TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

DIPLOMARBEIT

Entwicklung eines aktiven Einzelphotonen Lawinendioden Zeilensensors

ausgeführt zum Zwecke der Erlangung des
akademischen Grades eines Diplom - Ingenieurs

an der Technischen Universität Wien
Institut für Schaltungstechnik

unter der Leitung von

Univ.Prof. Mag.rer.nat. Dr.techn. Horst Zimmermann
Senior Scientist Dipl.-Ing. Dr.techn. Bernhard Goll
Projektass. Dipl.-Ing. Dr.techn. Michael Hofbauer, BSc

durch

Stefan Michael Kornfeld, BSc
Matr.-Nr. 01329126

Wien, Februar 2022

Statutory Declaration

I declare, that I have authored the present work independently according to the code of conduct, that I have not used other than the declared sources and that I have explicitly marked all material quoted either literally or by content from the used sources. This work was not yet submitted to any examination procedure neither in Austria, nor in any other country.

Erklärung zur Verfassung der Arbeit

Hiermit erkläre ich, dass die vorliegende Arbeit gemäß dem Code of Conduct - Regeln zur Sicherung guter wissenschaftlicher Praxis - ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel, angefertigt wurde. Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder in ähnlicher Form in anderen Prüfungsverfahren vorgelegt.

Wien, 28.02.2022 .....

Stefan Michael Kornfeld

Abstract

This thesis describes the conception and development of a line sensor IC, which was realized in a 0.35 µm Complementary Metal Oxide Semiconductor (CMOS) process. This chip, in which 32 detection elements are arranged in a one-dimensional array, is used to determine the time and location of the impact of individual photons. At the beginning, various detection technologies, their advantages, and disadvantages as well as typical application specifics are explained and a comparison of the relevant (detector) properties is carried out. After weighing and evaluating all aspects, single photon avalanche diodes (SPADs) and their operating circuits were selected as the basis for the line sensor. Based on a proven SPAD design, by considering and evaluating countless simulations, an attempt was made to find circuits by using bipolar transistors which would be superior to the field effect transistors used in the basic design for use in the line sensor mentioned, taking all aspects into account. While there were significant improvements in performance, i.e., increased quenching speed of the SPADs, the decision was made to use the field effect transistors mentioned in the basic design due to the increased current flow and thus increased power requirement. Derived from the start design, the use of a high quench voltage of 9.9 V and variable quench and reset times and the associated dead times in the range of 10 ns to 100 ns is made possible. This basic design was changed to obtain an optimal use of space in the layout. The 32 SPADs mentioned were positioned with a with a pitch of 60 µm on a chip area of around 6.45 mm^2. The filling factor could be increased to 53 % through intelligent partially overlapping positioning of the SPADs with an active area of around 1288 $µm^2$ per SPAD. The nested positioning of the detection elements used in the design allows for optimal utilization of the chip and provides a way to detect optical crosstalk using fewer digital signals. Adjustments also allow a switch to gated operation.

Kurzfassung

Diese Diplomarbeit beschreibt die Konzeption und Entwicklung eines Zeilensensor ICs, welcher in einem 0.35 µm Complementary Metal Oxide Semiconductor (CMOS) Prozess realisiert wurde. Dieser Chip, in dem 32 Detektionselemente in einem eindimensionalen Array angeordnet sind, dient zur zeitlichen und örtlichen Bestimmung des Auftreffens einzelner Photonen. Beginnend werden verschiedenste Detektionstechnologien, deren Vor- und Nachteile sowie typische Anwendungsspezifika erläutert und ein Vergleich der relevanten (Detektor)-Eigenschaften durchgeführt. Unter Abwägung und Bewertung sämtlicher Gesichtspunkte wurden Single Photon Avalanche Dioden (SPADs) bzw. deren Betriebsschaltungen als Grundlage für den Zeilensensor ausgewählt. Ausgehend von einem erprobten SPAD Design wurde, unter Betrachtung und Auswertung unzähliger Simulationen, versucht durch Verwendung von Bipolartransistoren Schaltungen zu finden, welche für die Verwendung in dem genannten Zeilensensor unter Beachtung sämtlicher Gesichtspunkte den im Basisdesign verwendeten Feldeffekttransistoren überlegen wären. Während sich im Bereich der Performance, d.h. Erhöhung Quenchgeschwindigkeit der SPADs, signifikante Verbesserungen ergaben, fiel die Entscheidung aufgrund von erhöhtem Stromfluss und damit erhöhten Leistungsbedarf darauf, die im Basisdesign genannten Feldeffekttransistoren zu verwenden. Abgeleitet vom Startdesign wird eine Nutzung einer hohen Quenchspannung von 9.9 V und variablen Quench und Resetzeiten und den damit verbunden Totzeiten im Bereich von 10 ns bis 100 ns ermöglicht. Dieses Basisdesign wurde verändert, um einerseits eine optimale Platzausnutzung im Layout zu erhalten. Auf einer Chipfläche von rund 6.45 mm^2 wurden die genannten 32 SPADs in einer örtlichen Auflösung von 60 µm positioniert. Durch intelligente teilweise überlappende Positionierung der SPADs mit einer aktiven Fläche von rund 1288 $µm^2$ pro SPAD konnte der Füllfaktor auf 53 % erhöht werden. Das verschachtelte Positionieren der Detektionselemente, welches beim Design verwendet wurde, ermöglicht eine optimale Ausnutzung des Chips, und stellt mit Hilfe weniger Digitalsignale eine Möglichkeit zur Erkennung von optischen Crosstalk dar. Anpassungen ermöglichen auch eine Umstellung auf einen Gater-Betrieb.

Acknowledgement

I would like to take this opportunity to thank everyone who supported me during my studies.

First of all, I would like to thank Professor Horst Zimmermann, who made my master's thesis possible in the first place, for his trust and support. I would also like to express my graditude for the excellent support from my supervisors Bernhard Goll and Michael Hofbauer.

Of course, I would also like to thank my family and my girlfriend Tanja, who made my studies possible with their support and who made a significant contribution to my success through their emotional support.

Contents

Chapter 1

Introduction

1.1 Problem-Motivation

In the course of the increasing research interest in areas and applications, whose functional principles are based on the detection of the smallest light quanta, such as light detection and ranging (LiDAR) [1], super-resolution microscopy, high-speed optical communication [2] and quantum cryptography [3], the need for fast and reliable single photon detection capabilities is accordingly increasing.

In many applications, for instance experiments from the field of research or real industrial applications from the mentioned fields, it is not only necessary to reliably detect an impact of a photon but also to make a statement about the location of the detection process. In other words: The impact location on a certain sensor of a photon has to be determined or measured.

The motivation for this work or the problem according to which this work should represent the solution can be defined in detail as follows: Design of a line sensor, i.e. a one-dimensional array for determining the appearance of individual photons. The spatial resolution should be less than or equal 100 µm. A single detection element, hereinafter referred to as a pixel, should be based on a tried and tested detection principle that is known within the institute. If the detection principle used allows it, options for parameterization should be provided which allow an influence on individual (measured) variables. The number of pixels should be at least 32 and the design should pay attention to scalability, or in other words it should offer a simple possibility for expansion. There should be an output interface to be defined

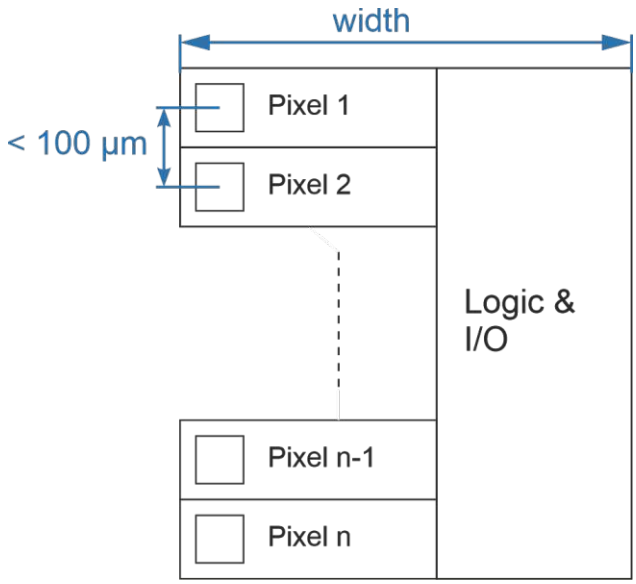

Figure 1.1: *Simple block diagram of the expected solution*

for connection to an existing measuring system, taking common communication standards into account. There should be the possibility of temporarily deactivating all pixels (gating mode) and, independently of this, individual elements should be able to be permanently deactivated. The design created should provide information with a comparatively low delay as to whether any detection has taken place. Since only one detection event can take place in the valid case when a photon hits, appropriate precautions should indicate via the interface mentioned whether a detection event of whatever type is valid or not. In the case of a valid single event, it must of course be clearly signaled where this detection took place.

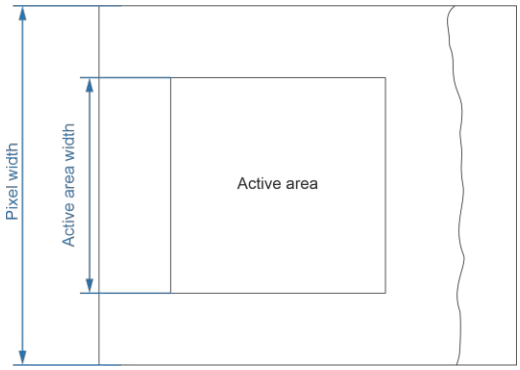

Figure 1.2: *Visualization of the relationship between sensor and pixel dimensions*

Figures 1.1 and 1.2 represent a schematic representation (and the detail section) of the required chip. The previous paragraph shows the limit of the resolution. These 100 µm are to be understood as the center-to-center distance of the active surfaces (= sensors). There are no special requirements for the width mentioned in Fig. 1.1, which represents

the total width of the chip. With the exception of the significant cost factor in the course of the production of this array, this dimension can theoretically be changed indefinitely. The block marked with a pixel contains not only the sensor, but also the functional and protective circuitry required for the respective sensor type. The shape of the active area can be selected as desired, but a high fill factor should be achieved. To maximize the fill factor, the difference between the pixel width and the width of the active area should be minimized. In the best case, the width of the pixel would correspond to the width or exactly the dimensions of the active area. These dimensions are shown schematically in Fig. 1.2.

1.2 State of the Art

Although some work already exists on the subject of single photon avalanche diode (SPAD) arrays, these works do not satisfy all the defined requirements or show deficiencies in specific areas. Therefore, some of these insufficiencies identified will be tempted to be eliminated in this design. Some works [4] use passive quenching techniques for simplicity. These are much easier to implement than actively quenched SPAD systems, both for space and complexity reasons (More details on functionality are explained in 1.3.2). In terms of performance and definability of quench and reset times, SPAD quenchers which are based on the active function model are better than passive ones. The disadvantage in those cases is the lower fill factor for standard complementary metal oxide semiconductor (CMOS) processes, since the quenching circuits require significantly more space compared to passive quenching circuits.

These basic spatial conditions result in "slim" SPAD/quencher designs whose spatial spread in one direction is dictated primarily by the dimensions of the real SPAD or optically sensitive areas. The other direction (when using a less modern BiCMOS process, i.e. planar design) is significantly longer, since the circuitry or the components inside it require correspondingly more amount of space [5]. Two-dimensional sensors, as far as they are manufactured in a classical process and not slim design, show application-related "blind" spots between the single SPADs [6].

Some works use special technologies, which involve higher financial outlays. For example, the papers by Charbon, et al. [7] and Lindner, et al. [8] show relatively high complexity due to the 3D stacked CMOS design used. The (additional) costs compared to a standard CMOS process are thus considerable.

Furthermore, regardless of the process type, previous works often show maximum photon

detection probabilities (PDP) similar or slightly lower than the maximum PDP values which are of the basic cell type [9] used for the sensor described subsequently. These (maximum) PDP values usually occur in the wavelength range from 400 nm to 700 nm. In the higher wavelength range (> 800 nm) previous work shows clear deficits compared to the single pixel element used (28.8 % @ 850 nm).

The combination of circumstances, which result from the need for (parameterizable) one-dimensional detectors of higher PDPs in this wavelength range, led to the creation of this master thesis.

1.3 Detection of Single Photons

Applications in quantum physics and photoelectrically integrated components that have already been briefly mentioned, require the detection of individual photons. In the following, parameters and implementations are explained.

A correlated relationship between optical input energy and measurable electrical quantity is the basis necessary for an optical detector. The physical basis for the optical detection of a photon is the photoelectric effect. A distinction is made between the external and internal photoelectric effect. The external photoelectric effect describes the release of electrons from semiconductor or metal surfaces, which is particularly useful for older technologies such as photomultiplier (PMT) or photocells. In more modern applications or technologies that work on the basis of an optically triggered charge carrier-hole pair generation, the internal photoelectric effect is considered as the physical background [10] [11].

1.3.1 Characteristic Parameters of Detectors

In order to be able to make a corresponding description of the detection behavior or to quantitatively compare the technologies with each other, a selection of characteristic values are explained in more detail.

Photon detection probability
A characteristic value that is used to quantify how many of the photons that have triggerd a measurable detection event is known as Photon Detection Probability or PDP.

The PDP which is mainly used in single photon detectors is defined as the ratio of the detected to detectable, i.e. incidents, photons. In the ideal case the value of the PDP is 1

or 100 %. However due to non-idealities, a PDP value of <1 must always be expected in applications.

Dead time
Depending on the detector technology, under functional or circuit-related circumstances a certain period of time is required after a detection to make the sensor element used or necessary operating circuits ready for a new detection. During this period of time, the sensor is "blind" and cannot be used to detect any light.

After pulsing probability
Sensors whose functional principle is based on triggering a charge carrier avalanche can be assigned the property of after pulsing with its the corresponding value After Pulsing Probabilty, APP. This property is used to quantify the probability of a post pulse with non-photon origin, which is correlated to a previous detector avalanche.

Jitter
Furthermore, optical detectors or, above all, single photon detectors are subject to an effect called jitter, which limits the time resolution of the detector. This statistical effect describes the variation of the time differences between the impact of a photon and the generation of the output signal. Depending on the detector or detector technology, various causes may be responsible. These range from a wavelength dependency (different penetration depth of photons) over statistical time and space deviations of the ionization process to variances in the corresponding detection circuit.

Dark count rate
In addition to optical triggers, free charge carriers, production-related inhomogeneities and other faults can also trigger a supposed detection process. One way to quantify these processes is through the Dark Count Rate, DCR. As the name suggests this parameter can be determined by operating the detector in the "dark", i.e. without active lighting and with the sensor being shaded. The DCR is indicated in counts per unit of time, typically in counts per second.

1.3.2 Detector Technologies

Following, the most common and well-known sensor technologies are briefly presented. Since SPADs are ultimately used as detector elements in the line sensor, this technology will be discussed in detail at the end of this chapter.

Photomultiplier Tubes

Among the first known single-photon detectors were photomultiplier tubes (PMT). As already mentioned their functional principle is based on the external photoelectric process. Individual photons entering the vacuum tube can strike out electrons from the photosensitive surface. These free electrons are accelerated towards the anode by a strong electric field. On their way to the anode, the electrons hit further electrodes (so-called dynodes) and release even further electrons from the dynodes, which are also accelerated towards the anode. The amplification factor grows exponentially with the number of dynodes. At the end, the electrons hit the anode and flow off to the ground. In the process, they generate a voltage drop across a resistor, which is then subsequently detected [12]. For a long time PMTs were the only sensors that could detect single photons with a maximum PDP of approximatly 40 %. The mechanical constraints (large, fragile design, necessary vacuum), the associated high production costs and the relatively high sensitivity to electromagnetic radiation are disadvantageous. Despite the age of the technology, commercially produced PMTs, often also in functional combination with newer semiconductors, can be found in a wide range of applications [13].

Superconducting Nanowires

Although destruction of superconductivity by laser light was reported as early as 1971[14] , work on single photon detectors was only published at the beginning of the last millennium [15]. Superconducting nanowire single-photon detectors (SNSPD) are lithographically manufactured (mostly meandering) in 100 nm to 200 nm wide structures which are operated current-biased below their critical temperature. If a photon hits these structures, the critical temperature is exceeded locally and a voltage drop can be measured. When using SNSPDs, high PDPs, exceeding 90 %, are possible with low DCR and jitter at the same time [16][17]. SNSPDs or nano wires in general are currently the subject of research. The necessary low temperatures in the single-digit Kelvin range and the expensive production are reasons for a difficult usage of this technology outside of laboratory setups.

APD

As already mentioned, Avalanche Photodiodes (APD) or p(i)n photodiodes in general are based on the functional principle of the internal photoelectric effect. A photon that strikes a pn or pin structure creates an electron-hole pair. These, either positively or negatively charged carriers move against the diffusion voltage or, depending on the application, applied reverse bias voltage into the similarly doped zones. This light-dependent charge transport then corresponds with the photocurrent, which is an indicator for a photon that has struck. Compared to pn or pin photodiodes, APDs have an additional multiplication zone which, in an avalanche-like manner, intensifies the photon-induced charge carrier separation. This release of additional charge carriers through already accelerated charge carriers is referred to as impcat ionization, not only in connection with APDs.

SPAD

Single Photon Avalanche Diodes (SPAD) are similar in structure and function to APDs. While APDs are operated in reversed biased operation below their breakdown voltage, SPADS are biased above the breakdown voltage. This operating mode is also known as Geiger mode. While APDs have a radiation-proportional operating behavior, which result from the photocurrent being proportional to the number of photons or the light power encountered, the SPAD has a digital behavior due to its "infinite" gain. In other words, a detected photon triggers a self-sustaining avalanche. Since the SPAD could be destroyed under certain circumstances, appropriate wiring and precautionary measures are required.

The operation of a SPAD or the sequence of a detection is shown in Fig 1.3. Starting from a pre-biased SPAD, no current flows through the SPAD. If a photon hits the sensitive area, a rapidly increasing current begins to flow, due to a charge separation and the associated multiplication (2). In order to prevent the SPAD from being destroyed and to enable a new detection of another photon, the avalanche triggered has to be stopped. This stopping process is referred to as "quenching" in connection with SPADs. During the quenching process the reverse voltage of the diode is reduced (3) to a level below the breakdown voltage of the specific device and then kept at this level for a defined time. In the course of the process called "recharge" (1) the bias voltage is recharged to the potential required above the breakdown voltage.

The dead time described in 1.3.1 is thus mainly composed by the time required for the detection, the quench phase, the waiting phase (time for staying below the breakdown voltage) and the reset time.

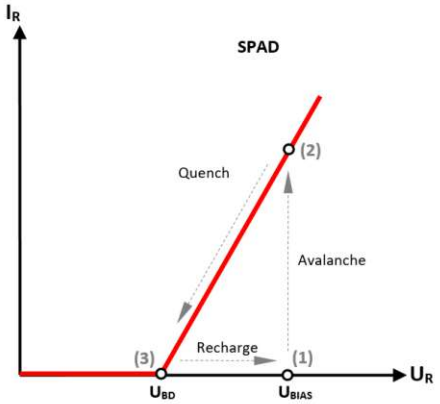

Figure 1.3: *Graphic representation of a photon detection process*

From the functional principle it can be seen that a dead time is necessary between two detections, due to quenching and recharging processes. Usually no photon can be detected during the transitions from one state to the other.

The functional cycle briefly described here is referred to as *Free running* mode. This means that a detection process is always possible within the operating period, with the exception of quench, hold off and reset times. In order to enable an operation only within a defined time window, e.g. only after or during the emission of a source light pulse, there is also a so-called gating operation. The name comes from the fact that analogous to an FET, the SPAD can be activated by a gate [18].

The photon detection probability depends on how much the reverse voltage is higher than the breakdown voltage. This voltage overhead is called excess bias voltage. Usually the PDP increases when raising the excess bias. With increasing excess bias voltage, i.e. the difference between the bias voltage and the voltage in the quench state, the PDP also increases.

Like all real components, SPADs are subject to noise phenomena. The causes, like after pulses, Jitter or dark counts have already been mentioned in chapter 1.3. For uncorrelated dark counts, the theory is based on interactions between lattice defects or charge traps and charge carriers, formulated by Shockley, Read and Hall [19] and serves as a descriptive basis. These charge traps, which enable energy states in the band gap are the sources for these false triggers generated in the dark. The DCR is strongly temperature dependent. Measurements have shown that an increase in the operating temperature of 10K can double the DCR [20]. Furthermore, high local electric field strengths can increase the emissivity of the local energy level. This and an increased probability of band-to-band tunneling increase the DCR [21] [22].

In addition to the uncorrelated noise effect, there are also correlated ones. These afterpulses or the probability of the occurrence of these undesired effects have already been briefly mentioned. After an avalanche has been triggered, charge traps are temporarily filled with charge carriers. Depending of the life time of this state, these carriers leave these traps again and may trigger secondary avalanches. There is a trade-off between dead time and APP. However, increasing dead time will reduce the maximum possible count rate [23].

Temperature dependency of the APP has not yet been clarified and is currently still the subject of research. On the one hand, the APP should be reduced with increasing temperature, on the other hand measurements have shown that there is negligible temperature dependency of the APP [20]. During the measurements mentioned above, the theory emerged that charge carriers in traps are not or, depending on the device, not the dominant effect for the APP. Without going into further detail, avalanche-induced photons are said to have triggered a charge carrier separation in the substrate, whereby the separation products diffuse into the absorption zone of the SPAD and can then trigger an afterpulse.

Since the occupancy of the charge traps is directly proportional to the number of charge carriers involved in an avalanche, in the course of reducing the number of "trapped" charge carriers it is of interest to only let the avalanche run for as little time as possible or to start the quenching process as soon as possible and carry it out as quickly as possible. Accordingly, the aim of the research is to reduce both the time in which the avalanche is active and the duration of the quenching process to a minimum [24].

Quenching Circuits for SPADs

In order to ensure the functionality of a SPAD as described above and to complete the detection process as mentioned, a corresponding quenching circuit is required. In addition to quenching, depending on the circuit variant, such operating circuits are also used for a defined reset after a quenching process, for signal conditioning and for driving the output pulse to subsequent system components.

Passive Quenching

The simplest form of a quenching circuit is the passive form. A simple functional diagram can be found in Fig. 1.4. Since no current flows through a pre-biased SPAD in the non-triggered case, the amount of voltage drop across the resistor is zero and the excess bias voltage is completely applied to the SPAD, as a first approximation. If an avalanche is triggered by a photon or another parasitic event, the voltage at the SPAD collapses, which is then recognized as a (supposed) detection event. The rapidly increasing current caused by the avalanche breakdown increases the voltage at the series resistor, which in turn reduces the voltage at the SPAD and thus the electric field strength in the multiplication zone and stops the avalanche when the reverse voltage drops below breakdown voltage [18]. It is also necessary to mention that the resistor has contradicting implementation requirements depending on the operating state. On the one hand, it should be low-resistive enough to ensure fast charging, on the other hand, high-ohmic enough to ensure fast quenching in the event of a SPAD triggering.

The disadvantageous behavior when a SPAD is reset, i.e. the slow charging after an quenching process, which occurs due to the function of circuits with a passive reset character, can lead to an avalanche being triggered in this reset phase, as the actual applied voltage can be above the breakdown voltage. However, since the required excess bias voltage has not yet been reached, this avalanche may not be detected, since the detection voltage level, which must be fallen below in the course of regular detection and therefore represents a detection event, is not reached. In any case, the reset time is extended. The reset time is very dominant here, which is particularly noticeable in SPADs with a large active area and the associated high capacity. The times that can be necessary for the reset can relatively quickly be orders of magnitude greater than the time required for the quenching process. Furthermore, in addition to the reset time mentioned, the quench

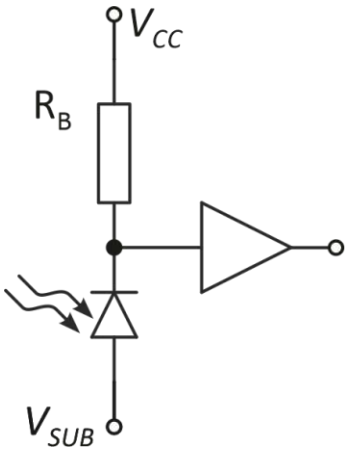

Figure 1.4: *Block diagram of a SPAD with passive quenching [18]*

time is also subject to strong components and material dependencies. A holdoff time is practically non-existent.

Active Quenching

An alternative approach is the active quenching approach. Active quenching circuits measure the voltage of the SPAD in a certain form or detect the voltage drop on it and then activate circuit elements (usually transistors) via appropriate logic to the quenching process and the reset start. The detection should take place as quickly as possible in order to keep the amount of charge involved as low as possible.

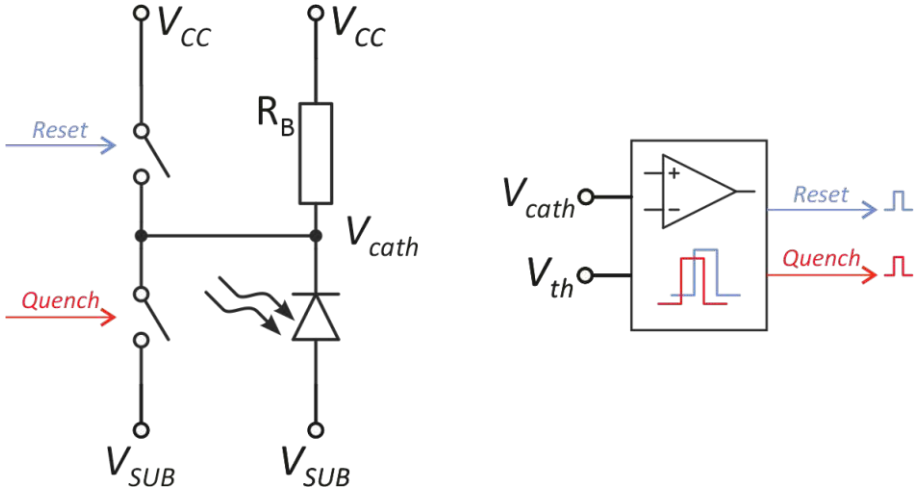

Figure 1.5: *Block diagram of a SPAD with active quenching [18] [24]*

There are innumerable variants and possible solutions for quenching circuits in current state of research. In addition to the basic types mentioned above, there are also mixed types - so-called Mixed Active-Passive Quenching Circuits - which try to combine the

advantages of the two variants or to enable compromises that are appropriate for the respective application. The quick detection often takes place by means of a comparator, also shown symbolically in the block diagram in Fig. 1.5, which compares the current cathode voltage with a defined threshold voltage. If this value falls below the threshold due to a voltage drop, the described quenching and resetting phase is started.

Gated Operation Mode
The two methods mentioned so far assume a so-called free running mode, i.e. the SPAD or the sensor element is ready to receive at any time, with the exception of the dead time after a detection. In gated operation mode or gate mode, detection is only possible during a defined phase. This is especially recommended for highly sensitive receivers, noisy environments or special applications with defined temporal "receive windows". The SPAD is kept below the breakdown voltage, i.e. permanently quenched, and an additional excess bias voltage is only applied during the gater phase. Depending on the implementation, the excess voltage is removed after detection of a breakdown or is quenched again after a fixed time or functional control, regardless of whether a detection has taken place or not.

SPAD Design Concepts
Due to the high sensitivity of SPADs, there are high demands on the purity and quality of the manufacturing process. Even slightest inhomogeneities in the substrate can lead to deterioration of paratsitic properties (DCR, APP, jitter). Shifts in the field distribution can locally reduce the breakdown voltage and decrease the active area. Furthermore, special measures are necessary in order to avoid high field strengths caused by edge effects, because these could lead to false triggers.

Basically, SPADs can be divided into two types. The reach-through design concept requires special manufacturing technologies with the possibility to optimize individual parameters. A thick absorption zone and a narrow multiplication zone with relatively high field strength are typical for such a design. First works on this technology date back to the 1970s/1980s of the last century [25] [26] [27]. This technology is still used today. Advantageous here is the low DCR (<10kcps) at relatively large volumes or diameters (500 µm) [28].

Alternatively, so-called planar design concepts have the advantage of beeing able to be realized more easily in CMOS technology which is understandably more favorable for mass production. Furthermore, the necessary circuitry (quencher or (readout) interface) can be realized in the same process, which is advantageous. Also, there were already first attempts with weakly p-doped Si wafers around the 1960s [29] [30]. Ghoni [31] and Cova [32], showed for the first time a construct consisting of an epitaxial p-doped layer on an n-substrate. This had, in addition to the higher quality or the associated lower number of defects or impurities, an additional boundary layer between substrate and epitaxial P-layer, which

reduced unwanted avalanche processes due to diffusing minority charge carriers.

Another possibility to stop diffusing (unwanted) charge carriers are guard rings. These also serve to reduce the electric field strength at the edges. Due to discontinuities (=edges) high local field strengths result, which consequently reduce the breakdown voltage of the SPAD and lead to edge breakdowns.

Chapter 2

Design

2.1 Process Technology

For this work the *XO035* process from *XFAB* [33] is used. It is a p-substrate 0.35 µm BiCMOS process. There is already some experience with this node size at the institute. The reasons are the possibilities offered by this process. Furthermore, SPADs functionally require high negative substrate voltages. This process allows to embed circuit parts in deep n-well regions and thus protects them from destruction.

2.2 Used SPAD Design

Figure 2.1 shows a section of the used SPAD (not to scale). It is a design which, as already described in 1.3.2, can be fabricated relatively easily by CMOS process. A p^- epi-layer of about 10 µm thickness is epitaxially grown on the used p-substrate. A p-well is placed in it, which forms the multiplication zone with the n^+-region located on it. The p-well, or rather the fact that the well does not extend to the end of the n-region at the edges, is also referred to as a "virtual guard ring". The relatively high doping of the p-well in the central region increases the field strength in the gain region and the field peaks caused in the edge region of the n+ region are thus relatively lower.

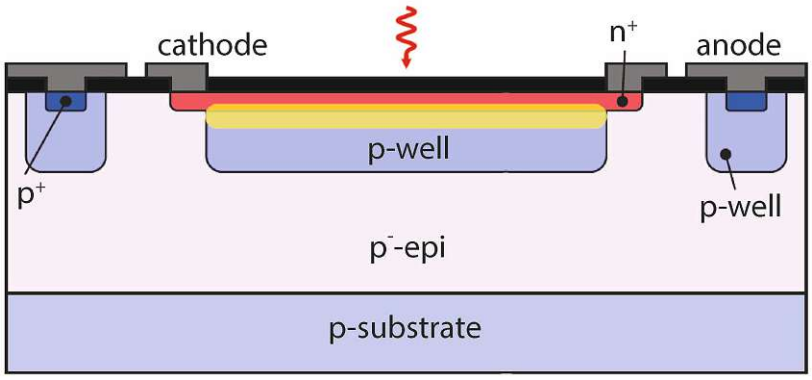

Figure 2.1: *SPAD cross-sectional (not to scale) [34]*

2.3 Basic Structure

From the problem described in 1.1 and the comparison of the detection technologies with the respective functional peculiarities mentioned in the previous chapter, as well as consultation with the relevant staff at the institute, the decision was made to use a SPAD-based receiver or a detector chip consisting of pixels, whose functional principle is based on SPAD behavior.

The design phase starts with the definition of a block diagram. The line sensor consists of individual line elements, also known as pixels, the necessary logic (consisting of input and output signal management) and the corresponding options for contacting downstream off-chip circuit technology. These line elements consist on the one hand of the quench/reset circuit and on the other hand of the SPAD which, as already explained in the previous chapter 1.3.2, represents the actual sensor. The numbering of the line elements shows that it is an interleaved pixel structure. This is based on the requirement to detect optical crosstalk and apart from that to minimize the dimensions of the finished sensor chip.

Special attention was paid to the interface of the chip during the design. Due to the high number of contacts for a single pixel (adjustability of quench and reset times, reference voltages, gater inputs, signal outputs, etc.), an appropriate interface and sophisticated signal routing is required. In order to reduce the number of bond operations necessary to a minimum, corresponding input and output interfaces have to be developed.

2.4 Pixel Base Design

A crucial building block for the design of a pixel element is the quenching circuit. As a starting point the quenching circuit from Dervic et al. [9] was used. Figure 2.2 shows the circuit diagram for this design.

For a better description, this triple voltage quencher circuit (TVQC) has been divided into blocks. In the course of the following functional description, the individual blocks will be discussed.

The following description assumes a detection ready SPAD. In the ready-to-receive state, neglecting the voltage drop of a few 10 mV across M_{A2} and M_{R2} due to compensating leakage currents, the potential of the SPAD cathode is at a value of about 3.3 V with respect to ground. If an avalanche is triggered, the voltage at the cathode breaks down. The differential amplifier located in block B detects this voltage dip and switches on the quenching transistor M_{Q1} located in block A via the level shifter (consisting of the MOSFETs M_{10} to M_{12}) and the voltage V_{Quench}. The cathode potential at V_{cath} is set to -6.6 V, resulting in a reverse voltage across the SPAD below the breakdown voltage. By varying the two supply voltages V_{SUP1} and V_{SUP2}, the excess bias voltage can be changed, at least within the limits of the application. The influence of this quantity has already been described in 1.3.1.

The differential amplifier consists of a two stage design, where the second stage is used as differential-to-single-ended conversion stage (M_3 to M_6). This is also the area in which an attempt was made to increase the performance, especially the detection speed, by changing the circuitry. More details will follow in the next section. The current mirror consisting of resistor R_{PQ}, p-MOSFETs M_{A1} and M_{A2} is used to bias the differential amplifier stage and compensate the leakage currents.

Besides the mentioned quenching transistor M_{Q1}, block A also contains the reset transistor M_{R1}. These two transistors, which have a switching function, are each provided with a cascoded series transistor M_{Q2} and M_{R0}, respectively. These transistors, which are operated with constant DC bias voltages V_{P4} and V_{P3}, in a gate circuit, serve on the one hand to increase the slew rate, the switching bandwidth and the gain, and on the other hand are necessary not to exceed the low maximum reverse voltages caused by the technology.

Since the differential amplifier or the voltage collapse detection cannot react infinitely fast, there is a phase of passive quenching (with M_{A2} as load) after the avalanche has

been triggered until the onset of active quenching. When designing this circuit, care was taken to ensure that the dead time is configurable. The dead time, neglecting the signal propagation times, consists mainly of the sum of the quench and reset times. For this purpose, special attention was paid to the adjustability of the reset and quench circuit sections. Sections E and D, respectively, are responsible for the durations of the quench and reset pulse, respectively.

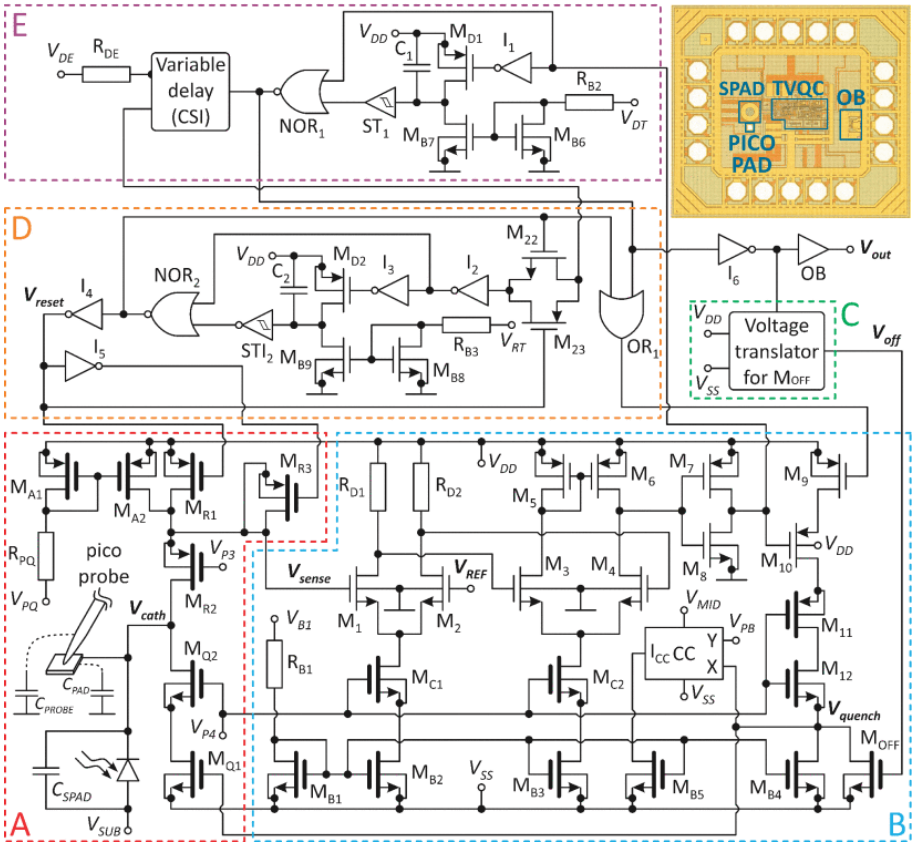

Figure 2.2: *Schematic of the triple voltage quench circuit [9]*

The circuit blocks for pulse shaping are similar for both pulses. The following description of the mode of operation for the pulse length definition of the quench pulse applies analogously also to the reset pulse. At the same time as the quenching transistor is driven, the transistor M_{D1} is switched off. The discharge current, defined by the current mirror M_{B6}/M_{B7} with resistor R_{B2} and voltage V_{DT}, leads to a discharge of capacitor C_1. The associated voltage drop is detected by the Schmitt trigger ST_1, thus terminating the quench pulse.

The variable delay element drawn in block E is a current starved inverter element whose task is to prevent overlapping of the quench and reset pulses.

By varying the quantities V_{DE}, V_{DT}, V_{RT} and R_{DE}, the dead time can be adjusted, ranging from 7.9 ns to 200 ns. The resistor R_{DE} is an optional resistor which is external, i.e. off-chip, and is therefore only installed after the bonding process if required.

For the sake of completeness, it should be mentioned that the transistors, which are shown with a wide solid line (e.g .: M_{Q1}, M_{Q2}) were implemented as 5V types, i.e. the typical maximum drain source voltage is 5V. In contrast to this, transistors with narrow solid lines (e.g .: M_1, M_2) are designed as 3V types. Analogous to the 5V types, the typical maximum voltage mentioned in the specification is 3V. In an application like the one shown in 2.1, which primarily requires a high switching speed, 3V types are to be preferred if possible, as they are superior to 5V types in this aspect. If a higher drain source voltage is required for functional reasons, the use of the higher-voltage types is preferable. Further information can be found in the specification [33].

2.4.1 Speed Improvement Using Bipolar Junction Transistors

As already described in 1.3, a reduction or a limitation of the charge carriers participating in an avalanche is desirable for various reasons (APP minimization, lower thermal load, etc.). A deactivation of the avalanche breakdrown as soon as possible is therefore essential. The comparator and its differential amplifier are signifcant for the detection and subsequent active quenching of an avalanche breakdown. In order to minimize the time between input voltage drop at V_{SENSE} and the level change at the output of the respective amplifier, there are a few parameters, which can be changed in the course of optimization. A reduction of the differential voltage (reference voltage minus sense voltage) necessary for detection comes to mind first. In the single pixel from [9], the base design for the quencher, this detection threshold was set to a level of 100mV. Reducing this amount is easier in single pixel applications than in array applications. It has to be noted that noise, whether internal or external, can cause the differential amplifier to trip incorrectly. A correspondingly small detection threshold level can only be selected if the signal to noise ratio is sufficient.

Another possibility is to increase the general switching speed of the differential amplifier stage. All components of the TVQC, with the exception of ohmic resistors, support and timing capacitors and other smaller applications, are based on field-effect transistors (FET), more precisely metal-oxide-semiconductor field-effect transistors (MOSFET). They have been, for some time, the dominant transistor type for CMOS IC manufacturing, as opposed to bipolar junction transistors (BJT). There are many reasons ranging from relative ease of fabrication to power-saving switching characteristics.

However, BJTs have better characteristics than FETs in certain areas. These include better bandwidth, higher gain and better (faster) driver capability. Furthermore, the gain of BJTs shows a proportional behavior with respect to the temperature. This means that with increasing temperature, which is inevitable in an array, the gain increases, which in turn increases the (switching) speed.

Unless otherwise indicated, BJTs with a single emitter input, an emitter length of 5 µm and a corresponding emitter area of 2 µm^2 were used for the following illustrations. This type of BJT is referred to as QNB2 due to its quasi-neutral base model behavior and two-base design.

In the course of the creation of this line sensor, attempts were made to use the advantages mentioned of the BJT in one form or another. Figure 2.3 illustrates the first point of attack for these improvement attempts. The MOS differential amplifier with the follow-up differential-to-single-ended conversion stage was replaced by a BJT differential amplifier. As BJTs, as already mentioned, require a significant current for amplification and switching, especially compared to MOS, a source follower is necessary for the sense input.

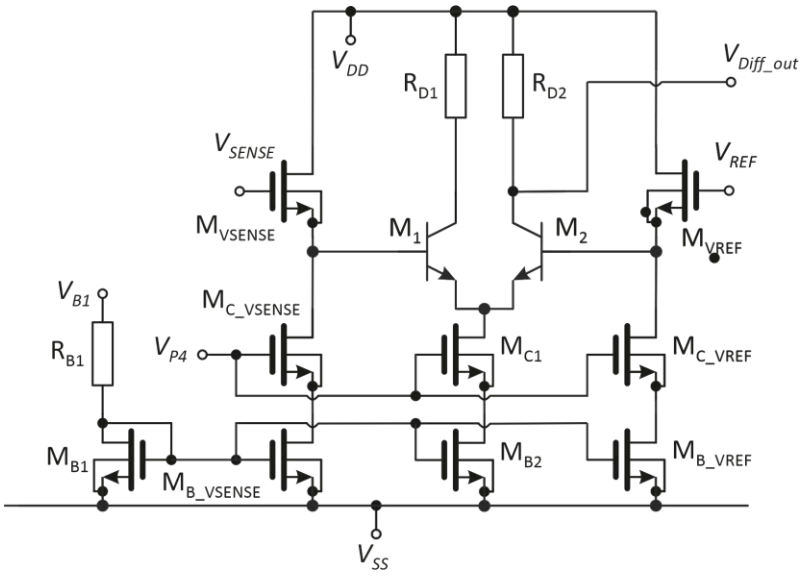

Figure 2.3: *Circuit diagram section of the BJT differential amplifier*

Another improvement of this circuit was to pull the two emitters of this differential amplifier stage to their own emitter potential $V_{EE,new}$ (V_{SENSE} - 0.7 V > $V_{EE,new}$ > -6.6 V). Again, a source follower is necessary, because if the emitter potential would be too low, e.g. 1 V, then the base pulls the cathode voltage (actually V_{SENSE}) to about 1.7 V.

Consequently the voltage drop between source and drain of transistor M_{A2} (see Fig. 2.2

amounts to 3.3 V - 1.7 V = 1.6 V thus operating in constant current range and reducing the adjustable range of the excess bias.

If a source follower is used, the overvoltage of the SPAD is not reduced. But the BJT (emitter to $V_{EE,new}$) then pulls the source to $V_{EE,new}$ +0.7 V, increases the current in the source follower by its base current and can go into saturation.

Using the BJT model provided, various simulations were performed for dimensioning. In certain constellations significant improvements in performance (improvement in slew rate; reduction in transit times) were achieved. The circuit shown in Fig. 2.5 gave the best simulation result in terms of detection performance. The cascode transistors M_{CAS1} and M_{CAS2} used to compensate for the Miller capacitance and to increase the gain have a positive side effect of reducing the maximum voltage that occurs at the transistors M_1 and M_2.

Using parameters that represent the typical component parameters results in an improvement in the response time with a constant reference voltage level of around 235 ps. The Fig. 2.6 and its excerpt in Fig. 2.7 show the voltage curve at the cathode.

However, this performance increase was not guaranteed over all operating ranges. Worst-case simulations showed that there were no significant advantages over the (proven) MOSFET solution because, depending on the boundary conditions, the circuit tends to trigger incorrectly (simulation with worst-power parameters) and not to converge to a stable operating point (worst-speed parameter set). The increased power requirement, resulting from the higher currents, could be counteracted (only rudimentarily) by creating a separate reference potential for the source follower and differential amplifier stage. However, the current requirement of the differential amplifier of the BJT variant is significantly higher than in the original solution with FET. The simulation using typical device parameters resulted in a differential amplifier current requirement of around 200 µA for the FET solution, while the BJT variant has a differential amplifier current consumption of around 1.7 mA under the same circumstances.

As a further possibility to increase the performance in addition or as an alternative to the replacement of the MOSFET-based differential amplifier with a BJT-based one, there was the possibility of replacing the MOSFET quenching transistor used in the original design also with a BJT one (M_{Q1} in Fig. 2.2).

Fig.2.10 shows the circuit diagram section with the relevant components. The signal that comes from the differential amplifier and a subsequent level shifter is called V_{Quench} in the original version (2.2) and serves as the gate signal for the FET M_{q1}. It is necessary to

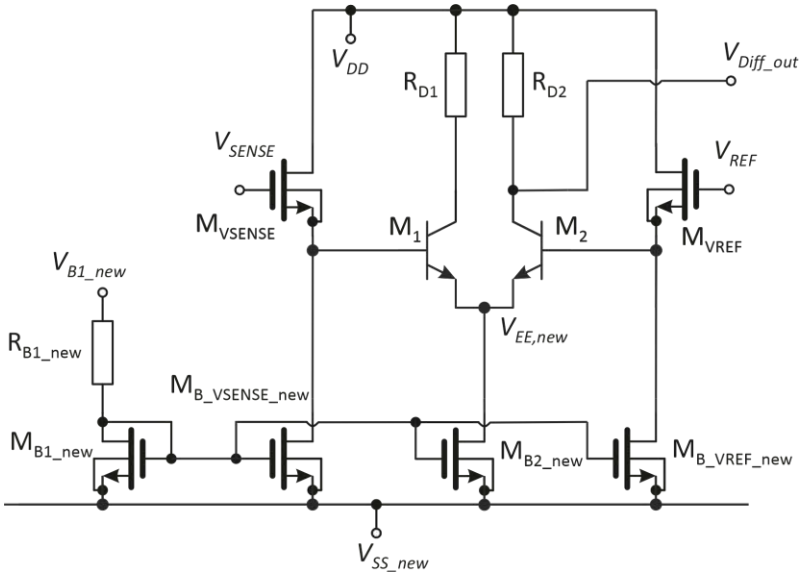

Figure 2.4: *Circuit diagram section of the BJT differential amplifier with new V_{SS}*

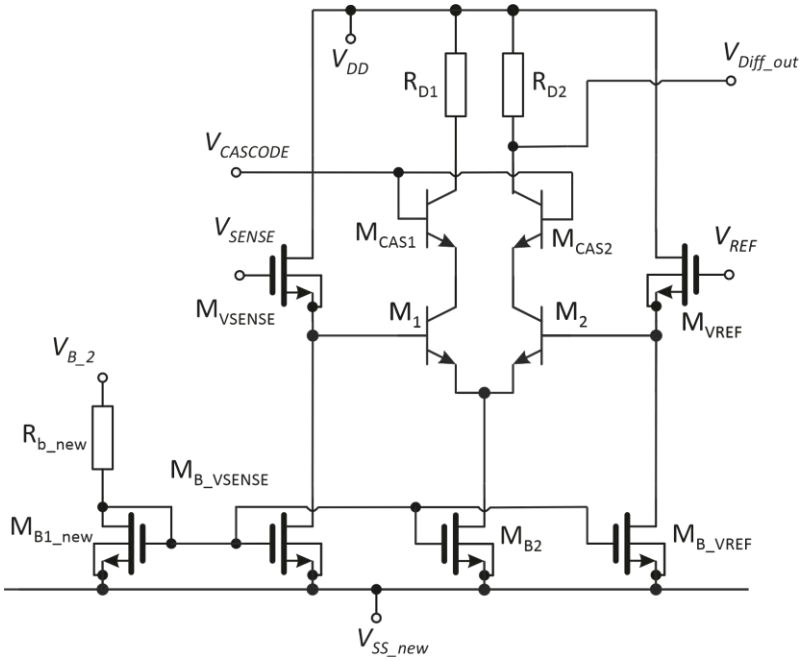

Figure 2.5: *Circuit diagram section of the BJT differential amplifier with new V_{SS} and Cascode*

provide a base current for the Quench BJT M_{Quench}, which is significantly higher compared with driving the gate of a MOSFET. Due to the high-impedance output character of the level converter from which the V_{Quench} is driven, direct control of M_{Quench} cannot achieve the required improvement in performance. M_{Quench} is quickly brought into saturation. In order to allow the SPAD to be recharged to the required detection potential, the BJT must

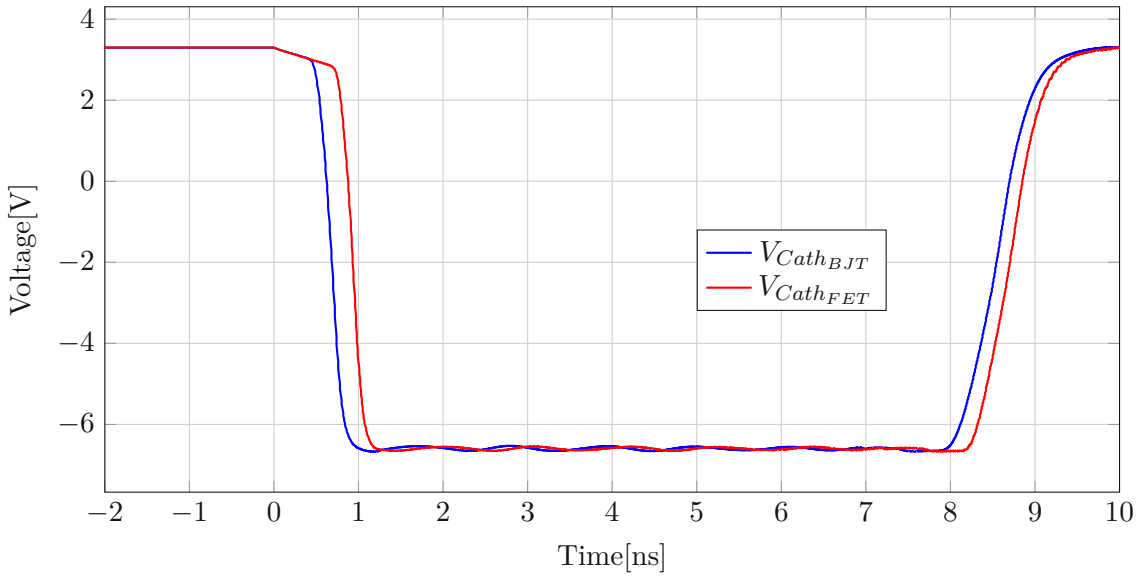

Figure 2.6: *The blue line represents the cathode voltage using the differential amplifier stage shown in 2.5. The red one shows the behavior in the original implementation (Fig. 2.2)*

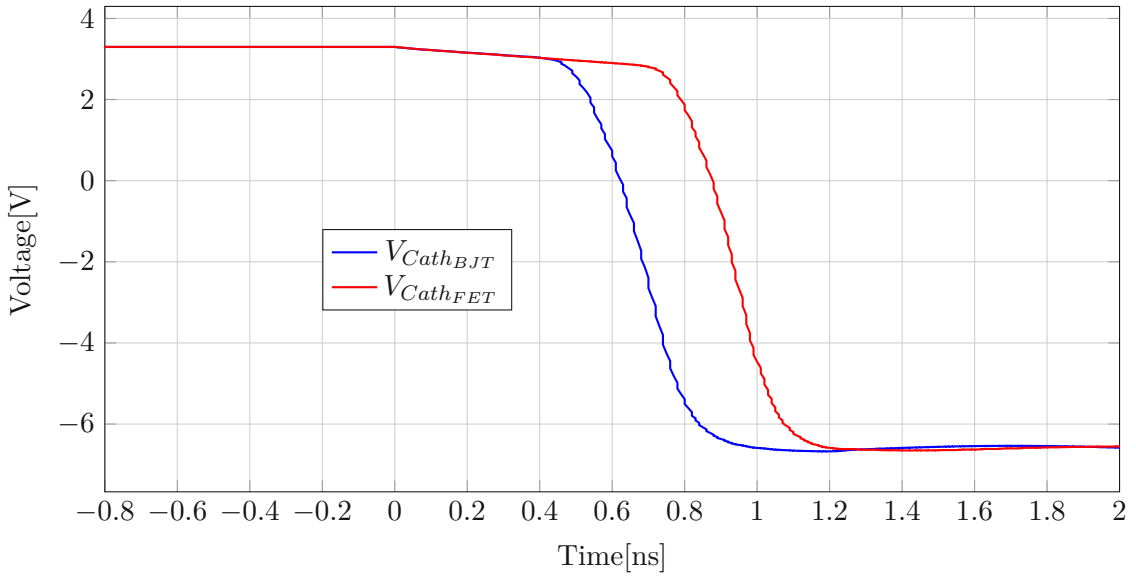

Figure 2.7: *Detail of the falling edge of the voltage curves shown in Fig. 2.6*

be brought back into the cut-off state as quickly as possible. Hence an intermediate stage with a separate supply voltage V_1 and consisting of transistors M_{Q1} and M_{Q2} is added.

The behavior of BJTs in the saturation range is problematic for the turn-off process. The

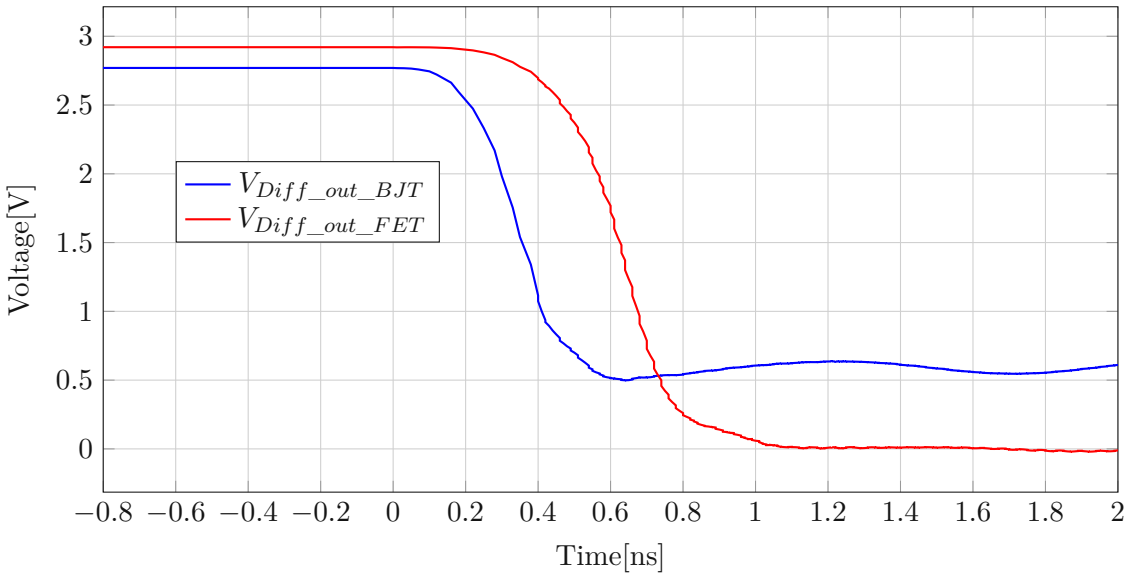

Figure 2.8: *Output of the detection unit depending on the implementation (falling edge). The blue line represents the output voltage using the differential amplifier stage shown in 2.5.The red one shows the voltage at the gate of transistor M_{10}, which is the output of the differential amplifier stage in the original implementation. (Fig. 2.2)*

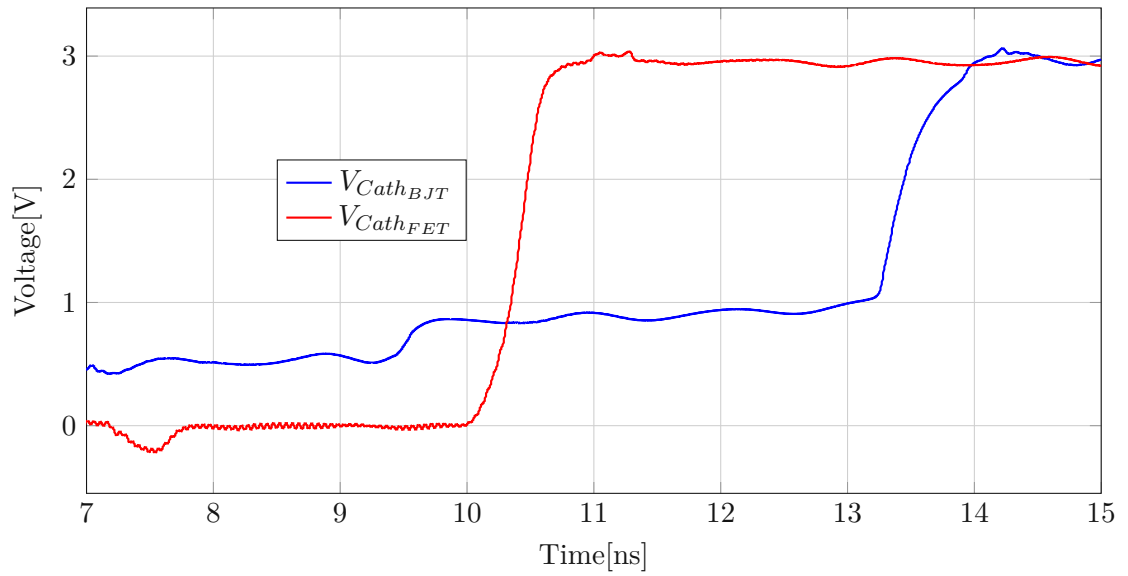

Figure 2.9: *Output of the detection unit depending on the implementation (rising edge). The description of the signal curves is analogous to that in Fig. 2.8*

high base current, which is useful for fast turn-on, i.e. the fast transition from the cut-off to the conducting saturation region, has a significant disadvantage. Many carriers are

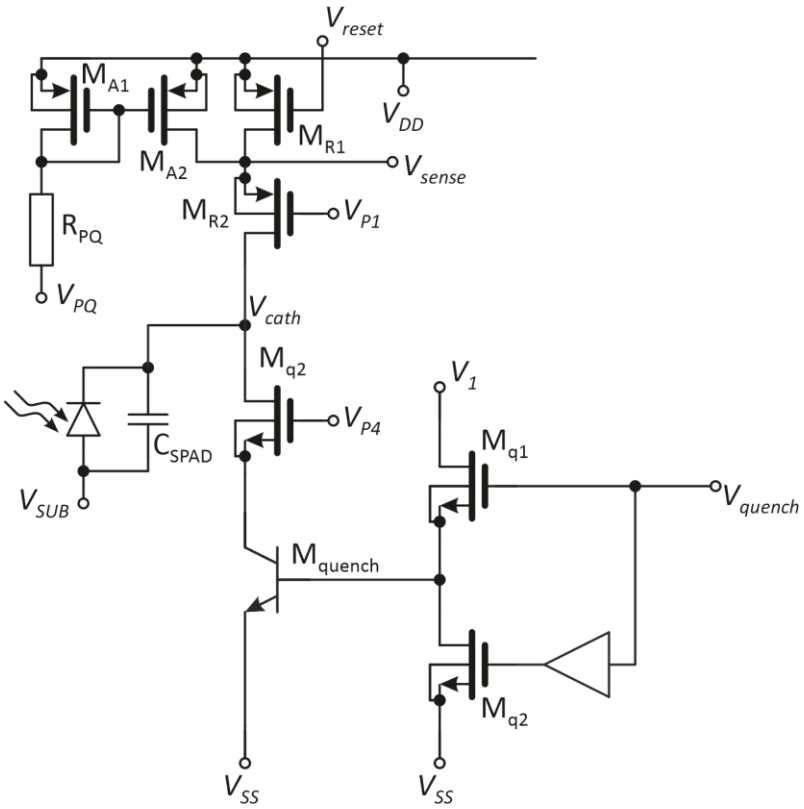

Figure 2.10: *Circuit diagram section of the BJT Quench version. The BJT considered here differs from those previously used. The emitter area is twice as large at* $4\,\mu m^2$*, this is the result of increasing the emitter length to* $10\,\mu m$.

accumulated in the base region in a short time. After the V_{Quench} switching pulse has dropped, the base potential or the base-emitter voltage would not change abruptly, but only slowly decrease until the base charges are removed.

Simulations have shown that this switch-off process would not take place quickly enough, i.e. the BJT M_{Quench} would not completely block if the reset process was started again by the FET M_{R1} after the quench and hold-off phase had ended. To speed up the BJT blocking, FET M_{q2} was added. This would significantly accelerate the removal of the base charge by using a negative base current.

As in the case of the BJT differential amplifier, simulations have shown that the desired increase in the slew rate or reduction in transitions time were not guaranteed over a sufficiently large range of boundary conditions or that the disadvantages and risks for this application are unacceptable. In order to identify an unfavorable constellation of component properties and parameters of different components, which in interaction would cause incorrect behavior or unstable functionality, consideration was given to simulating the

most diverse variants by means of a Monte Carlo simulation. Unfortunately there was no model library for the BJT model that could be used for such a simulation in the simulation environment used (in contrast to the FET models). Accordingly, it does not make sense to go into the results of Monte Carlo simulations, as they are not very informative for the designed circuit.

Above all, the risk of the BJT being destroyed by an excessively high collector-emitter voltage in the non-conductive state (worst case simulations have shown a long-term load of up to 7 V with an already very optimistic aussumed maximum voltage of the BJT of 6 V). Finally, in Fig. 2.11 all considered variants and their effects to the performance of active quenching of the SPAD are shown. The performance gain of around more than 200ps when using a differential amplifier consisting of bipolar junction transistors compared to a type consisting of field-effect transistors can be clearly seen. If a BJT is used as a quench switch instead of an FET, the slew rate is increased during the switching process, as shown in the figure above. The preliminary stage of the FET M_{Q1} required for the BJT implementation shown in 2.10 leads to a time delay, i.e. a shift in the switching time compared to the differential amplifier BJT and cross-switch FET implementation.

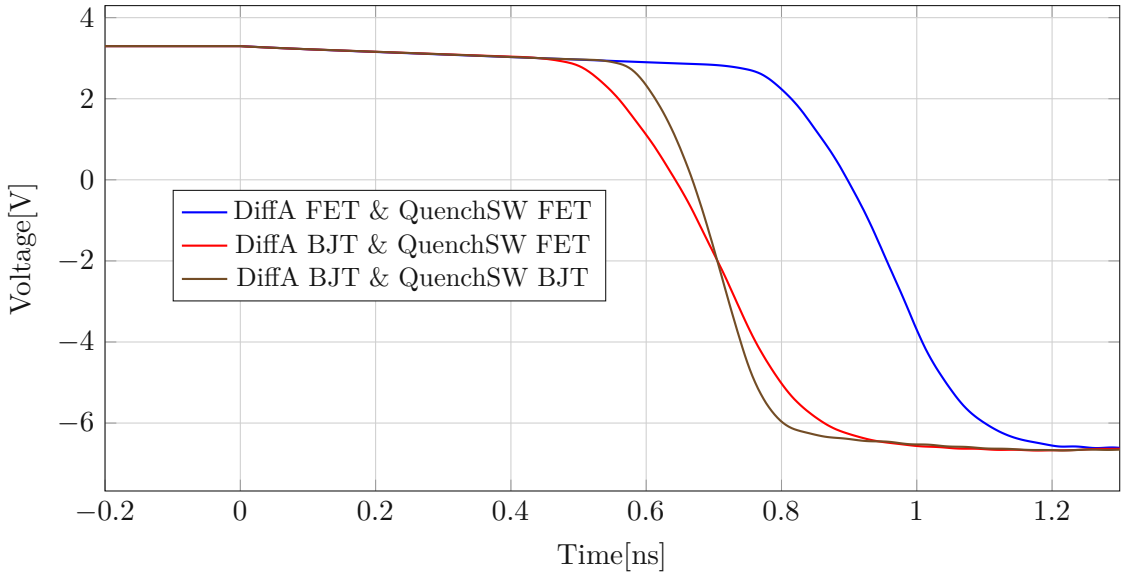

Figure 2.11: *Cathode voltage during active quenching for different design variants. DiffA stands for differential amplifier and represents the use of the respective solution (FET for Original in Fig. 2.2 or BJT in Fig 2.5). Similarly, QuenchSW stands for quench switch and associated implementation.*

2.5 Sensor Floorplaning

Figure 2.12 shows the floor plan or the basic scheme for the layout of a pixel. To achieve a high number of pixels and to minimize the pixel distance in one dimension, special care has to be taken to reduce the length parallel to the sensor (SPAD) line. While the length in the direction of detector line should be as short as possible, the chip area, needed for pixel electronics was extended to the orthogonal direction of the line sensor.

In the course of the design process, special attention was also paid to the possibility of simple expansion, i.e. scaling the number of individual pixels. Voltages that are made available directly, as references or indirectly as drivers of bias currents for all pixels, regardless of the state (triggered, untriggered or quenched), are brought up by vertical metal strips (orange). The circuit components can then be reached via corresponding vias on the metal layers relevant to them.

The Gater input (for temporary, permanent quenching or for resetting the latch) and the two output signals (latched for checking validity or unlatched for the fast trigger signal) for the individual pixels are on the outer edge, orthogonal to the general parameterization and supply lines attached.

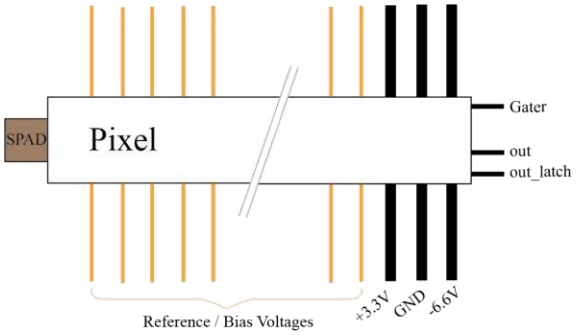

Figure 2.12: *Floorplan of a pixel*

The application of the interleaved design leads not only to a simplified interface but also to an increase in the pixel density. As depicted in 2.13, a small pixel pitch can be achieved through clever selection of the pixel height and the interwoven positioning of the even and odd elements. The aforementioned (input) voltages are made available to all pixels via the connected metal strips. Based on this design principle, the number of pixels can be increased as required if the metal strips are dimensioned sufficiently. As will be shown in the further course, it is advantageous for expandability if the number of pixels are an integer multiple of four.

Based on the assumed mean current flow from the voltage sources V_{SUP1}, V_{SUP2} and

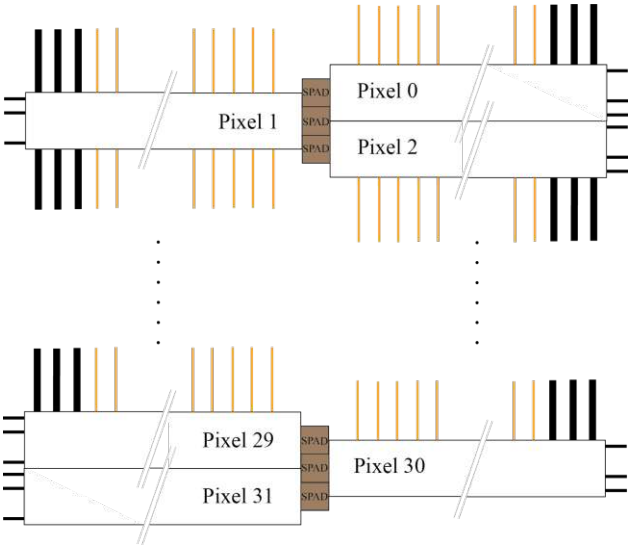

Figure 2.13: *Floorplan of the sensor line*

the current flow to the reference potential GND, the required widths were determined under the assumption that the maximum voltage drop over the assumed length of 1 mm (based on the fact that a half-chip, consisting of the named 16 pixels, is supplied from both sides) on the respective metal line should be less than or equal to 50 m V. Based on the manufacturing instructions, values for the width of the metal line of 35 µm were chosen for V_{SUP1} and 23 µm for the supply rails of V_{SUP2} and GND.

The maximum specific current load mentioned in the process specification document for a metal layer is 1 mA µm^{-1} width in the worst case.

With the widths used, the specified voltage rails V_{SUP1}, V_{SUP2} and GND with the considered current values of 12 m A, 10 m A and 11 m A result in an actual current load of 0.34 A µm^{-1}, 0.44 A µm^{-1} and 0.48 A µm^{-1}, respectively.

2.6 Chip Interface

From the requirements mentioned in the introduction for the smallest possible number of bond pads, the need for a corresponding interface arises. Due to the functionality, some (externally supplied) analog reference-, bias voltages and anaolg signals cannot be replaced by a digital interface, no matter what kind of interface is used. These quantities - beside the supply voltages these are voltages - which control the pulse lengths by the functionality described in 2.4, have to be applied to the chip directly via bond pads.

The following signals and the interface description consistently refer to half the sensor. The background of the design is the interleaved design approach. Beside a simplified implementation and better scalability, the solution also offers the possibility to detect (faulty) optical crosstalk.

Optical crosstalk can be better recognized if neighboring pixels are triggered one after the other at a short time interval. This will be recognized by the slightly time-shifted activation of the detection signals of both parity ranges and the subsequent evaluation of the address outputs. It is assumed that optical crosstalk is much more probable for neighboring pixels compared to pixels that are further apart. The highly unlikely case, according to which two pixels with the same parity undergo an optical interaction, i.e. crosstalk, is guaranteed by an extra signal, which represents the validity within a parity.

If the number of photons that can be detected is limited to a single one, the evaluation of the address signals is not necessary to be done if two or more pixels within a parity have been triggered.

A more detailed description of the signals can be seen under point 2.6.2.

2.6.1 Digital Input Interface

As already mentioned in the introduction, there should be a possibility to permanently deactivate individual pixels. Due to imperfect yield in the manufacturing process of the chip, possible damage from high avalanche currents and defects due to electron migration after a long period of operation, individual pixels can be impaired. On the one hand this damage may lead to blind detectors and, on the other hand, to pixels that have a permanent breakthrough. The secondly named issue is more problematic, thus a need to permanently "mute" individual pixels exists. Otherwise, these pixels, known as "screamer", would make the sensor unusable. With the option of deactivation, the array can continue to be used without the defective elements.

In addition to permanent deactivation, there should also be a way to deactivate all pixel elements together - or in other words, to enable a gating operation (as described in 1.3.2). This should be possible at any time during the operation. Accordingly, a fast reactivity is required instead of the permanent deactivation, in which a pixel-by-pixel activation/deactivation is done before the main experiment.

Regarding to the requirements for the possibility of permanent deactivation of any pixels and the attempt to need minimal communication to the chip from off-chip, there is a need

for a logic circuit that fulfills this functionality. The solution is a simple shift register used as a serial to parallel converter. Based on the operating requirements the deactivation of pixels in the initial configuration before an experiment starts, do not have any time requirements. Hence for input signals *Serial Data In* and *CLK* there exists relaxed demands of wireing for RF.

The high slew rate required for the "fast" gating input would pose a major problem without special precautions on the chip layout and PCB design. The oscillation tendency of the oscillating circuit consisting of bond wire inductance and capacitance of the components in the chip would increase dramatically and the effect of EM radiation on neighboring components could cause significant disturbances. To avoid this, a clock driver was implemented on the chip. This enables the required voltage swing from 0 V to 3 V on the chip to be generated by an externally supplied, zero symmetrical voltage signal with an amplitude of a few hundred millivolts. The use of a sinusoidal input signal is preferred in order to prevent problems with switching edges with a high slew rate, such as those that occur with square-wave signals.

2.6.2 Digital Output Interface

In order to integrate the outputs of the sensor chip into existing measurement equipment, an output driver designed for a characteristic output impedance of 120 Ω was provided for each signal. This is implemented by cascading CMOS inverters with increasing driver strength. The 120 Ω driver used was developed by R.Enne in the course of his works [35], [36], [37].

Trigger Output
The task of this output signal is to provide a fast trigger output signal for subsequent data processing. Not only the speed is important, but also the delay between an detection event and the corresponding trigger output. Additionally the debounced change of state at this output should be constant for each individual pixel.

Validation Output
Optical crosstalk between two adjacent pixels can be detected by comparing the two trigger outputs (even and odd). However, the trigger signal is only an OR of all pixels on the logical level. Thus, it cannot be distinguished if one or more pixels of a parity "fired". Therefore, a valdiation output is necessary. This output provides information about whether exactly one pixel of a sub array has a detection event or not.

Equation (2.1) shows the mathematical description of this output.

$$ValidOutput = (out_0 \cdot \overline{out_1} \cdot \overline{out_2} \cdot ... \cdot \overline{out_{15}}) + (\overline{out_0} \cdot out_1 \cdot \overline{out_2} \cdot ... \cdot \overline{out_{15}}) + ... \\ + (\overline{out_0} \cdot ... \cdot \overline{out_{13}} \cdot out_{14} \cdot \overline{out_{15}}) + (\overline{out_0} \cdot ... \cdot \overline{out_{13}} \cdot \overline{out_{14}} \cdot out_{15}) \quad (2.1)$$

Existing CMOS logic gates predefined in the associated design library were used for the implementation. Accordingly, gates with the necessary 16 inputs were not available. For this reason and taking into account the requirement for possible scalability of the chip, this logic was divided into blocks of four. Such a logic block can be seen in fig. 2.14. The outputs of these blocks are in turn summarized in a block shown and its output thus represents the validation signal for this parity group. Using one of *De-Morgan's* laws $\overline{a + b} = \overline{a} \cdot \overline{b}$, the logic block shown results in an application that uses only NAND gates.

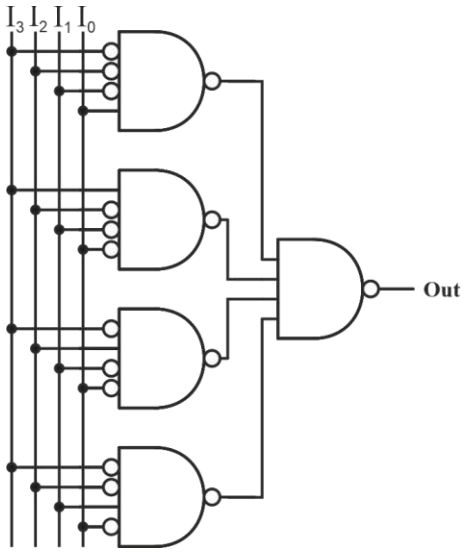

Figure 2.14: *Excerpt from the logic block for the Validation Output Block*

Output Addressing

Starting from one sensor half, i.e. from 16 pixels in the case of a valid detection, i.e. after trigger and validation signal, it still has to be output which of the pixels has triggered this event. In this case, a one-hot coding is used.

This functionality is best explained by means of the table 2.1. This shows with 16 states only a fraction of all possible combinations. However, the states not mentioned have no meaning, because they do not represent valid states which is indicated by the validation output.

Table 2.1: *One-Hot Decoder*

I_{15}	I_{14}	I_{13}	I_{12}	I_{11}	I_{10}	I_9	I_8	I_7	I_6	I_5	I_4	I_3	I_2	I_1	I_0	Out_3	Out_2	Out_1	Out_0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	1
0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0
0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	1
0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1
0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1

2.6.3 Analog Chip Interface

According to the original design, which is mentioned in 2.4, the need of lots of voltages for reference and parameterization tasks can be seen. Considerations of implementing a digital interface to set on-chip reference voltages instead of applying analog ones were discarde for reasons of complexity for resolutions demanded.

Figure 2.12 shows the procedure of the supplied analog values (orange) and the operating voltages, functionally also analog voltages (black) with external supply from "above".

2.7 Usage Adaptions

In order to simplify the subsequent signal processing, especially the validation of a single event (The event of detection of one and exactly one photon (or a detection event) by any array of more than one pixels.) by removing the influence of propagation jitter between the different pixel output signal, the output signal of every pixel is latched 2.15. This latch is set by a detection event. For reasons of simplicity, the reset is carried out by a falling

edge on the respective pixel. The use of an additional signal was discarded for reasons of complexity (more complex routing, additional interface module) and the lack of an explicit requirement.

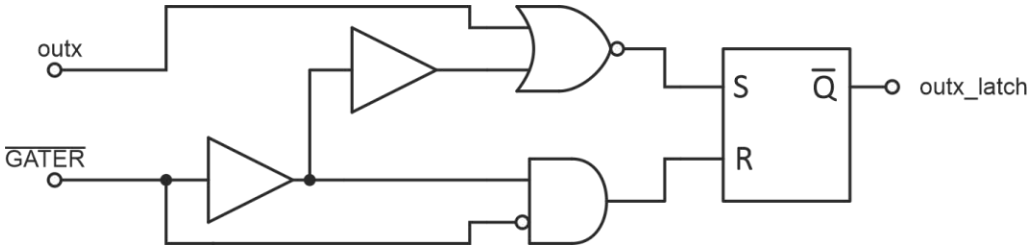

Figure 2.15: *Schematic representation of the adaptation*

The gater signal is low. Accordingly a high level leads to the deactivation of a pixel or to the permanent quenching of the respective element. If the gater signal, referred to as $\overline{GATER}$ in the following, is at the low level for a sufficiently long time, the output signal of the respective pixel element *outx_latch* (x stands for the pixel 0 to 31) is also set to low by default. If now a triggering event, i.e. a detection event occurs, the signal level is pulled high. This state is kept high until $\overline{GATER}$ has performed a transition from high to low again (edge-sensitive latch). Potential detection events are not detected by this signal. The quench in case of a detection is not influenced by this and takes place despite the already latched state.

In the original TVQC, the signal output, which signals a positive detection event, was positioned after the quench pulse shaping circuit part. Therefore there is always a delay which depends on the quench duration that has been set. In order to remove this dependency and to obtain a delay-free output signal, the signal tap was placed on the output of the differential amplifier. To incorporate the gate capability required and to ensure active quenching in the non-gated state, a small change is also implemented in the output stage of the differential amplifier. This reduces the speed or deteriorates the performance. However, the safe state of actively quenching is an accepted disadvantage.

The element shown in Fig. 2.2 for setting the delay, i.e. the current starved inverter, was set to a fixed potential, namely V_{SUP1}, due to problems in creating the layout. The delay is therefore fixed and cannot be varied by a current as shown in the circuit diagram.

2.8 Function Description

2.8.1 Functional Description of a Pixel

Figure 2.16 shows a typical photon detection process for a parity group.

Starting with a first pulse on the $\overline{GATER}$, exactly with the falling edge, the latch of this pixel is being reset. At the beginning the cathode is quenched. After the $\overline{GATER}$ pulse, the voltage at the cathode rises back on its detection-ready value. In case of a detection, which is already described in 1.3.2 and 2.4 the non-latched output *outx*, where x is to be seen as a placeholder for any pixel, is pulled to high for the duration of the sum of quench and reset time (which are controlled by external supplied voltages as mentioned in 2.4). After recharging the SPAD, this output is pulled back to its low value. As described in the last section, the latched signal *outx_latch* is provided to the validation circuitry (see 2.6.2). This logic block combines all latched outputs of a parity as already described in detail in the previous chapter. As shown in the signal curves this non-volatile output has to be aknowledged via the mentioned $\overline{GATER}$ pulse. Based on the previously described behavior and the signal curve in the corresponding figure (*outx_latch* in 2.16), it can be obtained that the non-latched output, in contrast to the latched one, may be used as a trigger in the duration between the first detection and the reset.

2.8.2 Output Behavior of the Chip

The behavior of the evaluation logic is presented in the following two figures and lists of pseudo signals. While the signals with the blue background in Fig. 2.17 show the selected outputs of the chip of one parity group, the others are the voltages at the cathodes x and y and the associated latched output signals *outx_latch* and *outy_latch* of two pixels of the same parity.

Starting from a reset state (falling edge of a $\overline{GATER}$ pulse as described in section 2.7 preceded), a detection event occurs at the SPAD x, recognizable by the voltage drop on the cathode signal x. This detection can be seen at the *Trigger Output* signal. As already explained in 2.6.2, the trigger is a disjunction of the fast unlatched outputs of a half sensor. The *Valdiation Output* also becomes active. This output is defined by the high level of the *outx_latch* signal and the (assumed) low levels of all the other signals stored. A subsequent detection event leads to another temporary transition to a high level at the *Trigger Output*, recognizable by the second spike of the signal mentioned. The valdiation output remains

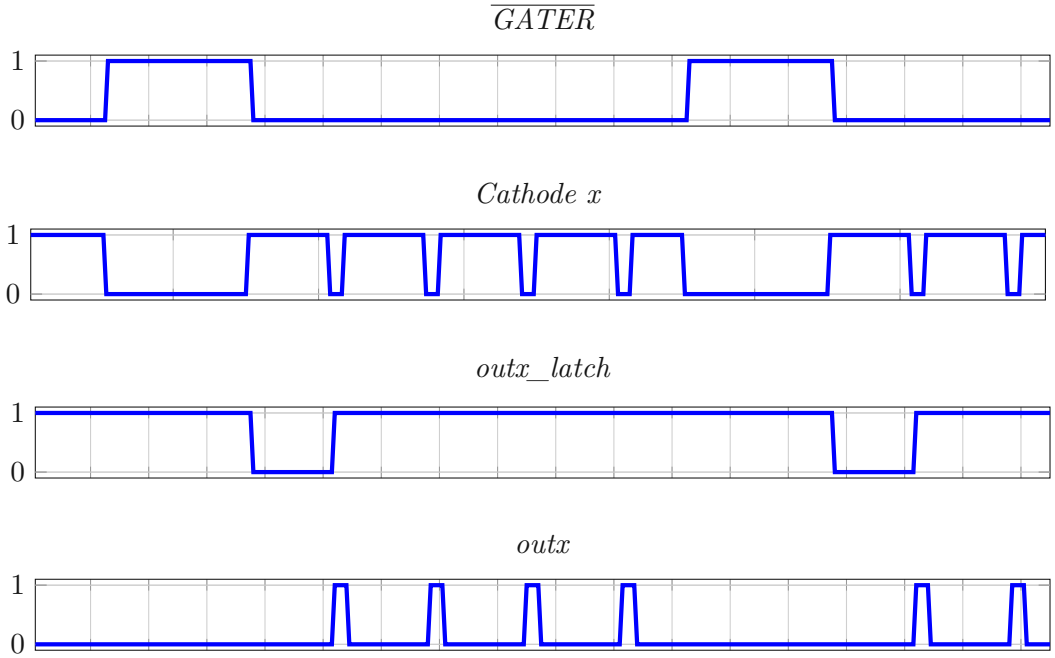

Figure 2.16: *Temporal behavior of internal chip (pseudo) signals in the case of a detection on a single pixel*

unchanged at the high level, since no reset has occurred. The next pulse at the *Trigger Output* is the event that occurred at element *y*. The reason for this is the fast *outy* output of the element named. Analogous to the processes already explained, the *outy_latch* signal is also set here. The level of the *Valdiation Output* signal is pulled to low by this high level, indicating that no permissible individual detection has occurred.

The four coding outputs, Out_3 to Out_0, show in descending order in Fig. 2.18, with active validation output information about where the valid single event took place. If the *Valdiation Output* is low, the coding represented by the signals mentioned does not correspond to the list in the table 2.1 and must be discarded.

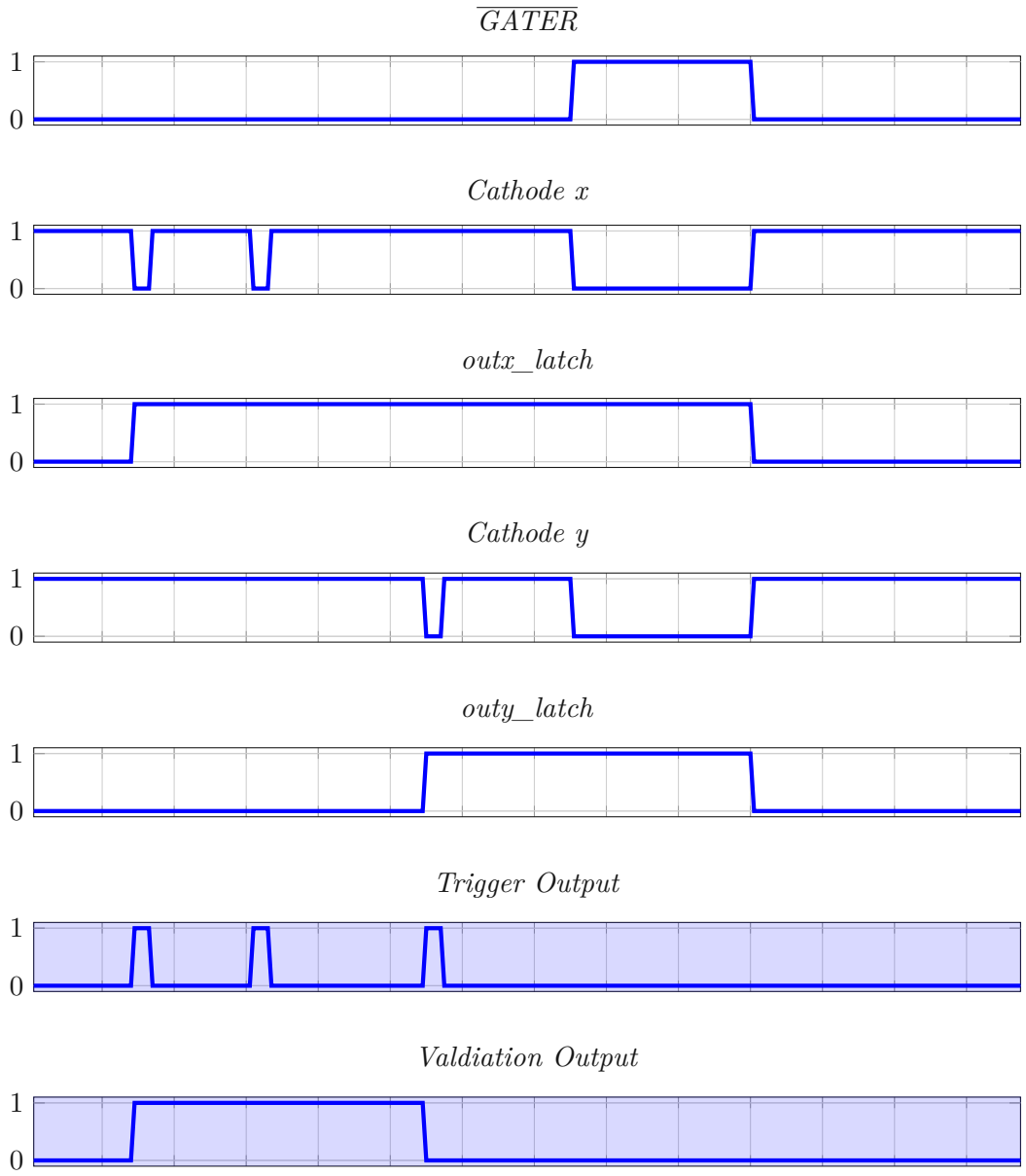

Figure 2.17: *Temporal behavior of internal and external (blue filling) chip (pseudo) signals in the case of a detection on two pixels in the same parity group*

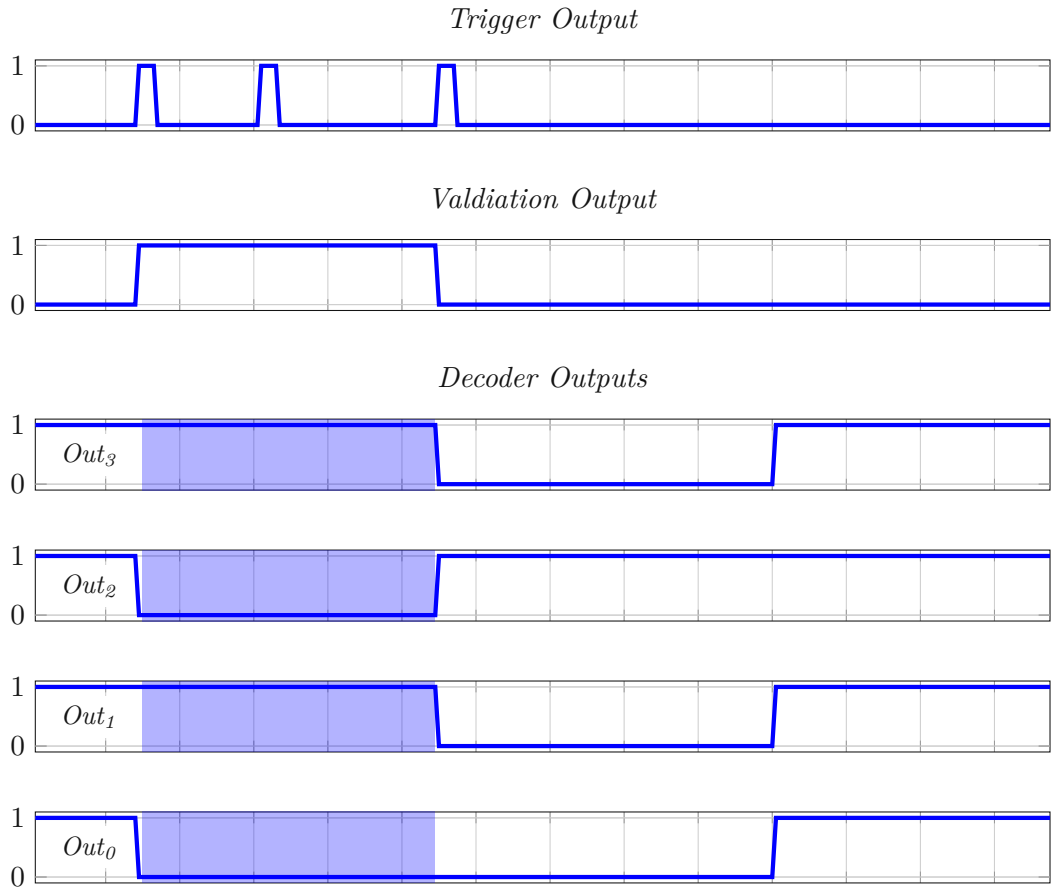

Figure 2.18: *Temporal behavior of chip coding output (pseudo) signals in the case of a detection on two pixels in the same parity group.*

Chapter 3

Results

3.1 Layout

3.1.1 Matching

The parameters and, to a large extent, the properties of circuit components are subject to large tolerances. Countless influences, which on the one hand result from ambient and environmental boundary conditions or their variance over time and space, and on the other hand are based on the spatial inhomogenity of the manufacturing process, can lead to significant restrictions or effects, especially in the area of circuit design. For this reason, appropriate measures to prevent or reduce these effects must be taken as early in the design phase and at latest when drawing a layout. In this context, these measures are referred to as *matching*.

As an example, the same alignment and spatially close placement of components can be seen here in order to compensate for the influence of anisotropic ion implantation angles. Furthermore, it may be useful to provide dummy components in order to bring the influence of unavoidable concentration differences, which occur in the course of the diffusion process steps, especially at the edges, on the relevant components to an insignificant level in the best case. The matching can be positively influenced by paying special attention to the variables relevant for the component behavior (W and L for MOSFETs, emitter area for BJTs) in the design. Local fluctuations and the influence of inhomogeneity in the materials used can be averaged out by using large dimensions. An important and proven method,

which is used to desensitize the circuit parts to relative variations, is called Common Centroid Layout (CCL). Circuit parts that require a high level of matching in relation to one another are split up into several identical parts of the same size and positioned to enable as many symmetries as possible. Influences such as doping and temperature gradients caused by the manufacturing process and local power emitters, are compensated by this arrangement.

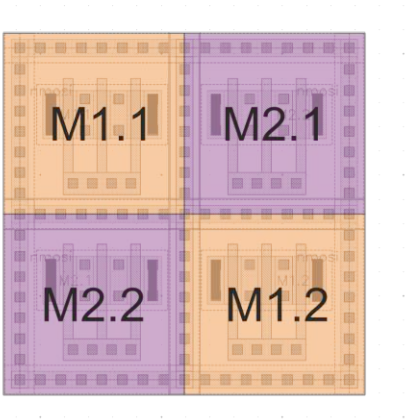

Figure 3.1: *Example of a CCL implementation based on the layout of the FETs M_1 and M_2 from 2.2*

3.1.2 Layout of a Pixel

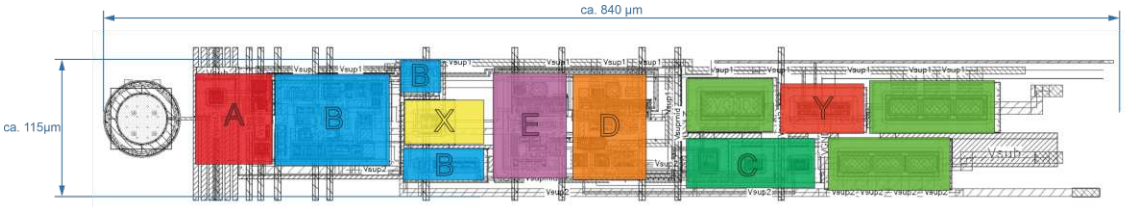

Figure 3.2: *Photo of the constructed pixel layout with selected comments*

Figure 3.2 displays the layout of a line element, which, as already mentioned in 2.4, is based on the TVQC. The individual function blocks A to E are colored analogous to the blocks in 2.2. On the far left, the SPAD is recognizable. The structure of this SPAD has already been mentioned in 2.2. The implementation of the diode and especially on the local positioning of the one-dimensional array is then added in 3.1.4. Then the high-voltage switch in block A, followed by the high-speed comparator are placed. The areas X and Y are the circuit adjustments or changes mentioned in 2.7. The blocks E and D illustrate the quenching and reset molding and timing blocks. Block C includes the voltage translator mentioned in 2.4.

The slim design presented in section 2.5 and in Fig. 2.12 has the shape of a rectangle with the outer dimensions of approx. 840 µm x 115 µm. It has to be noted that the depicted version is the one with additional support and smoothing capacitors in order to achieve an easy expandability (recognizable on the light green and non-named rectangles in the right area). From a functional point of view these support capacitors could be removed. However, the high number of externally supplied voltages for parameterization and the associated high number of pads (see 3.2) and their stringing together leads to a significant increase of chip area which would otherwise remain unobstructed.

3.1.3 Layout of the Line Sensor

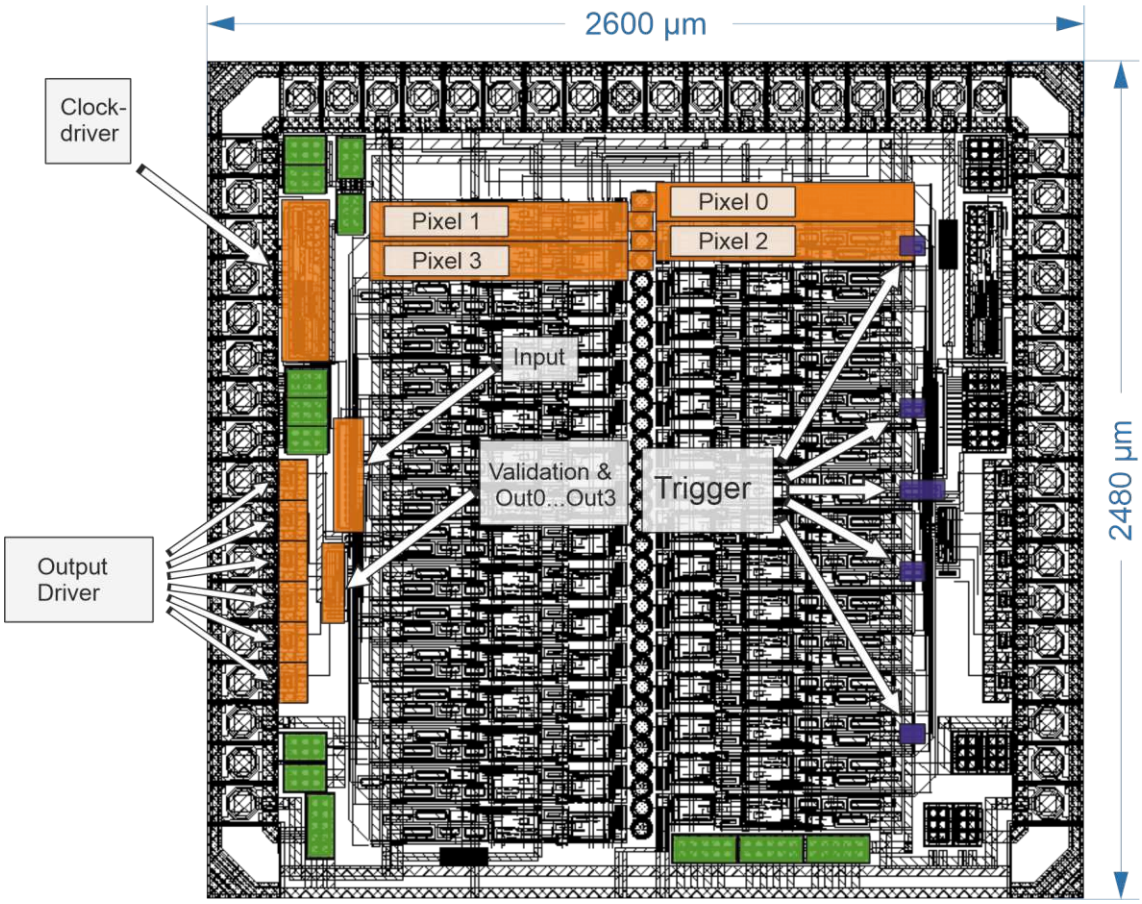

Figure 3.3: *Photo of the finished chip with selected comments*

The finished sensor can be seen in Fig. 3.3. The chip area is around 6.45 mm^2. The outer dimensions result as an integral multiple of the width of the pad element.

As already mentioned in 2.3, the even and odd pixel elements are arranged interleaved. The

left side is the one that contains the odd pixel elements and their logic components. The even elements can be found on the right side. On the side, the outputs of the individual pixels (latched and non-latched) are logically combined in groups of 16 inputs, equal parity with one another, in the form of asynchronous logic blocks named in 2.6. The same applies to the input interface mentioned in the same chapter for permanent and temporary deactivation of individual or all pixels. It is worth mentioning here that the input of the gate clock input has a clock driver, which enables the amplitude of the externally supplied clock sources to be reduced. These logic blocks are marked in an orange colour in Fig. 3.3 and provided with appropriate comments. For reasons of clarity, the logic function blocks have only been drawn on one side. For the other parity space in each case, these are also available in mirror symmetry.

There is remaining space within the chip that is not used by function or logic blocks, which results from the high number of externally supplied voltages and the pad elements required.These areas were filled with capacitors due to the lack of alternative functional occupancy. In Fig 3.3 these areas have been marked in green.

3.1.4 SPAD Line

The interleaved design described in 2.13 allows the distance between the SPAD elements (i.e. the resolution of the sensor) to be reduced to a value of 60 µm. In comparison, a non-interleaved design, which results from the non-nested stacked positioning of the pixels shown in Fig. 3.2, would result in a distance of 60 µm.

The aim is to maximize the active area relative to the total area required. Thus, as already mentioned in 1.3.2, a SPAD requires a corresponding edge area or further structural measures that require space in order to guarantee trouble-free operation in the desired states. In order to enable the relative active area to be maximized, these edge areas between two SPADs overlap. A single SPAD has an outer diameter of 64 µm, whereby the mentioned resolution results from the mentioned overlap of 2 µm per array direction. The active diameter, which also corresponds to the diameter of the associated p-well, of a SPAD is 40.5 µm. This results in an active area of 1288 μm^2 SPAD. The fill factor of a SPAD is thus calculated at 53 %. In order to realize the virtual guard ring described in 2.2, the n+ area shown in Fig. 2.1 with a diameter of 44.1 µm is required. [1]

[1] The fill factor is the quotient of the active area and the required total area. A rectangle with the width of the active area and the length of the pixel pitch was used as the total area.

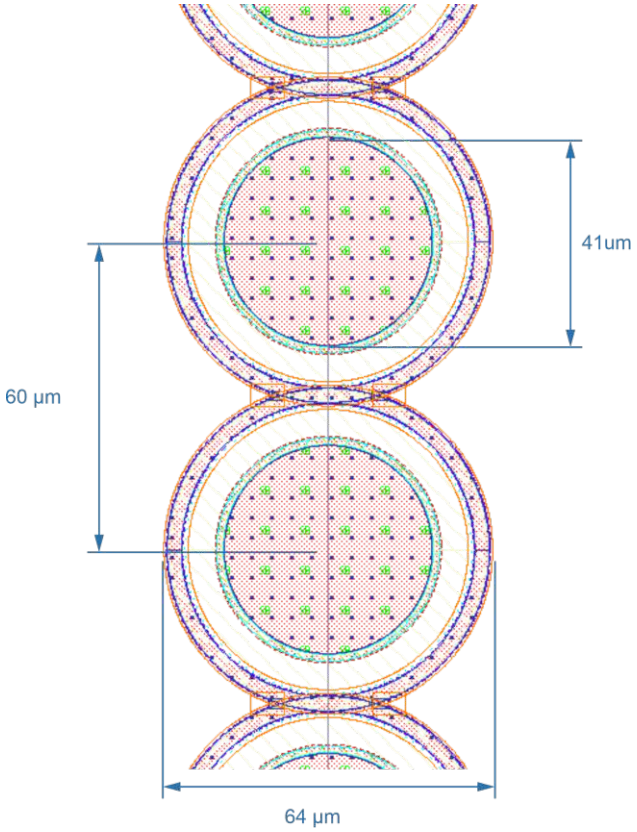

Figure 3.4: *Dimensions and relative positioning of the SPADs*

3.2 Post-Layout Simulation Results

3.2.1 Test Bench

All inputs and outputs were provided with a series inductance of 2 nH in the test and simulation environment. This inductance is intended to represent the bond wires or the behavior that is generated by them. The outputs have an impedance of 120 Ω. Accordingly, a terminating resistor with the same impedance was provided. All voltages mentioned in the following are measured at this load.

3.2.2 Typical Detection Event

Figure 3.5 shows a typical detection process on any element. It is based on a simulation with the approach mentioned in the previous paragraph for viewing the chip with the

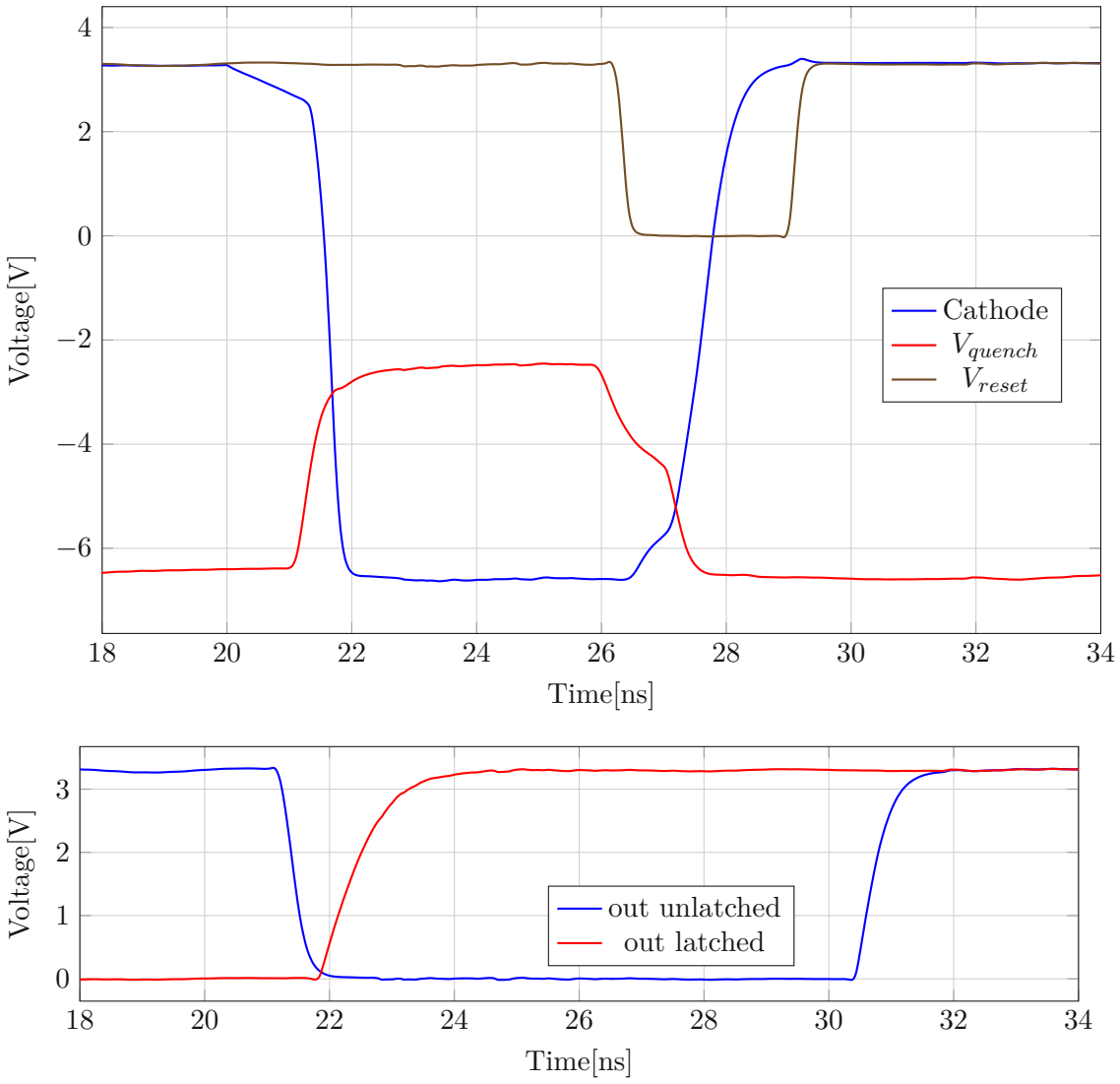

Figure 3.5: *Voltage curves of selected signals in the event of a detection event on a SPAD*

generated parasitic electrical quantities. It is of course not of importance for the further explanation, but for the sake of completeness it is mentioned that this detection event was carried out on element 14.

The upper part shows the voltage at the cathode as well as the voltage pulses for the quench or reset phase. The part at the bottom shows the outputs of the pixel, separated into latched and unlatched. It can be seen that the polarity, unlike in the design chapter, differs more precisely under 2.8.2 The reason for this is, on the one hand, the simpler

post-processing of the latched signal and, on the other hand, there are performance-related advantages.

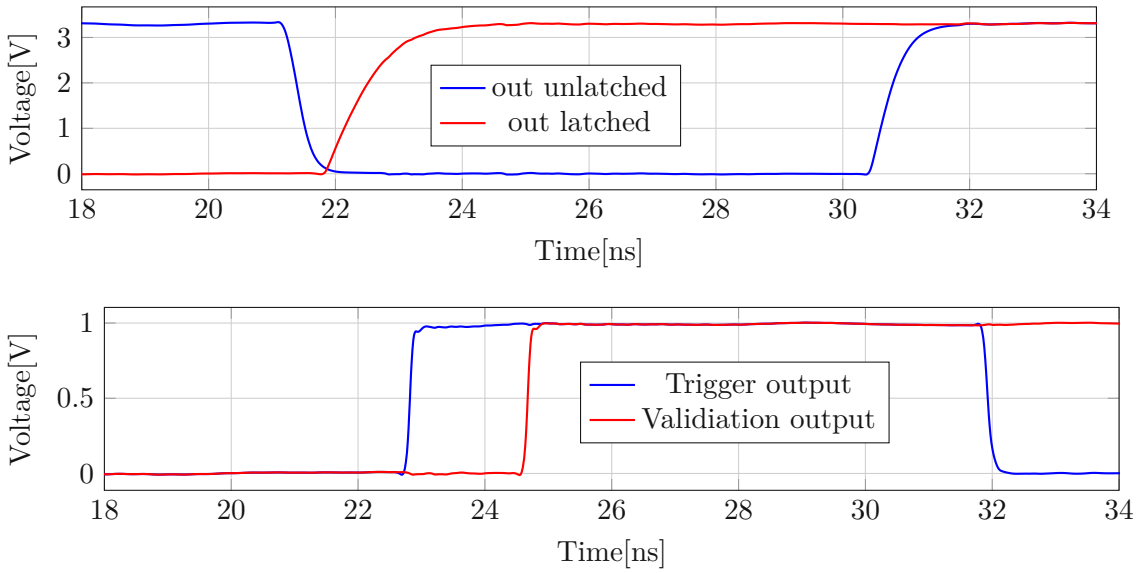

Figure 3.6: *Voltage curves of selected output signals in the event of a detection event on a SPAD*

Figure 3.6 shows the output of the digital output stage that logically combines the output signals of all pixels of a parity with one another. As in the previous figure, it was assumed that a single event is occured on one parity side. The time difference between the *Trigger* and the *Validiation output* is clearly recognizable. Based on this typical case, there is a delay of around 3 ns between a detection event at time t=20 ns and when the high level is reached at the *Trigger* output. For this delay a detailed description will follow in the further course of this chapter.

The necessary logic circuit to execute the behavior mentioned in the requirement for the one-to-one display of a single event, needs a relatively high number of logic elements, which results in the high throughput time and thus high delay compared to the logic gate for generating the *Trigger* output. The outputs for coding or naming the pixel and thus the identification of the location of the detection event (not shown here for reasons of clarity), as shown by the simulations, could have indicated that the pixels approach their clear valid state before the validation high level is reached. However, by definition, without corresponding signaling through the *Validiation* output, this cannot be seen as a reliable indication of a permissible single event detection.

3.2.3 Parameter for the Quench Time

In order to take into account the dependency of especially DCR and APP on the dead time in 1.3.1 and to adapt the aforementioned sources of noise and error during the operation of the chip developed to the corresponding application and the associated acceptance criteria, the already used circuit option described in 2.4 for quenching duration adjustment was used.

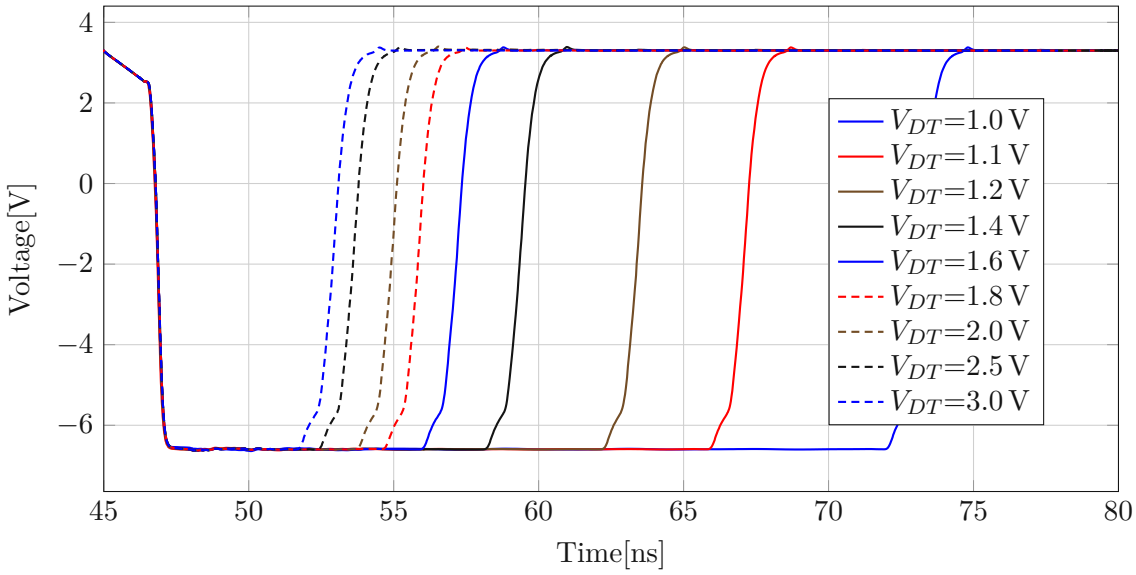

Figure 3.7: *Influence of the voltage V_{DT} on the quench duration shown on the course of the cathode voltage*

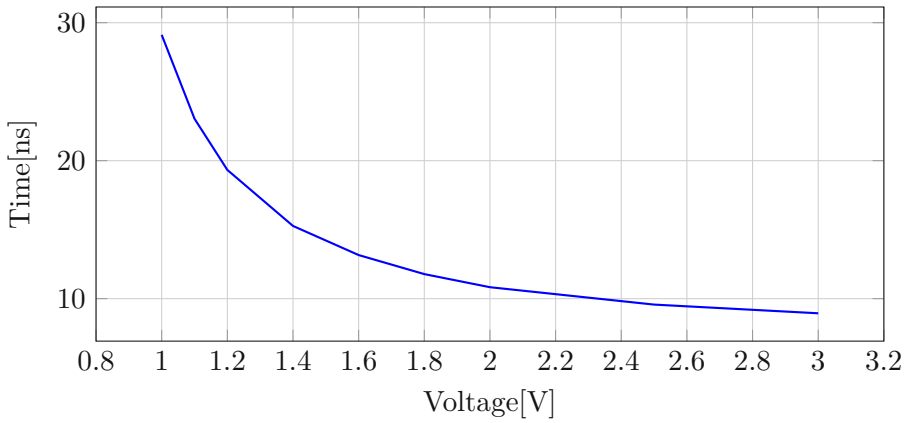

Figure 3.8: *Influence of the voltage V_{DT} on the quench duration*

Figure 3.7 shows, based on the simulation results, the influence of the voltage on the quench duration and thus on the deadtime. The voltage range shown does not represent the limits of the usable voltage, but represents a good approximation for a large part of the measurements that can be expected. The non-linear dependence of the quench duration on the set voltage V_{DT} is evident, as voltage differences in the lower range (comparison 1.0 V to 1.1 V) represent a significant change in the quench duration (approx. 10 ns).

3.2.4 Time Difference of the Pixels

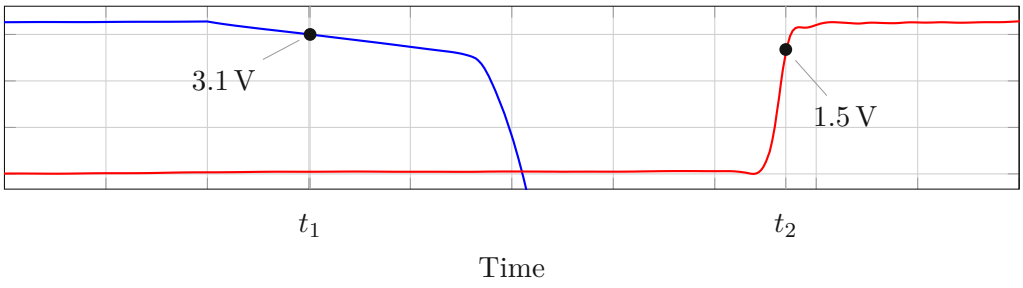

Figure 3.9: *Description of the measuring points in time and voltage domains. The blue curve shows the voltage at a cathode and the red one shows the voltage of the chip output signal of the associated trigger output under load with a* 120 Ω *resistor. Note: The vertical resolution of the two signals is different.*

As already mentioned in 2.6.2, it is essential that the delay between a detection event of a pixel and the output of a corresponding display or a noticeable signal change is as small as possible in the first place. More important than minimizing this delay is the constancy of the delay. Due to the relatively large dimensions (the outermost pixels are more than 2 mm apart from each other), the delay differences are the dominating factors.

In order to be able to evaluate the time difference and, above all, the variation in the time difference of the delay from the occurrence of a breakdown to the presence of the corresponding pulse, the following metric was selected. The time difference t_1-t_2 shown in 3.9 represents the value that is ought to be compared. The time t_1 represents the time at which any cathode (blue) reaches the voltage level of 3.1 V in the course of a detection event. When selecting the level, care was taken to ensure that, on the one hand, the reference voltage level is at the same level and, on the other hand, the circuit is still in the short phase of passive quenching. The time t_2 is the time at which an arbitrarily defined voltage level (in this case 1.5 V) of the respective associated (even or odd) trigger (red) signal was reached.

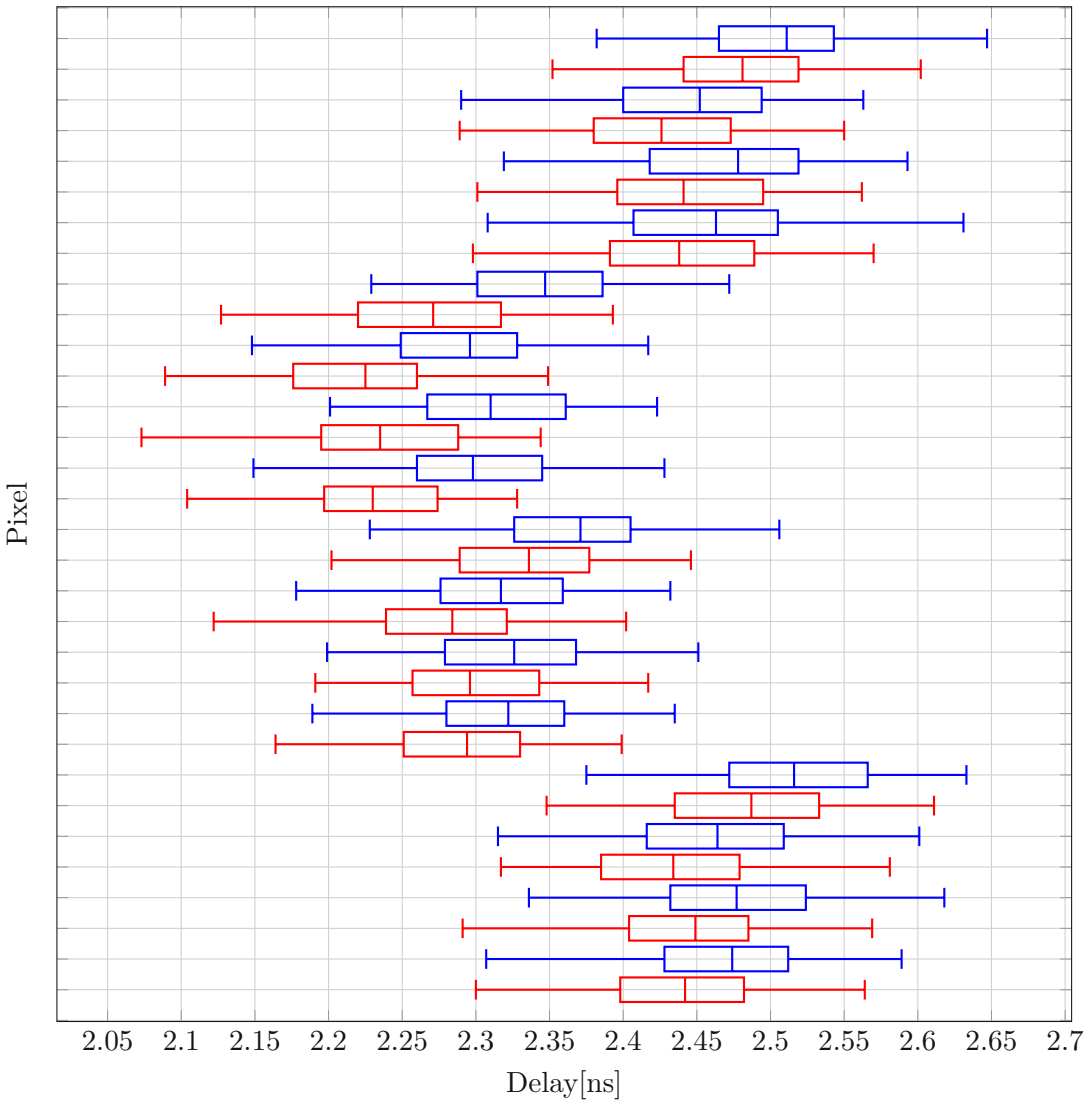

Figure 3.10: *Variation of the signal propagation times of the trigger signals in the case of a detection event at the respective pixel from the time of the voltage drop at the cathode until a defined voltage level is reached at the detection output of the chip.*

Based on the principle of operation, this output trigger signal is the result of cascading adjuncts. Four pixels each are combined (OR operation). Therefore, based on the given number of pixel elements per parity, there are four such ORs. In the layout of these sub-trigger signals, care was taken to ensure that the relevant components are positioned centrally between the individual pixels. These sub-signals are again combined with each other by means of OR gates to obtain the actual output trigger signal. This behavior can be seen very well in Fig. 3.10, which shows the simulation results of a Monte Carlo

simulation with 100 runs and the corresponding variation in the form of a box plot. Each of the 4 pixels of a parity (recognizable by the color coding red - even, blue - odd) show a very similar behavior in the form of similar delay times. It can also be clearly seen that the evaluation circuit and the power driver required for the output are located in the vertical center of the chip. The delays are less for all pixels located in the center than for detection elements located on the edge. However, it should also be noted that the maximum difference in the delay shown in 3.9 between the fastest pixel-logic-output driver combination and the slowest combination of one and the same simulation run (simulation process with 100 runs) has a value of 388 ps.

3.2.5 Power Consumption

Figures 3.12 and 3.11 show the temporal course of the input currents on the supply voltage levels mentioned as a function of time. For comparison the voltage curve of the cathode can be seen in dotted lines in the two curves. This is used to assign the current event to the two corresponding ones.

The first drop (in the range from 13 ns to approx. 26 ns) of the cathode voltage represents a temporary deactivation of the pixel caused by $\overline{GATER}$ with its associated quenching. The second pulse (60 ns to approx. 65 ns) represents a detection event.

The high current peak on the falling potential edge, which signals a quenching process (deactivation or breakdown event), is linked to the rapid reloading of the capacities and the associated rapid charge transport.

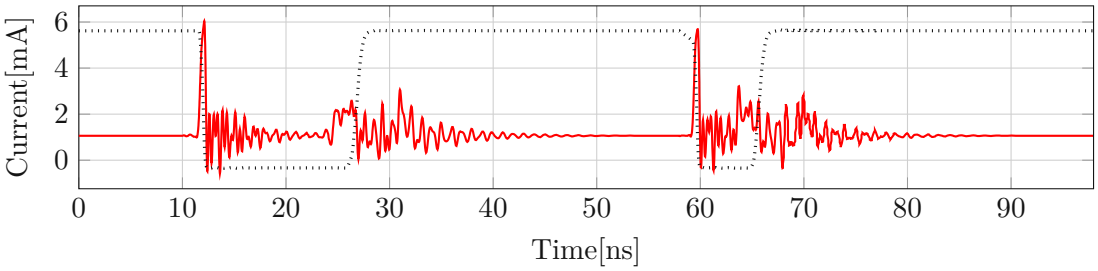

Figure 3.11: *Current consumption of one single pixel for supply voltage V_{SUP2}*

The power requirement of a pixel can be represented in a fairly good approximation to reality as the sum of the multiplication of the supply voltages (V_{SUP1} 3.3 V and V_{SUP2} −6.6 V) with the respective currents (figures 3.12 and 3.11).

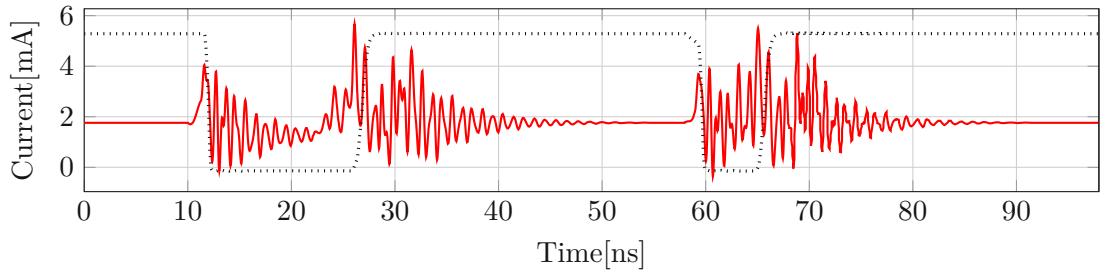

Figure 3.12: *Current consumption of one single pixel for supply voltage V_{SUP1}*

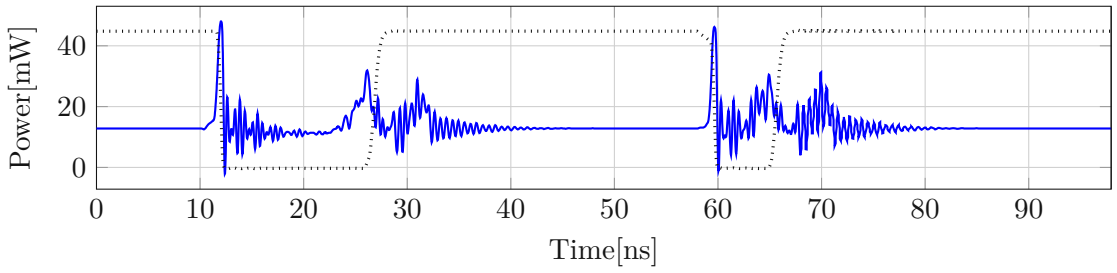

Figure 3.13: *Power consumption single pixel*

The power requirement is shown in Fig. 3.13. As depicted in that figure, the strong peak in the course of the potential transition from approximately V_{SUP1} to V_{SUP2} can be recognized in a few nanoseconds. This value is similar to the one in the original circuit (2.4), which was to be expected due to the extensive adoption of topology.

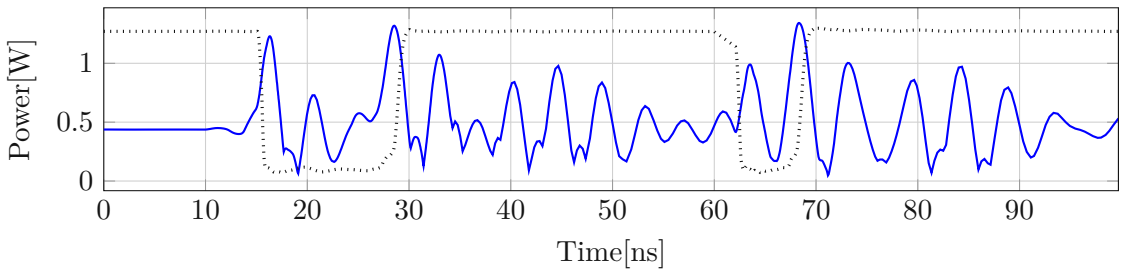

Figure 3.14: *Power consumption of the chip*

The maximum power requirement arises in the case of gate operation or especially when all 32 elements, i.e. both parities, are gated at the same time. The solid line in 3.14 shows the time course of the consumed power. The aforementioned externally triggered temporary quench pulse in the period 15 ns to 30 ns and the simulated detection event (from approx. 60 ns) of all 32 pixels show the maximum power requirement of approx. 1.8 W. The entire

chip was considered here, which, in addition to the 32 detection elements, also consists of the additional blocks mentioned, such as: clock driver, logic and output driver.

Chapter 4

Summary and Outlook

The chip presented in this thesis, which was designed in the *XO035* process by the manufacturer *XFAB* [33] and whose behavior was analyzed in detail and documented here in the course of the design process, is a line sensor with 32 pixels positioned at equal intervals, which makes it possible to determine the location of individual photons in a spatial dimension. A similarly high photon detection probability for the individual pixels can be expected by starting the design on an already given SPAD quencher [9], which achieves a photon detection probability of up to 53.1 % at a wavelength of 657 nm in a single configuration. Adjustments to this original design, which were necessary to increase stability, usability and connection to the pixel-internal digital interface, should not deteriorate the photon detection probability, at least not significantly. The circuit can be controlled by external analog reference voltages to e.g. adjust dead time and pulse duration of reset. Further requirements for the chip, which would result in diametrically opposed solutions when considered individually, were brought to what the author believes to be the best possible combination, taking into account individual weightings and the given boundary conditions. The requirement to signal a detection event, i.e. a photon hitting a pixel, as quickly as possible to the downstream off-chip circuitry is opposed to the requirement that the amount of pads to be bonded should be as small as possible with easy expandability of the pixel line. This discrepancy, and also the desire expressed during development to permanently deactivate individual pixels (in the event that individual pixels would make it impossible to use the sensor for various reasons) or temporarily (gater operation depending on the measurement) were to be deactivated to the decision to extend the largely analog-dominated pixel to a connection in digital form and to connect it to a suitable digital I/O system as a sensor chip backbone. Some detailed requirements for the digital detection system, such as a validation signal that provides information

about whether one and exactly one pixel was activated or the minimization of the signal propagation times in the case of detection processes at the various pixels, to name just two, posed further obstacles in the design. The layout, details can be seen in Chapter 3, which shows a sensor in an almost square shape, represents the result of this work. The 32 pixels, which were divided according to parity, are from two sides, left and right in the above view the digital interface mentioned in Chapter 2 can be contacted. Voltages whose function is not primarily used to supply energy, i.e. analog parameter values that specify the behavior of individual functions (reset, quench and hold off times, etc. in the pixel), are added via dedicated pads.

A possible starting point for further developments would also be mentioned here: Extensive post-layout simulations have shown that adapting the analog voltage feeds or replacing this analog interface with a bus system to be defined can significantly improve performance and consistency across the entire chip . This was not implemented due to time constraints, since a significant part of the work was already spent on improving the performance of the basic design by using bipolar transistors. Although this improvement was successful, since the disadvantages in the sensor application mentioned outweigh the advantages, it was not implemented.

List of Figures

List of Abbreviations

Symbol	Description
CMOS	Complementary Metal-Oxide-Semiconductor
APP	After Pulsing Probability
APD	Avalanche Photodiode
BiCMOS	Bipolar Complementary Metal Oxide Semiconductor
BJT	Bipolar Junction Transistor
CLK	Clock
CCL	Common Centroid Layout
CMOS	Complementary Metal Oxide Semiconductor
DCR	Dark Count Rate
EMR	Electromagnetic Radiation
IC	Integrated Circuit
MOSFET	Metal Oxide Semiconductor Field Effect Transistor
PMT	Photomultiplier
PDP	Photon Detection Probability
PCB	Printed Circuit Board
RF	Radio Frequency
SPAD	Single Photon Avalanche Diode
SNSPD	Superconducting Nanowire Single Photon Detectors
TVQC	Triple Voltage Quencher Circuit

Bibliography

[1] A. Katz, T. Blank, A. Fenigstein, T. Leitner, and Y. Nemirovsky. "Active-Reset for the N+P Single-Ended SPAD Used in the NIR LiDAR Receivers". In: *IEEE Transactions on Electron Devices* 66.12 (2019), pp. 5191–5195. DOI: 10.1109/TED.2019.2946015.

[2] E. Charbon. "Single-photon imaging in complementary metal oxide semiconductor processes". eng. In: *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences* 372.2012 (Feb. 2014), pp. 20130100–20130100. ISSN: 1364-503X. DOI: 10.1098/rsta.2013.0100.

[3] S. Cova, M. Ghioni, F. Zappa, A. Tosi, I. Rech, A. Gulinatti, and S. Tisa. "Single-photon avalanche detectors for quantum communications". In: *2010 Conference on Optical Fiber Communication (OFC/NFOEC), collocated National Fiber Optic Engineers Conference*. 2010, pp. 1–3. DOI: 10.1364/OFC.2010.OTuC2.

[4] S. Pellegrini, B. Rae, A. Pingault, D. Golanski, S. Jouan, C. Lapeyre, and B. Mamdy. "Industrialised SPAD in 40 nm technology". In: *2017 IEEE International Electron Devices Meeting (IEDM)*. 2017, pp. 16.5.1–16.5.4. DOI: 10.1109/IEDM.2017.8268404.

[5] Z. Ma, Z. Wu, and Y. Xu. "Compact SPAD pixels with fast and accurate photon counting in the analog domain". In: *Journal of Semiconductors* 42.5 (May 2021), p. 052402. DOI: 10.1088/1674-4926/42/5/052402.

[6] F. Guerrieri, S. Tisa, A. Tosi, and F. Zappa. "Two-Dimensional SPAD Imaging Camera for Photon Counting". In: *IEEE Photonics Journal* 2.5 (2010), pp. 759–774. DOI: 10.1109/JPHOT.2010.2066554.

[7] E. Charbon, C. Bruschini, and M.-J. Lee. "3D-Stacked CMOS SPAD Image Sensors: Technology and Applications". In: *2018 25th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*. 2018, pp. 1–4. DOI: 10.1109/ICECS.2018.8617983.

[8] S. Lindner, S. Pellegrini, Y. Henrion, B. Rae, M. Wolf, and E. Charbon. "A High-PDE, Backside-Illuminated SPAD in 65/40-nm 3D IC CMOS Pixel With Cascoded Passive Quenching and Active Recharge". In: *IEEE Electron Device Letters* 38.11 (2017), pp. 1547–1550. DOI: 10.1109/LED.2017.2755989.

[9] A. Dervic, M. Hofbauer, B. Goll, and H. Zimmermann. "Integrated Fast-Sensing Triple-Voltage SPAD Quenching/Resetting Circuit for Increasing PDP". In: *IEEE Photonics Technology Letters* 33.3 (2021), pp. 139–142. DOI: 10.1109/LPT.2020.3044484.

[10] H. Hertz. "Ueber einen Einfluss des ultravioletten Lichtes auf die electrische Entladung". In: *Annalen der Physik* 267.8 (1887), pp. 983–1000. DOI: https://doi.org/10.1002/andp.18872670827.

[11] A. Einstein. "Ueber einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt". In: *Annalen der Physik* 322.6 (1905), pp. 132–148. DOI: https://doi.org/10.1002/andp.19053220607.

[12] V. K. Zworykin, G. A. Morton, and L. Malter. "The Secondary Emission Multiplier-A New Electronic Device". In: *Proceedings of the Institute of Radio Engineers* 24.3 (1936), pp. 351–375. DOI: 10.1109/JRPROC.1936.226435.

[13] *Hamamatsu Photonics K.K Product Overview - Photomultiplier tubes (PMTs).* https://www.hamamatsu.com/eu/en/product/optical-sensors/pmt/index.html. Accessed: 2021-03-27.

[14] L. R. Testardi. "Destruction of Superconductivity by Laser Light". In: *Phys. Rev. B* 4 (7 Oct. 1971), pp. 2189–2196. DOI: 10.1103/PhysRevB.4.2189.

[15] G. N. Goltsman, O. Okunev, G. Chulkova, A. Lipatov, A. Semenov, K. Smirnov, B. Voronov, A. Dzardanov, C. Williams, and R. Sobolewski. "Picosecond superconducting single-photon optical detector". In: *Applied Physics Letters* 79.6 (2001), pp. 705–707. DOI: 10.1063/1.1388868.

[16] W. Zhang, L. You, H. Li, J. Huang, C. Lv, L. Zhang, X. Liu, J. Wu, Z. Wang, and X. Xie. "NbN superconducting nanowire single photon detector with efficiency over 90% at 1550 nm wavelength operational at compact cryocooler temperature". In: *Science China Physics, Mechanics & Astronomy* 60.12 (Oct. 2017), p. 120314. ISSN: 1869-1927. DOI: 10.1007/s11433-017-9113-4.

[17] D. V. Reddy, R. R. Nerem, S. W. Nam, R. P. Mirin, and V. B. Verma. "Superconducting nanowire single-photon detectors with 98% system detection efficiency at 1550 nm". In: *Optica* 7.12 (Dec. 2020), pp. 1649–1653. DOI: 10.1364/OPTICA.400751.

[18] A. Gallivanoni, I. Rech, and M. Ghioni. "Progress in Quenching Circuits for Single Photon Avalanche Diodes". In: *IEEE Transactions on Nuclear Science* 57.6 (2010), pp. 3815–3826. DOI: 10.1109/TNS.2010.2074213.

[19] W. Shockley and W. T. Read. "Statistics of the Recombinations of Holes and Electrons". In: *Phys. Rev.* 87 (5 Sept. 1952), pp. 835–842. DOI: 10.1103/PhysRev.87.835.

[20] M. Hofbauer, B. Steindl, and H. Zimmermann. "Temperature Dependence of Dark Count Rate and After Pulsing of a Single-Photon Avalanche Diode with an Integrated Active Quenching Circuit in 0.35 μm CMOS". In: *Journal of Sensors* 2018 (July 2018), p. 9585931. ISSN: 1687-725X. DOI: 10.1155/2018/9585931.

[21] S. Cova, M. Ghioni, A. Lotito, I. Rech, and F. Zappa. "Evolution and prospects for single-photon avalanche diodes and quenching circuits". In: *Journal of Modern Optics* 51.9-10 (2004), pp. 1267–1288. DOI: 10.1080/09500340408235272.

[22] S. Ganichev, E. Ziemann, I. Yassievich, W. Prettl, and A. Istratov. "Distinction Between the Poole-Frenkel and Tunneling Models of Electric-Field-Stimulated Carrier Emission From Deep Levels in Semiconductors". In: *Physical Review B* 61 (Apr. 2000). DOI: 10.1103/PhysRevB.61.10361.

[23] S. Cova, A. Lacaita, and G. Ripamonti. "Trapping phenomena in avalanche photodiodes on nanosecond scale". In: *IEEE Electron Device Letters* 12.12 (1991), pp. 685–687. DOI: 10.1109/55.116955.

[24] S. Cova, A. Longoni, and G. Ripamonti. "Active-Quenching and Gating Circuits for Single-Photon Avalanche Diodes (SPADs)". In: *IEEE Transactions on Nuclear Science* 29.1 (1982), pp. 599–601. DOI: 10.1109/TNS.1982.4335917.

[25] R. McIntyre. "The distribution of gains in uniformly multiplying avalanche photodiodes: Theory". In: *IEEE Transactions on Electron Devices* 19.6 (1972), pp. 703–713. DOI: 10.1109/T-ED.1972.17485.

[26] P. Webb, R. McIntyre, and J. Conradi. "Properties of avalanche photodiodes". In: *RCA Review* 35 (1974), pp. 234–278.

[27] R. McIntyre. "Recent developments in silicon avalanche photodiodes". In: *Measurement* 3.4 (Oct. 1985), pp. 146–152. DOI: 10.1016/0263-2241(85)90024-7.

[28] *Silicon Geiger Mode Avalanche Photodiode*. lc-apd&sap-series. V12. Laser Components GmbH. Jan. 2020.

[29] A. Goetzberger, B. McDonald, R. H. Haitz, and R. M. Scarlett. "Avalanche Effects in Silicon pn Junctions. II. Structurally Perfect Junctions". In: *Journal of Applied Physics* 34.6 (1963), pp. 1591–1600. DOI: 10.1063/1.1702640.

[30] R. H. Haitz. "Mechanisms Contributing to the Noise Pulse Rate of Avalanche Diodes". In: *Journal of Applied Physics* 36.10 (1965), pp. 3123–3131. DOI: 10.1063/1.1702936.

[31] M. Ghioni, S. Cova, A. Lacaita, and G. Ripamonti. "New silicon epitaxial avalanche diode for single-photon timing at room temperature". English. In: *Electronics Letters* 24 (24 Nov. 1988), 1476–1477(1). ISSN: 0013-5194.

[32] S. Cova, A. Lacaita, M. Ghioni, G. Ripamonti, and T. A. Louis. "20ps timing resolution with single photon avalanche diodes". In: *Review of Scientific Instruments* 60.6 (1989), pp. 1104–1110. DOI: 10.1063/1.1140324.

[33] *Process & Device Specification XO035*. Release 2.3.1. Technical Specification. X-FAB Semiconductor Foundries, Aug. 2019.

[34] W. Gaberl, B. Steindl, K. Schneider-Hornstein, R. Enne, and H. Zimmermann. "0.35 μm CMOS avalanche photodiode with high responsivity and responsivity-bandwidth product". In: *Opt. Lett.* 39.3 (Feb. 2014), pp. 586–589. DOI: 10.1364/OL.39.000586.

[35] R. Enne, B. Steindl, M. Hofbauer, and H. Zimmermann. "Fast Cascoded Quenching Circuit for Decreasing Afterpulsing Effects in 0.35-μm CMOS". In: *IEEE Solid-State Circuits Letters* 1.3 (2018), pp. 62–65. DOI: 10.1109/LSSC.2018.2827881.

[36] D. Milovancev, T. Jukic, B. Steindl, M. Hofbauer, R. Enne, K. Schneider-Hornstein, and H. Zimmermann. "Optical wireless communication with monolithic avalanche photodiode receivers". In: *2017 IEEE Photonics Conference (IPC)*. 2017, pp. 25–26. DOI: 10.1109/IPCon.2017.8115989.

[37] N. Mitrovic, R. Enne, and H. Zimmermann. "A Bootstrap circuit for DC-DC converters with a wide input voltage range in HV-CMOS". In: *2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. 2016, pp. 62–65. DOI: 10.1109/MIPRO.2016.7522111.