Applied Quantum Architectures
Mekelweg 4,
2628 CD Delft
The Netherlands
http://ens.ewi.tudelft.nl/

AQUA-2016-08

M.Sc. Thesis

High Resolution, Fully Digital Photon-Counting Image Sensors in DSM CMOS Technologies

Arin Can Ulku B.Sc.

Abstract

Single-Photon Avalanche Diodes (SPAD) have gradually become the top choice for time-resolved imaging applications thanks to their high timing resolution and single-photon sensitivity. However, a variety of factors complicate the implementation of SPAD sensors with large pixel arrays that achieve comparable specifications with competing technologies. The major issues that must be addressed to increase the scalability of SPAD sensors include fill factor, pixel array uniformity and power consumption. In addition, the integration of SPADs into deep sub-micron CMOS process technologies introduces its own challenges such as the lack of high voltage support and dead spaces that restrict pixel miniaturization.

In this thesis, a time-gated, fully digital pixel with an in-pixel memory was presented. The pixel functionality and basic parameters were tested in a 110 nm 4×4 array. In addition, the scalability of this architecture was demonstrated by designing a 512×512 sensor in 0.18 μm technology. Several performance boosting techniques were implemented in different variants of each chip. The impact of different SPAD structures on overall sensor performance was investigated. Finally, a 512×1 linear sensor with maximum 12 V excess bias was designed to operate the SPAD with high photon sensitivity.

TU Delft

Faculty of Electrical Engineering, Mathematics and Computer Science Delft University of Technology

High Resolution, Fully Digital Photon-Counting Image Sensors in DSM CMOS Technologies

Thesis

submitted in partial fulfillment of the
requirements for the degree of

Master of Science

in

Electrical Engineering

by

Arin Can Ulku B.Sc.
born in Adana, Turkey

This work was performed in:

Applied Quantum Architectures Group
Department of Quantum Engineering
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology

Delft University of Technology

Delft University of Technology
Department of
Quantum Engineering

The undersigned hereby certify that they have read and recommend to the Faculty of Electrical Engineering, Mathematics and Computer Science for acceptance a thesis entitled **"High Resolution, Fully Digital Photon-Counting Image Sensors in DSM CMOS Technologies"** by **Arin Can Ulku B.Sc.** in partial fulfillment of the requirements for the degree of **Master of Science**.

Dated: 31.08.2016

Chairman: _____
prof.dr. Edoardo Charbon

Advisor: _____
prof.dr. Edoardo Charbon

Committee Members: _____
dr. Daniele Cavallo

_____
dr. Carl Jackson

Abstract

Single-Photon Avalanche Diodes (SPAD) have gradually become the top choice for time-resolved imaging applications thanks to their high timing resolution and single-photon sensitivity. However, a variety of factors complicate the implementation of SPAD sensors with large pixel arrays that achieve comparable specifications with competing technologies. The major issues that must be addressed to increase the scalability of SPAD sensors include fill factor, pixel array uniformity and power consumption. In addition, the integration of SPADs into deep sub-micron CMOS process technologies introduces its own challenges such as the lack of high voltage support and dead spaces that restrict pixel miniaturization.

In this thesis, a time-gated, fully digital pixel with an in-pixel memory was presented. The pixel functionality and basic parameters were tested in a 110 nm 4×4 array. In addition, the scalability of this architecture was demonstrated by designing a 512×512 sensor in 0.18 μm technology. Several performance boosting techniques were implemented in different variants of each chip. The impact of different SPAD structures on overall sensor performance was investigated. Finally, a 512×1 linear sensor with maximum 12 V excess bias was designed to operate the SPAD with high photon sensitivity.

Acknowledgments

I would like to take this chance to express my gratitude to all people who helped me during this journey. Firstly, I wish to thank my thesis advisor, Prof. dr. Edoardo Charbon for providing me the opportunity to work in one of his projects. I am grateful to have worked with a professor who continuously motivates his students to achieve excellence. I would also like to thank Dr. Myung-Jae Lee and Ivan Michel Antolovic for supervising me during my thesis project.

My sincere thanks also goes to Dr. Carl Jackson for his time and guidance in my progress reviews.

This work would not have been possible without the invaluable help of Dr. Chockalingam Veerappan, Chao Zhang, Augusto Carimatto, Esteban Venialgo, Scott Lindner, Siddharth Sinha and Ting Gong. They never hesitated to share their knowledge and experiences when I needed. I would like to thank each of them for their selfless help.

I extend my special thanks to Antoon Frehe for his continuous technical support throughout the entire project, and offering me help in critical times.

I would like to recognize all of my colleagues in this research group: Bahador Valizadehpasha, Rosario Incandela, Bishnu Patra, Preethi Padmanabhan, Harald Homulle, Jeroen van Dijk, Pengfei Sun, Lin Song, Junjie Weng. I would like to thank them for sharing this great academic environment.

Finally, I would like to thank my parents for their unconditional love and support.

Arin Can Ulku B.Sc.
Delft, The Netherlands
31.08.2016

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

SPAD-based image sensors are important candidates to replace the existing technologies for single-photon, time-resolved imaging. They can achieve high time resolution and low noise in typical CMOS operating voltages. In addition, monolithic integration of SPADs with CMOS electronics allow the design of compact image sensors without compromise of functionality. Nonetheless, important milestones for the commercialization of SPAD sensors are still to be achieved. For instance, SPAD sensors with large pixel sizes were built only recently [4, 5]. These large-array sensors suffer from various problems related to scaling. Furthermore, it is a challenging task to combine high sensitivity, low noise, high uniformity and low power consumption using 2D CMOS integration techniques.

1.2 Background

This section provides background information related to single-photon time-resolved imaging. It is worth noting that even though several technologies that shaped the evolution of single-photon detection were discussed, this thesis mainly focuses on SPAD-based time-resolved imaging.

Single-photon counting image sensors are used in various applications such as positron emission tomography (PET), single-photon emission computed tomography (SPECT), fluorescence-lifetime imaging microscopy (FLIM), Förster resonance energy transfer (FRET), or fluorescence correlation spectroscopy (FCS) [6].

From the 1930s to the present day, different types of single-photon detectors have dominated the market, and were used for the aforementioned applications. Each newly invented single-photon detector type had relative advantages and disadvantages compared to the previous ones. In general, the trend is towards compact and reliable devices that can be integrated to CMOS processes. It is still a challenging task to build such devices without compromising the key figures of merit.

1.2.1 Single-Photon Counting Applications

- **Fluorescence Lifetime Imaging Microscopy (FLIM)**

The use of single-photon detectors in time-correlated single-photon counting (TCSPC) applications such as FLIM has become common over the last decades. Fluorescence is defined as the emission of photons from a material after absorbing light from another source. Fluorescence characteristics of a microscopic organism allows the detection of its various qualities in a non-invasive way [7]. Single-photon detectors are used to measure the fluorescent lifetime of a biological sample using TCSPC. This

technique is based on sending a repetitive laser signal to the target and recording its statistical distribution of response times in a histogram [8]. The crucial specifications of a single-photon detector device for FLIM are high sensitivity, low dark counts and high temporal resolution [9].

- **Time-of-Flight (ToF) 3D Imaging**

Recently, 3D imaging based on photon time-of-flight (ToF) detection has become a popular research field due to the demand from a wide range of commercial applications, including surveillance, automotive and robotics [10]. 3D range finding using photon detectors can be classified into two categories: direct and indirect ToF measurement [11]. In direct measurement, the distance between the sensor and the target is derived from the time delay between the emitted laser pulse and the reflected light. In the indirect measurement, the distance is extracted from the phase difference of a continuous sinusoidal wave, caused by the reflection from the target. In both methods, the spatial resolution of the output image is strongly dependent on the temporal resolution of the photon detector.

- **Positron Emission Tomography (PET)**

Positron emission tomography (PET) is a nuclear imaging technique where a series of photon detectors retrieve metabolic information about the human body [1]. They are used in a range of medical applications such as oncological diagnosis and brain functional analysis [12]. A PET system works based on the following principle [13]: firstly, a radioactive tracer is injected into the human body. As it moves inside the body, this tracer continuously emits a positron, which travels to a distance typically within a radius of 1 mm before annihilating. The tracer is usually a type of sugar and it is absorbed by a cancerous area at a higher rate, thus positron emission and annihilation tend to be localized in that area. Upon annihilation, two gamma photons with the same energy of 511 keV are emitted at $\sim$ 180 degrees towards opposite directions. A ring-shaped array of detectors placed around the body derives the location of the photon generation from the ToA of both gamma photons. The first PET detectors were formed by photomultiplier tubes (PMTs). Nowadays, PMTs are replaced by silicon photomultipliers (SiPMs) [14]. Aside from the generic standards, the detector used for PET applications must be robust to magnetic fields (so long as it needs to operate in an MRI), and gamma radiation.

- **Raman Spectroscopy**

Raman spectroscopy is an analysis that identifies structural and compositional characteristics of materials from the scattering of monochromatic light from their surfaces. Its applications include material science, archaeology, medical imaging and planetary science [15]. Resulting from the similarity of its measurement technique with FLIM, a major challenge of Raman spectroscopy is to filter out the fluorescence-related detections from the incoming photons. To this day, fluorescence background suppression in Raman spectroscopy is an active area of research. Time-resolved Raman spectroscopy, demonstrated in the work [16], is a technique that rejects the fluorescence background

by taking advantage of the separate response times of the two types of emissions. To achieve a detailed characterization of the material and to eliminate the fluorescence background, high timing resolution is essential for a photon detector in this application.

- **True Random Number Generation (TRNG)**

Another application of single-photon detectors is true random number generation (TRNG). There are two ways of generating random numbers. Pseudo random numbers are created using a variety of computational algorithms. Even though the distribution of the outputs achieves a high degree of randomness that meets the standards of many applications, the source of the numbers is deterministic; hence the sequence of the numbers is repeatable. To generate true random numbers, the source must be a physical mechanism that displays a quantum nature [17]. The work [18] demonstrates the use of a general-purpose SPAD-based single-photon detector as a TRNG. In this method, the randomness of the number generation results from the quantum nature of photon absorption in two or more identical detectors. A LED source was used as a photon generator, whose duration and the wavelength of were chosen to set sensitivity.

1.2.2 Single-Photon Detectors

- **Photomultiplier Tubes (PMT)**

A photomultiplier tube (PMT) generates relatively large currents by multiplying a photogenerated electron in a tube. Its structure is shown in Figure 1.1 [1]. The photo cathode, typically biased to hundreds of volts, generates a photoelectron upon photon arrival. This free electron impinges on several metal plates called dynodes, while travelling across the tube. After being hit by a photoelectron, each dynode multiplies the electrons; thus amplifies the current. Thanks to the geometrical placement of dynodes, this multiplication event is repeated several times before the electrons reach the anode. The magnitude of the current depends on the cathode bias voltage and the number of dynodes (number of amplification stages).

PMTs were the most popular choice of photodetectors for many decades, since being invented in 1930s [19]. Their major advantages are high timing resolution and high photon sensitivity. The last generation of PMTs achieves quantum efficiency (QE) of 32-36 % as opposed to less than 1 % QE in early devices [20]. However, PMTs have certain drawbacks in integration with CMOS electronics, especially in large-scale arrays with small pixel sizes. Furthermore, high power consumption due to high voltage operation (several hundred volts to several thousand volts) and sensitivity to magnetic fields are some factors that complicate the use of PMTs in certain applications.

- **Electron-Multiplying Charge Coupled Devices (EMCCD)**

Charge coupled device (CCD) is a prevalent solid-state imaging technology, introduced in 1969 [21]. These devices work by transporting the accumulated charge through a shift register formed by MOS capacitors. They are widely used in intensity imaging due to their high QE and scalability. On the other hand, CCDs suffer from slow

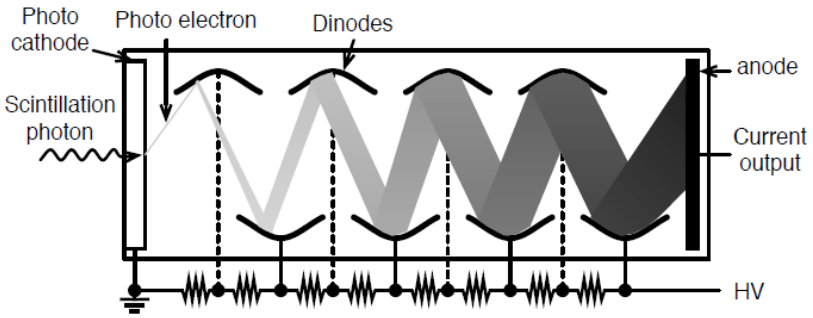


Figure 1.1: PMT structure [1]

readout due to a charge amplification stage at the end of the shift register and the inherent sequential operation. In early 2000s, a charge multiplication mechanism based on impact-ionization has been incorporated into the conventional CCD architecture, in order to achieve single-photon sensitivity [22]. These devices, called EMCCDs, combine the noiseless characteristic of electron multiplication mechanism with high photon sensitivity of the CCD technology. In addition, they eliminate the slow readout problem by amplifying charge before the readout.

EMCCD is the first single-photon detecting solid-state device. Its strongest advantage against PMT is the lack of image intensifier tube. As discussed earlier, these tubes are difficult to integrate into deep sub-micron CMOS image sensors, due to their sizes, high voltage operation and sensitivity to magnetic fields. EMCCD offers high photon sensitivity; its QE can exceed 90 % depending on photon wavelength [23]. However, it has a major drawback that restricts its compatibility with many applications. EMCCDs cannot work with fast gating mechanisms. Therefore, they cannot achieve timing resolution in the order of nanoseconds.

- **CMOS Active Pixel Sensors (CMOS APS)**

Starting from 1990s, CMOS image sensors have become popular in various commercial imaging applications, particularly in consumer electronics. CMOS APS are typically formed by p-n junctions operating in moderate reverse bias. CMOS APS devices have historically been synonymous with inexpensive sensors for low-demanding applications. They have never been widely adopted by scientific imaging applications, due to their high noise levels and pixel array non-uniformity [24]. Nevertheless, novel techniques are being developed to eliminate the major drawbacks of CMOS APS and to make them suitable for demanding scientific applications. Scientific CMOS (sCMOS) sensors, developed by Andor [24], achieve low noise and high dynamic range comparable to CCD technology. On the other hand, their photon sensitivity is still significantly below CCD devices, with a QE of only 60 %.

- **Proportional-Mode Avalanche Photodiodes**

An avalanche photodiode (APD) is a p-n or p-i-n junction that is reverse biased at or around its breakdown voltage. Its operation differs from pinned photodiodes of APS

sensors by exploiting a phenomenon called avalanche multiplication [25]. When a photon impinges on the junction of an APD, the photon-generated carriers are multiplied in the high electric field of its depletion region. This current amplification mechanism allows the detection of low levels of light. When the device is biased slightly below breakdown voltage, the current intensity increases linearly with photon count. Therefore, these devices are called proportional-mode APDs. When APDs are biased above breakdown, the avalanche mechanism continues until the device is damaged or driven below breakdown by quenching or other external factors. This mode is called Geiger-mode, and these APDs are called Geiger-mode APDs, or SPADs. Two major drawbacks of proportional-mode APDs are poor timing accuracy and multiplication noise.

- **Superconducting Single-Photon Detectors (SSPD)**

Some works in the literature explore different physical phenomena to build single-photon detectors. Most of these approaches that can potentially introduce fundamental changes to the imaging sensor field are in the development stage, and far behind reaching commercial standards.

Superconducting single-photon detectors exploit a physical property of superconducting wires [26]. When they generate a free carrier upon photon arrival, they switch to an insulating state for a brief period, typically within tens of picoseconds [27]. While this property allows detection of single-photons with the help of suitable electronic circuitry, the environmental conditions have to be right for superconductivity. Most suitable materials display superconductive characteristics at critical temperatures significantly below room temperature (4.2 K in [27]). Furthermore, having a quantum efficiency of only 20 %, their sensitivity is lower than the currently used technologies. Another important drawback of SSPDs is their incompatibility with large arrays, due to the difficulty of maintaining a low temperature for readout [28]. Consequently, further performance improvements are needed for the commercialization of SSPD.

- **Single-Photon Avalanche Diode (SPAD)**

Single-photon avalanche diode (SPAD) is a form of an APD that works in reverse bias above breakdown. This type of operation is called Geiger mode. When a photon impinges on an APD, it generates a carrier in the main junction depletion region through impact ionization. Contrary to conventional APDs, in SPADs the first photocarrier quickly multiplies inside the high electric field region and creates a very large current, whose magnitude can reach several milliamperes [29]. This event, exclusive to Geiger mode, is called avalanche breakdown. Since the current of a SPAD is not dependent on photon count, each avalanche event marks the count of a single photon.

SPADs are presented as potential alternatives to currently dominant single-photon counters. They offer very high temporal resolution and single-photon sensitivity [6]. Furthermore, they are resistant to magnetic fields, suitable for CMOS integration, and can operate in significantly lower voltages than PMTs. On the other hand, SPADs still suffer from low photon sensitivity and high noise. Also, due to the lack of large SPAD pixel arrays, their non-uniformity characteristics were not studied comprehensively.

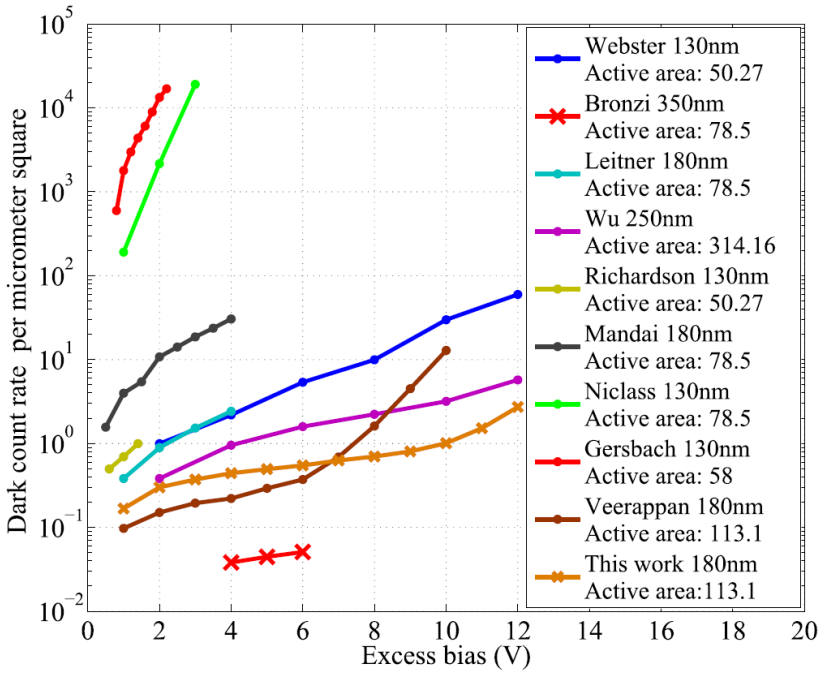


Figure 1.2: Dark Count Rate (DCR) per micrometer square of various SPAD structures in the literature [2]

SPADs have evolved over the years into devices compatible with single-photon applications. The first SPADs were built in custom processes; therefore their integration into CMOS electronics was challenging. Along with a gradual improvement in timing resolution to levels that can compete with PMTs, an important milestone in SPAD technology was its integration in CMOS processes [30]. This development eventually led to compact monolithic SPAD sensors in deep sub-micron technologies with large pixel arrays. To date, SPADs have been designed in technology nodes as small as 65 nm [31].

All chips presented in this thesis employ SPADs for single-photon detection. The structure and performance parameter of each device is discussed in the related chapters.

1.2.3 SPAD Figures of Merit

- **Dark Count Rate (DCR)**

The major source of noise in SPADs is current generation in the absence of photons, which is called dark noise. The two main causes of dark noise are trap-assisted thermal generation noise and band-to-band tunneling noise [32, 33]. Traps are defects in the silicon lattice that hold carrier charges during an avalanche. Trap-assisted noise generation strongly depends on temperature and process characteristics. Tunneling is a phenomenon in quantum mechanics, which implies that a particle has a probability to transcend a potential barrier without sufficient energy that classical mechanics principles require [34]. Tunneling-assisted noise does not have a strong dependence

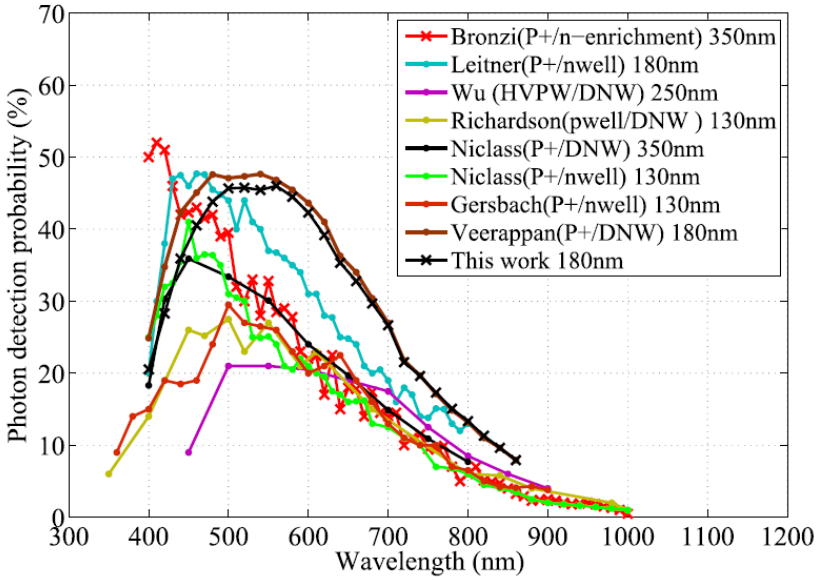


Figure 1.3: Photon Detection Probability (PDP) of various SPAD structures in the literature [2]

on temperature. Instead, its probability increases with the electric field of the junction. Therefore, doping concentration and excess bias are the main determinants of this noise [26]. In Figure 1.2, dark count rate (DCR) levels of various SPAD structures were compared at their operational excess bias voltages [2].

- **Photon Detection Probability (PDP)**

Photon detection probability (PDP) is the parameter used to determine the SPAD sensitivity. It is defined as the probability of photons impinging on the active region of the SPAD of generating a pulse. In SPAD-based image sensors, PDP can be expressed by Equation 1.1

$$PDP(\lambda, \beta, P) = T_s(\lambda, \beta, P) \times QE(\lambda) \times PA, \tag{1.1}$$

where λ is the wavelength, β is the angle of incidence to the surface, P is the polarization state of the incident light, T_s is optical transmittance through the surface, QE is quantum efficiency of the depletion region and PA is the probability of an excited photo-electron to start an avalanche [35]. In Figure 1.3, PDP levels of state-of-the-art SPADs were compared in the visible wavelength spectrum [2]. In SPAD devices, PDP increases with higher excess bias voltage. This chart shows the highest achievable PDP levels for each device; however, when integrated with electronics, certain configuration parameters might not be available to reach those numbers. For instance, most deep sub-micron CMOS processes do not support voltages above 3.3 V; whereas some devices in the list are tested around 10-12 V.

- **Timing Jitter**

As presented in previous sections, one of the strongest advantages of SPADs in single-photon detection is high timing resolution. SPAD-based sensors measure the photon time-of-arrival (ToA) with picosecond accuracy. This feature is characterized by a parameter called timing jitter, defined as the uncertainty of time between photon arrival and avalanche generation. To understand the elements that form jitter, the avalanche mechanism must be investigated. SPADs generate a pulse by exploiting a phenomenon called impact ionization, in which a photon-generated carrier multiplies quickly inside the high electric field region of the junction to start an avalanche current [36]. This process occurs in two stages: carrier build-up and lateral spread of current [37]. Particularly the carrier build-up stage includes random processes in its dynamics, which can be represented as a Gaussian distribution [26]. In addition, in the work [38] it was demonstrated that in avalanche photodiodes timing jitter increases with higher active area diameter.

In an image sensor, the timing jitter of the entire system depends on various factors in the readout stage, aside from the photodiode. In event-driven readout, the most deciding factor that determines timing resolution is the photodiode. Since the signal jitter increases with slower rise/fall times, the readout blocks should be designed to achieve high signal drivability in each stage. On the other hand, in a time-gated image system, gating window resolution is also a significant contributor to ToA measurement. Regardless of the SPAD jitter, a gated sensor cannot achieve higher timing accuracy than the minimum shift of gate edges. This parameter depends on the specifications of the FPGA; thus cannot be improved using integrated circuit design techniques.

- **Afterpulsing Probability and Crosstalk**

In SPADs, the main causes of correlated noise are afterpulsing and crosstalk [32]. During the avalanche process, some carrier charges are temporarily trapped inside the junction. When these carriers are released within nanoseconds, they may trigger another avalanche and therefore generate a second pulse. This phenomenon is called afterpulsing. The best method to minimize afterpulsing is to deactivate the SPAD shortly after a pulse generation, and to switch on after all trapped charges are released. This method is discussed in detail under the title "dead time".

Crosstalk covers a group of events where an avalanche in a diode causes another avalanche in a neighboring diode for a variety of reasons. Crosstalk events can be classified into two categories. Photons of different wavelengths, generated during an avalanche in a pixel, can impinge on neighboring pixels and generate extra pulses there. This event, called optical crosstalk, is a significant risk in diodes with narrow spacing [39]. Electrical crosstalk, on the other hand, occurs in pixels that share a common n-well. This phenomenon is due to hot carriers travelling between separate p+/n junctions inside a single n-well [32].

There are various ways to minimize crosstalk. Separation of neighboring pixels decreases the probability of photons or hot carriers to travel between pixels. Electrical isolation of adjacent junctions using oxide layers such as shallow/deep trench isolation (STI/DTI) significantly reduces the crosstalk probability. However, these measures generate extra photon insensitive area on the pixel surface, which adversely affects fill factor.

- **Pixel Non-Uniformity**

Image sensors with large pixel arrays are vulnerable to problems that become severe for the overall performance due to scaling. An essential requirement for the commercial endorsement of a new imaging technology is the performance uniformity across the pixel array. Recently, large SPAD-based pixel arrays are being fabricated, which provide an opportunity to conduct extensive non-uniformity analyses [40, 41].

Non-uniformity across a pixel array may be caused by a variety of elements such as supply IR-drop, process variations and other random factors. The non-uniformity cause of a certain parameter can be determined by the shape of the distribution curve.

The work [40] presents non-uniformity measurements for several figures of merit on a 512×128 SPAD-based image sensor. For some parameters, exhaustive techniques that measure each pixel in sequence can be time-consuming. Alternatively, some analytical models, such as the one proposed by [42] can estimate the non-uniformity of several performance parameters from simpler measurement results.

- **Fill Factor**

In monolithic solid-state sensors, the active area percentage of the entire chip surface has a major impact on photon sensitivity. In a pixel, the ratio of the active area to the entire pixel area is called fill factor. The photons that impinge on the CMOS electronics or the signal wires cannot be detected. The overall sensitivity, represented by photon detection efficiency (PDE), can be expressed in Equation 1.2.

$$PDE = PDP \times FF \tag{1.2}$$

Various design choices can be made to improve fill factor in an image sensor. Using photodiodes with large diameters in small CMOS technology nodes appears to be the simplest solution to fill factor related issues. However, as the technology node gets smaller, the dark count rate per area increases due to an increase in junction doping concentration. Therefore, in monolithic SPAD pixels there is a trade-off between fill factor and photon sensitivity, which is determined by the SPAD size. Some layers in the SPAD structure reduce the fill factor despite improving other FoMs. Guard rings that prevent premature edge breakdown (PEB), and STI/DTI layers that minimize afterpulsing increase the percentage of photon-insensitive area in the pixel.

In some architectures several SPADs share a deep n-well. Placing multiple junctions on a single well improves fill factor significantly by eliminating a major spacing requirement between two separate n-wells (usually close to 1 μm). Well sharing requires placement of mirrored pixels, which renders the photodiode spacing non-uniform across the array. This flaw is not desirable in many applications, and may require optical or digital correction. In addition, sensors with shared n-well are more vulnerable to electrical crosstalk, as explained in previous sections. Some pixels in the work [1] reach fill factor levels of 57 % thanks to shared n-well technique.

In addition to IC-design techniques, optical solutions are also available to achieve high fill factor. Microlenses are devices that concentrate the photons that fall into the pixel area onto the active region [43]. In the work [4], the use of microlenses increased fill factor by a factor of 6 (from 5 % to 30 %).

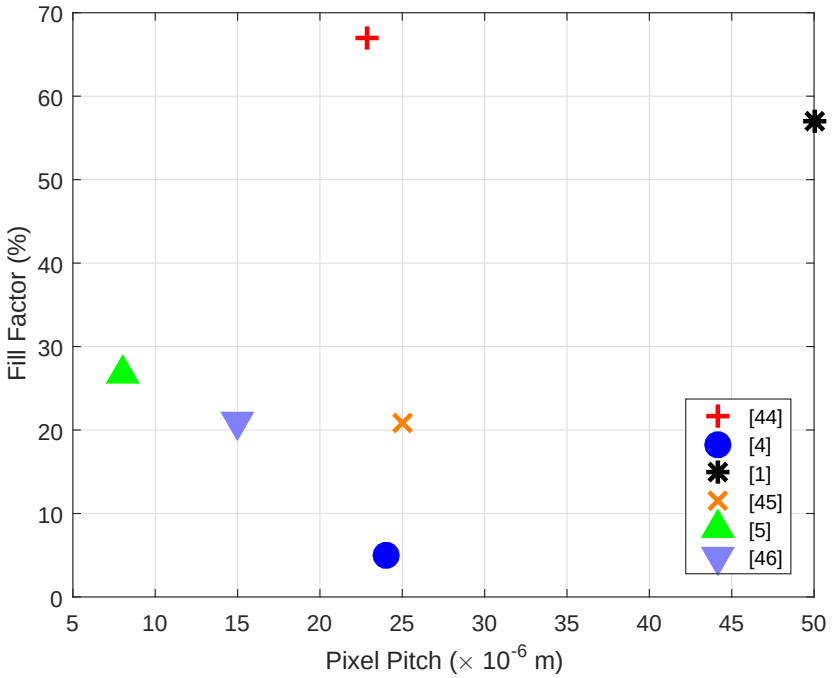


Figure 1.4: Fill factor vs. pixel pitch of state-of-the-art SPAD pixels

With the advent of state-of-the-art 3D processes, CMOS circuitry and signal wires will no longer restrict photon detection by occupying the chip surface. This technology allows the placement of SPADs in the top tier, and the electronics in the lower tiers that are not exposed to light. Different layers of substrates can be connected by through-silicon-vias (TSV). Some 3D processes even allow the integration of different process technologies, which can combine a low-noise SPAD with dense CMOS electronics that are fabricated in lower technology nodes.

In Figure 1.4, fill factor and pixel pitch of state-of-the-art SPAD pixels were compared [44, 4, 1, 45, 5, 46]. It is worth noting that the this chart provides the reader a simplified overview of the current progress in SPAD technology. It is a challenging task to make a fair comparison of pixel FoMs; because not all pixels presented in the chart has the same functionality, or same circuit complexity.

- **Dead Time**

During quenching and recharging periods following a photon-generated avalanche current, the SPAD is insensitive to new photons. This period is called dead time. The duration of dead time in a SPAD has an impact on several other FoMs. Long dead times prevent the sensor from detecting every photon. Too short dead times, triggered by active recharging, increase the afterpulsing rate.

1.3 Contribution

This thesis aims to contribute to SPAD imaging field in following ways:

1. Development of a 512×512 pixel array, the largest multichannel SPAD imaging sensor to our knowledge.
2. Analyzing the implementation techniques to minimize performance drop as a result of scaling.
3. Demonstration of a novel fill factor improvement technique that does not increase crosstalk in contrast to n-well sharing.
4. Design of a 2D monolithic pixel architecture that supports significantly greater excess bias voltages than CMOS processes can support.

Due to time constraints, the measurement of the 512×512 array is not included in this thesis. In the future, this device can provide valuable insight into the non-uniformity issues of the SPAD technology, a subject that must be comprehensively studied for the commercialization of SPAD imaging.

1.4 Overview

This thesis concentrates on digital SPAD pixel architectures and implementation of large pixel arrays. It includes 3 single-photon time-resolved SPAD sensors with different configurations and process technologies. The thesis also analyzes performance characteristics, including scalability-related challenges. Chapter 2 presents the pixels that were used in each of 3 chips in the project. It compares the strong and weak sides of each pixel type, discusses the function of each stage in a pixel, and finally explains their modes of operation. Chapter 3 discusses several optimization methodologies to improve figures of merit of a sensor. It also demonstrates some of these techniques on 2 4×4 time-gated SPAD sensors designed in 110 nm CMOS process. The measurement results of the main figures of merit are also presented. Chapter 4 presents a 512×1 event-driven linear SPAD array and a 512×512 time-gated image sensor, both designed in 0.18 μm CMOS technology. Both of these sensors employ a p-i-n diode-based SPAD with very high photon sensitivity and low noise. The linear sensor allows measuring the maximum photon sensitivity and time jitter of the p-i-n diode thanks to its high excess bias support and lack of time gating. 512×512 time-gated sensor is the largest SPAD-based multichannel pixel array designed to our knowledge. It is suitable for studying the scalability issues of SPAD-based imaging, and for the generation of high-resolution images. Finally, chapter 5 concludes the thesis by summarizing the work done for the project and proposing various ideas for the future work.

2 Pixel Architecture

A pixel is the most elementary circuit block of an image sensor. In large arrays designed in deep sub-micron CMOS technologies, the pixel contains almost all electronic functionality of the sensor. This chapter is dedicated to pixel architectures and functions in image sensors.

In this thesis, three image sensor chips, which were designed in the M.Sc. project, are introduced. Therefore, this chapter strongly focuses on the pixel types used in these chips. In two time-gated sensors, a fully-digital pixel model with in-pixel memory was used. This pixel employs a SPAD, quenching and recharging module, a 1-bit dynamic memory and a readout scheme. There are two variants of the pixel model. The pixel designed in 110 nm technology allows excess bias voltages up to 3.3 V, thanks to the thick-oxide transistors offered by the technology. On the other hand, the pixel designed in 0.18 μm technology is adapted to higher excess bias voltages than typical CMOS transistors can handle. The goal was to operate the p-i-n diode-based SPAD with the highest photon-sensitivity and noise performance by reaching 5-6 V. The 512×1 SPAD array employs a more basic pixel with less functionality, thanks to its event-driven readout configuration and availability of an output pad for each pixel in the array.

This chapter is organized as follows: Sections 2.1 and 2.2 are dedicated to the main pixel that forms the 512×512 time-gated image sensor. Section 2.1 analyzes the components with various functions in the pixel. Section 2.2 describes the main operation modes of the time-gated pixel. Section 2.3 introduces two other pixel types: another variant of the main pixel which was used for the 4×4 time-gated sensor, and an event-driven pixel architecture for the 512×1 linear array.

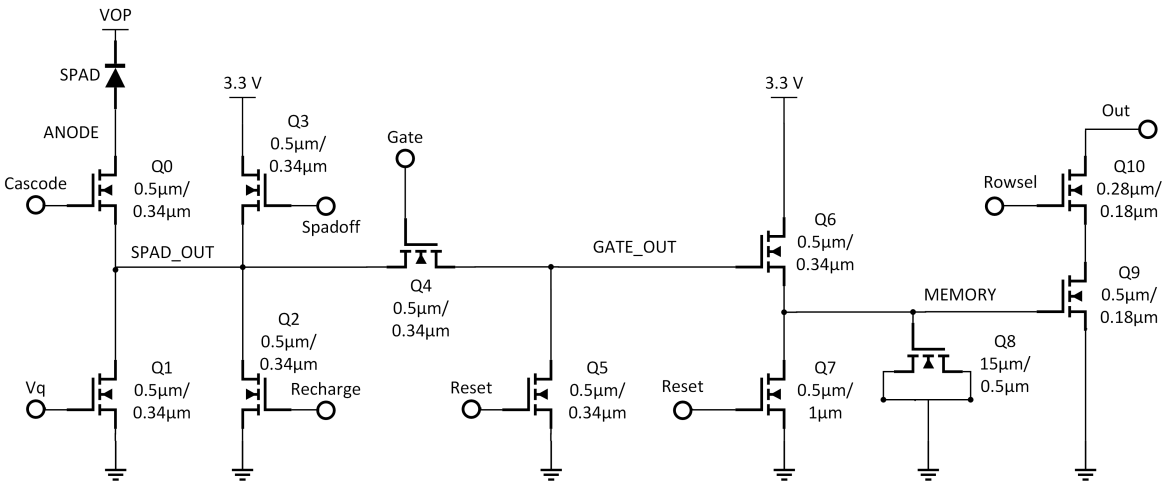


Figure 2.1: Pixel schematic view of the 512×512 pixel array designed in 0.18 μm technology

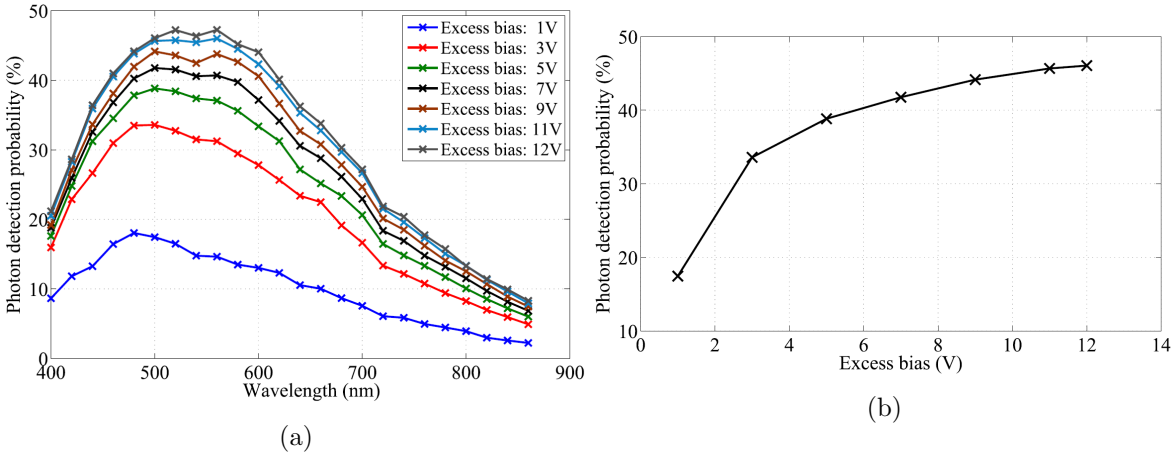


Figure 2.2: Photon detection probability (PDP) of the p-i-n diode-based SPAD for (a) various wavelengths and (b) various excess bias voltages [2]

2.1 Pixel Components

The main pixel is formed by a SPAD and 11 NMOS transistors. Its schematic view is shown in Figure 2.1. Q0-Q6 are thick-oxide transistors and Q7-Q10 are thin-oxide transistors with operating voltages of 3.3 V and 1.8 V, respectively. The electronics perform the following functions: quenching and recharging, 1-bit dynamic memory, time gating, memory reset and pixel readout. Each of these functions are described in detail in the related subsection.

2.1.1 Single-Photon Avalanche Diode (SPAD)

A SPAD is a photodiode operated typically in reverse bias, and above breakdown voltage. This mode of operation is called Geiger mode. In Geiger mode, through impact ionization mechanism, a SPAD generates a very large current upon photon arrival, which is converted to digital pulses using various methods by digital pixels. All photodiodes used in this thesis project are SPADs. Although all of them operate based on the same fundamental photoelectric principles, there are major structural variations among them.

In the main pixel, p-i-n diode based SPADs were used. This SPAD achieves among the highest levels of photon-detection-probability (PDP) and the lowest levels of dark-count rate (DCR) in the literature [2]. According to the data presented in Figure 2.2a and Figure 2.3a, this SPAD can reach PDP greater than 40 % from 460 to 600 nm and DCR of 1.5 $cps/\mu m^2$ at 11 V excess bias. The SPAD cross-section is presented in Figure 2.4. A variety of techniques that were used in the design of this photodiode lead to the high performance reported in [2]. Nevertheless, these techniques also decrease the fill factor, resulting from the trade-off between DCR, PDP and fill factor. To analyze the effects of design choices on each performance parameter, the structure of the p-i-n SPAD should be comprehensively studied.

The SPAD employs a p+/n main junction. A lightly doped p-well guard ring

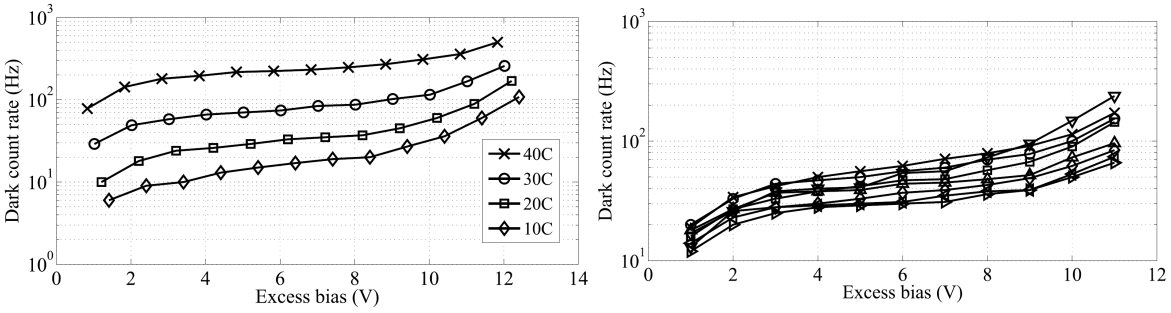


Figure 2.3: Dark count rate (DCR) vs. excess bias of the p-i-n diode-based SPAD for (a) eight different devices at 25° C and (b) various temperatures [2]

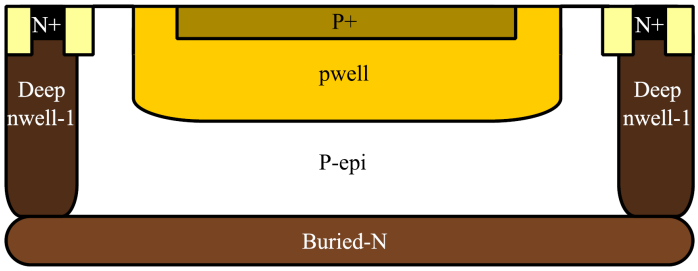


Figure 2.4: Cross-section view of the p-i-n diode-based SPAD [2]

outside the p+ region, and a more lightly doped p-epi layer outside the p-well increase the depletion region width of the lateral junction. That eliminates the possibility of PEB, which occurs when the lateral junction has a higher electric field level than the main junction at the same bias voltage. PEB restricts the utilization of the diode active region; therefore it makes the pixel less photon sensitive and more noisy. The electric field distribution of the SPAD in Figure 2.5 verifies that the main junction of the SPAD has the highest electric field; therefore it reaches the breakdown voltage earlier than the other junctions. On the other hand, the measures taken to widen the depletion region lead to a significantly large SPAD surface area that is insensitive to photons. As a result, the fill factor, the ratio of SPAD active area to the entire surface area, decreases. Fill factor is an essential FoM for image sensors, since it has a direct impact on the device photon sensitivity. Another challenge introduced by the p-i-n diode-based SPAD is the difficulty to reach 11 V excess bias with standard deep sub-micron CMOS technologies. This issue is addressed in the following sections and several techniques to reach voltages above the CMOS transistor capacity are proposed.

2.1.2 Quenching and Recharging

Quenching and recharging are two critical stages in the operation of a SPAD. When a SPAD fires upon photon arrival, a very high avalanche current flows through its terminals. If this current persists for a long time, it may lead to serious damage in the device. The most common way to stop the avalanche current is to lower the

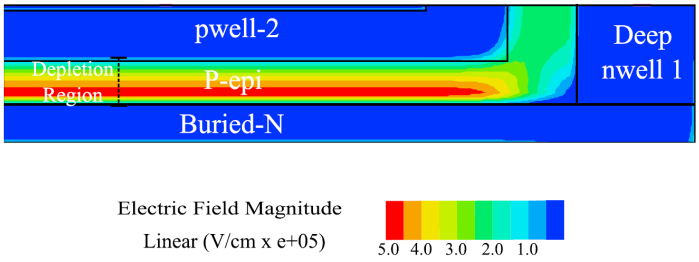


Figure 2.5: Electric field distribution simulation results of the p-i-n diode-based SPAD [2]

voltage across the SPAD terminals below the breakdown voltage. This process is called quenching. Quenching is usually implemented with a large resistor, whose terminals experience a voltage drop as a result of high current. Depending on the resistance and the value of the current, the SPAD can be quenched in less than a nanosecond. After quenching, the SPAD loses its photosensitivity until the voltage across its terminals is restored to the initial level above the breakdown voltage. This stage, called recharging, is usually the next event after quenching in the SPAD operation. Recharging can be performed passively or actively. After the avalanche current is quenched, the voltage across the resistor gradually drops to zero, which slowly brings the SPAD back to the Geiger mode. This is called passive recharging. The duration of passive recharge is limited by the minimum resistance requirement of quenching, which is usually in the order of hundreds of $k\Omega$. Consequently, the time required to reactivate the SPAD through passive recharging is usually greater than 50 ns. For certain readout modes, such as time-gating, the speed of passive recharging is not sufficient. Compared to a fixed resistance, a transistor controlled by an external signal can recharge the SPAD faster, thanks to its variable resistance. This method is called active recharging.

In the main pixel presented in Figure 2.1, both passive and active recharging configurations were employed. The transistors Q1 and Q2 function as quenching and recharging transistors, respectively. Q1 operates in the weak-inversion mode, effectively as a resistor. The current generated by the SPAD upon photon arrival leads to a voltage drop across the drain and source terminals of Q1, immediately driving the SPAD voltage below breakdown. During quenching, the anode voltage of the SPAD increases from 0 V to $VOP - V_{breakdown}$. The node $SPAD\_OUT$ reaches a lower voltage than the SPAD anode due to a limitation by the gate terminal of Q0. The function of the cascode transistor Q0 is discussed in the next sections. Immediately after quenching, the voltage of $SPAD\_OUT$ slowly drops due to passive recharging. The status of active recharging depends on the timing of the operating signals. The recharge transistor is only switched on for several nanoseconds shortly before the gating window. Its purpose is to ensure that the SPAD is active during the entire gating window. If the SPAD fires during the active recharging period, $SPAD\_OUT$ may rise due to voltage drop, and subsequently the gate may store the event in the memory even though the photon arrival occurred outside the gate window. Therefore, it is desirable that the active recharging period is as short as possible; since it has a major impact on gate uniformity, which is among the most important FoMs in this sensor.

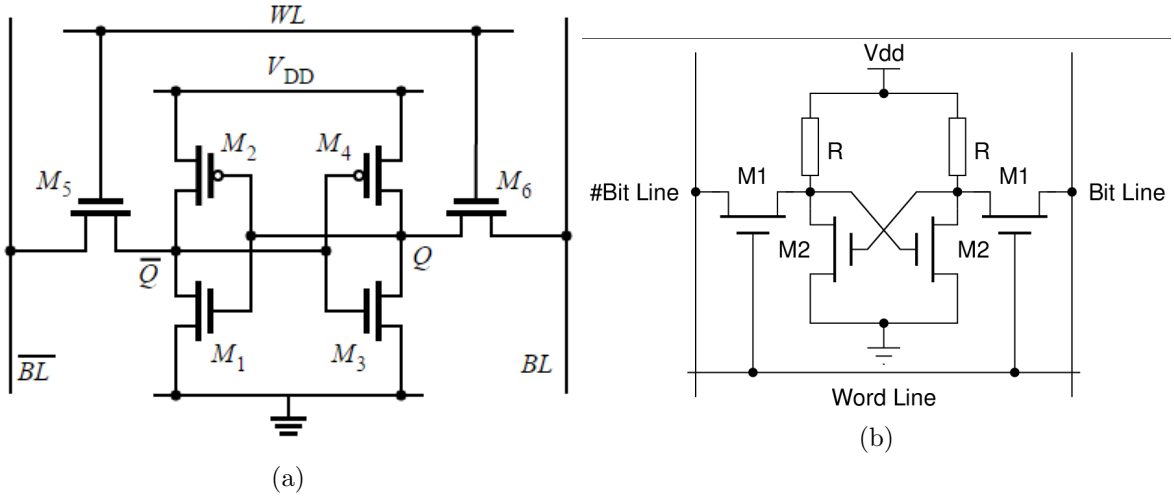


Figure 2.6: The architecture of a (a) 6 transistor and (b) 4 transistor static memory cell [3]

2.1.3 In-Pixel Memory

In-pixel memory is included in pixels which do not immediately send the photon arrival information outside the chip. In the 512×512 sensor, rolling-shutter based readout requires the pixels to store the photon arrival information for a period of time until the readout of the specific row where the pixel is located. This storage duration may last up to 6.4 μs in the typical mode of operation. However, in certain cases where the gate window must be open to detect photons under very low-light-level conditions, the memory state must be intact for several milliseconds.

In-pixel memory architectures can be classified by their operation principles. A static memory can maintain its state permanently as long as its supply is provided. It can only be toggled by an external signal. In contrast to this, a dynamic memory can only preserve its voltage for a finite period of time; hence it must be refreshed periodically to store information for long time intervals.

Despite their inherent advantages of reliability, the use of static memory in an image sensor pixel can be problematic. To discuss these problems, two popular static memory cells are evaluated. The first cell consists of 2 cross-coupled inverters and contains 4 NMOS and 2 PMOS transistors (Figure 2.6a). PMOS transistors can potentially increase the total pixel area due to their n-well spacing requirements; therefore, they are avoided in some SPAD pixels. An NMOS-only version is also available, as displayed in Figure 2.6b. This version performs pull-up using polysilicon resistors. Because of large area occupied by polysilicon resistors, pull-up devices can be implemented with NMOS transistors in weak-inversion mode. The drawback of this version is the static power consumption of the terminal with low voltage. The constant current can be minimized by increasing the resistance of the pull-up devices, at the expense of more pixel area.

An alternative in-pixel memory architecture is dynamic memory. Dynamic memory can be implemented with a single NMOS transistor whose source and drain terminals are grounded, and whose gate is connected to the related node. Although it consumes

no static power, it suffers from gradual voltage drop due to CMOS leakage current. Depending on the time requirement to store the voltage level, the size of the dynamic memory transistor must be carefully chosen and its performance must be verified using simulation tools.

In the rolling-shutter based chips designed for this project, dynamic memory cells were used to store photon arrival events. The sizes of the memory cells were determined by the minimum time requirement of the pixel to store high voltage. The 4×4 chip was designed mainly to test the SPAD structure, the pixel architecture and the novel techniques to improve pixel density. As a result, the in-pixel memory was implemented with minimum-sized transistors. On the other hand, the 512×512 array designed to operate fully as an image sensor. As discussed in the previous paragraphs, the size of the dynamic memory transistor was set to a value which is sufficient to preserve its high state for several milliseconds. This feature permits the sensor to operate at very low-light-levels with very long gate windows. The drawback of a high-capacity dynamic memory is that it occupies a significant portion of the pixel area. The leakage current that gradually discharges a dynamic memory usually flows through a reset transistor between the memory node and ground. Sizing of the reset transistor has a major impact on dynamic memory performance. A wide transistor discharges the memory with a high speed when switched on. This feature may be required to achieve high readout speeds. A narrow transistor, on the other hand, causes lower leakage current, thus allowing the memory to store its charge longer. The best strategy must be to choose the minimum-sized reset transistor that supports the desired chip readout speed.

2.1.4 Time Gating

All chips that are presented in this thesis support time-resolved imaging as a requirement of the target applications. A common method of time-resolved imaging is to measure photon ToA using a time-to-digital converter (TDC). In this architecture, a TDC block measures the time between the generation of a laser beam by the sensor and the incoming photon from the target. While TDC-based ToA measurement is very accurate with a timing resolution of several picoseconds, its implementation becomes complicated in large-sized arrays due to the size of a TDC block. An alternative method is to use time gating. This configuration includes a gating transistor which controls the connectivity between the SPAD and the in-pixel memory. An avalanche can only be recorded in the pixel memory if it falls inside the gate window, i.e. when the gate transistor is on.

In Figure 2.1, Q4 is the gating transistor. Controlled by the 3.3 V signal *Gate*, the gating transistor transfers the voltage of *SPAD_OUT* to the pixel memory. An important detail to be recognized is the source follower transistor Q6 between the gate and the memory. The purpose of Q6 is to ensure that the gate cannot discharge the memory if *SPAD_OUT* voltage is low inside the gate window.

The advantage of time gating is its relatively small size: it requires only 2 extra transistors per pixel. This makes it a more scalable option compared to TDCs. Nevertheless, the performance of the gated system cannot match the TDC-based timestamping. Its timing resolution is defined by the minimum shift in time of the signal leading edge

that an FPGA can support. In the work [4], the timing resolution of the gate was reported to be 20 ps.

2.1.5 Spadoff

In an image sensor pixel, total controllability of the SPAD is essential for a variety of reasons. Firstly, an extensive characterization and debugging of the entire pixel can only be possible if the pixel electronics can be tested independently from the SPAD. Secondly, keeping the SPAD in the off state outside the gate window prevents unnecessary power consumption and increased afterpulsing probability due to extra avalanche events.

In this pixel, the transistor Q3 is responsible for switching off the SPAD outside the gating window. In the absence of the cascode transistor Q0, Q3 would charge the SPAD anode to high voltage when the signal *Spadoff* would rise, thus bringing the voltage between SPAD terminals below breakdown until active recharging which occurs shortly before the gate window. *Spadoff* signal must be synchronous with *Recharge*: setting *Spadoff* and *Recharge* signals at the same time would disrupt pixel operation by forming a low resistance path through Q3 and Q2 and causing extra power consumption.

The effectiveness of the spadoff mechanism is dependent on a variety of factors. Spadoff can only be 100 % effective if it manages to switch the SPAD from Geiger mode to standard reverse bias mode. The main condition which must be met is that SPAD excess bias voltage is less than the V_{dd} of the CMOS process. As explained in previous sections, the p-i-n diode based SPADs achieve top performance at higher excess bias voltages (11 V) than CMOS processes can support. These two requirements cause a trade-off between low afterpulsing and high PDP, decided by SPAD excess bias voltage. Moreover, NMOS transistors are inherently less effective in voltage pull-up than PMOS transistors. This phenomenon can be explained by the fact that a MOS transistor enters cut-off mode when its gate-source voltage, V_{gs}, falls below the threshold voltage, V_{th}. In the pull-up configuration, V_{gs} of a PMOS is always equal to V_{dd}; therefore the conductive path never disappears until the voltage of the drain terminal equals V_{dd}. On the contrary, in an NMOS, the target node is connected to the source terminal; therefore, the transistor enters the cut-off mode as soon as the SPAD anode voltage reaches $V_{dd} - V_{th}$. Despite its low effectiveness, the choice of spadoff transistor in this pixel was NMOS due to extra area requirement of a PMOS transistor.

2.1.6 Readout

In the pixel shown in Figure 2.1, Q9 and Q10 are responsible for pixel readout. The readout of this pixel occurs by pulling down the output bus whose default voltage is high. For a pixel to discharge the output bus, both Q9 and Q10 must be conducting simultaneously. This requirement is fulfilled if the memory is in the high state when the pixel is selected by sending a pulse to the pin *Rowsel*.

The readout speed is dependent on the sizing of Q9 and Q10. For the fastest readout, the size of Q9 must be wider than the minimum configuration; because a large Q9 minimizes the effective series resistance between the bus and ground. The same rule applies to Q10; however, this is not the most dominant impact of the properties of Q10. Since the output bus is connected to all pixels in the same column, the effective

load capacitance of the pixel readout mechanism is dominated by the gate-source capacitance (C_{gs}) of all Q10s combined. While reducing the series resistance, a wide Q10 also generates a large load that is more difficult to discharge. Due to the dominance of its capacitive effect, a minimum-sized Q10 is the most optimal choice. There were also multiple factors limiting the maximum size of Q9: Firstly, due to the relatively larger series resistance of Q10, further reduction of Q9s resistance would have little effect on the total value. Secondly, the C_{gd} of a large Q9 would cause unwanted voltage fluctuations in the node *MEMORY* during readout due to capacitive AC coupling, which could potentially damage the transistor junctions permanently. The third constraint is based on area concerns due to increasing transistor size.

The output terminal of the pixel is connected to a column bus that is shared by the entire column. The structure and the operation modes of the chip readout mechanism are presented in chapter 4.

2.1.7 Cascode for High Excess Bias

The p-i-n diode based SPAD operates optimally at excess bias voltages up to 11 V, significantly higher than the CMOS technologies can support. To fully utilize these SPADs, a cascode transistor named Q0 was placed between the SPAD and the rest of the pixel. This transistor, permanently biased at 3.3 V, allows the SPAD anode voltage to vary between 0 V and 6.6 V. Despite the higher PDP and lower DCR offered by high excess-bias voltages, the cascode transistor has its drawbacks, too. Firstly, the cascode transistor increases the effective resistance between the SPAD anode and the ground. Therefore, the time required for active recharging process increases significantly. A fast fall time is required for the SPAD anode to minimize the transition stage where it is uncertain whether the SPAD is on or off. However, the SPAD model used in circuit simulation tools does not accurately predict the SPAD response to photon arrival. Therefore it is difficult to quantify the possible negative effects of the cascode transistor on gate jitter, based on simulation results. This effect can be compensated by increasing the width of the recharge transistor. Secondly, with increased excess bias voltage it is not possible to turn off the SPAD from the anode. Since Q3 can only increase the anode voltage by 2.6 V, the SPAD would still be photon sensitive during spadoff window. According to Figure 2.2b, in a pixel with no cascode transistor and maximum excess bias of 3.3 V, *Spadoff* signal reduces the SPAD PDP by 63 % (13 % PDP at 0.7 V vs. 35 % PDP at 3.3 V). In the cascoded pixel in Figure 2.1, however, the PDP decrease due to spadoff is only 14 % (42 % PDP at 6.6 V vs. 36 % PDP at 4 V). While not posing a fundamental threat to the pixel operation, the ineffectiveness of spadoff increases the overall pixel noise due to afterpulsing. Moreover, due to its weak impact on pixel sensitivity, the high power consumption of spadoff that dominates the overall chip power can no longer be justified. For many applications, the benefits of permanently deactivating the transistor Q3 can have more benefits than costs.

2.1.8 Reset

The in-pixel memory can be set and reset from outside through digital signals. The reset feature is essential for pixel operation: the memory must be reset after each pixel

readout in order to be sensitive to the next photon arrival. Furthermore, it improves the chip testability. For instance, in the pixel in Figure 2.1, the memory can be totally isolated from the SPAD. Consequently, the electronics of the pixel can still be tested if the SPAD does not function. The second role of the reset signal is to discharge *GATE_OUT* node while the memory is being reset. In the absence of Q5, the signal *Reset* would not be sufficient to discharge the memory completely when *GATE_OUT* is high. Instead, both Q6 an Q7 would be on simultaneously; and a high current would flow through them for a brief period of time. This undesirable event is prevented by simultaneously resetting *MEMORY* and *GATE_OUT*.

2.2 Pixel Operation

The pixel operation mode and parameters are controlled by and FPGA by means of several input signals. Figure 2.7 displays the typical operation mode of the pixel. Every pixel is connected to 5 input signals. 3 of them are high-voltage (3.3 V) global signals: *Spadoff*, *Gate* and *Recharge*. These signals ideally reach all pixels in the entire array simultaneously. Their activities continue during pixel readout. The order of global signals is as follows: First, *Spadoff* rises to deactivate the SPAD in order to avoid afterpulsing in case the SPAD fires shortly before the gate window. Then, *Spadoff* falls and shortly after that, *Recharge* rises. The purpose of *Recharge* is to restore the SPAD bias voltage to above breakdown. *Recharge* is deactivated as soon as the anode voltage reaches the initial level. This event is followed by the gating window, which is defined by the high level of the *Gate* signal. The pixel memory can only store photon arrivals events that occur inside the gating window. Depending on the application, the duration of the gate window may vary from 4 ns to several milliseconds. The remaining 2 signals are low-voltage (1.8 V) local signals, i.e. they are provided to each row at different times. The readout of the pixel array is typically done in a rolling shutter mode: each row of the final frame is captured at different time intervals. In a row, when *Rowsel* signal rises, all pixels in that row with a recorded photon event in their memory cells pull down the output bus of their columns. Then, the state of each column bus is sampled by d-flip-flops. Finally, *Rowsel* signal falls and the column bus is immediately pulled up using PMOS transistors. Once all column buses are sampled and pulled up, the next cycle starts with the next row. This readout procedure works continuously, independent of the global gating mechanism. In order to reach the target frame rate of 100 fps for 10-bit grayscale images, the entire readout cycle of 1 row must be below 20 ns.

2.3 Pixel Variants

In addition to the main pixel that was analyzed in the previous sections, several other pixels were also designed in this work. In this section, different versions of SPADs and pixel architectures are presented. Furthermore, the features of these SPAD and pixel types are compared, and finally their positions in design trade-offs are discussed.

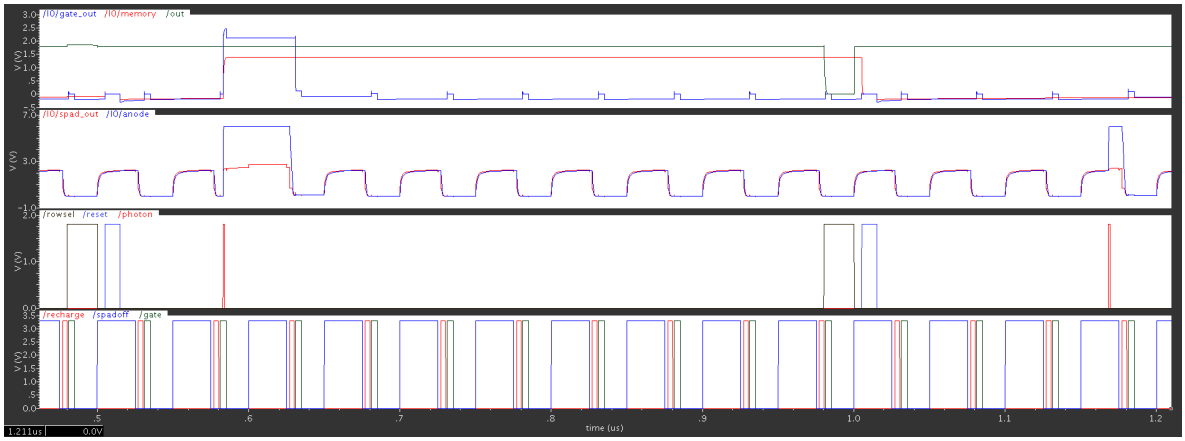


Figure 2.7: Timing diagram of the 512×512 time-gated image sensor

2.3.1 SPAD Variants

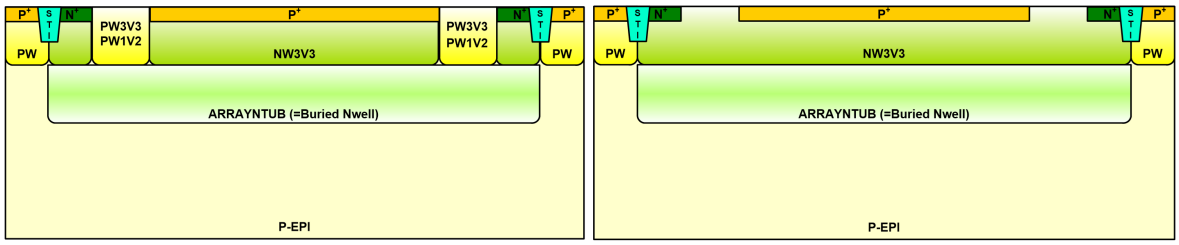
In this thesis, 5 different SPAD cells were used. These SPADs can be classified based on 2 criteria: their shapes and their guard ring types.

In SPAD design, there is a major trade-off between PDP and fill factor. A design choice that determines the position in this trade-off is the device shape. A round SPAD has a uniform electric field distribution; whereas a rectangular SPAD has higher electric field levels at the corners. These high electric field zones render the device vulnerable to PEB; hence they may reduce the PDP and increase DCR significantly, depending on the sharpness of the corners. While not suffering from non-uniform high electric field concentration, a round SPAD leads to sub-optimal utilization of the pixel area by introducing blank spaces with curved boundaries. Rectangular SPAD design with round corners can benefit from the relative advantages of both versions, achieving higher fill factor with no compromise of the noise performance.

Two versions of the sensor with 512×512 array were designed with round (Figure 4.4) and square (Figure 4.5) SPADs to compare the overall performance of the two shapes. The architectures and layouts of the 2 chips are identical except the SPAD cells. According to Table 4.2 , a square SPAD offers approximately 24 % higher fill factor than a round SPAD.

In SPAD design, a popular method to prevent PEB is to surround the active area with a guard ring. A guard ring is essentially a low-doped material that widens the depletion region of the lateral junction and lowers the electric field in that zone. This exact reason also makes the guard ring area less sensitive to photons in the Geiger mode. As a consequence, the designer has to consider this trade-off while designing the SPAD.

Three types of guard rings were employed in the SPADs in this thesis. The first type of guard ring, as displayed in Figure 2.8a, employs a p-well around the p+ region. The doping profile of the vertical p+/n junction is higher than the lateral p/n junction; hence PEB is avoided. A virtual guard ring, shown in Figure 2.8b, contains no p-well. Both vertical and lateral junctions are formed by p+ and n-well layers. However, the n-well layer does have retrograde doping profile, which means that its doping



(a) (b)

Figure 2.8: Cross-section view of 2 variants of p+/n diodes designed at 110 nm technology with (a) p-well guard ring and (b) retrograde n-well guard ring

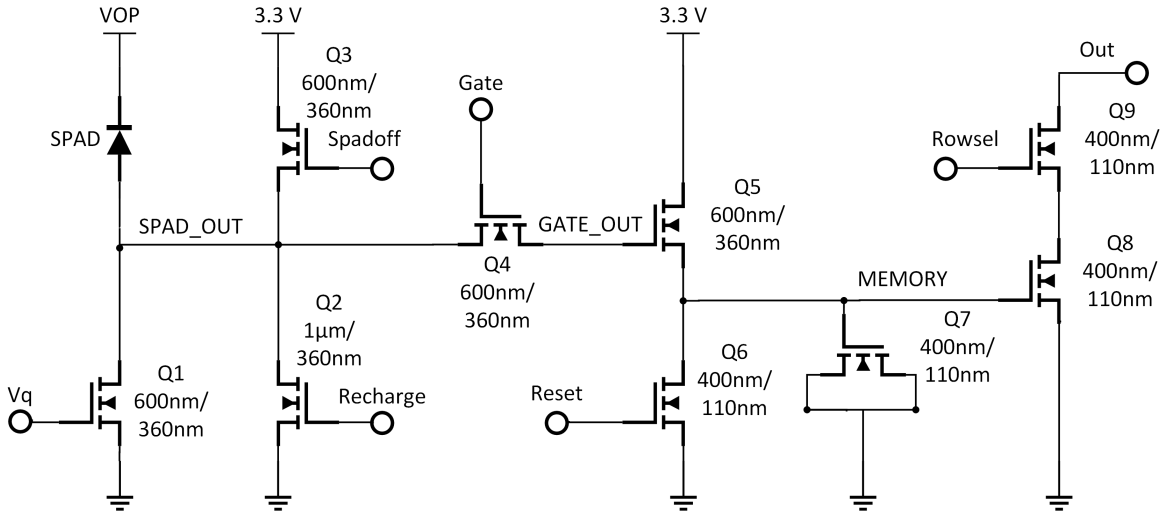


Figure 2.9: Pixel schematic view of the 4×4 pixel array designed in 110 nm technology

concentration decreases towards the wafer surface. The resulting effect is similar to the p-well guard ring: the doping profile of the vertical junction is higher than the lateral junction. The third type of guard ring is based on a phenomenon called p-well lateral diffusion. Due to the diffusion method during fabrication the doping profile of the p-well is not uniform across the entire layer; instead, the lateral junction has less doping concentration. In the p-i-n diode whose cross-section is shown in Figure 2.4, lateral diffusion and a lightly doped p-epi layer surrounding the p-well provide PEB protection.

2.3.2 A Time-Gated Low Excess Bias Pixel with In-Pixel Memory

For a 4×4 image sensor designed in 110 nm CMOS technology, a variant of the main pixel was designed. This pixel, displayed in Figure 2.9, is based on the same architecture as the main pixel; however, it lacks certain features that the main pixel contains.

The p+/n junction-based SPADs shown in Figure 2.8 are designed to operate at low excess bias voltages within the range of CMOS transistor operating voltages. In contrast to the main pixel, increasing the maximum excess bias via a cascode transistor was not

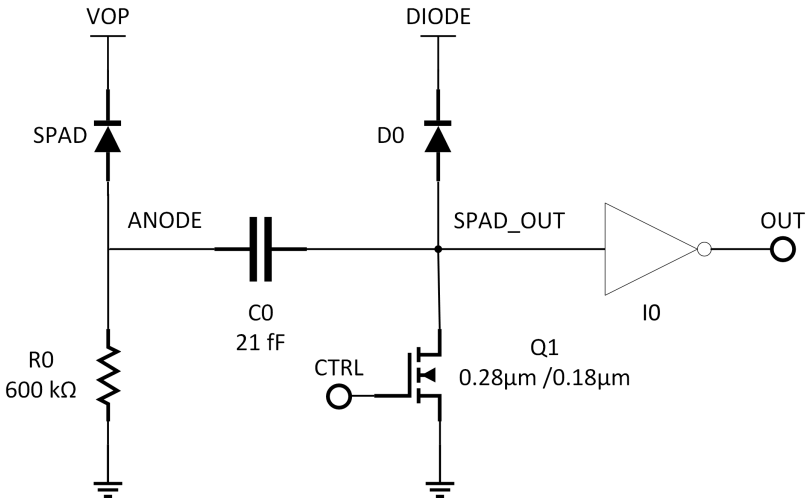


Figure 2.10: Schematic view of a pixel in the 512×1 image sensor

needed to achieve maximum SPAD performance in this pixel variant. Therefore this pixel supports excess bias voltages only up to 3.3 V. In addition, due to less demanding requirements of a small pixel array, the sizes of the in-pixel memory and the readout transistor are also smaller.

There are two layout versions of this pixel: a conventional layout and a non-symmetrical, high-fill factor layout. These layout variants and the techniques used to boost the performance are extensively discussed in chapter 3.

2.3.3 An Event-Driven High Excess Bias Pixel without In-Pixel Memory

The 512×1 linear sensor uses an event-driven readout configuration. The pixel designed for this sensor is based on a fundamentally different architecture from the ones presented earlier. Displayed in Figure 2.10, the pixel contains a p-i-n SPAD that is larger than the previously presented SPADs, and fewer number of components than the other pixels. The most important feature of this pixel is its ability to reach excess bias voltages up to 12 V. This was achieved by two techniques: implementation of passive quenching resistance with a poly resistor (R0) instead of a weak-inversion transistor, and the DC isolation of SPAD anode and *SPAD_OUT* using an AC coupling capacitor (C0).

2.4 SPAD Model for Circuit Simulations

The chips that are presented in this work were designed and simulated using CAD tools for integrated circuits. These tools are typically used for analog and digital electronic circuits, and they are not compatible with photodiodes. Since the SPADs were designed in standard CMOS process technologies, they had to be modeled using the available electronic blocks offered by these tools.

As shown in Figure 2.11, a SPAD was modeled with an SPST switch, an ideal DC supply and a capacitor. The DC voltage of the supply represents the breakdown voltage, the capacitor indicates the equivalent capacitance of the main SPAD junction,

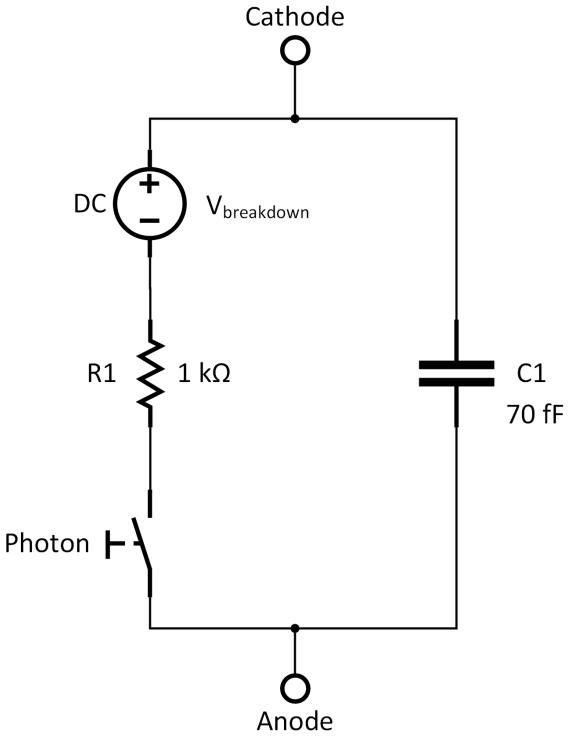


Figure 2.11: The SPAD model designed for simulations

and finally the on resistance of the switch determines the magnitude of the avalanche current. Impact ionization events are simulated by sending digital pulses to a virtual SPAD pin called *Photon*. The values chosen in the model are retrieved from empirical results during the characterization of the SPAD [2].

The accuracy of the SPAD model has a strong impact on the chip performance simulation results. For instance, the equivalent on resistance after impact ionization determines the risetime of the anode voltage due to avalanche current. Active recharging is essentially the discharging of the SPAD capacitance from high voltage to zero. Therefore, recharge time is affected by the equivalent capacitance of the SPAD, which contributes to the gate window uncertainty. It is worth noting that the model cannot simulate the SPAD behavior with 100 % accuracy; therefore the timing resolution and the gate uniformity of the chip can only be characterized through measurements.

3 Device Optimization

Image sensors combine photosensitive devices with electronic components in a very compact pixel area. Consequently, optimal design techniques have key significance in image sensor design to improve the specifications. It is a challenging task to achieve high standards in all figures of merit as the pixel sizes are shrinking and the number of pixels in a chip is increasing.

In this chapter, various optimization categories and techniques will be analyzed. In addition, some implementations of these techniques in a 4×4 pixel array will be presented.

3.1 Pixel Density Improvement

In previous discussions, it was mentioned that an important figure of merit for an image sensor was photon sensitivity. In a monolithic image sensor, the photodiode and photon-insensitive electronics must be placed on the same surface. As a result, the ratio of photosensitive area to the entire area, called the fill factor, becomes an important design concern. Along with SPAD structure choice, pixel density improvement techniques play a major role in cramming the most functionality into a unit chip area. In this section, implementations of several techniques are discussed.

Essentially the density improvement techniques can be classified into three categories. The first and the most effective method is to reduce the size of the electronics in the pixel. In other words, decreasing the width and the length of transistors or reducing the number of transistors in the pixel results in significant density improvements. However, doing that without compromising functionality is a major challenge, and such techniques are architecture-specific. Therefore, these techniques are only discussed in the related chapters for the specific pixel architecture. The second method is to choose the SPAD shape and set its size to large values relative to the electronics. For instance, choosing a rectangular SPAD avoids the non-usable arch-shaped blank spaces. However, this choice can lead to PEB if the corners are too sharp. On the other hand, the downside of using too large SPAD structures to achieve fill factor is high SPAD noise and nonuniform photon sensitivity. The third method uses various forms of resource sharing between pixels. The shared elements can be certain SPAD layers, such as n-well, that have stringent minimum spacing design rules. In addition, signal wires carrying global signals can also be shared between multiple pixels. An major negative effect introduced by resource sharing methods is pixel non-uniformity. When multiple pixels share one resource, the orientations of neighboring pixels are usually the opposite of each other. This non-uniformity could be partially compensated using non-symmetric microlenses.

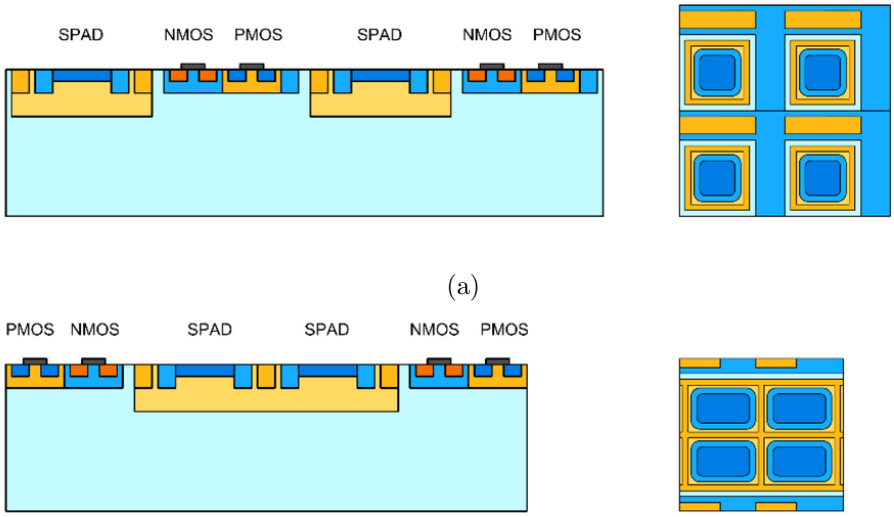


Figure 3.1: 2 different SPAD pixel layout types: (a) separate deep n-well and (b) SPADs sharing deep n-well

3.1.1 Deep N-well Sharing

In recent years, researchers have been investigating novel methods of SPAD placement in a pixel to improve fill factor without compromising functionality. Of these methods, deep n-well sharing, illustrated in Figure 3.1, yields promising results [47]. This method is based on eliminating the unutilized area in a pixel due to minimum n-well spacing DRC rule. In a conventional image sensor array, all pixels have identical layouts (Figure 3.1a). A photodiode, located on one pixel corner, is surrounded by signal and bias lines along x and y-axes. Usually, photodiodes are located on the blank space enclosed by 2 neighboring x and y-axes lines, preferably at equal distance to all 4 of them. The key advantage of this placement is a totally uniform distribution of the photodiodes across the pixel array, which is a desired feature for image quality. The alternative layout solution is n-well sharing between SPADs. Shown in Figure 3.1b, this method reduces the number of dead spaces between 2 n-wells, namely the n-well of the SPAD and the n-well of PMOS transistors. While increasing fill factor, n-well sharing requires the SPADs of adjacent pixels to be placed next to each other with no spacing. This requires the transistors to be moved outside, which can be only possible if the neighboring pixels have different layouts. Consequently, the distribution of photodiodes across the array becomes non-uniform. It should also be noted that deep n-well sharing is an effective method to boost the fill factor only in pixels with PMOS transistors.

3.1.2 Signal and Power Line Sharing

In addition to deep n-well sharing, various other pixel components can be shared, as well. As explained in previous discussions, signal and power lines occupy a large part of the photon-insensitive area; hence decreasing fill factor. Any technique that reduces

the overall area of these lines strongly contributes to pixel density.

There are several constraints of key importance to be taken into account while implementing this technique. The first factor that limits the level of sharing is the multichannel structure of an image sensor. In contrast to silicon photomultipliers (SiPM) where an entire pixel array is connected to a single output, image sensors contain structures to identify the location of a photon. As a result, in an image sensor there are lines that cannot be shared among pixels. Among them are output lines and local signal lines that transfer location-related information. On the other hand, all global signal and bias voltage lines can be shared. The second point is that the designer must recognize the dramatic increase in load capacitance after the signal and power lines are merged. Depending on the array size and parasitic parameters of the lines, the width of the shared wires may need to be widened to maintain the same drivability for signals and the same level of IR-drop for bias voltages.

The major drawback of this technique is similar to n-well sharing: the placement of SPADs across the array must be non-uniform. This technique can boost the fill factor with no major disadvantages in pixels that already are non-uniform due to n-well sharing.

3.1.3 Microlenses

Until this section, all performance improvement techniques that were presented were electrical design choices that exploited IC layout techniques. An alternative solution is to install optical devices called microlenses on each photodiode to collect incoming photons and to direct them to the photosensitive area. The advantages of microlenses are their capabilities to multiply the effective fill factor while not compromising from functionality nor pixel uniformity.

While having various advantages with no major drawbacks, microlenses can be installed on a sensor only if the pixels are physically compatible with microlens placement. The most important requirement is the minimum angle between signal lines and SPAD active area edges. The angle size is defined by two parameters: the height and the horizontal distance of the closest metal to active area.

An example of a SPAD image sensor with microlenses is a chip called SwissSPAD [4]. This chip, the predecessor of the 512×512 pixel array chip in chapter 2, was designed with an intention of optical fill factor enhancement. Microlenses with median concentration factor of 6 increase the fill factor from 5 % to 30 %. The choice of an optical solution provides more space for electronics and wires; thus allowing better overall chip performance and miniaturization.

3.2 Technology Node Adaptations

In device optimization, the role of the process technology is of key importance. Over the years, newer process technologies with smaller channel lengths are being developed. This fact raises the expectations for a trend in fill factor increase thanks to miniaturization of pixel electronics. However, there are several factors that complicate the adaptation of SPAD pixels in smaller CMOS technologies.

Since the introduction of pixels with integrated SPADs in a single die, it was repeatedly proven that monolithic pixels offer superior performance to SPADs built in custom processes. A monolithic approach requires the photodiode and the electronics to be designed in the same process technology. Therefore, technology node choice of these sensors has been restricted by the lack of low-noise SPADs in the latest deep sub-micron CMOS technologies [48]. An additional problem arises from SPAD operating voltage requirements. As discussed in previous chapters, the excess bias voltage for optimal SPAD performance is usually higher than the typical transistor operating voltages. For instance, the p-i-n diode used in one of the chips in in this project works best at 11 V as shown in Figure 2.2b. Due to a decrease in CMOS operating voltage with decreasing channel length, the gap between the SPAD performance allowed by the technology and the optimum performance becomes wider. A solution to partially overcome this difficulty is to place thick-oxide transistors that can operate at a higher voltage to the SPAD terminals. While allowing higher excess bias than the standard transistors, the presence of thick-oxide transistors in a pixel restricts the level of miniaturization due to their higher minimum channel lengths. Furthermore, in some pixel architectures, particularly the ones with gating and active recharge, a significant percentage of transistors have to be replaced with thick-oxide versions since they have contact with SPAD terminals. Due to these factors, the trend in pixel miniaturization is not progressing as fast as CMOS technology nodes.

In the near future, the development of 3D multi-wafer stacking technology can eliminate the aforementioned restrictions on pixel miniaturization by vertically integrating multiple dies. In 3D technology, the top tier can be designated only to the photodiode, and the lower tiers can be formed by electronics. Furthermore, the possibility of stacking dies fabricated in different processes allows the designer to choose a low-noise, more mature process for the SPAD and a more advanced process for the electronics.

3.3 A 4×4 Time-Gated SPAD-Based Image Sensor in 110 nm CMOS Technology

In this work, a 4×4 pixel array architecture was designed in 110 nm CMOS process technology. 2 variants of the chip layout were implemented with different SPAD modules. The first chip employs a more conventional, round-shaped SPAD with identical, uniform pixel structures. The second chip targets higher fill factor with the same level of functionality. In the second chip, the following techniques were applied to achieve that: Firstly, the round SPAD was replaced by a square SPAD with round corners. Secondly, 2 adjacent pixels on x and y-axes shared the signal wires, which reduced the number of wires per pixel. However, it should be noted that the second technique was suitable for this particular pixel, only because the area under the wires was sufficient for all transistors in a pixel. The specifications of both pixel versions are listed in Table 3.1. For both versions, the total chip dimensions including the pad ring are 874×874 μm (see Appendix A).

The main goal of this chip is to test the new pixel architecture, rather than to produce real images. The small size of the array makes the circuit less vulnerable to failures caused by circuit complexity. For instance, due to lack of high load capacitances,

Table 3.1: Specifications of pixels designed in 110 nm CMOS technology

	Pixel A	Pixel B
SPAD shape	round	square with round corners
Guard ring type	p-well	retrograde n-well (virtual)
Active area radius (μm)	2	2
Guard ring width (μm)	1.4	1.4
Cathode width (μm)	0.6	0.6
SPAD radius (μm)	4	4
Pixel pitch (μm)	9.8	9.1
Pixel fill factor	13 %	18 %

drivability of all input and output signals can be ensured. Therefore, small circuit size also improves the chance for more risky designs to work.

This chip consists of a 4×4 SPAD-based pixel array. The pixel architecture was comprehensively discussed in chapter 2. In this section, the emphasis will be on pixel layouts and the readout configuration.

3.3.1 Pixel Design

In this chip, two pixel variants with identical schematics (presented in Figure 2.9) are available. Pixel A, whose layout is shown in Figure 3.2, was designed using more conventional, low-risk methods, such as a round SPAD and a pixel with dedicated resources without sharing. This kind of an implementation protects the chip from various undesired effects. For instance, the lack of sharp cornered junctions in the SPAD prevents the formation of unwanted high electric field zones that may lead to premature edge breakdown (PEB). The use of p-well guard ring also strongly contributes to PEB prevention (Figure 2.8a). Furthermore, an independent pixel design isolates the internal signals of each pixel from each other; thus reducing the risk of crosstalk. On the other hand, these two methods limit the fill factor to only 13 %.

Pixel B, illustrated by Figure 3.3, aims to improve performance by using techniques that also increase the device failure chance. Firstly, the round SPAD was replaced by a square SPAD with rounded corners. For the same pixel size, a square shape increases the fill factor by 27 % compared to a round shape. Secondly, the choice of a virtual guard ring structure eliminates the minimum spacing requirement of 1 μm between the SPAD p-well and transistor p-wells, thus allowing the placement of NMOS transistors closer to the SPAD. Thirdly, Pixel B is indeed a 4-pixel unit with shared x and y-axes signal and power lines, as explained in detail in previous sections. The joint contribution of all three factors listed above creates a pixel with a fill factor of 18 % and pixel pitch of 9.1 μm. This corresponds to an 34 % increase in fill factor and a 7 % decrease in pixel pitch.

While offering significant pixel density improvement, the techniques listed above have their drawbacks, too. For instance, a SPAD with a perfect square shaped active area has a high risk of suffering from PEB. A solution that offers the optimal point in the trade-off between PEB risk and reduced fill factor is an square active area with

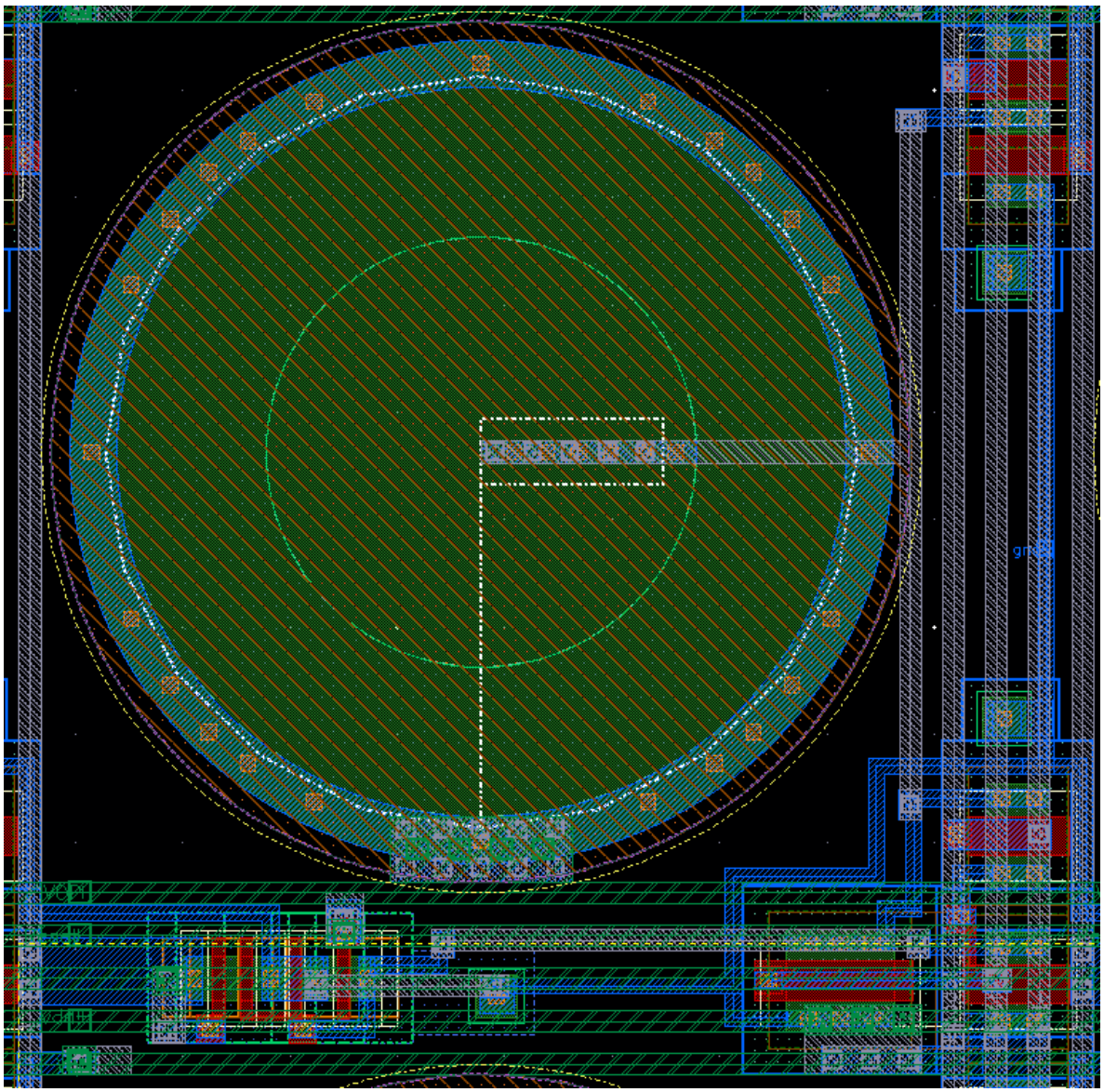


Figure 3.2: Layout view of Pixel A

rounded corners. As illustrated in Figure 3.3, the active area of the Pixel B SPAD has a form of a square whose corners are cropped by a quarter circle with a radius that is 25 % of the square side length. This shape offers an area increase of 20 % compared to a round shape, as opposed to 27 % for a complete square.

The process technology also had a major impact on the pixel characteristics. These chips were designed in 110 nm high voltage CMOS technology. The availability of both thin-oxide and thick-oxide transistors allows a maximum excess bias of 3.3 V. However, the use of 2 transistor types introduces 2 complications. Firstly, due to their significantly higher minimum channel lengths (360 nm vs. 110 nm), the thick-oxide transistors occupy more space, thus partially eliminating the miniaturization-related advantage of the 110 nm channel length. The second drawback is that thick and

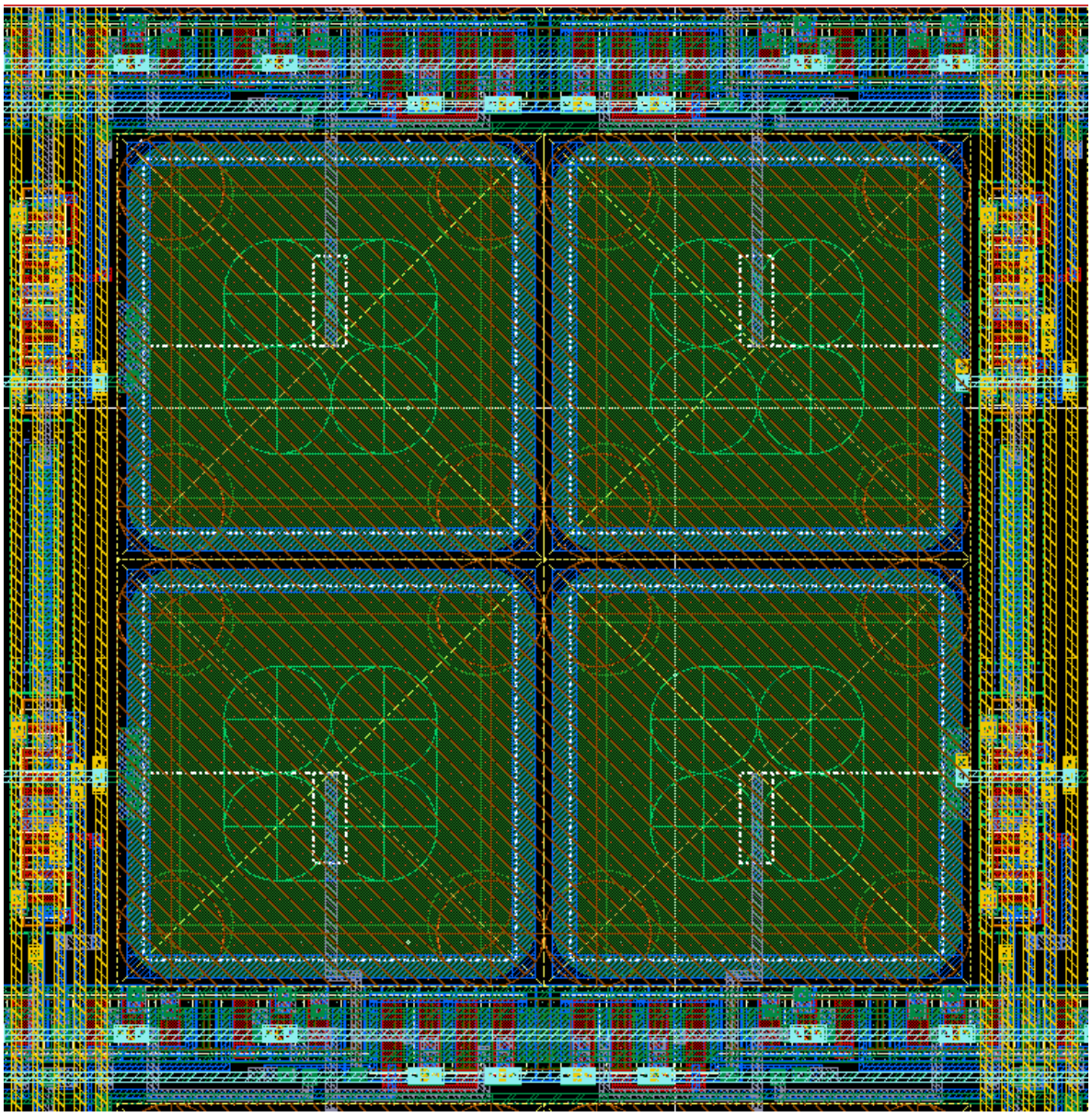


Figure 3.3: Layout view of Pixel B

thin oxide NMOS transistors contain different kinds of p-well layers. That introduces a requirement to place the thin and thick oxide transistor groups with a minimum spacing of 1 μm. This requirement leads to the creation of a dead space which occupies a considerable part of the pixel area. Several other technologies include thick and thin-oxide transistor cells that do not have to be separated due to p-well minimum spacing rule. The third problem related to having 2 NMOS types in a pixel is the difficulty of signal routing. As a result of the grouping requirements of thick and thin oxide transistors, the distance of any two transistors in the pixel layout is not necessarily based on the order in the schematic view. This leads to even more dead area, as well as undesirably long connection wires between two neighboring transistors. Although

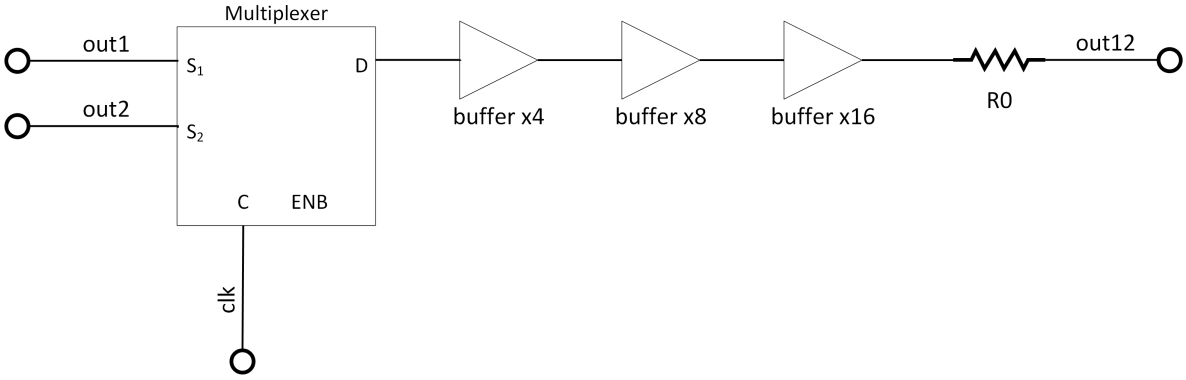


Figure 3.4: Schematic view of the readout block for the 4×4 array in 110 nm

these problems do not hinder the functionality of the pixels, they may reduce the performance particularly at high frequencies, where minimum signal propagation delay and maximum drivability is of key importance.

3.3.2 Readout and Pad Ring

As stated in the overview paragraphs, this chip was designed to test the operation and the performance parameters of the pixel architecture. Consequently, to minimize the risk of fundamental errors, the electronics outside the pixel array were kept as simple as possible.

In the 4×4 pixel array, each column has its own output bus, shared by 4 pixels in a column. Triggered by the signals *Rowsel* 1-4, only one row must be in the readout mode at any given time. During readout, if a pixel stores high voltage in its memory, it pulls down the output bus voltage from V_{dd} to 0 V. After readout, the column voltages are restored to V_{dd} through PMOS transistors that are permanently on. These PMOS transistors were sized carefully in such a way that their pull-up strengths are weaker than the pull-down strength of a pixel. The voltages of 4 output columns are transferred to the output pads through a chain of 3 buffers with increasing driving power along the chain. The best performance could be achieved by allocating an output pad for each output column bus. Due to overall chip area limitations, an output pad was shared by 2 column buses via a 2-to-1 multiplexer. Even though these multiplexers reduce the maximum achievable frame rate of the chip, they do not affect the maximum readout speed of a single row. By keeping the select signal at fixed voltage, the multiplexer can be by-passed during the measurement of a pixel. Finally, a resistor was placed after the buffer chain to strengthen ESD protection. The schematic view of the output configuration can be viewed in Figure 3.4.

Out of 4 row select signals *Rowsel*, only 1 is permitted to be high voltage. This requirement could be achieved by either on-chip or off-chip electronics. As a principle, on-chip electronics were kept as few and as simple as possible. Therefore, an input pad was dedicated to each row select signal to permit the transfer of these 4 signals independently. To generate an image, each row can be selected in a sequence by means of 4 *Rowsel* signals with shifted rise/falltimes. Meanwhile, the select signal of the mul-

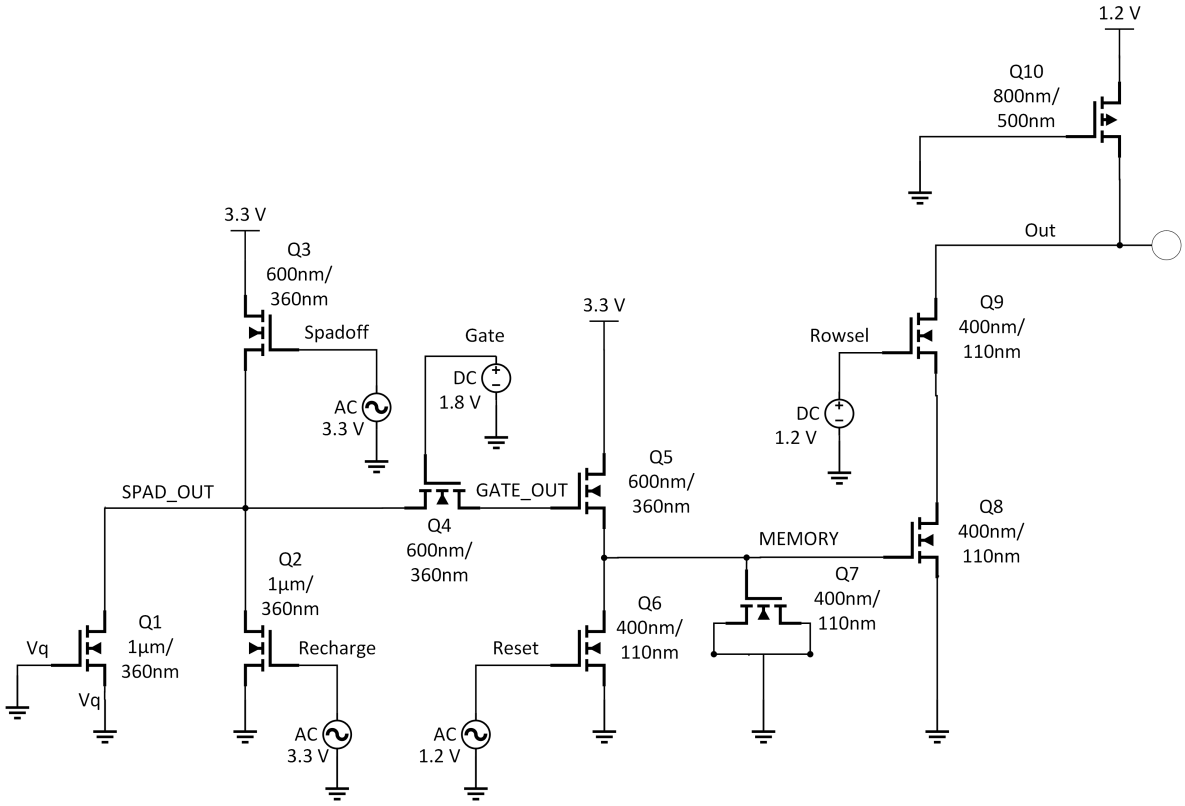


Figure 3.5: Experimental setup used for the characterization of the chip

tiplexer must toggle at the same speed with the readout to always select the currently active row. After reading all 4 rows, all in-pixel memories must be reset using *Reset* signal. Due to area constraints, all 4 rows are connected to the same *Reset* signal. That requires the row selection operation to stop temporarily during the reset process.

During the design of this chip, many design factors that could potentially hinder performance in large-scale arrays were not considered. Wire sizing to minimize IR-drop, adding buffer chains to enhance input signal drivability or adding decoupling capacitors to maintain stable voltages in the core were among these factors. In chapter 4, a larger-scale single-photon, time-resolved image sensor with a 512×512 pixel array is presented. That chip employs a more advanced version of the readout configuration presented here. Therefore, more in-depth performance analyses are available in chapter 4.

In this chip, no extra circuit blocks were placed to improve input signal drivability for the pixel array.

3.3.3 Measurement Results

In this subsection, the functionality of the pixel excluding the SPAD is presented. Figure 3.6 contains several plots that demonstrate the basic operation of different pixel components. The measurement setup is illustrated in Figure 3.5. In all plots, the gating window is permanently open with 1.8 V DC voltage, and all signals that are not displayed are grounded. Figure 3.6a tests the *Reset* signal which discharges the dynamic

memory. When *Reset* rises, the output bus is pulled up immediately. The behavior of the dynamic memory is displayed in Figure 3.6b. When *Recharge* rises, it discharges *SPAD_OUT* to 0 V, thus drives Q5 into cut-off mode. From that moment, the voltage of *MEMORY* gradually decreases. In the figure, the output rises approximately 200 μs after the rising edge of *Recharge*, which also defines the maximum storage duration of the memory. Output bus pull-up and pull-down durations are two parameters that determine the maximum readout speed in this chip. Figure 3.6c and Figure 3.6d plot the response of the output bus during pull-up and pull-down. According to these two figures, the pull-up and pull-down duration of a pixel were measured as 18.2 ns and 27.1 ns, respectively.

The results obtained from simulations and measurements are compared in Table 3.2. The simulation configuration does not include the parasitic capacitances and resistances. A significant drop in pull-up and pull-down performance was observed in the measurements compared to the simulation results. The memory storage duration, on the other hand, is 67 % higher in the measurement results than the simulation results.

Table 3.2: Performance parameters extracted from the simulations and measurements

	Simulation	Measurement
Output bus pull-up duration	4.8 ns	18.2 ns
Output bus pull-down duration	6.18 ns	27.1 ns
Dynamic memory voltage storage duration	120 μs	200 μs

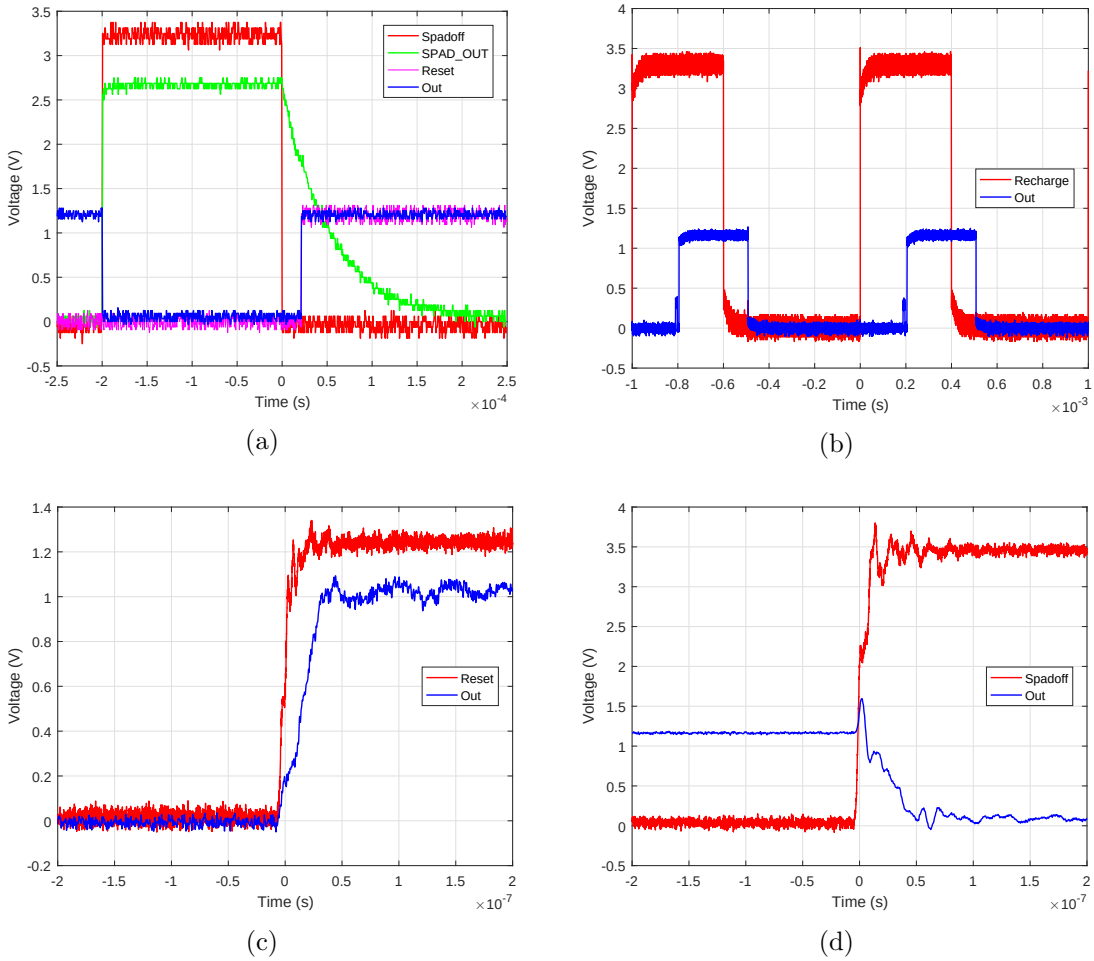


Figure 3.6: Measurement results of Pixel B: (a) Reset signal discharges the memory and the output bus is pulled up (b) Dynamic memory stores the high voltage after its connection with the voltage supply is cut off, (c) Output bus is pulled up by PMOS transistor Q10 after the pixel stops pulling down (d) Output bus is pulled down after dynamic memory is charged

4 Two Chip Variants

In this project, two chips, formed by large pixel arrays, were designed in 0.18 μm CMOS process. Both chips are composed of p-i-n diode-based SPADs; however, the pixel sizes and structures around the SPADs are different.

In this chapter, architectures and implementations of both circuits are described in detail. In section 4.1, a 512×1 pixel array built with dedicated output pads and event-driven readout scheme is presented. This linear sensor contains minimal on-chip functionality, and is highly dependent on an FPGA for its operation. Contrary to hard-wired electronics implemented in silicon, an FPGA allows the user to reconfigure the entire readout and processing circuitry. In addition, the assignment of an output pad for one pixel allows totally independent controllability of each pixel. In section 4.2, a 512×512 time-gated image sensor is analyzed. This sensor aims to generate high-resolution images by capturing photon ToA with its time-gating mechanism. In this chip, the pixels are designed with more functionality, at the expense of reduced fill factor. The readout electronics are placed inside the chip, which limits the flexibility of its operation settings.

4.1 A 512×1 Event-Driven SPAD-Based Line Sensor

A SPAD line sensor with 512 columns and a single row was designed to test the maximum SPAD performance in a pixel array. The sensor has a round p-i-n SPAD with an active area radius of 7.4 μm. The dimensions of the chip are 14.3×1 mm (see Appendix A). It operates based on an event-driven readout configuration. This linear array is suitable for testing the highest performance of the p-i-n based SPAD, thanks to its two features: high excess bias and direct connection between output pads and pixels. Compared to the the linear sensor in [49], it contains higher number of pixels and achieves improved time resolution. In the previous chapters, the pixel features of this chip were explained in detail. In this chapter, the focus will be on implementation techniques and performance characterization.

4.1.1 Chip Architecture

The smallest cell of the array is a unit formed by two adjacent pixels with opposite vertical orientations (Figure 4.1). This unit contains two round SPADs and pixels around them. The pixel, whose architecture is shown in Figure 2.10, consists of a poly resistor, an AC coupling capacitor, a diode, a control transistor and an inverter. The operation sequence of the pixel is as follows:

When the SPAD fires upon photon arrival, its anode voltage instantly rises from 0 V to the excess bias voltage, maximum 12 V. Since these voltage levels cause damage

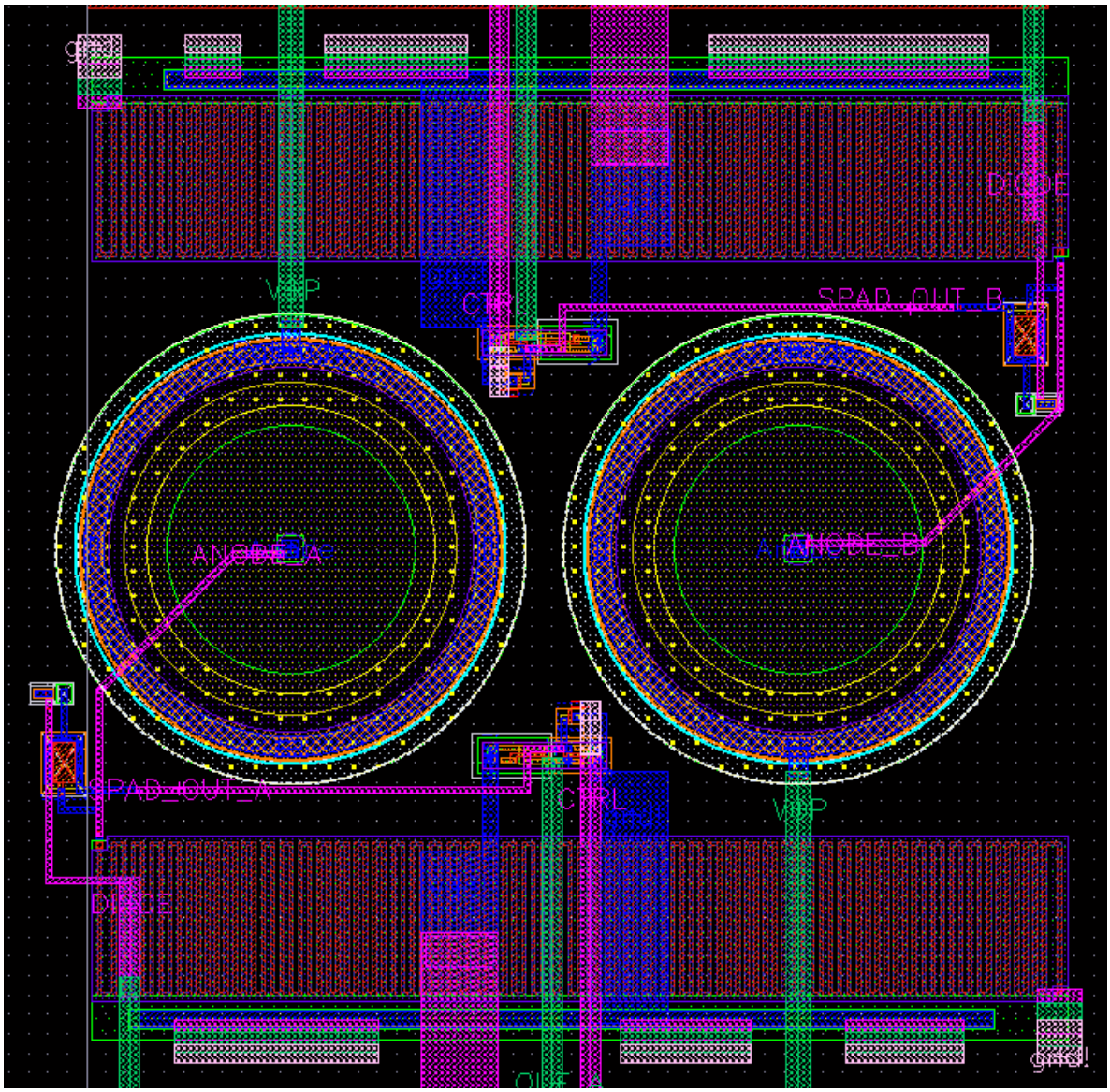


Figure 4.1: Layout view of a pixel in the 512×1 image sensor

to CMOS transistors, the *SPAD_OUT* node, which is connected to CMOS transistors, cannot undergo a voltage swing of 12 V. The AC coupling capacitor isolates the DC voltages of its 2 terminals and only permits the transfer of fast voltage transitions. Since the voltage transition in *ANODE* is very fast during an avalanche, this capacitance can be insufficient to keep the voltage level of *SPAD_OUT* below 1.8 V at all times. The diode, whose cathode is connected to a DC voltage adjustable from outside the chip, is added to the pixel to clamp the maximum voltage of *SPAD_OUT* at 1.8 V by switching to conducting mode if the voltage exceeds a certain threshold. A forward-biased diode starts conducting only if the voltage across its terminals is greater than a fixed value. This configuration exploits that feature of a diode. The NMOS transistor, biased by the signal *CTRL*, allows the adjustment of the SPAD dead time. Since the pixel is

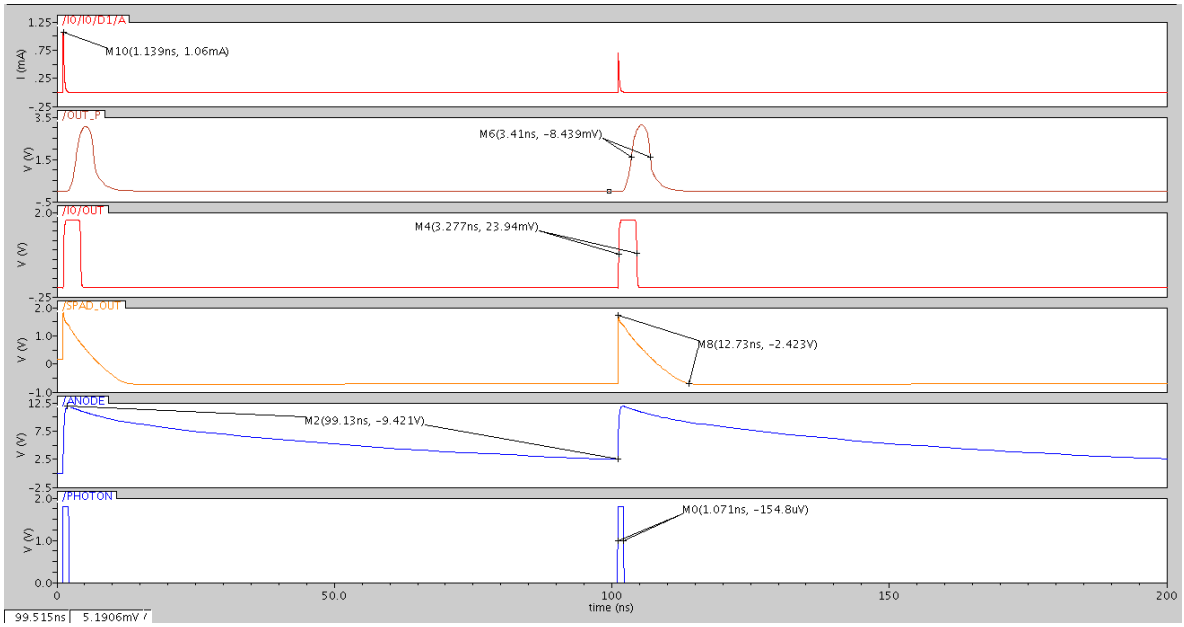


Figure 4.2: Simulation results of chip operation at 12 V excess bias

designed to operate in a wide range of excess bias voltages (3-12 V), the achievement of the desired dead time can only be possible with adjustable resistors. The effects of pixel dead time on performance were listed in previous chapters. Finally, the shape of the pulse is corrected by an inverter and the signal is sent to the output pads.

4.1.2 Performance Characterization

Compared to 2D arrays, a linear sensor is more suitable for higher performance, as previously explained. Therefore, a meaningful comparison for the performance evaluation of this chip would be with another linear sensor in the literature. The work [49] describes a 256×1 SPAD-based image sensor designed in 0.35 μm CMOS technology. In this work, thanks to the use of staggered pad arrays, the wire lengths between pixels and output pads are more uniform across the array, and shorter in average. In the 256×1 chip, due to higher pad pitch, some of the pads are placed away from the edges, which caused a larger overall chip area and non-uniformity. On the other hand, the fill factor of this work (25 %) is significantly less than the 256×1 chip (40 %). This difference is due to 2 factors. Firstly, a round SPAD leads to lower fill factor than a rectangular SPAD with round corners. Secondly, the presence of deep trench isolation (DTI) layers in our chip requires large minimum spacing requirements, which determines the minimum pixel pitch regardless of the size of pixel electronics.

The typical operation of this sensor is shown in Figure 4.2 for 12 V excess bias. OUT_P, a node that is not shown in the pixel schematic, is the output pad voltage with a load capacitance of 10 pF. It is worth noting that this value, which has a strong impact on the pad output signal risetime/falltime, is a conservative empirical estimation of the load. The characterization of the SPAD was presented in chapter 2, based on the measurement results in the work [2]. The most significant performance

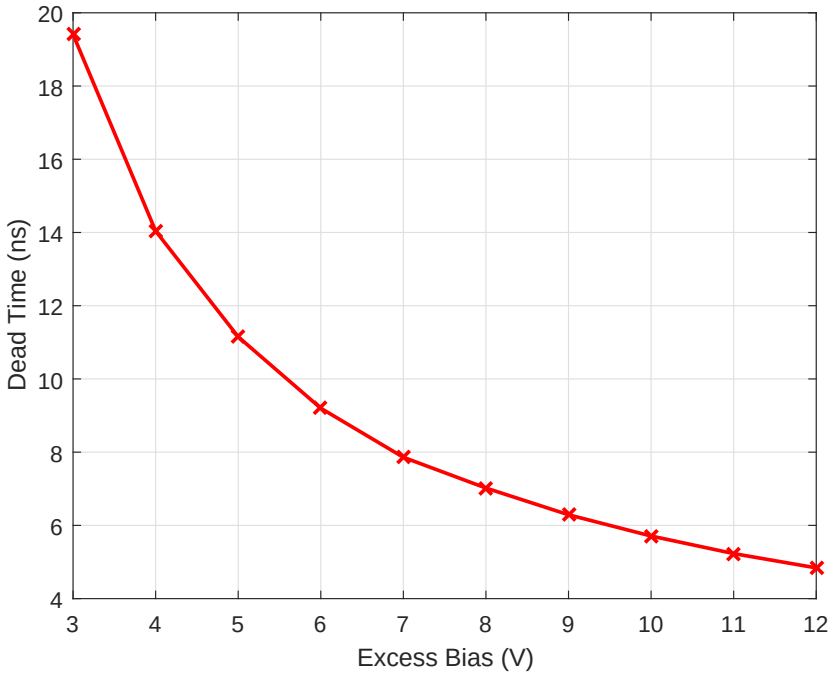


Figure 4.3: SPAD dead time of the 512×1 pixel array at various excess bias voltages

parameter that can be extracted from the simulations is the SPAD dead time. In this pixel, dead time is defined as the time between the start of an avalanche and the point where the SPAD is restored to its photosensitive state. This definition is open to interpretation; because the minimum PDP level to consider the SPAD photosensitive is unknown. Another factor that complicates this ambiguity is the lack of PDP data below 1 V excess bias in Figure 2.2b. In this report, the excess bias of 1 V was accepted as the point where SPAD gains its photosensitivity back, with a PDP of 17 %. With the aforementioned parameters, the SPAD dead time as a function of excess bias voltage is shown in Figure 4.3. The dominant factors that determine the dead time are the size of R0 and the capacitance of the SPAD junction. The size of R0 is 600 $k\Omega$, the minimum resistance that limits the avalanche current to a safe level. The SPAD junction capacitance was taken as 70 fF based on previous models of the device. However, the limited accuracy of SPAD models in IC simulation tools contribute to the uncertainty of the result.

4.2 A 512×512 Time-Gated SPAD-Based Image Sensor

The chip with the largest array designed in this project is the 512×512 time-gated SPAD image sensor. This device achieves a fill factor of 13 %, image frame rate of 200 kfps and gate edge uniformity of 150 ps across the entire array. Designed in 0.18 μm CMOS technology, the chip is formed by p-i-n SPADs and time-gated pixels. The SPADs can operate at excess bias voltages up to 6.6 V, above the capacity of CMOS transistors. The 512×512 array has a pixel pitch of 16.38 μm and the total size of the chip is 9.5×9.6 mm (see Appendix A).

Table 4.1: Specifications of the 512×1 image sensor designed in 0.18 μm CMOS technology

Readout configuration	Event-driven
Array size (column×row)	512×1
Process technology	0.18 μm CMOS
Pixel pitch (μm)	26.2
SPAD active area radius (μm)	7.4
P-epi layer width (μm)	2.0
Cathode (n-well) width (μm)	1.5
Maximum SPAD excess bias (V)	12
Fill factor	25 %

This sensor can be used in a variety of applications. Thanks to its large pixel array and fast readout, it is best suited for 3D ranging, such as automotive safety systems in low-light-levels, space based surveying or automated landing/docking. Its simulated gate edge precision of 20 ps achieves 3 mm spatial resolution, a sufficiently low number for this application. Nevertheless, the chip has two drawbacks that restrict its low-light-level performance: firstly, it misses all photons that impinge when the gate is off. Secondly, its low fill factor limits its photon sensitivity. The latter problem can be partially resolved by microlenses. The ability of the sensor to operate in global-shutter mode allows it to be used for TCSPC applications, such as FLIM. In addition to the aforementioned challenges, the global shutter mode reduces the photon efficiency of the sensor even more, due to the required time to process the entire array before a new exposure.

Table 4.2: Specifications of the 512×512 image sensor designed in 0.18 μm CMOS technology

	Pixel v1	Pixel v2
Fill factor	10.5 %	13 %
Readout configuration	Time-gated	
Array size (column×row)	512×512	
Process technology	0.18 μm CMOS	
Pixel pitch (μm)	16.38	
SPAD active area radius (μm)	3	
P-epi layer width (μm)	1.5	
Cathode (n-well) width (μm)	1.5	
Maximum SPAD excess bias (V)	6.6	
Timing resolution (ps)	20	
Image frame rate (kfps)	200	
Power consumption (W)	3.52	

4.2.1 Chip Architecture

The SPAD and the pixel used in the sensor were comprehensively analyzed in chapter 2. In this chapter, the composition of the entire device is discussed. The sensor detects the photon time-of-arrival (ToA) by a gated system. The in-pixel memory stores photon arrival information only if the SPAD fires inside the gating window. Photon ToA is recorded by generating the image of the same frame multiple times while shifting the gate window by multiples of 20 ps. For 3D applications, the image spatial resolution, which is a function of the minimum gate window shift, can be calculated from Equation 4.1

$$d = c \times t, \tag{4.1}$$

where d is the total distance, c is the speed of light and t is the elapsed time.

Considering that the speed of light is approximately 3×10^8 m/s and the parameter d is twice as large as the distance between the sensor and the object, the sensor can resolve distances as small as 3 mm.

The operation of the chip is controlled by 3 global high-voltage and 2 local low-voltage signals. The global signals determine the gate window length and deactivate the photodiode outside the gating window. The local signals read out each row in sequence and reset the in-pixel memory in every pixel in a row shortly after the row is read out. Due to the lack of sufficient area and power consumption limitations, each output pad is shared by 4 columns. This feature reduces the maximum frame-rate, limited by the I/O blocks and clock frequency, by a factor of 4.

The chip consists of two identical rectangular blocks of 512×256 pixels. One long side of this block is not filled with pads, which makes it possible to merge 2 identical blocks on this side. Even though the 512×512 integrated pixel array has totally uniform pixel spacing, the electronics of each half are independent from each other. All analyses related to this chip were performed on a single 512×256 block.

In addition to the main array and its supporting electronics, this chip includes several independent blocks that facilitate the testability of the circuit. The positions of these 3 structures in the chip layout are presented in Appendix B. Firstly, an independent SPAD cell was placed on the substrate (Figure B.1). This cell, controlled by its dedicated anode and cathode pads, can be used to measure SPAD parameters such as breakdown voltage, PDP and DCR. Secondly, 2 isolated pixels were added near the corners of the large array. These pixels, shown in Figure B.2, receive their dedicated input signals directly from the pads. This permits the testing of pixel operation without the risk factors caused by the large electronics. In addition to the pixel output, output pads were assigned to 2 in-pixel nodes (*MEMORY* and *SPAD_OUT*), as well. The observability of these 2 nodes allow the testing of each component of the pixel; thus making the diagnosis of a potential pixel-related error more probable. The only variation between the configurations of these 2 pixels is related to *Rowsel* and *Reset* signals. In one version, both signals are directly sent from the pads. In the second version, on the other hand, these 2 signals are sent to the pixel through a row driver block. The purpose of this setup is to test the functionality of the critical *Reset* signal which is derived from *Rowsel* in the row driver. The third testing setup aims to observe

the 3 global 3.3 V signals (*Gate*, *Recharge* and *Spadoff*) at the input of the farthest pixel from the signal source in a column (Figure B.3). By observing the signal shapes at that position, the capability of the signal distribution network and column signal wires can be analyzed.

4.2.2 Row Driver Block

In this chip, each row has a dedicated driver block, since the local signals must reach each row at different time intervals. As shown in Figure 4.6, a row driver block consists of an 8-bit decoder and a reset generator. The decoders interpret the 8-bit binary signal that the chip receives from an external counter. Each decoder returns high output for a different value of the 8-bit code. This system ensures that only one of 256 pixels in a column is pulling down the bus at a time. Reset generators produce the reset signal from the falling edge of the row select signal. Since reset is not provided by the FPGA as an independent signal, reset and row select signals are always synchronized.

In addition to discharging the dynamic memory, the reset signal also discharges the gate of Q5 to avoid unintentional photon counting outside the gating window. Therefore, the gate pulse width must be short enough to prevent permanent damage to the SPAD due to exposure to high current. As shown in Figure 4.7, *reset_e* signal can shorten the reset pulse width.

During the implementation stage of the row driver block, several design challenges had to be overcome in order to achieve a variety of performance requirements. The first task was to reach sufficient signal drivability to the load of an entire row. Row select and reset signals are sent to 512 pixels from the left side of the horizontal row signal wires. The signals had to be provided to all pixels with a small amount of skew, to ensure synchronous readout across the row. In addition, the rise and fall times of the signals had to be below a certain level, to avoid pulse shrinking and jitter. Considering the high load capacitance of the pixels and high RC constant of the signal wire, the row driver block had to have high drivability.

Another challenge was related to the reset signal. In this chip, the reset signal is generated inside the row driver block, triggered by the falling edge of the row select signal. The reset pulse width is a very critical parameter in the circuit: a too short reset may be insufficient to discharge the dynamic memory, failing to reset the states of the memory blocks across the array. It should be noted that the most difficult task is to reset the rightmost pixel in a row, as the signals are sent from the left side. On the other hand, a too long reset can make the pixel photon insensitive for a considerable amount of time, potentially leading to missed photons. Additionally, if the SPAD fires in the reset window, the quenching mechanism is ineffective due to a low resistive path to the ground. A too long reset signal can potentially harm the SPAD due to long exposure to high current. Taking into account all these factors, an adjustable reset mechanism was designed to fulfill all requirements.

4.2.3 Column Signal Distribution Network

A key performance parameter of a time-gated image sensor is gate uniformity. In order to ensure that the incoming gate signal reaches all pixels in a column simultaneously,

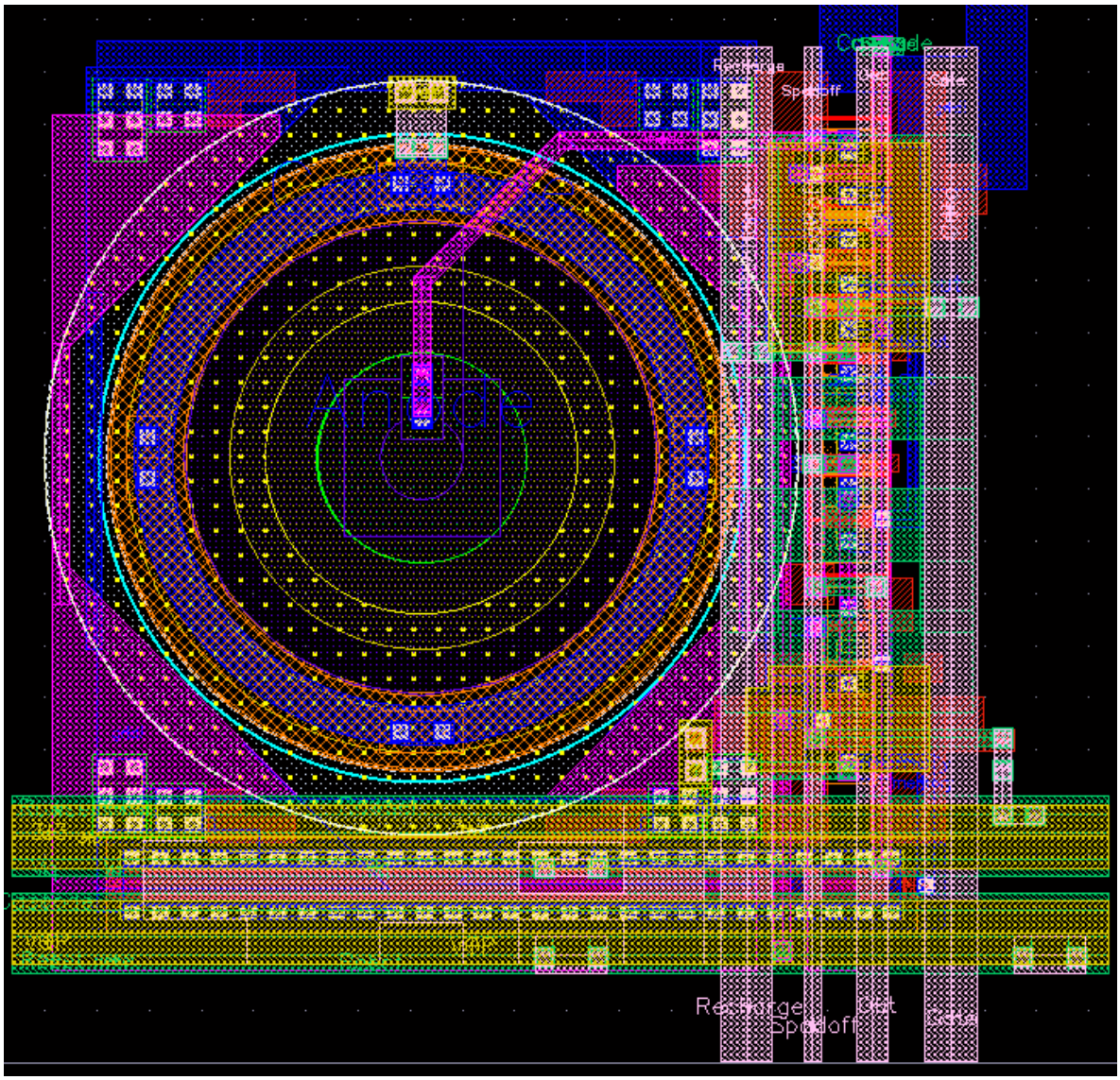


Figure 4.4: Layout view of Pixel A in 512×512 image sensor

a signal distribution tree block was placed between the global signal sources and pixel array columns. The 512-to-1 signal tree distributes the gate signal to 512 branches. The layout of the tree was designed in such a way that each branch has the same wire length and same driving strength.

Due to its strong impact on circuit performance, a signal tree is a special component in the sensor. From architecture to implementation, every design stage of these blocks was thoroughly calculated. The schematic view of a tree is presented in Figure 4.8. The circuit is formed by 5 stages of buffers. The buffers in the first 4 stages drive 4 or 8 other buffers, with a total load capacitance in the order of tens of fF. On the other hand, each last stage buffer drives a very high load formed by the equivalent capacitance of 256 pixels in a column as well as the signal wires. The equivalent load capacitance of each tree output equals several pF. Consequently, 2 different buffer cells were used in

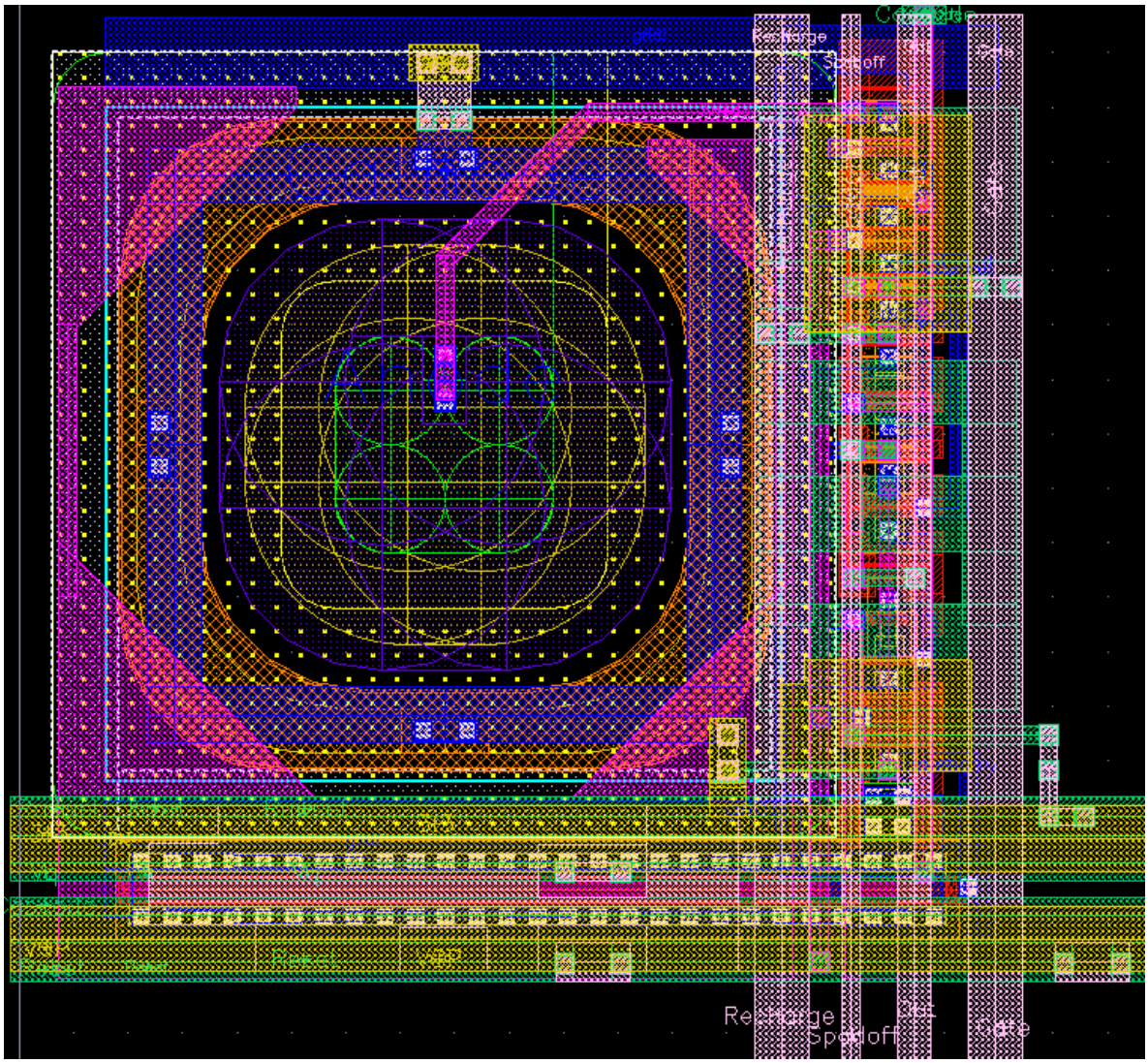


Figure 4.5: Layout view of Pixel B in 512×512 image sensor

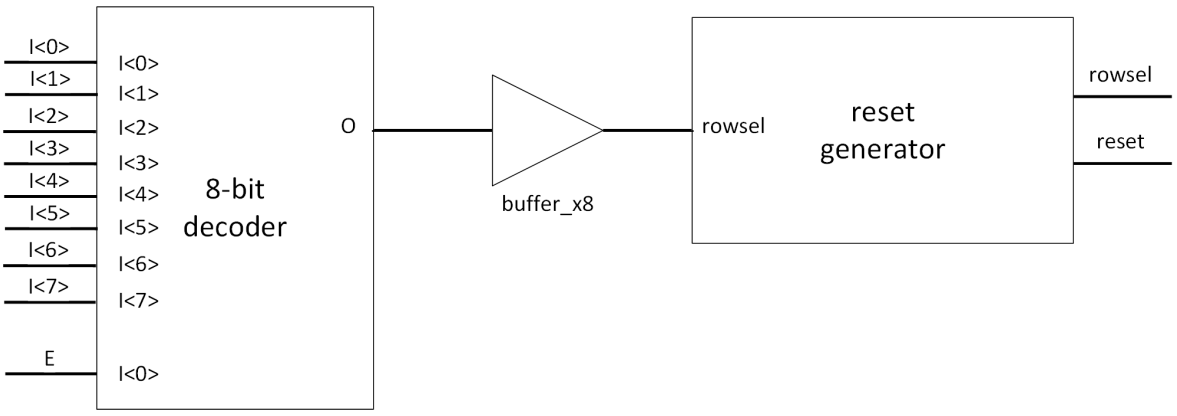


Figure 4.6: Block diagram of the row driver module

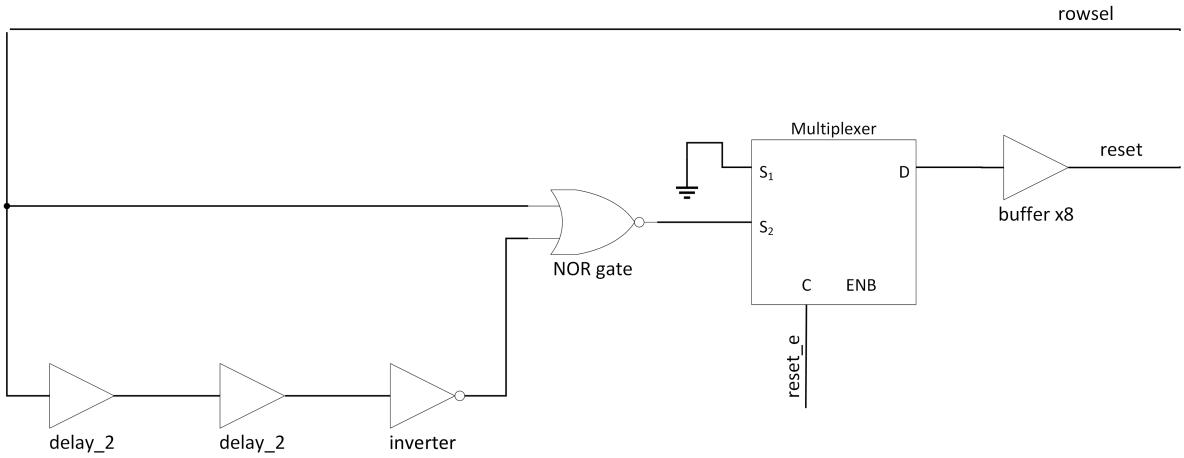


Figure 4.7: Block diagram of the reset generator module

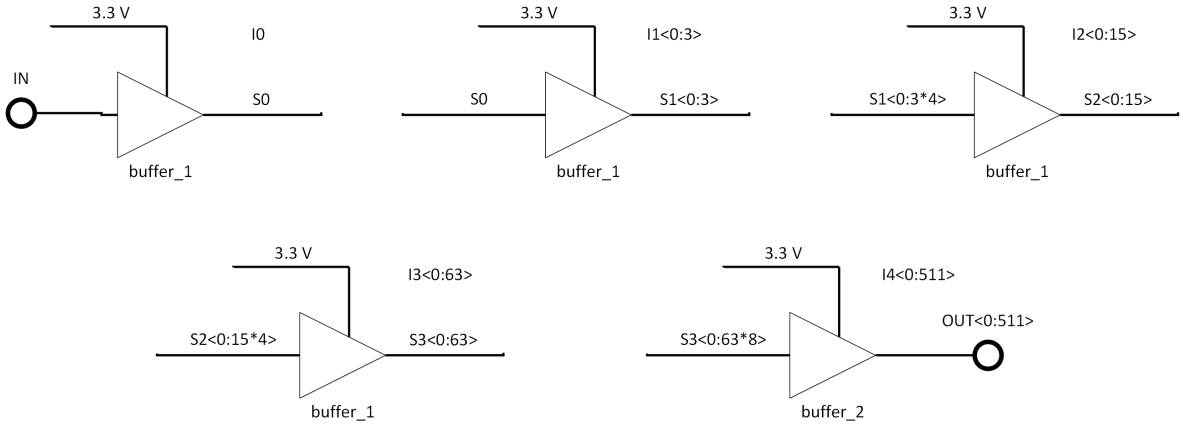


Figure 4.8: The schematic view of the column signal distribution network

the tree: the last stage buffers have considerably larger transistor size and drivability than the first four stages. Each buffer contains two inverters in series. While inverter 1 only drives inverter 2, inverter 2 drives the entire load of the buffer, whose capacitance is significantly higher. Therefore, second inverters have larger sizes than the first ones.

A major trade-off in the design of this structure was the choice between high performance and modularity. To achieve fastest signal propagation, the buffer in each stage must be sized in such a way that each stage will have equal contribution to the total delay of the combinational block. While being effective in boosting the performance, this design technique requires custom design of each block in the tree, which is a cumbersome method for a tree with 512 outputs. In this structure, fast design time was achieved using identical buffers in each stage except the last one. Furthermore, the number of branches in each stage was equalized to keep the load capacitance values at similar levels.

The performance of the tree is characterized by 2 parameters: skew and jitter. Although often used in a variety of meanings (sometimes even interchangeably) in

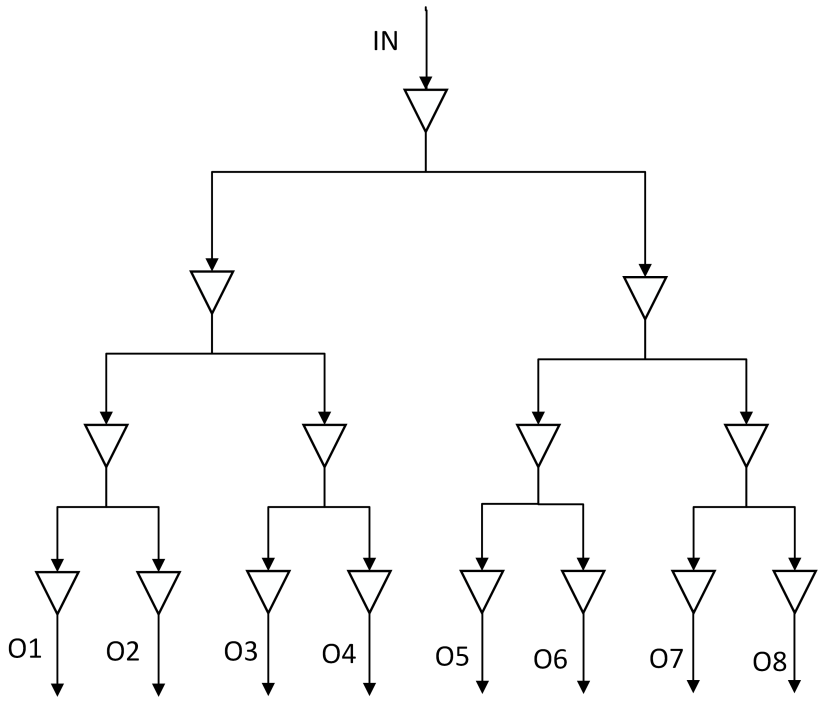


Figure 4.9: Diagram of a balanced clock tree

the literature, these two terms are independent. In this work, skew is defined as the maximum time difference between any 2 output of the block. Jitter indicates the time variation of a signal edge at the same output after a number of iterations. To minimize the skew, each route from the input to an output had to have identical parameters, particularly metal wire length and width. To achieve that, a clock tree distribution network model had to be chosen and adapted to the chip architecture. The architecture of this chip required all output pins to be placed along a single line with a pitch equal to the pixel pitch of the array. The topology which was most compatible with those requirements was called a balanced clock tree, represented in Figure 4.9. Jitter is caused by a variety of reasons, such as process variations, device mismatch and signal risetime/falltime. The first two factors are not dependent on IC design techniques; therefore a designer's focus must be on minimizing the third factor. Correct sizing of each stage to achieve a certain level of drivability minimizes the output signal risetime/falltime.

The waveforms of all levels of the tree are displayed in Figure 4.10. These waveforms were observed with a post-layout simulation setup comprising of the entire signal tree and the layout of one 256-pixel column. Several performance parameters can be inferred from this figure. The total propagation delay of the tree is 2.421 ns. This delay does not pose any threats to the overall chip performance, and can be compensated by the signal source outside the chip. The horizontal skew between 512 output channels of the tree is 12.45 ps. Finally, vertical skew, defined as the time variation of output signal rising/falling edge between the top and bottom pixels of a column, was observed as 109.9 ps. As a result of the delay of narrow and long signal wires and high total column load capacitance, vertical skew accounts for 90 % of the total skew.

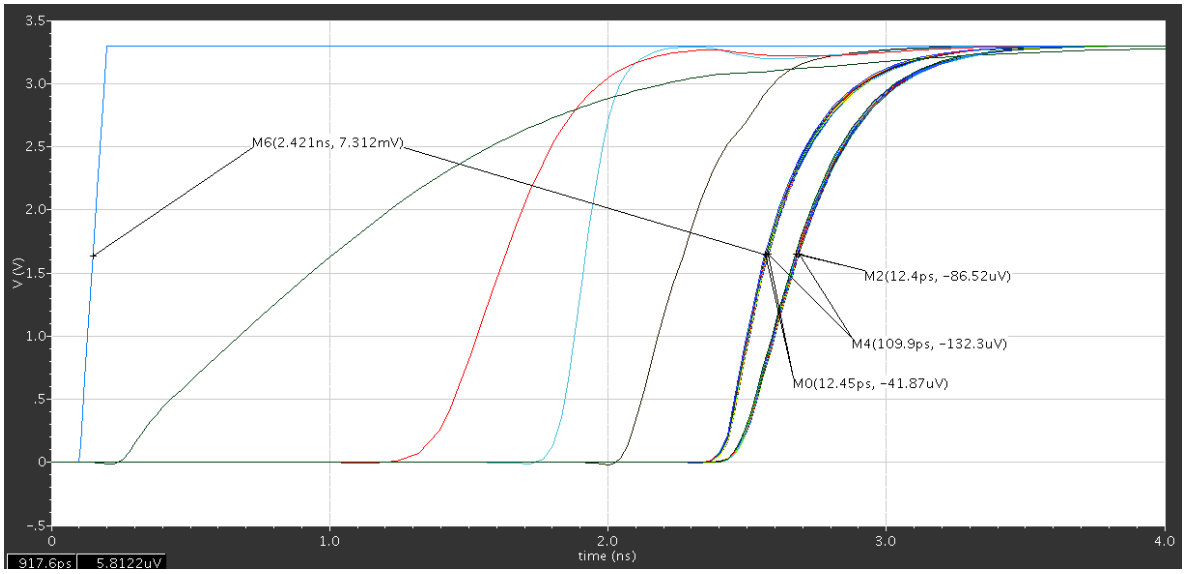


Figure 4.10: The waveform of the column signal distribution network

4.2.4 Readout Block

The design of the readout block for a large pixel array is a challenging task to achieve both high image frame-rate and low power consumption, while not exceeding the maximum area determined by the pixel array size. Since the pixel of this image sensor employs a 1-bit memory, a rolling-shutter based readout was used in this sensor.

Space constraints of the chip required the sharing of an output pad between multiple columns: the perimeter of the chip was not sufficient to place 512 output pads. Consequently, a readout topology with 128 pads and 4-to-1 multiplexers was chosen. The schematic view of the readout network is shown in Figure 4.11. Each output column bus was at first sampled by d-flip-flops, and then multiplexed by 4-to-1 multiplexers, whose outputs are connected to the pads through I/O blocks. The flip-flops improve the robustness of the readout circuit by securing the synchronous operation of each stage. In addition, a PMOS transistor was assigned to every column to keep the bus at high voltage in the idle state.

The readout network operates as follows: At any time 1 of 256 rows in a 512×256 array is in readout mode. This results from *Readout* signal being sent to each row with different phases from an 8-bit decoder. When a row is read out, one pixel in each column has a possibility of pulling down the output bus, depending on the current state of its in-pixel memory. During readout, PMOS pull-up transistors are in cut-off mode to facilitate the bus pull-down. The voltage of each output bus is sampled by a d-flip-flop at the end of the readout process. Right after sampling, the *Readout* signal falls and the all column voltages are pulled up by switching on the PMOS transistors. When the pull-up is complete, the next row switches to the readout mode and the same cycle continues. Meanwhile, a multiplexer in the next stage sends all 4 column data to the output pad within one row readout cycle. It is worth noting that this requires the multiplexer to operate 4 times faster than the previous stages, namely the d-flip-flop

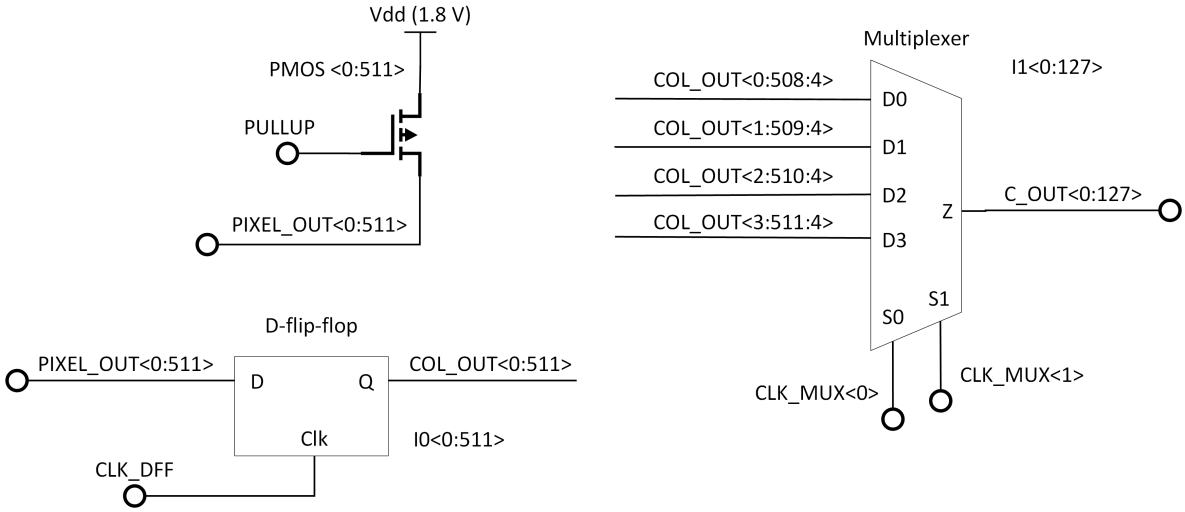


Figure 4.11: The schematic view of the readout circuit

and the 8-bit decoder.

An important stage of designing the readout topology was to derive the target readout parameters from the overall chip specifications. The important constraints to include in the computations were the following: a 10-bit grayscale output at 100 frames-per-second (fps) and maximum output pad data rate of 200 Mbps. Readout setup can be configured by adjusting the following 4 signals: switching frequency of the 8-bit counter in the FPGA that drives the decoder, clock speed of the d-flip-flops, clock speed of the PMOS pull-up transistors and 2-bit select signal of 4-to-1 multiplexers. The speed limit of readout results from the maximum data rate of the output pads, which was empirically measured to be 200 Mbps. According to the formula $time = frequency^{-1}$, this speed requires the multiplexers to switch input channels in every 5 ns. In contrast to sequential circuits such as flip-flops, multiplexers are controlled by level sensitive select signals. In this case, the LSB and MSB of the 2-bit select signal must receive a 100 MHz and 50 MHz clock signal with 50 % duty cycle, respectively. The d-flip-flops must sample the column bus voltage in every 20 ns. Since these devices are only rising edge-sensitive, they must receive 50 MHz clock signals. The 8-bit counter must switch rows at equally same speed as the sampling frequency of the flip-flops. Therefore, the 8-bit counter must increment in every 20 ns. Finally, the PMOS transistors must receive a signal that switches them on in every 20 ns. The duty cycle of this signal is determined by the duration of the pull-down process. All of the aforementioned signals can be provided by a standard FPGA.

4.2.5 IR-Drop and Decoupling Capacitors

In SPAD-based image sensors, certain design challenges become crucial as a result of scaling. In other words, some problems that can be overlooked in small pixel arrays can pose serious threats to circuit functionality in a 512×512 array. IR-drop in the supply lines is among the most critical of these problems; because it can potentially make the sensor totally unfunctional, if certain measures are not taken. The supply

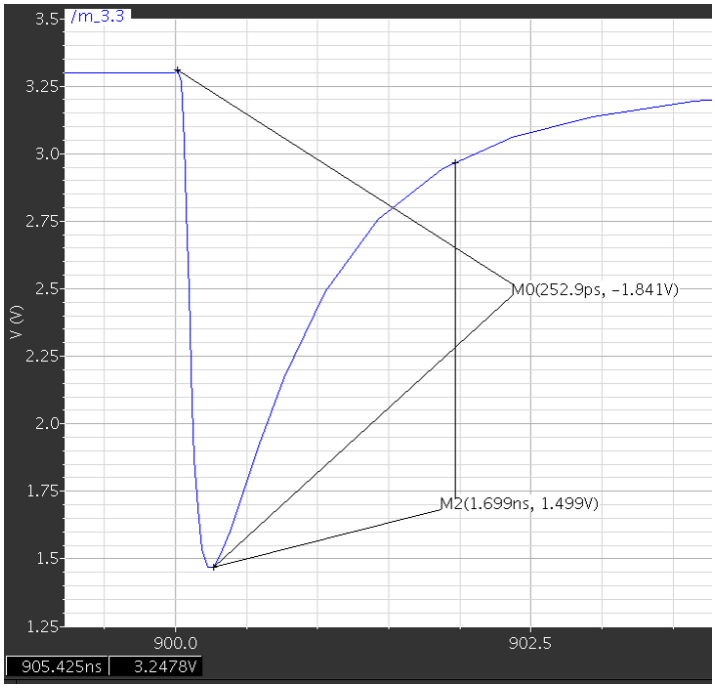


Figure 4.12: IR-drop of the 3.3 V supply bus at the central pixel

lines that provide power to the electronics in the entire chip can suffer from momentary voltage drops due to high current flow through the parasitic resistances of the metal. This phenomenon is called IR-drop.

In this chip, power was supplied to all electronic components outside the pixel array from close distances with wide metal wires. That reduces the equivalent resistance between the power source and the target. Furthermore, the placement of multiple supply and ground pads divides the current flowing through a single pad. These two measures are essential to decrease the IR-drop in a chip. In addition, the power rails of these components are connected to large decoupling capacitors to maintain the voltage during a high current flow. The common drawback of the aforementioned techniques is the extra chip area requirement. Since pixel density was a key feature, these techniques were not applicable inside the 512×512 pixel array; thus exposing the pixels to high IR-drop risk.

The methods to minimize IR-drop under stringent area requirements can be classified in two main categories: to decrease parasitic resistance or to increase parasitic capacitance. Given the same instantaneous current, the wire experiences less voltage drop with decreasing resistance. Therefore, the supply and ground wires inside the pixel array must be as wide as possible. The second method takes advantage of the parasitic capacitances formed between the wire and the ground. Those capacitors store charges during the steady state, and generate currents by transferring these charges to the pixels. Since each pixel draws current from the closest parasitic capacitors, those charges flow through a very low resistive path compared to the actual power sources; thus reducing the IR-drop. A capacitor, charged to a given voltage, can provide current for a longer time if its capacitance is bigger. As soon as its charge storage is depleted,

the pixel continues to transfer charges from the actual sources through a high resistive path.

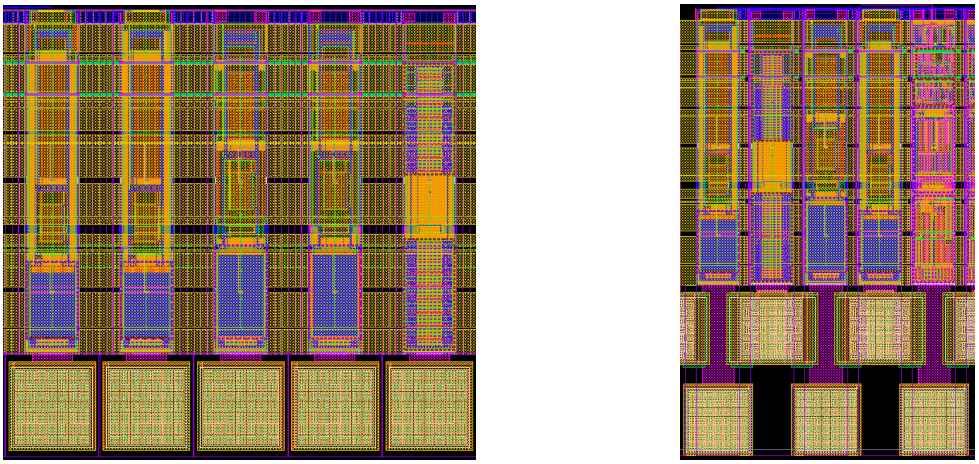
Another commonly used method to compensate IR-drop is called a decoupling capacitor. Placed between the wire and ground, a decoupling capacitor works on the same principle as described in the previous paragraph. In fact, metal parasitic capacitances mostly operate as decoupling capacitors, though with significantly small sizes. Due to the lack of available space, no decoupling capacitors were placed inside the pixels. Decoupling capacitors were only placed around the pixel array, to avoid IR-drops in the supply sources.

The IR-drop of the 3.3 V supply wire at the central pixel of a row is displayed in Figure 4.12. This waveform was generated in the post-layout simulation of an entire pixel row. The steep voltage drop occurs while turning off the SPAD via *Spadoff* signal. The maximum instantaneous voltage drop on the supply bus was recorded as 56 %, which was observed right after *Spadoff* process was completed. This seemingly threatening voltage drop has no significant impact on the pixel performance; because the supply voltage is restored to 90 % of its initial value in a very short time, approximately 1.7 ns. Based on the pixel operation mode described in chapter 2, the next use of the 3.3 V supply in the pixel occurs during the memory charging. Gating window, which starts tens of nanoseconds later than the *Spadoff* rising edge is the only time interval where memory charging is permitted. Therefore, based on the data in Figure 4.12, the supply bus always provides 3.3 V to the pixel when needed, even in the most exhaustive operation modes.

4.2.6 Pad Ring Design

In a multichannel image sensor that contains more than 250,000 pixels, the pad ring design is crucial from various perspectives. Decreasing perimeter vs. area ratio with increasing number of pixels requires the placement of pad ring in a very compact area. Power consumption and heating constraints limit the highest number of pads that can be read simultaneously. In addition, for the modularity of the design process, the pad pitch must be a specific value, which is not necessarily the smallest pitch that the technology allows.

In the layout of the sensor, two 512×256 modules were mirrored with respect to x-axis. Therefore, in a 512×256 chip, one of the long sides were left blank, and the pads were placed along one long side and two short sides. To ensure that the I/O blocks are biased correctly, a series of bias pads had to be scattered between signal pads. In this technology, a bias set consisted of 6 pads: core V_{dd}, core gnd!, pad V_{dd}, pad gnd!, substrate voltage and core 3.3 V. Another feature that allows a compact pad placement is the availability of staggered pads. When the pads are placed in 2 rows, as shown in Figure 4.13b, the factor that determines the minimum pitch is the I/O block width, a significantly lower value than pad width. In this technology, a staggered pad ring can have minimum pad pitch of 40.32 μm. However, to achieve modularity, the total width of 20 signal pads and 6 bias pads is equalized to the total width of 80 pixels. In this case, the pad pitch was set to 50.4 μm. Contrary to the long side, 2 short sides were filled with regular pad distribution, with a pad pitch of 70.4 μm Figure 4.13a.



(a) (b)

Figure 4.13: The layout view of (a) regular pad distribution with 70.4 μm pad pitch and (b) staggered pad distribution with 50.4 μm pad pitch

The frequency of biasing pads had to be chosen in such a way that prevents excessive steady current flow from a single pad. To ensure that the output signals are sent off the chip in the right form, the steady current through each pad had to be less than 50 mA. The average current flow through a bias pad can be calculated as follows:

The most dominant current consumption in the chip occurs while deactivation and activation of the SPAD using *Spadoff* and *Recharge* signals, respectively. This event can be modeled by charging an equivalent capacitance of 70 fF to 2.25 V per pixel. Based on Equation 4.2

$$\Delta Q = C \times \Delta V, \tag{4.2}$$

the total charge drawn per pixel in every 50 ns can be computed as 1.58×10^{-13} C. To calculate the average current per pixel, Equation 4.3 can be used.

$$\Delta Q = I \times \Delta t \tag{4.3}$$

The resulting average current consumption per pixel is equal to 3.15 μA, and the total average current consumption for 512×512 pixel array is equal to 826 mA. Assuming a totally uniform distribution of the 30 biasing pads throughout the pad ring, the average static current through a 3.3 V bias pad is $826mA/30 = 27.5mA$, significantly below the 50 mA limit.

4.2.7 Performance Characterization

The performance of this chip is characterized by gate uniformity. There are two parameters that define gate uniformity. Gate jitter is the time variation of the gate risetime/falltime between multiple iterations. Gate skew is the maximum time variation of the gate risetime/falltime in a single iteration between pixels across the array. The target specifications for maximum gate skew and gate jitter were both 150 ps.

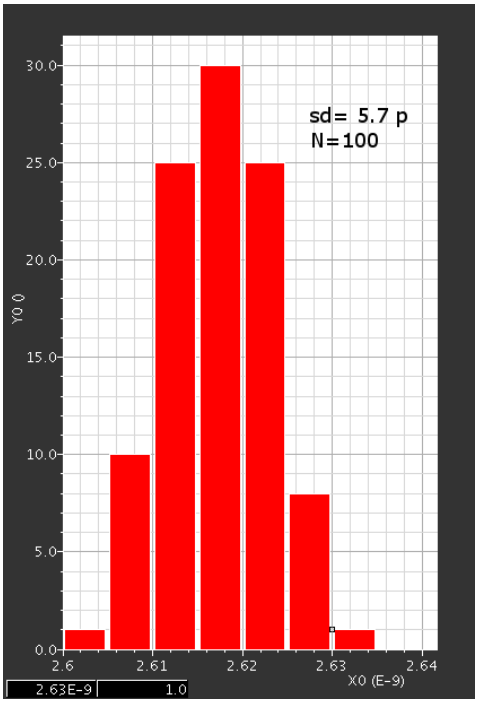


Figure 4.14: The jitter of the gate signal

Since the chip is yet to be fabricated, the characterization data in this thesis is based on post-layout simulation results.

The signal with the most critical jitter performance is *Gate*. Therefore, in this thesis the gate jitter measurement will be described as an example. However, it should be noted that the jitter of all input signals were simulated using similar methodologies. Gate jitter was computed by a Monte Carlo simulation. The target data was the time variation of signal edges in a single die. Therefore, process variation was disabled and only mismatch variation was enabled in the simulation settings. Figure 4.14 displays the signal edge distribution of the signal tree block with 100 iterations, in the typical process corner. The resulting shape is a clear Gaussian distribution with a standard deviation, represented by the letter σ, of 5.7 ps. In the literature, the conventional definition of signal jitter is equal to 2σ; as a result the gate jitter caused by the elements in the signal tree is equal to 11.4 ps. The contribution of other components to the jitter is negligible compared to the signal tree. For the computation of jitter, the importance of signal rise/fall times have to be considered. Signal jitter increases if the signal rises/falls slowly. Therefore, the output drivability of every block in the signal route was essential for the chip performance, thus had to be computed and simulated extensively during the design process. The column signal distribution tree architecture and implementation details were analyzed in the previous section.

Another important property of a signal is skew. In this chip, signal skew has two components: horizontal and vertical. Horizontal skew occurs only due to imperfections in the column signal tree. If different branches of the tree receive the input signal at different times, the gate window in each pixel of a row starts at different times. Since

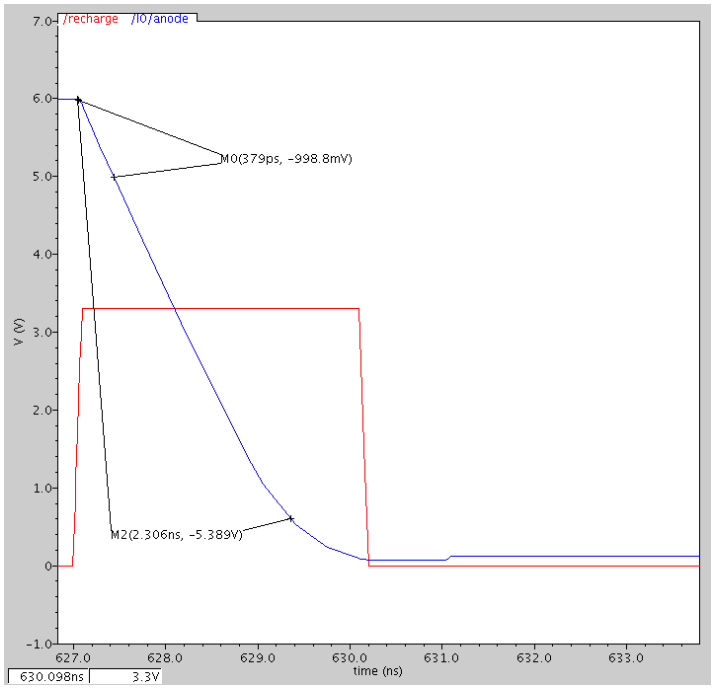


Figure 4.15: The response of the SPAD to the *Recharge* signal

the layout of the tree is designed to achieve uniformity, horizontal skew is expected to be negligible. On the other hand, vertical skew is generated by the parasitic capacitances and resistances of the vertical signal wires in a column. All global signals are sent to the pixel array from the bottom, and they travel in the vertical wires to reach the top pixels. Since the top pixels are the last ones in a column to receive the signals, vertical skew is defined by the time difference between the top and bottom pixels.

The horizontal and vertical gate skew in this chip were presented in Figure 4.10. This graph was generated from a post-layout simulation of the combination of a signal tree and a 256-pixel column. The resulting vertical skew is equal to 109.9 ps and horizontal skew is equal to 12.5 ps. These values, generated in the typical process corner, are below the target skew of 150 ps.

There are two defining points for the gating window of the pixel: rising edge of *Recharge* and falling edge of *Gate* for the beginning and the end, respectively. *Gate* skew and jitter, which are presented in previous paragraphs, are the only factors that determine the gating window end uniformity. The beginning uniformity, however, depends on the SPADs response to *Recharge*, instead of *Recharge* itself. The gating window effectively starts whenever the SPAD becomes sensitive to photons due to the voltage decrease in its anode terminal. Two factors complicate the detection of this moment. Firstly, there is no existing standard for the minimum PDP level that is considered "sensitive". In the preparation of Figure 4.3, 1 V excess bias was defined as the threshold. For the sake of consistency, in this analysis the same threshold was adopted. Secondly, low accuracy of the existing SPAD model shown in Figure 2.11 poses a risk of producing unreliable simulation results of the SPAD response. In contrast to the low excess bias version of the pixel where the recharge transistor is directly connected

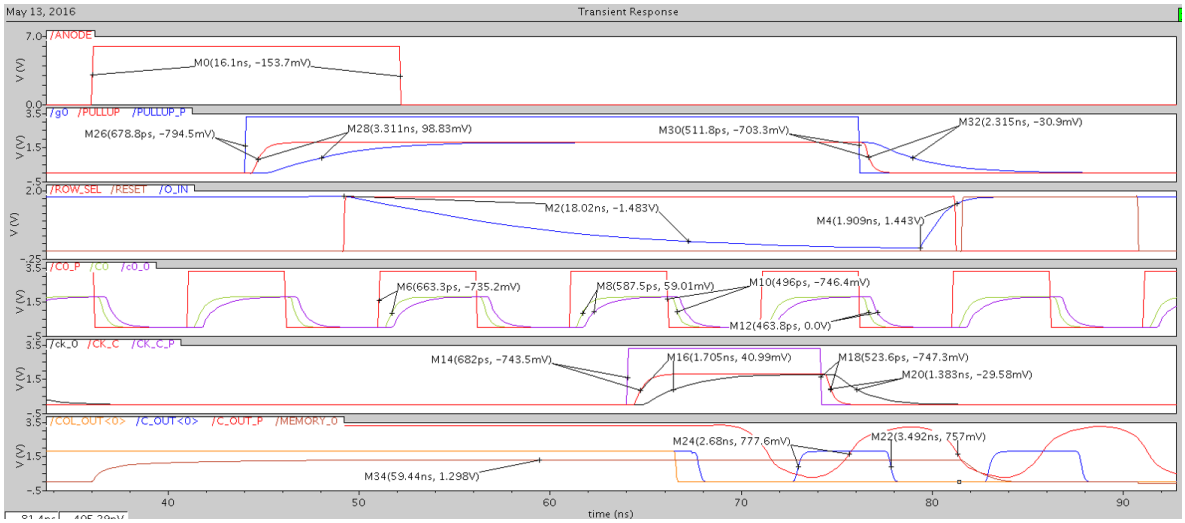


Figure 4.16: Post-layout simulation waveforms of the low-voltage signals

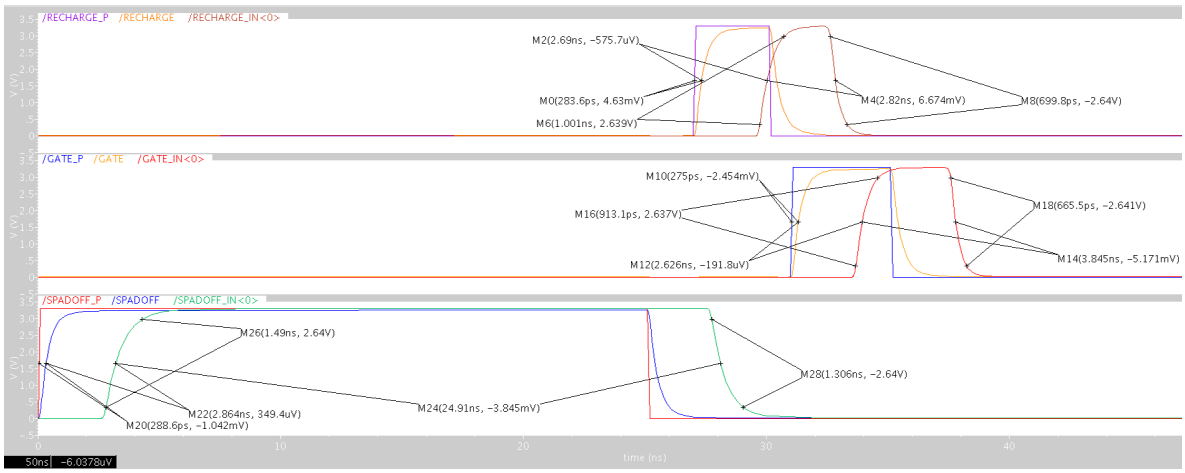


Figure 4.17: Post-layout simulation waveforms of the high-voltage signals

to the anode terminal, slower SPAD response to *Recharge* can be expected in the high voltage version due to the increased resistance introduced by the cascode transistor. SPAD anode voltage as a response to *Recharge* is provided in Figure 4.15. After an avalanche, the recharge signal restores the SPAD back to the sensitive state in 379 ps. However, at this point PDP is only 17 %. The maximum sensitivity of 40 % can only be achieved at 2.3 ns after the recharge risetime. Based on these two parameters, the uncertainty of gating window end can be approximated as 1.92 ns. Compared to the beginning of the window, there is a more than tenfold less precision in the end of the window. In the sensor operation mode where ToA is captured by gate shifting, only the beginning of the window determines the performance; hence the recharge speed can be ignored.

In various sections of this chapter, the importance of signal drivability in large arrays was emphasized, and several techniques to transfer high-frequency signals to and

from the chip. The effectiveness of these techniques are tested by tracking every signal through their paths in post-layout simulations. The results, shown in Figure 4.16 and Figure 4.17 demonstrate that all signals are successfully transferred to their destinations. Figure 4.16 also shows the highest readout speed that is achievable by the sensor. The pixel output, represented by the node *O_IN*, can be pulled down in 18.02 ns, and subsequently pulled up in 1.9 ns. This proves that the entire readout cycle of a row is below 20 ns. To capture an entire frame, 256 rows in a column should be read out in a sequence; which results in an overall frame rate of approximately 200 kfps. It should be noted that the farthest pixel from the signal source in a column was chosen for this readout simulation; hence the resulting frame rate is a pessimistic number.

4.2.8 Power Consumption

In large chips, power consumption is among the major concerns that limit functionality. In this subsection, a thorough analysis of this sensor's power consumption is presented.

In Table 4.3, the main components of total simulated power consumption are listed. Some power consumption values vary with sensor operation parameters or settings.

Table 4.3: Simulated power consumption list of the blocks in the 512×512 time-gated SPAD image sensor

Circuit block	Power consumption
Spadoff/recharge	2.65 W
Signal trees	432 mW
I/O blocks	420 mW
Readout	15.8 mW
Dynamic memory	4.5 mW

The highest power consuming mechanism in the circuit occurs while charging and discharging the SPAD anode using *Spadoff/Recharge*. In the typical operation settings with the minimum gating window length of 4 ns, global signals are repeated every 50 ns. In other words, in each pixel a SPAD with an equivalent capacitance of 70 fF is being charged and discharged to 2.5 V. According to the results reported in Table 4.3, for an array of 512×512 pixels the total power consumption equals 2.65 W.

The user has an option to disable *Spadoff*, which eliminates this power consumption component. However, this choice may make the sensor more vulnerable to noise. Particularly when a SPAD fires shortly before the gating window and is recharged actively in several nanoseconds, the afterpulsing probability is very high. Therefore, this is a trade-off between noise performance and power consumption.

The second largest contributors to power are the signal trees. In the sensor, there are 6 signal trees (2 for each global signal) whose total power consumption is approximately 432 mW, so the average consumption of a single tree equals 72 mW. This estimate was based on the standard global signal repetition period of 50 ns. The major capacitances that contribute to the power consumption are the readout transistors of each pixel and the parasitic capacitances of signal lines.

The power consumption of the I/O blocks is the most unpredictable element in the table, because the parameters used in the computations are empirical. The two most important parameters in this calculation are the average photon flux and the effective load capacitance of an I/O block. The load capacitance of each I/O is estimated to be between 1-10 pF. The most pessimistic computation is done using 10 pF load capacitance. In addition, the multiplexer output was assumed to be switching for every input switch. In other words, any adjacent pixels are assumed to be of the opposite color. Considering the low probability of this scenario, the I/O power consumption of 1.66 W is unrealistic. Instead, when the calculation is repeated with an average photon flux of 50 kcps/pixel, which is an empirical figure, the power consumption becomes 0.42 W for 10 pF load capacitance.

The contributions of readout and dynamic memory are insignificant compared to the first three elements. During row readout, the total equivalent capacitance of the column bus is discharged from 1.8 V to 0 V, and later pulled up to 1.8 V again. The total capacitance is composed of the total parasitic capacitance of the wire and the junction capacitances of 255 readout transistors in the off state. The dynamic memory causes the smallest power consumption even though its capacitance is almost equal to the SPAD capacitance (around 75 fF). The reason is that compared to the SPAD which is charged every 50 ns, the dynamic memory is charged every 6.4 μs.

5 Conclusion

5.1 Summary

This thesis aimed to explore the capability of SPAD-based pixels to form very large arrays and still meet the demanding requirements of various time-resolved imaging applications. To this end, several imaging sensor chips were designed in deep sub-micron CMOS technologies to demonstrate the peak performance of SPAD pixels under various conditions. In each stage of this work, a different property of the SPAD sensor was under focus.

The first stage of the thesis was to design a pixel that allows the SPAD cell to reach its top performance. The pixel structure was derived from the architecture described in [50]. Two major modifications performed on the existing version were to replace the static memory with a smaller sized dynamic memory and to replace some of the transistors with thick-oxide models to improve the maximum excess bias. The first version of the new pixel was placed into a 4×4 pixel array in 110 nm technology. This array was intended to comprise only the essential CMOS electronics to enable basic pixel operation.

The second stage was to demonstrate the scalability of the pixel. This time, the same pixel was redesigned for a 512×512 image sensor in 0.18 μm technology. In this chip, the SPAD choice was a p-i-n diode based SPAD whose optimal excess bias was approximately 10 V [2]. To increase the maximum excess bias from 3.3 V to 6.6 V, a cascode transistor was added between the SPAD anode and the pixel. Since this chip was intended to operate as a functional sensor, the complexity of the electronics inside it was higher than the 4×4 chip. This allowed a more comprehensive characterization based on more performance parameters. The analyses of the following parameters were conducted: IR-drop, gating signal skew and jitter, maximum bias pad current, maximum gate window period allowed by the dynamic memory and SPAD response to active recharge. In addition, this chip is suitable for non-uniformity analyses that are essential for the commercial adoption of SPAD image sensors. Based on post-layout simulation results, the 512×512 chip achieved 13 % fill factor, 16.38 μm pixel pitch, 3.52 W maximum power consumption, 20 ps timing resolution and a 200 kfps image frame rate.

Subsequently, a large event-driven pixel that supports 12 V excess bias was designed. This chip was also designed in 0.18 μm technology with p-i-n diode based SPADs. The pixel in this chip had a fundamentally different architecture from the previous pixels: it offered significantly less controllability, did not have an in-pixel ToA detection mechanism, and occupied more space due to the placement of a very large polysilicon quenching resistor. The major advantage of this chip was its fully parallel circuit structure which allows independent operation of each pixel and achieves a faster readout

than the previous chips.

During the design procedure of each chip, multiple layout versions of the same architecture were tried. Usually, the first version employed a more conventional, low-risk design method. The second version, on the other hand, exploited novel design techniques that aim to boost various performance parameters. The varying elements of 2 chip versions include SPAD structure, SPAD active area shape and component sharing level between pixels.

5.2 Future Work

The work that was presented in this thesis can be extended in the future in the following directions:

1. The 512×512 sensor can be characterized based on measurement results. Due to timing constraints, the characterization of this chip was presented in this thesis based on post-layout simulation data.

2. More effective solutions to actively turn off the SPAD can be investigated. The current spadoff mechanism only reduces SPAD sensitivity, which does not eliminate the noise due to afterpulsing.

3. To enhance single-photon sensitivity, two independently controlled gates can be placed inside a pixel. In the current configuration, the photon arrival events that fall outside the gating window are missed by the sensor. That limits the low-light-level performance of a sensor.

4. The option of designing the sensor in a 3D stacking technology can be investigated. This technology can boost the fill factor significantly without increasing pixel pitch or reducing pixel functionality.

5. More compact and controllable solutions to reach SPAD excess bias above 10 V with a standard CMOS pixel can be discovered. This can have two advantages: Firstly, it can permit the SPAD to be operated in the optimal configuration. Secondly, pixel miniaturization level can be proportional to the technology node by eliminating the thick-oxide high voltage transistors.

Layout View of Circuit Blocks

A

A.1 512×512 Time-Gated SPAD-Based Image Sensor

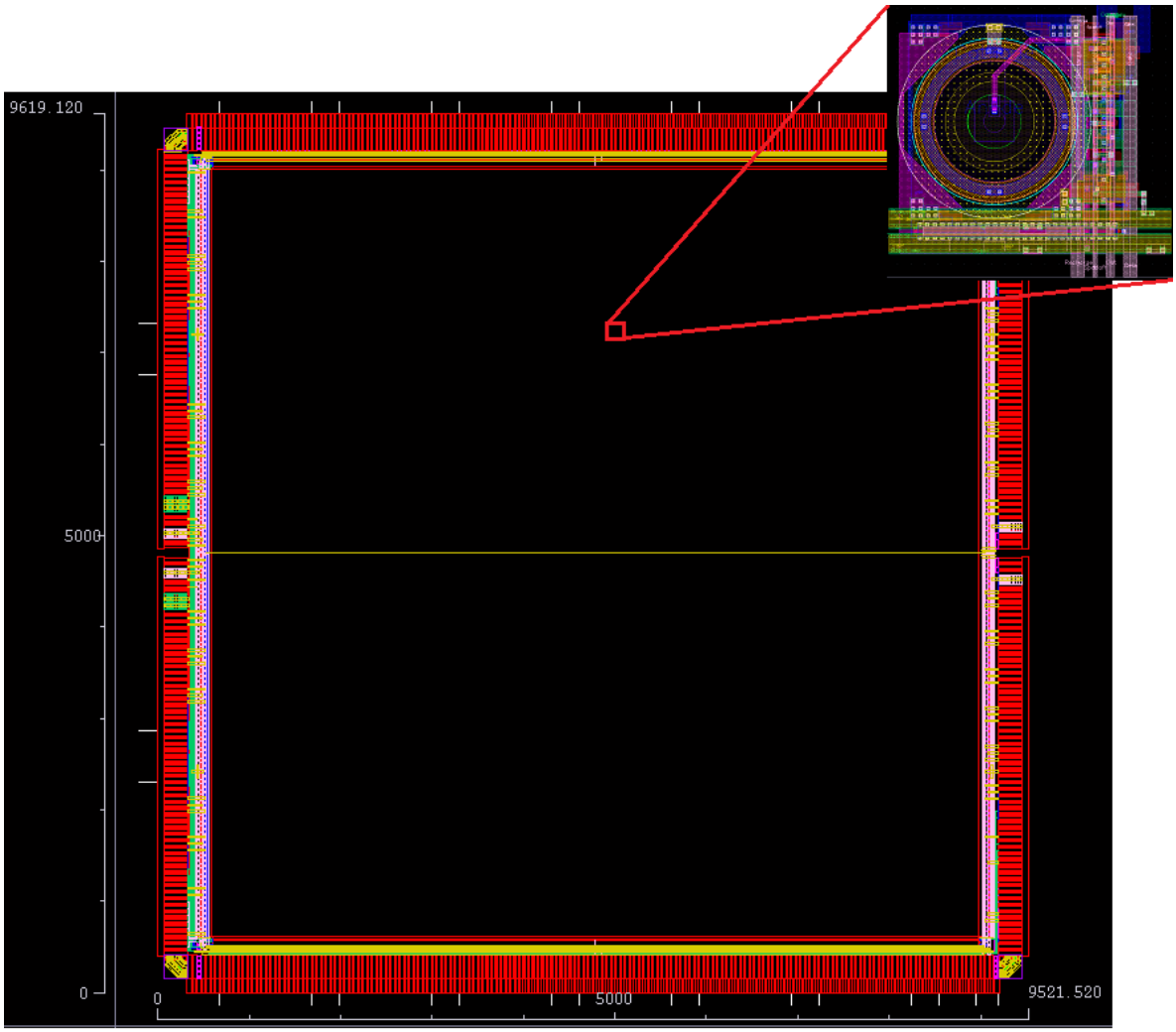


Figure A.1: Top view layout (512×512)

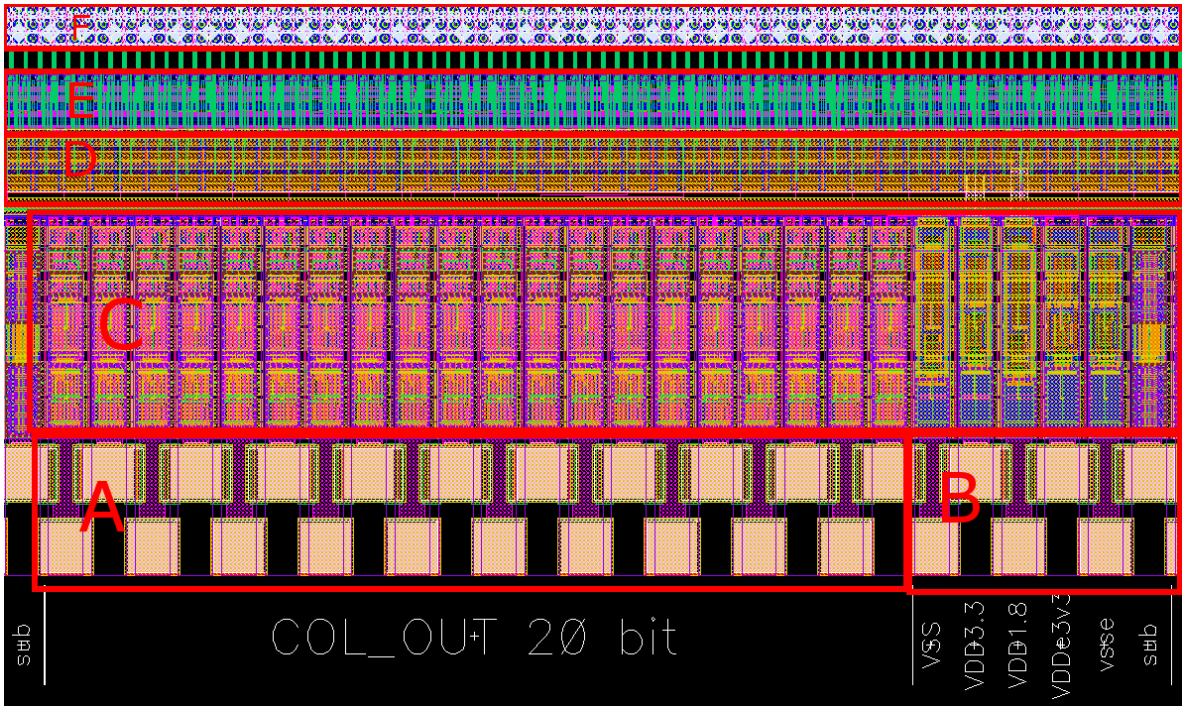


Figure A.2: Column control circuit layout (512×512). A: Column output pads, B: Bias pads, C: I/O cells, D: Decoupling capacitors for signal distribution network supply, E: 3 signal distribution network blocks for 3.3 V global signals, F: Pixel array

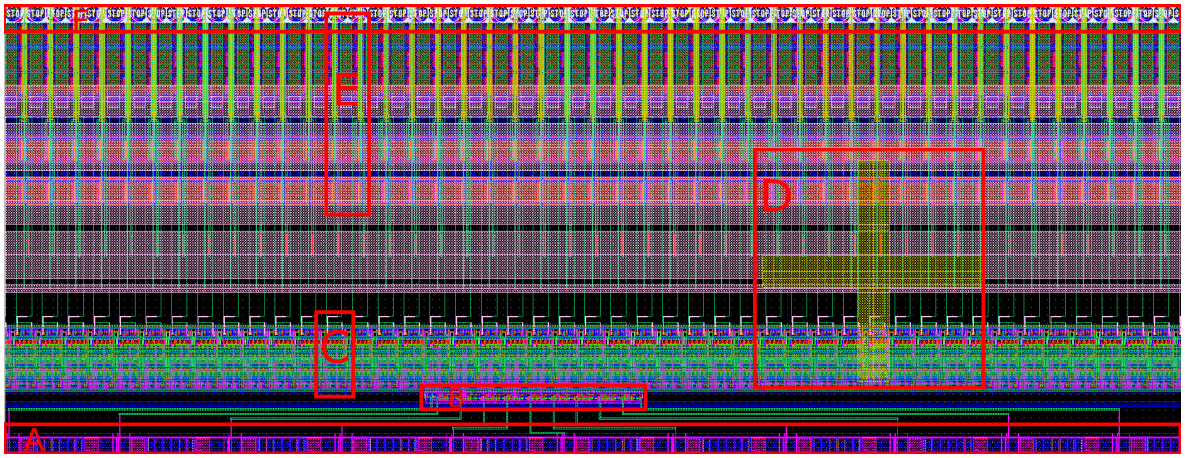


Figure A.3: Row control circuit layout (512×512). A: I/O and pad cells, B: Buffers for the incoming signals, C: Row driver block, D: Top metal marker for microlens placement, E: Decoupling capacitors for to prevent temporary voltage drops in DC supply wires, F: Pixel array

A.2 A 512×1 Event-Driven SPAD-Based Line Sensor

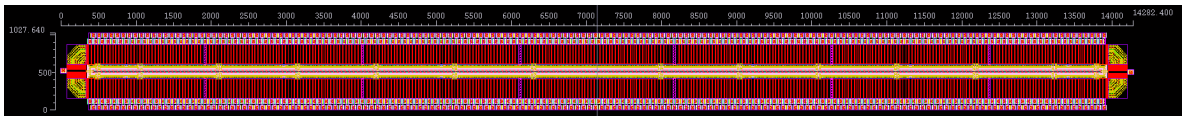


Figure A.4: Top view layout (512×1)

A.3 A 4×4 Time-Gated SPAD-Based Image Sensor in 110 nm CMOS Technology

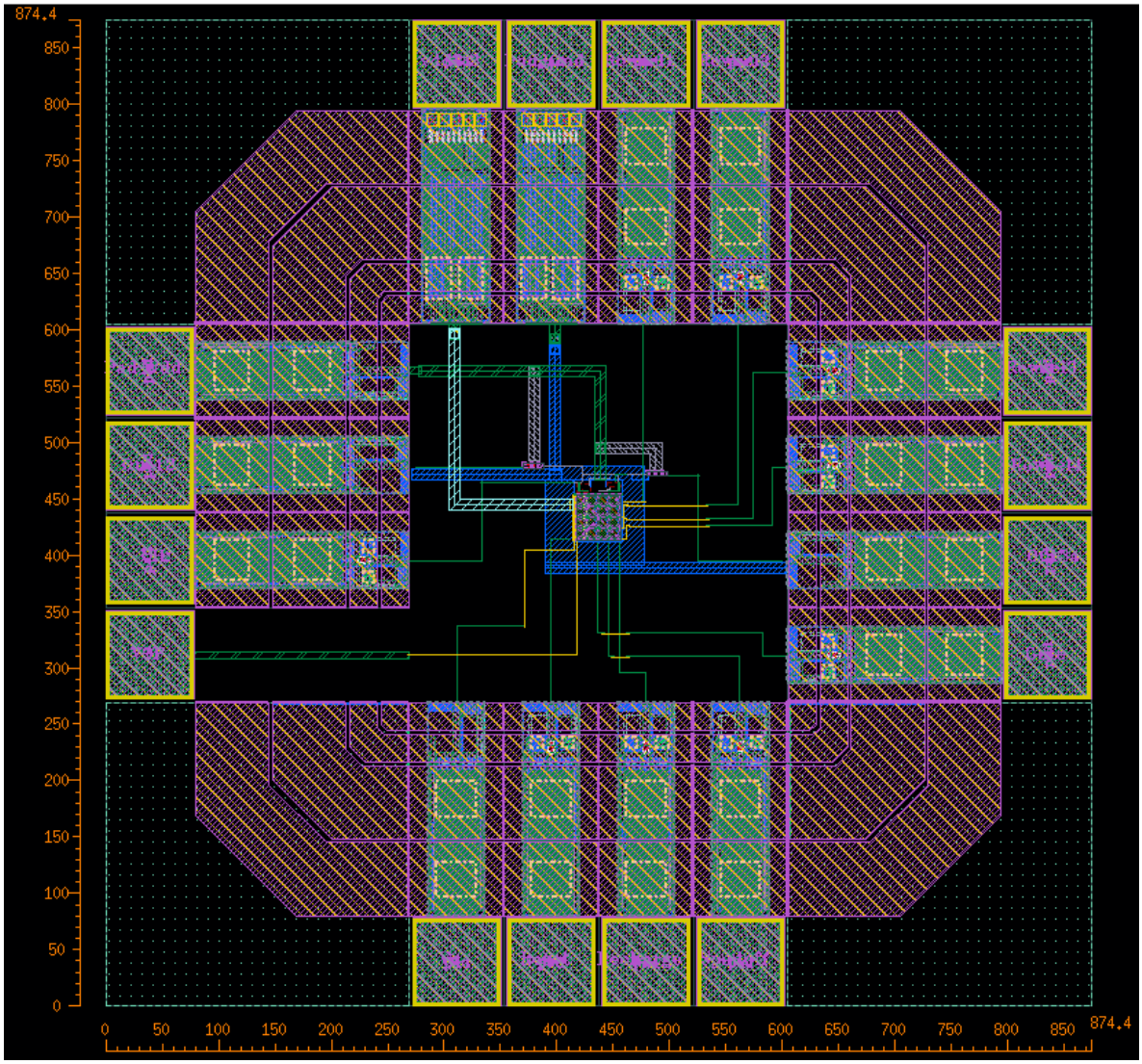


Figure A.5: Top view layout (4×4)

Test Structures in the 512×512 Sensor

B

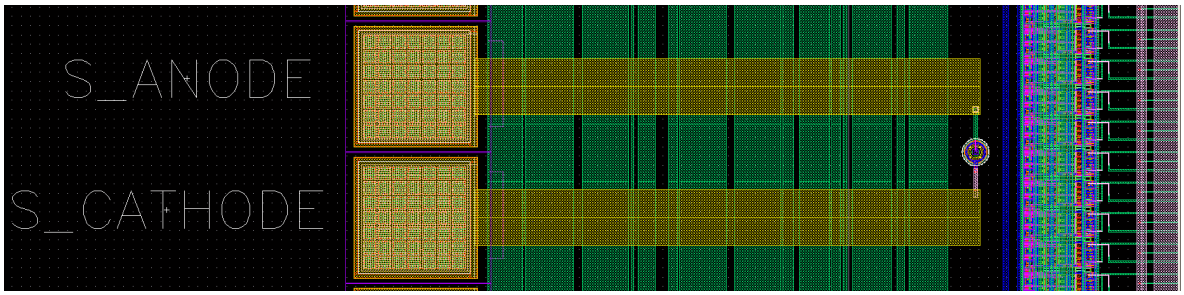


Figure B.1: Position of a test SPAD cell (512×512)

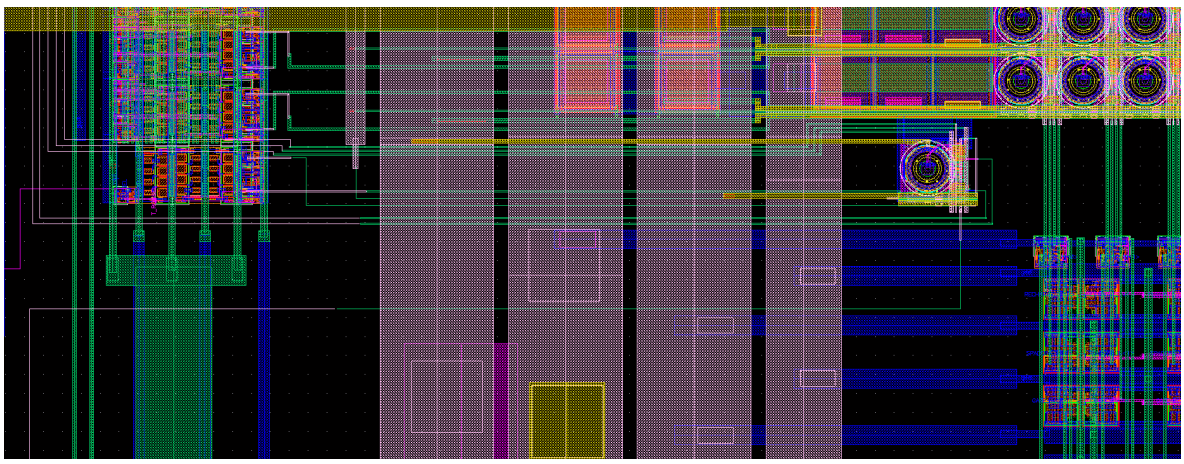


Figure B.2: Position of a test pixel (512×512)

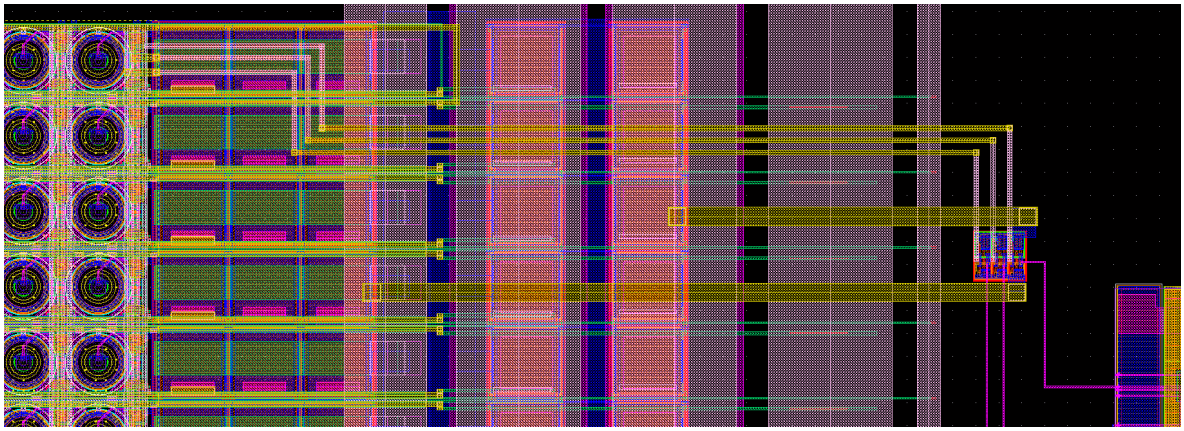


Figure B.3: Position of the testing configuration for column signals (512×512)

Bibliography

[1] S. Mandai. *Multichannel Digital Silicon Photomultipliers for Time-of-Flight PET*. PhD thesis, TU Delft, Fac. EEMCS, July 2014. ISBN 9789462592346.

[2] C. Veerappan and E. Charbon. A low dark count p-i-n diode based spad in cmos technology. *IEEE Transactions on Electron Devices*, 63(1):65–71, Jan 2016.

[3] Wikipedia. Static random access memory — Wikipedia, the free encyclopedia, 2016. [Online; accessed 22-August-2016].

[4] Samuel Burri, Yuki Maruyama, Xavier Michalet, Francesco Regazzoni, Claudio Bruschini, and Edoardo Charbon. Architecture and applications of a high resolution gated spad image sensor. *Opt. Express*, 22(14):17573–17589, Jul 2014.

[5] N. A. W. Dutton, L. Parmesan, A. J. Holmes, L. A. Grant, and R. K. Henderson. 320 x 240 oversampled digital single photon counting image sensor. In *2014 Symposium on VLSI Circuits Digest of Technical Papers*, pages 1–2, June 2014.

[6] C. Veerappan. *Single Photon Avalanche Diodes for Cancer Diagnosis*. PhD thesis, TU Delft, Fac. EEMCS, March 2016.

[7] M. Gersbach, R. Trimananda, Y. Maruyama, M. Fishburn, D. Stoppa, J. Richardson, R. Walker, R.K. Henderson, and E. Charbon. High frame-rate TCSPC-FLIM readout system using a SPAD-based image sensor. In *Proc. SPIE Optics+Photonics Single-Photon Imaging*, San Diego (CA), August 2010.

[8] Edinburgh Instruments. What is TCSPC: Time-Correlated Single-Photon Counting. Technical report, Edinburgh Instruments, July 2013.

[9] Anand Pratap Singh, Jan Wolfgang Krieger, Jan Buchholz, Edoardo Charbon, Jörg Langowski, and Thorsten Wohland. The performance of 2d array detectors for light sheet based fluorescence correlation spectroscopy. *Opt. Express*, 21(7):8652–8668, Apr 2013.

[10] S. Bellisai, F. Villa, S. Tisa, D. Bronzi, and F. Zappa. Indirect time-of-flight 3d ranging based on spads. *Proc. SPIE*, 8268:82681C–82681C–8, 2012.

[11] Edoardo Charbon, Matt Fishburn, Richard Walker, Robert K. Henderson, and Cristiano Niclass. *SPAD-Based Sensors*, pages 11–38. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[12] Miles N. Wernick and John N. Aarsvold. *Emission Tomography: The Fundamentals of PET and SPECT*. Academic Press, New York, NY, USA, 2004.

[13] Sibylle I. Ziegler. Proceedings of the 22nd international nuclear physics conference (part 2) positron emission tomography: Principles, technology, and recent developments. *Nuclear Physics A*, 752:679 – 687, 2005.

[14] L. H. C. Braga, L. Gasparini, L. Grant, R. K. Henderson, N. Massari, M. Perenzoni, D. Stoppa, and R. Walker. A fully digital 8 x 16 sipm array for pet applications with per-pixel tdcs and real-time energy output. *IEEE Journal of Solid-State Circuits*, 49(1):301–314, Jan 2014.

[15] Jordana Blacksberg, Yuki Maruyama, Edoardo Charbon, and George R. Rossman. Fast single-photon avalanche diode arrays for laser raman spectroscopy. *Opt. Lett.*, 36(18):3672–3674, Sep 2011.

[16] Y. Maruyama, J. Blacksberg, G.R. Rossman, and E. Charbon. A time-resolved 128x128 SPAD camera for laser Raman spectroscopy. In *Proc. SPIE DSS Single-Photon Imaging*, April 2012.

[17] D. Stucki, S. Burri, E. Charbon, C. Chunnilall andA. Meneghetti, and F. Regazzoni. Towards a high-speed quantum random number generator. In *Proc. SPIE Conference on Defense and Security*, September 2013.

[18] S. Burri, D. Stucki, Y. Maruyama, C. Bruschini, E. Charbon, and F. Regazzoni. Spads for quantum random number generators and beyond. In *2014 19th Asia and South Pacific Design Automation Conference (ASP-DAC)*, pages 788–794, Jan 2014.

[19] V. K. Zworykin, G. A. Morton, and L. Malter. The secondary emission multiplier-a new electronic device. *Proceedings of the Institute of Radio Engineers*, 24(3):351–375, March 1936.

[20] R. Mirzoyan, M. Laatiaoui, and M. Teshima. Very high quantum efficiency {PMTs} with bialkali photo-cathode. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 567(1):230 – 232, 2006. Proceedings of the 4th International Conference on New Developments in PhotodetectionBEAUNE 2005Fourth International Conference on New Developments in Photodetection.

[21] Albert J P Theuwissen. *Solid-State Imaging with Charge-Coupled Devices* . Kluwer Academic Publishers, Dordrecht, 1995.

[22] J. Hynecek. Impactron-a new solid state image intensifier. *IEEE Transactions on Electron Devices*, 48(10):2238–2241, Oct 2001.

[23] Donal J. Denvir and Emer Conroy. Electron-multiplying ccd: the new iccd. *Proc. SPIE*, 4796:164–174, 2003.

[24] Colin Coates, Boyd Fowler, and Gerhard Holst. sCMOS: Scientific CMOS Technology. Technical report, Andor Technology, Fairchild Imaging and PCO AG, June 2009.

[25] S. M. Sze and K. K. Ng. *Physics of Semiconductor Devices*. John Wiley & Sons, 2006.

[26] M.W. Fishburn. *Fundamentals of CMOS single-photon avalanche diodes*. PhD thesis, TU Delft, Fac. EEMCS, September 2012. ISBN 978-94-91030-29-1.

[27] G. N. Goltsman, O. Okunev, G. Chulkova, A. Lipatov, A. Semenov, K. Smirnov, B. Voronov, A. Dzardanov, C. Williams, and Roman Sobolewski. Picosecond superconducting single-photon optical detector. *Applied Physics Letters*, 79(6):705–707, 2001.

[28] T. Yamashita, S. Miki, H. Terai, K. Makise, and Z. Wang. Parallel bias and readout techniques toward realization of large-scale sspd array with sfq circuit. *IEEE Transactions on Applied Superconductivity*, 23(3):2500804–2500804, June 2013.

[29] Myung-Jae Lee, Pengfei Sun, and Edoardo Charbon. A first single-photon avalanche diode fabricated in standard soi cmos technology with a full characterization of the device. *Opt. Express*, 23(10):13200–13209, May 2015.

[30] A. Rochas, M. Gani, B. Furrer, P. A. Besse, R. S. Popovic, G. Ribordy, and N. Gisin. Single photon detector fabricated in a complementary metaloxidesemiconductor high-voltage technology. *Review of Scientific Instruments*, 74(7):3263–3270, 2003.

[31] E. Charbon, H. J. Yoon, and Y. Maruyama. A geiger mode apd fabricated in standard 65nm cmos technology. In *2013 IEEE International Electron Devices Meeting*, pages 27.5.1–27.5.4, Dec 2013.

[32] Peter Seitz and Albert J P Theuwissen. *Single-Photon Imaging*. Springer, Heidelberg, 2011.

[33] Roland H. Haitz. Mechanisms contributing to the noise pulse rate of avalanche diodes. *Journal of Applied Physics*, 36(10):3123–3131, 1965.

[34] Donald Neamen. *Semiconductor Physics And Devices*. McGraw-Hill, Inc., New York, NY, USA, 3 edition, 2003.

[35] Balázs Játékos, Ferenc Ujhelyi, Emőke Lőrincz, and Gábor Erdei. Investigation of photon detection probability dependence of spadnet-i digital photon counter as a function of angle of incidence, wavelength and polarization. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 769:59 – 64, 2015.

[36] S. Cova, M. Ghioni, A. Lacaita, C. Samori, and F. Zappa. Avalanche photodiodes and quenching circuits for single-photon detection. *Appl. Opt.*, 35(12):1956–1976, Apr 1996.

[37] M. Assanelli, A. Ingargiola, I. Rech, A. Gulinatti, and M. Ghioni. Photon-timing jitter dependence on injection position in single-photon avalanche diodes. *IEEE Journal of Quantum Electronics*, 47(2):151–159, Feb 2011.

[38] A. Lacaita and M. Mastrapasqua. Strong dependence of time resolution on detector diameter in single photon avalanche diodes. *Electronics Letters*, 26:2053–2054(1), November 1990.

[39] I. Rech, A. Ingargiola, R. Spinelli, I. Labanca, S. Marangoni, M. Ghioni, and S. Cova. A new approach to optical crosstalk modeling in single-photon avalanche diodes. *IEEE Photonics Technology Letters*, 20(5):330–332, March 2008.

[40] I. M. Antolovic, S. Burri, C. Bruschini, R. Hoebe, and E. Charbon. Nonuniformity analysis of a 65-kpixel cmos spad imager. *IEEE Transactions on Electron Devices*, 63(1):57–64, Jan 2016.

[41] N. A. W. Dutton, I. Gyongy, L. Parmesan, S. Gnecchi, N. Calder, B. R. Rae, S. Pellegrini, L. A. Grant, and R. K. Henderson. A spad-based qvga image sensor for single-photon counting and quanta imaging. *IEEE Transactions on Electron Devices*, 63(1):189–196, Jan 2016.

[42] V. Savuskan, I. Brouk, M. Javitt, and Y. Nemirovsky. An estimation of single photon avalanche diode (spad) photon detection efficiency (pde) nonuniformity. *IEEE Sensors Journal*, 13(5):1637–1640, May 2013.

[43] Silvano Donati, Giuseppe Martini, and Michele Norgia. Microconcentrators to recover fill-factor in image photodetectors with pixel on-board processing circuits. *Opt. Express*, 15(26):18066–18075, Dec 2007.

[44] A. Vilá, E. Vilella, O. Alonso, and A. Dieguez. Crosstalk-free single photon avalanche photodiodes located in a shared well. *IEEE Electron Device Letters*, 35(1):99–101, Jan 2014.

[45] L. Pancheri, N. Massari, F. Borghetti, and D. Stoppa. A 32x32 spad pixel array with nanosecond gating and analog readout. In *International Image Sensor Workshop (IISW)*, Hokkaido, Japan, 2011.

[46] M. Perenzoni, N. Massari, D. Perenzoni, L. Gasparini, and D. Stoppa. 11.3 a 160 x 120-pixel analog-counting single-photon imager with sub-ns time-gating and self-referenced column-parallel a/d conversion for fluorescence lifetime imaging. In *2015 IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers*, pages 1–3, Feb 2015.

[47] Matteo Perenzoni, Lucio Pancheri, and David Stoppa. Compact spad-based pixel architectures for time-resolved image sensors. *Sensors*, 16(5):745, 2016.

[48] M. A. Karami, H. J. Yoon, and E. Charbon. Single-photon Avalanche Diodes in sub-100nm Standard CMOS Technologies. In *Proc. Intl. Image Sensor Workshop (IISW)*, 2011.

[49] Samuel Burri, Harald Homulle, Claudio Bruschini, and Edoardo Charbon. LinoSPAD : a time-resolved 256 x 1 CMOS SPAD line sensor system featuring 64 FPGA-based TDC channels running at up to 8 . 5 giga-events per second. *SPIE Optical Sensing and Detection*, 9899:1–10, 2016.

[50] Y. Maruyama and E. Charbon. An all-digital, time-gated 128x128 spad array for on-chip, filter-less fluorescence detection. In *2011 16th International Solid-State Sensors, Actuators and Microsystems Conference*, pages 1180–1183, June 2011.