g_m/I_D]_1$:

$$P \propto V_{DD}I_{D1} = V_{DD} \left[\frac{g_m}{I_D}\right]_1^{-1} g_{m1}. \quad (3.19)$$

The impact on the size is obtained through the normalized current I_{Dn} . The size ratio of the i^{th} transistor can be expressed as:

$$\left(\frac{W}{L}\right)_i = \frac{I_D}{I_{Dn,i}} \propto \left[\frac{g_m}{I_D}\right]_1^{-1} \frac{g_{m1}}{I_{Dn,i}}. \quad (3.20)$$

From Equation 3.20, one should decrease $[g_m/I_D]_i$ to reduce the transistor size since it lowers $I_{Dn,i}$ (Figure 3.3a). There is an exception with $[g_m/I_D]_1$, as increasing its value reduces the size of the subsequent transistors. The overall size tends to decrease until $(W/L)_1$ becomes dominant compared to the other sizes.

In terms of noise, increasing the size of M1-M2 reduces the flicker noise (Equation 3.13). Additionally the increase of $[g_m/I_D]_1$ reduces the thermal noise, as highlighted by adapting Equation 3.12:

$$S_{a,th}^2 = \frac{8kTn}{g_{m1}} \left(\Gamma_1 + \Gamma_3 \left[\frac{g_m}{I_D}\right]_3 \left[\frac{g_m}{I_D}\right]_1^{-1} + \Gamma_9 \left[\frac{g_m}{I_D}\right]_9 \left[\frac{g_m}{I_D}\right]_1^{-1} \right). \quad (3.21)$$

Considering all these aspects, high $[g_m/I_D]_1$ gives higher gain, lower power, lower noise, and generally smaller transistor sizes. On the other hand, it reduces the slew rate:

$$SR \approx \frac{2I_{D1}}{C_{sf}} = \left[\frac{g_m}{I_D}\right]_1^{-1} \frac{2g_{m1}}{C_{sf}} = \left[\frac{g_m}{I_D}\right]_1^{-1} \frac{GBW}{\pi}, \quad (3.22)$$

which set an upper limit on $[g_m/I_D]_1$.

M3-M4 and M9-M10 current mirror design

Lower $[g_m/I_D]_3$ and lengths are preferred to increase the phase margin as it reduces the transistor size. It results in a lower parasitic capacitance (C_2), setting further away the doublet $\omega_{p2} - \omega_{z1}$ (Equation 3.5). In terms of noise, reducing $[g_m/I_D]_3$ lowers the thermal noise (Equation 3.21) but increases the flicker noise as it decreases the transistor width (Equation 3.13). This effect can be mitigated by increasing the transistor lengths.

Similar trends are retrieved from M9-M10 in terms of noise. Moreover, a higher value of $[g_m/I_D]_9$ and lengths give better matching compared to the current reference.

M5-M6 and M7-M8 cascode design

One should increase $[g_m/I_D]_5$ and $[g_m/I_D]_7$ to raise the output impedance and obtain a higher gain (Equation 3.18). As for M3-M4, reducing $[g_m/I_D]_5$, $[g_m/I_D]_7$ and their lengths improves the phase margin by reducing C_3 and C_1 , respectively (Equation 3.5).

Table 3.2 presents the main trends in terms of g_m/I_D and lengths on the performance of the amplifier. It provides general guidelines and trade-offs regarding the parametrization of each transistor to optimize the overall performance. However, it is worth noting the DC operating point restricts the choice of g_m/I_D values as all transistors must remain in saturation. The maximum transistor size and area constraints also set an upper limit.

	M1-M2	M3-M4	M5-M6	M7-M8	M9-M10	M11
↑ Gain A_0	$\frac{g_m}{I_D} \uparrow$		$\frac{g_m}{I_D} \uparrow$	$\frac{g_m}{I_D} \uparrow$		
↓ Power	$\frac{g_m}{I_D} \uparrow$					
↓ Area	$\frac{g_m}{I_D} \uparrow^*, L \downarrow$	$\frac{g_m}{I_D} \downarrow, L \downarrow$				
↑ Phase margin	$g_m \downarrow$	$\frac{g_m}{I_D} \downarrow, L \uparrow$	$\frac{g_m}{I_D} \downarrow, L \uparrow$	$\frac{g_m}{I_D} \downarrow, L \uparrow$		
↑ Matching	$\frac{g_m}{I_D} \uparrow, L \uparrow$				$\frac{g_m}{I_D} \uparrow, L \uparrow$	$\frac{g_m}{I_D} \uparrow, L \uparrow$
↑ SR	$\frac{g_m}{I_D} \downarrow$					
↓ Thermal noise	$\frac{g_m}{I_D} \uparrow$	$\frac{g_m}{I_D} \downarrow$			$\frac{g_m}{I_D} \downarrow$	
↓ 1/f noise	$\frac{g_m}{I_D} \uparrow, L \uparrow$	$\frac{g_m}{I_D} \uparrow, L \uparrow$			$\frac{g_m}{I_D} \uparrow, L \uparrow$	

Table 3.2: General trends of the folded-cascode parametrization. *Until $(W/L)_1$ becomes dominant compared to the subsequent ratios.

3.3 Low-voltage bias circuit

3.3.1 Architecture

The bias circuit imposes the reference current in the different branches of the circuit and generates the gate voltage of transistors M5-M6 and M7-M8. While it is simply done for M11 and M9-M10 with current mirrors, an alternative is preferred for the cascode transistors to increase the output swing. The naive implementation for the NMOS transistors is shown in Figure 3.6a. Assuming strong inversion, the condition to get M8 in saturation is set by:

$$V_{DS8} \geq V_{GS8} - V_{TH8} \quad \Leftrightarrow \quad V_{OUT} \geq V_{GS4b} + V_{GS6b} - V_{TH8}. \quad (3.23)$$

Considering that all the transistors have the same size for analysis purposes, It gives the lower bound of the output swing as:

$$V_{OUT} \geq 2V_{OV} + V_{TH}, \quad (3.24)$$

with $V_{OV} = V_{GS} - V_{TH}$, the overdrive voltage. In contrast, Figure 3.6b illustrates a low-voltage bias alternative. The ratio $(W/L)_{4b}$ is set four times smaller than $(W/L)_{5b}$ to obtain:

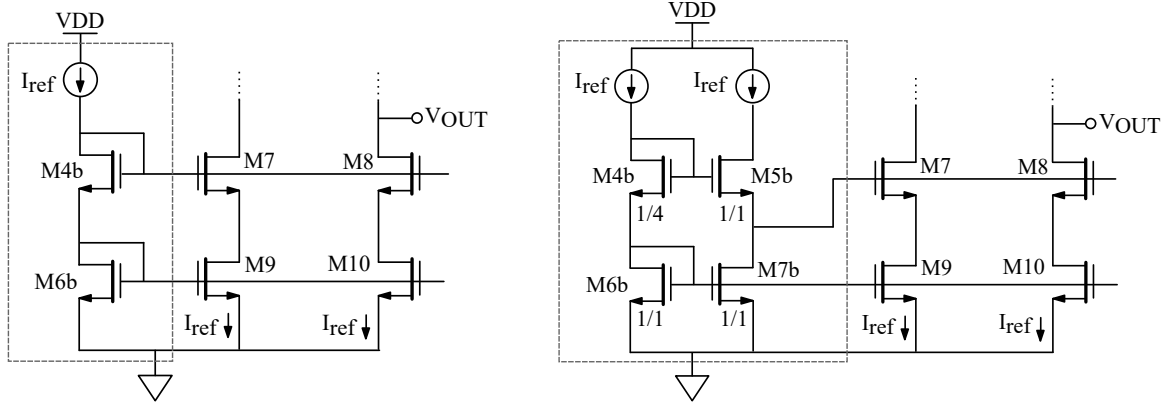
$$V_{OV4b} = \sqrt{\frac{2I_D}{k'_n \left(\frac{W}{L}\right)_{4b}}} = 2\sqrt{\frac{2I_D}{k'_n \left(\frac{W}{L}\right)_{5b}}} = 2V_{OV5b} \quad \text{if SI}, \quad (3.25)$$

with $k'_n = \mu_n C_{ox}$. The condition to get M8 in saturation is now given by:

$$V_{OUT} \geq V_{GS6b} + V_{GS4b} - V_{GS5b} - V_{TH8} \quad \Leftrightarrow \quad V_{OUT} \geq 2V_{OV}. \quad (3.26)$$

Hence, the second structure enhances the output swing by one V_{TH} compared to the standard bias circuit. A similar implementation is done to bias M5-M6 using PMOS. It gives the upper limit of the output swing as:

$$V_{OUT} \leq V_{DD} - 2V_{OV}. \quad (3.27)$$



(a) Standard bias circuit using current mirrors.

(b) Low-voltage bias circuit.

Figure 3.6: Bias structure (grey box) for the NMOS transistors of the folded-cascode amplifier.

3.3.2 Design flow

The sizing of the low-voltage bias circuit is determined by the transistors of the folded cascode, where the size ratios are set to satisfy Equation 3.25. Additionally, the drain currents of M7b and M13b (Figure 3.2) are halved compared to the reference current to reduce power consumption. Table 3.3 displays the transistor size configurations for the bias circuit compared to the folded-cascode transistors.

Reference transistors	1:1	4:1	2:1
M3-M4	M8b	-	M9b
M5-M6	-	M10b	M11b
M7-M8	-	M4b	M5b
M9-M10	M6b, M12b	-	M7b, M13b
M11	M1b, M2b	-	M3b

Table 3.3: Summary of the sizing of the bias circuit. The reference transistors set the ratio in each row.

3.3.3 DC point

As the bias circuit sets the DC point of the amplifier, one can investigate the design choices associated with the bias point. All the transistors must be saturated to ensure the validity of the small-signal analysis. The condition for achieving saturation is given by:

$$V_{DS} \geq V_{DSsat}, \quad \text{where in SI: } V_{DSsat} = V_{DS} - V_{TH}. \quad (3.28)$$

Imposing the saturation condition on the different transistors of the structure allows to estimate the input and output dynamic ranges. It also establishes the required bias values for the cascode transistors M7-M8 and M5-M6. A summary of the transistors setting the upper and lower bounds of the bias points is found in Table 3.4.

Lower bound [V]		bias point [V]	Upper bound [V]	
-	0	V_{IN}	$V_{DD} - V_{SDsat11} - V_{SG1}$	M11
M7*	$V_{GS6b} + V_{GS4b} - V_{GS5b} - V_{TH7}$	V_{OUT}	$V_{DD} - V_{SG8b} - V_{SG10b} + V_{SG11b}$	M6
M7*	$V_{DD} - V_{SG3} + V_{TH7}$	$V_{B7} = V_{GS6b} + V_{GS4b} - V_{GS5b}$	$V_{DSsat9} + V_{GS7}$	M9
M4	$V_{SDsat5} + V_{DD} - V_{SG3} - V_{SG5}$	$V_{B5} = V_{DD} - V_{SG8b} - V_{SG10b} + V_{SG11b}$	$V_{DD} - V_{SG5} - V_{SDsat4}$	M4

Table 3.4: Summary of the DC point and the conditions to have all transistors in saturation. At the edges of the table, the transistors setting the conditions are shown. *Assuming strong inversion.

A Parametric analysis can be conducted around an operating point to determine the acceptable range of g_m/I_D values that ensure all transistors are in saturation. To simplify this analysis, it is considered that $[g_m/I_D]_3 = [g_m/I_D]_5$ and $[g_m/I_D]_7 = [g_m/I_D]_9$. The parameters of the bias circuit are set accordingly, with the relationships given by Table 3.4. Unless noted otherwise, the g_m/I_D values are given by Table 3.5.

$[g_m/I_D]_1$ [/V]	$[g_m/I_D]_{3,5}$ [/V]	$[g_m/I_D]_{7,9}$ [/V]	$[g_m/I_D]_{11}$ [/V]
15	15	20	10

Table 3.5: $[g_m/I_D]$ operating point used to obtain the bias range curves.

Figure 3.7 shows the conditions set on the input and output voltages. The grey areas indicate zones where the saturation condition is not met for all transistors. The dashed lines represent the shift of the grey areas when sweeping a given g_m/I_D . Increasing $[g_m/I_D]_{1,11}$ increases the input dynamic range (Figure 3.7a), while increasing $[g_m/I_D]_{7,9}$ or $[g_m/I_D]_{3,5}$ enhances the output swing (Figure 3.7b). Because the TIA works in a feedback loop, one should ensure that the input dynamic range matches the output one. Additionally, $[g_m/I_D]_{7,9}$ should be high to obtain a sufficient lower bound of the output dynamic range (Figure 3.7b). It is even more critical as adding the output buffer reduces the output swing by one V_{TH} .

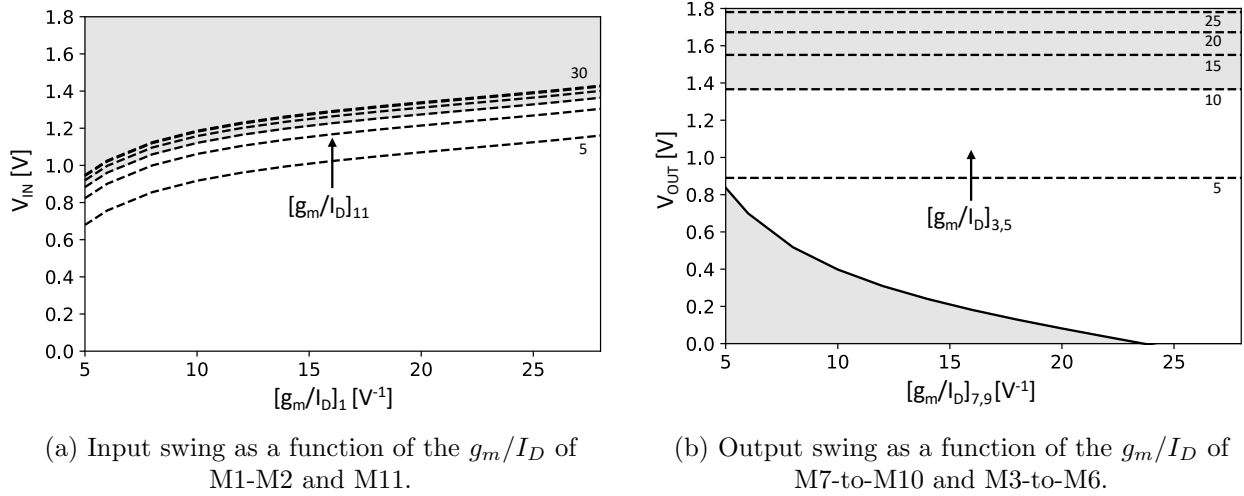


Figure 3.7: Estimation of the input and output dynamic ranges (white), using the *Python* script and the technological parameter curves.

The analysis can also be conducted for the bias voltages V_{B7} and V_{B5} (Figure 3.8). The reference voltage induced by the bias circuit is represented by a blue line. As seen in Figure 3.8a, the saturation condition on V_{B7} is almost always satisfied, except for both $[g_m/I_D]_{3,5}$ and $[g_m/I_D]_{7,9}$ lower than 5 V⁻¹ (Figure 3.8a, left). Choosing smaller $[g_m/I_D]_{7,9}$ increases the margin on the saturation condition, making the design more robust to variations. In contrast, the bias conditions on V_{B5} are more restrictive (Figure 3.8b). The window where M3-M4 and M5-M6 are both saturated is limited by the architecture of the PMOS current mirror.

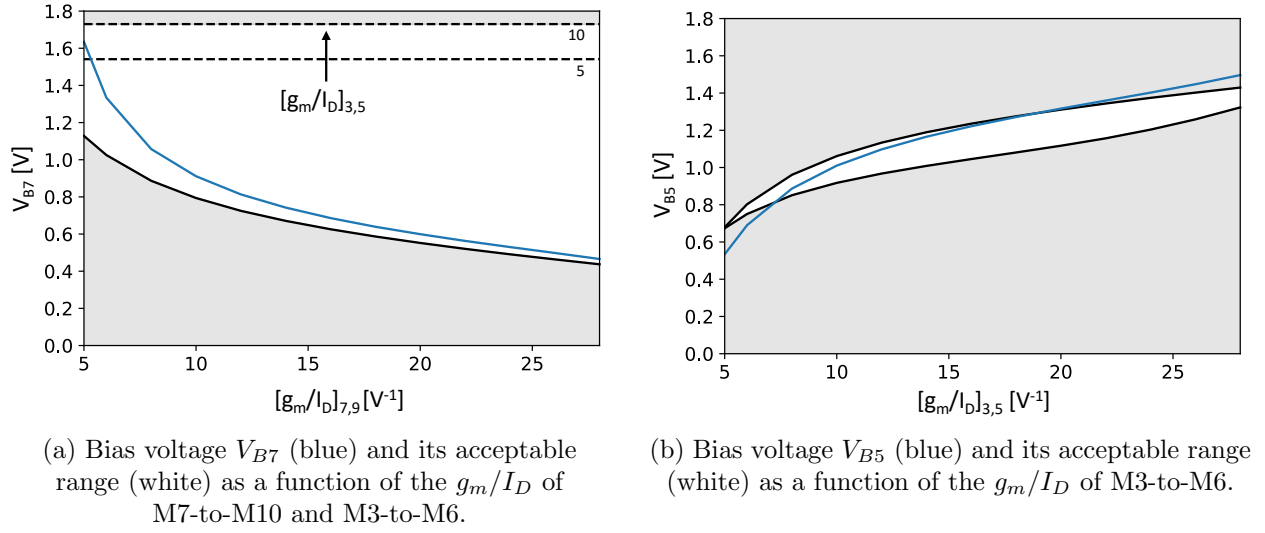


Figure 3.8: Acceptable range in terms of the DC point of the bias voltages (white), using the *Python* script and the technological parameter curves.

The results presented in Figure 3.7 and Figure 3.8 only provide general trends, as they rely on analytical expressions using parameters from simulated curves. In particular, the lower bounds of V_{OUT} and V_{B7} are underestimated at high g_m/I_D due to the assumption of strong inversion when deriving the saturation condition for M7.

3.4 Output buffer

3.4.1 Architecture

The output buffer serves two purposes. First, it reduces the output impedance to limit the impact of the load on the transfer function. Then, it provides a sufficient current to drive the input one. A typical implementation is done using a source follower (Figure 3.9), which benefits from unitary gain and a relatively low output impedance, as determined by small-signal analysis:

$$A_{v0}^{sf} = \frac{g_{m1,sf} \cdot r_{o2,sf}}{g_{m1,sf} \cdot r_{o2,sf} + 1} \approx 1 \quad \text{if } g_{m1,sf} \cdot r_{o2,sf} \gg 1, \quad (3.29)$$

$$R_{sf} = \frac{r_{o2,sf}}{g_{m1,sf} \cdot r_{o2,sf} + 1} \approx \frac{1}{g_{m1,sf}} \quad \text{if } g_{m2,sf} \cdot r_{o1,sf} \gg 1. \quad (3.30)$$

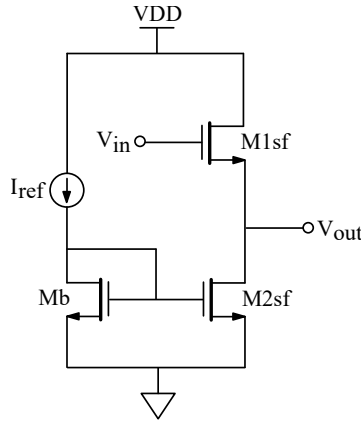


Figure 3.9: Voltage buffer implemented with a source follower.

As the transfer function of the output buffer is set, the overall gain of the operational amplifier A_v can be assessed by cascading the folded cascode and the source follower. Assuming that the buffer has unity gain and an infinite load, the transfer function of the operational amplifier corresponds to the one obtained for the folded-cascode (A_{v0} , Equation 3.7). However, the output load is finite as the buffer is connected to the feedback structure and the pulse shaper. They introduce extra poles and zeros that affect the transfer function. Considering the total load impedance, A_v is expressed as:

$$A_v = A_{v0} \frac{(Z_f || Z_{hp})}{(Z_f || Z_{hp}) + R_{sf}}, \quad (3.31)$$

with R_{sf} is the output impedance of the source follower. Z_f is the equivalent impedance of the feedback structure, and Z_{hp} the input impedance of the pulse shaper. Their expressions are respectively given by:

$$Z_f = \frac{R_f}{j\omega R_f C_f + 1} + \frac{1}{j\omega C_{in}} \quad \text{and} \quad Z_{hp} = \frac{j\omega C_{hp} R_{hp} + 1}{j\omega C_{hp}}. \quad (3.32)$$

From Equation 3.31, R_{sf} should be small compared to the load impedance to limit the impact of the latter on the transfer function.

3.4.2 Design flow

As a practical rule, the buffer output impedance should be one order of magnitude below the load resistance. If the feedback resistance R_f is small compared to the pulse shaper input impedance, one can write:

$$g_{m1,sf} \geq \frac{10}{R_f}. \quad (3.33)$$

The sizing of the output buffer is summarized in Figure 3.10. $g_{m1,sf}$ is found by taking the equality from Equation 3.33. The current flowing in the buffer branch $I_{D1,sf}$ is retrieved for a given $[g_m/I_D]_{1,sf}$. The latter should be high to obtain a large $g_{m1,sf}$ while limiting the power

consumption. The value of $[g_m/I_D]_{1,sf}$ is upper limited by the maximum transistor width, whereas $I_{D1,sf}$ should be larger than the maximum input current to drive the signal. M2sf can be linked to the main bias circuit by mirroring the current from M6b (Figure 3.2). The sizing of the bias circuit sets $(W/L)_b$ and I_{ref} . It gives the size of M2sf to obtain the desired current $I_{D1,sf}$. The load capacitance of the folded cascode is set by the parasitic capacitance of the buffer input gate:

$$C_{sf} \approx C_{gs1,sf}, \quad (3.34)$$

where increasing $[g_m/I_D]_{1,sf}$ enlarges the transistor width and the load capacitance. As a result, the amplifier requires more power to maintain the same GBW .

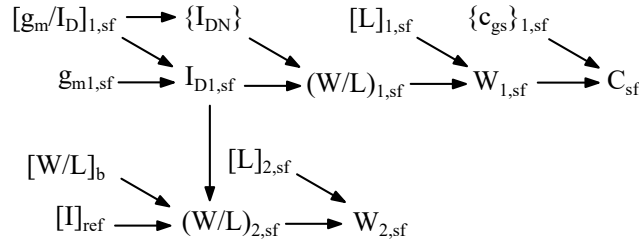


Figure 3.10: Design flow using the g_m/I_D methodology to obtain the buffer parameters. In brackets ([]): fixed parameters. In braces ({ }): parameters recovered from the simulated curves.

Although the source follower benefits from a simple implementation, it has drawbacks. It reduces the dynamic swing by one V_{TH} , and the power consumption increases linearly with the current $I_{D1,sf}$ [73]. More advanced architectures exist, such as the class AB buffer or push-pull amplifier. Both rely on a dynamic current which lowers the static power consumption, at the expense of a more complex structure and lower dynamic range [74].

3.5 Feedback structure

3.5.1 Feedback resistance

The feedback resistor fixes the transimpedance gain and directly impacts the bandwidth. Considering the amplifier as a black box, its value should be one order of magnitude greater than the buffer output impedance. The maximum input current sets the upper limit of the resistance, as increasing the gain reduces the input dynamic range. Figure 3.11 shows its impact on the transient and frequency responses. In the time domain, a higher resistance increases the gain and setup time. In the frequency domain, it leads to a decrease in bandwidth and peaking. In further works, it could be possible to implement a voltage-controlled feedback structure [75], enabling more versatility of the transimpedance amplifier.

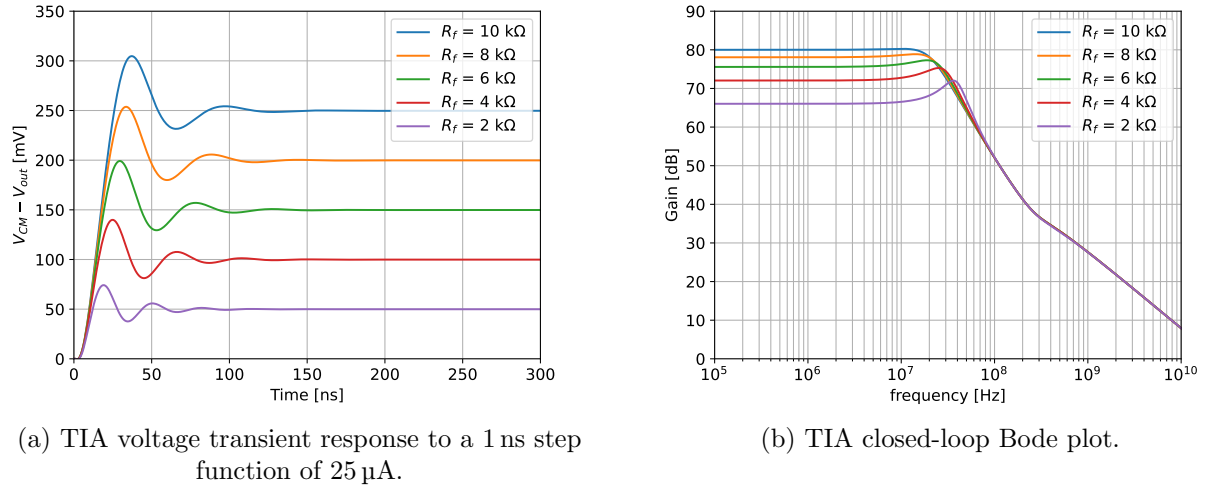


Figure 3.11: Parametrization of the feedback resistance. Considering an operational amplifier with $GBW = 100$ MHz, $A_v = 60$ dB and with $C_f = 1$ pF.

3.5.2 Feedback capacitance

The feedback capacitance C_f is added for stabilization purposes. It adds a zero in the transfer function to counteract the pole associated with the feedback resistance and the input capacitance. Figure 3.12 illustrates the trade-off of C_f on the transient and frequency responses. In the time domain (Figure 3.12a), a lower capacitance results in a shorter setup time but in a greater overshoot. In the frequency domain (Figure 3.12b), it increases the bandwidth but causes peaking. The latter can be an issue as it amplifies the noise of the TIA.

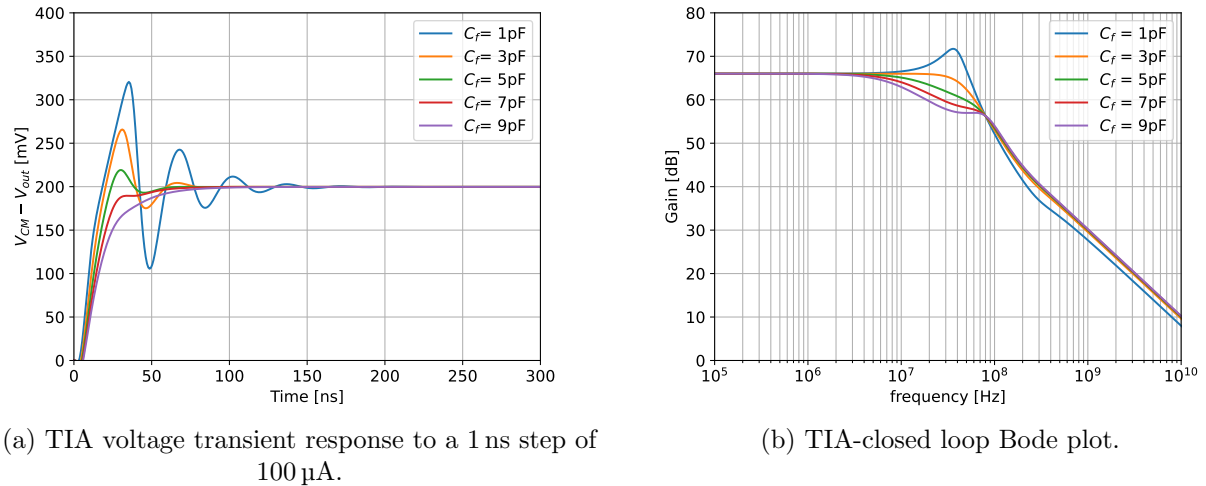


Figure 3.12: Parametrization of the feedback capacitance. Considering an operational amplifier with $GBW = 100$ MHz, $A_v = 60$ dB and with $R_f = 2$ kΩ.

3.6 Conclusion of the chapter

The specifications of the preamplifier are assessed for the application of fast SiPM readout. The requirements include high bandwidth, the capability to drive the input current, low input impedance, and low power consumption. A suitable architecture is identified accordingly: the transimpedance amplifier benefits from high bandwidth and low input impedance, while the folded-cascode topology provides a sufficient gain with reasonable power consumption. A buffer stage is added using a source follower to reduce the output impedance and increase the driven current. The main drawback of the proposed structure is its limited output swing. Hence, low-voltage threshold transistors and low-voltage bias circuits are employed to mitigate this effect.

This chapter provides a comprehensive design flow for the first amplification stage at the transistor level using the g_m/I_D methodology. It explores the tendencies and trade-offs related to the transistor parameters and their impact on the amplifier performance. Moving forward, the amplifier reference current still needs to be designed. The buffer structure is also inefficient regarding power consumption. If the latter is critical, more advanced architecture could be explored, such as class AB or push-pull buffers. The flexibility of the transimpedance gain could also be improved by relying on a tunable feedback structure. Finally, the subsequent blocks of the chain need to be designed, including the pulse shaper.

Chapter 4

Readout Implementation

After establishing the methodology and design guidelines, the amplifier can be sized. The initial estimation of the dimensions is performed using a *Python* script. Then, the values are finely tuned at the schematic level using *Cadence Virtuoso*. In the second step, the amplifier design is validated and fully characterized. In particular, its performances for photon counting applications are assessed. The simulations of the readout are conducted by coupling the designed transimpedance amplifier with the proposed SPICE SiPM model.

4.1 Amplifier characterization

4.1.1 Final design

To start the design flow, the length of each transistor, the minimum feedback resistance, and the GBW have to be fixed. The XT018 technology sets the minimum transistor size to $0.220\text{ }\mu\text{m}$ and the maximum size to $200\text{ }\mu\text{m}$. By default, L is set to $0.5\text{ }\mu\text{m}$. The transistor length should be low to minimize the overall size, but it lowers the Early voltage and increases the mismatch. An idea of the required GBW is given by the relation linking the bandwidth to the GBW , which is recalled as:

$$BW_{TIA} = \sqrt{\frac{GBW}{2\pi R_f C_{in}}}. \quad (4.1)$$

As mentioned in the specifications (Section 3.1.3), the closed-loop bandwidth should be higher than 10 MHz. Considering a maximum feedback resistance of $10\text{ k}\Omega$ ($80\text{ dB}\Omega$) and $C_{in} = 4.3\text{ pF}$, the lower bound of the GBW is set to 30 MHz. To achieve a wider bandwidth and allow for additional margins, the GBW is set to 100 MHz.

The output stage is designed first, as it provides the load capacitance for the folded cascode. The minimum feedback resistance is set to $2\text{ k}\Omega$. It limits the pressure on requiring a high $g_{m1,sf}$ and high bias current. The sizing of the output stage is summarized in Table 4.1. A large $[g_m/I_D]_{1,sf}$ is set to lower the power consumption, which induces a large width close to the technology limit. The buffer sets the maximum current that the amplifier can drive to $226\text{ }\mu\text{A}$.

	g_m/I_D [/V]	g_m [μ S]	I_D [μ A]	W [μ m]	L [μ m]
M1sf	18.2	4.12	226	150	0.5
M2sf	17.1	3.87	226	120	0.5

Table 4.1: Sizing of the output buffer retrieved from the *Cadence* schematic.

The design of the folded cascode is then achieved and summarized in Table 4.2. In particular, a low $[g_m/I_D]_1$ is preferred to obtain a large slew rate since the application requires sensing high peak pulses. It is favored at the cost of a smaller gain contribution of the first stage. Consequently, $[g_m/I_D]_{5,7}$ are higher to enhance the overall gain. The length of the current reference M11 is enlarged for better matching, as well as M9-M10 for noise purposes. One can see that the current mirrored by M9-M10 is reduced compared to the reference current. The mismatch is due to the Early effect. A relatively low Early voltage and the difference of V_{DS} among the transistors lead to a mismatch between the reference and the copied current. A lower current in the output branch results in a higher gain by increasing the output impedance. However, it is less optimal regarding the slew rate. If the latter is critical, one could either adjust $(W/L)_9$ to retrieve the desired current or increase the Early voltage by increasing the length of M9-M10.

	g_m/I_D [/V]	g_m [μ S]	I_D [μ A]	W [μ m]	L [μ m]
M1-M2	13	196	15	10	0.5
M3-M4	14.4	148	10.3	10	0.5
M5-M6	17.8	182	10.3	20	0.5
M7-M8	19.2	192	10.3	8	0.5
M9-M10	16.4	416	25.3	30	1
M11	9.7	291	30	50	2

Table 4.2: Sizing of the folded-cascode amplifier retrieved from the *Cadence* schematic.

Finally, the (W/L) ratios of the bias circuit are obtained from the sizing of the folded cascode. In particular, the sizes are enlarged to enhance robustness against eventual variations and mismatches.

	g_m/I_D [/V]	g_m [μ S]	I_D [μ A]	W [μ m]	L [μ m]
M1b-M2b	9.7	291	30	50	2
M3b	6.1	92	15	25	2
M4b	8	238	30	8	2
M5b	14	208	15	4	0.5
M6b	16.4	489	30	60	2
M7b	16.4	245	15	30	2
M8b	8.8	266	30	30	1.5
M9b	8.5	128	15	30	3
M10b	6.2	186	30	20	2
M11b	13	194	15	10	0.5
M12b	16.3	490	30	60	2
M13b	15.8	237	15	30	2

Table 4.3: Sizing of the bias circuit retrieved from the *Cadence* schematic.

The voltage output swing from the simulated *Cadence* schematic ranges from 0.3 V to 1 V. The common mode is set at about the midpoint of this range, with $V_{CM} = 0.6$ V. The extraction of the power supply rejection ratio (PSRR) and common mode rejection ratio (CMRR) are described in Section [A.6](#). At DC, they equals 63 dB and 120 dB, respectively. Power consumption is mainly governed by the output buffer, as it requires a substantial current to achieve high $g_{m1,sf}$ and drive the input. The total power of the amplifier is about 760 μ W, with the buffer accounting for 57 %.

4.1.2 Frequency response

The loop gain of the TIA is obtained from the *Cadence* schematic and depicted in Figure [4.1](#). It is compared to the transfer function retrieved from the *Python* script considering the same values of g_m/I_D . The feedback structure is done with $R_f = 2$ k Ω and $C_f = 3$ pF. The feedback capacitance is set in accordance with the resistance, considering the trade-off discussed in Section [3.5.2](#). Without any specification on the pulse shaper, the values of R_{hp} and C_{hp} are set to one order of magnitude larger than the components of the feedback structure ($R_{hp} = 10$ k Ω , $C_{hp} = 10$ pF).

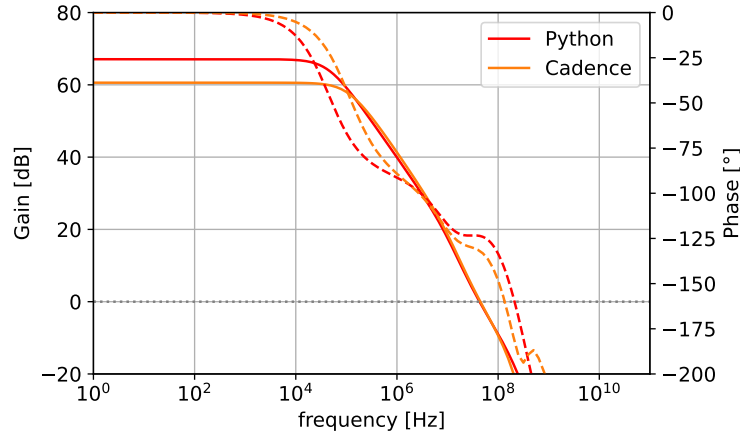


Figure 4.1: Comparison of the loop gain $A_v\beta$ between the *Python* script and the *Cadence* schematic simulation.

The loop gain $A_v\beta$ extracted from the schematic simulation has a DC gain of 61.2 dB with a GBW of about 43 MHz. The script presents a similar value of GBW but shows an upper shift of the DC gain by 7 dB and a larger dominant pole. It is due to an overestimation of the folded-cascode output impedance. This difference can be associated with the estimation of the Early voltage from the simulated curves. The V_{EA} is retrieved considering an equal V_{DS} amongst all the transistors of the output branch, that is, 0.45 V. It is only an approximation, and the distribution of V_{DS} is rather unbalanced within the branch. The Early voltage depends on V_{DS} (Figure 4.2), where the slope is steeper at low V_{DS} . Hence large unbalanced values of V_{DS} lead in a lower equivalent Early voltage, reducing both output impedance and gain.

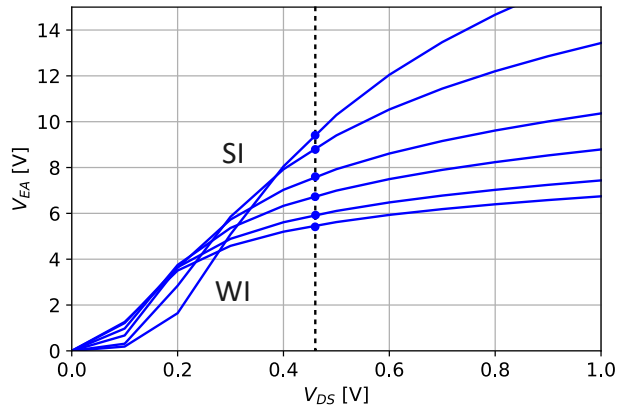


Figure 4.2: Simulated Early voltage for NMOS with $L = 0.5 \mu\text{m}$. $g_m/I_D = 8 \text{ V}^{-1}$ to 28 V^{-1} per 4 V^{-1} step. The blue dots indicate the values used in the *Python* script.

The non-dominant poles of the folded cascode extracted from the *Python* script are:

$$\omega_{p1} = 3.5 \text{ GHz}, \quad \omega_{p2} = 0.5 \text{ GHz}, \quad \omega_{p3} = 1.2 \text{ GHz}, \quad \omega_{z3} = 2.4 \text{ GHz}. \quad (4.2)$$

All of them are above f_T . Considering the pole and zero introduced by the feedback structure, the estimated phase margin of the loop gain is 56.4° . In contrast, the phase margin in *Cadence* is about 46.4° , reaching the good practice of 45° mentioned in Section 3.2.1.

The closed-loop transfer function (Figure 4.3a) has a DC transimpedance gain of $66 \text{ dB}\Omega$ with a bandwidth of about 48 MHz. This value is closed from the predicted bandwidth of 43 MHz, indicated by Equation 4.1. Figure 4.3b illustrates the input impedance of the TIA. It benefits from a low DC impedance of 2Ω . However, the feedback structure induces a peaking of the impedance at a higher frequency. At 43 Mhz, it reaches 700Ω , which will affect the fast component of the input signal.

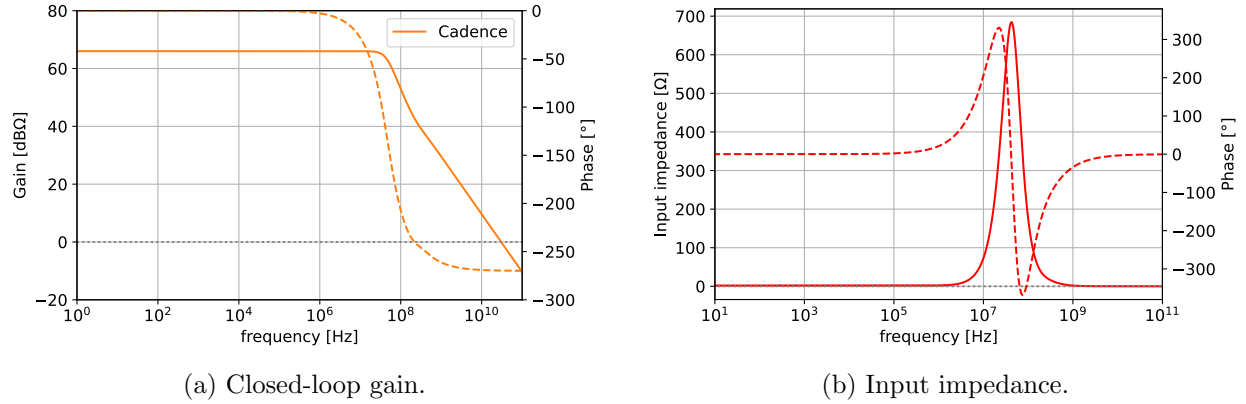


Figure 4.3: Frequency response of the closed-loop TIA, with $R_f = 2 \text{ k}\Omega$ and $C_f = 3 \text{ pF}$.

4.1.3 Transient response

Figure 4.4 depicts the TIA transient response to different input currents with a 1 ns step function. The response is underdamped. It exhibits a slew rate of 10 mV/ns and an overshoot of about 32% at the maximum considered current ($100 \mu\text{A}$). The amplifier stabilizes to the desired value characterized by $V_{CM} - V_{out} = I_{in}/R_f$. However, as the amplifier operates in pulse amplification, one should ensure its transient response is still stable in this configuration.

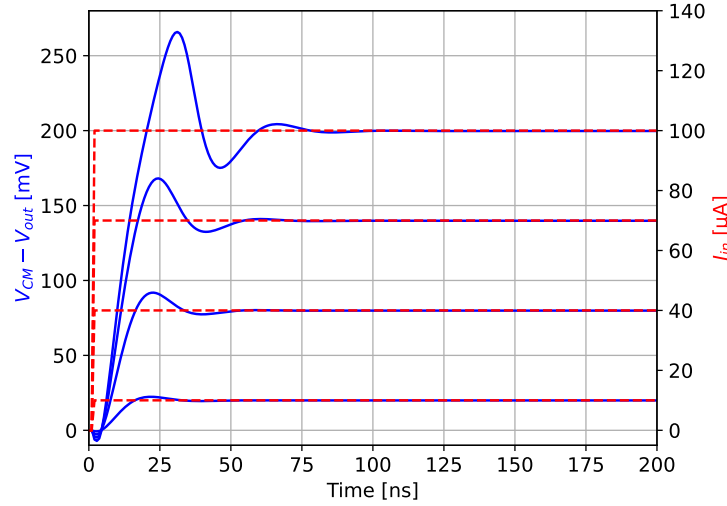


Figure 4.4: TIA voltage transient response (blue) to a 1 ns step function from 10 μA to 100 μA by step of 30 μA (red). $R_f = 2\text{ k}\Omega$ and $C_f = 3\text{ pF}$.

4.1.4 PVT corners

In practice, process, voltage, and temperature variations occur. The circuit should be robust to these deviations, and the design should ensure its functionality within a given variation range.

Process variation refers to the overall variability induced by the fabrication process, which causes differences at the wafer-to-wafer or die-to-die level [76]. This variability arises due to tolerances in various process aspects, such as sizes, oxide thickness, and impurity concentration. These tolerances directly impact the transistor parameters, including the threshold voltage V_{TH} and mobility. Four worst-case process scenarios are compared to the typical case (TT). These scenarios involve both NMOS and PMOS transistors being either fast (FF), both slow (SS), NMOS fast, and PMOS slow (SF), or conversely (FS). The process variation affects all PMOS and NMOS transistors in the same manner across the wafer. The transistors exhibit higher mobility and a lower threshold voltage in the fast corner scenario. In contrast, the transistors have lower mobility and a higher threshold voltage in the slow corner scenario. Table 4.4 presents the performance of the amplifier for these four scenarios.

Corner	TIA bandwidth [MHz]	TIA loop gain [dB]	Phase margin [$^\circ$]
TT	48	61.2	46.4
FF	49.8	61.3	45.9
SF	48.9	61.8	46.1
SS	45.6	60.8	46.5
FS	47	60.3	46.6

Table 4.4: Summary of process corner performances.

On top of that, there is a random statistical variability among the individual transistors in the circuit. A Monte Carlo simulation is employed to capture these variations. It considers that each transistor parameter follows a normal distribution, where a realization is performed at each run. Figure 4.5 illustrates the distribution of the amplifier performance obtained through a Monte Carlo simulation conducted over 1000 runs in the case of a typical process (TT).

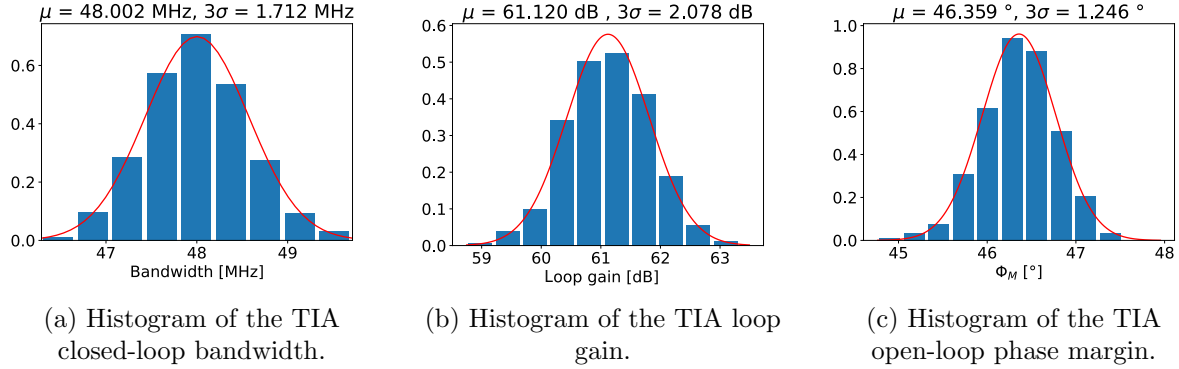


Figure 4.5: Monte Carlo mismatch simulation over 1000 runs.

The default temperature is set to 27°C. As for the process variation, temperature impacts both the threshold voltage and the mobility. However, its effect on the $I_D - V_{GS}$ relationship differs (Figure 4.6). Reducing the temperature increases both the threshold voltage and the mobility, having opposite effects on the drain current. The zero temperature coefficient (ZTC) refers to the point at which these two contributions compensate each other. Before reaching this value, the effect of the threshold voltage dominates, resulting in a lower drain current when T decreases. After the ZTC, the relationship is dominated by mobility, resulting in a higher drain current.

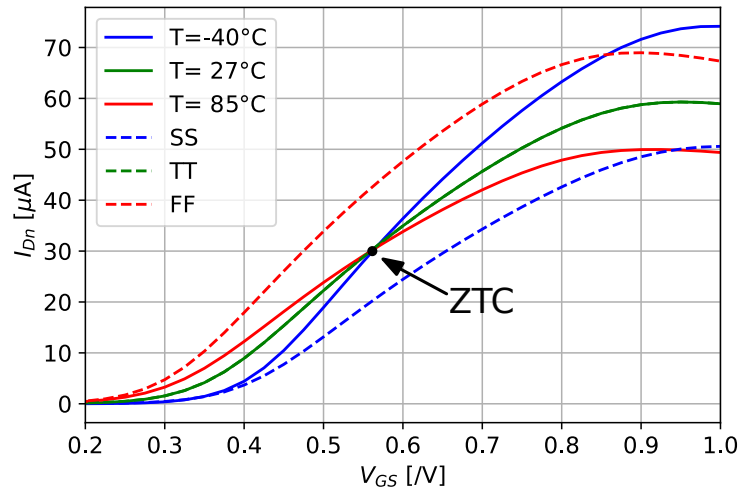


Figure 4.6: Impact of the process and temperature on the drain current (in case of NMOS, $L = 0.5 \mu\text{m}$, $V_{DS} = 0.45$ V).

One can compare the impact of process and temperature variations on the loop gain as they impact similar transistor parameters. At low temperatures (Table 4.5), the gain is higher than in the typical case (27 °C), whereas the gain is lower in the SS scenario (Table 4.4). It can be associated with $V_{GS,1}$, where its value is above the ZTC point. It results in a higher g_{m1} at low temperatures, increasing the gain. In contrast, in the SS scenario, g_{m1} is lower than in the typical case, leading to a lower gain.

T [°C]	TIA bandwidth [MHz]	TIA loop gain [dB]	Phase margin [°]
-40	48.2	61.6	47.3
27	48	61.2	46.4
85	47.8	60.5	45.3

Table 4.5: Summary of temperature corner performances.

Finally, Table 4.6 summarizes the sensitivity of the amplifier regarding the supply voltage. The considered range is $\pm 5\%V_{DD}$.

V_{DD} [V]	TIA bandwidth [MHz]	TIA loop gain [dB]	Phase margin [°]
1.71	47.7	60.9	47.7
1.8	48	61.2	46.4
1.89	48.7	61.33	46.3

Table 4.6: Summary of voltage corner performances.

The PVT corners are evaluated individually, but one should ensure that the amplifier still operates when combining the worst scenarios. In particular, the critical cases regarding the loop gain and phase margin are assessed in Table 4.7. The lowest loop gain shows a drop of 1.7 dB compared to the typical case, whereas the critical case in terms of phase margin is 44.38°.

Worst case	Setup	TIA bandwidth [MHz]	TIA loop gain [dB]	Phase margin [°]
Gain	FS, 85 °C, 1.71 V	46.5	59.5	45.5
Phase margin	FF, 85 °C, 1.89 V	51.1	60.72	44.38

Table 4.7: Performances when combining the worst corners regarding gain and phase margin.

4.1.5 Noise

The expression of the input noise PSD of the operational amplifier S_a is given by Equation 3.11. The associated output noise PSD for the TIA is found using the following equation:

$$S_{a,o} = \frac{1}{\beta} \frac{S_a}{1 + \frac{1}{A_v\beta}}, \quad (4.3)$$

where calculation details are provided in Section A.5. Equation 4.3 shows that the thermal noise is enhanced by a peaking effect due to the $1/\beta$ factor, as illustrated in Figure 4.7a. The low-pass filter

induced by the amplifier cuts the noise at higher frequencies. Although flicker noise dominates at low frequencies, its overall contribution is negligible compared to the thermal noise due to the high bandwidth operation and the peaking effect. The total root mean square (RMS) output noise σ_{out} is computed by integrating the output noise PSD:

$$\sigma_{out} = \sqrt{\int_{f_{min}}^{f_{max}} S_{a,o}(f) df} \quad \text{with} \quad \begin{cases} f_{max} = 10 \text{ GHz} \gg f_T, \\ f_{min} = 1 \text{ Hz}. \end{cases} \quad (4.4)$$

The total RMS output noise of the TIA is 392 μVrms . To better understand the impact of each noise contribution, the PSD can be integrated by part, as shown in Figure 4.7a. The integration bound between flicker and thermal noise is set by the corner frequency f_c , at which the two noise sources have equal contributions. Its value is about 100 kHz. The flicker noise accounts for 2 μVrms (0.5 %), the filtered thermal noise for 88 μVrms (22.5 %), and the enhanced noise due to peaking for 302 μVrms (77 %) on the total RMS output noise. Figure 4.7b shows that the noise contribution of the mirror pairs (M3-M4 and M8-M9) is in the same order of magnitude compared to M1-M2. If the noise is critical and based on the design guidelines (Table 3.2), enlarging the size of M1-M2 while reducing $[g_m/I_D]_{3,9}$ would decrease the thermal noise.

One should also consider the impact of the feedback structure on the noise. The thermal noise induced by the feedback resistor is given by:

$$S_R = 4kTR_f, \quad (4.5)$$

where a higher resistance increases its associated thermal noise. It is possible to reduce the noise enhancement by adjusting the components of the feedback structure. Recalling the expression of $1/\beta$:

$$\frac{1}{\beta} = \frac{1 + j\omega R_f(C_{in} + C_f)}{1 + j\omega R_f C_f}. \quad (4.6)$$

Larger C_f reduces the high-frequency gain of the feedback structure ($= 1 + C_{in}/C_f$) at the cost of lower bandwidth. Similarly, a lower input capacitance reduces the noise enhancement.

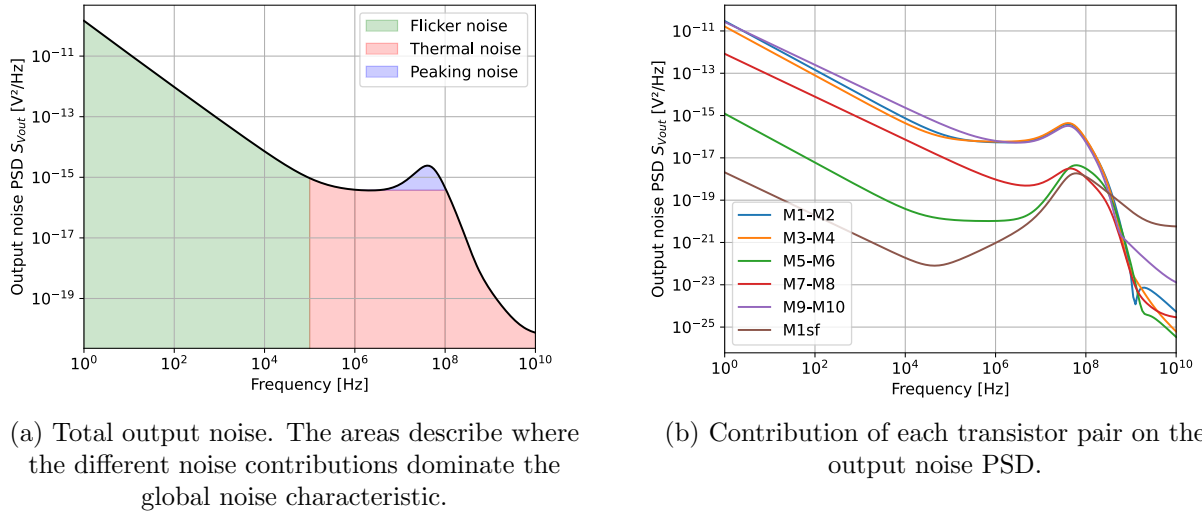


Figure 4.7: TIA output noise power spectral density (PSD) from the *Cadence* schematic simulation.

4.1.6 Comparison with the literature

Table 4.8 compares the performance of the proposed amplifier with TIAs used for SiPM readout found in the literature. Considering a feedback resistance of $2\text{ k}\Omega$, the presented TIA achieves a transimpedance gain of $66\text{ dB}\Omega$ and a bandwidth of 48 MHz , where the operational amplifier consumes $760\text{ }\mu\text{A}$. The main difference between the proposed design and the literature lies in the considered transimpedance gain. It is attributed to the maximum input current that the device should handle. This work accommodates currents up to $100\text{ }\mu\text{A}$ without falling out of the dynamic range. In contrast, the other amplifiers are designed to read lower pulse currents, which justifies a higher transimpedance gain. Compared to the other design with sub-mW power consumption, this circuit reaches higher bandwidth thanks to the lower feedback resistance.

For better comparison, the gain of the proposed TIA can be set to a similar value using a resistance of $10\text{ k}\Omega$ ($80\text{ dB}\Omega$). Considering a similar output range as for the former case, I_{max} is set to $25\text{ }\mu\text{A}$. With a feedback capacitance of 1 pF , it achieves a bandwidth of 24.4 MHz , a phase margin of 51.7° , and a output noise of $506\text{ }\mu\text{Vrms}$. However, this configuration is sub-optimal regarding the power consumption as the internal design is done for $R_f = 2\text{ k}\Omega$. Redesigning the source follower accounting for lower pulses would alleviate the need for a large current in the output branch. It would decrease the overall required power, making the amplifier more competitive.

Refs.	Process	VDD	Topology	Input capacitance	Max. input current	Transimpedance gain	Bandwidth	Phase Margin	Noise	Power
[31]	0.35 μm	3.3 V	Regulated cascode	10 pF	2 μA	106 dB Ω	15.9 MHz	-	6.9 mVrms	660 μW
[42]	0.13 μm	1.2 V	Regulated CG	300 pF	22.5 μA	86 dB Ω	12 MHz	60°	2.7 mVrms	336 μW
[43]	0.18 μm	1.8 V	Three stage CS	2 pF	-	83 dB Ω	1.75 GHz	66°	476.6 μVrms	48.6 mW
This work	0.18 μm	1.8 V	FC	4.3 pF	100 μA	66 dB Ω	48 MHz	46.3°	392 μVrms	760 μW
	0.18 μm	1.8 V	FC	4.3 pF	25 μA	80 dB Ω	24.4 MHz	51.7°	506 μVrms	760 μW

Table 4.8: Comparison of TIA performances for SiPM readout application.

4.2 Modeled SiPM readout

The SiPM readout can be simulated by combining the amplifier with the proposed SPICE model from Chapter 2. It allows the characterization of the amplifier response to the signal induced by the SiPM after a light pulse.

4.2.1 Setup

Figure 4.8 presents the front-end architecture associated with the SiPM model. The latter comprises $N = 234$ cells, with n fired cells. A summary of all the considered parameters is presented in Table 4.9. The values of the SiPM model are retrieved from Chapter 2, while the values associated with the amplification chain are obtained from Chapter 4. The circuit is directly connected to the anode of the SiPM. It sets the value of the SiPM anode voltage to the common mode of the amplifier ($V_{CM} = 0.6 \text{ V}$). To keep the same excess voltage as previously considered for the SiPM of about 3.1 V, V_A is set to 10.1 V. The output signal V_{out} is taken directly from the preamplifier output. V_{hp} refers to the signal after the passive first-order high-pass filter from the pulse shaper.

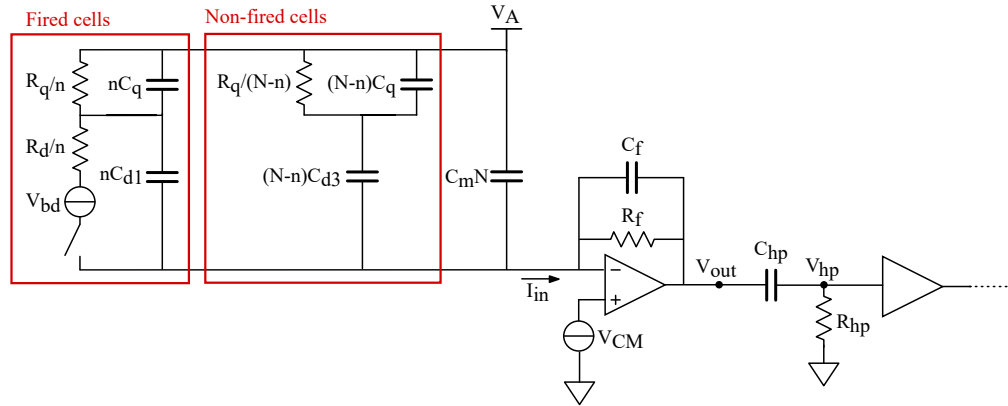


Figure 4.8: Schematic of the front-end circuit considered to simulate and read out the transient response of the SiPM to a light pulse.

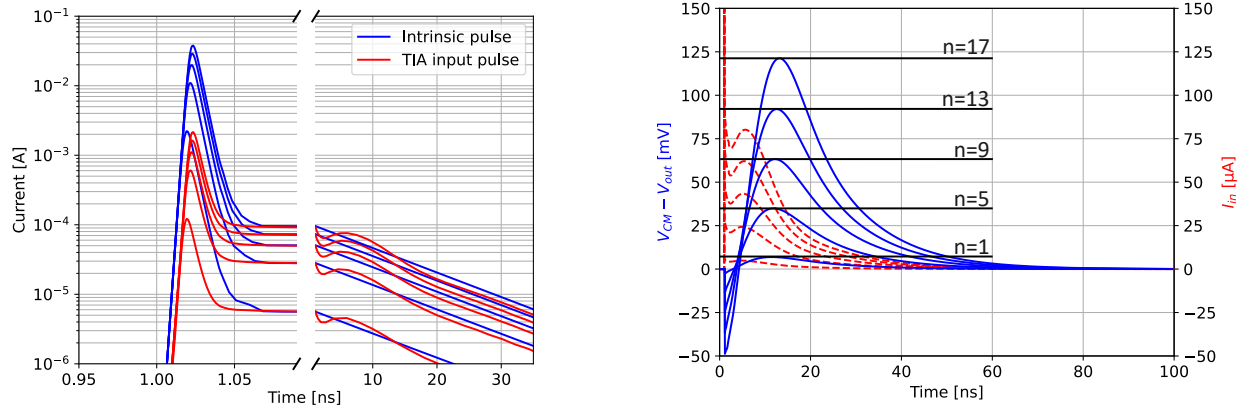
Voltage	[V]	Resistance	[Ω]	Capacitance	[F]
V_A	10.1	R_d	142	C_{d1}	22f
V_{bd}	6.3	R_q	440k	C_{d3}	0.6p
V_{CM}	0.6	R_f	2k	C_m	13.5f
V_{DD}	1.8	R_{hp}	10k	C_q	6f
				C_f	3p
				C_{hp}	10p

Table 4.9: Overview of the parameter values used in the front-end schematic.

4.2.2 Pulse response

Figure 4.9b shows the impact of the amplifier input impedance on the input current pulses, compared to pulses generated without any load. Because of the TIA structure, its input impedance is enhanced at high frequency (Figure 4.3b), affecting the fast response of the signal. It decreases the rising slope and amplitude of the initial peak (Figure 4.9b, left). Regarding the slower decay of the pulses, it shows underdamping (Figure 4.9b, right). It is due to the feedback structure, as already observed in the transient response to a step function (Figure 3.12a).

The response of the transimpedance amplifier for several simultaneously fired cells is presented in Figure 4.9b. The first pulse width is too short to be sensed, but the amplifier can detect the slower decay of the signal. Due to the restricted slew rate, the oscillation observed at the input is not transcribed in the readout signal. The signal peak occurs at 10ns after the initial light pulse, with an associated FMWH of about 15.7ns. The response is proportional to the number of simultaneously fired cells. The minimum detectable resolution, equivalent to detecting a single photon, is about 6.75 mV. As the amplifier 3σ output noise is around 1.18 mVrms, the signal-to-noise ratio is sufficient to distinguish a single photon event from the output noise.



(a) Pulses without any load (blue) and affected by the input impedance of the readout circuit (red).

(b) Pulses at the input (red) and output (blue) of the transimpedance amplifier.

Figure 4.9: Transient response of the SiPM model and front-end circuit (Figure 3.1), before (a) and after (b) the first amplification stage. $N = 234$, from $n = 1$ to $n = 17$ fired cells by step of 4.

As concurrent photons could happen in practice, one can evaluate the maximum number of fired cells the amplifier can drive. As seen in Figure 4.10a, the slew rate and pulse shape degrade after about $n = 60$. One can better investigate the linearity of the response from Figure 4.10b. The red line shows the ideal relationship between the number of fired cells and the output voltage. In the first step, the output peak amplitudes display a slight distortion, showing a higher slope than the linear curve. It is associated with the non-linear trend of the SiPM response itself, as highlighted by the normalized input current peaks (green curve). The SiPM impedance depends on the number of fired cells. More fired cells decrease the overall capacitance as they exhibit a lower capacitance (C_{d1}) compared to the quiescent cells (C_{d3}), resulting in higher pulses. In the second step, two phenomena related to the operational amplifier dominate the distortion. After about 40 photons, the output node fall below the output swing. Consequently, some transistors no longer remain in saturation, degrading the linearity of the gain. At 70 photons, the signal experiments saturation. It is associated with the maximum current the output buffer can drive, which corresponds to 226 μA .

Given the maximum current peak of 100 μA , 28 photons can be simultaneously detected. At this value, the error compared to the ideal curve is about 4.5 %.

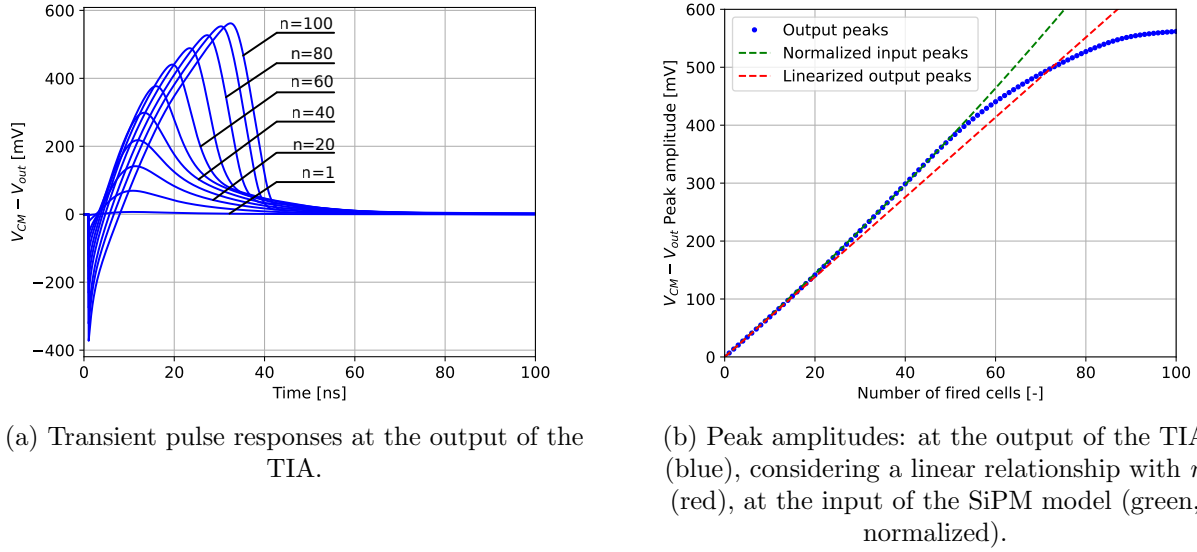


Figure 4.10: Characterization of the TIA output response regarding the number of simultaneously fired cells n from the SPICE model.

Figure 4.11 illustrates the amplifier response following the pulse shaper high-pass filter. This stage reduces the tail of the pulse, resulting in a narrower width for improved timing precision. It cuts the DC component, lowers the noise, and limits the pile-up effect. Due to its passive implementation, the filter slightly decreases the amplitude of the signal peak. Still, the signal can be further amplified by the next stages of the pulse shaper. The response presents an undershoot caused by the zero associated with C_{hp} , which can be mitigated using a pole-zero cancellation resistor. Note that the presented filtering uses an arbitrary cut-off frequency of 1.6 MHz. The bandwidth should be set in accordance with the requirements in terms of pulse detection frequency.

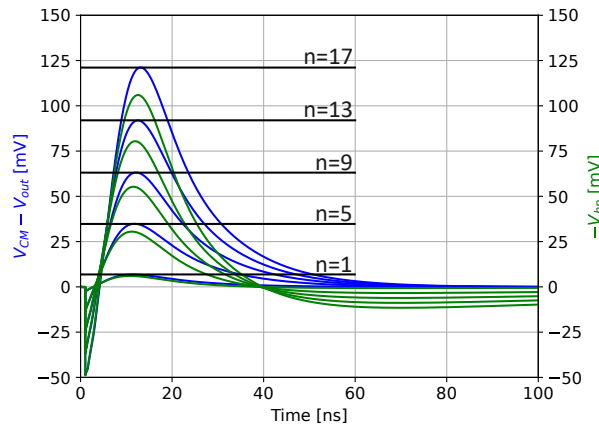


Figure 4.11: Pulses at the input (blue) and output (green) of the first-order passive high-pass filter of the pulse shaper.

4.3 Conclusion of the chapter

The amplifier is sized considering the specifications for a fast readout of the studied SiPM. The performance of the simulated preamplifier stage is assessed using *Cadence* schematic simulations. It exhibits a transimpedance gain of $66 \text{ dB}\Omega$ with a bandwidth of 48 MHz and a slew rate of 10 mV/ns . The amplifier generates an output referred noise of $392 \text{ }\mu\text{Vrms}$, dominated by thermal noise and the peaking of the feedback structure. It consumes $760 \text{ }\mu\text{W}$, where half of the power is associated with the buffer. Power consumption could be mitigated by considering similar current peaks reported in the literature, resulting in a lower input current and a higher load impedance. It would reduce the output branch current to the same order of magnitude as the other branches of the circuit, limiting the impact of the buffer on power consumption. Compared to the literature, the presented TIA sets a lower transimpedance gain to trade a larger bandwidth. Amplifiers with comparable power consumption rather reach the tens of MHz with gains above $80 \text{ dB}\Omega$. Additionally, the circuit proves to be robust against PVT corners, as they do not significantly impact the performance of the amplifier. Other non-idealities still need to be tested. Further simulations, including at the layout level, should be conducted to ensure that the performance still meets the requirements.

The amplifier successfully detects and amplifies the simulated SiPM pulses. It allows the detection of single trigger events with an associated output peak of 6.75 mV . This pulse is higher than the 3σ output noise of 1.18 mVrms , showing a sufficient signal-to-noise ratio to guarantee the detection at the photon level. It can detect 28 simultaneously fired cells with a maximum error of 4.5% compared to the ideal characteristic. The primary distortion is associated with the dependency of the SiPM impedance on the number of fired cells. However, obtaining more precise application specifications is crucial. A better knowledge of the SiPM response, such as the pulse peaking and the number of simultaneous photons, would lead to more stringent specifications in terms of gain, maximum current, and bandwidth. The determination of the rate of event detection would also allow the design of the dedicated pulse shaper.

Conclusion and perspectives

In conclusion, this work presents the modeling of a SPAD and the design of a low-power analog readout circuit specifically tailored for photon counting applications.

In the first step, TCAD simulations are implemented to characterize the studied SPAD. It enables the extraction of its parameters through DC, AC, and transient simulations to set up an equivalent model. Based on these simulations, a novel SPICE model for SPADs is introduced to simulate the intrinsic transient response to a light pulse. It accurately reproduces the fast double-response of a photon excitation associated with drift and diffusion mechanisms. This compact model results in a conference paper that will be presented at the ESSDERC 2023 [77]. A SiPM model is derived to design a dedicated integrated readout circuit. Subsequent works could involve tuning the TCAD model parameters to fit the measured diode, whether in terms of dark current, gain factor, or breakdown voltage. It could also imply characterizing the transient response of the fabricated SPAD. The model parameters extracted from the measurements could be compared to the values obtained from the TCAD simulations. It could lead to a better estimation of the generated current, giving more realistic specifications for the readout circuit design.

In the second step, this work presents the design of the first amplification stage of the readout structure. The preamplification is done using a transimpedance amplifier with a folded-cascode topology. The former offers high bandwidth and small input impedance, while the latter benefits from sufficient gain for a limited power and acceptable output swing. The design flow relies on the g_m/I_D methodology, allowing working with the complete range of inversion levels. The design flow assesses the trade-off and the critical choices to achieve the desired performances. However, the structure suffers from a high static current requirement due to the passive output buffer. Relying on an active configuration or obtaining the exact specifications in terms of input current would leverage this need. The remaining tasks for the preamplifier include designing the current reference and implementing a tunable feedback structure. The subsequent steps of the readout chain also need to be developed, such as the pulse shaper and digitization stage.

Finally, the design choices of the preamplifier are derived to meet the specifications. The design is validated, and its performances are stated through schematic simulation. The readout of the SiPM is simulated and characterized using the proposed SPICE model. Yet, a better knowledge of the application is desired to ensure that the design meets the requirements. It includes the maximum number of simultaneous photons and the timing precision required between successive detections. It would allow better optimization of the preamplifier and establish the specifications of the dedicated pulse shaper. Concerning the validation of the circuit, the subsequent steps involve simulations at the layout level to ensure its robustness, considering additional parasitic elements. After the circuit

proves its reliability in simulation, the device can be fabricated and measured to assess its actual performance.

Appendix A

Appendix

A.1 SPAD equivalent width

The SPAD has a 3D circular shape, but the simulation in *Silvaco* is 2D. An equivalent width has to be retrieved from the circular shape to obtain similar proprieties between both structures. Since the SPAD has the shape of a rounded square (Figure A.1), the approximation of a quarter of the perimeter is analytically given by:

$$y(x, r) = r \sqrt[4]{1 - \left(\frac{x}{r}\right)^4}, \quad (\text{A.1})$$

where r is a given radius. The associated perimeter, $W(r)$, is set as the length of the curve $4y(x, r)$ from $x = 0$ to $x_1 = r$. Considering R the distance between the two electrodes of the diode (Figure A.1), the equivalent width $W_{eff}(R)$ is obtained by integrating $W(r)$ from 0 to R :

$$W_{eff}(R) = \frac{\int_0^R W(r) dr}{R}. \quad (\text{A.2})$$

The studied diode has a radius R of $19.5 \mu\text{m}$. In case of a complete revolution, $W_{eff_{360}} = 68.42 \mu\text{m}$, which is round up to $70 \mu\text{m}$. As the 2D diode layout is two-sided, the considered width is set to $W_{eff_{180}} = 35 \mu\text{m}$.

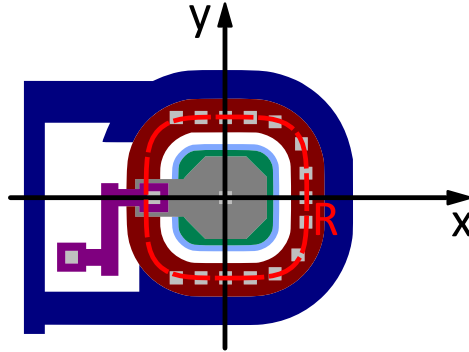


Figure A.1: Layout of the actual studied SPAD with the geometrical consideration to extract the equivalent width. Inspired from [4].

A.2 Implantation level

Figure A.2 demonstrates the effect of the implantation energy on the doping concentration. Increasing the energy increases the deep presence of impurities within the structure. As a result, it decreases the resistivity in the deeper region of the polysilicon, reducing the rod resistance. Additionally, it reduces the doping gradient within the material.

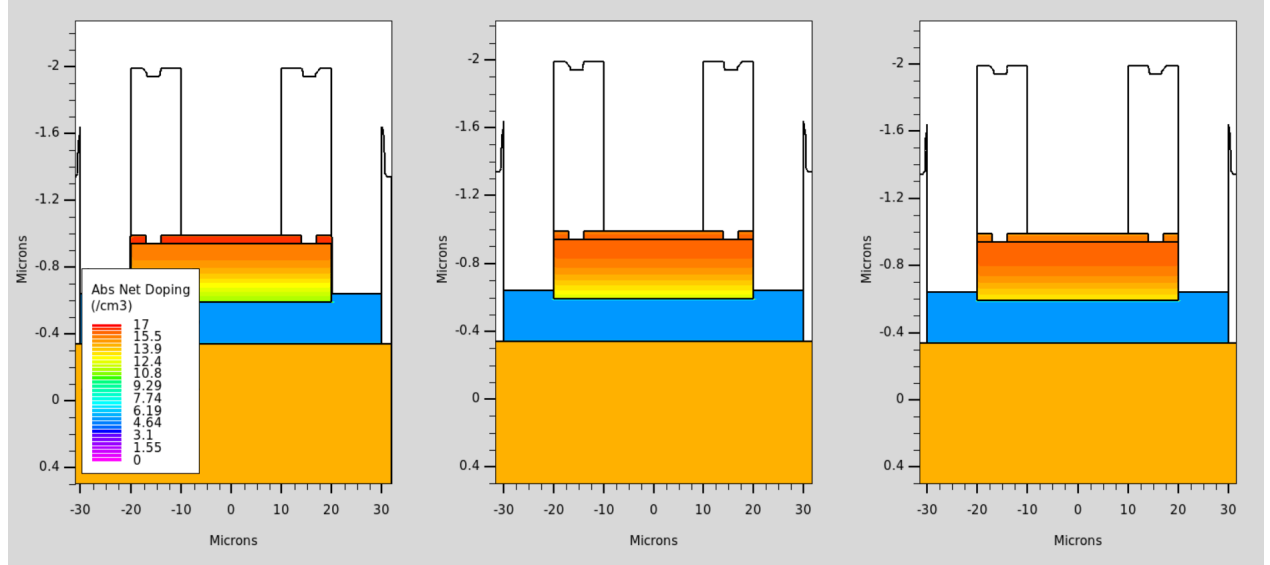


Figure A.2: TCAD simulation of the polysilicon resistor doping implementation level $\log(\text{cm}^3)$. $t_{ox} = 250 \text{ nm}$, Dose $= 4.5 \times 10^{11} \text{ cm}^{-3}$. From left to right, implantation energy of 10 keV, 20 keV and 30 keV, respectively.

A.3 Time constant vs. junction capacitance

The time constants of the transient response can be used to obtain the equivalent capacitances of the model. Although small-signal analysis could be used, a transient point of view would be preferable since the pulse is more characterized by an abrupt variation of voltage rather than a sine variation. From the equivalent model in Figure A.3:

$$\frac{V_{in}(t) - V_K(t)}{R_d} + \frac{V_A - V_K}{R_q} + C_q \frac{d(V_A - V_K(t))}{dt} = C_d \frac{dV_K(t)}{dt},$$

$$\Leftrightarrow \quad (A.3)$$

$$\frac{V_{in}(t)}{R_d} + \frac{V_A}{R_q} = (C_d + C_q) \frac{dV_K(t)}{dt} + \left(\frac{1}{R_q} + \frac{1}{R_d} \right) V_K(t).$$

The solution of this EDO is given by:

$$V_K(t) = A + B e^{-\frac{t}{\tau}} \quad \text{with} \quad \tau = (C_q + C_d)(R_q || R_d). \quad (A.4)$$

The initial conditions of the system are set by:

$$\begin{cases} V_K(t = 0^-) &= V_A, \\ V_K(t = \infty) &= V_{bd}. \end{cases} \quad (\text{A.5})$$

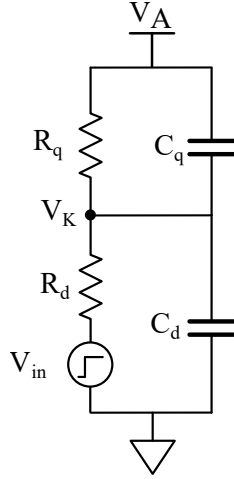


Figure A.3: Equivalent representation of the SPAD, where the voltage switch is considered as a step function to perform the transient analysis.

Putting it all together, it gives the expression of the cathode voltage $V_K(t)$:

$$V_K(t) = V_{bd} + V_{ex} e^{\frac{-t}{\tau}}. \quad (\text{A.6})$$

Once the cathode voltage is known, the output current $I(t)$ is retrieved as:

$$I(t) = C_q \frac{d(V_A - V_K(t))}{dt} + \frac{V_A - V_k(t)}{R_q}. \quad (\text{A.7})$$

By substituting $V_K(t)$ by its expression from Equation [A.6](#), it gives:

$$I(t) = V_{ex} \left(\frac{1}{R_q} + \frac{C_q}{\tau} \right) e^{\frac{-t}{\tau}} + \frac{V_{ex}}{R_q}. \quad (\text{A.8})$$

This last equation can be used to determine the time constant of the fast (τ_1) and slower (τ_2) characteristics of the pulse. By subtracting the constant term of V_{ex}/R_q from Equation [A.8](#), the slope of the natural logarithm directly gives these time constants. This methodology only applies before the diode recharge since the condition that $V_K(t = \infty) = V_{bd}$ is no longer valid afterward. By using points in the middle of the fast response and slower slope (Figure [A.4](#)), it is possible to estimate the fast and slow characteristics of the curve, respectively τ_1 and τ_2 , using:

$$\begin{cases} \tau_1 = -\frac{t_2 - t_1}{\ln(I'(t_2)) - \ln(I'(t_1))} \\ \tau_2 = -\frac{t_4 - t_3}{\ln(I'(t_4)) - \ln(I'(t_3))} \end{cases} \quad \text{with} \quad I'(t) = I(t) - \frac{V_{ex}}{R_q}. \quad (\text{A.9})$$

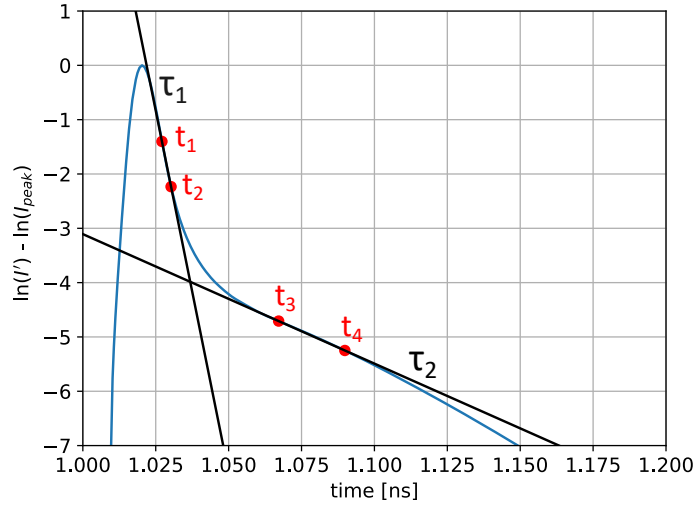


Figure A.4: Extraction of the fast and slower pulse time constants through the transient response. The current amplitude is normalized by the current peak I_{peak} .

Considering $R_q=440\text{ k}\Omega$ and $R_d=142\text{ }\Omega$, R_q is negligible: $(R_q||R_d) \approx R_d$. It gives $(C_q + C_d) = \tau/R_d$:

$$\begin{cases} \tau_1 = 3.5\text{ ps} & \implies C_q + C_{d1} \approx 18.7\text{ fF}, \\ \tau_2 = 42\text{ ps} & \implies C_q + C_{d2} \approx 295\text{ fF}. \end{cases} \quad (\text{A.10})$$

As the value of C_q is set to 6 fF, an idea of the internal capacitances can be obtained. C_{d1} is around 12.7 fF while C_{d2} is about 289 fF.

A.4 Quenching circuit

The impact of the quenching resistor on the response of the SPAD is depicted in Figure [A.5](#). Because of the RC delay induced by the parasitic capacitance and the quenching resistor, minimizing the latter reduces the recharge time. When the SPAD is quenched, the resistance does not impact the pulse shape, as it is short-circuited by the quenching capacitance. However, as seen in Figure [A.5a](#), a too low resistance is not able to quench the SPAD. The voltage drop induced by the resistor during quenching is not large enough to lower the diode cathode voltage down to the breakdown voltage. Hence, the generated current still ensures the self-sustaining mechanism.

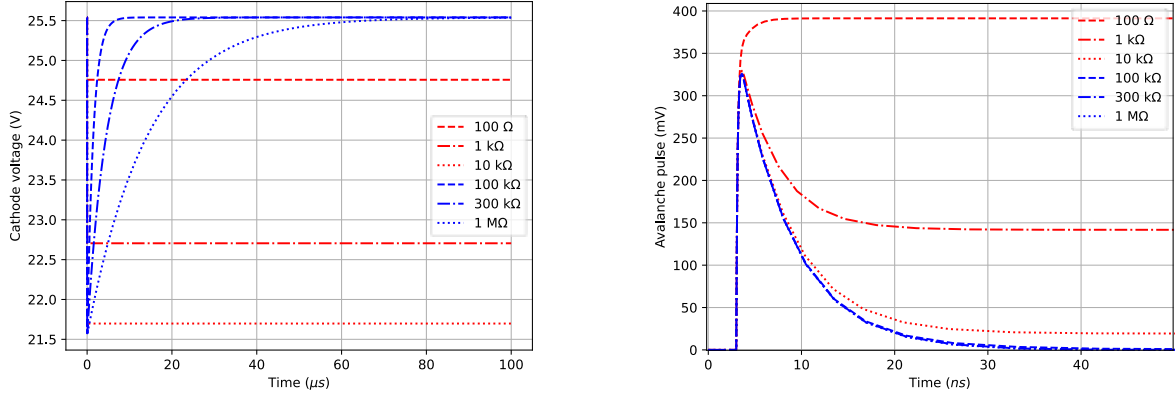
(a) Impact of R_q on the recharge time.(b) Impact of R_q on the transient response to a light pulse.

Figure A.5: Parametrization of the quenching resistor R_q based on a SPICE model of the SPAD. In red, the resistance is not able to quench the avalanche process.

A.5 Output referred noise

Considering a given transimpedance amplifier and its equivalent input noise voltage e_n (Figure A.6), one can obtain the equivalent output noise source e_o as follows:

$$-A_v(v_x + e_n) = e_o, \quad (\text{A.11})$$

with A_v the gain of the amplifier. v_x can be expressed as:

$$v_x = e_o \frac{j\omega R_f C_f + 1}{j\omega R_f (C_{in} + C_f)} = \beta e_o. \quad (\text{A.12})$$

The expression for e_o can be obtained by substituting the value of v_x found in Equation A.12 into Equation A.11:

$$A_v e_n = e_o (1 + A_v \beta) \quad \Leftrightarrow \quad e_o = e_n \frac{A_v}{1 + A_v \beta}. \quad (\text{A.13})$$

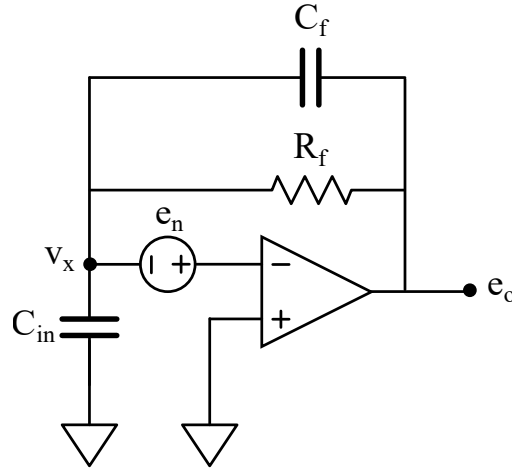


Figure A.6: Small-signal circuit analysis of the TIA to retrieve the equivalent output noise from the input equivalent noise voltage of the amplifier.

A.6 PSRR and CMRR

The power supply rejection ratio (PSRR) expresses the ability of the amplifier to suppress the power supply variation v_{dd} to its output signal v_{out} . It is expressed as:

$$PSRR[dB] = 20 \log_{10} \left(\frac{A_v}{A_{vdd}} \right) \quad \text{with} \quad A_{vdd} = \frac{v_{out}}{v_{dd}}. \quad (\text{A.14})$$

Because the negative supply of the amplifier is connected to the ground, only the value of the positive PSRR is required. Similarly, the common mode rejection ratio (CMRR) refers to the ability of the circuit to suppress the common mode variation v_{cm} to its output signal v_{out} . It is given by:

$$CMRR[dB] = 20 \log_{10} \left(\frac{A_v}{A_{cm}} \right) \quad \text{with} \quad A_{cm} = \frac{v_{out}}{v_{cm}}. \quad (\text{A.15})$$

The PSRR and CMRR are derived from the response of the amplifier to the small signals v_{dd} and v_{cm} , respectively. The setup is presented in Figure A.7. At DC, the simulated PSRR is about 63 dB (Figure A.8a), whereas the CMRR is about 120 dB (Figure A.8b).

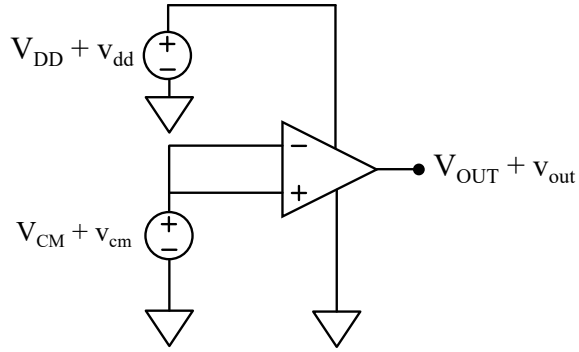
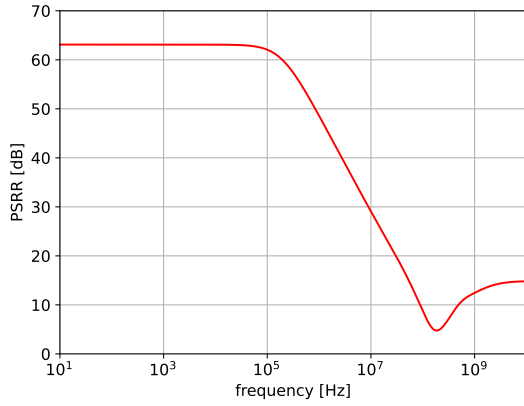
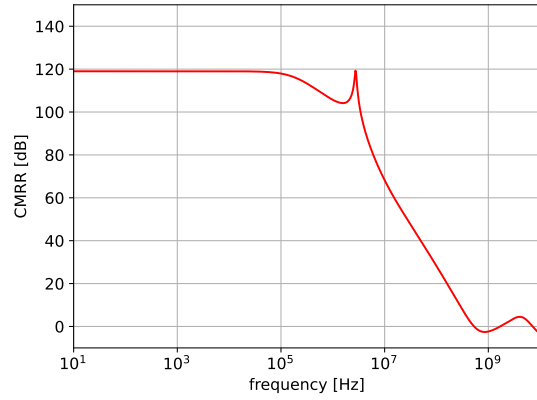


Figure A.7: Setup employed to extract the PSRR and CMRR of the operational amplifier.



(a) PSRR.



(b) CMRR.

Figure A.8: Frequency responses of the PSRR and CMRR, respectively, from the schematic simulation of the operational amplifier.

Bibliography

- [1] ChatGPT. OpenAI. <https://chat.openai.com/> (accessed May 5, 2023).
- [2] Danilo Bronzi, Federica Villa, Simone Tisa, Alberto Tosi, and Franco Zappa. Spad figures of merit for photon-counting, photon-timing, and imaging applications: A review. *IEEE Sensors Journal*, 16(1):3–12, 2016.
- [3] Virginia Ch Spanoudaki and Craig S. Levin. Photo-detectors for time of flight positron emission tomography (ToF-PET). *Sensors (Basel, Switzerland)*, 10(11):10484–10505, 2010.
- [4] Iman Sabri Alirezaei, Nicolas Andre, Amor Sedki, Pierre Gerard, and Denis Flandre. An ultra-thin ultraviolet enhanced backside-illuminated single-photon avalanche diode with 650 nm-thin silicon body based on soi technology. *IEEE Journal of Selected Topics in Quantum Electronics*, 28(2: Optical Detectors):1–10, 2022.
- [5] Diego Haya Enriquez. Simulation and characterization of an ultraviolet enhanced backside-illuminated single-photon avalanche diode. Master’s thesis, Ecole polytechnique de Louvain, Université catholique de Louvain, 2022.
- [6] Iris Cusini, Davide Berretta, Enrico Conca, Alfonso Incoronato, Francesca Madonini, Arianna Adelaide Maurina, Chiara Nonne, Simone Riccardo, and Federica Villa. Historical perspectives, state of art and research trends of single photon avalanche diodes and their applications (part 1: Single pixels). *Frontiers in Physics*, 10, 2022.
- [7] Broadcom. AFBR-S4N66P024M 2×1 NUV-MT Silicon Photomultiplier Array.
- [8] Xudong Jiang, Mark Itzler, Rafael Ben-Michael, and Krystyna Slomkowski. Ingaasp–inp avalanche photodiodes for single photon detection. *Selected Topics in Quantum Electronics, IEEE Journal of*, 13:895 – 905, 08 2007.
- [9] Dylan Issartel, Remy Cellier, Francis Calmon, Shaochen Gao, D. Golanski, Patrick Pittet, and A. Cathelin. Architecture optimization of spad integrated in 28nm fd-soi cmos technology to reduce the dcr. *Solid-State Electronics*, 191:108297, 03 2022.
- [10] Ivan Rech, Antonino Ingargiola, Roberto Spinelli, Ivan Labanca, Stefano Marangoni, Massimo Ghioni, and Sergio Cova. Optical crosstalk in single photon avalanche diode arrays: a new complete model. *opt. express* 16, 8381–8394. *Optics express*, 16:8381–94, 07 2008.
- [11] Andrea Gallivanoni, Ivan Rech, and Massimo Ghioni. Progress in quenching circuits for single photon avalanche diodes. *IEEE Transactions on Nuclear Science*, 57(6):3815–3826, 2010.

- [12] Piatek Slawomir S. What is an SiPM and how does it work? | Hamamatsu Photonics, October 2016.
- [13] Marzieh Ameri. Controllable front-end circuit for geiger mode avalanche photodiode. Master's thesis, École Polytechnique de Montréal, février 2014.
- [14] Willem Johannes Kindt. *Geiger mode avalanche photodiode arrays: For spatially resolved single photon counting*. PhD thesis, Technische Universiteit te Delft, October 1999.
- [15] Silvaco. *ATLAS User's Manual Device Simulation Software*.
- [16] A. G. Chynoweth. Ionization rates for electrons and holes in silicon. *Phys. Rev.*, 109:1537–1540, Mar 1958.
- [17] Francesco Licciulli and Cristoforo Marzocca. An active compensation system for the temperature dependence of sipm gain. *IEEE Transactions on Nuclear Science*, 62(1):228–235, 2015.
- [18] Fabio Acerbi and Stefan Gundacker. Understanding and simulating SiPMs. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 926:16–35, 2019.
- [19] G. Gallina, F. Retiere, P. Giampa, J. Kroeger, P. Margetak, S. Byrne Mamahit, A. De St. Croix, F. Edaltafar, L. Martin, N. Massacret, M. Ward, and G. Zhang. Characterization of SiPM avalanche triggering probabilities. *IEEE Transactions on Electron Devices*, 66(10):4228–4234, oct 2019.
- [20] Aymeric Panglosse, Philippe Martin-Gonthier, Olivier Marcelot, Cédric Virmondois, Olivier Saint-Pé, and Pierre Magnan. Dark count rate modeling in single-photon avalanche diodes. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 67(5):1507–1515, 2020.
- [21] S. Cova, M. Ghioni, A. Lacaita, C. Samori, and F. Zappa. Avalanche photodiodes and quenching circuits for single-photon detection. *Applied Optics*, 35(12):1956, April 1996.
- [22] F. Zappa, A. Tosi, A. Dalla Mora, and S. Tisa. Spice modeling of single photon avalanche diodes. *Sensors and Actuators A: Physical*, 153(2):197–204, 2009.
- [23] S. Tisa, F. Zappa, Alberto Tosi, and S. Cova. Electronics for single photon avalanche diode arrays. *Sensors and Actuators A: Physical*, 140(1):113–122, 2007.
- [24] A. Arbat, A. Dièguez, D. Gascon, J. Trenado, and Ll. Garrido. Avalanche photodiodes for high energy particle tracking in 130 nm CMOS technology. In *2009 16th IEEE International Conference on Electronics, Circuits and Systems - (ICECS 2009)*, pages 579–582, December 2009.
- [25] D. Marano, M. Belluso, G. Bonanno, S. Billotta, A. Grillo, S. Garozzo, G. Romeo, O. Catalano, G. La Rosa, G. Sottile, D. Impiombato, and S. Giarrusso. Silicon photomultipliers electrical model extensive analytical analysis. *IEEE Transactions on Nuclear Science*, 61(1):23–34, 2014.
- [26] Federica Villa, Yu Zou, Alberto Dalla Mora, Alberto Tosi, and Franco Zappa. Spice electrical models and simulations of silicon photomultipliers. *IEEE Transactions on Nuclear Science*, 62(5):1950–1960, 2015.

- [27] Davide Marano, Giovanni Bonanno, Salvatore Garozzo, Alessandro Grillo, and Giuseppe Romeo. A new simple and effective procedure for sipm electrical parameter extraction. *IEEE Sensors Journal*, 16(10):3620–3626, 2016.
- [28] Anna Vil, Anna Arbat, Eva Vilella, and Angel Dieguez. *Geiger-Mode Avalanche Photodiodes in Standard CMOS Technologies*, chapter 9. IntechOpen, 03 2012.
- [29] Majeed M. Hayat, Mark A. Itzler, David A. Ramirez, and Graham J. Rees. Model for passive quenching of SPADs. In Manijeh Razeghi, Rengarajan Sudharsanan, and Gail J. Brown, editors, *Quantum Sensing and Nanophotonic Devices VII*, volume 7608, page 76082B. International Society for Optics and Photonics, SPIE, 2010.
- [30] Ion Vornicu, Ricardo Carmona-Galan, Belén Pérez-Verdú, and Ángel Rodríguez Vázquez. Compact cmos active quenching/recharge circuit for spad arrays. *International Journal of Circuit Theory and Applications*, 44, 07 2015.
- [31] Inyong Kwon, Taehoon Kang, Byron T. Wells, Lawrence D’Aries, and Mark D. Hammig. A high-gain 1.75-ghz dual-inductor transimpedance amplifier with gate noise suppression for fast radiation detection. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 63:356–360, 2016.
- [32] Elisabetta Nocerino. *The Semiconductor Multiplication System for Photoelectrons in a Vacuum Silicon Photomultiplier Tube and Related Front End Electronics*. PhD thesis, Università degli Studi di Napoli Federico II, 11 2016.
- [33] Shahram Hatefi Hesari, Mohammad Aminul Haque, and Nicole McFarlane. A comprehensive survey of readout strategies for sipms used in nuclear imaging systems. *Photonics*, 8(7), 2021.
- [34] Mingchen Wang, Yonggang Wang, Qiang Cao, Liwei Wang, Jie Kuang, and Yong Xiao. Comparison of three pre-amplifier circuits for time readout of sipm in tof-pet detectors. In *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5, 2019.
- [35] J. Huizenga, S. Seifert, F. Schreuder, H.T. van Dam, P. Dendooven, H. Löhner, R. Vinke, and D.R. Schaart. A fast preamplifier concept for sipm-based time-of-flight pet detectors. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 695:379–384, 2012. New Developments in Photodetection NDIP11.
- [36] F. Corsi, M. Foresta, C. Marzocca, G. Matarrese, and A. Del Guerra. Current-mode front-end electronics for silicon photo-multiplier detectors. In *2007 2nd International Workshop on Advances in Sensors and Interface*, pages 1–6, 2007.
- [37] Pietro P. Calò, Fabio Ciciriello, Savino Petrignani, and Cristoforo Marzocca. Sipm readout electronics. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 926:57–68, 2019. Silicon Photomultipliers: Technology, Characterisation and Applications.
- [38] Zhen Gu and Xiaojun Bi. A charge amplifier with noise peaking suppression and gain drop compensation utilizing a quasi-miller rc network. *AEU - International Journal of Electronics and Communications*, 107:252–256, 2019.

- [39] Zhi Deng, Yinong Liu, Lan Zhang, Yulan Li, Yuanjing Li, Jin Li, Allan Kejian Lan, and Ed Hungerford. A multi-channel front-end ASIC for pixellated detectors. In *2006 IEEE Nuclear Science Symposium Conference Record*, volume 1, pages 337–340, 2006.
- [40] Mohammad Beikahmadi and Shahriar Mirabbasi. A low-power low-noise CMOS charge-sensitive amplifier for capacitive detectors. In *2011 IEEE 9th International New Circuits and Systems conference*, pages 450–453, 2011.
- [41] Francesco Corsi, Maurizio Foresta, Cristoforo Marzocca, Gianvito Matarrese, and Alberto Del Guerra. ASIC development for SiPM readout. *Journal of Instrumentation*, 4(03):P03004, mar 2009.
- [42] Luis B. Oliveira, Carlos M. Leitao, and M. Medeiros Silva. Noise performance of a regulated cascode transimpedance amplifier for radiation detectors. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 59(9):1841–1848, 2012.
- [43] Manuel de Medeiros Silva and Luis B. Oliveira. Regulated common-gate transimpedance amplifier designed to operate with a silicon photo-multiplier at the input. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 61(3):725–735, 2014.
- [44] Luis Henrique Rodovalho. Schmitt trigger based single-ended voltage amplifier with positive feedback control for ultra-low-voltage supplies. In *2020 IEEE Nordic Circuits and Systems Conference (NorCAS)*, pages 1–6, 2020.
- [45] M.Yu. Kantor and A.V. Sidorov. Shaping pulses of radiation detectors into a true gaussian form. *Journal of Instrumentation*, 14(01):P01004, jan 2019.
- [46] Pawel Grybos and Robert Szczygiel. Pole-Zero Cancellation Circuit With Pulse Pile-Up Tracking System for Low Noise Charge-Sensitive Amplifiers. *IEEE Transactions on Nuclear Science*, 55(1):583–590, 2008.
- [47] H. Wu, H. Xu, Q. Zhang, X. Meng, R. Di, and Y. Liu. Implementation of a straight-line shaper on an FPGA for a UAVs. *Journal of Instrumentation*, 15(10):T10011, oct 2020.
- [48] P.R. Gray. *Analysis and Design of Analog Integrated Circuits, 5th edition*, ISV. Wiley India Pvt. Limited, 2009.
- [49] Zihong Liu, Chao Bian, Zhihua Wang, and Chun Zhang. Full custom design of a two-stage fully differential CMOS amplifier with high unity-gain bandwidth and large dynamic range at output. In *48th Midwest Symposium on Circuits and Systems, 2005.*, pages 984–987 Vol. 2, 2005.
- [50] Ka Nang Leung and P.K.T. Mok. Analysis of multistage amplifier-frequency compensation. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 48(9):1041–1056, 2001.
- [51] Simone Becchetti, Anna Richelli, Luigi Colalongo, and Zsolt Kovacs-Vajna. A comprehensive comparison of EMI immunity in CMOS amplifier topologies. *Electronics*, 8(10), 2019.
- [52] H. Finkelstein, M.J. Hsu, and S.C. Esener. Si-bounded single-photon avalanche diode in a deep-submicrometer CMOS technology. *IEEE Electron Device Letters*, 27(11):887–889, 2006.

- [53] Jun Ma and Sanjiv Sinha. Thermoelectric properties of highly doped n-type polysilicon inverse opals. *Journal of Applied Physics*, 112, 10 2012.
- [54] Paul Drude, Albert Einstein, and Max Planck. *Annalen der Physik*. J.A. Barth, 1901.
- [55] Hamamatsu. Si apd. Technical note, Hamamatsu, October 2021.
- [56] Gwladys Perrais, O. Gravrand, J. Baylet, Gerard L. Destefanis, and Johan Rothman. Gain and dark current characteristics of planar hgcdte avalanche photo diodes. *Journal of Electronic Materials*, 36:963–970, 2007.
- [57] Yux Xu, Ping Xiang, and Xiaopeng Xie. Comprehensive understanding of dark count mechanisms of single-photon avalanche diodes fabricated in deep sub-micron cmos technologies. *Solid-State Electronics*, 129:168–174, 2017.
- [58] Yue Xu, Ping Xiang, Xiaopeng Xie, and Yang Huang. A new modeling and simulation method for important statistical performance prediction of single photon avalanche diode detectors. *Semiconductor Science and Technology*, 31(6):065024, may 2016.
- [59] Shuai Wang, Han Ye, Liyan Geng, Ziqing Lu, Feng Xiao, Fan Xiao, and Qin Han. Design, Fabrication, and Characteristic Analysis of 64×64 InGaAs/InP Single-Photon Avalanche Diode Array. *Journal of Electronic Materials*, 51(5):2692–2697, May 2022.
- [60] G.A.M. Hurkx, D.B.M. Klaassen, and M.P.G. Knuvers. A new recombination model for device simulation including tunneling. *IEEE Transactions on Electron Devices*, 39(2):331–338, 1992.
- [61] Lucio Pancheri, David Stoppa, and Gian-Franco Dalla Betta. Characterization and modeling of breakdown probability in sub-micrometer cmos spads. *IEEE Journal of Selected Topics in Quantum Electronics*, 20(6):328–335, 2014.
- [62] D. Neamen. *Semiconductor Physics And Devices*. McGraw-Hill Series in Electrical and Computer Engineering. McGraw-Hill Education, 2003.
- [63] P. Jarron, E. Auffray, S.E. Brunner, M. Despeisse, E. Garutti, M. Goettlich, H. Hillemanns, P. Lecoq, T. Meyer, F. Powolny, W. Shen, H.C. Schultz-Coulon, and M.C. S. Williams. Time based readout of a silicon photomultiplier (sipm) for time of flight positron emission tomography (tof-pet). In *2009 IEEE Nuclear Science Symposium Conference Record (NSS/MIC)*, pages 1212–1219, 2009.
- [64] F. Corsi, A. Dragone, C. Marzocca, A. Del Guerra, P. Delizia, N. Dinu, C. Piemonte, M. Boscardin, and G.F. Dalla Betta. Modelling a silicon photomultiplier (sipm) as a signal source for optimum front-end design. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 572(1):416–418, 2007. Frontier Detectors for Frontier Physics.
- [65] D. Marano, G. Bonanno, M. Belluso, S. Billotta, A. Grillo, S. Garozzo, G. Romeo, O. Catalano, G. La Rosa, G. Sottile, D. Impiombato, and S. Giarrusso. Improved spice electrical model of silicon photomultipliers. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 726:1–7, 2013.

- [66] T. Chaves de Albuquerque, D. Issartel, R. Clerc, P. Pittet, R. Cellier, W. Uhring, A. Cathelin, and F. Calmon. Body-biasing considerations with spad fdsoi: advantages and drawbacks. In *ESSDERC 2019 - 49th European Solid-State Device Research Conference (ESSDERC)*, pages 210–213, 2019.
- [67] XT018. Xfab. <https://www.xfab.com/xt018> (accessed May 5, 2023).
- [68] Federica Villa, Danilo Bronzi, Michele Vergani, Yu Zou, Alessandro Ruggeri, Franco Zappa, and Alberto Dalla Mora. Analog sipm in planar cmos technology. In *2014 44th European Solid State Device Research Conference (ESSDERC)*, pages 294–297, 2014.
- [69] Guillaume Pollissard-Quatremère, Geoffroy Gosset, and Denis Flandre. A modified g m /i d design methodology for deeply scaled cmos technologies. *Analog Integrated Circuits and Signal Processing*, 78, 03 2014.
- [70] K.R. Laker and W.M.C. Sansen. *Design of Analog Integrated Circuits and Systems*. McGraw-Hill series in electrical and computer engineering. Electronics and VLSI circuits. McGraw-Hill, 2002.
- [71] David L. Terrell. Chapter ten - nonideal op amp characteristics. In David L. Terrell, editor, *Op Amps (Second Edition)*, pages 383–403. Newnes, Burlington, second edition edition, 1996.
- [72] Jean-Paul Eggermont. *Study and realization of SOI operational amplifiers for high-temperature or high-frequency applications*. PhD thesis, UCL - Université Catholique de Louvain, 1998.
- [73] J.G. Kenney, G. Rangan, K. Ramamurthy, and G. Temes. An enhanced slew rate source follower. *IEEE Journal of Solid-State Circuits*, 30(2):144–146, 1995.
- [74] R. Jacob Baker, Harry W. Li, and David E. Boyce. *CMOS. Circuit design layout and simulation*. IEEE Press, 1998.
- [75] Vikas Vinayaka. *Analysis and Design of Analog Front-end Circuitry for Avalanche Photodiodes (APD) and Silicon Photo-multipliers (SiPM) in Time-of-flight Applications*. PhD thesis, University of Nevada, Las Vegas, 2018.
- [76] Kiyoshi Takeuchi, Akio Nishida, and Toshiro Hiramoto. Random fluctuations in scaled mos devices. In *2009 International Conference on Simulation of Semiconductor Processes and Devices*, pages 1–7, 2009.
- [77] Tom Klauner, Iman Sabri Alirezaei, Nicolas Roisin, Nicolas Andre, and Denis Flandre. Spice model for spad transient intrinsic response using mixed-mode tcad simulations. In *ESSDERC 2023 - 53th European Solid-State Device Research Conference (ESSDERC)*, (accepted) 2023.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
École polytechnique de Louvain

Rue Archimède, 1 bte L6.11.01, 1348 Louvain-la-Neuve, Belgique | www.uclouvain.be/epl